PHASE-SPACE OPTICS

Fundamentals and Applications

Markus Testorf
Bryan Hennelly
Jorge Ojeda-Castañeda

Phase-Space Optics

About the Authors

Markus Testorf received his doctorate in physics from the University of Erlangen in Germany. He is currently an assistant professor at the Thayer School of Engineering at Dartmouth College. Dr. Testorf has written numerous articles on the use of phase-space optics. He has taught optics courses throughout his professional career, and phase-space concepts have also become part of his standard repertoire in the classroom.

Bryan Hennelly received his doctorate in optical physics from the University College Dublin in Ireland in 2005. He is currently a research fellow at the National University of Ireland, Maynooth. Dr. Hennelly has written or coauthored numerous articles based on Wigner or phase-space optics relating to optical metrology systems and the sampling and numerical simulation of optical systems.

Jorge Ojeda-Castañeda earned his doctorate in applied optics, at University of Reading, UK. As an Alexander von Humboldt fellow, he worked at the University of Erlangen, in Germany. Dr. Ojeda-Castañeda has been a visiting professor at the Institute of Optics in Madrid, the University of Valencia, the Autonomous University of Barcelona, and the University James I in Spain. He is currently professor of applied optics, in the University of Guanajuato, México. Dr. Ojeda-Castañeda has written more than 200 papers in academic journals and conference proceedings. In many of his oral and written contributions, he has pioneered the use of phase-space representations of optical systems. Dr. Ojeda-Castañeda is a fellow of the SPIE and the OSA.

Phase-Space Optics

Fundamentals and Applications

Markus E. Testorf
Bryan M. Hennelly
Jorge Ojeda-Castañeda

New York Chicago San Francisco
Lisbon London Madrid Mexico City
Milan New Delhi San Juan
Seoul Singapore Sydney Toronto

ISBN: 978-0-07-159799-9

MHID: 0-07-159799-9

The material in this eBook also appears in the print version of this title: ISBN: 978-0-07-159798-2, MHID: 0-07-159798-0.

Contents

Preface

It is no simple task to characterize the importance of phase-space methods in the optical sciences. Geometrical optics, formally related to classical mechanics, has benefited implicitly and explicitly from phase-space concepts since Hamiltonian and Lagrangian optics were first formulated. In comparison, phase-space optics of coherent wavefronts, namely, the use of the Wigner distribution functions and of the ambiguity function, constitutes a more recent development, and the Wigner distribution remains far from being integrated into the canon of standard tools used by the optics community.

Optical engineers and researchers are polarized on the use of phase-space optics. Many remain intrigued, but skeptical toward a mathematical formalism that appears theoretically demanding, without providing obvious complementary information for describing optical phenomena. On the other end of the spectrum one can find a small, but fast-growing community that is enchanted by the beauty and simplicity of phase-space optics, revealing itself even with only a scant familiarity with the theoretical framework.

To understand this devotion, it is important to recognize the unique position that optics holds in science and engineering. Optics is both a subject of basic research and an enabling technology. Fundamental questions about the quantum nature of light, and its interaction with matter, are at the core of modern physics. At the same time, there is a rich history of optical instruments pivotal to ground-breaking discoveries in astronomy, biology, communications, and many other disciplines. In the past half century, the optical sciences have developed at an astounding pace. Perhaps with the exception of microelectronics, optics has become the most vibrant technology resting at the intersection of different brands of research.

As a consequence, different optical sciences have developed unique and effective models to describe light propagation and the interaction of light with matter. Notwithstanding the universal validity of Maxwell's equations, or quantum electrodynamics, it is often more effective to describe light propagation based on specific models (rays, scalar waves, or Gaussian beams) than to consider the full complexity of the electrodynamic wave field. All models of light propagation

are well explored, as are their relationships with one another. However, it is, without question, desirable to have a simple, common platform with which to unify these models, while preserving their unique features.

From our viewpoint, such a platform can provide a powerful tool for understanding and applying the physics of light propagation through optical systems. Ideally, this common platform should be a superior model, with all other models reducing to different facets of a common framework. The contributions collected in this book emphasize the fact that this model may be best implemented by what we term *phase-space optics*.

Phase-space optics refers to a representation of optical signals in an artificial configuration space simultaneously providing information about spatial properties of the signal and its angular spectrum, or equivalently in terms of its temporal and time-frequency characteristics. In coherent optics, this concept has also been popularized as "Wigner optics" since the properties of the Wigner distribution function are often used to motivate the use of a joint signal representation. In the signal processing community, the Wigner-Ville distribution is recognized as a relevant member of a larger class of joint time-frequency transforms. Closely connected with the Wigner distribution function through a double Fourier transform, the ambiguity function is used by the radar community for representing signals simultaneously carrying information about the down range of the target and its velocity.

In contrast, the term *phase space*, while being based on the same conceptual and formal mathematical tools, rather emphasizes the dynamics of the physical system. Phase space, and in particular the Wigner distribution, can be recognized as one common platform for understanding and applying the physics of more traditional models for describing electromagnetic signals as they evolve and propagate through an optical system.

By compiling this book, it was our desire to create a mosaic of phase-space optics. Each contribution constitutes a self-consistent perspective on one particular aspect of optical signals in phase space, while revealing its full beauty and importance only as part of this entire collection. We owe it to the authors who contributed to this effort that the result has far exceeded our expectations.

Each of the chapters illustrates original ways to gain physical insight and to develop novel engineering applications. All chapters are written by authors who are pioneers in using phase-space optics in their fields of expertise. As a consequence, the topics are discussed with unique depth, without losing sight of the necessity to embed phase-space optics in a broader context.

We believe that the book will be helpful for researchers and graduate students alike, who wish to familiarize themselves with phase-space concepts in optics, but also want to move beyond a mere introductory level of discussion. We are sure that the number of applications derived from phase-space optics will grow, and we hope that this collection will help to facilitate this development.

This book would not have been possible without the guidance and encouragement of McGraw-Hill senior editor Taisuke Soda. We are indebted to him and the helpful staff at McGraw-Hill.

Markus E. Testorf
Dartmouth College, Hanover, New Hampshire, United States

Bryan M. Hennelly
National University of Ireland, Maynooth, Ireland

Jorge Ojeda-Castañeda
Universidad de Guanajuato, México

This page intentionally left blank

Phase-Space Optics

This page intentionally left blank

CHAPTER 1

Wigner Distribution in Optics

Martin J. Bastiaans

Technische Universiteit Eindhoven, Faculteit Elektrotechniek
Eindhoven, Netherlands

1.1 Introduction

In 1932 Wigner[1] introduced a distribution function in mechanics that permitted a description of mechanical phenomena in a phase space. Such a Wigner distribution was introduced in optics by Dolin[2] and Walther[3,4] in the 1960s, to relate partial coherence to radiometry. A few years later, the Wigner distribution was introduced in optics again[5–11] (especially in the area of Fourier optics), and since then, a great number of applications of the Wigner distribution have been reported.

While the mechanical phase space is connected to classical mechanics, where the movement of particles is studied, the phase space in optics is connected to geometrical optics, where the propagation of optical rays is considered. Whereas the position and momentum of a particle are the two important quantities in mechanics, in optics we are interested in the position and the direction of an optical ray. We will see that the Wigner distribution represents an optical field in terms of a ray picture, and that this representation is independent of whether the light is partially coherent or completely coherent.

We will observe that a description by means of a Wigner distribution is, in particular, useful when the optical signals and systems can be described by quadratic-phase functions, i.e., when we are in the realm of first-order optics: spherical waves, thin lenses, sections of free space in the paraxial approximation, etc. Although formulated

in Fourier-optical terms, the Wigner distribution will form a link to such diverse fields as geometrical optics, ray optics, matrix optics, and radiometry.

Sections 1.2 through 1.7 mainly deal with optical signals and systems. We treat the description of completely coherent and partially coherent light fields in Sec. 1.2. The Wigner distribution is introduced in Sec. 1.3 and elucidated with some optical examples. Properties of the Wigner distribution are considered in Sec. 1.4. In Sec. 1.5 we restrict ourselves to the one-dimensional case and observe the strong connection of the Wigner distribution to the fractional Fourier transformation and rotations in phase space. The propagation of the Wigner distribution through Luneburg's first-order optical systems is the topic of Sec. 1.6, while the propagation of its moments is discussed in Sec. 1.7. The final Sec. 1.8 is devoted to the broad class of bilinear signal representations known as the Cohen class, of which the Wigner distribution is an important representative.

1.2 Elementary Description of Optical Signals and Systems

We consider scalar optical signals, which can be described by, say, $\tilde{f}(x, y, z, t)$, where x, y, z denote space variables and t represents the time variable. Very often we consider signals in a plane $z =$ constant, in which case we can omit the longitudinal space variable z from the formulas. Furthermore, the transverse space variables x and y are combined into a two-dimensional column vector $\mathbf{r}$. The signals with which we are dealing are thus described by a function $\tilde{f}(\mathbf{r}, t)$.

Although real-world signals are real, we will not consider these signals as such. The signals $\tilde{f}(\mathbf{r}, t)$ that we consider in this chapter are *analytic signals*, and our real-world signals follow as the real part of these analytic signals.

Throughout we denote column vectors by boldface lowercase symbols, while matrices are denoted by boldface uppercase symbols; transposition of vectors and matrices is denoted by the superscript t. Hence, for instance, the two-dimensional column vectors $\mathbf{r}$ and $\mathbf{q}$ represent the space and spatial-frequency variables $[x, y]^t$ and $[u, v]^t$, respectively, and $\mathbf{q}^t\mathbf{r}$ represents the inner product $ux + vy$. Moreover, in integral expressions, $d\mathbf{r}$ and $d\mathbf{q}$ are shorthand notations for $dx\,dy$ and $du\,dv$, respectively.

1.2.1 Impulse Response and Coherent Point-Spread Function

The input-output relationship of a general linear system $\tilde{f}_i(\mathbf{r}, t) \rightarrow \tilde{f}_o(\mathbf{r}, t)$ reads

$$\tilde{f}_o(\mathbf{r}_o, t_o) = \int\int \tilde{h}(\mathbf{r}_o, \mathbf{r}_i, t_o, t_i)\, \tilde{f}_i(\mathbf{r}_i, t_i)\, d\mathbf{r}_i\, dt_i \tag{1.1}$$

where $\tilde{h}(\mathbf{r}_o, \mathbf{r}_i, t_o, t_i)$ is the impulse response, i.e., the system's response to a Dirac function:

$$\delta(\mathbf{r} - \mathbf{r}_i)\delta(t - t_i) \rightarrow \tilde{h}(\mathbf{r}, \mathbf{r}_i, t, t_i)$$

We restrict ourselves to a time-invariant system $\tilde{h}(\mathbf{r}_o, \mathbf{r}_i, t_o, t_i) =: \tilde{h}(\mathbf{r}_o, \mathbf{r}_i, t_o - t_i)$, in which case the input-output relationship takes the form of a convolution (as far as the time variable is concerned):

$$\tilde{f}_o(\mathbf{r}_o, t_o) = \int\int \tilde{h}(\mathbf{r}_o, \mathbf{r}_i, t_o - t_i)\, \tilde{f}_i(\mathbf{r}_i, t_i)\, d\mathbf{r}_i\, dt_i \tag{1.2}$$

The temporal Fourier transform of the impulse response $\tilde{h}(\mathbf{r}_o, \mathbf{r}_i, \tau)$

$$h(\mathbf{r}_o, \mathbf{r}_i, \nu) = \int \tilde{h}(\mathbf{r}_o, \mathbf{r}_i, \tau)\, \exp(i2\pi\nu\tau)\, d\tau =: h(\mathbf{r}_o, \mathbf{r}_i) \tag{1.3}$$

is known as the coherent point-spread function; note that throughout we omit the explicit expression of the temporal frequency ν. If the temporal Fourier transform of the signal exists

$$f(\mathbf{r}, \nu) = \int \tilde{f}(\mathbf{r}, t)\, \exp(i2\pi\nu t)\, dt =: f(\mathbf{r}) \tag{1.4}$$

we can formulate the input-output relationship in the temporal-frequency domain as[12]

$$f_o(\mathbf{r}_o) = \int h(\mathbf{r}_o, \mathbf{r}_i)\, f_i(\mathbf{r}_i)\, d\mathbf{r}_i \tag{1.5}$$

1.2.2 Mutual Coherence Function and Cross-Spectral Density

How shall we proceed if the temporal Fourier transform of the signal does not exist? This happens in the general case of partially coherent light, where the signal $\tilde{f}(\mathbf{r}, t)$ should be considered as a stochastic process. We then start with the mutual coherence function[13–16]

$$\tilde{\Gamma}(\mathbf{r}_1, \mathbf{r}_2, t_1, t_2) = \mathrm{E}\{\tilde{f}(\mathbf{r}_1, t_1)\, \tilde{f}^*(\mathbf{r}_2, t_2)\} =: \tilde{\Gamma}(\mathbf{r}_1, \mathbf{r}_2, t_1 - t_2) \tag{1.6}$$

where we have assumed that the stochastic process is temporally stationary. After Fourier transforming the mutual coherence function $\tilde{\Gamma}(\mathbf{r}_1, \mathbf{r}_2, \tau)$, we get the mutual power spectrum[15,16] or cross-spectral density:[17]

$$\Gamma(\mathbf{r}_1, \mathbf{r}_2, \nu) = \int \tilde{\Gamma}(\mathbf{r}_1, \mathbf{r}_2, \tau) \exp(i2\pi\nu\tau)\, d\tau =: \Gamma(\mathbf{r}_1, \mathbf{r}_2) \tag{1.7}$$

The basic property[16,17] of $\Gamma(\mathbf{r}_1, \mathbf{r}_2)$ is that it is a nonnegative definite Hermitian function of $\mathbf{r}_1$ and $\mathbf{r}_2$, i.e.,

$$\Gamma(\mathbf{r}_1, \mathbf{r}_2) = \Gamma^*(\mathbf{r}_2, \mathbf{r}_1) \qquad \text{and} \qquad \int\int g(\mathbf{r}_1)\Gamma(\mathbf{r}_1, \mathbf{r}_2)g^*(\mathbf{r}_2)\, d\mathbf{r}_1\, d\mathbf{r}_2 \geq 0 \tag{1.8}$$

for any function $g(\mathbf{r})$. The input-output relationship can now be formulated in the temporal-frequency domain as

$$\Gamma_o(\mathbf{r}_1, \mathbf{r}_2) = \int\int h(\mathbf{r}_1, \boldsymbol{\rho}_1)\, \Gamma_i(\boldsymbol{\rho}_1, \boldsymbol{\rho}_2)\, h^*(\mathbf{r}_2, \boldsymbol{\rho}_2)\, d\boldsymbol{\rho}_1\, d\boldsymbol{\rho}_2 \tag{1.9}$$

which expression replaces Eq. (1.5). Note that in the completely coherent case, for which $\Gamma(\mathbf{r}_1, \mathbf{r}_2)$ takes the product form $f(\mathbf{r}_1) f^*(\mathbf{r}_2)$, the coherence is preserved and Eq. (1.9) reduces to Eq. (1.5).

1.2.3 Some Basic Examples of Optical Signals

Important basic examples of coherent signals, as they appear in a plane $z =$ constant, are as follows:

1. An impulse in that plane at position $\mathbf{r}_o$, $f(\mathbf{r}) = \delta(\mathbf{r}-\mathbf{r}_o)$. In optical terms, the impulse corresponds to a point source.
2. The crossing with that plane of a plane wave with spatial frequency $\mathbf{q}_o$, $f(\mathbf{r}) = \exp(i2\pi\mathbf{q}_o^t\mathbf{r})$. The plane wave example shows us how we should interpret the spatial-frequency vector $\mathbf{q}_o$. We assume that the wavelength of the light equals λ_o, in which case the length of the wave vector $\mathbf{k}$ equals $2\pi/\lambda_o$. If we express the wave vector in the form $\mathbf{k} = [k_x, k_y, k_z]^t$, then $2\pi\mathbf{q}_o = 2\pi[q_x, q_y]^t = [k_x, k_y]^t$ is simply the transversal part of $\mathbf{k}$, that is, its projection onto the plane $z =$ constant. Furthermore, if the angle between the wave vector $\mathbf{k}$ and the z axis equals θ, then the length of the spatial-frequency vector $\mathbf{q}_o$ equals $\sin\theta/\lambda_o$.
3. The crossing with that plane of a spherical wave (in the paraxial approximation), $f(\mathbf{r}) = \exp(i\pi\mathbf{r}^t\mathbf{H}\mathbf{r})$, whose curvature is

described by the real symmetric 2×2 matrix $\mathbf{H} = \mathbf{H}^t$. We use this example to introduce the *instantaneous* frequency of a signal $|f(\mathbf{r})| \exp[i2\pi\phi(\mathbf{r})]$ as the derivative $d\phi/d\mathbf{r} = \nabla\phi(\mathbf{r}) = [\partial\phi/\partial x, \partial\phi/\partial y]^t$ of the signal's argument. In the case of a spherical wave we have $d\phi/d\mathbf{r} = \mathbf{H}\mathbf{r}$, and the instantaneous frequency corresponds to the normal on the spherical wavefront.

Basic example of partially coherent signals include

4. Completely incoherent light with intensity $p(\mathbf{r})$, $\Gamma(\mathbf{r}_1, \mathbf{r}_2) = p(\mathbf{r}_1)\delta(\mathbf{r}_1 - \mathbf{r}_2)$. Note that $p(\mathbf{r})$ is a nonnegative function.
5. Spatially stationary light, $\Gamma(\mathbf{r}_1, \mathbf{r}_2) = s(\mathbf{r}_1 - \mathbf{r}_2)$. We will see later that the Fourier transform of $s(\mathbf{r})$ is a nonnegative function.

1.3 Wigner Distribution and Ambiguity Function

In this section we introduce the Wigner distribution and its Fourier transform, the ambiguity function.

1.3.1 Definitions

We introduce the spatial Fourier transforms of $f(\mathbf{r})$ and $\Gamma(\mathbf{r}_1, \mathbf{r}_2)$:

$$\bar{f}(\mathbf{q}) = \int f(\mathbf{r}) \exp(-i2\pi\mathbf{q}^t\mathbf{r})\, d\mathbf{r} \tag{1.10}$$

$$\bar{\Gamma}(\mathbf{q}_1, \mathbf{q}_2) = \int\int \Gamma(\mathbf{r}_1, \mathbf{r}_2) \exp\left[-i2\pi\left(\mathbf{q}_1^t\mathbf{r}_1 - \mathbf{q}_2^t\mathbf{r}_2\right)\right] d\mathbf{r}_1\, d\mathbf{r}_2 \tag{1.11}$$

Throughout we use the generic form $\Gamma(\mathbf{r}_1, \mathbf{r}_2)$, even in the case of completely coherent light, where we could use the product form $f(\mathbf{r}_1)f^*(\mathbf{r}_2)$. We thus elaborate on Eq. (1.11) and apply the coordinate transformation

$$\begin{aligned} \mathbf{r}_1 &= \mathbf{r} + \tfrac{1}{2}\mathbf{r}' \qquad & \mathbf{r} &= \tfrac{1}{2}(\mathbf{r}_1 + \mathbf{r}_2) \\ \mathbf{r}_2 &= \mathbf{r} - \tfrac{1}{2}\mathbf{r}' \qquad & \mathbf{r}' &= \mathbf{r}_2 - \mathbf{r}_1 \end{aligned} \tag{1.12}$$

and similarly for $\mathbf{q}$. Note that the jacobian equals 1, so that $d\mathbf{r}_1\, d\mathbf{r}_2 = d\mathbf{r}\, d\mathbf{r}'$. The Wigner distribution[1] $W(\mathbf{r}, \mathbf{q})$ and ambiguity function[18] $A(\mathbf{r}', \mathbf{q}')$ now arise "midway" between the cross-spectral density

$\Gamma(\mathbf{r}_1, \mathbf{r}_2)$ and its Fourier transform $\bar{\Gamma}(\mathbf{q}_1, \mathbf{q}_2)$,

$$\begin{aligned}\bar{\Gamma}\left(\mathbf{q}+\tfrac{1}{2}\mathbf{q}', \mathbf{q}-\tfrac{1}{2}\mathbf{q}'\right) &= \int\int \Gamma\left(\mathbf{r}+\tfrac{1}{2}\mathbf{r}', \mathbf{r}-\tfrac{1}{2}\mathbf{r}'\right) \\ &\quad \times \exp[-i2\pi(\mathbf{q}^t\mathbf{r}' + \mathbf{r}^t\mathbf{q}')]\, d\mathbf{r}\, d\mathbf{r}' \\ &= \int W(\mathbf{r}, \mathbf{q})\, \exp(-i2\pi\mathbf{r}^t\mathbf{q}')\, d\mathbf{r} \\ &= \int A(\mathbf{r}', \mathbf{q}')\, \exp(-i2\pi\mathbf{q}^t\mathbf{r}')\, d\mathbf{r}' \qquad (1.13)\end{aligned}$$

and their definitions follow as

$$\begin{aligned}W(\mathbf{r}, \mathbf{q}) &= \int \Gamma\left(\mathbf{r}+\tfrac{1}{2}\mathbf{r}', \mathbf{r}-\tfrac{1}{2}\mathbf{r}'\right) \exp(-i2\pi\mathbf{q}^t\mathbf{r}')\, d\mathbf{r}' \\ &= \int \bar{\Gamma}\left(\mathbf{q}+\tfrac{1}{2}\mathbf{q}', \mathbf{q}-\tfrac{1}{2}\mathbf{q}'\right) \exp(i2\pi\mathbf{r}^t\mathbf{q}')\, d\mathbf{q}' \qquad (1.14)\end{aligned}$$

$$\begin{aligned}A(\mathbf{r}', \mathbf{q}') &= \int \Gamma\left(\mathbf{r}+\tfrac{1}{2}\mathbf{r}', \mathbf{r}-\tfrac{1}{2}\mathbf{r}'\right) \exp(-i2\pi\mathbf{r}^t\mathbf{q}')\, d\mathbf{r} \\ &= \int \bar{\Gamma}\left(\mathbf{q}+\tfrac{1}{2}\mathbf{q}', \mathbf{q}-\tfrac{1}{2}\mathbf{q}'\right) \exp(i2\pi\mathbf{q}^t\mathbf{r}')\, d\mathbf{q} \qquad (1.15)\end{aligned}$$

We immediately notice the realness of the Wigner distribution, and the Fourier transform relationship between the Wigner distribution and the ambiguity function:

$$\begin{aligned}A(\mathbf{r}', \mathbf{q}') &= \int W(\mathbf{r}, \mathbf{q})\, \exp[-i2\pi(\mathbf{r}^t\mathbf{q}' - \mathbf{q}^t\mathbf{r}')]\, d\mathbf{r}\, d\mathbf{q} \\ &= \mathcal{F}[W(\mathbf{r}, \mathbf{q})](\mathbf{r}', \mathbf{q}') \qquad (1.16)\end{aligned}$$

This Fourier transform relationship implies that properties for the Wigner distribution have their counterparts for the ambiguity function and vice versa: moments for the Wigner distribution become derivatives for the ambiguity function, convolutions in the Wigner domain become products in the ambiguity domain, etc.

We like to present the cross-spectral density Γ, its spatial Fourier transform $\bar{\Gamma}$, the Wigner distribution W, and the ambiguity function A at the corners of a rectangle (see Fig. 1.1). Along the sides of this rectangle we have Fourier transformations $\mathbf{r}' \to \mathbf{q}$ and $\mathbf{r} \to \mathbf{q}'$ and their inverses, while along the diagonals we have double Fourier transformations $(\mathbf{r}, \mathbf{r}') \to (\mathbf{q}', \mathbf{q})$ and $(\mathbf{r}, \mathbf{q}) \to (\mathbf{q}', \mathbf{r}')$.

A distribution according to the definitions in (1.14) was introduced in optics by Dolin[2] and Walther[3,4] in the field of radiometry; Walther called it the *generalized radiance*. A few years later it was reintroduced, mainly in the field of Fourier optics.[5–11] The ambiguity function was

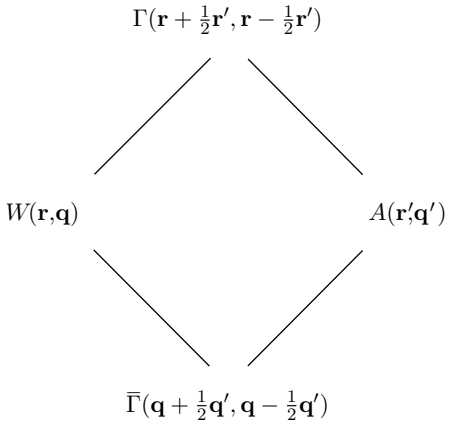


FIGURE 1.1 Schematic representation of the cross-spectral density Γ, its spatial Fourier transform $\bar{\Gamma}$, the Wigner distribution W, and the ambiguity function A, on a rectangle.

introduced in optics by Papoulis.[19] The ambiguity function is treated in greater detail in Chap. 2 by Jean-Pierre Guigay; in this chapter we concentrate on the Wigner distribution.

1.3.2 Some Basic Examples Again

Let us return to our basic examples. The space behavior $f(\mathbf{r})$ or $\Gamma(\mathbf{r}_1, \mathbf{r}_2)$, the spatial-frequency behavior $\bar{f}(\mathbf{q})$ or $\bar{\Gamma}(\mathbf{q}_1, \mathbf{q}_2)$, and the Wigner distribution $W(\mathbf{r}, \mathbf{q})$ of (1) a point source, (2) a plane wave, (3) a spherical wave, (4) an incoherent light field, and (5) a spatially stationary light field are represented in Table 1.1.

Example*	$f(\mathbf{r})$ or $\Gamma(\mathbf{r}_1, \mathbf{r}_2)$	$\bar{f}(\mathbf{q})$ or $\bar{\Gamma}(\mathbf{q}_1, \mathbf{q}_2)$	$W(\mathbf{r}, \mathbf{q})$
(1)	$\delta(\mathbf{r}-\mathbf{r}_o)$	$\exp(-i2\pi\mathbf{r}_o^t\mathbf{q})$	$\delta(\mathbf{r}-\mathbf{r}_o)$
(2)	$\exp(i2\pi\mathbf{q}_o^t\mathbf{r})$	$\delta(\mathbf{q}-\mathbf{q}_o)$	$\delta(\mathbf{q}-\mathbf{q}_o)$
(3)	$\exp(i\pi\mathbf{r}^t\mathbf{H}\mathbf{r})$	$[\det(-i\mathbf{H})]^{-1/2}$ $\exp(-i\pi\mathbf{q}^t\mathbf{H}^{-1}\mathbf{q})$	$\delta(\mathbf{q}-\mathbf{H}\mathbf{r})$
(4)	$p(\mathbf{r}_1)\,\delta(\mathbf{r}_1-\mathbf{r}_2)$	$\bar{p}(\mathbf{q}_1-\mathbf{q}_2)$	$p(\mathbf{r})$
(5)	$s(\mathbf{r}_1-\mathbf{r}_2)$	$\bar{s}(\mathbf{q}_1)\,\delta(\mathbf{q}_1-\mathbf{q}_2)$	$\bar{s}(\mathbf{q})$

*(1) Point source, (2) plane wave, (3) spherical wave, (4) incoherent light, and (5) spatially stationary light.

TABLE 1.1 Wigner Distribution of Some Basic Examples

We remark the clear physical interpretations of the Wigner distributions.

1. The Wigner distribution of a point source $f(\mathbf{r}) = \delta(\mathbf{r} - \mathbf{r}_o)$ reads $W(\mathbf{r}, \mathbf{q}) = \delta(\mathbf{r} - \mathbf{r}_o)$, and we observe that all the light originates from one point $\mathbf{r} = \mathbf{r}_o$ and propagates uniformly in all directions $\mathbf{q}$.
2. Its dual, a plane wave $f(\mathbf{r}) = \exp(i2\pi\mathbf{q}_o^t\mathbf{r})$, also expressible in the frequency domain as $\bar{f}(\mathbf{q}) = \delta(\mathbf{q} - \mathbf{q}_o)$, has as its Wigner distribution $W(\mathbf{r}, \mathbf{q}) = \delta(\mathbf{q} - \mathbf{q}_o)$, and we observe that for all positions $\mathbf{r}$ the light propagates in only one direction $\mathbf{q}_o$.
3. The Wigner distribution of the spherical wave $f(\mathbf{r}) = \exp(i\pi\mathbf{r}^t\mathbf{H}\mathbf{r})$ takes the simple form $W(\mathbf{r}, \mathbf{q}) = \delta(\mathbf{q} - \mathbf{H}\mathbf{r})$, and we conclude that at any point $\mathbf{r}$ only one frequency $\mathbf{q} = \mathbf{H}\mathbf{r}$, the instantaneous frequency, manifests itself. This corresponds exactly to the ray picture of a spherical wave.
4. Incoherent light, $\Gamma(\mathbf{r}+\frac{1}{2}\mathbf{r}', \mathbf{r}-\frac{1}{2}\mathbf{r}') = p(\mathbf{r})\,\delta(\mathbf{r}')$, yields the Wigner distribution $W(\mathbf{r}, \mathbf{q}) = p(\mathbf{r})$. Note that it is a function of the space variable $\mathbf{r}$ only, and that it does not depend on the frequency variable $\mathbf{q}$: the light radiates equally in all directions, with intensity profile $p(\mathbf{r}) \geq 0$.
5. Spatially stationary light, $\Gamma(\mathbf{r} + \frac{1}{2}\mathbf{r}', \mathbf{r} - \frac{1}{2}\mathbf{r}') = s(\mathbf{r}')$, is dual to incoherent light: its frequency behavior is similar to the space behavior of incoherent light and vice versa, and $\bar{s}(\mathbf{q})$, its intensity function in the frequency domain, is nonnegative. The duality between incoherent light and spatially stationary light is, in fact, the Van Cittert-Zernike theorem.

   The Wigner distribution of spatially stationary light reads as $W(\mathbf{r}, \mathbf{q}) = \bar{s}(\mathbf{q})$; note that it is a function of the frequency variable $\mathbf{q}$ only, and that it does not depend on the space variable $\mathbf{r}$. It thus has the same form as the Wigner distribution of incoherent light, except that it is rotated through 90° in the space-frequency domain. The same observation can be made for the point source and the plane wave; see examples (1) and (2), which are also each other's duals.

We illustrate the Wigner distribution of the one-dimensional spherical wave $f(x) = \exp(i\pi hx^2)$, (see example (3) above), by a numerical simulation. To calculate $W(x, u)$ practically, we have to restrict the integration interval for x'. We model this by using a window function $w(\frac{1}{2}x')$, so that the Wigner distribution takes the form

$$P(x, u; w) = \int f\left(x + \tfrac{1}{2}x'\right) w\left(\tfrac{1}{2}x'\right) w^*\left(-\tfrac{1}{2}x'\right) f^*\left(x - \tfrac{1}{2}x'\right) \times \exp\left(-i2\pi ux'\right) dx' \tag{1.17}$$

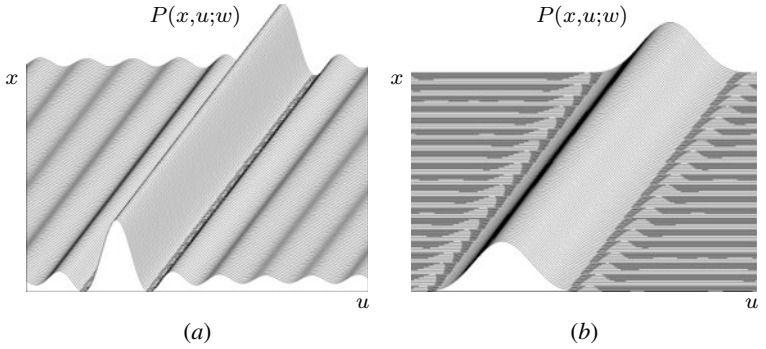


FIGURE 1.2 Numerical simulation of the (pseudo) Wigner distribution $P(x, u; w) \simeq W(x, u) = \delta(u - hx)$ of the spherical wave $f(x) = \exp(i\pi hx^2)$, for the case that $w(\frac{1}{2}x')$ is (*a*) a rectangular window and (*b*) a Hann(ing) window.

The function $P(x, u; w)$ is called the pseudo-Wigner distribution. It is common to choose an even window function $w(\frac{1}{2}x') = w(-\frac{1}{2}x')$, so that we have $w(\frac{1}{2}x')w^*(-\frac{1}{2}x') = |w(\frac{1}{2}x')|^2$. Figure 1.2 shows the (pseudo) Wigner distribution of the signal $f(x) = \exp(i\pi hx^2)$, which reads as

$$\int \left|w\left(\tfrac{1}{2}x'\right)\right|^2 \exp[-i2\pi(u-hx)]\,dx' = \mathcal{F}\left[\left|w\left(\tfrac{1}{2}x'\right)\right|^2\right](u-hx) \simeq \delta(u-hx)$$

where we have chosen a rectangular window of width X in Fig. 1.2*a*

$$w\left(\tfrac{1}{2}x'\right) = \text{rect}\left(\frac{x'}{X}\right)$$

and a Hann(ing) window of width X in Fig. 1.2*b*

$$w\left(\tfrac{1}{2}x'\right) = \cos^2\left(\frac{\pi x'}{X}\right)\text{rect}\left(\frac{x'}{X}\right)$$

Note the effect of $\mathcal{F}\left[|w(\frac{1}{2}x')|^2\right]$, which results in a sinc-type behavior in the case of the rectangular window, $P(x, u; w) = \sin[\pi(u - hx)X]/\pi(u - hx)$, and in a nonnegative but smoother version in the case of the Hann(ing) window.

1.3.3 Gaussian Light

Gaussian light is an example that we will treat in greater detail. The cross-spectral density of the most general partially coherent Gaussian

light can be written in the form

$$\Gamma(\mathbf{r}_1, \mathbf{r}_2) = 2\sqrt{\det \mathbf{G}_1}\exp\left(-\frac{\pi}{2}\begin{bmatrix}\mathbf{r}_1+\mathbf{r}_2\\ \mathbf{r}_1-\mathbf{r}_2\end{bmatrix}^t\begin{bmatrix}\mathbf{G}_1 & -i\mathbf{H}\\ -i\mathbf{H}^t & \mathbf{G}_2\end{bmatrix}\begin{bmatrix}\mathbf{r}_1+\mathbf{r}_2\\ \mathbf{r}_1-\mathbf{r}_2\end{bmatrix}\right) \tag{1.18}$$

where we have chosen a representation that enables us to determine the Wigner distribution of such light in an easy way. The exponent shows a quadratic form in which a four-dimensional column vector $[(\mathbf{r}_1+\mathbf{r}_2)^t, (\mathbf{r}_1-\mathbf{r}_2)^t]^t$ arises together with a symmetric 4×4 matrix. This matrix consists of four real 2×2 submatrices $\mathbf{G}_1$, $\mathbf{G}_2$, $\mathbf{H}$, and $\mathbf{H}^t$, where, moreover, matrices $\mathbf{G}_1$ and $\mathbf{G}_2$ are positive definite symmetric. The special form of the matrix is a direct consequence of the fact that the cross-spectral density is a nonnegative definite Hermitian function. The Wigner distribution of such Gaussian light takes the form[20,21]

$$W(\mathbf{r}, \mathbf{q}) = 4\sqrt{\frac{\det \mathbf{G}_1}{\det \mathbf{G}_2}}\exp\left(-2\pi\begin{bmatrix}\mathbf{r}\\ \mathbf{q}\end{bmatrix}^t\begin{bmatrix}\mathbf{G}_1+\mathbf{H}\mathbf{G}_2^{-1}\mathbf{H}^t & -\mathbf{H}\mathbf{G}_2^{-1}\\ -\mathbf{G}_2^{-1}\mathbf{H}^t & \mathbf{G}_2^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{r}\\ \mathbf{q}\end{bmatrix}\right) \tag{1.19}$$

In a more common way, the cross-spectral density of general Gaussian light (with 10 degrees of freedom) can be expressed in the form

$$\begin{aligned}\Gamma(\mathbf{r}_1, \mathbf{r}_2) = {} & 2\sqrt{\det \mathbf{G}_1}\exp\left[-\tfrac{1}{2}\pi(\mathbf{r}_1-\mathbf{r}_2)^t\,\mathbf{G}_0\,(\mathbf{r}_1-\mathbf{r}_2)\right]\\ & \times\exp\left\{-\pi\mathbf{r}_1^t\left[\mathbf{G}_1 - i\tfrac{1}{2}(\mathbf{H}+\mathbf{H}^t)\right]\mathbf{r}_1\right\}\\ & \times\exp\left\{-\pi\mathbf{r}_2^t\left[\mathbf{G}_1 + i\tfrac{1}{2}(\mathbf{H}+\mathbf{H}^t)\right]\mathbf{r}_2\right\}\\ & \times\exp\left[-\pi\mathbf{r}_1^t\, i(\mathbf{H}-\mathbf{H}^t)\,\mathbf{r}_2\right]\end{aligned} \tag{1.20}$$

where we have introduced the real, positive definite symmetric 2×2 matrix $\mathbf{G}_0 = \mathbf{G}_2 - \mathbf{G}_1$. Note that the asymmetry of matrix $\mathbf{H}$ is a measure for the twist[22–26] of Gaussian light, and that general Gaussian light reduces to zero-twist Gaussian Schell-model light[27,28] if the matrix $\mathbf{H}$ is symmetric, $\mathbf{H}-\mathbf{H}^t = \mathbf{0}$. In that case, the light can be considered as spatially stationary light with a Gaussian cross-spectral density $2\sqrt{\det \mathbf{G}_1}\exp[-\frac{1}{2}\pi(\mathbf{r}_1-\mathbf{r}_2)^t\,\mathbf{G}_0\,(\mathbf{r}_1-\mathbf{r}_2)]$, modulated by a Gaussian modulator with modulation function $\exp[-\pi\mathbf{r}^t\,(\mathbf{G}_1 - i\mathbf{H})\,\mathbf{r}]$. We remark that such Gaussian Schell-model light (with nine degrees of freedom) forms a large subclass of Gaussian light; it applies, for instance, in

- the completely coherent case ($\mathbf{H} = \mathbf{H}^t$, $\mathbf{G}_0 = \mathbf{0}$, $\mathbf{G}_1 = \mathbf{G}_2$)
- the (partially coherent) one-dimensional case ($g_0 = g_2 - g_1 \geq 0$)

- the (partially coherent) rotationally symmetric case ($\mathbf{H} = h\mathbf{I}$, $\mathbf{G}_1 = g_1\mathbf{I}$, $\mathbf{G}_2 = g_2\mathbf{I}$, $\mathbf{G}_0 = (g_2 - g_1)\mathbf{I}$, with $\mathbf{I}$ the 2×2 identity matrix)

Gaussian Schell-model light reduces to so-called symplectic Gaussian light,[21] if matrices $\mathbf{G}_0$, $\mathbf{G}_1$, and $\mathbf{G}_2$ are proportional to one another. Now $\mathbf{G}_1 = \sigma\mathbf{G}$, $\mathbf{G}_2 = \sigma^{-1}\mathbf{G}$, and thus $\mathbf{G}_0 = (\sigma^{-1} - \sigma)\mathbf{G}$, with $\mathbf{G}$ a real, positive definite symmetric 2×2 matrix and $0 < \sigma \leq 1$. The Wigner distribution then takes the form

$$W(\mathbf{r}, \mathbf{q}) = 4\sigma^2 \exp\left(-2\pi\sigma \begin{bmatrix}\mathbf{r}\\ \mathbf{q}\end{bmatrix}^t \begin{bmatrix}\mathbf{G} + \mathbf{H}\mathbf{G}^{-1}\mathbf{H} & -\mathbf{H}\mathbf{G}^{-1}\\ -\mathbf{G}^{-1}\mathbf{H} & \mathbf{G}^{-1}\end{bmatrix} \begin{bmatrix}\mathbf{r}\\ \mathbf{q}\end{bmatrix}\right) \tag{1.21}$$

The name *symplectic Gaussian light* (with six degrees of freedom) originates from the fact that the 4×4 matrix that arises in the exponent of the Wigner distribution (1.21) is symplectic. We will return to symplecticity later in this chapter. We remark that symplectic Gaussian light forms a large subclass of Gaussian Schell-model light; it applies again, for instance, in the completely coherent case, in the (partially coherent) one-dimensional case, and in the (partially coherent) rotationally symmetric case. And again, symplectic Gaussian light can be considered as spatially stationary light with a Gaussian cross-spectral density, modulated by a Gaussian modulator [cf. Eq. (1.20)], but now with the real parts of the quadratic forms in the two exponents described—up to a positive constant—by the same real, positive definite symmetric matrix $\mathbf{G}$.

1.3.4 Local Frequency Spectrum

The Wigner distribution can be considered as a local frequency spectrum; the marginals are correct

$$\Gamma(\mathbf{r}, \mathbf{r}) = \int W(\mathbf{r}, \mathbf{q})\, d\mathbf{q} \qquad \text{and} \qquad \bar{\Gamma}(\mathbf{q}, \mathbf{q}) = \int W(\mathbf{r}, \mathbf{q})\, d\mathbf{r} \tag{1.22}$$

Integrating over all frequency values $\mathbf{q}$ yields the intensity $\Gamma(\mathbf{r}, \mathbf{r})$ of the signal's representation in the space domain, and integrating over all space values $\mathbf{r}$ yields the intensity $\bar{\Gamma}(\mathbf{q}, \mathbf{q})$ of the signal's representation in the frequency domain. To operate easily in the mixed $\mathbf{rq}$ plane, the so-called phase space, we will benefit from normalization to dimensionless coordinates $\mathbf{W}^{-1}\mathbf{r} =: \mathbf{r}$ and $\mathbf{W}\mathbf{q} =: \mathbf{q}$, where $\mathbf{W}$ is a

diagonal matrix with positive diagonal entries

$$\mathbf{W} = \begin{bmatrix} w_x & 0 \\ 0 & w_y \end{bmatrix} \tag{1.23}$$

In subsequent sections, we will often work with these normalized coordinates; it will be clear from the context whether normalization is necessary.

1.4 Some Properties of the Wigner Distribution

Let us consider some of the important properties of the Wigner distribution. We consider in particular properties that are specific for partially coherent light. Additional properties of the Wigner distribution, especially of the Wigner distribution in the completely coherent case, can be found elsewhere; see, for instance, Refs. 29 to 40 and the many references cited therein.

1.4.1 Inversion Formula

The definition (1.14) of the Wigner distribution $W(\mathbf{r}, \mathbf{q})$ has the form of a Fourier transformation of the cross-spectral density $\Gamma(\mathbf{r} + \frac{1}{2}\mathbf{r}', \mathbf{r} - \frac{1}{2}\mathbf{r}')$ with $\mathbf{r}'$ and $\mathbf{q}$ as conjugated variables and with $\mathbf{r}$ as a parameter. The cross-spectral density can thus be reconstructed from the Wigner distribution simply by applying an inverse Fourier transformation.

1.4.2 Shift Covariance

The Wigner distribution satisfies the important property of space and frequency shift covariance: if $W(\mathbf{r}, \mathbf{q})$ is the Wigner distribution that corresponds to $\Gamma(\mathbf{r}_1, \mathbf{r}_2)$, then $W(\mathbf{r} - \mathbf{r}_o, \mathbf{q} - \mathbf{q}_o)$ is the Wigner distribution that corresponds to the space- and frequency-shifted version $\Gamma(\mathbf{r}_1 - \mathbf{r}_o, \mathbf{r}_2 - \mathbf{r}_o) \exp[i2\pi\mathbf{q}_o^t(\mathbf{r}_1 - \mathbf{r}_2)]$.

1.4.3 Radiometric Quantities

Although the Wigner distribution is real, it is not necessarily nonnegative; this prohibits a direct interpretation of the Wigner distribution as an energy density function (or radiance function). Friberg has shown[41] that it is not possible to define a radiance function that satisfies all the physical requirements from radiometry; in particular, as we mentioned, the Wigner distribution has the physically unattractive property that it may take negative values.

Nevertheless, several integrals of the Wigner distribution have clear physical meanings and can be interpreted as radiometric quantities.

We mentioned already that the integral over the frequency variable $\int W(\mathbf{r}, \mathbf{q})\, d\mathbf{q} = \Gamma(\mathbf{r}, \mathbf{r})$ represents the intensity of the signal, whereas the integral over the space variable $\int W(\mathbf{r}, \mathbf{q})\, d\mathbf{r} = \bar{\Gamma}(\mathbf{q}, \mathbf{q})$ yields the intensity of the signal's Fourier transform; the latter is, apart from the usual factor $\cos^2\theta$ (where θ is the angle of observation with respect to the z axis), proportional to the radiant intensity.[42,43] The total energy E of the signal follows from the integral over the entire space-frequency domain:

$$E = \int\int W(\mathbf{r}, \mathbf{q})\, d\mathbf{r}\, d\mathbf{q} \tag{1.24}$$

The real symmetric 4×4 matrix $\mathbf{M}$ of normalized second-order moments, defined by

$$\begin{aligned}\mathbf{M} &= \frac{1}{E}\int\int \begin{bmatrix}\mathbf{r}\\ \mathbf{q}\end{bmatrix} [\mathbf{r}^t, \mathbf{q}^t] W(\mathbf{r}, \mathbf{q})\, d\mathbf{r}\, d\mathbf{q} \\ &= \frac{1}{E}\int\int \begin{bmatrix}\mathbf{r}\mathbf{r}^t & \mathbf{r}\mathbf{q}^t\\ \mathbf{q}\mathbf{r}^t & \mathbf{q}\mathbf{q}^t\end{bmatrix} W(\mathbf{r}, \mathbf{q})\, d\mathbf{r}\, d\mathbf{q} \\ &= \begin{bmatrix}\mathbf{M}_{\mathbf{rr}} & \mathbf{M}_{\mathbf{rq}}\\ \mathbf{M}_{\mathbf{rq}}^t & \mathbf{M}_{\mathbf{qq}}\end{bmatrix} = \begin{bmatrix} m_{xx} & m_{xy} & m_{xu} & m_{xv}\\ m_{xy} & m_{yy} & m_{yu} & m_{yv}\\ m_{xu} & m_{yu} & m_{uu} & m_{uv}\\ m_{xv} & m_{yv} & m_{uv} & m_{vv}\end{bmatrix}\end{aligned} \tag{1.25}$$

yields such quantities as the effective width $d_x = \sqrt{m_{xx}}$ of the intensity $\Gamma(\mathbf{r}, \mathbf{r})$ in the x direction

$$m_{xx} = \frac{1}{E}\int\int x^2\, W(\mathbf{r}, \mathbf{q})\, d\mathbf{r}\, d\mathbf{q} = \frac{1}{E}\int x^2\, \Gamma(\mathbf{r}, \mathbf{r})\, d\mathbf{r} = d_x^2 \tag{1.26}$$

and similarly the effective width $d_u = \sqrt{m_{uu}}$ of the intensity $\bar{\Gamma}(\mathbf{q}, \mathbf{q})$ in the u direction, but it also yields all kinds of mixed moments. It will be clear that the main-diagonal entries of the moment matrix $\mathbf{M}$, being interpretable as squares of effective widths, are positive. As a matter of fact, it can be shown that the matrix $\mathbf{M}$ is positive definite; see, for instance, Refs. 44 to 46

The radiant emittance[42,43] is equal to the integral

$$j_z(\mathbf{r}) = \int \frac{\sqrt{k^2 - (2\pi)^2 \mathbf{q}^t\mathbf{q}}}{k}\, W(\mathbf{r}, \mathbf{q})\, d\mathbf{q} \tag{1.27}$$

where $k = 2\pi/\lambda_o$ represents the usual wave number. When we combine the radiant emittance j_z with the two-dimensional vector

$$\mathbf{j}_r(\mathbf{r}) = \int \frac{2\pi\mathbf{q}}{k}\, W(\mathbf{r}, \mathbf{q})\, d\mathbf{q} \tag{1.28}$$

we can construct the three-dimensional vector $[\mathbf{j}_r^t,\ j_z]^t$, which is known as the geometrical vector flux.[47] The total radiant flux[42] $\int j_z(\mathbf{r})\,d\mathbf{r}$ follows from integrating the radiant emittance over the space variable $\mathbf{r}$. More on radiometry can be found in Chap. 7 by Arvind Marathay.

1.4.4 Instantaneous Frequency

The Wigner distribution $W_f(\mathbf{r}, \mathbf{q})$ satisfies the nice property that for a coherent signal $f(\mathbf{r}) = |f(\mathbf{r})| \exp[i2\pi\phi(\mathbf{r})]$, the instantaneous frequency $d\phi/d\mathbf{r} = \nabla\phi(\mathbf{r})$ follows from $W_f(\mathbf{r}, \mathbf{q})$ through

$$\frac{d\phi}{d\mathbf{r}} = \frac{\displaystyle\int \mathbf{q}\, W_f(\mathbf{r}, \mathbf{q})\, d\mathbf{q}}{\displaystyle\int W_f(\mathbf{r}, \mathbf{q})\, d\mathbf{q}} \tag{1.29}$$

To prove this property, we proceed as follows. From $f(\mathbf{r}) = |f(\mathbf{r})| \exp[i2\pi\phi(\mathbf{r})]$, we get $\ln f(\mathbf{r}) = \ln|f(\mathbf{r})| + i2\pi\phi(\mathbf{r})$, hence $\mathrm{Im}\{\ln f(\mathbf{r})\} = 2\pi\phi(\mathbf{r})$, which then leads to the identity

$$\begin{aligned}
2\pi \frac{d\phi(\mathbf{r})}{d\mathbf{r}} &= \mathrm{Im}\left\{\frac{d \ln f(\mathbf{r})}{d\mathbf{r}}\right\} = \mathrm{Im}\left\{\frac{\nabla f(\mathbf{r})}{f(\mathbf{r})}\right\} \\
&= \frac{1}{2i}\left[\frac{\nabla f(\mathbf{r})}{f(\mathbf{r})} - \left(\frac{\nabla f(\mathbf{r})}{f(\mathbf{r})}\right)^*\right] \\
&= \frac{1}{2i}\,\frac{[\nabla f(\mathbf{r})] f^*(\mathbf{r}) - f(\mathbf{r})[\nabla f(\mathbf{r})]^*}{f(\mathbf{r})\, f^*(\mathbf{r})} \\
&= -i\,\frac{1}{|f(\mathbf{r})|^2}\,\frac{\partial}{\partial \mathbf{r}'}\left[f\left(\mathbf{r} + \tfrac{1}{2}\mathbf{r}'\right) f^*\left(\mathbf{r} - \tfrac{1}{2}\mathbf{r}'\right)\right]\Big|_{\mathbf{r}'=\mathbf{0}}
\end{aligned}$$

On the other hand, we have the identity

$$\begin{aligned}
&2\pi \int \mathbf{q}\, W_f(\mathbf{r}, \mathbf{q})\, d\mathbf{q} \\
&= 2\pi \int \left[\int f\left(\mathbf{r} + \tfrac{1}{2}\mathbf{r}'\right) f^*\left(\mathbf{r} - \tfrac{1}{2}\mathbf{r}'\right) \exp(-i2\pi\mathbf{q}^t\mathbf{r}')\, d\mathbf{r}'\right] \mathbf{q}\, d\mathbf{q} \\
&= \int f\left(\mathbf{r} + \tfrac{1}{2}\mathbf{r}'\right) f^*\left(\mathbf{r} - \tfrac{1}{2}\mathbf{r}'\right) \left[2\pi \int \mathbf{q} \exp(-i2\pi\mathbf{q}^t\mathbf{r}')\, d\mathbf{q}\right] d\mathbf{r}' \\
&= i \int f\left(\mathbf{r} + \tfrac{1}{2}\mathbf{r}'\right) f^*\left(\mathbf{r} - \tfrac{1}{2}\mathbf{r}'\right) [\nabla\delta(\mathbf{r}')]\, d\mathbf{r}' \\
&= -i\,\frac{\partial}{\partial \mathbf{r}'}\left[f\left(\mathbf{r} + \tfrac{1}{2}\mathbf{r}'\right) f^*\left(\mathbf{r} - \tfrac{1}{2}\mathbf{r}'\right)\right]\Big|_{\mathbf{r}'=\mathbf{0}}
\end{aligned}$$

and when we combine these two results, we immediately get Eq. (1.29). It is this property in particular that made the Wigner distribution a popular tool for the determination of the instantaneous frequency.

1.4.5 Moyal's Relationship

An important relationship between the Wigner distributions of two signals and the cross-spectral densities of these signals, which is an extension to partially coherent light of a relationship formulated by Moyal[48] for completely coherent light, reads as

$$\begin{aligned}\int\int W_1(\mathbf{r},\mathbf{q})\, W_2(\mathbf{r},\mathbf{q})\, d\mathbf{r}\, d\mathbf{q} &= \int\int \Gamma_1(\mathbf{r}_1,\mathbf{r}_2)\, \Gamma_2^*(\mathbf{r}_1,\mathbf{r}_2)\, d\mathbf{r}_1\, d\mathbf{r}_2 \\ &= \int\int \bar{\Gamma}_1(\mathbf{q}_1,\mathbf{q}_2)\, \bar{\Gamma}_2^*(\mathbf{q}_1,\mathbf{q}_2)\, d\mathbf{q}_1\, d\mathbf{q}_2\end{aligned} \tag{1.30}$$

This relationship has an application in averaging one Wigner distribution with another one, which averaging always yields a nonnegative result.

1.5 One-Dimensional Case and the Fractional Fourier Transformation

Let us for the moment restrict ourselves to coherent light and to the one-dimensional case, and let us use normalized coordinates. The signal is now written as $f(x)$.

1.5.1 Fractional Fourier Transformation

An important transformation with respect to operations in a phase space is the fractional Fourier transformation, which reads as[49–53]

$$f_o(x_o) = F_\gamma(x_o) = \frac{\exp\left(i\frac{1}{2}\gamma\right)}{\sqrt{i\sin\gamma}} \int \exp\left[i\pi\,\frac{\left(x_i^2+x_o^2\right)\cos\gamma - 2x_o x_i}{\sin\gamma}\right] \times f_i(x_i)\, dx_i \qquad (\gamma \neq n\pi) \tag{1.31}$$

where $\sqrt{i\sin\gamma}$ is defined as $|\sin\gamma|\exp[i(\frac{1}{4}\pi)\,\mathrm{sgn}(\sin\gamma)]$. We mention the special cases $F_0(x) = f(x)$, $F_\pi(x) = f(-x)$, and the common Fourier transform $F_{\pi/2}(x) = \bar{f}(x)$. Two realizations of an optical fractional Fourier transformer have been proposed by Lohmann[50] (see Fig. 1.3). For both cases we have $\sin^2(\frac{1}{2}\gamma) = d/2f$; the normalization

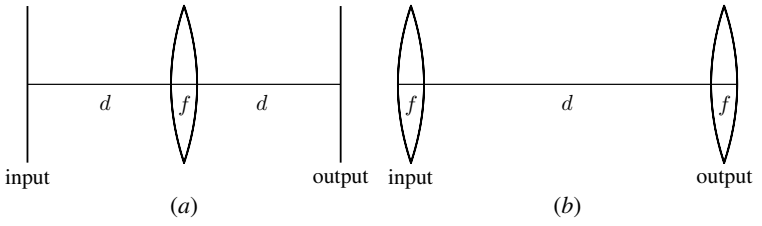


FIGURE 1.3 Two optical realizations of the fractional Fourier transformer.

width w is related to the distance d and the focal length of the lens f by $w^2 \tan(\frac{1}{2}\gamma) = \lambda_o d$ for Fig. 1.3*a* and by $w^2 \sin\gamma = \lambda_o d$ for Fig. 1.3*b*.

1.5.2 Rotation in Phase Space

In terms of the ray transformation matrix, which is introduced and treated in greater detail in Sec. 1.6, the fractional Fourier transformer is represented by

$$\begin{bmatrix} x_o \\ u_o \end{bmatrix} = \begin{bmatrix} w & 0 \\ 0 & w^{-1} \end{bmatrix} \begin{bmatrix} \cos\gamma & \sin\gamma \\ -\sin\gamma & \cos\gamma \end{bmatrix} \begin{bmatrix} w^{-1} & 0 \\ 0 & w \end{bmatrix} \begin{bmatrix} x_i \\ u_i \end{bmatrix} \tag{1.32}$$

and after normalization, $w^{-1}x =: x$ and $wu =: u$, we have the form

$$\begin{bmatrix} x_o \\ u_o \end{bmatrix} = \begin{bmatrix} \cos\gamma & \sin\gamma \\ -\sin\gamma & \cos\gamma \end{bmatrix} \begin{bmatrix} x_i \\ u_i \end{bmatrix} \tag{1.33}$$

The input-output relation of a fractional Fourier transformer in terms of the Wigner distribution is remarkably simple; if W_f denotes the Wigner distribution of $f(x)$ and W_{F_γ} denotes that of $F_\gamma(x)$, we have

$$W_{F_\gamma}(x, u) = W_f(x\cos\gamma - u\sin\gamma,\ x\sin\gamma + u\cos\gamma) \tag{1.34}$$

and we conclude that a fractional Fourier transformation corresponds to a rotation in phase space.

1.5.3 Generalized Marginals—Radon Transform

On the analogy of the two special cases $|f(x)|^2 = \int W_f(x, u)\, du$ and $|\bar{f}(u)|^2 = \int W_f(x, u)\, dx$, which correspond to projections along the u

and the x axes, respectively, we can now get an easy expression for the projection along an axis that is tilted through an angle γ.

$$\begin{aligned} |F_\gamma(x)|^2 &= \int W_{F_\gamma}(x, u)\, du \\ &= \int W_f(x\cos\gamma - u\sin\gamma,\, x\sin\gamma + u\cos\gamma)\, du \\ &= \int\int W_f(\xi, u)\, \delta(\xi\cos\gamma + u\sin\gamma - x)\, d\xi\, du \end{aligned} \tag{1.35}$$

We thus conclude that not only are the marginals for $\gamma = 0$ and $\gamma = \frac{1}{2}\pi$ correct, but in fact any marginal for an arbitrary angle γ is correct. We observe a strong connection between the Wigner distribution $W_f(x, u)$ and the intensity $|F_\gamma(x)|^2$ of the signal's fractional Fourier transform. Note also the relation to the Radon transform.

Since the ambiguity function is the two-dimensional Fourier transform of the Wigner distribution, we could also represent $|F_\gamma(x)|^2$ in the form[54–56]

$$|F_\gamma(x)|^2 = \int A_{F_\gamma}(\rho\sin\gamma,\, -\rho\cos\gamma)\, \exp(-i2\pi x\rho)\, d\rho \tag{1.36}$$

and we conclude that the values of the ambiguity function along the line defined by the angle γ and the projections of the Wigner distribution for the same angle γ are related to each other by a Fourier transformation. Note that the ambiguity function in Eq. (1.36) is represented in a quasi-polar coordinate system (ρ, γ).

We recall that the signal $f(x) = |f(x)| \exp[i2\pi\phi(x)]$ can be reconstructed by using the intensity profiles of the fractional Fourier transform $F_\gamma(x)$ for two close values of the fractional angle γ.[56] The reconstruction procedure is based on the property[54–56]

$$\left.\frac{\partial |F_\gamma(x)|^2}{\partial\gamma}\right|_{\gamma=0} = -\frac{d}{dx}\left[|f(x)|^2\, \frac{d\phi(x)}{dx}\right] \tag{1.37}$$

which can be proved by first differentiating Eq. (1.35) with respect to γ and using the identity

$$\begin{aligned} &\left.\frac{\partial\delta(\xi\cos\gamma + u\sin\gamma - x)}{\partial\gamma}\right|_{\gamma=0} \\ &= (-\xi\sin\gamma + u\cos\gamma)\, \delta'(\xi\cos\gamma + u\sin\gamma - x)\big|_{\gamma=0} = u\delta'(\xi - x) \end{aligned}$$

leading to

$$\left.\frac{\partial|F_\gamma(x)|^2}{\partial\gamma}\right|_{\gamma=0} = \int\int u\, W_f(\xi, u)\, \delta'(\xi - x)\, d\xi\, du$$
$$= -\frac{d}{dx}\left[\int u\, W_f(x, u)\, du\right]$$

and then substituting, from Eq. (1.29), $\int uW_f(x, u)\, du = |f(x)|^2\, d\phi(x)/dx$. By measuring two intensity profiles around $\gamma = 0$, $|F_{\gamma_o}(x)|^2$ and $|F_{-\gamma_o}(x)|^2$ for instance, approximating $\partial|F_\gamma(x)|^2/\partial\gamma$ by $[|F_{\gamma_o}(x)|^2 - |F_{-\gamma_o}(x)|^2]/2\gamma_o$, and integrating the result, we get $|f(x)|^2\, d\phi(x)/dx$. After dividing this by the intensity $|f(x)|^2 = |F_0(x)|^2$, which can be approximated by $[|F_{\gamma_o}(x)|^2 + |F_{-\gamma_o}(x)|^2]/2$, we find an approximation for the phase derivative $d\phi(x)/dx$, which after a second integration yields the phase $\phi(x)$. Together with the modulus $|f(x)|$, the signal $f(x)$ can thus be reconstructed. This procedure can be extended to other members of the class of Luneburg's first-order optical systems, to be considered next, in particular by using a section of free space instead of a fractional Fourier transformer.[57]

1.6 Propagation of the Wigner Distribution

In this section, we study how the Wigner distribution propagates through linear optical systems. We therefore consider an optical system as a black box, with an input plane and an output plane, and focus on the important class of first-order optical systems. A continuous medium, in which the signal must satisfy a certain differential equation, is considered in Sec. 1.6.5, but without going into much detail.

1.6.1 First-Order Optical Systems—Ray Transformation Matrix

An important class of optical systems is the class of Luneburg's first-order optical systems.[58] This class consists of a section of free space (in the Fresnel approximation), a thins lens, and all possible combinations of these. A first-order optical system can most easily be described in terms of its (normalized) ray transformation matrix[59]

$$\begin{bmatrix}\mathbf{r}_o\\ \mathbf{q}_o\end{bmatrix} = \begin{bmatrix}\mathbf{W} & \mathbf{0}\\ \mathbf{0} & \mathbf{W}^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{bmatrix}\begin{bmatrix}\mathbf{W}^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{W}\end{bmatrix}\begin{bmatrix}\mathbf{r}_i\\ \mathbf{q}_i\end{bmatrix} \tag{1.38}$$

which relates the position $\mathbf{r}_i$ and direction $\mathbf{q}_i$ of an incoming ray to the position $\mathbf{r}_o$ and direction $\mathbf{q}_o$ of the outgoing ray. In normalized

coordinates, $\mathbf{W}^{-1}\mathbf{r} =: \mathbf{r}$ and $\mathbf{W}\mathbf{q} =: \mathbf{q}$, we have

$$\begin{bmatrix} \mathbf{r}_o \\ \mathbf{q}_o \end{bmatrix} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{r}_i \\ \mathbf{q}_i \end{bmatrix} \tag{1.39}$$

We recall that the ray transformation matrix is symplectic. Using the matrix $\mathbf{J}$,

$$\mathbf{J} = i \begin{bmatrix} \mathbf{0} & -\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} = \mathbf{J}^{-1} = \mathbf{J}^{\dagger} = -\mathbf{J}^{t} \tag{1.40}$$

where $\mathbf{J}^{-1}$, $\mathbf{J}^{\dagger} = (\mathbf{J}^{*})^{t}$, and $\mathbf{J}^{t}$ are the inverse, the adjoint, and the transpose of $\mathbf{J}$, respectively, symplecticity can be elegantly expressed as $\mathbf{T}^{-1} = \mathbf{J}\mathbf{T}^{t}\mathbf{J}$. In detail we have

$$\mathbf{T}^{-1} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{D}^{t} & -\mathbf{B}^{t} \\ -\mathbf{C}^{t} & \mathbf{A}^{t} \end{bmatrix} = \mathbf{J}\mathbf{T}^{t}\mathbf{J} \tag{1.41}$$

If $\det \mathbf{B} \neq 0$, the coherent point-spread function of the first-order optical system reads

$$h(\mathbf{r}_o, \mathbf{r}_i) = (\det i\mathbf{B})^{-1/2} \exp\left[i\pi\left(\mathbf{r}_o^{t}\mathbf{D}\mathbf{B}^{-1}\mathbf{r}_o - 2\mathbf{r}_i^{t}\mathbf{B}^{-1}\mathbf{r}_o + \mathbf{r}_i^{t}\mathbf{B}^{-1}\mathbf{A}\mathbf{r}_i\right)\right] \tag{1.42}$$

see also Refs. 60 and 61. In the limiting case that $\mathbf{B} \to \mathbf{0}$, we have

$$h(\mathbf{r}_o, \mathbf{r}_i) = |\det \mathbf{A}|^{-1/2} \exp\left(i\pi\mathbf{r}_o^{t}\mathbf{C}\mathbf{A}^{-1}\mathbf{r}_o\right) \delta\left(\mathbf{r}_i - \mathbf{A}^{-1}\mathbf{r}_o\right) \tag{1.43}$$

In the degenerate case $\det \mathbf{B} = 0$ but $\mathbf{B} \neq \mathbf{0}$, a representation in terms of the coherent point-spread function can also be formulated.[62] The relationship between the input Wigner distribution $W_i(\mathbf{r}, \mathbf{q})$ and the output Wigner distribution $W_o(\mathbf{r}, \mathbf{q})$ takes the simple form

$$W_o(\mathbf{A}\mathbf{r} + \mathbf{B}\mathbf{q}, \mathbf{C}\mathbf{r} + \mathbf{D}\mathbf{q}) = W_i(\mathbf{r}, \mathbf{q}) \tag{1.44}$$

and this is independent of the possible degeneracy of submatrix $\mathbf{B}$.

1.6.2 Phase-Space Rotators—More Rotations in Phase Space

If the ray transformation matrix is not only symplectic but also orthogonal, $\mathbf{T}^{-1} = \mathbf{T}^{t}$, the system acts as a general phase-space rotator,[53] as we will see shortly. We then have $\mathbf{A} = \mathbf{D}$ and $\mathbf{B} = -\mathbf{C}$, and $\mathbf{U} = \mathbf{A} + i\mathbf{B}$ is a unitary matrix: $\mathbf{U}^{\dagger} = \mathbf{U}^{-1}$. We thus have

$$\mathbf{T} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ -\mathbf{B} & \mathbf{A} \end{bmatrix} \qquad \text{and} \qquad (\mathbf{A} - i\mathbf{B})^{t} = \mathbf{U}^{\dagger} = \mathbf{U}^{-1} = (\mathbf{A} + i\mathbf{B})^{-1} \tag{1.45}$$

and hence

$$W_o(\mathbf{Ar} + \mathbf{Bq}, -\mathbf{Br} + \mathbf{Aq}) = W_i(\mathbf{r}, \mathbf{q}) \tag{1.46}$$

In the one-dimensional case, such a system reduces to a fractional Fourier transformer ($\mathbf{A} = \cos\gamma$, $\mathbf{B} = \sin\gamma$); the extension to a higher-dimensional separable fractional Fourier transformer (with diagonal matrices $\mathbf{A}$ and $\mathbf{B}$, and different fractional angles for the different coordinates) is straightforward.

In the two-dimensional case, the three basic systems with an orthogonal ray transformation matrix are (1) the separable fractional Fourier transformer $\mathcal{F}(\gamma_x, \gamma_y)$, (2) the rotator $\mathcal{R}(\varphi)$, and (3) the gyrator $\mathcal{G}(\varphi)$, with unitary representations $\mathbf{U} = \mathbf{A} + i\mathbf{B}$ equal to

$$\begin{bmatrix} \exp(i\gamma_x) & 0 \\ 0 & \exp(i\gamma_y) \end{bmatrix} \quad \begin{bmatrix} \cos\varphi & \sin\varphi \\ -\sin\varphi & \cos\varphi \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \cos\varphi & i\sin\varphi \\ i\sin\varphi & \cos\varphi \end{bmatrix} \tag{1.47}$$

respectively. All three systems correspond to rotations in phase space, which justifies the name *phase-space rotators*!

From the many decompositions of a general phase-space rotator into the more basic ones, we mention $\mathcal{F}(\frac{1}{2}\gamma, -\frac{1}{2}\gamma)\,\mathcal{R}(\varphi)\,\mathcal{F}(-\frac{1}{2}\gamma, \frac{1}{2}\gamma)\,\mathcal{F}(\gamma_x, \gamma_y)$, which follows directly if we represent the unitary matrix as

$$\mathbf{U} = \begin{bmatrix} \exp(i\gamma_x)\cos\varphi & \exp[i(\gamma_y + \gamma)]\sin\varphi \\ -\exp[i(\gamma_x - \gamma)]\sin\varphi & \exp(i\gamma_y)\cos\varphi \end{bmatrix} \tag{1.48}$$

Note that we have the relationship $\mathcal{F}(\frac{1}{4}\pi, -\frac{1}{4}\pi)\,\mathcal{R}(\varphi)\,\mathcal{F}(-\frac{1}{4}\pi, \frac{1}{4}\pi) = \mathcal{G}(\varphi)$, which is just one of the many similarity-type relationships that exist between a rotator $\mathcal{R}(\alpha)$, a gyrator $\mathcal{G}(\beta)$, and an antisymmetric fractional Fourier transformer $\mathcal{F}(\gamma, -\gamma)$:

$$\mathcal{F}\big(\pm\tfrac{1}{4}\pi, \mp\tfrac{1}{4}\pi\big)\,\mathcal{G}(\pm\varphi)\,\mathcal{F}\big(\mp\tfrac{1}{4}\pi, \pm\tfrac{1}{4}\pi\big) = \mathcal{R}(-\varphi) \tag{1.49a}$$

$$\mathcal{F}\big(\pm\tfrac{1}{4}\pi, \mp\tfrac{1}{4}\pi\big)\,\mathcal{R}(\pm\varphi)\,\mathcal{F}\big(\mp\tfrac{1}{4}\pi, \pm\tfrac{1}{4}\pi\big) = \mathcal{G}(\varphi) \tag{1.49b}$$

$$\mathcal{R}\big(\pm\tfrac{1}{4}\pi\big)\,\mathcal{F}(\pm\varphi, \mp\varphi)\,\mathcal{R}\big(\mp\tfrac{1}{4}\pi\big) = \mathcal{G}(-\varphi) \tag{1.49c}$$

$$\mathcal{G}\big(\pm\tfrac{1}{4}\pi\big)\,\mathcal{F}(\pm\varphi, \mp\varphi)\,\mathcal{G}\big(\mp\tfrac{1}{4}\pi\big) = \mathcal{R}(\varphi) \tag{1.49d}$$

$$\mathcal{R}\big(\pm\tfrac{1}{4}\pi\big)\,\mathcal{G}(\pm\varphi)\,\mathcal{R}\big(\mp\tfrac{1}{4}\pi\big) = \mathcal{F}(\varphi, -\varphi) \tag{1.49e}$$

$$\mathcal{G}\big(\pm\tfrac{1}{4}\pi\big)\,\mathcal{R}(\pm\varphi)\,\mathcal{G}\big(\mp\tfrac{1}{4}\pi\big) = \mathcal{F}(-\varphi, \varphi) \tag{1.49f}$$

If we separate from $\mathbf{U}$ the scalar matrix $\mathbf{U}_f(\vartheta, \vartheta) = \exp(i\vartheta)\,\mathbf{I}$ with $\exp(2i\vartheta) = \det\mathbf{U}$, which matrix corresponds to a symmetric fractional Fourier transformer $\mathcal{F}(\vartheta, \vartheta)$, the remaining matrix is a so-called quaternion, and thus a 2×2 unitary matrix with unit determinant;

expressed in the form of Eq. (1.48), this would mean $\gamma_y = -\gamma_x$. Note that the matrices $\mathbf{U}_r(\alpha)$, $\mathbf{U}_g(\beta)$, and $\mathbf{U}_f(\gamma, -\gamma)$, corresponding to a rotator $\mathcal{R}(\alpha)$, a gyrator $\mathcal{G}(\beta)$, and an antisymmetric fractional Fourier transformer $\mathcal{F}(\gamma, -\gamma)$, respectively, are quaternions, and that every separable fractional Fourier transformer $\mathcal{F}(\gamma_x, \gamma_y)$ can be decomposed as $\mathcal{F}(\vartheta, \vartheta)\,\mathcal{F}(\gamma, -\gamma)$.

We easily verify—for instance, by expressing the unitary matrix $\mathbf{U}$ in the form of Eq. (1.48)—that the input-output relation for a phase-space rotator can be expressed in the form

$$\mathbf{r}_o - i\,\mathbf{q}_o = \mathbf{U}(\mathbf{r}_i - i\,\mathbf{q}_i) \tag{1.50}$$

which is an easy alternative for Eq. (1.39). Phase-space rotators are considered in greater detail in Chap. 3 by Tatiana Alieva.

1.6.3 More General Systems—Ray-Spread Function

First-order optical systems are a perfect match for the Wigner distribution, since their point-spread function is a quadratic-phase function. Nevertheless, an input-output relationship can always be formulated for the Wigner distribution. In the most general case, based on the relationships (1.5) and (1.9), we write

$$W_o(\mathbf{r}_o, \mathbf{q}_o) = \int\int K(\mathbf{r}_o, \mathbf{q}_o, \mathbf{r}_i, \mathbf{q}_i) W_i(\mathbf{r}_i, \mathbf{q}_i)\, d\mathbf{r}_i\, d\mathbf{q}_i \tag{1.51}$$

with

$$\begin{aligned} K(\mathbf{r}_o, \mathbf{q}_o, \mathbf{r}_i, \mathbf{q}_i) = \int\int & h\big(\mathbf{r}_o + \tfrac{1}{2}\mathbf{r}'_o, \mathbf{r}_i + \tfrac{1}{2}\mathbf{r}'_i\big) h^*\big(\mathbf{r}_o - \tfrac{1}{2}\mathbf{r}'_o, \mathbf{r}_i - \tfrac{1}{2}\mathbf{r}'_i\big) \\ & \times \exp\big[-i2\pi\big(\mathbf{q}_o^t\mathbf{r}'_o - \mathbf{q}_i^t\mathbf{r}'_i\big)\big]\, d\mathbf{r}'_o\, d\mathbf{r}'_i \end{aligned} \tag{1.52}$$

Relation (1.52) can be considered the definition of a double Wigner distribution; hence, the function K has all the properties of a Wigner distribution, for instance, the property of realness.

Let us think about the physical meaning of the function K. In a formal way, the function K is the response of the system in the space-frequency domain when the input signal is described by a product of two Dirac functions $W_i(\mathbf{r}, \mathbf{q}) = \delta(\mathbf{r}-\mathbf{r}_i)\delta(\mathbf{q}-\mathbf{q}_i)$; only in a formal way, since an actual input signal yielding such a Wigner distribution does not exist. Nevertheless, such an input signal could be considered as a single ray entering the system at the position $\mathbf{r}_i$ with direction $\mathbf{q}_i$. Hence, the function K might be called the ray-spread function of the system.

1.6.4 Geometric-Optical Systems

Let us start by studying a modulator, described—in the case of partially coherent light—by the input-output relationship $\Gamma_o(\mathbf{r}_1, \mathbf{r}_2) = m(\mathbf{r}_1)\,\Gamma_i(\mathbf{r}_1, \mathbf{r}_2)\,m^*(\mathbf{r}_2)$. The input and output Wigner distributions are related by

$$W_o(\mathbf{r}, \mathbf{q}) = \int W_m(\mathbf{r}, \mathbf{q} - \mathbf{q}_i)\, W_i(\mathbf{r}, \mathbf{q}_i)\, d\mathbf{q}_i \tag{1.53}$$

where $W_m(\mathbf{r}, \mathbf{q})$ is the Wigner distribution of the modulation function $m(\mathbf{r})$.

We now confine ourselves to the case of a pure phase modulation function $m(\mathbf{r}) = \exp[i2\pi\phi(\mathbf{r})]$. We then get

$$\begin{aligned} m\left(\mathbf{r} + \tfrac{1}{2}\mathbf{r}'\right) m^*\left(\mathbf{r} - \tfrac{1}{2}\mathbf{r}'\right) &= \exp\left\{i2\pi\left[\phi\left(\mathbf{r} + \tfrac{1}{2}\mathbf{r}'\right) - \phi\left(\mathbf{r} - \tfrac{1}{2}\mathbf{r}'\right)\right]\right\} \\ &= \exp\{i2\pi[(d\phi/d\mathbf{r})^t\mathbf{r}' + \text{higher-order terms}]\} \end{aligned} \tag{1.54}$$

If we consider only the first-order derivative in relation (1.54), we get $W_m(\mathbf{r}, \mathbf{q}) \simeq \delta(\mathbf{q} - d\phi/d\mathbf{r})$, and the input-output relationship of the pure phase modulator becomes $W_o(\mathbf{r}, \mathbf{q}) \simeq W_i(\mathbf{r}, \mathbf{q} - d\phi/d\mathbf{r})$, which is a mere coordinate transformation. We conclude that a single input ray yields a single output ray.

The ideas described above have been applied to the design of optical coordinate transformers[63,64] and to the theory of aberrations.[65] Now, if the first-order approximation is not sufficiently accurate, i.e., if we have to take into account higher-order derivatives in relation (1.54), the Wigner distribution allows us to overcome this problem. Indeed, we still have the exact input-output relationship (1.53), and we can take into account as many derivatives in relation (1.54) as necessary. We thus end up with a more general form[66] than $W_o(\mathbf{r}, \mathbf{q}) \simeq W_i(\mathbf{r}, \mathbf{q} - d\phi/d\mathbf{r})$. This will yield an Airy function instead of a Dirac function, for instance, when we take not only the first but also the third derivative into account.

We concluded that a single input ray yields a single output ray. This may also happen in more general—not just modulation-type—systems; we call such systems geometric-optical systems. These systems have the simple input-output relationship $W_o(\mathbf{r}, \mathbf{q}) \simeq W_i[\mathbf{g}_x(\mathbf{r}, \mathbf{q}), \mathbf{g}_u(\mathbf{r}, \mathbf{q})]$, where the $\simeq$ sign becomes an $=$ sign in the case of linear functions $\mathbf{g}_x$ and $\mathbf{g}_u$, i.e., in the case of Luneburg's first-order optical systems. There appears to be a close relationship to the description of such geometric-optical systems by means of the Hamilton characteristics.[6]

1.6.5 Transport Equations

With the tools of this section, we could study the propagation of the Wigner distribution through free space by considering a section of free space as an optical system with an input plane and an output plane. It is possible, however, to find the propagation of the Wigner distribution through free space directly from the differential equation that the signal must satisfy. We therefore let the longitudinal variable z enter into the formulas and remark that the propagation of coherent light in free space (at least in the Fresnel approximation) is governed by the differential equation (see, for instance, Ref. 15, p. 358)

$$-i\,\frac{\partial f}{\partial z} = \left(k + \frac{1}{2k}\,\frac{\partial^2}{\partial \mathbf{r}^2}\right) f \tag{1.55}$$

with $\partial^2/\partial \mathbf{r}^2$ representing the scalar operator $\partial^2/\partial x^2 + \partial^2/\partial y^2$ and with k the wave number. The propagation of the Wigner distribution is now described by a so-called transport equation[7,8,67–70] which in this case takes the form

$$\frac{2\pi \mathbf{q}^t}{k}\,\frac{\partial W}{\partial \mathbf{r}} + \frac{\partial W}{\partial z} = 0 \tag{1.56}$$

with $\partial/\partial \mathbf{r} = \nabla$. The transport equation (1.56) has the solution

$$W(\mathbf{r}, \mathbf{q}; z) = W\left(\mathbf{r} - \frac{2\pi \mathbf{q}}{k} z, \mathbf{q}; 0\right) \tag{1.57}$$

which is equivalent to the result Eq. (1.44) in Sec. 1.6.1, with the special choice $\mathbf{A} = \mathbf{D} = \mathbf{I}$.

In a weakly inhomogeneous medium, the optical signal must satisfy the Helmholtz equation

$$-i\,\frac{\partial f}{\partial z} = \sqrt{k^2(\mathbf{r}, z) + \frac{\partial^2}{\partial \mathbf{r}^2}}\, f \tag{1.58}$$

with $k = k(\mathbf{r}, z)$. In this case, we can again derive a transport equation for the Wigner distribution; the exact transport equation is rather complicated, but in the geometric-optical approximation, i.e., restricting ourselves to first-order derivatives, it takes the simple form

$$\frac{2\pi \mathbf{q}^t}{k}\,\frac{\partial W}{\partial \mathbf{r}} + \frac{\sqrt{k^2 - (2\pi)^2 \mathbf{q}^t \mathbf{q}}}{k}\,\frac{\partial W}{\partial z} + \left(\frac{\partial k}{2\pi\,\partial \mathbf{r}}\right)^t \frac{\partial W}{\partial \mathbf{q}} = 0 \tag{1.59}$$

which, in general, cannot be solved explicitly. With the method of characteristics, however, we conclude that along a path defined by

$$\frac{d\mathbf{r}}{ds} = \frac{2\pi \mathbf{q}}{k} \qquad \frac{dz}{ds} = \frac{\sqrt{k^2 - (2\pi)^2 \mathbf{q}^t \mathbf{q}}}{k} \qquad \frac{d\mathbf{q}}{ds} = \frac{\partial k}{2\pi\,\partial \mathbf{r}} \tag{1.60}$$

the Wigner distribution has a constant value. When we eliminate the frequency variable $\mathbf{q}$ from Eqs. (1.60), we are immediately led to

$$\frac{d}{ds}\left(k\,\frac{d\mathbf{r}}{ds}\right)=\frac{\partial k}{\partial \mathbf{r}} \qquad \frac{d}{ds}\left(k\,\frac{dz}{ds}\right)=\frac{\partial k}{\partial z} \tag{1.61}$$

which are the equations for an optical ray in geometrical optics.[71] We are thus led to the general conclusion that in the geometric-optical approximation, the Wigner distribution has a constant value along the geometric-optical ray paths, which conforms to our conclusions in Sec. 1.6.4: $W_o(\mathbf{r},\mathbf{q}) \simeq W_i[\mathbf{g}_x(\mathbf{r},\mathbf{q}),\mathbf{g}_u(\mathbf{r},\mathbf{q})]$. For a more detailed treatment of rays, see Chap. 8 by Miguel Alonso.

1.7 Wigner Distribution Moments in First-Order Optical Systems

The Wigner distribution moments provide valuable tools for the characterization of optical beams (see, for instance, Ref. 37). First-order moments, defined as

$$[m_x, m_y, m_u, m_v] = \frac{1}{E}\int\!\!\int\!\!\int\!\!\int [x, y, u, v]\, W(x, y, u, v)\, dx\, dy\, du\, dv \tag{1.62}$$

yield the position of the beam (m_x and m_y) and its direction (m_u and m_v). Second-order moments, defined by Eq. (1.25), give information about the spatial width of the beam (the shape m_{xx} and m_{yy} of the spatial ellipse and its orientation m_{xy}) and the angular width in which the beam is radiating (the shape m_{uu} and m_{vv} of the spatial-frequency ellipse and its orientation m_{uv}). Moreover, they provide information about its curvature (m_{xu} and m_{yv}) and its twist (m_{xv} and m_{yu}), with a possible definition of the twistedness as[46] $m_{yy}m_{xv} - m_{xx}m_{yu} + m_{xy}(m_{xu} - m_{yv})$. Many important beam characterizers, such as the overall beam quality[72]

$$\left(m_{xx}m_{uu} - m_{xu}^2\right) + \left(m_{yy}m_{vv} - m_{yv}^2\right) + 2(m_{xy}m_{uv} - m_{xv}m_{yu})$$

(see also Sec. 1.7.1), are based on second-order moments. Also the longitudinal component of the orbital angular momentum $\Lambda = \Lambda_a + \Lambda_v \propto (m_{xv} - m_{yu})$ [see Eq. (3) in Ref. 73] and its antisymmetrical part Λ_a and vortex part Λ_v,

$$\Lambda_a \propto \frac{(m_{xx} - m_{yy})(m_{xv} + m_{yu}) - 2m_{xy}(m_{xu} - m_{yv})}{m_{xx} + m_{yy}}$$

$$\Lambda_v \propto 2\,\frac{m_{yy}m_{xv} - m_{xx}m_{yu} + m_{xy}(m_{xu} - m_{yv})}{m_{xx} + m_{yy}}$$

[see Eqs. (22) and (21) in Ref. 73] are based on these moments.[74] Higher-order moments are used, for instance, to characterize the beam's symmetry and its sharpness.[37]

Because the Wigner distribution of a two-dimensional signal is a function of four variables, it is difficult to analyze. Therefore, the signal is often represented not by the Wigner distribution itself, but by its moments. Beam characterization based on the second-order moments of the Wigner distribution thus became the basis of an International Organization for Standardization standard.[75]

In this section we restrict ourselves mainly to second-order moments. The propagation of the matrix $\mathbf{M}$ of second-order moments of the Wigner distribution through a first-order optical system with ray transformation matrix $\mathbf{T}$ can be described by the input-output relationship[9,76] $\mathbf{M}_o = \mathbf{T}\mathbf{M}_i\mathbf{T}^t$. This relationship can be readily derived by combining the input-output relationship (1.39) of the first-order optical system with the definition (1.25) of the moment matrices of the input and the output signal. Since the ray transformation matrix $\mathbf{T}$ is symplectic, we immediately conclude that a possible symplecticity of the moment matrix (to be discussed later) is preserved in a first-order optical system: if $\mathbf{M}_i$ is proportional to a symplectic matrix, then $\mathbf{M}_o$ is proportional to a symplectic matrix as well, with the same proportionality factor.

1.7.1 Moment Invariants

If we multiply the moment relation $\mathbf{M}_o = \mathbf{T}\mathbf{M}_i\mathbf{T}^t$ from the right by $\mathbf{J}$, and use the symplecticity property (1.41) and the properties of $\mathbf{J}$, the input-output relationship can be written as[77] $\mathbf{M}_o\mathbf{J} = \mathbf{T}(\mathbf{M}_i\mathbf{J})\mathbf{T}^{-1}$. From the latter relationship we conclude that the matrices $\mathbf{M}_i\mathbf{J}$ and $\mathbf{M}_o\mathbf{J}$ are related to each other by a similarity transformation. As a consequence of this similarity transformation, and writing the matrix $\mathbf{MJ}$ in terms of its eigenvalues and eigenvectors according to $\mathbf{MJ} = \mathbf{S}\mathbf{\Lambda}\mathbf{S}^{-1}$, we can formulate the relationships $\mathbf{\Lambda}_o = \mathbf{\Lambda}_i$ and $\mathbf{S}_o = \mathbf{T}\mathbf{S}_i$. We are thus led to the important property[77] that the eigenvalues of the matrix $\mathbf{MJ}$ (and any combination of these eigenvalues) remain invariant under propagation through a first-order optical system, while the matrix of eigenvectors $\mathbf{S}$ transforms in the same way as the ray vector $[\mathbf{r}^t, \mathbf{q}^t]^t$ does.

It can be shown[77] that the eigenvalues of $\mathbf{MJ}$ are real. Moreover, if λ is an eigenvalue of $\mathbf{MJ}$, then $-\lambda$ is an eigenvalue, too; this implies that the characteristic polynomial $\det(\mathbf{MJ} - \lambda\mathbf{I})$, with the help of which we determine the eigenvalues, is a polynomial of λ^2. Indeed, the characteristic equation takes the form

$$\det(\mathbf{MJ} - \lambda\mathbf{I}) = 0 = \lambda^4 - a_2\lambda^2 + a_4$$

with $a_4 = \det \mathbf{M}$ and

$$a_2 = \left(m_{xx}m_{uu} - m_{xu}^2\right) + \left(m_{yy}m_{vv} - m_{yv}^2\right) + 2(m_{xy}m_{uv} - m_{xv}m_{yu})$$

Since the eigenvalues of **MJ** are invariants, the same holds for the coefficients of the characteristic equation. And since the characteristic equation is an equation in λ^2, we have only two such independent eigenvalues ($\pm\lambda_x$ and $\pm\lambda_y$, say) and thus only two independent invariants (such as λ_x and λ_y, or a_2 and a_4).

An interesting property follows from Williamson's theorem:[78,79] For any real, positive definite symmetric matrix **M**, there exists a real symplectic matrix $\mathbf{T}_o$ such that $\mathbf{M} = \mathbf{T}_o\boldsymbol{\Delta}_o\mathbf{T}_o^t$, where $\boldsymbol{\Delta}_o = \mathbf{T}_o^{-1}\mathbf{M}(\mathbf{T}_o^{-1})^t$ takes the normal form

$$\boldsymbol{\Delta}_o = \begin{bmatrix} \boldsymbol{\Lambda}_o & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Lambda}_o \end{bmatrix} \quad \text{with} \quad \boldsymbol{\Lambda}_o = \begin{bmatrix} \lambda_x & 0 \\ 0 & \lambda_y \end{bmatrix} \quad \text{and} \quad \lambda_x, \lambda_y > 0 \tag{1.63}$$

From the similarity transformation $\mathbf{MJ} = \mathbf{T}_o(\boldsymbol{\Delta}_o\mathbf{J})\mathbf{T}_o^{-1}$, we conclude that $\boldsymbol{\Delta}_o$ follows directly from the eigenvalues $\pm\lambda_x$ and $\pm\lambda_y$ of **MJ** and that $\mathbf{T}_o$ follows from the eigenvectors of $(\mathbf{MJ})^2$: $(\mathbf{MJ})^2\mathbf{T}_o = \mathbf{T}_o\boldsymbol{\Delta}_o^2$. Any moment matrix **M** can thus be brought into the diagonal form $\boldsymbol{\Delta}_o$ by means of a realizable first-order optical system with ray transformation matrix $\mathbf{T}_o^{-1}$.

1.7.2 Moment Invariants for Phase-Space Rotators

In the special case that we are dealing with a phase-space rotator, for which the ray transformation matrix satisfies the orthogonality relation $\mathbf{T}^{-1} = \mathbf{T}^t$, we have not only the similarity transformation $\mathbf{M}_o\mathbf{J} = \mathbf{T}(\mathbf{M}_i\mathbf{J})\mathbf{T}^{-1}$ but also the similarity transformation $\mathbf{M}_o = \mathbf{T}\mathbf{M}_i\mathbf{T}^{-1}$. The eigenvalues of **M** are now also invariants, and the same holds for the coefficients of the corresponding characteristic equation

$$\det(\mathbf{M} - \mu\mathbf{I}) = 0 = \mu^4 - b_1\mu^3 + b_2\mu^2 - b_3\mu + b_4$$

Since $b_4 = \det \mathbf{M}$ is already a known invariant ($= a_4$), this yields at most three new independent invariants.

Another way to find moment invariants for phase-space rotators is to consider the Hermitian matrix

$$\begin{aligned}\mathbf{M}' &= \frac{1}{E}\int\int(\mathbf{r}-i\mathbf{q})(\mathbf{r}-i\mathbf{q})^{\dagger}W(\mathbf{r},\mathbf{q})\,d\mathbf{r}\,d\mathbf{q}\\ &= \mathbf{M_{rr}}+\mathbf{M_{qq}}+i(\mathbf{M_{rq}}-\mathbf{M_{rq}^{t}})\\ &= \begin{bmatrix} m_{xx}+m_{uu} & m_{xy}+m_{uv}+i(m_{xv}-m_{yu})\\ m_{xy}+m_{uv}-i(m_{xv}-m_{yu}) & m_{yy}+m_{vv}\end{bmatrix}\\ &= \begin{bmatrix} Q_0+Q_1 & Q_2+iQ_3\\ Q_2-iQ_3 & Q_0-Q_1\end{bmatrix}\end{aligned} \tag{1.64}$$

and to use Eq. (1.50) to get the relation

$$\mathbf{M}_o' = \mathbf{U}\mathbf{M}_i'\mathbf{U}^{\dagger} = \mathbf{U}\mathbf{M}_i'\mathbf{U}^{-1} \tag{1.65}$$

which is again a similarity transformation. Note that the moments m_{xu} and m_{yv}, i.e., the diagonal entries of submatrix $\mathbf{M_{rq}}$, do not enter matrix $\mathbf{M}'$ and that we have introduced the four moment combinations Q_j $(j = 0, 1, 2, 3)$ as

$$Q_0 = \tfrac{1}{2}[(m_{xx}+m_{uu})+(m_{yy}+m_{vv})] \tag{1.66a}$$

$$Q_1 = \tfrac{1}{2}[(m_{xx}+m_{uu})-(m_{yy}+m_{vv})] \tag{1.66b}$$

$$Q_2 = m_{xy}+m_{uv} \tag{1.66c}$$

$$Q_3 = m_{xv}-m_{yu} \tag{1.66d}$$

The characteristic equation with which the eigenvalues of $\mathbf{M}'$ can be determined reads

$$\det(\mathbf{M}'-\nu\mathbf{I}) = 0 = \nu^2-2Q_0\nu+Q_0^2-Q^2 = (\nu-Q_0)^2-Q^2,$$

where we have also introduced

$$Q = \sqrt{Q_1^2+Q_2^2+Q_3^2} \tag{1.66e}$$

The eigenvalues are real and we can write $\nu_{1,2} = Q_0 \pm Q$. Since the eigenvalues are invariant, we immediately get that $\nu_1-\nu_2 = 2Q$ is an invariant,[80] and we also get the invariants $\nu_1+\nu_2 = 2Q_0 = b_1$, which is the trace of $\mathbf{M}'$ and of $\mathbf{M}$, and $\nu_1\nu_2 = Q_0^2-Q^2 = b_2-a_2$, which is the determinant of $\mathbf{M}'$. We remark that Q_3 corresponds to the longitudinal component of the orbital angular momentum of a paraxial beam propagating in the z direction. From the invariance of Q, we conclude that the three-dimensional vector $(Q_1, Q_2, Q_3) = (Q\cos\vartheta,\, Q\sin\vartheta\cos\gamma,\, Q\sin\vartheta\sin\gamma)$ lives on a sphere with radius Q.

It is not difficult to show now that $\mathbf{M}'$ can be represented in the general form

$$\mathbf{M}' = Q_0 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + Q \begin{bmatrix} \cos\vartheta & \exp(i\gamma)\sin\vartheta \\ \exp(-i\gamma)\sin\vartheta & -\cos\vartheta \end{bmatrix} \tag{1.67}$$

where the angles ϑ and γ follow from the relations $Q\cos\vartheta = Q_1$ (with $0 \le \vartheta \le \pi$) and $Q\exp(i\gamma)\sin\vartheta = Q_2 + iQ_3$.

A phase-space rotator will only change the values of the angles ϑ and γ, but does not change the invariants Q_0 and Q. To transform a diagonal matrix $\mathbf{M}'$ with diagonal entries $Q_0 + Q$ and $Q_0 - Q$ into the general form (1.67), we can use, for instance, the phase-space rotating system[81] $\mathcal{F}(\frac{1}{2}\gamma, -\frac{1}{2}\gamma)\,\mathcal{R}(-\frac{1}{2}\vartheta)\,\mathcal{F}(-\frac{1}{2}\gamma, \frac{1}{2}\gamma)$; see also Sec. 1.6.2 and Eq. (1.48). Moreover, from Eq. (1.65), we easily derive[80] that for a separable fractional Fourier transformer $\mathcal{F}(\gamma_x, \gamma_y)$, Q_1 is an invariant and $Q_2 + iQ_3$ undergoes a rotation-type transformation: $(Q_2 + iQ_3)_o = \exp[i(\gamma_x - \gamma_y)]\,(Q_2 + iQ_3)_i$. Similar properties hold for a gyrator $\mathcal{G}(\varphi)$, for which Q_2 is an invariant and $(Q_3 + iQ_1)_o = \exp(i2\varphi)\,(Q_3 + iQ_1)_i$, and for a rotator $\mathcal{R}(-\varphi)$, for which Q_3 is an invariant and $(Q_1 + iQ_2)_o = \exp(i2\varphi)\,(Q_1 + iQ_2)_i$.

1.7.3 Symplectic Moment Matrix—The Bilinear ABCD Law

If the moment matrix $\mathbf{M}$ is proportional to a symplectic matrix, it can be expressed in the form[77]

$$\mathbf{M} = m \begin{bmatrix} \mathbf{G}^{-1} & \mathbf{G}^{-1}\mathbf{H} \\ \mathbf{H}\mathbf{G}^{-1} & \mathbf{G} + \mathbf{H}\mathbf{G}^{-1}\mathbf{H} \end{bmatrix} \tag{1.68}$$

with m a positive scalar, $\mathbf{G}$ and $\mathbf{H}$ real symmetric 2×2 matrices, and $\mathbf{G}$ positive definite; the two positive eigenvalues of $\mathbf{MJ}$ are now equal to $+m$, and the two negative eigenvalues are equal to $-m$.

We recall that for a symplectic moment matrix, the input-output relation $\mathbf{M}_o = \mathbf{T}\mathbf{M}_i\mathbf{T}^t$ can be expressed equivalently in the form of the bilinear relationship

$$\mathbf{H}_o \pm i\mathbf{G}_o = [\mathbf{C} + \mathbf{D}(\mathbf{H}_i \pm i\mathbf{G}_i)][\mathbf{A} + \mathbf{B}(\mathbf{H}_i \pm i\mathbf{G}_i)]^{-1} \tag{1.69}$$

This bilinear relationship, together with the invariance of $\det \mathbf{M}$, completely describes the propagation of a symplectic matrix $\mathbf{M}$ through a first-order optical system. Note that the bilinear relationship (1.69) is identical to the **ABCD** law for spherical waves: for spherical waves we have $\mathbf{H}_o = [\mathbf{C} + \mathbf{D}\mathbf{H}_i][\mathbf{A} + \mathbf{B}\mathbf{H}_i]^{-1}$, and we have only replaced the (real) curvature matrix $\mathbf{H}$ by the (generally complex) matrix $\mathbf{H} \pm i\mathbf{G}$. We are thus led to the important result that if matrix $\mathbf{M}$ of second-order

moments is symplectic (up to a positive constant) as described in Eq. (1.68), its propagation through a first-order optical system is completely described by the invariance of this positive constant and the **ABCD** law (1.69).

1.7.4 Measurement of Moments

Several optical schemes to determine all 10 second-order moments have been described.[72,82–87] We mention in particular Ref. 87, which is based on a general scheme that also can be used for the determination of arbitrary higher-order moments μ_{pqrs} with

$$\mu_{pqrs}\, E = \iiiint W(x, y, u, v)\, x^p u^q y^r v^s\, dx\, dy\, du\, dv \qquad (p, q, r, s \geq 0) \tag{1.70}$$

Note that for $q = s = 0$ we have intensity moments

$$\begin{aligned} \mu_{p0r0}\, E &= \iiiint W(x, y, u, v) x^p y^r\, dx\, dy\, du\, dv \\ &= \iint x^p\, y^r\, \Gamma(x, x; y, y)\, dx\, dy \qquad (p, r \geq 0) \end{aligned} \tag{1.71}$$

which can easily be measured. The 10 second-order moments can be determined from the knowledge of the output intensities of four first-order optical systems, where one of them has to be anamorphic. For the determination of the 20 third-order moments, for instance, we thus find the need of using a total of six-first-order optical systems: four isotropic systems and two anamorphic systems. For the details of how to construct appropriate measuring schemes, we refer to Ref. 87.

1.8 Coherent Signals and the Cohen Class

The Wigner distribution belongs to a broad class of space-frequency functions known as the Cohen class.[30] Any function of this class is described by the general formula

$$\begin{aligned} C_f(\mathbf{r}, \mathbf{q}) = \iiint & f\left(\mathbf{r}_o + \tfrac{1}{2}\mathbf{r}'\right) f^*\left(\mathbf{r}_o - \tfrac{1}{2}\mathbf{r}'\right) k(\mathbf{r}, \mathbf{q}, \mathbf{r}', \mathbf{q}') \\ & \times \exp\left[-i2\pi\left(\mathbf{q}^t\mathbf{r}' - \mathbf{r}^t\mathbf{q}' + \mathbf{r}_o^t\mathbf{q}'\right)\right] d\mathbf{r}_o\, d\mathbf{r}'\, d\mathbf{q}' \end{aligned} \tag{1.72}$$

and the choice of the kernel $k(\mathbf{r}, \mathbf{q}, \mathbf{r}', \mathbf{q}')$ selects one particular function of the Cohen class. The Wigner distribution, for instance, arises for $k(\mathbf{r}, \mathbf{q}, \mathbf{r}', \mathbf{q}') = 1$, whereas $k(\mathbf{r}, \mathbf{q}, \mathbf{r}', \mathbf{q}') = \delta(\mathbf{r} - \mathbf{r}')\delta(\mathbf{q} - \mathbf{q}')$ yields the ambiguity function. In this chapter we restrict ourselves to the case that $k(\mathbf{r}, \mathbf{q}, \mathbf{r}', \mathbf{q}')$ does not depend on the space variable $\mathbf{r}$ and the

spatial-frequency variable $\mathbf{q}$, hence $k(\mathbf{r}, \mathbf{q}, \mathbf{r}', \mathbf{q}') = \bar{K}(\mathbf{r}', \mathbf{q}')$, in which case the resulting space-frequency distribution is shift-covariant (see Sec. 1.4.2).

1.8.1 Multicomponent Signals—Auto-Terms and Cross-Terms

The Wigner distribution, like the mutual coherence function and the cross-spectral density, is a bilinear signal representation. In the case of completely coherent light, however, we usually deal with a linear signal representation. Using a bilinear representation to describe coherent light thus yields cross-terms if the signal consists of multiple components. The two-component signal $f(\mathbf{r}) = f_1(\mathbf{r}) + f_2(\mathbf{r})$ yields the Wigner distribution

$$W_f(\mathbf{r}, \mathbf{q}) = W_{f_1}(\mathbf{r}, \mathbf{q}) + W_{f_2}(\mathbf{r}, \mathbf{q}) + 2\,\mathrm{Re}\left[\int f_1\left(\mathbf{r} + \tfrac{1}{2}\mathbf{r}'\right) f_2^*\left(\mathbf{r} - \tfrac{1}{2}\mathbf{r}'\right) \exp(-i2\pi\mathbf{q}^t\mathbf{r}')\, d\mathbf{r}'\right] \quad (1.73)$$

and we notice a cross-term in addition to the auto-terms $W_{f_1}(\mathbf{r}, \mathbf{q})$ and $W_{f_2}(\mathbf{r}, \mathbf{q})$. In the case of two point sources $\delta(\mathbf{r} - \mathbf{r}_1)$ and $\delta(\mathbf{r} - \mathbf{r}_2)$, for instance, the cross-term reads

$$2\delta\left[\mathbf{r} - \tfrac{1}{2}(\mathbf{r}_1 + \mathbf{r}_2)\right] \cos[2\pi(\mathbf{r}_1 - \mathbf{r}_2)^t\mathbf{q})]$$

It appears at the position $\frac{1}{2}(\mathbf{r}_1 + \mathbf{r}_2)$, i.e., in the middle between the two auto-terms $W_{f_1}(\mathbf{r}, \mathbf{q}) = \delta(\mathbf{r} - \mathbf{r}_1)$ and $W_{f_2}(\mathbf{r}, \mathbf{q}) = \delta(\mathbf{r} - \mathbf{r}_2)$, and is modulated in the $\mathbf{q}$ direction. We can get rid of this cross-term when we average the Wigner distribution with a kernel that is narrow in the $\mathbf{r}$ direction and broad in the $\mathbf{q}$ direction. We thus remove the cross-term without seriously disturbing the auto-terms.

The occurrence of cross-terms is also visible from the general condition[45,88]

$$\begin{aligned} &W_f\left(\mathbf{r} + \tfrac{1}{2}\mathbf{r}'', \mathbf{q} + \tfrac{1}{2}\mathbf{q}''\right) W_f\left(\mathbf{r} - \tfrac{1}{2}\mathbf{r}'', \mathbf{q} - \tfrac{1}{2}\mathbf{q}''\right) \\ &\quad = \int\int W_f\left(\mathbf{r} + \tfrac{1}{2}\mathbf{r}', \mathbf{q} + \tfrac{1}{2}\mathbf{q}'\right) W_f\left(\mathbf{r} - \tfrac{1}{2}\mathbf{r}', \mathbf{q} - \tfrac{1}{2}\mathbf{q}'\right) \\ &\qquad \times \exp[-i2\pi(\mathbf{q}''^t\mathbf{r}' - \mathbf{q}'^t\mathbf{r}'')]\, d\mathbf{r}'\, d\mathbf{q}' \end{aligned} \quad (1.74)$$

which, for $\mathbf{r}'' = \mathbf{q}'' = \mathbf{0}$, reduces to

$$W_f^2(\mathbf{r}, \mathbf{q}) = \int\int W_f\left(\mathbf{r} + \tfrac{1}{2}\mathbf{r}', \mathbf{q} + \tfrac{1}{2}\mathbf{q}'\right) W_f\left(\mathbf{r} - \tfrac{1}{2}\mathbf{r}', \mathbf{q} - \tfrac{1}{2}\mathbf{q}'\right) d\mathbf{r}'\, d\mathbf{q}' \quad (1.75)$$

From the latter equality we conclude that the value of the Wigner distribution at some phase-space point $(\mathbf{r}, \mathbf{q})$ is related to the values of all those pairs of points $(\mathbf{r} \pm \frac{1}{2}\mathbf{r}', \mathbf{q} \pm \frac{1}{2}\mathbf{q}')$ for which $(\mathbf{r}, \mathbf{q})$ is the midpoint. Using, as we generally do, the analytic signal $f(\mathbf{r})$ instead of the real signal $\frac{1}{2}[f(\mathbf{r}) + f^*(\mathbf{r})]$ avoids the cross-terms that otherwise would automatically appear around $\mathbf{q} = \mathbf{0}$.

The requirement of removing cross-terms without seriously affecting the auto-terms has led to the Cohen class of bilinear signal representations. All members $C_f(\mathbf{r}, \mathbf{q})$ of this class can be generated by a convolution (for both $\mathbf{r}$ and $\mathbf{q}$) of the Wigner distribution with an appropriate kernel $K(\mathbf{r}, \mathbf{q})$:

$$C_f(\mathbf{r}, \mathbf{q}) = K(\mathbf{r}, \mathbf{q}) \underset{\mathbf{r}}{*} \underset{\mathbf{q}}{*} W_f(\mathbf{r}, \mathbf{q})$$

$$= \int\int K(\mathbf{r} - \mathbf{r}_o, \mathbf{q} - \mathbf{q}_o)\, W_f(\mathbf{r}_o, \mathbf{q}_o)\, d\mathbf{r}_o\, d\mathbf{q}_o \qquad (1.76)$$

Note that a convolution keeps the important property of shift covariance! After Fourier transforming the latter equation, we are led to an equation in the "ambiguity domain," and the convolution becomes a product:

$$\bar{C}_f(\mathbf{r}', \mathbf{q}') = \bar{K}(\mathbf{r}', \mathbf{q}')\, A_f(\mathbf{r}', \mathbf{q}') \qquad (1.77)$$

with

$$\bar{C}_f(\mathbf{r}', \mathbf{q}') = \mathcal{F}[C_f(\mathbf{r}, \mathbf{q})](\mathbf{r}', \mathbf{q}') \qquad (1.78a)$$

$$A_f(\mathbf{r}', \mathbf{q}') = \mathcal{F}[W_f(\mathbf{r}, \mathbf{q})](\mathbf{r}', \mathbf{q}') \qquad (1.78b)$$

$$\bar{K}(\mathbf{r}', \mathbf{q}') = \mathcal{F}[K(\mathbf{r}, \mathbf{q})](\mathbf{r}', \mathbf{q}') \qquad (1.78c)$$

The product form (1.77) offers an easy way in the design of appropriate kernels.

Again, cf. Fig. 1.1, we position the different signal and kernel representations at the corners of a rectangle, see Fig. 1.4. For completeness we have also introduced the kernels $R(\mathbf{r}_1, \mathbf{r}_2)$ and $\bar{R}(\mathbf{q}_1, \mathbf{q}_2)$ that operate on the product $\Gamma_f(\mathbf{r}_1, \mathbf{r}_2) = f(\mathbf{r}_1) f^*(\mathbf{r}_2)$ and $\bar{\Gamma}_f(\mathbf{q}_1, \mathbf{q}_2) = \bar{f}(\mathbf{q}_1)\, \bar{f}^*(\mathbf{q}_2)$, respectively, by means of a convolution for $\mathbf{r}$ or $\mathbf{q}$. Again, we have Fourier transformations along the sides of the rectangle, and we readily see that the kernel $K(\mathbf{r}, \mathbf{q})$ is related to the kernels $R(\mathbf{r}_1, \mathbf{r}_2)$ and $\bar{R}(\mathbf{q}_1, \mathbf{q}_2)$ as

$$K(\mathbf{r}, \mathbf{q}) = \int R\big(\mathbf{r} + \tfrac{1}{2}\mathbf{r}', \mathbf{r} - \tfrac{1}{2}\mathbf{r}'\big) \exp(-i2\pi \mathbf{q}^t \mathbf{r}')\, d\mathbf{r}' \qquad (1.79a)$$

$$K(\mathbf{r}, \mathbf{q}) = \int \bar{R}\big(\mathbf{q} + \tfrac{1}{2}\mathbf{q}', \mathbf{q} - \tfrac{1}{2}\mathbf{q}'\big) \exp(i2\pi \mathbf{r}^t \mathbf{q}')\, d\mathbf{q}' \qquad (1.79b)$$

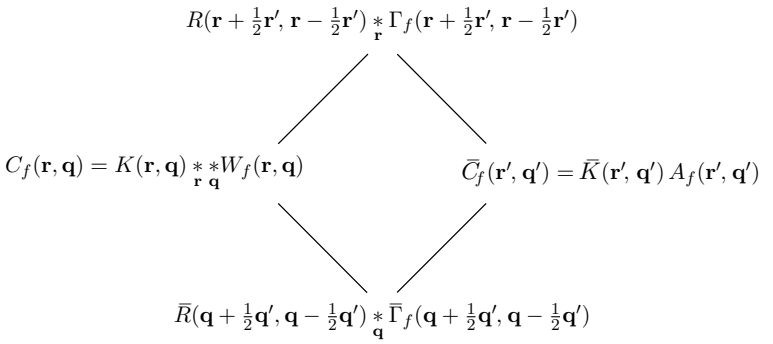


FIGURE 1.4 Schematic representation of the cross-spectral density Γ, its spatial Fourier transform $\bar{\Gamma}$, the Wigner distribution W, and the ambiguity function A, together with the corresponding kernels R, $\bar{R}$, K, and $\bar{K}$, on a rectangle.

As an example, we mention that the kernel $K(\mathbf{r}, \mathbf{q}) = \delta(\mathbf{r})\,\delta(\mathbf{q})$, for which $C_f(\mathbf{r}, \mathbf{q}) = W_f(\mathbf{r}, \mathbf{q})$ is the Wigner distribution, corresponds to the kernels $\bar{K}(\mathbf{r}', \mathbf{q}') = 1$, $R(\mathbf{r} + \frac{1}{2}\mathbf{r}, \mathbf{r}' - \frac{1}{2}\mathbf{r}) = \delta(\mathbf{r})$, and $\bar{R}(\mathbf{q} + \frac{1}{2}\mathbf{q}', \mathbf{q} - \frac{1}{2}\mathbf{q}') = \delta(\mathbf{q})$.

1.8.2 One-Dimensional Case and Some Basic Cohen Kernels

Many kernels have been proposed in the past, and some already existing bilinear signal representations have been identified as belonging to the Cohen class with an appropriately chosen kernel. Table 1.2 mentions some of them.[30,31,36]

In designing kernels, one may try to keep the interesting properties of the Wigner distribution; this reflects itself in conditions for the kernel. We recall that shift covariance is already maintained. To keep also the properties of realness, x marginal, and u marginal, for instance, the kernel $\bar{K}(x', u')$ should satisfy the conditions $\bar{K}(x', u') = \bar{K}^*(-x', -u')$, $\bar{K}(0, u') = 1$, and $\bar{K}(x', 0) = 1$, respectively. To keep the important property that for a signal $f(x) = |f(x)| \exp[i2\pi\phi(x)]$ the instantaneous frequency $d\phi/dx$ should follow from the bilinear representation through

$$\frac{d\phi}{dx} = \frac{\int u C_f(x, u)\, du}{\int C_f(x, u)\, du}$$

Bilinear Signal Representation	$\bar{K}(x', u')$
Wigner $W(x, u)$, Eq. (1.14)	1
Pseudo-Wigner $P(x, u; w)$, Eq. (1.17)	$w(\frac{1}{2}x')w^*(-\frac{1}{2}x')$
Page	$\exp(-i\pi u'\|x'\|)$
Kirkwood-Rihaczek	$\exp(-i\pi u'x')$
w-Rihaczek	$w(x')\exp(-i\pi u'x')$
Levin	$\cos(\pi u'x')$
w-Levin	$w(x')\cos(\pi u'x')$
Born-Jordan (sinc)	$\sin(\alpha\pi u'x')/\alpha\pi u'x'$
Zhao-Atlas-Marks (cone/ windowed sinc)	$w(x')\|\alpha x'\|\sin(\alpha\pi u'x')/\alpha\pi u'x'$
Choi-Williams (exponential)	$\exp[-(u'x')^2/\sigma]$
Generalized exponential	$\exp[-(u'/u_o)^{2N}]\exp[-(x'/x_o)^{2M}]$
Spectrogram $\|S(x, u; w)\|^2$, Eq. (1.86)	$A_w(-x', -u')$

TABLE 1.2 Kernels $\bar{K}(x', u')$ of Some Basic Cohen-Class Bilinear Signal Representations

as it does for the Wigner distribution, the kernel should satisfy the condition

$$\bar{K}(0, u') = \text{constant} \qquad \text{and} \qquad \left.\frac{\partial \bar{K}}{\partial x'}\right|_{x'=0} = 0$$

The Levin, Born-Jordan, and Choi-Williams representations clearly satisfy these conditions.

1.8.3 Rotation of the Kernel

In the case of two point sources $\delta(x - x_1)$ and $\delta(x - x_2)$, the cross-term

$$2\delta\left[x - \tfrac{1}{2}(x_1 + x_2)\right]\cos[2\pi(x_1 - x_2)u]$$

was located such that we needed averaging in the u direction when we wanted to remove it. In other cases, the cross-term may be located such that we need averaging in a different direction; for two plane waves $\exp(i2\pi u_1 x)$ and $\exp(i2\pi u_2 x)$, for instance, the cross-term reads

$$2\delta\left[u - \tfrac{1}{2}(u_1 + u_2)\right]\cos[2\pi(u_1 - u_2)x]$$

and we need averaging in the x direction. We may thus benefit from a rotation of the kernel, or let the original kernel operate on the Wigner distribution of the fractional Fourier transform of the signal,

$$C_f(x,u) = K(x\cos\gamma + u\sin\gamma, -x\sin\gamma + u\cos\gamma) \underset{x}{*}\underset{u}{*} W_f(x,u) \tag{1.80a}$$

$$C_{F_\gamma}(x,u) = K(x,u) \underset{x}{*}\underset{u}{*} W_{F_\gamma}(x,u) \tag{1.80b}$$

To find the optimal rotation angle γ_o, we may proceed as follows. Let m_x^γ and m_{xx}^γ be the first- and second-order moments of the intensity $|F_\gamma(x)|^2$ of the fractional Fourier transform $F_\gamma(x)$,

$$m_x^\gamma = \frac{1}{E}\int\int x W_{F_\gamma}(x,u)\,dx\,du = \frac{1}{E}\int x\,|F_\gamma(x)|^2\,dx \tag{1.81a}$$

$$m_{xx}^\gamma = \frac{1}{E}\int\int x^2 W_{F_\gamma}(x,u)\,dx\,du = \frac{1}{E}\int x^2\,|F_\gamma(x)|^2\,dx \tag{1.81b}$$

and let m_{xu}^γ be the mixed moment

$$m_{xu}^\gamma = \frac{1}{E}\int\int xu\, W_{F_\gamma}(x,u)\,dx\,du \tag{1.81c}$$

The propagation laws for the first- and second-order moments through a rotator read

$$\begin{bmatrix} m_x^\gamma \\ m_u^\gamma \end{bmatrix} = \begin{bmatrix} \cos\gamma & \sin\gamma \\ -\sin\gamma & \cos\gamma \end{bmatrix} \begin{bmatrix} m_x \\ m_u \end{bmatrix} \tag{1.82a}$$

$$\begin{bmatrix} m_{xx}^\gamma & m_{xu}^\gamma \\ m_{xu}^\gamma & m_{uu}^\gamma \end{bmatrix} = \begin{bmatrix} \cos\gamma & \sin\gamma \\ -\sin\gamma & \cos\gamma \end{bmatrix} \begin{bmatrix} m_{xx} & m_{xu} \\ m_{xu} & m_{uu} \end{bmatrix} \begin{bmatrix} \cos\gamma & -\sin\gamma \\ \sin\gamma & \cos\gamma \end{bmatrix} \tag{1.82b}$$

Note that $m_u = m_x^{\pi/2}$, $m_{uu} = m_{xx}^{\pi/2}$, and $m_{xu} = m_{xx}^{\pi/4} - \frac{1}{2}(m_{xx} + m_{xx}^{\pi/2})$, and that all second-order moments follow directly from the measurement of the intensity profiles of only three fractional Fourier transforms: $F_0(x) = f(x)$, $F_{\pi/2}(x) = \bar{f}(x)$, and $F_{\pi/4}(x)$. While the second-order moment m_{xx}^γ can be expressed as

$$m_{xx}^\gamma = m_{xx}\cos^2\gamma + m_{uu}\sin^2\gamma + m_{xu}\sin 2\gamma \tag{1.83}$$

the second-order central moment $\mu_{xx}^\gamma = m_{xx}^\gamma - (m_x^\gamma)^2$ can be expressed as

$$\mu_{xx}^\gamma = \mu_{xx}\cos^2\gamma + \mu_{uu}\sin^2\gamma + (m_{xu} - m_x m_u)\sin 2\gamma \tag{1.84}$$

and extremum values of μ_{xx}^{γ} arise for the angle γ_o, defined by

$$\tan 2\gamma_o = 2\,\frac{m_{xu} - m_x m_u}{\mu_{xx} - \mu_{uu}} = 2\,\frac{m_{xx}^{\pi/4} - \frac{1}{2}\left(m_{xx}^{0} + m_{xx}^{\pi/2}\right) - m_x^0 m_x^{\pi/2}}{m_{xx}^{0} - m_{xx}^{\pi/2} - \left(m_x^0\right)^2 + \left(m_x^{\pi/2}\right)^2} \tag{1.85}$$

Note that γ_o corresponds to the minimum value of μ_{xx}^{γ}, if γ_o is chosen such that $\cos 2\gamma_o$ has the same sign as $\mu_{xx}^{\pi/2} - \mu_{xx}^{0}$; then $\gamma_o + \frac{1}{2}\pi$ corresponds to the maximum value of μ_{xx}^{γ}. The angles γ_o and $\gamma_o + \frac{1}{2}\pi$ determine the principal axes of the moment ellipse in phase space. Kernels can be optimized by rotating them and aligning them to these principal axes.[89]

1.8.4 Rotated Version of the Smoothed Interferogram

We will apply the aligning of the kernel to the smoothed interferogram, which can be best derived from the pseudo-Wigner distribution. With

$$S_f(x, u; w) = \int f(x + x_o) w^*(x_o)\, \exp(-i2\pi u x_o)\, dx_o \tag{1.86}$$

denoting the windowed Fourier transform, the pseudo-Wigner distribution $P_f(x, u; w)$, i.e., the Wigner distribution with the additional window $w(\frac{1}{2}x')w^*(-\frac{1}{2}x')$ in its defining integral [see Eq. (1.17)] can also be represented as

$$P_f(x, u; w) = \int S_f\left(x, u + \tfrac{1}{2}t; w\right) S_f^*\left(x, u - \tfrac{1}{2}t; w\right) dt \tag{1.87}$$

The smoothed interferogram, also known as the S method, is now defined as[90]

$$P_f(x, u; w, z) = \int S_f\left(x, u + \tfrac{1}{2}t; w\right) z(t) S_f^*\left(x, u - \tfrac{1}{2}t; w\right) dt \tag{1.88}$$

It is based on the pseudo-Wigner distribution written in the form (1.87), but with an additional smoothing window $z(t)$ in the u direction. The resulting distribution is of the Wigner distribution form, with significantly reduced cross-terms of multicomponent signals, while the auto-terms are close to those in the pseudo-Wigner distribution. For $z(t) = \delta(t)$, the bilinear representation $P_f(x, u; w, z) = |S_f(x, u; w)|^2$ is known as the *spectrogram*: the squared modulus of the windowed Fourier transform. For $z(t) = 1$, $P_f(x, u; w, z)$ reduces to the pseudo-Wigner distribution (1.87).

Since the window $z(t)$ controls the behavior of $P_f(x, u; w, z)$—more Wigner-type or more spectrogram-type—we spend one paragraph on

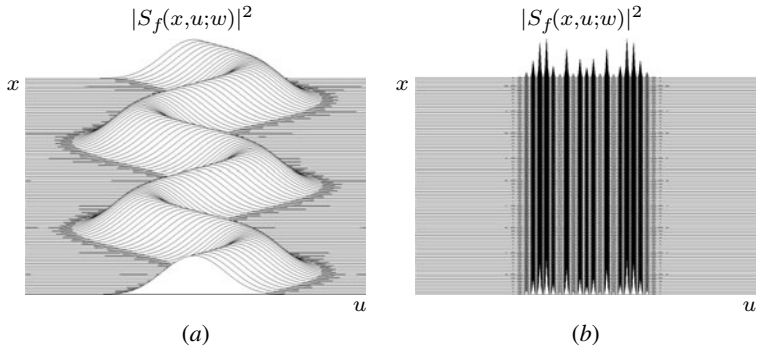


FIGURE 1.5 Spectrogram of a sinusoidal FM signal $\exp\{i[2\pi u_0 x + a_1 \sin(2\pi u_1 x)]\}$ with (*a*) a medium-sized window, leading to a space-frequency representation with smearing, and (*b*) a long window, leading to a pure frequency representation.

the spectrogram. Although the spectrogram is a quadratic signal representation $|S_f(x, u; w)|^2$, the squaring is introduced only in the final step and therefore does not lead to undesirable cross-terms that are present in other bilinear signal representations. This freedom from artifacts, together with simplicity, robustness, and ease of interpretation, has made the spectrogram a popular tool for speech analysis since its invention in 1946.[91] The price that has to be paid, however, is that the auto-terms are smeared by the window $w(x)$. Note that for $w(x) = \delta(x)$, the spectrogram yields the pure space representation $|S_f(x, u; w)|^2 = |f(x)|^2$, whereas for $w(x) = 1$, it yields the pure frequency representation $|S_f(x, u; w)|^2 = |\bar{f}(u)|^2$. This is illustrated in Fig. 1.5 on the sinusoidal FM signal

$$\exp\{i[2\pi u_0 x + a_1 \sin(2\pi u_1 x)]\}$$

and a rectangular window $w(x) = \text{rect}(x/X)$ of variable width X. Note in particular the smearing that appears in Fig. 1.5*a*.

Based on Eq. (1.88), but replacing the signal $f(x)$ by its fractional Fourier transform $F_\gamma(x)$, the γ-rotated version $P_f^\gamma(x, u; w, z)$ of the smoothed interferogram $P_f(x, u; w, z)$ was defined subsequently as[89,92]

$$\begin{aligned} P_f^\gamma(x, u; w, z) &= P_{F_\gamma}(x, u; w, z) \\ &= \int S_{F_\gamma}\left(x, u + \tfrac{1}{2}t; w\right) z(t)\, S^*_{F_\gamma}\left(x, u - \tfrac{1}{2}t; w\right) dt \end{aligned} \quad (1.89)$$

A definition directly in terms of the signal $f(x)$ reads

$$P_f^{\gamma}(x, u; w, z) = \int S_f\left(x + \tfrac{1}{2}t\sin\gamma,\, u + \tfrac{1}{2}t\cos\gamma;\, W_{-\gamma}\right) z(t) \times \exp(-i2\pi u t \sin\gamma) \times S_f^*\left(x - \tfrac{1}{2}t\sin\gamma,\, u - \tfrac{1}{2}t\cos\gamma;\, W_{-\gamma}\right) dt \quad (1.90)$$

where the fractional Fourier transform $W_{-\gamma}(x)$ of the window $w(x)$ arises and where we have used the relationship

$$S_{F_\gamma}(x_2, u_2; W_\gamma) = \exp[i\pi(u_2 x_2 - u_1 x_1)]\, S_f(x_1, u_1; w)$$
$$\text{with} \quad \begin{bmatrix} x_2 \\ u_2 \end{bmatrix} = \begin{bmatrix} \cos\gamma & \sin\gamma \\ -\sin\gamma & \cos\gamma \end{bmatrix} \begin{bmatrix} x_1 \\ u_1 \end{bmatrix} \quad (1.91)$$

The γ-rotated smoothed interferogram $P_f^{\gamma}(x, u; w, z)$ is related to the Wigner distribution $W_f(x, u)$ with the kernels[89,92]

$$K(x, u) = W_w(-x, -u)\bar{z}(-x\cos\gamma + u\sin\gamma) \quad (1.92a)$$

$$\bar{K}(x', u') = \int A_w(-x' + t\sin\gamma, -u' + t\cos\gamma) z(t)\, dt \quad (1.92b)$$

Note that for $\gamma = 0$, the distribution $P_f^{\gamma}(x, u; w, z)$ reduces to the one originally introduced, which was based on a combination of windowed Fourier transforms in the u direction, while for $\gamma = \frac{1}{2}\pi$ it reduces to the version that combines these windowed Fourier transforms in the x direction.[90]

The rotated version of the smoothed interferogram is a versatile method to remove cross-terms. To illustrate this, we show two numerical examples. Consider first the signal

$$f(x) = \exp\left[-\left(\frac{3x}{x_o}\right)^8\right] \{\exp[i\phi_1(x)] + \exp[i\phi_2(x)]\}$$
$$\text{with} \quad \begin{aligned} \phi_1(x) &= \pi h_1 x^2 + a_1 \cos(2\pi u_1 x) \\ \phi_2(x) &= \pi h_2 x^2 + a_2 \cos(2\pi u_2 x) \end{aligned}$$

consisting of two components with instantaneous frequency $h_1 x - a_1 u_1 \sin(2\pi u_1 x)$ and $h_2 x - a_2 u_2 \sin(2\pi u_2 x)$, respectively; note that the instantaneous frequencies cross at $x = 0$. In the numerical simulation, the variables take the values $x_o = 128$, $h_1 = 192$, $h_2 = 64$, $u_1 = u_2 = 2$, $-a_1 = a_2 = 8/\pi$. A Hann(ing) window $w(x) = \cos^2(\pi x/X)\ \mathrm{rect}(x/X)$ with width $X = 128$ is used for the calculation of the windowed Fourier transform $S_f(x, u; w)$. The values of the normalized second-order central moments are $\mu_{xx}^0 = 1$, $\mu_{xx}^{\pi/2} = 1.38$, and $\mu_{xx}^{\pi/4} = 0.07$.

According to Eq. (1.85), and using the fact that $\mu_{xx}^{\pi/2} - \mu_{xx}^{0} > 0$, we get $\gamma_o = 41^\circ$. The second-order moment in this direction, $\mu_{xx}^{41^\circ} = 0.057$, is smaller than in any other direction, while the second-order moment in the orthogonal direction, $\mu_{xx}^{-49^\circ} = 2.01$, is the largest. The fractional Fourier transform $F_\gamma(x)$ of the signal $f(x)$ for the angle $\gamma = \gamma_o - \frac{1}{2}\pi = -49^\circ$ can now be calculated by using a discrete fractional Fourier transformation algorithm. The next step is to calculate the windowed Fourier transform $S_{F_\gamma}(x, u; w)$ of the fractional Fourier transform $F_\gamma(x)$ and to use it in Eq. (1.89).

The results of this analysis are presented in Fig. 1.6. The pseudo-Wigner distribution $P_f(x, u; w)$ is shown in Fig. 1.6*a*. The smoothed interferogram $P_f(x, u; w, z)$, calculated by the standard definition (1.88), i.e., combining terms along the u axis, with a rectangular window $z(t) = \text{rect}(t/T)$ and $T = 15$, is presented in Fig. 1.6*b*. We see that some cross-terms already appear, although the auto-terms are still very different from those in the Wigner distribution in Fig. 1.6*a*. The reason lies in the very significant spread of one component along the

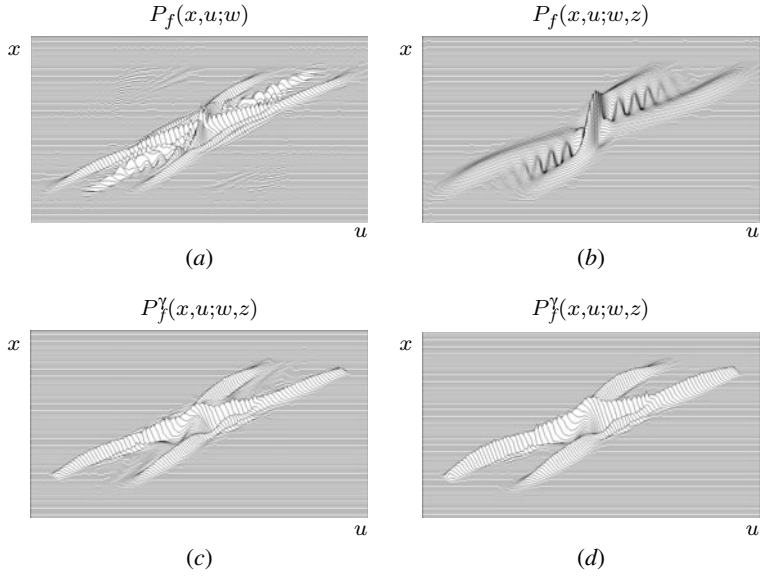


FIGURE 1.6 (*a*) Pseudo-Wigner distribution $P_f(x, u; w)$ of the signal $f(x)$; (*b*) smoothed interferogram $P_f(x, u; w, z)$ calculated in the frequency domain, with a rectangular window z; (*c*) smoothed interferogram $P_f^\gamma(x, u; w, z)$ calculated in the optimal frequency domain, with a rectangular window z; (*d*) smoothed interferogram $P_f^\gamma(x, u; w, z)$ calculated in the optimal frequency domain, with a Hann(ing) window z.

u axis. The γ-rotated smoothed interferogram $P_f^{\gamma}(x, u; w, z)$ for the optimal fractional angle $\gamma = -49^{\circ}$ is presented in Fig. 1.6*c* for a rectangular window with $T = 9$ and in Fig. 1.6*d* for a Hann(ing) window with $T = 15$. We can see that as a consequence of the high concentration of the components along the optimal fractional angle, we almost achieved the goal of getting the auto-terms of the Wigner distribution without any cross-terms.

Similar results are obtained for the signal

$$f(x) = \exp\left[-\left(\frac{3x}{x_o}\right)^8\right](\exp\{i[\phi(x) + 50\pi x]\} + \exp\{i[\phi(x) - 50\pi x]\})$$

$$\text{with} \qquad \phi(x) = \int_{-\infty}^{x} 15\,\pi\ \text{arcsinh}(100\,\xi)\,d\xi,$$

where $x_o = X = 128$ again and $T = 21$ see Fig. 1.7.

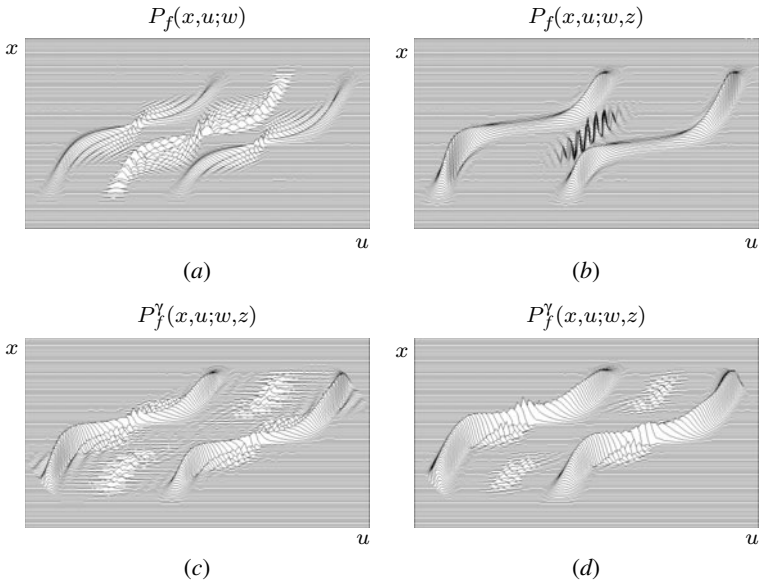


FIGURE 1.7 (*a*) Pseudo-Wigner distribution $P_f(x, u; w)$ of the signal $f(x)$; (*b*) smoothed interferogram $P_f(x, u; w, z)$ calculated in the frequency domain, with a rectangular window z; (*c*) smoothed interferogram $P_f^{\gamma}(x, u; w, z)$ calculated in the optimal frequency domain, with a rectangular window z; (*d*) smoothed interferogram $P_f^{\gamma}(x, u; w, z)$ calculated in the optimal frequency domain, with a Hann(ing) window z.

1.9 Conclusion

We have presented an overview of the Wigner distribution and of some of its properties and applications in an optical context. The Wigner distribution describes a signal in space (i.e., position) and spatial frequency (i.e., direction) simultaneously and can be considered as the local frequency spectrum of the signal, like the score in music and the phase space in mechanics. Although it is derived in terms of Fourier optics, the description of a signal by means of its Wigner distribution closely resembles the ray concept in geometrical optics. It thus presents a link between Fourier optics and geometrical optics. Moreover, the concept of the Wigner distribution is not restricted to deterministic signals (i.e., completely coherent light); it can be applied to stochastic signals (i.e., partially coherent light) as well, thus presenting a link between partial coherence and radiometry.

Properties of the Wigner distribution and its propagation through linear systems have been considered; the corresponding description of signals and systems can be directly interpreted in geometric-optical terms. For first-order optical systems, the propagation of the Wigner distribution is completely determined by the system's ray transformation matrix, thus presenting a strong interconnection with matrix optics.

We have studied the second-order moments of the Wigner distribution and some interesting combinations of these moments, together with the propagation of these moment combinations through first-order optical systems. Special attention has been paid to systems that perform rotations in phase space.

In the case of completely coherent light, the Wigner distribution is a member of a broad class of bilinear signal representations, known as the Cohen class. Each member of this class is related to the Wigner distribution by means of a convolution with a certain kernel. Because of the quadratic nature of such signal representations, they suffer from unwanted cross-terms, which one tries to minimize by a proper choice of this kernel. Some members of the Cohen class have been reviewed, and special attention was devoted to the smoothed interferogram in combination with the optimal angle in phase space in which the smoothing takes place.

References

1. E. Wigner, "On the quantum correction for thermodynamic equilibrium," *Phys. Rev.* **40**, 749–759 (1932).
2. L. S. Dolin, "Beam description of weakly-inhomogeneous wave fields," *Izv. Vyssh. Uchebn. Zaved. Radiofiz.* **7**, 559–563 (1964).

3. A. Walther, "Radiometry and coherence," *J. Opt. Soc. Am.* **58**, 1256–1259 (1968).
4. A. Walther, "Propagation of the generalized radiance through lenses," *J. Opt. Soc. Am.* **68**, 1606–1610 (1978).
5. M. J. Bastiaans, "The Wigner distribution function applied to optical signals and systems," *Opt. Commun.* **25**, 26–30 (1978).
6. M. J. Bastiaans, "The Wigner distribution function and Hamilton's characteristics of a geometric-optical system," *Opt. Commun.* **30**, 321–326 (1979).
7. M. J. Bastiaans, "Transport equations for the Wigner distribution function," *Opt. Acta* **26**, 1265–1272 (1979).
8. M. J. Bastiaans, "Transport equations for the Wigner distribution function in an inhomogeneous and dispersive medium," *Opt. Acta* **26**, 1333–1344 (1979).
9. M. J. Bastiaans, "Wigner distribution function and its application to first-order optics," *J. Opt. Soc. Am.* **69**, 1710–1716 (1979).
10. M. J. Bastiaans, "The Wigner distribution function of partially coherent light," *Opt. Acta* **28**, 1215–1224 (1981).
11. M. J. Bastiaans, "Application of the Wigner distribution function to partially coherent light," *J. Opt. Soc. Am. A* **3**, 1227–1238 (1986).
12. J. W. Goodman, *Introduction to Fourier Optics, 2nd ed.*, McGraw-Hill, New York, 1996.
13. E. Wolf, "A macroscopic theory of interference and diffraction of light from finite sources. I. Fields with a narrow spectral range," *Proc. R. Soc. London Ser. A* **225**, 96–111 (1954).
14. E. Wolf, "A macroscopic theory of interference and diffraction of light from finite sources. II. Fields with a spectral range of arbitrary width," *Proc. R. Soc. London Ser. A* **230**, 246–265 (1955).
15. A. Papoulis, *Systems and Transforms with Applications in Optics*, McGraw-Hill, New York, 1968.
16. M. J. Bastiaans, "A frequency-domain treatment of partial coherence," *Opt. Acta* **24**, 261–274 (1977).
17. L. Mandel and E. Wolf, "Spectral coherence and the concept of cross-spectral purity," *J. Opt. Soc. Am.* **66**, 529–535 (1976).
18. P. M. Woodward, *Probability and Information Theory with Applications to Radar*, Pergamon, London, 1953, Chap. 7.
19. A. Papoulis, "Ambiguity function in Fourier optics," *J. Opt. Soc. Am.* **64**, 779–788 (1974).
20. R. Simon, E. C. G. Sudarshan, and N. Mukunda, "Anisotropic Gaussian Schell-model beams: Passage through optical systems and associated invariants," *Phys. Rev. A* **31**, 2419–2434 (1985).
21. M. J. Bastiaans, "ABCD law for partially coherent Gaussian light, propagating through first-order optical systems," *Opt. Quant. Electron.* **24**, 1011–1019 (1992).
22. R. Simon and N. Mukunda, "Twisted Gaussian Schell-model beams," *J. Opt. Soc. Am. A* **10**, 95–109 (1993).
23. R. Simon, K. Sundar, and N. Mukunda, "Twisted Gaussian Schell-model beams. I. Symmetry structure and normal-mode spectrum," *J. Opt. Soc. Am. A* **10**, 2008–2016 (1993).
24. K. Sundar, R. Simon, and N. Mukunda, "Twisted Gaussian Schell-model beams. II. Spectrum analysis and propagation characteristics," *J. Opt. Soc. Am. A* **10**, 2017–2023 (1993).
25. A. T. Friberg, B. Tervonen, and J. Turunen, "Interpretation and experimental demonstration of twisted Gaussian Schell-model beams," *J. Opt. Soc. Am. A* **11**, 1818–1826 (1994).
26. D. Ambrosini, V. Bagini, F. Gori, and M. Santarsiero, "Twisted Gaussian Schell-model beams: A superposition model," *J. Mod. Opt.* **41**, 1391–1399 (1994).
27. A. C. Schell, "A technique for the determination of the radiation pattern of a partially coherent aperture," *IEEE Trans. Antennas Propag.* **AP-15**, 187–188 (1967).

28. F. Gori, "Collett-Wolf sources and multimode lasers," *Opt. Commun.* **34**, 301–305 (1980).
29. T. A. C. M. Claasen and W. F. G. Mecklenbräuker, "The Wigner Distribution—A tool for time-frequency signal analysis; Part 1: Continuous-time signals," *Philips J. Res.* **35**, 217–250 (1980).
30. L. Cohen, "Time-frequency distributions—A review," *Proc. IEEE* **77**, 941–981 (1989).
31. F. Hlawatsch and G. F. Boudreaux-Bartels, "Linear and quadratic time-frequency signal representations," *IEEE Signal Processing Magazine* **9** (2), 21–67 (1992).
32. L. Cohen, *Time-Frequency Analysis*, Prentice-Hall, Englewood Cliffs, N.J., 1995.
33. H. W. Lee, "Theory and applications of the quantum phase-space distribution functions," *Phys. Rep.* **259**, 147–211 (1995).
34. W. Mecklenbräuker and F. Hlawatsch (eds.), *The Wigner Distribution—Theory and Applications in Signal Processing*, Elsevier Science, Amsterdam, 1997.
35. D. Dragoman, "The Wigner distribution function in optics and optoelectronics," in E. Wolf (ed.), *Progress in Optics*, Vol. 37, North-Holland, Amsterdam, 1997, pp. 1–56.
36. B. Boashash (ed.), "*Time-Frequency Signal Analysis and Processing: A Comprehensive Reference*," Elsevier, Oxford, UK, 2003; in particular, Part 1: "Introduction to the concepts of TFSAP."
37. D. Dragoman, "Applications of the Wigner distribution function in signal processing," *EURASIP J. Appl. Signal Process.* **2005**, 1520–1534 (2005).
38. A. Torre, *Linear Ray and Wave Optics in Phase Space*, Elsevier, Amsterdam, 2005.
39. M. E. Testorf, J. Ojeda-Castañeda, and A. W. Lohmann (eds.), *Selected Papers on Phase-Space Optics, SPIE Milestone Series*, vol. MS 181, SPIE, Bellingham, Wash., 2006.
40. M. J. Bastiaans, "Applications of the Wigner distribution to partially coherent light beams," in A. Friberg and R. Dändliker (eds.), *Advances in Information Optics and Photonics*, SPIE, Bellingham, Wash., 2008, pp. 27–56.
41. A. T. Friberg, "On the existence of a radiance function for finite planar sources of arbitrary states of coherence," *J. Opt. Soc. Am.* **69**, 192–198 (1979).
42. W. H. Carter and E. Wolf, "Coherence and radiometry with quasi-homogeneous planar sources," *J. Opt. Soc. Am.* **67**, 785–796 (1977).
43. E. Wolf, "Coherence and radiometry," *J. Opt. Soc. Am.* **68**, 6–17 (1978).
44. M. J. Bastiaans, "Wigner distribution function applied to partially coherent light," in P. M. Mejías, H. Weber, R. Martínez-Herrero, and A. González-Ureña (eds.), *Proceedings of the Workshop on Laser Beam Characterization*, SEDO, Madrid, 1993, pp. 65–87.
45. M. J. Bastiaans, "Application of the Wigner distribution function in optics," in W. Mecklenbräuker and F. Hlawatsch (eds.), *The Wigner Distribution—Theory and Applications in Signal Processing*, Elsevier Science, Amsterdam, 1997, pp. 375–426.
46. M. J. Bastiaans, "Wigner distribution function applied to twisted Gaussian light propagating in first-order optical systems," *J. Opt. Soc. Am. A* **17**, 2475–2480 (2000).
47. R. Winston and W. T. Welford, "Geometrical vector flux and some new nonimaging concentrators," *J. Opt. Soc. Am.* **69**, 532–536 (1979).
48. J. E. Moyal, "Quantum mechanics as a statistical theory," *Proc. Cambridge Philos. Soc.* **45**, 99–132 (1949).
49. A. C. McBride and F. H. Kerr, "On Namias' fractional Fourier transforms," *IMA J. Appl. Math.* **39**, 159–175 (1987).
50. A. W. Lohmann, "Image rotation, Wigner rotation, and the fractional Fourier transform," *J. Opt. Soc. Am. A* **10**, 2181–2186 (1993).
51. R. Simon and K. B. Wolf, "Fractional Fourier transforms in two dimensions," *J. Opt. Soc. Am. A* **17**, 2368–2381 (2000).

52. H. M. Ozaktas, Z. Zalevsky, and M. A. Kutay, *The Fractional Fourier Transform with Applications in Optics and Signal Processing*, Wiley, New York, 2001.
53. K. B. Wolf, *Geometric Optics on Phase Space*, Springer, Berlin, 2004.
54. T. Alieva and M. J. Bastiaans, "On fractional Fourier transform moments," *IEEE Signal Process. Lett.* **7**, 320–323 (2000).
55. T. Alieva and M. J. Bastiaans, "Phase-space distributions in quasi-polar coordinates and the fractional Fourier transform," *J. Opt. Soc. Am. A* **17**, 2324–2329 (2000).
56. T. Alieva, M. J. Bastiaans, and L. Stanković, "Signal reconstruction from two close fractional Fourier power spectra," *IEEE Trans. Signal Process.* **51**, 112–123 (2003).
57. M. J. Bastiaans and K. B. Wolf, "Phase reconstruction from intensity measurements in linear systems," *J. Opt. Soc. Am. A* **20**, 1046–1049 (2003).
58. R. K. Luneburg, *Mathematical Theory of Optics*, University of California Press, Berkeley, 1966.
59. G. A. Deschamps, "Ray techniques in electromagnetics," *Proc. IEEE* **60**, 1022–1035 (1972).
60. S. A. Collins, Jr., "Lens-system diffraction integral written in terms of matrix optics," *J. Opt. Soc. Am.* **60**, 1168–1177 (1970).
61. M. Moshinsky and C. Quesne, "Linear canonical transformations and their unitary representations," *J. Math. Phys.* **12**, 1772–1780 (1971).
62. T. Alieva and M. J. Bastiaans, "Alternative representation of the linear canonical integral transform," *Opt. Lett.* **30**, 3302–3304 (2005).
63. O. Bryngdahl, "Geometrical transformations in optics," *J. Opt. Soc. Am.* **64**, 1092–1099 (1974).
64. J.-Z. Jiao, B. Wang, and H. Liu, "Wigner distribution function and geometrical transformation," *Appl. Opt.* **23**, 1249–1254 (1984).
65. A. W. Lohmann, J. Ojeda-Castañeda, and N. Streibl, "The influence of wave aberrations on the Wigner distribution," *Opt. Appl.* **13**, 465–471 (1983).
66. A. J. E. M. Janssen, "On the locus and spread of pseudo-density functions in the time-frequency plane," *Philips J. Res.* **37**, 79–110 (1982).
67. H. Bremmer, "General remarks concerning theories dealing with scattering and diffraction in random media," *Radio Sci.* **8**, 511–534 (1973).
68. J. J. McCoy and M. J. Beran, "Propagation of beamed signals through inhomogeneous media: A diffraction theory," *J. Acoust. Soc. Am.* **59**, 1142–1149 (1976).
69. I. M. Besieris and F. D. Tappert, "Stochastic wave-kinetic theory in the Liouville approximation," *J. Math. Phys.* **17**, 734–743 (1976).
70. H. Bremmer, "The Wigner distribution and transport equations in radiation problems," *J. Appl. Science Eng. A* **3**, 251–260 (1979).
71. M. Born and E. Wolf, *Principles of Optics*, Pergamon, Oxford, 1975.
72. J. Serna, R. Martínez-Herrero, and P. M. Mejías, "Parametric characterization of general partially coherent beams propagating through ABCD optical systems," *J. Opt. Soc. Am. A* **8**, 1094–1098 (1991).
73. A. Ya. Bekshaev, M. S. Soskin, and M. V. Vasnetsov, "Optical vortex symmetry breakdown and decomposition of the orbital angular momentum of the light beams," *J. Opt. Soc. Am. A* **20**, 1635–1643 (2003).
74. T. Alieva and M. J. Bastiaans, "Evolution of the vortex and the asymmetrical parts of orbital angular momentum in separable first-order optical systems," *Opt. Lett.* **29**, 1587–1589 (2004).
75. International Organization for Standardization, Technical Committee / Subcommittee 172 / SC9, "Lasers and laser-related equipment—test methods for laser beam parameters—beam widths, divergence angle and beam propagation factor," ISO Doc. 11146: 1999, International Organization for Standardization, Geneva, Switzerland, 1999.
76. R. Simon, N. Mukunda, and E. C. G. Sudarshan, "Partially coherent beams and a generalized ABCD-law," *Opt. Commun.* **65**, 322–328 (1988).
77. M. J. Bastiaans, "Second-order moments of the Wigner distribution function in first-order optical systems," *Optik* **88**, 163–168 (1991).

78. J. Williamson, "On the algebraic problem concerning the normal forms of linear dynamical systems," *Am. J. Math.* **58**, 141–163 (1936).
79. K. Sundar, N. Mukunda, and R. Simon, "Coherent-mode decomposition of general anisotropic Gaussian Schell-model beams," *J. Opt. Soc. Am. A* **12**, 560–569 (1995).
80. T. Alieva and M. J. Bastiaans, "Invariants of second-order moments of optical beams under phase-space rotations," in *ICO-21 Congress Proceedings 2008*, International Commission for Optics ICO 21, Book of Proceedings, Sydney, Australia, 7–10 July, 2008, p. 103.
81. T. Alieva and M. J. Bastiaans, "Two-dimensional signal representation on the angular Poincaré sphere," in *Proc. Topical Meeting on Optoinformatics 2008*, St. Petersburg, Russia.
82. M. J. Bastiaans and T. Alieva, "Wigner distribution moments in fractional Fourier transform systems," *J. Opt. Soc. Am. A* **19**, 1763–1773 (2002).
83. G. Nemes and A. E. Siegman, "Measurement of all ten second-order moments of an astigmatic beam by the use of rotating simple astigmatic (anamorphic) optics," *J. Opt. Soc. Am. A* **11**, 2257–2264 (1994).
84. B. Eppich, C. Gao, and H. Weber, "Determination of the ten second order intensity moments," *Opt. Laser Technol.* **30**, 337–340 (1998).
85. C. Martínez, F. Encinas-Sanz, J. Serna, P. M. Mejías, and R. Martínez-Herrero, "On the parametric characterization of the transversal spatial structure of laser pulses," *Opt. Commun.* **139**, 299–305 (1997).
86. J. Serna, F. Encinas-Sanz, and G. Nemes, "Complete spatial characterization of a pulsed doughnut-type beam by use of spherical optics and a cylindrical lens," *J. Opt. Soc. Am. A* **18**, 1726–1733 (2001).
87. M. J. Bastiaans and T. Alieva, "Wigner distribution moments measured as intensity moments in separable first-order optical systems," *EURASIP J. Appl. Signal Process.* **2005**, 1535–1540 (2005).
88. S. R. de Groot and L. G. Suttorp, *Foundations of Electrodynamics*, North-Holland, Amsterdam, 1972, Chap. 6.
89. M. J. Bastiaans, T. Alieva, and L. Stanković, "On rotated time-frequency kernels," *IEEE Signal Process. Lett.* **9**, 378–381 (2002).
90. L. Stanković, "A method for time-frequency analysis," *IEEE Trans. Signal Process.* **42**, 225–229 (1994).
91. W. Koenig, H. K. Dunn, and L. Y. Lacy, "The sound spectrograph," *J. Acoust. Soc. Am.* **18**, 19–49 (1946).
92. L. Stanković, T. Alieva, and M. J. Bastiaans, "Time-frequency signal analysis based on the windowed fractional Fourier transform," *Signal Process.* **83**, 2459–2468 (2003).

CHAPTER 2

Ambiguity Function in Optical Imaging

Jean-Pierre Guigay

European Synchrotron Radiation Facility (ESRF) Grenoble, France

2.1 Introduction

The concept of the *ambiguity function* (AF) was introduced by Woodward[1] in the theory of signal processing of radar or sonar measurements; it bears in its name the idea that it is impossible to perform arbitrarily accurate measurements of both the distance and the velocity of a moving target. The well-known reason is that the width of a signal in the time domain is inversely proportional to its width in the frequency domain (a narrower signal has a wider spectrum and inversely). From the mathematical point of view, this is just a property of the Fourier transformation; from the physical point of view, this is a common property in wave mechanics, for any kind of waves, and is the basis of the uncertainty relations in quantum physics.

This concept can be introduced in optics as an extension of the spectral analysis of images; the images are two-dimensional intensity patterns $I(x, y)$ which can often be analyzed in a convenient way by considering their Fourier transform (intensity spectrum) written as

$$\tilde{I}(u, v) = \iint dx\, dy\, I(x, y) \exp\left[-i2\pi(ux + vy)\right]$$

or

$$\tilde{I}(\mathbf{f}) = \int d\mathbf{x}\, I(\mathbf{x}) \exp\left[-i2\pi\mathbf{x} \cdot \mathbf{f}\right] \tag{2.1}$$

where two-dimensional vectors are used; the variables (u, v) are the spatial frequencies conjugate to the coordinates (x, y).

The intensity distribution is nevertheless insufficient to describe the optical field; the phase correlation between any pair of points must be included in the description. For this purpose, it seems natural to generalize $I(\mathbf{x})$ by the mutual intensity $\rho(\mathbf{x}, \mathbf{x})$* which contains 2 times more variables and is reduced to $I(\mathbf{x})$ in the particular case $\mathbf{x} = \mathbf{x}'$. The ambiguity function AF is defined, in the frame of *phase-space optics* (PSO), as the Fourier transform with respect to $\mathbf{x}$ of the mutual intensity written as $\rho(\mathbf{x} + \mathbf{a}/2, \mathbf{x} - \mathbf{a}/2)$:

$$A(\mathbf{f}, \mathbf{a}) = \int d\mathbf{x} \exp(-i2\pi\mathbf{x} \cdot \mathbf{f}) \rho\left(\mathbf{x} + \frac{\mathbf{a}}{2}, \mathbf{x} - \frac{\mathbf{a}}{2}\right) \tag{2.2}$$

The equivalent representation

$$A(\mathbf{f}, \mathbf{a}) = \int d\mathbf{m} \exp(i2\pi\mathbf{m} \cdot \mathbf{a}) \tilde{\rho}\left(\mathbf{m} + \frac{\mathbf{f}}{2}, \mathbf{m} - \frac{\mathbf{f}}{2}\right) \tag{2.3}$$

is obtained by replacing the mutual intensity by its double Fourier expansion

$$\rho\left(\mathbf{x} + \frac{\mathbf{a}}{2}, \mathbf{x} - \frac{\mathbf{a}}{2}\right) = \iint d\mathbf{g}\, d\mathbf{h} \exp\left\{i2\pi\left[\mathbf{g} \cdot \left(\mathbf{x} + \frac{\mathbf{a}}{2}\right) - \mathbf{h} \cdot \left(\mathbf{x} - \frac{\mathbf{a}}{2}\right)\right]\right\} \tilde{\rho}(\mathbf{g}, \mathbf{h}) \tag{2.4}$$

Formula (2.2) shows that $A(\mathbf{f}, 0)$ is equal to the intensity spectrum $\tilde{I}(\mathbf{f})$ which, as well as $I(\mathbf{x})$, represents the experimental data recorded by a digital detector. Formula (2.3) shows that $A(0, \mathbf{a})$ is the inverse Fourier transform of the intensity distribution $\tilde{\rho}(\mathbf{m}, \mathbf{m})$ in Fourier space.

The *Wigner distribution function* (WDF)† is defined as the Fourier transform of $\rho(\mathbf{x} + \mathbf{a}/2, \mathbf{x} - \mathbf{a}/2)$ with respect to $\mathbf{a}$, instead of $\mathbf{x}$; the WDF formulas similar to Eqs. (2.2) and (2.3) are

$$\begin{aligned} W(\mathbf{x}, \mathbf{g}) &= \int d\mathbf{a} \exp(-i2\pi\mathbf{a} \cdot \mathbf{g})\, \rho\left(\mathbf{x} + \frac{\mathbf{a}}{2}, \mathbf{x} - \frac{\mathbf{a}}{2}\right) \\ &= \int d\mathbf{m} \exp(i2\pi\mathbf{m} \cdot \mathbf{x}) \tilde{\rho}\left(\mathbf{f} + \frac{\mathbf{m}}{2}, \mathbf{f} - \frac{\mathbf{m}}{2}\right) \end{aligned} \tag{2.5}$$

The mutual intensity can be obtained from the AF or from the WDF as

$$\begin{aligned} \rho\left(\mathbf{x} + \frac{\mathbf{a}}{2}, \mathbf{x} - \frac{\mathbf{a}}{2}\right) &= \int d\mathbf{f} \exp(i2\pi\mathbf{f} \cdot \mathbf{x}) A(\mathbf{f}, \mathbf{a}) \\ &= \int d\mathbf{g} \exp(i2\pi\mathbf{g} \cdot \mathbf{a}) W(\mathbf{x}, \mathbf{g}) \end{aligned} \tag{2.6}$$

*See for instance Refs. [1] and [14] for a detailed definition of the mutual intensity function.

†The WDF is discussed in detail in Chapter 1 by Martin Bastiaans.

The AF and the WDF are related to each other by double Fourier transformation over the position and frequency variables:

$$A(\mathbf{f}, \mathbf{a}) = \int d\mathbf{x} \int d\mathbf{g} \exp\left[i2\pi(\mathbf{a}\cdot\mathbf{g} - \mathbf{f}\cdot\mathbf{x})\right] W(\mathbf{x}, \mathbf{g})$$
$$W(\mathbf{x}, \mathbf{g}) = \int d\mathbf{a} \int d\mathbf{f} \exp\left[i2\pi(\mathbf{f}\cdot\mathbf{x} - \mathbf{a}\cdot\mathbf{g})\right] A(\mathbf{f}, \mathbf{a}) \qquad (2.7)$$

The property $\rho(\mathbf{x}', \mathbf{x}) = \rho^*(\mathbf{x}, \mathbf{x}')$ of the mutual intensity shows that the WDF is real and that the AF, which is in general complex, satisfies the relation $A(-\mathbf{f}, -\mathbf{a}) = [A(\mathbf{f}, \mathbf{a})]^*$.

In the exit plane of an object of transmittance $T(\mathbf{x})$, under uniform coherent illumination, the mutual intensity $\rho(\mathbf{x} + \mathbf{a}/2, \mathbf{x} - \mathbf{a}/2)$ is equal to the product $T(\mathbf{x} + \mathbf{a}/2)T^*(\mathbf{x} - \mathbf{a}/2)$ which is, according to Chap. 5, the *product-space representation* of the signal $T(\mathbf{x})$. The corresponding AF and WDF, which are referred to as the AF and the WDF associated to $T(\mathbf{x})$, were introduced in optics by Papoulis[2] and Bastiaans,[3] respectively. They are redundant representations of $T(\mathbf{x})$.

PSO representations are useful tools to characterize the performances of optical systems. They provide elegant approaches to the description and processing of optical signals or images. It has been shown recently by Nugent[4] (see also Ref. 5) that the concept of AF can be used to unify the various noninterferometric approaches to X-ray phase retrieval.

To simplify the formulation of the following sections, we most often consider one-dimensional fields, in which case the two-dimensional vectors are replaced by scalars, without a real loss of generality, because the extension to the general case is usually straigthforward.

2.2 Intensity Spectrum of a Fresnel Diffraction Pattern Under Coherent Illumination

2.2.1 General Formulation

For simplicity, let us consider a plane wave, of wavelength λ, incident along the z direction on a thin object of transmittance $T(x)$ in the plane $z = 0$. In the conditions of Fresnel diffraction, the wave function in a plane $z = D$ is

$$\psi_D(x) = |\lambda D|^{-1/2} \exp\left(-i\frac{\pi}{4}\right) \int d\eta \exp\left[i\pi\frac{(x-\eta)^2}{\lambda D}\right] T(\eta) \qquad (2.8)$$

The corresponding intensity spectrum can be expressed by the multiple integral

$$\tilde{I}_D(f) = \int dx \exp(-i2\pi xf) \iint \frac{d\eta\, d\eta'}{\lambda D} \exp\left[i\pi \frac{(x-\eta)^2 - (x-\eta')^2}{\lambda D}\right] T(\eta) T^*(\eta') \tag{2.9}$$

As the integration over x results in the following delta function

$$\int dx \exp\left[-i2\pi x\left(f + \frac{\eta - \eta'}{\lambda D}\right)\right] = \lambda D \delta(\lambda D f + \eta - \eta') \tag{2.10}$$

the intensity spectrum can thus be reduced to a single integration.[6, 7]

$$\begin{aligned}\tilde{I}_D(f) &= \exp(-i\pi\lambda D f^2) \int d\eta \exp(-i2\pi f\eta) T(\eta) T^*(\eta + \lambda D f) \\ &= \int dx \exp(-i2\pi f x) T\left(x - \frac{\lambda D f}{2}\right) T^*\left(x + \frac{\lambda D f}{2}\right)\end{aligned} \tag{2.11}$$

Similar expressions also exist in terms of $\tilde{T}(f)$:

$$\begin{aligned}\tilde{I}_D(f) &= \exp(-i\pi\lambda D f^2) \int dh \exp(-i2\pi\lambda D h f) \tilde{T}(h + f) \tilde{T}^*(h) \\ &= \int dh \exp(-i2\pi\lambda D h f) \tilde{T}\left(h + \frac{f}{2}\right) \tilde{T}^*\left(h - \frac{f}{2}\right)\end{aligned} \tag{2.12}$$

It is interesting to note that the AF associated with $T(x)$ is apparent in this formulation if the intensity spectrum is formally considered as a function of f and $a = \lambda D f$.

2.2.2 Application to Simple Objects

This formulation can provide interesting results for some typical Fresnel diffraction patterns. For instance, in the case of a slit of full width w, we obtain[7]

$$\tilde{I}_D(\mathbf{f}) = \int_{-(w-|\lambda D f|)/2}^{(w-|\lambda D f|)/2} dx\, e^{-i2\pi f x} = \begin{cases} \frac{\sin[\pi f (w - |\lambda D f|)]}{\pi f} & \text{for } |f| \le \left|\frac{w}{\lambda D}\right| \\ 0 & \text{otherwise} \end{cases} \tag{2.13}$$

which is analytically much simpler than the intensity distribution $I(x)$ in terms of Fresnel integrals represented geometrically by the Cornu spiral.

2.2.3 Contrast Transfer Functions

Considering $T(x) = \exp[-B(x) + i\varphi(x)]$, where $\exp[-B(x)]$ and $\varphi(x)$ are the absorption and the phase modulations, respectively, of the object, we can introduce in formula (2.11) the approximation

$$T^*\left(x + \frac{\lambda D f}{2}\right) T\left(x - \frac{\lambda D f}{2}\right) \simeq 1 - B\left(x + \frac{\lambda D f}{2}\right) - B\left(x - \frac{\lambda D f}{2}\right) - i\left[\varphi\left(x + \frac{\lambda D f}{2}\right) - \varphi\left(x - \frac{\lambda D f}{2}\right)\right] \tag{2.14}$$

which should be valid under the conditions $0 < B(x) \ll 1$ (weak absorption) and $|\varphi(x) - \varphi(x - \lambda D f)| \ll 1$. This last condition is the slowly varying phase condition[6] which is less restrictive and more precise than the weak phase condition $|\varphi(x)| \ll 1$. Under such conditions, the intensity spectrum takes a simple linear form

$$\tilde{I}_D(f) = \delta(f) - 2\cos(\pi\lambda D f^2)\tilde{B}(f) + 2\sin(\pi\lambda D f^2)\tilde{\varphi}(f) \tag{2.15}$$

where the factors of $\tilde{B}(f)$ and $\tilde{\varphi}(f)$ are called the *absorption-transfer function* (ATF) and the *phase-transfer function* (PTF) respectively.

Formula (2.14) can be generalized to the case of an imaging system with aberrations other than defocusing. In electron microscopy, for which primary spherical aberration characterized by the coefficient C_S is unavoidable, the following formula is to be used.[8, 9]

$$\tilde{I}_D(f) = \delta(f) - 2\cos[\omega(f)]\tilde{B}(f) + 2\sin[\omega(f)]\tilde{\varphi}(f) \tag{2.16}$$

with

$$\omega(f) = \pi\left(\lambda D f^2 + \frac{C_S \lambda^3 f^4}{2}\right)$$

2.3 Propagation through a Paraxial Optical System in Terms of AF

2.3.1 Propagation in Free Space

Let us consider the propagation in free space, with mean direction along the z axis, of a partially coherent beam. The mutual intensity in the $z = D$ plane is given in terms of the mutual intensity in the $z = 0$ plane as

$$\rho_D\left(x + \frac{a}{2}, x - \frac{a}{2}\right) = \frac{1}{\lambda D}\int d\eta \exp\left[i\pi\frac{(x + a/2 - \eta)^2}{\lambda D}\right]\int d\xi \times \exp\left[-i\pi\frac{(x - a/2 - \xi)^2}{\lambda D}\right]\rho_0(\eta, \xi) \tag{2.17}$$

After insertion of this expression in (2.2), it can be seen that the integration over x results in the delta function $\lambda D\delta(\xi - \eta + a - \lambda Df)$. The multiple integral is thus reduced to a single integral

$$A_D(f,a) = \int d\eta \exp(-i2\pi f\eta)\rho\left(\eta + \frac{a-\lambda Df}{2}, \eta - \frac{a-\lambda Df}{2}\right)$$
$$= A_0(f, a - \lambda Df) \tag{2.18}$$

This result can be more readily obtained by recalling that the Fresnel diffraction integral is the convolution of the input function $T(x)$ by the propagator $G(x) = \exp(i\pi x^2/\lambda D - i\pi/4)/\sqrt{\lambda D}$; the output spectrum is therefore the product of the input spectrum $\tilde{T}(f)$ by the spectrum $\tilde{G}(f) = \exp(-i\pi\lambda Df^2)$ of $G(x)$; this is translated in terms of mutual intensity as

$$\tilde{\rho}_D\left(m + \frac{f}{2}, m - \frac{f}{2}\right) = \exp\left\{-i\pi\lambda D\left[\left(m + \frac{f}{2}\right)^2 - \left(m - \frac{f}{2}\right)^2\right]\right\}$$
$$\times \tilde{\rho}_0\left(m + \frac{f}{2}, m - \frac{f}{2}\right)$$
$$= \exp(-i2\pi m\lambda Df)\tilde{\rho}_0\left(m + \frac{f}{2}, m - \frac{f}{2}\right) \tag{2.19}$$

Inserting this expression in (2.3), we indeed obtain directly[2,11]

$$A_D(f,a) = \int dm \exp[i2\pi m(a - \lambda Df)]\tilde{\rho}_0\left(m + \frac{f}{2}, m - \frac{f}{2}\right)$$
$$= A_0(f, a - \lambda Df) \tag{2.20}$$

As shown in Ref. 3, a similar formula also exists for the WDF:

$$W_D(x,g) = W_0(x - \lambda Dg, g) \tag{2.21}$$

According to Eqs. (2.18) and (2.21), the AF and the WDF propagate in a uniform medium without a change of their functional forms; only the variables are linearly transformed. This is an elegant representation of Fresnel diffraction phenomena.

2.3.2 Transmission through a Thin Object

In this case, the incident mutual intensity $\rho_{\text{inc}}(x, x')$ is multiplied by $T(x)T^*(x')$, where $T(x)$ is the object transmittance. The AF of the incident beam is then to be convoluted with the object AF $A_T(f,a)$ as follows:

$$A(f,a) = \int dh\, A_{\text{inc}}(h,a)\, A_T(f-h,a) \tag{2.22}$$

where

$$A_T(f,a) = \int dx \exp(-i2\pi xf) T\left(x + \frac{a}{2}\right) T^*\left(x - \frac{a}{2}\right)$$

The transmission by a thin lens of focal length F is of special interest.[2] This lens behaves as an object of transmittance $\exp(-i\pi x^2/\lambda F)$. With $\rho_{\text{inc}}(x + a/2, x - a/2)$ being the mutual intensity in the lens entrance surface, the AF in the lens exit surface is calculated as

$$\begin{aligned} A(f,a) &= \int dx \exp\left[-i2\pi xf - i\pi \frac{(x + a/2)^2 - (x - a/2)^2}{\lambda F}\right] \\ &\quad \times \rho_{\text{inc}}\left(x + \frac{a}{2}, x - \frac{a}{2}\right) \\ &= \int dx \exp\left[-i2\pi x\left(f + \frac{a}{\lambda F}\right)\right] \rho_{\text{inc}}\left(x + \frac{a}{2}, x - \frac{a}{2}\right) \\ &= A_{\text{inc}}\left(f + \frac{a}{\lambda F}, a\right) \end{aligned} \tag{2.23}$$

The corresponding formula for the WDF is easily found as

$$W(x,g) = W_{\text{inc}}\left(x, g + \frac{x}{\lambda F}\right) \tag{2.24}$$

2.3.3 Propagation in a Paraxial Optical System

Consider the process of propagation from an input plane to a thin lens over a distance D_1, then transmission by this lens of focal length F, and finally propagation to the output plane over a distance D_2. Performing the corresponding transformations successively, according to Eqs. (2.20) and (2.22), it is easy to obtain the output AF in terms of the input AF:

$$\begin{aligned} A_{\text{out}}(f,a) = A_{\text{in}}\Bigg(&f - f\frac{D_1}{F} + \frac{a}{\lambda F}, a - a\frac{D_2}{F} \\ &- f\lambda\left(D_1 + D_2 - \frac{D_1 D_2}{F}\right)\Bigg) \end{aligned} \tag{2.25}$$

This linear transformation of variables can be represented by the matrix

$$\mathbf{M} = \begin{pmatrix} 1 - \frac{D_1}{F} & \frac{1}{\lambda F} \\ \lambda\left(\frac{D_1 D_2}{F} - D_1 - D_2\right) & 1 - \frac{D_2}{F} \end{pmatrix} \tag{2.26}$$

which can also be obtained by multiplication of the matrices corresponding to the elementary successive transformations. This matrix shows a remarkable similarity with the ray-transfer matrix of geometrical optics (see, for instance, Ref. 10) which transforms an ingoing

ray to the corresponding outgoing ray, in the paraxial approximation, with each ray being represented by a column vector $\begin{pmatrix} \alpha/\lambda \\ x \end{pmatrix}$, where x and α denote the position and the direction of the ray, respectively. It turns out that the ray-transfer matrix is equal to the inverse of matrix **M** of Eq. (2.26).

Therefore, the correspondance between geometrical optics and the PSO formulation pointed out in Ref. 3 for the WDF appears to be also valid for the AF.

This PSO method is thus a convenient and elegant tool to describe the propagation of a coherent or partially coherent beam in any system comprising coaxial lenses, and it is possible to use the WDF instead of the AF. The PSO method is much simpler than the method based on the propagation of mutual intensity which would involve convolution integrals or Fourier transformations.

2.4 The AF in Isoplanatic (Space-Invariant) Imaging

The mutual intensity $\rho_{\text{im}}(x, x')$ in the image plane is given in terms of the mutual intensity $\rho_{\text{ob}}(x, x')$ in the object plane (for convenience, the magnification is set equal to 1) as

$$\rho_{\text{im}}(x, x') = \iint d\eta\, d\eta' G(x-\eta) G^*(x'-\eta') \rho_{\text{ob}}(\eta, \eta') \tag{2.27}$$

where $G(x)$ is the coherent *point-spread function* (PSF) of the imaging system. The image AF is therefore

$$\begin{aligned} A_{\text{im}}(f, a) = & \int dx \exp(-i2\pi x f) \int d\eta\, G\left(x + \frac{a}{2} - \eta\right) \\ & \times \int d\eta G^*\left(x - \frac{a}{2} - \eta'\right) \rho_{\text{ob}}(\eta, \eta') \end{aligned} \tag{2.28}$$

By introducing new variables $\sigma = (\eta + \eta')/2$, $\tau = \eta - \eta'$, and $t = x - \sigma$ in this integral expression, we obtain directly[11–13]

$$\begin{aligned} A_{\text{im}}(f, a) = & \iint dt\, d\sigma \exp\left[-i2\pi f(t+\sigma)\right] \\ & \times \int d\tau\, G\left(t + \frac{a-\tau}{2}\right) G^*\left(t - \frac{a-\tau}{2}\right) \rho_{\text{ob}}\left(\sigma + \frac{\tau}{2}, \sigma - \frac{\tau}{2}\right) \\ = & \int d\tau\, A_G(f, a-\tau) A_{\text{ob}}(f, \tau) \end{aligned} \tag{2.29}$$

This is the convolution integral, with respect to the variable a, of the AF in the object plane

$$A_{\text{ob}}(f,\tau) = \int d\sigma \exp(-i2\pi f\sigma)\, \rho_{\text{ob}}\left(\sigma + \frac{\tau}{2}, \sigma - \frac{\tau}{2}\right) \tag{2.30}$$

with the function

$$\begin{aligned} A_G(f,a) &= \int dx \exp(-i2\pi xf)\, G\left(x + \frac{a}{2}\right) G^*\left(x - \frac{a}{2}\right) \\ &= \int dg \exp(i2\pi ga)\, \tilde{G}\left(g + \frac{f}{2}\right) \tilde{G}^*\left(g - \frac{f}{2}\right) \end{aligned} \tag{2.31}$$

where $\tilde{G}(g)$ is known as the *pupil function* of the imaging system. Also, $A_G(f,a)$ is to be considered equivalently as the AF of the coherent PSF or as the pupil AF.

The image intensity spectrum, which is a quantity of special interest, is obtained by setting a equal to 0 in Eq. (2.29):

$$\tilde{I}_{\text{im}}(f) = A_{\text{im}}f, 0) = \int d\tau\, A_G(f,-\tau) A_{\text{ob}}(f,\tau) \tag{2.32}$$

Formula (2.31) shows that, in the particular case of free-space propagation for which $\tilde{G}(f) = \exp(-i\pi\lambda Df^2)$, the pupil AF is a delta function $\delta(a - \lambda Lf)$. The pupil AF of an ideal (stigmatic) system is equal to $\delta(a)$.

2.5 The AF of the Image of an Incoherent Source

2.5.1 Derivation of the Zernike-Van Cittert Theorem from the Propagation of the AF

The mutual intensity of a planar incoherent source, in the source plane, is of the form $\rho(x, x') = S(x)\delta(x - x')$, where $S(x)$ is the source density [note that, for obvious dimensionality considerations, $S(x)$ is not an intensity distribution in the sense of the present formulation].

Formula (2.2) shows that the AF in the source plane is $\tilde{S}(f)\delta(a)$, where $\tilde{S}(f)$ is the spectrum of the source density. According to formula (2.18), at distance L from the source, the AF is therefore $\tilde{S}(f)\delta(a - \lambda Lf)$, and the mutual intensity can be obtained by formula (2.6) as

$$\begin{aligned} \rho\left(x + \frac{a}{2}, x - \frac{a}{2}\right) &= \int df \exp(i2\pi xf)\tilde{S}(f)\delta(a - \lambda Lf) \\ &= \tilde{S}\left(\frac{a}{\lambda L}\right) \exp\left(\frac{i2\pi xa}{\lambda L}\right)(\lambda L)^{-1} \end{aligned} \tag{2.33}$$

This result, which is the expression of the Van Cittert-Zernike theorem,[14] can also be obtained from the WDF which is equal to $S(x - \lambda L f)$, since the WDF is easily found to be equal to $S(x)$ in the plane of the incoherent source.

2.5.2 Partial Coherence Properties in the Image of an Incoherent Source[11]

The image of an incoherent source (or equivalently in the image of an object under incoherent illumination) by a nonideal optical system shows some degree of coherence because the light from each point of the primary source is spread over a finite area in the image; it is convenient to introduce the image AF, which characterizes this image completely, including its coherence properties, and has a simple expression.[11]

$$A_{\text{im}}(f, a) = \int d\tau A_G(f, a - \tau)\tilde{S}(f)\delta(\tau) = A_G(f, a)\tilde{S}(f) \qquad (2.34)$$

2.5.3 The Pupil AF as a Generalization of the OTF

In the case $a = 0$, formula (2.32) gives the image intensity spectrum as

$$\tilde{I}_{\text{im}}(f) = A_{\text{im}}(f, 0) = A_G(f, 0)\tilde{S}(f) \qquad (2.35)$$

This formula shows that $A_G(f, 0)$ is identical to the well-known *optical transfer function* (OTF). Since formula (2.34) is obviously a generalization of formula (2.35), the pupil AF is to be considered as a generalization of the OTF.

According to Eq. (2.18), if the image is recorded at a distance D from its normal position (defocusing), a is to be replaced by $a - \lambda D f$ in Eq. (2.34). This means[11] that the defocused pupil AF is $A_G(f, a - \lambda D f)$. The defocused OTF is consequently $A_G(f, -\lambda D f)$. Therefore, the pupil AF contains the values of the OTF for any value of the defocusing distance. More precisely, as first pointed out in Ref. 15, the pupil AF can be seen as a polar display of the OTF for variable defocusing distance: the OTF is displayed, as represented schematically in Fig. 2.1, along lines going through the origin of coordinates in the (f, a) representation (see Chapt. 5).

This connection between the OTF and the pupil AF has been used[16–24] for designing pupil phase masks (phase apodizers) which increase the depth of focus without losing lateral resolution and light-gathering power. Furthermore, various effects such as the behavior of the Strehl ratio and the sensitivity to spherical aberration,[25] or the

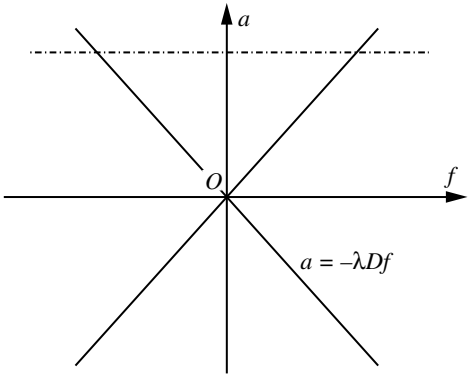


FIGURE 2.1 Schematic PSO representation. The AF along the line $a = -\lambda D f$ corresponds to the intensity spectra $\tilde{I}_D(f)$ at distance D from the reference plane. The integration of formula (2.37) is to be performed along lines parallel to the f axis.

focal shift,[26] have been studied by considerations based on the behavior of the pupil AF. These applications are detailed in Chapt. 5.

2.6 Phase-Space Tomography

The idea of phase-space tomography[4, 5, 27–33] is to reconstruct the AF (or the WDF) in the plane $z = 0$ from a set of intensity measurements $I_D(x)$ in planes $z = D_n$. The mutual intensity can then be derived from the reconstructed AF (or WDF) by using formula (2.6). This is of interest for the characterization of the optical field in the plane $z = 0$. If an object of transmittance $T(x)$ is placed in this plane, we obtain

$$\rho_{\text{inc}}\left(x + \frac{a}{2}, x - \frac{a}{2}\right) T\left(x + \frac{a}{2}\right) T^*\left(x - \frac{a}{2}\right) = \int df \exp(i2\pi f x) A(f, a) \tag{2.36}$$

where ρ_{inc} is the mutual intensity of the incident beam. With $x = a/2$, this is reduced to

$$\rho_{\text{inc}}(a, 0) T(a) T^*(0) = \int df \exp(i\pi f a) A(f, a) \tag{2.37}$$

As $\rho_{\text{inc}}(a, 0)$ and the modulus of $T(0)$ can be measured independently, this last formula allows the determination (up to a constant phase factor) of the complex function $T(a)$ from the AF.

The WDF tomographic reconstruction is based on the formula

$$I_D(x) = \int df\, W(x - \lambda D f,\, f) \tag{2.38}$$

which shows that $I_D(x)$ is the projection of the WDF in the (x, f) space along a direction that can be varied by the position $z = D$ of the recording plane. The operation that allows the WDF reconstruction is an inverse Radon transform.[27, 28] The feasibility of the tomographic WDF reconstruction has been discussed in Refs. 27 to 31.

The AF reconstruction is considered[28–31] to be simpler, because there is no need of inverse Radon transform (the term *phase-space tomography* is therefore not really appropriate in this case); we only need to perform the Fourier transformation of the measured $I_D(x)$. The relation $\tilde{I}_D(f) = A(f, -\lambda D f)$ shows that the intensity spectra represent the variations of the AF along the radial lines $a = -\lambda D f$ in the (f, a) space, as depicted in Fig. 2.1. To sample the AF over the complete (f, a) space, it is necessary to use negative as well as positive values of the distance D; this is not possible presently in X-ray optics because appropriate lenses are not available.

The process of AF reconstruction was first considered in the case of one-dimensional structures.[28, 29] A two-dimensional structure can be considered as an ensemble of one-dimensional y structures $T(x_0, y)$; the corresponding AFs are $A(x_0; f_y, a_y)$, from which $T(x_0, y)$ can be derived as a function of y according to Eq. (2.35). For this purpose, a convenient setup, actually a one-dimensional propagator system, has been proposed by Liu and Brenner[32] (see also Ref. 33): A cylindrical lens of focal length F produces an exact image (corresponding to no effective propagation) in the x direction, while there is propagation over the object image distance in the y direction; this distance can be varied by using several cylindrical lenses.

2.7 Another Possible Approach to AF Reconstruction

The AF in the exit plane of an object illuminated by a tilted plane wave of tilting angle α is

$$\int dx \exp(-i2\pi f x) T\left(x + \frac{a}{2}\right) \exp\left[i2\pi \frac{\alpha}{\lambda}\left(x + \frac{a}{2}\right)\right] T^*\left(x - \frac{a}{2}\right)$$
$$\times \exp\left[-i2\pi \frac{\alpha}{\lambda}\left(x - \frac{a}{2}\right)\right] = A_{\text{ob}}(f, a) \exp\left(i2\pi \frac{\alpha a}{\lambda}\right) \tag{2.39}$$

where $A_{\text{ob}}(f, a)$ is the AF associated to the transmittance $T(x)$. By setting $\omega = \alpha/\lambda$, the intensity spectrum of the image delivered by an

imaging system with pupil AF $A_G(f, \tau)$ is

$$\tilde{I}_{\text{im}}(f, \omega) = \int d\tau A_G(f, -\tau) A_{\text{ob}}(f, \tau) \exp(i2\pi\omega\tau) \tag{2.40}$$

which is a Fourier transform. Consequently,

$$A_{\text{ob}}(f, \tau) A_G(f, -\tau) = \int d\omega \exp(-i2\pi\omega\tau) \tilde{I}_{\text{im}}(f, \omega) \tag{2.41}$$

Supposing $A_G(f, -\tau)$ to be known and $\tilde{I}_{\text{im}}(f, \omega)$ to be measured as a function of ω, we see this last formula provides the possibility to obtain the object AF.

This approach was suggested[34] in the context of X-ray analyzer-based imaging which is a Schlieren-type technique (See Fig. 2.2) based on Bragg diffraction by a perfect crystal (analyzer) acting as a filter in Fourier space.[34] The image spectrum is equal to $\tilde{T}(f)\tilde{G}(f)$, where $\tilde{T}(f)$ is the object spectrum and $\tilde{G}(f)$ is the complex reflectivity of the crystal for a plane wave of offset angular position $\alpha = \lambda f$ with respect to the exact Bragg angular position. Therefore $\tilde{G}(f)$ is the pupil function of the system.

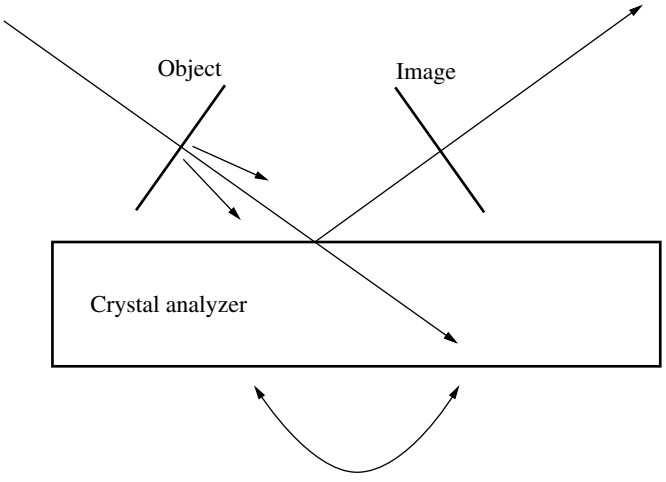


FIGURE 2.2 Principle of the X-ray analyzer-based imaging system. A quasi-parallel and quasi-monochromatic beam is diffracted, after transmission through the object, by a plate of perfect silicon crystal. The arrangement is such that Bragg diffraction occurs, corresponding to reflecting planes parallel to the crystal surface. The Bragg diffraction process is highly sensitive to the direction of the rays. Images are recorded across the diffracted beam for different angular settings of the crystal.

If the crystal is rotated by an angle $\delta\theta$ from its peak position in the incident beam, this pupil function is changed into $\tilde{G}(f - \delta\theta/\lambda)$. It is easy to show from formula (2.31) that the pupil AF $A_G(f, a)$ is then changed to $A_G(f, a) \exp(i2\pi a\,\delta\theta/\lambda)$; we then obtain for the image intensity spectrum the same formula as Eq. (2.41) with $\omega = \delta\theta/\lambda$. This shows that, as expected, the rotation of the crystal analyzer is equivalent to a change in the direction of the incident beam.

This technique is sensitive to the object structure in one dimension. To overcome this limitation, it should be just necessary, for each angular position of the crystal analyzer, to perform a 90° rotation of the object in its plane, to obtain finally two-dimensional information.

2.8 Propagation-Based Holographic Phase Retrieval from Several Images

2.8.1 Fresnel Diffraction Images as In-Line Holograms

The holographic features of Fresnel diffraction images are clearly shown by writing formula (2.11), with $T(\eta) = 1 + \psi(\eta)$, as

$$\exp(i\pi\lambda D f^2)\tilde{I}_D(f) = \delta(f) + \Psi(f) + \exp(2i\pi\lambda D f^2)\Psi^*(-f) + \int dh \exp(-i2\pi\lambda D h f)\Psi(h + f)\Psi^*(h) \quad (2.42)$$

The sum of the two first terms corresponds to the reconstructed object; the next term corresponds to the out-of-focus image (at distance $2D$) of the conjugate object $\psi^*(x)$; the integral term is negligible if $|\psi(x)| = 1$ (weak object). The importance of these perturbation terms can be strongly reduced by performing the following summation based on N images recorded at different distances D_n:

$$\frac{1}{N}\sum_{n=1}^{N} \tilde{I}_{D_n}(f) \exp\left(i\pi\lambda D_n f^2\right) = \delta(f) + \Psi(f) + \frac{\Psi^*(-f)}{N} \times \sum_{n=1}^{N} \exp\left(2i\pi\lambda D_n f^2\right) + N^{-1} \times \int dh \Psi(h + f)\Psi^*(h) \times \sum_{n=1}^{N} \exp(-i2\pi\lambda D_n h f) \quad (2.43)$$

The quantity on the left-hand side may be calculated from digitally recorded images; this allows a good reconstruction of the object if the perturbation terms are nearly canceled by this summation procedure.

2.8.2 Application to Phase Retrieval and X-Ray Holotomography

In the case of a phase object, such that $T(\eta) = \exp[i\varphi(\eta)]$, formula (2.42) can be written in the following form:

$$\tilde{I}_{D_n}(f) - \exp(-i\pi\lambda D f^2) \int dh \exp(-i2\pi\lambda D_n h f)\, \Psi(h+f)\Psi^*(h) = \delta(f) + 2\sin(\pi\lambda D_n f^2)\tilde{\varphi}(f) \tag{2.44}$$

Neglecting the nonlinear term (the integral term), denoted as NLT_{D_n}, in the left-hand side of this equation, we obtain the following estimation of the phase spectrum[35] from the experimentally known $\tilde{I}_{D_n}(f)$ by a least-squares fitting as

$$\tilde{\varphi}(f) = \frac{\sum_n \sin(\pi\lambda D_n f^2)\tilde{I}_{D_n}(f)}{2\sum_n \sin^2(\pi\lambda D_n f^2)} \tag{2.45}$$

From this result, it is possible to calculate the nonlinear terms to check whether they could indeed be neglected. If necessary, it is possible to take them into account recursively: the calculated NLT_{D_n} can be substracted from the experimentally known $\tilde{I}_{D_n}(f)$ to obtain a new estimate of the phase spectrum

$$\tilde{\varphi}(f) = \frac{\sum_n \sin(\pi\lambda D_n f^2)[\tilde{I}_{D_n}(f) - \mathrm{NLT}_{D_n}]}{2\sum_n \sin^2(\pi\lambda D_n f^2)} \tag{2.46}$$

and this process can be continued recursively.

If a single image ($N = 1$) were used in formula (2.43), the phase spectrum could not be obtained for the spatial frequencies f such that $\sin(\pi\lambda D f^2) \simeq 0$. Using several images (typically four or five images are used[36]) allows one to eliminate this defect and to reduce the influence of the nonlinear terms.

This phase retrieval approach, which has some similarity to the focus variation method used in electron microscopy,[37] has been implemented in synchrotron X-ray optics (see Fig. 2.3) to provide two-dimensional phase maps, with micrometer resolution, of objects showing a nearly uniform absorption but introducing an important phase modulation. Advantage is taken of the high degree of spatial coherence (due to the small lateral size of the source and the long source-specimen distance) and the good monochromaticity available on modern synchrotron beam lines. The phase maps obtained for different orientations of the object are used as input for a tomographic reconstruction of the three-dimensional distribution of the electron density in the sample. This technique named *holotomography*[35,36,38] is of particular interest in the case of objects opaque to visible light. It has been

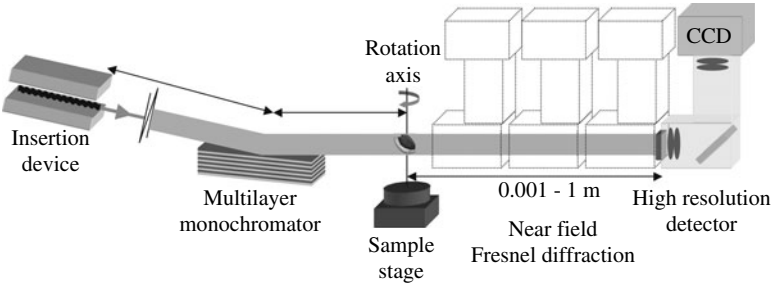


FIGURE 2.3 Scheme of the setup for X-ray holotomography operated on the ID19 beamline of the ESRF (Grenoble, France). The X-ray beam is monochromatized by a crystal monochromator or a multilayer, the energy used being typically around 15 keV (wavelength around 0.08 nm). The sample is mounted on a rotating table. The detector ensemble is a scintillator screen coupled by light optics to a CCD camera. This dectector can be moved to record images close to the object or at different distances from it.

applied to a large variety of objects of interest in biological or material sciences.

2.9 Conclusion

The AF shares with the WDF the ability to describe the propagation of a partially coherent beam in free space and through a paraxial optical system in a simple and elegant way. The AF is a generalization of the intensity spectra which are the basis of important phase retrieval methods. The images given by a space-invariant system are conveniently described in terms of AF of the object and of the pupil AF, which is a generalization of the OTF and has important applications in the design of phase apodizers. Phase-space tomography is a growing field of research, in which the AF reconstruction may be more practical than the WDF tomographic reconstruction.

References

1. P. M. Woodward, *Probability and Information Theory with Application to Radar*, Pergamon, New York, 1953.
2. A. Papoulis, "Ambiguity function in Fourier optics," *J. Opt. Soc. Am.* **64**: 779–788 (1974).
3. M. J. Bastiaans, "The Wigner distribution function applied to optical signals and systems," *Opt. Comm.* **25**(1): 274–278 (1978).

4. K. A. Nugent, "X-ray noninterferometric phase imaging: A unified picture," *J. Opt. Soc. Am. A* **24**: 536–547 (2007).
5. A. Semichaevsky and M. Testorf, "Phase-space interpretation of deterministic phase retrieval," *J. Opt. Soc. Am. A* **21**: 2173–2179 (2004).
6. J. P. Guigay, "Fourier transform analysis of Fresnel diffraction patterns and in-line holograms," *Optik* **49**: 121–125 (1977).
7. J. P. Guigay, "Analyse spectrale (frequences spatiales) d'une image de diffraction de Fresnel," C. R. Acad. Sc. Paris, **284B**: 193–196 (1977).
8. K. J. Hanszen, "In-line-holographische Erfahrungen mit Radialgittern als Testobjekten in lichtoptischen Modellanordungen für das Elektronenmikroskop," *Optik* **36**: 41–54 (1972).
9. R. W. Wade, "Spectral analysis of holograms and reconstructed images," *Optik* **40**: 201 (1974).
10. M. V. Klein and T. E. Furtak, *Optics*, 2d ed., John Wiley & Sons, New York, 1986.
11. J. P. Guigay, "The ambiguity function in diffraction and isoplanatic imaging by partially coherent beams," *Opt. Comm.* **26**: 136–138 (1978).
12. K. Dutta and J. W. Goodman, "Reconstruction of images of partially coherent objects from samples of mutual intensity," *J. Opt. Soc. Am.* **67**(6): 796–803 (1977).
13. J. Ojeda-Castañeda and E. Sicre, "Bilinear systems: Wigner distribution function and ambiguity function representations," *Optica Acta* **31**(3): 255–260 (1984).
14. M. Born and E. Wolf, "Principles of Optics," 7th ed., Cambridge University Press, 1999, Chapt. 2.
15. K-H. Brenner, A. Lohmann, and J. Ojeda-Castañeda, "The ambiguity function as a display of the OTF," *Opt. Comm.* **44**(5): 323–326 (1983).
16. K. H. Brenner and J. Ojeda-Castañeda, "Ambiguity function and Wigner distribution function applied to partially coherent imagery," *Optica Acta* **31**(2): 213–223 (1984).
17. E. R. Dowski and W. T. Cathey, "Extended depth of field through wave-front coding," *Appl. Opt.* **34**(11): 1859–1866 (1995).
18. A. R. Fitzgerrell, E. R. Dowski, and W. T. Cathey, "Defocus transfer function for circularly symmetric pupils," *Appl. Opt.* **36**(23): 5796–5804 (1997).
19. A. Castro and J. Ojeda-Castañeda, "Asymmetric phase masks for extended depth of field," *Appl. Opt.* **43**(17): 3474–3479 (2004).
20. J. Ojeda-Castañeda and L. R. Berriel-Valdos, "Ambiguity function as a design tool for high focal depth," *Appl. Opt.* **27**: 790–795 (1988).
21. J. Ojeda-Castañeda, J. E. A. Landgrave, and H. M. Escamilla, "Annular phase-only mask for high focal depth," *Opt. Lett.* **30**: 1647–1649 (2005).
22. A. Castro, J. Ojeda-Castañeda, and A. Lohmann, "Bow-tie effect: differential operator," *Appl. Opt.* **45**(30): 7878–7884 (2006).
23. Q. Yang, L. Liu, and J. Sun, "Optimized phase pupil masks for extended depth of field," *Opt. Comm.* **272**: 56–66 (2007).
24. N. Caron and Y. Sheng, "Polynomial phase masks for extending the depth of field of a microscope," *Applied Optics*, feature issue on phase-space representations in *Optics*, **47**(22): E39–E41 (2008).
25. J. Ojeda-Castañeda, P. Andrés, and E. Montes, "Phase-space representation of the Strehl ratio: Ambiguity function," *J. Opt. Soc. Am.* **4**(2): 313–317 (1987).
26. C. J. R. Sheppard and K. G. Larkin, "Focal shift, optical transfer function and phase-space representations," *J. Opt. Soc. Am. A* **17**(4): 772–779 (2000).
27. M. G. Raymer, M. Beck, and D. F. McAlister, "Complex wave-field reconstruction using phase-space tomography," *Phys. Rev. Lett.* **72**(8): 1137–1140 (1994).
28. J. Tu and S. Tamura, "Wave field determination using tomography of the ambiguity function," *Phys. Rev. E.* **55**(2): 1946–1949 (1997).
29. J. Tu and S. Tamura, "Analytic relation for recovering the mutual intensity by means of intensity information," *J. Opt. Soc. Am.* **15**(1): 202–206 (1998).
30. D. Dragoman, M. Dragoman, and K. H. Brenner, "Tomographic amplitude and phase recovery of vertical-cavity surface-emitting lasers by use of the ambiguity function," *Opt. Lett.* **27**(17): 1519–1521 (2002).

31. C. Q. Tran, A. G. Peele, A. Roberts, K. A. Nugent, D. Paterson, and I. McNulty, "X-ray imaging: A generalized approach using phase-space tomography," *J. Opt. Soc. Am.* **22**(8): 1691–1700 (2005).
32. X. Liu and K. H. Brenner, "Reconstruction of two-dimensional complex amplitudes from intensity measurements," *Opt. Comm.* **225**: 19–30 (2003).
33. X. Liu, C. Hruscha, and K. H. Brenner, "Efficient reconstruction of two-dimensional complex amplitudes utilizing the redundancy of the ambiguity function," *Applied Optics*, feature issue on phase-space representations in *Optics*, **47**(22): E1–E7 (2008).
34. J. P. Guigay, E. Pagot, and P. Cloetens, "Fourier optics approach to X-ray analyser-based imaging," *Opt. Comm.* **270**: 180–188 (2007).
35. P. Cloetens, W. Ludwig, J. Baruchel, D. Van Dyck, J. Van Landuyt, J. P. Guigay, and M. Schlenker, "Holotomography: Quantitative phase tomography with micrometer resolution using hard synchrotron radiation x rays," *Appl. Phys. Lett.* **75**: 2912–2914 (1999).
36. S. Zabler, P. Cloetens, J. P. Guigay, J. Baruchel, and M. Schlenker, "Optimization of phase contrast imaging using hard x rays," *Rev. Sci. Instr.* **76**: 073705 (2005).
37. M. Op de Beeck, D. Van Dyck, and W. Coene, "Wave-function reconstruction in HRTEM: The parabola method," *Ultramicroscopy* **64**: 167–183 (1996).
38. J. P. Guigay, M. Langer, R. Boistel, and P. Cloetens, "Mixed transfer function and transport of intensity approach for phase retrieval in the Fresnel region," *Opt. Lett.* **32**(12): 1617–1619 (2007).

CHAPTER 3
Rotations in Phase Space

Tatiana Alieva

Universidad Complutense de Madrid, Facultad de Ciencias Físicas
Ciudad Universitaria s/n, Madrid, Spain

3.1 Introduction

The *canonical* (integral linear) *transforms* (CTs) are widely used in signal and image processing, optics, quantum mechanics, etc.[1–8] The CTs are related to the affine transformations of the Wigner distribution, discussed in Chap. 1. The affine transformations in phase space, defined by the position and spatial frequency (momentum) coordinates, include scaling, shearing, rotation, etc. As we will see below, it is a rotation that plays an important role for different applications in information acquisition and processing, beam characterization, etc.

A well-known phase-space rotator is the *Fourier transform* (FT), which produces a rotation in the position (time)–spatial (temporal) frequency plane of $\pi/2$. The FT together with closely related convolution and correlation operations[9,10] forms the basis for information processing. The fractionalization of the FT[11–14] has opened new perspectives in this field. Thus the fractional Fourier transform, which produces the rotation in the position-frequency plane at arbitrary angle, has been used for shift-variant filtering, noise reduction, chirp localization, encryption, phase retrieval, etc.[4,6,15–19] The fractional FT is the only possible phase-space rotator for one-dimensional signals. If the dimension of a signal is larger than 1, there exist other phase-space rotators.

In coherent optics the FT of a two-dimensional signal, associated with complex field amplitude, can be performed by application of a

thin convergent spherical lens. It will be shown that all other phase-space rotators also can be realized experimentally by using thin lenses, but in this case some of them must be cylindrical.

In this chapter, we briefly summarize the main properties of the two-dimensional CTs, corresponding to rotations in four-dimensional phase space; we also consider in detail the basic phase-space rotators: symmetric and antisymmetric fractional FTs, signal (image) rotator and gyrator, as well as the first-order optical systems performing these transforms. Finally we discuss the applications of the phase-space rotators.

3.2 First-Order Optical Systems and Canonical Integral Transforms

3.2.1 Canonical Integral Transforms and Ray Transformation Matrix Formalism

In paraxial approximation of the scalar diffraction theory, the propagation of a coherent monochromatic light through a first-order system is described by a canonical integral transform.[1,2,4] Thus starting from the complex field amplitude $f_i(\mathbf{r}_i)$ at the input plane of the system, we have its CT at the output plane $f_o(\mathbf{r}_o)$

$$f_o(\mathbf{r}_o) = \int_{-\infty}^{\infty} f_i(\mathbf{r}_i) K^{\mathbf{T}}(\mathbf{r}_i, \mathbf{r}_o)\, d\mathbf{r}_i \tag{3.1}$$

The kernel $K^{\mathbf{T}}(\mathbf{r}_i, \mathbf{r}_o)$ is parameterized by the wavelength λ and the real symplectic ray transformation 4×4 matrix $\mathbf{T}$ that relates the position $\mathbf{r}_i$ and direction $\mathbf{p}_i$ of an incoming ray to the position $\mathbf{r}_o$ and direction $\mathbf{p}_o$ of the outgoing ray

$$\begin{bmatrix} \mathbf{r}_o \\ \mathbf{p}_o \end{bmatrix} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{r}_i \\ \mathbf{p}_i \end{bmatrix} = \mathbf{T} \begin{bmatrix} \mathbf{r}_i \\ \mathbf{p}_i \end{bmatrix} \tag{3.2}$$

where $\mathbf{r} = (x, y)^t$ and $\mathbf{p} = (p_x, p_y)^t$. The superscript t denotes transposition. Note that the term related to the time dependence and the phase accumulation $\exp(i2\pi z/\lambda)$ due to propagation at distance z will be omitted. Here and further in this chapter we use the normalized dimensionless variables and the matrix parameters. The normalized variable $\mathbf{p}$ can also be interpreted as spatial frequency or ray momentum. To convert them to real position $\boldsymbol{r}$ and ray direction $\boldsymbol{p}$, the following relations have to be used: $\boldsymbol{r} = \mathbf{r}\sqrt{\lambda w}$, $\boldsymbol{p} = \mathbf{p}\sqrt{\lambda/w}$, $\mathbf{a} = \mathbf{A}$, $\mathbf{b} = \mathbf{B}w$, $\mathbf{c} = \mathbf{C}w$, and $\mathbf{d} = \mathbf{D}$, where w is some length factor defined by the used optical system and the beam width.

The canonical integral transform associated with matrix $\mathbf{T}$ is represented by the operator $\mathcal{R}^{\mathbf{T}}$

$$f_o(\mathbf{r}_o) = \mathcal{R}^{\mathbf{T}}[f_i(\mathbf{r}_i)](\mathbf{r}_o) = F_{\mathbf{T}}(\mathbf{r}_o) \tag{3.3}$$

In the often used case $\det \mathbf{B} \neq 0$, the CT takes the form of Collins' integral[1]

$$f_o(\mathbf{r}_o) = \mathcal{R}^{\mathbf{T}}[f_i(\mathbf{r}_i)](\mathbf{r}_o) = (\det i\mathbf{B})^{-1/2} \int_{-\infty}^{\infty} f_i(\mathbf{r}_i) \times \exp\left[i\pi\left(\mathbf{r}_i^t\mathbf{B}^{-1}\mathbf{A}\mathbf{r}_i - 2\mathbf{r}_i^t\mathbf{B}^{-1}\mathbf{r}_o + \mathbf{r}_o^t\mathbf{D}\mathbf{B}^{-1}\mathbf{r}_o\right)\right] d\mathbf{r}_i \tag{3.4}$$

The kernel is a two-dimensional generalized chirp function since its phase is a polynomial of second degree of variables $\mathbf{r}_i$ and $\mathbf{r}_o$. For $\mathbf{A} = \mathbf{D} = \mathbf{0}$ and $\mathbf{B} = -\mathbf{C} = \mathbf{I}$, with $\mathbf{I}$ throughout denoting the identity matrix, we obtain, apart from a constant phase factor $\exp(-i\pi/2)$, the Fourier transform $\mathcal{F}[f(\mathbf{r}_i)](\mathbf{r}_o)$

$$\mathcal{F}[f(\mathbf{r}_i)](\mathbf{r}_o) = \int_{-\infty}^{\infty} f(\mathbf{r}_i)\, \exp(-i2\pi\mathbf{r}_o^t\mathbf{r}_i)\, d\mathbf{r}_i \tag{3.5}$$

known in optics as an angular spectrum of the complex field amplitude f. The matrix parameters $\mathbf{A} = \mathbf{D} = \mathbf{I}$, $\mathbf{C} = \mathbf{0}$, and $\mathbf{B} = z\mathbf{I}$ correspond to the Fresnel transform

$$f_o(\mathbf{r}_o) = \frac{1}{iz} \int_{-\infty}^{\infty} f_i(\mathbf{r}_i) \exp\left[i\frac{\pi}{z}(\mathbf{r}_i - \mathbf{r}_o)^2\right] d\mathbf{r}_i \tag{3.6}$$

which describes the propagation of the paraxial beams in homogeneous medium.

The case $\mathbf{B} = \mathbf{0}$ corresponds to the generalized imaging condition

$$f_o(\mathbf{r}) = (|\det \mathbf{A}|)^{-1/2} \exp(i\pi\mathbf{r}^t\mathbf{C}\mathbf{A}^{-1}\mathbf{r})\, f_i(\mathbf{A}^{-1}\mathbf{r}) \tag{3.7}$$

which includes a possible scaling and rotation of the input function accompanied by an additional phase modulation.

The CT is a linear transform: $\mathcal{R}^{\mathbf{T}}[f(\mathbf{r}_i) + g(\mathbf{r}_i)](\mathbf{r}) = \mathcal{R}^{\mathbf{T}}[f(\mathbf{r}_i)](\mathbf{r}) + \mathcal{R}^{\mathbf{T}}[g(\mathbf{r}_i)](\mathbf{r})$. It is additive in the sense that $\mathcal{R}^{\mathbf{T}_2}\mathcal{R}^{\mathbf{T}_1} = \mathcal{R}^{\mathbf{T}_2\times\mathbf{T}_1}$. The ray transformation matrix $\mathbf{T}$ is symplectic [see also Eq. (1.41)]

$$\begin{aligned} &\mathbf{A}\mathbf{B}^t = \mathbf{B}\mathbf{A}^t \qquad \mathbf{C}\mathbf{D}^t = \mathbf{D}\mathbf{C}^t \qquad \mathbf{A}\mathbf{D}^t - \mathbf{B}\mathbf{C}^t = \mathbf{I} \\ &\mathbf{A}^t\mathbf{C} = \mathbf{C}^t\mathbf{A} \qquad \mathbf{B}^t\mathbf{D} = \mathbf{D}^t\mathbf{B} \qquad \mathbf{A}^t\mathbf{D} - \mathbf{C}^t\mathbf{B} = \mathbf{I} \end{aligned} \tag{3.8}$$

and therefore it has only 10 free parameters. The inverse transformation is parametrized by the matrix $\mathbf{T}^{-1}$, which, since $\mathbf{T}$ is symplectic,

is given by

$$\mathbf{T}^{-1} = \begin{bmatrix} \mathbf{D}^t & -\mathbf{B}^t \\ -\mathbf{C}^t & \mathbf{A}^t \end{bmatrix} \tag{3.9}$$

3.2.2 Modified Iwasawa Decomposition of Ray Transformation Matrix

Any proper normalized symplectic ray transformation matrix can be decomposed in the modified Iwasawa form as[20–22]

$$\mathbf{T} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{G} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{S} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{X} & \mathbf{Y} \\ -\mathbf{Y} & \mathbf{X} \end{bmatrix} = \mathbf{T}_L \mathbf{T}_S \mathbf{T}_O \tag{3.10}$$

in which the first matrix represents a lens transform described by the symmetric matrix

$$\mathbf{G} = -(\mathbf{C}\mathbf{A}^t + \mathbf{D}\mathbf{B}^t)(\mathbf{A}\mathbf{A}^t + \mathbf{B}\mathbf{B}^t)^{-1} = \mathbf{G}^t \tag{3.11}$$

The second matrix corresponds to a scaler described by the positive definite symmetric matrix

$$\mathbf{S} = (\mathbf{A}\mathbf{A}^t + \mathbf{B}\mathbf{B}^t)^{1/2} = \mathbf{S}^t \tag{3.12}$$

and the third matrix, $\mathbf{T}_O$, is orthogonal[21,22] and due its symmetry can be shortly represented by the unitary 2×2 matrix

$$\mathbf{U} = \mathbf{X} + i\mathbf{Y} = (\mathbf{A}\mathbf{A}^t + \mathbf{B}\mathbf{B}^t)^{-1/2}(\mathbf{A} + i\mathbf{B}) \tag{3.13}$$

The action of the CTs described by the first two matrices is obvious. The lens transform produces the second-order polynomial phase modulation of the signal, and the scaler is responsible for the magnification of the signal. Therefore the most significant signal transformations are related to the last orthosymplectic matrix $\mathbf{T}_O$. They correspond to phase-space rotators and will be denoted as *rotational canonical* (integral) *transforms* (RCTs). The phase-space rotators include the signal rotator, separable fractional FT, and gyrator among others.

The signal rotator [ray transformation matrix $\mathbf{T}_r(\alpha)$] can be expressed by the unitary matrix, Eq. (3.13),

$$\mathbf{U}_r(\alpha) = \begin{bmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{bmatrix} \tag{3.14}$$

associated with the clockwise rotation in the xy and $p_x p_y$ planes at angle α.

The separable fractional FT [ray transformation matrix $\mathbf{T}_f(\gamma_x, \gamma_y)$] described by the unitary matrix

$$\mathbf{U}_f(\gamma_x, \gamma_y) = \begin{bmatrix} \exp(i\gamma_x) & 0 \\ 0 & \exp(i\gamma_y) \end{bmatrix} \tag{3.15}$$

corresponds to rotations in the xp_x and yp_y planes through angles γ_x and γ_y, respectively.

The gyrator transform associated with $\mathbf{T}_g(\vartheta)$ or

$$\mathbf{U}_g(\vartheta) = \begin{bmatrix} \cos\vartheta & i\sin\vartheta \\ i\sin\vartheta & \cos\vartheta \end{bmatrix} \tag{3.16}$$

produces twisting, i.e., rotations in the mixed xp_y and yp_x planes of phase space.

It has been shown[23] that any orthosymplectic matrix can be decomposed in the form

$$\mathbf{T}_O = \mathbf{T}_r(\beta)\,\mathbf{T}_f(\gamma_x, \gamma_y)\,\mathbf{T}_r(\alpha) \tag{3.17}$$

It means that $R^{\mathbf{T}_O}$ is a separable fractional Fourier transformer $R^{\mathbf{T}_f}$ embedded between two rotators $R^{\mathbf{T}_r}$. In particular for the gyrator matrix, we obtain $\mathbf{T}_g(\vartheta) = \mathbf{T}_r(-\pi/4)\,\mathbf{T}_f(\vartheta, -\vartheta)\,\mathbf{T}_r(\pi/4)$.

Based on the modified Iwasawa decomposition Eq. (3.10) and Eq. (3.17), we can write a general representation of the CT, which is valid for any ray transformation matrix, including a singular submatrix $\mathbf{B}$, $\det\mathbf{B} = 0$.[23]

$$\begin{aligned} f_o(\mathbf{r}_o) = \mathcal{R}^{\mathbf{T}}\left[f_i(\mathbf{r}_i)\right](\mathbf{r}_o) = (\det\mathbf{S})^{-1/2}\exp(-i\pi\mathbf{r}_o^t\mathbf{G}\mathbf{r}_o) \\ \times\,\mathcal{R}^{\mathbf{T}_f(\gamma_x,\gamma_y)}\left[f_i\left(\mathbf{X}_r(\alpha)\,\mathbf{r}_i\right)\right]\left(\mathbf{X}_r(-\beta)\mathbf{S}^{-1}\mathbf{r}_o\right) \end{aligned} \tag{3.18}$$

3.3 Canonical Transformations Producing Phase-Space Rotations

3.3.1 Matrix and Operator Description

The ray transformation matrix $\mathbf{T}_O$, which describes the phase-space rotations, is symplectic, Eq. (3.8),

$$\begin{aligned} \mathbf{X}\mathbf{Y}^t = \mathbf{Y}\mathbf{X}^t \qquad \mathbf{X}\mathbf{X}^t + \mathbf{Y}\mathbf{Y}^t = \mathbf{I} \\ \mathbf{X}^t\mathbf{Y} = \mathbf{Y}^t\mathbf{X} \qquad \mathbf{X}^t\mathbf{X} + \mathbf{Y}^t\mathbf{Y} = \mathbf{I} \end{aligned} \tag{3.19}$$

and orthogonal, $\mathbf{T}_O = (\mathbf{T}_O^t)^{-1}$, and therefore it has only four free parameters.

If a complex field amplitude $f(\mathbf{r})$ is canonically transformed with the matrix $\mathbf{T}$, then its Fourier spectrum is canonically transformed with $(\mathbf{T}^t)^{-1}$. It is easy to see that for the case of phase-space rotators both transforms coincide.

For the description of phase-space rotations we can use the unitary matrix $\mathbf{U} = \mathbf{X} + i\mathbf{Y}$ instead of the ray transformation matrix $\mathbf{T}_O$. Indeed, by introducing the following complex vector $\mathbf{q} = \mathbf{r} - i\mathbf{p}$, it is easy to check that the ray transformation equation, expressed in dimensionless variables, Eq. (3.2), for the case of orthogonal matrix

$$\begin{bmatrix} \mathbf{r}_o \\ \mathbf{p}_o \end{bmatrix} = \begin{bmatrix} \mathbf{X} & \mathbf{Y} \\ -\mathbf{Y} & \mathbf{X} \end{bmatrix} \begin{bmatrix} \mathbf{r}_i \\ \mathbf{p}_i \end{bmatrix} \tag{3.20}$$

can be rewritten as

$$(\mathbf{r} - i\mathbf{p})_o = (\mathbf{X} + i\mathbf{Y})(\mathbf{r} - i\mathbf{p})_i \tag{3.21}$$

or in more compact form as $\mathbf{q}_o = \mathbf{U}\mathbf{q}_i$. This presentation underlines the similarity in the description of phase-space rotators and polarization rotators of the monochromatic paraxial beams defined by corresponding Jones matrices.[10]

Further, the RCT operator associated with unitary matrix $\mathbf{U}$ will be denoted as $R^{\mathbf{U}}$. As well as for the ray transformation matrix $\mathbf{T}$, the additivity of the phase-space rotators is expressed as $R^{\mathbf{U}_2} R^{\mathbf{U}_1} = R^{\mathbf{U}_2 \times \mathbf{U}_1}$.

Since matrix $\mathbf{U}$ has four free parameters, there are four uniparametric groups of phase-space rotators: symmetric fractional FT, Eq. (3.15), $\gamma_x = \gamma_y$; antisymmetric fractional FT, Eq. (3.15), $\gamma_x = -\gamma_y$; gyrator, Eq. (3.16); and signal rotator, Eq. (3.14). These transforms are often written in the form of Hermitian operators[20,24–26]

$$\begin{aligned} \widehat{J}_0 &= \tfrac{1}{4}\left[\widehat{x}^2 + \widehat{y}^2 + \widehat{p_x}^2 + \widehat{p_y}^2\right] \\ \widehat{J}_1 &= \tfrac{1}{4}\left[\widehat{x}^2 - \widehat{y}^2 + \widehat{p_x}^2 - \widehat{p_y}^2\right] \\ \widehat{J}_2 &= \tfrac{1}{2}\left[\widehat{xy} + \widehat{p_x p_y}\right] \\ \widehat{J}_3 &= \tfrac{1}{2}\left[\widehat{x p_y} - \widehat{y p_x}\right] \end{aligned} \tag{3.22}$$

where $\widehat{x}$ and $\widehat{y}$, $\widehat{p_x} = -i\partial/\partial x$, and $\widehat{p_y} = -i\partial/\partial y$ are position and momentum operators. The operators $\widehat{J}_0$, $\widehat{J}_1$, $\widehat{J}_2$, and $\widehat{J}_3$ are associated with symmetric and antisymmetric fractional Fourier transforms, gyrator, and rotator, respectively. Note that the operator $\widehat{J}_0$ commutes with all others and that $[\widehat{J}_i, \widehat{J}_j] = i\varepsilon_{ijk}\widehat{J}_k$, where $i, j, k = 1, 2, 3$ and ε_{ijk} is the totally antisymmetric symbol, normalized through $\varepsilon_{123} = 1$.

Due to these commutation relations and by analogy with spin angular momentum, the operators $\widehat{J}_1$, $\widehat{J}_2$, and $\widehat{J}_3$ are often associated with *orbital angular momentum* (OAM) defined in phase space. Note that only $\widehat{J}_3$ produces the rotation in configuration space (xy plane) and relates to the beam OAM projection on the propagation direction. Moreover, the OAM operators, as we will see below, provide an elegant signal representation on the sphere, called, again by analogy with polarization description, the *orbital Poincaré sphere*. This presentation permits easy identification of the signal symmetry, its z-OAM projection, and defines the geometric phase accumulated by the Gaussian beams during their propagation through the first-order optical systems, etc. Let us consider these basic transforms in detail.

3.3.2 Signal Rotator

The signal rotator transform associated with unitary matrix $\mathbf{U}_r(\alpha)$, Eq. (3.14); $\mathbf{Y}_r = \mathbf{0}$, and

$$\mathbf{X}_r = \begin{bmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{bmatrix} \tag{3.23}$$

produces a clockwise rotation of f_i in the xy plane and, correspondingly, its FT (the angular spectrum) $F_i(p_x, p_y) = \mathcal{F}[f(\mathbf{r}_i)](\mathbf{p}_i)$ in the $p_x p_y$ plane at angle α.

$$\begin{aligned} f_o(x, y) &= f_i(x\cos\alpha - y\sin\alpha,\, x\sin\alpha + y\cos\alpha) \\ F_o(p_x, p_y) &= F_i(p_x\cos\alpha - p_y\sin\alpha,\, p_x\sin\alpha + p_y\cos\alpha) \end{aligned} \tag{3.24}$$

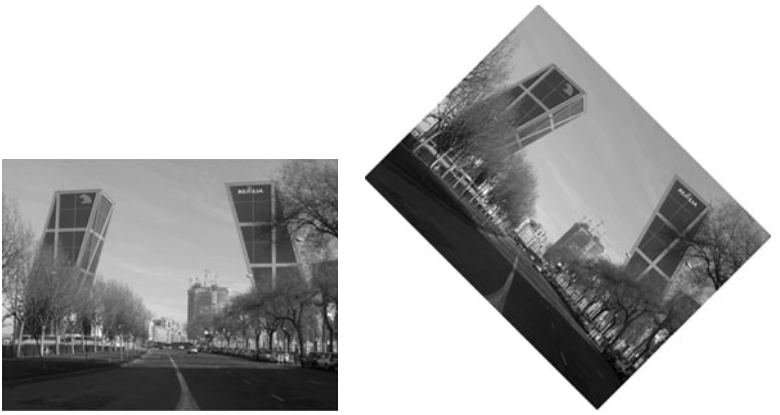
This transformation is additive with respect to angle parameter α. Thus $R^{\mathbf{U}_r(\alpha)} R^{\mathbf{U}_r(\beta)} = R^{\mathbf{U}_r(\alpha+\beta)}$, and therefore the inverse transform for $R^{\mathbf{U}_r(\alpha)}$ is a signal rotator at angle $-\alpha$. Note that $\det\mathbf{U} = \det\mathbf{X} = 1$.

The action of the signal rotator is easy to understand, and it is demonstrated in Fig. 3.1, where the original signal (real image, photograph of Madrid street) is seen in Fig. 3.1*a* and its transformation after the rotation at angle $\alpha = \pi/4$ in Fig. 3.1*b*.

The signal rotator is an important tool for the study of signal symmetry.

3.3.3 Fractional Fourier Transform

We call the transform associated with ray transformation matrix $\mathbf{T}$ *separable* if the block matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$ are diagonal. The only possible separable phase-space rotator is the fractional FT, as it is easy to see from the orthosymplectic conditions [Eq. (3.19)]. Indeed for



(*a*) (*b*)

FIGURE 3.1 (*a*) The test real positive image and (*b*) its transformation after the rotation at angle $\alpha = \pi/4$.

diagonal block matrices **X** and **Y**, they lead to the relations

$$X_{11}^2 + Y_{11}^2 = 1$$
$$X_{22}^2 + Y_{22}^2 = 1 \tag{3.25}$$

which are satisfied only if $X_{11,22} = \cos\gamma_{x,y}$ and $Y_{11,22} = \sin\gamma_{x,y}$. Thus the associated unitary matrix corresponds to the separable fractional FT one, Eq. (3.15).

The kernel of the two-dimensional separable fractional FT is a product of two one-dimensional fractional FT kernels, $K^{\mathbf{U}_f(\gamma_x,\gamma_y)}(\mathbf{r}_i, \mathbf{r}_o) = K_f^{\gamma_x}(x_i, x_o)K_f^{\gamma_y}(y_i, y_o)$, with $K_f^{\gamma_x}(x_i, x_o)$ given by

$$K_f^{\gamma_x}(x_i, x_o) = (i\sin\gamma_x)^{-1/2}\exp\left[i\pi\,\frac{\left(x_i^2 + x_o^2\right)\cos\gamma_x - 2x_o x_i}{\sin\gamma_x}\right] \tag{3.26}$$

There are two main definitions of the fractional FT kernel which differ by the phase factor $\exp(i\gamma_x/2)$; see, for example, Refs. 4 and 5. Indeed, to obtain the ordinary FT for $\gamma_x = \pi/2$ and rigorously satisfy the angle additivity, the kernel $\exp(i\gamma_x/2)K_f^{\gamma_x}(x_i, x_o)$ has to be used. Nevertheless here we consider one, Eq. (3.26), that describes the complex field amplitude propagation through the related first-order optical systems as well as time evolution of the harmonic oscillator. In general the difference in the kernel definition is not important for the applications of the RCTs, except in such particular cases as the definition of the Gouy phase,[27] where Eq. (3.26) is preferable. Moreover, the matrix formalism widely used for the description of phase-space rotators permits one to avoid the differences in the fractional FT kernel definition.

If $\gamma_x = \gamma_y = \varphi$, then the fractional FT is symmetric. The symmetric fractional FT produces the rotation in phase planes xp_x and yp_y at the same angle φ. Its kernel is written as

$$K^{\mathbf{U}_f(\varphi,\varphi)}(\mathbf{r}_i, \mathbf{r}_o) = \frac{1}{i \sin\varphi} \exp\left[i\pi \frac{\left(\mathbf{r}_i^2 + \mathbf{r}_o^2\right)\cos\varphi - 2\mathbf{r}_o^t \mathbf{r}_i}{\sin\varphi} \right] \tag{3.27}$$

For $\varphi = 0$ it corresponds to the identity transform $K^{\mathbf{U}_f(0,0)}(\mathbf{r}_i, \mathbf{r}_o) = \delta(\mathbf{r}_i - \mathbf{r}_o)$; for $\varphi = \pi/2$ to the common Fourier transform [Eq. (3.5)] apart from constant $-i$; for $\varphi = \pi$, to the reverse transform $K^{\mathbf{U}_f(\pi,\pi)}(\mathbf{r}_i, \mathbf{r}_o) = -\delta(\mathbf{r}_i + \mathbf{r}_o)$, which coincides, except for the sign, with signal rotation at angle π; and for $\varphi = 3\pi/2$ to the inverse FT apart from constant i. We observe that the symmetric fractional FT is periodic, in the strict sense, with 4π and not with 2π as the rest of basic phase-space rotators: signal rotator, antisymmetric fractional FT, and gyrator.

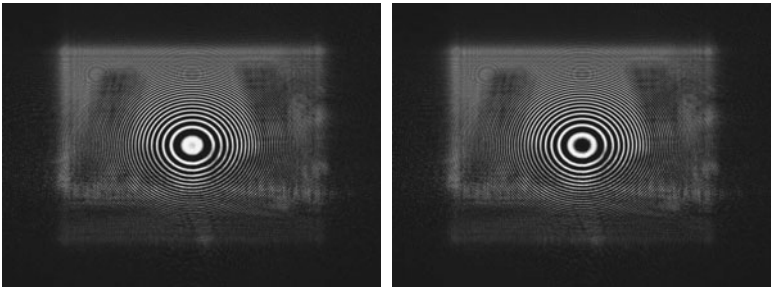
The unitary matrix associated with the symmetric fractional FT is a scalar matrix $\mathbf{U}_f(\varphi, \varphi) = \exp(i\varphi)\mathbf{I}$ with determinant $\exp(i2\varphi)$. From the scalar form of $\mathbf{U}_f(\varphi, \varphi)$, it is easy to see that the symmetric fractional FT commutes with any phase-space rotator: $\mathbf{U}_f(\varphi, \varphi)\mathbf{U} = \mathbf{U}\mathbf{U}_f(\varphi, \varphi)$.

The signal transformation under the symmetric fractional FT, obtained by numerical simulations, is demonstrated in Fig. 3.2, where the real Fig 3.2*a* and imaginary Fig.3.2*b* parts of the symmetric fractional FT at angle $\varphi = \pi/4$ of the signal shown in Fig. 3.1*a* are displayed.

If $\gamma_x = -\gamma_y = \gamma$, then the kernel is given by

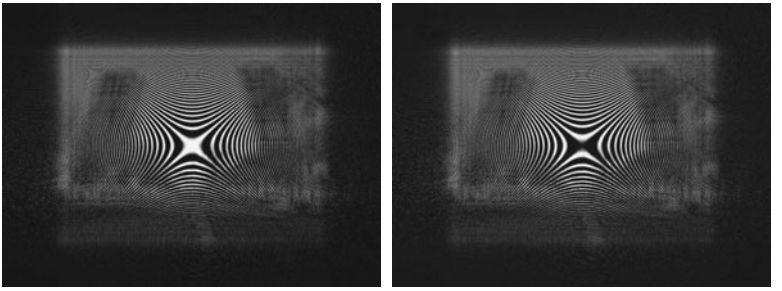
$$K^{\mathbf{U}_f(\gamma,-\gamma)}(\mathbf{r}_i, \mathbf{r}_o) = \frac{1}{\sin\gamma} \exp\left[i\pi \frac{\left(x_i^2 + x_o^2 - y_i^2 - y_o^2\right)\cos\gamma - 2\left(x_o x_i - y_o y_i\right)}{\sin\gamma} \right] \tag{3.28}$$

and corresponds to the antisymmetric fractional FT, which also produces the rotation in phase planes xp_x and yp_y but at the angles γ



(*a*) (*b*)

FIGURE 3.2 (*a*) Real and (*b*) imaginary parts of the symmetric fractional FT at angle $\varphi = \pi/4$ of the test signal (Fig. 3.1*a*).



(*a*) (*b*)

FIGURE 3.3 (*a*) Real and (*b*) imaginary parts of the antisymmetric fractional FT at angle $\varphi = \pi/4$ of the test signal (Fig. 3.1*a*) are displayed.

and $-\gamma$, respectively. The determinant of the associated unitary matrix equals 1: $\det \mathbf{U}_f(\gamma, -\gamma) = 1$. In Fig. 3.3 the real (part *a*) and the imaginary (part *b*) parts of the numerically simulated antisymmetric fractional FT at angle $\pi/4$ of the signal shown in Fig. 3.1*a* are displayed. Here as well as in Fig. 3.2, the chirp phase modulation is clearly observed.

The combination of the symmetric $\mathcal{R}^{\mathbf{U}_f(\varphi,\varphi)}$ and antisymmetric $\mathcal{R}^{\mathbf{U}_f(\gamma,-\gamma)}$ fractional FTs defines the separable fractional FT $\mathcal{R}^{\mathbf{U}_f(\gamma_x,\gamma_y)}$ at angles $\gamma_x = \varphi + \gamma$ and $\gamma_y = \varphi - \gamma$ because

$$\mathbf{U}_f(\gamma_x, \gamma_y) = \mathbf{U}_f(\varphi, \varphi)\mathbf{U}_f(\gamma, -\gamma) = \exp(i\varphi)\mathbf{U}_f(\gamma, -\gamma) \tag{3.29}$$

If $\gamma_x = 0$ and $\gamma_y = \pi$, and correspondingly $\varphi = -\gamma = \pi/2$, the separable fractional FT reduces to y reflector described by the unitary matrix

$$\mathbf{U}_{ref_y} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \tag{3.30}$$

For $\gamma_x = \pi$ and $\gamma_y = 0$, the x reflector is obtained

$$\mathbf{U}_{ref_x} = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \tag{3.31}$$

The cascade of two identical reflectors leads to the identity transform $\mathbf{U}_{ref_y}\mathbf{U}_{ref_y} = \mathbf{U}_{ref_x}\mathbf{U}_{ref_x} = \mathbf{I}$; meanwhile the cascade of the different reflectors produces the signal rotation at π, $\mathbf{U}_{ref_x}\mathbf{U}_{ref_y} = \mathbf{U}_{ref_y}\mathbf{U}_{ref_x} = -\mathbf{I}$.

As we mentioned before in Eq. (3.17), any phase-space rotator can be presented as the separable fractional FT embedded into two signal

rotators[23] that can be expressed in matrix form as

$$\mathbf{U} = \exp(i\varphi)\,\mathbf{U}_r(\beta)\,\mathbf{U}_f(\gamma, -\gamma)\,\mathbf{U}_r(\alpha) \tag{3.32}$$

where the parameters φ, γ, α, and β are defined from the components of the matrix $\mathbf{U} = \mathbf{X} + i\mathbf{Y}$ by the following relations:

$$\begin{aligned} \det \mathbf{U} &= \exp(i2\varphi) \\ \det \mathbf{X} + \det \mathbf{Y} &= \cos 2\gamma \end{aligned} \tag{3.33}$$

and

$$\begin{aligned} X_{11} + X_{22} - Y_{12} + Y_{21} &= 2\cos(\alpha + \beta + \varphi)\cos\gamma \\ X_{12} - X_{21} + Y_{11} + Y_{22} &= 2\sin(\alpha + \beta + \varphi)\cos\gamma \\ -X_{11} + X_{22} + Y_{12} + Y_{21} &= 2\sin(\alpha - \beta + \varphi)\sin\gamma \\ X_{12} + X_{21} + Y_{11} - Y_{22} &= 2\cos(\alpha - \beta + \varphi)\sin\gamma \end{aligned} \tag{3.34}$$

Note that $\det \mathbf{X} = \cos(\varphi+\gamma)\cos(\varphi-\gamma)$ and $\det \mathbf{Y} = \sin(\varphi+\gamma)\sin(\varphi-\gamma)$.

As we will see below, the fractional FT is widely used in signal and image processing, phase retrieval, tomographic reconstruction of the Wigner distribution, etc. More information about the fractional FT can be found in Refs. 4 to 6, 13, 28, and 29.

3.3.4 Gyrator

As well as the signal rotator, symmetric and antisymmetric fractional FTs, the gyrator defined by the unitary matrix $\mathbf{U}_g(\vartheta)$ with determinant equal to 1 [see Eq. (3.16)] also forms a uniparametric group of transformations. The kernel of the gyrator transform at angle ϑ has a form of a hyperbolic wave

$$K^{\mathbf{U}_g(\vartheta)}(\mathbf{r}_i, \mathbf{r}_o) = \frac{1}{|\sin\vartheta|} \exp\left[i2\pi \frac{(x_o y_o + x_i y_i)\cos\vartheta - (x_i y_o + x_o y_i)}{\sin\vartheta}\right] \tag{3.35}$$

which reduces to $\delta(\mathbf{r}_i - \mathbf{r}_o)$ for $\vartheta = 0$, to $\delta(\mathbf{r}_i + \mathbf{r}_o)$ for $\vartheta = \pi$, and to the twisted FT kernel $\exp[\mp i2\pi(x_i y_o + x_o y_i)]$ for $\vartheta = \pm\pi/2$. It is periodic with 2π. The inverse transform is the gyrator at angle $-\vartheta$. The gyrator produces rotations in the twisted xp_y and yp_x planes of phase space at angle ϑ.

The gyrator plays an important role in two-dimensional signal processing, orbital angular momentum manipulation, and beam conversion. Thus by applying the gyrator transform at angle $\pm\pi/4$ to the properly normalized Hermite-Gaussian beam, the helicoidal

Laguerre-Gaussian mode is obtained. A detailed analysis of the gyrator can be found in Refs. 30 to 33.

3.3.5 Other Phase-Space Rotators

The four uniparametric transforms—signal rotator, symmetric fractional FT, antisymmetric fractional FT, and gyrator—are additive with respect to their angle parameters and form the basis for the presentation of other phase-space rotators. Thus, for example, the cascade of the reflector, Eq. (3.31), and signal rotator produces a reflector with rotation described by the matrix

$$\mathbf{U}_1(\alpha) = \mathbf{U}_r(\alpha)\mathbf{U}_{ref_x} = \begin{bmatrix} -\cos\alpha & \sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix} \tag{3.36}$$

This transformation is not additive with respect to parameter α, because $\mathbf{U}_1(\alpha)\mathbf{U}_1(\beta) = \mathbf{U}_r(\alpha - \beta) \neq \mathbf{U}_1(\alpha + \beta)$.

The combination of this transform at angle $\pi/2$ and the fractional FT leads to the phase-space rotator described by the antidiagonal unitary matrix

$$\mathbf{U}_2(\gamma_x, \gamma_y) = \mathbf{U}_1\left(\frac{\pi}{2}\right)\mathbf{U}_f(\gamma_x, \gamma_y) = \begin{bmatrix} 0 & \exp(i\gamma_x) \\ \exp(i\gamma_y) & 0 \end{bmatrix} \tag{3.37}$$

This RCT for $\gamma_x = \gamma_y$ has been considered in Ref. 34. It is also not additive with respect to the angles $\mathbf{U}_2(\gamma_x, \gamma_y)\mathbf{U}_2(\theta_x, \theta_y) = \mathbf{U}_f(\gamma_x + \theta_x, \gamma_y + \theta_y) \neq \mathbf{U}_2(\gamma_x + \theta_x, \gamma_y + \theta_y)$.

The cascades of signal rotators and reflector correspond to all possible phase-space rotators with $\mathbf{Y} = 0$. Nevertheless there exist phase-space rotators with $\det \mathbf{Y} = 0$, but $\mathbf{Y} \neq \mathbf{0}$. Since $\det \mathbf{Y} = \sin\gamma_x \sin\gamma_y$, it is easy to see that in this case $\gamma_x = \pi n_x$ or/and $\gamma_y = \pi n_y$ and $n_{x,y}$ are integers. It means that for one coordinate, the fractional Fourier transformer in the decomposition Eq. (3.32) acts as an identity or π rotation system. As an example of such a system, we mention one considered in Ref. 35 and described by the unitary matrix

$$\mathbf{U} = \begin{bmatrix} \cos\alpha & \sin\alpha \\ -i\sin\alpha & i\cos\alpha \end{bmatrix} \tag{3.38}$$

3.4 Properties of the Phase-Space Rotators

In this section we consider the basic properties of the RCTs that are useful for the application of these transformations for the signal processing tasks and for the description of the related first-order optical systems.

In the case where $\det \mathbf{Y} \neq 0$, the RCT parameterized by $\mathbf{U}$ of function $f_i(\mathbf{r}_i)$ takes the form

$$f_o(\mathbf{r}_o) = \mathcal{R}^{\mathbf{U}}[f_i(\mathbf{r}_i)](\mathbf{r}_o) = F_U(\mathbf{r}_o) = (\det i\mathbf{Y})^{-1/2} \int_{-\infty}^{\infty} f_i(\mathbf{r}_i) \times \exp\left[i\pi\left(\mathbf{r}_i^t \mathbf{Y}^{-1}\mathbf{X}\mathbf{r}_i - 2\mathbf{r}_i^t \mathbf{Y}^{-1}\mathbf{r}_o + \mathbf{r}_o^t \mathbf{X}\mathbf{Y}^{-1}\mathbf{r}_o\right)\right] d\mathbf{r}_i \quad (3.39)$$

For $\mathbf{X} = \mathbf{0}$ the kernel reduces to a plane wave function, and the transform can be denoted as a Fourier type. As it follows from Eqs. (3.32), $\mathbf{X} = \mathbf{0}$ only if $\gamma_{x,y} = \pi/2 + \pi k_{x,y}$, where $k_{x,y}$ is an integer.

If $\det \mathbf{Y} = 0$, the decomposition Eq. (3.18), which is also valid for the nonsingular case, has to be used

$$f_o(\mathbf{r}_o) = \mathcal{R}^{\mathbf{U}}[f_i(\mathbf{r}_i)](\mathbf{r}_o) = \mathcal{R}^{\mathbf{U}_f(\gamma_x,\gamma_y)}[f_i(\mathbf{X}_r(\alpha)\,\mathbf{r}_i)][\mathbf{X}_r(-\beta)\mathbf{r}_o] \quad (3.40)$$

where the parameters $\gamma_{x,y}$, α, and β are defined from Eqs. (3.33) and (3.34). In particular, if $\mathbf{Y} = \mathbf{0}$, then $f_o(\mathbf{r}) = f_i(\mathbf{X}^{-1}\mathbf{r})$. Since $|\det \mathbf{U}| = |\det \mathbf{X}| = 1$, these RCTs correspond to signal rotation or signal rotation with reflection and may be denoted as imaging-type rotators.

3.4.1 Some Useful Relations for Phase-Space Rotators

Based on the analysis of the canonical integral transform performed in Ref. 7, it is easy to formulate the main theorems for the phase-space rotators.

The complex conjugation of the RCT parameterized by $\mathbf{U}$ of $f(\mathbf{r}_i)$ is equivalent to the RCT parameterized by $\mathbf{U}^* = \mathbf{X} - i\mathbf{Y}$ of $f^*(\mathbf{r}_i)$, that is, $\{\mathcal{R}^{\mathbf{U}}[f(\mathbf{r}_i)](\mathbf{r})\}^* = \mathcal{R}^{\mathbf{U}^*}[f^*(\mathbf{r}_i)](\mathbf{r})$.

As shown in Ref. 7, the gradient of the RCT for $\det \mathbf{Y} \neq 0$ can be written as

$$\nabla_o f_o(\mathbf{r}_o) = \nabla_o \left\{\mathcal{R}^{\mathbf{U}}[f_i(\mathbf{r}_i)](\mathbf{r}_o)\right\} = i2\pi(\mathbf{Y}^t)^{-1}\left\{\mathbf{X}^t\mathbf{r}_o\, f_o(\mathbf{r}_o) - \mathcal{R}^{\mathbf{T}}\left[\mathbf{r}_i\, f_i(\mathbf{r}_i)\right](\mathbf{r}_o)\right\} \quad (3.41)$$

where $\nabla = (\partial/\partial x,\, \partial/\partial y)^t$ and therefore

$$\mathcal{R}^{\mathbf{U}}\left[\mathbf{r}_i\, f_i(\mathbf{r}_i)\right](\mathbf{r}_o) = \left\{\mathbf{X}^t\mathbf{r}_o + i\frac{\mathbf{Y}^t}{2\pi}\nabla_o\right\} f_o(\mathbf{r}_o)$$

$$\mathcal{R}^{\mathbf{U}}\left[\nabla_i\, f_i(\mathbf{r}_i)\right](\mathbf{r}_o) = \left(i2\pi\mathbf{Y}^t\mathbf{r}_o + \mathbf{X}^t\nabla_o\right) f_o(\mathbf{r}_o) \quad (3.42)$$

The well-known Parseval theorem holds for the entire class of the CTs, and therefore for the phase-space rotators

$$\int_{-\infty}^{\infty} f(\mathbf{r}_i)\, g^*(\mathbf{r}_i)\, d\mathbf{r}_i = \int_{-\infty}^{\infty} \mathcal{R}^{\mathbf{U}}[f(\mathbf{r}_i)](\mathbf{r}_o)\, \mathcal{R}^{\mathbf{U}^*}[g^*(\mathbf{r}_i)](\mathbf{r}_o)\, d\mathbf{r}_o \quad (3.43)$$

which, in particular, yields the energy conservation law

$$\int_{-\infty}^{\infty} |f(\mathbf{r}_i)|^2 \, d\mathbf{r}_i = \int_{-\infty}^{\infty} |\mathcal{R}^{\mathbf{U}}[f(\mathbf{r}_i)](\mathbf{r}_o)|^2 \, d\mathbf{r}_o \tag{3.44}$$

We also remind (see Sec. 1.6) that the Wigner distribution is rotated in phase space under the RCT

$$W_f(\mathbf{r}, \mathbf{p}) = W_{R^{\mathbf{U}}[f]}(\mathbf{X}\mathbf{r} + \mathbf{Y}\mathbf{p}, -\mathbf{Y}\mathbf{r} + \mathbf{X}\mathbf{p}) \tag{3.45}$$

Moreover its projection corresponds to the squared modulus of the appropriated RCT, which can be registered experimentally. These properties are crucial for phase-space tomography, which permits one to reconstruct the Wigner distribution from its projections. The details of this method for the case of the fractional FT are clarified in Chap. 4.

3.4.2 Similarity to the Fractional Fourier Transform

It has been shown[36,37] that any unitary matrix $\mathbf{U}_s$ is similar to one $\mathbf{U}_f$ associated with the fractional FT. Indeed, the unitary matrix has unimodular eigenvalues and can be diagonalized. The diagonal unitary matrix corresponds to the fractional FT, Eq. (3.15). Moreover, the matrix that diagonalizes the matrix is also unitary, and therefore we can write

$$\mathbf{U}_s = \mathbf{U}\mathbf{U}_f(\gamma_x, \gamma_y)\mathbf{U}^{-1} \tag{3.46}$$

where γ_x and γ_y and the matrix $\mathbf{U}$ are defined from the eigenvalues and eigenvectors of $\mathbf{U}_s$, correspondingly. Then we can conclude that any phase-space rotator is similar to the fractional FT. For example, the signal rotator and gyrator are similar to the antisymmetric fractional FT because

$$\begin{aligned} \mathbf{U}_r(\gamma) &= \mathbf{U}_g\left(\frac{\pi}{4}\right) \mathbf{U}_f(\gamma, -\gamma)\, \mathbf{U}_g\left(-\frac{\pi}{4}\right) \\ \mathbf{U}_g(\gamma) &= \mathbf{U}_r\left(-\frac{\pi}{4}\right) \mathbf{U}_f(\gamma, -\gamma)\, \mathbf{U}_r\left(\frac{\pi}{4}\right) \end{aligned} \tag{3.47}$$

Note that due to the symmetry of the phase-space rotator matrices, such as $\mathbf{U}_g(\pi \pm \pi/4) = -\mathbf{U}_g(\pm\pi/4)$, there exist various similarity relations (see Sec. 1.6.2). For example, we can also write for the signal rotator

$$\begin{aligned} \mathbf{U}_r(\pm\gamma) &= \mathbf{U}_g\left(\pm\frac{\pi}{4}\right) \mathbf{U}_f(\gamma, -\gamma)\, \mathbf{U}_g\left(\mp\frac{\pi}{4}\right) \\ &= \mathbf{U}_g\left(\pi \pm \frac{\pi}{4}\right) \mathbf{U}_f(\gamma, -\gamma)\, \mathbf{U}_g\left(\pi \mp \frac{\pi}{4}\right) \end{aligned} \tag{3.48}$$

3.4.3 Shift Theorem

A shift of the input function by a vector $\mathbf{v}$, $f_i(\mathbf{r}) \rightarrow f_i(\mathbf{r}-\mathbf{v})$, leads to a shift of the output signal by the vector $\mathbf{Xv}$ and to an additional quadratic phase factor

$$\mathcal{R}^{\mathbf{U}}[f_i(\mathbf{r}_i-\mathbf{v})](\mathbf{r}_o) = \exp[-i\pi(2\mathbf{r}_o-\mathbf{Xv})^t\mathbf{Yv}]\ \mathcal{R}^{\mathbf{U}}[f_i(\mathbf{r}_i)](\mathbf{r}_o-\mathbf{Xv}) \quad (3.49)$$

where we have used the symplecticity conditions [Eq. (3.19)] and the fact that $\mathbf{v}^t\mathbf{Zq} = \mathbf{q}^t\mathbf{Z}^t\mathbf{v}$. This implies that the squared modulus of the RCT, associated in optics with intensity distribution, does not change due to a displacement by $\mathbf{v}$, but is merely shifted by $\mathbf{Xv}$:

$$\left|\mathcal{R}^{\mathbf{U}}[f(\mathbf{r}_i-\mathbf{v})](\mathbf{r}_o)\right|^2 = |F_{\mathbf{U}}(\mathbf{r}_o-\mathbf{Xv})|^2 \quad (3.50)$$

Equation (3.49) reduces to $\mathcal{R}^{\mathbf{U}}[f(\mathbf{r}_i-\mathbf{v})](\mathbf{r}_o) = F_{\mathbf{U}}(\mathbf{r}_o-\mathbf{Xv})$ and to $\mathcal{R}^{\mathbf{U}}[f(\mathbf{r}_i-\mathbf{v})](\mathbf{r}_o) = \exp(-i\pi 2\mathbf{r}_o^t\mathbf{Yv})\,F_{\mathbf{U}}(\mathbf{r}_o)$ for $\mathbf{Y}=\mathbf{0}$ and $\mathbf{X}=\mathbf{0}$, respectively. The shift theorem underlines the position-variant nature of signal processing in the phase-space domains if $\mathbf{X}\neq\mathbf{0}$.

3.4.4 Convolution Theorem

Using the shift theorem, the RCT of the convolution between f and h

$$\mathcal{C}_{f,h}(\mathbf{r}) = (f*h)(\mathbf{r}) = \int_{-\infty}^{\infty} f(\mathbf{r}-\mathbf{v})h(\mathbf{v})\,d\mathbf{v} = \int_{-\infty}^{\infty} h(\mathbf{r}-\mathbf{v})\,f(\mathbf{v})\,d\mathbf{v} \quad (3.51)$$

can be written in the form

$$\mathcal{R}^{\mathbf{U}}\left[(f*h)(\mathbf{r}_i)\right](\mathbf{r}_o) = \int_{-\infty}^{\infty} \exp[-i\pi(2\mathbf{r}_o-\mathbf{Xv})^t\mathbf{Yv}]\,F_{\mathbf{U}}(\mathbf{r}_o-\mathbf{Xv})\,h(\mathbf{v})\,d\mathbf{v} \quad (3.52)$$

In the case where $\mathbf{X}=\mathbf{0}$ (and thus also $\mathbf{Y}^t=\mathbf{Y}^{-1}$), it reduces to

$$\mathcal{R}^{\mathbf{U}}\left[(f*h)(\mathbf{r}_i)\right](\mathbf{r}_o) = (\det i\mathbf{Y})^{1/2}\,F_{\mathbf{U}}(\mathbf{r}_o)\,H_{\mathbf{U}}(\mathbf{r}_o) \quad (3.53)$$

For imaging-type systems $\mathbf{Y}=\mathbf{0}$, we have

$$\mathcal{R}^{\mathbf{U}}\left[(f*h)(\mathbf{r}_i)\right](\mathbf{r}_o) = \int_{-\infty}^{\infty} F_{\mathbf{U}}(\mathbf{r}_o-\mathbf{Xv})\,h(\mathbf{v})\,d\mathbf{v} \quad (3.54)$$

3.4.5 Scaling Theorem

The scaling, as we discussed above, belongs to the class of CTs. Therefore, as it follows from the additivity property of the CTs, the scaling of the input function leads to a change of the parameterizing matrix. Thus the phase-space rotation associated with matrix $\mathbf{U}$ of the scaled

function $(\det \mathbf{W})^{1/2} f(\mathbf{W}\mathbf{r}_i)$, $\mathcal{R}^{\mathbf{U}}[(\det \mathbf{W})^{1/2} f(\mathbf{W}\mathbf{r}_i)](\mathbf{r}_o)$ with $\mathbf{W} = \mathbf{W}^t$, corresponds to the CT of $f(\mathbf{r}_i)$ itself, $\mathcal{R}^{\widehat{\mathbf{T}}}[f(\mathbf{r}_i)](\mathbf{r}_o)$, parameterized by the matrix

$$\widehat{\mathbf{T}} = \begin{bmatrix} \mathbf{X} & \mathbf{Y} \\ -\mathbf{Y} & \mathbf{X} \end{bmatrix} \begin{bmatrix} \mathbf{W}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{W} \end{bmatrix} = \begin{bmatrix} \mathbf{X}\mathbf{W}^{-1} & \mathbf{Y}\mathbf{W} \\ -\mathbf{Y}\mathbf{W}^{-1} & \mathbf{X}\mathbf{W} \end{bmatrix} \tag{3.55}$$

The scaling theorems for

$$\mathbf{W} = \begin{bmatrix} w_x & \mathbf{0} \\ \mathbf{0} & w_y \end{bmatrix} \tag{3.56}$$

can be formulated for important RCTs—signal rotator, fractional FT, and gyrator transforms, relatively—as follows.[4, 29, 30]

$$\begin{aligned}
&\mathcal{R}^{\mathbf{U}_r(\alpha)}[f(\mathbf{W}\mathbf{r}_i)](\mathbf{r}_o) = f(\mathbf{X}_r(-\alpha)\mathbf{W}\mathbf{r}_o) \\
&\mathcal{R}^{\mathbf{U}_f(\gamma_x,\gamma_y)}[f(\mathbf{W}\mathbf{r}_i)](\mathbf{r}_o) \\
&\quad = \left(\frac{\cos\beta_x}{\cos\gamma_x}\right)^{1/2} \exp\left\{i\pi x_o^2 \left[1 - \left(\frac{\cos\beta_x}{\cos\gamma_x}\right)^2\right] \cot\gamma_x\right\} \\
&\qquad \times \left(\frac{\cos\beta_y}{\cos\gamma_y}\right)^{1/2} \exp\left\{i\pi y_o^2 \left[1 - \left(\frac{\cos\beta_y}{\cos\gamma_y}\right)^2\right] \cot\gamma_y\right\} \\
&\qquad \times \mathcal{R}^{\mathbf{U}_f(\beta_x,\beta_y)}[f(\mathbf{r}_i)]\left(\frac{\cos\beta_x}{\cos\gamma_x} w_x x_o, \frac{\cos\beta_y}{\cos\gamma_y} w_y y_o\right) \\
&\mathcal{R}^{\mathbf{U}_g(\vartheta)}[f(\mathbf{W}\mathbf{r}_i)](\mathbf{r}_o) \\
&\quad = \left|\frac{\cos\zeta}{\cos\vartheta}\right| \exp\left\{i2\pi x_o y_o \left[1 - \left(\frac{\cos\zeta}{\cos\vartheta}\right)^2\right] \cot\vartheta\right\} \\
&\qquad \times \mathcal{R}^{\mathbf{U}_g(\zeta)}[f(\mathbf{r}_i)]\left(\frac{\cos\zeta}{\cos\vartheta}\mathbf{W}\mathbf{r}_o\right)
\end{aligned} \tag{3.57}$$

where $\cot\gamma_{x,y} = w_{x,y}^2 \cot\beta_{x,y}$, and $\cot\vartheta = w_x w_y \cot\zeta$. Note that if $w_x = w_y^{-1} = w$, then $\mathcal{R}^{\mathbf{U}_g(\vartheta)}[f(\mathbf{W}\mathbf{r}_i)](\mathbf{r}_o) = \mathcal{R}^{\mathbf{U}_g(\vartheta)}[f(\mathbf{r}_i)](\mathbf{W}\mathbf{r}_o)$.

The scaling property for the fractional FT has been used for the analysis of fractal signals.[38]

3.4.6 Phase-Space Rotations of Selected Functions

Phase-space rotation of only a limited number of functions can be expressed analytically. Among them there is the function

$$f_i(\mathbf{r}) = \exp\left(i2\pi \mathbf{k}_i^t \mathbf{r} - \pi \mathbf{r}^t \mathbf{L}_i \mathbf{r}\right) \tag{3.58}$$

where $\mathbf{L}_i$ is a symmetric matrix with nonnegative definite real part and $\mathbf{k}_i$ is a real vector. Following the calculations done in Refs. 7 and 42, one can find that the RCT of the function (3.58) takes the form

$$f_o(\mathbf{r}) = \mathcal{R}^{\mathbf{U}}[f_i(\mathbf{r}_i)](\mathbf{r}) = [\det(\mathbf{X} + i\mathbf{Y}\mathbf{L}_i)]^{-1/2} \times \exp\left[-i\pi\mathbf{k}_i^t(\mathbf{X} + i\mathbf{Y}\mathbf{L}_i)^{-1}\mathbf{Y}\mathbf{k}_i + i2\pi\mathbf{k}_o^t\mathbf{r} - \pi\mathbf{r}^t\mathbf{L}_o\mathbf{r}\right] \quad (3.59)$$

where $\mathbf{k}_o^t = \mathbf{k}_i^t(\mathbf{X} + i\mathbf{Y}\mathbf{L}_i)^{-1}$ and $i\mathbf{L}_o = (-\mathbf{Y} + i\mathbf{X}\mathbf{L}_i)(\mathbf{X} + i\mathbf{Y}\mathbf{L}_i)^{-1}$.

If $\mathbf{L}_i = -i\mathbf{H}_i$ is imaginary, then $\mathbf{L}_o = -i\mathbf{H}_o = -i(-\mathbf{Y} + \mathbf{X}\mathbf{H}_i)\times(\mathbf{X} + \mathbf{Y}\mathbf{H}_i)^{-1}$ is imaginary, too, which implies that $f_i(\mathbf{r}_i)$ [Eq. (3.58)] and $f_o(\mathbf{r})$ [Eq. (3.59)] are the generalized chirp functions, which include as particular cases the plane, elliptic, hyperbolic, and parabolic waves.

For $\mathbf{k}_i = \mathbf{0}$, and thus $f_i(\mathbf{r}) = \exp(-\pi\mathbf{r}^t\mathbf{L}_i\mathbf{r})$, Eq. (3.59) reduces to

$$f_o(\mathbf{r}) = [\det(\mathbf{X} + i\mathbf{Y}\mathbf{L}_i)]^{-1/2}\exp\left(-\pi\mathbf{r}^t\mathbf{L}_o\mathbf{r}\right) \quad (3.60)$$

A Gaussian beam $\exp[-\pi(l_{11}x^2 + 2l_{12}xy + l_{22}y^2)]$ appears when $\mathbf{L}_i$ is real and positive definite.

For plane wave $f_i(\mathbf{r}) = \exp(i2\pi\mathbf{k}_i^t\mathbf{r})$ ($\mathbf{L}_i = \mathbf{0}$) we obtain

$$f_o(\mathbf{r}) = (\det\mathbf{X})^{-1/2}\exp\left(-i\pi\mathbf{k}_i^t\mathbf{X}^{-1}\mathbf{Y}\mathbf{k}_i + i2\pi\mathbf{k}_i^t\mathbf{X}^{-1}\mathbf{r} - i\pi\mathbf{r}^t\mathbf{Y}\mathbf{X}^{-1}\mathbf{r}\right) \quad (3.61)$$

Then a plane wave remains a plane wave only under the imaging-type phase-space rotations ($\mathbf{Y} = \mathbf{0}$).

Equation (3.61) can be used for the calculation of the phase-space rotations of periodic functions. Thus by representing a periodic function $f_i(\mathbf{r})$ with periods p_x and p_y with respect to the x and y coordinates as a superposition of plane waves,

$$f_i(\mathbf{r}) = \sum_{m,n=-\infty}^{\infty} a_{mn}\exp(i2\pi\mathbf{k}_{mn}^t\mathbf{r}) \quad (3.62)$$

with $\mathbf{k}_{mn}^t = (m/p_x, n/p_y)$ and using Eq. (3.61), we get after the RCT

$$f_o(\mathbf{r}) = (\det\mathbf{X})^{-1/2}\exp(-i\pi\mathbf{r}^t\mathbf{Y}\mathbf{X}^{-1}\mathbf{r}) \times \sum_{m,n=-\infty}^{\infty} a_{mn}\exp\left(-i\pi\mathbf{k}_{mn}^t\mathbf{X}^{-1}\mathbf{Y}\mathbf{k}_{mn} + i2\pi\mathbf{k}_{mn}^t\mathbf{X}^{-1}\mathbf{r}\right) \quad (3.63)$$

If $\mathbf{k}_{mn}^t\mathbf{X}^{-1}\mathbf{Y}\mathbf{k}_{mn} = j$, where j is an even integer, then the generalized Talbot imaging[7] is obtained

$$f_o(\mathbf{r}) = (\det\mathbf{X})^{-1/2}\exp\left(-i\pi\mathbf{r}^t\mathbf{Y}\mathbf{X}^{-1}\mathbf{r}\right) f_i(\mathbf{X}^{-1}\mathbf{r}) \quad (3.64)$$

It includes a rotation, scaling of the coordinates described by the matrix $\mathbf{X}^{-1}$, and phase modulation associated with the matrix product $-\mathbf{Y}\mathbf{X}^{-1}$.

Phase-space rotation of the Dirac δ function leads to the generalized chirp function, which corresponds to the point-spread function of the related first-order optical system.

$$\mathcal{R}^{\mathbf{U}}[\delta(\mathbf{r}_i - \mathbf{v})](\mathbf{r}_o) = \frac{1}{\sqrt{\det i\mathbf{Y}}} \exp\left[i\pi\left(\mathbf{v}^t\mathbf{Y}^{-1}\mathbf{X}\mathbf{v} - 2\mathbf{v}^t\mathbf{Y}^{-1}\mathbf{r}_o + \mathbf{r}_o^t\mathbf{X}\mathbf{Y}^{-1}\mathbf{r}_o\right)\right] \tag{3.65}$$

Correspondingly applying the inverse transform parameterized by the matrix $\mathbf{U}^{-1}$ to this chirp function, we obtain $\delta(\mathbf{r}_i - \mathbf{v})$. Therefore the phase-space rotations can be used for the localization of certain chirp signals, as will be discussed further.

3.5 Eigenfunctions for Phase-Space Rotators

3.5.1 Some Relations for the Eigenfunctions

It is known that the Hermite-Gaussian (HG) functions $\mathcal{H}_{m,n}(\mathbf{r}) = \mathcal{H}_m(x)\,\mathcal{H}_n(y)$, where $\mathcal{H}_n(x) = 2^{1/4}\,(2^n\,n!)^{-1/2}\,H_n(\sqrt{2\pi}\,x)\,\exp(-\pi x^2)$ and $H_n(\cdot)$ denotes the Hermite polynomials, are eigenfunctions for the separable fractional FT for any angles γ_x and γ_y with eigenvalues $\exp[-i(m+\frac{1}{2})\gamma_x - i(n+\frac{1}{2})\gamma_y]$ (see, for example, Ref. 4).

To find the eigenfunctions for the RCT parameterized by the unitary matrix $\mathbf{U}_s$, first we perform the similarity decomposition $\mathbf{U}_s = \mathbf{U}\mathbf{U}_f(\gamma_x, \gamma_y)\mathbf{U}^{-1}$, where γ_x and γ_y and the matrix $\mathbf{U}$ are defined from the eigenvalues and eigenvectors of $\mathbf{U}_s$ correspondingly. Then it is clear that[39] the functions obtained from $\mathcal{H}_{m,n}(\mathbf{r}_i)$ by the RCT parameterized by $\mathbf{U}$: $\mathcal{H}^{\mathbf{U}}_{m,n}(\mathbf{r}) = \mathcal{R}^{\mathbf{U}}[\mathcal{H}_{m,n}(\mathbf{r}_i)](\mathbf{r})$ are eigenfunctions for the RCT described by the matrix $\mathbf{U}_s$ with eigenvalues $\exp[-i(m+\frac{1}{2})\gamma_x - i(n+\frac{1}{2})\gamma_y]$.

There are various names for $\mathcal{H}^{\mathbf{U}}_{m,n}(\mathbf{r})$ modes: two-dimensional Hermite-Gaussian functions,[40] Hermite-Laguerre Gaussian functions,[35] and orthosymplectic modes.[41] Here we will use the last one since $\mathcal{H}^{\mathbf{U}}_{m,n}(\mathbf{r})$ is an eigenfunction for the RCT parameterized by the orthosymplectic ray transformation matrix associated with $\mathbf{U}_s$.

As well as the HG functions, the modes $\mathcal{H}^{\mathbf{U}}_{m,n}(\mathbf{r})$ for the same $\mathbf{U}$ and different indices $m, n \in [0, \infty)$ form a complete orthonormal set, and therefore, any function can be represented as their linear superposition.

From the fact that $\mathbf{U}_f(\varphi, \varphi)$ commutes with any unitary matrix $\mathbf{U}$ follows that the mode $\mathcal{H}^{\mathbf{U}}_{m,n}(\mathbf{r})$ is an eigenfunction for the symmetric fractional FT with eigenvalue $\exp[-i(m+n+1)\varphi]$. Then the kernel of the symmetric fractional FT, Eq. (3.27), can be alternatively presented as a series of products of the orthosymplectic modes.

$$K^{\mathbf{U}_f(\varphi,\varphi)}(\mathbf{r}_i, \mathbf{r}_o) = \sum_{m,n=0}^{\infty} \exp[-i(m+n+1)\varphi]\, \mathcal{H}^{\mathbf{U}}_{m,n}(\mathbf{r}_0)\mathcal{H}^{\mathbf{U}^{-1}}_{m,n}(\mathbf{r}_i) \tag{3.66}$$

The orthosymplectic modes $\mathcal{H}^{\mathbf{U}}_{m,n}(\mathbf{r})$, modes obtained from the HG ones $\mathcal{H}_{m,n}(\mathbf{r})$ by the RCT associated with matrix $\mathbf{U}$, have the following generating function[40,42]

$$\begin{aligned}\sqrt{\frac{2}{\det \mathbf{U}}}\, &\exp\left(-\mathbf{s}^t\mathbf{U}^{-1}\mathbf{U}^*\mathbf{s} + 2\mathbf{s}^t\mathbf{U}^{-1}\mathbf{r}\sqrt{2\pi} - \pi\mathbf{r}^t\mathbf{r}\right) \\ &= \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} \mathcal{H}^{\mathbf{U}}_{m,n}(\mathbf{r}) \left(\frac{2^{m+n}}{m!n!}\right)^{1/2} s_x^m s_y^n \end{aligned} \tag{3.67}$$

where $\mathbf{s}^t = (s_x, s_y)$. They can be expressed as[40,42]

$$\begin{aligned}\mathcal{H}^{\mathbf{U}}_{m,n}(\mathbf{r}) = &\frac{(-1)^{m+n}\exp[\pi(x^2+y^2)]}{2^{m+n-1/2}\,(\pi^{m+n}m!n!\det\mathbf{U})^{1/2}} \\ &\times \left(U_{11}^*\frac{\partial}{\partial x} + U_{21}^*\frac{\partial}{\partial y}\right)^m \left(U_{12}^*\frac{\partial}{\partial x} + U_{22}^*\frac{\partial}{\partial y}\right)^n \\ &\times \exp[-2\pi(x^2+y^2)]\end{aligned} \tag{3.68}$$

where U_{jk} $(j,k=1,2)$ are parameters of the unitary matrix $\mathbf{U}$. Thus for the separable fractional FT $\mathbf{U} = \mathbf{U}_f(\gamma_x, \gamma_y)$, this formula for any angles γ_x and γ_y reduces to the HG functions up to the constant phase $\exp[-i(m+\frac{1}{2})\gamma_x - i(n+\frac{1}{2})\gamma_y]$.

The orthosymplectic modes satisfy the symmetry relations

$$\begin{aligned}\mathcal{H}^{\mathbf{U}}_{m,n}(-\mathbf{r}) &= (-1)^{m+n}\mathcal{H}^{\mathbf{U}}_{m,n}(\mathbf{r}) \\ \left[\mathcal{H}^{\mathbf{U}}_{m,n}(\mathbf{r})\right]^* &= \mathcal{H}^{\mathbf{U}^{-1}}_{m,n}(\mathbf{r})\end{aligned} \tag{3.69}$$

the derivative relations[39]

$$\begin{aligned}\left[\frac{\partial}{\partial x}, \frac{\partial}{\partial y}\right]^t \mathcal{H}^{\mathbf{U}}_{m,n}(\mathbf{r}) = &2\sqrt{\pi}\,\mathbf{U}^*\left[\sqrt{m}\,\mathcal{H}^{\mathbf{U}}_{m-1,n}(\mathbf{r}), \sqrt{n}\,\mathcal{H}^{\mathbf{U}}_{m,n-1}(\mathbf{r})\right]^t \\ &-2\pi\,\mathcal{H}^{\mathbf{U}}_{m,n}(\mathbf{r})\,[x, y]^t\end{aligned} \tag{3.70}$$

and the recurrence relations

$$2\sqrt{\pi}\,[x,\,y]^t\,\mathcal{H}^{\mathbf{U}}_{m,n}(\mathbf{r}) = \mathbf{U}\left[\sqrt{m+1}\,\mathcal{H}^{\mathbf{U}}_{m+1,n}(\mathbf{r}),\,\sqrt{n+1}\,\mathcal{H}^{\mathbf{U}}_{m,n+1}(\mathbf{r})\right]^t + \mathbf{U}^*\left[\sqrt{m}\mathcal{H}^{\mathbf{U}}_{m-1,n}(\mathbf{r}),\,\sqrt{n}\,\mathcal{H}^{\mathbf{U}}_{m,n-1}(\mathbf{r})\right]^t \quad (3.71)$$

Based on Eqs. (3.70) and (3.71), we can determine[41] the z-OAM component for the orthosymplectic mode $\mathcal{H}^{\mathbf{U}}_{m,n}(\mathbf{r})$.

$$\mathrm{L}_z^{m,n} = \int_{-\infty}^{\infty} \mathrm{Im}\left\{\mathcal{H}^{\mathbf{U}}_{m,n}(\mathbf{r})^*\left(x\,\frac{\partial}{\partial y} - y\,\frac{\partial}{\partial x}\right)\mathcal{H}^{\mathbf{U}}_{m,n}(\mathbf{r})\right\} dx\,dy = 2\,\mathrm{Im}\left\{mU_{11}U_{21}^* - nU_{22}U_{12}^*\right\} \quad (3.72)$$

As an example, let us find the eigenfunctions for the signal rotator. Using the similarity transformation Eq. (3.48), we observe that the orthosymplectic mode $\mathcal{H}^{\mathbf{U}_g(\mp\pi/4+\pi k)}_{m,n}(\mathbf{r})$ is an eigenfunction for the signal rotator for any angle. Note that $\mathcal{H}^{\mathbf{U}_g(\mp\pi/4+\pi k)}_{m,n}$ with integer k corresponds to the helicoidal Laguerre-Gaussian (LG) modes $\mathcal{L}^{\pm}_{m,n}(\mathbf{r})$ apart from the constant phase factor

$$\mathcal{H}^{\mathbf{U}_g(\mp\pi/4+\pi k)}_{m,n}(\mathbf{r}) \propto \mathcal{L}^{\pm}_{m,n}(\mathrm{r}) = 2^{1/2}\left[\frac{(\min\{m,\,n\})!}{(\max\{m,\,n\})!}\right]^{1/2}(\sqrt{2\pi}r)^{|m-n|} \times \exp[\pm i(m-n)\psi]\,L^{(|m-n|)}_{\min\{m,n\}}(2\pi r^2) \times \exp(-\pi r^2) \quad (3.73)$$

where $L_n^{(\alpha)}(\cdot)$ denotes the generalized Laguerre polynomials, and spatial coordinates are represented by the two-dimensional column vector $\mathbf{r} = (x,\,y)^t = (r\cos\psi,\,r\sin\psi)^t$. Therefore, the LG mode $\mathcal{L}^{\pm}_{m,n}(\mathbf{r})$ is an eigenfunction for the signal rotator. From Eq. (3.72) it follows that $\mathcal{L}^{\pm}_{m,n}(\mathbf{r})$ possesses the integer OAM projection $\mathrm{L}_z^{m,n} = \pm(m-n)$, also known as a topological charge.

Correspondingly using the similarity transformation Eq. (3.47) for the gyrator, we conclude that its eigenfunctions are the HG ones rotated at $\mp\pi/4 + \pi k$.

3.5.2 Mode Presentation on Orbital Poincaré Sphere

We have emphasized that a phase-space rotator is described by the unitary matrix $\mathbf{U}$, which has 4 degrees of freedom. Nevertheless $\mathcal{H}^{\mathbf{U}}_{m,n}(\mathbf{r})$ is characterized by only two parameters because it is an eigenfunction for the symmetric fractional FT $R^{\mathbf{U}_f(\varphi,\varphi)}$ and for the RCT associated with matrix $\mathbf{U}_s = \mathbf{U}\mathbf{U}_f(\gamma,-\gamma)\mathbf{U}^{-1}$ for any φ and γ. Let us demonstrate this statement with the following example.

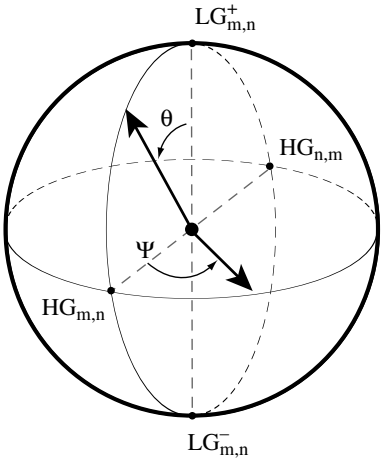


FIGURE 3.4 (m, n)-Poincaré sphere for orthosymplectic mode presentation.

The orthosymplectic mode $\mathcal{H}^{\mathbf{U}}_{m,n}(\mathbf{r})$, Eq. (3.68), can be obtained from the HG one, $\mathcal{H}_{m,n}(\mathbf{r})$, by application of the RCT associated with the matrix $\mathbf{U}$ or alternatively from the LG mode as $\mathcal{H}^{\mathbf{U}}_{m,n}(\mathbf{r}) = R^{\mathbf{U}_l}[\mathcal{L}^{\pm}_{m,n}(\mathbf{r}_i)](\mathbf{r})$, where $\mathbf{U}_l = \mathbf{U} \times \mathbf{U}_g(\pm\pi/4)$ [see Eq. (3.73)]. Since $\mathbf{U}_l$ can be written as $\mathbf{U}_l = \mathbf{U}_r(\alpha)\mathbf{U}_f(\gamma, -\gamma)\mathbf{U}_r(\beta)\mathbf{U}_f(\varphi, \varphi)$ and the LG modes are eigenfunctions for the symmetrical fractional FT and the signal rotator, we conclude that all different orthosymplectic modes $\mathcal{H}^{\mathbf{U}}_{m,n}(\mathbf{r})$ can be generated from the LGs by two-parameter RCTs described by matrix $\mathbf{U}_r(\alpha)\mathbf{U}_f(\gamma, -\gamma)$.

It has been proposed to present all different orthosymplectic modes $\mathcal{H}^{\mathbf{U}}_{m,n}(\mathbf{r})$ (here a constant phase factor of the mode is ignored) for fixed indices m and n on the sphere called the orbital (m, n)-Poincaré sphere,[43–45] which is similar to the one used for characterization of polarized light (see Fig. 3.4).

For example, by starting from the LG mode $\mathcal{L}^{+}_{m,n}(\mathbf{r}) = \mathcal{L}^{(0,.)}_{m,n}(\mathbf{r})$, living on the north pole of the (m, n)-Poincaré sphere, and applying the RCT associated with two-parameter matrix $\mathbf{U}(\theta, \psi) = \mathbf{U}_r(-\pi/4 + \psi/2)\,\mathbf{U}_f(\theta/2, -\theta/2)\,\mathbf{U}_r(\pi/4 - \psi/2)$ to this mode, the entire sphere can be populated by the different orthosymplectic modes $\mathcal{L}^{(\theta,\psi)}_{m,n}(\mathbf{r}) = R^{\mathbf{U}(\theta,\psi)}[\mathcal{L}^{+}_{m,n}(\mathbf{r}_i)](\mathbf{r})$, where the parameters $\theta \in [0, \pi]$ and $\psi \in [-\pi, \pi]$ indicate the colatitude of a parallel and the longitude of a meridian on the sphere, respectively. The HG modes $\mathcal{H}_{m,n}(\mathbf{r})$ and $\mathcal{H}_{n,m}(\mathbf{r})$ are located at the intersection of the main meridian and equator at points $(\theta,\psi) = (\pi/2, 0)$ and $(\pi/2, \pi)$, respectively. Moreover, it has been shown[43] that the transformation along the main meridian

$\psi = 0$ corresponds to the gyrator transform Eq. (3.16), along the meridian with $\psi = \pi/2$—to the antisymmetric fractional FT—and along the equator—to the signal rotator transform Eq. (3.14). Thus the HG mode $\mathcal{H}_{m,n}(\mathbf{r})$ rotated counterclockwise at angle $\psi/2$ lives on the equator at longitude ψ.

It is easy to see from Eq. (3.72) that the modes from the same colatitude have the same projection of the OAM along the propagation direction $\mathrm{L}_z^{m,n} = (m-n)\cos\theta$. Then for the LG mode $\mathcal{L}_{m,n}^{\pm}(\mathbf{r})$ the value of the projection $\mathrm{L}_z^{m,n} = \pm(m-n)$ is an integer; meanwhile $\mathrm{L}_z^{m,n} = 0$ for HG mode $\mathcal{L}_{m,n}^{(\pi/2,\psi)}(\mathbf{r})$.

Correspondingly, any orthonormal set $\{\mathcal{H}_{m,n}^{\mathrm{U}}(\mathbf{r})\}$ with integer $m, n \in [0, \infty)$ is characterized by two parameters and can be associated with a certain direction (θ,ψ) in three-dimensional parametric space.

3.6 Optical Setups for Basic Phase-Space Rotators

It is well known (e.g., see Ref. 10) that in paraxial optics the Fourier transform can be performed using a convergent thin lens. Thus the complex field amplitude at the back focal plane of the lens corresponds to the FT of one at the front focal plane. As derived in Ref. 28 and discussed in Sec. 1.5, the symmetric fractional FT at angle φ can be also performed by the same scheme if the distance z between the input/output plane and the lens of focal distance f equals $z = 2f\,\sin^2(\varphi/2)$. Another proposed scheme[28] consists of two identical spherical convergent lenses of focal distance f located at the input and output system planes with the distance $z = 2f\,\sin^2(\varphi/2)$ between them. Moreover, the propagation of the optical beam through the optical fiber with a quadratic refractive index profile also produces the symmetric fractional FT at angles defined by the propagation distance and the refractive index gradient.[14,29]

To perform the separable fractional FT, as well as signal rotator and gyrator, the cylindrical lenses are needed. Several setups for separable fractional FT,[46–50] antisymmetric fractional FT,[51] and signal rotator[21,22] have been proposed. Nevertheless most are difficult to adapt to the often needed change of transformation parameter. Moreover, the parameter turning is usually accompanied by additional scaling which depends on the transformation parameter.

The main objective of the system design is to find a minimal lens–free-space configuration that is flexible for transformation parameter changing. A setup with fixed free-space intervals between the generalized lenses is a promising candidate for this task.

The generalized lens,[52,53] which can be mathematically described by the CT parameterized by ray transformation matrix $\mathbf{T}_L$, Eq. (3.10),

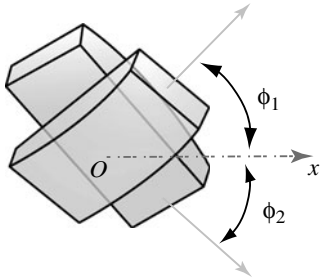


FIGURE 3.5 Generalized lens constructed from two cylindrical lenses rotated at different angles.

with the block matrix

$$\mathbf{G} = \begin{bmatrix} g_{xx} & g_{xy} \\ g_{xy} & g_{yy} \end{bmatrix} \tag{3.74}$$

produces the quadratic phase modulation of the input wavefront

$$f_o(x, y) = \exp[-i\pi(g_{xx}x^2 + 2g_{xy}xy + g_{yy}y^2)]\, f_i(x, y) \tag{3.75}$$

In practice, the generalized lens can be implemented by a *spatial light modulator* (SLM) that allows one to change the lens parameters almost in real time. Also it can be constructed as a combination of n aligned cylindrical lenses of power p_j ($p_j > 0$ for convergent lens), which are attached one to another and rotated counterclockwise with respect to the transversal OX axis at angles ϕ_j. Then $g_{xx} = \sum_{j=1}^{n} p_j \cos^2 \phi_j$, $g_{xy} = -\sum_{j=1}^{n} p_j (\sin 2\phi_j)/2$, and $g_{yy} = \sum_{j=1}^{n} p_j \sin^2 \phi_j$. Depending on the angles and the powers of the cylindrical lenses, we obtain the elliptic (including spherical), hyperbolic, or parabolic phase modulations. In Fig. 3.5 the generalized lens that contains only two cylindrical lenses is displayed.

Below we will consider flexible optical schemes with fixed location of the generalized lenses which implement the basic phase-space rotators.

3.6.1 Flexible Optical Setups for Fractional FT and Gyrator

Based on the matrix formalism, flexible optical setups, which perform the fractional FT $\mathcal{R}^{\mathbf{U}_f(\gamma_x, \gamma_y)}$, and the gyrator $\mathcal{R}^{\mathbf{U}_g(\vartheta)}$ have been designed.[21,33,54] These optical schemes contain three generalized lenses $\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{L}_3$; the last is identical to $\mathcal{L}_1$, with fixed equal

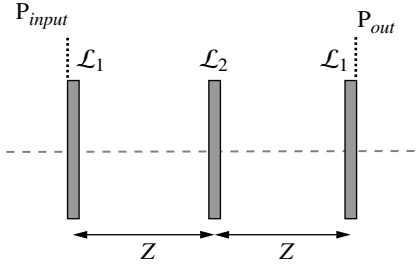


FIGURE 3.6 Scheme for the flexible setup performing the fractional FT and gyrator.

distances between them denoted by z. Lenses $\mathcal{L}_1$ and $\mathcal{L}_3$ are located at the input and output planes, as indicated in Fig. 3.6.

The block matrices for lenses $\mathcal{L}_1$ and $\mathcal{L}_2$ corresponding to the separable fractional Fourier transformer $\mathcal{R}^{\mathbf{U}_f(\gamma_x,\gamma_y)}$ have the diagonal forms

$$\mathbf{G}_1 = \frac{1}{z}\begin{bmatrix} 1-\frac{1}{2}\cot(\gamma_x/2) & 0 \\ 0 & 1-\frac{1}{2}\cot(\gamma_y/2) \end{bmatrix}$$

$$\mathbf{G}_2 = \frac{2}{z}\begin{bmatrix} 1-\sin\gamma_x & 0 \\ 0 & 1-\sin\gamma_y \end{bmatrix} \tag{3.76}$$

In the case of the separable fractional or symmetric fractional FT, the required generalized lenses can be implemented only by the SLM. Note that this scheme simplifies the acquisition of the Wigner distribution projections, needed for its reconstruction by the phase-space tomography method, discussed in Chap. 4.

For the gyrator transform $\mathcal{R}^{\mathbf{U}_g(\vartheta)}$ the generalized lenses are expressed as

$$\mathbf{G}_1 = \frac{1}{z}\begin{bmatrix} 1 & -\cot(\vartheta/2) \\ -\cot(\vartheta/2) & 1 \end{bmatrix}$$

$$\mathbf{G}_2 = \frac{2}{z}\begin{bmatrix} 1 & -\frac{1}{2}\sin\vartheta \\ -\frac{1}{2}\sin\vartheta & 1 \end{bmatrix} \tag{3.77}$$

In the case of the gyrator and antisymmetric fractional FT, the required generalized lenses can be obtained as a superposition of ordinary cylindrical lenses. Then the transformation angle is changed by rotation of the cylindrical lenses which form the generalized lenses.

A generalized lens constructed from two convergent cylindrical lenses of the same power p rotated at angles ϕ_1 and ϕ_2 is characterized by

$$\mathbf{G} = p \begin{bmatrix} \cos^2\phi_1 + \cos^2\phi_2 & -\frac{1}{2}[\sin(2\phi_1) + \sin(2\phi_2)] \\ -\frac{1}{2}[\sin(2\phi_1) + \sin(2\phi_2)] & \sin^2\phi_1 + \sin^2\phi_2 \end{bmatrix} \tag{3.78}$$

which reduces to

$$\mathbf{G} = p \begin{bmatrix} 1 & -\sin(2\phi_1) \\ -\sin(2\phi_1) & 1 \end{bmatrix} \tag{3.79}$$

for $\phi_2 = -\phi_1 \pm \pi/2$. Comparing Eqs. (3.77) and (3.79), we conclude that in the gyrator setup every generalized lens $\mathcal{L}_j$ $(j = 1, 2)$ is a combination of two convergent cylindrical lenses of equal focal distance z/j rotated counterclockwise at angles $\phi_1^{(j)} = \phi^{(j)}$ and $\phi_2^{(j)} = -\phi^{(j)} \pm \pi/2$ with respect to the OX axis. The gyrator at angle ϑ is achieved if $\sin(2\phi^{(1)}) = \cot(\vartheta/2)$ and $2\sin(2\phi^{(2)}) = \sin\vartheta$. We observe that this setup is able to perform the gyrator for the angles from the π interval $[\pi/2, 3\pi/2]$. The experimental implementation of this optical system has been demonstrated in Refs. 31 and 33 on the example of orthosymplectic mode conversion. The experimental results are in good agreement with theoretic predictions.

If the angles in Eq. (3.78) are chosen as $\phi_1^{(j)} = \phi^{(j)} + \pi/4$ and $\phi_2^{(j)} = \phi^{(j)} - \pi/4$, we obtain the generalized lenses suitable for the antisymmetric fractional FT setup.

$$\mathbf{G}_j = \frac{j}{z} \begin{bmatrix} 1 - \sin(2\phi^{(j)}) & 0 \\ 0 & 1 + \sin(2\phi^{(j)}) \end{bmatrix} \tag{3.80}$$

Indeed, comparing Eqs. (3.76) and (3.80), we observe that these lens combinations perform the antisymmetric fractional FT at angle $(\gamma, -\gamma)$, where $2\sin(2\phi^{(1)}) = \cot(\gamma/2)$ and $2\phi^{(2)} = \gamma$. It is easy to see from the last relation that this setup is able to perform the antisymmetric fractional FT for the angles $\gamma \in [\pi/2, 3\pi/2]$ that cover a π interval needed for the different applications, discussed in Sec. 3.7.

3.6.2 Flexible Optical Setup for Image Rotator

Usually, Dove prisms are used for optical signal rotator realization. But the diffraction effects during the propagation through the prisms require additional optical elements for their compensation. Here we consider the optical signal rotator based on the application of cylindrical lenses.[21, 22, 55]

A flexible optical scheme performing a rotation at angle α by only the appropriate rotation of cylindrical lenses composing the setup has

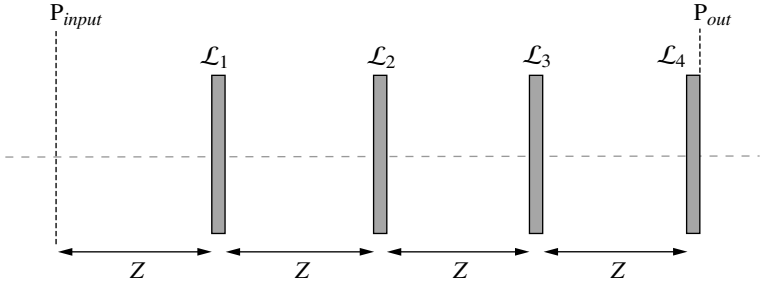


FIGURE 3.7 Scheme for the flexible setup performing the signal rotator.

been recently proposed.[54] It has been shown that four is a minimal number of generalized lenses located in fixed positions needed to preform the signal rotator. The optical scheme of the signal rotator at angle α is displayed in Fig. 3.7. If the distances between all elements equal z, then the block matrices of the applied lenses are given by

$$\begin{aligned}
\mathbf{G}_1(\alpha) &= \mathbf{G}_3(-\alpha) = \frac{1}{2z}\begin{bmatrix} 3+\cos\alpha & \sin\alpha \\ \sin\alpha & 5-\cos\alpha \end{bmatrix} \\
\mathbf{G}_2(\alpha) &= \frac{4}{z}\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \\
\mathbf{G}_4(\alpha) &= \frac{2}{z}\begin{bmatrix} 1+\cos\alpha & -\sin\alpha \\ -\sin\alpha & 1-\cos\alpha \end{bmatrix}
\end{aligned} \tag{3.81}$$

This scheme can be realized by implementation of analog generalized lenses, which consist of three cylindrical lenses for $\mathcal{L}_1$ and $\mathcal{L}_3$ and one cylindrical lens for $\mathcal{L}_2$ and $\mathcal{L}_4$.

By using the optical setups performing the basic phase-space rotations, other phase-space rotators can be constructed as their cascades. Nevertheless there is no guarantee that the obtained setup is optimal.

3.7 Applications of Phase-Space Rotators

3.7.1 Generalized Convolution

As mentioned before, the well-known phase-space rotator—the Fourier transform—plays a crucial role in signal and image processing. It forms a base for shift-invariant filtering which is used for pattern recognition, denoising, encryption, etc. Many good books are devoted to this subject; see, e.g., Refs. 9 and 10. Here we consider the application

of other phase-space rotators for signal processing tasks, which can be elegantly expressed in the framework of generalized convolution.

The convolution operation between signals f and h, Eq. (3.51), can be alternatively expressed via the Fourier transform as[9]

$$\mathcal{C}_{f,h}(\mathbf{r}) = (f * h)(\mathbf{r}) = \mathcal{F}^{-1}\left\{\mathcal{F}[f(\cdot)](\mathbf{u})\,\mathcal{F}[h(\cdot)](\mathbf{u})\right\}(\mathbf{r}) \tag{3.82}$$

By analogy we can introduce the *generalized* (canonical) *convolution* (GC) operation as[6,8,56]

$$\mathcal{GC}_{f,h}(\mathbf{T}_1, \mathbf{T}_2, \mathbf{T}_3, \mathbf{r}) = \mathcal{R}^{\mathbf{T}_3}\{\mathcal{R}^{\mathbf{T}_1}[f(\cdot)](\mathbf{u})\,\mathcal{R}^{\mathbf{T}_2}[h(\cdot)](\mathbf{u})\}(\mathbf{r}) \tag{3.83}$$

where the FT operators $\mathcal{F}$ are substituted by the CT ones $\mathcal{R}^{\mathbf{T}}$. In the widely used GCs, $\mathcal{R}^{\mathbf{T}}$ corresponds to the RCT $\mathcal{R}^{\mathbf{U}}$ and can be denoted by $\mathcal{GC}_{f,h}(\mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3, \mathbf{r})$.

It is easy to see[6] that Eq. (3.83) reduces to common convolution, Eq. (3.82), if the ray transformation matrices correspond to the direct/inverse FT ones $\mathbf{U}_1 = \mathbf{U}_2 = \mathbf{U}_3^{-1} = \mathbf{U}_f(\pi/2, \pi/2)$. Besides that the GC includes, as particular cases, the correlation operation

$$\mathrm{Cor}_{f,h}(\mathbf{r}) = \mathcal{GC}_{f,h^*}\left[\mathbf{U}_f\left(\frac{\pi}{2}, \frac{\pi}{2}\right), \mathbf{U}_f\left(-\frac{\pi}{2}, -\frac{\pi}{2}\right), \mathbf{U}_f\left(-\frac{\pi}{2}, -\frac{\pi}{2}\right)\mathbf{r}\right] \tag{3.84}$$

used as a measure of similarity between two signals f and h;[9] the fractional convolution

$$\mathcal{GC}_{f,h}[\mathbf{U}_f(\gamma_x, \gamma_y), \mathbf{U}_f(\beta_x, \beta_y), \mathbf{U}_f(\alpha_x, \alpha_y), \mathbf{r}] \tag{3.85}$$

applied for shift-variant filtering and pattern recognition;[4,15] the Wigner distribution

$$2\mathcal{GC}_{f,f^*}\left[\mathbf{U}_f\left(\gamma_x + \frac{\pi}{2}, \gamma_y + \frac{\pi}{2}\right), \mathbf{U}_f\left(-\gamma_x + \frac{\pi}{2}, -\gamma_y + \frac{\pi}{2}\right), \mathbf{U}_f\left(-\frac{\pi}{2}, -\frac{\pi}{2}\right), 2\boldsymbol{\rho}\right] \tag{3.86}$$

expressed in polar coordinates $\boldsymbol{\rho} = (\sqrt{x^2 + p_x^2}, \sqrt{y^2 + p_y^2})$; and the RCT power spectrum $\mathcal{GC}_{f,f^*}(\mathbf{U}, \mathbf{U}^{-1}, \mathbf{I}, \mathbf{r}) = |\mathcal{R}^{\mathbf{U}}[f](\mathbf{r})|^2$, corresponding to the squared modulus of the RCT of the signal or Wigner distribution projection, which in the case $\mathbf{U} = \mathbf{U}_f(\gamma_x, \gamma_y)$ is denoted as the Radon-Wigner transform (see Chap. 4)

$$\mathcal{GC}_{f,f^*}[\mathbf{U}_f(\gamma_x, \gamma_y), \mathbf{U}_f(-\gamma_x, -\gamma_y), \mathbf{I}, \mathbf{r}] = |\mathcal{R}^{\mathbf{U}_f(\gamma_x, \gamma_y)}[f](\mathbf{r})|^2 \tag{3.87}$$

The generalized convolution $\mathcal{GC}_{f,h}(\mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3, \mathbf{r})$ of two-dimensional signals f and h is a function of 2 variables ($\mathbf{r}$) and 12

parameters, defined by the matrices $\mathbf{U}_1$, $\mathbf{U}_2$, and $\mathbf{U}_3$. Some of the parameters can also play the role of variables. The choice of the parameters and the number of variables of the GC depends on the particular application. Thus, if we are interested in the improvement of image quality or in its manipulation for some feature extraction (e.g., edge enhancement or image deblurring), then we have to choose $\mathbf{U}_3 = \mathbf{U}_1^{-1}$ to represent the result of filtering in the position domain.

A typical optical scheme for GC, $\mathcal{GC}_{f,h}(\mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3, \mathbf{r})$, is a straightforward generalization of the Van der Lugt processor[9] and consists of a cascade of two first-order systems described by the matrices $\mathbf{U}_1$ and $\mathbf{U}_3$ (the flexible schemes for the phase-space rotators were considered in Sec. 3.6) with a diffraction/reflection screen between them corresponding to multiplication of the passing/reflecting beam by $\mathcal{R}^{\mathbf{U}_2}[h(\cdot)]$. Then with $f(\cdot)$ in the input of this system, we have its GC with $h(\cdot)$ at the output plane. The common convolution operation $\mathcal{C}_{f,h}(\mathbf{r})$, Eg. (3.82), arises when $\mathbf{U}_1$ and $\mathbf{U}_2$ correspond to the direct FT matrices and $\mathbf{U}_3$ to the inverse one. In optical realization, $\mathbf{U}_3$ is usually the direct FT, and then we have $\mathcal{C}_{f,h}(-\mathbf{r})$ at the output plane.

3.7.2 Pattern Recognition

The correlation operation $\mathrm{Cor}_{f,h}(\mathbf{r})$ is a measure of the similarity between two signals f and h. The mathematical verification of this statement is related to the inequality of Schwarz, which permits one to discriminate two signals of equal energy, since in this case the autocorrelation peak $|\mathrm{Cor}_{f,f}(\mathbf{0})|$ is larger than the cross-correlation one $\left|\mathrm{Cor}_{f,h}(\mathbf{0})\right|$. Note that $|\mathrm{Cor}_{f,f}(\mathbf{0})|$ has a maximum in the origin of the coordinates $\mathbf{r} = \mathbf{0}$. Then by applying the appropriate threshold to the correlation map $|\mathrm{Cor}_{f,h}(\mathbf{r})|$, the pattern associated with h can be found on the investigated scene f. Moreover, because the correlation is shift-invariant, $\mathrm{Cor}_{f(\mathbf{r}_i-\mathbf{v}),h(\mathbf{r}_i)}(\mathbf{r}) = \mathrm{Cor}_{f(\mathbf{r}_i),h(\mathbf{r}_i)}(\mathbf{r}-\mathbf{v})$, the positions of all patterns h, if there are several, can be localized. This operation is also performed by the Van der Lugt processor using $\mathcal{F}^{-1}[h^*(\cdot)](\mathbf{u})$ as a filter mask.

Let us consider as an example a set of numbers presented in Fig. 3.8*a*. The amplitude of the numerically simulated cross-correlation between this image and the reference one (Fig. 3.8*b*), is given in Fig. 3.8*c*. The largest peaks are observed at the positions where 0 is written, which permits its localization. Note that the value of the cross-correlation peaks depends on the similarity between 0 and other numbers. Thus, a relatively large peak is also observed in the end of the middle line where 8 is written. More sophisticated filters are usually used for better object discrimination.

If the pattern has to be detected only in a certain region of the scene, then we must apply the fractional FT convolution,[4,15,19,48] Eq. (3.85),

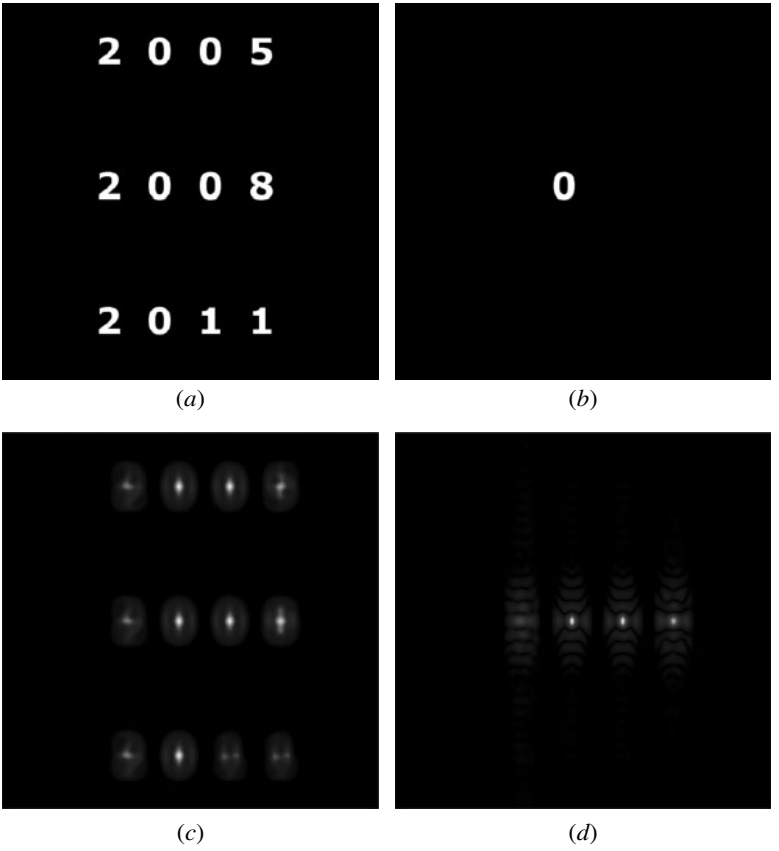


FIGURE 3.8 (*a*) The analyzed image, (*b*) the reference image, (*c*) the amplitude of their correlation in the Fourier domain, and (*d*) the amplitude of their correlation in the fractional Fourier domain for $\gamma_x = \pi/2$ and $\gamma_y = \pi/4$.

with $\mathbf{U}_1 = \mathbf{U}_2^{-1} = \mathbf{U}_f(\gamma_x, \gamma_y)$ and $\mathbf{U}_3 = \mathbf{U}_f(-\pi/2, -\pi/2)$, which provides the shift-variant pattern recognition. The shift tolerance condition is usually expressed in the form[4,15] $\pi v_{x,y}\sigma_{x,y}\cot\gamma_{x,y} \ll 1$, where $v_{x,y}$ is the allowed shift of the pattern on the scene with respect to the reference one used for filter design and $\sigma_{x,y}$ is the pattern width in the x and y directions, correspondingly.

Thus, if we choose different fractional angles for two orthogonal coordinates, such that $\gamma_x = \pi/2$ and $\gamma_y = \pi/4$ and the same filter as we have used before, then 0 will be recognized only on the middle line of the number set, as shown in Fig. 3.8*d*, where the amplitude of the fractional correlation is displayed. Therefore, the fractional correlation

is a useful tool for shift-variant pattern recognition. This operation can be performed by a fractional Van der Lugt correlator, which is a modification of the common one, where the first part is replaced by the fractional FT system.

To maximize the Horner efficiency of the correlation operation, phase-only filters are often used. It was shown in Ref. 57 that, in general, the phase of the fractional FT for $\varphi \neq n\pi$ with integer n contains more information about the signal than the amplitude, and therefore, the phase-only filters can also be applied in the fractional FT domains. To demonstrate, let us analyze the reconstruction of the test image, Fig. 3.1*a*, from only the phase or only the amplitude of its symmetrical fractional FT for different angles φ. Thus if we introduce the notation $\mathcal{R}^{\mathbf{U}_f(\varphi,\varphi)}[f_i(\mathbf{r}_i)](\mathbf{r}_o) = A_\varphi(\mathbf{r}_o)\exp[i\psi_\varphi(\mathbf{r}_o)]$, where $A_\varphi \geq 0$ and ψ_φ are the amplitude and the phase of the fractional FT of the image, then the considered operations are expressed as $\mathcal{R}^{\mathbf{U}_f(-\varphi,-\varphi)}[\exp[i\psi_\varphi(\mathbf{r}_i)]](\mathbf{r}_o)$ and $\mathcal{R}^{\mathbf{U}_f(-\varphi,-\varphi)}[A_\varphi(\mathbf{r}_i)](\mathbf{r}_o)$. In Fig. 3.9 the amplitudes of the image

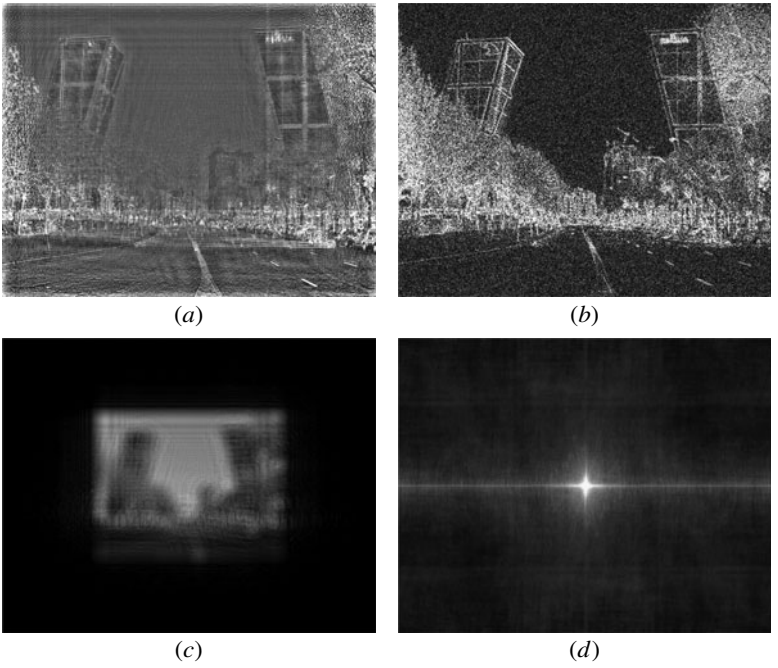


FIGURE 3.9 The amplitudes of the image reconstructed from the phase-only data (parts *a* and *b*) and the amplitude-only data (parts *c* and *d*) of the symmetric fractional FT of the test image for two transformation angles φ: $\varphi = \pi/4$ (parts *a* and *c*) and $\varphi = \pi/2$ (parts *b* and *d*).

reconstructed numerically from the phase-only data (parts *a* and *b*) and the amplitude-only data (parts *c* and *d*) of the symmetric fractional FT of the test image are displayed for two transformation angles φ: $\varphi = \pi/4$ (parts *a* and *c*) and $\varphi = \pi/2$ (parts *b* and *d*). We observe that the information about the structure of the image codified in the amplitude of the fractional FT transform is poor for $\varphi \neq n\pi$. Meanwhile the fractional FT phase contains essential information about the signal almost for all ranges of φ. Thus we can conclude that the phase information of the fractional FT is more relevant than the amplitude one, where we exclude rather exotic images whose fractional FTs have constant phase. The same results are also valid for other phase-space rotators[32] excluding ones of the imaging type.

If the pattern on the scene is rotated with respect to the reference one, then we can apply the GC where $\mathbf{U}_1 = \mathbf{U}_f(\pi/2, \pi/2)\mathbf{U}_r(\alpha)$ and $\mathbf{U}_2 = \mathbf{U}_3 = \mathbf{U}_f(-\pi/2, -\pi/2)$, or $\mathbf{U}_1 = \mathbf{U}_3^{-1} = \mathbf{U}_f(\pi/2, \pi/2)$ and $\mathbf{U}_2 = \mathbf{U}_f(-\pi/2, -\pi/2)\mathbf{U}_r(\alpha)$. The identification of the largest correlation peak as a function of α indicates the right orientation of the pattern. This analysis for the two mentioned cases, respectively, can be done by adding the flexible rotator system[55] just before the common correlator with an invariable filter mask or using the Van der Lugt correlator with a variable filter mask, which can be obtained by application of the spatial light modulator.

For rotation-invariant pattern recognition, the reference image *h*, presented in the polar coordinates, is decomposed into a linear sum of the circular harmonics.[58]

$$h(r, \varphi) = \sum_{l=-\infty}^{\infty} h_l(r) \exp(il\varphi) = \sum_{l=-\infty}^{\infty} c_l(r, \varphi)$$

$$h_l(r) = \frac{1}{2\pi} \int_0^{2\pi} h(r, \varphi) \exp(-il\varphi)\, d\varphi \tag{3.88}$$

Since the Laguerre-Gaussian functions [see Eq. (3.73)] $\mathcal{L}^{\pm}_{m,n}(\mathbf{r}) = \mathcal{L}^{l}_{p}(r, \varphi)$, where $l = m - n$ and $p = \min\{m, n\}$, form the complete orthonormal set, then $h(r, \varphi)$ can be also represented as their linear superposition

$$h(r, \varphi) = \sum_{l=-\infty}^{\infty} \sum_{p=0}^{\infty} b_{l,p} \mathcal{L}^{l}_{p}(r, \varphi) \tag{3.89}$$

and therefore the circular harmonic $c_l(r, \varphi)$ is a linear superposition of the LG modes with the same index $l = m - n$

$$c_l(r, \varphi) = \sum_{p=0}^{\infty} b_{l,p} \mathcal{L}^{l}_{p}(r, \varphi) \tag{3.90}$$

Thus for rotation-invariant pattern recognition, only one circular harmonic $c_l(r, \varphi)$ (usually with $l = \pm 1$) substitutes the reference image. The application of the combination of the various harmonics limits the rotation invariance for a certain angle range.

3.7.3 Chirp Signal Analysis

Chirp, given, e.g., by Eq. (3.65), is often a part of medical and industrial signals. It may contain valuable information or may correspond to a noise. Then chirp detection, localization, estimation, and, if necessary, elimination are important tasks in signal processing. The chirp, Eq. (3.65), can be easily localized applying the RCT parameterized by $\mathbf{U}^{-1}$ because the output signal becomes a δ function. In particular, the application of the FT, the fractional FT, and the gyrator allows one to localize plane, elliptic, and hyperbolic waves, respectively. Thus, the GC $\mathcal{GC}_{f,f^*}(\mathbf{U}, \mathbf{U}^{-1}, \mathbf{I}, \mathbf{r})$ corresponding to the RCT spectra $|\mathcal{R}^{\mathbf{U}}[f(\mathbf{r}_i)](\mathbf{r})|^2$ with modifying parameters of $\mathbf{U}$, associated with the intensity distributions of the output signal, is suitable for the detection of chirps presented in the signal $f(\mathbf{r}_i)$. Here $\mathbf{r}$ and the parameters of $\mathbf{U}$ are variables of the GC $\mathcal{GC}_{f,f^*}(\mathbf{U}, \mathbf{U}^{-1}, \mathbf{I}, \mathbf{r})$.

For example, if $\mathbf{U} = \mathbf{U}_f(\gamma_x, \gamma_y)$, then elliptic-type chirps can be detected as a local maxima of the Radon-Wigner transform map[59] $|\mathcal{R}^{\mathbf{U}_f(\gamma_x,\gamma_y)}[f(\mathbf{r}_i)](\mathbf{r})|^2$ for $\gamma_x, \gamma_y \in [0, \pi]$. The appropriate filtering in the fractional FT domains has been used for elimination of elliptic chirplike noise and, therefore, image quality improvement.[4] Analogously, the hyperbolic chirps can be localized by analyzing the gyrator power spectra $|\mathcal{R}^{\mathbf{U}_g(\vartheta)}[f(\mathbf{r}_i)](\mathbf{r})|^2$ (Ref. 32).

3.7.4 Signal Encryption

The phase-space rotators are also used for signal encryption. The simple algorithm for optical image encryption consists of random phase filtering in the position and FT domains.[60] It has been recently generalized to the case of random phase filtering in different fractional Fourier[16] and gyrator[32] domains. In these cases, not only the random phase masks but also the orders of the phase-space domains (fractional or gyrator angles) where they are located play the role of encryption keys. It was demonstrated that it is impossible to reconstruct the image by using the correct masks but the wrong phase-space domains.

In general, other phase-space rotators can also be used for signal encryption. Indeed the simple encryption procedure of signal f using phase-space rotators consists of a cascade of N operations: the RCT transform parameterized by matrix $\mathbf{U}_n$ with further resultant multiplication at a random phase mask $\exp(i\phi_n)$ for $n = 1, 2, \ldots, N$, which

can be summarized as

$$F = \exp(i\phi_N) R^{\mathbf{U}_N}[\cdots[\exp(i\phi_2) R^{\mathbf{U}_2}[\exp(i\phi_1) R^{\mathbf{U}_1}[f]]]] \tag{3.91}$$

The decryption procedure is written correspondingly as

$$f = R^{\mathbf{U}_1^{-1}}[\exp(-i\phi_1)\cdots[R^{\mathbf{U}_{N-1}^{-1}}[\exp(-i\phi_{N-1}) R^{\mathbf{U}_N^{-1}}[\exp(-i\phi_N) F]]]] \tag{3.92}$$

The randomness of the phase masks together with a large number of encryption parameters $\mathbf{U}_n$ provides a high security of the encryption procedure. More sophisticated algorithms for signal encryption applying phase-space rotators have been developed in Refs. 17 and 18.

3.7.5 Mode Converters

The Hermite-Gaussian and the helical Laguerre-Gaussian (LG) modes are probably the best-known functions used in optics. Indeed, the transversal field distributions for widely applied laser cavities are described by these modes. Moreover, they, as well as all orthosymplectic modes, are structurally stable which means that, ignoring the scaling, their intensity profiles remain the same during the propagation in homogeneous medium. This is a consequence of the fact that the modes $\mathcal{H}^{\mathbf{U}}_{m,n}(\mathbf{r})$ are eigenfunctions for the symmetric fractional FT, which appear in the Iwasawa decomposition of the Fresnel ray transformation matrix, Eq. (3.10), $\mathbf{T}_O = \mathbf{T}_f(\varphi, \varphi)$.

Although LG and HG modes can be produced directly from laser cavities, it is often needed to switch from one type of mode to another. The simplest and cheapest way to do it is based on cylindrical lens application. Since a Laguerre-Gaussian beam is rotationally symmetric and is an eigenfunction of a symmetric fractional Fourier transformer, there are several first-order optical systems which produce this operation.[25,35,42,52,61,62] Any of the phase-space rotators associated with the matrix[42]

$$\mathbf{U} = \frac{1}{\sqrt{2}} \begin{bmatrix} \exp(i\theta_1) & \pm i \exp(i\theta_2) \\ \pm i \exp(i\theta_1) & \exp(i\theta_2) \end{bmatrix} \tag{3.93}$$

can serve as HG-to-LG mode converter. The LG beams at the output of these systems differ one from another by only a constant phase shift. The special case $\theta_1 = 0, \theta_2 = \pi/2$ has been considered in Ref. 35; meanwhile for $\theta_1 = \theta_2 = 0$, matrix $\mathbf{U}$ reduces to the gyrator matrix $\mathbf{U}_g(\pm\pi/4)$. Moreover, the gyrator transform of HG mode at other angles ϑ generates all possible orthosymplectic modes, beside their rotated replicas.

Therefore the flexible scheme proposed for the gyrator implementation can serve as a tunable mode converter. In Fig. 3.10 the

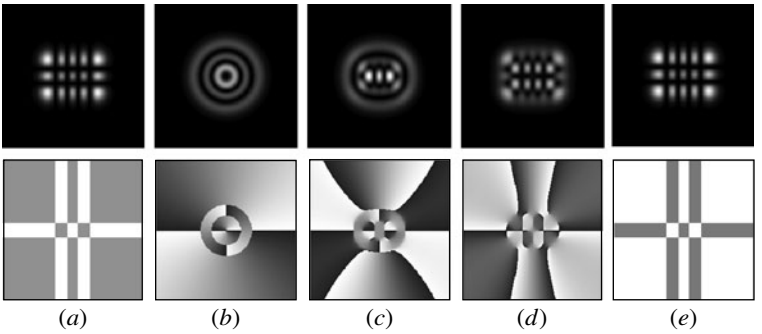


FIGURE 3.10 The amplitude (upper row) and the phase (lower row) of the orthosymplectic modes obtained from the HG $\mathcal{H}_{m,n}(\mathbf{r})$ by gyrator at angles (a) $\vartheta = 0^\circ$, (b) $\vartheta = 135^\circ$, (c) $\vartheta = 150^\circ$, (d) $\vartheta = 165^\circ$, and (e) $\vartheta = 180^\circ$.

transformation of the HG mode by gyrator is illustrated. There the amplitude (upper row) and phase (lower row) of $\mathcal{H}_{m,n}^{\mathbf{U}_g(\vartheta)}(\mathbf{r})$ are displayed for angles (*a*) $\vartheta = 0^\circ$, (*b*) $\vartheta = 135^\circ$, (*c*) $\vartheta = 150^\circ$, (*d*) $\vartheta = 165^\circ$, and (e) $\vartheta = 180^\circ$ correspondingly. The experimental realization of mode conversion by the flexible gyrator setup[31] demonstrates good agreement with numerical calculations.

While the HG and LG modes are widely used in various areas of science and technology, including metrology, interferometry, laser surgery, etc, the application of the other orthosymplectic modes is still under development. It seems that as well as the HG and LG modes they can used for microparticle manipulation.[63] We also emphasize that systems used as mode converters serve for the orbital angular momentum management of coherent as well as partially coherent paraxial light.

3.7.6 Beam Characterization

Since in optics the measurements of the intensity distribution are the only feasible ones, the phase recovering from intensity information is one of the important problems.[64] As mentioned above, the phase-space rotators produce the rotation of the Wigner distribution, which completely characterizes the signal up to the constant phase factor. Moreover, the squared modulus of the RCT of the signal, associated with intensity distribution, corresponds to a certain projection of the Wigner distribution. After exploration of this connection, a method of phase-space tomography was proposed.[65,66] It permits one to reconstruct the Wigner distribution and therefore the complex field amplitude or the mutual intensity for the case of coherent or partially coherent fields, respectively, from the measurements of the intensity

distributions. Note that this method is noninterferometric and noniterative. The reconstruction of the Wigner distribution from the separable fractional FT power spectra, known as the Radon-Wigner transform, is discussed in detail in Chap. 4. Here we only mentioned that the flexible optical setup for the fractional FT considered in Sec. 3.6 permits the almost real-time measurements of the Radon-Wigner transform. It has been shown in Ref. 67 that for the reconstruction of the Wigner distribution of the optical field separable in the Cartesian coordinates, the setup performing the antisymmetric fractional FT can be used. The numerical simulations show that a large number of the Wigner distribution projections, which can be acquired only by a flexible setup, are required for the correct field identification. This stresses the importance of the flexible phase-space rotator setups considered in the previous section.

We mentioned that the orthosymplectic modes $\mathcal{H}^{\mathbf{U}}_{m,n}(\mathbf{r})$ for fixed $\mathbf{U}$ form a complete orthonormal set that permits one to represent a signal as their linear superposition $f(\mathbf{r}) = \sum_{m,n} a^{\mathbf{U}}_{mn} \mathcal{H}^{\mathbf{U}}_{m,n}(\mathbf{r})$. For the determination of the spatial mode spectrum $|a^{\mathbf{U}}_{mn}|^2$ of a complex field amplitude, the two uniparametric phase-space rotators for which $\mathcal{H}^{\mathbf{U}}_{m,n}(\mathbf{r})$ are eigenfunctions are needed. Thus in Ref. 68 the symmetric fractional FT and the signal rotator were applied for the LG spectrum measurements. Another scheme based on the symmetric and anti symmetric fractional FT has been proposed for the determination of the HG spectrum.[69]

The optical field is often represented not by the Wigner distribution itself, which for the two-dimensional case is a function of four variables, but by its global moments (see Sec. 1.7). Thus beam characterization by means of 10 the second-order WD moments,[26,54,70,71] defined in Eq. (1.25), became the basis of the international standard. All these moments can be found from the measurements of the four fractional FT power spectra[72] $|\mathcal{R}^{\mathbf{U}_f(\gamma_x,\gamma_y)}[f(\mathbf{r}_i)](\mathbf{r})|^2$ with at least one of them corresponding to the different angles $\gamma_x \neq \gamma_y$.

It has been shown in Ref. 73 [see also Eq. (1.66)] that the eight normalized second-order moments can be combined into four linear superpositions

$$\begin{aligned} Q_0 &= \tfrac{1}{2}[m_{xx} + m_{uu} + m_{yy} + m_{vv}] \\ Q_1 &= \tfrac{1}{2}[m_{xx} + m_{uu} - (m_{yy} + m_{vv})] \\ Q_2 &= m_{xy} + m_{uv} \\ Q_3 &= m_{xv} - m_{yu} \end{aligned} \tag{3.94}$$

which are related to the four basic phase space rotators: symmetric and antisymmetric fractional FTs, gyrator, and signal rotator, respectively.

Parameters Q_0 and $Q^2 = Q_1^2 + Q_2^2 + Q_3^2$ are invariant under the phase-space rotations. The components Q_j ($j = 1, 2, 3$), which can be organized as a vector $\mathbf{Q} = (Q_1, Q_2, Q_3)$, define the degree of vorticity of the beam,[46] which can be demonstrated on the example of the orthosymplectic modes.

For the HG modes only the diagonal second-order moments [see Eq. (1.25)] differ from zero: $m_{xx} = m_{uu} = m + \frac{1}{2}$ and $m_{yy} = m_{vv} = n + \frac{1}{2}$, and therefore

$$\begin{aligned} Q_0 &= m + n + 1 \\ \mathbf{Q} &= (m - n, 0, 0) \end{aligned} \tag{3.95}$$

Using the transformation relations for Q_j under the phase-space rotators[73] (see also Sec. 1.7.2), we find that the orthosymplectic mode $\mathcal{L}_{m,n}^{(\theta,\psi)}(\mathbf{r})$ presented on the (m, n)-orbital Poincaré sphere, defined in Sec. 3.5.2, is characterized by the parameters

$$\begin{aligned} Q_0 &= m + n + 1 \\ \mathbf{Q} &= Q(\sin\theta\cos\psi, \sin\theta\sin\psi, \cos\theta) \end{aligned} \tag{3.96}$$

where $Q = m - n$. Thus for the LG mode $\mathcal{L}_{m,n}^{\pm}(\mathbf{r})$ we obtain $Q_1 = Q_2 = 0$, and $Q_3 = \pm(m - n)$; meanwhile the HG modes rotated counterclockwise at $\pm\pi/4$ are characterized by $Q_1 = Q_3 = 0$ and $Q_2 = \pm(m - n)$. It has been mentioned that Q_3 corresponds to the z component of the OAM of the beam propagating in the z direction. A beam with nonzero integer Q_3 is referred as a *vortex beam*. Among the orthosymplectic modes $\mathcal{H}_{m,n}^{\mathrm{U}}(\mathbf{r})$ ($m \neq n$), only the LG modes are usually mentioned as vortex beams. Nevertheless others with $Q^2 = (m - n)^2 \neq 0$ can be considered as potential vortices, since they are converted to the LG modes by phase-space rotations. Note that for the modes with symmetric indices $m = n$, we obtain $Q_1 = Q_2 = Q_3 = 0$. This is the case of the fundamental Gaussian mode, for which Q_0 takes a minimal value $Q_0 = 1$.

The orbital Poincaré sphere introduced for the presentation of the orthosymplectic modes can also be used for the characterization of other two-dimensional signals,[74] which may be coherent or partially coherent. It has been shown in Ref. 20 that there exists such an optical first-order optical system associated with ray transformation matrix $\mathbf{T}_c$ that brings the moment matrix $\mathbf{M}$ to the diagonal form $\mathbf{M}_c = \mathbf{T}_c\mathbf{M}\mathbf{T}_c^t$, called *canonical* f_c, where $m_{xx} = m_{uu}$, $m_{yy} = m_{vv}$ (see details in Sec. 1.7.1). The signal f_c has the parameters

$$\begin{aligned} Q_0 &= m_{xx} + m_{yy} \\ \mathbf{Q} &= (m_{xx} - m_{yy}, 0, 0) \end{aligned} \tag{3.97}$$

and as well as the HG mode can be associated with the point $(\pi/2, 0)$ on the orbital Poincaré sphere. By performing the appropriate phase-space rotations, the entire sphere can be populated. The signal $f_{\theta,\psi}$ at point (θ, ψ) is characterized by the vector $\mathbf{Q} = (m_{xx} - m_{yy})(\sin\theta\cos\psi, \sin\theta\sin\psi, \cos\theta)$, which defines the signal symmetry. Thus signal with $\mathbf{Q} = (Q_1, 0, 0)$ is better described in the Cartesian coordinates; meanwhile for $\mathbf{Q} = (0, 0, Q_3)$ polar coordinates are the best choice for signal analysis. Moreover, in the last case the signal $f_{0,\psi}$ possesses a longitudinal component of the OAM $(m_{xx} - m_{yy})$.

Let us consider a signal $f(x, y)$ presented as a superposition of the HG modes

$$f(x, y) = \sum_{m,n} a_{mn}\mathcal{H}_{mn}(x, y) \qquad \sum_{m,n} |a_{mn}|^2 = 1 \tag{3.98}$$

Based on the expressions for the signal second-order moments,[75] we derive its Q components.

$$\begin{aligned}
Q_0 &= \sum_{m,n} |a_{mn}|^2 (m + n + 1) = 1 + \sum_{m,n} |a_{mn}|^2 (m + n) \\
Q_1 &= \sum_{m,n} |a_{mn}|^2 (m - n) \\
Q_2 &= 2\,\mathrm{Re}\left[\sum_{m,n} a_{m,n+1} a^*_{m+1,n} \sqrt{(m+1)(n+1)}\right] \\
Q_3 &= 2\,\mathrm{Im}\left[\sum_{m,n} a_{m,n+1} a^*_{m+1,n} \sqrt{(m+1)(n+1)}\right]
\end{aligned} \tag{3.99}$$

We observe that $Q_2 = Q_3 = 0$ if in the signal decomposition Eq. (3.98) there are no subsequent terms $a_{m,n+1}$ and $a_{m+1,n}$. Moreover $Q_3 = 0$ if the signal is real.

The stable beams that do not change their form under the propagation in free space contain in the decomposition Eq. (3.98) only the HG modes with the same index sum $m + n$. Then since the signal decomposition is normalized, they have integer parameter $Q_0 = m + n + 1$, which is related to the Gouy phase of the beam.

If a signal corresponds to a circular harmonic $c_l(r, \varphi)$ [see Eq. (3.90)], it has the same parameter $\mathbf{Q} = (0, 0, l)$ as the LG mode $\mathcal{L}^l_p$.

Representing signals on the orbital Poincaré spheres, we simplify the analysis and processing as well as the comparison with other signals. Since the search of the CT $\mathbf{T}_c$ is related to the diagonalization of the moment matrix, the signal presentation on the Poincaré sphere is valid for coherent as well as partially coherent beams.

3.7.7 Gouy Phase Accumulation

It was found by Gouy[76] more than 100 years ago that the Gaussian beam accumulates the additional constant phase during its free-space propagation. Now it is known that other transversal modes with a Gaussian envelope, undergoing a cycle of transformations while propagating through a paraxial optical system, accumulate Gouy phase,[77,78] usually divided into two parts: a dynamic part and a geometric part.[44,45,79,80] The identification of the Gouy phase is important in resonator theory,[81] in optical trapping,[82] and in possible applications of its geometric part for quantum computation.[83,84] A simple method for the determination of the Gouy phase—and in particular its dynamic and geometric parts—accumulated by an appropriate Gaussian-type mode during its propagation through a first-order optical system has been proposed in Ref. 27. It is based on the analysis of the eigenvalues and eigenvectors of the ray transformation matrix associated with the first-order optical system as it is briefly summarized below.

Strictly speaking, a beam of light $\Psi(\mathbf{r})$ propagating through an optical system, described by operator R, accumulates a phase shift only if it is an eigenfunction of R with eigenvalue in the form of complex exponent: $R[\Psi(\mathbf{r}_i)](\mathbf{r}_o) = \exp(i\psi)\,\Psi(\mathbf{r}_o)$. We recall (see Sec. 3.5.1), that an orthosymplectic mode $\mathcal{H}^{\mathrm{U}}_{m,n}(\mathbf{r}_i)$ is an eigenfunction for the phase-space rotator associated with unitary matrix $\mathbf{U}_s = \mathbf{U}\mathbf{U}_f(\gamma_x, \gamma_y)\mathbf{U}^{-1}$ with eigenvalue $\exp[-i(m+\frac{1}{2})\gamma_x - i(n+\frac{1}{2})\gamma_y]$, which corresponds to the Gouy phase. The decomposition Eq. (3.29) is crucial for the identification of the dynamic and geometric parts of the Gouy phase. Indeed, during the propagation through the symmetric fractional FT system, the mode $\mathcal{H}^{\mathrm{U}}_{m,n}(\mathbf{r}_i)$ acquires the dynamic phase $\psi_d = -(m+n+1)\varphi$, defined by the sum of mode indices; meanwhile in the case of a system similar to the antisymmetric fractional FT, the accumulated phase, known as the geometric one, is proportional to the index difference $\psi_g = -(m-n)\gamma$. Note that the dynamic and geometric phases are also defined by the second-order moments of $\mathcal{H}^{\mathrm{U}}_{m,n}(\mathbf{r}_i)$ through parameters $Q_0 = -\psi_d$ and $Q = -\psi_g$. This emphasizes that the geometric phase accumulation is related to the orbital angular momentum operators defined in phase space.

If the eigenvalues of the ray transformation matrix are not unimodular, we can speak about phase accumulation only in a wide sense, where scaling and quadratic-phase modulation of the field amplitude at the output system plane are present. In this case we permit the beam to be an eigenfunction of the transformation described by the orthosymplectic matrix in the Iwasawa decomposition Eq. (3.10).

We conclude that the Gouy phase accumulation of Gaussian-type beams is associated with rotations in phase space. Dynamic phase is acquired in symmetric, rotationally invariant systems, whose $\mathbf{T}_O$ in the decomposition Eq. (3.10) corresponds to the symmetric fractional

FT. The example of this system is a free-space propagation. Geometric phase accumulation requires a system with astigmatic elements, and $\mathbf{T}_O$ is similar to the antisymmetric fractional FT, as well as mode asymmetry $m \neq n$.

3.8 Conclusion

In this chapter we have considered the phase-space rotators—the transformations that produce the rotation of the Wigner distribution in phase space. Several approaches to the description of phase-space rotators by integral transforms, Hermitian operators, and ray transformation matrices have been discussed. There are four basic phase-space rotators for two-dimensional signals: symmetric and antisymmetric fractional FTs, gyrator, and signal rotator. The others can be obtained as their cascades. The fractional FT certainly plays a main role in the phase-space rotator description since it is associated with the diagonal unitary matrix, and any unitary matrix describing a phase-space rotator is similar to it.

We have seen that the eigenfunctions for the phase-space rotators are Gaussian functions modulated by the orthogonal polynomials, with Hermite-Gaussian and Laguerre-Gaussian modes among them. These modes are widely used for the description of optical beams. During the beam propagation through the optical system related to certain phase-space rotators for which it is an eigenfunction, the Gouy phase is acquired, because the eigenvalue is unimodular. We have stressed that the phase-space rotators similar to the symmetric (antisymmetric) fractional FTs are responsible for the accumulation of the dynamic (geometric) phase, correspondingly.

The application of the phase-space rotators to the different signal and image processing tasks such as filtering, pattern recognition, chirp detection, and signal encryption has been discussed. We also note that phase-space rotators play an important role in signal characterization, orbital angular momentum manipulation, and beam conversion. Thus the fractional FT is crucial for the phase-space tomography reconstruction of the Wigner distribution and therefore the complex field amplitude or mutual intensity of the coherent or partially coherent beam correspondingly.

The flexible optical setups for the experimental realization of basic phase-space rotators have been considered. We mention that the systems constructed from the generalized lenses located at the fixed position permit one to easily change the transformation parameters, which is required in various applications of phase-space rotators.

In this chapter we have considered the application of phase-space rotators to classical optical, beams, but they are also widely used in quantum physics and signal processing.

Acknowledgments

I am pleased to express my gratitude to M. J. Bastiaans, J. A. Rodrigo, and M. L. Calvo for fruitful collaboration in many of the discussed topics and to E. G. Abramochkin for careful reading of the manuscript and valuable comments. I also thank J. A. Rodrigo and A. Camara Iglesias for the help in figure preparation.

References

1. Jr. S. A. Collins, "Lens-system diffraction integral written in terms of matrix optics," *J. Opt. Soc. Am.* **60**: 1168–1177 (1970).
2. M. Moshinsky and C. Quesne, "Linear canonical transformations and their unitary representations," *J. Math. Phys.* **12**: 1772–1780 (1971).
3. K. B. Wolf, "*Integral Transforms in Sience and Engineering*," Plenum Publishing Corp., New York, 1979.
4. H. M. Ozaktas, Z. Zalevsky, and M. A. Kutay, *The Fractional Fourier Transform with Applications in Optics and Signal Processing*, Wiley, New York, 2001.
5. A. Torre, *Linear Ray and Wave Optics in Phase Space*, Elsevier, Amsterdam, 2005.
6. T. Alieva, M. J. Bastiaans, and M. L. Calvo, "Fractional transforms in optical information processing," *EURASIP J. Appl. Signal Process.* **2005**: 1498–1519 (2005).
7. T. Alieva and M. J. Bastiaans, "Properties of the linear canonical integral transformation," *J. Opt. Soc. Am. A* **24**: 3658–3665 (2007).
8. T. Alieva, "First-order optical systems for information processing," in A. Friberg and R. Dandliker (eds.), *Advances in Information Optics and Photonics*, SPIE, Bellingham, Wash., 2008, pp. 1–26.
9. A. Van der Lugt (ed.), *Optical Signal Processing*, Wiley, New York, 1992.
10. J. W. Goodman, *Introduction to Fourier Optics*, McGraw-Hill, New York, 1996.
11. V. Namias, "The fractional order Fourier transform and its applications to quantum mechanics," *J. Inst. Math. and Appl.* **25**: 241–265 (1980).
12. L. B. Almeida, "The fractional Fourier transform and time-frequency representations," *IEEE Trans. Sign. Proc.* **42**: 3084–3091 (1994).
13. D. Mendlovic and H. M. Ozaktas, "Fractional Fourier transforms and their optical implementation: I," *J. Opt. Soc. Am. A* **10**: 1875–1881 (1993).
14. H. M. Ozaktas and D. Mendlovic, "Fractional Fourier transforms and their optical implementation: II," *J. Opt. Soc. Am. A* **10**: 2522–2531 (1993).
15. J. García, D. Mendlovic, Z. Zalevsky, and A. W. Lohmann, "Space-variant simultaneous detection of several objects by the use of multiple anamorphic fractional-Fourier transform filters," *Appl. Opt.* **35**: 3945–3952 (1996).
16. G. Unnikrishnan, J. Joseph, and K. Singh, "Optical encryption by double-random phase encoding in the fractional Fourier domain," *Opt. Lett.* **25**: 887–889 (2000).
17. N. K. Nishchal, G. Unnikrishnan, J. Joseph, and K. Singh, "Optical encryption using a localized fractional Fourier transform," *Opt. Engin.* **42**: 3566–3571 (2003).
18. B. Hennelly and J. T. Sheridan, "Fractional Fourier transform-based image encryption: Phase retrieval algorithm," *Opt. Comm.* **226**: 61–80 (2003).
19. S. Q. Zhang and M. A. Karim, "Fractional correlation filter for fuzzy associative memories," *Opt. Eng.* **41**: 126–129 (2002).
20. K. Sundar, N. Mukunda, and R. Simon, "Coherent mode decomposition of general anisotropic Gaussian Schell-model beams," *J. Opt. Soc. Am. A* **12**: 560-569 (1995).
21. R. Simon and K. B. Wolf, "Structure of the set of paraxial optical systems," *J. Opt. Soc. Am. A* **17**: 342–355 (2000).

22. K. B. Wolf, *Geometric Optics in Phase Space*, Springer, New York, 2004.
23. T. Alieva and M. J. Bastiaans, "Alternative representation of the linear canonical integral transform," *Opt. Lett.* **30**: 3302–3304 (2005).
24. R. Simon, K. Sundar, and N. Mukunda, "Twisted Gaussian Schell-model beams. I. Symmetry structure and normal-mode spectrum," *J. Opt. Soc. Am. A* **10**: 2008–2016 (1993).
25. R. Simon and G. S. Agarwal, "Wigner representation of Laguerre-Gaussian beams," *Opt. Lett.* **25**: 1313–1315 (2000).
26. R. Simon and N. Mukunda, "Optical phase space, Wigner representation, and invariant quality parameters," *J. Opt. Soc. Am. A* **17**: 2440–2463 (2000).
27. T. Alieva and M. J. Bastiaans, "Dynamic and geometric phase accumulation by Gaussian-type modes in first-order optical systems," *Opt. Lett.* **33**: 1659–1661 (2008).
28. A. Lohmann, "Image rotation, Wigner rotation, and the fractional order Fourier transform," *J. Opt. Soc. Am. A* **10**: 2181–2186 (1993).
29. T. Alieva, V. Lopez, F. Agullo-Lopez, and L. B. Almeida, "The fractional Fourier transform in optical propagation problems," *J. Mod. Opt.* **41**: 1037–1044 (1994).
30. J. A. Rodrigo, T. Alieva, and M. L. Calvo, "Gyrator transform: Properties and applications," *Opt. Express* **15**: 2190–2203 (2007).
31. J. A. Rodrigo, T. Alieva, and M. L. Calvo, "Experimental implementation of the gyrator transform," *J. Opt. Soc. Am. A* **24**: 3135–3139 (2007).
32. J. A. Rodrigo, T. Alieva, and M. L. Calvo, "Applications of gyrator transform for image processing," *Opt. Comm.* **278**: 279–284 (2007).
33. J. A. Rodrigo, "First-Order Optical Systems in Information Processing and Optronic Devices," Ph.D. thesis, Universidad Complutense de Madrid, Spain, 2008.
34. H.-Y. Fan and H.-L. Lu, "Eigenmodes of fractional Hankel transform derived by the entangled-state method," *Opt. Lett.* **28**: 680–683 (2003).
35. E. G. Abramochkin and V. G. Volostnikov, "Generalized Gaussian beams," *J. Opt. A.: Pure Appl. Opt.* **6**: S157–S161 (2004).
36. M. J. Bastiaans and T. Alieva, "First-order optical systems with unimodular eigenvalues," *J. Opt. Soc. Am. A* **23**: 1875–1883 (2006).
37. M. J. Bastiaans and T. Alieva, "Classification of lossless first-order optical systems and the linear canonical transformation," *J. Opt. Soc. Am. A* **24**: 1053–1062 (2007).
38. T. Alieva, "Fractional Fourier transform as a tool for investigation of fractal objects," *J. Opt. Soc. Am. A* **13**: 1189–1192 (1996).
39. T. Alieva and M. J. Bastiaans, "Mode mapping in paraxial lossless optics," *Opt. Lett.* **30**: 1461–1463 (2005).
40. A. Wünsche, "General Hermite and Laguerre two-dimensional polynomials," *J. Phys. A: Math. Gen.* **33**: 1603–1629 (2000).
41. T. Alieva M. J. Bastiaans, "Orthonormal mode sets for the two-dimensional fractional Fourier transformation," *Opt. Lett.* **32**: 1226–1228 (2007).
42. M. J. Bastiaans and T. Alieva, "Propagation law for the generating function of Hermite-Gaussian-type modes in first-order optical systems," *Opt. Express* **13**: 1107–1112 (2005).
43. M. J. Padgett and J. Courtial, "Poincaré-sphere equivalent for light beams containing orbital angular momentum," *Opt. Lett.* **24**: 430–432 (1999).
44. G. F. Calvo, "Wigner representation and geometric transformations of optical orbital angular momentum spatial modes," *Opt. Lett.* **30**: 1207–1209 (2005).
45. G. S. Agarwal, "SU(2) structure of the Poincaré sphere for light beams with orbital angular momentum," *Opt. Lett.* **16**: 2914–2916 (1999).
46. D. Mendlovic, Y. Bitran, R. G. Dorsh, C. Ferreira, J. García, and H. M. Ozaktaz, "Anamorphic fractional Fourier transform: Optical implementation and applications," *Appl. Opt.* **34**: 7451–7456 (1995).
47. J. García, R. G. Dorsch, A. W. Lohmann, C. Ferreira, and Z. Zalevsky, "Flexible optical implementation of fractional Fourier transform processors, applications to correlation and filtering," *Opt. Comm.* **133**: 393–400 (1997).

48. M. F. Erden, H. M. Ozaktaz, A. Sahin, and D. Mendlovic, "Design of dynamically adjustible anamorphic fractional Fourier transformer," *Opt. Comm.* **136**: 52–60 (1997).
49. A. Sahin, H. M. Ozaktaz, and D. Mendlovic, "Optical implementations of two-dimensional fracional Fourier transforms and linear canonical transforms with arbitrary parameters," *Appl. Opt.* **37**: 2130–2141 (1998).
50. I. Moreno, J. A. Davis, and K. Crabtree, "Fractional Fourier transform optical system with programmable diffractive lenses," *Appl. Opt.* **42**: 6544–6548 (2003).
51. A. A. Malutin, "Tunable Fourier transformer of fractional order," *Quant. Elect.* **36**: 79–83 (2006).
52. B. Macukow and H. H. Arsenault, "Matrix decompositions for nonsymmetrical optical systems," *J. Opt. Soc. Am.* **73**: 1360–1366 (1983).
53. G. Nemes and A. E. Siegman, "Measurements of all ten second-order moments of an astigmatic beam by the use of rotating simple astigmatic (anamorphic) optics," *J. Opt. Soc. Am. A* **11**: 2257–2264 (1994).
54. J. A. Rodrigo, T. Alieva, and M. L. Calvo, "Optical system design for orthosymplectic transformations in phase space," *J. Opt. Soc. Am. A* **23**: 2494–2500 (2006).
55. H. Braunecker, O. Bryngdahl, and B. Schnell, "Optical system for image rotation and magnification," *J. Opt. Soc. Am.* **70**: 137–141 (1980).
56. O. Akay and G. F. Boudreaux-Bartels, "Fractional convolution and correlation via operator methods and an application to detection of linear FM signals," *IEEE Trans. Signal Process*. **49**: 979–993 (2001).
57. T. Alieva and M. L. Calvo, "Importance of the phase and amplitude in the fractional Fourier domain," *J. Opt. Soc. Am. A* **20**: 533–541 (2003).
58. J. C. Wood and D. T. Berry, "Tomographic time-frecuency analysis and its application toward time-varying filtering and adaptive kernel design for multicomponent linear-FM signals," *IEEE Trans. Signal Process*. **42**: 2094-2104 (1994).
59. Y. N. Hsu and H. H. Arsenault, "Optical pattern recognition using circular harmonic expansion," *Appl. Opt*. **21**: 4016–4019 (1982)
60. N. Towghi, B. Javidi, and Z. Luo, "Fully phase encrypted image processor," *J. Opt. Soc. Am. A* **16**: 1915–1927 (1999).
61. J. Courtial and M. Padgett, "Performance of a cylindrical lens mode converter for producing Laguerre-Gaussian laser modes," *Opt. Comm.* **159**: 13–18 (1999).
62. M. W. Beijersbergen, L. Allen, H. E. L. O. van der Veen, and J. P. Woerdman, "Astigmatic laser mode converters and transfer of orbital angular momentum," *Opt. Comm.* **96**: 123–132 (1993).
63. D. G. Grier, "A revolution in optical manipulation," *Nature* **424**: 810-816 (2003).
64. K. A. Nugent, D. Paganin, and T. E. Gureyev, "A phase odyssey," *Phys. Today* **8**: 27–32 (2001).
65. K. A. Nugent, "Wave field determination using 3-dimensional intensity information," *Phys. Rev. Lett.* **68**: 2261–2264 (1992).
66. M. G. Raymer, M. Beck, and D. F. McAlister, "Complex wave field reconstruction using phase space tomography," *Phys. Rev. Lett.* **72**: 1137–1140 (1994).
67. A. Cámara-Iglesias and T. Alieva, "Phase-space tomography for separable optical fields," in ICO-21 Congress Proceeding 2008, (2008), p.102.
68. R. Zambrini and S. M. Barnett, "Quasi-intristic angular momentum and the measurements of its specrum," *Phys. Rev. Lett.* **96**: 113901–4 (2006).
69. G. F. Calvo, A. Picon, and R. Zambrini, "Measuring the complete transverse spatial mode spectrum of a wave field," *Phys. Rev. Lett.* **100**: 173902–4 (2008).
70. M. J. Bastiaans, "Second-order moments of the Wigner distribution function in first-order optical systems," *Optik* **88**: 163–168 (1991).
71. J. Serna, R. Martínez-Herrero, and P. M. Mejías, "Parametric characterization of general partially coherent beams propagating through ABCD optical systems," *J. Opt. Soc. Am. A* **8**: 1094–1098 (1991).
72. M. J. Bastiaans and T. Alieva, "Wigner distribution moments in fractional Fourier transform systems," *J. Opt. Soc. Am. A* **19**: 1763–1773 (2002).

73. T. Alieva and M. J. Bastiaans, "Invariants of second-order moments of optical beams under phase-space rotations," in *ICO-21 Congress Proceeding 2008*, (2008), p. 103.
74. T. Alieva and M. J. Bastiaans, "Two-dimensional signal representation on the angular Poincaré sphere," in *Proc. Topical Meeting on Optoinformatics 2008*, St. Petersburg, Russia.
75. A. Ya. Bekshaev, "Intensity moments of a laser beam formed by superposition of Hermite-Gaussian modes," in *Fotoelektronika* **8**: 22–25 (1999), Odessa University.
76. L. G. Gouy, "Sur une propriété nouvelle des ondes lumineuses," *Compt. Rendue Acad. Sci.* (Paris) **110**: 1251–1253 (1890).
77. M. F. Erden and H. M. Ozaktas, "Accumulated Gouy phase shift in Gaussian beam propagation through first-order optical systems," *J. Opt. Soc. Am. A* **14**: 2190–2194 (1997).
78. R. Borghi, M. Santarsiero and R. Simon, "Shape invariance and a universal form for the Gouy phase," *J. Opt. Soc. Am. A* **21**: 572–579 (2004).
79. S. J. van Enk, "Geometric phase, transformations of Gaussian light beams and angular momentum transfer," *Opt. Comm.* **102**: 59–64 (1993).
80. E. J. Galvez, P. R. Crawford, H. I. Sztul, M. J. Pysher, P. J. Haglin, and R. E. Williams,"Geometric phase associated with mode transformations of optical beams bearing orbital angular momentum," *Phys. Rev. Lett.* **90**: 203901 (2003).
81. A. E. Siegman, *Lasers*, University Science Books, Sausalito, Calif., 1986.
82. F. Gittes and C. F. Schmidt, "Interference model for back-focal-plane displacement detection in optical tweezers," *Opt. Lett.* **23**: 7–9 (1998).
83. A. Vaziri, G. Weihs, and A. Zeilinger, "Experimental two-photon, three-dimensional entanglement for quantum communication," *Phys. Rev. Lett.* **89**: 240401 (2002).
84. A. Ekert, M. Ericsson, P. Hayden, H. Inamori, J. A. Jones, D. K. L. Oi, and V. Vedral, "Geometric quantum computation," *J. Mod. Opt.* **47**: 2501–2513 (2000).

This page intentionally left blank

CHAPTER 4

The Radon-Wigner Transform in Analysis, Design, and Processing of Optical Signals

Walter D. Furlan and Genaro Saavedra

Universitat de València, Optics Department
Burjassot, Spain

4.1 Introduction

One of the main features of phase space is that its conjugate coordinates are noncommutative and cannot be simultaneously specified with absolute accuracy. As a consequence, there is no phase-space joint distribution that can be formally interpreted as a joint probability density. Indeed, most of the classic phase-space distributions, such as the Wigner distribution function (WDF), the ambiguity function (AF), or the complex spectrogram, have difficult interpretation problems due to the complex, or negative, values they have in general. Besides, they may be nonzero even in regions of the phase space where either the signal or its Fourier transform vanishes. This is a critical issue, especially for the characterization of nonstationary or nonperiodic signals. As an alternative, the projections (*marginals*) of the phase-space distributions are strictly positive, and as we will see later, they give information about the signal on both phase-space variables. These projections can be formally associated with probability functions, avoiding all interpretation ambiguities associated with the original phase-space distributions. This is the case of the Radon-Wigner transform (RWT),

closely related to the projections of the WDF in phase space and also intimately connected with AF, as will be shown.

The general structure of this chapter is as follows. In Sec. 4.2, a general overview of mathematical properties of the RWT is given, and a summary of different optical setups for achieving it is presented. Next, the use of this representation in the analysis of optical signals and systems is developed in several aspects, namely, the computation of diffraction intensities, the optical display of Fresnel patterns, the amplitude and phase reconstruction of optical fields, and the calculation of merit function in imaging systems. Finally, in Sec. 4.4, a review of design techniques, based on the utilization of the RWT, for these imaging systems is presented, along with some techniques for optical signal processing.

4.2 Projections of the Wigner Distribution Function in Phase Space: The Radon-Wigner Transform (RWT)

The RWT was first introduced for the analysis and synthesis of frequency-modulated time signals, and it is a relatively new formalism in optics.[1,2] However, it has found several applications in this field during the last years. Many of them, such as the analysis of diffraction patterns, the computation of merit functions of optical systems, or the tomographic reconstruction of optical fields, are discussed in this chapter. We start by presenting the definition and some basic properties of the RWT. The optical implementation of the RWT which is the basis for many of the applications is discussed next.

Note, as a general remark, that for the sake of simplicity most of the formal definitions for the signals used hereafter are restricted to one-dimensional signals, that is, functions of a single variable $f(x)$. This is mainly justified by the specific use of these properties that we present in this chapter. The generalization to more than one variable is in most cases straightforward. We will refer to the dual variables x and ξ as *spatial* and *spatial-frequency* coordinates, since we will deal mainly with signals varying on space. Of course, if the signal is a function of time instead of space, the terms *time* and *frequency* should be applied.

4.2.1 Definition and Basic Properties

We start this section by recalling the definition of the WDF associated with a complex function $f(x)$, namely,

$$\mathcal{W}\{f(x), x, \xi\} = W_f(x, \xi) = \int_{-\infty}^{+\infty} f\left(x + \frac{x'}{2}\right) f^*\left(x - \frac{x'}{2}\right) \exp(-i2\pi\xi x')\, dx' \quad (4.1)$$

which also can be defined in terms of the Fourier transform (FT) of the original signal

$$\mathcal{F}\{f(x), \xi\} = F(\xi) = \int_{-\infty}^{+\infty} f(x) \exp(-i2\pi\xi x)\, dx \tag{4.2}$$

as

$$W_f(x, \xi) = \int_{-\infty}^{+\infty} F\left(\xi + \frac{\xi'}{2}\right) F^*\left(\xi - \frac{\xi'}{2}\right) \exp(i2\pi\xi' x)\, d\xi' \tag{4.3}$$

It is interesting to remember that any WDF can be inverted to recover, up to a phase constant, the original signal or, equivalently, its Fourier transform. The corresponding inversion formulas are[3]

$$f(x) = \frac{1}{f^*(x')} \int_{-\infty}^{+\infty} W_f\left(\frac{x + x'}{2}, \xi\right) \exp[i2\pi\xi(x - x')]\, d\xi \tag{4.4}$$

$$F(\xi) = \frac{1}{F^*(\xi')} \int_{-\infty}^{+\infty} W_f\left(x, \frac{\xi + \xi'}{2}\right) \exp[-i2\pi(\xi - \xi')x]\, dx \tag{4.5}$$

Note that these equations state the uniqueness of the relationship between the signal and the corresponding WDF (except for a phase constant). It is straightforward to deduce from these formulas that the integration of the WDF on the spatial or spatial-frequency coordinate leads to the modulus square of the signal or its Fourier transform, respectively, i.e.,

$$|f(x)|^2 = \int_{-\infty}^{+\infty} W_f(x, \xi)\, d\xi \tag{4.6}$$

$$|F(\xi)|^2 = \int_{-\infty}^{+\infty} W_f(x, \xi)\, dx \tag{4.7}$$

These integrals, or *marginals*, can be viewed as the projection of the function $W_f(x, \xi)$ in phase space along straight lines parallel to the ξ axis [in Eq. (4.6)] or to the x axis [in Eq. (4.7)]. These cases are particular ones of all possible projections along straight lines of a given function in phase space. In fact, for any function of (at least) two coordinates,

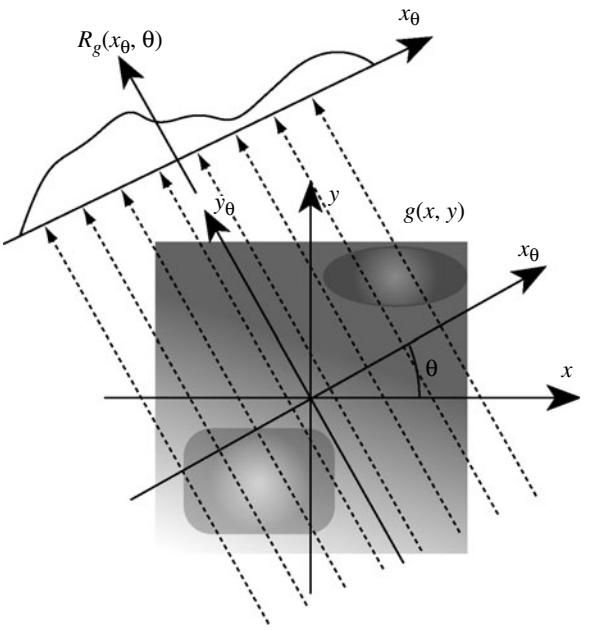


FIGURE 4.1 Projection scheme for the definition of the Radon transform.

say, $g(x, y)$, its Radon transform is defined as a generalized marginal

$$\mathcal{R}\{g(x, y), x_\theta, \theta\} = R_g(x_\theta, \theta) = \int_{-\infty}^{+\infty} g(x, y)\, dy_\theta \tag{4.8}$$

where, as presented in Fig. 4.1, x_θ and y_θ are the coordinates rotated by an angle θ. It is easy to see from this figure that

$$R_g(x_\theta, \theta + \pi) = R_g(-x_\theta, \theta) \tag{4.9}$$

Thus, the reduced domain $\theta \in (0, \pi)$ is used for $R_g(x_\theta, \theta)$. Note that the integration in the above definition is performed along straight lines characterized, for a given pair (x_θ, θ), by

$$\begin{aligned} y &= \frac{x_\theta}{\sin\theta} - \frac{x}{\tan\theta} && \text{for } \theta \neq 0, \frac{\pi}{2} \\ x &= x_\theta && \text{for } \theta = 0 \\ y &= x_\theta && \text{for } \theta = \frac{\pi}{2} \end{aligned} \tag{4.10}$$

and therefore Eq. (4.8) can be reformulated as

$$R_g(x_\theta, \theta) = \begin{cases} \displaystyle\int_{-\infty}^{+\infty} g\left(x, \frac{x_\theta}{\sin\theta} - \frac{x}{\tan\theta}\right) dx & \text{for } \theta \neq 0, \frac{\pi}{2} \\ \displaystyle\int_{-\infty}^{+\infty} g\,(x_\theta, y)\, dy & \text{for } \theta = 0 \\ \displaystyle\int_{-\infty}^{+\infty} g\,(x, x_\theta)\, dx & \text{for } \theta = \frac{\pi}{2} \end{cases} \tag{4.11}$$

Thus, when we consider as projected function $W_f\,(x, \xi)$, we can define the *generalized marginals* as the Radon transform of this WDF, namely,

$$\mathcal{R}\{W_f(x, \xi), x_\theta, \theta\} = R_{W_f}(x_\theta, \theta) = \int_{-\infty}^{+\infty} W_f\,(x, \xi)\, d\xi_\theta$$

$$= \int_{-\infty}^{+\infty} W_f\,(x_\theta \cos\theta - \xi \sin\theta, x_\theta \sin\theta + \xi \cos\theta)\, d\xi$$

$$= \begin{cases} \displaystyle\int_{-\infty}^{+\infty} W_f\left(x, \frac{x_\theta}{\sin\theta} - \frac{x}{\tan\theta}\right) dx & \text{for } \theta \neq 0, \frac{\pi}{2} \\ \displaystyle\int_{-\infty}^{+\infty} W_f\,(x_\theta, \xi)\, d\xi & \text{for } \theta = 0 \\ \displaystyle\int_{-\infty}^{+\infty} W_f\,(x, x_\theta)\, dx & \text{for } \theta = \frac{\pi}{2} \end{cases} \tag{4.12}$$

where, in the last expression, we have explicitly considered the equations for the integration lines in the projection. In terms of the original signal, this transform is called its Radon-Wigner transform. It is easy to show that

$$R_{W_f}(x_\theta, \theta) = RW_f(x_\theta, \theta) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f\left(x_\theta \cos\theta - \xi \sin\theta + \frac{x'}{2}\right)$$
$$\times f^*\left(x_\theta \cos\theta - \xi \sin\theta - \frac{x'}{2}\right)$$
$$\times \exp[-i2\pi(x_\theta \sin\theta + \xi \cos\theta)x']\, dx' d\xi \tag{4.13}$$

By performing a proper change in the integration variables, the following more compact expression can be obtained

$$RW_f(x_\theta, \theta) = \begin{cases} \left|\frac{1}{\sin\theta}\right| \left| \int_{-\infty}^{+\infty} f(x) \exp\left(i\pi \frac{x^2}{\tan\theta}\right) \exp\left(-i2\pi \frac{x_\theta x}{\sin\theta}\right) dx \right|^2 & \text{for } \theta \neq 0, \frac{\pi}{2} \\ |f(x_0 = x)|^2 & \text{for } \theta = 0 \\ \left|F\left(x_{\pi/2} = \xi\right)\right|^2 & \text{for } \theta = \frac{\pi}{2} \end{cases} \tag{4.14}$$

From this equation it is clear that

$$RW_f(x_\theta, \theta) \geq 0 \tag{4.15}$$

This is a very interesting property, since the WDF cannot be positive in whole phase space (except for the particular case of a Gaussian signal). Note also that from Eq. (4.14) a symmetry condition can be stated, namely,

$$RW_f(x_\theta, \pi - \theta) = RW_{f^*}(-x_\theta, \theta) \tag{4.16}$$

so that for real signals, that is, $f(x) = f^*(x) \forall x \in \mathbb{R}$, one finds

$$RW_f(x_\theta, \pi - \theta) = RW_f(-x_\theta, \theta) \tag{4.17}$$

and, therefore, for this kind of signal the reduced domain $\theta \in [0, \pi)$ in the Radon transform is clearly redundant. In this case, the range $\theta \in [0, \pi/2]$ contains in fact all the necessary values for a full definition of the RWT.

Equation (4.14) also allows one to link the RWT with another integral transform defined directly from the original signal, namely, the fractional Fourier transform (FrFT). This transformation, often considered a generalization of the classic Fourier transform, is given by

$$\mathcal{F}^p\{f(x), \alpha\} = F_p(\alpha) = \begin{cases} \frac{\exp\left[i\left(\theta + \pi\alpha^2/\tan\theta\right)\right]}{\sqrt{i\sin\theta}} \int_{-\infty}^{+\infty} f(x) \\ \quad \times \exp\left(i\pi \frac{x^2}{\tan\theta}\right) \exp\left(-i2\pi \frac{\alpha x}{\sin\theta}\right) dx & \text{for } \theta \neq 0, \frac{\pi}{2} \\ f(\alpha) & \text{for } \theta = 0 \\ F(\alpha) & \text{for } \theta = \frac{\pi}{2} \end{cases} \tag{4.18}$$

where $\theta = p\pi/2$. From this definition, it is easy to see that

$$RW_f(x_\theta, \theta) = |F_{2\theta/\pi}(x_\theta)|^2 \tag{4.19}$$

so that the RWT can be also interpreted as a two-dimensional representation of all the FrFTs of the original function.

Another interesting relationship can be established between the RWT and the AF associated with the input signal. For our input signal the AF is defined as

$$\begin{aligned}\mathcal{A}\{f(x), \xi', x'\} &= A_f(\xi', x') \\ &= \int_{-\infty}^{+\infty} f\left(x + \frac{x'}{2}\right) f^*\left(x - \frac{x'}{2}\right) \exp(-i2\pi\xi' x)\, dx\end{aligned} \tag{4.20}$$

which can be understood as the two-dimensional FT of the WDF, i.e.,

$$\begin{aligned}\mathcal{F}_{2D}\{W_f(x, \xi), \xi', x'\} &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} W_f(x, \xi) \exp\left[-i2\pi(\xi' x + x'\xi)\right] dx\, d\xi \\ &= A_f(\xi', -x')\end{aligned} \tag{4.21}$$

There is a well-known relationship between the two-dimensional FT of a function and the one-dimensional Fourier transformation of its projections. This link is established through the *central slice theorem*, which states that the values of the one-dimensional FT of a projection at an angle θ give a central profile—or *slice*—of the two-dimensional FT of the original signal at the same angle. If we apply this theorem to the WDF, it is straightforward to show that

$$\mathcal{F}\{RW_f(x_\theta, \theta), \xi_\theta\} = A_f(\xi_\theta \cos\theta, -\xi_\theta \sin\theta) \tag{4.22}$$

i.e., the one-dimensional FT of the RWT for a fixed projection angle θ provides a central profile of the AF $A_f(\xi', x')$ along a straight line forming an angle $-\theta$ with the ξ' axis. These relationships together with other links between representations in the phase space are summarized in Fig. 4.2.

To conclude this section, we consider the relationship between the RWT of an input one-dimensional signal $f(x)$ and the RWT of the same signal but after passing through a first-order optical system. In this case, the input signal undergoes a canonical transformation defined through four real parameters (a, b, c, d) or, equivalently, by a 2×2 real matrix

$$M = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \tag{4.23}$$

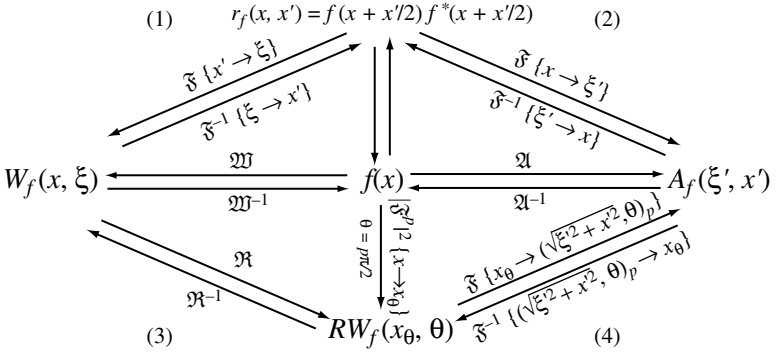


FIGURE 4.2 Relationship diagram between the original signal $f(x)$ and different phase-space representations. $\mathcal{F}$, $\mathcal{F}^p$, $\mathcal{W}$, $\mathcal{A}$, and $\mathcal{R}$ stand for FT, FrFT, WDF integral, AF transform, and Radon transformation, respectively, while -1 represents the corresponding inverse operator. (1) WDF and inverse transform; (2) AF and inverse transform; (3) projection (Radon) transformation and tomographic reconstruction operator; (4) expression of the *central slice theorem* applied to Radon transform and AF. $(\rho, \theta)_p$ represents polar coordinates in phase space.

in such a way that the transformed signal $g(x)$ is given by

$$g(x) = \begin{cases} \frac{1}{\sqrt{ib}} \exp\left(\frac{-i\pi d x'^2}{b}\right) \int_{-\infty}^{+\infty} f(x') \exp\left(\frac{-i\pi a x'^2}{b}\right) \exp\left(\frac{i2\pi}{b} x\, x'\right) dx' & b \neq 0 \\ \exp\left(\frac{-i\pi c x'^2}{a}\right) \frac{1}{\sqrt{a}} f\left(\frac{x}{a}\right) & b = 0 \end{cases} \tag{4.24}$$

which are the one-dimensional counterparts of Eqs. (3.4) and (3.7). We are restricting our attention to nonabsorbing systems corresponding to the condition $\det M = ad - bc = 1$.

The application of a canonical transformation on the signal produces a distortion on the corresponding WDF according to the general law

$$W_g(x, \xi) = W_f(ax + b\xi, cx + d\xi) = W_f(x', \xi') \tag{4.25}$$

where the mapped coordinates are given by

$$\begin{pmatrix} x' \\ \xi' \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ \xi \end{pmatrix} \tag{4.26}$$

By applying the definition in Eq. (4.12), it is straightforward to obtain

$$RW_g(x_\theta, \theta) = \begin{cases} \int_{-\infty}^{+\infty} W_g\left(x, \frac{x_\theta}{\sin\theta} - \frac{x}{\tan\theta}\right) dx & \text{for } \theta \neq 0, \frac{\pi}{2} \\ \int_{-\infty}^{+\infty} W_g(x_\theta, \xi)\, d\xi & \text{for } \theta = 0 \\ \int_{-\infty}^{+\infty} W_g(x, x_\theta)\, dx & \text{for } \theta = \frac{\pi}{2} \end{cases}$$

$$= \left\{ \begin{array}{ll} \int_{-\infty}^{+\infty} W_f\left(ax + b\left(\frac{x_\theta}{\sin\theta} - \frac{x}{\tan\theta}\right), cx \right. & \\ \quad \left. + d\left(\frac{x_\theta}{\sin\theta} - \frac{x}{\tan\theta}\right)\right) dx & \text{for } \theta \neq 0, \frac{\pi}{2} \\ \int_{-\infty}^{+\infty} W_f(ax_\theta + b\xi, cx_\theta + d\xi)\, d\xi & \text{for } \theta = 0 \\ \int_{-\infty}^{+\infty} W_f(ax + bx_\theta, cx + dx_\theta)\, dx & \text{for } \theta = \frac{\pi}{2} \end{array} \right\}$$

$$\propto RW_f(x_{\theta'}, \theta') \tag{4.27}$$

where the mapped coordinates for the original RWT are given by

$$\tan\theta' = -\frac{a\tan\theta - b}{c\tan\theta - d}, \quad x_{\theta'} = \frac{x_\theta}{a\sin\theta - b\cos\theta}\sin\theta' \tag{4.28}$$

Let us consider in the following examples a spatially coherent light distribution $f(x)$, with wavelength λ, that travels along a system that imposes a transformation in the input characterized by an *abcd* transform. Special attention is usually paid to the cases $\theta = 0, \pi/2$ since, according to Eqs. (4.6) and (4.7), the modulus squared of the *abcd* transform in Eq. (4.24) and its FT are then obtained, respectively.

1. *Coherent propagation through a (cylindrical) thin lens.* In this case the associated M matrix for the transformation of the light field is given by

$$M_L = \begin{pmatrix} 1 & 0 \\ \frac{1}{\lambda f} & 1 \end{pmatrix} \tag{4.29}$$

with f being the focal length of the lens. Thus, the RWT for the transformed amplitude light distribution is given in this case by

$$RW_g(x_\theta, \theta) \propto RW_f\left(x_{\theta'}, \theta'\right), \quad \tan\theta' = -\lambda f \frac{\tan\theta}{\tan\theta - \lambda f},$$
$$x_{\theta'} = \frac{x_\theta}{\sin\theta}\sin\theta' \tag{4.30}$$

A careful calculation for the case of $\theta = 0$ leads to

$$RW_g(x_0, 0) = |g(x_0)|^2 \propto RW_f(x_0, 0) \tag{4.31}$$

while for the value $\theta = \pi/2$ the following result is obtained

$$RW_g\left(x_{\pi/2}, \frac{\pi}{2}\right) \propto RW_f\left(x_{\pi/2} \sin\theta', \theta'\right), \qquad \tan\theta' = -\lambda f \tag{4.32}$$

Note that the effect of this propagation through a thin lens of focal length f is also physically equivalent to the illumination of the incident light distribution by a spherical wavefront whose focus is located at a distance $\eta = f$ from the input plane. Thus, the same results discussed here can be applied straightforwardly to that case.

2. *Free-space (Fresnel) propagation.* If we consider now the Fresnel approximation for the propagation of a transverse coherent light distribution $f(x)$ by a distance z, namely,

$$g(x) = \int_{-\infty}^{+\infty} f(x') \exp\left[\frac{i\pi}{\lambda z}(x' - x)^2\right] dx' \tag{4.33}$$

the transformation matrix M is given by

$$M_F = \begin{pmatrix} 1 & -\lambda z \\ 0 & 1 \end{pmatrix} \tag{4.34}$$

and, therefore, the transformed RWT can be calculated through the expression

$$RW_g(x_\theta, \theta) \propto RW_f(x_{\theta'}, \theta'), \qquad \tan\theta' = \tan\theta - \lambda z,$$
$$x_{\theta'} = \frac{x_\theta}{\sin\theta + \lambda z \cos\theta} \sin\theta' \tag{4.35}$$

For the projection with $\theta = 0$, one obtains

$$RW_g(x_0, 0) = |g(x_0)|^2 \propto RW_f(x_{\theta'}, \theta'), \qquad \tan\theta' = -\lambda z,$$
$$x_{\theta'} = \frac{x_0}{\lambda z} \sin\theta' \tag{4.36}$$

and for the orthogonal projection $\theta = \pi/2$ the following result is achieved

$$RW_g\left(x_{\pi/2}, \frac{\pi}{2}\right) \propto RW_f\left(x_{\pi/2}, \frac{\pi}{2}\right) \tag{4.37}$$

3. *Magnifier.* If a uniform scale factor m is applied to the input function, the associated M matrix is given by

$$M_m = \begin{pmatrix} \frac{1}{m} & 0 \\ 0 & m \end{pmatrix} \tag{4.38}$$

In this case, the RWT is transformed according to the law

$$RW_g(x_\theta, \theta) \propto RW_f(x_{\theta'}, \theta'), \qquad \tan\theta' = \frac{1}{m^2}\tan\theta,$$
$$x_{\theta'} = \frac{m x_\theta}{\sin\theta}\sin\theta' \tag{4.39}$$

The vertical and horizontal projections are given here simply by the following formulas.

$$RW_g(x_0, 0) = |g(x_0)|^2 \propto RW_f\left(\frac{x_0}{m}, 0\right)$$
$$RW_g\left(x_{\pi/2}, \frac{\pi}{2}\right) \propto RW_f\left(m x_{\pi/2}, \frac{\pi}{2}\right) \tag{4.40}$$

4.2.2 Optical Implementation of the RWT: The Radon-Wigner Display

Like any other phase-space function, the RWT also enables an optical implementation that is desirable for applications in the analysis and processing of optical signals. The correct field identification requires a large number of Wigner distribution projections, which raises the necessity to design flexible optical setups to obtain them. The relationship between the RWT and the FrFT, expressed mathematically by Eq. (4.19), suggests that the optical computation of the RWT is possible directly from the input function, omitting the passage through its WDF. In fact, the RWT for a given projection angle is simply the intensity registered at the output plane of a given FrFT transformer. For one-dimensional signals, the RWT for all possible projection angles simultaneously displays a continuous representation of the FrFT of a signal as a function of the fractional Fourier order p, and it is known as the *Radon-Wigner display* (RWD). This representation, proposed by Wood and Barry for its application to the detection and classification of linear FM components,[1] has found several applications in optics as we will see later in this chapter.

Different and simple optical setups have been suggested to implement the FrFT, and most have been the basis for designing other systems to obtain the RWD. The first one described in the literature, designed to obtain the RWD of one-dimensional signals, was proposed by Mendlovic et al.[4] It is based on Lohmann's bulk optics systems for obtaining the FrFT.[5] In this method, the one-dimensional input

function is converted to a two-dimensional object by the use of cylindrical lenses to allow the construction of a multichannel processor that optically implements the calculations of the RWD. The setup consists of three phase masks separated by fixed distances in free space. The masks consist of many strips, each one representing a different channel that performs an FrFT with a different order over the input signal. Each strip is a Fresnel zone plate with a different focal length that is selected for obtaining the different fractional order p. Thus, the main shortcoming of the RWD chart produced by this setup is that it has a limited number of projection angles (or fractional orders). Besides the very poor angular resolution, the experimental results obtained in the original paper are actually very far from the theoretical predictions.

A truly continuous display, i.e., a complete RWD setup, was proposed by Granieri et al.[6] This approach is based on the relationship between the FrFT and Fresnel diffraction,[7,8] which establishes that every Fresnel diffraction pattern of an input object is univocally related to a scaled version of a certain FrFT of the same input. Therefore, if the input function $f(x)$ is registered in a transparency with amplitude transmittance $t(x/s)$, with s being the construction scale parameter, then the FrFT of the input can be optically obtained by free-space propagation of a spherical wavefront impinging on it. Actually, the Fresnel diffraction field $U(x, R_p)$ obtained at distance R_p from the input, which is illuminated with a spherical wavefront of radius z and wavelength λ, is related to the FrFT of order p of the input function $\mathcal{F}^p\{t(x), \alpha\}$ as follows:[9]

$$U(x, R_p) = \exp\left\{\frac{i\pi x^2}{\lambda}\left[\frac{z(1-M_p)-R_p}{z\,R_p\,M_p^2}\right]\right\}\mathcal{F}^p\left\{t\left(\frac{x'}{M_p}\right), x\right\} \quad (4.41)$$

where M_p is the magnification of the optical FrFT. For each fractional order, the values of M_p and R_p are related to the system parameters s, λ, and z through

$$R_p = \frac{s^2\lambda^{-1}\tan(p\pi/2)}{1+s^2(z\lambda)^{-1}\tan(p\pi/2)} \quad (4.42)$$

$$M_p = \frac{1+\tan(p\pi/2)\tan(p\pi/4)}{1+s^2(z\lambda)^{-1}\tan(p\pi/2)} \quad (4.43)$$

These last equations allow us to recognize that by illumination of an input transparency with a spherical wavefront converging to an axial point S, all the FrFTs in the range [0, 1] can be obtained simultaneously, apart from a quadratic-phase factor and a scale factor. The FrFTs are axially distributed between the input transparency plane ($p = 0$) and the virtual source (S) plane ($p = 1$) in which the optical FT of the input is obtained. For one-dimensional input signals, instead of a spherical

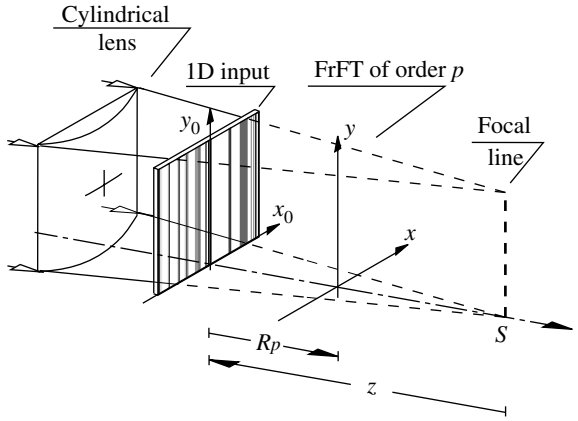


FIGURE 4.3 Implementation of the FrFT by free-space propagation.

wavefront, we can use a cylindrical one to illuminate the input (see Fig. 4.3).

Keeping in mind Eq. (4.19), we see the next step is to obtain the RWD from this setup. To do this, we have to find an optical element to form the image of the axially distributed FrFT channels, at the same output plane simultaneously. Therefore, the focal length of this lens should be different for each fractional order p. Since in this case the different axially located FrFTs present no variations along the vertical coordinate, we can select a different one-dimensional horizontal slice of each one and use it as a single and independent fractional-order channel (see Fig. 4.4).

The setup of Fig. 4.4 takes advantage of the one-dimensional nature of the input, and it behaves as a multichannel parallel FrFT transformer, provided that the focal length of the lens L varies with the y coordinate in the same way as it varies with p. In this way, the problem can be addressed as follows. For each value of p (vertical coordinate y) we want to image a different object plane at a distance a_p from the lens onto a fixed output plane located at a' from the lens. To obtain this result, it is straightforward to deduce from the Gaussian lens equation and from the distances in Fig. 4.4 that it is necessary to design a lens with a focal length that varies with p (vertical coordinate y) according to

$$f(p) = \frac{a' a_p}{a' + a_p} = \frac{a' l + (1 + l z^{-1}) a' s^2 \lambda^{-1} \tan(p\pi/2)}{a' - l - (a' + l + z) z^{-1} s^2 \lambda^{-1} \tan(p\pi/2)} \quad (4.44)$$

On the other hand, this focal length should provide the exact magnification at each output channel. The magnification given by the system

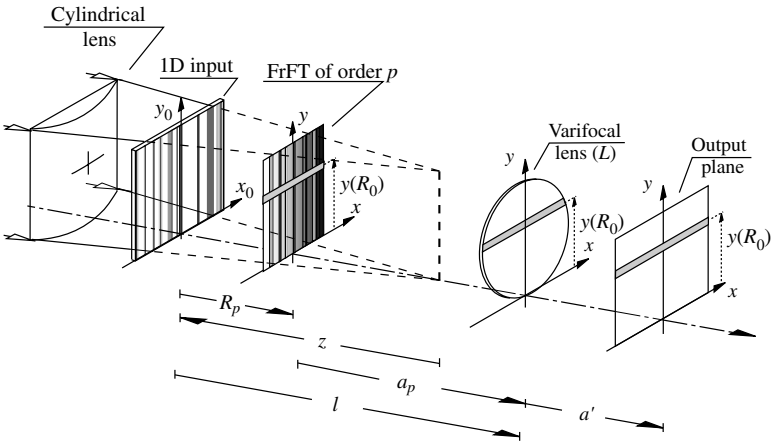


FIGURE 4.4 RWD setup (multichannel continuous FrFT transformer).

for each fractional order p is

$$M_L(p) = \frac{-a'}{a_p} = \frac{a'}{\frac{s^2\lambda^{-1}\tan(p\pi/2)}{1+s^2(z\lambda)^{-1}\tan(p\pi/2)} - l} \tag{4.45}$$

However, for the p-order slice of the RWT of the input function to be achieved, the lens L should counterbalance the magnification of the FRT located at R_p to restore its proper magnification at the output plane. Therefore, by using Eq. (4.43), the magnification provided by L should be

$$M_L(p) = \frac{-1}{M_p} = -\frac{1 + s^2(z\lambda)^{-1}\tan(p\pi/2)}{1 + \tan(p\pi/2)\quad\tan(p\pi/4)} \tag{4.46}$$

Comparing Eqs. (4.45) and (4.46), we note that the functional dependence of both equations on p is different, and, consequently, we are unable to obtain an exact solution for all fractional orders. However, an approximate solution can be obtained by choosing the parameters of the system, namely, s, z, l, λ, and a', in such a way that they minimize the difference between these functions in the interval $p \in [0, 1]$. One way to find the optimum values for these parameters is by a least-square method. This optimization[6] leads to the following constraint conditions.

$$a' = l\left(\frac{1}{2} + \frac{\pi}{4}\right), \qquad z = \frac{-ls^2}{\lambda l + s^2} \tag{4.47}$$

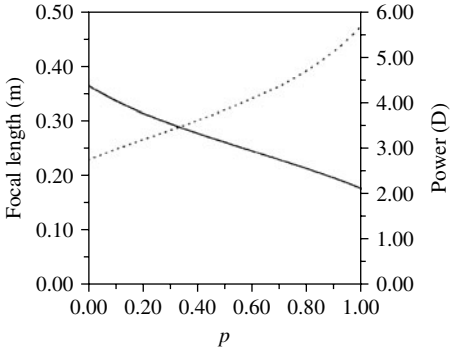


FIGURE 4.5 Focal length (solid curve) and optical power (dotted curve) of the designed varifocal lens L for the values $z = 426$ mm, $l = 646$ mm, and $a = 831$ mm.

The variation of the focal distance of the lens L with p according to Eq. (4.44) and its optical power, under the constraints given by Eqs. (4.47), are represented in Fig. 4.5 for the following values: $z = 426$ mm, $l = 646$ mm, and $a = 831$ mm.

For this particular combination of parameters, the optical power is nearly linear with p, except for values close to $p = 1$. This linearity is also accomplished by some designs of ophthalmic progressive addition lenses in which there is a continuous linear transition between two optical powers that correspond to the *near portion* and *distance portion*. In the experimental verification of the system, a progressive lens of +2.75 D spherical power and +3 D of addition was used in the setup of Fig. 4.4 with the above-mentioned values for the parameters z, l, and a. Figure 4.6 illustrates a comparison between the numerical

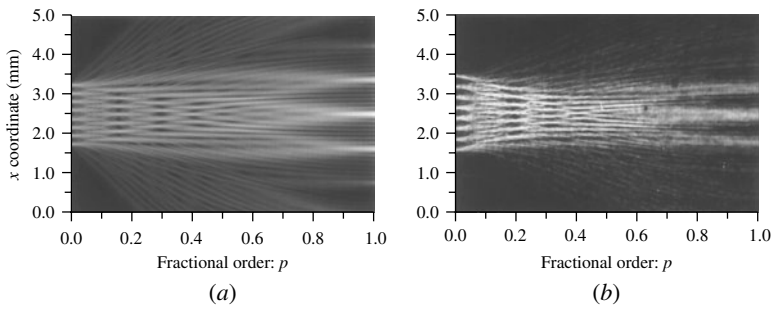


FIGURE 4.6 RWD of a Ronchi grating of 3 lines/mm: (*a*) exact numerical simulation; (*b*) experimental result.

simulations and the experimental results obtained using a Ronchi grating as input object.

Interestingly, in Fig. 4.6 the values of p that correspond to the self-images, both positive and negative, can be clearly identified. The optical setup designed for the experimental implementation of the RWD was successfully adapted to several applications, as we show later in this chapter.

In searching for an RWD with an exact scale factor for all the fractional orders, this approach also inspired another proposal[10] in which a bent structure for the detector was suggested. The result is an exact, but unfortunately impractical, setup to obtain the RWD. This drawback was partially overcome in other configurations derived by the same authors using the *abcd* matrix formalism. There, the free propagation distances are designed to be fixed or to vary linearly with the transverse coordinate,[11] so the input plane and/or the output plane should be tilted instead of bent, resulting in a more realistic configuration, provided that the tilt angles are measured very precisely.

4.3 Analysis of Optical Signals and Systems by Means of the RWT

4.3.1 Analysis of Diffraction Phenomena

4.3.1.1 Computation of Irradiance Distribution along Different Paths in Image Space

Determination of the irradiance at a given point in the image space of an imaging system is a classic problem in optics. The conventional techniques carry out a finite partition of the pupil of the system to sum all these contributions at the observation point.[12–16] This time-consuming procedure needs to be completely repeated for each observation point, or if the aberration state of the system changes. In this section we present a useful technique, based on the use of the RWT of a mapped version of the pupil of the system, for a much more efficient analysis of the irradiance in the image space of imaging systems. This technique has been successfully applied to the analysis of different optical systems with circular[17] as well as square,[18] elliptical,[19] triangular,[19] and even fractal pupils.[20] The method has also been applied to the study of multifaceted imaging devices.[21]

Let us consider a general imaging system, characterized by an exit pupil function with generalized amplitude transmittance $P(\mathbf{x})$. The distance from this pupil to the Gaussian imaging plane is denoted by f. Note that the function $P(\mathbf{x})$ includes any arbitrary amplitude variation $p(\mathbf{x})$ and any phase aberration that the imaging system may suffer from.

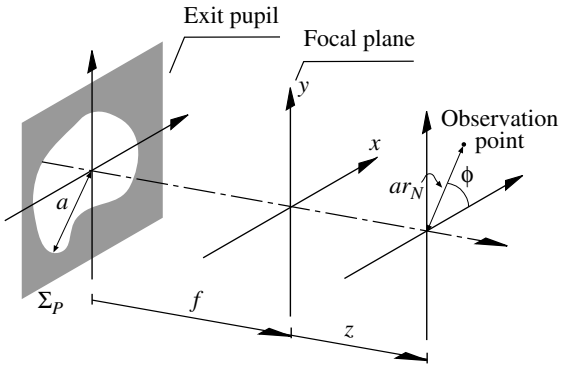


FIGURE 4.7 The imaging system under study.

We now describe the monochromatic scalar light field at any point of the image space of the system in the Fresnel approximation.[21] It is straightforward to show that, within this approach, the field irradiance is given by

$$I(\mathbf{x}, z) = \frac{1}{\lambda^2(f+z)^2} \times \left| \iint_{\Sigma_P} P(\mathbf{x}') \exp\left[\frac{-i\pi z|\mathbf{x}'|^2}{\lambda f(f+z)}\right] \exp\left[\frac{-i2\pi}{\lambda(f+z)}\mathbf{x}\cdot\mathbf{x}'\right] d^2\mathbf{x}' \right|^2 \tag{4.48}$$

where λ is the field wavelength, $\mathbf{x}'$ and z stand for the transverse and axial coordinates of the observation point, respectively, and Σ_P represents the pupil plane surface. The origin for the axial distances is fixed at the axial Gaussian point, as shown in Fig. 4.7.

It is convenient to express all transverse coordinates in normalized polar form, namely,

$$x = ar_N \cos\phi, \qquad y = ar_N \sin\phi \tag{4.49}$$

where x and y are Cartesian coordinates and a stands for the maximum radial extent of the pupil. By using these explicit coordinates in Eq. (4.48), we obtain

$$I(r_N, \phi, z) = \frac{1}{\lambda^2(f+z)^2} \left| \int_0^{2\pi}\int_0^1 \bar{p}(r'_N, \phi') \exp\left[\frac{i2\pi W(r'_N, \phi')}{\lambda}\right] \exp\left[\frac{i2\pi W_{20}(z)\, r'^2_N}{\lambda}\right] \times \exp\left[\frac{-i2\pi}{\lambda(f+z)} r'_N r_N \cos(\phi' - \phi)\right] r'_N\, dr'_N\, d\phi' \right|^2 \tag{4.50}$$

where the bar denotes the polar coordinate expression for the corresponding function and where we have split out the generalized pupil $P(\mathbf{x}')$ to explicitly show the dependence on the amplitude pupil variations $p(\mathbf{x}')$ and the aberration function $W(r'_N, \phi')$ of the system. The classic defocus coefficient has also been introduced in this equation, namely,

$$W_{20}(z) = -\frac{za^2}{2f(f+z)} \tag{4.51}$$

In many practical situations the most important contribution to the aberration function is the primary *spherical aberration* (SA), whose dependence on the pupil coordinates is given by

$$W_{40}(r'_N, \phi') = W_{40} {r'_N}^4 \tag{4.52}$$

where W_{40} is the SA coefficient design constant. In the following reasoning, we will consider this term explicitly, splitting the generalized pupil of the system as follows:

$$\bar{p}(r'_N, \phi') \exp\left[\frac{i2\pi W(r'_N, \phi')}{\lambda}\right] = Q(r'_N, \phi') \exp\left[\frac{i2\pi W_{40} {r'_N}^4}{\lambda}\right] \tag{4.53}$$

Thus $Q(r'_N, \phi')$ includes the amplitude variations on the pupil plane and the aberration effects except for SA. Note that if no aberrations different from SA are present in the system, $Q(r'_N, \phi')$ reduces simply to the pupil mask $\bar{p}(r'_N, \phi')$.

By substituting Eq. (4.53) into Eq. (4.50), we finally obtain

$$\begin{aligned}\bar{I}(r_N, \phi, z) &= \frac{1}{\lambda^2(f+z)^2} \left| \int_0^{2\pi} \int_0^1 Q(r'_N, \phi') \exp\left(\frac{i2\pi W_{40} {r'_N}^4}{\lambda}\right) \exp\left[\frac{i2\pi W_{20}(z) {r'_N}^2}{\lambda}\right] \right. \\ &\quad \left. \times \exp\left[\frac{-i2\pi}{\lambda(f+z)} r'_N r_N \cos(\phi' - \phi)\right] r'_N \, dr'_N \, d\phi' \right|^2 \end{aligned} \tag{4.54}$$

Let us now consider explicitly the angular integration in this equation, namely,

$$Q(r'_N, r_N, \phi, z) = \int_0^{2\pi} Q(r'_N, \phi') \exp\left[\frac{-i2\pi}{\lambda(f+z)} r'_N r_N \cos(\phi' - \phi)\right] d\phi' \tag{4.55}$$

Thus we arrive at a compact form for the irradiance at a point in the image space

$$\bar{I}(r_N, \phi, z) = \frac{1}{\lambda^2 (f+z)^2} \times \left| \int_0^1 Q(r'_N, r_N, \phi, z) \exp\left(\frac{i2\pi W_{40}\, r'^4_N}{\lambda} \right) \exp\left[\frac{i2\pi W_{20}(z)\, r'^2_N}{\lambda} \right] r'_N\, dr'_N \right|^2 \tag{4.56}$$

By using the mapping transformation

$$r'^2_N = s + \frac{1}{2}, \qquad Q(r'_N, r_N, \phi, z) = q(s, r_N, \phi, z) \tag{4.57}$$

we finally obtain

$$\bar{I}(r_N, \phi, z) = \frac{1}{\lambda^2 (f+z)^2} \times \left| \int_{-0.5}^{0.5} q(s, r_N, \phi, z) \exp\left(\frac{i2\pi W_{40}\, s^2}{\lambda} \right) \exp\left\{ \frac{i2\pi [W_{40} + W_{20}(z)]\, s}{\lambda} \right\} ds \right|^2 \tag{4.58}$$

Note that in this expression all the dependence on the observation coordinates is concentrated in the mapped pupil $q(s, r_N, \phi, z)$ and the defocus coefficient $W_{20}(z)$. If we expand the modulus square in this equation, we find

$$\bar{I}(r_N, \phi, z) = \frac{1}{\lambda^2 (f+z)^2} \times \int_{-0.5}^{0.5} \int_{-0.5}^{0.5} q(s, r_N, \phi, z) q^*(s', r_N, \phi, z) \exp\left[\frac{i2\pi W_{40}\, (s^2 - s'^2)}{\lambda} \right] \times \exp\left\{ \frac{i2\pi\, [W_{40} + W_{20}(z)\,]\, (s - s')}{\lambda} \right\} ds\, ds' \tag{4.59}$$

which by using the change of variables

$$t = \frac{s + s'}{2}, \qquad u = s - s' \tag{4.60}$$

can be rewritten as

$$\bar{I}(r_N, \phi, z) = \frac{1}{\lambda^2 (f+z)^2} \times \int_{-1}^{1} \int_{-0.5}^{0.5} q\left(t + \frac{u}{2}, r_N, \phi, z\right) q^*\left(t - \frac{u}{2}, r_N, \phi, z\right) \times \exp\left\{\frac{i2\pi}{\lambda}\left[W_{40} + W_{20}(z) + 2\xi W_{40}\right] u\right\} dt\, du \tag{4.61}$$

The above integration over the variable u can be clearly identified as the WDF of $q(s, r_N, \phi, z)$ with respect to the first variable, as stated in Eq. (4.1). Thus, it is straightforward to show that

$$\bar{I}(r_N, \theta, z) = \frac{1}{\lambda^2 (f+z)^2} \int_{-0.5}^{0.5} W_q\left(t, -2\frac{W_{40}}{\lambda} t - \frac{W_{40} + W_{20}(z)}{\lambda}\right) dt \tag{4.62}$$

This expression relates the irradiance at any observation point to the line integral of the function $W_q(x, \xi)$ along a straight line in phase space described by the equation

$$\xi = -2\frac{W_{40}}{\lambda} x - \frac{W_{40} + W_{20}(z)}{\lambda} \tag{4.63}$$

as depicted in Fig. 4.8. One can identify this integration as a projection of the WDF at an angle θ given by [see Eq. (4.10)]

$$\tan\theta = -\frac{\lambda}{2W_{40}} \tag{4.64}$$

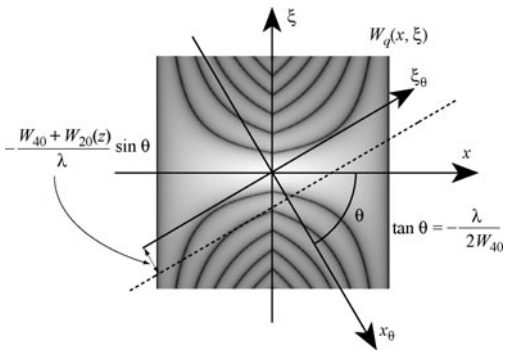


FIGURE 4.8 Integration line in phase space.

and at an oriented distance from the origin

$$x_\theta(z) = -\frac{W_{40} + W_{20}(z)}{\sqrt{4W_{40}^2 + \lambda^2}} \tag{4.65}$$

in such a way that it is possible to express

$$\bar{I}(r_N, \theta, z) = \frac{1}{\lambda^2 (f+z)^2} RW_q\,(x_\theta(z), \theta) \tag{4.66}$$

The main conclusion of all this is that it is possible to obtain the irradiance at any point in image space through the values of the RWT of a given function $q(s, r_N, \phi, z)$ related to the pupil of the system. Note, however, that this function depends in general on the particular coordinates r_N, ϕ, and z of the observation point. Thus, a different function $RW_q(x_\theta, \theta)$ has to be considered for different points in image space. This major drawback can be overcome for special sets of points or trajectories in image space that share the same associated mapped pupil $q(s, r_N, \phi, z)$.

To describe such trajectories in image space, let us express these lines in parametric form $r_N(z)$, $\phi(z)$. By substituting Jacobi's identity

$$\exp(i\gamma\cos\chi) = \sum_{n=-\infty}^{+\infty} i^n J_n(\gamma)\exp(-in\chi) \quad \gamma, \chi \in \mathbb{R} \tag{4.67}$$

where $J_n(x)$ represents the Bessel function of the first kind and order n, into Eq. (4.55), it is straightforward to obtain

$$Q(r'_N, r_N(z), \phi(z), z) = \sum_{n=-\infty}^{+\infty} i^n J_n\left(\frac{-2\pi}{\lambda(f+z)} r'_N r_N(z)\right) Q_n(r'_N) \times \exp[in\phi(z)] \tag{4.68}$$

where $Q_n(r'_N)$ stands for the n-order circular harmonic of $Q_n(r'_N, \phi')$, that is,

$$Q_n(r'_N) = \int_0^{2\pi} Q(r'_N, \phi')\exp\left(-in\phi'\right) d\phi' \tag{4.69}$$

Note that the dependence on the position parameter z in Eq. (4.68) is established exclusively in the argument of the Bessel functions—through $r_N(z)$—and the phase exponentials—through $\phi(z)$. Thus, the only way to strictly cancel this dependence is to consider the trajectories

$$r_N(z) = K(f+z), \qquad \phi(z) = \phi_o \tag{4.70}$$

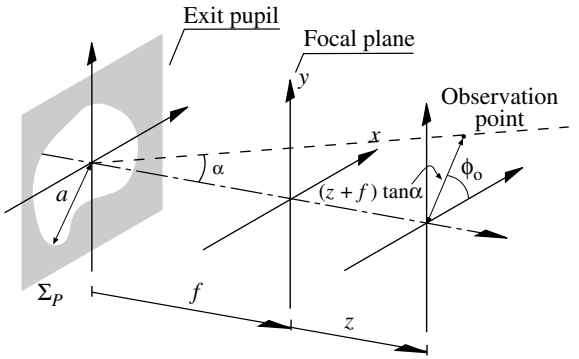


FIGURE 4.9 Trajectories in image space.

These curves correspond to straight lines passing through the axial point at the plane of the exit pupil. Together with the optical axis, each line defines a plane that forms an angle ϕ_o with the x axis, as depicted in Fig. 4.9. Note that the angle α of any of these lines with the optical axis is given by

$$\tan\alpha = Ka \tag{4.71}$$

For these subsets of observation points, the mapped pupil of the system can be expressed as

$$\begin{aligned} Q(r'_N, r_N(z), \phi(z), z) &= Q^{\alpha,\phi_o}(r'_N) \\ &= \sum_{n=-\infty}^{+\infty} i^n J_n\left(\frac{-2\pi a \tan\alpha}{\lambda} r'_N\right) Q_n(r'_N)\exp(in\phi_o) \end{aligned} \tag{4.72}$$

and analogously

$$r'^2_N = s + \frac{1}{2}, \quad q(s, r_N(z), \phi(z), z) = Q^{\alpha,\phi_o}(r'_N) = q^{\alpha,\phi_o}(s) \tag{4.73}$$

in such a way that now the corresponding RWT $RW_{q^{\alpha,\theta_o}}(x_\theta, \theta)$ is independent of the propagation parameter z. This is a very interesting issue since the calculation of the irradiance at any observation point lying on the considered line can be achieved from this single two-dimensional display by simply determining the particular coordinates $(x_\theta(z), \theta)$ through Eqs. (4.64) and (4.65). Furthermore, the proper choice of these straight paths allows one to obtain any desired partial feature of the whole three-dimensional image irradiance distribution. Note also that since W_{40} is just a parameter in these coordinates

and does not affect the function $RW_{q^{\alpha,\theta_0}}(x_\theta, \theta)$, this single display can be used for the determination of the irradiance for different amounts of SA. Thus, compared to classic techniques, the reduction in computation time is evident. The axial irradiance distribution is often used as a figure of merit for the performance of optical systems with aberrations. This distribution can be obtained here as a particular case with $\alpha = 0$, namely,

$$I(0, 0, z) = \frac{1}{\lambda^2(f + z)^2} RW_{q^{0,0}}(x_\theta(z), \theta) \tag{4.74}$$

where

$$q^{0,0}(s) = Q^{0,0}(r'_N) = Q_0(r'_N) \tag{4.75}$$

This result is especially interesting since this mapped pupil, and thus the associated RWT, is also independent of the wavelength λ. This fact represents an additional advantage when a polychromatic assessment of the imaging system is needed, as will be shown in forthcoming sections. Some quantitative estimation of these improvements is presented in Ref. 19.

To prove the performance of this computation method, next we present the result of the computation of the irradiance distribution along different lines in image space of two imaging systems, labeled system I and system II. For the sake of simplicity, we consider only purely absorbing pupils and no aberrations apart from SA in both cases. Thus, $Q(r'_N, \phi')$ reduces to the normalized pupil function $\bar{p}(r'_N, \phi')$. A gray-scale representation for the pure absorbing masks considered for each system is shown in Fig. 4.10.

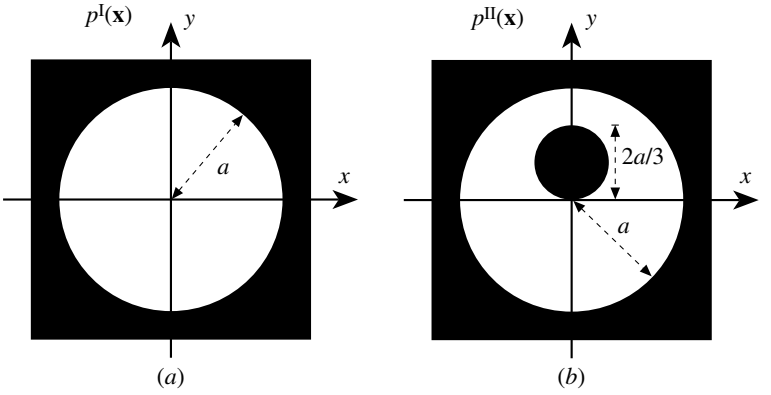


FIGURE 4.10 Gray-scale picture of the pupil functions for (*a*) system I and (*b*) system II.

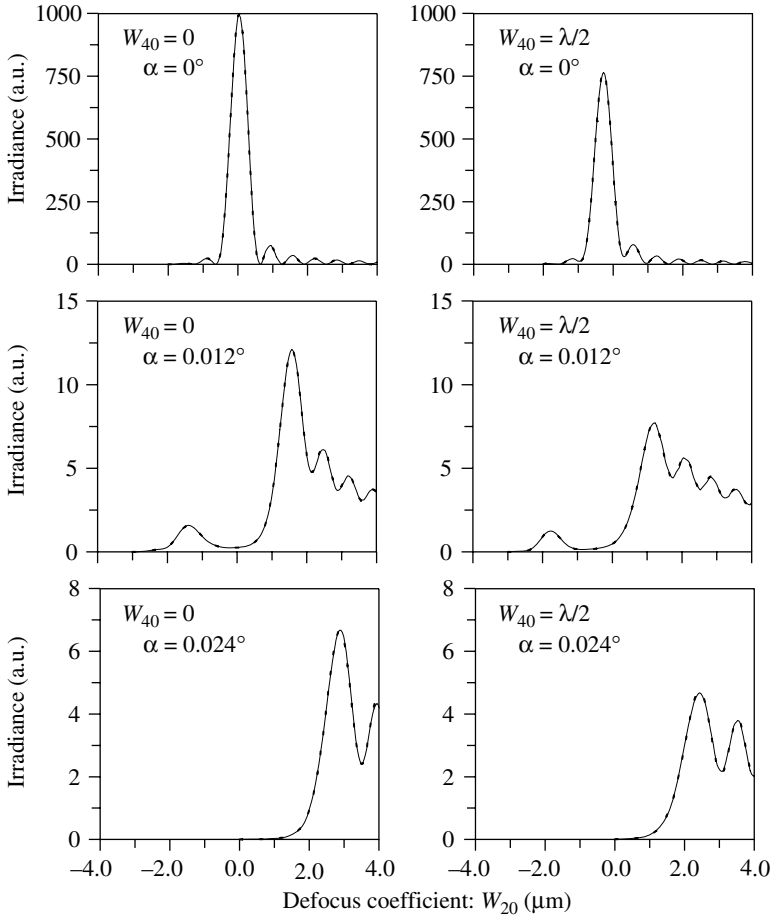


FIGURE 4.11 Irradiance values provided by system I, along different lines containing the axial point of the pupil and for two different amounts of SA. Continuous lines represent the result by the proposed RWT method while dotted lines stand for the computation by the classic method.

We compute the irradiance values for 256 points along three different lines passing through the axial point of the pupil, all characterized by an azimuthal angle $\phi_o = \pi/2$. These trajectories are chosen with tilt angles $\alpha = 0.024°$, $0.012°$, and $0°$ (optical axis). We set $a = 10$ mm, $z = 15.8$ mm, and $\lambda = 638.2$ nm. The function $RW_q^{\alpha,\theta_o}(x_\theta, \theta)$ was computed for 4096×4096 points, and for comparison purposes, the same irradiance values were computed by using the classic method[12,13] by partitioning the exit pupil of the imaging system into 1024×1024 radial-azimuthal elements. Figure 4.11 shows a joint representation of

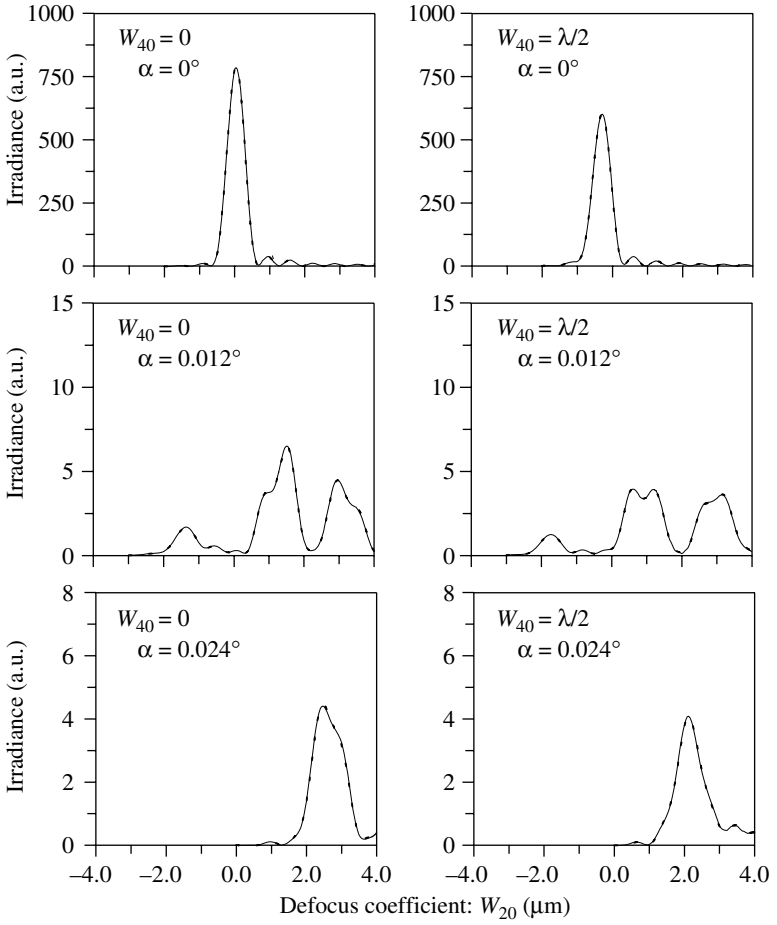


FIGURE 4.12 Irradiance values provided by system II, as in Fig. 4.11.

the numerical calculation for system I, when two different values of the SA are considered. The same results applied now to system II are presented in Fig. 4.12.

The analysis of these pictures shows that the results obtained with the RWT method match closely those obtained with the classic technique. In fact, both results differ by less than 0.03 percent. However, the RWT is much more efficient in this computation process. This is so because the basic RWT does not require recalculation for any point in each of the curves. This is also true for any amount of SA. Obviously, the greater the number of observation points, or SA values, that

have to be considered, the greater the resultant savings in computation time.

As a final remark on this subject, we want to point out that this approach can also be applied to other trajectories of interest in image space. For instance, short paths parallel to the optical axis in the neighborhood of the focal plane[17] or straight lines crossing the focal point can be considered.[22]

4.3.1.2 Parallel Optical Display of Diffraction Patterns

In Sec. 4.2.2 we mentioned that the mathematical relationship between Fresnel diffraction and the FrFT is given by Eq. (4.41). This means that the RWD is itself a continuous display of the evolution of diffraction patterns of one-dimensional objects, and this property is extremely useful from a pedagogical point of view. In fact, calculations of Fresnel and Fraunhofer diffraction patterns of uniformly illuminated one-dimensional apertures are standard topics in undergraduate optics courses. These theoretical predictions are calculated analytically for some typical apertures, or, more frequently, they are computed numerically. The evolution of these diffraction patterns under propagation is often represented in a two-dimensional display of gray levels in which one axis represents the transverse coordinate—pattern profile—and the other axis is related to the axial coordinate—evolution parameter.[23] This kind of representation illustrates, e.g., how the geometrical shadow of the object transforms into the Fraunhofer diffraction pattern as it propagates, and that the Fraunhofer diffraction simply is a limiting case of Fresnel diffraction.[24] In addition to the qualitative physical insight that the RWD provides about diffraction, it can provide a quantitative measurement of a variety of terms. These include the precise location y_s and the relative magnification M_s of each diffraction pattern. These two terms are quantitatively defined in terms of the maximum φ_h and minimum φ_0 powers of the varifocal lens L of the system represented in Fig. 4.5, i.e.,

$$y_s = \frac{h\sigma}{\sigma + l^2(\varphi_h - \varphi_0)}, \qquad M_s = 1 + \frac{\sigma}{l^2(\varphi_h - \varphi_0)} \tag{4.76}$$

where σ is the axial coordinate at which the corresponding diffraction pattern is localized under parallel illumination and h is the extent of the so-called progression zone of the varifocal lens. Figure 4.13 illustrates the experimental results registered by a CCD camera using a double slit as an input object. It can be seen that the RWD is a nice representation of the evolution by propagation of the interference phenomena. In fact, the Fraunhofer region of the diffracted field clearly shows the characteristic Young fringes modulated by a sinc function. To compare the theoretical and experimental results, a cross section of

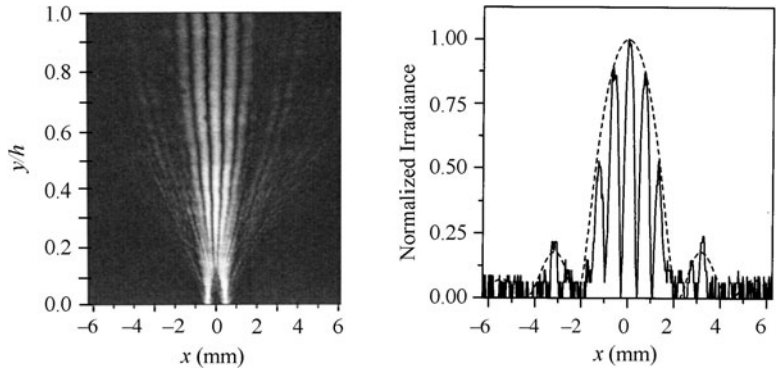


FIGURE 4.13 RDW showing the evolution of the field diffracted by a double slit. (*a*) Experimental result. (*b*) Cross section of the RWD showing the intensity profile near the Fraunhofer region. For comparison purposes, the theoretical sinc envelope of the Young fringes is also shown by the dotted line.

the experimental RWD for values of y/h close to 1 is also represented in Fig. 4.13.

Another classic example is diffraction by periodic objects. Here, self-imaging phenomena, such as the Talbot effect, are interesting and stimulating and usually attract the students' attention. As illustrated earlier in Fig. 4.6, which shows the diffraction patterns of a Ronchi grating, several self-imaging planes can be identified. It can be clearly seen that, due to the finite extent of the grating at the input, the number of Talbot images is limited by the so-called walk-off effect. Self imaging phenomena are discussed in more detail in Chap. 9 by Markus Testorf.

In addition to its use as an educational tool for displaying diffraction patterns, the RWD has been used to investigate diffraction by a variety of different interesting structures, including fractal diffraction screens. In fact, the properties of diffraction patterns produced by fractal objects and their potential applications have attracted the attention of several researchers during recent years because many natural phenomena and physical structures, such as phase transition, turbulence, or optical textures, can be analyzed and described by assuming fractal symmetry. Most research has been devoted to the study of diffraction patterns obtained from fractal objects in the Fraunhofer region,[25] yet it is in the Fresnel region where interesting features appear. For instance, Fresnel diffraction of a Cantor set[26] shows an irradiance distribution along the optical axis having a periodicity that depends on the level of the set. Furthermore, the intensity distributions at transverse planes

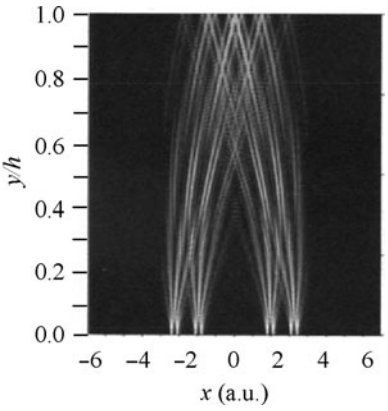


FIGURE 4.14 RWD as a display of all diffraction patterns generated by a Cantor grating of level 3.

show a partial self-similar behavior that is increased when moving toward the Fraunhofer region. For this reason, it is useful to represent the evolution of the complex amplitude of one-dimensional fractals propagating through free space represented on a two-dimensional display, especially if such a display can be obtained experimentally. In this case one axis represents the transversal coordinate, and the other is a function of the axial coordinate. In fact, according to the analysis carried out in Ref. 27, the evolution of the diffraction patterns allows one to determine the main characteristic parameters of the fractal. Therefore, one of the most important applications of the RWD has been in this field.[28] The RWD obtained for a triadic Cantor grating developed up to level 3 is shown in Fig. 4.14. Moreover, this result can be favorably compared with the results obtained with other displays.[27] The magnification provided by the lens L in the experimental setup (see Fig. 4.4) enables the RWD representation to provide an optimum sampling of the diffracted field. Near the object, where the diffraction patterns change rapidly, the mapping of the propagation distance provides a fine sampling, whereas the sampling is coarse in the far field where the variation of the diffraction patterns with the axial distance is slow. We note that sampling is the subject of Chap. 10.

4.3.2 Inverting RWT: Phase-Space Tomographic Reconstruction of Optical Fields

The WDF is an elegant and graphical way to describe the propagation of optical fields through linear systems. Since the WDF of a complex field distribution contains all the necessary information to retrieve the field itself,[29,30] many of the methods to obtain the WDF (and the AF)

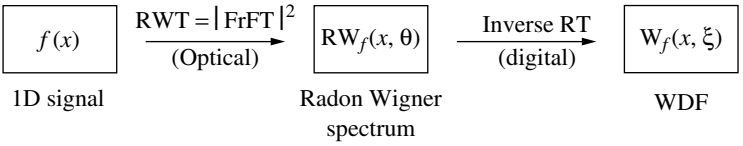


FIGURE 4.15 Diagram of the proposed hybrid optodigital method.

could be adapted to solve the phase retrieval problem. Optical or optoelectronic devices are the most commonly employed systems to obtain a representation of phase-space functions of one-dimensional or two-dimensional complex signals.[31,32] However, because most detectors used to this end are only sensitive to the incident intensity, interferometric or iterative methods are necessary in general to avoid loss of information. This is true even for the optically obtained WDF, which is real but has, in general, negative values and therefore is obtained from an intensity detector with an uncertainty in its sign. On the other hand, obtaining the WDF of wave fields is also possible indirectly through other representations such as the Radon transform.[33] In this particular case, a tomographic reconstruction is needed to synthesize the WDF. With this information it is possible to recover the amplitude and the phase of the original field distribution solely by means of intensity measurements. With most experimental setups for phase retrieval,[29,30] these measurements have to be taken sequentially in time while varying the distances between some components in each measurement. In this way the potential advantage of optics, i.e., parallel processing of signal information, is wasted. Consequently, another interesting application of the setup discussed in Sec. 4.2.2 to obtain the RWD is the experimental recovery of the WDF by means of an inverse Radon transformation.

The technique to obtain the WDF from projections is divided into two basic stages, sketched in Fig. 4.15. In the first stage, the experimental Radon-Wigner spectrum of the input function is obtained from a two-dimensional single-shot intensity measurement by the use of the experimental setup in Fig. 4.4. This optical method benefits from having no moving parts.

The second part of the proposed method is the digital computation of the inverse Radon transforms of the experimental Radon-Wigner spectrum. The most common algorithms used in tomographic reconstruction are based on the technique known as filtered backprojection. This algorithm is based on the *central slice theorem* discussed in Sec. 4.2.1. Thus, from Eqs. (4.21) and (4.22) we have

$$\mathcal{F}\{RW_f(x_\theta, \theta), \xi_\theta\} = \mathcal{F}_{2D}\{W_f(x, \xi), (\xi_\theta \cos\theta, \xi_\theta \sin\theta)\} \qquad (4.77)$$

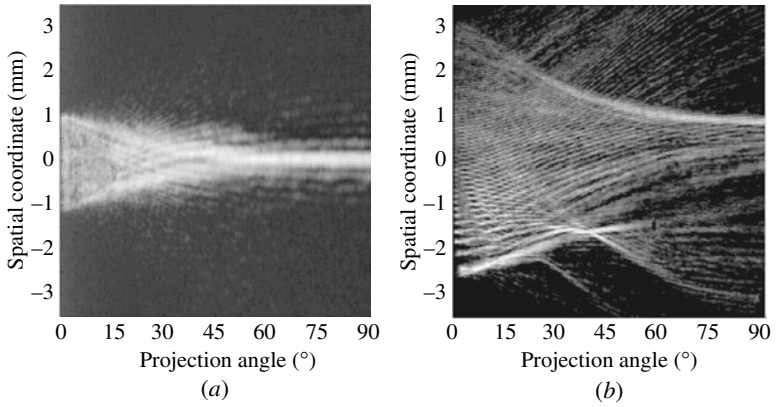


FIGURE 4.16 Experimental Radon-Wigner spectrum of (*a*) a single slit of 2.2 mm and (*b*) a binary grating with a linearly increasing spatial frequency.

where the one-dimensional FT is performed on the first argument of $RW_f(x_\theta, \theta)$. The inversion of this last transformation allows the recovery of $W_f(x, \xi)$ from its projections. Explicitly[34]

$$W_f(x, \xi) = \int_0^{\pi} C_f(x\cos\theta + \xi\sin\theta, \theta)\, d\theta \tag{4.78}$$

with

$$C_f(u, \theta) = \int_{-\infty}^{+\infty} \mathcal{F}\{RW_f(x_\theta, \theta), \xi_\theta\}\, |\xi_\theta| \exp(i2\pi\xi_\theta u)\, d\xi_\theta \tag{4.79}$$

Equation (4.79) can be clearly identified as a filtered version of the original RWT. In this way, from Eq. (4.78), $W_f(x, \xi)$ is reconstructed for each phase-space point as the superposition of all the projections passing through this point.

The experimental RWD of different one-dimensional functions has been used to reconstruct the WDF from projections. In Fig. 4.16 we show the RWD obtained with the optical device described in Sec. 4.2.2 for two different functions, namely, a rectangular aperture (single slit) and a grating with a linearly increasing spatial frequency (*chirp* signal).

To undertake the reconstruction of the WDF through the filtered backprojection algorithm, it is necessary to consider the complete angular region of the RWD, that is, $\theta \in [0, \pi)$. Although we only obtain optically the RWT for $\theta \in [0, \pi/2]$, the symmetry property in Eq. (4.17) has been used to complete the spectrum. From the experimental RWD

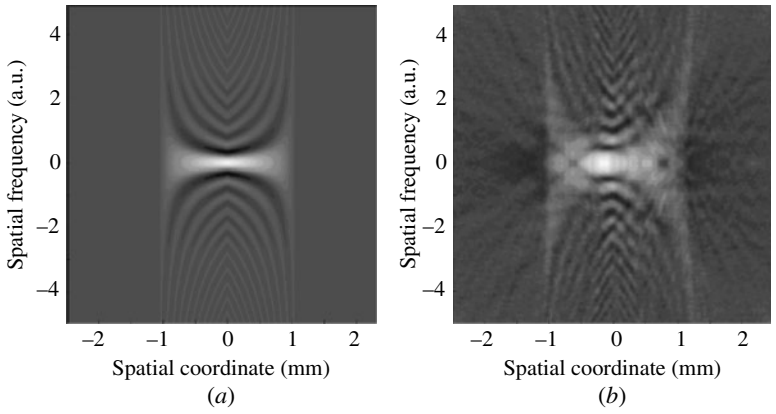


FIGURE 4.17 (*a*) Theoretical WDF of a single slit. (*b*) Experimental result for the tomographic reconstruction of the WDF of the same slit.

in Fig. 4.16, the corresponding WDFs have been obtained using the filtered backprojection algorithm. For comparison purposes, Figs. 4.17 and 4.18 show both the theoretical and the experimentally reconstructed WDF of the single slit and the chirp grating, respectively. Note that in Figs. 4.17*b* and 4.18*b* some artifacts appear. The lines radiating from the center and outward are typical artifacts (*ringing effect*) associated with the filtered backprojection method.[35] In spite

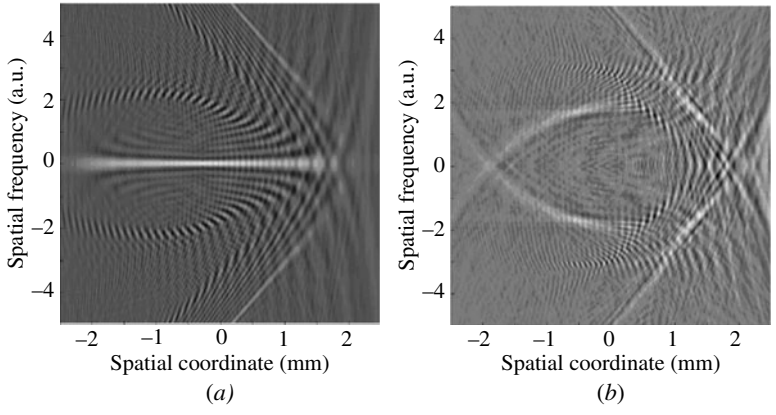


FIGURE 4.18 (*a*) Theoretical WDF of a binary grating with a linearly increasing spatial frequency. (*b*) Experimental tomographic reconstruction of the WDF of the same grating.

of this effect, a very good qualitative agreement can be observed between the results obtained with the theoretical and experimental data. The asymmetry in Fig. 4.17 is a consequence of the noise in Fig. 4.16*a*, reflecting also the asymmetry on the spatial coordinate in this figure. In Fig. 4.18 the typical arrow-shaped WDF of a chirp function can be observed in both cases. The slope in the arrowhead that characterizes the chirp rate of the signal is the same for the theoretical and the experimental results.

Several extensions of the proposed method are straightforward. On one hand, a similar implementation proposed here for the WDF can be easily derived for the AF, by virtue of Eq. (4.22). Note also that it is easy to extend this method to obtain two-dimensional samples of the four-dimensional WDF of a complex two-dimensional signal by use of a line scanning system. Moreover, since complex optical wave fields can be reconstructed from the WDF provided the inversion formulas, this approach can be used as a phase retrieval method that is an alternative to the conventional interferometric or iterative-algorithm-based techniques. In fact, as demonstrated,[36] phase retrieval is possible with intensity measurements at two close FrFT domains. This approach, however, requires some a priori knowledge of the signal bandwidth. In our method, a continuous set of FrFTs is available simultaneously, and this redundancy should avoid any previous hypothesis about the input signal.

4.3.3 Merit Functions of Imaging Systems in Terms of the RWT

4.3.3.1 Axial Point-Spread Function (PSF) and Optical Transfer Function (OTF)

There are several criteria for analyzing the performance of an optical imaging system for aberrations and/or focus errors in which the on-axis image intensity, or axial point-spread function (PSF), is the relevant quantity. Among them we mention:[37] Rayleigh's criterion, Marechal's treatment of tolerance, and the Strehl ratio (SR). As Hopkins suggested,[38] the analysis of Marechal can be reformulated to give a tolerance criterion based on the behavior of the optical transfer function (OTF) (spatial frequency information) instead of the PSF (space information). Phase-space functions were also employed to evaluate some merit functions and quality parameters.[39–41] This point of view equally emphasizes both the spatial and the spectral information contents of the diffracted wave fields that propagate in the optical imaging systems. Particularly, since the information content stored in the FrFT of an input signal changes from purely spatial to purely spectral as p varies from $p = 0$ to $p = 1$, that is, in the domain of the RWT, it is expected that the imaging properties of a given optical system, in

both the space and spatial frequency domains, could also be evaluated from the RWD.

To derive the formal relationship between the PSF (and the OTF) and the RWT, let us consider the monochromatic wave field, with wavelength λ, generated by an optical imaging system characterized by a one-dimensional pupil function $t(x)$, when a unit amplitude point source is located at the object plane. In the neighborhood of the image plane, located at $z = 0$, the field amplitude distribution can be written, according to the Fresnel scalar approximation, as

$$U(x, z) = \int_{-\infty}^{+\infty} t(x') \exp\left(\frac{-i\pi}{\lambda f} x'^2\right) \exp\left[\frac{i\pi}{\lambda(f+z)}(x'-x)^2\right] dx' \tag{4.80}$$

where f is the distance from the pupil to the image plane. The transformation of $t(x)$ to obtain the field $U(x, z)$ is given by a two-step sequence of elementary *abcd* transforms, namely, a spherical wavefront illumination (with focus at $\eta = f$) and a free-space propagation (for a distance $f + z$). Considering the results presented in Sec. 4.2.1, the *abcd* matrix associated with this transform can be found to be

$$M = \begin{pmatrix} 1 & 0 \\ \frac{1}{\lambda f} & 1 \end{pmatrix} \begin{pmatrix} 1 & -\lambda(f+z) \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & -\lambda(f+z) \\ \frac{1}{\lambda f} & -\frac{z}{f} \end{pmatrix} \tag{4.81}$$

and, therefore, the equivalent relationships to that given by Eq. (4.80) in terms of the corresponding RWTs can be expressed as [see Eq. (4.25)]

$$RW_{U(x,z)}(x_\theta, \theta) \propto RW_t(x_{\theta'}, \theta') \tag{4.82}$$

$$\tan\theta' = \frac{\lambda f \tan\theta - \lambda^2 f(f+z)}{\tan\theta - \lambda z}, \quad x_{\theta'} = \frac{x_\theta}{\sin\theta + \lambda(f+z)\cos\theta} \sin\theta'$$

In particular, the value $\theta = 0$ provides the irradiance distribution at the considered observation point, as stated in Sec. 4.2.1. This function is the PSF of the imaging system, as a function of the distance z to the image plane. Thus,

$$RW_{U(x,z)}(x_0, 0) = |U(x_0, z)|^2 = I(x_0, z) \propto RW_t(x_{\theta'_0}, \theta'_0) \tag{4.83}$$

$$\tan\theta'_0 = \frac{\lambda f(f+z)}{z}, \quad x_{\theta'_0} = \frac{x_0}{\lambda(f+z)} \sin\theta'_0$$

For the optical axis ($x_0 = 0$) the PSF can be expressed as

$$RW_{U(x,z)}(0, 0) = |U(0, z)|^2 = I(0, z) \propto RW_t\left(0, \arctan\left[\frac{\lambda f(f+z)}{z}\right]\right) \tag{4.84}$$

A normalized version of this axial irradiance is often used as a figure of merit of the performance of the imaging system, namely, the SR versus defocus, defined as

$$S(W_{20}) = \frac{I(0, z)}{I(0, 0)} \propto RW_t\left(0, \arctan\left(-\frac{\lambda h^2}{2W_{20}}\right)\right) \tag{4.85}$$

where h is the maximum lateral extent of the one-dimensional pupil and W_{20} stands for the one-dimensional version of the defocus coefficient defined in Eq. (4.51). Thus, the function $S(W_{20})$ can be analyzed in a polar fashion in the two-dimensional domain of the WDF associated with the pupil function $t(x)$ or, equivalently, in terms of its associated RWT.

To illustrate this approach, the defocus tolerance of different kinds of one-dimensional pupils was investigated, namely, a clear aperture (slit) and a pupil with a central obscuration (double slit). The general form of these pupils can be written as $t(x) = \text{rect}(x/h) - \text{rect}(x/b)$, with $b = 0$ for the uniform aperture. Figure 4.19 shows the RWD

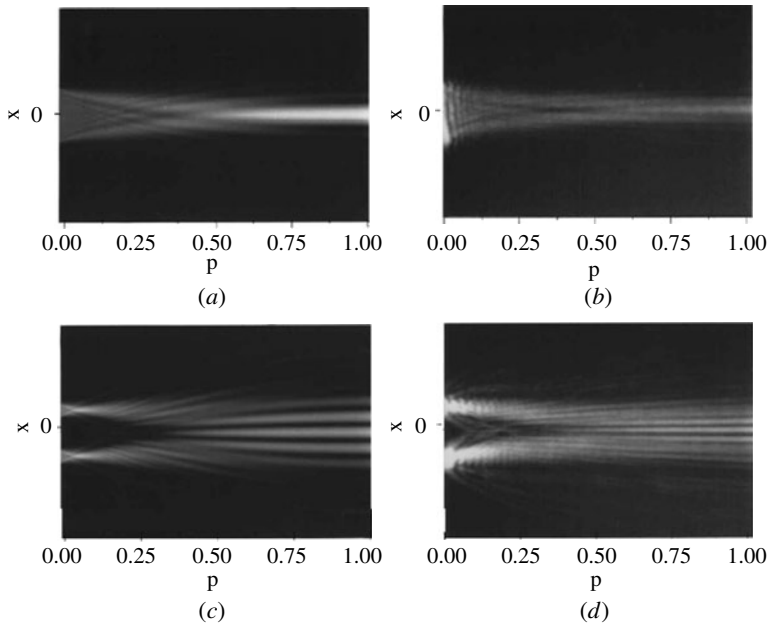


FIGURE 4.19 RWTs: (*a*) Computer simulation for an aperture with $h = 2.5$ mm and $b = 0$ mm. (*b*) Experimental result for (*a*). (*c*) Computer simulation for an aperture with $h = 2.5$ mm and $b = 1.3$ mm. (*d*) Experimental result for (*c*). The horizontal axis corresponds to the parameterization of the projection angle $\theta = p\pi/2$.

numerically (parts *a* and *c*) and experimentally (parts *b* and *d*) obtained, for two different values of the obscuration b.

According to our previous discussion, the slices of the RWD for $x = 0$ give rise to the SR for variable W_{20}. These profiles are plotted in Fig. 4.20 for three different pupils. From these results, it can be observed that, as expected, annular apertures have higher tolerance to defocus.

The knowledge of the SR is useful to characterize some basic features of any optical system, such as the depth of focus. However, the main shortcoming of the SR as a method of image assessment is that although it is relatively easy to calculate for an optical design prescription, it is normally difficult to measure for a real optical system. Moreover, the quality of the image itself is better described through the associated OTF. Fortunately, this information can also be obtained from the RWD via its relationship with the AF established in Sec. 4.2.1, since the AF contains all the OTFs $H(\xi; W_{20})$ associated with the optical system with varying focus errors according to the formula[42]

$$H_\lambda(\xi; W_{20}) = A_t\left(-\lambda(f+z)\xi, \frac{2W_{20}(f+z)}{h^2}\xi\right) \tag{4.86}$$

In this way, the AF of the pupil function $t(x)$ can be interpreted as a continuous polar display of the defocused OTFs of the system. Conversely,

$$A_t\left(x', \xi'\right) = H_\lambda\left(-\frac{x'}{\lambda(f+z)}; W_{20} = -\frac{\lambda h^2 \xi'}{2x'}\right) \tag{4.87}$$

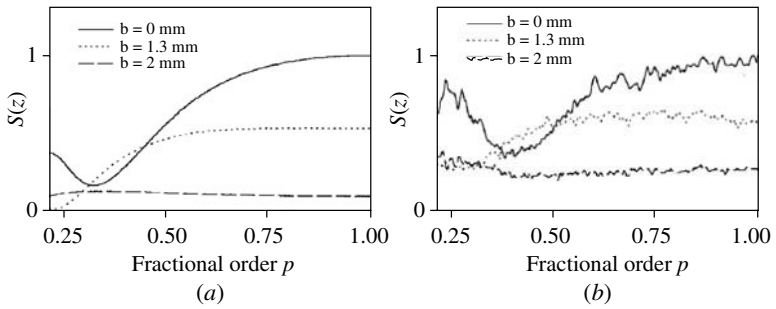


FIGURE 4.20 SR versus defocus for circular pupils with pupil function $t(x) = \text{rect}(x/h) - \text{rect}(x/b)$: (*a*) Computer simulation; (*b*) experimental results. Again, the projection angle $\theta = p\pi/2$.

Thus, by using Eq. (4.22) it is easy to find that

$$\mathcal{F}\{RW_t(x_\theta, \theta), \xi_\theta\} = A_t(\xi_\theta \cos\theta, -\xi_\theta \sin\theta)$$
$$= H_\lambda\left(-\frac{\xi_\theta \cos\theta}{\lambda(f+z)}; W_{20} = \frac{\lambda h^2}{2}\tan\theta\right) \quad (4.88)$$

Therefore, the one-dimensional FT of the profile of the RWD for a given value of the fractional order $\theta = p\pi/2$ corresponds to a defocused (scaled) OTF. This representation is quite convenient to visualize Hopkins' criterion.[39]

Figure 4.21 shows the one-dimensional Fourier transforms, taken with respect to the x variable, of the RWT illustrated in Fig. 4.19. From the previous analysis, the defocused OTFs are displayed along the vertical or spatial-frequency axis. These results for the clear aperture are shown in Fig. 4.22.

The RWD can also be used for calculating the OTF of an optical system designed to work under polychromatic illumination. In this case, as we will discuss next, a single RWD can be used to obtain the set of monochromatic OTFs necessary for its calculation.

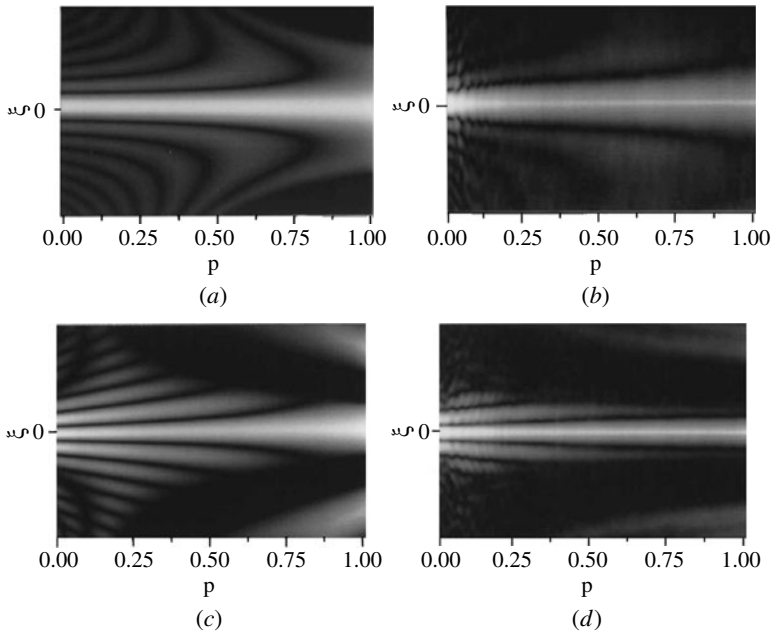


FIGURE 4.21 Computed one-dimensional FT of the RWDs shown in Fig. 4.19.

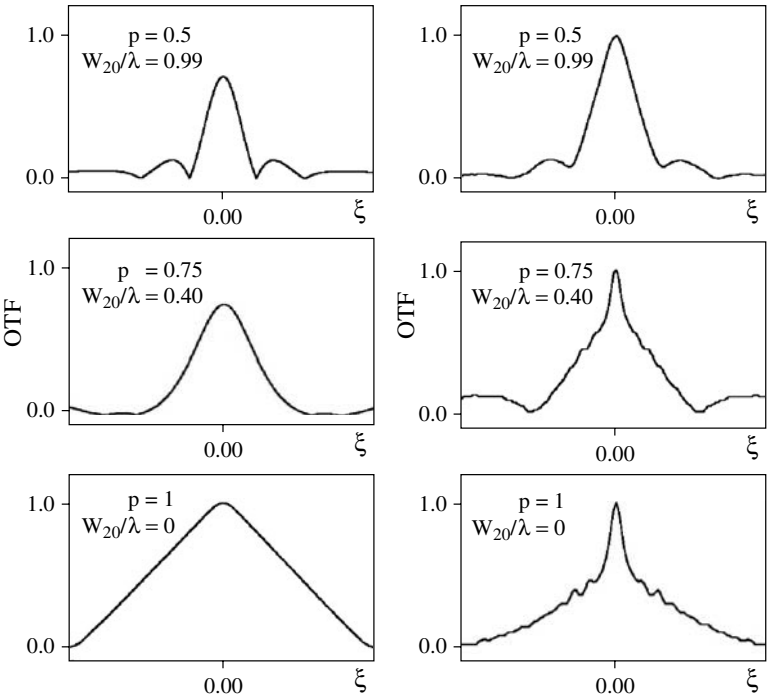


FIGURE 4.22 OTFs obtained from different slices of the intensity distributions shown in Fig. 4.21 for the case of a uniform aperture, for different amount of defocus.

4.3.3.2 Polychromatic OTF

As stated above, the RWT associated with the one-dimensional pupil of an imaging system can be used to obtain the OTF of the device, as a function of the defocus coefficient, through Eq. (4.88). It is worth noting that in this equation the wavelength λ of the incoming light acts as a parameter in the determination of the particular coordinates of the FT of the RWT, but it does not affect the RWT itself. Thus, changing the value of λ simply resets the position inside the same two-dimensional display for the computation of the OTF. The calculation procedure used in the previous section can be used, therefore, to compute the transfer function for any wavelength by means of the same RWD. This approach is based on previous work, where it was shown that the AF of the generalized pupil function of the system is a display of all the monochromatic OTFs with longitudinal chromatic aberration.[43]

An especially interesting application of this technique is the evaluation of the spatial-frequency behavior of optical systems working

under polychromatic illumination. In fact, the proper generalization of the OTF-based description to this broadband illumination case allows one to define quality criteria for imaging systems working with color signals.[44,45] This extension presents, however, some difficulties. The direct comparison of the incoming and the outgoing polychromatic irradiance distributions does not allow, in general, a similar relationship to the monochromatic case to be established. It can be shown, in fact, that only when the input signal is spectrally uniform can the frequency contents of both signals be related through a single polychromatic OTF function, providing the imaging system does not suffer from any chromatic aberrations regarding magnification.[46,47] Under these restrictions, a single polychromatic OTF can be used for relating input and output polychromatic irradiances in spatial-frequency space. This function is defined as

$$\mathcal{H}(\xi; W_{20}) = \frac{\int_{\Lambda} H_{\lambda}(\xi; W'_{20}(\lambda)) S(\lambda) V(\lambda)\, d\lambda}{\int_{\Lambda} S(\lambda) V(\lambda)\, d\lambda} \tag{4.89}$$

where Λ and $S(\lambda)$ are the spectral range and the spectral power of the illumination, respectively. The function $V(\lambda)$ represents the spectral sensitivity of the irradiance detector used to record the image. Note also that a new wavelength-dependent defocus coefficient has been defined, to account for the longitudinal chromatic aberration $\delta W_{20}(\lambda)$ that the system may suffer from, namely,

$$W'_{20}(\lambda) = W_{20} + \delta W_{20}(\lambda) \tag{4.90}$$

where W_{20} is the defocus coefficient defined in the previous section.

This OTF cannot account, however, for the chromatic information of the image, since only a single detector is assumed.[48] Indeed, by following the trichromacy of the human eye, three different chromatic channels are usually employed to properly describe color features in irradiance distributions, and, consequently, three different polychromatic OTFs are used, namely,[44,45]

$$\begin{aligned}
\mathcal{H}_X(\xi; W_{20}) &= \frac{\int_{\Lambda} H_{\lambda}(\xi; W'_{20}(\lambda)) S(\lambda) x_{\lambda}\, d\lambda}{\int_{\Lambda} S(\lambda) x_{\lambda}\, d\lambda} \\
\mathcal{H}_Y(\xi; W_{20}) &= \frac{\int_{\Lambda} H_{\lambda}(\xi; W'_{20}(\lambda)) S(\lambda) y_{\lambda}\, d\lambda}{\int_{\Lambda} S(\lambda) y_{\lambda}\, d\lambda} \\
\mathcal{H}_Z(\xi; W_{20}) &= \frac{\int_{\Lambda} H_{\lambda}(\xi; W'_{20}(\lambda)) S(\lambda) z_{\lambda}\, d\lambda}{\int_{\Lambda} S(\lambda) z_{\lambda}\, d\lambda}
\end{aligned} \tag{4.91}$$

where x_λ, y_λ, and z_λ are three spectral sensitivity functions associated with the measured chromaticity. These functions depend, obviously, on the specific color detector actually used. In the case of a conventional digital color camera, these channels can be associated with the R, G, and B bands of the three pixel families in the detector array. On the other hand, when a visual inspection of the final image is considered, these sensitivity functions are the well-known spectral tristimulus values of the human eye.[49]

Equations (4.91) establish the formulas to describe completely the response of a system from a spatial-frequency point of view. To numerically compute the functions described there, the evaluation of the monochromatic OTFs for a sufficient number of wavelengths inside the illumination spectrum has to be performed. Since any of these monochromatic transfer functions can be obtained from a same single RWD, as stated in the previous section, these computations can be done in a much more efficient way by use of this two-dimensional display. Furthermore, the same imaging system (i.e., the same pupil function) but suffering from different longitudinal chromatic aberration can be assessed as well, with no additional computation of the RWD. This is a critical issue in the saving of computation time which provides this technique with a great advantage compared to other classic techniques, as cited above.

To illustrate this technique, we present the result of the computation of the polychromatic OTFs associated with a conventional one-dimensional clear-pupil optical system (slit of width h) but suffering from two different chromatic aberration states (systems I and II from now on), as shown in Fig. 4.23. We assume that no other aberrations

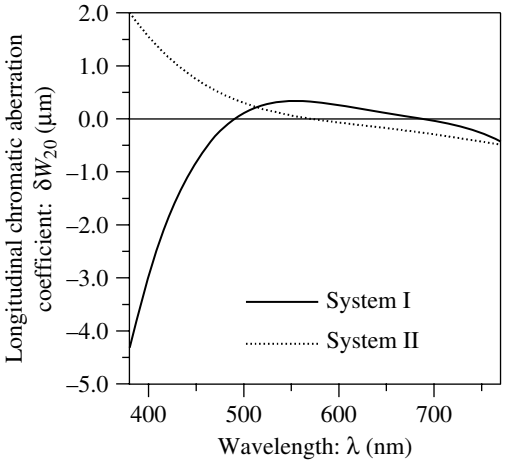


FIGURE 4.23 Longitudinal chromatic aberration coefficient associated with the two different correction states of the system under study.

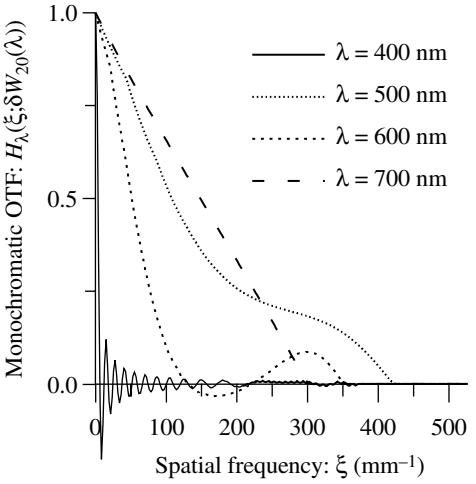


FIGURE 4.24 Monochromatic OTFs for system I in Fig. 4.23, corresponding to the imaging plane ($W_{20} = 0$).

are present. This assumption does not imply any restriction of the method, and the same applies to the one-dimensional character of the imaging system, as can be easily shown. Regarding the geometric parameters of the system, we fixed $h/f = 0.2$.

The evaluation of the corresponding monochromatic OTFs for both aberration states is achieved through the same computation method as in the previous section, namely, through the sequential one-dimensional FT of the two-dimensional display of the RWT $RW_t(x_\theta, \theta)$. Some of these results are shown in Fig. 4.24.

The computation of the polychromatic OTFs is performed next for both correction states, through the superposition of the monochromatic ones stated in Eqs. (4.91) for uniform sampling of 36 wavelengths in the range between 400 and 700 nm. The x_λ, y_λ, and z_λ functions are set to be the spectral tristimulus values of the standard human observer CIE 1931, while the spectral power for the illumination corresponds to the standard illuminant C[49]. The results for system I, corresponding to a defocused plane, and for system II, at the image plane, are shown in Fig. 4.25. Note that in both cases the same RWD is used in the computation, as stated above.

4.3.3.3 Polychromatic Axial PSF

In this section we propose the use of a single two-dimensional RWD to compute the axial irradiance in image space provided by an imaging

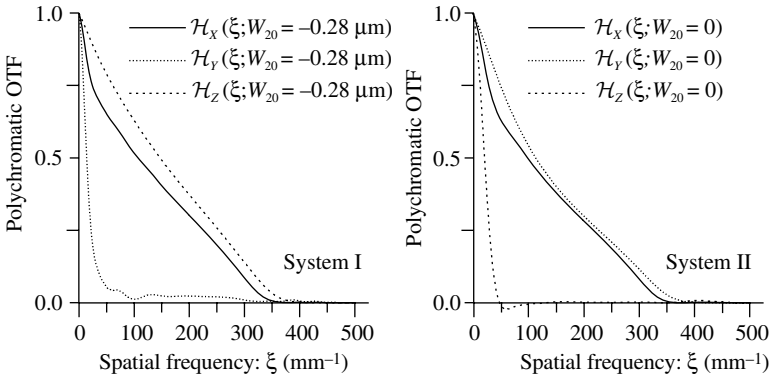


FIGURE 4.25 Polychromatic OTFs for (*a*) system I and (*b*) system II in Fig. 4.23, corresponding to a defocused plane ($W_{20} = -0.28$ μm) and image plane, respectively.

system with polychromatic illumination. In fact, the proposed technique is a straightforward extension of what is stated in Sec. 4.3.1.1, namely, that the axial irradiance distribution $I(0, 0, z)$ provided by a system with an arbitrary value of SA can be obtained from the single RWT $RW_{q^{0,0}}(x, \xi)$ of the mapped pupil $q^{0,0}(s)$ in Eq. (4.75). When an object point source is used, this irradiance distribution corresponds, of course, to the on-axis values of the three-dimensional PSF of the imaging system. For notation convenience we denote $I_\lambda(z) = I(0, 0, z)$ in this section.

According to the discussion in Sec. 4.3.3.2, the account for chromaticity information leads to a proper generalization of the monochromatic irradiances to the polychromatic case through three functions, namely,

$$\begin{aligned} X(W_{20}) &= \int_\Lambda I_\lambda(z)\, S(\lambda) x_\lambda \, d\lambda \\ Y(W_{20}) &= \int_\Lambda I_\lambda(z)\, S(\lambda) x_\lambda \, d\lambda \\ Z(W_{20}) &= \int_\Lambda I_\lambda(z)\, S(\lambda) x_\lambda \, d\lambda \end{aligned} \tag{4.92}$$

where $S(\lambda)$, $V(\lambda)$, x_λ, y_λ, and z_λ stand for the magnitudes used in the previous section. The defocus coefficient is defined in Eq. (4.51). However, it is often more useful to describe a chromatic signal through

a combination of these basic functions. A conventional choice for these new parameters is the set

$$x(W_{20}) = \frac{X(W_{20})}{X(W_{20}) + Y(W_{20}) + Z(W_{20})}$$
$$y(W_{20}) = \frac{Y(W_{20})}{X(W_{20}) + Y(W_{20}) + Z(W_{20})} \tag{4.93}$$

known as *chromaticity coordinates*, along with the parameter $Y(W_{20})$. If the sensitivity functions are selected to be the spectral tristimulus values of the human eye, the $Y(W_{20})$ parameter is known as illuminance and it is associated basically with the brightness of the chromatic stimulus. On the other hand, in this case the chromaticity coordinates provide a joint description for the hue and saturation of the colored signal.[49]

Anyway, as in the previous section, the evaluation of these magnitudes requires the computation of the monochromatic components for a sufficient number of spectral components. The use of conventional techniques, as stated earlier, is not very efficient at this stage, since the computation performed for a fixed axial point, a given wavelength, and a given aberration state cannot be applied to any other configuration. The method proposed in Sec. 4.3.1.1 represents a much more efficient solution since all the monochromatic values of the axial irradiance can be obtained, for different aberration correction states, from a single two-dimensional display associated with the pupil of the system.

To describe this proposal in greater detail, let us consider the system presented in Fig. 4.9 with $\alpha = 0$. According to the formulas in Sec. 4.3.1.1, the axial irradiance distribution in image space, for a given spectral component, can be expressed as

$$I_\lambda(z) = \frac{1}{\lambda^2(f+z)^2} RW_{q^{0,0}}(x_\theta(z), \theta) \tag{4.94}$$

where $q^{0,0}(s)$ represents the zero-order circular harmonic of the pupil $Q(r_N, \phi)$, with $s = r_N^2 + \frac{1}{2}$. The normalized coordinates r_N and ϕ are implicitly defined in Eq. (4.49). The specific coordinates $(x_\theta(z), \theta)$ for the RWT are given by Eqs. (4.64) and (4.65). Note that for systems with longitudinal chromatic aberration, the defocus coefficient W_{20} is substituted for the wavelength-dependent coefficient in Eq. (4.90). Note that now the whole dependence of the axial irradiance on λ, W_{40}, and z is established through these coordinates if the function $Q(r_N, \phi)$ itself does not depend on wavelength. This is the case when all the aberrations of the system, apart from SA and longitudinal chromatic aberration, have a negligible chromatic dependence. This is a very usual situation in well-corrected systems, and in this case, every axial

position, SA and chromatic aberration state, and wavelength can be studied from the same two-dimensional RWD.

Thus, providing that these kinds of systems are analyzed, the polychromatic description for the axial image irradiance can be assessed by the formulas

$$
\begin{aligned}
X(W_{20}) &= \int_{\Lambda} \frac{RW_{q^{0,0}}(x_\theta(z), \theta)}{\lambda^2(f+z)^2} S(\lambda) x_\lambda \, d\lambda \\
Y(W_{20}) &= \int_{\Lambda} \frac{RW_{q^{0,0}}(x_\theta(z), \theta)}{\lambda^2(f+z)^2} S(\lambda) y_\lambda \, d\lambda \\
Z(W_{20}) &= \int_{\Lambda} \frac{RW_{q^{0,0}}(x_\theta(z), \theta)}{\lambda^2(f+z)^2} S(\lambda) z_\lambda \, d\lambda
\end{aligned}
\tag{4.95}
$$

where the values of $(x_\theta(z), \theta)$ for every wavelength, axial position, and SA amount are given by Eqs. (4.64) and (4.65). Thus, once the RWD of the function $q^{0,0}(s)$ of the system is properly computed, these weighted superpositions can be quickly and easily calculated.[19,50,51]

As an example for testing this technique, we evaluate the axial response of a clear circular pupil imaging system, affected by spherical and longitudinal chromatic aberrations as shown in Fig. 4.26. Without loss of generality we assume here that the SA coefficient has a flat behavior for the considered spectral range. Once again, for the sake of simplicity, we assume that no other aberrations are present.

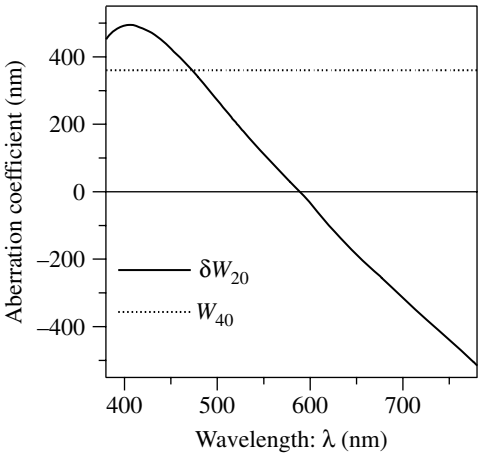


FIGURE 4.26 Aberration coefficients associated with the system under issue.

We consider 36 axial positions characterized by defocus coefficient values in a uniform sequence. We follow the same procedure as in earlier sections for the digital calculation of the RWD $RW_{q^{0,0}}(x_\theta, \theta)$. It is worth mentioning that for this pupil is possible to achieve an analytical result for the monochromatic axial behavior of the system for any value of W'_{20}, W_{40}, and λ, namely,[12]

$$I_\lambda(z) = \left[\frac{\pi a^2}{2\lambda f(f+z)}\right]^2 \frac{1}{W_{40}} \left| F\left[\frac{W'_{20}(\lambda) + 2W_{40}}{\sqrt{\lambda W_{40}}}\right] - F\left[\frac{W'_{20}(\lambda)}{\sqrt{\lambda W_{40}}}\right]\right|^2 \tag{4.96}$$

where

$$F(z) = \int_0^z \exp\left(\frac{i\pi t^2}{2}\right) dt \tag{4.97}$$

is the complex form of Fresnel integral. This analytical formula is used here to evaluate the results obtained by the proposed method. Figure 4.27 presents a comparison of these approaches for three different wavelengths in the visible spectral range. Excellent agreement can be observed in this figure.

Finally, we performed the calculation of the axial values for the chromaticity coordinates and the illuminance, by assuming the same

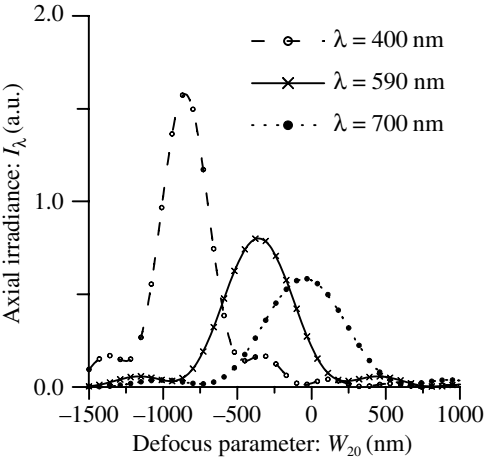


FIGURE 4.27 Axial irradiance values for the system under study. Solid lines represent the results by analytical calculation, while superimposed symbols correspond to the computation through the single RWD technique.

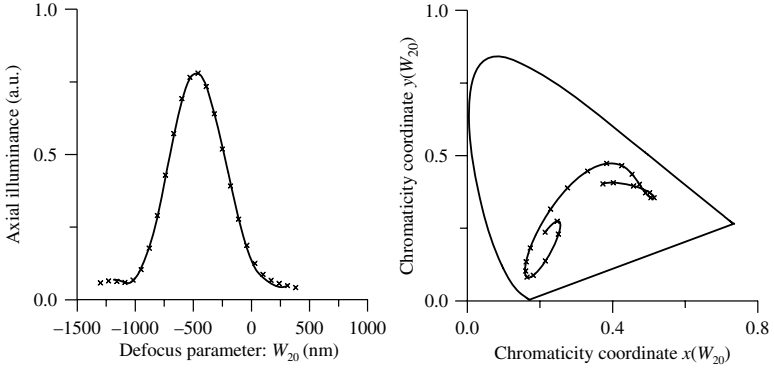


FIGURE 4.28 Axial illuminance and chromaticity coordinates for the system under study. Here solid lines represent the results obtained by the conventional method, while superimposed symbols correspond to the computation through the single RWD technique.

settings for the sensitivity functions and the illuminant as in the previous section. The values obtained with the method presented here are compared in Fig. 4.28 with the ones obtained by applying the same classic technique as in Sec. 4.3.3.2. Again, a very good agreement between them can be seen. A more detailed comparison of both methods is presented in Ref. 19.

4.4 Design of Imaging Systems and Optical Signal Processing by Means of RWT

4.4.1 Optimization of Optical Systems: Achromatic Design

We now present a design method for imaging systems working under polychromatic illumination on a RWT basis. In particular, we fix our attention on the optimal compensation of the axial chromatic dispersion of the Fresnel diffraction patterns of a plane object. Although this proposal can be applied to a wide variety of systems, we concentrate on an optical system specially designed for this purpose. This device allows us to obtain the image of any arbitrary diffraction pattern with very low residual chromatic aberration.[52,53] The optical system, sketched in Fig. 4.29, works under planar broadband illumination. The only two optical elements in this device are an achromatic lens, with focal length f, and an on-axis kinoform zone plate. This element acts,

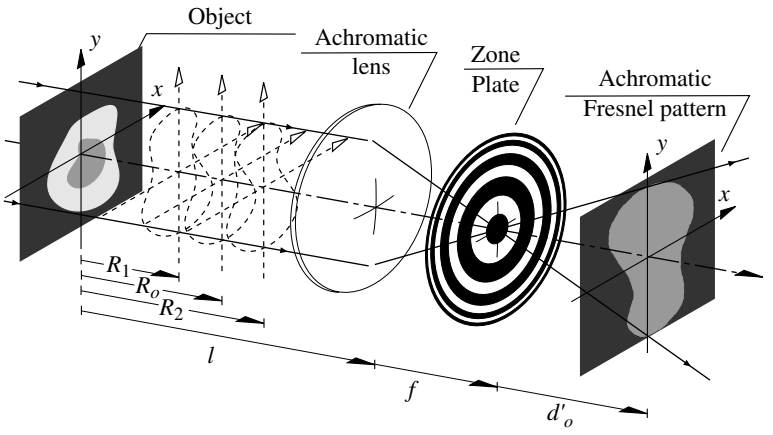


FIGURE 4.29 Achromatic imaging system under study.

from a scalar paraxial diffraction point of view, as a conventional thin lens with a focal length proportional to the inverse of the wavelength λ of the incoming light, i.e.,

$$Z(\lambda) = Z_o \frac{\lambda_o}{\lambda} \tag{4.98}$$

λ_o being a reference design wavelength and $Z_o = Z(\lambda_o)$. We note that although the effect of residual focuses can be significant for those wavelengths that are different from the design wavelength, we do not consider it here.

Our goal in this section is to achieve the optimal relationship between the geometric distances in the imaging system to obtain an output image corresponding to a given Fresnel pattern with minimum chromatic aberration. Thus, let us consider a given diffraction pattern located at a distance R_o from the object for the reference chromatic component of wavelength λ_o. It is well known that with parallel illumination the same diffraction pattern appears for any other spectral component at a distance from the input mask given by

$$R(\lambda) = R_o \frac{\lambda_o}{\lambda} \tag{4.99}$$

In this way, if the limits of the spectrum of the incoming radiation are λ_1 and λ_2, the same diffraction pattern is replicated along the optical axis between the planes characterized by distances $R_1 = R(\lambda_1)$ and $R_2 = R(\lambda_2)$, providing a *dispersion volume* for the diffraction pattern under study. However, if we fix our attention on the reference plane located at a distance R_o from the object, for $\lambda \neq \lambda_o$ we obtain a different structure

for the diffracted pattern, and, therefore, the final superposition of all the spectral components of the incoming light produces a chromatic blur of the monochromatic result. To analyze this effect, we again use the RWT approach to describe the spectral components of this Fresnel pattern. Since the above dispersion volume is transformed, by means of the imaging system, into a different image volume, it is interesting to derive the geometric conditions that provide a minimum value for its axial elongation in the image space. Equivalently, the same RWT analysis will be performed at the output of the system to analyze the chromatic blur at the output plane for the reference wavelength λ_o.

For the sake of simplicity, we consider a one-dimensional amplitude transmittance $t(x)$ for the diffracting object. Let us now apply our approach to calculate the irradiance free-space diffraction pattern under issue through the RWT of the object mask. If we recall the result in Eq. (4.36) for $z = R_o$, we obtain that each spectral component of this Fresnel pattern is given by

$$I_o(x;\lambda) \propto RW_t\left(x_\theta(\lambda), \theta(\lambda)\right), \quad \tan\theta(\lambda) = -\lambda R_o,$$
$$x_\theta(\lambda) = \frac{x}{\lambda R_o}\sin\theta(\lambda) \tag{4.100}$$

In this equation the chromatic blur is considered through the spectral variation of the coordinates in the RWT for any given transverse position x in the diffraction pattern. Thus, for a fixed observation position there is a region in the Radon space that contains all the points needed to compute the polychromatic irradiance. If we define $\theta_i = \arctan(\lambda_i R_o)$, for $i = 1, 2$, the width of this region in both Radon-space directions can be estimated as

$$\Delta\theta = |\theta_1 - \theta_2|, \qquad \Delta x_\theta = \left|\frac{x}{R_o}\left(\frac{\sin\theta_1}{\lambda_1} - \frac{\sin\theta_2}{\lambda_2}\right)\right| \tag{4.101}$$

Note that the smaller this region is, the less is the effect of the chromatic blur affecting the irradiance at the specified observation point. To achieve an *achromatization* of the selected diffraction pattern, this region has to be reduced in the output plane of the optical setup.

Let us now fix our attention on the effect of the imaging system on the polychromatic diffraction pattern under issue. Again, we use the RWT approach to achieve this description by simply noting that the system behaves as an *abcd* device that links the object plane and the selected output plane. The transformation matrix M_{achr} can be obtained as a sequence of elemental transformations (see Fig. 4.29), namely, free propagation at a distance l, propagation through the achromatic lens, free propagation to the focal plane of that element, passage through the zone plate, and, finally free propagation at a distance d_o'. The output plane is selected as the image plane of the diffraction pattern

under study for $\lambda = \lambda_o$. Thus, by using the results in Sec. 4.2.1, it is straightforward to obtain

$$M(\lambda) = \begin{pmatrix} 1 - \dfrac{l}{f} - \dfrac{f}{Z(\lambda)} & -\lambda\left(f - m_o f + m_o l + \dfrac{f^2 m_o}{Z(\lambda)}\right) \\ \dfrac{1}{\lambda f} & m_o \end{pmatrix} \tag{4.102}$$

where the following restriction applies to the fixed desired output plane

$$l = R_o + f - \frac{f}{m_o} - \frac{f^2}{Z_o} \tag{4.103}$$

where $m_o = -d_o'/f$ is the magnification obtained at the fixed image plane (for $\lambda = \lambda_o$). The relationship between the RWTs of the input object and the output Fresnel pattern for each spectral channel now can be established by application of Eqs. (4.27) and (4.28). In particular, by setting $\theta = 0$ we find

$$I_o'(x;\lambda) \propto RW_t\left(x_{\theta'}(\lambda), \theta'(\lambda)\right)$$

$$\tan\theta'(\lambda) = \lambda\left[R_o + \frac{f^2}{Z_o}\left(\frac{\lambda}{\lambda_o} - 1\right)\right], \qquad x_{\theta'} = \frac{x}{m_o}\cos\theta'(\lambda) \tag{4.104}$$

Therefore, for the polychromatic description of the output diffraction pattern we have to sum values of the RWT of the transmittance of the object in a region in the Radon domain whose size in both dimensions is given by

$$\Delta\theta' = |\theta'_{\max} - \theta'_{\min}|, \qquad \Delta x_{\theta'} = \left|\frac{x}{m_o}(\cos\theta'_{\max} - \cos\theta'_{\min})\right| \tag{4.105}$$

where

$$\theta'_{\max} = \max\{\theta'(\lambda)|\lambda \in [\lambda_1, \lambda_2]\}, \qquad \theta'_{\min} = \min\{\theta'(\lambda)|\lambda \in [\lambda_1, \lambda_2]\} \tag{4.106}$$

The specific values of these limits, which define the extension of the integration region in Radon space in the polychromatic case, depend on the particular values of the geometric design parameters f and Z_o of the imaging system. We now try to find a case that minimizes the chromatic blur in the output pattern. It is worth mentioning that exact achromatization of the pattern is achieved only when $\theta'(\lambda) = \theta'(\lambda_o)\ \forall\lambda \in [\lambda_1, \lambda_2]$, which cannot be fulfilled in practice, as can be seen from Eq. (4.104). However, a first-order approximation to that ideal correction can be achieved by imposing a stationary behavior for $\theta'(\lambda)$ around $\lambda = \lambda_o$. Mathematically, we impose

$$\left.\frac{d\theta'(\lambda)}{d\lambda}\right|_{\lambda_o} = 0 \tag{4.107}$$

or, equivalently,

$$\left.\frac{d\tan\theta'(\lambda)}{d\lambda}\right|_{\lambda_o} = 0 \tag{4.108}$$

which leads to the optimal constraint

$$R_o = -\frac{f^2}{Z_o} \tag{4.109}$$

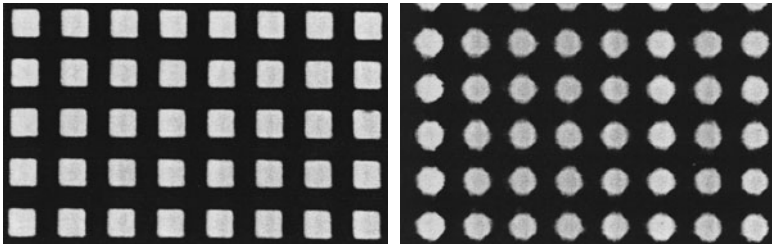
This condition transforms Eq. (4.103) into

$$l = 2R_o + f - \frac{f}{m_o} \tag{4.110}$$

Thus, the choice of a set of geometric parameters l, f, Z_o, and d_o' fulfilling the two above equations provides a design prescription for a first-order compensation of the chromatic blur in the diffraction pattern located, for $\lambda = \lambda_o$, at distance R_o from the object.[54]

To illustrate this design procedure and to check the predicted results, we present an experimental verification by using a two-dimensional periodic transmittance as an object, with the same period $p = 0.179$ mm in both orthogonal directions. As a Fresnel pattern to be achromatized, a self-imaging distribution is selected. In particular, after parallel illumination with $\lambda_o = 546.1$ nm, the distance $R_o = 11.73$ cm is selected. Figure 4.30*a* shows a picture of the irradiance distribution in that situation. In Fig. 4.30*b*, the irradiance distribution over the same plane, but when a polychromatic collimated beam from a high-pressure Hg lamp is used, is presented. The chromatic blur is clearly seen by comparing these two figures.

To optimally achromatize this diffraction pattern, we follow the prescriptions given in the above paragraphs. We use a kinoform lens



(*a*) (*b*)

FIGURE 4.30 Gray-scale display of the irradiance distribution to be achromatized: (*a*) Monochromatic pattern for $\lambda_o = 546.1$ nm. (*b*) Broadband (Hg lamp) irradiance distribution.

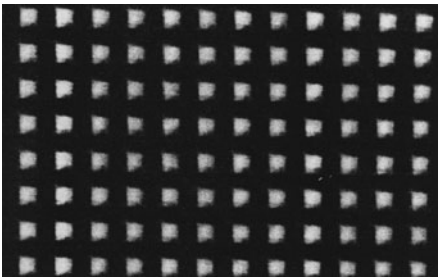


FIGURE 4.31 Gray-scale display of the achromatized irradiance distribution.

with $Z_o = -12$ cm, and we choose a value $d'_o = 10.00$ cm. Therefore, we select a focal distance for the achromatic lens $f = \sqrt{-Z_o R_o} = 11.86$ cm, and we place that object at a distance $l = 2R_o + f + f^2/d'_o = 49.39$ cm from that lens. A gray-scale display of the output irradiance is presented in Fig. 4.31. The comparison between this result and the monochromatic one in Fig. 4.30*a* shows the high achromatization level obtained with the optimized system.

4.4.2 Controlling the Axial Response: Synthesis of Pupil Masks by RWT Inversion

In Sec. 4.3.1.1 we showed that the axial behavior of the irradiance distribution provided by a system with an arbitrary value of SA can be obtained from the single RWT of the mapped pupil $q^{0,0}(s)$ of the system. In fact, Eq. (4.74) can be considered the keystone of a pupil design method[55] in which the synthesis procedure starts by performing a tomographic reconstruction of $W_{q^{0,0}}(x, \xi)$ from the projected function $I(0, 0, z)$ representing the irradiance at the axial points—variable W_{20}—for a sufficient set of values of W_{40}. Thus, the entire two-dimensional Wigner space can be sampled on a set of lines defined by these parameters. The backprojection algorithm converts the desired axial irradiance for a fixed value of W_{40}, represented by a one-dimensional function, to a two-dimensional function by smearing it uniformly along the original projection direction (see Fig. 4.8). Then the algorithm calculates the summation function that results when all backprojections are summed over all projection angles θ, i.e., for all the different values of W_{40}. The final reconstructed function $W_{q^{0,0}}(x, \xi)$ is obtained by a proper filtering of the summation image.[55] Once the WDF is synthesized with the values of the input axial irradiances, the pupil function is obtained by use of Eq. (4.4). Finally, the geometric mapping in Eq. (4.57) is inverted to provide the desired pupil function.

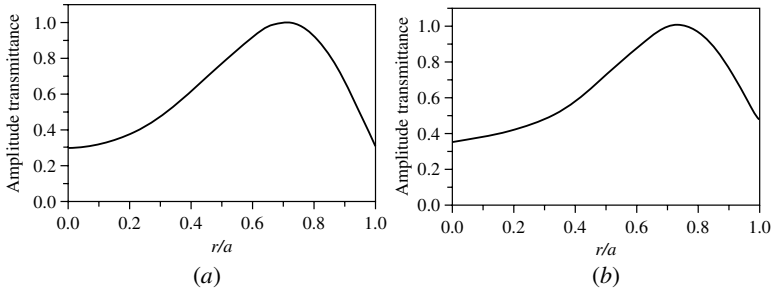


FIGURE 4.32 (*a*) Amplitude transmittance of a desired pupil function. (*b*) Phase-space tomographic reconstruction of the same pupil.

To illustrate the method, we numerically simulated the synthesis of an annular apodizer represented in Fig. 4.32. It has been shown that its main features are to increase the focal depth and to reduce the influence of SA. From this function we numerically determined first the $W_{q^{0,0}}(x, \xi)$ function, using the WDF definition, and thereby the projected distributions defined by the RWT, obtaining the axial irradiance distribution for different values of SA. In this case, we used 1024 values for both W_{40}/λ and W_{20}/λ, ranging from -16 to $+16$. We treated these distributions as if they represented the desired axial behavior for a variable SA, and we reconstructed the WDF by using a standard filtered backprojection algorithm for the inverse Radon transform. From the reconstructed WDF we obtained the synthesized pupil function $p(\mathbf{x})$ by performing the discrete one-dimensional inverse FT of $W_{q^{0,0}}(x, \xi)$. The result is shown in Fig. 4.32*b*. As can be seen, the amplitude transmittance of the synthesized pupil function closely resembles the original apodizer in Fig. 4.32*a*.

4.4.3 Signal Processing through RWT

Throughout this chapter we have discussed the RWT as a mathematical tool that allows us to develop novel and interesting applications in optics. Among several mathematical operations that can be optically implemented, correlation is one of the most important because it can be used for different applications, such as pattern recognition and object localization. Optical correlation can be performed in coherent systems by use of the fact that the counterpart of this operation in the Fourier domain is simply the product of both signals. To implement this operation, several optical architectures were developed, such as the classic VanderLugt and joint transform correlators.[56,57] Because

conventional correlation is a shift-invariant operation, the correlation output simply moves if the object translates at the input plane. In many cases this property is necessary, but there are situations in which the position of the object provides additional information such as in image coding or cryptographic applications, and so shift invariance is a disadvantage.

The fractional correlation[58,59] is a generalization of the classic correlation that employs the optical FrFT of a given fractional order p instead of the conventional FT. Conventionally, the fractional correlation is obtained as the inverse Fourier transform of the product of the FrFT of both the reference and the input objects, but for a single fractional order p at a time. The fractional order involved in the FrFT controls the amount of shift variance of the correlation. As is well known, the shift-variance property modifies the intensity of the correlation output when the input is shifted. In several pattern recognition applications this feature is useful, for example, when an object should be recognized in a relevant area and rejected otherwise, or when the recognition should be based on certain pixels in systems with variable spatial resolution. However, the optimum amount of variance for a specific application is frequently difficult to predict, and therefore more complete information would certainly be attained from a display showing several fractional correlations at the same time. Ideally, such a display should include the classic shift-invariant correlation as the limiting case. In this section we will show that such a multichannel fractional correlator could be easily implemented from the RWD system presented in Sec. 4.2.2. The resulting optical system generates a simultaneous display of fractional correlations of a one-dimensional input for a continuous set of fractional orders in the range $p \in [0, 1]$.

We start by recalling[58] the definition of the fractional correlation between two one-dimensional functions $f(x)$ and $f'(x)$

$$C_p(x) = \mathcal{F}^{-1}\{F_p(\alpha)F_p'^{*}(\alpha), x\} \tag{4.111}$$

It is important to note that with the above definition the classic correlation is obtained if we set $p = 1$. The product inside the brackets of Eq. (4.111) can be optically achieved simultaneously for all fractional orders, ranging between $p = 0$ and $p = 1$, following a two-step process. In the first stage, the RWD of the input is obtained with the experimental configuration shown in Sec. 4.2.2. A matched filter can be obtained at the output plane if, instead of recording the intensity, we register a hologram of the field distribution at this plane with a reference wavefront at an angle θ. (see Fig. 4.33).

In the second stage, the obtained multichannel matched filter is located at the filter plane, and the input function to be correlated is located at the input plane (see Fig. 4.34).

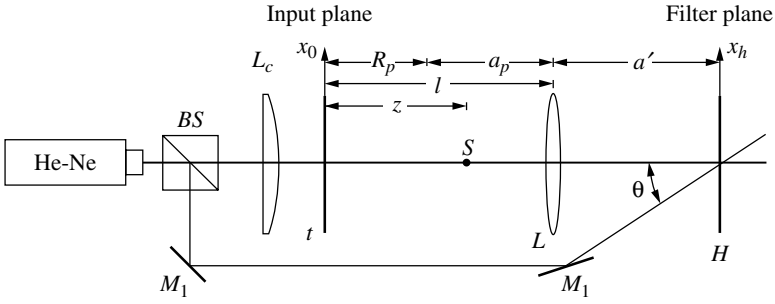


FIGURE 4.33 Multichannel matched filter registration for a fractional correlation. The elements are the same as in Sec. 4.2.2, except for BS (beam splitter) and $M1$ and $M2$ (plane mirrors).

Because the transmittance of the holographic filter has one term proportional to the complex conjugate of the reference field in Eq. (4.111), for each fractional order channel the field immediately behind the filter plane has one term proportional to the product of the complex conjugate of the FrFT of the reference function $f'(x)$ and the same FrFT of the input function $f(x)$. Thus the multiplicative phase factor in this equation and the corresponding one of the matched filter cancel out. Besides, although the experimental FrFT for a given order p is approximated owing to the scale error discussed in Sec. 4.2.2, the experimental fractional correlation can be obtained exactly because this error affects both $F_p(\alpha)$ and $F_p'^*(\alpha)$. Finally, the diffracted field at angle θ is collected by the lens L_c, which performs a one-dimensional FT. Because each fractional order $p \in [0, 1]$ has an independent one-dimensional correlation channel, all the fractional correlations for this

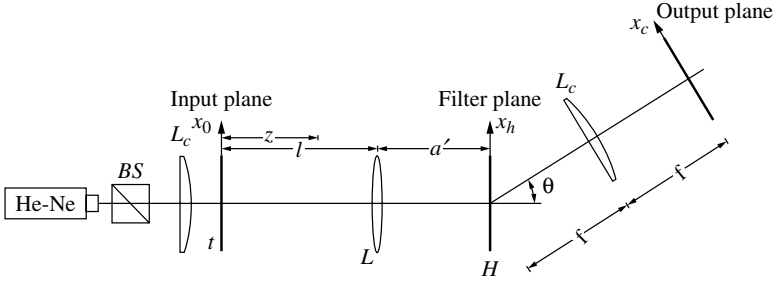


FIGURE 4.34 Multichannel fractional correlator. The filter H corresponds to the one obtained in the setup of Fig. 4.33.

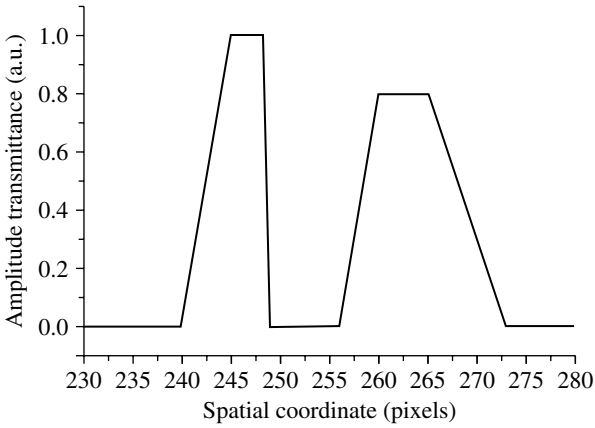


FIGURE 4.35 Amplitude transmittance of an input object selected to perform multichannel fractional correlation.

range of fractional orders are obtained simultaneously at the output plane. Thus a two-dimensional display is obtained in which the fractional correlations are ordered in a continuous display along the axis normal to the plane shown in Fig. 4.34.

The shift-variant property of the FrFT correlation was confirmed experimentally in Ref. 60. Here we present a numerical simulation using an input object whose amplitude transmittance is shown in Fig. 4.35. It represents a double nonsymmetric slit with a continuous gray-level amplitude transmittance. The continuous transition between the shift-variant case $p = 0$ and the shift-invariant case $p = 1$ is confirmed in Fig. 4.36. In this figure the fractional autocorrelation of the input is considered, but the reference objects are shifted at the input plane.

Figure 4.36*a* shows the fractional correlations when the input is shifted an amount of one-half of the object size, and Fig. 4.36*b* shows the fractional correlation when the input is shifted an amount equal to the size of the object. The variant behavior of the fractional correlation can be clearly seen by the comparison of these figures. Both displays coincide near to $p = 1$ (except for the location of the maxima), but for lower values of p the fractional correlation is highly dependent on the magnitude of the shift. As can be seen in the three-dimensional plot in this figure, for a fixed displacement the correlation peak increases with p. As expected for $p = 1$, the correlation peak is the classic one located at the input position. For values ranging between $p = 0.5$ and $p = 1$, the correlation peak did not change appreciably. The

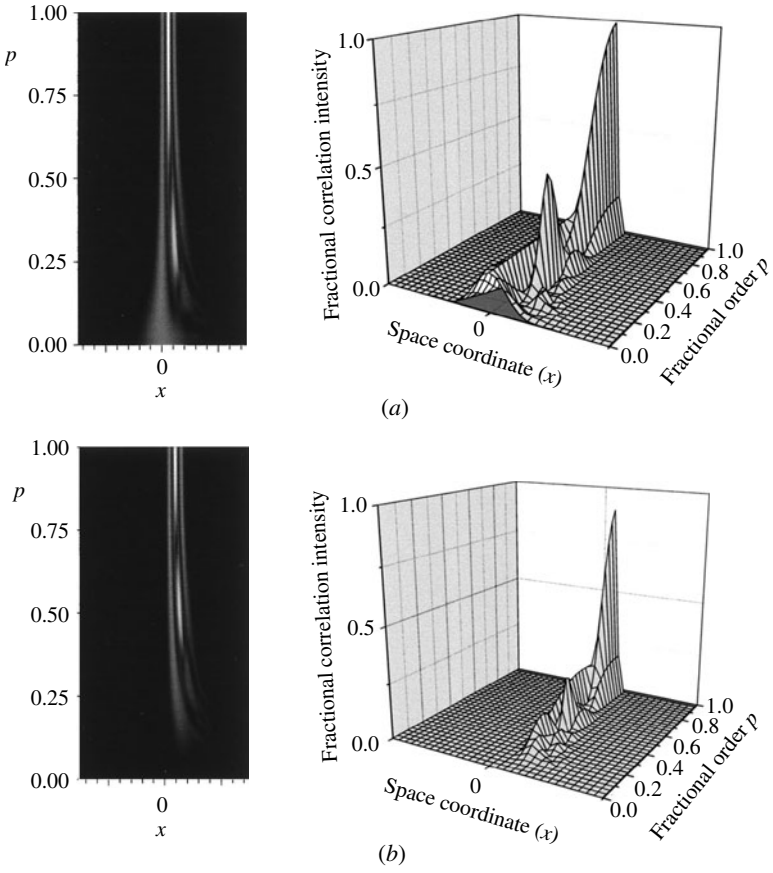


FIGURE 4.36 Multichannel fractional autocorrelation of the function represented in Fig. 4.35 with (*a*) a shift in the input plane of one-half of the object size and (*b*) a shift of the whole size of the object.

shift-variant property becomes evident for values close to $p = 0.25$. It can be seen that as the fractional order becomes lower, the peak degenerates and shifts disproportionately toward the object position. Thus, the output of the system shows a variable degree of space variance ranging from the pure shift variance case $p = 0$ to the pure shift invariance case $p = 1$, that is, the classic correlation. This kind of representation provides information about the object, such as classic correlation, but also quantifies its departure from a given reference position.

Acknowledgments

The authors would like to express their gratitude to P. Andrés, S. Granieri, E. Sicre, and E. Silvestre for their contributions in the research revisited in this chapter. They also acknowledge financial support of Ministerio de Ciencia y Tecnología, Spain (Grants DPI 2006-8309 and DPI 2008-02953).

References

1. J. C. Wood and D. T. Barry, "Radon transformation of time-frequency distributions for analysis of multicomponent signals," *Proc. Int. Conf. Acoust. Speech Signal Process.* **4**: 257–261 (1992).
2. J. C. Wood and D. T. Barry, "Linear signal synthesis using the Radon-Wigner transform," *IEEE Trans. Signal Process.* **42**: 2105–2111 (1994).
3. L. Cohen, *Time-Frequency Analysis*, Prentice-Hall, Upper Saddle River, N.J., 1995.
4. D. Mendlovic, R. G. Dorsch, A. W. Lohmann, Z. Zalevsky, and C. Ferreira, "Optical illustration of a varied fractional Fourier-transform order and the Radon-Wigner display," *Appl. Opt.* **35**: 3925–3929 (1996).
5. A. W. Lohmann, "Image rotation, Wigner rotation, and the fractional Fourier transform," *J. Opt. Soc. Am. A* **10**: 2181–2186 (1993).
6. S. Granieri, W. D. Furlan, G. Saavedra, and P. Andrés, "Radon-Wigner display: A compact optical implementation with a single varifocal lens," *Appl. Opt.* **36**: 8363–8369 (1997).
7. P. Pellat-Finet, "Fresnel diffraction and the fractional-order Fourier transform," *Opt. Lett.* **19**: 1388–1390 (1994).
8. P. Andrés, W. D. Furlan, G. Saavedra, and A. W. Lohmann, "Variable fractional Fourier processor: A simple implementation," *J. Opt. Soc. Am. A* **14**: 853–858 (1997).
9. E. Tajahuerce, G. Saavedra, W. D. Furlan, E. E. Sicre, and P. Andrés, "White-light optical implementation of the fractional fourier transform with adjustable order control," *Appl. Opt.* **39**: 238–245 (2000).
10. Y. Zhang, B. Gu, B. Dong, and G. Yang, "Optical implementations of the Radon-Wigner display for one-dimensional signals," *Opt. Lett.* **23**: 1126–1128 (1998).
11. Y. Zhang, B.-Y. Gu, B.-Z. Dong, and G.-Z. Yang, "New optical configurations for implementing Radon-Wigner display: Matrix analysis approach," *Opt. Comm.* **160**: 292–300 (1999).
12. H. H. Hopkins and M. J. Yzuel, "The computation of diffraction patterns in the presence of aberrations," *Optica Acta* **17**: 157–182 (1970).
13. M. J. Yzuel and F. Calvo, "Point-spread function calculation for optical systems with residual aberrations and non-uniform transmission pupil," *Optica Acta* **30**: 233–242 (1983).
14. J. J. Stamnes, B. Spjelkavik, and H. M. Pedersen, "Evaluation of diffraction integrals using local phase and amplitude approximations," *Optica Acta* **30**: 207–222 (1983).
15. H. G. Kraus, "Finite element area and line integral transforms for generalization of aperture function and geometry in Kirchhoff scalar diffraction theory," *Opt. Eng.* **32**: 368–383 (1993).
16. L. A. D'Arcio, J. J. M. Braat, and H. J. Frankena, "Numerical evaluation of diffraction integrals for apertures of complicated shape," *J. Opt. Soc. Am. A* **11**: 2664–2674 (1994).
17. G. Saavedra, W. D. Furlan, E. Silvestre, and E. Sicre, "Analysis of the irradiance along different paths in the image space using the Wigner distribution function," *Opt. Comm.* **139**: 11–16 (1997).

18. W. D. Furlan, G. Saavedra, E. Silvestre, and M. Martínez-Corral, "On-axis irradiance for spherically aberrated optical systems with obscured rectangular apertures: A study using the Wigner distribution function," *J. Mod. Opt.* **45**: 69–77 (1998).
19. W. D. Furlan, G. Saavedra, E. Silvestre, J. A. Monsoriu, and J. D. Patrignani, "Assessment of a Wigner-distribution-function-based method to compute the polychromatic axial response given by an aberrated optical system," *Opt. Eng.* **42**: 753–758 (2003).
20. W. D. Furlan, G. Saavedra, J. A. Monsoriu, and J. D. Patrignani, "Axial behaviour of Cantor ring diffractals," *J. Opt. A: Pure Appl. Opt.* **5**: S361–S364 (2003).
21. W. D. Furlan, M. Martínez-Corral, B. Javidi, and G. Saavedra, "Analysis of 3-D integral imaging displays using the Wigner distribution," *J. Disp. Technol.* **2**: 180–185 (2006).
22. P. Andrés, M. Martínez-Corral, and J. Ojeda-Castañeda, "Off-axis focal shift for rotationally nonsymmetric screens," *Opt. Lett.* **18**: 1290–1292 (1993).
23. A. Dubra and J. A. Ferrari, "Diffracted field by an arbitrary aperture," *Am. J. Phys.* **61**: 87–92 (1999).
24. W. D. Furlan, G. Saavedra, and S Granieri, "Simultaneous display of all the Fresnel diffraction patterns of one-dimensional apertures," *Am. J. Phys.* **69**: 799 (2001).
25. C. Allain and M. Cloitre, "Optical diffraction on fractals," *Phys. Rev.* **B33**: 3566–3569 (1986).
26. Y. Sakurada, J. Uozumi, and T. Asakura, "Fresnel diffraction by one-dimensional regular fractals," *Pure Appl. Opt.* **1**: 29–40 (1992).
27. T. Alieva and F. Agulló-López, "Optical wave propagation of fractal fields," *Opt. Comm.* **125**: 267–274 (1996).
28. O. Trabocchi, S. Granieri, and W. D. Furlan, "Optical propagation of fractal fields: Experimental analysis in a single display," *J. Mod. Opt.* **48**: 1247–1253 (2001).
29. M. G. Raymer, M. Beck, and D. F. McAlister, "Complex wave-field reconstruction using phase-space tomography," *Phys. Rev. Lett.* **72**: 1137–1140 (1994).
30. D. F. McAlister, M. Beck, L. Clarke, A. Mayer, and M. G. Raymer, "Optical phase retrieval by phase-space tomography and fractional-order Fourier transforms," *Opt. Lett.* **20**: 1181–1183 (1994).
31. Y. Li, G. Eichmann, and M. Conner, "Optical Wigner distribution and ambiguity function for complex signals and images," *Opt. Comm.* **67**: 177–179 (1988).
32. G. Shabtay, D. Mendlovic, and Z. Zalevsky, "Proposal for optical implementation of the Wigner distribution function," *Appl. Opt.* **37**: 2142–2144 (1998).
33. R. L. Easton, Jr., A. J. Ticknor, and H. H. Barrett, "Application of the Radon transform to optical production of the Wigner distribution," *Opt. Eng.* **23**: 738–744 (1984).
34. W. D. Furlan, C. Soriano, and G. Saavedra, "Opto-digital tomographic reconstruction of the Wigner distribution function of complex fields," *Appl. Opt.* **47**: E63–E67 (2008).
35. A. C. Kak and M. Slaney, *Principles of Computerized Tomographic Imaging*, IEEE Press, New York, 1988.
36. U. Gopinathan, G. Situ, T. J. Naughton, and J. T. Sheridan, "Noninterferometric phase retrieval using a fractional Fourier system," *J. Opt. Soc. Am. A* **25**: 108–115 (2008).
37. M. Born and E. Wolf, *Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light*, (7th ed., Cambridge University Press, Cambridge, 1999, Chap. 9.
38. H. H. Hopkins, "The aberration permissible in optical systems," *Proc. Phys. Soc.* **B70**: 449–470 (1957).
39. H. Bartelt, J. Ojeda-Castañeda, and E. E. Sicre, "Misfocus tolerance seen by simple inspection of the ambiguity function," *Appl. Opt.* **23**: 2693–2696 (1984).

40. W. D. Furlan, G. Saavedra, and J. Lancis, "Phase-space representations as a tool for the evaluation of the polychromatic OTF," *Opt. Comm.* **96**: 208–213 (1993).
41. W. D. Furlan, M. Martínez-Corral, B. Javidi, and G. Saavedra, "Analysis of 3-D integral imaging display using the Wigner distribution," *J. Disp. Technol.* **2**: 180–185 (2006).
42. K. H. Brenner , A. W. Lohmann, and J. Ojeda-Castañeda, "The ambiguity function as a polar display of the OTF," *Opt. Comm.* **44**: 323 (1983).
43. W. D. Furlan, G. Saavedra, and J. Lancis, "Phase-space representations as a tool for the evaluation of the polychromatic OTF," *Opt. Comm.* **96**: 208–213 (1993).
44. J. Bescós and J. Santamaría, "Formation of color images: Optical transfer functions for the tristimulus values," *Photogr. Sci. and Eng.* **21**: 355–362 (1977).
45. J. Bescós, J. H. Altamirano, A. Santisteban, and J. Santamaría, "Digital restoration models for color imaging," *Appl. Opt.* **27**: 419–424 (1988).
46. R. Barnden, "Calculation of axial polychromatic optical transfer function," *Optica Acta* **21**: 981–1003 (1974).
47. R. Barnden, "Extra-axial polychromatic optical transfer function," *Optica Acta* **23**: 1–24 (1976).
48. M. Takeda, "Chromatic aberration matching of the polychromatic optical transfer function," *Appl. Opt.* **20**: 684–687 (1981).
49. G. Wyszecki and W. S. Stiles, *Color Science*, Wiley, New York, 1982.
50. W. D. Furlan, G. Saavedra, E. Silvestre, M. J. Yzuel, and P. Andrés, "Polychromatic merit functions in terms of the Wigner distribution function," *Proc. SPIE* **2730**: 252–255 (1996).
51. W. D. Furlan, G. Saavedra, E. Silvestre, P. Andrés, and M. J. Yzuel, "Polychromatic axial behavior of aberrated optical systems: Wigner distribution function approach," *Appl. Opt.* **36**: 9146–9151 (1997).
52. P. Andrés, J. Lancis, E. E. Sicre, and E. Bonet, "Achromatic Fresnel diffraction patterns," *Opt. Comm.* **104**: 39–45 (1993).
53. P. Andrés, J. Lancis, E. Tajahuerce, V. Climent, and G. Saavedra, "White-light optical information processing with achromatic processors," *1994 OSA Tech. Digest Series* **11**: 220–223 (1994).
54. J. Lancis, E. E. Sicre, E. Tajahuerce, and P. Andrés, "White-light implementation of the Wigner-distribution function with an achromatic processor," *Appl. Opt.* **34**: 8209–8212 (1995).
55. W. Furlan, D. Zalvidea, and G. Saavedra, "Synthesis of filters for specified axial irradiance by use of phase-space tomography," *Opt. Comm.* **189**: 15–19 (2001).
56. A. Vander Lugt, *Optical Signal Processing*, Wiley, New York, 1992.
57. J. W. Goodman, *Introduction to Fourier Optics*, McGraw Hill, New York, 1996.
58. D. Mendlovic, H. M. Ozaktas, and A. W. Lohmann, "Fractional correlation," *Appl. Opt.* **34**: 303–309 (1995).
59. S. Granieri, R. Arizaga, and E. E. Sicre, "Optical correlation based on the fractional Fourier transform," *Appl. Opt.* **36**: 6636–6645 (1997).
60. S. Granieri, M. Tebaldi, and W. D. Furlan, "Parallel fractional correlation: An optical implementation," *Appl. Opt.* **40**: 6439–6444 (2001).

CHAPTER 5

Imaging Systems: Phase-Space Representations

Jorge Ojeda-Castañeda

DICIS, University of Guanajuato, Salamanca, México

5.1 Introduction

For designing optical imaging systems, one is often faced with an inevitable tradeoff between two figures of merit, which are expressed as two Fourier conjugate variables. For example, for image acquisition there is a tradeoff between the size of the pupil aperture and the depth of field of the optical system. Phase-space representations may be useful for suggesting novel solutions to these types of tradeoffs.

The purpose of this chapter is to put the reader in touch with the use of phase-space representation for analyzing and designing novel imaging systems. To that end, we selected a group of imaging devices that highlight key features on the use of phase-space representations. We illustrate the use of the Wigner distribution function[1] (WDF) and the ambiguity function[2] (AF) by considering equivalent space-invariant, coherent optical processors, with unit magnification, in a similar fashion to the approach discussed in Refs. 3 and 4 for other applications. Our examples are invitations rather than endpoints. For the sake of clarity, our discussions restrict mainly to one-dimensional optical systems.

5.2 The Product-Space Representation and Product Spectrum Representation

In early studies on the use of coherent optical systems, Leith and his colleagues[5–7] recognized that several one-dimensional signals can be processed in parallel by using suitable two-dimensional masks. As is depicted in Fig. 5.1, one can implement optically a spectrum analyzer of several one-dimensional signals, if one is able to generate a suitable two-dimensional mask. Here, we consider the following two-dimensional, complex amplitude transmittance

$$p(x, y) = u\left(x + \frac{y}{2}\right) u^*\left(x - \frac{y}{2}\right) \tag{5.1}$$

In this chapter, the complex amplitude transmittance in Eq. (5.1) is referred to (see Wood and Barry[8]) as the *product-space representation* of the signal $u(x)$. For a rectangular window, $u(x) = \text{rect}\,(x/X)$, the *product-space representation* is

$$\begin{aligned} p(x, y) &= \text{rect}\left(\frac{x + \frac{y}{2}}{X}\right) \text{rect}\left(\frac{x - \frac{y}{2}}{X}\right) \\ &= \text{rect}\left(\frac{x}{X - |y|}\right) \text{rect}\left(\frac{\frac{y}{2}}{X}\right) \end{aligned} \tag{5.2}$$

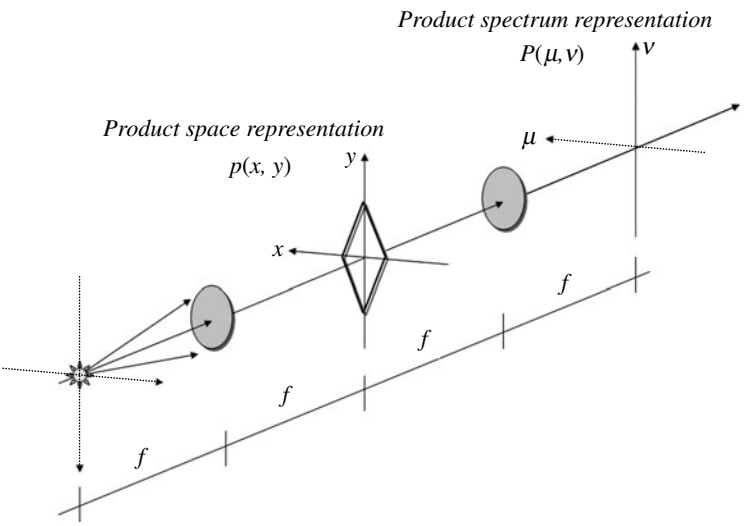


FIGURE 5.1 Optical setup for mapping the product space representation into the product spectrum representation.

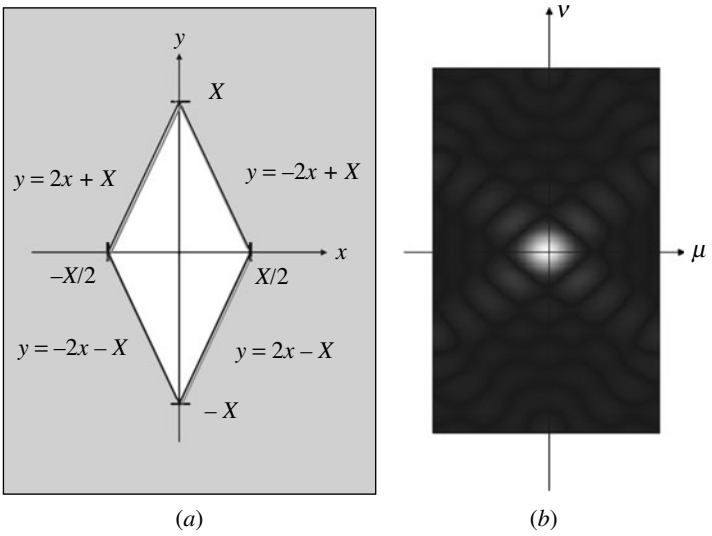


FIGURE 5.2 Product space representation and product spectrum representation of a rectangular window.

The result in Eq. (5.2) is a binary screen with transparent rhomboid that is depicted in Fig. 5.2*a*. We note that for phase-space representations, the above-mentioned rhomboid describes the support of any signal that is space-bound. We denote the Fourier spectrum of an optical signal $u(x)$ as $U(\mu)$,

$$U(\mu) = \int_{-\infty}^{\infty} u(x) \exp(-i2\pi\mu x)\, dx \tag{5.3}$$

Thus, the two-dimensional complex amplitude distribution at the Fraunhofer diffraction plane of the product-space representation, $p(x, y)$ is

$$\begin{aligned} P(\mu, \nu) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x, y) \exp[-i2\pi(\mu x + \nu y)]\, dx\, dy \\ &= U\left(\nu + \frac{\mu}{2}\right) U^*\left(\nu - \frac{\mu}{2}\right) \end{aligned} \tag{5.4}$$

In other words, if the mask in Eq. (5.1) is placed at the input of an optical spectrum analyzer (as in Fig. 5.1), the output is the product spectrum representation $P(\mu, \nu)$, as defined in Eq. (5.4). In Fig. 5.2*b* we display $|P(\mu, \nu)|$ for the example in Eq. (5.2).

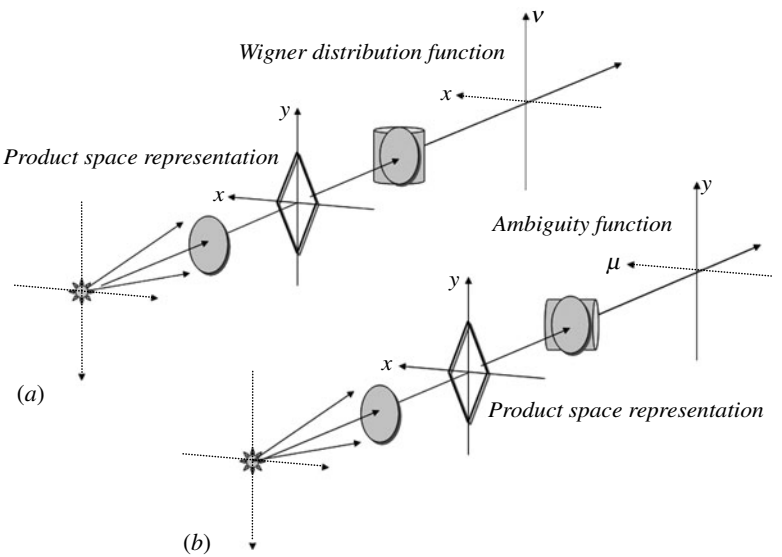


FIGURE 5.3 Anamorphic processors for visualizing: (a) the Wigner distribution function, (b) the ambiguity function.

Apparently, Ville[9] was the first person to explore mathematically two possible variations of the result in Eq. (5.4). His explorations are rephrased here in terms of optical anamorphic processors.

In Fig. 5.3*a* we depict an anamorphic processor that is obtained by adding (to the spectrum analyzer in Fig. 5.1) a cylindrical lens with the same focal length as the spherical lens. Due to the presence of the cylindrical lens, now the anamorphic processor images the complex amplitude along the horizontal axis, while it implements a Fourier transform along the vertical axis. In mathematical terms, the anamorphic processor is able to generate the WDF, $W(x, \nu)$, by implementing over the product-space representation the two-dimensional operation

$$W(x, \nu) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x_0, y)\delta(x - x_0) \exp(-i2\pi\nu y)\, dx_0\, dy$$

$$= \int_{-\infty}^{\infty} p(x, y) \exp(-i2\pi\nu y)\, dy \tag{5.5}$$

Next, we analyze the anamorphic processor depicted in Fig. 5.3b. Now, the anamorphic processor images the complex amplitude along the vertical axis, while it implements a Fourier transform along the

horizontal axis. In mathematical terms, the second anamorphic processor generates the AF, $A(\mu, y)$, by implementing over the product-space representation the two-dimensional operation

$$A(\mu, y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x, y_0)\delta(y - y_0)\exp(-i2\pi\mu x)\,dx\,dy_0$$
$$= \int_{-\infty}^{\infty} p(x, y)\exp(-i2\pi\mu x)\,dx \tag{5.6}$$

It is straightforward to extend the above results to similar cases. For example, if we add a spherical lens to the anamorphic processors in Fig. 5.3*b*, we find that the complex amplitude distribution of the Fraunhofer diffraction pattern of $A(\mu, y)$ is $W(x, \nu)$. That is,

$$W(x, \nu) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} A(\mu, y)\exp[i2\pi(x\mu - \nu y)]\,d\mu\,dy \tag{5.7}$$

The above results are summarized pictorially in Fig. 5.4. This type of diagram was introduced, by Brenner and Ojeda-Castaneda,[10] as a

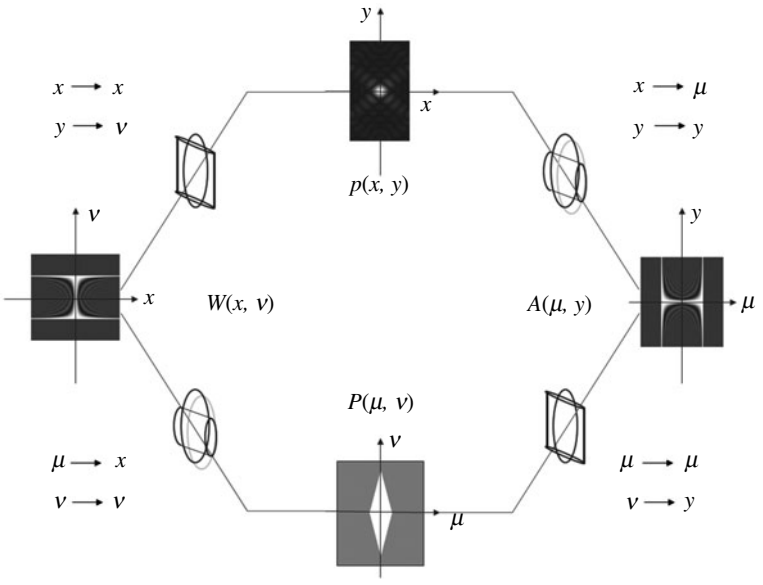


FIGURE 5.4 Phase-space representation: a roadmap for coherent illumination.

road map for visualizing the basic integral transformations that define phase-space representations. See Chaps. 1 and 2 of this book for the conceptual developments sketched in Fig. 5.4.

5.3 Optical Imaging Systems

Optical imaging devices are commonly analyzed using linear system theory. Under coherent illumination, a space-invariant optical system maps linearly the complex amplitude distribution at the input $u_0(x)$ into the complex amplitude distribution at the output $u(x)$. If the input is a pinhole-size source, then the complex amplitude distribution at the output is the coherent point-spread function, or the coherent impulse response, $q(x)$. Hence, if one has a space-invariant optical system, then the imaging process is represented by the convolution integral

$$u(x) = \int_{-\infty}^{\infty} q(x - x_0) u_0(x_0)\, dx_0 \tag{5.8}$$

In what follows we explore the use of the WDF for linking ray optics with wave optics. To that end, we rephrase the linear mapping, in Eq. (5.8), in terms of the WDF of the input, $W_0(x, \nu)$, and the WDF of the impulse response, $W_q(x, \nu)$. Then we rewrite Eq. (5.8) as

$$W(x, \nu) = \int_{-\infty}^{\infty} W_0(x_0, \nu) W_q(x - x_0, \nu)\, dx_0 \tag{5.9}$$

Now, if the input is a pinhole-size source $u_0(x) = \delta(x)$, then in the paraxial regime the WDF of the input represents a bundle of rays with the same amplitude, for any possible angle $\theta = \lambda\nu$, that is, $W_0(x, \nu) = \delta(x)$. Then, according to Eq. (5.9), the output WDF is, $W_q(x, \nu)$. By adding the amplitudes of any possible ray, associated to the output WDF, we obtain the output irradiance distribution

$$I(x) = \int_{-\infty}^{\infty} W_q(x, \nu)\, d\nu \tag{5.10}$$

We illustrate the use of Eq. (5.10) by considering the WDF of a clear pupil aperture with cutoff spatial frequency Ω

$$W_q(x, \nu) = 4(\Omega - |\nu|)\, \mathrm{sinc}[4(\Omega - |\nu|)x]\, \mathrm{rect}\left(\frac{\nu}{2\Omega}\right) \tag{5.11}$$

We note that several interesting features are apparent from Eq. (5.11). First, independently of the value of x, the rays coming from the edge of the pupil (marginal rays) can be considered as having zero amplitude, $W_q(x, \Omega) = 0$. Second, in a lax manner, one can consider that all the rays coming to the optical axis ($x = 0$) add constructively. However, the amplitude of the rays decreases as $4(\Omega - |\nu|)$. Third, in a lax manner, one can consider that outside the optical axis some of the rays add destructively. Specifically, at the point $x = 1/2\Omega$, the amplitude of the rays varies as $(2\Omega/\pi)\sin[2\pi(1 - |\nu/\Omega|)]$.

According to Ref. 11, under the influence of wave aberrations, the WDF of the diffraction-limited system $W_0(x, \nu)$ changes as follows:

$$W(x, \nu) = \exp\left[-\sum_{m=1}^{[M/2]} C_m(\nu)\left(\frac{\partial^{2m+1}}{\partial x^{2m+1}}\right)\right] W_0\left(x - f\left(\frac{d\psi}{d\nu}\right), \nu\right) \tag{5.12}$$

In Eq. (5.12) we employ the letter f for denoting the focal length of the optical processor; $\psi(\nu)$ denotes the wave aberration polynomial; $[M/2] = M/2 - 1$ if M is an even integer number; and $[M/2] = (M-1)/2$ if M is an odd integer number. In the differential operator, the coefficients are

$$C_m(\nu) = \frac{f^{2m+1}}{\left(\frac{4\pi}{\lambda}\right)^{2m}(2m+1)!}\frac{d^{2m+1}\psi}{d\nu^{2m+1}} \tag{5.13}$$

It is apparent from Eq. (5.12) that the WDF $W_0(x, \nu)$ suffers from lateral displacements, which are predicted by geometrical optics. In addition, the WDF modifies its amplitude distribution, as predicted by wave optics.[11]

Next we discuss the use of an equivalent optical processor, as depicted in Fig. 5.5, for implementing filtering operations in the phase-space representation. To that end, we represent the equivalent optical image processor, by the superposition integral

$$W(x, \nu) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} W_0(x_0, \nu_0)\, W_b(x, \nu; x_0, \nu_0)\, dx_0\, d\nu_0 \tag{5.14}$$

The impulse response $W_b(x, \nu; x_0, \nu_0)$ is the Bastiaans[12] double WDF. By simple comparison between Eqs. (5.9) and (5.14), one finds that for a space-invariant imaging system

$$W_b(x, \nu; x_0, \nu_0) = W_q(x - x_0, \nu_0)\delta(\nu - \nu_0) \tag{5.15}$$

Next, at the pupil aperture of the equivalent optical processor, we note that the complex amplitude distribution is the AF $A_0(\mu, y)$.

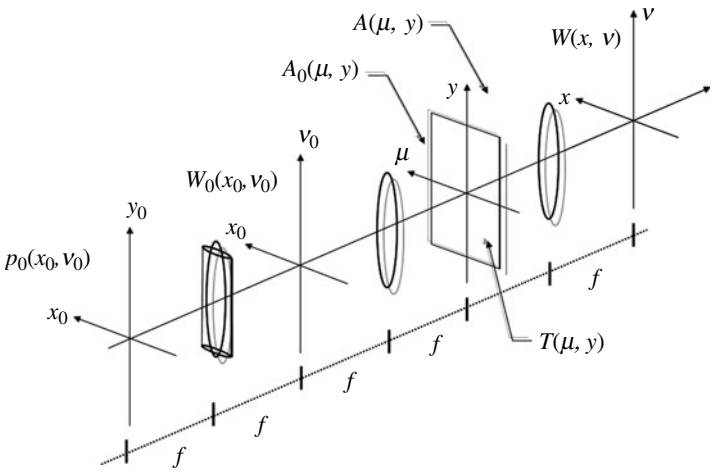


FIGURE 5.5 Optical setup for processing the Wigner distribution function.

If the complex amplitude transmittance of the pupil mask is $T(\mu, y)$, then just after the mask the complex amplitude distribution is the AF $A(\mu, y) = T(\mu, y)A_0(\mu, y)$. The above result can also be rephrased as follows. If one uses a pupil mask with complex amplitude transmittance $T(\mu, y)$, then one generates the impulse response $t(x, \nu)$, which implements the mapping

$$W(x, \nu) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} W_0(x_0, \nu_0) t(x - x_0, \nu - \nu_0)\, dx_0\, d\nu_0 \tag{5.16}$$

By comparison of Eqs. (5.14) and (5.16) we know that the impulse response of the equivalent processor is $W_b(x, \nu; x_0, \nu_0)$. For an ideal system, the impulse response is $W_b(x, \nu; x_0, \nu_0) = \delta(x - x_0)\delta(\nu - \nu_0)$. This WDF cannot be obtained from a product-space representation. Hence, we comment on the two following approaches.

On the one hand, it is possible to use masks that can be expressed in terms of the *product spectrum representation*, say, $T(\mu, y) = P(y/\lambda f, \mu)$. And consequently, the impulse response is expressed in terms of the product-space representation. And in this manner, the equivalent optical processor effectively implements Eq. (5.9).

On the other hand, one can use masks that cannot be expressed in terms of the product spectrum representation for implementing optically nonconventional transformations in phase-space. For example, if at the Fraunhofer plane of the equivalent processor one sets

$T(\mu, y) = \delta(\mu)$, then at the output plane one obtains the WDF

$$W(x, \nu) = \int_{-\infty}^{\infty} W_0(x_0, \nu)\, dx_0 = |U_0(\nu)|^2 \tag{5.17}$$

And consequently, since the signal $u(x)$ can be recovered from its WDF,

$$u(x)u^*(0) = p\left(\frac{x}{2}, x\right) = \int_{-\infty}^{\infty} W\left(\frac{x}{2}, \nu\right) \exp(i2\pi x\nu)\, d\nu \tag{5.18}$$

we can rewrite Eq. (5.17) as the nonlinear mapping

$$u(x)u^*(0) = \int_{-\infty}^{\infty} u_0\left(x_0 + \frac{x}{2}\right) u_0^*\left(x_0 - \frac{x}{2}\right)\, dx_0 \tag{5.19}$$

From Eq. (5.19) it is apparent that the equivalent optical processor is useful for visualizing a correlation operation. Of course a similar result is obtained by setting $T(\mu, y) = \delta(y)$. For this latter example, at the output plane, the WDF is

$$W(x, \nu) = \int_{-\infty}^{\infty} W_0(x, \nu_0)\, d\nu_0 = |u_0(x)|^2 \tag{5.20}$$

And now Eq. (5.19) becomes

$$u(x)u^*(0) = |u_0(0)|^2 \delta(x) \tag{5.21}$$

The above results remind us that phase-space representations have an inherent nonlinear nature, caused by using as input either the product-space representation or the product spectrum representation. The nonlinear attribute is linked next to the Saleh bilinear transformations.[13]

5.4 Bilinear Optical Systems

A bilinear transformation relates the complex amplitude distribution at the input to the irradiance distribution at the output. Hence, in terms of the third term of a Volterra series expansion, a bilinear transformation is defined as

$$|u(x)|^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} V(x; x_1, x_2) u_0(x_1) u_0^*(x_2)\, dx_1\, dx_2 \tag{5.22}$$

In Eq. (5.22) the kernel of the Volterra transformation, $V(x; x_1, x_2)$, is the bilinear impulse response.[13] The concept of bilinearity has found practical applications in some optical problems.[14] In what follows we show that bilinearity, in the above sense, is related to a phase-space transformation.[15] To our end, we use the following center and difference coordinates:

$$\hat{y} = \frac{x_1 + x_2}{2}, \qquad y = x_1 - x_2, \qquad B(x; \hat{y}, y) = V\left(x, \hat{y} + \frac{y}{2}, \hat{y} - \frac{y}{2}\right) \tag{5.23}$$

By employing Eq. (5.23), the definition of the product-space representation, and the definition of the WDF, we rewrite Eq. (5.22) as

$$\begin{aligned} |u(x)|^2 &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} B(x; \hat{y}, y) u_0\left(\hat{y} + \frac{y}{2}\right) u_0^*\left(\hat{y} - \frac{y}{2}\right) d\hat{y}\, dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} B(x; \hat{y}, y)\, p_0(\hat{y}, y)\, d\hat{y}\, dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} B(x; \hat{y}, y) \exp(i2\pi\hat{\nu}y)\, dy\right] W_0(\hat{y}, \hat{\nu}) d\hat{y}\, d\hat{\nu} \end{aligned} \tag{5.24}$$

Equation (5.24) tells one how to tailor the output irradiance distribution if one takes as input either the product-space representation, $p_0(x, y)$ or the WDF $W_0(\hat{y}, \hat{\nu})$. Here it is convenient to recall the result in Eq. (5.14).

$$\begin{aligned} |u(x)|^2 &= \int_{-\infty}^{\infty} W(x, \nu)\, d\nu \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} W_0(\hat{y}, \hat{\nu}) W_b(x, \nu; \hat{y}, \hat{\nu}) d\hat{y}\, d\hat{\nu}\, d\nu \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} W_b(x, \nu; x_0, \nu_0)\, d\nu\right] W_0(\hat{y}, \hat{\nu})\, d\hat{y}\, d\hat{\nu} \end{aligned} \tag{5.25}$$

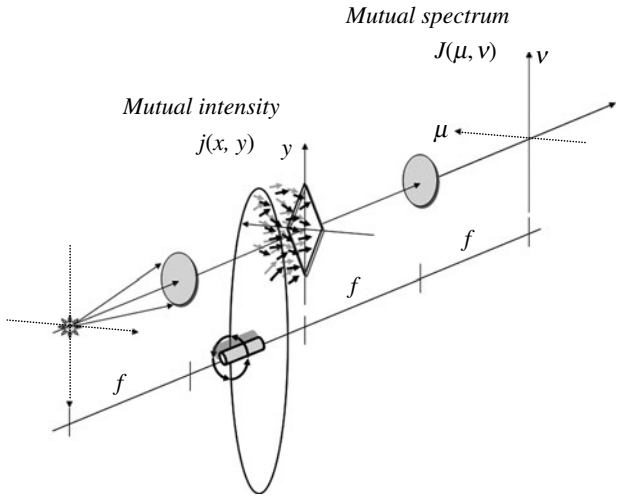


FIGURE 5.6 Same as Fig. 5.1, but with a rotating scatter plate for reducing the degree of spatial coherence.

Comparing Eqs. (5.24) and (5.25) and using Eq. (5.15), we obtain

$$\begin{aligned} B(x; \hat{y}, y) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} W_b(x, \nu; \hat{y}, \hat{\nu}) \exp(-i2\pi y\hat{\nu})\, d\hat{\nu}\, d\nu \\ &= \int_{-\infty}^{\infty} W_q(x - \hat{y}, \nu) \exp[i2\pi(-y)\nu]\, d\nu \\ &= q\left[x - \left(\hat{y} + \frac{y}{2}\right)\right] q^*\left[x - \left(\hat{y} - \frac{y}{2}\right)\right] \\ &= p_q(x - \hat{y}, -y) \end{aligned} \tag{5.26}$$

And therefore, in suitable coordinates, the Volterra kernel is related to the double WDF. Equivalently, one can say that phase-space representations are bilinear transformations. For space-invariant systems, the Volterra kernel is the product-space representation of the impulse response. Here it is relevant to note that for optical systems working with partially coherent illumination, one needs to substitute the product-space representation for the mutual coherence function (the mutual intensity for monochromatic illumination), as depicted in Fig. 5.6. Equivalently, the product spectrum representation should be substituted by the cross-spectral density (the mutual power spectrum or the mutual spectrum for monochromatic illumination). Thus, the road map changes as depicted in Fig. 5.7.

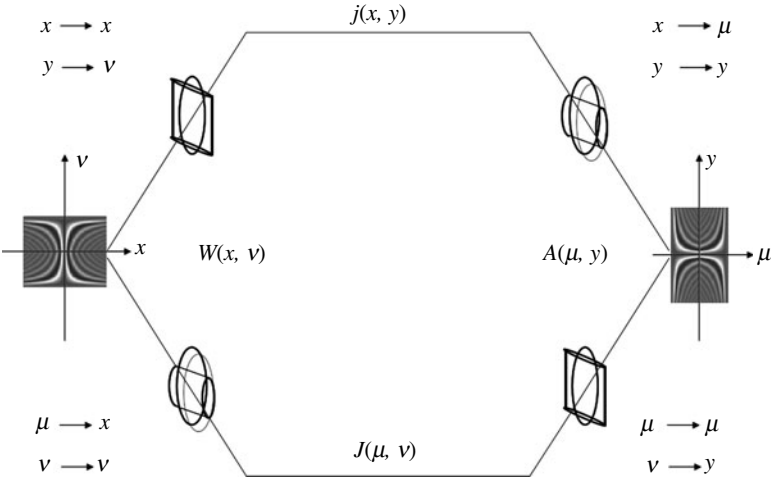


FIGURE 5.7 Same as Fig. 5.4, but for the mutual intensity and the mutual spectrum.

5.5 Noncoherent Imaging Systems

Under noncoherent illumination, a space-invariant optical system is represented by the noncoherent impulse response $|q(x)|^2 = h(x)$. Except for a normalization factor, the Fourier transform of $h(x)$ is denoted as the optical transfer function (OTF), $H(\mu)$. The modulus of the OTF, $|H(\mu)|$, is known as the *modulation transfer function* (MTF). Next, we analyze the impact of focus errors on the MTF of an optical processor working under noncoherent illumination. Hence, the generalized pupil function of the optical system is

$$Q(\mu; W_{2,0}) = T(\mu) \exp\left[i2\pi\left(\frac{W_{2,0}}{\lambda}\right)\left(\frac{\mu}{\Omega}\right)^2\right] \tag{5.27}$$

In Eq. (5.27), $T(\mu)$ is the complex amplitude transmittance of a mask located over the pupil aperture. We denote as $W_{2,0}$ the wavefront aberration coefficient for describing focus error.[16–18] Consequently, except for a normalization factor, the out-of-focus MTF is

$$|H(\mu, W_{2,0})| = \left|\int_{-\infty}^{\infty} T\left(\nu + \frac{\mu}{2}\right) T^*\left(\nu - \frac{\mu}{2}\right) \exp\left[i2\pi\left(\frac{2W_{2,0}\mu}{\lambda\Omega^2}\right)\nu\right] d\nu\right| \tag{5.28}$$

The result in Eq. (5.28) can be readily connected with the ambiguity function of $T(\mu)$, as follows. We note that the AF of the pupil mask is

$$A_T(\mu, y) = \int_{-\infty}^{\infty} T\left(\nu + \frac{\mu}{2}\right) T^*\left(\nu - \frac{\mu}{2}\right) \exp(i2\pi y\nu)\, d\nu \tag{5.29}$$

If we evaluate Eq. (5.29) at $y = [2\mu/(\lambda\Omega^2)]W_{2,0}$, then

$$\left| A_T\left(\mu, \frac{2W_{2,0}\mu}{\lambda\Omega^2}\right)\right| = |H(\mu, W_{2,0})| \tag{5.30}$$

In other words, if one evaluates the modulus of the AF, $|A_T(\mu, y)|$, along the straight line $y = m\mu$, then one obtains the MTF $|H(\mu, W_{2,0})|$ with focus error coefficient $W_{2,0} = (\lambda/2)m\Omega^2$. Hence, we recognize that $|A_T(\mu, y)|$ contains all the possible MTFs $|H(\mu; W_{2,0})|$ for variable focus error $W_{2,0}$. See Ref. 19.

Next we note that the simple result in Eq. (5.30) has the two following applications. First, one can visualize (in a single picture) the impact of variable focus error on the MTF. Second, one can seek pupil masks that generated rotationally symmetric AF, for reducing the influence of focus error on the MTF. And in this manner, one can extend the depth of field of an optical system.

In Fig. 5.8*a*, we display the modulus of the ambiguity function of a clear pupil aperture

$$|A_T(\mu, y)| = [2\Omega - |\mu|]\,\mathrm{rect}\left(\frac{\mu}{4\Omega}\right) |\,\mathrm{sinc}\,[(2\Omega - |\mu|)y]| \tag{5.31}$$

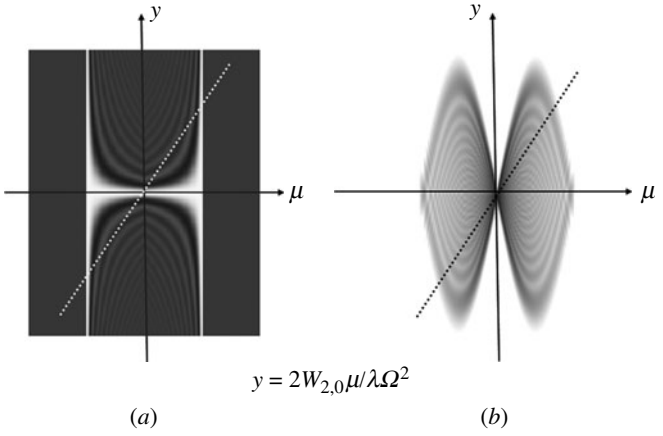


FIGURE 5.8 Modulus of the ambiguity function for: (a) a rectangular pupil aperture, (b) a phase-only mask with a phase function that has odd symmetry.

The zero-loci curves of Eq. (5.31) obey the relationship

$$n\left(\frac{\lambda}{2}\right) = \left(2 - \frac{\mu}{\Omega}\right)\left(\frac{\mu}{\Omega}\right) W_{2,0} \tag{5.32}$$

where $n = 1, 2, 3, 4, \ldots$. The zero-loci curves are useful for identifying the points where the values of $W_{2,0}$ and μ have a MTF equal to zero, and in this manner for setting tolerance values for the focus error, in terms of $W_{2,0}$. See Refs. 20 and 21. In Ref. 22, this approach is extended to polychromatic illumination.

5.6 Tolerance to Focus Errors and to Spherical Aberration

It is a misconception to assume that, under noncoherent illumination, the phase distribution of the generalized pupil function does not influence the quality of an image. If the wave leaving the exit pupil has departures from sphericity, the image quality is deteriorated. Hence, one can expect that by modifying the phase of the generalized pupil function, one can reduce the impact of aberrations on the MTF.

In other words, heuristically, one expects that by preventing an image from becoming badly degraded (due to the presence of aberrations) its digitally restored version might have higher quality than the restored picture of the nonpreventive image.[23–25] This approach has found applications for extending the depth of field, of an optical system, by using a suitable phase mask that preserves both lateral resolution and light-gathering power.

There are several phase masks that are able to generate a MTF with low sensitivity to focus error.[26–37] A suitable phase mask generates a MTF that, inside its passband, does not have zero values for certain amounts of focus error. However, inside its passband, the generated MTF has reduced visibility.

Since one simultaneously records the images of planar scenes located at different depths of the object field, by using a suitable phase mask we ensure that each recorded image will suffer from virtually the same amount of contrast reduction. For this reason, later on, by digital processing, the image contrast can be simultaneously corrected for all the recorded images. Next, we discuss a simple model for describing this approach.

In Sec. 5.5 we indicate that for any pupil mask, its AF conveniently contains all the defocused MTFs. Hence, an MTF with reduced sensitivity to focus errors must be visualized as an AF with rotational symmetry. If you will, the AF exhibits a "bow tie" effect, as depicted in Fig. 5.8*b*. This type of AF was obtained early by using a parabolic FM, in radar engineering.[38] However, for extending the depth of field, the

phase mask that produces this type of AF was discovered by Dowski and Cathey.[39]

Now, we follow the treatment in Ref. 40. The complex amplitude transmittance of the pupil aperture is

$$Q(\mu) = T(\mu) \exp\left[i2\pi \left(\frac{W_{2,0}}{\lambda} \right) \left(\frac{\mu}{\Omega} \right)^2 \right] \tag{5.33}$$

In Eq. (5.33) we denote as $T(\mu)$ the complex amplitude transmittance of the mask located at the pupil aperture. Now, as discussed previously except for a normalization factor, the OTF of $T(\mu)$ with variable focus error is

$$H(\mu; W_{2,0}) = \int_{-\infty}^{\infty} T\left(\nu + \frac{\mu}{2} \right) T^*\left(\nu - \frac{\mu}{2} \right) \exp\left[i2\pi \left(2W_{2,0} \frac{\mu}{\lambda\Omega^2} \right) \nu \right] d\nu \tag{5.34}$$

Next, we note that if $T(\mu)$ is a continuous function in μ, then the OTF $H(\mu; W_{2,0})$ is also continuous in both μ and $W_{2,0}$. Consequently, we can express Eq. (5.34) as a Maclaurin power series expansion. That is,

$$H(\mu; W_{2,0}) = H(\mu; 0) + W_{2,0} \left(\frac{\partial H}{\partial W_{2,0}} \right) + \left(\frac{W_{2,0}^2}{2!} \right) \left(\frac{\partial^2 H}{\partial W_{2,0}^2} \right) + \cdots \tag{5.35}$$

In the above power series, the nth coefficient

$$\begin{aligned} \partial'' H / \partial W_{2,0}^n &= \left(\frac{i2\pi\mu}{\lambda\Omega^2} \right)^n \int_{-\infty}^{\infty} \nu^n T\left(\nu + \frac{\mu}{2} \right) T^*\left(\nu - \frac{\mu}{2} \right) d\nu \\ &= \left(\frac{i2\pi\mu}{\lambda\Omega^2} \right)^n \int_{-\infty}^{\infty} \nu^n P_T(\mu, \nu)\, d\nu \end{aligned} \tag{5.36}$$

In Eq. (5.36), $P_T(\mu, \nu)$ is the product spectrum representation of the mask $T(\mu)$. Here, it is relevant to recognize the following. If the complex amplitude transmittance of the pupil mask is a Hermitian function $T(\mu) = T^*(-\mu)$, then the product spectrum representation $P_T(\mu, \nu)$ is an even function in the integrating variable ν. That is,

$$P_T(\mu, -\nu) = P_T(\mu, \nu) = T\left(\nu + \frac{\mu}{2} \right) T^*\left(\nu - \frac{\mu}{2} \right) \tag{5.37}$$

Hence, the integrand in Eq. (5.36) is an odd function, provided that the power order n is an odd integer number ($n = 2s + 1$). And consequently, the odd-order coefficients are equal to zero for Hermitian

pupil masks. If this symmetry condition is fulfilled, then the Maclaurin series expansion becomes

$$H(\mu; W_{2,0}) = H(\mu; 0) + \sum_{n=1}^{\infty} \frac{W_{2,0}^{2n}}{2n!} \left(\frac{\partial^{2n} H}{\partial W_{2,0}^{2n}} \right) \tag{5.38}$$

The result in Eq. (5.38) has two powerful consequences. First, the OTF exhibits a symmetrical behavior before and after the in-focus image $W_{2,0} = 0$. Second, the optical system can have a large tolerance to focus error $W_{2,0}$, provided that the second-order coefficient has small values.

We note that if the pupil mask is a phase-only filter, then the condition for Hermitan symmetry implies that the phase profile must exhibit odd symmetry. That is, if $\phi(\mu) = -\phi(-\mu)$, then

$$T^*(-\mu) = \exp[-i\phi(-\mu)] = T(\mu) \tag{5.39}$$

In Fig. 5.9, we display numerically simulated images of a spoke pattern. Except for the clear pupil aperture, along the columns of Fig. 5.9, the phase masks obey the relationship

$$\exp[i\phi(\mu)] = \exp\left[i \operatorname{sign}(\mu) \Theta \left| \frac{\mu}{\Omega} \right|^n \right] \tag{5.40}$$

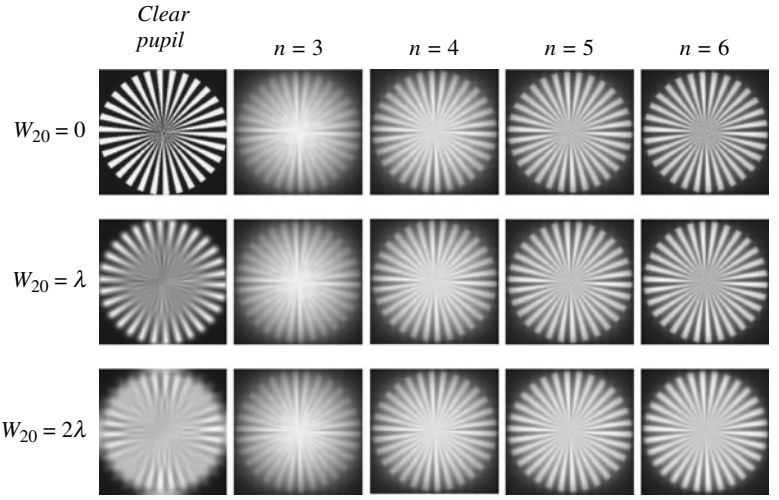


FIGURE 5.9 Numerical simulations of in-focus and out-of-focus images of a spoke pattern, when using a clear pupil aperture, and a mask with phase variations that exhibits odd symmetry.

In Eq. (5.40) we denote as sign(μ) the signum function. We use Θ for denoting the maximum phase delay, at the edge of the pupil aperture. And n is any integer number. However, we note that n can also represent a real number. For further discussion of this topic, see Refs. 30, 31, and 40 to 43.

Next, we discuss how these results are also useful for reducing the impact of spherical aberration in two-dimensional optical systems with radial symmetry. To that end, we discuss a simplified version of McCutchen theorem[44] that is useful for analyzing the Strehl ratio of rotationally symmetric systems, which suffer from wave aberrations.[45–50]

The generalized pupil function of an optical system, with rotationally symmetry, that suffers from focus error $W_{2,0}$ and from spherical aberration $W_{4,0}$ is

$$S(\rho; W_{2,0}; W_{4,0}) = T(\rho) \exp\left\{ i2\pi \left[\left(\frac{W_{2,0}}{\lambda} \right) \left(\frac{\rho}{\Omega} \right)^2 + \left(\frac{W_{4,0}}{\lambda} \right) \left(\frac{\rho}{\Omega} \right)^4 \right] \right\} \tag{5.41}$$

In Eq. (5.41) the complex amplitude transmittance of the pupil mask is $T(\rho)$. We denote as ρ the radial spatial frequency, and its maximum value is the cutoff spatial frequency Ω. In polar coordinates, the impulse response of the optical system is obtained by taking the two-dimensional Fourier transform of Eq. (5.41). That is,

$$s(r; W_{2,0}; W_{4,0}) = 2\pi \int_0^{\Omega} S(\rho; W_{2,0}; W_{4,0}) J_0(2\pi r \rho) \rho \, d\rho \tag{5.42}$$

At the optical axis $r = 0$, Eq. (5.42) reduces to

$$s(0; W_{2,0}; W_{4,0}) = 2\pi \int_0^{\Omega} S(\rho; W_{2,0}; W_{4,0}) \rho \, d\rho \tag{5.43}$$

Now, we map the radially symmetric, two-dimensional pupil into a one-dimensional pupil aperture, by using the geometrical transformation

$$\zeta = \left(\frac{\rho}{\Omega} \right)^2 - 0.5; \qquad Q(\zeta) = T(\Omega(\zeta + 0.5)^{1/2}) \tag{5.44}$$

From Eq. (5.44), we observe that the geometrical transformation defines an effective pupil function $Q(\zeta)$ from the physical pupil mask $T(\rho)$. And therefore, by using Eq. (5.43) at the optical axis, we can define an effective impulse response as the square modulus of

$s(0; W_{2,0}; W_{4,0})$. That is,

$$h(W_{2,0}; W_{4,0}) = |s(0; W_{2,0}; W_{4,0})|^2 = (\pi\Omega)^2 \left| \int_{-0,5}^{0.5} Q(\zeta) \exp\left\{ i2\pi \left[\left(\frac{W_{4,0}}{\lambda} \right) \zeta^2 + (W_{4,0} + W_{2,0})\zeta \right] \right\} d\zeta \right|^2 \tag{5.45}$$

Of course, the effective impulse response, in Eq. (5.45), can be related to an effective OTF

$$h(W_{2,0}; W_{4,0}) = \int_{-\infty}^{\infty} H(\mu; W_{4,0}) \exp\left\{ i2\pi \left[\frac{(W_{4,0} + W_{2,0})}{\lambda} \right] \mu \right\} d\mu \tag{5.46}$$

Except for a normalization factor, the effective OTF is

$$H(\mu; W_{4,0}) = \int_{-\infty}^{\infty} Q\left(\zeta + \frac{\mu}{2}\right) Q^*\left(\zeta - \frac{\mu}{2}\right) \exp\left\{ i2\pi \left[\left(\frac{2W_{4,0}}{\lambda} \right) \mu \right] \zeta \right\} d\zeta = A_q\left(\mu; 2W_{4,0}\frac{\mu}{\lambda}\right) \tag{5.47}$$

Now, in a similar fashion to the discussion in Sec. 5.5, it is apparent from Eq. (5.47) that the effective OTF is related to the AF of the effective pupil function $Q(\zeta)$. Hence, if the effective pupil mask has complex amplitude transmittance with Hermitian symmetry (odd phase distribution for phase-only mask), then the effective OTF has low sensitivity to spherical aberration.

Of course, after the effective pupil function is selected, it is necessary to find the complex amplitude transmittance of the physical mask $T(\rho)$, by using the inverse of the geometrical mapping in Eq. (5.44). The effective OTF is not to be confused with the OTF of the physical mask. We illustrate this result with the following simple example. We consider an effective pupil mask that has a cubic phase profile. That is,

$$Q(\zeta) = \exp(i2\pi\Theta\zeta^3)\,\mathrm{rect}(\zeta) \tag{5.48}$$

The effective pupil function has an odd phase profile. However, as expected, the physical pupil mask has a complex amplitude transmission

that exhibits rotational symmetry. That is,

$$T(\rho) = \exp\left\{ i2\pi\Theta \left[\left(\frac{\rho}{\Omega} \right)^2 - \frac{1}{2} \right]^3 \right\} \operatorname{circ}\left(\frac{\rho}{\Omega} \right) \tag{5.49}$$

If you will, in Eq. (5.49), the physical pupil mask has an annularly distributed phase profile, which has odd phase variation provided that (on the pupil aperture) we take the circle with radius $\rho = \Omega/\sqrt{2}$ as the center of symmetry. The ambiguity function of the effective pupil mask exhibits the bow tie effect. And consequently, the optical system has low sensitivity to spherical aberration; see Ref. 33. Yet, the ambiguity function of the effective pupil is not the OTF of the physical mask.

Related to the previous discussion, we emphasize the following. There is a difference between the use of radially symmetric phase masks (axiconlike elements) for generating large axial irradiance distributions[51–57] and the use of annularly distributed odd phase masks for reducing the impact of aberrations on the MTF, as in Ref. 58. So it may be useful to use the term *high focal depth* for describing axiconlike elements and the term *extended depth of field* for describing optical elements that reduce the impact of focus error on the MTF.

5.7 Phase Conjugate Plates

An optical system with variable focal length is commonly designated as a varifocal, or zoom, system. For tuning the focal length, one changes the longitudinal separation between two quadratic-phase components, which usually have opposite-sign powers. Alvarez[59] and Lohmann[60–62] independently proposed a varifocal system that consists of a pair of cubic phase elements, which are laterally displaced. This type of optical device is also useful for generating, in a tunable fashion, wavefront aberrations.[63,64]

In Fig. 5.10 we depict schematically the use of a pair of phase elements with opposite-sign powers. We assume that the pair is located at the pupil aperture (Fourier domain) of a $4f$ optical processor. The pupil aperture has a rectangular shape. We use the Greek letters α and β to represent the spatial frequencies along the horizontal axis and the vertical axis, respectively, in the pupil aperture. For any phase element, the cutoff spatial frequency is Ω. The complex amplitude transmittance of a single-phase plate is

$$Q(\alpha, \beta) = T(\alpha)\ \operatorname{rect}\left(\frac{\beta}{2\Omega} \right) = \exp\left[i\varphi(\alpha) \right] \operatorname{rect}\left(\frac{\alpha}{2\Omega} \right) \operatorname{rect}\left(\frac{\beta}{2\Omega} \right) \tag{5.50}$$

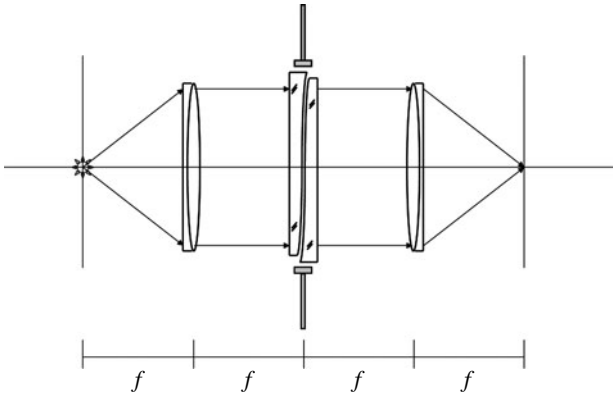


FIGURE 5.10 Optical setup for using a phase conjugate pair, at the Fourier plane.

Now, let us suppose that the pupil mask is formed with a pair of conjugate phase elements, and that these elements are laterally displaced in opposite directions, say, by the spatial frequency $\mu/2$ along the horizontal axis. Then the complex amplitude transmittance of the pupil mask can be expressed in terms of the product spectrum representation $P_T(\mu, \alpha)$ as

$$\begin{aligned} S(\alpha, \beta; \mu) &= T\left(\alpha + \frac{\mu}{2}\right) T^*\left(\alpha - \frac{\mu}{2}\right) \text{rect}\left(\frac{\beta}{2\Omega}\right) \\ &= P_T(\mu, \alpha) \,\text{rect}\left(\frac{\beta}{2\Omega}\right) \end{aligned} \tag{5.51}$$

Equation (5.51) tells us how a generalized pupil function can be synthesized from a pair of conjugate phase elements. To find the corresponding synthesized PSF, we take the two-dimensional inverse Fourier transform of $S(\alpha, \beta; \mu)$. Except for the normalization factor $1/(2\Omega)$, this gives

$$s(x, y; \mu) = \text{sinc}\,(2\Omega y)\left[\int_{-\infty}^{\infty} T\left(\alpha + \frac{\mu}{2}\right) T^*\left(\alpha - \frac{\mu}{2}\right) \exp\,(i2\pi x\alpha)\, d\alpha\right] \tag{5.52}$$

The integral in Eq. (5.52) can be readily recognized as the AF, $A_T(\mu, x)$, of the mask $T(\alpha)$. Hence, the modulus of $s(x, y; \mu)$ can be written as

$$|s(x, y; \mu)| = |\text{sinc}\,(2\Omega y)||A_T(\mu, x)| \tag{5.53}$$

In Sec. 5.4 we note that the ambiguity function of any single phase mask T(μ) contains all the MTFs, $|H(\mu; W_{2,0})|$, for variable focus error $W_{2,0}$. That is,

$$\left| A_T \left(\mu, x = \frac{2W_{2,0}\mu}{(\lambda\Omega)^2} \right) \right| = |H(\mu; W_{2,0})| \tag{5.54}$$

Now, by setting $y = 0$ in Eq. (5.53) and using Eq. (5.54), we get

$$|s(x, 0; \mu)| = |H(\mu; W_{2,0})| \tag{5.55}$$

This is a remarkable result. By using a pair of conjugate phase elements that are laterally displaced with respect to each other, we select a tunable spatial frequency μ at the $4f$ optical system. With the displaced elements as a pupil mask, we generate a PSF whose modulus would display optically the variation of the MTF with focus error of a single-phase element, for the spatial frequency μ that was previously selected.

We use the result in Eq. (5.55) for relating the Alvarez-Lohmann technique to the wavefront coding technique of Dowski and Cathey. For a cubic phase element, the complex amplitude transmittance along the horizontal axis is

$$T(\alpha) = \exp\left[i2\pi\Theta \left(\frac{\alpha}{\Omega} \right)^3 \right] \operatorname{rect}\left(\frac{\alpha}{2\Omega} \right) \tag{5.56}$$

In Eq. (5.56), Θ denotes the maximum optical path difference, which is introduced by the element at the edge of the pupil aperture. If we use a mask that is composed of two laterally displaced cubic phase elements, from Eqs. (5.51) and (5.56) we obtain

$$\begin{aligned} S(\alpha, \beta; \mu) = \exp\left[i\pi \left(\frac{\Theta}{2} \right) \left(\frac{\mu}{\Omega} \right)^3 \right] \exp\left[i2\pi(3\Theta) \left(\frac{\mu}{\Omega} \right)^2 \right] \\ \times \operatorname{rect}\left(\frac{\alpha}{2\Omega - |\mu|} \right) \operatorname{rect}\left(\frac{\beta}{2\Omega} \right) \end{aligned} \tag{5.57}$$

Therefore, except for the normalization factor $1/(2\Omega)$, the modulus of the synthesized PSF is

$$\begin{aligned} &|s(x, y; \mu)| \\ &\quad = |\operatorname{sinc}(2\Omega y)| \\ &\qquad \times \left| \int_{-\infty}^{\infty} \exp\left[i2\pi(3\Theta) \left(\frac{\mu}{\Omega} \right) \left(\frac{\alpha}{\Omega} \right)^2 \right] \operatorname{rect}\left(\frac{\alpha}{2\Omega - |\mu|} \right) \exp(i2\pi x\alpha)\, d\alpha \right| \end{aligned} \tag{5.58}$$

On the other hand, if we use the complex amplitude transmittance of Eq. (5.56) in the case of a single phase element, except for a normalization factor, the corresponding defocused MTF is

$$|H(\mu; W_{2,0})| = \left| \int_{-\infty}^{\infty} \exp\left\{ i2\pi \left[(3\Theta) \left(\frac{\alpha}{\Omega}\right)^2 + 2\left(\frac{W_{2,0}}{\lambda}\right)\left(\frac{\alpha}{\Omega}\right) \right] \left(\frac{\mu}{\Omega}\right) \right\} \times \operatorname{rect}\left(\frac{\alpha}{2\Omega - |\mu|}\right) d\alpha \right| \tag{5.59}$$

We discuss next the relationship between Eqs. (5.58) and (5.59). On one hand, we can set a coherent optical processor that uses as a spatial filter a pair of cubic conjugate phase elements. According to Eq. (5.58), by introducing a displacement between both elements, we generate a quadratic-phase delay within the integral, which is used for evaluating the PSF.

On the other hand, under noncoherent illumination, we can gather images using a single cubic phase element as the spatial filter. According to Eq. (5.59), due to the autocorrelation operation, we also generate a quadratic-phase delay within the integral, which is used for evaluating the MTF.

Hence, in the above two cases, we are able to generate a quadratic-phase delay within a Fourier integral. In this manner, we transform the Fourier integral into a Fresnel integral. Of course, in each case the Fresnel integral appears for a different physical reason. However, it is convenient to exploit this similarity with the purpose of visualizing the defocused MTF of a single-phase element by using a pair of conjugate phase elements. It is worth remarking that the expression in Eq. (5.59), for the AF in terms of a Fresnel integral, was discovered early by radar engineers.[58] More recently, it has been used by Somayaji and Christensen.[65]

Finally, we discuss a method for implementing optically a tunable wavefront coding mask. We assume that the complex amplitude transmittance of a single-phase element is

$$T(\alpha) = \exp\left[i2\pi\Theta \left(\frac{\alpha}{\Omega}\right)^4 \right] \operatorname{rect}\left(\frac{\alpha}{2\Omega}\right) \tag{5.60}$$

We employ Θ again to represent the maximum phase delay, at the edge of the pupil aperture. The two-dimensional version of Eq. (5.60) was presented by Lopez-Gil et al. for generating spherical aberration.[64] Here we consider that at the pupil aperture we have a pair of quadratic-phase elements, which are laterally displaced in opposite directions, say, by $\mu/2$. We also assume that the optical system suffers

from focus error, and then the generalized pupil function is

$$S(\alpha;\mu;W_{2,0}) = \exp\left[i2\pi(4\Theta)\left(\frac{\mu}{\Omega}\right)\left(\frac{\alpha}{\Omega}\right)^3\right]\exp\left[i2\pi\Theta\left(\frac{\mu}{\Omega}\right)^3\left(\frac{\alpha}{\Omega}\right)\right] \times \operatorname{rect}\left(\frac{\alpha}{2\Omega-|\mu|}\right)\exp\left[i2\pi\left(\frac{W_{2,0}}{\lambda}\right)\left(\frac{\alpha}{\Omega}\right)^2\right] \quad (5.61)$$

From Eq. (5.61) we find that, except for a normalization factor, the MTF for the pair of quartic phase conjugates is

$$|H(\nu;W_{2,0};\mu)| = \left|\int_{-\infty}^{\infty}\exp\left\{i2\pi\left[\left(\frac{12\Theta\mu\nu}{\Omega^2}\right)\left(\frac{\alpha}{\Omega}\right)^2+2\left(\frac{W_{2,0}}{\lambda}\right)\left(\frac{\nu}{\Omega}\right)\left(\frac{\alpha}{\Omega}\right)\right]\right\} \times \operatorname{rect}\left(\frac{\alpha}{2\Omega-|\mu|}\right)d\alpha\right| \quad (5.62)$$

From Eq. (5.62) we can see that again the integral for evaluating the MTF is a Fresnel integral, as in the case of the wavefront coding technique. However, the phase "strength" of the element is now proportional to the product $\Theta\mu$. Thus, by changing the lateral displacement μ between the phase elements, we can choose the strength of the cubic phase term of the wavefront coding technique.

In Fig. 5.11 we show four graphs of the MTF versus ν/Ω and $W_{2,0}/\lambda$ for lateral displacements $\mu/\Omega = 0.00$, 0.02, 0.06, and 0.30, with $\Theta = 12$. It is apparent from Fig. 5.11 that the MTF has low sensitivity to focus errors $W_{2,0}/\lambda$ with increasing values of the plates' lateral displacement μ/Ω.

Summarizing, we described the use of the optical anamorphic processor for relating the product-space representation, the WDF, the AF, and the product spectrum representation. We indicated that by adding a rotating ground glass, for reducing the degree of spatial coherence, the anamorphic processors are useful for linking the mutual intensity with the WDF, the AF, and the mutual spectrum. We reported two road maps for visualizing the basic integral transformations for phase-space representations.

We explored the use of the WDF for linking geometrical optics and wave optics. We discussed a method for analyzing the impact of wave aberrations from the viewpoint of the WDF. Then we related bilinear transformations with phase-space representations. We have shown that for space-invariant systems, the Volterra kernel is the product-space representation of the coherent impulse response.

We revisited the link between the OTF of an optical system that suffers from focus errors and the AF of the pupil mask. We employed

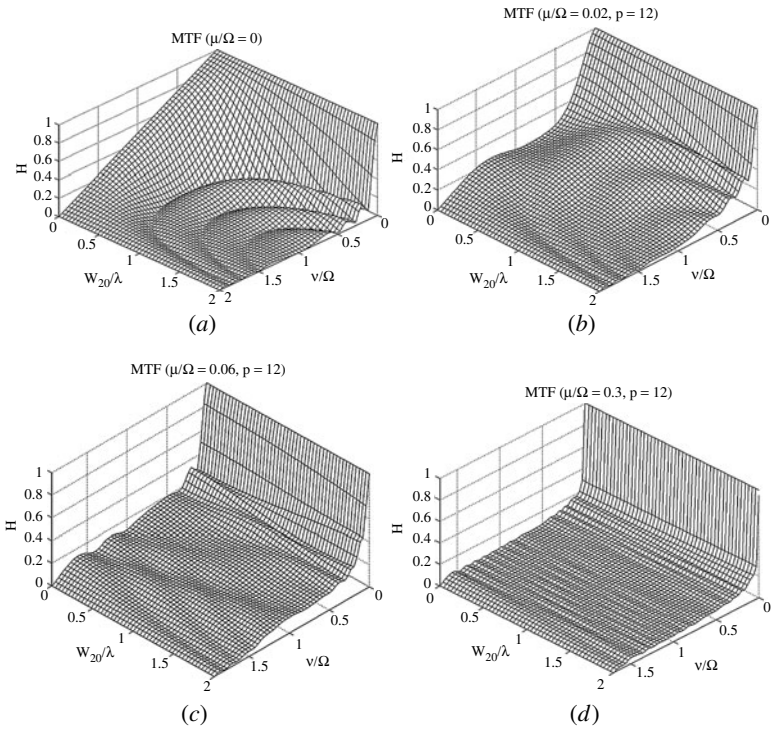


FIGURE 5.11 Modulation transfer function vs. focus errors for a phase conjugate pair. Each element of the pair has a quartic phase profile. The lateral displacements are: (a) $\mu/\Omega = 00.0$, (b) $\mu/\Omega = 02.0$, (c) $\mu/\Omega = 06.0$, and (d) $\mu/\Omega = 30.0$.

a Maclaurin series expansion, of the defocused OTF, for achieving an optical system with low sensitivity to focus errors. We have shown computer-simulated image for visualizing the extended depth of field, which can be achieved by using phase-only masks that have phase variations with odd symmetry. This analysis was extended to spherical aberration by using a simplified version of the McCutchen theorem.

We considered a coherent optical processor that uses as a spatial filter a phase mask which includes two phase elements, with opposite-sign powers. We indicated that by introducing a lateral displacement between the two elements, we generate a PSF that represents the ambiguity function of a single element. We indicated that the varifocal technique proposed by Alvarez and Lohmann can be used to visualize the defocused MTF of a cubic phase mask, as used in the wavefront coding technique. We applied a pair of phase conjugates, with quartic-phase profile, for proposing a tunable wavefront coding technique.

It is pleasure to acknowledge useful discussions with A. W. Lohmann, K. H. Brenner, N. Streibl, H. O. Bartelt, E. E. Sicre, P. Andrés, L. R. Berriel-Valdos, W. H. Steel, M. E. Testorf, W. T. Cathey, E. R. Dowski, J. Santamaría, M. J. Yzuel, A. Sauceda-Carvajal, Albertina Castro, W. T. Rhodes, M. Martinez-Corral, L. Mertz, J. P. Guigay, C. Sheppard, H. Escamilla-Taylor, J. E. A. Landgrave, and Cristina M. Gómez Sarabia.

References

1. M. J. Bastiaans, "Wigner distribution and its relatives," Chap. 1, *Phase Space Representations in Optics*, McGraw-Hill, New York, 2008.
2. J.-P. Guigay, "Ambiguity function in optical imaging," Chap. 2, *Phase Space Representations in Optics*, McGraw-Hill, New York, 2008.
3. W. T. Rhodes and J. M. Florence, "Frequency variant optical signal analysis," *Appl. Opt.*, **15**: 3073–3079 (1975).
4. B. E. Saleh, "Bilinear processing of one-dimensional signals by use of linear two-dimensional coherent optical processors," *Appl. Opt.*, **17**: 3408–3411 (1978).
5. L. J. Cutrona, E. N. Leith, C. J. Palermo, and L. J. Porcello, "Optical data processing and filtering systems," *IRE Trans. Inf. Theory* **IT6**: 386 (1960).
6. E. N. Leith, A. Kozma, and J. Upatnieks, "Coherent optical systems for data processing, spatial filtering, and wavefront reconstruction," *Optical and Electro-Optical Information Processing,* J. T. Tippet et al. (eds.), Massachusetts Institute of Technology Press, Cambridge, 1965, pp. 143–158.
7. L. J. Cutrona, "Recent developments in coherent optical technology," *Optical and Electro-Optical Information Processing,* J. T. Tippet et al. (eds.), Massachusetts Institute of Technology Press, Cambridge, 1965, pp. 83–123.
8. J. Wood and D. Barry, "Linear signal synthesis using the Radon-Wigner transform," *IEEE Trans. Signal Process.* **42**: 2101–2111 (1995).
9. J. Ville, "Théorie et applications de la notion de signal analytique," *Câblés and Transmission* **2**: 61–74 (1948). Reprinted in English in *Selected Papers on Phase-Space Optics,* M. E. Testorf et al. (eds.), *SPIE Milestone Series 181*, Bellingham, Washington, USA, 2006, pp. 149–181.
10. K.-H. Brenner and J. Ojeda-Castaneda, "Ambiguity function and Wigner distribution function applied to partially coherent imagery," *Opt. Acta*, **31**: 213–223 (1984).
11. A. W. Lohmann, J. Ojeda-Castaneda, and N. Streibl, "The influence of wave aberrations on the Wigner distribution," *Opt. Appl.*, **13**: 467–471 (1983).
12. M. J. Bastiaans, "The Wigner distribution function and the Hamiltonian's characteristics of a geometrical optical system," *Opt. Comm.*, **30**: 321–326 (1979).
13. B. E. A. Saleh, "Optical bilinear transformations: General properties," *Opt. Acta* **26**: 777–799 (1979).
14. B. E. Saleh and W. C. Goeke, "Linear restoration of bilinearly distorted images," *J. Opt. Soc. Amer.*, **70**: 506–514 (1980).
15. J. Ojeda-Castaneda and E. E. Sicre, "Bilinear optical systems: Wigner distribution function and ambiguity function representations," *Opt. Acta* **31**: 255–260 (1984).
16. H. H. Hopkins, *Wave Theory of Aberrations*, Oxford University Press, Oxford, 1950.
17. H. H. Hopkins, "The frequency response of a defocused optical system," *Proc. Roy. Soc., A* **231**: 91–103 (1955).
18. P. Mouroulis and J. MacDonald, *Geometrical Optics and Optical Design,* Oxford University Press, Oxford, 1999.

19. K. H. Brenner, A. Lohmann, and J. Ojeda-Castaneda, "The ambiguity function as a polar display of the OTF," *Opt. Commu.* **44**: 323–326 (1983).
20. H. H. Hopkins, "21st Thomas Young Oration: The applications of frequency response techniques in optics," *Proc. Physical Soc.*, **79**: 889–919 (1962).
21. H. Bartelt, J. Ojeda-Castaneda, and E. E. Sicre, "Misfocus tolerance seen by simple inspection of the ambiguity function," *Appl. Opt.* **23**: 2693–2696 (1984).
22. W. D. Furlan, G. Saavedra, and J. Lancis, "Phase-space representation as a tool for the evaluation of the polychromatic OTF," *Opt. Comm.*, **96**: 208–213 (1993).
23. J. Ojeda-Castaneda and Arturo Noyola-Isgleas, "High focal depth by apodization and digital restoration," *Appl. Opt.*, **27**: 2583–2586 (1988).
24. B. H. Saleh and N. S. Subotic, "Pre-inverse filtering of distorted images," *Appl. Opt.*, **20**: 3912–3916 (1981).
25. W. T. Cathey and E. R. Dowski, "New paradigm for imaging systems," *Appl. Opt.*, **41**: 6080–6092 (2002).
26. E. R. Dowski and T. W. Cathey, "Extended depth of field through wave-front coding," *Appl. Opt.* **34**: 1859–1865 (1995).
27. H. Wang and F. Gan, "High focal depth with a pure-phase apodizer," *Appl. Opt.* **40**: 5658–5662 (2001).
28. Nicholas George and W. Chi, "Extended depth of field using a logarithmic asphere," *J. Opt. Pure & Applied* **5**: s157–s163 (2003).
29. S. Mezouari and A. R. Harvey, "Phase reduction of defocus and spherical aberrations," *Opt. Lett.* **28**: 771–773 (2003).
30. Angel Sauceda and J. Ojeda-Castaneda, "High focal depth with fractional-power wave fronts," *Opt. Lett.* **29**: 560–562 (2004).
31. Albertina Castro and J. Ojeda-Castaneda, "Asymmetric phase masks for extended depth of field," *Appl. Opt.* **43**: 3474–3479 (2004).
32. S. S. Sherif, W. T. Cathey, and E. R. Dowski, "Phase plate to extend depth of field of incoherent hybrid imaging system," *Appl. Opt.* **43**: 2709–2721 (2004).
33. Jorge Ojeda-Castaneda, J. E. A. Landgrave, and H. M. Escamilla, "Annular phase-only mask for high focal depth," *Opt. Lett.* **30**: 1647–1649 (2005).
34. Albertina Castro, Jorge Ojeda-Castaneda, and Adolf W. Lohmann, "Bow-tie effect: Differential operator," *Appl. Opt.*, **45**: 7878–7884 (2006).
35. M. Somayaji and M. P. Christensen, "Frequency analysis of the wavefront-coding odd-symmetric quadratic phase mask," *Appl. Opt.* **46**: 216–226 (2007).
36. J. Ojeda-Castaneda, J. E. A. Landgrave, and Cristina M. Gomez-Sarabia, "Conjugated phase plate use in analysis of the frequency response of optical systems designed for extended depth of field," *Appl. Opt.* **47**: E99–E105 (2008).
37. N. Caron and Y. Sheng, "Polynomial phase masks for extending the depth of field of a microscope," *Appl. Opt.* **47**: E39–E43 (2008).
38. C. E. Cook and M. Bernfeld, *Radar Signals: An Introduction to Theory and Applications*, Artech House Inc., Norwood, Mass., 1993, p. 120.
39. E. R. Dowski and W. T. Cathey, U.S. Patent **5**, 748, 371.
40. J. Ojeda-Castaneda, L. R. Berriel-Valdos, and E. Montes, "Ambiguity function as a design tool for high focal depth," *Appl. Opt.* **27**: 790–795 (1988).
41. S. Prasad, T. Torgersen, V. Pauca, R. Plemmons, and J. van der Gracht, "Engineering the pupil phase to improve image quality," *Proc. SPIE* **5108**: 1–12 (2003).
42. D. S. Barwick, "Efficient metric for pupil-phase engineering," *Appl. Opt.* **46**: 7258–7261 (2007).
43. D. S. Barwick, "Defocus sensitivity optimization using the defocus Taylor expansion of the optical transfer function," *Appl. Opt.*, **47**: 5893–5902 (2008).
44. C. W. McCutchen, "Generalized aperture and three-dimensional diffraction image," *J. Opt. Soc. Amer.* **54**: 240–244 (1964).
45. J. Ojeda-Castaneda, P. Andrés, and A. Díaz, "Annular apodizers for low sensitivity to defocus and to spherical aberration," *Opt. Lett.*, **11**: 487–489 (1986).
46. J. Ojeda-Castaneda, P. Andrés, and E. Montes, "Phase-space representation of the Strehl ratio: Ambiguity function," *J. Opt. Soc. Amer. A* **4**: 313–317 (1987).

47. J. Ojeda-Castaneda, L. R. Berriel-Valdos, and E. Montes, "Bessel annular apodizers: Imaging characteristics," *Appl. Opt.* **26**: 2770–2772 (1987).
48. J. Ojeda-Castaneda, P. Andrés, and A. Díaz, "Strehl ratio with low sensitivity to spherical aberration," *J. Opt. Soc. Amer. A* **5**: 1233–1236 (1988).
49. J. Ojeda-Castaneda, E. Tepichin, and A. Pons, "Apodization of annular apertures: Strehl ratio," *Appl. Opt.*, **27**: 5140–5145 (1988).
50. J. Ojeda-Castaneda and C. M. Gómez-Sarabia, "Aberration balancing for shade annular apertures," *Microwave & Opt. Technol. Lett.* **1**: 226–228 (1988).
51. W. H. Steel, "Axicons with spherical surfaces," in *L'Optique en métrologie*, P. Mollet (ed.), Pergamon Press, New York, 1960, pp. 181–192.
52. J. Ojeda-Castaneda and L. R. Berriel-Valdos, "Zone plate for arbitrarily high focal depth," *Appl. Opt.* **29**: 994–997 (1990).
53. N. Davidson, A. A. Friesem, and E. Hasman, "Holographic axilens: High resolution and long focal depth," *Opt. Lett.* **16**: 523–525 (1991).
54. J. Sochacki, S. Bara, J. Jaroszewicz, and A. Kolodziejcyk, "Phase retardation of the uniform-intensity axilens," *Opt. Lett.* **17**:7–9 (1992).
55. J. Ojeda-Castaneda, M. Martinez-Corral, and P. Andrés, "Zero axial irradiance by annular screens with angular variation," *Appl. Opt*. **31**: 4600–4602 (1992).
56. J. Ojeda-Castaneda and G. Ramirez, "Zone plates for zero axial irradiance," *Opt. Lett.* **18**: pp. 87–89 (1993).
57. A. W. Lohmann, J. Ojeda-Castaneda, and G. Ramirez, "Zone plates encoding limaçonal variations", *Opt. Comm.* **114**: 30–36 (1995).
58. I. Escobar, G. Saavedra, and M. Martinez-Corral, "Reduction of the spherical aberration effect in high-numerical-aperture optical scanning instruments," *J. Opt. Soc. Amer. A* **23**: 3150–3155 (2006).
59. L. W. Alvarez, "Two-element variable-power spherical lens," U.S. Patent 3,305,294, Dec. 3, 1964.
60. A. W. Lohmann, "Lente focale variabile," Italian Patent **727**, 848, June 19, 1964.
61. A. W. Lohmann, "Improvements relating to lenses and to variable optical lens systems formed by such lenses," Patent Specification 998, 191, Patent Office, London, 1965.
62. Adolf W. Lohmann, "A new class of varifocal lenses," *Appl. Opt.* **9**: 1669–1671 (1970).
63. I. A. Palusinski, J. M. Sasián, and J. E. Greivenkamp, "Lateral shift variable aberration generators," *Appl. Opt.* **38**: 86–90 (1999).
64. N. López-Gil, H. C. Howland, B. Howland, N. Charman, and R. Applegate, "Generation of third-order spherical and coma aberrations by the use of radially symmetrical fourth-order lenses," *J. Opt. Soc. Amer. A* **15**: 2563–2571 (1998).
65. M. Somayaji and M. P. Christensen, "Enhancing form factor and light collection of multiplex imaging systems by using a cubic phase mask," *Appl. Opt.* **45**: 2911–2923 (2006).

This page intentionally left blank

CHAPTER 6

Super Resolved Imaging in Wigner-Based Phase Space

Zeev Zalevsky

School of Engineering, Bar-Ilan University
Ramat-Gan, Israel

6.1 Introduction

Super resolution (SR) is a field integrating the sciences of optics with the expertise of image processing and computer vision science.[1–7] Basically any imaging system, digital as well as a human eye, has a limited capability to separate close spatial features. This limitation can be related either to the diffraction or to the geometry of the imaging sensing array.[1] In the case of diffraction, given an imaging lens with limited aperture size, not all the rays reflected from the object are collected by the lens. According to Rayleigh criteria,[8–10] this limitation is proportional to the product of the wavelength of the illumination and the F number of the optics (the ratio of the focal length to the diameter of the lens). Thus, the smaller the F number, the better the spatial separation becomes. In the case of the geometry of the sensing array, the smaller the pixels are, i.e., the denser the spatial sampling of the space, the better the capability to reconstruct closer point sources originated from the imaged object.[1]

To overcome the limitation of a given imaging system, one may convert the spatial degrees of freedom, which before could not pass through the limited spectral bandwidth of the imaging system, into other domains that the imaging system can transmit, and then after

passing them through the system, one can convert them back to their proper location in the space domain. The process of converting to and converting back of the degrees of freedom is also called *encoding and decoding* or *multiplexing and demultiplexing*. The improvement of resolution requires "payment." The payment needed to improve the resolution in the space domain is the devotion of other domain or other subspaces into which the required spatial degrees of freedom of the input signal can be converted. The conversion of spatial degrees of freedom can be done to a single subspace or to a plurality of several such subspaces.

To do this properly without losing the desired spatial information, one needs to have a priori information on the signal. Having a priori knowledge that a certain domain is not used by the signal may allow one to designate it for the use of spatial resolution improvement. For instance, knowing that the object does not vary in the time domain may assist in using the time domain for the process of converting to and converging back the degrees of freedom.

The pluralities of other domains that may be used for this temporary conversion of degrees of freedom are the time[11–15] wavelength,[16–18] polarization,[19,20] code,[21–25] gray levels,[26] field of view,[27–32] and even light's coherence[33–35] domain.

To better understand how this adaptation of degrees of freedom may be done, one may describe the space-frequency distribution, i.e., a phase space, of the signal (SWI) and the one of the system (SWY) and perform the adaptation to all the degrees of freedom of SWI that are not graphically overlapping with the SWY representation.[36–38] The phase space that is simple for presenting the space-frequency distributions of both the signal and the system is the Wigner transformation.[39,40] Although bilinear, this transformation has interesting properties of representing basic optical modules as simple mathematical operations in this domain.[40–42]

In this chapter we provide a schematic description and explanation for how the process of SR may be understood in the Wigner space. In general, a more heuristic explanation for the SR process may involve any other phase-space diagrams. However, the advantage of using the Wigner as part of our mathematical description is related to the fact that the Wigner, although it is a bilinear transformation, can be related mathematically to the spatial degrees of freedom of a signal. In our presentation we mainly focus on the diffraction-related limitation of resolution.

The chapter is constructed as follows: In Sec. 6.2 we mathematically define the *space bandwidth* (SW), i.e., the space-frequency distribution, while separating the distribution of the signal from that of the system (SWI versus SWY). In Sec. 6.3 we focus on five ways of performing SR while explaining how those operations are represented in the Wigner

phase space. We discuss the use of Wigner for code, time, polarization, wavelength, and gray-level multiplexing SR. Section 6.4 concludes the chapter.

6.2 General Definitions

The *Wigner phase space* is a bilinear distribution defined as

$$W_u(x, \nu_x) = \int u\left(x + \frac{x'}{2}\right) u^*\left(x - \frac{x'}{2}\right) \exp\left(-2\pi i \nu_x x'\right) dx' \quad (6.1)$$

where $u(x)$ is the signal that is being transformed while x and ν_x are the space and spatial-frequency coordinates, respectively.

The projection of the Wigner distribution contains the spatial and the spectral distribution of the signal u:

$$\begin{aligned} |u(x)|^2 &= \int W_u(x, \nu_x)\, d\nu_x \\ |U(\nu_x)|^2 &= \int W_u(x, \nu_x)\, dx \end{aligned} \quad (6.2)$$

where $U(\nu_x)$ is the Fourier transform of $u(x)$. Another important property is that the area of the Wigner equals the total energy of the signal:

$$\int |u(x)|^2 dx = \int |U(\nu_x)|^2 d\nu_x = \iint W_u(x, \nu_x)\, d\nu_x\, dx \quad (6.3)$$

Because of those important and fundamental properties, as well as the fact that basic optical modules can be represented as simple and well-defined geometric operations over the Wigner chart,[41] this representation became a very useful tool for analyzing optical systems and especially when dealing with optical SR.

One important parameter is the number of degrees of freedom, also called the *space-bandwidth product* (SW). This number equals

$$N = \Delta x \cdot \Delta \nu \quad (6.4)$$

where Δx is the spatial size of the signal and $\Delta \nu$ is its spectral width. In the general case also when the Wigner distribution is not a rectangular function, it can still be shown that the number of degrees of freedom is related to the area of the Wigner.[36]

$$N = \iint W_u(x, \nu_x)\, dx\, d\nu_x$$

The proof is done by dividing the Wigner into infinitesimal rectangles with each representing a single degree of freedom and showing that their overall contribution equals the overall area of the distribution.

Note that this number is preserved in various manipulations that we may wish to apply to the signal. If the signal is compressed in the space domain (due to magnification) by a given factor, then its spectral bandwidth will be increased by the same factor in order to preserve the product.

Instead of a number, one may define an **SW function** that can be formulated by following the next steps. First, to define a binary Wigner distribution

$$\mathrm{SW}_B(x, \nu_x) = \begin{cases} 1 & \langle W_u(x, \nu_x) \rangle > W_{Tresh} \\ 0 & \text{otherwise} \end{cases} \tag{6.5}$$

where $\langle \dots \rangle$ stands for ensemble averaging over all possible signals that are relevant to the investigated problem. Now we define the SW distribution according to

$$\mathrm{SW}(x, \nu_x) = \eta \cdot \mathrm{SW}_B(x, \nu_x) \tag{6.6}$$

where η is a normalizing constant equal to

$$\eta = \frac{\iint \mathrm{SW}_B(x, \nu_x) W_u(x, \nu_x)\, dx\, d\nu_x}{\iint \mathrm{SW}_B(x, \nu_x)\, dx\, d\nu_x} \tag{6.7}$$

Now the SW is a two-dimensional distribution that, on one hand, has the heuristics of a space and frequency representation since it is evolved from a phase space and, on the other hand, has the rigorous relation to the degrees of freedom due to the inherent property related to the Wigner distribution.[36]

To realize SR, one may need to adapt or adjust the SW distribution (in the Wigner phase space) of the signal to that of the system. This process of adaptation is relevant only in the case where the system can contain at least the number of degrees of freedom that is available in the signal:

$$N_{\text{signal}} \leq N_{\text{system}} \tag{6.8}$$

This means that the SR process is relevant only if the area of SWI is smaller than the area of SWY.

As schematically depicted in Fig. 6.1 resolution may still be lost even when Eq. (6.8) holds, if the SW chart of the signal (denoted as SWI) is graphically not contained in the SW of the system (denoted by SWY). To have the signal's full resolution transmitted through the system, one needs

$$\mathrm{SWI}(x, \nu_x) \subseteq \mathrm{SWY}(x, \nu_x) \tag{6.9}$$

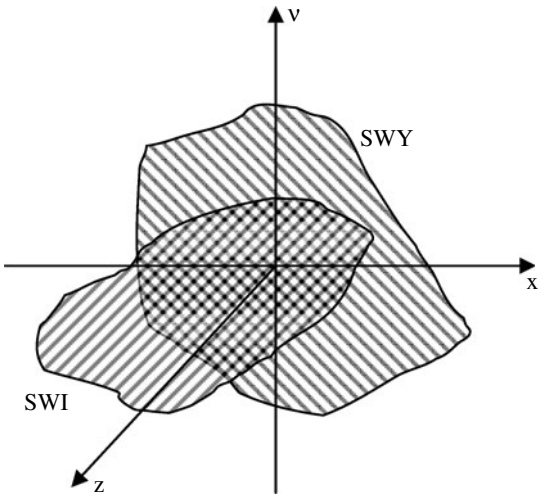


FIGURE 6.1 Example of a case where the SW adaptation process may lead to resolution improvement.

The process of adaptation includes conversion of spatial degrees of freedom contained in the SWI into other domains such as time, wavelength, polarization, coherence, gray level, field of view, or unused spatial regions/domain, etc., that are part of a hyperspectral and multidimensional Wigner distribution representing the system (SWY).

Figure 6.1 presents an example where the SR process may be implemented. It is the case where the area of the SWI is smaller than the area of SWY; i.e., the signal's number of degrees of freedom is smaller than the number of degrees of freedom that the system can transmit, and yet graphically SWI is not contained within SWY and thus resolution is reduced. By proper adaptation between the two charts, improvement of resolution is feasible in this case. The adaptation process itself may be implemented by using additional domains (denoted by z in the figure) rather than using only the space and frequency plane.

6.3 Description of SR

We will assume a general imaging system, as presented in Fig. 6.2. In this setup an encoding mask is positioned near the input object that is imaged by a finite imaging lens, which is schematically described as an aperture positioned at the optical Fourier plane, and later a decoding

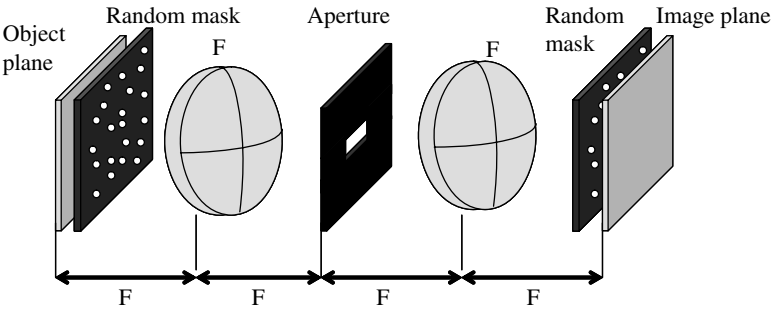


FIGURE 6.2 Schematic sketch of the imaging setup.

mask is positioned near the detector at which the output image is obtained. The detector performs time, wavelength, or other types of averaging to decode the spatial information that was encrypted by the encoding input mask. The encoding mask can be a random spatial distribution, time-varying spatial distribution, or wavelength-dependent filter, etc. The dependency of this mask is related to the domain into which we aim to convert the degrees of freedom.

As mentioned before, we assume that the spatial encoding mask which has some polarization, wavelength, or temporal dependency is positioned near the input object $u(x)$. We denote this mask as $g(x, p(t, \lambda), \lambda, t)$. The meaning of this notation is that at the different spatial positions along the mask we may have wavelength dependencies (denoted by λ), time dependencies (denoted by t), or polarization dependency (denoted by p) while the polarization dependency can have time and wavelength dependency as well: $p(t, \lambda)$.

We denote the spatial spectrum of this mask as

$$G(\nu_x, p(\lambda, t), \lambda, t) = \int g(x, p(\lambda, t), \lambda, t) \exp(-2\pi i x \nu_x x)\, dx \quad (6.10)$$

and the spatial spectrum of the object by

$$U(\nu_x) = \int u(x) \exp(-2\pi i x \nu_x x)\, dx \quad (6.11)$$

Since the encoding mask is attached to the object (i.e., mathematically there is a multiplication operation between the two distributions in the space domain), in the spectrum we obtain a convolution operation between the spectrum of the input object and the Fourier transform of the mask:

$$\int U(\nu_x') G(\nu_x - \nu_x', p(\lambda, t), \lambda, t)\, d\nu_x' \quad (6.12)$$

After the coded spatial information is passed through the finite aperture of the imaging lens, we have a multiplication of this aperture by the overall spectral distribution:

$$\text{rect}\left(\frac{\nu_x}{\Delta\nu_x}\right)\int U(\nu_x')G(\nu_x-\nu_x',\,p(\lambda,\,t),\,\lambda,\,t)\,d\nu_x' \tag{6.13}$$

We denoted the spatial width of the aperture by $\Delta\nu_x$. The decoding mask is attached to the output plane, and thus its Fourier transform performs an additional convolution operation with the overall expression of Eq. (6.13).

$$\begin{aligned}\int G_d\,(\nu_x-\nu_x',\,p(\lambda,\,t),\,\lambda,\,t)\,\text{rect}\left(\frac{\nu_x'}{\Delta\nu_x}\right)\\ \times\int U(\nu_x'')G(\nu_x'-\nu_x'',\,p(\lambda,\,t),\,\lambda,\,t)\,d\nu_x''\,d\nu_x'\end{aligned} \tag{6.14}$$

where G_d is the Fourier transform of the decoding mask g_d:

$$G_d(\nu_x,\,p(\lambda,\,t),\,\lambda,\,t)=\int g_d\,(x,\,p(\lambda,\,t),\,\lambda,\,t)\exp\left(-2\pi i x\nu_x x\right)dx \tag{6.15}$$

Assuming that the decoding mask is the complex conjugate of the encoding mask, we have

$$g_d(x)=g^*(x) \tag{6.16}$$

which means that

$$G_d(\nu_x)=G^*(-\nu_x) \tag{6.17}$$

and thus the expression of Eq. (6.14) becomes

$$\begin{aligned}U_R(\nu_x)=\int G^*(-\nu_x+\nu_x',\,p(\lambda,\,t),\,\lambda,\,t)\,\text{rect}\left(\frac{\nu_x'}{\Delta\nu_x}\right)\\ \times\int U(\nu_x'')G(\nu_x'-\nu_x'',\,p(\lambda,\,t),\,\lambda,\,t)\,d\nu_x''\,d\nu_x'\end{aligned} \tag{6.18}$$

where U_R is the spectrum of the reconstructed image u_R:

$$U_R(\nu_x)=\int u_R(x)\exp\left(-2\pi i x\nu_x x\right)dx \tag{6.19}$$

6.3.1 Code Division Multiplexing

The expression of Eq. (6.18) can be reformulated as follows:

$$U_R(\nu_x) = \int U(\nu_x'') \left[\int \text{rect}\left(\frac{\nu_x'}{\Delta\nu_x} \right) G^*(\nu_x' - \nu_x) G(\nu_x' - \nu_x'')\, d\nu_x' \right] d\nu_x'' \tag{6.20}$$

In the case of code division multiplexing there are no time, polarization, or wavelength variations. For a random encoding mask having small spatial features, the internal integral may be approximated as follows:

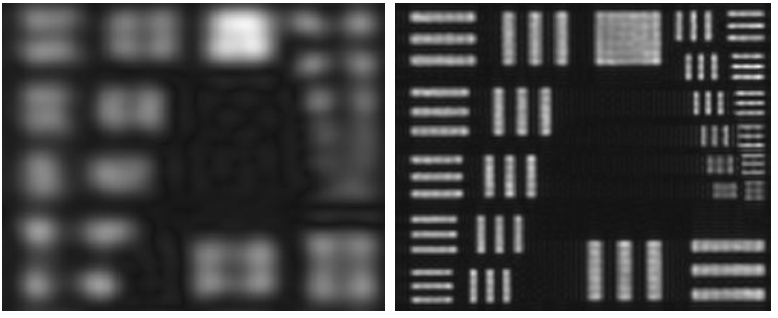
$$\int \text{rect}\left(\frac{\nu_x'}{\Delta\nu_x} \right) G^*(\nu_x' - \nu_x)\, G(\nu_x' - \nu_x'')\, d\nu_x' \approx \delta(\nu_x - \nu_x'') \tag{6.21}$$

which leads to

$$U_R(\nu_x) = \int U(\nu_x'')\, \delta(\nu_x - \nu_x'')\, d\nu_x'' = U(\nu_x) \tag{6.22}$$

Note that the assumption of Eq. (6.21) is an approximation and in practice when $\Delta\nu_x$ is getting narrower, the right wing can better be approximated with a spectral function which is wider than a delta. In addition the right wing can contain another additive term that may be approximated by a constant. The widening of the delta will blur the spectral expression of U_R which means that the super resolved image will become field of view limited. The addition of the constant level will reduce the contrast or the signal to noise ratio of the obtained reconstruction.

In Fig. 6.3 we simulate numerically the proposed approach for code division multiplexing SR. A resolution target was lowpass filtered



(*a*) (*b*)

FIGURE 6.3 (*a*) Low-resolution USAF target; (*b*) super resolved reconstruction using code multiplexing.

while being imaged through a band-limited lens aperture. The input object is a U.S. Air Force (USAF) resolution target. Its spatial blurring is clear in Fig. 6.3*a*. However, after the addition of the random encoding and decoding code multiplexing mask, one obtains the image of Fig. 6.3*b* where the high-resolution features of the resolution target are almost completely reconstructed. The resolution improvement here is by more than a factor of 5.

Obviously the image of Fig. 6.3*b* is obtained after image processing that includes some reduction of background noises.

6.3.2 Time Multiplexing

In this case we will perform time averaging (thus we do not have to assume spatial randomality of the encoding and decoding masks), so Eq. (6.18) becomes

$$U_R(\nu_x) = \int\int U(\nu_x'') \operatorname{rect}\left(\frac{\nu_x'}{\Delta\nu_x}\right) \times \left[\int G(\nu_x' - \nu_x'',\, p(\lambda, t),\, \lambda,\, t)\, G^*(-\nu_x + \nu_x',\, p(\lambda, t),\, \lambda,\, t)\, dt\right] \times d\nu_x''\, d\nu_x' \tag{6.23}$$

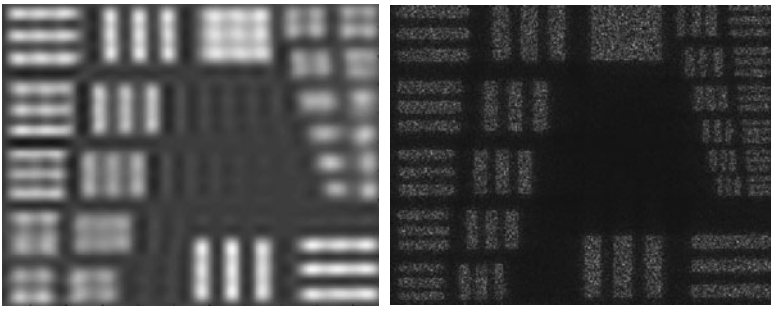
Since we have a time-varying mask we can approximate that

$$\int G(\nu_x' - \nu_x'',\, p(\lambda, t),\, \lambda,\, t) G^*(-\nu_x + \nu_x',\, p(\lambda, t),\, \lambda,\, t)\, dt \approx \delta(\nu_x - \nu_x'') \tag{6.24}$$

We obtain the final expression for the reconstructed spectrum:

$$U_R(\nu_x) = \left[\int \operatorname{rect}\left(\frac{\nu_x'}{\Delta\nu_x}\right) d\nu_x'\right] \int U(\nu_x'')\delta(\nu_x - \nu_x'') d\nu_x'' = \Delta\nu_x \cdot U(\nu_x) \tag{6.25}$$

Numerical demonstration of the time averaging SR approach may be seen in Fig. 6.4. In Fig. 6.4*a* we present the blurred (lowpass) resolution target (USAF) and in Fig. 6.4*b* its reconstruction after the addition of the time-varying encoding and decoding random mask and the performing of the averaging operation in the time domain. In the simulation we averaged 800 images. One can clearly see the resolution improvement demonstrated in this simulation. The obtained improvement factor is about 3.



(*a*) (*b*)

FIGURE 6.4 (*a*) Low-resolution USAF target; (*b*) super resolved reconstruction using time multiplexing.

6.3.3 Polarization Multiplexing

In this case we perform averaging over the time-varying polarization state, and thus Eq. (6.18) becomes

$$\begin{aligned} U_R(\nu_x) = &\int\int U(\nu_x'') \operatorname{rect}\left(\frac{\nu_x'}{\Delta\nu_x}\right) \\ &\times \left[\int G(\nu_x' - \nu_x'', p(\lambda, t), \lambda, t) G^*(-\nu_x + \nu_x', p(\lambda, t), \lambda, t)\, dp\right] \\ &\times d\nu_x''\, d\nu_x' = \int\int U(\nu_x'') \operatorname{rect}\left(\frac{\nu_x'}{\Delta\nu_x}\right) \\ &\times \left[\int G(\nu_x' - \nu_x'', p(\lambda, t), \lambda, t) G^*(-\nu_x + \nu_x', p(\lambda, t), \lambda, t)\right. \\ &\left.\times \left(\frac{dp}{dt}\right) dt\right] d\nu_x''\, d\nu_x' \end{aligned} \tag{6.26}$$

since

$$\begin{aligned} &\int G(\nu_x' - \nu_x'', p(\lambda, t), \lambda, t) G^*(-\nu_x + \nu_x', p(\lambda, t), \lambda, t) \left(\frac{dp}{dt}\right) dt \\ &\quad \approx \delta(\nu_x - \nu_x'') \end{aligned} \tag{6.27}$$

We obtain once again as the final expression for the reconstructed spectrum

$$U_R(\nu_x) = \left[\int \operatorname{rect}\left(\frac{\nu_x'}{\Delta\nu_x}\right) d\nu_x'\right] \int U(\nu_x'')\delta(\nu_x - \nu_x'')\, d\nu_x'' = \Delta\nu_x \cdot U(\nu_x) \tag{6.28}$$

6.3.4 Wavelength Multiplexing

In this case we will perform wavelength averaging (so again we do not have to assume the spatial randomality of the encoding and decoding masks), so Eq. (6.18) becomes

$$U_R(\nu_x) = \int\int U(\nu_x'') \operatorname{rect}\left(\frac{\nu_x'}{\Delta\nu_x}\right) \times \left[\int G(\nu_x' - \nu_x'',\, p(\lambda, t),\, \lambda, t) G^*(-\nu_x + \nu_x',\, p(\lambda, t),\, \lambda, t)\, d\lambda\right] \times d\nu_x''\, d\nu_x' \tag{6.29}$$

since

$$\int G(\nu_x' - \nu_x'',\, p(\lambda, t),\, \lambda, t) G^*(-\nu_x + \nu_x',\, p(\lambda, t),\, \lambda, t)\, d\lambda \approx \delta(\nu_x - \nu_x'') \tag{6.30}$$

We obtain once again as the final expression for the reconstructed spectrum

$$U_R(\nu_x) = \left[\int \operatorname{rect}\left(\frac{\nu_x'}{\Delta\nu_x}\right) d\nu_x'\right] \int U(\nu_x'')\delta(\nu_x - \nu_x'')\, d\nu_x'' = \Delta\nu_x \cdot U(\nu_x) \tag{6.31}$$

Note that in this case the meaning of the encoding mask is that every spatial pixel of the input object is "painted" with a different color. The decoding mask is identical to the encoding one, and it is picking out, from the blurred image, the right color in every high-resolution spatial location. The realization of such a mask can be straightforward by placing a chromatic spatially varying filter (e.g., chromatic absorption filter) in front of the input object which is being illuminated by a white light source. Otherwise this may also be realized by illuminating a dispersion grating with a white light source while this grating is positioned before the object, and thus the input object will be illuminated with the dispersed colors.

6.3.5 Gray-Level Multiplexing

Instead of using the domains previously discussed, for the coding of the spatial degrees of freedom, one may use the gray-level or the dynamic range domain as well. We assume that we have a priori information that the dynamic range of the input object is limited. Once again we attach a gray-level coding mask to the input object. Thus, prior to the blurring due to the low-resolution imaging, we attach a different transmission value to each pixel of the input, while the ratio between every one of those values is 2^M, where M is the a priori known and limited number of bits spanning the dynamic range of the

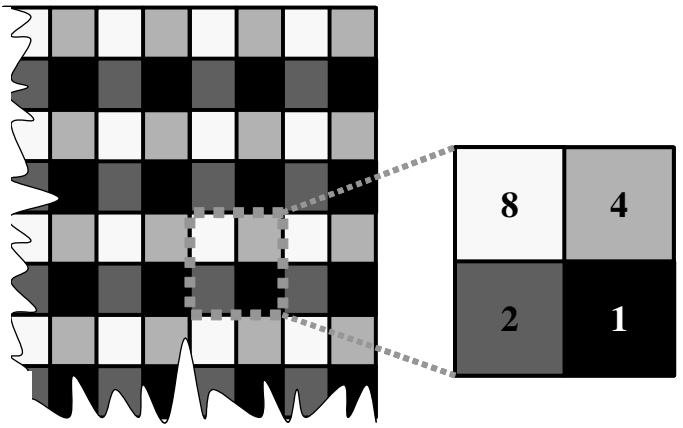


FIGURE 6.5 Schematic sketch of gray-level coding mask for binary objects.

imaged object. Obviously, the imager should have a sufficient number of dynamic range bits. It should be at least $M \times K^2$, where K is the SR factor in every spatial dimension.

To clarify this concept, the schematic description of the gray level coding mask appears in Fig. 6.5. In this figure the resolution improvement is by a factor of 2 in each axis; thus the dynamic range of the detector should have 4 times more bits than the number of bits in the original object. So if the sensor has a dynamic range of 12 bits, the imaged object should not have more than 3 bits of gray level. This coding causes a spatial blurring. However, since every high resolution lateral feature is mapped to a different region in the dynamic range axis, it may be recovered later on. This is so because a priori we know the encoding/decoding conversion map that converts between every high-resolution spatial pixel and its corresponding bits region in the dynamic range axis.

Specifically referring again to the schematic sketch of Fig. 6.5, since the ratio of two adjacent pixels of the gray-level coding mask is 2, the original object should have 1 bit of dynamic range (a binary object).

An important comment related to this approach is that it is more suitable to deal with the reduction of the imaging resolution due to geometric limitation (the number and the size of detector's pixels) than the diffraction limitation since the proper conversion between space resolution and dynamic range bits is done not continuously in space but only for spatially adjacent blocks of pixels (in Fig. 6.5 those are blocks of 2×2 pixels). The approach will not perform proper gray-level coding for spatial sampling in regions of transition between two adjacent blocks (e.g., the spatial transition sample which is the blurred value that averages the right column of pixels of one block with the left column of the next adjacent block positioned on its right side).

(*a*) (*b*)

FIGURE 6.6 (*a*) Low-resolution Lena image containing 3 bits of dynamic range; (*b*) super resolved reconstruction using gray-level multiplexing.

In Fig. 6.6 we simulated this approach by taking Lena image containing 3 bits of gray level and coded it with gray-level mask similar to the one presented in Fig. 6.5 (except that since the original object has 3 bits, the values of the coding mask should be 1, 8, 64, 512). We assume that the dynamic range of the sensor has 12 bits (maximal value of 4096). In Fig. 6.6*a* we present the low-resolution and dynamic range-limited image. In Fig. 6.6*b* we present the reconstruction. Clearly a resolution improvement of close to a factor of 2 in each axis is obtained. This is especially evident by observing the borders (e.g., the borders of the hat of Lena).

6.3.6 Description in the Phase-Space Domain

In this subsection we describe the previously discussed SR principles, using the Wigner transformation. As previously mentioned, a more heuristic phase-space diagram can also do the job of describing the SR principles. However, the advantages of using the Wigner transformation are connected to the relation between this distribution and the spatial degrees of freedom.

In Fig. 6.7 we schematically present the various steps of the setup of Fig. 6.2 for the case of time and polarization (which is time-varying) SR approaches where the degrees of freedom are converted from the spatial domain to the time or polarization domains.

In our schematic representations to come we deal with the case in which the spectral bandwidth of the signal is 3 times larger than the bandwidth that may be transmitted through the aperture of the imaging lens. The maximal bandwidth that may fit through the aperture of the lens is denoted by $\Delta\nu$, where ν and x designate the spectral and the spatial domain coordinates, respectively.

In Fig. 6.7*a* we present the phase-space diagram of a randomly varied distribution having high spatial resolution. This chart presents the time-varying random encoding mask that we will use. Every different spatial value is designated with a different color. Since we are talking

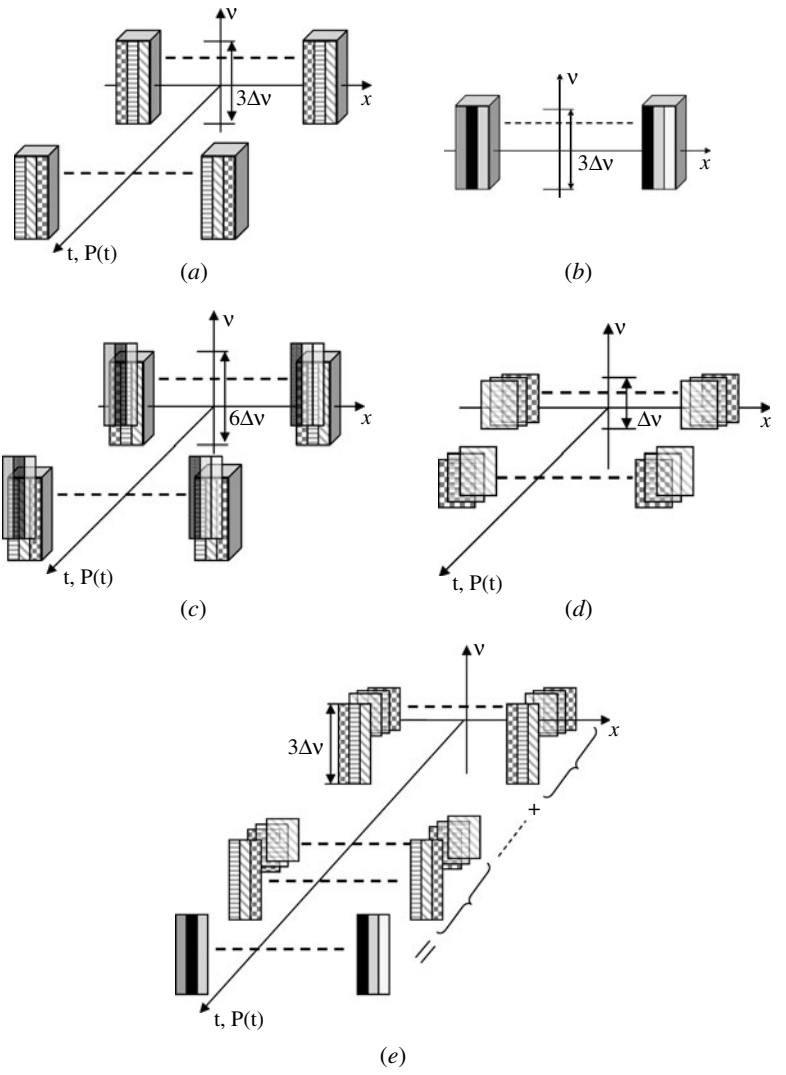


FIGURE 6.7 Schematic description of the adaptation of degrees of freedom in the phase-space plane for time and polarization multiplexing. (*a*) Distribution of the encoding mask. (*b*) Distribution of the high-resolution signal. (*c*) Distribution of the product of the signal and the encoding mask. (*d*) Blurring due to reduced resolution of the imaging system. (*e*) Decoding procedure.

about time (or time-varied polarization) multiplexing, this spatial distribution is varied with time. Thus, in Fig. 6.7*a* the order of the different colors is changed versus time. This indicates that the spatial distribution of the encoding mask is time-varying. The small spatial pixels of the chart occupy a size of $3\,\Delta\nu$ in the spectral axis because in the space domain the mask has pixels which are 3 times smaller than the imaging resolution. To simplify our explanation, if we let δx denote by δx the spatial resolution that corresponds to spectral bandwidth of $\Delta\nu$, then

$$\delta x = \frac{1}{\Delta\nu} \tag{6.32}$$

When we have 3 times finer resolution of $\delta x/3$, the spectral bandwidth will be 3 times larger, or $3\,\Delta\nu$, because the product of the spatial resolution and the spectral bandwidth equals to 1 (a well-known property of the Fourier transform).

In Fig. 6.7*b* we present the phase-space diagram of the signal that has a bandwidth of $3\Delta\nu$ (3 times larger than the bandwidth that may be transmitted through the aperture of the imaging lens).

In Fig. 6.7*c* we present the schematic sketch of the phase-space diagram of the product of the random coding mask and the signal. The phase-space distribution of the signal does not vary with time, but the encoding mask does. The bandwidth of the product equals $6\Delta\nu$ since a well-known Fourier relation dictates that the product in the space domain will be a convolution in the spectrum domain. A convolution of two spectral functions having spectral width of $3\Delta\nu$ yields a result with width of $6\Delta\nu$.

In Fig. 6.7*d* we show what happens when the high-resolution (and time-varying) product distribution is passed through a size-limited aperture. There is spatial blurring which reduces the spatial resolution (the thin rectangles became 3 times wider in the spatial axis and 3 times narrower in the spectral axis), and thus the various colors that designated different gray levels of the spatial pixels are mixed together (which reduces their dimension in the spectral axis ν). Note that the area of each rectangle denotes a single degree of freedom, and this area remains constant: increasing its dimension in the space domain reduces the dimension in the spectral axis while preserving the product.

The decoding process is described by Fig. 6.7*e*, and it includes multiplication of the captured information by the same high-resolution and time-varying decoding spatial mask distribution and then performing time averaging. The time averaging that is performed after the multiplication will extract, from every high-resolution spatial degree of freedom (that occupied spatial size of $3\Delta\nu$) the original value while averaging to zero, the undesired information that was added to it due to the spatial blurring.

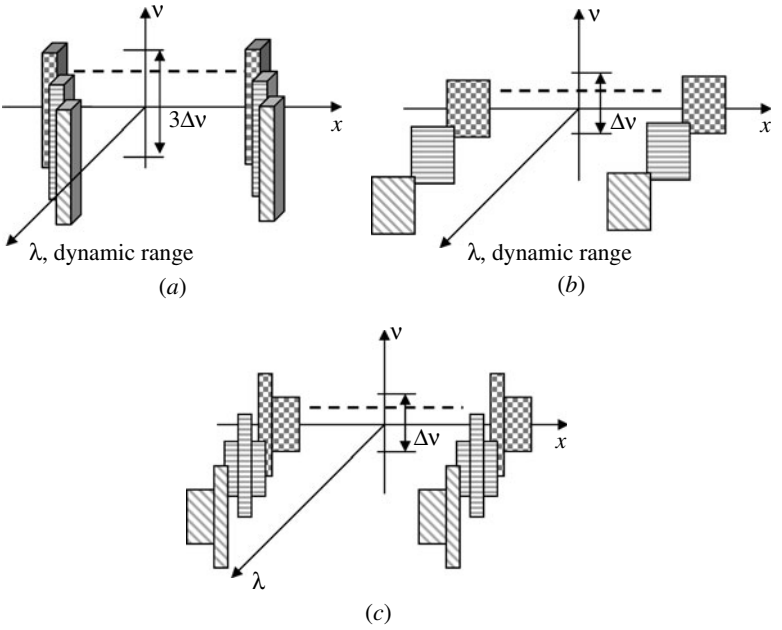


FIGURE 6.8 Schematic description of the adaptation of degrees of freedom in the phase-space plane for wavelength and dynamic range multiplexing. (*a*) Encoding. (*b*) Blurring due to reduced resolution of the imaging system. (*c*) Decoding for wavelength multiplexing.

Note that the dashed line in Fig. 6.7 represents a continuation of pixel values along x of which we show three at the start and three at the end.

In Fig. 6.8 we present the schematic sketch of the phase-space diagram in the case of wavelength or dynamic range encoding. In Fig. 6.8*a* we present the schematic sketch of the encoding process. There, once again the object contains small spatial features that occupy 3 times more bandwidth that the lens can transmit. Each of the spatial pixels is "painted" with a different wavelength or is associated with different regions along the dynamic range axis.

In Fig. 6.8*b* we show how the spatial low passing affects the phase-space diagram of Fig. 6.8*a*. The spatial information is blurred, and thus the spatial degrees of freedom become 3 times wider in the space axis x but also 3 times narrower in the spatial-frequency domain ν_x.

In Fig. 6.8*c* we present the schematic effect of the decoding. Here in each one of the spatial degrees of freedom is multiplied by a proper spatially high-resolution distribution which corresponds to the spatial

distribution used to encode the information as in Fig. 6.8*a*. The decoding is designated by thin, long rectangles having spectral dimensions of $3\Delta\nu$, which correspond to the highest spatial resolution that we aim to image. Those rectangles select out or filter out of the spatially blurred information (that is designated with short, wide rectangles having spectral width of $\Delta\nu$) the relevant gray-level information of each high-resolution pixel (having dimension of $\delta x/3$). Obviously the decoding that is presented in Fig. 6.8*c* is relevant only to the wavelength multiplexing case since in the dynamic range case the decoding is trivial: one only needs to pick up the relevant bits, knowing that every group of bits is related to a different high-resolution spatial allocation.

In Fig. 6.9 we present the schematic explanation for the case of field of view multiplexing SR and how it is seen in the phase-space diagram space. In Fig. 6.9*a* and 6.9*b* we see the phase-space diagram of an object

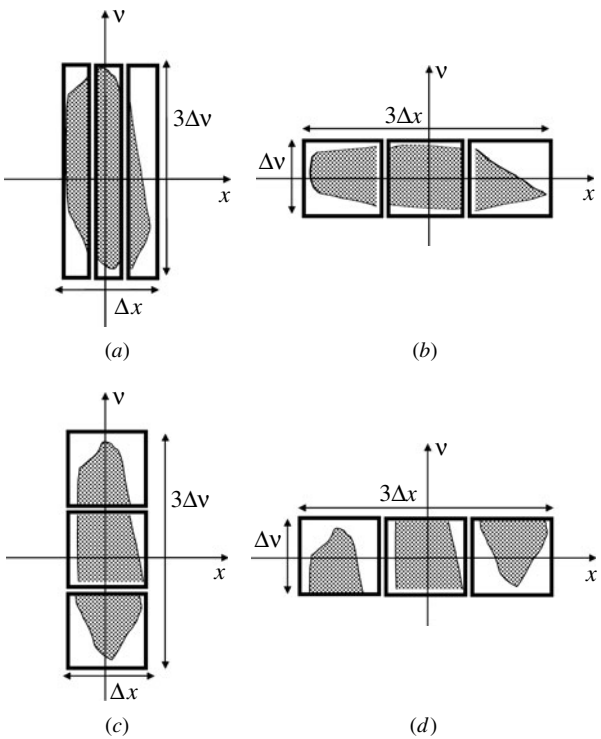


FIGURE 6.9 Schematic description of field of view multiplexing. (*a*), (*b*) Spatial separation of the information while reducing its spectral resolution and increasing its spatial bandwidth. (*c*), (*d*) Every spectral bandwidth is multiplexed to a different spatial position.

occupying the field of view of Δx and having spectral bandwidth of $3\Delta\nu$. If we divide the high spatial resolution over a 3 times larger field of view (i.e., reduces the spatial resolution by a factor of 3 by spreading it along the field of view), we obtain $3\Delta x$ for the spatial region and 3 times smaller spectral bandwidth of only $\Delta\nu$. In this figure every one out of the three spatial regions is reduced by 3 times in its resolution (thus, the spectral bandwidth of each one of the three spatial regions is reduced by a factor of 3 while their spatial region is increased by a factor of 3 and the entire area of each one of them is preserved). An effect similar to that is obtained in the case of optical magnification or zooming.

Another type of field of view multiplexing approach is the technique in which every one out of the three spectral slots (each one of the three slots has the bandwidth of $\Delta\nu$) is multiplexed by being shifted to different spatial positions, transmitted through the resolution-limiting imager, and later on demultiplexed back to compose the high-resolution image. This multiplexing/demultiplexing is done using proper gratings. The grating can redirect or reposition the different spectral slots (modulation and demodulation operation). This operation is described in Fig. 6.9*c* and 6.9*d*. In this case the various spectral slots are not changed in their shape as before (reduced in the spectral domain and expanded in the space domain), but rather only repositioned along the spatial axis. The optical realization of the setup that is using the grating to perform the relocation of the spectral slots while sacrificing the field of view can be achieved by positioning 2 or 3 gratings in predetermined locations along the imaging system, as described in Ref. 28.

In many cases such as those presented in Refs. 31 and 32 where the object is a one-dimensional object, one may use the second spatial axis to improve the imaging resolution. In this case the schematic sketch of the phase-space distribution is very similar to the one presented in Fig. 6.8, while in this case the axis that is denoted as λ or as the dynamic range axis in Fig. 6.8 (e.g., in Fig. 6.8*a*) will be now the spectral axis ν corresponding to the second spatial dimension (y instead of x).

In Fig. 6.10 we perform a true numerical simulation for the Wigner distribution in the case of a time multiplexing super resolution approach. In Fig. 6.10*a* we see the Wigner transform of a Gaussian signal. In Fig. 6.10*b* we plot the Wigner distribution of the lowpass Gaussian signal being low passed with a rectangular spectral window that is approximately 3 times narrower than the original width occupied by the input Gaussian.

In Fig. 6.10*c* we show the Wigner transform of the lowpass signal after it is multiplied by the time-varying random decoding mask. The chart in Fig. 6.10*d* is the computed Wigner distribution of the reconstruction, i.e., the signal after being averaged in the time domain.

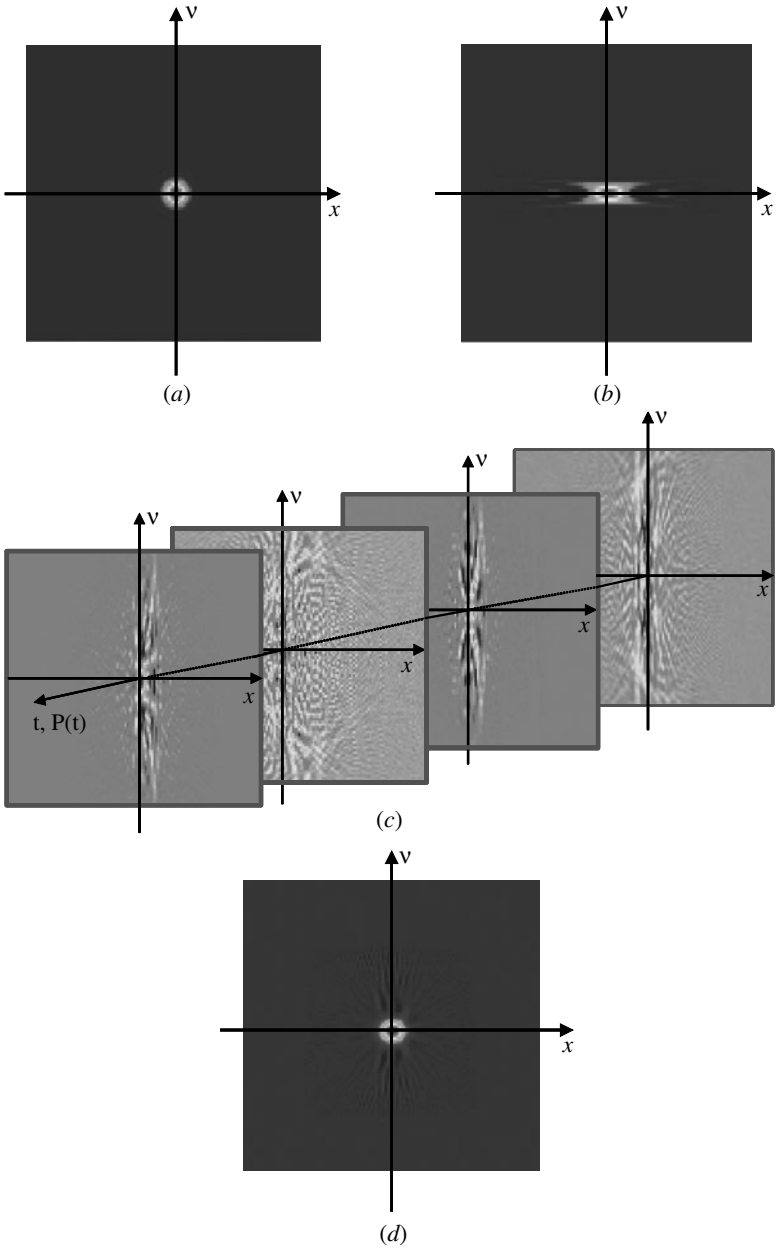


FIGURE 6.10 The Wigner transform of (*a*) Gaussian signal, (*b*) Gaussian signal after being low passed with a spectral rectangular aperture, (*c*) the low passed signal being multiplied by the time-varying random decoding mask, and (*d*) the resulting Wigner distribution of the time-averaged decoded signal.

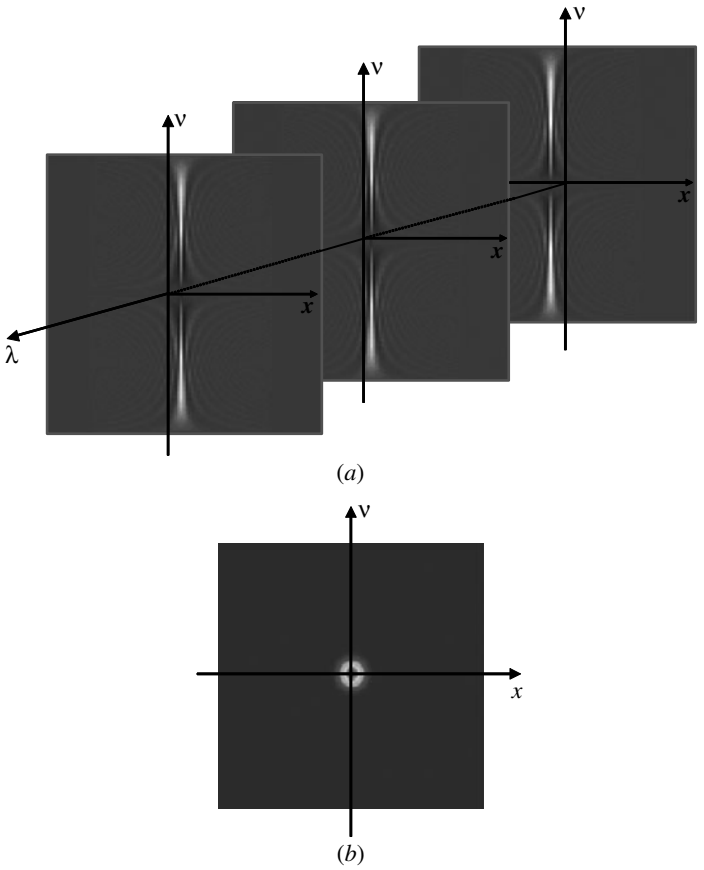


FIGURE 6.11 The Wigner of the reconstruction obtained using wavelength multiplexing for the Gaussian signal of Fig. 6.10*a*. (*a*) The lowpass signal of wavelength coded Gaussian. (*b*) The Wigner chart of the obtained reconstruction after averaging over the wavelength axis.

The averaging is performed over the spatial signals that are obtained after multiplying the blurred time-varying image by the time-varying spatial decoding mask. The Wigner distributions of those time-varying signals are presented in Fig. 6.10*c*.

We see that the original high-resolution distribution of the Gaussian that is presented in Fig. 6.10*a* is fully reconstructed in Fig. 6.10*d*.

In Fig. 6.11 we present simulations of the Wigner distribution for the case of wavelength coding. In this simulation each one of the spatial pixels was coded with a different wavelength before the spatial

blurring. The wavelength coding can be realized by using an optical element performing spatial dispersion of colors which are used as the color-space coding map (e.g., by a grating) or just by placing a space-varying color transmission filter in the setup.

The object that we used for the simulation was the Gaussian signal having the Wigner distribution presented in Fig. 6.10*a*. The Wigner chart of the wavelength coded signal is presented in Fig. 6.11*a*. This numerical simulation corresponds to the schematic sketch of Fig. 6.8*a*. Since every spatial high-resolution pixel is coded with a different wavelength, in the Wigner chart every such coded pixel is a narrow rectangle having spatial width of one pixel and maximal spectral width of $3\Delta\nu$. This narrow rectangle is shifted according to the spatial position of the coded pixel.

The result of the reconstruction obtained after averaging over the wavelength domain and using the same decoding color distribution (i.e., realization of an inverse color-space mapping), yields the result seen in Fig. 6.11*b*. We see that the obtained result is very similar to the Wigner of the original nonblurred Wigner distribution presented in Fig. 6.10*a*.

6.4 Conclusion

In this chapter we presented the usage of Wigner phase space for the description and the understanding of the field of super resolution. We showed that the Wigner phase space is more than just a heuristic representation. It is rather a chart that simplifies the understanding of the optical system and provides a representation that aids the physical comprehension of the optical behavior of the imaging system.

We focused in our description on five ways of performing super resolution by using a priori knowledge on other domains into which we could convert the spatial degrees of freedom such that they will not be lost during transmission through the band-limited optical imaging system. The five ways included code, time, polarization (which is time-varying), wavelength, and gray-level multiplexing.

The additional domain used for the conversion of spatial degrees of freedom generates a hyper phase space having more complete representation than a conventional Wigner chart or a conventional phase-space diagram.

In this chapter we presented both the schematic description of the various super resolving approaches in the Wigner phase space and the accurate numerical simulation of those approaches.

References

1. Z. Zalevsky and D. Mendlovic, *Optical Super Resolution*, Springer, 2002.
2. Z. Zalevsky, D. Mendlovic, and A. W. Lohmann, "Optical system with improved resolving power," *Progress Opt.*, vol. xl, Ch. 4 (1999).
3. W. T. Freeman, T. R. Jones, and E. C. Pasztor, "Example-based super-resolution," *IEEE Comput. Graphics & Applic*. **22**: 56–65 (2002).
4. H. Chang, D-Y. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," *IEEE Computer Vision and Pattern Recognition* (CVPR), pp. 275–282 (2004).
5. M. Elad and A. Feuer, "Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images," *IEEE Trans. Image Process.* **6**: 1646–1658 (1997).
6. A. Zomet and S. Peleg, "Multi-sensor super-resolution," *Sixth IEEE Workshop on Applications of Computer Vision* (WACV02), Orlando, Florida, US, Digital Object Identifier: 10.1109/ACV.2002.1182134 2002, pp. 27–31.
7. E. Gur and Z. Zalevsky, "Single image digital super resolution: A revised Gerschberg-Papoulis algorithm," *Int. J. Comput. Sci.*, **32**: 2 (2008).
8. E. Abbe, "Beitrage zur theorie des mikroskops und der mikroskopischen wahrnehmung," *Arch. Mikrosk. Anat*. **9**: 413-468 (1873).
9. W. Gartner and A. W. Lohmann, "An experiment going beyond Abbe's limit of diffraction," *Z. Physik* **174**: 18 (1963).
10. C. W. McCutchen, "Superresolution in Microscopy and the Abbe Resolution Limit," *J. Opt. Soc. Am.* **57**: 1190 (1967).
11. W. Lukosz, "Optical systems with resolving powers exceeding the classical limits." *J. Opt. Soc. Am.* **56**: 1463–1472 (1967).
12. M. Francon, "Amelioration de resolution d'optique," *Nuovo Cimento, Suppl*. **9**: 283–290 (1952).
13. D. Mendlovic, A. W. Lohmann, N. Konforti, I. Kiryuschev, and Z. Zalevsky, "One dimensional superresolution optical system for temporally restricted objects," *Appl. Opt*. **36**: 2353–2359 (1997).
14. D. Mendlovic, I. Kiryuschev, Z. Zalevsky, A. W. Lohmann, and D. Farkas, "Two dimensional super resolution optical system for temporally restricted objects," *Appl. Opt.* **36**: 6687–6691 (1997).
15. A. Shemer, D. Mendlovic, Z. Zalevsky, J. Garcia, and P. G. Martinez, "Super resolving optical system with time multiplexing and computer decoding," *Appl. Opt.* **38**: 7245–7251 (1999).
16. A. I. Kartashev, "Optical systems with enhanced resolving power," *Opt. Spectrosc*. **9**: 204–206 (1960).
17. J. D. Armitage, A. W. Lohmann, and D. P. Parish, "Superresolution image forming systems for objects with restricted lambda dependence," *Jpn. J. Appl. Phys*. **4**: 273–5 (1965).
18. S. A. Alexandrov and D. D. Sampson, "Spatial information transmission beyond a system's diffraction limit using optical spectral encoding of the spatial frequency," *J. Opt. A: Pure Appl. Opt*. **10**: 025304 (2008).
19. A. W. Lohmann and D. Paris, "Superresolution for nonbirefringent objects," *J. Opt. Soc. Am*. **3**: 1037–43 (1964).
20. A. Zlotnik, Z. Zalevsky, and E. Marom, "Superresolution with nonorthogonal polarization coding," *Appl. Opt*. **44**: 3705–3715 (2005).
21. J. Solomon, Z. Zalevsky, and D. Mendlovic, "Super resolution using code division multiplexing," *Appl. Opt*. **42**: 1451–1462 (2003).
22. Z. Zalevsky, J. Solomon, and D. Mendlovic, "Geometrical super resolution using code division multiplexing," *Appl. Opt.* **42**: 32–40 (2005).
23. Z. Zalevsky and A. Zlotnik, "Single Snap-Shot Double Field Optical Zoom," *Opt. Exp.* **13**: 9858–9868 (2005).
24. Z. Zalevsky, E. Leith, and K. Mills, "Optical implementation of code division multiplexing for super resolution. Part I. Spectroscopic method," *Opt. Comm*. **195**: 93–100 (2001).

25. Z. Zalevsky, E. Leith, and K. Mills, "Optical implementation of code division multiplexing for super resolution. Part II. Temporal method," *Opt. Comm.* **195**: 101–106 (2001).
26. Z. Zalevsky, P. García-Martínez, and J. García, "Superresolution using gray level coding," *Opt. Exp.* **14**: 5178–5182 (2006).
27. W. Lukosz, "Optical systems with resolving powers exceeding the classical limits II," *J. Opt. Soc. Am.* **57**: 932–41 (1967).
28. Z. Zalevsky, D. Mendlovic, and A. W. Lohmann, "Super resolution optical systems using fixed gratings," *Opt. Comm.* **163**: 79–85 (1999).
29. E. Sabo, Z. Zalevsky, D. Mendlovic, N. Konforti, and I. Kiryuschev, "Super resolution optical system using two fixed generalized Dammann gratings," *Appl. Opt.* **39**: 5318–5325 (2000).
30. E. Sabo, Z. Zalevsky, D. Mendlovic, N. Konforti, and I. Kiryuschev, "Super resolution optical system using three fixed generalized gratings: Experimental results," *J. Opt. Soc. Am. A* **18**: 514–520 (2001).
31. Z. Zalevsky, V. Eckhouse, N. Konforti, A. Shemer, D. Mendlovic, and J. Garcia, "Super resolving optical system based on spectral dilation," *Opt. Comm.* **241**: 43–50 (2004).
32. M. A. Grim and A. W. Lohmann, "Super resolution image for 1-D objects," *J. Opt. Soc. Am.* **56**: 1151–1156 (1966).
33. A. Zlotnik, Z. Zalevsky, and E. Marom, "Optical encryption using synthesized mutual intensity function," *Appl. Opt.* **43**: 3455–3465 (2004).
34. Z. Zalevsky, J. Garcia, P. Garcia-Martinez, and C. Ferreira, "Spatial information transmission using orthogonal mutual coherence coding," *Opt. Lett.* **20**: 2837–2839 (2005).
35. V. Mico, J. García, C. Ferreira, D. Sylman, and Zeev Zalevsky, "Spatial information transmission using axial temporal coherence coding," *Opt. Lett.* **32**: 736–738 (2007).
36. A. W. Lohmann, R. G. Dorsch, D. Mendlovic, Z. Zalevsky, and C. Ferreira, "About the space bandwidth product of optical signal and systems," *J. Opt. Soc. Am. A* **13**: 470–473 (1996).
37. D. Mendlovic and A. W. Lohmann, "Space-bandwidth product adaptation and its applications to super resolution: Fundamentals," *J. Opt. Soc. Am. A* **14**: 558–562 (1997).
38. D. Mendlovic, A. W. Lohmann, and Z. Zalevsky, "SW—Adaptation and its application for super resolution—Examples," *J. Opt. Soc. Am.* **14**: 563–567 (1997).
39. Z. Zalevsky, D. Mendlovic, and A. W. Lohmann, "Understanding super resolution in Wigner space," *J. Opt. Soc. Am. A* **17**: 2422–2430 (2000).
40. A. Peer, D. Wang, A. W. Lohmann, and A. A. Friesem, "Wigner formulation of optical processing with light of arbitrary coherence," *Appl. Opt.* **40**: 249–256 (2001).
41. D. Mendlovic, H. M. Ozaktas, and A. W. Lohmann, "Graded-index fibers, Wigner distribution functions and the fractional Fourier transform," *Appl. Opt.* **33**: 6188–6193 (1994).
42. K. B. Wolf, D. Mendlovic, and Z. Zalevsky, "The generalized Wigner function for analysis of super resolution systems," *Appl. Opt.* **37**: 4374–4379 (1998).

This page intentionally left blank

CHAPTER 7

Radiometry, Wave Optics, and Spatial Coherence

Arvind S. Marathay

College of Optical Sciences, University of Arizona, Tucson, Arizona, USA

John F. McCalmont

Air Force Research Laboratory, Sensors Directorate, Wright-Patterson AFB, Ohio, USA

David B. Pollock

Center for Applied Optics, University of Alabama, Huntsville, Alabama, USA

7.1 Introduction

Radiometry is the science of measurement or detection of radiation. It has a long history, starting from the works of Bouguer (1760) and Lambert. It is widely used today. The ideas and concepts of this science are based on geometrical or ray optics. However, radiation is an electromagnetic wave. Waves diffract and have states of partial coherence and polarization. Therefore, it is important to include the wave nature of radiation and formulate radiometry in the framework of wave theory. We refer to radiometry based on ray theory as *conventional* radiometry and that based on wave theory as *generalized* radiometry.

Section 7.2 reviews conventional radiometry and defines key radiometric quantities. Section 7.3 discusses the unique radiometric

qualities of Lambertian sources. Section 7.4 introduces the mutual coherence function and statistical quantities that play a central role in connecting the radiometric quantities of conventional radiometry to those of generalized radiometry. Section 7.5 examines the concept of stationary phase, an important tool in determining the radiometry of diffracting systems. Section 7.6 brings together the radiometric concepts of the previous sections to establish generalized radiometry. Section 7.7 examines specific examples of generalized radiometry in the context of blackbody radiation, partially coherent sources, and coherent sources.

7.2 Conventional Radiometry

This science is largely empirical. Several workers having to deal with the detection of radiation under different experimental conditions have found it necessary to define and use quantities applicable to their cases. A notable attempt to unify these concepts was made by Jones[1] (1963).

In this chapter, we define the radiometric quantities in common use and show their interrelationships. Basically, we have sources, illumination, and detection.

Can make visual observations of radiation in the visible region and/or quantitative measurement by radiation detectors. Radiometry provides a defined vocabulary for describing the properties of sources and various experimental arrangements for observations or detections.

We begin with the (total) radiant power Φ in units of watts (W). The *spectral radiant power* $\hat{\Phi}$ is generally expressed as a function of wavelength λ or frequency ν. We define it as a function of frequency with units of W Hz^{-1}. Its integral over all frequencies yields the total radiant power

$$\Phi = \int_0^\infty \hat{\Phi}(\nu)\, d\nu \tag{7.1}$$

With respect to a source, we speak of radiation emanating from it. We use the term *radiant exitance* M with units of W cm^{-2}. It is a function of position $\vec{r}$,

$$M(\vec{r}) = \int_0^\infty \hat{M}(\vec{r}, \nu)\; d\nu \tag{7.2}$$

In this equation, $\hat{M}(\vec{r}, \nu)$ is the *spectral radiant exitance* with units of W $\text{cm}^{-2}\text{Hz}^{-1}$.

For a planar source of area A and in the plane $z =$ constant, the total radiant power is defined by the integral over the area of the source

$$\Phi(z) = \iint_A M(x, y, z)\, dx\, dy \tag{7.3}$$

On the other hand, the radiation from the source is received on a surface. It is described by the term *radiant incidence* or *irradiance* E with units of W cm^{-2}. It too is a function of position $\vec{r}$. The *spectral irradiance* $\hat{E}(\vec{r}, \nu)$ has units of W cm^{-2}Hz^{-1}.

Let us reconsider the *radiant exitance* M. It is a function of position

$$\vec{r} = \hat{\imath}x + \hat{\jmath}y + \hat{k}z \tag{7.4}$$

From every point (x, y) on the source, it may channel different amounts of radiation energy in different directions $\hat{n}$ with components

$$\begin{aligned} \hat{n} &= \hat{\imath}p + \hat{\jmath}q + \hat{k}m \\ &= \hat{\imath}(\sin\theta\cos\phi) + \hat{\jmath}(\sin\theta\sin\phi) + \hat{k}\cos\theta \end{aligned} \tag{7.5}$$

where θ and ϕ are the polar angles, θ is measured from the z axis, and the azimuthal angle ϕ is measured from the x axis in the xy plane. The differential element of solid angle $d\Omega$ with units of *steradians* (sr) about the direction $\hat{n}$ is

$$d\Omega = \frac{dp\, dq}{m} = \sin\theta\, d\theta\, d\phi \tag{7.6}$$

This equality is established by using the Jacobian of the change of variables from (p, q) to (θ, ϕ).

To account for the variation of the source properly as a function of position and direction, a term called *radiance* $L(\vec{r}, \hat{n})$ is used, with units of W cm^{-2}sr^{-1}. An observer looking at the source in the direction of $\hat{n}$ sees the projected area $dA_{\text{proj}} = dA\cos\theta$ as shown in Fig. 7.1.

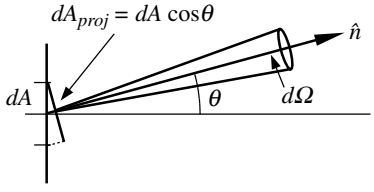


FIGURE 7.1 Projected area.

Total radiant power in terms of L is given by

$$\Phi(z) = \int_A \int_{1/2} L(\vec{r}, \hat{n}) \cos\theta \, d\Omega \, dA \tag{7.7}$$

In this expression, the z value indicates the source location on the z axis. The symbol A under the integral stands for the integration over the area of the source, and the $1/2$ under the integral implies the angle integration is limited to the right half space, $0 \leq \theta \leq \pi/2$ and $0 \leq \phi \leq 2\pi$. The product $\cos\theta \, d\Omega$ is called the projected solid angle differential element. Following Eq. (7.7), the spectral radiant power $\hat{\Phi}(z, \nu)$ is found by using the spectral radiance $\hat{L}(\vec{r}, \hat{n}, \nu)$ with units of W cm^{-2}sr^{-1}Hz^{-1}.

Far enough away from the source, the details of the source structure become less important but the characteristic angular distribution of radiation assumes importance. To describe this situation, the term *radiant intensity* $I(z, \hat{n})$ with units W sr^{-1} is used. Its solid angle integral over the right half-space yields the total radiant power

$$\Phi(z) = \int_{1/2} I(z, \hat{n}) \, d\Omega \tag{7.8}$$

The solid angle integral over the *spectral radiant intensity* $\hat{I}(z, \hat{n}, \nu)$ with units of W sr^{-1}Hz^{-1} yields the total spectral radiant power.

Among the various functions defined above, the basic one is the spectral radiance $\hat{L}(\vec{r}, \hat{n}, \nu)$. In terms of it, the other functions may be derived by regrouping and performing the appropriate integral. The interrelationships among the spectral functions are displayed in Table 7.1.

<table>
<tr><td colspan="2">$\hat{L}(\vec{r}, \hat{n}, \nu)$, [W cm^{-2}sr^{-1}Hz^{-1}]
$\hat{\Phi}(z, \nu) = \int_A \int_{1/2} \hat{L}(\vec{r}, \hat{n}, \nu) \cos\theta \, d\Omega \, dA$</td></tr>
<tr><td>$\hat{M}(\vec{r}, \nu) = \int_{1/2} \hat{L}(\vec{r}, \hat{n}, \nu) \cos\theta \, d\Omega$</td><td>$\hat{I}(z, \hat{n}, \nu) = \cos\theta \int_A \hat{L}(\vec{r}, \hat{n}, \nu) \, dA$</td></tr>
<tr><td>$\hat{\Phi}(z, \nu) = \int_A \hat{M}(\vec{r}, \nu) \, dA$</td><td>$\hat{\Phi}(z, \nu) = \int_{1/2} \hat{I}(z, \hat{n}, \nu) \, d\Omega$</td></tr>
</table>

TABLE 7.1 Interrelationships among the Spectral Functions of Conventional Radiometry

7.3 Lambertian Sources

Although sources in general do have position-dependent and/or angle-dependent properties, it is often advantageous to consider the limiting case of a source whose properties are independent of position and direction. This is an idealization, never realized in practice in the strict sense but approached in practice to a good approximation. Consider a source whose spectral radiance is independent of position $\vec{r}$ on the source and the direction of observation $\hat{n}$, that is,

$$\hat{L}(\vec{r}, \hat{n}, \nu) = \hat{L}_0(\nu) \tag{7.9}$$

A source described by Eq. (7.9) is called a *Lambertian* source. For a planar Lambertian source radiating in the right half space, it follows from the relationships shown in Table 7.1 that,

$$\begin{aligned} \hat{\Phi} &= \pi A \hat{L}_0 \\ \hat{M} &= \pi \hat{L}_0 \\ \hat{I} &= \cos\theta \, A \hat{L}_0 \end{aligned} \tag{7.10}$$

In the first two relationships, the factor π (sr) is the value of the hemispherical solid angle with proper account of the $\cos\theta$ weighting. The spectral radiant intensity relationship correctly accounts for the projection $A\cos\theta$ along the viewing direction of the source area A. Such a source would appear uniformly bright to an observer viewing it from different directions.

7.4 Mutual Coherence Function

Radiometry and wave optics can be brought together by use of the *mutual coherence function* (MCF). It is a statistical quantity, defined in terms of the wave function. In principle, it can be studied through optical experiments, and hence it is regarded as an "observable." We give a very brief introduction of the mutual coherence function in this section before proceeding to the radiometry of wave optics. The details of the theory may be found in Born and Wolf,[2] Beran and Parrent,[5] and Marathay.[3]

The MCF is defined as a cross-correlation:

$$\Gamma(\vec{r}_1, \vec{r}_2, \tau) = \langle \psi(\vec{r}_1, t)\psi^*(\vec{r}_2, t+\tau) \rangle = \lim_{T\to\infty} \frac{1}{T} \int_0^T \psi(\vec{r}_1, t)\psi^*(\vec{r}_2, t+\tau) dt \tag{7.11}$$

The field ψ is any solution of the time-dependent wave equation. The MCF contains the field evaluated at two different points at two different times. We have assumed that the field is stationary in time,

i.e., independent of the time origin. In Eq. (7.11) the angular brackets represent the time average. It is useful to define a normalized MCF.

$$\gamma_{12}(\tau) \equiv \frac{\Gamma(\vec{r}_1, \vec{r}_2, \tau)}{\sqrt{\Gamma(\vec{r}_1, \vec{r}_1, 0)\Gamma(\vec{r}_2, \vec{r}_2, 0)}} \tag{7.12}$$

By the Cauchy-Schwarz[4] inequality, it can be shown to have the property

$$0 \leq |\gamma_{12}(\tau)| \leq 1 \tag{7.13}$$

The Fourier transform of the MCF is called the *cross-spectral density function* or *mutual spectral density* (MSD) and may be defined with an ensemble average of the transformed fields.

$$\Gamma(\vec{r}_1, \vec{r}_2, \nu) = \langle \psi(\vec{r}_1, \nu)\psi^*(\vec{r}_2, \nu)\rangle \tag{7.14}$$

The same symbol is used for the transformed function provided they are identified by the arguments τ or ν as the case may be. In Eq. (7.14), the angular brackets represent the ensemble average. A normalized MSD function is defined by,

$$\gamma_{12}(\nu) \equiv \gamma(\vec{r}_1, \vec{r}_2, \nu) = \frac{\Gamma(\vec{r}_1, \vec{r}_2, \nu)}{\sqrt{\Gamma(\vec{r}_1, \vec{r}_1, \nu)\Gamma(\vec{r}_2, \vec{r}_2, \nu)}} \tag{7.15}$$

Again by the Cauchy-Schwarz inequality we have

$$0 \leq |\gamma(\vec{r}_1, \vec{r}_2, \nu)| \leq 1 \tag{7.16}$$

The MCF is a convenient theoretical quantity to describe the quality of the fringes in an interference experiment. A term such as *contrast* of the fringes is used, but a better term is the *visibility* of fringes and is denoted by V, defined by

$$V = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}} \tag{7.17}$$

In this expression, the symbol I is used for the time average of the squared modulus of the optical field $\psi(r, t)$, that is, $I = \langle|\psi(r, t)|^2\rangle$. Clearly when $I_{\max} = I_{\min}$, the visibility is zero; and when $I_{\min} = 0$, the visibility is unity.

$$0 \leq V \leq 1 \tag{7.18}$$

It is a good estimate of how dark the minimum of the fringes is.

A typical setup of a simple interference experiment is shown in Fig. 7.2. The source plane shows a noncoherent[5] source in the shape of a slit of width $2a$. The source slit is placed at a distance d_0 from the double-slit plane with the slits labeled s_1 and s_2 separated by a

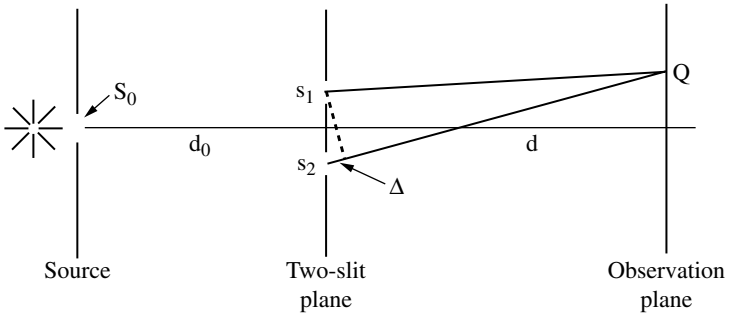


FIGURE 7.2 Two-slit interference experiment to measure spatial coherence.

distance s and symmetrically placed on either side of the horizontal z axis. The plane of observation is at a distance d behind the double-slit plane.

A noncoherent source produces a uniform illumination on the two-slit plane. By the property of noncoherence of the source, the spatial coherence function depends only on the distance between two points in the two-slit plane: $r_{12} = |\vec{r}_1 - \vec{r}_2| = s$, that is, $\gamma(\vec{r}_1, \vec{r}_2, \nu) = \gamma(r_{12}, \nu)$.

For the case of a non-coherent slit source, the normalized spatial coherence function takes the form

$$\gamma(r_{12}, \nu) = \frac{\sin u}{u}, \qquad u \equiv \frac{2\pi a r_{12}}{\lambda d_0} \tag{7.19}$$

The state of mutual spatial coherence of the radiation from the two slits is described by Eq. (7.19), with the distance $r_{12} = s$ between the two slits.

To understand the orders of magnitude of the distances involved, we assume $\lambda = 500.0$ nm, $d_0 = 1.0$ m, and $2a = 0.05$ mm. The second zero ($u = 2\pi$) of the coherence function of Eq. (7.19) is at $r_{12} = s = 20.0$ mm. Thus when the two slits are separated by a distance of 20.0 mm, the visibility of fringes in the observation plane will be zero. In the same way, when they are separated by 10.0 mm (corresponding to the first zero), the visibility will again be zero. When the slits are separated by a distance in between these two values, the visibility will be small but not zero and the fringes will exhibit reverse contrast. That is, the fringe at zero optical path difference (OPD) will be dark instead of bright. For $2a$ much less than 10.0 mm, the visibility of fringes is higher, approaching unity.

In an interference experiment, the ensemble average of the squared modulus of the optical field is detected. At a point Q in the plane of

observation, the squared modulus of the field is given by

$$I(Q) = I_1(Q) + I_2(Q) + 2\sqrt{I_1(Q)I_2(Q)}\,\frac{\sin u}{u}\cos\left(\frac{2\pi}{\lambda}\frac{sx}{d}\right) \tag{7.20}$$

In this expression, d is the distance of the plane of observation from the two-slit plane, and x is the coordinate of the point Q above the z axis. The fringe spatial frequency is $f = s/(\lambda d)$. In Eq. (7.20), $I_i(Q)$, for $i = 1$ or 2, is the squared modulus of the field at Q contributed by either of the two slits, individually. For simplicity we let $I_1(Q) = I_2(Q) = I_0$ and

$$I(Q) = 2I_0\left[1 + \frac{\sin u}{u}\cos\left(\frac{2\pi}{\lambda}\frac{sx}{d}\right)\right] \tag{7.21}$$

In this form, it is clear that the normalized spatial coherence function

$$\gamma(s, \nu) = \frac{\sin u}{u} = V \tag{7.22}$$

plays the role of visibility of the fringes. In an experiment, it may be better to leave the slit separation s and the distance d fixed, so that the fringe period is kept fixed. The spatial coherence function may be varied by changing the distance d_0 of the source plane from the two-slit plane.

7.5 Stationary Phase Approximation

The diffracted field as described by the Rayleigh-Sommerfeld diffraction theory is given by

$$\psi(x, y, z) = \iint_A \psi(x_s, y_s, 0)\left[\frac{1}{2\pi}\frac{z}{R}(1 - ikR)\frac{\exp(ikR)}{R^2}\right]dx_s dy_s \tag{7.23}$$

In this expression, $\psi(x_s, y_s, 0)$ is the amplitude distribution of the field in the diffracting aperture centered at the origin, and R is the distance between the aperture point and the point of observation $R = |\vec{r} - \vec{r}_s| = \sqrt{(x - x_s)^2 + (y - y_s)^2 + z^2}$.

Consider the two-dimensional spatial Fourier transform of the diffracted field

$$\psi(x, y, z) = \iint \psi(p, q, 0)\,\exp\left[i\frac{2\pi}{\lambda}(px + qy + mz)\right]dp\,dq \tag{7.24}$$

The stationary phase approximation is carried out on the phase function

$$\sigma(p, q) = px + qy + mz = px + qy + z\sqrt{1 - p^2 - q^2}. \tag{7.25}$$

The direction cosines (p, q) are chosen to make the first partial derivatives of $\sigma(p, q)$ equal to zero. These values are substituted in the second partial derivatives of the phase function, and the higher-order terms are neglected. This approximated phase function is used in the evaluation of the double integral of Eq. (7.24). Alternatively, it is simpler to substitute the values of (p, q) in the formula given in Born and Wolf,[2] Eq. (20) of Section 3 on double integrals, contained in Appendix III, entitled "Asymptotic Approximations to Integrals." This procedure leads to the following expression for the diffracted field.[6]

$$\psi(x, y, z) = \frac{-i}{\lambda}\frac{z}{r}\frac{\exp(ikr)}{r} \times \iint_A \psi(x_s, y_s, 0)\, \exp\left[-i\frac{2\pi}{\lambda}\left(\frac{xx_s + yy_s}{r}\right)\right] dx_s\, dy_s \tag{7.26}$$

The form of this expression suggests that we can rewrite it in the form

$$\psi(\vec{r}, \nu) = \psi(r, p, q, \nu) = \frac{-i}{\lambda} m \frac{\exp(ikr)}{r} \times \iint_A \psi(x_s, y_s, 0)\, \exp\left[-i\frac{2\pi}{\lambda}(px_s + qy_s)\right] dx_s\, dy_s \tag{7.27}$$

Here, we have used the direction cosines $p = x/r, q = y/r$, and $m = z/r$.

The diffracted field on a hemisphere is simply the spatial Fourier transform of the field distribution in the aperture as long as the distance r to the observation point satisfies the far-field condition

$$r >> \frac{2m^2a^2}{\lambda} \tag{7.28}$$

In this expression $m = z/r = \cos\theta$ which is the third direction cosine. The symbol a is the radius of the aperture. The z axis is perpendicular to the aperture plane and θ is measured from the z axis.

As the angle increases, the far-field condition becomes weaker. For observation points not satisfying the far-field condition, the higher-order terms of the stationary-phase approximation cannot be neglected. This fact was realized by Harvey and Shack[7] and they developed an aberration theory inherent to the diffraction process. This theory was applied to near-field diffraction whereby Fresnel diffraction is interpreted as an aberrated form of Fraunhofer diffraction. In the stationary-phase approximation, there are no restrictions on the direction cosines p, q and m. Hence, the diffracted field amplitude of Eq. (7.27) is valid over the entire hemisphere, free from any paraxial restrictions other than $p^2 + q^2 + m^2 = 1$. For an in-depth discussion on the application of the stationary-phase approximation to optical diffraction and imaging, see Mansuripur.[8]

Later Harvey and coworkers[9] used it to describe diffraction grating behavior and surface-scattering effects. Next we define radiometric quantities, starting with the basic relationship given in Eqs. (7.26) and (7.27).

7.6 Radiometry and Wave Optics

Radiation incident from the left on an aperture in a plane at $z = 0$ diffracts radiation into the right half-space. The diffracted field resides on a hemisphere of radius r. The origin of the coordinate system is in the open aperture. Let $d\Omega$ denote a differential element of solid angle. The differential element of area on the hemisphere is $r^2 d\Omega$ (cm^2 sr). A radiation detector responds to the ensemble average of the squared modulus of the optical field; the output is in watts (W). To find the total power radiated into the right half-space, we integrate over the hemisphere. The *spectral radiant power* in the right half-space is given by

$$\Phi(\nu) = \iint_{1/2} \langle |\psi(\vec{r}, \nu)|^2 \rangle r^2 \sin\theta \, d\theta \, d\phi = \iint_{1/2} \langle |\psi(\vec{r}, \nu)|^2 \rangle r^2 \, d\Omega \quad (7.29)$$

In this expression, the angular brackets denote the ensemble average. The linear frequency of the radiation is ν. The integrand contains the diffracted field $\psi(\vec{r}, \nu)$ as given in Eqs. (7.26) and (7.27). From Eq. (7.29), it is clear that the *ensemble average of the squared modulus of the diffracted field* $\langle |\psi(\vec{r}, \nu)|^2 \rangle$ *plays the role of spectral radiance with units* $\text{W cm}^{-2}\text{sr}^{-1}\text{Hz}^{-1}$ for radiation detection on the surface of the hemisphere.

To reduce the complexity of the resulting equations and for notational convenience, it is useful to introduce vectors defined in the following way: The coherence function may involve the

variables $(x_{s1}, y_{s1}, 0)$, $(x_{s2}, y_{s2}, 0)$, for which we introduce average and difference variables:

$$x_s \equiv \frac{1}{2}(x_{s1} + x_{s2}), \qquad y_s \equiv \frac{1}{2}(y_{s1} + y_{s2}), \qquad x_{s12} \equiv (x_{s1} - x_{s2}),$$
$$y_{s12} \equiv (y_{s1} - y_{s2})$$

Next we introduce vectors $\vec{r}_s \equiv \hat{i}x_s + \hat{j}y_s$, $\vec{r}_{s12} \equiv \hat{i}x_{s12} + \hat{j}y_{s12}$, and we can also have combination vectors $\vec{r}_s + \frac{1}{2}\vec{r}_{s12} = \hat{i}x_{s1} + \hat{j}y_{s1}$ and $\vec{r}_s - \frac{1}{2}\vec{r}_{s12} = \hat{i}x_{s2} + \hat{j}y_{s2}$. On the receiving side use, $\vec{r} \equiv \hat{i}x + \hat{j}y + \hat{k}z$ and the integration over the area elements $dx_{s1}\,dy_{s1}\,dx_{s2}\,dy_{s2} = dx_s\,dy_s\,dx_{s12}\,dy_{s12} = d^2\vec{r}_s\,d^2\vec{r}_{s12}$. The unit normal vector is defined by $\hat{n} = (\hat{i}x + \hat{j}y + \hat{k}z)/r = \hat{i}p + \hat{j}q + \hat{k}m$, where p, q, and m are the actual direction cosines. Thus, $(xx_{s12} + yy_{s12})/r = \hat{n} \cdot \vec{r}_{s12}$.

Armed with this symbolic notation, we are now ready to proceed with the formulation of radiometry for wave optics. In the computation of Eq. (7.29), we need to define the spatial coherence function

$$\Gamma(x_{s1}, y_{s1}, x_{s2}, y_{s2}, 0, \nu) = \langle \psi(x_{s1}, y_{s1}, 0, \nu)\psi^*(x_{s2}, y_{s2}, 0, \nu)\rangle,$$

$$\Gamma\left(\vec{r}_s + \frac{1}{2}\vec{r}_{s12}, \vec{r}_s - \frac{1}{2}\vec{r}_{s12}, 0, \nu\right) = \left\langle \psi\left(\vec{r}_s + \frac{1}{2}\vec{r}_{s12}, 0, \nu\right) \times \psi^*\left(\vec{r}_s - \frac{1}{2}\vec{r}_{s12}, 0, \nu\right)\right\rangle \tag{7.30}$$

The angular brackets denote the ensemble average. The coherence function can describe a coherent, partially coherent, or noncoherent input field in the $z = 0$ plane. In the computation, we can use the average and difference variables, as done in Eq. (7.30). The total spectral radiant power of Eq. (7.29) may be expressed as

$$\Phi(\nu) = \iint_{1/2} r^2\,d\Omega \left[\frac{1}{\lambda^2 r^2}\left(\frac{z}{r}\right)^2 \iint_A \Gamma\left(\vec{r}_s + \frac{1}{2}\vec{r}_{s12}, \vec{r}_s - \frac{1}{2}\vec{r}_{s12}, 0, \nu\right) \times \exp\left(-ik\frac{xx_{s12} + yy_{s12}}{r}\right) d^2\vec{r}_s d^2\vec{r}_{s12}\right] \tag{7.31}$$

This expression can be regrouped in the following way,

$$\Phi(\nu) = \iint_A d^2\vec{r}_s \iint_{1/2} m\,d\Omega \left[\frac{m}{\lambda^2}\iint_A \Gamma\left(\vec{r}_s + \frac{1}{2}\vec{r}_{s12}, \vec{r}_s - \frac{1}{2}\vec{r}_{s12}, 0, \nu\right) \times \exp\left(-ik\hat{n} \cdot \vec{r}_{s12}\right) d^2\vec{r}_{s12}\right] \tag{7.32}$$

Now following Walther,[10] we identify the expression in the square brackets as the *spectral radiance* (W $\text{cm}^{-2}\text{sr}^{-1}\text{Hz}^{-1}$) function and denote it as follows:

$$\begin{aligned} B(\vec{r}_s, 0, \hat{n}, \nu) &= \frac{m}{\lambda^2} \iint_A \Gamma\left(\vec{r}_s + \frac{1}{2}\vec{r}_{s12}, \vec{r}_s - \frac{1}{2}\vec{r}_{s12}, 0, \nu\right) \\ &\quad \times \exp\left(-ik\hat{n} \cdot \vec{r}_{s12}\right) d^2\vec{r}_{s12} \\ &= \frac{m}{\lambda^2} \iint_A \left\langle \psi\left(\vec{r}_s + \frac{1}{2}\vec{r}_{s12}, 0, \nu\right) \psi^*\left(\vec{r}_s - \frac{1}{2}\vec{r}_{s12}, 0, \nu\right) \right\rangle \\ &\quad \times \exp\left(-ik\hat{n} \cdot \vec{r}_{s12}\right) d^2\vec{r}_{s12} \end{aligned} \tag{7.33}$$

With this definition, the total spectral radiant power is

$$\Phi(\nu) = \iint_A \left[\iint_{1/2} B(\vec{r}_s, 0, \hat{n}, \nu) m \, d\Omega \right] d^2\vec{r}_s \tag{7.34}$$

Observe that we were able to do this by defining average and difference variables. Following this first step, Marchand and Wolf[11] developed the remaining functions of radiometry, as we shall now do.

The definition of *spectral radiance* given in Eq. (7.33) almost looks like an expression of a Wigner distribution[12] W_f of a function $f(x)$

$$W_f(x, \xi) = \int_{-\infty}^{\infty} f\left(x + \frac{1}{2}x'\right) f^*\left(x - \frac{1}{2}x'\right) \exp(-i2\pi\xi x') \, dx' \tag{7.35}$$

In this definition, the variables x and ξ are Fourier conjugate variables. The spectral radiance $B(\vec{r}_s, 0, \hat{n}, \nu)$ defined in Eq. (7.33) is similar to the Wigner distribution, but its arguments $\vec{r}_s$ and $\hat{n}$ are *not* Fourier conjugate variables; $\hat{n}$ is a Fourier conjugate to $\vec{r}_{s12}$.

The *spectral radiant exitance* of a source or the *spectral irradiance* on a receiving plane is defined by

$$M(\vec{r}_s, 0, \nu) = \iint_{1/2} B(\vec{r}_s, 0, \hat{n}, \nu) m \, d\Omega \tag{7.36}$$

In this way, the total *spectral radiant power* takes the form

$$\Phi(\nu) = \iint_A M(\vec{r}_s, 0, \nu) \, d^2\vec{r}_s \tag{7.37}$$

Alternatively, we can regroup Eq. (7.34) to define the *spectral radiant intensity*

$$J(\hat{n}, \nu) = m \iint_A B(\vec{r}_s, 0, \hat{n}, \nu)\, d^2\vec{r}_s \tag{7.38}$$

The total spectral radiant power now reads

$$\Phi(\nu) = \iint_{1/2} J(\hat{n}, \nu)\, d\Omega \tag{7.39}$$

The definitions given in Eqs. (7.32) to (7.39) form the structure of wave-theoretic radiometry. We based it on the diffracted field as derived from the stationary-phase expression of Eqs. (7.26) and (7.27). It is valid over the whole hemisphere of large radius centered on the open aperture containing the field distribution. Thus, the radiometry formulated in this way is free from any paraxial restrictions.

Comparison of Eqs. (7.39) and (7.29), suggests the relationship

$$J(\hat{n}, \nu) = r^2 \langle |\psi(r, p, q, \nu)|^2 \rangle \tag{7.40}$$

Alternatively, we can start with the definition of spectral radiant intensity $J(\hat{n}, \nu)$ of Eq. (7.38) and use in it the expression of spectral radiance $B(\vec{r}_s, 0, \hat{n}, \nu)$ of Eq. (7.33). The integrals on $d^2\vec{r}_s$ and $d^2\vec{r}_{s12}$ are identified as the ensemble average of the squared modulus of the diffracted field of Eq. (7.27), thus reestablishing the relationship shown in Eq. (7.40). We have just established that for radiation detection on a hemisphere the spectral radiant intensity is directly related to the ensemble average of the squared modulus of the diffracted field multiplied by the square of the radius of the hemisphere. Observe that the squared modulus of the field depends on $1/r^2$ and the factor r^2 in Eq. (7.40) indicates that the *spectral radiant intensity is a conserved quantity from one hemisphere to the next concentric hemisphere*.

The ensemble average of the squared modulus of the optical field always plays a role in optical detection. However, which radiometric quantity it represents depends very much on the geometry and experimental arrangement. In relation to Eq. (7.40), the squared modulus corresponds to the spectral radiant intensity, since the measurement was presumably made on the surface of a hemisphere. If the geometry and the experimental arrangements are changed, the above conclusion will not hold. For example, suppose the measurement is made on a plane tangent to the hemisphere and perpendicular to the z axis. When the detector explores the points on the plane, it will subtend a projected-solid angle $m\, d\Omega$ at the origin in the plane of the diffracting aperture. Let us convert the diffraction pattern on a hemisphere to the

one on the tangent plane. To do this we begin with Eq. (7.26), namely,

$$\psi(x, y, z) = \frac{-i}{\lambda}\frac{z}{r}\frac{\exp(ikr)}{r} \times \iint_A \psi(x_s, y_s, 0)\, \exp\left[-i\frac{2\pi}{\lambda}\left(\frac{xx_s + yy_s}{r}\right)\right] dx_s\, dy_s$$

Make the substitution

$$r = \frac{z}{m} = \frac{z}{\cos\theta}; \qquad r = \sqrt{x^2 + y^2 + z^2}$$

We obtain

$$\psi_P(x, y, z_0) = \frac{-i}{\lambda} m^2 \frac{\exp(ikz_0/m)}{z_0} \times \iint_A \psi(x_s, y_s, 0)\, \exp\left[-i\frac{2\pi(xx_s + yy_s)}{\lambda\sqrt{x^2 + y^2 + z_0^2}}\right] dx_s\, dy_s \tag{7.41}$$

The symbol ψ_P is the distribution of the complex amplitude diffracted onto the tangent plane perpendicular to the z axis at a distance $z = z_0$ from the diffracting aperture. The total spectral radiant power (W Hz^{-1}) now reads

$$\Phi(z_0, \nu) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \langle|\psi_P(x, y, z_0)|^2\rangle\, dx\, dy \tag{7.42}$$

The units for $\langle|\psi_P(x, y, z_0)|^2\rangle$ are W $\text{cm}^{-2}\text{Hz}^{-1}$. We have just established that *for radiation detection (measurement) on a plane parallel to the aperture plane, the ensemble average of the squared modulus of the diffracted field* $\langle|\psi_P(x, y, z_0)|^2\rangle$ *is the spectral irradiance.*

Also, it can be established that, with respect to the aperture plane, the spectral radiant exitance is given by

$$M(\vec{r}_s, 0, \nu) = \iint_{1/2} B(\vec{r}_s, 0, \hat{n}, \nu) m\, d\Omega = \Gamma(\vec{r}_s, \vec{r}_s, 0, \nu) = \langle|\psi(\vec{r}_s, \nu)|^2\rangle \tag{7.43}$$

The details of this calculation are not included here. It involves the substitution of the definition of the spectral radiance $B(\vec{r}_s, 0, \hat{n}, \nu)$ from Eq. (7.33) and evaluation of the angle integrals.

In this section, we discussed the various roles played by the ensemble average of the squared modulus of the optical field and the radiometric quantity it represents relative to the experimental arrangement.

7.7 Examples

7.7.1 Blackbody Radiation

Mehta and Wolf[13] have calculated the spatial coherence function of radiation in thermal equilibrium with the walls of a cavity. We can write it in the form

$$\Gamma\left(\vec{r}_s + \frac{1}{2}\vec{r}_{s12}, \vec{r}_s - \frac{1}{2}\vec{r}_{s12}, 0, \nu\right) = 2\pi S \frac{\sin kr_{s12}}{kr_{s12}} \tag{7.44}$$

In this expression $r_{s12} = \sqrt{(x_{s1} - x_{s2})^2 + (y_{s1} - y_{s2})^2}$ is the magnitude of the vector $\vec{r}_{s12}$. The function S is defined by

$$S \equiv S(\nu, T) = \frac{8\pi h\nu^3}{c^3}\left[\frac{1}{\exp\left(h\nu/k_B T\right) - 1}\right] \tag{7.45}$$

This is in fact the spectral density (properly denoted as $du/d\nu = S$) and has units J cm^{-3}Hz^{-1}; u is the energy density (J cm^{-3}), and the frequency ν is in hertz (Hz) of the blackbody radiation in frequency space.

Now we follow Palmer and Grant[14] and observe that radiation in thermal equilibrium with the walls of a cavity escapes (through a small hole in the wall) with velocity c and spreads in free space over 4π sr. We multiply Eq. (7.45) by $c/4\pi$ (cm s^{-1}sr^{-1}) to obtain the spectral density as accessible to measurement in free space,

$$\begin{aligned} S_{BB} \equiv S_{BB}(\nu, T) &= \frac{c}{4\pi}\frac{8\pi h\nu^3}{c^3}\left[\frac{1}{\exp\left(h\nu/k_B T\right) - 1}\right] \\ &= \frac{2h\nu^3}{c^2}\left[\frac{1}{\exp\left(h\nu/k_B T\right) - 1}\right] \end{aligned} \tag{7.46}$$

It has the units of W cm^{-2}sr^{-1}Hz^{-1}. In this way, the spatial coherence of blackbody radiation in free space may be written

$$R_{BB}\left(r_{s12}, \nu\right) = 2\pi S_{BB}(\nu, T)\frac{\sin\left(kr_{s12}\right)}{kr_{s12}} \tag{7.47}$$

We use this spatial coherence function in the definition of spectral radiance in Eq. (7.33). By use of table of integrals, this expression can be evaluated to give

$$B(\vec{r}_s, 0, \hat{n}, \nu) = \frac{m}{\lambda^2} 2\pi S_{BB} \cdot \frac{\lambda^2}{2\pi}\frac{1}{m} = S_{BB}(\nu, T) \tag{7.48}$$

This is the spectral radiance of blackbody radiation in free space and has the units of W cm^{-2}sr^{-1}Hz^{-1} as pointed out in relation to Eq. (7.46). It is independent of the average variable $\vec{r}_s$.

The total spectral radiant power is found to be

$$\Phi = \pi A S_{BB}(\nu, T) \tag{7.49}$$

The constant A is the area of the elementary hole in the blackbody cavity to provide the radiation to escape. The spectral radiant exitance evaluates to

$$M(\vec{r}_s, 0, \nu) = \pi S_{BB}(\nu, T) \tag{7.50}$$

Finally, the spectral radiant intensity can be shown to be

$$J(\hat{n}, \nu) = \cos\theta A S_{BB}(\nu, T) \tag{7.51}$$

We observe from Eqs. (7.48) to (7.51) that blackbody radiation is Lambertian; see Eq. (7.10).

7.7.2 Non-coherent Source

The spatial coherence function for a noncoherent[3,5] source is defined by

$$\Gamma\left(\vec{r}_s + \frac{1}{2}\vec{r}_{s12}, \vec{r}_s - \frac{1}{2}\vec{r}_{s12}, 0, \nu\right) = \frac{\lambda^2}{\pi}\hat{I}_0(\nu)\delta(\vec{r}_{s12}) \tag{7.52}$$

In this expression, $\hat{I}_0(\nu)$ is the squared modulus of the optical field at frequency ν.

The use of the Dirac delta function of Eq. (7.52) permits us to evaluate the spectral radiance function of Eq. (7.33); it gives

$$B(\vec{r}_s, 0, \hat{n}, \nu) = \frac{m}{\lambda^2}\frac{\lambda^2}{\pi}\hat{I}_0(\nu) = \frac{m}{\pi}\hat{I}_0(\nu) \tag{7.53}$$

Since the noncoherent source is assumed to be spatially stationary, the spectral radiance is independent of the average variable $\vec{r}_s$.

The spectral radiant power is

$$\Phi(\nu) = \frac{2}{3}A \cdot \hat{I}_0(\nu) \tag{7.54}$$

The spectral radiant exitance is given by

$$M(\vec{r}_s, 0, \nu) = \frac{2}{3}\hat{I}_0(\nu) \tag{7.55}$$

Finally, the spectral radiant intensity takes the form

$$J(\hat{n}, \nu) = m^2\frac{A}{\pi}\hat{I}_0(\nu) \tag{7.56}$$

Owing to the factor $m^2 = \cos^2\theta$, the noncoherent source will appear *darker* when viewed at large angles from the normal.

7.7.3 Coherent Wave Fields

The spatial coherence function for coherent fields assumes a factored form[15]

$$\Gamma\left(\vec{r}_s + \frac{1}{2}\vec{r}_{s12}, \vec{r}_s - \frac{1}{2}\vec{r}_{s12}, 0, \nu\right) = U\left(\vec{r}_s + \frac{1}{2}\vec{r}_{s12}\right) U^*\left(\vec{r}_s - \frac{1}{2}\vec{r}_{s12}\right) \times \delta(\nu - \nu_0) \tag{7.57}$$

The first factor $U(\vec{r}_s + \frac{1}{2}\vec{r}_{s12}) = U(\vec{r}_{s1})$ is any solution of the Helmholtz equation at frequency ν_0. The second factor is the complex conjugate of the first evaluated at a different point $\vec{r}_{s2}$. The delta function simply emphasizes that the coherent wave field is monochromatic at frequency ν_0.

The spectral radiance function of Eq. (7.33) also assumes a factored form

$$B\left(\vec{r}_s, 0, \hat{n}, \nu_0\right) = \frac{m}{\lambda^2} \int U\left(\vec{r}_{s1}\right) \exp\left(-i2\pi \frac{\hat{n}}{\lambda_0} \cdot \vec{r}_{s1}\right) d^2\vec{r}_{s1} \times \int U^*\left(\vec{r}_{s2}\right) \exp\left(+i2\pi \frac{\hat{n}}{\lambda_0} \cdot \vec{r}_{s2}\right) d^2\vec{r}_{s2} \tag{7.58}$$

Each integral is a spatial Fourier transform. We can redefine this expression as

$$B(\vec{r}_s, 0, \hat{n}, \nu_0) = \frac{m}{\lambda_0^2} \tilde{U}\left(\frac{\hat{n}}{\lambda_0}\right) \tilde{U}^*\left(\frac{\hat{n}}{\lambda_0}\right) = \frac{m}{\lambda_0^2} \left|\tilde{U}\left(\frac{\hat{n}}{\lambda_0}\right)\right|^2 \tag{7.59}$$

In this expression the argument $\hat{n}/\lambda_0$ is in fact a two-dimensional spatial frequency variable.

The spectral radiant power takes the form

$$\Phi(\nu) = A \iint_{\frac{1}{2}} \left|\frac{m}{\lambda_0} \tilde{U}\left(\frac{\hat{n}}{\lambda_0}\right)\right|^2 d\Omega \tag{7.60}$$

The spectral radiant exitance reads

$$M(\vec{r}_s, 0, \nu_0) = \iint_{\frac{1}{2}} \left|\frac{m}{\lambda_0} \tilde{U}\left(\frac{\hat{n}}{\lambda_0}\right)\right|^2 d\Omega \tag{7.61}$$

Finally, the spectral radiant intensity takes the form

$$J\left(\hat{n}, \nu_0\right) = A \left|\frac{m}{\lambda_0} \tilde{U}\left(\frac{\hat{n}}{\lambda_0}\right)\right|^2 \tag{7.62}$$

As before, A is the effective area of the open aperture in the $z = 0$ plane containing the coherent field.

7.7.4 Quasi-Homogeneous Wave-Field

The spatial coherence of a quasi-homogeneous plane wave field has a factored form

$$\Gamma_s(\vec{r}_{s1}, \vec{r}_{s2}, \nu) = I_s(\vec{r}_s, \nu)\, g_s(\vec{r}_{s12}, \nu) \tag{7.63}$$

where I_s is the ensemble average of the squared modulus of the optical field and is assumed to be very broad and slowly varying compared to the coherence function $g_s(\vec{r}_{s12}, \nu)$.

As a simple example, let us use Gaussians to represent the wave field in the initial plane $z = 0$,

$$\begin{aligned} I_s(\vec{r}_s, \nu) &= I_Q \exp\left(-\frac{r_s^2}{2\sigma_Q^2}\right) \\ g_s(\vec{r}_{s12}, \nu) &= \exp\left(-\frac{r_{s12}^2}{2\sigma_g^2}\right) \end{aligned} \tag{7.64}$$

The widths σ_Q and σ_g are both frequency ν-dependent and $\sigma_Q >> \sigma_g$.

The first order of business is to obtain the spectral radiance

$$B(\vec{r}_s, \hat{n}, \nu) = \frac{1}{\pi} I_s(\vec{r}_s, \nu)\, \frac{m}{2}(k\sigma_g)^2 \exp\left[-\frac{1}{2}(k\sigma_g)^2(p^2 + q^2)\right] \tag{7.65}$$

It is a Gaussian in angular spectrum space.

The spectral radiant power is given by

$$\Phi(\nu) = \int I_s(\vec{r}_s, \nu)\, d^2\vec{r}_s \left\{(k\sigma_g)^2 \int_0^1 \exp\left[-\frac{(k\sigma_g)^2}{2}(1 - \mu^2)\right] \mu^2\, d\mu\right\} \tag{7.66}$$

In doing the angle integral, we substituted $\mu = \cos\theta$. The angle integral, contained within curly braces, can be evaluated by using *Mathematica*. It can be plotted as a function of $(k\sigma_g)^2/2$. For values of $\sigma_g > \lambda$ and for values $\sigma_g >> \lambda$, the angle integral is rather approximately unity to a high degree of accuracy. Hence, we may write

$$\Phi(\nu) = \int I_s(\vec{r}_s, \nu)\, d^2\vec{r}_s \tag{7.67}$$

In like manner, the spectral radiant exitance can be shown to be

$$M(\vec{r}_s, \nu) = I_s(\vec{r}_s, \nu) \tag{7.68}$$

The spectral radiant intensity takes the form

$$J(\hat{n}, \nu) = \frac{1}{2\pi}(k\sigma_g)^2 \exp\left[-\frac{1}{2}(k\sigma_g)^2\left(p^2+q^2\right)\right] \int I_s(\vec{r}_s, \nu)\, d^2\vec{r}_s \tag{7.69}$$

The factor $(p^2+q^2) = (\sin\theta)^2$ shows the dependence on angle θ which varies from 0 to $\pi/2$.

The set of Eqs. (7.65) to (7.69) describes the radiometry with respect to the initial plane $z = 0$. After propagation the Gaussians retain their form but the scale changes. For example, I_s after propagation takes the form

$$I(\vec{r}, \nu) = I_Q\left(\frac{1}{1+z'^2}\right) \exp\left(-\frac{r^2}{2\sigma_Q^2}\frac{1}{1+z'^2}\right) \tag{7.70}$$

The parameter z' is defined by $z' \equiv z/(k\sigma_Q\sigma_g)$ in which z is the location of the plane parallel to the initial plane $z = 0$. The peak value is reduced by the factor $1/(1+z'^2)$, and the variance σ_g^2 is increased by the factor $1+z'^2$. The spatial coherence function scales in an analogous manner but acquires a linear phase factor. For details, see Ref. 3 where more references to the literature are included.

Acknowledgments

Randal Johnson, MacAulay-Brown, Inc., for help in the design of the figures. Juliet Hughes, University of Arizona, for help in the layout of the manuscript

References

1. R. C, Jones, "Terminology in photometry and radiometry," *J. Opt. Soc. Am.* **53**(11): 1314–1315 (1963).
2. Max Born and Emil Wolf, *Principles of Optics*, Pergamon Press, New York, 1959.
3. Arvind S. Marathay, *Elements of Optical Coherence Theory*, Wiley, New York, 1982, Chapter 3, pp. 29–32.
4. I. S. Sokolnikoff and E. M. Redheffer, *Mathematics of Physics and Modern Engineering*, McGraw-Hill, New York, 1958, p. 322.
5. M. J. Beran and G. B. Parrent, Jr., *Theory of Partial Coherence*, Prentice-Hall, Englewood Cliffs, N.J., 1964; see also "incoherent source," ref. 3, Marathay, p. 78.
6. The details of stationary-phase calculation are described in Chapter 6, "Diffraction," by A. S. Marathay, J. F. McCalmont, and J. Shiefman, *Optical Engineer's Desk Reference*, Edited by William L. Wolfe, Pub. Optical Society of America and The International Society for Optical Engineering, 2003.
7. James Harvey and Roland Shack, "Aberrations of diffracted wave fields," *Appl. Opt*. **17**: 3003 (1978).

8. Masud Mansuripur, *The Physical Principles of Magneto-Optical Recording*, Cambridge University Press, New York, 1995, Chapter 3, Section 3.1, Stationary phase approximation.
9. James Harvey, "Fourier treatment of near-field scalar diffraction theory," *Am. J. Phys.* **47**(11): 974–980 (1979). James E. Harvey, Cynthia L. Vernold, Andrey Krywonos, and Patrick L. Thompson, "Diffracted radiance: A fundamental quantity in nonparaxial scalar diffraction theory," *Appl. Opt.* **38**(31): 6469–6481 (1999). James E. Harvey, Cynthia L. Vernold, Andrey Krywonos, and Patrick L. Thompson, "Diffracted radiance: A fundamental quantity in nonparaxial scalar diffraction theory: Errata," *Appl. Opt.* **39**(34): 6374–6375 (2000). James E. Harvey, Andrey Krywonos, and Cynthia L. Vernold, "Modified Beckmann-Kirchhoff scattering model for rough surfaces with large incident and scattering angles," *Opt. Eng.* **46**(7): 078002-[1-9] (2007).
10. A, Walther, "Radiometry and coherence," *J. Opt. Soc. Am.* **58**(9): 1256–1259 (1968).
11. E. W. Marchand, and E. Wolf, "Radiometry with sources of any state of coherence," *J. Opt. Soc. Am.* **64**(9): 1219–1226 (1974).
12. Harrison H. Barrett and Kyle J. Myers, *Foundations of Image Science*, Wiley Interscience, New York, 2004, Section 5.2, p. 227.
13. C. L. Mehta and E. Wolf, "Coherence properties of blackbody radiation, I. Correlation tensors of the classical field," *Phys. Rev. A* **134**(5): 1143–1149 (1964).
14. J. M. Palmer and B. Grant, *The Art of Radiometry* (based on J. M. Palmer, Lecture Notes, University of Arizona) to be published.
15. C. L. Mehta E. Wolf, and A. P. Balachandran, "Some theorems on the unimodular complex degree of coherence," *J. Math. Phys.* **7**(1): 133–138 (1966).

CHAPTER 8
Rays and Waves

Miguel A. Alonso

The Institute of Optics, University of Rochester, New York, USA

8.1 Introduction

From a conceptual point of view, the ray model for the propagation of light is an outdated theory. Yet, it is still perhaps the most important tool for the design and modeling of imaging and illumination optical instruments due to its simplicity, intuitiveness, and often sufficient accuracy. (An analogous although significantly more extreme situation occurs for mechanical systems: machines and tools are designed and modeled using classical mechanics, which is also conceptually an outdated theory; quantum effects are important only for very small or very special mechanical systems.) It turns out that even when wave effects are important, they can often be modeled based on the ray-optical description of the system in question. There are a variety of methods for modeling wave propagation based on rays. Phase space is a natural framework for studying the link between the ray and wave models. Using phase-space representations, a wave field can be described as a function of both position and direction of propagation.

In general, the use of rays leads only to approximate wave propagation models. However, the laws of wave propagation can be expressed exactly in terms of rays in three limits. The first is the paraxial limit for the case of propagation through the so-called ABCD or first-order systems. (See Chaps. 1 and 3 by Martin Bastiaans and Tatiana Alieva, respectively.) These systems include free-space and homogeneous media, thin quadratic lenses, and transversely linear and quadratic gradient-index media. The propagation of waves in these systems can be described exactly in ray terms, either by employing a point-spread function like that in Eq. (1.42), or in terms of the Wigner function as in Eq. (1.44). The second limit is the so-called quasi-homogeneous

limit, corresponding to fields of low spatial coherence. In this case, the Wigner function and other bilinear phase-space representations acquire all the defining properties of the radiance, which is essentially a ray-weighting distribution whose propagation is ruled by the laws of geometrical optics. The radiometric description of wave fields and the quasi-homogeneous limit are discussed in Chap. 7. The third limit is that of small wavelength. This limit is discussed in many optics textbooks[1–3] and is the main topic of the two books by Kravtsov and Orlov.[4,5] As it turns out, there are a variety of ways in which this third limit can be enforced, all leading to the same laws for the rays, but different ray-based descriptions of the wave field. These various approaches are the topic of this chapter. Also discussed briefly here is the mathematically analogous semiclassical limit of quantum mechanics, where instead of rays one uses classical particle trajectories to estimate the wave aspects of particle motion.

For simplicity, the propagation of scalar fields is considered in what follows. We also limit our attention to the case of monochromatic light (of frequency ω), where the field's time dependence can be factored as $E(\vec{r}, t) = U(\vec{r}) \exp(-i\omega t)$, with $\vec{r} = (x, y, z)$ being the position vector. The basic equation that describes the propagation of a monochromatic scalar field $U(\vec{r})$ is the Helmholtz equation

$$[\nabla^2 + k^2 n^2(\vec{r})]U(\vec{r}) = 0 \tag{8.1}$$

where $k = \omega/c$ is the wave number (with c representing the speed of light in vacuum) and $n(\vec{r})$ is the position-dependent refractive index. It is assumed throughout that this refractive index is real, i.e., that the medium presents no absorption or gain.

8.2 Small-Wavelength Limit in the Position Representation. I: Geometrical Optics

The standard procedure for studying the connection between wave and ray optics relies on the assumption that the field U consists of a slowly varying amplitude and a rapidly oscillating phase proportional to the wave number, i.e.,

$$U(\vec{r}) = A(\vec{r}) \exp[ik\phi(\vec{r})] \tag{8.2}$$

where it is assumed that at least ϕ is real. The substitution of Eq. (8.2) into Eq. (8.1) gives, after some reordering and multiplication by $\exp(-ik\phi)$,

$$k^2 A(n^2 - |\nabla\phi|^2) + ik(2\nabla A \cdot \nabla\phi + A\nabla^2\phi) + \nabla^2 A = 0 \tag{8.3}$$

The goal now is to separate this equation into two or more equations that are amenable to a simple solution or that at least lead to an intuitive interpretation. Two alternative approaches are considered in what follows:

1. Assume that the wave number k is large, and use an asymptotic treatment.
2. Assume that both A and ϕ are real, and separate Eq. (8.3) into real and imaginary parts.

In this section and the next we explore the first approach, which leads to geometrical optics. The second approach is discussed in Sec. 8.4.

8.2.1 The Eikonal and Geometrical Optics

Since k is assumed to be large, the leading part of Eq. (8.3) is the term proportional to k^2. Therefore, as a first step toward enforcing Eq. (8.3), we choose to make the coefficient of this leading part vanish. This results in the expression

$$|\nabla\phi(\vec{r})|^2 = n^2(\vec{r}) \tag{8.4}$$

This equation is the well-known eikonal equation, which is formally equivalent to other formulations of geometrical optics. The function ϕ is called the eikonal function or simply the eikonal. This formulation was proposed by Bruns[6] in 1895, although an equivalent formalism was proposed by Hamilton[7] almost seventy years earlier.

The eikonal equation can be solved (or at least written in a form that is better suited for numerical solution) by parameterizing the position vector in terms of three independent parameters τ, ξ_1, ξ_2 as $\vec{r} = \vec{R}(\tau, \xi_1, \xi_2)$. These parameters must be chosen so that $\vec{R}$ moves in three different directions with variations in each of them (i.e., the three vectors corresponding to the partial derivatives of $\vec{R}$ with respect to each of the parameters must be linearly independent). To solve the eikonal equation, the partial derivative with respect to τ (denoted by an overdot) of $\vec{R}$ is chosen as parallel to the gradient of the eikonal, i.e.,

$$\frac{\partial\vec{R}}{\partial\tau} = \dot{\vec{R}} = \alpha\nabla\phi(\vec{R}) \tag{8.5}$$

where the proportionality function $\alpha(\tau, \xi_1, \xi_2)$ is assumed to be positive. Substituting the gradient of the eikonal as given in Eq. (8.5) into

Eq. (8.4) evaluated at $\vec{R}$, we find

$$\alpha(\tau, \xi_1, \xi_2) = \frac{v(\tau, \xi_1, \xi_2)}{n[\vec{R}(\tau, \xi_1, \xi_2)]} \tag{8.6}$$

where

$$v(\tau, \xi_1, \xi_2) = |\dot{\vec{R}}(\tau, \xi_1, \xi_2)| \tag{8.7}$$

is the speed of the parameterization in τ. The trajectories traced by the vector $\vec{R}$ for increasing τ are precisely the rays of geometrical optics.

To find the equations that determine the rays, it is convenient to define the *optical momentum* vector $\vec{P}$ as

$$\vec{P}(\tau, \xi_1, \xi_2) = \nabla\phi[\vec{R}(\tau, \xi_1, \xi_2)] = \frac{n(\vec{R})}{v}\dot{\vec{R}} \tag{8.8}$$

where Eqs. (8.5) and (8.6) were used in the last step. The equation for the propagation of the rays is found by considering the derivative with respect to τ of the first two parts of Eq. (8.8):

$$\dot{\vec{P}} = (\dot{\vec{R}}\cdot\nabla)\nabla\phi = \frac{v}{n}(\nabla\phi\cdot\nabla)\nabla\phi = \frac{v}{n}\frac{\nabla(\nabla\phi\cdot\nabla\phi)}{2} = \frac{v}{n}\frac{\nabla(n^2)}{2} = v\nabla n(\vec{R}) \tag{8.9}$$

where the chain rule was used in the first step, the second part of Eq. (8.8) was used in the second step, and Eq. (8.4) was used in the fourth step. We must also find how the eikonal ϕ evolves along a ray. For this purpose, let L denote the eikonal evaluated along a parameterized ray, i.e.,

$$L(\tau, \xi_1, \xi_2) = \phi[\vec{R}(\tau, \xi_1, \xi_2)] \tag{8.10}$$

This function increases with τ according to the expression

$$\dot{L} = \dot{\vec{R}}\cdot\nabla\phi = n(\vec{R})v \tag{8.11}$$

where the chain rule, as well as Eqs. (8.7) and (8.8), was used. The value of the eikonal is then found by integrating this expression in τ.

The expressions in Eqs. (8.8), (8.9), and (8.11) are the basic equations that rule the propagation of the rays. Let us summarize these equations:

$$\dot{\vec{R}} = \frac{v}{n(\vec{R})}\vec{P} \tag{8.12a}$$

$$\dot{\vec{P}} = v\,\nabla n(\vec{R}) \tag{8.12b}$$

$$\dot{L} = v\,n(\vec{R}) \tag{8.12c}$$

(Recall that $v = |\dot{\vec{R}}|$ is the speed of the parameterization.) The first two of these equations rule the propagation of each ray, and the third implies that the eikonal corresponds to the *optical path length* along the ray. Equation (8.12*a*) is the geometrical definition of the optical momentum as a vector that is locally tangent to the ray and whose magnitude is given by the local refractive index (see Fig. 8.1), while Eq. (8.12*b*) states that local changes in the refractive index modify the direction of propagation of the ray. For a homogeneous medium, Eq. (8.12*b*) implies that $\vec{P}$ remains constant, so rays are straight lines and, due to Eq. (8.12*c*), L increases linearly with the propagation length. At the interface between two homogeneous media, on the other hand, ∇n is deltalike, pointing in the direction of the interface's normal. In this case, Eq. (8.12*b*) states that the component of $\vec{P}$ locally parallel to the interface remains constant. This fact, combined with the requirement that $|\vec{P}| = n$ at either side of the interface, leads to Snell's law. When the refractive index changes continuously and smoothly, the rays change direction gradually and become curved.

While varying the parameter τ causes $\vec{R}$ to move along a ray, variations of ξ_1 or ξ_2 make $\vec{R}$ move from one ray to another. That is, the above set of equations describes the evolution of a two-parameter family of rays, with each ray corresponding to a set of values of the parameters $\boldsymbol{\xi} = (\xi_1, \xi_2)$. However, the evolution of each ray is completely autonomous, as can be appreciated from Eqs. (8.12*a*) and (8.12*b*), which involve no operation on the parameters ξ_1, ξ_2. Nevertheless, the optical path lengths of the different rays in the family are interconnected, as one can see from considering the derivative of Eq. (8.10) with respect to one of these parameters,

$$\frac{\partial L}{\partial \xi_j} = \nabla\phi \cdot \frac{\partial \vec{R}}{\partial \xi_j} = \vec{P} \cdot \frac{\partial \vec{R}}{\partial \xi_j} \tag{8.13}$$

This relation is a consequence of the fact that in Eq. (8.5) the rays were chosen to be perpendicular to the surfaces of constant Φ (and therefore L). That is, all along their propagation, the rays constitute what is known as a *normal congruence*, which means that there exists a continuous set of surfaces that intersect perpendicularly all rays in the family. Then L corresponds to the optical path length along the rays measured from one of these normal surfaces. For example, for a

FIGURE 8.1 Definition of the optical momentum.

point source in free space, the rays are straight lines radiating away from the source, and the normal surfaces are spheres centered at the source. The reference surface is usually chosen to be the sphere of zero radius, i.e., the source itself.

8.2.2 Choosing z as the Parameter

The equations for the rays given above are general in the sense that the parameterization along the rays is arbitrary. There are, however, particular parameterizations that are convenient in certain situations, leading to several different forms for the ray equations.[8–10] Some common choices are the ones that make τ equal to the arclength of the ray (so $v = 1$), the optical path length (so $v = 1/n$), or the length divided by the local refractive index (so $v = n$). In what follows, we concentrate on a fourth particular parameterization which, while more limited in application, is convenient for the type of problems studied in this book. This parameterization is only valid when one can choose a "main direction of propagation" such that the component of the momentum in this direction for all rays in the family is always positive. Let us align the z axis with this direction of propagation. Then the condition for the application of this parameterization is that the rays do not turn around in z, so that their positions are single-valued functions of z. Under these circumstances, z itself can be used as the parameter of propagation.

For this parameterization, it is convenient to separate the z and transverse components of the position and momentum vectors as

$$\vec{R}(z, \boldsymbol{\xi}) = [X(z, \boldsymbol{\xi}), Y(z, \boldsymbol{\xi}), z] = [\mathbf{X}(z, \boldsymbol{\xi}), z] \tag{8.14}$$

$$\vec{P}(z, \boldsymbol{\xi}) = [P_x(z, \boldsymbol{\xi}), P_y(z, \boldsymbol{\xi}), H(z, \boldsymbol{\xi})] = [\mathbf{P}(z, \boldsymbol{\xi}), H(z, \boldsymbol{\xi})] \tag{8.15}$$

where $\mathbf{X} = (X, Y)$ and $\mathbf{P} = (P_x, P_y)$. As stated earlier, it is assumed that the longitudinal component of the momentum, namely H, is always positive. As we know from the eikonal equation [i.e., Eq. (8.4)], this function is related to $\mathbf{P}$ and the refractive index by

$$H(z, \boldsymbol{\xi}) = \sqrt{n^2(\mathbf{X}, z) - |\mathbf{P}|^2} \tag{8.16}$$

Let us now find the form that the ray equations take for this parameterization. First, the longitudinal part (i.e., the z component) of Eq. (8.12*a*) gives $1 = vH/n$, that is,

$$v = \frac{n}{H} \tag{8.17}$$

By using this result, the transverse part of Eq. (8.12*a*) gives

$$\dot{\mathbf{X}} = \frac{\mathbf{P}}{H} \tag{8.18a}$$

where the overdot now denotes a derivative with respect to z. Similarly, the transverse and longitudinal parts of Eq. (8.12*b*) become, respectively,

$$\dot{\mathbf{P}} = \frac{n}{H}\frac{\partial n}{\partial \mathbf{x}}(\mathbf{X}, z) = \frac{1}{2H}\frac{\partial n^2}{\partial \mathbf{x}}(\mathbf{X}, z) \tag{8.18b}$$

$$\dot{H} = \frac{n}{H}\frac{\partial n}{\partial z}(\mathbf{X}, z) = \frac{1}{2H}\frac{\partial n^2}{\partial z}(\mathbf{X}, z) \tag{8.18c}$$

where $\partial/\partial \mathbf{x}$ is the transverse (or x, y) part of the gradient. Finally, Eq. (8.12*c*) becomes

$$\dot{L} = \frac{n^2(\mathbf{X}, z)}{H} \tag{8.18d}$$

8.2.3 Ray-Optical Phase Space and the Lagrange Manifold

At this point, it is convenient to recall that rays can be represented by points in phase space. For simplicity, let us consider fields propagating in only two dimensions. That is, there is only one transverse coordinate x plus the longitudinal coordinate z. In this case, the ray equations become

$$\dot{X}(z, \xi) = \frac{P(z, \xi)}{H(z, \xi)} \tag{8.19a}$$

$$\dot{P}(z, \xi) = \frac{1}{2H(z, \xi)}\frac{\partial n^2}{\partial x}[X(z, \xi), z] \tag{8.19b}$$

$$\dot{H}(z, \xi) = \frac{1}{2H(z, \xi)}\frac{\partial n^2}{\partial z}[X(z, \xi), z] \tag{8.19c}$$

Notice that there is now only one parameter ξ that labels the rays. The equations for the path length L become

$$\dot{L}(z, \xi) = \frac{n^2(X, z)}{H} \tag{8.20}$$

$$L'(z, \xi) = P X' \tag{8.21}$$

where the primes denote derivatives in ξ.

At any fixed z, each ray is fully characterized by its transverse position X and transverse momentum P; knowing where a ray is and in what direction it is propagating is enough to trace it away from this plane. This ray can then be represented by a point in the plane of x versus p. This plane is called *phase space*. The complete ray family is therefore represented by a curve, traced by the points for each ray by varying ξ. This curve is called the *phase-space curve* (PSC) or *Lagrange manifold*. Notice that the integral of Eq. (8.21) gives

$$L(z, \xi_1) - L(z, \xi_0) = \int_{\xi_0}^{\xi_1} P(z, \xi) \frac{\partial X}{\partial \xi}(z, \xi)\, d\xi \tag{8.22}$$

That is, the area under a segment of the PSC equals the difference in optical path length between the rays that correspond to the ends of the PSC segment, as shown in Fig. 8.2. This means that, given the knowledge of the PSC for a given z and the value of L for only one ray, the value of L for all the other rays can be determined. As mentioned earlier, this relation is a consequence of the fact that L corresponds to the optical path length along the rays, measured from a common normal.

For three-dimensional fields, phase space is four-dimensional, since there are two transverse directions and two transverse momenta. The Lagrange manifold is then a two-parameter surface embedded in this four-dimensional space.

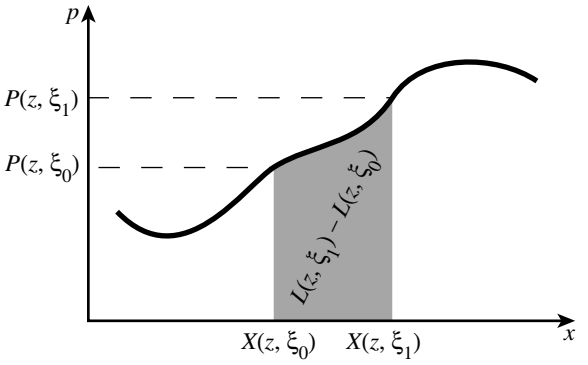


FIGURE 8.2 The phase-space area under any segment of the PSC corresponds to the difference in optical path length L for the corresponding two rays.

8.3 Small-Wavelength Limit in the Position Representation. II: The Transport Equation and the Field Estimate

After we chose ϕ to satisfy the eikonal equation, the remainder of Eq. (8.3) can be written as

$$2\nabla A \cdot \nabla\phi + A\nabla^2\phi + \frac{1}{ik}\nabla^2 A = 0 \tag{8.23}$$

Again, the key to the asymptotic treatment that follows is to regard k as a parameter that takes very large values.

8.3.1 The Debye Series Expansion

The amplitude A is now written as a so-called Debye series of the form

$$A(\vec{r}) = \sum_{j=0}^{\infty} \frac{A_j(\vec{r})}{(ik)^j} \tag{8.24}$$

Then, upon substitution of Eq. (8.24), Eq. (8.23) can be written as

$$2\nabla A_0 \cdot \nabla\phi + A_0\nabla^2\phi + \sum_{j=1}^{\infty} \frac{1}{(ik)^j}\left(2\nabla A_j \cdot \nabla\phi + A_j\nabla^2\phi + \nabla^2 A_{j-1}\right) = 0 \tag{8.25}$$

Since k is used as an asymptotic parameter, the coefficient of each power of k is made to vanish independently. This gives a hierarchy of linked equations for each of the A_j of the form

$$2\nabla A_0 \cdot \nabla\phi + A_0\nabla^2\phi = 0 \tag{8.26a}$$

$$2\nabla A_j \cdot \nabla\phi + A_j\nabla^2\phi = -\nabla^2 A_{j-1}, \qquad j = 1, 2, \ldots \tag{8.26b}$$

8.3.2 The Transport Equation and Its Solution

Equation (8.26*a*) can be solved to find A_0, and in principle each A_j can be found successively in terms of the previous one by solving Eq. (8.26*b*). For sufficiently large k, however, $A \approx A_0$, so only Eq. (8.26*a*) will be considered. Notice that by multiplying both sides by A_0, this equation can be rewritten as

$$\nabla \cdot \left(A_0^2\nabla\phi\right) = 0 \tag{8.27}$$

This expression is known as the *transport equation*. To solve it, consider integrating both sides over the volume occupied by a segment of an

infinitesimally thin bundle of rays $\mathcal{B}$ corresponding to small intervals $\Delta\xi_1$, $\Delta\xi_2$ around a central ray ξ_1, ξ_2, and to z between z_0 and z_1, as shown in Fig. 8.3. By using Gauss' theorem, this volume integral can be reduced to a surface integral, i.e.,

$$\int_{\mathcal{B}} \nabla \cdot \left(A_0^2 \nabla\phi\right) d^3r = \int_{\partial\mathcal{B}} A_0^2 \nabla\phi \cdot d\vec{a} = 0 \tag{8.28}$$

where $\partial\mathcal{B}$ refers to the outer surface of the bundle $\mathcal{B}$, and $d\vec{a}$ is the outward-pointing differential area element. It is easy to see that the only contributions to the surface integral come from the infinitesimal end faces of the bundle, since $d\vec{a}$ is perpendicular to the ray momentum $\nabla\phi$ at the sides of the bundle. Let the infinitesimally small area elements at both ends of the bundle be called Δa_0 and Δa_1, respectively, so Eq. (8.28) can be written as

$$A_0^2[\mathbf{X}(z_0, \boldsymbol{\xi}), z_0] H(z_0, \boldsymbol{\xi}) (-\Delta a_0) + A_0^2[\mathbf{X}(z_1, \boldsymbol{\xi}), z_1] H(z_1, \boldsymbol{\xi}) \, \Delta a_1 = 0 \tag{8.29}$$

where the minus sign in the area element for the first term comes from the fact that $\nabla\phi$ points into $\mathcal{B}$ at the beginning of the bundle segment and out of $\mathcal{B}$ at its end. In getting to this expression, we also used the fact that the z component of $\nabla\phi[\mathbf{X}(z, \boldsymbol{\xi}), z]$ is simply $H(z, \boldsymbol{\xi})$. The intensity of the field is given by $|A|^2 \approx A_0^2$. The product of this intensity and H (which is the refractive index times an obliquity factor) is proportional to the flux density traversing the area element. Therefore

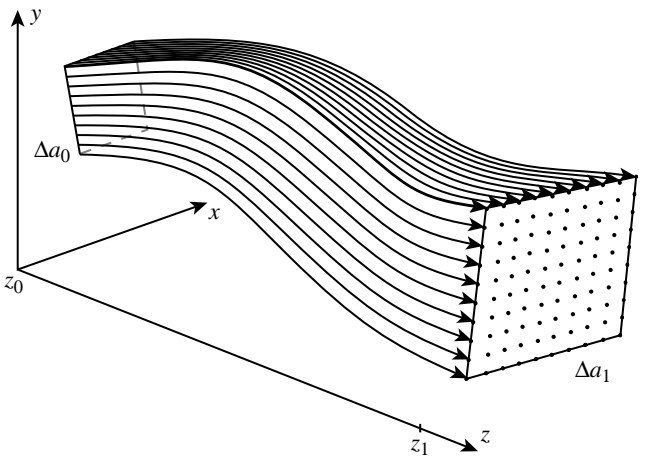


FIGURE 8.3 Volume $\mathcal{B}$, occupied by a segment of an infinitesimally thin bundle of rays.

Eq. (8.29) has an intuitive interpretation: the total flux entering the bundle segment at one end equals the flux exiting at the other end. This means that the rays behave as infinitesimal conduits of power. When the bundle expands, causing its transverse area at the exit to be bigger than that at the entrance, the flux density (and therefore the intensity) becomes smaller in the same proportion, since the conserved power spreads over a larger area.

Now, notice that the area elements are given by

$$\Delta a_j = \left.\frac{\delta(\mathbf{X})}{\delta(\boldsymbol{\xi})}\right|_{z_j} \Delta\xi_1 \Delta\xi_2 = \left.\left(\frac{\partial X}{\partial \xi_1}\frac{\partial Y}{\partial \xi_2} - \frac{\partial X}{\partial \xi_2}\frac{\partial Y}{\partial \xi_1}\right)\right|_{z_j} \Delta\xi_1 \Delta\xi_2 \tag{8.30}$$

where the Jacobian $\delta(\mathbf{X})/\delta(\boldsymbol{\xi})$ is the determinant of the stability matrix $\partial\mathbf{X}/\partial\boldsymbol{\xi}$. By replacing z_1 with z, we can solve Eq. (8.29) for $A_0(\mathbf{X}, z)$ which, after using Eq. (8.30), gives

$$A_0[\mathbf{X}(z, \boldsymbol{\xi}), z] = \sqrt{\frac{H(z_0, \boldsymbol{\xi})}{H(z, \boldsymbol{\xi})}\frac{\delta[\mathbf{X}(z_0, \boldsymbol{\xi})]}{\delta(\boldsymbol{\xi})}\left\{\frac{\delta[\mathbf{X}(z, \boldsymbol{\xi})]}{\delta(\boldsymbol{\xi})}\right\}^{-1}} A_0[\mathbf{X}(z_0, \boldsymbol{\xi}), z_0] \tag{8.31}$$

This expression is the final piece that is needed to estimate the field.

8.3.3 The Field Estimate and Its Problems at Caustics

All the pieces are now put together for the construction of the field estimate. The estimate of the solution to the Helmholtz equation is found in terms of the parameteric equation

$$\begin{aligned} U[\mathbf{X}(z, \boldsymbol{\xi}), z] &\approx A_0[\mathbf{X}(z, \boldsymbol{\xi}), z] \exp[ikL(z, \boldsymbol{\xi})] \\ &= \sqrt{\frac{H(z_0, \boldsymbol{\xi})}{H(z, \boldsymbol{\xi})}\frac{\delta[\mathbf{X}(z_0, \boldsymbol{\xi})]}{\delta(\boldsymbol{\xi})}\left\{\frac{\delta[\mathbf{X}(z, \boldsymbol{\xi})]}{\delta(\boldsymbol{\xi})}\right\}^{-1}} \\ &\quad \times A_0[\mathbf{X}(z_0, \boldsymbol{\xi}), z_0] \exp[ikL(z, \boldsymbol{\xi})] \end{aligned} \tag{8.32}$$

This approximation results from neglecting all terms but the first in the Debye series for A, which is justified by the fact that k is large.

To use the formula in Eq. (8.32), one must determine, from the initial conditions of the field, the initial conditions for the rays. Once this is done, the rays can be traced by using the ray equations, and from there the amplitude and phase of the field can be computed. A distinctive aspect of this formula is that the field information is transmitted through the rays: the value of the field at any point is determined only by the infinitesimal bundle(s) of rays surrounding

the ray(s) passing through this point. That is, if we only integrate the ray equations over a thin bundle of rays, we can estimate the field along this bundle over a long distance, regardless of what the rest of the rays do, as long as rays do not cross. Notice also that if rays from a point source are considered, and the medium is an ABCD system in the paraxial approximation, the formula in Eq. (8.32) can be shown to give the point-spread function in Eq. (1.42) after solving for $\boldsymbol{\xi}$ in terms of the rays' final position.

Notice, however, that the expression for the amplitude of this estimate given in Eq. (8.31) diverges when either of the two following conditions is met:

$$H(z, \boldsymbol{\xi}) = 0 \tag{8.33}$$

$$\frac{\delta[\mathbf{X}(z, \boldsymbol{\xi})]}{\delta(\boldsymbol{\xi})} = 0 \tag{8.34}$$

The first of these conditions happens when rays turn around in z. As mentioned at the outset, it is assumed that this does not occur. (If it does, this problem can be alleviated by using a different ray parameterization.) The second condition, on the other hand, occurs when the bundle collapses, i.e., when the rays that make up the bundle cross. This crossing of contiguous rays is what is known as a *caustic*. While the field at a caustic is indeed large, it is not infinite as Eq. (8.31) predicts. This signals a problem in this formalism, which arises from the fact that in the vicinity of the caustic, the right-hand side of Eq. (8.26*b*) for $j = 1$ becomes large due to the fast variation of A_0. This causes A_1, as well as the rest of the A_j, to be large too, up to a point where the Debye series cannot be approximated by its leading term. The field estimate in Eq. (8.32) therefore breaks down in the vicinity of caustics.

Remarkably, though, if after the caustic the rays become sufficiently spread and uniform, this field estimate becomes valid again. Note, however, that the passage through the caustic gives rise to changes in sign for the Jacobian inside the square root in Eq. (8.31). Due to the square root, these sign changes cause phase shifts that are integer multiples of $\pi/2$, where the integer is called the Maslov or Morse index. The determination of these indices can be complicated,[11,12] as the appropriate sign of the square root must be chosen. (*Note:* This phase shift is consistent with the Gouy phase shift undergone by a focused beam.) In many situations, after a caustic or caustics, there are several bundles of rays occupying the same volume. The field estimate then results from summing their contributions, each with a suitable Maslov phase. (This is one example of how ray optics can account for interference effects.) Also, there might be regions to one side of a caustic where no rays arrive. The simple estimate presented above then offers no access to the field in these regions, suggesting

that the field must vanish there. This is not true, however, since the field is small but not zero in these regions. There are generalizations of the scheme presented here that give accurate estimates in these dark regions by considering rays that leave the real space (the so-called complex rays).[13] This more complicated approach is beyond the scope of this chapter. As will be shown later, the methods described in Secs. 8.4, 8.6, 8.7, 8.8, and 8.9 do lead to nonvanishing field estimates in the dark regions.

The type of ray-based field estimation presented in this section is useful, e.g., for modeling standard imaging systems, as long as it is complemented appropriately by a simple wave-based computation. Consider the case where the object is a point source. Then the two-parameter family of rays corresponds to the rays emanating from this point. The field at the image plane cannot be estimated with this scheme, as the image is a caustic. The field must instead be reconstructed at the exit pupil, where the rays are uniformly spread. Then a wave-based computation must be used for the final stage of free-space propagation from the exit pupil to the image plane. This computation can be performed in a numerically efficient fashion through the use of fast Fourier transforms. An additional advantage of this approach is that the effects of diffraction from the aperture stop are automatically accounted for in this last step by setting the field outside the exit pupil to zero.

8.4 Flux Lines versus Rays

The second approach mentioned at the beginning of Sec. 8.2, which consists of the assumption that both A and ϕ are purely real, is discussed in this section. In this case, Eq. (8.3) can be separated into two equations, corresponding to its real and imaginary parts, which can be written as

$$|\nabla\phi|^2 = n^2 - \frac{\nabla^2 A}{k^2 A} \tag{8.35a}$$

$$\nabla \cdot (A^2 \nabla\phi) = 0 \tag{8.35b}$$

Notice that Eq. (8.35*b*) is identical to Eq. (8.27), except for the fact that it includes the full amplitude A instead of A_0, which is only the leading term in a series. That is, in this case it is not necessary to express A as a Debye series, and the full A can, in principle, be found by integrating Eq. (8.35*b*) over an infinitesimal bundle of trajectories. These trajectories must be found by solving Eq. (8.35*a*) parameterically. The nature of these trajectories becomes apparent from the substitution of

Eq. (8.2) with real A and ϕ into the definition of the field's flux:

$$\vec{F} = \frac{\mathrm{Im}(U^* \nabla U)}{ik} = A^2 \nabla \phi \tag{8.36}$$

That is, $\nabla\phi$ points now in the direction of the local flux. [Notice that Eq. (8.35*b*) is simply the scalar version of Poynting's theorem for a stationary field.] Therefore, the trajectories that result from this approach are the flux lines of the field.

Equation (8.35*a*) strongly resembles the eikonal equation in Eq. (8.4), except for the presence of the correction term $-\nabla^2 A/k^2 A$. Since this correction contains A, Eqs. (8.35*a*) and (8.35*b*) are coupled: one can no longer first solve the eikonal equation to find the rays and then use these rays to solve the transport equation. The correction term is proportional to k^{-2}, so for large k this correction is usually very small, meaning that the flux lines are very similar to the rays almost everywhere. This correction has a similar effect to that of the refractive index: its variation causes the flux lines to bend. This extra bending is negligible except in regions where A varies quickly, e.g., near a focus. There, this term causes the flux lines to deflect away from one another instead of crossing. (It also changes the spacing of the wavefronts, giving rise to the Gouy phase shift.) Therefore, flux lines never cross, and there is always only one at any point in space. It would therefore appear that the estimation of the field in terms of flux lines is a better alternative than the one in terms of rays, since there are no problems with caustics. However, this approach has two disadvantages. First, the fact that Eqs. (8.35*a*) and (8.35*b*) are coupled complicates the determination of the trajectories, which must be found numerically even in very simple cases such as propagation in free space. (A result of this coupling is that, unlike rays, flux lines do not propagate in a mutually independent way.) Second, the method has problems at zeros of the field, since the correction term in Eq. (8.35*a*) can diverge when $A = 0$.

8.5 Analogy with Quantum Mechanics

All the ideas presented so far can also be applied to the study of quantum dynamics.[14] Let us concentrate on the case of nonrelativistic quantum mechanics for a single particle of mass m moving in a potential $V(\vec{r}, t)$. This problem is ruled by the Schrödinger equation

$$i\hbar \frac{\partial \Psi}{\partial t}(\vec{r}, t) = -\frac{\hbar^2}{2m} \nabla^2 \Psi(\vec{r}, t) + V(\vec{r}, t)\Psi(\vec{r}, t) \tag{8.37}$$

where $\hbar$ is the reduced Planck constant. The wavefunction Ψ plays the role of the wave field U. As with the Helmholtz equation, let us

write the wavefunction in terms of a slowly varying amplitude and a rapidly oscillating phase, i.e.,

$$\Psi(\vec{r}, t) = A(\vec{r}, t) \exp\left[\frac{i}{\hbar}\phi(\vec{r}, t)\right] \tag{8.38}$$

Notice that the phase was chosen to be proportional to the inverse of $\hbar$. Therefore, like k for the Helmholtz case, $\hbar^{-1}$ plays the role of the large asymptotic parameter. The resulting approximate results are then valid within the so-called semiclassical regime, i.e., when $\hbar$ is small compared to all other quantities (or variations of quantities) in the problem under study that present the same units (action = length × mass × speed).

After substitution of Eq. (8.38) and reordering, Eq. (8.37) can be written as

$$\left(\frac{\partial\phi}{\partial t} + \frac{|\nabla\phi|^2}{2m} + V\right) - i\hbar\left(\frac{\partial A}{\partial t} + \frac{2\nabla A\cdot\nabla\phi + A\nabla^2\phi}{2m}\right) - \hbar^2\nabla^2 A = 0 \tag{8.39}$$

Again, we face a decision between two approaches:

1. Assume that $\hbar$ is very small, and expand A in a Debye series.
2. Assume that both A and ϕ are real, and separate the real and imaginary parts of Eq. (8.39).

As before, let us consider approach 1 first.

8.5.1 Semiclassical Mechanics

We start by setting the leading term in Eq. (8.39) to zero. This leads to the equation

$$\frac{\partial\phi}{\partial t} + \frac{|\nabla\phi|^2}{2m} + V = 0 \tag{8.40}$$

This is the equation for the action in classical mechanics. Like the eikonal equation, it can be solved parameterically. We start by parameterizing the position as $\vec{r} = \vec{R}(t, \xi_1, \xi_2, \xi_3)$. We then define the momentum $\vec{\mathcal{P}}$ and the Hamiltonian $\mathcal{H}$, respectively, as the spatial and (minus the) temporal derivatives of the action

$$\vec{\mathcal{P}} = \nabla\phi(\vec{R}, t) \tag{8.41}$$

$$\mathcal{H} = -\frac{\partial\phi}{\partial t}(\vec{R}, t) \tag{8.42}$$

and choose for the time derivative of the parameterized position to be proportional to the momentum, i.e.,

$$\vec{\mathcal{P}} = m\dot{\vec{R}} \tag{8.43}$$

where, in this section, the overdot denotes a temporal derivative (since here t plays a role analogous to that of z in the rest of this chapter). The equation for the evolution of the "rays" for this problem results from considering the time derivative of Eq. (8.41) [in a step analogous to that in Eq. (8.9)]:

$$\dot{\vec{\mathcal{P}}} = (\dot{\vec{R}} \cdot \nabla)\nabla\phi + \nabla\frac{\partial\phi}{\partial t} = \left(\frac{\nabla\phi}{m} \cdot \nabla\right)\nabla\phi + \nabla\frac{\partial\phi}{\partial t} = \nabla\left(\frac{\nabla\phi \cdot \nabla\phi}{2m} + \frac{\partial\phi}{\partial t}\right)$$
$$= -\nabla V(\vec{R}, t) \tag{8.44}$$

where Eq. (8.40) was used in the last step. This equation is clearly Newton's second law of motion, so the "rays" of the Schrödinger equation are classical trajectories. In other words, the mathematical relation between wave and ray optics is analogous to that between quantum and classical mechanics. (In fact, it was the analogy between ray optics and classical mechanics that inspired Schrödinger to postulate his wave equation for mechanics.) Like the eikonal, the action can be parameterized as $\mathcal{S} = \phi(\vec{R}, t)$, and its equation of evolution results from considering

$$\dot{\mathcal{S}} = \dot{\vec{R}} \cdot \nabla\phi + \frac{\partial\phi}{\partial t} = \frac{|\vec{\mathcal{P}}|^2}{2m} - V(\vec{R}, t) \tag{8.45}$$

Notice that the right-hand side of this expression is the Lagrangian.[15]

Let us now consider the rest of Eq. (8.39). As mentioned earlier, we propose a Debye expansion for the amplitude of the form

$$A(\vec{r}, t) = \sum_{j=0}^{\infty} (i\hbar)^j A_j(\vec{r}, t) \tag{8.46}$$

After substituting this into Eq. (8.39) (recalling that the first term has been made to vanish), separating the different powers of $\hbar$, and rearranging, we get the equations

$$\frac{\partial A_0^2}{\partial t} + \nabla \cdot \left(A_0^2 \frac{\nabla\phi}{m}\right) = 0 \tag{8.47a}$$

$$\frac{\partial A_j}{\partial t} + \frac{2\nabla A_j \cdot \nabla\phi + A_j\nabla^2\phi}{2m} = -\nabla^2 A_{j-1} \tag{8.47b}$$

for $j \geq 1$. Notice that Eq. (8.47*a*) is a continuity equation. As in the optical case, it can be solved through integration to give

$$A_0[\vec{R}(t, \vec{\xi}), t] = \sqrt{\frac{\delta[\vec{R}(t_0, \vec{\xi})]}{\delta(\vec{\xi})} \left\{ \frac{\delta[\vec{R}(t, \vec{\xi})]}{\delta(\vec{\xi})} \right\}^{-1}} A_0[\vec{R}(t_0, \vec{\xi}), t_0] \quad (8.48)$$

where $\vec{\xi} = (\xi_1, \xi_2, \xi_3)$. These results are the basis of many semiclassical techniques used to understand and model quantum dynamics based on classical mechanics. These techniques include the WKB (or JWKB) method[16] and the Van Vleck-Gutzwiller propagator.[17,18] However, the amplitude estimate in Eq. (8.48) diverges when classical trajectories cross. This problem is analogous to the caustic problem in optics.

8.5.2 Bohmian Mechanics and the Hydrodynamic Model

Now let us consider approach 2, where both A and ϕ are assumed to be real. After simple manipulation, the real and imaginary parts of Eq. (8.39) can be written as

$$\frac{\partial \phi}{\partial t} + \frac{|\nabla \phi|^2}{2m} + V - \hbar^2 \frac{\nabla^2 A}{A} = 0 \quad (8.49a)$$

$$\frac{\partial A^2}{\partial t} + \nabla \cdot \left(A^2 \frac{\nabla \phi}{m} \right) = 0 \quad (8.49b)$$

This form of separating Schrödinger's equation is the basis of Louis deBroglie's and David Bohm's pilot wave interpretation for quantum mechanics.[19] Notice that Eq. (8.49*a*) is almost identical to the classical equation for the action, except for the extra term $-\hbar^2 \nabla^2 A/A$. This term is referred to as the *quantum potential*, and like the last term in Eq. (8.35*a*), it has the effect of steering the trajectories away from the classical ones in order to keep them from crossing. The interpretation of deBroglie and Bohm is that there is a directly undetectable "pilot wave" whose behavior is ruled by Schrödinger's equation and which guides the motion of the detectable particle.

Besides the philosophical interpretation of these results, Eqs. (8.49*a*) and (8.49*b*) serve as the basis for computational methods. This formalism is referred to as the hydrodynamic model[20] since, as seen from Eq. (8.49*b*), the square modulus of the wave function satisfies a continuity equation akin to that of a fluid. However, as in the optical case, the fact that Eqs. (8.49*a*) and (8.49*b*) are coupled makes their solution difficult, both algebraically and computationally, especially when the wave function presents zeros.

8.6 Small-Wavelength Limit in the Momentum Representation

In the asymptotic approach presented in Secs. 8.2 and 8.3, the field estimate at a point is only due to the rays that go through that point. That is, rays are completely local entities. This approach leads to problems at caustics, where there is an infinite density of rays. An alternative asymptotic approach is presented in this section, where instead of working with the field as a function of $\mathbf{r}$, we use its Fourier transform over the transverse coordinates, defined as

$$\tilde{U}(p_x, p_y, z) = \frac{k}{2\pi} \iint U(x, y, z) \exp(-ik\mathbf{x} \cdot \mathbf{p})\, dx\, dy \tag{8.50}$$

where $\mathbf{x} = (x, y)$ and $\mathbf{p} = (p_x, p_y)$. While the derivation presented in the next few pages is in many ways analogous to that in Secs. 8.2 and 8.3, it is appreciably more cumbersome. Those readers who prefer to do so can jump directly to the final result, given in Eq. (8.74).

8.6.1 The Helmholtz Equation in the Momentum Representation

In free space, the transverse Fourier transform in Eq. (8.50) is known as the *angular spectrum representation*.[21] In a smoothly inhomogeneous medium, $\tilde{U}$ satisfies an equation corresponding to the Fourier transformation of both sides of Eq. (8.1):

$$\left[-k^2|\mathbf{p}|^2 + \frac{\partial^2}{\partial z^2} + k^2 n^2 \left(\frac{i}{k} \frac{\partial}{\partial \mathbf{p}}, z \right) \right] \tilde{U} = 0 \tag{8.51}$$

Here a function evaluated at a derivative is to be interpreted in terms of its Taylor expansion:

$$n^2 \left(\frac{i}{k} \frac{\partial}{\partial \mathbf{p}}, z \right) \tilde{U}(\mathbf{p}, z) = \sum_{j=0}^{\infty} \frac{1}{j!} \left(\frac{i}{k} \right)^j \times \left[n^2(\mathbf{x}, z) \left(\frac{\overleftarrow{\partial}}{\partial \mathbf{x}} \cdot \frac{\overrightarrow{\partial}}{\partial \mathbf{p}} \right)^j \tilde{U}(\mathbf{p}, z) \right] \Bigg|_{\mathbf{x}=(0,0)} \tag{8.52}$$

where the arrows indicate the direction in which the derivatives act. Notice that it is assumed here that n^2 is an analytic function and, for convenience, the Taylor expansion is carried out around $\mathbf{x} = (0, 0)$. It turns out, however, that the asymptotic results of this section are independent of the point of expansion, and they hold as long as n is continuous.

We now write $\tilde{U}$ as a slowly varying amplitude times a phase factor:

$$\tilde{U} = B(\mathbf{p}, z)\, \exp[ik\Gamma(\mathbf{p}, z)] \tag{8.53}$$

Notice that the substitution of this form in Eq. (8.52) gives

$$\begin{aligned}
& n^2\left(\frac{i}{k}\frac{\partial}{\partial \mathbf{p}}, z\right)[B\exp(ik\Gamma)] \\
&\quad = \sum_{j=0}^{\infty}\left(\frac{i}{k}\right)^j n^2(\mathbf{x}, z)\left[\frac{1}{j!}\left(ik\frac{\overleftarrow{\partial}}{\partial \mathbf{x}}\cdot\frac{\partial\Gamma}{\partial \mathbf{p}}\right)^j B\right. \\
&\qquad + \frac{1}{(j-1)!}\left(ik\frac{\overleftarrow{\partial}}{\partial \mathbf{x}}\cdot\frac{\partial\Gamma}{\partial \mathbf{p}}\right)^{j-1}\left(ik\frac{\overleftarrow{\partial}}{\partial \mathbf{x}}\cdot\frac{\partial B}{\partial \mathbf{p}}\right) \\
&\qquad + \frac{1}{2(j-2)!}\left(ik\frac{\overleftarrow{\partial}}{\partial \mathbf{x}}\cdot\frac{\partial\Gamma}{\partial \mathbf{p}}\right)^{j-2}\left(ik\frac{\overleftarrow{\partial}}{\partial \mathbf{x}}\cdot\frac{\partial^2 B}{\partial \mathbf{p}\,\partial \mathbf{p}}\cdot\frac{\overleftarrow{\partial}}{\partial \mathbf{x}}\right) \\
&\qquad \left. + \mathcal{O}(k^{j-2})\right]\Bigg|_{\mathbf{x}=(0,0)} \exp(ik\Gamma) \\
&\quad = \left[Bn^2\left(-\frac{\partial\Gamma}{\partial \mathbf{p}}, z\right) + \frac{i}{k}\frac{\partial B}{\partial \mathbf{p}}\cdot\frac{\partial n^2}{\partial \mathbf{x}}\left(-\frac{\partial\Gamma}{\partial \mathbf{p}}, z\right)\right. \\
&\qquad \left. - \frac{iB}{2k}\mathrm{Tr}\left\{\frac{\partial^2 n^2}{\partial \mathbf{x}\,\partial \mathbf{x}}\left(-\frac{\partial\Gamma}{\partial \mathbf{p}}, z\right)\cdot\frac{\partial^2\Gamma}{\partial \mathbf{p}\,\partial \mathbf{p}}\right\} + \mathcal{O}(k^{-2})\right] \\
&\qquad \times \exp(ik\Gamma) \\
&\quad = \left[Bn^2\left(-\frac{\partial\Gamma}{\partial \mathbf{p}}, z\right) + \frac{i}{k}\frac{\partial B}{\partial \mathbf{p}}\cdot\frac{\partial n^2}{\partial \mathbf{x}}\left(-\frac{\partial\Gamma}{\partial \mathbf{p}}, z\right)\right. \\
&\qquad \left. + \frac{iB}{2k}\frac{\partial}{\partial \mathbf{p}}\cdot\frac{\partial n^2}{\partial \mathbf{x}}\left(-\frac{\partial\Gamma}{\partial \mathbf{p}}, z\right) + \mathcal{O}(k^{-2})\right]\exp(ik\Gamma)
\end{aligned} \tag{8.54}$$

where only the two leading orders in powers of k were written explicitly. With this, Eq. (8.51) can be written, after dividing by $-k^2\exp(ik\Gamma)$, as

$$\begin{aligned}
& B\left[|\mathbf{p}|^2 + \left(\frac{\partial\Gamma}{\partial z}\right)^2 - n^2\left(-\frac{\partial\Gamma}{\partial \mathbf{p}}, z\right)\right] \\
&\quad + \frac{1}{ik}\left[2\frac{\partial B}{\partial z}\frac{\partial\Gamma}{\partial z} + B\frac{\partial^2\Gamma}{\partial z^2} + \frac{\partial B}{\partial \mathbf{p}}\cdot\frac{\partial n^2}{\partial \mathbf{x}}\left(-\frac{\partial\Gamma}{\partial \mathbf{p}}, z\right)\right. \\
&\quad \left. + \frac{B}{2}\frac{\partial}{\partial \mathbf{p}}\cdot\frac{\partial n^2}{\partial \mathbf{x}}\left(-\frac{\partial\Gamma}{\partial \mathbf{p}}, z\right)\right] + \mathcal{O}(k^{-2}) = 0
\end{aligned} \tag{8.55}$$

8.6.2 Asymptotic Treatment and Ray Equations

Approaches 1 and 2 mentioned in Sec. 8.2 can be used to separate Eq. (8.55). For approach 2, it is assumed that both B and Γ are real, and Eq. (8.55) is separated into real and imaginary parts. This leads to a momentum-space flux-line formalism where no two trajectories ever have the same optical momentum at a given z, although they can cross freely in position. (For quantum mechanics, the equivalent procedure would lead to a momentum-space Bohmian formalism.) However, except for very simple refractive index distributions, the two resulting equations would be very complicated and would involve terms of many orders in k. This approach is therefore not considered further.

On the other hand, approach 1, which corresponds to the asymptotic treatment, is tractable. We start by expanding B as a Debye series:

$$B(\mathbf{p}, z) = \sum_{j=0}^{\infty} \frac{B_j}{(ik)^j} \tag{8.56}$$

The substitution of this series into Eq. (8.55) gives analogs of the eikonal and transport equations

$$|\mathbf{p}|^2 + \left(\frac{\partial \Gamma}{\partial z}\right)^2 = n^2\left(-\frac{\partial \Gamma}{\partial \mathbf{p}}, z\right) \tag{8.57a}$$

$$\frac{\partial}{\partial z}\left(B_0^2 \frac{\partial \Gamma}{\partial z}\right) + \frac{1}{2}\frac{\partial}{\partial \mathbf{p}} \cdot \left[B_0^2 \frac{\partial n^2}{\partial \mathbf{x}}\left(-\frac{\partial \Gamma}{\partial \mathbf{p}}, z\right)\right] = 0 \tag{8.57b}$$

Equation (8.57a) can be solved by parameterizing $\mathbf{p} = \bar{\mathbf{P}}(z, \boldsymbol{\xi})$ and defining

$$\bar{\mathbf{X}}(z, \boldsymbol{\xi}) = -\frac{\partial \Gamma}{\partial \mathbf{p}}(\bar{\mathbf{P}}, z) \tag{8.58}$$

$$\bar{H}(z, \boldsymbol{\xi}) = \frac{\partial \Gamma}{\partial z}(\bar{\mathbf{P}}, z) \tag{8.59}$$

Notice that the derivative with respect to z of both sides of Eq. (8.58) gives

$$\dot{\bar{\mathbf{X}}} = -\left(\dot{\bar{\mathbf{P}}} \cdot \frac{\partial}{\partial \mathbf{p}}\right)\frac{\partial \Gamma}{\partial \mathbf{p}} - \frac{\partial^2 \Gamma}{\partial \mathbf{p}\, \partial z} \tag{8.60}$$

To eliminate the cross-derivative term, let us consider the vector derivative with respect to the transverse momentum of both sides

of Eq. (8.57*a*), i.e.,

$$2\mathbf{p} + 2\frac{\partial \Gamma}{\partial z}\frac{\partial^2 \Gamma}{\partial \mathbf{p}\,\partial z} = -\left(\frac{\partial n^2}{\partial \mathbf{x}} \cdot \frac{\partial}{\partial \mathbf{p}}\right)\frac{\partial \Gamma}{\partial \mathbf{p}} \tag{8.61}$$

The evaluation of this expression at $\mathbf{p} = \bar{\mathbf{P}}(z, \boldsymbol{\xi})$ gives, after reordering and the use of Eq. (8.59),

$$\frac{\bar{\mathbf{P}}}{\bar{H}} = -\frac{1}{2\bar{H}}\left(\frac{\partial n^2}{\partial \mathbf{x}} \cdot \frac{\partial}{\partial \mathbf{p}}\right)\frac{\partial \Gamma}{\partial \mathbf{p}} - \frac{\partial^2 \Gamma}{\partial \mathbf{p}\,\partial z} \tag{8.62}$$

From the comparison of Eqs. (8.60) and (8.62), we see that it is convenient to choose

$$\dot{\mathbf{X}} = \frac{\bar{\mathbf{P}}}{\bar{H}} \tag{8.63a}$$

$$\dot{\bar{\mathbf{P}}} = \frac{1}{2\bar{H}}\frac{\partial n^2}{\partial \mathbf{x}}(\bar{\mathbf{X}}, z) \tag{8.63b}$$

These equations are identical to Eqs. (8.18*a*) and (8.18*b*). Also, notice that the substitution of Eqs. (8.58) and (8.59) into Eq. (8.57*a*) evaluated at $\mathbf{p} = \bar{\mathbf{P}}(z, \boldsymbol{\xi})$ implies that

$$\bar{H} = \sqrt{n^2(\bar{\mathbf{X}}, z) - |\bar{\mathbf{P}}|^2} \tag{8.64}$$

so, remarkably, the parameterized trajectories that result from the asymptotic treatment in the momentum representation are the standard rays. From now on, the bars over $\mathbf{X}$, $\mathbf{P}$, and H are dropped.

The phase function is again obtained parameterically. Let us define

$$T(z, \boldsymbol{\xi}) = \Gamma(\mathbf{P}, z) \tag{8.65}$$

The evolution of this function results from considering its derivative with respect to z:

$$\dot{T} = \frac{\partial \Gamma}{\partial \mathbf{p}} \cdot \dot{\mathbf{P}} + \frac{\partial \Gamma}{\partial z} = -\frac{1}{2H}\mathbf{X} \cdot \frac{\partial n^2}{\partial \mathbf{x}}(\mathbf{X}, z) + H \tag{8.66}$$

where Eqs. (8.58), (8.59), and (8.63*b*) were used in the second step. The relation between the values of T for contiguous rays is found similarly by taking the partial derivative with respect to ξ_j of both sides of Eq. (8.65):

$$\frac{\partial T}{\partial \xi_j} = \frac{\partial \Gamma}{\partial \mathbf{p}} \cdot \frac{\partial \mathbf{P}}{\partial \xi_j} = -\mathbf{X} \cdot \frac{\partial \mathbf{P}}{\partial \xi_j} \tag{8.67}$$

This relation, combined with Eq. (8.13), implies that

$$T(z, \boldsymbol{\xi}) = L(z, \boldsymbol{\xi}) - \mathbf{X}(z, \boldsymbol{\xi}) \cdot \mathbf{P}(z, \boldsymbol{\xi}) \tag{8.68}$$

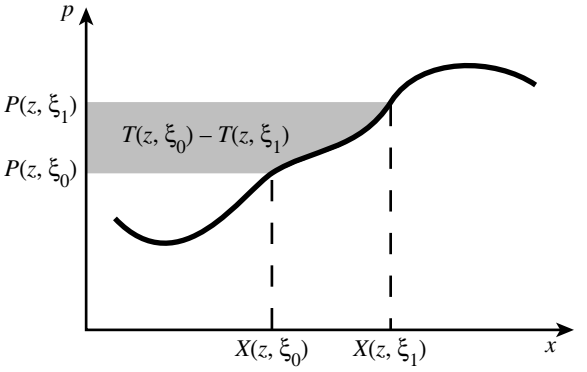


FIGURE 8.4 The phase-space area between the p axis and any segment of the PSC corresponds to the difference in optical path length T for the corresponding two rays.

(Notice that an additive constant could have been included in this expression. However, this constant can be absorbed by L.) It is easy to show that Eqs. (8.66) and (8.68) are consistent with Eq. (8.12*c*).

As for L in the position representation, the geometrical interpretation of the phase T as an area in phase space is easier to understand in the case of two-dimensional propagation, where only one parameter ξ labels the rays. Equation (8.67) then becomes

$$T'(z, \xi) = -XP' \tag{8.69}$$

where the primes denote derivatives in ξ. It can be seen from Eq. (8.69) that the area to the left of a segment of the PSC (up to the p axis) corresponds to the difference in T for the corresponding end rays, as shown in Fig. 8.4, so the knowledge of T for one ray and of the PSC is sufficient to find T for any other ray.

8.6.3 Transport Equation in the Momentum Representation

Now, to solve the momentum transport equation given by Eq. (8.57*b*), let us define the differential operator in the mixed space $(\mathbf{p}, z)$ as

$$\tilde{\nabla} = \left(\frac{\partial}{\partial \mathbf{p}}, \frac{\partial}{\partial z} \right) \tag{8.70}$$

Let us also define the vector

$$\vec{\Theta} = (H\dot{\mathbf{P}}, H) = H\frac{\partial}{\partial z}(\mathbf{P}, z) \tag{8.71}$$

With this, the transport equation in Eq. (8.57*b*) can be written as

$$\tilde{\nabla} \cdot \left(B_0^2 \vec{\Theta}\right) = 0 \tag{8.72}$$

This equation can be solved by integrating it over the volume in the $(\mathbf{p}, z)$ space occupied by an infinitesimal bundle of rays between z_0 and z, and using Gauss' theorem. Equation (8.71) states that $\vec{\Theta}$ is locally parallel to the rays in this space, so the contributions to the surface integral from the sides of the bundle vanish, as in the position representation case. By following steps analogous to those in Sec. 8.3, we find

$$B_0[\mathbf{P}(z, \boldsymbol{\xi}), z] = \sqrt{\frac{H(z_0, \boldsymbol{\xi})}{H(z, \boldsymbol{\xi})} \frac{\delta[\mathbf{P}(z_0, \boldsymbol{\xi})]}{\delta(\boldsymbol{\xi})} \left\{\frac{\delta[\mathbf{P}(z, \boldsymbol{\xi})]}{\delta(\boldsymbol{\xi})}\right\}^{-1}} B_0[\mathbf{P}(z_0, \boldsymbol{\xi}), z_0] \tag{8.73}$$

Notice that this solution has problems when the Jacobian between braces vanishes. This happens when contiguous rays in the family have the same transverse momentum. (For a homogeneous medium, this means that the rays are locally parallel.) That is, the field estimate that results from this derivation also has problems, but these are different from those for the estimate found in the position representation, associated with caustics. The location of these new problems, i.e., the places where contiguous rays have the same momentum, are called *momentum caustics*. As in the case of the amplitude of the position representation estimate, one must be careful when choosing the sign of the square root in Eq. (8.73).

8.6.4 Field Estimate

The field estimate is obtained by approximating $B \approx B_0$, that is, $\tilde{U}(\mathbf{P}, z) \approx B_0 \exp(ikT)$. To obtain the field in the position representation, the inverse Fourier transform of this estimate must be taken. Because $\mathbf{p}$ is parameterized, the Fourier transform integral must be done parameterically by inserting a Jacobian factor:

$$\begin{aligned} U(\vec{r}) &\approx \frac{k}{2\pi} \iint B_0(\mathbf{P}, z) \exp(ikT) \exp(ik\mathbf{x} \cdot \mathbf{P}) \frac{\delta(\mathbf{P})}{\delta(\boldsymbol{\xi})} d\xi_1 \, d\xi_2 \\ &= \frac{k}{2\pi} \iint B_0[\mathbf{P}(z_0, \boldsymbol{\xi}), z_0] \sqrt{\frac{H(z_0, \boldsymbol{\xi})}{H(z, \boldsymbol{\xi})} \frac{\delta[\mathbf{P}(z_0, \boldsymbol{\xi})]}{\delta(\boldsymbol{\xi})} \frac{\delta[\mathbf{P}(z, \boldsymbol{\xi})]}{\delta(\boldsymbol{\xi})}} \\ &\quad \times \exp(ik\{L(z, \boldsymbol{\xi}) + [\mathbf{x} - \mathbf{X}(z, \boldsymbol{\xi})] \cdot \mathbf{P}(z, \boldsymbol{\xi})\}) d\xi_1 \, d\xi_2 \end{aligned} \tag{8.74}$$

Notice that the contribution for each ray now extends over all the configuration space. That is, in this picture, rays do not contribute to the field as infinitesimally thin conduits of power. Instead, they are infinitely extended waves that interfere to make up the wave field. In fact, for the case of homogeneous media, these waves are plane waves. This estimate has no problems at position caustics (focal points), but fails near momentum caustics.

While in this section we considered Fourier transforms over both transverse coordinates, it is also possible to find estimates where Fourier transforms over only one transverse coordinate are performed. These could be useful in problems with specific asymmetries, or in the implementation of the method discussed next.

8.7 Maslov's Canonical Operator Method

The ray-based schemes discussed in previous sections present problems at either position or momentum caustics. These problematic situations have geometrical interpretations in terms of phase space. Again, for simplicity, let us consider the case of two-dimensional propagation, where there is only one transverse position and one transverse momentum, so that phase space is a plane. In this case, the formulas for the amplitudes of the estimates in Eqs. (8.31) and (8.73) become, respectively,

$$A_0[X(z,\xi),z] = \sqrt{\frac{H(z_0,\xi)}{H(z,\xi)}\frac{X'(z_0,\xi)}{X'(z,\xi)}}\,A_0[X(z_0,\xi),z_0] \quad (8.75a)$$

$$B_0[X(z,\xi),z] = \sqrt{\frac{H(z_0,\xi)}{H(z,\xi)}\frac{P'(z_0,\xi)}{P'(z,\xi)}}\,B_0[P(z_0,\xi),z_0] \quad (8.75b)$$

The field estimate resulting from using the position-dependent approach fails at caustics, i.e., when $X' = 0$. In phase space, caustics correspond to segments of the PSC that are locally vertical (see Fig. 8.5). On the other hand, the momentum-representation-based estimate fails at momentum caustics, when $P' = 0$, i.e., at segments of the PSC that are locally horizontal. One could formulate field estimates based on other representations associated, e.g., with a fractional Fourier transform over the transverse variable of the field.[22] These field estimates would be well behaved at both position and momentum caustics, but would fail around rays associated with segments of the PSC with a given inclination (depending on the degree of the fractional Fourier transform).

When a PSC is sufficiently complicated, the caustic problems are unavoidable, regardless of representation. Based on this fact, Maslov[11]

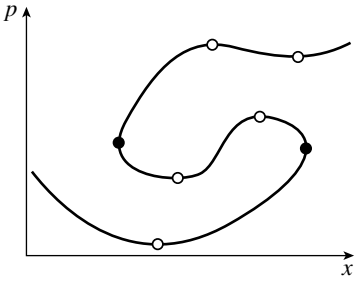


FIGURE 8.5 Vertical segments (indicated by black dots) of the PSC correspond to position caustics, while horizontal segments (indicated by open dots) correspond to momentum caustics.

proposed a scheme where the PSC is subdivided into several segments, each free of at least one type of caustic (position or momentum). A small transition region is left between the segments to avoid errors introduced by the abrupt cuts. A field estimate is then performed for each individual segment, using the appropriate prescription. The total estimate is found by adding the contributions due to all the segments. Notice that one must be careful in choosing the correct Maslov index phase for each contribution. This approach is known as Maslov's canonical operator method, or simply Maslov's method.[5,11,12] Its implementation can be quite complicated, however, especially for three-dimensional fields where instead of a PSC one has a two-dimensional Lagrange manifold embedded in a four-dimensional phase space. In this case, the global estimate can involve contributions of not only the position- and momentum-representation-based estimates, but also of those in mixed representations mentioned at the end of Sec. 8.6.

8.8 Gaussian Beams and Their Sums

In this section, a different scheme for connecting rays and waves is discussed. Here, field contributions with finite effective extent in both position and momentum are considered. For simplicity, the analysis is performed in two-dimensional space.

8.8.1 Parabasal Gaussian Beams

According to the uncertainty relation, a field contribution cannot be simultaneously arbitrarily localized in position and direction (i.e., momentum), since the product of the rms widths in these two

representations must be equal to or greater than $1/(2k)$. Since the minimum product of widths is achieved by Gaussians, field contributions with transverse Gaussian amplitude profile are considered here. Let $X(z)$ and $P(z)$ be the centroids in x and p of such a field contribution. This contribution, called a Gaussian beam, can then be written as

$$U_G(x, z) = u(z) \exp\left[-\frac{(x - X)^2}{2w^2} \right] \exp\{ik[L(z) + P(x - X)]\} \quad (8.76)$$

where $u(z)$ is a complex amplitude and $L(z)$ is a phase accumulated under propagation. In what follows, it is shown that the beam centroids X and P evolve according to geometrical optics.

The transverse rms width of the Gaussian beam in Eq. (8.76) is $w/\sqrt{2}$. It is easy to show that the Fourier transform in x of this beam is indeed a Gaussian in p centered at P, with rms width equal to $1/(\sqrt{2}kw)$. Since the case of large k is considered, we choose $w = 1/\sqrt{k\gamma}$, where γ has units of inverse length. This way, the width of the beam is proportional to $1/\sqrt{k}$ in both the position and momentum representations, leading to comparable levels of localization in phase space in the x and p directions. The Gaussian beam in Eq. (8.76) can then be written as

$$U_G(x, z) = u(z) g_\gamma(x, z) \quad (8.77)$$

where

$$g_\gamma(x, z) = \exp\left(-\frac{k\gamma}{2}[x - X(z)]^2 + ik\{L(z) + P(z)[x - X(z)]\} \right) \quad (8.78)$$

Of course, the real part of γ must be positive.

The next step is to substitute U_G into the Helmholtz equation. First, the second partial derivatives of U_G can be found to be

$$\frac{\partial^2 U_G}{\partial x^2} = [-k\gamma + k^2(iP - \gamma\Delta)^2] u g_\gamma \quad (8.79a)$$

$$\begin{aligned} \frac{\partial^2 U_G}{\partial z^2} = \Bigg\{ & \ddot{u} + \dot{u}k[2iH + 2(\gamma\dot{X} + i\dot{P})\Delta - \dot{\gamma}\Delta^2] \\ & + uk\left[i\dot{H} - (\gamma\dot{X} + i\dot{P})\dot{X} + (\gamma\ddot{X} + i\ddot{P} + 2\dot{\gamma}\dot{X})\Delta - \ddot{\gamma}\frac{\Delta^2}{2} \right] \\ & + uk^2\left[iH + (\gamma\dot{X} + i\dot{P})\Delta - \dot{\gamma}\frac{\Delta^2}{2} \right]^2 \Bigg\} g_\gamma \end{aligned} \quad (8.79b)$$

where $\Delta = x - X$ and $H = \dot{L} - P\dot{X}$. Also, it is convenient to expand the refractive index in a Taylor series around the ray position as

$$n^2(x, z) = \sum_{j=0}^{\infty} \frac{1}{j!} \frac{\partial^j n^2}{\partial x^j}(X, z)\Delta^j \tag{8.80}$$

Equations (8.79*a*), (8.79*b*), and (8.80) are now substituted into the Helmholtz equation for U_G. The result can be grouped in powers of k, with each coefficient itself being separated into powers of Δ:

$$[k^2(C_{20}+C_{21}\Delta+C_{22}\Delta^2+\cdots)+k(C_{10}+C_{11}\Delta+\cdots)+C_{00}]\, g_\gamma = 0 \tag{8.81}$$

where the first few coefficients of each subseries are

$$C_{20} = u[n^2(X, z) - P^2 - H^2] \tag{8.82a}$$

$$C_{21} = u\left[\frac{\partial n^2}{\partial x}(X, z) - 2H\dot{P} + 2i\gamma(H\dot{X} - P)\right] \tag{8.82b}$$

$$C_{22} = u\left[\frac{1}{2}\frac{\partial^2 n^2}{\partial x^2}(X, z) + \gamma^2 + (\gamma\dot{X} + i\dot{P})^2 - iH\dot{\gamma}\right] \tag{8.82c}$$

$$C_{10} = 2i\dot{u}H + u(-\gamma - \gamma\dot{X}^2 - i\dot{X}\dot{P} + i\dot{H}) \tag{8.82d}$$

$$C_{11} = 2\dot{u}(\gamma\dot{X} + i\dot{P}) + u(\gamma\ddot{X} + i\ddot{P} + 2\dot{\gamma}\dot{X}) \tag{8.82e}$$

$$C_{00} = \ddot{u} \tag{8.82f}$$

We now assume that k is large, and we go on to perform the asymptotic treatment. Notice, however, that the coefficients of each power of k involve terms with different powers of Δ, which cannot be mixed to lead to an asymptotic constraint, because Δ depends on the spatial variable x. Since g_γ is a Gaussian of width $1/\sqrt{k\gamma}$ centered at $x = X$, and $\Delta = x - X$, the importance of each term in Eq. (8.81) decreases with increasing powers of Δ. The main contribution, then, is the one with the largest power of k and the smallest power of Δ, or k^2C_{20}. As can be seen from Eq. (8.82*a*), this contribution vanishes if X, P, and H are chosen to follow the rules of ray optics, i.e., if the two-dimensional version of Eq. (8.16) is satisfied. The ray propagation equations for X and P, given in Eqs. (8.19*a*) and (8.19*b*), are also found from setting to zero the next contribution in importance, or $k^2C_{21}\Delta$, as can be easily seen from Eq. (8.82*b*). That is, as in the position- and momentum-based derivations presented earlier, forcing the leading orders of the asymptotic form of the wave equation to vanish leads to the laws of ray optics.

In analogy with those derivations, it might be expected that the amplitude u should be expanded into a Debye series and that this would lead to a hierarchy of equations. However, given that the coefficients of different powers of both k and Δ must be made to vanish independently, this would lead to a set of mutually inconsistent equations. For this reason, the strategy employed here is to leave u as one function. The equations that describe its evolution as well as that of γ follow from considering only the next two most significant terms in Eq. (8.80), i.e., from forcing C_{22} and C_{10} to vanish. The resulting equations are

$$\begin{aligned}\dot{\gamma} &= -\frac{i}{H}\left[\frac{1}{2}\frac{\partial^2 n^2}{\partial x^2}(X,z) + \gamma^2 + (\gamma\dot{X} + i\dot{P})^2\right] \\ &= -\frac{i}{H}\left\{\gamma^2\frac{n^2(X,z)}{H^2} + i\gamma\frac{P}{H^2}\frac{\partial n^2}{\partial x}(X,z) + \frac{1}{2}\frac{\partial^2 n^2}{\partial x^2}(X,z)\right. \\ &\quad \left. - \left[\frac{1}{2H}\frac{\partial n^2}{\partial x}(X,z)\right]^2\right\} \qquad (8.83a)\end{aligned}$$

$$\begin{aligned}\dot{u} &= \frac{u}{2H}[\dot{X}\dot{P} - \dot{H} - i\gamma(1+\dot{X}^2)] \\ &= \frac{u}{4H^3}\left[P\frac{\partial n^2}{\partial x}(X,z) - H\frac{\partial n^2}{\partial z}(X,z) - 2i\gamma n^2(X,z)\right] \qquad (8.83b)\end{aligned}$$

Equation (8.83*a*) is a nonlinear first-order differential equation of the Riccati type, which can be expressed in terms of the solution to a second-order linear differential equation.[23] Equation (8.83*b*) requires the use of the solution of Eq. (8.83*a*). These equations do not lead to closed-form solutions except for certain simple refractive index distributions such as free space, where the solutions are the standard paraxial Gaussian beams. In three dimensions, these equations are more complicated, since γ must be replaced by a 2×2 matrix.

Setting the remaining coefficients to zero leads to constraints that are inconsistent with the ones found earlier. Therefore, the fields U_G that result from the substitution of the solutions of Eqs. (8.83*a*) and (8.83*b*) into Eq. (8.77) are only approximate solutions to the Helmholtz equation. These beams are sometimes called *parabasal* Gaussian beams, as they are the result of an expansion around a "base" ray with phase-space coordinates (X, P).

8.8.2 Sums of Gaussian Beams

The parabasal propagation of Gaussian beams can be calculated through the methods outlined earlier. A large body of work has been produced in the last few decades with the goal of modeling the propagation of arbitrary fields by expressing them as superpositions of

Gaussian beams. These schemes are known collectively as Gaussian beam summation methods. They were proposed independently in many areas starting with geophysics, for the description of seismic waves.[24–28] Similar methods followed for quantum mechanics[29–31] (where the beams are referred to as Gaussian wave packets) and electromagnetic waves[32–37] (particularly in optics and radio science). One of the main advantages of this approach is that it is free of problems at caustics.

Gaussian beam summation methods rely on the expression of an initial field as a Gabor representation,[38] i.e., as a weighted superposition of Gaussian functions of a given width, with different locations and linear phase factors. These superpositions can be discrete or continuous. A discrete Gabor representation, for example, allows us to write a function $f(x)$ in the form

$$f(x) = \sum_{m,n} a_{m,n} \exp\left[\frac{-k\gamma_0(x - m\Delta_X)^2}{2} + ikn\Delta_P(x - m\Delta_X)\right] \quad (8.84)$$

Here, the sampling spacings Δ_X and Δ_P must satisfy the relation $\Delta_X\Delta_P \leq 2\pi/k$, where the basis is complete if the equality holds and is overcomplete otherwise. (The standard complete Gabor basis results from choosing $\Delta_X = \sqrt{2\pi/k\gamma_0}$ and $\Delta_P = \sqrt{2\pi\gamma_0/k}$.) This basis of Gaussians is not orthogonal, so the expansion coefficients $a_{m,n}$ must be found through the use of a biorthogonal basis.[32,33] Each Gaussian roughly occupies a phase-space area of size $\sqrt{2\pi/k\gamma_0}$ by $\sqrt{2\pi\gamma_0/k}$ centered at $(x, p) = (m\Delta_X, n\Delta_P)$, so the Gabor representation can be thought of as a subdivision of phase space into a Cartesian grid of cells whose size is smaller than or equal to the minimum-uncertainty phase-space area. The procedure for the propagation of a field $U(x, z)$ in two dimensions is as follows: First the coefficients $a_{m,n}$ for the specified initial field $U(x, 0)$ are found. Then each Gaussian is propagated as a Gaussian beam with $u(0) = a_{m,n}$, $\gamma(0) = \gamma_0$, $X(0) = m\Delta_X$, $P(0) = n\Delta_P$, and $L(0) = 0$. Finally, the field away from the initial plane is approximated as the sum of the propagated Gaussian beams. This method relies on the validity of the parabasal approximation for each beam.

Continuous Gaussian beam summation methods result from expressing the field as a continuous superposition of Gaussian beams. These can involve all initial positions and directions, although then the superposition is not unique, since the continuous set of Gaussians constitutes an overcomplete basis. Another form of continuous superposition can be used for fields associated with a known initial Lagrange manifold. In this case, the initial field is not composed of all possible Gaussians but only of those whose central initial position and direction fall within this Lagrange manifold. For fields in two

dimensions, the corresponding field estimate then takes the form of an integral over the PSC

$$U(x, z) \approx \int u(z, \xi) g_{\gamma(z,\xi)}(x, z, \xi)\, d\xi \tag{8.85}$$

where

$$g_\gamma(x, z, \xi) = \exp\left(-\frac{k\gamma}{2}[x - X(z, \xi)]^2 + ik\{L(z, \xi) + P(z, \xi)[x - X(z, \xi)]\}\right) \tag{8.86}$$

and $u(0, \xi)$ and $\gamma(0, \xi)$ must be chosen so that the superposition matches the initial field at $z = 0$. Again, this method is valid as long as the parabasal approximation holds for each independent Gaussian beam.

8.9 Stable Aggregates of Flexible Elements

A different approach for building wave field estimates based on rays through the superposition of Gaussian contributions is now discussed. Like Eq. (8.85), this framework, referred to as *stable aggregates of flexible elements* (SAFE),[39–43] takes the form of a continuous superposition of Gaussian components around the rays in the Lagrange manifold. However, in SAFE, the Gaussian contributions are not independently propagating parabasal Gaussian beams; rather, they are interrelated contributions, where γ is not constrained to have a specific dependance on z or ξ:

$$U(x, z) = \int u(z, \xi) g_\gamma(x, z, \xi)\, d\xi \tag{8.87}$$

In what follows, it is assumed for simplicity that γ is a real and positive constant, although the results remain valid if this parameter has an imaginary part and/or is allowed to vary slowly with z, ξ, or even x.

8.9.1 Derivation of the Estimate

To find the specific form of this estimate, we substitute Eq. (8.87) into Eq. (8.1). By following the same steps as in Sec. 8.8, we obtain a result that is identical to that in Eq. (8.81), except for the presence of an integral in ξ:

$$\int [k^2(C_{20} + C_{21}\Delta + C_{22}\Delta^2 + \cdots) + k(C_{10} + C_{11}\Delta) + C_{00}]\, g_\gamma\, d\xi = 0 \tag{8.88}$$

where the functional coefficients C_{mn} are those in Eqs. (8.82*a*) through (8.82*f*), with all derivatives of γ set to zero.

The main feature of SAFE is that the Gaussians are not independent beams but interrelated contributions. This is due to the presence of the integral in Eq. (8.88), which plays a fundamental role in the derivation. The key for this is a trick based on the fact that the derivative of g_γ with respect to ξ is given by

$$g'_\gamma = k[\gamma X'\Delta + i(L' - PX' + P'\Delta)]g_\gamma = k(\gamma X' + iP')\Delta g_\gamma \qquad (8.89)$$

where, in the last step, we chose L to be the area under the PSC, so that Eq. (8.21) is satisfied. Equation (8.89) can be written as

$$\Delta g_\gamma = \frac{g'_\gamma}{k\mathcal{Y}'} \qquad (8.90)$$

where the shorthand $\mathcal{Y} = \gamma X + iP$ is used in what follows. By using this expression, a factor of Δ multiplying g_γ can be removed at the cost of turning g_γ into g'_γ (and including an extra factor of $1/k\mathcal{Y}'$). The ξ derivative in g_γ can then be removed through integration by parts. This process can be repeated to remove higher powers of Δ in the form

$$\begin{aligned}\int k^m C_{mn}\Delta^n g_\gamma\, d\xi &= \int k^{m-1}\frac{C_{mn}}{\mathcal{Y}'}\Delta^{n-1} g'_\gamma\, d\xi = -\int k^{m-1}\left(\frac{C_{mn}}{\mathcal{Y}'}\Delta^{n-1}\right)' \\ &\quad \times g_\gamma\, d\xi \\ &= \int k^{m-1}\left[C_{mn}\frac{X'}{\mathcal{Y}'}\Delta^{n-2} - \left(\frac{C_{mn}}{\mathcal{Y}'}\right)'\Delta^{n-1}\right] g_\gamma\, d\xi\end{aligned} \qquad (8.91)$$

where the integrated terms resulting from the integration by parts are dropped by assuming that the magnitude of the integrands goes to zero at the limits. By using this trick repeatedly, we can rewrite Eq. (8.88) as

$$\int \left\{k^2 C_{20} + k\left[\frac{X'}{\mathcal{Y}'}C_{22} + C_{10} - \left(\frac{C_{21}}{\mathcal{Y}'}\right)'\right] + \mathcal{O}(k^0)\right\} g_\gamma\, d\xi = 0 \qquad (8.92)$$

As for the Gaussian beams, setting the leading term, $k^2 C_{20}$, to zero leads to the laws of ray optics, which also make C_{21} vanish. Then forcing the remaining part of the next term to vanish amounts to choosing u so that $X'C_{22} + \mathcal{Y}'C_{10} = \mathcal{O}(k^{-1})$. The derivation that follows requires a few steps. Let us start by substituting Eqs. (8.82*c*) and (8.82*d*)

(recalling that $\dot{\gamma} = 0$) into this expression:

$$\begin{aligned} X'C_{22} + \mathcal{Y}'C_{10} &= uX'\left(\frac{1}{2}\frac{\partial^2 n^2}{\partial x^2} + \gamma^2 + \dot{\mathcal{Y}}^2\right) \\ &\quad + \mathcal{Y}'[2i\dot{u}H + u(-\gamma - \dot{\mathcal{Y}}\dot{X} + i\dot{H})] \\ &= u\left[\left(\frac{1}{2}\frac{\partial n^2}{\partial x}\right)' + \gamma(\gamma X' - \mathcal{Y}') + \dot{\mathcal{Y}}(\dot{\mathcal{Y}}X' - \dot{X}\mathcal{Y}')\right] \\ &\quad + i\mathcal{Y}'(2\dot{u}H + u\dot{H}) \end{aligned} \tag{8.93}$$

By using Eq. (8.19*b*) as well as the fact that $\mathcal{Y} = \gamma X + iP$, this expression can be simplified to

$$X'C_{22} + \mathcal{Y}'C_{10} = u[(H\dot{P})' - i\gamma P' + i\dot{\mathcal{Y}}(\dot{P}X' - \dot{X}P')] + i\mathcal{Y}'(2\dot{u}H + u\dot{H}) \tag{8.94}$$

By noticing that

$$H' = (\sqrt{n^2(X,z) - P^2})' = \frac{1}{2H}\left(\frac{\partial n^2}{\partial x}X' - 2PP'\right) = \dot{P}X' - \dot{X}P' \tag{8.95}$$

where Eqs. (8.19*a*) and (8.19*b*) were used in the last step, Eq. (8.94) can be simplified further as

$$\begin{aligned} X'C_{22} + \mathcal{Y}'C_{10} &= u[(H\dot{P} - i\gamma P)' + i\dot{\mathcal{Y}}H'] + i\mathcal{Y}'(2\dot{u}H + u\dot{H}) \\ &= iu[\dot{\mathcal{Y}}H' - (\gamma H\dot{X} + iH\dot{P})'] + i\mathcal{Y}'(2\dot{u}H + u\dot{H}) \\ &= iu[\dot{\mathcal{Y}}H' - (H\dot{\mathcal{Y}})'] + i\mathcal{Y}'(2\dot{u}H + u\dot{H}) \\ &= -iuH\dot{\mathcal{Y}}' + i\mathcal{Y}'(2\dot{u}H + u\dot{H}) \\ &= 2i\sqrt{H\mathcal{Y}'^3}\frac{\partial}{\partial z}\left(u\sqrt{\frac{H}{\mathcal{Y}'}}\right) \end{aligned} \tag{8.96}$$

where Eq. (8.19*a*) was used in the second step. Therefore, Eq. (8.92) is satisfied up to the two leading orders if the result of Eq. (8.96) is asymptotically negligible. This can be enforced by writing u as a Debye series of the form

$$u = \sqrt{\frac{k}{2\pi}}\sqrt{\frac{\mathcal{Y}'(z,\xi)}{H(z,\xi)}}\sum_{j=0}^{\infty}\frac{a_j}{(ik)^j} \tag{8.97}$$

where the dominant term of the sum, that is, $a_0(\xi)$, is independent of z, and the constant factor in front was added for convenience. That is, for SAFE, the transport equation takes the simple form $\dot{a_0} = 0$.

The field can be estimated by approximating $a \approx a_0$, giving

$$\begin{aligned} U(\vec{r}) &\approx U_\gamma(\vec{r}) \\ &= \sqrt{\frac{k}{2\pi}} \int a_0(\xi) \sqrt{\frac{\gamma X' + iP'}{H}} \\ &\quad \times \exp\left\{ -\frac{k\gamma}{2}(x - X)^2 + ik[L + (x - X)P] \right\} d\xi \end{aligned} \tag{8.98}$$

This result is SAFE's basic field estimate. Like the Gaussian beam summation methods, this estimate does not fail at caustics of any kind. Its only divergence occurs when rays turn around in z, that is, if H vanishes. Unlike the Gaussian beam summation methods, this estimate does not depend on the parabasal approximation. In fact, its results have been shown[43] to remain valid beyond the point at which the parabasal approximation fails, and the corresponding continuous Gaussian beam superposition [as given in Eq. (8.85)] breaks down. The generalization of this result to three dimensions is straightforward:

$$\begin{aligned} U(\vec{r}) \approx U_\gamma(\vec{r}) = \frac{k}{2\pi} \iint a_0(\boldsymbol{\xi}) \sqrt{\frac{1}{H} \frac{\delta(\boldsymbol{\gamma} \cdot \mathbf{X} + i\mathbf{P})}{\delta(\boldsymbol{\xi})}} \\ \times \exp\left\{ -\frac{k}{2}(\mathbf{x} - \mathbf{X}) \cdot \boldsymbol{\gamma} \cdot (\mathbf{x} - \mathbf{X}) + ik[L + (\mathbf{x} - \mathbf{X}) \cdot \mathbf{P}] \right\} d\xi_1 \, d\xi_2 \end{aligned} \tag{8.99}$$

In the most general form of this result, $\boldsymbol{\gamma}$ is a 2×2 matrix (whose eigenvalues must have positive real parts).

Related propagation methods have been proposed in the quantum-mechanical context. Heller[44] proposed estimating the temporal evolution of a wave function as a superposition of "frozen Gaussians." Later, Herman and Kluk[45] found that a prefactor was missing from Heller's formulation. Their corrected result, known as the HK-IVR (initial value representation) method, has become a standard tool in quantum chemistry. Unlike SAFE, these methods involve integration over all phase space, and not only over a Lagrange manifold.

8.9.2 Insensitivity to γ

It would seem that SAFE's estimate depends strongly on the choice of the width parameter γ. However, it is easy to show that this is not the case. Consider the derivative of Eq. (8.98) with respect to γ, that is,

$$\begin{aligned} \frac{\partial U_\gamma}{\partial \gamma} &= \int a_0 \sqrt{\frac{\mathcal{Y}'}{H}} \left(\frac{X'}{2\mathcal{Y}'} - k\frac{\Delta^2}{2} \right) g \, d\xi \\ &= \int a_0 \sqrt{\frac{\mathcal{Y}'}{H}} \left[\frac{X'}{2\mathcal{Y}'} - \frac{X'}{2\mathcal{Y}'} + \mathcal{O}(k^{-1}) \right] g \, d\xi = \mathcal{O}(k^{-1}) U_\gamma \end{aligned} \tag{8.100}$$

where in the second step the integration-by-parts trick in Eq. (8.91) was used to remove the Δ's. This means that the variation of SAFE's estimate due to changes of the contributions' widths is asymptotically small, provided that a_0 is chosen to be independent of γ. This is the reason for the name of the method: while each of the contributions (or "elements") is flexible in width, their superposition (or "aggregate") is stable.

8.9.3 Phase-Space Interpretation

Insight into SAFE and the insensitivity of the estimate on γ (as well as the limitations of this insensitivity) can be gained through a phase-space picture. The *windowed Fourier transform* (WFT) like the one defined in Eq. (1.86) is a linear phase-space representation of a function. Here, let us consider a WFT where the window is chosen as a Gaussian of width $(k\mu)^{-1/2}$ (with $\mu > 0$):

$$S_f(x,\, p;\, \mu) = \sqrt{\frac{k\mu^{1/2}}{2\pi}} \int f(x') \exp\left[-k\mu \frac{(x'-x)^2}{2}\right] \times \exp\left[-ikp\left(x' - \frac{x}{2}\right)\right] dx' \tag{8.101}$$

The squared modulus of this transform is known as the *spectrogram* or *Husimi function*.[46] The application of this transformation to the field estimate in Eq. (8.98) gives

$$S_{U_\gamma}(x,\, p;\, z,\, \mu) = \sqrt{\frac{k\mu^{1/2}}{2\pi(\gamma+\mu)}} \int a_0 \sqrt{\frac{\gamma X' + iP'}{H}} \exp\left[-k\frac{(x-X)^2}{2(\gamma^{-1}+\mu^{-1})} - k\frac{(p-P)^2}{2(\gamma+\mu)}\right] \times \exp\left\{-\frac{ik}{\gamma+\mu}\left[\mu x\left(P - \frac{p}{2}\right) - \gamma p\left(X - \frac{x}{2}\right) - \mu XP\right]\right\} d\xi \tag{8.102}$$

That is, the contribution from each ray is a Gaussian (times a linear phase factor) localized around the corresponding phase-space point, with rms widths $\sqrt{(\gamma^{-1}+\mu^{-1})/2k}$ in the x direction and $\sqrt{(\gamma+\mu)/2k}$ in the p direction, as shown in Fig. 8.6*a*. This leads to the following intuitive picture illustrated in Fig. 8.6*b*: the evaluation of Eq. (8.102) is like spray-painting the wave field's phase-space distribution over the rays' traced line (the PSC). The characteristic footprint of the spray can has the widths mentioned earlier. However, the appearance of the final thicker fuzzy line painted over the PSC is roughly independent of the widths of the footprint,[47] as long as this footprint is fine enough to resolve the sections of the PSC where the curvature is tight.

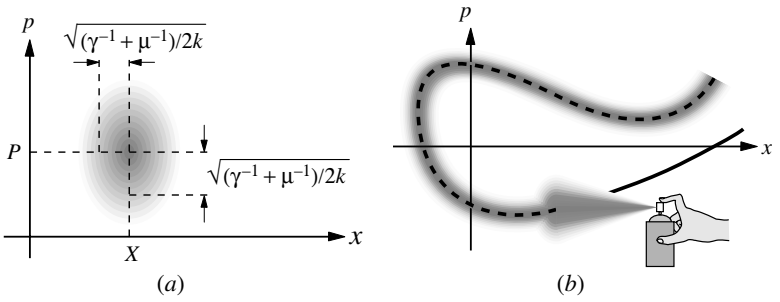


FIGURE 8.6 (*a*) The rms widths in phase space of the Gaussian-windowed Fourier transform of a field contribution in Eq. (8.102) due to a ray. (*b*) Picture of the Gaussian-windowed Fourier transform of SAFE's field estimate as the result of spray-painting over the PSC, where the footprint of the spray can has the shape shown in part *a*.

For example, near a position caustic, the size of the footprint in the p direction should not be excessively large, placing an upper bound on γ. On the other hand, near a momentum caustic, the x direction width of the footprint should be sufficiently small to resolve the corresponding curvature, and this results in a lower bound on γ. This implies that the geometry of the PSC (or the Lagrange manifold) dictates the valid ranges for the width of the contributions.

It is worth noticing that, in the limit $\gamma \to 0$, SAFE's estimate reduces to the momentum representation estimate in Eq. (8.74), as the contributions become infinitely wide. On the other hand, in the limit $\gamma \to \infty$, the Gaussian contributions become delta functions, and the estimate reduces to that resulting from the position representation treatment, given in Eq. (8.32). The fact that, in the presence of position or momentum caustics, these limiting values of γ violate the limitations outlined in the previous paragraph is consistent with the failure of the estimates discussed in Secs. 8.3 and 8.6 in these situations.

8.10 A Simple Example

Some of the methods described earlier are applied here to a simple example, corresponding to a mirage like reflection in two-dimensional space. Consider a hot surface at $x = 0$ heating up a transparent medium in the half-space $x > 0$. Away from this surface, the medium's refractive index is n_0, while near it, the refractive index decreases smoothly due to thermal expansion. Let us use the following

simplified model for the square of the refractive index

$$n^2(x) = \begin{cases} n_0^2 & x \geq x_0 \\ n_0^2[1 - \nu^2(x - x_0)^2] & x < x_0 \end{cases} \tag{8.103}$$

where x_0 is the height at which the refractive index becomes constant and ν determines the speed of the variation of the refractive index near $x = 0$. Consider a point source located at $(x_s, 0)$, where $x_s > x_0$, emitting light uniformly in all directions. Let us choose the ray parameter ξ as the angle (with respect to the z axis) at which each ray leaves the source. Since the refractive index is independent of z, H is invariant under propagation, according to Eq. (8.19*c*). From geometry, its value is found to be

$$H(\xi) = n_0 \cos \xi \tag{8.104}$$

The solutions to Eqs. (8.19*a*), (8.19*b*), and (8.20) take different forms in three regions:

1. If the ray has not entered the inhomogeneous region $x < x_0$, then it is a straight line defined by

$$X(z, \xi) = x_s + z \tan \xi \tag{8.105a}$$

$$P(z, \xi) = n_0 \sin \xi \tag{8.105b}$$

$$L(z, \xi) = n_0 z \sec \xi \tag{8.105c}$$

2. If the ray is inside the inhomogeneous region, then

$$X(z, \xi) = x_0 + \frac{\sin \xi}{\nu} \sin\{\nu[z - Z_1(\xi)] \sec \xi\} \tag{8.106a}$$

$$P(z, \xi) = n_0 \sin \xi \cos\{\nu[z - Z_1(\xi)] \sec \xi\} \tag{8.106b}$$

$$L(z, \xi) = n_0 \sec \xi \left\{ Z_1(\xi) + \left(1 + \frac{\sin \xi}{2}\right)[z - Z_1(\xi)] \right\} + \frac{n_0}{4\nu} \sin \xi \sin\{2\nu[z - Z_1(\xi)] \sec \xi\} \tag{8.106c}$$

where $Z_1(\xi) = (x_s - x_0) \cot \xi$ is the value of z at which the ray enters the inhomogeneous region.

3. If the ray has entered and exited the inhomogeneous region, then

$$X(z, \xi) = x_0 - [z - Z_2(\xi)] \tan \xi \tag{8.107a}$$

$$P(z, \xi) = -n_0 \sin \xi \tag{8.107b}$$

$$L(z, \xi) = n_0[z + Z_1(\xi) - Z_2(\xi)] \sec \xi + \frac{\pi n_0}{\nu} \left(1 + \frac{\sin \xi}{2}\right) \tag{8.107c}$$

where $Z_2(\xi) = Z_1(\xi) + \pi n_0 \cos\xi/\nu$ is the value of z at which the ray exits the inhomogeneous region. From the equations above, as well as their substitution in any of the wave field estimation formulas, it is easy to see that the solution to the problem depends on the dimensionless parameters kn_0/ν, νx_0, and νx_s, as well as on the dimensionless variables $(\nu x, \nu z)$. For SAFE, the estimate also depends on γ/ν.

Field estimates can be constructed from the equations for the rays given above, once they are supplemented with initial weights for the rays. In this example, only the position- and momentum-based field estimates as well as that corresponding to SAFE (with $\gamma/\nu = 1$) are calculated. Since we are considering a point source emitting light equally in all directions, the initial ray weight for SAFE is chosen as a constant, $a_0 = U_0$. For propagation in two dimensions, the position-representation-based estimate in Eq. (8.32) can be written as

$$U[X(z,\xi), z] \approx \sqrt{\frac{H(z_0,\xi)\,X'(z_0,\xi)}{H(z,\xi)\,X'(z,\xi)}}\; A_0[X(z_0,\xi), z_0]\,\exp[ikL(z,\xi)] \tag{8.108}$$

Notice that, to plot the estimate as a function of x, the equation $x = X(z,\xi)$ has to be solved for ξ as a function of x and z. Except for very simple cases such as free-space propagation, this cannot be done in closed form, so it is necessary to use a numerical root-search procedure. For the example considered here, this procedure is straightforward. However, for more complicated optical systems, this root-search procedure can be computationally demanding, as each iteration involves tracing a ray across the system. Also notice that, as written, Eq. (8.108) is not suitable for the evaluation of a field generated by a point source, since $X'(z_0,\xi) = 0$. This problem can be solved by substituting $\sqrt{H(z_0,\xi)\,X'(z_0,\xi)}\,A_0[X(z_0,\xi), z_0] = U_0$, as this substitution would give the asymptotic estimation of a circular wave in free space. The momentum-based estimate for two-dimensional propagation is given by

$$U(\vec{r}) \approx \sqrt{\frac{k}{2\pi}} \int B_0[P(z_0,\xi), z_0] \sqrt{\frac{H(z_0,\xi)\,P'(z_0,\xi)\,P'(z,\xi)}{H(z,\boldsymbol{\xi})}} \times \exp(ik\{L(z,\xi) + [x - X(z,\xi)]P(z,\xi)\})\,d\xi \tag{8.109}$$

For calculating the field due to a point source, we use $B_0[P(z_0,\xi), z_0] = \sqrt{i}U_0/H(z_0,\xi)$ (where the phase factor is inserted so that this estimate is in phase with the other two).

Figure 8.7 shows (*a*) some of the rays and (*b*) a segment of the PSC for $\nu x_s = 3$, $\nu x_0 = 0.8$, and $\nu z = 10$. Notice that these rays present both a position caustic (at $\nu x \approx 0.55$) and a momentum caustic (at $\nu x \approx 0.78$). The intensity estimates at $\nu z = 10$ for $kn_0/\nu = 1000$ are

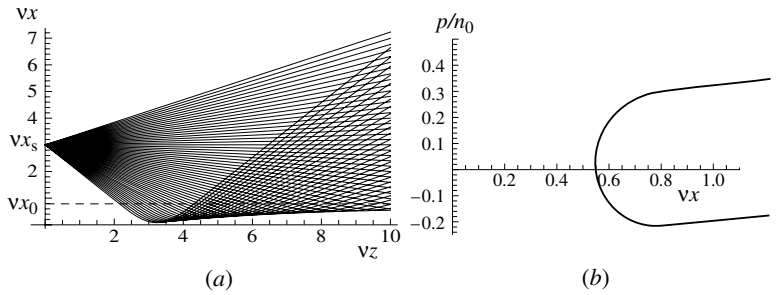


FIGURE 8.7 (*a*) Rays from a point source, reflected by a quadratic gradient index, and (*b*) a segment of the PSC for this family of rays at $\nu z = 10$.

shown in Fig. 8.8*a*, and the square modulus of their difference is shown (over a larger region) in Fig. 8.8. The position-based field estimate diverges at the caustic, but approaches SAFE's estimate elsewhere. The momentum-based estimate, on the other hand, differs from the other two over a wide interval around the momentum caustic. The error of this estimate is not divergent like that of the position-based one, but it is considerably more extended. This is so because the momentum-based estimate has a divergent but localized error in momentum space and, due to the uncertainty relation of the Fourier transform, this error becomes broad in the position representation.

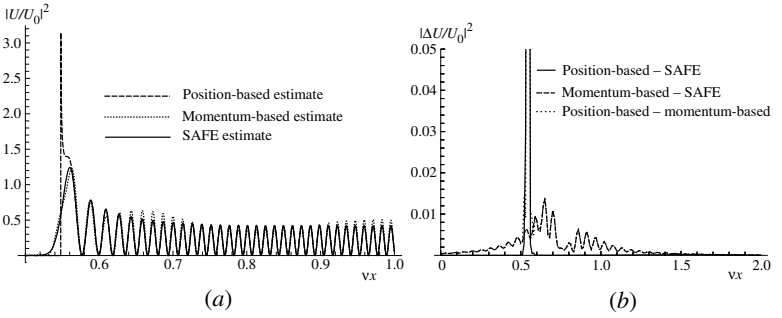


FIGURE 8.8 (*a*) Intensity estimates resulting from Eqs. (8.108) (dashed line), (8.109) (dotted line), and (8.98) (solid line) for the ray family in Fig. 8.7 with $\nu z = 10$ and $kn_0/\nu = 1000$. (*b*) Square modulus of the difference between the estimates in Eqs. (8.108) and (8.98) (solid line), Eqs. (8.109) and (8.98) (dashed line), and Eqs. (8.108) and (8.109) (dotted line), for the same values of the parameters.

8.11 Concluding Remarks

The field estimates presented in this chapter were derived through the same generic procedure: an ansatz is substituted into the Helmholtz equation, leading to conditions that the functions in the ansatz must satisfy. Note, however, that an alternative procedure for deriving some of these results exists where, instead of using the (differential) Helmholtz equation, one considers an infinite succession of propagation integrals over infinitesimally short distances. For each infinitesimally thin slice of space, the propagation integral amounts to a Rayleigh-Sommerfeld-like superposition of secondary waves, where for each of these waves the refractive index is approximated as that at the secondary source position. This description of wave propagation is analogous to Feynman's path integral formulation of quantum mechanics[48] as a "sum of all possible histories." The ray-based estimates can then be obtained by approximating the integrals asymptotically, where k is used as the asymptotic parameter. The position-based estimate in particular results from applying the method of stationary phase[49] to all the integrals. The resulting leading contribution to the field at a point is associated with a trajectory (or a set of trajectories) arriving at this point from the initial plane. The optical path length of this trajectory is stationary with respect to infinitesimal variations, so it satisfies Fermat's theorem. These stationary paths are therefore the rays of geometrical optics. Since the method of stationary phase consists of approximating the rapidly varying phase as a quadratic polynomial around a stationary point, the short-wavelength limit studied here has something in common with the paraxial and quasi-homogeneous limits mentioned in Sec. 8.1: they all lead to field estimates consistent with ray optics through the quadratic approximation of phases inside integrals, so that these integrals can be approximated analytically.

The author acknowledges support from the National Science Foundation under grant 0449708, and from an anonymous donor under an internal grant from The Institute of Optics.

References

1. M. Born and E. Wolf, *Principles of Optics*, 7th ed., Cambridge University Press, Cambridge, 1999.
2. R. K. Luneburg, *Mathematical Theory of Optics*, University of California Press, Berkeley, 1964.
3. A. Sommerfeld, *Optics*, *Lectures on Theoretical Physics*, vol. IV, Academic Press, New York, 1964.
4. Yu. A. Kravtsov and Yu. I. Orlov, *Geometrical Optics of Inhomogeneous Media*, Springer-Verlag, Heidelberg, 1990.
5. Yu. A. Kravtsov and Yu. I. Orlov, *Caustics, Catastrophes, and Wave Fields*, 2d ed., Springer-Verlag, Heidelberg, 1998.

6. H. Bruns, "Das Eikonal," *Leipz. Sitzgsber.* **21**: 321–436 (1895).
7. W. R. Hamilton, *The Mathematical Papers of Sir William Rowan Hamilton,* vol. I. *Geometrical Optics*, Cambridge University Press, Cambridge, 1931.
8. G. W. Forbes, "On variational problems in parametric form," *Am. J. Phys.* **59**: 1130–1140 (1991).
9. K. B. Wolf, "Las tres caras the Hamilton en la óptica geométrica y en la mecánica," *Rev. Mex. Fis.* **37**: 136–146 (1991).
10. K. B. Wolf, *Geometric Optics on Phase Space*, Springer, Berlin, 2004, pp. 5–13.
11. V. P. Maslov, *Perturbation Theory and Asymptotic Methods*, Moskov UP, Moscow, 1965.
12. J. B. Delos, "Semiclassical calculation of quantum mechanical wave functions," *Adv. Chem. Phys.* **65**: 161–213 (1986).
13. Yu. A. Kravtsov, G. W. Forbes, and A. A. Asatryan, "Theory and applications of complex rays," *Progress Opt.* **34**: 1–62 (1999).
14. A comprehensive annotated list of references is given in M. C. Gutzwiller, "Resource letter ICQM-1: The interplay between classical and quantum mechanics," *Am. J. Phys.* **66**: 304–324 (1998).
15. H. Goldstein, C. Poole, and J. Safko, *Classical Mechanics*, 3d ed., Addison-Wesley, San Francisco, 2002, p. 21.
16. See, e.g., D. Bohm, *Quantum Theory*, Dover, Mineola, N.Y., 1989, pp. 264–295.
17. J. H. Van Vleck, "The correspondence principle in the statistical interpretation of quantum mechanics," *Proc. N.A.S.* **14**: 178–188 (1928).
18. M. C. Gutzwiller, *Chaos in Classical and Quantum Mechanics*, Springer-Verlag, New York, 1990.
19. D. Bohm, "A suggested interpretation of the quantum theory in terms of hidden variables, I and II," *Phys. Rev.* **85**: 166–193 (1952).
20. R. E. Wyatt, *Quantum Dynamics with Trajectories: Introduction to Quantum Hydrodynamics*, Springer-Verlag, New York, 2005.
21. See, e.g., Ref. 1, pp. 639–642.
22. M. A. Alonso and G. W. Forbes, "Uniform asymptotic expansions for wave propagators via fractional transformations," *J. Opt. Soc. Am. A* **14**: 1279–1292 (1997).
23. E. L. Ince, *Ordinary Differential Equations*, Dover, New York, 1956.
24. M. M. Popov, "A new method of computation of wave fields using Gaussian beams," *Wave Motion* **4**: 85–97 (1982).
25. V. M. Babich and M. M. Popov, "The Gaussian summation method (review),"*Radiophys. Q. Elec.* **32**: 1063-1081 (1989).
26. M. M. Popov, *Ray Theory and Gaussian Beam for Geophysics*, EDUFBA, Salvador-Bahia, 2002.
27. L. Klimeš, "Gaussian packets in the computation of seismic wavefields," *Geophys. J. Int.* **99**: 421–433 (1989).
28. B. S. White, A. Norris, A. Bayliss, and R. Burridge, "Some remarks on the Gaussian beam summation method," *Geophys. J. R. Astr. Soc.* **89**: 579–636 (1987).
29. E. J. Heller, "Time-dependent approach to semiclassical dynamics," *J. Chem. Phys.* **62**: 1544–1555 (1975).
30. R. G. Littlejohn, "The semiclassical evolution of wave packets," *Phys. Rep.* **138**: 193–291 (1986).
31. S. W. McDonald, "Phase space representations of wave equations with applications to the eikonal approximation for short wavelength waves," *Phys. Rep.* **158**: 337–416 (1988).
32. M. J. Bastiaans, "The expansion of an optical signal into a discrete set of Gaussian beams," *Optik* **57**: 95–102 (1980).
33. A. Shlivinski, E. Heyman, A. Boag, and C. Letrou,"A phase-space beam summation formulation for ultra wideband radiation," *IEEE Trans. Antennas Propag.* **53**: 2042–2053 (2004).
34. A. N. Norris, "Complex point-source representation of real point sources and the Gaussian beam summation method," *J. Opt. Soc. Am. A* **3**: 2005–2010 (1987).

35. P. D. Einziger and S. Raz, "Beam-series representation and the parabolic approximation: The frequency domain," *J. Opt. Soc. Am. A* **5**: 1883–1892 (1998).
36. J. M. Arnold, "Rays, beams and diffraction in a discrete phase space: Wilson bases," *Opt. Express* **10**: 716–727 (2002).
37. B. Z. Steinberg, E. Heyman, and L. B. Felsen, "Phase-space beam summation for time-harmonic radiation from large apertures," *J. Opt. Soc. Am. A* **8**: 41–59 (1991).
38. D. Gabor, "Theory of communication," *J. Inst. Electr. Eng.* **20**: 594–598 (1946).
39. G. W. Forbes and M. A. Alonso, "Using rays better. I. Theory for smoothly varying media," *J. Opt. Soc. Am. A* **18**: 1132–1145 (2001).
40. M. A. Alonso and G. W. Forbes, "Using rays better. II. Ray families to match prescribed wave fields," *J. Opt. Soc. Am. A* **18**: 1146–1159 (2001).
41. M. A. Alonso and G. W. Forbes, "Using rays better. III. Error estimates and illustrative applications in smooth media," *J. Opt. Soc. Am. A* **18**: 1357–1370 (2001).
42. G. W. Forbes, "Using rays better. IV. Theory for refraction and reflection," *J. Opt. Soc. Am. A* **18**: 2557–2564 (2001).
43. M. A. Alonso and G. W. Forbes, "Stable aggregates of flexible elements give a stronger link between rays and waves," *Opt. Express* **10**: 728–739 (2002).
44. E. J. Heller, "Frozen Gaussians: A very simple semiclassical approximation," *J. Chem. Phys.* **75**: 2923–2931 (1981).
45. M. F. Herman and E. Kluk, "A semiclassical justification for the use of non-spreading wavepackets in dynamics calculations," *J. Chem. Phys.* **91**: 27–35 (1984).
46. W. P. Schleich, *Quantum Optics in Phase Space*, Wiley, Berlin, 2001, p. 324.
47. M. A. Alonso and G. W. Forbes, "New approach to semiclassical analysis in mechanics," *J. Math. Phys.* **40**: 1699–1718 (1999).
48. R. P. Feynman and A. R. Hibbs, *Quantum Mechanics and Path Integrals*, McGraw-Hill, New York, 1965.
49. See, e.g., Ref. 1, pp. 883–891.

This page intentionally left blank

CHAPTER 9
Self-Imaging in Phase Space

Markus E. Testorf

Dartmouth College, Hanover, New Hampshire, U.S.A

9.1 Introduction

There is little doubt that Fourier optics has shaped optical engineering in ways only comparable to geometrical optics. Understanding wavefronts and optical hardware in terms of linear system theory has been pivotal to integrating optical sciences with signal processing and numerical computing. Topics such as diffractive optics design and computational imaging are almost unimaginable without the theoretical foundations of Fourier mathematics.

An emerging and fascinating alternative to Fourier optics is phase-space optics. Forged by the marriage of joint time-frequency analysis and the phase-space formalism of quantum mechanics, phase-space optics is a platform for describing ray optics, radiometry, coherent Fourier optics, and coherence theory with a single consistent framework.

The phase-space interpretation is often perceived as a highly mathematical exercise with little additional information to complement the standard treatment in terms of Fourier optics. This perception is perhaps justified when looking at the mathematical properties of basic phase-space tools, namely, the Wigner distribution function (WDF). The WDF "inflates" the complex amplitude into an apparently redundant multidimensional function with transverse spatial position and spatial frequency as independent variables. In addition, the WDF is a bilinear transformation, and hence the linearity of signal superposition is lost.

Closer study, however, reveals a powerful representation of optical signals and systems adding unprecedented insight and intuition to well-known optical phenomena, which in turn forms the basis for new system designs and applications.

While not all effects in classical optics find a preferable interpretation in phase space, the self-imaging phenomenon can be regarded as a poster child for promoting phase-space optics as a true alternative to Fourier optics.

In this chapter, we revisit the self-imaging effect and its closest relatives, including the Talbot effect, the fractional Talbot effect, and the Lau effect. All essential relationships are derived from simple diagrams of the associated phase space. We show that much of the mathematics involving the WDF can be avoided once a small set of relationships has been established. Compared to the Fourier optics treatment, we will then barely need any mathematical instrument, except basic algebra and geometry. This is only possible if the mathematical tools of phase-space optics are not applied in a mechanistic way, but are customized to each situation. Thus, the study of self-imaging may not provide a foolproof recipe. The chapter rather is intended as a teaser to illustrate the beauty of phase-space optics. While intellectual pleasure is guaranteed, the phase-space interpretation of self-imaging may also have the potential of pointing toward new effects and applications not immediately obvious from a Fourier optics perspective.

9.2 Phase-Space Optics Minimum Tool Kit

Phase-space optics represents N-dimensional signals in a $2N$-dimensional configuration space. Since we want to use diagrams not merely for illustration, but also to obtain quantitative results, we restrict our discussion to one-dimensional optical signals. A phase-space distribution suitable to represent the one-dimensional complex amplitude distribution $u(x)$ is the WDF [1–3]

$$W(x,\nu)=\int_{-\infty}^{\infty}u\left(x+\frac{x'}{2}\right)u^{*}\left(x-\frac{x'}{2}\right)\exp(-i2\pi\nu x')\,dx' \qquad (9.1)$$

The signal function enters the transform twice which results in a bilinear transformation. The properties of the WDF are highly symmetric with respect to the two conjugate variables x and ν. This is reflected by the alternative definition

$$W(x,\nu)=\int_{-\infty}^{\infty}\tilde{u}\left(\nu+\frac{\nu'}{2}\right)\tilde{u}^{*}\left(\nu-\frac{\nu'}{2}\right)\exp(i2\pi\nu' x)\,d\nu' \qquad (9.2)$$

based on the Fourier transform of the signal

$$\tilde{u}(\nu) = \int_{-\infty}^{\infty} u(x) \exp(-i2\pi\nu x)\, dx \tag{9.3}$$

Intensity and power spectrum of the complex signal can be recovered as the marginals of the WDF, i.e., the projections parallel to the phase-space axes

$$\int_{-\infty}^{\infty} W(x, \nu)\, d\nu = |u(x)|^2 \quad \text{and} \quad \int_{-\infty}^{\infty} W(x, \nu)\, dx = |\tilde{u}(\nu)|^2 \tag{9.4}$$

In fact, it is possible to regain the original signal, apart from a constant factor, as a Fourier transformation of the WDF, proving that the WDF is a complete representation of the complex amplitude.

To gain intuition, we consider the WDF of two copropagating plane waves (Fig. 9.1). The complex amplitude

$$u_{\text{tpw}}(x) = \exp(i2\pi\nu_1 x) + \exp(i2\pi\nu_2 x) \tag{9.5}$$

is translated to

$$W_{\text{tpw}}(x, \nu) = \delta(\nu-\nu_1)+\delta(\nu-\nu_2)+2\cos[2\pi(\nu_1-\nu_2)x]\,\delta\left(\nu - \frac{\nu_1+\nu_2}{2}\right) \tag{9.6}$$

Figure 9.1*b* shows a schematic representation of the WDF which we will call the phase-space diagram (PSD). The phase-space interpretation of optical rays associates each ray with a single point in the $x\nu$ plane. This means, for a given plane z, along the optical axis a ray is represented by its transverse coordinate x and its propagation

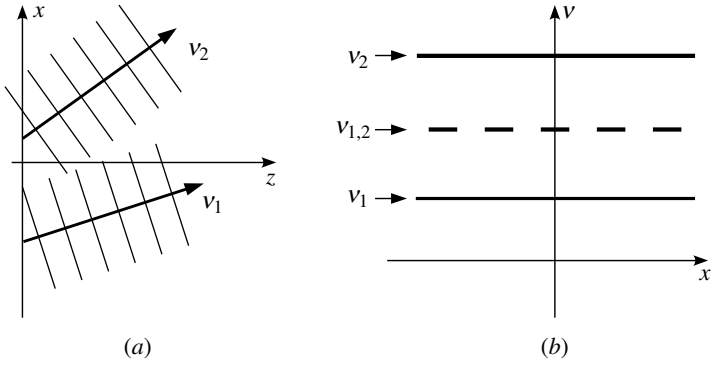


FIGURE 9.1 Interference in phase space: *(a)* Two propagating coherent plane waves and *(b)* the corresponding phase-space diagram.

angle θ, where $\lambda\nu = \sin\theta$, with λ being the wavelength of the coherent wavefront.

The WDF can be regarded as a generalized ray distribution. If we interpret each of the two plane waves as a bundle of rays, where each bundle has a different propagation angle, we can map each plane wave into the PSD. The two horizontal δ-lines, marked as ν_1 and ν_2 in Fig. 9.1*b*, correspond to the ray coordinates which we would intuitively expect as the phase-space distribution of two plane waves. The additional cosine modulated line at the intermediate frequency $\nu_{1,2}$ is the so-called interference term or cross-term related to the bilinearity of the WDF. The fundamental period of the interference term is represented as the period of the dashed line in the PSD. This additional term ensures proper encoding of interference effects is not considered by geometrical optics. It is also implicit that the interference term carries the information about the mutual coherence of the two plane waves.[4] For mutually incoherent waves the interference term of the WDF vanishes, and for partially coherent signals it is a weighted contribution related to the degree of mutual coherence of the two plane waves. While not corresponding to rays in a geometrical optics sense, the phase-space points associated with interference terms behave exactly as points associated with ordinary rays.

This means that the WDF allows us to study the phase space of rays and how it changes as the light signal propagates through a paraxial optical system. Then the same rules are applied to the generalized phase-space distribution of the WDF to propagate wavefronts through the optical system. Paraxial ray tracing is conveniently described with matrix optics. In fact, matrix optics, which appears in many textbooks (see, e.g., Ref. 5) is phase-space optics in disguise. Each optical element or system can be represented by a 2×2 matrix, and the coordinates transform according to

$$\begin{pmatrix} x \\ \nu \end{pmatrix}_{\text{out}} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} x \\ \nu \end{pmatrix}_{\text{in}} \tag{9.7}$$

Thus any paraxial optical system that can be described by an $ABCD$ matrix amounts to a geometrical transformation of the WDF

$$W_{\text{out}}(x, \nu) = W_{\text{in}}(Ax + B\nu, Cx + D\nu) \tag{9.8}$$

modifying the location of each point of the WDF, but not its value.

For instance, paraxial free-space propagation or Fresnel diffraction corresponds to a shear of the WDF parallel to the x axis

$$W_{\text{Fr}}(x, \nu) = W_0(x - \lambda z\nu, \nu) \tag{9.9}$$

The Fourier dual operation is the phase modulation with a linear chirp function, i.e., the function of a parabolic lens. For a convex lens of focal

length f we find

$$W_L(x, \nu) = W_0\left(x, \nu + \frac{x}{\lambda f}\right) \tag{9.10}$$

In Fig. 9.2 both operations are applied to a generic phase-space volume of rectangular shape. In addition, Fig. 9.2*d* shows the effect of a Fourier transformation that corresponds to exchanging both phase-space coordinates with a clockwise rotation of the WDF by 90°.

Important operations are modulation and convolution of two signals. For the product of two functions $u(x) = g(x)\,h(x)$, the corresponding WDFs are convolved with respect to the frequency variable

$$W_u(x, \nu) = \int_{-\infty}^{\infty} W_g(x, \nu')\,W_h(x, \nu - \nu')\,d\nu' = W_g(x, \nu) *_\nu W_h(x, \nu) \tag{9.11}$$

The symmetry between x and ν implies that a convolution between the two signals is translated to a convolution between the corresponding WDFs with respect to x.

Finally, we will also need the phase-space representation of a linear chirp function

$$u_{\text{ch}}(x) = \exp[i2\pi(\alpha x^2 + \beta x + \gamma)] \tag{9.12}$$

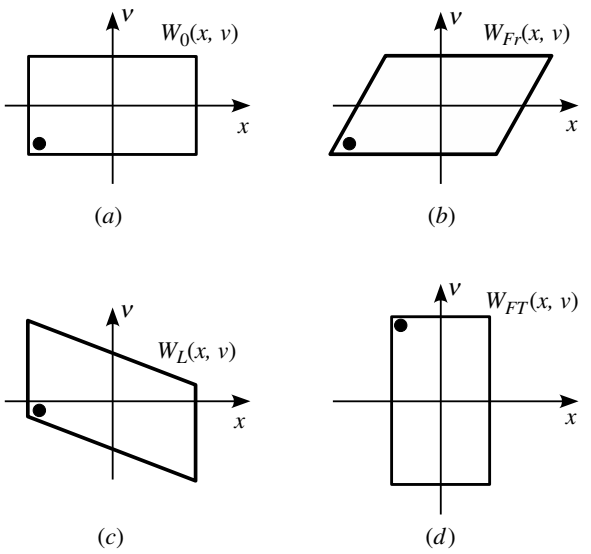


FIGURE 9.2 Paraxial optics in phase space: (*a*) Generic phase-space distribution of an optical signal, (*b*) signal after Fresnel diffraction, (*c*) after modulation with a quadratic phase function, and (*d*) after Fourier transformation.

with the WDF reading

$$W_{\mathrm{ch}}(x, \nu) = \delta(\nu - 2\alpha x - \beta) \tag{9.13}$$

which includes as a limiting case a line parallel to x for each off-axis plane wave and a line parallel to ν as the representation of a single point source.

The WDF of the chirp function also provides us with an alternative interpretation of the affine transformations of phase space associated with Fresnel diffraction and a thin lens. Fresnel diffraction can be understood as a convolution of the complex amplitude distribution with the point response function of free space

$$h_{\mathrm{Fr}}(x, z) = \frac{1}{\sqrt{i\lambda z}} \exp\left(\frac{i\pi}{\lambda z} x^2\right) \tag{9.14}$$

This translates to

$$W_h(x, \nu) = \frac{1}{|\lambda z|} \delta\left(\nu - \frac{x}{\lambda z}\right) = \delta(x - \lambda z \nu) \tag{9.15}$$

which is a straight line in phase space. From this we obtain Eq. (9.9) straightforwardly as the convolution in x between the input WDF and $W_h(x, \nu)$. Similarly, convolution of the oblique line in ν with the WDF of the input signal corresponds to the operation in Eq. (9.10).

9.3 Self-Imaging of Paraxial Wavefronts

Self-imaging was first observed by Henry Fox Talbot[6] in 1836 and theoretically explained by Lord Rayleigh[7] in 1881. In modern language, the Talbot effect is concerned with Fresnel diffraction of a coherent monochromatic wavefront that is strictly periodic in the transverse direction. Then the physics of wave propagation ensures strict periodicity along the axis of propagation z as well.

It was not until 1967 that Montgomery proved lateral periodicity to be a sufficient, but not a necessary, condition for self-imaging.[8] In fact it is possible to construct signals with a discrete plane wave spectrum, which exhibit self-imaging not only within the bounds of paraxial optics, but also for the nonparaxial domain of propagation.

With few exceptions the Talbot effect was ignored until affordable coherent light sources became available and triggered a wave of research related to coherent optical signal processing. Since then, the Talbot effect has become a standard tool of Fourier optics. For a detailed survey of the self-imaging phenomenon and its applications, refer to the 1989 review by Patorski.[9]

The scope of self-imaging was dramatically expanded by the study of Fresnel diffraction of periodic signals at rational fractions of the Talbot self-imaging period. Namely, the work by Winthrop and Worthington[10] identified Fresnel images, i.e., the diffraction patterns at so-called fractional Talbot planes, as cases, where the Fresnel diffraction integral can be expressed in simple analytic form. Subsequent investigations further simplified the analytic expressions,[11–13] culminating in a discrete matrix formulation of near-field diffraction, which relates the amplitudes of sampled periodic signals in different fractional Talbot planes via linear transformations.[14–16] Interest in studying the fractional Talbot effect largely increased by the invention of the Talbot array illuminator,[17,18] a diffractive optical element to convert a homogeneous wavefront to an array of high-intensity spots. Today, a vast number of studies can be found in the literature that describe design procedures, experimental work, and applications of Talbot array generators (see Refs. 19–23 as only a small set of related work).

A close relative of the Talbot effect is the Lau effect which is concerned with incoherent periodic optical signals.[24–26] While not receiving the same attention as coherent self-imaging, perhaps due to its less intuitive nature and a more difficult experimental implementation, the Lau effect was shown to be useful for a number of applications and remains a vivid member of the family of self-imaging phenomena.

Self-imaging in phase space was first studied by Ojeda-Castañeda and Sicre[27] and applied to both the Talbot effect and the Lau effect. The phase-space analysis was later extended to include the fractional Talbot effect[28] and the design of Talbot array illuminators.[29]

The remainder of the chapter, in part, is a review of previously published results. In part, however, it contains original contributions to highlight self-imaging as a phenomenon that is exceptionally suited to be explored with phase-space optics.

9.4 The Talbot Effect

The setup to observe the Talbot effect is schematically depicted in Fig. 9.3. An infinitely extended grating is illuminated with a monochromatic coherent plane wave of wavelength λ. Figure 9.3 shows the simulated intensity pattern behind a Ronchi grating. After some propagation distance z_T we find the exact intensity distribution which is observable immediately behind the grating, and we call z_T the self-imaging distance or Talbot distance. The fractional Talbot effect, which is discussed in detail in Sec. 9.6, is associated with Fresnel diffraction at rational fractions M/N of the Talbot distance, where M and N are integer numbers. Our discussion will exclusively assume

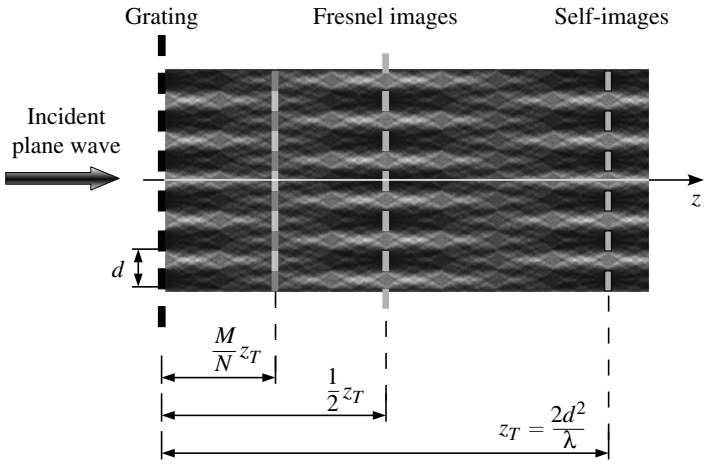


FIGURE 9.3 Configuration to observe the Talbot effect and the fractional Talbot effect.

paraxial wave propagation and will be limited to a single transverse coordinate.

To analyze the Talbot effect in phase space, we need to add the WDF of a periodic signal to our phase-space toolbox. Transverse periodicity implies a complex amplitude $u(x) = u(x + d)$ with d being the transverse period of the signal. It is commonly assumed that the periodic amplitude distribution is the result of illuminating a grating with a plane wave. It should be emphasized, however, that the Talbot effect is not concerned with the interaction between the incident wave and the diffraction screen, but exclusively with the evolution of a periodic paraxial wavefront.

The periodic signal can be expanded into a Fourier series

$$u_p(x) = \sum_{-\infty}^{\infty} u_n \exp\left(\frac{i2\pi n x}{d}\right) \tag{9.16}$$

which can be used to compute the corresponding WDF as

$$W_p(x, \nu) = \sum_{n=-\infty}^{\infty} \sum_{n'=-\infty}^{\infty} u_n u_{n'}^* \, \delta\left(\nu - \frac{n + n'}{2d}\right) \exp\left(i2\pi \frac{n - n'}{d} x\right) \tag{9.17}$$

The PSD of the periodic signal is shown in Fig. 9.4. Equation (9.17) expresses the WDF as a set of modulated δ lines at integer multiples of frequency $1/(2d)$. For $n = n'$ we obtain the so-called self-terms of the WDF associated with the discrete frequencies of the Fourier series in

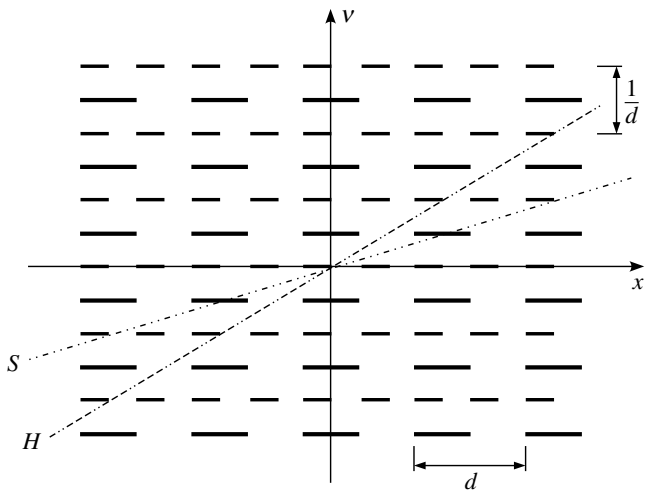


FIGURE 9.4 PSD of a periodic function.

Eq. (9.16). The self-terms correspond to the WDF of individual discrete frequency components and form lines without modulation located at multiples of the base interval $1/d$. All other terms $n \neq n'$, in Eq. (9.17), are cross-terms equivalent to the cross-term in Eq. (9.6). Note that the choice of what term to identify as self-term depends on the particular expansion we use to express the signal as a linear superposition of signal components.

We can further observe that the modulation in x can again be interpreted as a Fourier series; i.e., at each frequency $1/(2d)$ we find a δ line, with periodic modulation in x. It is important to note that the base frequency of this periodicity is d for terms $\nu_m = (2m+1)/(2d)$, but is $d/2$ for $\nu_m = m/d$, with m being an integer number.

While this can be readily verified from Eq. (9.17), we can also interpret this as an inherent property of the cross-terms. We can construct any cross-term by considering all possible pairs of self-terms in turn. For each pair we expect a cross-term to appear at half distance in between the two self-terms. The cross-terms are generally modulated periodically to ensure they do not contribute to the marginal if we integrate the WDF along its axes. This modulation frequency is proportional to the distance in phase space between the respective self-terms.

In Fig. 9.4 the cross-terms at $\nu_m = (2m+1)/(2d)$, which are interlaced with the self-terms, can only be constructed from self-terms separated by an odd multiple of $1/d$, and the base period of the modulation

becomes d. In contrast, the cross-terms superposed with self-terms at $\nu_m = m/d$ can only be associated with pairs of self-terms separated by integer multiples of $2/d$, and as a consequence the modulation we observe has a period of $d/2$, that is, one-half of what we obtain for interlaced cross-terms.

Fresnel diffraction corresponds to a horizontal shear of the phase-space distribution, and we see without difficulty that the PSD does not change if the shear equals d at $\nu = 1/(2d)$. All other discrete δ lines are automatically sheared by an integer multiple of the period d, and we obtain the original phase-space distribution. The line S in Fig. 9.4 corresponds to the tilt we would observe for a vertical line at the input and is equivalent to the WDF of the associated point-spread function of free space in Eq. (9.14). Using Eq. (9.15), the Talbot length can now be deduced from $d = \lambda z_T/(2d)$, or

$$z_T = \frac{2d^2}{\lambda} \tag{9.18}$$

It is immediately clear from the PSD in Fig. 9.4 that self-images can be found at any integer multiple of the Talbot distance. This is related to the fact that the WDFs (and not only their projections) before and after shearing are identical, which proves that self-imaging recovers not only the intensity distribution, but also the complex amplitude of the input signal. Also note that once the WDF of a periodic function is known, the self-imaging condition can be deduced with a minimum of mathematical formalism. Furthermore, the quantitative result can only be obtained correctly by including the cross-terms, namely, the interlaced terms at half intervals, into our analysis. This is of significance as PSDs are often constructed as heuristic notions of phase space rather than from the results of a rigorous evaluation of the WDF.

This rigorous analysis of the WDF also provides access to the Fresnel diffraction amplitude at $z_T/2$. The diffraction patter is often described as an additional grating image which is reversed in contrast.[30] To understand this notion, we again turn to Fig. 9.4. If we consider the shear associated with line H, which is only one-half the shear necessary for Talbot self-imaging, we again recover the phase-space distribution of the input signal, however shifted in x by $d/2$ compared to the distribution in Fig. 9.4. By inspecting the points of intersection between line H and the horizontal delta lines, we can verify that the lateral shift at $\nu_m = m/d$ in fact is a multiple of d, while it is an odd multiple of $d/2$ for the interlaced frequencies. Thus the terms at even multiples of $1/(2d)$ and at odd multiples of $1/(2d)$ only register because the base period of the modulation at $\nu_m = m/d$ is $d/2$, that is, one-half of what we might expect intuitively for a phase-space distribution of a periodic signal. For typical grating profiles, including, for instance, the Ronchi

grating, we indeed observe a self-image of the input grating pattern at $z_T/2$, but with the bright and dark lines exchanged.

For practical applications the transverse shift by $d/2$ may be negligible, and it seems almost justified to use $z_T/2$ as the self-imaging distance. In Sec. 9.6, however, we expand the notion of self-imaging to include rational fractions of the Talbot distance. In this context it will be more suitable to count only precise replicas of the input complex amplitude as self-images. The Talbot image at half distance is then identified as a Fresnel image, i.e., the Fresnel diffraction pattern at the fractional Talbot plane $z_{1,2} = z_T/2$.

9.5 The "Walk-off" Effect

The Talbot effect is the result of the in-phase superposition of all plane waves. The discrete nature of the spectrum guarantees equal phase delays between adjacent frequencies; i.e., the in-phase condition can readily be achieved.

This has to change as soon as we consider a finite grating aperture. On one hand, the δ lines will be replaced by the WDF of a sinc function along ν, and the quadratic propagator of the Fresnel diffraction transform is no longer sampled at the appropriate intervals only. As a consequence the line shape of the Talbot grating image will be perturbed.

On the other hand, in a quasi-geometrical sense, each discrete plane wave mode is now truncated while propagating off-axis. This is observable as the so-called walk-off effect.[30–32] At the boundaries of the propagating window a transition region develops, where no self-images can be observed, which cuts into the domain of self-imaging as a function of propagation distance. We can interpret this effect in terms of truncated plane waves, which "walk off" the window defined by the grating aperture until they no longer contribute to the self-images. In close analogy to the Abbe theory of coherent image formation, we need at least two interfering plane waves to observe a self-image with structural information about the original grating.

It is possible to estimate the maximum distance for self-imaging with para-geometrical optics. The analysis is conveniently executed with the help of the PSD in Fig. 9.5. We assume a periodic signal with a finite bandwidth $\Delta\nu$ and thus a finite number of Fourier coefficients. This does not impose any severe restrictions since in practice all signals are essentially band-limited due to the frequency cutoff of systems for signal generation and transport. We now assume a quasi-ray optical perspective by further assuming a truncation of the periodic signal to M periods without impact on its plane wave spectrum.

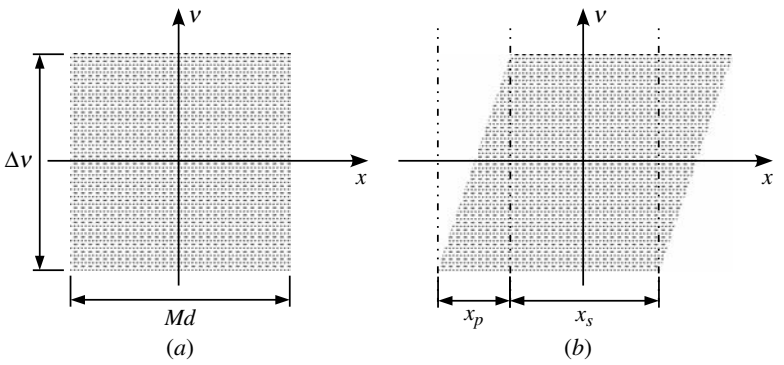


FIGURE 9.5 Phase-space interpretation of the walk-off effect.

Fresnel diffraction corresponds to a horizontal shear of the phase-space distribution, and only over a region x_s, in Fig. 9.5*b*, where all plane waves can interfere, we expect the self-image to resemble the input signal. The region x_p marks a transition region that can be associated with edge diffraction from the grating aperture as the dominating effect.

To derive a limiting-conditions analog to Abbe's theory of the microscope, we consider a cosine pattern as input signal with $\Delta\nu = 2/d$, giving rise to three propagating plane waves only (the analysis does not change if we assume a periodic pattern with higher-order nonzero Fourier coefficients; in this case $\Delta\nu$ defines the frequency band for which truncated plane waves have not yet completely moved out of the signal window).

With a grating aperture of size Md we can now estimate the maximum distance over which self-imaging can be observed as the point where $x_s = 0$. With $x_p = \lambda z_{\max} \Delta\nu = Md/2$ we find

$$z_{\max} = \frac{Md^2}{\lambda} \tag{9.19}$$

which corresponds to the estimate given in Ref. 30.

9.6 The Fractional Talbot Effect

While forming the basis for various applications,[9] the Talbot effect is also interesting from a mathematical perspective. For the majority of functions that are commonly used to model optical systems, the Fresnel diffraction integral has no simple analytic solution. The Talbot

effect is exceptional, because it defines a case in which the Fresnel diffraction integral indeed has a trivial solution regardless of the grating's groove shape.

A similar characterization applies to the fractional Talbot effect, which also defines a set of cases in which the Fresnel diffraction integral can be expressed in simple closed form. The study of Fresnel images, i.e., Fresnel diffraction patterns at rational fractions of the Talbot distance, has revealed a formal structure of the fractional Talbot effect, which appeals to the experimentalist as well as to the theoretician.

While the first systematic study of Fresnel images can be found in the seminal paper by Winthrop and Worthington,[10] it is worth pointing out the extent to which the formulation of the fractional Talbot effect was simplified during the past four decades. Formulating the fractional Talbot effect in terms of phase-space optics allows us to appreciate this progress in a particularly satisfying way.

Instead of postulating rational fractions of the Talbot distance as diffraction planes worthy of our attention, the phase-space interpretation effortlessly finds the fractional Talbot planes as the result of searching for cases with interesting properties.

The analysis of the fractional Talbot effect requires some preparation, which also serves as a demonstration as to how problems can be dissected for applying a phase-space analysis most effectively. For Fresnel diffraction of periodic complex amplitudes, we have to express the Fresnel diffraction amplitude at distance z from the input plane as

$$u(x, z) = u_p(x) * \sum_{n=-\infty}^{\infty} \delta(x - nx_0) * h_{\mathrm{Fr}}(x, z) \tag{9.20}$$

with $*$ denoting a convolution. This means that we formulate the periodic input signal as the function describing a single period $u_p(x)$ convolved with an infinite comb function. The propagation of the signal corresponds to a second convolution with the pulse response of free space, in Eq. (9.14).

The convolution operation is associative; i.e., we are at liberty to interpret the output signal as a convolution of the comb function with the propagator, then followed by a second convolution with the groove shape of the grating. It is the propagation of the array factor that is most conveniently discussed in phase space. This analysis can be carried out without explicitly considering the groove shape for which the WDF may be hard or impossible to calculate analytically.

However, one more time, we need to extend our phase-space toolbox to include the WDF of the comb function. Substituting the infinite

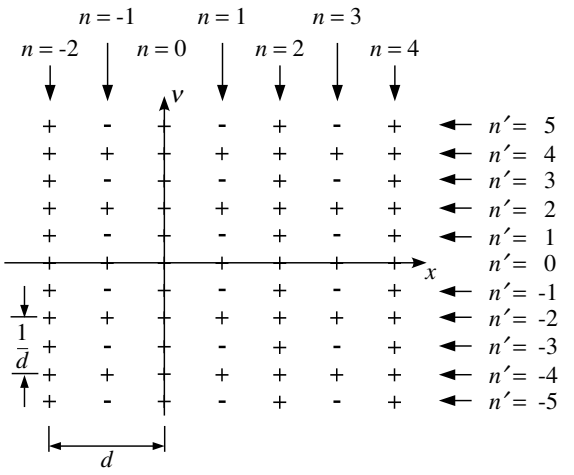


FIGURE 9.6 The PSD of the comb function. The δ functions with a positive (+) sign are interlaced with δ functions with a negative (−) sign.

δ comb into the definition of the WDF in Eq. (9.1), we find

$$W_{\text{comb}}(x, \nu) = \frac{1}{2d} \sum_{n=-\infty}^{\infty} \sum_{n'=-\infty}^{\infty} (-1)^{nn'} \delta\left(x - \frac{nd}{2}\right) \delta\left(\nu - \frac{n'}{2d}\right) \quad (9.21)$$

The corresponding PSD is shown in Fig. 9.6. The δ functions at locations $(x_m, \nu_{m'}) = (md, m'/d)$, with m and m' being integer numbers, are interlaced with a grid of δ functions of alternating sign. The alternating sign ensures that these interlaced terms do not contribute to the marginals of WDF, i.e., the intensity and the power spectrum of the comb function.

We can now use this PSD to study Fresnel diffraction by applying a linear shear in x. As we increase the shear, we can identify cases where points with a spatial frequency coordinate $N/(2d)$ are laterally shifted by a multiple of the period Md. Figure 9.7 shows the PSD corresponding to $M = 1$, $N = 3$. We observe registration of the horizontal positions of δ functions forming columns of δ points with only positive sign, interlaced with columns of alternating sign. Without further analysis we can deduce that the intensity distribution has to be again a comb function. For N an odd integer, we find N delta functions within an interval of size d. The corresponding diffraction plane is

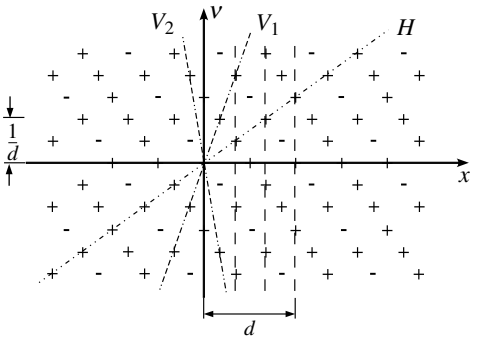


FIGURE 9.7 Fractional Talbot effect of the comb function at $z_T/3$.

straightforwardly calculated from the shearing operation, in Eq. (9.9), with $z_{M,N}\lambda N/(2d) = Md$, or

$$z_{M,N} = \frac{M}{N}\frac{2d^2}{\lambda} \tag{9.22}$$

which we identify as the set of fractional Talbot planes.

For the present we restrict the discussion to cases $M = 1$ and odd numbers N. The WDF in Fig. 9.7 does not strictly resemble a comb function. However, instead of interpreting the distribution of points as a Cartesian grid which was sheared horizontally, we arrive at the same distribution by interpreting this as a (different) Cartesian grid sheared in the vertical direction.

This vertical shear is not unique. In Fig. 9.7 line H corresponds to the WDF of the point response of free space, while V_1 and V_2 illustrate the WDF of two chirp functions which can be used to modulate a comb function of period $d/3$ to obtain the distribution in Fig. 9.7. Any suitable line V has to run through the origin and the location of a δ function with positive sign at $x = d/(2N)$. For N an odd integer, this is only the case for multiples n of $\nu = 1/d$.

We find n by determining the slope of H in Fig. 9.7 as $N/(2d^2)$ and seeking the discrete frequency for which the lateral shift caused by the horizontal shear equals $kd/2 + d/(2N)$, with k being an integer multiplier. This allows us to find $n = (1+N)/4$ and $k = 1$ for $(1+N)/2$ being even, and $n = (1+N)/4$, $k = -1$ for $(1-N)/2$ being even.

This allows us to deduce the phase function of the chirp with the help of Eqs. (9.12) and (9.13) by evaluating the slope of V. As solutions we find

$$2\alpha = \frac{(1+N)N}{2d^2} \tag{9.23}$$

for even numbers $(1 + N)/2$ and

$$2\alpha = \frac{(1 - N)N}{2d^2} \tag{9.24}$$

for even numbers $(1 - N)/2$. We can substitute α back into the chirp signal to obtain the phase of the chirp function

$$\phi(x) = 2\pi\alpha x^2 = \pi\frac{(1 \pm N)N}{2d^2}x^2 \tag{9.25}$$

This modulates a comb function which samples the chirp at equidistant intervals d/N; that is, the sheared distribution of δ functions can be expressed as

$$\begin{aligned} u_{\text{comb}}\left(x, \frac{z_T}{N}\right) &= \frac{1}{\sqrt{N}}\exp[i\phi(x)]\sum_{n=-\infty}^{\infty}\delta\left(x - \frac{nd}{N}\right) \\ &= \frac{1}{\sqrt{N}}\sum_{l=0}^{N-1}\exp(i\phi_l)\sum_{n=-\infty}^{\infty}\delta\left(x - nd - \frac{ld}{N}\right) \end{aligned} \tag{9.26}$$

where we used the fact that the samples

$$\phi_l = \phi\left(\frac{ld}{N}\right) = \pi\frac{1 \pm N}{2N}l^2 \tag{9.27}$$

are N-periodic. The factor $1/\sqrt{N}$ in Eq. (9.26) can be deduced as a consequence of intensity conservation. The result is equivalent to the expressions given in Ref. 14.

We have arrived at a remarkable result, which recognizes the Fresnel diffraction amplitude of a periodic function as the superposition of N replicas of the input function, each replica being laterally shifted by a multiple of d/N and modulated with a constant phase factor. Coefficients $c_l = \exp(i\phi_l)/\sqrt{N}$, called the Talbot coefficients, are obtained as the samples of a chirp function. The importance of using schematic, yet rigorous PSDs to study optical systems is perhaps best illustrated by comparing the analysis of the fractional Talbot effect given in this section with a rather formal application of the WDF to the same problem.[33] While the phase-space analysis in both cases provides a rather compact formulation of the fractional Talbot effect, the analysis assisted by the PSD of comb functions at each step facilitates the interpretation of the phase-space expressions in the signal domain with simple and explicit relationships for both the complex amplitude and the Talbot coefficients.

It is possible to extend the analysis[28] to include all fractional Talbot planes, namely, those with $M \neq 1$ and even numbers N. However, in what follows we will assume a slightly different perspective, which

will allow us to extend the context of our discussion and further highlight the importance of the fractional Talbot.

9.7 Matrix Formulation of the Fractional Talbot Effect

In Sec. 9.6 we made no assumption regarding the groove shape of the grating structure to illustrate that the fractional Talbot effect can be expressed as a superposition of shifted and modulated copies of the grating's transmission function.

We will now consider the case in which each period of the input function can be written as

$$u_p(x) = u_{\text{int}}(x) * \sum_{q=0}^{Q-1} a_q \delta\left(x - \frac{qd}{Q}\right) \tag{9.28}$$

with Q being an integer number. We can interpret Eq. (9.28) as a sampling expansion with $u_{\text{int}}(x)$ defining the interpolation function. In particular, with $u_{\text{int}}(x) = \text{sinc}(xQ/d)$, Eq. (9.28) turns into the well-known Shannon-Whitaker sampling theorem. For $u_{\text{int}}(x) = \text{rect}(xQ/d)$ we obtain a model most suitable to describe binary diffractive optical elements or images with rectangular pixels.

We are again at liberty to explore Fresnel diffraction without specifying the interpolation function explicitly. It is sufficient to investigate Fresnel propagation of the discrete sampled version of the input function. Given our discussion in Sec. 9.6, we expect to find fractional Talbot planes where the diffraction amplitude is described as a superposition of Q copies laterally shifted by a multiple of d/Q. This, in turn, implies that for an input function consisting of modulated δ functions at intervals d/Q, the output function also has to be a modulated comb function with the same spacing between pulses. With the interpolation functions remaining unaffected, each period of the Fresnel diffracted wave can be expressed as

$$u_p(x, z_{M,N}) = u_{\text{int}}(x) * \sum_{q=0}^{Q-1} b_q \delta\left(x - \frac{qd}{Q}\right) \tag{9.29}$$

In other words, we can solve the propagation problem by expressing coefficients b_q as a linear superposition of the coefficients a_q. This, in turn, means we can interpret the a_q and b_q as vectors in a Q-dimensional vector space and determine a linear discrete transformation to relate both sets of coefficients.[14,15] The transformation matrix has to be a function of the respective Talbot coefficients which

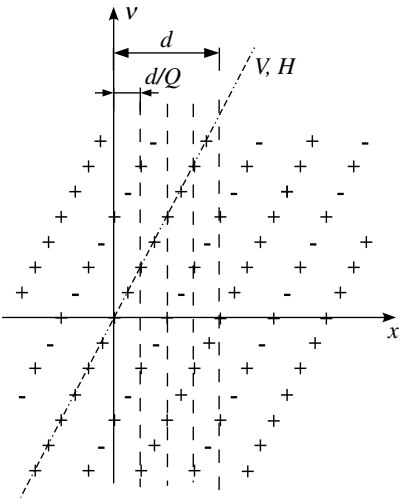


FIGURE 9.8 Fractional Talbot in phase space: comb function at $z_T/8$.

are sufficient to describe diffraction from one fractional Talbot plane to another. Here, we again turn to phase-space optics for constructing the transformation matrix.

We restrict our attention to the case of even numbers Q and argue that this will allow us to obtain the transformation matrix for any fractional Talbot plane. Figure 9.8 illustrates the shear of the comb function for the case $Q = 4$. From this special case it is easy to verify that we obtain exactly Q columns with only positive δ's if we shear the WDF such that the row at $\nu = 1/d$ moves by $d/(2Q)$. This shear corresponds to the fractional Talbot plane

$$z_{1,2Q} = \frac{1}{2Q} z_T \tag{9.30}$$

We can also generalize the fact that lines H and V (as defined in Sec. 9.6) coincide; i.e., the chirp function which can be thought of modulating the modified comb function is

$$\phi(x) = \exp\left(i\pi \frac{Q}{d^2} x^2\right) \tag{9.31}$$

and the Talbot coefficients are samples of this chirp function at $x_q = qd/Q$,

$$c_q = \frac{1}{\sqrt{Q}} \exp\left(i\pi \frac{q^2}{Q}\right) \tag{9.32}$$

We now use the superposition principle and interpret the PSD of the comb function after shearing, in Fig. 9.8, as the PSD of an input signal with $a_0 = 1$ and $a_q = 0$ for $q = 1, \ldots, Q-1$. Then the output signal is the set of Talbot coefficients $b_q = c_q$. Next we investigate the case $a_1 = 1$ and $a_q = 0$ for $q \neq 1$. Thus the PSD in Fig. 9.8 merely has to be shifted in x by d/Q. The same shift has to be considered for the phase function of the modulating chirp, i.e.,

$$\phi(x) = \exp\left[i\pi\frac{Q}{d^2}\left(x - \frac{d}{Q}\right)^2\right] \tag{9.33}$$

Sampling the shifted chirp again at $x_q = qd/Q$ results in a reordering of the Talbot coefficients. We obtain the output coefficients $b_q = c_{q-1}$, where the subscript of the Talbot coefficients needs to be evaluated modulo Q.

This procedure can be repeated by considering all input samples a_q in turn. The solution of the diffraction problem is obtained as a superposition of diffraction amplitudes associated with all input coefficients, i.e.,

$$b'_q = \sum_{q=0}^{Q-1} c_{q'-q} a_q \tag{9.34}$$

where again the subscript $q' - q$ has to be calculated modulo Q. This linear transformation is completely determined by the matrix $\mathrm{C} = \{c_{q'-q}\}$, and Eq. (9.34) can be used to obtain the Fresnel diffraction amplitude in any fractional Talbot plane (M, N). We can extend the matrix description effortlessly to planes $(M, 2Q)$, by applying the linear transformation M times; i.e., the system matrix becomes C^M. Planes defined by odd numbers Q are identical to fractional Talbot planes $(2M, 2L)$, with $L = 2Q$. This implies, however, that we need to consider twice as many samples per period to cover this case with the matrix formalism in Eq. (9.34). This is necessary because for odd Q, the Q copies of the input signal are shifted laterally by $qd/Q + d/(2Q)$.[28,29,34] Doubling of the sampling frequency automatically incorporates this shift. At least for formal explorations this seems to be a small disadvantage if compared to the fact that all fractional Talbot planes are covered with one single compact expression, while other definitions need to distinguish carefully between different cases (see, for example, Ref. 35).

Finally, we can also include planes specified by odd numbers N, which were discussed in Sec. 9.6, by computing the transfer matrix for fractional Talbot planes $(4M, 4N)$. Again, we accept twice as many samples as necessary to describe the diffraction problem in order to apply the formalism without modification. Note, however, that the

oversampling of the signal can be avoided in this case by modifying the expression for the Talbot coefficients in Eq. (9.32).[15]

It may seem that the vector-matrix formulation of the fractional Talbot effect carries a rather severe restriction: The base period has to be expressed by a sampling expansion and the fractional Talbot planes which we can access are linked with the number of samples per period. However, the interpretation as Fresnel propagation of sampled functions dramatically extends the scope of the fractional Talbot effect in general.

In particular, the matrix formalism provides the link between a continuous description of Fresnel diffraction and discrete computations. It is a well-understood fact that numerical simulations of Fresnel diffraction have to be carried out with care, because the pulse response is neither space- nor band-limited. The matrix expression in Eq. (9.34) defines a set of cases where a numerical computation of Fresnel diffraction can be carried out rigorously (see also Refs. 13 and 16). The restriction to periodic functions, in this context, is similar to that of the discrete Fourier transformation which is also used to approximate the corresponding continuous transform.

For numerical applications it is also interesting that C has the structure of a unitary matrix.[15] This means, that the inverse Fresnel transform can be carried out without difficulty as well, which is important for the performance of iterative methods, e.g., Gerchberg-Saxton type of algorithms for phase retrieval and diffractive optics design.[36,37]

9.8 Point Source Illumination

So far we only considered self-imaging of periodic wave fields, which can be interpreted as the result of illuminating a diffraction screen with a coherent plane wave. If diffraction can be described with Kirchhoff's approximation, the complex amplitude at the input plane is completely equivalent to the transmission function of the diffractive element.

A small yet important generalization is the illumination with spherical wavefronts (or parabolic wavefront, if the paraxial approximation is invoked). This means that the strictly periodic wavefront is modulated with the chirp function $h_{\mathrm{Fr}}(x, R)$ in Eq. (9.14). With the sign convention in Eq. (9.14) we obtain a diverging wave for $R > 0$ and a converging wave for $R < 0$.

While the solution of this diffraction problem presents no fundamental difficulty if solved with standard Fourier optics (see, e.g., Ref. 9), it requires some bookkeeping effort, in particular when dealing with the quadratic-phase terms of the Fresnel diffraction integral.

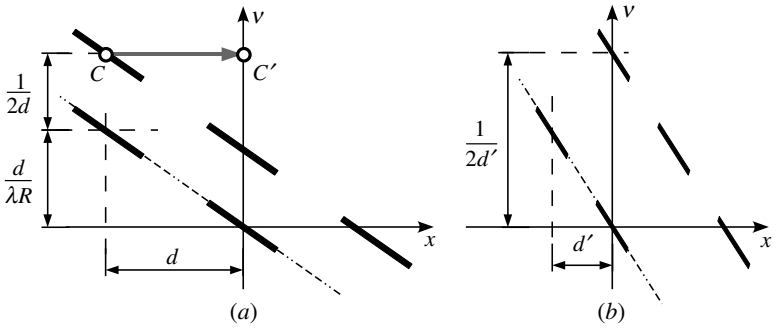


FIGURE 9.9 Self-imaging of a periodic object under spherical illumination: (*a*) input plane and (*b*) first self-imaging plane.

By contrast, the phase-space analysis again only requires basic algebra to obtain all fundamental relationships.

Figure 9.9 shows two cross-terms of the WDF associated with the grating structure. The oblique orientation of the modulated δ lines is the result of a (negative) vertical shear caused by a chirp function with $R < 0$. Both cross-terms are shown with a modulation of period d, which is adequate since we are only interested in identifying the fundamental self-imaging distance. For inspecting intermediate planes, we would again need to assign $d/2$ for the base period of interference terms at every second δ line.

It is immediately clear that we observe self-imaging after a propagation distance that moves point C, in Fig. 9.9*a*, to point C'. This automatically causes the maxima of all other interference terms to line up vertically. The result of this horizontal shearing operation is shown in Fig. 9.9*b*.

The new phase-space distribution has the same qualitative shape as the distribution in the input plane. However, the vertical projection of the WDF indicates a reduced base period d'. The phase-space distribution in Fig. 9.9*b* can again be interpreted as a chirp modulated function, where the radius of curvature is smaller than the one in the input plane.

We can deduce the self-imaging distance from the horizontal shear of point C. Taking the sign convention for R into account, we find

$$\overline{C\,C'} = d = \lambda z_P \left(\frac{1}{2d} - \frac{d}{\lambda R} \right) \tag{9.35}$$

or

$$z_P = \frac{2d^2}{\lambda} \frac{1}{1 - 2d^2/\lambda R} = m_R z_T \tag{9.36}$$

The change of the self-imaging length, compared to the case of plane wave illumination, can be expressed with the help of a magnification factor

$$m_R = \frac{1}{1 - z_T/R} \tag{9.37}$$

For diverging waves $R > 0$, the self-imaging distance is increased, while for converging wavefronts the self-imaging distance is reduced compared to plane wave illumination. The Talbot distance can be recovered for $R \to \infty$.

The modified period of the first self-image can be deduced from the new frequency spacing of the δ lines; i.e., we can write

$$\frac{1}{2d'} = \frac{1}{2d} - \frac{d}{\lambda R} \tag{9.38}$$

or

$$d' = m_R d \tag{9.39}$$

Finally, the new radius of curvature at the first self-imaging plane becomes

$$R' = R + z_p = m_R R \tag{9.40}$$

Additional self-imaging planes can be found by substituting R' back into Eq. (9.36). For a converging wavefront we find self-imaging planes with increasing density along the optical axis, the closer they are located to the focal point of the illuminating spherical wave. At the focal point we expect to see self-imaging replaced by the Fourier spectrum of the grating. In phase space this corresponds to a horizontal shear which turns all δ lines vertical; i.e., the vertical projections will no longer contain any information about the interference terms, but will only show the distribution of discrete self-terms.

The case of self-imaging under spherical illuminations also serves as an example to highlight other important generalizations of Talbot self-imaging. For investigating diverging and converging wavefronts, we had to drop the requirement of obtaining a perfect replica of the complex amplitude. Instead, we accepted a scaled replica as a generalized self-image. It should be mentioned that the geometric scaling under spherical illuminations also applies to the fractional Talbot effect and was used to design Talbot array illuminators.[38]

A further generalization is self-imaging in arbitrary *ABCD* optical systems. Fresnel diffraction from a grating under spherical illumination is equivalent to plane wave illumination of the grating followed by a system consisting of a parabolic lens and free-space propagation. In fact, conditions for obtaining self-images in arbitrary *ABCD* systems have already been investigated.[39]

A further variant is the Fourier dual case, where the input signal is an array of discrete equidistant pulses and the system is composed of only a parabolic lens. Then it is possible to describe the equivalent of self-imaging in the frequency domain, which has been termed *spectral self-imaging*,[40] which was shown to be of interest for time-domain signals and optical fiber communication.

9.9 Another Path to Self-Imaging

So far we have assumed a periodic signal and explored the astounding richness of Fresnel diffraction as a consequence of the signal's periodicity. We now change our perspective by using the self-imaging phenomenon as the starting point. The task, then, is to identify those signals for which self-imaging can be observed. In other words, instead of analyzing a specific set of signals to discover self-imaging, we now consider self-imaging as a given phenomenon and construct signals with this property.

This design-oriented approach was suggested by Montgomery,[8] in 1967 when seeking a necessary condition for signals to exhibit self-imaging. The construction again can be carried out straightforwardly in phase space.

We assume the signal to be composed of discrete frequencies. The corresponding self-terms of the WDF all have the form of a δ line parallel to the x axis. This means that the information about permissible spatial frequencies has to be obtained from the associated cross-terms, all having the form

$$W_{n,m}(x, \nu) = 2a_n a_m^* \cos\left[2\pi(\nu_n - \nu_m)x\right] \delta\left(\nu - \frac{\nu_n + \nu_m}{2}\right) \tag{9.41}$$

where n and m are integer numbers for labeling the spatial frequencies and a_n and a_m are the associated Fourier amplitudes. Assuming $m = 0$ and $\nu_0 = 0$, we can now use Eq. (9.41) to establish the condition for self-imaging. The horizontal shift associated with Fresnel diffraction has to be a multiple of the frequency $\nu_n - \nu_m$ which modulates each cross-term. Without loss of generality we assume $m = 0$ to find

$$\lambda z_M \frac{\nu_n}{2} = \frac{n}{\nu_n} \tag{9.42}$$

from which follows the set of permissible frequencies

$$\nu_n = \sqrt{\frac{2n}{\lambda z_M}} \tag{9.43}$$

Equation (9.43) can be recognized as the Montgomery condition for paraxial signals. Figure 9.10 illustrates the case of three discrete frequencies and the associated cross-terms. It is a simple exercise to show

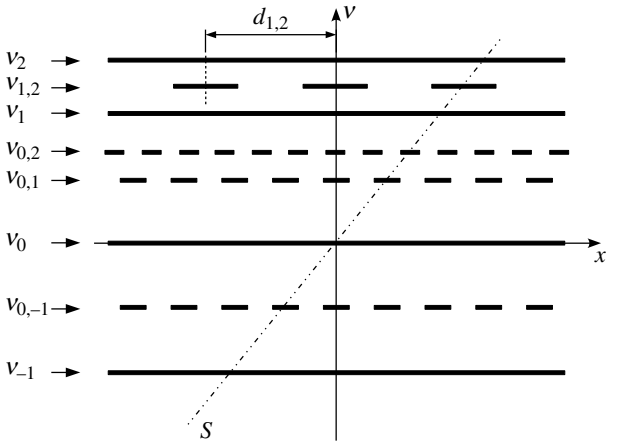


FIGURE 9.10 Self-imaging based on the Montgomery condition.

that the condition in Eq. (9.43) automatically fulfills the self-imaging condition for all other cross-terms with $m \neq 0$.

9.10 Self-Imaging and Incoherent Illumination

Phase-space optics allows one to describe both coherent and partially coherent signals with one consistent formalism. While it is not the purpose of this discussion to explore partially coherent optics in phase space, the study of self-imaging conditions allows us to take a sneak peak at how problems involving incoherent signals can be addressed.

To this end we consider the setup in Fig. 9.11, which is a double-grating configuration. The first grating G_1 is illuminated with incoherent quasi-monochromatic light. At distance z_L a second grating is located in the front focal plane of a Fourier lens f. This is one of various systems for studying the Lau effect.[24] The first grating G_1 serves as a periodically modulated incoherent light source, which is used to illuminate a the second grating G_2 located at some distance z_L along the optical axis. For a discrete set of grating separations it is possible to observe a pattern of high-contrast fringes in the Fraunhofer diffraction plane of the 2-f system.

The Lau effect easily rivals the Talbot effect in terms of its mathematical beauty and its potential for applications. Perhaps due to its slightly higher complexity, however, the Lau effect has unquestionably received much less attention than coherent self-imaging.

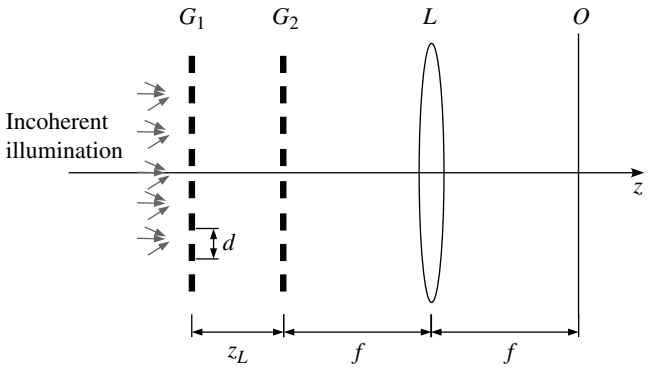


FIGURE 9.11 Setup for observing the Lau effect.

The discussion of the Lau effect focuses on the phase-space interpretation and is aimed at establishing the basic condition for observing Lau fringes. To this end we return to Fig. 9.4, the PSD of a periodic (coherent) signal. If grating G_2 were illuminated with a totally incoherent light signal, no features would be observed in the output plane of the 2-f system. Each grating line would act as a separate incoherent light source with a homogeneous far-field diffraction pattern, and incoherent superposition of all intensities would not show any fringe pattern. Thus the first grating G_1 plus free-space propagation acts as a system to modify the coherence properties of the source, ensuring the formation of a far-field fringe pattern.

The phase-space interpretation allows us to compute the far-field intensity by first determining the intensity for a point source in the plane of grating G_1 followed by the convolution with the source distribution in that plane. This resembles the procedure we would apply by using elementary coherence theory. For the phase-space analysis it is advantageous that this incoherent signal summation is a linear operation (i.e., bilinear interference terms vanish).

The result of illuminating grating G_2 with a point source was discussed in Sec. 9.8. The incident wave is described by a chirp, and the WDF of a periodic structure is convolved with the δ line of the chirp signal. To analyze the Lau effect, however, we do not consider near-field diffraction, but a Fourier transformation corresponding to a rotation of the WDF by 90°. In fact, we do not even need to execute this rotation explicitly, but it is sufficient to consider the projection of the WDF along the x axis to obtain the power spectrum (i.e., the desired intensity of the far-field diffraction amplitude).

Then the Lau condition corresponds to a vertical shear of the phase-space distribution for which we can identify a distinct modulation of the far-field patter. Figure 9.12 shows the distribution in Fig. 9.4 after a

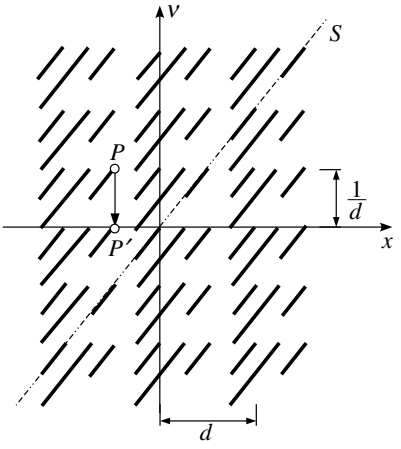


FIGURE 9.12 PSD of a periodic function under spherical illumination [radius $z = d^2/(2\lambda)$].

vertical shear. The representation includes the frequency doubling of the terms at multiples of the base frequency interval $1/d$. A modulation in the horizontal projection can be ensured if the maxima of all terms register in rows parallel to the x axis.

The first occurrence of this condition corresponds to a shear that moves point P to point P'. This means that the point with coordinate $x = -d/2$ is moved in frequency $\Delta\nu = -1/d$. From Eq. (9.10) we can deduce the radius of the corresponding spherical wave as

$$z_L = \frac{d^2}{2\lambda} \tag{9.44}$$

which is the well-known Lau condition for observing high-contrast fringes in the far field.[24]

We can now consider the convolution with the source distribution. Note that the phase of the far-field modulation depends on the transverse source location, which would determine the interference between different source points for the case of coherent illumination. For an incoherent source, this mutual phase shift between source points is irrelevant, however.

In Fig. 9.12 line S refers to the WDF of the point source, and its intersection with the x axis marks the center of the shear which is applied to the phase-space distribution. Thus as the source moves in x, the sheared distribution in Fig. 9.12 moves vertically, and for a shift of the source by one grating period d, the phase-space distribution has moved vertically by $2/d$.

For a continuous incoherent source illuminating grating G_2, we effectively integrate the phase-space distribution in the vertical direction, with the obvious result that no fringe pattern can be observed in the far field. For a spatially inhomogeneous, but continuous source, each shifted copy of the phase-space distribution is weighted with the respective source strength, and we obtain as the far-field intensity the expected convolution (or correlation) between the distribution for a single source point and the source distribution.

For the case where G_1 consists of an array of equidistantly spaced pinholes with period d, all copies again register perfectly in the vertical direction, and we observe the Lau effect for quasi-monochromatic illumination. In this context it is again emphasized that incoherent signal superposition is a linear operation in phase space, and no cross-terms are observed. This also means that the PSD of an array of point sources (at G_1) corresponds to an array of vertical lines, rather than the two-dimensional distribution of δ functions, in Fig. 9.6, which represent an array of mutually coherent pulses.

While we will not attempt a more detailed analysis of the Lau effect, we note that condition [Eq. (9.44)] is not the only configuration for which far-field fringes can be observed. In fact it follows from our analysis that the fringes can also be observed if grating G_1 is composed of narrow slits with a period $d/2$.

A more general analysis was performed by using the formalism of coherence theory predicting fringe patterns at $z_L = (M/N)d^2/\lambda$, with M and N being relative prime integer numbers.[25] Coherence theory has a formal interpretation in phase space,[41,42] and it was shown that this can also be used for a more rigorous derivation of the conditions for observing the Lau effect.[27]

9.11 Summary

Not all optical phenomena find a natural interpretation in phase space. Self-imaging, however, is exceptionally well suited to be analyzed with the help of the WDF and PSDs. In part, this is due to the strict periodicity of the signals. As a consequence, the WDF is discrete in at least one of its coordinates.

Not only did the use of phase-space optics allow us to find a qualitative interpretation, but more importantly we were able to perform a quantitative analysis requiring only a minimum of mathematical formalism. Pivotal to this analysis is the use of PSDs as mediators between the mathematical formalism of the WDF and an intuitive interpretation of physical optics.

The discussion presented in this chapter was aimed at recovering well-known relationships. However, we never even came close

to exhausting the use of phase-space optics for exploring the self-imaging phenomenon. A number of related phenomena still await a thorough phase-space interpretation. It would be even more exciting, however, if the phase-space interpretation helped us to discover new self-imaging phenomena and applications not obvious from a Fourier optics perspective.

References

1. M. Bastiaans, "Application of the Wigner distribution function in optics," in W. Mecklenbräuker and F. Hlawatsch (eds.), *The Wigner Distribution—Theory and Applications in Signal Processing*, Elsevier, Amsterdam, 1997, pp. 375–426.
2. A. Torre, *Linear Ray and Wave Optics in Phase Space*, Elsevier, Amsterdam, 2005.
3. M. Testorf, J. Ojeda-Castañeda, and A. W. Lohmann (eds.), *Selected Papers on Phase-Space Optics*, volume MS 181 of *SPIE Milestone Series*. SPIE, Bellingham, Wash., 2006.
4. R. Castaneda, "Phase space representation of spatially partially coherent imaging," *Appl. Opt.* **47**: E53–E62 (2008).
5. B. E. A. Saleh and M. C. Teich, *Fundamentals of Photonics*, Wiley, New York, 1991.
6. H. F. Talbot, "Facts relating to optical science. No IV," *Philosophical Mag., J. Sci.*, **9**: 401–407 (1836).
7. Lord Rayleigh, "On copying diffraction gratings, and on some phenomena connected therewith," *Philosophical Mag.*, **11**: 196–205 (1881).
8. W. D. Montgomery, "Self-imaging objects of infinite aperture," *J. Opt. Soc. Am.* **57**: 772–778 (1967).
9. K. Patorski, "The self-imaging phenomenon and its applications," in E. Wolf (ed.), *Progress in Optics*, vol. XXVII, Elsevier Science, Amsterdam, 1989, pp. 1–108.
10. J. T. Winthrop and C. R. Worthington, "Theory of Fresnel images. I. Plane periodic objects in monochromatic light," *J. Opt. Soc. Am.* **55**: 373–381 (1965).
11. J. P. Guigay, "On Fresnel diffraction by one-dimensional periodic objects, with application to structure determination of phase objects," *Opt. Comm.* **18**: 677–682 (1971).
12. J. Westerholm, J. Turunen, and J. Huttunen, "Fresnel diffraction in fractional Talbot planes: A new formulation," *J. Opt. Soc. Am. A* **11**: 1283–1290 (1994).
13. H. Hamam, "Simplified linear formulation of Fresnel diffraction," *Opt. Comm.* **144**: 89–98 (1997).
14. V. Arrizón and J. Ojeda-Castañeda, "Fresnel diffraction of substructured gratings: Matrix description," *Opt. Lett.* **20**: 118–120 (1995).
15. V. Arrizón, J. G. Ibarra, and J. Ojeda-Castañeda, "Matrix formulation of the Fresnel transform of complex transmittance gratings," *J. Opt. Soc. Am. A* **13**: 2414–2422 (1996).
16. S. B. Tucker, J. Ojeda-Castañeda, and W. T. Cathey, "Matrix description of near field diffraction and the fractional Fourier transform," *J. Opt. Soc. Am. A* **16**: 316–322 (1999).
17. A. W. Lohmann, "An array illuminator based on the Talbot effect," *Optik* **79**: 41–45 (1988).
18. A. W. Lohmann and J. A. Thomas, "Making an array illuminator based on the Talbot effect," *Appl. Opt.* **29**: 4337–4340 (1990).
19. J. R. Leger and G. J. Swanson, "Efficient array illuminator using binary-optics phase plates at fractional Talbot planes," *Opt. Lett.* **15**: 288–290 (1990).
20. V. Arrizón and J. Ojeda-Castañeda, "Multilevel phase gratings for array illuminators," *Appl. Opt.* **33**: 5925–5931 (1994).

21. T. J. Suleski, "Generation of Lohmann images from binary-phase Talbot array illuminators," *Appl. Opt.* **36**: 4686–4691 (1997).
22. W. Klaus, Y. Arimoto, and K. Kodate, "High performance Talbot array illuminators," *Appl. Opt.* **37**: 4357–4365 (1998).
23. M. Testorf, V. Arrizón, and J. Ojeda-Castañeda, "Numerical optimization of phase-only elements based on the fractional Talbot effect," *J. Opt. Soc. Am. A* **16**: 97–105 (1999).
24. J. Jahns and A. W. Lohmann, "The Lau effect (a diffraction experiment with incoherent light)," *Opt. Comm.* **28**: 263–267 (1979).
25. F. Gori, "Lau effect and coherence theory," *Opt. Comm.* **31**: 4–8 (1979).
26. J. Jahns, A. W. Lohmann, and J. Ojeda-Castañeda, "Talbot and Lau effects, a parageometrical approach," *Optica Acta* **31**: 313–324 (1984).
27. J. Ojeda-Castañeda and E. E. Sicre, "Quasi ray-optical approach to longitudinal periodicities of free and bounded wavefields," *Optica Acta* **32**: 17–26 (1985).
28. M. Testorf and J. Ojeda-Castañeda, "Fractional Talbot effect: Analysis in phase space," *J. Opt. Soc. Am. A* **13**: 119–125 (1996).
29. M. Testorf, "Designing Talbot array illuminators with phase-space optics," *J. Opt. Soc. Am. A* **23**: 187–192 (2006).
30. A. W. Lohmann, *Optical Information Processing*, Universitätsverlag Illmenau, Germany, 2006.
31. D. E. Silva, "Talbot interferometer for radial and lateral derivatives," *Appl. Opt.* **11**: 2613–2624 (1972).
32. E. Keren and O. Kafri, "Diffraction effects in moiré deflectometry," *J. Opt. Soc. Am. A* **2**: 111–120 (1985).
33. K. Banaszek, K. Wódkiewicz, and W. Schleich, "Fractional Talbot effect in phase-space: A compact summation formula," *Opt. Express* **2**: 169–172 (1998).
34. K. Patorski, "Self-imaging phenomenon, lateral shift of Fresnel images," *Optica Acta* **30**: 1255–1258 (1983).
35. V. Arrizón, G. Rojo-Valázquez, and J. G. Ibarra, "Fractional Talbot effect: Compact description," *Opt. Rev.* **7**: 129–131 (2000).
36. Z. Zalevsky, D. Mendlovic, and R. G. Dorsch, "Gerchberg-Saxton algorithm in the fractional Fourier or the Fresnel domain," *Opt. Lett.* **21**: 842–844 (1996).
37. R. G. Dorsch, A. W. Lohmann, and S. Sinzinger, "Fresnel ping-pong algorithm for two-plane computer-generated hologram display," *Appl. Opt.* **33**: 869–875 (1994).
38. H. Hamam, "Design of array illuminators under spherical illumination," *Appl. Opt.* **37**: 1393–1400 (1998).
39. C. R. Fernández-Pousa, M. T. Flores-Arias, C. Bao, M. V. Pérez, and C. Gómez-Reino, "Talbot conditions, Talbot resonators, and first-order systems," *J. Opt. Soc. Am. A* **20**: 638–643 (2003).
40. J. Azaña, "Spectral Talbot phenomena of frequency combs induced by cross-phase modulation in optical fibers," *Opt. Lett.* **30**: 227–229 (2005).
41. M. J . Bastiaans, "The Wigner distribution function of partially coherent light," *Optica Acta* **28**: 1215–1224 (1981).
42. K.-H. Brenner and J. Ojeda-Castañeda, "Ambiguity function and Wigner distribution function applied to partially coherent imagery," *Optica Acta* **31**: 213–223 (1984).

This page intentionally left blank

CHAPTER 10

Sampling and Phase Space

Bryan M. Hennelly

National University of Ireland, Maynooth, Ireland

John J. Healy and John T. Sheridan

University College Dublin, Ireland

10.1 Introduction

Sampling a continuous signal in order to represent or approximate it with a discrete one is of enormous importance in today's digital world. In the optical sciences we are often interested in recording optical signals with discrete photosensitive devices such as CCD or CMOS cameras. Such devices are sensitive to the intensity of an incident optical wave field and bring about the spatial sampling of this intensity pattern. By using interferometry it is possible to recover phase information from the recorded intensity pattern, and so we may say that we are effectively sampling the complex wavefront with our digital camera. The operation of discrete display devices, such as liquid crystal displays (LCDs) and electrically addressed spatial light modulators (SLMs), are also governed by sampling theory and are of increasing interest in diffractive optics. Thus, the discrete signal processing of digitally captured data plays a central role in modern optoelectronics, and this science is anchored in sampling theory. In the past decade the Wigner distribution function (WDF) and, moreover, its simplified version, the phase-space diagram (PSD), have been shown to be effective tools in gaining considerable insight into the discrete sampling of signals. Not only does the PSD elegantly account for known sampling

theorems, but also it paves the way for rich new theorems and algorithm designs. This is the subject of this chapter.

Until recently, sampling of electromagnetic signals was performed primarily using the theorems devised in the first half of the last century by Nyquist,[1] Shannon[2] Whittaker,[3] and Kotelnikov.[4] We note that the sampling theory of diffraction patterns[5] was introduced by Francia in 1955. Modern sampling theory has evolved far beyond Nyquist-Shannon sampling, and a thorough account of contemporary sampling theory and discrete signal processing can be found in Refs. 6 and 7. In this chapter we focus primarily on Shannon sampling (and a recently generalized derivative of it) and its simplified description using phase space. In this chapter we limit the scope of the discussion to classical sampling. The rules of classical sampling and interpolation may be summarized as follows: It is assumed that the continuous signal is band-limited in frequency. Sampling this signal with a sampling rate at least as fast as twice the maximum frequency (the Nyquist rate) allows for the continuous signal to be "interpolated" from the discrete sampled values by applying the Shannon interpolation formula. This can be explained using the Fourier transform.[8–11] Sampling creates an infinite number of copies of the signal's Fourier transform, all adjacent along the frequency axis, with a separation equal to the inverse of the sampling interval. If the sampling rate is high enough, the copies will not overlap with one another due to their assumed finite support. The continuous signal may be interpolated by isolating one of these copies, achieved by multiplying by an appropriate rect function, which amounts to convolving the sampled signal with a sinc function. This convolution is known as Shannon interpolation.[2]

Recently there has been considerable interest[12–25] in the literature on the subject of sampling certain optical signals at rates below the Nyquist rate and still managing to interpolate the continuous signal. To the best of our knowledge, the first demonstration of this was by Gori[12] in 1981 in which he investigated the sampling of Fresnel diffraction patterns. This collective body of work[12–25] has demonstrated that the requirement of imposing the property of band-limitedness in the Fourier domain is too strict a requisite for interpolation to be achievable. It is sufficient that the signal be bounded in any one of an infinite set of domains which are output domains of the linear canonical transform (LCT).[13,19–21,23–33] The Fourier transform[8–11] and the Fresnel transform[9,12,14,18] are special cases of the LCT. If the signal is bounded within some finite support in such a domain, then the signal can be sampled at a rate, proportional to the finite LCT support width. The Nyquist sampling rate, which is proportional to the Fourier support width (bandwidth), is merely a special case of this more generalized sampling theorem. Importantly, the phase-space investigation of these concepts that follows unearths an interesting insight that

this generalized sampling can be based entirely on the assumption of a chirped signal.

It is also possible to deduce a more general interpolation formula. This amounts to multiplying the signal samples by an appropriate chirp function (the scale of the chirp is dependent on the LCT domain in which the signal has finite support), followed by standard Shannon interpolation; and this is in turn followed by multiplying by the conjugate of the fore-mentioned chirp. If a signal has finite support in some LCT domain, the generalized sampling theorem predicts a finite sampling rate, although this is associated with a signal that does not exhibit a band limit in a classical sense. In this chapter we show how phase-space diagrams allow us to understand and interpret generalized sampling in a most elegant manner as well as to calculate specifics such as the most appropriate sampling rate for a given LCT bounded signal. The generalized sampling theorem is of great importance for digital holography.[22,34–42] It is of considerable interest to this research area because it implies that one may place the object to be recorded at a distance much closer to the camera than previously predicted by the Nyquist-Shannon theorem. A shorter distance between object and camera implies a greater numerical aperture, which in turn should allow reconstruction of the object at a resolution previously thought impossible and which is greater than the resolution of the recording CCD. Furthermore the reduced camera-object distance implies that a far greater range of three-dimensional perspectives may be reconstructed as a result.

To the best of our knowledge, the Wigner distribution function (WDF),[27,43–47] was introduced to the optical community by Bastiaans[45] in 1979, and since then it has found application in describing numerous applications to which this book is a testament. The WDF transforms a one-dimensional spatial signal into a two-dimensional space–spatial frequency distribution. Besides being bilinear, the WDF has a number of rich properties that are shared with both the spatial representation of the signal and its Fourier transform. For example, the integral projections of the WDF along the k and x axes yield the space and frequency marginals, respectively. In the proceeding section we review those properties of the WDF that are of interest in the context of this chapter. One very useful method of visualizing the WDF and conceptualizing its properties is by using Wigner charts or phase-space diagrams[48–52] (PSDs). These PSDs were popularized by Lohmann, Mendlovic, and Zalevsky in graphically describing signal propagation through quadratic-phase systems (QPSs) and also superresolution systems. PSDs are plan view style diagrams of a signal's WDF. They do not include any information about the actual values of the WDF other than the support in the xk plane. By endowing the PSD with many of the properties of the WDF, such as the convolution property, we can conceptualize many optical processes. In some

cases it is even useful not to apply some WDF properties (in particular, the property of bilinearity) to the PSD in order to simplify our understanding. Such omissions must be done with care and with good reason. In Sec. 10.2 we discuss the PSD in greater detail, and then we demonstrate the application of the PSD in understanding sampling theory and in simulating optical systems.

The central theme of this chapter is to show that the WDF and the PSD are useful tools in understanding the sampling of signals with an LCT of finite support. This subject can be further complicated if we consider signals which are sampled, then transformed by a LCT, and then sampled again. The topic is of considerable interest because it is central to the numerical implementation (or simulation) of optical processes.[34–42,53–88] The volume of publications on the subject in the last 10 years highlights its relevance to contemporary optics as does the industrial application of these algorithms in today's optoelectronic world. The double sampling considerably complicates matters, and it forces us to consider sampling criteria of a signal in two transformation domains sequentially. The first sampling operation considerably affects the second, and vice versa; i.e., the second sample must be considered as also shaping the first sampling operation. In this case a new type of aliasing can be encountered which is discussed for the first time in this chapter. Again we find that the most intuitive approach to this subject is through the PSD.

This chapter is broken down as follows. In Sec. 10.2 we discuss some initial concepts that are utilized in the following sections. In Sec. 10.3 we review how a signal can have a finite support in some LCT domain. In Sec. 10.4 we discuss how Nyquist sampling and generalized sampling may be discussed both qualitatively and quantitatively in an elegant fashion using the WDF and PSD. In Sec. 10.5 we progress to discuss sampling of a signal in two domains for the purposes of simulating a quadratic-phase system, and finally in Sec. 10.6 we offer a brief conclusion.

10.2 Notation and Some Initial Concepts

10.2.1 The Wigner Distribution Function and Properties

The WDF is a *time-frequency distribution* and is mathematically defined in terms of this spatial (x) distribution as follows

$$\psi\{u(x)\}(x,\ k) = \int_{-\infty}^{\infty} u\left(x - \frac{\xi}{2}\right) u^*\left(x + \frac{\xi}{2}\right) \exp(-j2\pi k\xi)\, d\xi = W_{uu^*}(x, k) \tag{10.1}$$

where k represents the spatial frequency, * denotes complex conjugation, and $\psi\{u(x)\}(x, k)$ denotes the WDF operator. The WDF can also be equivalently defined in terms of the Fourier transform (FT) of $u(x)$, which is denoted $U(k)$.

$$\psi\{u(x)\}(x,\ k) = \psi\{U(k)\}(x, k) = \int_{\infty}^{\infty} U\left(k - \frac{\xi}{2}\right) U^*\left(k + \frac{\xi}{2}\right) \times \exp(+j2\pi k\xi)\, d\xi \tag{10.2}$$

$$U(k) = \int_{-\infty}^{\infty} u(x) \exp(-j2\pi kx)\, dx \tag{10.3}$$

The real-valued WDF of a function has double the number of dimensions of the function. To find the intensity $I(x) = |u(x)|^2$, we integrate $\psi\{u(x)\}(x, k)$ over k; similarly, to find the spatial frequency distribution $\tilde{I}(k) = |U(k)|^2$, we integrate over x. The WDF is real-valued and it is reversible with the exception of a constant phase factor. The WDF of a shifted signal is given by a simple shift $\psi\{u(x - \beta)\}(x, k) = \psi\{u(x)\}(x - \beta, k)$. Similarly, if we multiply $u(x)$ by a harmonic function $\exp(j2\pi\alpha x)$, the resultant WDF is shifted in k, $\psi\{u(x) \exp(+j2\pi\alpha x)\}(x, k) = \psi\{u(x)\}(x, k - \alpha)$. If two signals u and h are convolved along the x axis, the WDF of the resultant signal is given by the convolution of the individual WDFs along the same x axis. Conversely, if two signals u and v are multiplied in the x domain, the WDF of the resultant signal is given by the convolution of the individual WDF along the same k axis. The WDF of a convolution and a product are given by Eqs. (10.4) and (10.5) respectively,

$$\psi\left\{\int u(\xi)\dot{v}(x - \xi)\, d\xi\right\}(x,\ k) = \int \psi\{u(x)\}(x - x', k)\psi\{v(x)\}(x', k)\, dx' = \psi\{u(x)\}(x, k) *^x \psi\{v(x)\}(x, k) \tag{10.4}$$

$$\psi\{u(x)v(x)\}(x, k) = \int \psi\{u(x)\}(x, k - k')\psi\{v(x)\}(x, k')\, dk' = \psi\{u(x)\}(x, k) *^k \psi\{v(x)\}(x, k) \tag{10.5}$$

In the above equations we have introduced the notation $*^x$ and $*^k$ to denote convolution along the x and k axes, respectively. The similarities between these properties and the convolution property of the FT are obvious. Another property of the WDF is that it is bilinear;

$$\psi\{u(x) + v(x)\}(x, k) = W_{uu^*}(x, k) + W_{vv^*}(x, k) + W_{uv^*}(x, k) + W_{vu^*}(x, k) \tag{10.6}$$

10.2.2 The Linear Canonical Transform and the WDF

A property of central importance in this chapter is the relationship of the WDF to the LCT. When an optical signal is input to a QPS, the LCT describes the relationship between the signal at the output and input to the system. The parameters of the LCT depend on the type of system. QPSs are systems made up of any number of sequential thin lenses and free space as well as many other lossless optical elements for which the paraxial approximation is valid. The LCT also has meaning in quantum mechanics. The LCT is mathematically defined as

$$u_M(x') = L_M\{u(x)\}(x') = \frac{\exp(-j\pi/4)}{\sqrt{B}} \times \int_{-\infty}^{\infty} u(x) \exp\left[j\pi\left(\frac{A}{B}x^2 - \frac{2}{B}xx' + \frac{D}{B}x'^2\right)\right] dx \quad (10.7)$$

where $L_M\{u(x)\}(x')$ is the operator notation for the LCT and M is a matrix that contains the parameters of the LCT

$$\begin{pmatrix} x' \\ k' \end{pmatrix} = M \begin{pmatrix} x \\ k \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} x \\ k \end{pmatrix} \quad (10.8)$$

where $AD - BC = 1$. This is simply the ray transfer matrix that is commonly applied in geometrical optics. It maps the position and angle of an input ray to those of the output. Collins[26] first pointed out the relationship between the ray transfer matrix and the LCT. Remarkably, this relationship can be extended to include the WDF as defined in Eq. (10.9).

$$\psi\{u_M(x')\}(x', k') = \psi\{u(x)\}(Ax + Bk, Cx + Dk) \quad (10.9)$$

Therefore, if an LCT is applied to a signal, the WDF of the signal undergoes a simple coordinate transformation. This operation is *affine*,[29–31] meaning that a given area on the WDF plane is conserved under this coordinate shift. A noticeable and very useful property of the LCT-WDF matrix relationship is that the combined matrix of several optical systems placed in series, each with its own matrix, can be found by multiplying the individual system matrices. Thus rather than calculate a series of LCTs to determine the output of the constituent subsystems, a single LCT can be determined that approximates the entire system.

10.2.3 The Phase-Space Diagram

The PSD is an illustrative plan-view outline of the WDF of a signal. This diagrammatic approximation can be a very useful source of

insight when we bestow it with many of the WDF properties. For example, a signal with a finite bandwidth might have a PSD shown in Fig. 10.1.

In Fig. 10.1*a* we show the PSD of a signal that has a finite bandwidth. It is well known that in the strict mathematical sense any signal that has such a property must, as a consequence, have infinite spatial support. However, in many practical cases we assume that the signal has an approximately finite support in both domains, and the PSD of such a signal is shown in Fig. 10.1*b*. The subject of finite support is addressed in greater detail in Sec. 10.3. The PSD that is bounded in both x and k is much easier to use for illustrating some of the properties of the WDF. For example, the application of the FT to our signal brings about a 90° rotation of the WDF and therefore the PSD as shown in Fig. 10.1*c*. The Fresnel transform causes a horizontal shearing of the

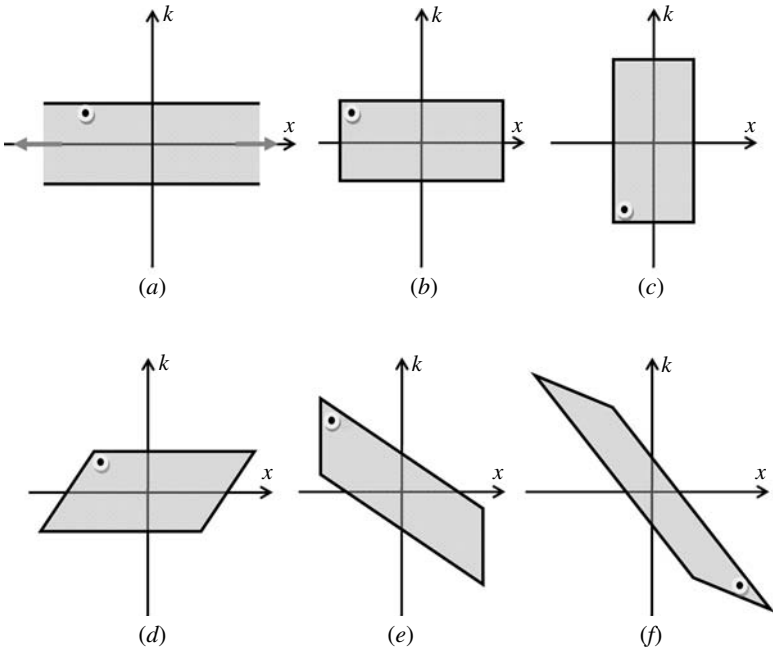


FIGURE 10.1 The PSD of a signal (*a*) having finite bandwidth and therefore an infinite spatial support, (*b*) having approximately finite spatial support and finite bandwidth, (*c*) which is a Fourier transform of signal represented in part *b*, (*d*) which is a Fresnel transform of the signal represented in part *b*, (*e*) which is a signal represented in part *b* after being multiplied by a chirp function, and (*f*) which is a linear canonical transform of the signal represented in part *b*.

signal's WDF. In Fig. 10.1*d* we illustrate the PSD of the Fresnel transformed signal. In Fig. 10.1*e* we show the PSD of a chirp modulated signal. This is what happens when the signal passes through an ideal thin lens. Multiplication by the chirp causes a vertical shearing of the WDF. All these linear transformations that effect some change on the WDF are special cases of the LCT. They all have matrices associated with them that map each x-k coordinate on the WDF (and PSD) to a new position. This coordinate shift is defined in Eqs. (10.8) and (10.9). It is very important to note that all these mappings are affine; the shaded area inside the PSD is conserved under the mapping. In Fig. 10.1*f* we show the PSD after the signal has been transformed by an arbitrary LCT. We also note that in the case of the x-k bounded signal shown in Fig. 10.1*b* the area of the PSD is exactly equal to the number of samples required to represent the signal in the Nyquist limit. In the next section we describe some more properties of the WDF and PSD that are used in later sections.

10.2.4 Harmonics and Chirps and Convolutions

The WDF of a harmonic function $\exp(+j2\pi k_0 x)$, which in optics represents a plane wave with wavelength λ traveling at an angle $\theta = \sin^{-1}(\lambda k_0)$, has a WDF given by $\delta(k - k_0)$, where δ represents the Dirac delta function. The PSD for this harmonic is shown in Fig. 10.2*a*. The arrows indicate that it extends over infinity in x. Similarly a point source at a position x_0 has a WDF given by $\delta(x - x_0)$. This is a further example of the FT bringing about a 90 degree rotation of the WDF.

It is well known from Fourier theory that if a signal is modulated by a harmonic function, it is shifted in the frequency domain. The same

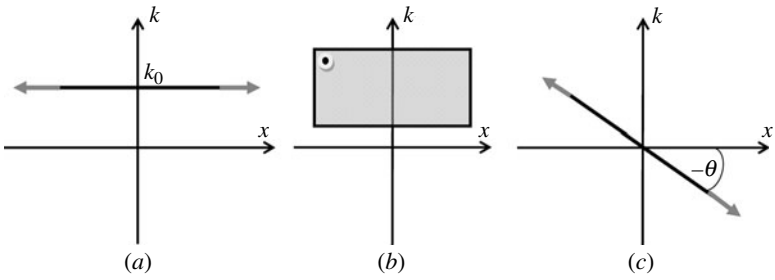


FIGURE 10.2 PSD of (*a*) a harmonic function, (*b*) signal represented in Fig. 10.1*b* after being multiplied by the harmonic in Fig. 10.2*a* and (*c*) a chirp signal.

effect may be observed using the WDF. When the signal with PSD shown in Fig. 10.1*b* is multiplied by the harmonic with PSD shown in Fig. 10.2*a* the resultant PSD is given by the convolution of the two along the k axis as defined in Eq. (10.5). The result of this convolution is an exact replica of the signal's WDF centred at $k = k_0$. Chirp functions are mathematically expressed as $\exp(+j\pi\alpha x^2)$. In the paraxial approximation such functions represent spherical waves with curvature α. The effect of a thin lens with focal length f is modeled by multiplying by a chirp function where $\alpha = 1/\lambda f$. In the case of convex and concave lenses α is negative and positive, respectively. The WDF of the chirp signal can be shown to be $\delta(k - \alpha x)$. Thus at any point x only one local frequency exists at $k = \alpha x$. The PSD for the chirp signal is shown in Fig. 10.2*c*. In this case $\alpha = 1/\tan(-\theta)$. Again the arrows indicate that this Dirac delta line extends outward infinitely.

From Eq. (10.4) we know that if a signal is convolved in space with a chirp $\exp(+j\pi\alpha x^2)$, their WDFs are also convolved along x. Hence the PSD shown in Fig. 10.1*d*, is given by the convolution along x of the PSDs shown in Fig. 10.1*b* and Fig. 10.2*c*. This can be interpreted as the paraxial approximation of spherical waves with curvature α. This is actually equivalent to a LCT with $A = D = 1, B = 1/\alpha$, and $C = 0$. Therefore if a signal is convolved with a chirp function $\exp(j\pi\alpha x^2)$, the signal's WDF undergoes the following coordinate transformation.

$$\psi\{u(x)\}(x, k) \rightarrow \psi\{u(x)\}(x + k/a, k) \tag{10.10}$$

This follows from Eq. (10.9) and is known as a horizontal shearing. We note that as $\alpha = 1/\lambda z$ in such a convolution, the result is the Fresnel transform for a distance z.

Similarly from Eq. (10.5) we can see that if a signal is multiplied in space with a chirp $\exp(j\pi\alpha x^2)$, their WDFs are convolved along k. Hence the PSD shown in Fig. 10.1*e* is given by the convolution along k of the PSDs in Fig. 10.1*b* and Fig. 10.2*c*. Again it is equivalent to a LCT, this time with $A = D = 1, C = 1/\alpha$, and $\beta = 0$. Therefore if a signal is multiplied with a chirp function $\exp(j\pi\alpha x^2)$, the signal's WDF undergoes the following coordinate transformation.

$$\psi\{u(x)\}(x, k) \rightarrow \psi\{u(x)\}\left(x, k + \frac{x}{a}\right) \tag{10.11}$$

This is known as vertical shearing. We note that if $\alpha = 1/\lambda f$, such a convolution, and the resultant WDF coordinate transformation, describes the result of a signal passing through a lens of focal length f.

10.2.5 The Comb Function and Rect Function

10.2.5.1 Comb Functions

In sampling theory one cannot avoid encountering both comb functions and rect functions. For example, the physical sampling of a signal is modeled by multiplying by a train of Dirac delta functions, sometimes called a comb function $\delta_T(x)$,

$$\delta_T(x) = \sum_{n=-\infty}^{\infty} \delta(x - nT) = \frac{1}{T} \sum_{n=-\infty}^{\infty} \exp\left(\frac{j2\pi nx}{T}\right) \tag{10.12}$$

where the rightmost part of Eq. (10.12) comes from a Fourier series expansion. The WDF of this comb function can be expressed as[47]

$$\psi\{\delta_T(x)\}(x,k) = \frac{1}{T^2} \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \delta\left(k - \frac{n+m}{2T}\right) \exp\left[j2\pi\left(\frac{n-m}{T}\right)x\right] \tag{10.13}$$

Equation (10.13) can be expanded out into the following form:

$$\psi\{\delta_T(x)\}(x,k) = \frac{1}{2T} \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} (-1)^{nm} \delta\left(k - \frac{n}{2T}\right) \delta\left(x - \frac{mT}{2}\right) \tag{10.14}$$

In Fig. 10.3 we show the WDF of the comb function. In Fig. 10.3*a* we show the actual WDF of the comb function defined in Eq. (10.14). In Fig. 10.3*b* we show the same WDF, but this time we ignore all the even

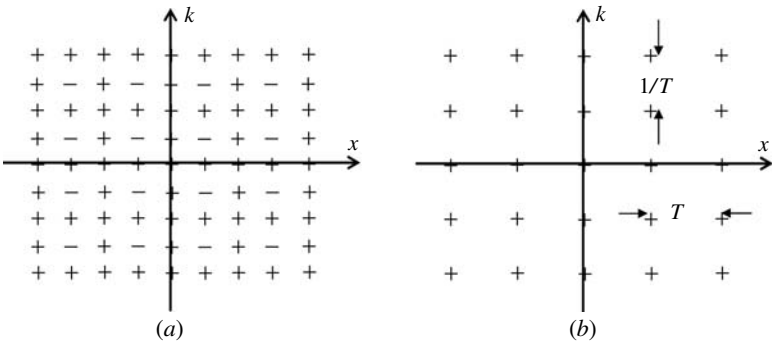


FIGURE 10.3 WDF of a comb function. (*a*) The actual WDF of the comb functions where we include all the interfering terms. The plus and minus terms represent positive and negative Dirac delta function. (*b*) Here we show only those terms that manifest themselves in the marginals of the WDF, i.e., at regular intervals of $x = mT/2$, $k = n/2T$ for all even integers m, n.

m and n terms. At this point the reader might fairly ask why one might ever wish to use this incomplete description of the comb function in any analysis. We can argue that this description is actually sufficient for the analysis presented in this chapter. Those terms that we have omitted in Fig. 10.3*b* do not manifest themselves in the marginals of the WDF; i.e., if we integrate $\psi\{\delta_T(x)\}(x, k)$ over k, all those terms for even m in Eq. (10.14) integrate to zero (those lines containing consecutive positive and negative Dirac delta functions will integrate to zero). Similarly, if we integrate $\psi\{\delta_T(x)\}(x, k)$ over x, all those terms for even n in Eq. (10.14) integrate to zero. Importantly, it is possible to multiply the comb function by the signal, thereby convolving their WDFs along the k axis and to employ an incomplete version of the comb function's WDF in our analysis. This is true if (1) the signal's bandwidth is less than or equal to $1/T$ and (2) we intend only to view the signal from the x or the k domain. For such a band-limited signal, the resultant convolution will create copies of strips of the signal's WDF that do not overlap looking parallel to the x axis. Therefore there will be no interference between adjacent copies in the Fourier domain, and the cross-terms of the comb function can be ignored *before* the convolution with the signal's WDF. We note that viewing the signal from domains other than the x or k domain will require us to include the cross-terms from the outset. The idea becomes much clearer as we proceed to the discussion on sampling in Sec. 10.4.

If the comb function defined in Eq. (10.12) is multiplied by a chirp function $\exp(+j\pi\alpha x^2)$, such as that illustrated in Fig. 10.2*c*, we get

$$\exp(+j\pi\alpha x^2)\,\delta_T(x) = \exp(+j\pi\alpha x^2) \sum_{n=-\infty}^{\infty} \delta(x - nT)$$

$$= \sum_{n=-\infty}^{\infty} \exp[+j\pi\alpha(nT)^2]\,\delta(x - nT) \tag{10.15}$$

The result is a coordinate shift of the WDF given by Eq. (10.11). Thus the actual WDF of a sheared comb function can be found by setting $k \to k + x/\alpha$ in Eq. (10.14). One may envisage the process as a convolution along k of the PSDs illustrated in Fig. 10.2*c* and Fig. 10.3. The results shown in Fig. 10.4*a* and *b* correspond to Fig. 10.3*a* and *b*. In Fig. 10.4*a* we show the actual WDF of the sheared comb function, and in Fig. 10.4*b* we show the case when we have removed every second delta function in both the x and k dimensions, just as we did in Fig. 10.3*b*. We recall that in relation to Fig. 10.3*b* we stated that we could use the incomplete comb function in our future analysis because all those terms integrated to zero along the x and k projections. Thus as long as our analysis is interested only in the spatial and FT distributions, we may employ the incomplete WDF. We need to amend this

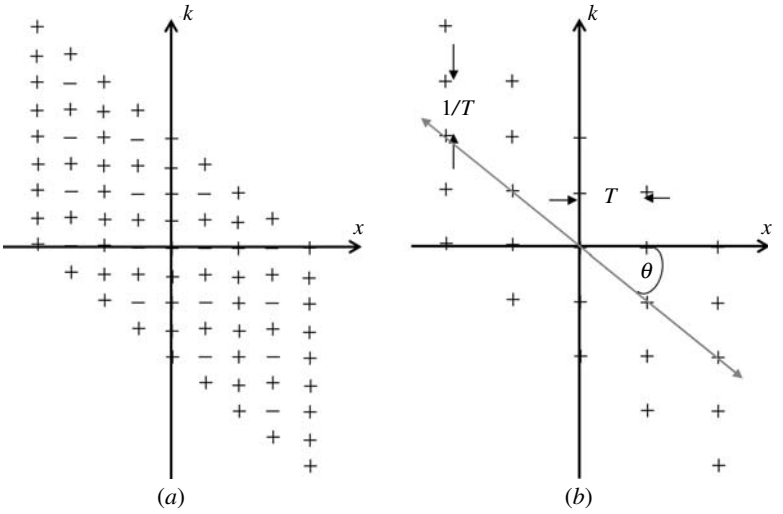


FIGURE 10.4 WDF of a vertically sheared comb function, i.e., a comb function that has been multiplied by a chirp function. (*a*) The actual WDF of the vertically sheared comb function where we include all the interfering terms. (*b*) Here, as in Fig. 10.3*b*, we show only those terms $x = mT/2$, $k = n/2T$ for all even integers m, n.

argument a little if we are to apply it to Fig. 10.4*b*. It is true that the shearing of the comb function means that the projection of the WDF along the k axis will be unchanged. However, the projection of the sheared comb function along the x axis (i.e., into the FT domain) will be altered significantly. This alteration is dependent on α. Thus our previous statement "as long as our analysis is interested only in the spatial and FT distributions, we may employ the incomplete WDF" is no longer valid. If we wish to argue a case for the use of this incomplete sheared comb WDF, we alter this statement accordingly: As long as our analysis is interested only in the spatial distribution and that projection along the arrow line (at angle θ), we may employ the incomplete WDF in this future analysis. We note that the relevance of this vertically sheared comb function PSD will become clear in the context of sampling a signal that is bounded in some LCT domain.

10.2.5.2 Rect Functions

Another signal often encountered in sampling theory is the rectangular window function, denoted by rect. We often multiply the sampled signal's Fourier transform by a rect function to recover the original continuous signal. This process is of course equivalent to a convolution

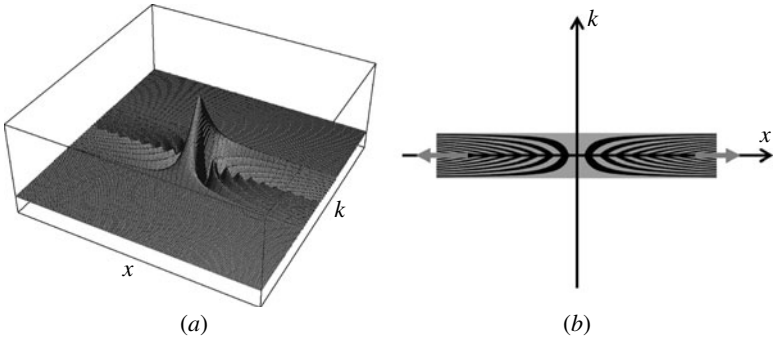


FIGURE 10.5 The rect function: (*a*) Wigner distribution function of rect(k) and (*b*) the phase-space diagram of that function.

of the signal samples in the space domain with a sinc function. Since this process is so central to sampling, it is important that we define the WDF of a rect function in the frequency domain, defined as

$$\text{rect}(Tk) = \begin{cases} 1 & \forall |k| \leq 1/2T \\ 0 & \forall |k| > 1/2T \end{cases} \tag{10.16}$$

Using Eq. (10.16) as input to the WDF integral defined in terms of the Fourier transform [see Eq. (10.2)] results in the following WDF for the rect function.

$$\psi\{\text{rect}(Tk)\}(x,k) = 2T(1-2T|k|)\,\text{sinc}\left[\frac{2x}{T}(1-2T|k|)\right] \tag{10.17}$$

In Fig. 10.5*a* we illustrate this function. We see that it is bounded in the frequency domain. We note that integrating this function along the k axis results in sinc(x/T) while integrating along the x axis results in rect(Tk). The PSD for this function is shown in Fig. 10.5*b*.

10.3 Finite Supports

10.3.1 Band-limitedness in Fourier Domain

If a signal $u(x)$ is zero-valued outside of some finite range, that is, $u(x) = 0$ for $|x| < D$, it is said to have compact support. If the FT of a signal has compact support, i.e., if $U(k)$ for $|k| < B$, where $U(k)$ is the FT of $u(x)$, we say that $u(x)$ is *band-limited*. Such a signal has a PSD shown in Fig. 10.1. We furthermore refer to $u(x)$ as having bandwidth B. This concept of band-limitedness is very important in sampling

theory. The Shannon sampling theorem assumes that a signal has this property. A theorem that appears in many textbooks states that a signal and its Fourier transform cannot both have compact support (Ref. 9, p. 26; Ref. 89). We will refer to this as the *Fourier transform compact support theorem*. This theorem is a corollary of the Paley-Wiener theorem.[90–91] However, for numerical work, we must assume that a signal may be approximately represented using a finite number of samples. This is achieved in the discrete Fourier transform by pretending that the signal is periodic in both space and frequency. Inevitably, this will result in some error referred to as aliasing.

10.3.2 Band-limitedness and the LCT

The FT is a special case of the LCT. Given that a signal and its FT cannot both have compact support, it is natural to ask, Can a signal and its LCT ever have compact support, and if so, when? The question of compact support and band-limitedness is important in relation to the LCT[92] as the development of a generalized sampling theory[93] is one of the topics we address in this chapter. We now present theorems that describe how the LCT preserves, destroys, or transforms compact support or band-limitedness. The proofs of these theorems are omitted, but can be found in Ref. 92.

We first consider the case of a LCT with none of the *ABCD* parameters equal to zero. Such a LCT is entirely destructive of compact support and band-limitedness. Given a quadratic-phase system (QPS) characterized by an *ABCD* matrix with no elements equal to zero, and given an input waveform $u(x)$ that either has compact support or is band-limited, the output waveform $L_M\{u(x)\}(x')$ neither is band-limited nor has compact support. This case is represented by the first two lines in Table 10.1. Equivalently, if the *ABCD* matrix has no entries equal to zero and the output waveform $L_M\{u(x)\}(x')$ has compact support or is band-limited, then the input waveform $u(x)$ neither is band-limited nor has compact support. This case is represented by lines 3 and 4 in Table 10.1. A number of other cases are also shown in Table 10.1. For example, if $C = 0$, as it does in the case of the Fresnel transform ($A = D = 1$, $B = \lambda z$, $C = 0$), the property of finite bandwidth will be conserved through the transform. Similarly the property of infinite bandwidth will also be conserved. For a Fourier transform $A = D = 0$, $B = 1$, and $C = -1$. Since $A = 0$, a property of finite bandwidth will not be preserved through the Fourier transform; rather, it will produce a property of finite support. If $D = 0$, the reverse also holds true.

At this point the reader might reasonably ask, What will the PSD of signal with a finite support w in an LCT domain (with parameters *ABCD*) look like? As a simple illustration we show such a signal in Fig. 10.6*a*. The signal will have a local bandwidth which is equal to b. We show two lines, parallel to the boundary support of the signal,

Input Signal		LCT				Output Signal	
Finite Support	Finite Bandwidth	A	B	C	D	Finite Support	Finite Bandwidth
$\surd$	X	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	X	X
X	$\surd$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	X	X
X	X	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\surd$	X
X	X	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	X	$\surd$
X	$\surd$	$= 0$				$\surd$	X
	X	$= 0$				X	
$\surd$	X		$= 0$			$\surd$	X
X			$= 0$			X	
X	$\surd$			$= 0$		X	$\surd$
	X			$= 0$			X
$\surd$	X				$= 0$	X	$\surd$
X					$= 0$		X

Note: If a signal is to have a finite support in some domain, it must belong to one of the above possible sets.

TABLE 10.1 Properties of Finite Support and Band-limitedness for Linear Canonically Transformed Signals

defined as follows:

$$\begin{aligned} x + k\tan\theta &= b\tan\frac{\theta}{2} \\ x + k\tan\theta &= -b\tan\frac{\theta}{2} \end{aligned} \tag{10.18}$$

After we apply the LCT, there will occur a simple coordinate transformation to the PSD shown in Fig. 10.6*b*. This coordinate shift is given by Eqs. (10.8) and (10.9). In this domain the signal clearly has a finite support in x equal to w. Clearly our two lines have now become

$$\begin{aligned} x &= -\frac{w}{2} \\ x &= \frac{w}{2} \end{aligned} \tag{10.19}$$

After applying the coordinate shift to Eq. (10.19), we get

$$(A + C\tan\theta)x + (B + D\tan\theta)k = b\tan\frac{\theta}{2}$$

$$(A + C\tan\theta)x + (B + D\tan\theta)k = b\tan\frac{\theta}{2} \tag{10.20}$$

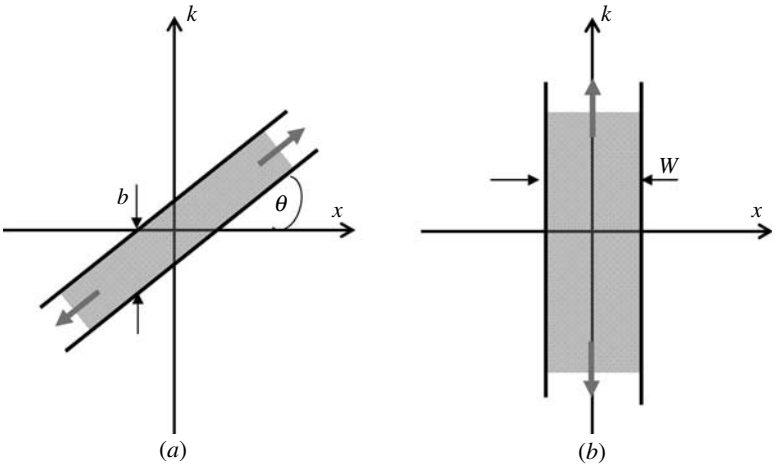


FIGURE 10.6 The phase-space diagram of (*a*) a signal with finite local support in the LCT domain and (*b*) a signal with finite support in x (obtained by applying LCT to signal in part *a*).

which, when we make use of the fact that $AD - BC = 1$, reduces to give

$$\begin{aligned} \tan\theta &= -\frac{B}{D} \\ w &= \frac{b}{B} \end{aligned} \tag{10.21}$$

10.3.3 Finite Space-Bandwidth Product-Compact Support in *x* and *k*

In many practical problems it is assumed that a signal is bounded within some finite region in both the spatial and spatial-frequency domains. The spatial extent w and the frequency extent b are defined such that

$$\begin{aligned} &u(x) \approx 0 \qquad |x| > \frac{w}{2} \\ &U(k) = \int_{-\infty}^{\infty} u(x)\exp(-j2\pi kx)\,dx \qquad |k| > \frac{b}{2} \end{aligned} \tag{10.22}$$

and therefore, the signal energy is negligible outside these spatial and spatial-frequency regions. For all signals discussed here, w and b may also be defined as

$$\int_{-w/2}^{w/2} |u(x)|^2\,dx = \eta E \qquad \int_{-b/2}^{b/2} |U(k)|^2\,dk = \eta E \tag{10.23}$$

where $\eta \cong 1$ and E represents the total signal energy.

$$E = \int_{-\infty}^{\infty} |u(x)|^2 \, dx = \int_{-\infty}^{\infty} |U(k)|^2 \, dk \tag{10.24}$$

The dual equality in Eq. (10.24) follows from Rayleigh's theorem. Earlier in Fig. 10.1*b* we showed the PSD of a signal $u(x)$ in which the signal energy lies within a rectangular area. In the next section we discuss sampling and interpolation. As a prelude we note that the signal $u(x)$ is completely determined if it is sampled equidistantly in x with sample space δx such that the Nyquist criterion is satisfied. Therefore the number of samples N required to completely describe $u(x)$ is $N = d/\delta x \geq db$. Clearly, for the most efficient uniform sampling $\delta x = 1/b$ and $N = db$, the space-bandwidth product (SBP) of the signal. In general, signals may have an irregularly bounded WDF, and one such case is shown in Fig. 10.1*d*.

10.4 Sampling a Signal

In the last section we discussed signals that had the properties of compact support and band-limitedness and the effect that different types of LCT would have on these two properties. It is well known that a signal that has the property of band-limitedness can be sampled and interpolated exactly from these samples. This is true only if the signal has been sampled at a rate greater than or equal to the Nyquist rate, which is determined by the bandwidth of the signal. In this section we review this sampling theorem, using phase-space diagrams. As shown in the last section, often the LCT of a band-limited signal is no longer band-limited. Therefore, if one were to rigorously follow the laws of Nyquist and Shannon, one would arrive at the conclusion that such a signal could not be sampled and interpolated from these samples. Recent work suggests otherwise. It has been shown that a more general sampling theorem must be employed for signals of this type. As we show in this section, by far the simplest way that one can deduce this generalized sampling theorem is to again employ phase-space diagrams. In fact we shall see that we need only heuristically apply some of the rules we have learned thus far on the PSD in order to fully derive the generalized sampling theorem in the briefest of fashions. We begin with a discussion of Nyquist.

10.4.1 Nyquist-Shannon Sampling

The comb function was discussed in Sec. 10.2.5.1. We begin with a signal $u(x)$ with finite bandwidth equal to B with the PSD shown

in Fig. 10.7*a*. Sampling this signal is modeled by multiplying by the comb function defined in Eq. (10.12) to get a sampled function $u_T(x)$.

$$u_T(x) = u(x)\delta_T(x) \tag{10.25}$$

Employing the Fourier series of the comb function in Eq. (10.12), we may deduce that the Fourier transform of $u_T(x)$ is given by an infinite sum of shifted replicas of $U(k)$

$$U_T(k) = \frac{1}{T}\sum_{n=-\infty}^{\infty} U\left(k - \frac{n}{T}\right) \tag{10.26}$$

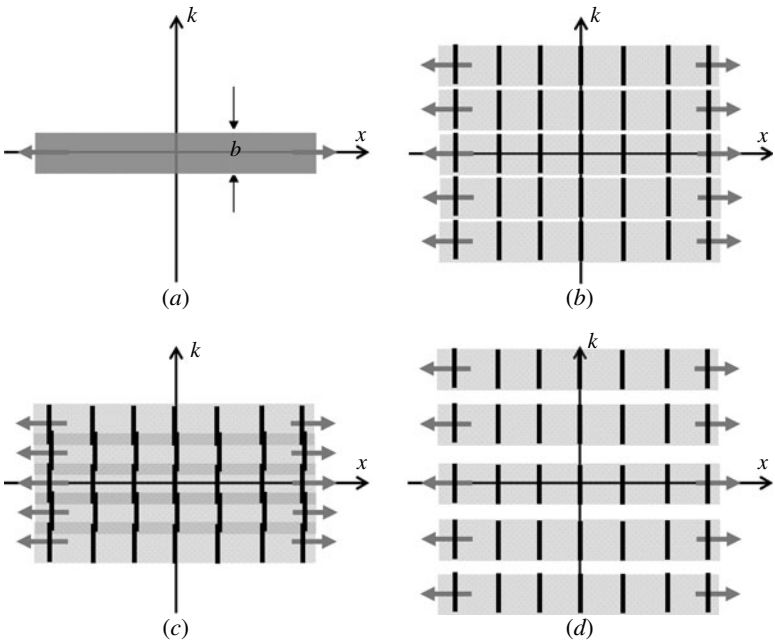


FIGURE 10.7 Phase-space diagrams that demonstrate Nyquist-Shannon sampling. (*a*) A band-limited signal with infinite support. The thick arrows denote that the phase-space diagram extends to infinity along the x axis. (*b*) Phase-space diagram for the sampled signal where the sampling rate is equal to the Nyquist rate. The thick black lines correspond to strips of the original WDF separated in x by T. These strips are actually the only nonzero values in the WDF, although we show a light copy of the original WDF beneath the strips for illustrative purposes. (*c*) The rate is less than the Nyquist rate, and aliasing occurs. (*d*) The rate is greater than the Nyquist rate, and the signal is oversampled.

From our previous discussions the WDF of $u_T(x)$ may be defined as

$$\psi\{u_T(x)\}(x,k) = \psi\{u(x)\}(x,k) *^k \psi\{\delta_T(x)\}(x,k) \tag{10.27}$$

and therefore the PSD for $u_T(x)$ is given by the convolution along k of the PSD of $u(x)$ and the PSD of the comb function shown in Fig. 10.3*b*. The result of convolving the signal's PSD with these Dirac delta lines is that we obtain multiple copies, in k, of the original signal PSD. More accurately we obtain strips of the original WDF separated by the sampling period, and we obtain multiple copies of these strips in k. It is very important that the distance between the center of adjacent strip copies in k (which from our earlier comb function analysis is equal to $1/T$) be greater than or equal to b. When $b = 1/T$, we get the PSD shown in Fig. 10.7*b*, where we can see that the replicas are just shy of overlapping one another. This is the optimum sampling case known as the Nyquist limit. The signal may be reconstructed by multiplying $U_T(k)$ by $T\text{rect}(Tk)$ defined in Eq. (10.17). From Fourier theory this is equivalent to convolving $u_T(x)$ with $\text{sinc}(x/T)$. All three equivalent expressions for reconstruction are given below in terms of the spatial representation, FT and WDF.

$$U(k) = U_T(k)\text{rect}(Tk)$$

$$u(x) = u_T(x) * \frac{1}{T}\,\text{sinc}\left(\frac{x}{T}\right) = \sum_{n=-\infty}^{\infty} u(nT)\,\text{sinc}\left(\frac{x-nT}{T}\right)$$

$$\psi\{u(x)\}(x,k) = \psi\{u_T(x)\}(x,k) *^x \psi\{\text{rect}(Tk)\}(x,k) \tag{10.28}$$

We can visualise this reconstruction as follows: The rect function WDF illustrated in Fig. 10.5 convolves along the x axis in Fig. 10.7*b*. This rect WDF has zero values for all values of k outside of the center order of the sample's signal WDF. Thus we need only visualize the convolution of rect WDF with those strips of our signal's WDF which lie crossing the x axis.

Fascinatingly, the resulting convolution must be equal to the original signal's WDF in accordance with the Nyquist-Shannon sampling theorem. If we do not sample quickly enough, such that $T > 1/b$, aliasing will occur where the copies of the signal will overlap with one another and our reconstructions in Eq. (10.28) will be invalid. We illustrate this case in Fig. 10.7*c*. In Fig. 10.7*d* we also illustrate the case where we have oversampled, that is, $T < 1/b$. In this case we can see our replicas have moved farther apart. Equation (10.28) still holds, and indeed in this case it must hold for a range of different rect widths.

10.4.2 Generalized Sampling

We now consider signals that have finite support in some LCT domain with parameters $ABCD$. In the case of these signals it is possible to derive a rigorous sampling theory that can be interpreted and derived in the simplest possible way by employing phase-space diagrams. We take the signal $u(x)$ with PSD shown in Fig. 10.6*a*. We define this signal to have a finite support w in some LCT domain with parameters $ABCD$ and to have local bandwidth b. From the third row of Table 10.1 we know that $u(x)$ must have an infinite bandwidth and an infinite spatial support. Therefore Nyquist-Shannon sampling cannot be applied in the conventional sense. If this signal can be sampled at some rate T and reconstructed from these samples $u(nT)$, we must derive a new sampling criterion and a new interpolation formula. In fact we already have derived both. We refer the reader to Sec. 10.2.5.1 where we discussed the sheared comb function $\exp(j\pi\alpha x^2)\delta_T(x)$. The PSD for this function is given in Fig. 10.4, where $\alpha = 1/\tan(-\theta)$. The WDF of this sheared-sampled signal is given by

$$\psi\left\{u_T(x)\,\exp(j\pi\alpha x^2)\right\}(x,k) = \psi\{u(x)\}(x,k) *^k \psi\left\{\delta_T(x)\exp(j\pi\alpha x^2)\right\}(x,k) \quad (10.29)$$

If we match up the values of θ, then the convolution along k of the PSD of $u(x)$ shown in Fig. 10.6 and the sheared comb function that result will be an infinite number of replicas of a band-limited signal with bandwidth b similar to that shown in Fig. 10.7*b*. From Eq. (10.21) matching up the values of θ, this implies that for the comb function $\alpha = 1/\tan(-\theta) = D/B$. When $b = 1/T$, we get the PSD shown in Fig. 10.7*b*, where we can see that the replicas are just shy of overlapping one another. To avoid aliasing, we must have $T \leq 1/b$. From Eq. (10.21) this forces the following condition: The signal may be reconstructed by multiplying the resultant signal by $T\,\mathrm{rect}(Tk)$ in the Fourier domain. This step in the reconstruction is identical to Shannon interpolation described by Eq. (10.28).

$$\psi\{u(x)\exp(j\pi\alpha x^2)\}(x,k) = \psi\{u_T(x)\exp(j\pi\alpha x^2)\}(x,k) *^x \psi\{\mathrm{rect}(Tk)\}(x,k) \quad (10.30)$$

The final part of the reconstruction is accomplished by multiplying by the conjugate of the original shear. The overall reconstruction algorithm is given by

$$\begin{aligned} u(x) &= \exp\left(-j\pi x^2\frac{D}{B}\right)\left\{\left[u_T(x)\exp\left(j\pi x^2\frac{D}{B}\right)\right] * \frac{1}{T}\mathrm{sinc}\frac{x}{T}\right\} \\ &= \exp\left(-j\pi x^2\frac{D}{B}\right)\sum_{n=-\infty}^{\infty} u(nT)\exp\left[j\pi(nT)^2\frac{D}{B}\right]\mathrm{sinc}\frac{x-nT}{T} \end{aligned} \quad (10.31)$$

We note for the first time an interesting observation. Regardless of what LCT caused phase space to be compact in some direction, the sampling representation can always be based on the assumption of a chirped signal. In addition, this does not change the number of samples despite any bandwidth compression or expansion. From Eq. (10.31) we can see that only two parameters of the LCT are employed in interpolation.

10.5 Simulating an Optical System: Sampling at the Input and Output

So far, we have discussed how to reconstruct a signal from its samples. Often we encounter the case where a signal is sampled, an LCT is applied to this discrete signal and the result of this is again sampled. This arises in numerical simulations, where the input and output are necessarily discrete, and when modeling paraxial optical systems with discrete elements such as SLMs and CCD cameras. In this section, we demonstrate how to sample a wave field which then undergoes a LCT, how to sample the output of this LCT and then reconstruct from these samples the analog LCT of the original analog wave field. Our major goal here is to make sure that the LCT of the sampled signal actually looks like the LCT of the continuous signal. This should obviously be the case if we are to effectively simulate an optical system. This is described by the block diagram in Fig. 10.8. We make two assumptions about the input wave field: (1) It has approximately finite bandwidth, and (2) it has approximately finite support. Both of these assumptions are described by Eq. (10.23).

In the case of these signals it is possible to derive a rigorous sampling theory that can be interpreted and derived in the simplest possible way by employing PSDs.

Consider the first process in Fig. 10.8, sampling of the input. When a signal is sampled, its phase-space diagram is altered by the

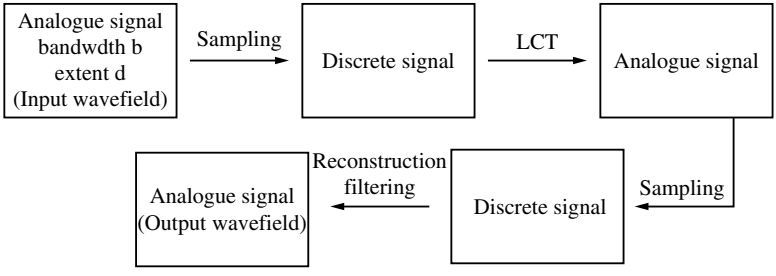


FIGURE 10.8 Block diagram of the problem considered in this section.

addition of periodic replicas, as illustrated by Fig. 10.7. When a signal is transformed by a LCT, its phase-space diagram undergoes an area-preserving (affine) coordinate transformation, as illustrated in Fig. 10.1. The case for a sampled signal (which is produced by the second step in Fig. 10.8) is illustrated in Fig. 10.9.

In Fig. 10.9*a* is the LCT of an analog function. In Fig. 10.9*b* is the LCT of the sampled version. We have chosen our sampling rate sufficiently high that the generalized sampling theory described in Sec. 10.4.2 allows us to recover the LCT of the analog function from the LCT of the sampled one by filtering operations. Step 3 in Fig. 10.8 is to sample the LCT of the sampled function, a sampling in the domain x. This produces replicas in the orthogonal domain k. If we haven't chosen our sampling rates correctly, then we have replicas overlapping one another in the PSD, as shown in Fig. 10.10*a*. This overlap illustrates that aliasing has occurred, and our recovered signal will be degraded.

In the situation shown in Fig. 10.10*b*, we chose our first sampling rate to be a little higher when sampling the input wave field, so that the replicas don't overlap, and we can reconstruct the output by truncation (getting rid of everything outside the two dashed lines). This reduces the problem to that discussed in the preceding section on generalized filtering. These two operations are the reconstruction filtering process indicated in Fig. 10.8.

We now determine the sampling condition that guarantees the situation shown in Fig. 10.10*b*. There are other ways to prevent aliasing,

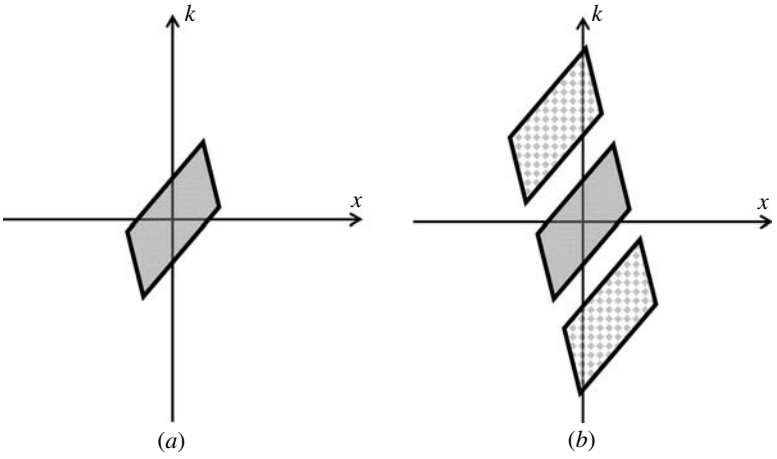


FIGURE 10.9 (*a*) PSD of the LCT of the wave field after a LCT. (*b*) PSD of the wave field after sampling and a LCT. The zeroth order is shaded and the replicas are checkered.

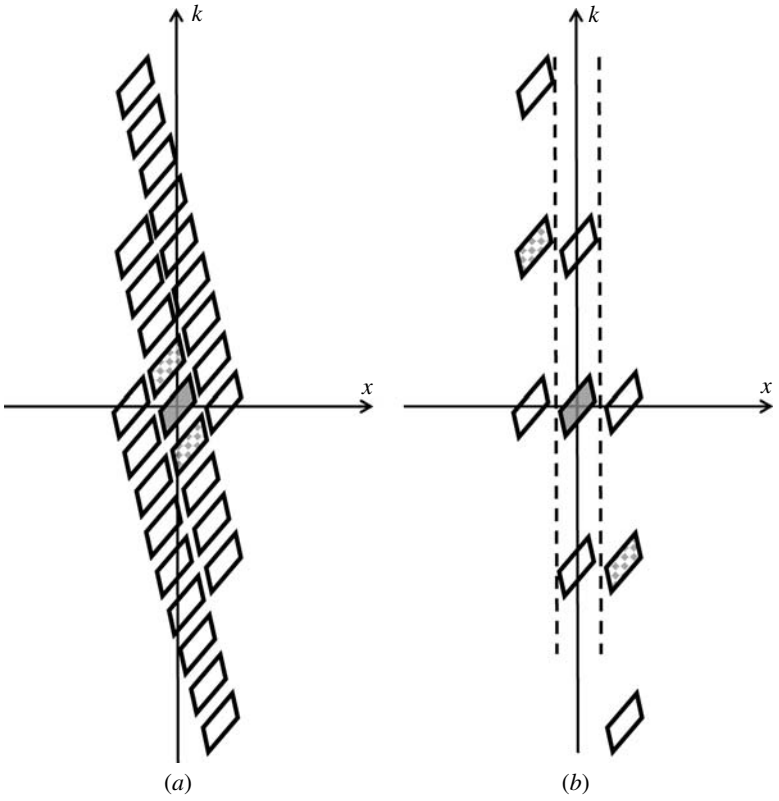


FIGURE 10.10 (*a*) The phase-space diagram of the sampled LCT of the undersampled input. The zeroth-order term is shaded. The terms created by sampling the input (as in Fig. 10.9) are checkered. (*b*) The phase-space diagram of the sampled LCT of the correctly sampled input.

some of which will be more useful than this one in specific cases. The key idea to remember is to prevent any overlap of the replicas. If this is achieved, we can recover the zeroth-order term by truncation and generalized filtering. More importantly the new sampled signal will be a sampled version of the LCT of the original continuous signal. Consider a point in the phase-space footprint of a sampled function $p_1(x_1, k_1)$. The equivalent point in one of the two nearest replicas of the footprint is given by $p_2(x_1, k_1 + 1/T_x)$, where T_x is the sampling rate used in the first sampling operation. After the LCT operation, these points are transformed as shown in Eq. (10.9). This results in the points $p_1(Ax_1 + Bk_1, Cx_1 + Dk_1)$ and $p_2'(Ax_1 + Bk_1, +B/T_x, Cx_1 + Dk_1, +D/T_x)$. These points are separated

by a horizontal distance of $|B|/T_x$. This distance must be greater than the extent of the transformed (unsampled) function in x. Similar results can be found for all other replicas, but this defines the lower bound on the sampling rate.

$$T_x \leq \frac{|B|}{w} \tag{10.32}$$

When this bound is used in addition to the Nyquist criterion, the PSD of the output waveform—after this is also sampled at the generalized Nyquist rate—will look like Fig. 10.10*b*. Equation (10.32) fails for $B = 0$, which is a special case of the LCT where the output is already discrete and no second sampling process takes place.

10.6 Conclusion

In conclusion we have demonstrated the usefulness of the WDF in both qualitatively and quantitatively understanding sampling theory. In particular, the ability of the WDF to describe generalized sampling is extraordinary. At no point is it necessary to employ a wave integral in these derivations. In fact, almost all the equations in this chapter serve as a pretext to the underlying idea; by visualizing a signal's energy as a bounded shape in phase space and endowing this shape with the basic properties of the WDF, we may derive complicated sampling criteria and interpolation formula.

References

1. H. Nyquist, "Certain topics in telegraph transmission theory," *Trans. AIEE*, **47**: 617–644 (1928).
2. C. E. Shannon, "Communication in the presence of noise," *Proc. Inst. Radio Eng.* **37**: 10–21 (1949).
3. E. Whittaker, "On the functions which are represented by the expansions of the interpolation theory," *Proc. Royal Soc. Edinburgh*, Sec. *A*, **35**: 181–194 (1915).
4. V. A. Kotelnikov, "On the carrying capacity of the ether and wire in telecommunications," Material for the First All-Union Conference on Questions of Communication, *Izd. Red. Upr. Svyazi RKKA*, Moscow, 1933.
5. G. T. Di Francia, "Resolving power and information," *J. Opt. Soc. Am.* **45**: 497–499 (1955).
6. S. Haykin, *Communication Systems*, 4th ed., Wiley, New York, 2001.
7. A. V. Oppenheim, R. Schafer, and J. R. Buck, *Discrete Time Signal Processing*, 2d ed., Prentice-Hall, Englewood Cliffs, N.J., 1999.
8. R. N. Bracewell, *The Fourier Transform and Its Applications*, 2d Ed., McGraw-Hill, New York, 1978.
9. J. W. Goodman, *Introduction to Fourier Optics*. 2d Ed., McGraw-Hill, New York, 1996.
10. A. Papoulis, *The Fourier Integral and Its Applications*, McGraw-Hill, New York, 1962.

11. A. Papoulis, *Signal Analysis*, McGraw-Hill, New York, 1977.
12. F. Gori, "Fresnel transform and sampling theorem," *Opt. Eng.* **39**: 293–297 (1981).
13. J. J. Ding., "Research of Fractional Fourier Transform and Linear Canonical Transform," Ph.D. National Taiwan University, Taipei, Taiwan, R.O.C thesis, 2001.
14. L. Onural, "Sampling of the diffraction field," *Appl. Opt.* **39**: 5929–5935 (2000).
15. C. Candan and H. M. Ozaktas, "Sampling and series expansion theorems for fractional Fourier and other transforms," *Signal Proc.* **83**: 2455–2457 (2003).
16. A. Stern and B. Javidi, "Sampling in the light of Wigner distribution," *J. Opt. Soc. Am. A* **21**: 360–366 (2004).
17. A. Stern and B. Javidi, "Sampling in the light of Wigner distribution: Errata," *J. Opt. Soc. Am.* JOSA A, *A* **21**(9): 1602–1612 (2004).
18. A. Stern and B. Javidi, "Analysis of practical sampling and reconstruction from Fresnel fields," *Opt. Eng.* **43**: 239–250 (2004).
19. B. M. Hennelly and J. T. Sheridan, "Fast numerical algorithm for the linear canonical transform," *J. Opt. Soc. Am. A* **22**: 928–937 (2005).
20. A. Stern, "Sampling of linear canonical transformed signals," *Signal Proc.* **86**: 1421–1425 (2006).
21. B. Deng, R. Tao, and Y. Wang, "Convolution theorems for the linear canonical transform and their applications," *Science in China* (Ser. F, *Info. Sci.*), **49**: 592–603 (2006).
22. A. Stern and B. Javidi, "Improved-resolution digital holography using the generalized sampling theorem for locally band-limited fields," *J. Opt. Soc. Am. A* **23**: 1227–1235 (2006).
23. R. T. B. Li and Y. Wang, "New sampling formulae related to linear canonical transform," *Signal Proc.* **86**: 983–990 (2007).
24. B. Z. Li, R. Tao, and Y. Wang, "New sampling formulae related to linear canonical transform," *Signal Proc.* **87**: 983–990 (2007).
25. J. J. Healy and J. T. Sheridan, "Cases where the linear canonical transform of a signal has compact support or is band-limited," *Opt. Lett.* **33**: 228–230 (2008).
26. S. A. Collins, "Lens-system diffraction integral written in terms of matrix optics," *J. Opt. Soc. Am.* **60**: 1168–1177 (1970).
27. H. M. Ozaktas, Z. Zalevsky, and M. A. Kutay, *The Fractional Fourier Transform with Applications in Optics and Signal Processing*. Wiley, Chichester, 2001.
28. K. B. Wolf, "Canonical transforms," *Integral Transforms in Science and Engineering*, K. B. Wolf (ed.), Plenum, New York, 1979, Chap. 9, pp. 381–416.
29. S. Abe and J. T. Sheridan, "Generalization of the fractional Fourier transformation to an arbitrary linear lossless transformation: An operator approach," *J. Phys. A* **27**: 4179–4187 (1994).
30. S. Abe and J. T. Sheridan, "Corrigenda: Generalization of the fractional Fourier transformation to an arbitrary linear lossless transformation: An operator approach," *J. Phys. A* **27**: 7937 (1994).
31. S. Abe and J. T. Sheridan, "Optical operations on wave functions as the Abelian subgroups of the special affine Fourier transformation," *Opt. Lett.* **9**: 1801–1803 (1994).
32. S. C. Pei and J. J. Ding, "Eigenfunctions of linear canonical transform," *IEEE Trans. Signal Proc.* **50**: 11–26 (2002).
33. S. C. Pei and J. J. Ding, "Generalised eigenvectors and fractionalisation of offset DFTs and DCTs," *IEEE Trans. Signal Proc.* **52**: 2032–2046 (2004).
34. J. W. Goodman and R. Lawrence, "Digital image formation from electronically detected holograms," *Appl. Phys. Lett.* **11**: 77–79 (1967).
35. L. P. Yaroslavskii and N. S. Merzlyakov. *Methods of Digital Holography*. Consultants Bureau, New York, 1980.
36. T. M. Kreis, M. Adams, and W. P. O. Juptner, "Methods of digital holography: A comparison," *Proc. SPIE* **3098**: 224–233 (1997).
37. I. Yamaguchi and T. Zhang, "Phase-shifting digital holography," *Opt. Lett.* **22**: 1268–1270 (1997).

38. T. Kreis, M. Adams, and W. Juptner, "Digital in-line holography in particle measurement," *Proc. SPIE* **3744**: 54–64 (1999).
39. G. Pedrini, P. Frning, H. Tiziani, and F. Santoyo, "Shape measurement of microscopic structures using digital holograms," *Appl. Opt.* **164**: 257–268 (1999).
40. S. Schedin, G. Pedrini, H. Tiziani, A. Aggarwal, and M. Gusev, "Highly sensitive pulsed digital holography for built-in defect analysis with a laser excitation," *Appl. Opt.* **40**: 100–117 (2001).
41. U. Schnars and W. Juptner, "Digital recording and numerical reconstruction of holograms," *Meas. Sci. Technol.* **13**: 85–101 (2002).
42. E. Wigner, "On the quantum correction for thermodynamic equilibrium," *Phys. Rev.* **40**: 749–759 (1932).
43. A. Papoulis, "Ambiguity function in Fourier optics," *J. Opt. Soc. Am.* **64**: 779–788 (1974).
44. L. Cohen, "Time-frequency distributions—A review," *Proc. IEEE* **77**: 941–981 (1989).
45. M. J. Bastiaans, "Wigner distribution function and its application to first order optics," *J. Opt. Soc. Am.* **69**: 1710–1716 (1979).
46. M. J. Bastiaans, "Application of the Wigner distribution function in optics," *The Wigner Distribution—Theory and Applications in Signal Processing*. W. Mecklenbrauker and F. Hlawatsch (eds.), Elsevier Science, Amsterdam, 1997.
47. M. Testorf and J. Ojeda-Castaneda, "Fractional Talbot effect: Analysis in phase space, *J. Opt. Soc. Am. A* **13**: 119–125 (1996).
48. A. W. Lohmann, "Image rotation, Wigner rotation and the fractional Fourier transform," *J. Opt. Soc. Am. A* **10**: 2181–2186 (1993).
49. A. W. Lohmann, R. G. Dorsch, D. Mendelovic, Z. Zalevsky, and C. Ferreira, "Space-bandwidth product of optical signals and systems," *J. Opt. Soc. Am. A* **13**: 470–473 (1996).
50. D. Mendelovic and A. W. Lohmann, "Space-bandwidth product adaptation and its application to superresolution: Fundamentals," *J. Opt. Soc. Am. A* **14**: 558–562 (1997).
51. D. Mendelovic, A. W. Lohmann, and Z. Zalevsky, "Space-bandwidth product adaptation and its application to superresolution: Examples," *J. Opt. Soc. Am. A* **14**: 563–567 (1997).
52. Z. Zalevsky, D. Mendelovic, and A. W. Lohmann, "Understanding superresolution in Wigner space," *J. Opt. Soc. Am. A* **17**: 2422–2429 (2000).
53. J. W. Cooley and J. W. Tukey, "An algorithm for the machine calculation of complex Fourier series," *Math. Comput.* **19**: 297–301 (1965).
54. S. Nishiwaki, "Calculations of optical field by fast Fourier transform analysis," *Appl. Opt.* **27**: 3518–3521 (1988).
55. J. A. Hudson, "Fresnel-Kirchhoff diffraction in optical systems: An approximate computational algorithm," *Appl. Opt.* **23**: 2292–2295 (1984).
56. H. Hamam and J. Tocnaye, "Efficient Fresnel-transform algorithm based on fractional Fresnel diffraction," *J. Opt. Soc. Am. A* **12**: 1920–1931 (1995).
57. J. Garcia, D. Mas, and R. Dorsch, "Fractional Fourier transform calculation through the fast Fourier transform algorithm," *Appl. Opt.* **35**: 7013–7018 (1996).
58. M. Sypek, "Light propagation in the Fresnel region. New numerical approach," *Opt. Comm.* **116**: 43–48 (1995).
59. X-G. Xia, "On bandlimited signals with fractional Fourier transform," *IEEE Signal Proc. Lett.* **3**: 72–74 (1996).
60. Z. Zalevsky, D. Mendelovic, and R. G. Dorsch, "Gerchberg-Saxton algorithm applied in the fractional Fourier or Fresnel domain," *Opt. Lett.* **21**: 842–844 (1996).
61. Z. Zalevsky, D. Mendelovic, and R. G. Dorsch, "Gerchberg-Saxton algorithm applied in the fractional Fourier or Fresnel domain," *Opt. Lett.* **21**: 842–844 (1996).
62. H. M. Ozaktas, O. Arikan, M. A. Kutay, and G. Bozdagi, "Digital computation of the fractional Fourier transform," *IEEE Trans. Signal Proc.* **44**: 2141–2150 (1996).

63. D. Mendelovic, Z. Zalevsky, and N. Konforti, "Computation considerations and fast algorithms for calculating the diffraction integral," *J. Mod. Opt.* **44**: 407–414 (1997).
64. F. J. Marinho, J. Francisco, and L. Bernardo, "Numerical calculation of fractional Fourier transforms with a single fast Fourier transform algorithm," *J. Opt. Soc. Am. A* **15**: 2111–2116 (1998).
65. W. Cong, N. Chen, and B. Gu, "Recursive algorithm for phase retrieval in the fractional Fourier transform domain," *Appl. Opt.* **37**: 6906–6910 (1998).
66. Y. Zhang, B. Dong, B. Gu, and G. Yang, "Beam shaping in the fractional Fourier transform domain," *J. Opt. Soc. Am. A* **15**: 1114–1120 (1998).
67. N. Delen and B. Hooker, "Free-space beam propagation between arbitrarily oriented planes based on full diffraction theory: A fast Fourier transform approach," *J. Opt. Soc. Am. A* **15**: 857–867 (1998).
68. C. Kopp and P. Meyrueis, "Near-field Fresnel diffraction: improvement of a numerical propagator," *Opt. Comm.* **158**: 7–10 (1998).
69. D. Mas, J. Garcia, C. Ferriera, and L. M. Bernardo, "Fast algorithms for free space diffraction patterns calculation," *Opt. Comm.* **164**: 233–245 (1999).
70. T. Erseghe, P. Kraniauskas, and G. Cariolaro, "Unified fractional Fourier transform and sampling theorem," *IEEE Trans. Signal Proc.* **47**: 3419–3423 (1999).
71. C. Candan, M. A. Kutay, and H. M. Ozaktas, "The discrete fractional Fourier transform," *IEEE Trans. Signal Proc.* **48**: 1329–1337 (2000).
72. X. Deng, B. Bihari, J. Gang, F. Zhao, and R. T. Chen, "Fast algorithm for chirp transforms with zooming-in ability and its applications," *J. Opt. Soc. Am. A* **17**: 762–771 (2000).
73. N. Delen and B. Hooker, "Verification and comparison of a fast Fourier transform-based full diffraction method for tilted and offset planes," *Appl. Opt.* **40**: 3525–3531 (2001).
74. W. T. Rhodes, "Numerical simulation of Fresnel-regime wave propagation: The light tube model," *Proc. SPIE* **4436**: 21–26 (2001).
75. W. T. Rhodes, "Light tubes, Wigner diagrams and optical signal propagation simulation," *Optical Information Processing: A Tribute to Adolf Lohmann*, H. J. Caulfield (ed.), SPIE Press, Bellingharm, Wash., 2002.
76. J. Li, Z. Fan, and Y. Fu, "The FFT calculation for Fresnel diffraction and energy conservation criterion of sampling quality," *Proc. SPIE* **4915**: 180–186 (2002).
77. U. Schnars and W. P. O. Juptner, "Digital recording and numerical reconstruction of holograms," *Meas. Sci. Technol.* **13**: 85–101 (2002).
78. D. Mas, J. Perez, C. Hernandez, C. Vazquez, J. J. Miret, and C. Illueca, "Fast numerical calculation of Fresnel patterns in convergent systems," *Opt. Comm.* **227**: 245–258 (2003).
79. B. M. Hennelly and J. T. Sheridan, "The fast linear canonical transform," *Proc. SPIE* **5456**: 71–82 (2004).
80. B. M. Hennelly and J. T. Sheridan, "Generalizing, optimizing, and inventing numerical algorithms for the fractional Fourier, Fresnel, and linear canonical transforms," *J. Opt. Soc. Am. A* **22**: 917–927 (2005).
81. B. M. Hennelly and J. T. Sheridan, "Efficient algorithms for the linear canonical transform," *Proc. SPIE* **5557**: 191–199 (2004).
82. K. Matsushima, "Computer-generated holograms for three-dimensional surface objects with shade and texture," *Appl. Opt.* **44**: 4607–4614 (2005).
83. B. Gombkötő, P. Koppa, P. Maák, and E. Lőrincz, "Application of the fast-Fourier-transform-based volume integral equation method to model volume diffraction in shift-multiplexed holographic data storage," *J. Opt. Soc. Am. A* **23**: 2954–2960 (2006).
84. F. Shen and A. Wang, "Fast-Fourier-transform based numerical integration method for the Rayleigh-Sommerfeld diffraction formula," *Appl. Opt.* **45**: 1102–1110 (2006).
85. R. P. Muffoletto, J. M. Tyler, and J. E. Tohline, "Shifted Fresnel diffraction for computational holography," *Opt. Express* **15**: 5631–5640 (2007).

86. T. Shimobaba, Y. Sato, J. Miura, M. Takenouchi, and T. Ito, "Real-time digital holographic microscopy using the graphic processing unit," *Opt. Express* **16**: 11776–11781 (2008).
87. A. Koc, H. M. Ozaktas, C. Candan, and M. A. Kutay, "Digital computation of linear canonical transforms," *IEEE Trans. Signal Proc.* **56**: 2383–2394 (2008).
88. J. J. Healy and J. T. Sheridan, "Sampling and discretization of the linear canonical transform," *Signal Proc.* **89**: 641–648 (2009).
89. H. Baher, *Analog & Digital Signal Processing*, Wiley, Chichester, 1990, p. 121.
90. Papoulis, *The Fourier Integral and Its Applications*, McGraw-Hill, New York, 1962, p. 215.
91. R. E. A. C. Paley and N. Wiener, *Fourier Transforms in the Complex Domain*, American Mathematical Society, New York, 1934, p. 12.
92. J. J. Healy and J. T. Sheridan, "Bandwidth, compact support, apertures and the linear canonical transform in ABCD systems," *Proc. SPIE* **6994**: 69940W (2008).
93. J. J. Healy, B. M. Hennelly, and J. T. Sheridan, "Additional sampling criterion for the linear canonical transform," *Opt. Lett.* **33**: 2599–2601 (2008).

CHAPTER 11

Phase Space in Ultrafast Optics

Christophe Dorrer

Laboratory for Laser Energetics, University of Rochester, USA

Ian Wamsley

Department of Physics, University of Oxford, USA

11.1 Introduction

Time scales for dynamical measurements of optical pulses now cover both the femto- and the attosecond domain, due to the rapid evolution of ultrafast technology in the last decade.[1–4] Optical frequency synthesis enables the generation of single-cycle optical pulses with spectral bandwidths of several octaves.[5] Novel light sources, based on frequency mixing in nonlinear optical structures, are being developed for imaging applications such as optical coherence tomography for medical diagnostics. Further, table-top laser systems are now capable of generating subpicosecond pulsed radiation in the XUV region, so that attosecond pulses are routinely generated and characterized.[6] These developments presage a host of new applications including attoscience of atoms and molecules and new types of time standards based on precision frequency measurement from the microwave to the UV. The need for measurement methods for ultrashort optical pulses is therefore as compelling as ever. New spectral regions need to be accessed, along with ever briefer durations, as well as pulses with increasingly complex space-time structure, in order to access new phenomena and develop new technologies.

Phase-space descriptions of optical pulses, ultrafast processes, pulse manipulation, and measurement provide both a convenient

framework for developing intuition about how such pulses propagate and interact with matter and a set of rigorous calculation tools that enable information to be extracted efficiently and accurately from experimental data. In this chapter, we develop the phase-space description of ultrafast processes and its application to the characterization of ultrashort optical pulses. We make use of the strong analogy between the propagation of pulses in time through dispersive optical elements with the propagation of beams in space through paraxial optical systems, since this has played an important role in developing concepts for measurement.

11.2 Phase-Space Representations for Short Optical Pulses

The fundamental quantity describing an isolated, individual pulse of light is its electric field vector $\vec{\varepsilon}(\vec{x}, t)$. This is a function of time t and space $\vec{x}$ or equivalently optical frequency ω and transverse wave vector $\vec{k}$. In all but the most intense pulses, the magnetic field does not affect the interaction of the pulse with matter and can be estimated directly from the electric field. Characterizing an optical pulse therefore involves estimating the space-time dependence of the electric field.

Since the electric field $\vec{\varepsilon}(\vec{x}, t)$ is the fundamental entity in Maxwell's theory, the ability to measure it precisely not only provides the ne plus ultra of diagnostics, but also enables new experimental methods. When electromagnetic radiation interacts with matter, both its amplitude and its phase can be altered. The changes induced by the interaction can yield important information about the material dynamics: in fact, proper characterization of the temporal amplitude and phase of the field can potentially lead to complete reconstruction of the response function of the system. The spatiotemporal structure of the input and output fields provides all the available information from an optical experiment and therefore provides the data for the most exacting tests of models of the process under consideration.

11.2.1 Representation of Pulsed Fields

Typically in ultrafast optics the problem is simplified by taking a scalar approximation to the field vector. Simultaneous measurements of two orthogonal polarizations can then be combined to give the full vector field. Within this approximation, the real electric field, $\varepsilon(t)$ underlying an optical pulse (suppressing, for brevity, the spatial dependence)* is twice the real part of its analytic signal $E(t)$: $\varepsilon(t) = 2 \times Re[E(t)]$.

*The appropriate dispersion relation for the medium or structure in which the pulse propagates provides a connection that reduces the number of variables to three.

The analytic signal is the single-sided inverse Fourier transform of the Fourier transform of the field

$$E(t) = \frac{1}{\sqrt{2\pi}} \int_0^\infty d\omega\, \tilde{\varepsilon}(\omega)\, \exp(-i\omega t) \tag{11.1}$$

where

$$\tilde{\varepsilon}(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty dt\, \varepsilon(t)\, \exp(i\omega t) \tag{11.2}$$

The electric field is considered to have compact support in the time domain and is further assumed to have no spectral component at $\omega = 0$, so $\tilde{\varepsilon}(0) = 0$ (the electric field of a pulse propagating in a charge-free region of space must have zero area). The analytic signal is complex and therefore can be expressed uniquely in terms of an amplitude and phase

$$E(t) = |E(t)| \exp[i\phi_t(t)]\, \exp(i\phi_0)\, \exp(-i\omega_0 t) \tag{11.3}$$

where $|E(t)|$ is the time-dependent envelope, ω_0 is the carrier frequency (usually chosen near the center of the pulse spectrum), $\phi_t(t)$ is the time dependent phase, and ϕ_0 is a constant. The square of the envelope $I(t) = |E(t)|^2$ is the time-dependent instantaneous power of the pulse, which can be measured if a square-law photodetector of sufficient bandwidth is available. The derivative of the time-dependent phase accounts for the occurrence of different frequencies at different times; that is, $\Omega(t) = -\partial\phi_t/\partial t$ is the instantaneous frequency of the pulse that describes the oscillations of the electric field around that time, although such interpretation can be difficult.[7,8]

The frequency representation of the analytic signal is the Fourier transform of $E(t)$

$$\begin{aligned} \tilde{E}(\omega) = |\tilde{E}(\omega)|\, \exp\left[i\phi_\omega(\omega)\right] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty dt\, E(t)\, \exp(i\omega t) \\ &= \begin{cases} \tilde{\varepsilon}(\omega) & \omega > 0 \\ 0 & \omega \le 0 \end{cases} \end{aligned} \tag{11.4}$$

Here $|\tilde{E}(\omega)|$ is the spectral amplitude and $\phi_\omega(\omega)$ is the spectral phase. The square of the spectral amplitude $\tilde{I}(\omega) = |\tilde{E}(\omega)|^2$ is the spectral intensity (strictly speaking, this quantity is the spectral density—the quantity measured in the familiar way by means of a spectrometer followed by a photodetector). The spectral phase describes the relative phases of the optical frequencies composing the pulse, and

its derivative $\partial\phi_\omega/\partial\omega$ is the group delay $T(\omega)$ at the corresponding frequency, i.e., the time of arrival of a group of optical frequencies around ω.

To reconstruct the temporal electric field, it is necessary and sufficient to measure its Fourier transform for a finite set of frequencies. The Whittaker-Shannon sampling theorem[9] asserts that if the field has compact support contained in a range Δt, a sampling of $\tilde{E}(\omega)$ at the Nyquist frequency interval of $2\pi/\Delta t$ is sufficient for reconstructing the analytic signal $E(t)$ and consequently the electric field $\varepsilon(t)$ exactly.

These equivalent representations of the field in terms of the complementary variables t and ω suggest that an appropriate phase space for representing the fields is the two-dimensional chronocyclic phase space (t, ω) in which these variables are arguments of joint time-frequency distributions describing the pulse field. Such distributions provide a description that is relevant for measuring pulses using standard photodetectors. This is because they account properly for the fluctuations in the set of pulses that contribute to the detected signal. Further, they also provide an intuitive representation of the temporal and spectral structure of the pulses, such as a time-dependent frequency or chirp.

11.2.2 Pulse Ensembles and Correlation Functions

Applications and experiments involving ultrashort optical pulses often rely on a train or ensemble of pulses rather than a single pulse. In this case the pulses must be characterized using quantities related to the ensemble of which they are realizations. The mean electric field may be defined in some cases for the ensemble as the square root of the mean intensity times the exponential of the mean phase. This may be measured directly provided a single-shot measurement is possible for each pulse of the ensemble; that is, the mean intensity and phase are obtained respectively by independently averaging the measured intensities and the measured phases. However, the mean quantities might not give an insightful picture of the pulse ensemble (e.g., one could have large variations of the spectral phase from pulse to pulse and obtain a mean spectral phase that is identically zero). Furthermore, it is more usual for a multishot measurement to be made and the detected signal averaged over this sample set of pulses. Even assuming that the ensemble is ergodic, it is not the case that the field reconstructed from the averaged signal is the mean field of the ensemble. In certain cases, though, it is possible to show this directly. Such a mean electric field is not the most general or useful quantity—often the fluctuations of the pulses are important. When this is the case, the electric field amplitude and phase of an individual pulse may be meaningless.

One quantity, perhaps the simplest, that describes some of the statistical properties of the ensemble is the nonstationary two-time field correlation function

$$\Gamma(t_1, t_2) = \langle E(t_1)E^*(t_2)\rangle \tag{11.5}$$

where the angle brackets indicate an average over the ensemble of pulses and time is referenced to a frame moving at the pulse velocity. If each pulse in the train is an independent realization of a stochastic ensemble, the time average is equivalent to an ensemble average, by definition. This enables the coherence of the train to be defined operationally in a reasonable way. For a train of identical pulses $\Gamma(t_1, t_2)$ factorizes into $E(t_1)E^*(t_2)$, and the analytic signal $E(t)$ is proportional to $\Gamma(t, t_2)$, where t_2 is such that $E(t_2)$ is nonzero. Thus, any pulse measurement method capable of reconstructing the correlation function is also capable of returning the electric field when such a description will suffice.

This definition of the ensemble is suitable for describing a train of optical pulses for which the pulse-to-pulse temporal phase (and indeed amplitude) varies more or less randomly. It is always adequate for situations where pulses are measured individually. However, when averages are taken over trains of pulses, especially those for which the carrier-envelope phase is fixed,[10] the proper description of the ensemble must involve consideration of the field of the entire train, which has a nonstationary correlation function. It is formally quite difficult to formulate rigorously even the simplest of concepts, such as the spectrum, for a nonstationary field such as this. Procedures along the lines of those developed by Wiener and Khintchine[11] must then be extended to define properly the correlation function of nonstationary fields.[12]

The correlation function $\Gamma(t_1, t_2)$ provides a quantitative description of fluctuations from pulse to pulse in the electric field at time t_1 relative to those at time t_2. This is a complete description of the pulse ensemble as long as the fluctuations obey normal (or Gaussian) statistics. If not, then it is the simplest of a hierarchy of multitime correlation functions defining the ensemble. The degree to which an ensemble consists of identical pulses may be obtained from $\Gamma(t_1, t_2)$ in terms of an integral degree of temporal coherence μ, where μ is readily derived from the time-domain analog of Born and Wolf's[11] degree of coherence $\gamma(t_1, t_2)$, by first redefining the two-time correlation function in terms of a center-time coordinate t and a difference-time coordinate Δt

$$C(t, \Delta t) = \Gamma(t_1, t_2) \tag{11.6}$$

where $t = (t_1 + t_2)/2$ and $\Delta t = t_1 - t_2$. Then $\gamma(t_1, t_2)$ is defined as

$$\gamma\left(t + \frac{\Delta t}{2}, t - \frac{\Delta t}{2}\right) = \frac{C(t, \Delta t)}{[C(t + \Delta t/2, 0)\, C(t - \Delta t/2, 0)]^{1/2}} \tag{11.7}$$

Using the Schwarz inequality, it is straightforward to show that $0 \leq |\gamma(t + \Delta t/2, t - \Delta t/2)| \leq 1$. This leads directly to the inequality

$$0 \leq |C(t, \Delta t)|^2 \leq C\left(t + \frac{\Delta t}{2}, 0\right) C\left(t - \frac{\Delta t}{2}, 0\right) \tag{11.8}$$

The upper and lower bounds on the degree of coherence follow from Eq. (11.8). However, it is difficult to determine $\gamma(t + \Delta t/2, t - \Delta t/2)$ experimentally since it becomes singular for times at which $C(t, \Delta t)$ is zero. A more practically useful definition is offered by integrating Eq. (11.8) over the entire $(t, \Delta t)$ space and dividing by the quantity on the right-hand side, leading to the integral degree of coherence μ

$$0 \leq \mu = \iint \frac{dt\, d\Delta t |C(t, \Delta t)|^2}{\left[\int dt\, C(t, 0)\right]^2} \leq 1 \tag{11.9}$$

Here and in the remainder of this chapter, all integrals are understood to be from $-\infty$ to $+\infty$. An integral degree of coherence strictly smaller than 1 corresponds to a partially coherent train in which the pulse amplitude and/or phase fluctuates, in which case $C(t, \Delta t)$ is the fundamental quantity of interest. When $\mu = 1$, the ensemble is said to be fully coherent (identical pulses) and $C(t, \Delta t)$ factorizes. In the latter case the electric field becomes the fundamental quantity of interest and is readily retrieved from the two-time correlation function using

$$|E(t)| = \sqrt{C(t, 0)} \tag{11.10}$$

and, with t_2 held fixed,

$$\text{Arg}[E(t)] = \tan^{-1}\left\{\frac{\text{Im}\,[C[(t + t_2)/2, t - t_2]]}{\text{Re}\,[C[(t + t_2)/2, t - t_2]]}\right\} + \phi_0 \tag{11.11}$$

where ϕ_0 is an undetermined constant. It is important to note that Eqs. (11.10) and (11.11) are valid only if the integral degree of coherence has been explicitly demonstrated to be equal to unity, which of course requires that the two-time correlation function or equivalent representation in frequency or phase space be measured. Thus, in cases where an ensemble or train of pulses, rather than an individual pulse, is used for application or experimentation, pulse-shape characterization efforts must ultimately be directed toward measurement of the ensemble statistics.

The two-frequency correlation function is linked to the two-time correlation function of the ensemble by a double Fourier transform

$$\tilde{\tilde{C}}(\Delta\omega, \omega) = \left\langle \tilde{E}\left(\omega + \frac{\Delta\omega}{2}\right) \tilde{E}^*\left(\omega - \frac{\Delta\omega}{2}\right) \right\rangle$$
$$= \frac{1}{2\pi} \iint dt\, d\Delta t\, C(t, \Delta t) \exp[i(t\Delta\omega + \Delta t\omega)] \quad (11.12)$$

The center-frequency and difference-frequency coordinates in Eq. (11.12) are given by $\omega = (\omega_1 + \omega_2)/2$ and $\Delta\omega = \omega_1 - \omega_2$, respectively. Similar arguments to those mentioned for the two-time-correlation function apply to the two-frequency correlation function.

For a coherent train of pulses, Eqs. (11.10) and (11.11), and their equivalent in the frequency domain, indicate that the time or frequency representation of the analytic signal can be reconstructed from a single line of the corresponding correlation function. Therefore, if the ensemble is assumed a priori to be coherent, the amount of collected data can be greatly reduced. This is a luxury afforded only to those measurement techniques that directly measure one of the correlation functions.

11.2.3 The Time-Frequency Phase Space

Time-frequency distributions are central to the characterization of pulses in the optical domain, since they are straightforwardly related to the measured data. In optics, direct measurement of the waveform is not possible. This is in contrast to the more usual application of the distributions in signal processing, where they are commonly used as mathematical tools for signal representation. It is frequently useful to work with a representation of the correlation functions in the chronocyclic phase space. The intuitive concept of chirp (that is, time-dependent frequency in the pulse) can be most easily seen within this space. The pulse ensemble may also be represented within the chronocyclic phase spaces defined by the complementary variables (t, ω) and $(\Delta\omega, \Delta t)$. The chronocyclic Wigner function $W(t, \omega)$ and ambiguity or Wigner characteristic function $A(\Delta\omega, \Delta t)$ provide two particularly useful descriptions of the pulse train statistics in these spaces. The relationship between the various representations of the correlation function has been discussed in the context of spatially localized fields,[13] and the Wigner function was originally applied to problems in ultrafast optics.[14–16] Examples of applications of the Wigner function, ambiguity function, and other time-frequency distributions in ultrafast optics can be found in Refs. 17 to 29. General properties of the Wigner and ambiguity functions can, for example, be found in Ref. 30.

The Wigner function is obtained by taking the one-dimensional Fourier transform of $C(t, \Delta t)$ over the time-difference coordinate

$$\begin{aligned} W(t, \omega) &= \frac{1}{\sqrt{2\pi}} \int d\Delta t\, C(t, \Delta t) \exp(i\omega\Delta t) \\ &= \frac{1}{\sqrt{2\pi}} \int d\Delta t \left\langle E\left(t + \frac{\Delta t}{2}\right) E^*\left(t - \frac{\Delta t}{2}\right) \right\rangle \exp(i\omega\Delta t) \\ &= \frac{1}{\sqrt{2\pi}} \int d\Delta\omega \left\langle \tilde{E}\left(\omega + \frac{\Delta\omega}{2}\right) \tilde{E}^*\left(\omega - \frac{\Delta\omega}{2}\right) \right\rangle \exp(-i\Delta\omega t) \end{aligned} \tag{11.13}$$

The ambiguity function is obtained from $C(t, \Delta t)$ by performing the Fourier transform over the average-time coordinate

$$\begin{aligned} A(\Delta\omega, \Delta t) &= \frac{1}{\sqrt{2\pi}} \int dt\, C(t, \Delta t) \exp(i\Delta\omega t) \\ &= \frac{1}{\sqrt{2\pi}} \int dt \left\langle E\left(t + \frac{\Delta t}{2}\right) E^*\left(t - \frac{\Delta t}{2}\right) \right\rangle \exp(i\Delta\omega t) \\ &= \frac{1}{\sqrt{2\pi}} \int d\omega \left\langle \tilde{E}\left(\omega + \frac{\Delta\omega}{2}\right) \tilde{E}^*\left(\omega - \frac{\Delta\omega}{2}\right) \right\rangle \exp(-i\omega\Delta t) \end{aligned} \tag{11.14}$$

These representations are uniquely and invertibly related to one another by Fourier transformations.

The Wigner function has some features that make it useful in representing short optical pulses. For example, in contrast to the field and correlation representations, it is real-valued. Moreover, its time and frequency marginals (i.e., projections on the corresponding axis) are the temporal and spectral intensity, respectively. The average time-dependent intensity is obtained from the two-time correlation function by setting $\Delta t = 0$. This corresponds to a projection of the Wigner function onto the time axis, or the Fourier transform of the $\Delta t = 0$ section of the ambiguity function

$$\begin{aligned} I(t) = C(t, 0) &= \frac{1}{\sqrt{2\pi}} \int d\omega\, W(t, \omega) \\ &= \frac{1}{\sqrt{2\pi}} \int d\Delta\omega\, A(\Delta\omega, 0) \exp(-i\Delta\omega t) \end{aligned} \tag{11.15}$$

Furthermore, the average pulse spectral intensity is obtained from the two-frequency correlation function by setting $\Delta\omega = 0$, by projecting

the Wigner function onto the frequency axis, or by taking the Fourier transform of the $\Delta\omega = 0$ section of the ambiguity function

$$\tilde{I}(\omega) = \tilde{\tilde{C}}(0, \omega) = \frac{1}{\sqrt{2\pi}} \int dt\, W(t, \omega)$$
$$= \frac{1}{\sqrt{2\pi}} \int d\Delta t\, A(0, \Delta t)\, \exp(i\omega\Delta t) \quad (11.16)$$

As shown in Fig. 11.1, the Wigner function provides an intuitive representation of the pulse field, in particular the notion that different frequencies may occupy different time slots in the pulse. Figure 11.1*a* displays the Wigner function of a Fourier-transform-limited pulse

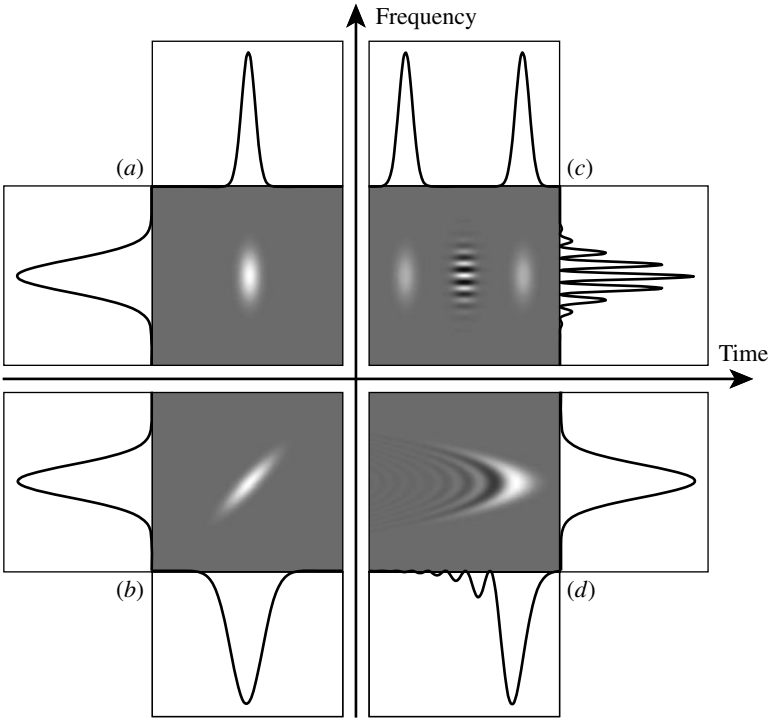


FIGURE 11.1 Wigner functions of (*a*) a Fourier-transform-limited Gaussian pulse, (*b*) a pulse with Gaussian spectrum and quadratic spectral phase, (*c*) a pair of identical Fourier-transform-limited Gaussian pulses, and (*d*) a pulse with Gaussian spectrum and third-order spectral phase. In each case, the temporal and spectral marginals are plotted. Light and dark regions, respectively, correspond to positive and negative values of the Wigner function.

(i.e., the spectral phase is at most a linear function of the optical frequency), indicating no correlation between time and frequency localization. Figure 11.1*b* is the Wigner function of a pulse with the same spectrum and a parabolic spectral phase, i.e., a linear chirp. This leads to stretching of the pulse in the time domain, as indicated by the temporal marginal. Correlation between the temporal and spectral components of the pulse can be inferred from the slope of the Wigner function in the chronocyclic space. The amount of such chirp may be quantified using the Wigner distribution in a way that respects Fourier's theorem, which may be thought to preclude the simultaneous specification of time and frequency. The *instantaneous frequency* $\Omega(t)$ may be defined as the instantaneous mean value of the frequency of the distribution

$$\Omega(t) = \frac{\int d\omega\, \omega W(t, \omega)}{\int d\omega\, W(t, \omega)} \tag{11.17}$$

This can be evaluated by using integration by parts, making use of the compact support of the pulse field in the formula for the Wigner function. This leads to the relation

$$\Omega(t) = -\frac{d\phi_t}{dt}(t) = -\phi_t'(t) \tag{11.18}$$

which embodies the intuitive result that the instantaneous frequency is the temporal derivative of the temporal phase. The group delay can be calculated similarly as

$$T(\omega) = \frac{\int dt\, tW(t, \omega)}{\int dt\, W(t, \omega)} = \frac{d\phi_\omega}{d\omega}(\omega) = \phi_\omega'(\omega) \tag{11.19}$$

which corresponds to the common interpretation of the group delay. The Wigner function also has some less intuitive features. For example, it is tempting to consider W as a joint probability distribution of the time at which different frequencies occur within the pulse ensemble. But since W is not a positive definite function, it cannot play the role of a probability distribution. Negativity of the Wigner function is a common phenomenon. For example, Fig. 11.1*c* displays the Wigner function of a pair of temporally delayed identical pulses. Interference between the two pulses is indicated at the center of the chronocyclic space by an alternation of positive and negative regions of the Wigner function. Note that the frequency marginal, i.e., the optical spectrum, is positive, since negative regions of the cross-terms of the Wigner function at the center of the chronocyclic space are canceled by the positive Wigner functions of each individual pulse. Finally, the Wigner function of a pulse with a third-order spectral phase is plotted in Fig. 11.1*d*. This Wigner function also takes negative values, but its shape remains

indicative of the group delay of the pulse, i.e., a parabolic function of the optical frequency.

There exist an infinite number of time-frequency distributions that can potentially be used to represent a signal in the chronocyclic space. One approach to obtain bilinear time-frequency distributions is to use a signal-independent kernel to generate the set of functions[30]

$$R(t, \omega) = \frac{1}{(2\pi)^{3/2}} \iiint d\Delta t\, du\, d\theta\, E\left(u + \frac{\Delta t}{2}\right) E^*\left(u - \frac{\Delta t}{2}\right) \times K(\theta, \Delta t)\, \exp(i\omega\Delta t + i\theta t - i\theta u) \tag{11.20}$$

For example, K uniformly equal to 1 leads to the Wigner function, while K equal to $\exp(-i\theta\,|\Delta t|)$ leads to the Page distribution which has also been used in the context of representing linear optical systems.[31] In the coherent case, if K is chosen as the ambiguity function A_g of an ancillary signal g as defined by Eq. (11.14), the time-frequency distribution of Eq. (11.20) becomes

$$R_g(t, \omega) = \left| \frac{1}{\sqrt{2\pi}} \int du\, E(u) g(u - t)\, \exp(i\omega u) \right|^2 \tag{11.21}$$

Equation (11.21) indicates that $R_g(t, \omega)$ can be interpreted as the optical spectrum of the field after gating by the function g, represented as a function of the optical frequency ω and the relative delay t between the pulse and the gate. This representation is known as a (Gabor) spectrogram. This particular time-frequency distribution is evidently positive and can be measured directly by applying a time gate g to the optical test pulse and measuring the resulting spectrum. The marginals of the spectrogram are convolutions of the corresponding intensity of the test pulse with the intensity of the ancillary function. Therefore the marginals are not equal to the temporal and spectral intensities of the pulse. It is interesting to note that Eq. (11.21) can also be written as

$$R_g(t, \omega) = \frac{1}{2\pi} \iint du\, d\Omega\, W_E(u, \Omega) W_g(u - t, \omega - \Omega) \tag{11.22}$$

which is valid regardless of the coherence of the ensemble of optical pulses. This indicates that the spectrogram is obtained by convolution of the Wigner function of the pulse with the Wigner function of the gate in the chronocyclic space, and that its measurement requires a full scan of the chronocyclic space with the Wigner function of the gate. There exist other bilinear chronocyclic representations of pulses that are positive definite, and therefore may be used as joint probability distributions. However, these do not have the property that their marginal distributions are the temporal and spectral intensity. Phase-space representations are not limited to the bilinear functions

generated by Eq. (11.20). For example, the time-frequency representation $P(t, \omega) = I(t)\tilde{I}(\omega)$ is positive, and its marginals are the temporal and spectral intensity of the pulse.[30] However, it is not uniquely related to a field and does not represent chirp properly since it is not phase-dependent.

Coming back to bilinear time-frequency distributions, an entire class of chronocyclic representations may be derived from the Wigner function by means of a convolution

$$P_s(t, \omega) = \frac{1}{2\pi} \iint d\omega' \, dt' \, W(t', \omega') G_s(t - t', \omega - \omega') \tag{11.23}$$

where

$$G_s(t, \omega) = \frac{4}{s} \exp\left[-\frac{1}{s}\left(\frac{\omega^2}{\gamma^2} + 4\gamma^2 t^2\right)\right] \tag{11.24}$$

This class is analogous to the commonly used phase-space representations of the optical field in quantum optics.[32,33] For $s = 0$, the convolving function is a Dirac function, and P_0 is the Wigner function of the pulse. Positive values of s correspond to smoothing in the chronocyclic space, analogous to the Q function used in quantum physics. The time-frequency distribution defined by Eq. (11.23) is positive for s larger than 2. In the particular case of $s = 2$, G_s is the Wigner function of a coherent state, and P_2 corresponds to the spectrogram of the pulse defined for coherent ensembles by Eq. (11.21), the gating function being the Gaussian function

$$g(t) = \sqrt{2\gamma} \exp(-\gamma^2 t^2) \tag{11.25}$$

A set of smoothed Wigner functions corresponding to a Gaussian pulse with third-order spectral phase is plotted in Fig. 11.2. The phase-space representation of the pulse evolves from the Wigner function

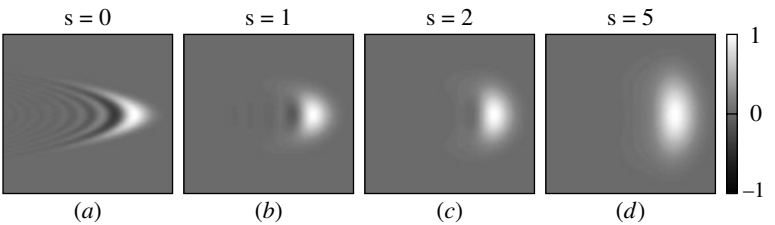


FIGURE 11.2 Smoothed Wigner functions P_s of a pulse with Gaussian spectrum and third-order spectral phase for (*a*) $s = 0$, (*b*) $s = 1$, (*c*) $s = 2$, and (*d*) $s = 5$. Part *a* corresponds to the Wigner function and part *c* corresponds to a spectrogram.

($s = 0$, Fig. 11.2*a*) to the spectrogram ($s = 2$, Fig. 11.2*c*) and is positive for values of s larger than 2. As s increases, the Wigner function is smoothed out, but the resulting time-frequency distribution loses its ability to display chirp.

In fact, it is possible to cast all measurement strategies in terms of phase-space distributions, in a form that ensures the positivity of the signal distribution. The requirement that the signal be positive arises from the way in which optical detectors respond to the field. They are square-law detectors in which the intensity or energy (depending on the response time of the detector) gives rise to a photocurrent or charge. This implies that the detected signal is positive. As we have seen, not all phase-space distributions are positive, but a theorem due to Jordan shows that the distribution

$$S(T, \Omega) = \int d\omega\, dt\, W(t, \omega) W_M(t - T, \omega - \Omega) \tag{11.26}$$

with $W_M(t, \omega)$ the Wigner representation of the measurement apparatus, is always positive. Apparatuses for which the measurement function is a convolution, no matter how complex and negative the Wigner function of the pulse and the measurement apparatus are, yield a signal that is positive.

In general the measurement function may be written in the form

$$S(T, \Omega) = \int d\omega\, dt\, W(t, \omega) W_M(t, \omega; \Omega, T) \tag{11.27}$$

where the measurement Wigner function contains the apparatus parameters. This cannot always be written as a convolution, as in Eq. (11.26), but nonetheless is always constrained to give a positive signal. Further, it is clear from these formulas that any measurement technique attempting to reconstruct either the Wigner or the ambiguity function must be capable of exploring the entire two-dimensional chronocyclic phase space.

11.2.4 Phase-Space Representation of Paraxial Optical Systems

The representation of optical pulses in phase space can be understood by analogy to the representation of optical ray trajectories in geometrical optics.[34] This provides an important first-order design framework for ultrafast optical systems as well as an intuitive appreciation of the more formal representations of the chronocyclic distributions. Paraxial optical systems are specified by a 2×2 real transfer matrix $\mathbf{T}$,

$$\mathbf{T} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \tag{11.28}$$

with $\det(\mathbf{T}) = 1$. This matrix relates the input and output properties of the ray trajectories, i.e., the ray height y with respect to the principal ray of a bundle and its angle u with respect to this ray.† The refractive index of the medium n in which the ray propagates is usually appended to the ray angle, so that the specification of the ray is the column vector

$$\vec{Y} = \begin{pmatrix} y \\ u \end{pmatrix} \tag{11.29}$$

The output and input rays are related by the equation

$$\vec{Y}_{\text{out}} = \mathbf{T}\,\vec{Y}_{\text{in}} \tag{11.30}$$

This relation may be represented in the phase space consisting of a transverse coordinate (the ray height) and the corresponding transverse wave vector (proportional to the ray angle). A single ray is a point in this space, a ray bundle emanating from a single point occupies a region of constant height, and a plane wave occupies a region of the phase space of constant angle (Fig. 11.3).

The elements of the transfer matrix may be derived by using Hamilton's characteristic function in the paraxial approximation, i.e., using the Fresnel approximation to the propagation kernel in the Kirchhoff formula. The output and input scalar electric fields for such a system are related by

$$E_{\text{out}}(x) = \int dx'\, K(x, x') E_{\text{in}}(x') \tag{11.31}$$

The most general form of the Fresnel kernel K is (for light of wave number $k_0 = 2\pi/\lambda$)

$$K(x, x') = \sqrt{\frac{ik_0}{2\pi B}}\, \exp\left[-\frac{ik_0}{2B}(Ax^2 - 2xx' + Dx'^2)\right] \tag{11.32}$$

The parameters of the transfer matrix determine the action of the optical system on the input field through this kernel. This may be illustrated by some important simple paraxial optical elements.

†The notation used for specifying the phase-space coordinates of a ray in geometrical optics is conventionally in terms of the ray height y and the ray angle u with respect to the principal ray of the bundle. This notation is used, e.g., in paraxial ray tracing for first-order system design, which is sometimes known as *ynu* tracing. The corresponding coordinates used for specifying the electric field are more usually the transverse coordinate $x = y$ and the transverse wave vector $k_x = k_0 \tan u \cong k_0 u$ with $k_0 = 2\pi/\lambda$, where λ is the optical wavelength.

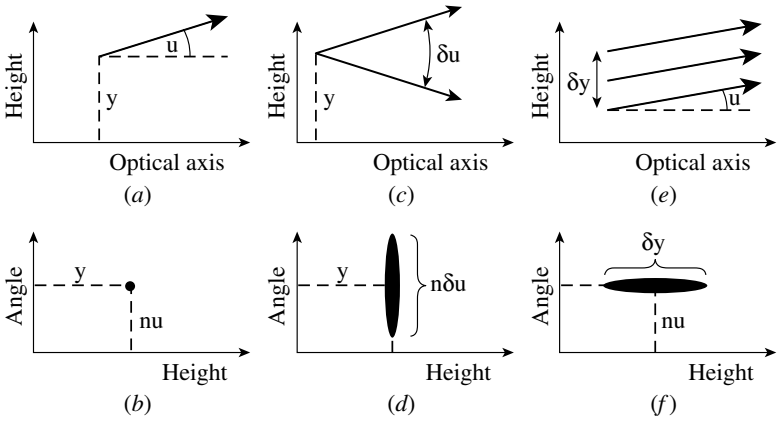


FIGURE 11.3 Representations of various ray bundles in geometrical optics and in the associated phase space. The first row corresponds to the geometrical optics representations of (*a*) a single ray located at height y with angle u, (*c*) a bundle of rays emanating from a point of height y in a range of angles δu, and (*e*) a plane wave covering a range of heights δy propagating at an angle u. The second row [plots (*b*), (*d*), and (*f*)] displays the corresponding phase-space representations.

The transfer matrix for free-space propagation over a distance L is

$$\mathbf{T}_{\text{prop.}} = \begin{pmatrix} 1 & L \\ 0 & 1 \end{pmatrix} \tag{11.33}$$

with the corresponding space-shift-invariant Fresnel kernel

$$K(x, x') = \sqrt{\frac{ik_0}{2\pi L}} \exp\left[-\frac{ik_0}{2L}(x - x')^2\right] \tag{11.34}$$

Free-space propagation therefore increases the spatial coordinate proportionally to the propagation distance and the associated wave vector $y_{\text{out}} = y_{\text{in}} + Lu$, but does not modify the wave vector, so $u_{\text{out}} = u_{\text{in}}$ (Fig. 11.4*a*). This can be seen as a shear along the position direction in the phase space (Fig. 11.4*b*).

The matrix describing propagation through a thin lens is

$$\mathbf{T}_{\text{lens}} = \begin{pmatrix} 1 & 0 \\ -1/f & 1 \end{pmatrix} \tag{11.35}$$

with the corresponding kernel

$$K(x, x') = \exp\left(\frac{ik}{2f}x^2\right)\delta(x - x') \tag{11.36}$$

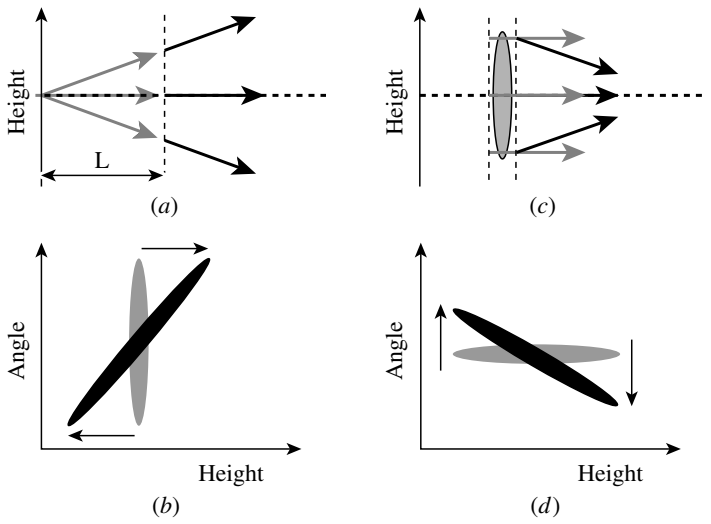


FIGURE 11.4 Representations of free-space propagation and propagation in a thin lens in geometrical optics and in the associated phase space. The first row corresponds to the geometrical optics representations of (*a*) three different rays propagating in free space, where it can be seen that the rays before (depicted in gray) and after (depicted in black) conserve their angle and acquire an angle-dependent height, and (*c*) of three different rays propagating through a thin lens, where it can be seen that the rays before and after the lens conserve their height and acquire a position-dependent angle. The second row displays the corresponding phase-space representations before (in gray) and after (in black), where the effect of free-space propagation is seen in (*b*) as a shear along the height direction, and the effect of the thin lens is seen in (*d*) as a shear along the angle direction.

Because this element is not space-shift-invariant, but rather angle-shift-invariant, its form in the conjugate angle space is simpler.

$$\tilde{\tilde{K}}(k_x, k'_x) = \sqrt{\frac{f}{k}} \exp\left[-\frac{if}{2k}(k_x - k'_x)^2\right] \tag{11.37}$$

The thin lens therefore modifies the wave vector proportionally to the lateral position of the ray on the lens, that is, $u_{\text{out}} = u_{\text{in}} - y_{\text{in}}/f$, but the ray height does not change, that is, $y_{\text{out}} = y_{\text{in}}$ (Fig. 11.4*c*). This is a shear in the phase space along the wave vector direction (Fig. 11.4*d*). A lens can be combined with free-space propagation to rotate the phase-space representation. The transfer matrix directly describes modifications of the Wigner function through optical systems.[35]

An important feature of the paraxial approximation is that it is straightforward to propagate Gaussian beams using the transfer matrices acting on the complex beam parameter

$$\frac{1}{q} = \frac{1}{R} - \frac{i\lambda}{\pi w^2} \tag{11.38}$$

In Eq. (11.38), R is the radius of curvature of the beam at the reference plane, w is the corresponding beam size, and λ is the index-dependent wavelength in the medium. The complex beam parameters before and after propagation, q and q', respectively, are linked by the formula

$$\frac{q'}{n'} = \frac{Aq/n + B}{Cq/n + D} \tag{11.39}$$

where n and n' are the optical index in the medium before and after propagation, respectively. The elements of the transfer matrix of Eq. (11.28) are then interpreted as modifying the beam waist and radius of curvature, accordingly.

11.2.5 Temporal Paraxiality and the Chronocyclic Phase Space

An optical pulse may be represented in a manner similar to optical rays in geometrical optics. The analogy between space and time, the space-time duality, has been very fruitful.[36–38] Consider the action of a linear filter on a pulsed field. The relationship between input and output fields for the filter is

$$E_{\text{out}}(t) = \int dt'\, H(t, t') E_{\text{in}}(t') \tag{11.40}$$

which can also be written in the frequency domain as

$$\tilde{E}_{\text{out}}(\omega) = \int d\omega' \tilde{H}(\omega, -\omega') \tilde{E}_{\text{in}}(\omega') \tag{11.41}$$

Let us define a temporally paraxial approximation and postulate a general linear filter function in the form of a *temporal Fresnel kernel*

$$H(t, t') = \frac{1}{\sqrt{2\pi b}} \exp\left[-\frac{i}{2b}(at^2 - 2tt' + dt'^2)\right] \tag{11.42}$$

where a, b, and d are real numbers. Here H is unitary and verifies

$$\int dt\, H(t, t') H^*(t, t'') = \delta(t' - t'') \tag{11.43}$$

Then the parameters describing the filter form a temporal transfer matrix

$$\mathbf{T} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \tag{11.44}$$

which is again unimodular [$\det(\mathbf{T}) = ad - bc = 1$]. Define also a column vector

$$\vec{\Omega} = \begin{pmatrix} t \\ \omega \end{pmatrix} \tag{11.45}$$

which describes time of occurrence and frequency of a temporal "ray." The properties of the pulse are described by a "bundle" of such rays, which can be represented in the chronocyclic phase space to interpret the effects of various linear filters. As in the case of the geometrical optical rays, the output and input temporal rays are related by

$$\vec{\Omega}_{\text{out}} = \mathbf{T}\vec{\Omega}_{\text{in}} \tag{11.46}$$

Optical elements entirely analogous to free-space propagation, and imaging may also be defined. The kernel for propagation in a dispersive medium, i.e., with a frequency-dependent index of refraction, or for double-passing a two-grating compressor as represented in Fig. 11.5*a*, is

$$\tilde{\tilde{H}}(\omega, -\omega') = \exp\left(i\frac{\phi''_\omega}{2}\omega^2\right)\delta(\omega - \omega') \tag{11.47}$$

in the limit where only second-order dispersion ϕ''_ω is considered in the development of the introduced spectral phase ϕ_ω. The corresponding transfer matrix is

$$\mathbf{T}_{\text{dispersion}} = \begin{pmatrix} 1 & \phi''_\omega \\ 0 & 1 \end{pmatrix} \tag{11.48}$$

A pulse with such linear chirp is represented by a collection of ray vectors in which t is a linear function of ω

$$\{\vec{\Omega}_i\} = \left\{ \begin{pmatrix} t_i = t + \phi''_\omega \omega_i \\ \omega_i \end{pmatrix} \right\} \tag{11.49}$$

This manifests itself in the phase-space representation of the ray bundle by a slope related to ϕ''_ω (Fig. 11.5*b*).

Transfer through a quadratic temporal phase modulator (Fig. 11.5*c*) is the temporal analog of the action of a paraxial lens on a ray. The corresponding kernel is

$$H(t, t') = \exp\left(i\frac{\phi''_t}{2}t^2\right)\delta(t - t') \tag{11.50}$$

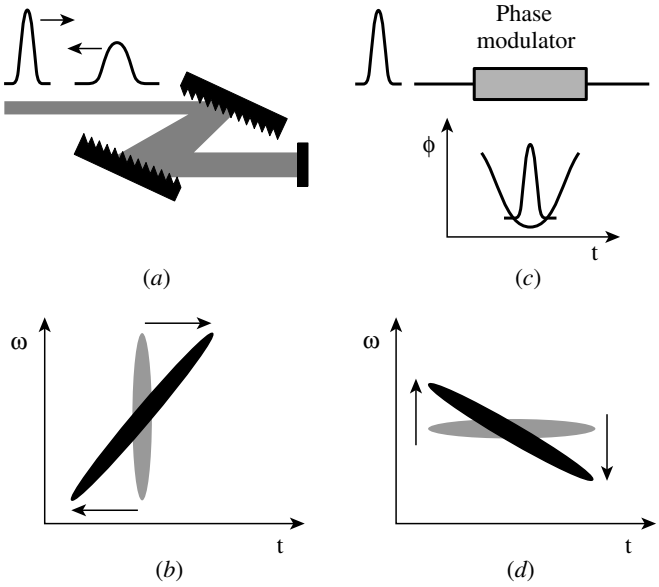


FIGURE 11.5 Representations of the effect of (*a*) dispersive propagation in a two-grating compressor and (*c*) propagation in a quadratic temporal phase modulator. (*b*) Dispersive propagation leads to a shear of the chronocyclic representation along the time axis. (*d*) The quadratic temporal phase modulator leads to a shear of the chronocyclic representation along the frequency axis.

belonging to the transfer matrix

$$\mathbf{T}_{\text{modulator}} = \begin{pmatrix} 1 & 0 \\ -\phi_t'' & 1 \end{pmatrix} \tag{11.51}$$

This adds a time-dependent instantaneous frequency $-\phi_t'' t$ to the pulse, as is observed in Fig. 11.5*d*. Analogy to the combined action of free-space propagation followed by a lens suggests that propagation in a time-stationary dispersive element (such as a grating pair compressor) and passage through a time-nonstationary phase modulator (such as an electrooptic crystal driven by a temporal parabolic voltage) will cause a rotation of the phase-space distribution of the temporal ray bundle in the chronocyclic phase space. The combination of dispersive propagation $\phi_{\omega,1}''$ followed by a temporal lens ϕ_t'', again followed by dispersive propagation $\phi_{\omega,2}''$, leads

to the transfer matrix

$$\mathbf{T} = \begin{pmatrix} 1 + \phi''_{\omega,2}\phi''_t & \phi''_{\omega,1} + \phi''_{\omega,2} + \phi''_{\omega,1}\phi''_{\omega,2}\phi''_t \\ \phi''_t & 1 + \phi''_{\omega,1}\phi''_t \end{pmatrix} \tag{11.52}$$

The particular case of $\phi''_{\omega,1} = \phi''_{\omega,2} = -1/\phi''_t$ yields the remarkable relations $t_{\text{out}} = -\omega_{\text{in}}/\phi''_t$ and $\omega_{\text{out}} = \phi''_t t_{\text{in}}$ between temporal and spectral coordinates of a ray bundle, in a fashion similar to the Fourier transform relation induced by a lens of focal length f between two planes located a distance f apart from the lens. This is known as the time-to-frequency converter because the temporal intensity of the input pulse can be recovered from a simple measurement of the spectral intensity of the output pulse.[39] Another interesting result is that with $1/\phi''_{\omega,1} + 1/\phi''_{\omega,2} + \phi''_t = 0$ (a condition referred to as *temporal imaging*), the upper right quadrant of Eq. (11.52) is zero, which leads to $t_{\text{out}} = (1 + \phi''_{\omega,2}\phi''_t)t_{\text{in}}$. Such assembly therefore magnifies the ray bundle in the time domain by the quantity $1 + \phi''_{\omega,2}\phi''_t$. Temporal magnification following this formalism has been used to decrease the resolution required to measure an optical waveform.[36,40–43]

A further useful analogy is that it is straightforward to propagate Gaussian pulses using the temporal transfer matrices. The complex pulse parameter

$$\Gamma^2 = \beta - i\frac{1}{\tau^2} \tag{11.53}$$

where β is the chirp parameter of the pulse at the reference plane and τ is the corresponding pulse duration, is modified simply according to the less well-known formula

$$\frac{\omega_0}{c\Gamma'^2} = \frac{A\omega_0/(c\Gamma^2) + B}{C\omega_0/(c\Gamma^2) + D} \tag{11.54}$$

The elements of the temporal transfer matrix are then interpreted as modifying the chirp and duration accordingly.

This formulation is useful in visualizing both the optical pulses and the strategies that are used to measure them. They are in broad agreement with formal definitions of phase-space distributions of the pulsed fields, although they only agree in detail in cases, such as for incoherent ensembles, when all quantities are positive definite. Further, it is useful in system design and analysis, because it provides a simple way to understand the first-order space-time couplings that occur when geometrical dispersion (such as happens in a prism or grating) is used to build a temporally dispersive delay line. In this case the paraxial and temporal transfer matrices are combined into 4×4 matrices that describe the coupling between the spectral and temporal

properties of the beam with the spatial and angular properties:

$$\mathbf{T} = \begin{pmatrix} A & B & \alpha & \beta \\ C & D & \gamma & \delta \\ \alpha' & \beta' & a & b \\ \gamma' & \delta' & c & d \end{pmatrix} \tag{11.55}$$

This matrix acts on the ray representation (x, k, t, ω) in the (space, wave vector, time, frequency) space. The 2×2 block diagonal elements are the spatial and temporal transfer matrices described above, and the off-diagonal elements are those involved in space-time coupling. For example, the matrix describing a diffraction grating or prism introducing angular dispersion (linear relation between wave vector and optical frequency) has a nonzero δ. A restricted form of this matrix, where 7 of the 16 variables of the matrix of Eq. (11.55) are constrained, has been applied to time-stationary filters by Kostenbauder.[44] Space-time Wigner functions have been used, e.g., to describe the space-time coupling induced by zero-dispersion line pulse shapers.[45,46]

11.3 Metrology of Short Optical Pulses

11.3.1 Measurement Strategies

In principle, one can use an antenna to directly measure the oscillating electric field. For instance, low-temperature GaAs gated antennas are routinely used to measure the oscillating field of electromagnetic pulses whose carrier frequencies are on the order of 1 THz. But the fastest antennas are far too slow to resolve the oscillations of optical fields (the period of one cycle of an optical field in the visible spectram is less than 3 fs). Detectors for the optical regime are square-law (or energy) detectors which only respond to the intensity of the field. State-of-the-art commercial photodiodes have response times on the order of 10 ps, while streak cameras, by far the fastest electronic detection devices, offer a temporal resolution of about 1 ps. Herein lies the problem of ultrashort pulse characterization: it is not possible to directly measure the temporal intensity of optical pulses with durations less than 1 ps or so. The problem is especially acute for the few-cycle optical regime and the XUV attosecond regime. Most conventional photodetection schemes are also not sensitive to the phase of the electric field. These problems may be circumvented, however, by passing the unknown (test) pulses through filters of known response functions and then recording the average output energy as a function of the parameters characterizing the filter response functions. As a general proposition, pulse measurement techniques may be categorized

according to the inversion algorithm they employ for reconstructing the ensemble statistics from the measured experimental trace. Phase-space techniques estimate either the Wigner function or the ambiguity function. There are two classes of phase-space techniques, spectrographic and tomographic. Methods from which the inversion returns either the two-time or two-frequency correlation function may be classified as direct techniques: these are interferometric.

11.3.2 Pulse Characterization Apparatuses as Linear Systems

Although currently most methods for pulse characterization are based on nonlinear optical processes, it is informative to consider a general framework for analyzing measurement methods based on a linear filter analysis. This approach has two benefits. First, it proves that nonlinearities are not necessary for determining the pulse field and shows why they are often used. Second, it specifies the necessary and sufficient conditions that must be fulfilled by any apparatus that is capable of determining the field. This leads to a convenient classification scheme for most methods.

Consider the general interferometer shown in Fig. 11.6. It consists of four causal filters and a square-law, integrating detector. Therefore, we may consider a two-beam interferometer for the analysis without loss of generality because a multibeam interferometer can always be decomposed into a linear combination of two-beam interferometers. Each filter is characterized by a time-domain response function $H_k(t, t')$. We take the filters, and therefore the interferometer, to be linear systems in the broad sense that the output field $E_{\text{out}}(t)$ after the filters can be expressed as a function of the input field $E_{\text{in}}(t)$ as

$$E_{\text{out}}(t) = \int dt' \, H_k(t, t') E_{\text{in}}(t') \tag{11.56}$$

for the kth filter in the sequence. For measurement schemes in which no interference of the filtered versions of the input pulse is required, H_3 and H_4 may be set to zero.

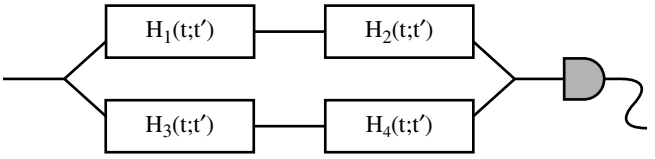


FIGURE 11.6 General interferometer for optical pulse characterization where H_k ($k = 1$ to 4) represents the action of linear filters on the electric field of the input pulse before a square-law time-integrating detector.

It turns out to be necessary to consider only two classes[47] of linear filter: time-stationary, in which the time of incidence of the input pulse does not affect the output, and frequency stationary, the output of which is unchanged by arbitrary frequency shifts of the input. A linear filter of arbitrary response function may be synthesized from these two classes. Moreover, they are the only classes of filter that have been used to date in pulse shape measurement and are the easiest to implement in practice. For a time-stationary filter, the output field is related to the input field by

$$E_{\text{out}}(t) = \int dt'\, S(t - t') E_{\text{in}}(t') \tag{11.57}$$

where the filter response function $H(t, t')$ is a function only of the difference in its arguments $t - t'$. A frequency-stationary filter is defined in an analogous manner in the spectral domain

$$\tilde{E}_{\text{out}}(\omega) = \int d\omega'\, \tilde{N}(\omega - \omega') \tilde{E}_{\text{in}}(\omega') \tag{11.58}$$

where the filter transfer function $\tilde{N}(\omega - \omega')$ is a function only of the difference in its arguments and the tilde represents a Fourier transform. Frequency-stationary filters are time-nonstationary, since their output depends on the time at which the pulse arrives at the input. We use S and $\tilde{S}$, and N and $\tilde{N}$, to denote the response functions and transfer functions of time-stationary and time-nonstationary filters, respectively.

There are two further important filter specializations: amplitude-only and phase-only. These filters behave as their names suggest; the former provides amplitude modulation while the later modulates only the phase. We distinguish amplitude-only and phase-only filters with the superscripts A and P, respectively. To be specific, we identify six filters to be used in this analysis and their corresponding response or transfer functions.

Time gate:

$$N^A(t; \tau) = \exp\left[\frac{-\Gamma^2 (t - \tau)^2}{2}\right] \tag{11.59}$$

Quadratic temporal phase modulator:

$$N_Q^P(t; \phi_t'') = \exp\left(\frac{i \phi_t'' t^2}{2}\right) \tag{11.60}$$

Linear temporal phase modulator or frequency shifter:

$$N_L^P(t; \phi_t') = \exp\left(i \phi_t' t\right) \tag{11.61}$$

Spectral filter:

$$\tilde{S}^{A}(\omega;\omega_C) = \exp\left[\frac{-(\omega-\omega_C)^2}{2\gamma^2}\right] \tag{11.62}$$

Quadratic spectral phase modulator or dispersive delay line:

$$\tilde{S}^{P}_{Q}(\omega;\phi''_\omega) = \exp\left(\frac{i\phi''_\omega\omega^2}{2}\right) \tag{11.63}$$

Linear spectral phase modulator or delay line:

$$\tilde{S}^{P}_{L}(\omega;\phi'_\omega) = \exp(i\phi'_\omega\omega) \tag{11.64}$$

The filter of Eq. (11.59) is a time gate, or shutter, where τ is the time of maximum transmission and Γ^{-1} is the duration for which it is open. Both the quadratic and linear temporal phase modulators of Eqs. (11.60) and (11.61), denoted by the subscripts Q and L, respectively, impose new time-dependent phase on the input field. Temporal phase is imparted on the input pulses, e.g., by passing the pulses through an electrooptic phase modulator. Quadratic phase is imposed if the input pulses are passed through the phase modulator near one of the maxima of a sinusoidal driving signal, whereas linear phase results if the input pulses arrive near a zero crossing. Therefore ϕ''_t is proportional to the modulator's modulation depth and the square of the modulator's driving frequency while ϕ'_t is proportional to both the modulation depth and driving frequency. The spectral filter of Eq. (11.62) is simply represented by an idealized spectrometer transfer function, where ω_C is the center frequency and γ is the frequency passband. The quadratic and linear spectral phase modulators [Eqs. (11.63) and (11.64)] impose new frequency-dependent phase on the input field. Spectral phase is easily imparted by a delay line. Quadratic spectral phase is imposed if the delay line is dispersive with group-delay dispersion of ϕ''_ω. Linear spectral phase is the result of a nondispersive delay line of temporal delay ϕ'_ω.

Using this formulation, the output of such a general measurement apparatus is given by

$$D(\{p_k\}) = \int d\omega\, dt\, W(t,\omega) W_M(t,\omega;\{p_k\}) \tag{11.65}$$

where the set $\{p_k\}$ specifies the filter parameters. Therefore any apparatus measures a smoothed out, positive definite version of the Wigner function—the pulse Wigner function $W(t,\omega)$ integrated with an apparatus Wigner function $W_M(t,\omega;\{p_k\})$.

It is easy to show that if all the filters are time-stationary, then

$$W_M(t, \omega; \{p_k\}) = \prod_k |\tilde{S}_k(\omega; \{p_k\})|^2 \tag{11.66}$$

and for nonstationary filters

$$W_M(t, \omega; \{p_k\}) = \prod_k |N_k(t; \{p_k\})|^2 \tag{11.67}$$

In both cases, it is clear that D consists of an overlap of a marginal of the pulse Wigner function with the measurement Wigner function, and therefore it returns no information on the phase of the field. An apparatus consisting of only one class of filters will not work. What is surprising is that an apparatus consisting of at least one time-stationary and one time-nonstationary filter yields a signal from which the field *can* be reconstructed.

Because of the need to explore the entire region of the time-frequency phase space occupied by the pulse, a measurement scheme in which the smallest possible number of elements is present must therefore be an apparatus containing at least two filters of the classes described previously, each characterized by one parameter. From Fig. 11.6 it is clear that there are two general two-filter strategies. The first consists of two filters in sequence, say, in the upper arm of the interferometer, with the lower arm not used at all. This class of devices may be called phase-space methods, since it turns out that they make measurements directly on a phase-space representation of the test pulse. The second category may be labeled interferometric or in-parallel methods, since these devices use one filter in each of the upper and lower arms of the interferometer of Fig. 11.6.

11.3.3 Phase-Space Methods

The analysis of phase-space techniques is found in Ref. 47. Our discussion follows this framework. There are two subclasses of phase-space techniques—those that make simultaneous measurements of the complementary variables ω and t, recording thereby one of the phase-space distributions, and those that record marginals of the Wigner function, following a rotation in the phase space, leading to a set of spectral or temporal intensities parameterized by the rotation angle. The former method is known as spectrographic while the latter is referred to as tomographic. For each of these subclasses there are two possible filter orderings, resulting in a total of four types of phase-space measurement.

Taking into consideration the amplitude- and phase-only filter subclasses, there are a number of possible ways to arrange the filters to make up a minimalist scheme. But it is completely ineffective to allow

the last filter before a square-law detector to be a phase-only filter. The final filter must be an amplitude-only filter, but it may be either time-stationary or time-nonstationary. If it is the former, then the first filter must be a time-nonstationary filter, but it can be either an amplitude or a phase filter. If the last filter is time-nonstationary, the first must be a time-stationary filter, but again it may be of either the amplitude or the phase variety. Thus there are only four possible configurations of two filters that can be used to measure the complete Wigner function of the input pulse.

11.3.3.1 Spectrographic Techniques

The two spectrographic techniques make use of two sequential amplitude-only filters, one time-stationary (spectral filter) and one time-nonstationary (time gate) followed by a square-law detector, as shown in Fig. 11.7. The recorded signal is either a measure of the spectrum of a series of time slices (type I, Fig. 11.7*a*) or a measure of the

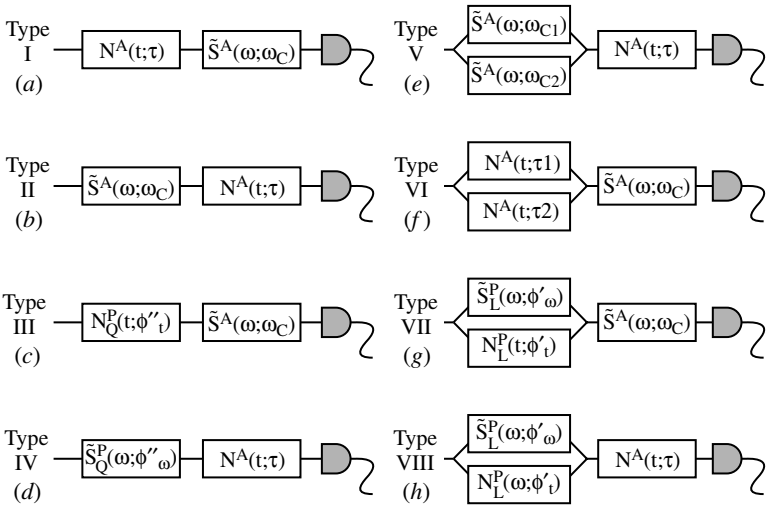


FIGURE 11.7 Linear filter description of type I to type VIII devices. Spectrographic devices, based on two serial amplitude filters in conjugate variables, correspond to (*a*) type I and (*b*) type II. Tomographic devices, based on a quadratic-phase modulation followed by an amplitude filter in the conjugate variable, correspond to (*c*) type III and (*d*) type IV. Interferometric techniques related to Young's double-slit experiment, with two amplitude filters in parallel followed by one amplitude filter in the conjugate variable, that correspond to (*e*) type V and (*f*) type VI. Interferometric techniques related to shearing interferometry, with two linear phase modulations in conjugate domains in parallel, correspond to (*g*) type VII and (*h*) type VIII.

time of arrival of a series of spectral slices (type II, Fig. 11.7*b*) depending on the ordering of the filters. There is no difference in principle between the two possible filter orderings, and thus this type of apparatus should be thought of as one that makes simultaneous measurements of the conjugate variables rather than sequential measurements. Fourier's principle precludes precise simultaneous measurements of the conjugate variables. Some of the earliest developments in the representation and measurement of short optical pulses are based on the concepts of spectrography in the time-frequency space.[48]

The Wigner function of the measurement apparatus for the type I device, e.g., is

$$W_M(t, \omega; \{\omega_C, \tau\}) = \int d\omega' |\tilde{S}^A(\omega'; \omega_C)|^2 \int dt' N^A\left(t + \frac{t'}{2}; \tau\right) N^{A^*} \times \left(t - \frac{t'}{2}; \tau\right) \exp[i(\omega' - \omega)t'] \tag{11.68}$$

In fact, for near-transform-limited input pulses, the apparatus Wigner function has nearly the same area as the pulse Wigner function itself. In principle the Wigner function can be retrieved from the data by deconvolution, but because of severe signal-to-noise requirements this approach is impractical. Thus, spectrographic phase-space pulse characterization techniques supply only qualitative insight into pulse train statistics. However, in the limit of narrowband filtering, that is, $|\tilde{S}^A(\omega'; \omega_C)|^2 \to \delta(\omega' - \omega_C)$, and if the pulses in the ensemble are assumed to be identical, the experimental trace is a simple convolution of Wigner functions, which can be expressed as a function of the electric field of the pulse

$$D(\omega_C, \tau) = \int d\omega\, dt\, W(t, \omega) W_M(t - \tau, \omega - \omega_C) = \left| \int dt\, E_{\text{in}}(t) N^A(t - \tau) \exp(i\omega_C t) \right|^2 \tag{11.69}$$

This set of conditions gives an apparatus function that occupies the minimum possible area of phase space, and therefore minimally "smoothes" the signal Wigner function. In this case, the experimental trace is the Gabor spectrogram with a window N^A.

Type I devices are popular in ultrafast optics. The time-nonstationary filter required for these devices can be implemented using nonlinear interactions or temporal modulators, and the high-resolution spectral measurements can be performed by an optical spectrum analyzer (OSA). As an example of the use of a nonlinear interaction to implement a type I device, let's consider sum-harmonic generation frequency resolved optical gating (SHG-FROG), as schematized in

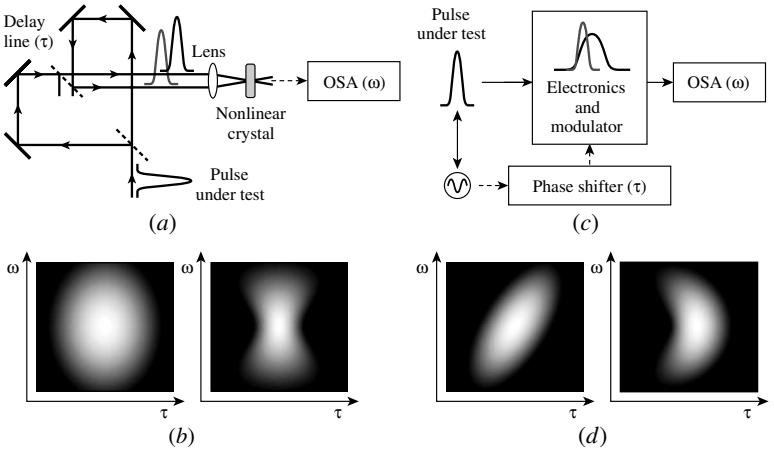


FIGURE 11.8 Schematics of spectrographic techniques and representative spectrograms. (*a*) In SHG-FROG, the pulse under test is replicated, and the spectrum of the field obtained by nonlinear mixing of the two replicas is measured by an optical spectrum analyzer (OSA) as a function of the optical frequency and delay between the two replicas modified by translation of a pair of mirrors. (*b*) The experimental trace of a pulse with second-order (left) and third-order (right) spectral phase does not give an intuitive representation of the group delay. (*c*) In linear spectrography, the pulse under test is modulated by a modulator driven by an electric drive signal. The spectrum of the modulated pulse is measured by an OSA as a function of the optical frequency and delay between the pulse under test and the modulation, which is controlled in the electrical domain. (*d*) The experimental trace of a pulse with second-order (left side) and third-order (right side) spectral phase gives an intuitive representation of the group delay.

Fig. 11.8*a*.[49,50] The pulse under test is sent into a symmetric Michelson interferometer that generates two replicas of the pulse with a variable relative delay τ controlled by translation of one pair of mirrors. Interaction in a nonlinear crystal provides a gating function; i.e., the electric field of the up-converted pulse is essentially proportional to the product of the fields of the two interacting pulses $E(t)E(t-\tau)$. The high-resolution optical spectrum analyzer measures the optical spectrum of the up-converted pulse, which leads to the experimental trace

$$S(\omega, \tau) = \left| \int dt\, E(t) E(t-\tau) \exp(i\omega t) \right|^2 \tag{11.70}$$

This identifies the unknown electric field of the pulse under test E as the function N^A. Figure 11.8*b* displays the SHG-FROG spectrogram

of a pulse with second- and third-order spectral phase. Although the corresponding electric fields and their Wigner functions are asymmetric in time, the corresponding nonlinear spectrograms are not, which implies that information on the direction of time cannot be recovered directly from the measured spectrogram. Versions of frequency resolved gating based on other nonlinearities, e.g., third-order nonlinearities, are characterized by gates N^A corresponding to higher-order products of the unknown electric field E and do not have this ambiguity.[50] Nonlinear interactions with a known ancillary pulse can also be used, in which case the experimental trace becomes the spectrogram of the electric field of the pulse under test, measured using the electric field of the known pulse as the gate.[51–53] In these cases, more intuitive nonlinear spectrograms are usually obtained, allowing, e.g., the visualization of chirp and the interpretation of linear and nonlinear propagation effects. The previous examples show that the formalism of type I devices is not limited to amplitude only filters, as the electric field is in most cases a complex quantity. Scanning of a phase-only modulation and recording of the associated spectrogram have also been demonstrated.[54]

Figure 11.8*c* represents a schematic of a type I device based on linear optics.[55] A temporal modulator is a useful time-nonstationary device for pulse measurement if its transfer function changes significantly during the time scale of the test pulse. While this condition is difficult to meet for sub-100-fs pulses, the development of fast temporal modulators for optical telecommunications has made possible the characterization of short optical pulses in the range of hundreds of femtoseconds to hundreds of picoseconds with completely linear techniques. Lithium-niobate electrooptic modulators and electroabsorption modulators driven by sinusoidal drives at 10 GHz or higher frequencies provide suitable speeds and magnitudes of phase or amplitude modulation. Since nonlinear optics requires high optical intensities or large nonlinearities, linear techniques are advantageous in terms of sensitivity. The temporal modulator is driven by an electric signal synchronized to the pulse under test and has a gating function $N^A = g$ which does not depend on the pulse under test. The relative delay between the pulse and the gate is scanned in the rf domain using a phase shifter. The OSA after the modulator measures the spectrum of the modulated pulse as a function of the optical frequency, leading to

$$S(\omega, \tau) = \left| \int dt E(t) g(t - \tau) \exp(i\omega t) \right|^2 \tag{11.71}$$

As can be seen in Fig. 11.8*d*, spectrograms measured with a signal-independent gate give a better representation of the chirp present on

the pulse, although the properties of the spectrogram are identically impacted by the chirp characteristics of E and g.

Type II devices are also common in ultrafast optics. Implementation of a stationary amplitude filter can be done using filters such as a slit in a zero-dispersion line. A high-resolution time-nonstationary amplitude filter is more difficult to implement since the pulse under test is usually the shortest event available in the laboratory, and that filter should in theory be significantly shorter than that to ensure that the experimental trace is the equivalent of Eq. (11.69) in the frequency domain. Photodetection has been used for pulses in the picosecond range,[56,57] and nonlinear cross-correlation with the pulse under test has been used for shorter pulses.[58–60]

Reconstruction of the electric field of the pulse under test from the experimental trace of Eq. (11.69) can, in principle, be performed directly by deconvolution of the Wigner function of the pulse using the Wigner function of the gate. However, this requires a good knowledge of the function N^A (or equivalently, its Wigner function), which is somewhat available for linear techniques but not available for nonlinear techniques with unknown gate pulses. Estimates of the group delay or instantaneous frequency in the pulse can also be obtained in some cases, using the properties of the spectrogram.[30] For a type I device with a narrow function N^A, the weighted average time as a function of frequency obtained using the spectrogram as the weight function is the group delay of the unknown pulse as a function of frequency. The practical use of this property, which is valid when the gating function is significantly shorter than the pulse under test, is hindered by the fact that the precision on the determination of the group delay can be poor since the width of the spectrogram increases dramatically in these conditions. The most practical approach to signal reconstruction for type I and type II devices is based on iterative phase retrieval. Electric field reconstruction from Eq. (11.69) is equivalent to phase reconstruction of the two-dimensional quantity $\int dt\, E_{\text{in}} \times (t) N^A(t-\tau) \exp(i\omega_C t)$ from its measured modulus $|\int dt\, E_{\text{in}}(t) N^A \times (t-\tau) \exp(i\omega_C t)|$. Projections between ensembles of electric fields matching different constraints allow, in most cases, convergence to one possible solution of Eq. (11.69), whether the function N^A is known, unknown, or a function of the unknown electric field itself. The principal component generalized projections algorithm can be used to invert experimental trace obtained with type I and type II devices.[60,61]

11.3.3.2 Tomographic Techniques

As with spectrographic methods, the so-called tomographic techniques require in-series, time-stationary, and time-nonstationary filters so that the entire phase space can be explored. However, unlike spectrographic techniques, the first filter in a tomographic apparatus is

a phase-only filter, as illustrated in Fig. 11.7 [either a quadratic temporal phase modulator (type III, Fig. 11.7*c*) or a quadratic spectral phase modulator (type IV, Fig. 7*d*)]. The inclusion of a quadratic phase-only filter results in a distinctly different interpretation of the measurement, leading to a fundamentally different inversion algorithm. To see this, notice that a phase-only filter does not provide any information on the frequency or the arrival time of a pulse ensemble and hence does not constitute a measurement of either frequency or time. Therefore, a tomographic apparatus does not make a simultaneous measurement of these incompatible variables. Rather, the quadratic-phase modulation acts to rotate the phase space. The square-law detector in combination with the amplitude-only filter records the resulting intensity distribution.

The measurement function takes the form

$$W_M(t, \omega; \{\omega_C, \phi_t''\}) = \int d\omega' \left|\tilde{S}^A(\omega'; \omega_C)\right|^2 \int dt' N_Q^P \left(t + \frac{t'}{2}; \phi_t''\right) N_Q^{P*} \times \left(t - \frac{t'}{2}; \phi_t''\right) \exp[i(\omega' - \omega)t'] \tag{11.72}$$

When the phase-nonstationary filter takes the form of Eq. (11.60) and the amplitude stationary filter that of Eq. (11.62), this reduces to the form

$$W_M(t, \omega; \{\omega_C, \phi_t''\}) = \exp\left[-\frac{(\omega - \omega_C - \phi_t'' t)^2}{\gamma^2}\right] \tag{11.73}$$

This function is different from the filter function of the Gabor transform. Its location in phase space is not determined by the filter parameters—rather its orientation is. A change in ω_C translates the entire function along the frequency axis, and a change in ϕ_t'' alters the orientation of the function about $\omega = \omega_C$. The detected signal is, in the limit as $\gamma \to 0$,

$$D(\omega_C, \phi_t'') = \int dt\, W(t, \omega_C + \phi_t'' t) \tag{11.74}$$

This is easily interpreted by transforming the variables to the form

$$D(\omega_\theta, \theta) = \int dt_\theta W(\omega_\theta \sin\theta + t_\theta \cos\theta, \omega_\theta \cos\theta - t_\theta \sin\theta) \tag{11.75}$$

where we have defined $\tan\theta = -\phi_t''$ and $\omega_\theta = \omega_C \cos\theta$ and scaled the measured trace by $\cos(\theta)$. The signal $D(\omega_\theta, \theta)$ is, therefore, a set of distributions that are marginals of a rotated version of the pulse Wigner function. This is the essence of tomographic measurements; indeed, the above formula may be inverted to give the pulse Wigner

distribution using the inverse Radon transform.[62] The quadratic temporal phase modulator of a type III device rotates the input pulses so as to map time into frequency, and a spectrometer is used to resolve the spectrum of the output pulses. Similarly, the quadratic spectral phase modulator of a type IV device rotates the input pulses so as to map frequency into time, and a time gate is used to resolve the time-dependent intensity of the output pulses. A combination of quadratic temporal and spectral phase modulations in series can also be used to rotate the Wigner function, which avoids the requirement for large amplitude of a single quadratic modulation.[63] The Wigner function of an optical source can, in principle, be completely reconstructed from a set of its projections determined by using a type III or type IV device. If the optical source is a train of identical optical pulses, the electric field can be determined from the Wigner function reconstructed using a large number of its projections.

Simplified versions of chronocyclic tomography have been experimentally demonstrated. The first one rotates the Wigner function by 90° and measures its frequency marginal using a spectrometer, which is a scaled representation of the temporal intensity of the pulse under test. This is known as the time-to-frequency converter because it is effectively a procedure to map the temporal intensity of the input pulse onto the spectral intensity of another signal.[39] The second one rotates the Wigner function by a small angle,[64,65] so that Eq. (11.75) becomes

$$D(\omega_C, \delta\theta) = \int dt_\theta W(t_\theta + \omega_C \delta\theta, \omega_C - t_\theta \delta\theta) \tag{11.76}$$

Development of this equation leads to the result

$$\frac{\partial D}{\partial \theta}(\omega_C, 0) = -\frac{\partial}{\partial \omega_C}\left[\tilde{I}(\omega_C)\phi'_\omega(\omega_C)\right] \tag{11.77}$$

Equation (11.77) indicates that the spectrally resolved changes of the optical spectrum of the pulse after small amounts of quadratic temporal phase modulation are algebraically linked to the optical spectrum and spectral phase of the pulse. Since $\tilde{I}(\omega_C)$ can be directly measured with a spectrometer [e.g., the spectrometer used to measure the signal of Eq. (11.76) when the modulation is turned off], the spectral phase ϕ_ω can be obtained by solving the second-order differential equation, Eq. (11.77). Figure 11.9 is a schematic of an implementation of simplified chronocyclic tomography. The left-hand side of Eq. (11.77) is obtained as a finite difference of the optical spectra obtained after two small quadratic temporal phase modulations of opposite signs. Quadratic modulation is obtained by synchronization with the maximum and minimum of a sinusoidal phase modulation obtained in a lithium niobate electrooptic phase modulator. In this symmetric configuration, the optical spectrum is simply obtained as the average of

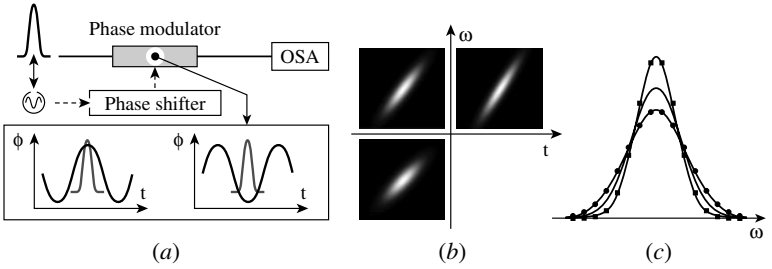


FIGURE 11.9 (*a*) Schematics of simplified chronocyclic tomography based on a temporal phase modulator. Quadratic temporal phase modulations of opposite signs are provided by a sinusoidally driven phase modulator synchronized such that the pulse under test coincides with the minimum or maximum of the phase modulation. (*b*) The Wigner function of a chirped pulse is shown in the absence of modulation (upper left plot) and with quadratic temporal modulations of opposite signs (upper right and lower left plots). The effect of the shear along the frequency axis can be seen. (*c*) The initial spectrum of the pulse (continuous line) and the spectra after the quadratic temporal phase modulations of opposite signs (continuous lines with round and square markers) are seen, from which the electric field of the test pulse can be reconstructed.

the two measured optical spectra. Figure 11.9*b* represents the Wigner function of a chirped pulse in the absence of modulation, and after quadratic temporal phase modulation of opposite signs. The shear effect due to the modulation can be observed, and it results in a change of the frequency marginal. The spectral phase can directly be reconstructed from the corresponding optical spectra represented in Fig. 11.9*c*. The combination of a quadratic spectral phase modulation (i.e., chromatic dispersion) and temporal gating by photodetection has also been demonstrated based on the same formalism.[66] These one-dimensional determinations of the electric field are closely related to phase retrieval using the transport-of-intensity equation in the spatial domain.[67]

11.3.4 Interferometric or Direct Techniques

Direct techniques, as the name indicates, reconstruct the correlation function in either the time domain or frequency domain directly (i.e., noniteratively) from the recorded intensity distributions. In such schemes each pulse in the ensemble is split into two replicas at a beam splitter, and each replica is independently filtered before being recombined. The interference of the field from the parallel pathways introduces structure on the output intensity distribution, which then carries information about both the amplitude and the phase of the correlation function of the input field. Direct techniques of this

kind are known as self-referencing. It is also possible to reconstruct the electric field of an unknown pulse from the interferogram resulting from the overlap of the unknown pulse with a well-characterized reference.[68–70] Of course, this approach requires one to first characterize a reference, which begs the question. We thus confine this discussion to self-referencing techniques.

One significant advantage of direct techniques compared to phase-space techniques is that the entire space over which the phase-space or correlation functions are defined need not be explored if the pulse train is assumed to consist of identical pulses. Only a single section of one quadrature of the (complex) correlation function is required to obtain the electric field amplitude and phase, and this is precisely what is recorded by direct techniques. Thus, while phase-space techniques must explore the entire chronocyclic space even when the electric field, rather than the correlation function, is the fundamental quantity of interest, direct techniques need only return a single slice of the correlation function in order to construct the simpler quantity. Roughly speaking, if one wishes to reconstruct the complex electric field at N temporal points, at least $2N$ independent real data points are required. While direct techniques are capable of reconstructing the field by recording only the necessary $2N$ points, phase-space techniques essentially require the measurement of N^2 points. The excess data of the phase-space methods can be advantageous as a means of refining the estimate of the pulse shape. Of course, the overcomplete data set is available from direct measurement of the entire correlation function as well, or from any extended sampling of it.

11.3.4.1 Two-Pulse Double-Slit Interferometry

This class consists of an in-parallel pair of amplitude-only filters, followed by an additional amplitude-only filter, as shown in Fig. 11.7. The in-parallel amplitude-only filters select either two frequency slices or two time slices of the pulse which beat together at the output of the interferometer to provide information for a single point of the respective correlation function. Thus these two types of direct techniques are the time-domain analog of Young's double-slit interferometer. In a type V apparatus, the in-parallel pair of amplitude-only filters is time-stationary (spectral filters), and the final filter is a time-nonstationary amplitude-only filter (time gate) (see Fig. 11.7*e*). Each pulse in the ensemble under consideration is split into identical replicas at a beam splitter, and a single frequency from each replica is selected by the spectral filters. The center frequencies of the spectral filters ω_{C1} and ω_{C2} are independently controllable, and usually the two spectral filters have the same passband γ. The selected frequency components are recombined, and the resulting temporal interferogram is subsequently resolved by a time gate. The signal recorded by the square-law

detector is a function of the center frequencies and the time of maximum transmission τ of the time gate

$$D(\{\omega_{C1}, \omega_{C2}, \tau\}) = \left\langle \int dt \left| N^A(t-\tau) \int d\omega \left[\tilde{S}^A(\omega - \omega_{C1}) + \tilde{S}^A(\omega - \omega_{C2}) \right] \tilde{E}(\omega) \exp(-i\omega t) \right|^2 \right\rangle \quad (11.78)$$

The detected signal takes on a particularly useful form when certain assumptions regarding the filters are valid. The first assumption is that the passband of the spectral filters is much narrower than the spectrum of the input pulses, so that the spectral filter transfer functions become

$$\tilde{S}^A(\omega - \omega_C) \rightarrow \delta(\omega - \omega_C) \quad (11.79)$$

The second assumption is that the duration over which the time gate is open is much shorter than the temporal period of the beat note to be measured, so that the time-gate response function becomes

$$N^A(t-\tau) \rightarrow \delta(t-\tau) \quad (11.80)$$

The integration time of the square-law detector must be long enough that an average over a sufficiently large number of pulses is obtained. Changing the frequency variables to the center- and difference-frequency coordinates, the detected signal of Eq. (11.78) simplifies to

$$D\left(\left\{\omega - \frac{\Delta\omega}{2}, \omega + \frac{\Delta\omega}{2}, \tau\right\}\right) = \tilde{I}\left(\omega - \frac{\Delta\omega}{2}\right) + \tilde{I}\left(\omega + \frac{\Delta\omega}{2}\right) + 2|\tilde{\tilde{C}}(\Delta\omega, \omega)| \cos\{\arg[\tilde{\tilde{C}}(\Delta\omega, \omega)] + \Delta\omega\tau\} \quad (11.81)$$

where $\omega = (\omega_{C1} + \omega_{C2})/2$ and $\Delta\omega = \omega_{C1} - \omega_{C2}$.

The detected signal is an interferogram measuring sections of the two-frequency correlation function of the pulse train. The inversion procedure for reconstructing the correlation function from type V measurements is apparent from the form of Eq. (11.81) and is illustrated in Fig. 11.10. The visibility of the fringes, occurring with temporal period $2\pi/\Delta\omega$, provides a measure of the magnitude of $\tilde{\tilde{C}}(\Delta\omega, \omega)$. The location of the fringes along the delay axis τ provides a relative measure of the phase of $\tilde{\tilde{C}}(\Delta\omega, \omega)$. Each temporal beat note supplies enough information to reconstruct a single point of the two-frequency correlation function. Therefore, the temporal fringe visibility and relative fringe position need to be recorded for every pair of frequencies contained within the pulse spectrum if one wishes to reconstruct the entire two-frequency correlation function. This procedure is experimentally intensive and demands the recording of a prodigious amount of data.

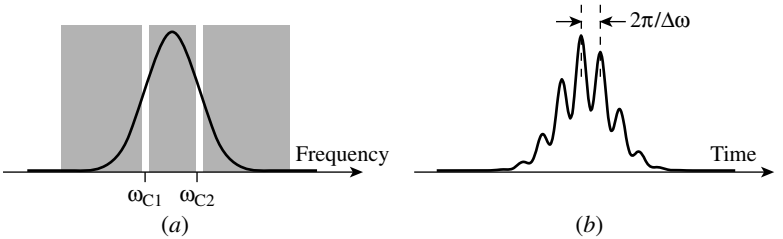


FIGURE 11.10 (*a*) Schematics of a double-slit experiment in the spectral domain and (*b*) resulting experimental trace for one particular setting of the central frequencies of the two filters. The gray regions in part *a* depict the filtering of two narrowband spectral slices at the center frequencies ω_{C1} and ω_{C2}.

In many implementations, especially for femtosecond duration pulses, the time gate consists of a nonlinear optical mixing process, such as up-conversion with a portion of the pulse being characterized, which sets the temporal resolution Γ^{-1} to be close to the duration of the input pulses. Consequently, the narrow time-gate assumption of Eq. (11.80) is not valid for frequency separations $\Delta\omega$ greater than a small fraction of the pulse bandwidth, since the temporal beat note is too fast to resolve.

The narrow time-gate approximation does hold for small frequency separations so that slices of the two-frequency correlation function near $\Delta\omega = 0$ can be recorded. If the pulses in the train are assumed to be identical, a sampling of one such slice is sufficient for reconstructing the pulse electric field. When coherence is assumed, the phase of the two-frequency correlation function is no more than the phase difference between the selected spectral components. Coupled with knowledge of the pulse spectrum, the spectral phase differences for a set of frequencies separated by $\Delta\omega$ provide ample information for reconstructing the pulse electric field. This is precisely the approach of direct optical spectral phase measurement (DOSPM).[71] DOSPM uses an apparatus in which a pair of adjustable slits is placed in the Fourier transform plane of a zero-dispersion line. This spectral filter with dual passbands of adjustable center frequencies is equivalent to a pair of in-parallel single-frequency spectral filters. The beating with pairs of optical frequencies was recorded by nonlinear interaction with the pulse under test. This work was extended to the measurement of the spectral phase difference between a reference optical frequency and a set of other frequencies in the pulse, these frequencies being filtered by a mask with multiple slits placed at the Fourier plane of a zero-dispersion line.[72] A version of the DOSPM that does not require the

isolation of discrete optical frequencies in the pulse under test can be implemented, provided that a large amount of chirp is added to the pulse.[73] The phase difference between two adjacent optical frequencies is measured in the time domain by interfering two chirped replicas of the pulse under test, and the time-frequency dependence in the chirped pulses naturally links the temporal axis of the measured intensity to the optical frequency in the pulse under test. This method can be combined with a Fourier processing algorithm identical to that of spectral shearing interferometry.[74] Another approach to measuring the spectral phase difference between different optical frequencies relies on a fast electronic detector. The beat note between pairs of adjacent frequencies, instead of the beat note between frequencies, was measured to reduce the bandwidth requirement of the detector.[75] The overlap of two spectrally dispersed and spatially sheared replicas of the pulse under test can also be used so that quantity of Eq. (11.81) can be measured at different spatial locations.[76] Finally, let's note that for a periodic source of period T, it suffices to measure the phase difference between spectral modes separated by $2\pi/T$ with a detector of sufficient resolution. This is of particular interest for optical sources used in optical telecommunications.[77]

A similar interferometric approach has been employed to reconstruct the electric field amplitude and phase by Rothenberg and Grischkowsky.[78] In their technique, generically referred to as time-domain interferometry, a spectral filter is placed in only one arm of the interferometer. The monochromatic frequency component resulting from the spectrally filtered path provides an effective reference with which to compare the pulse that passes through the unfiltered arm of the interferometer. Constraints on the available temporal resolution limit this method to the measurement of stretched pulses of relatively long duration.

A complementary approach to the temporally resolved two-pulse interferometry is spectrally resolved two-pulse interference. This approach consists of two in-parallel time-nonstationary amplitude-only filters (time gates), followed by a time-stationary amplitude-only filter (spectral filter). The two replicas of the pulse are independently filtered by time gates with variable times of maximum transmission τ_1 and τ_2, before being recombined. The spectral beats, resulting from the overlap of the two time slices, are resolved by a spectrometer. The spectrum for each pair of time settings of the time gates is recorded. The visibility of the spectral fringes, occurring at the spectral period $2\pi/\Delta\tau = 2\pi/(\tau_1 - \tau_2)$, is a measure of the magnitude of the two-time correlation function at these two times. The position of the fringes along the frequency axis is a relative measure of the phase. Thus, each recorded spectrum returns one point of the two-time correlation function so that a simple point-by-point reconstruction algorithm is

possible. In practice, for femtosecond-duration pulses it is difficult to satisfy the narrow time-gate approximation, so that iterative reconstruction algorithms are needed to deconvolve the response functions of these filters. Given the current state of technology, this class of direct devices is not practical for pulses of duration less than several tens of picoseconds.

11.3.4.2 Shearing Interferometry

Shearing interferometers consist of a time-nonstationary linear phase filter and a time-stationary linear phase filter in parallel, followed by an amplitude-only filter. The action of linear phase filters is to shift the electric field in either time or frequency. For instance, consider the spectral linear phase modulator of Eq. (11.64). The action of this filter is a translation of the pulse in time, which can easily be obtained with a nondispersive delay line. Likewise, imparting a temporal linear phase on the input field is equivalent to a translation, or shift, of the frequency axis. The resulting interferogram contains information about an entire section of the correlation function, as opposed to sampling a single point of the function, as is the case with the two-slit types.

In spectral shearing interferometry, the amplitude-only filter following the in-parallel linear phase filter arrangement is a spectral filter (see Fig. 11.7*g*). Since the spectral filter is a time-stationary device, the key filter is the time-nonstationary linear temporal phase modulator that provides a shift, or shear, of the spectrum of one replica of the input pulse.[79,80] The detected signal is a function of the linear temporal phase modulator parameter ϕ_t' as well as the center frequency of the spectrometer ω_C,

$$D(\{\phi_t', \omega_C; \phi_\omega'\}) = \left\langle \int d\omega \left| \tilde{S}^A(\omega - \omega_C) \left[\int d\omega' \tilde{N}_L^P(\omega - \omega', \phi_t') \tilde{E}(\omega') + \tilde{S}_L^P(\omega, \phi_\omega') \tilde{E}(\omega) \right] \right|^2 \right\rangle \tag{11.82}$$

where the temporal linear phase filter's response function and the spectral linear phase filter's transfer function inherently depend on the variables ϕ_t' and ϕ_ω', respectively. Therefore, the detected signal is also a function of the amount of spectral phase modulation ϕ_ω', although this dependence plays a secondary role which will be described below. It is easy to see from Eq. (11.61) that the transfer function of the temporal linear phase modulator is

$$\tilde{N}_l^P(\omega', \phi_t') = \delta(\omega' - \phi_t') \tag{11.83}$$

Again the spectral filter is taken to have a passband much narrower than the spectrum of the input pulses. Upon substitution of the

appropriate forms for the filter functions, recognizing that ϕ'_t plays the role of a spectral shear $\Delta\omega$ and changing variables to the center- and difference-frequency coordinates, the recorded distribution simplifies to

$$D\left(\left\{\Delta\omega, \omega + \frac{\Delta\omega}{2}; \phi'_\omega\right\}\right) = \tilde{I}\left(\omega - \frac{\Delta\omega}{2}\right) + \tilde{I}\left(\omega + \frac{\Delta\omega}{2}\right) + 2|\tilde{\tilde{C}}(\Delta\omega, \omega)| \cos\{\arg[\tilde{\tilde{C}}(\Delta\omega, \omega)] + \phi'_\omega\omega\} \tag{11.84}$$

The measured signal $D(\omega) = D(\{\Delta\omega, \omega + \Delta\omega/2; \phi'_\omega\})$ for a given shear $\Delta\omega$ is related simply to a section of the two-frequency correlation function. This section may be extracted using a simple and direct inversion algorithm that separates the interference term [the third term in Eq. (11.84)] from the noninterferometric terms.[68,81] This is easily accomplished by means of Fourier transforms, in a manner described in Fig. 11.11. The spectral interferogram is first Fourier transformed to separate the noninterferometric terms (located around $t = 0$) from the interferometric terms (located around $t = \pm\phi'_\omega$). One of the interferometric terms is then filtered out and Fourier transformed back to the frequency domain, where the amplitude and phase of the correlation function are obtained. The key point of spectral shearing interferometry is that the spectral phase of the test pulse, $\arg[\tilde{\tilde{C}}(\Delta\omega, \omega)]$, is encoded on the *spacing* of the fringes in the interference term.

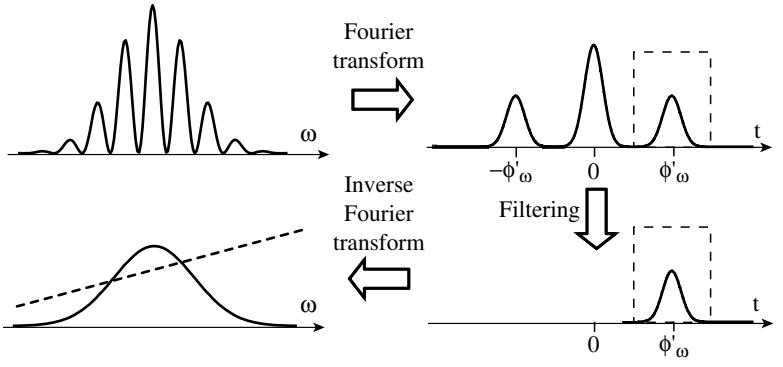


FIGURE 11.11 Extraction of a slice of the two-frequency correlation function in spectral shearing interferometry. The spectral interferogram is Fourier transformed to separate the interferometric components from the noninterferometric terms. The slice of the two-frequency correlation function is obtained by Fourier transforming back to the spectral domain one of the sidebands.

An entirely analogous argument may be made for temporal shearing interferometers. In this case, the delay line in one arm of the interferometer causes the pulses on recombining at the second beam splitter to exhibit temporal beats in their intensity that may be resolved by a fast time gate. This latter element is the amplitude-only filter that replaces the spectrometer required in the spectral shearing interferometer. In this arrangement, a temporal linear phase modulator may be used to provide a "temporal carrier" for the two-time correlation function in the interference term. This is accomplished by frequency-shifting one of the pulses with respect to the other by a shear ϕ'_t. The detected signal is then

$$D\left(\left\{t-\frac{\Delta t}{2},\Delta t;\phi'_t\right\}\right)=I\left(t-\frac{\Delta t}{2}\right)+I\left(t+\frac{\Delta t}{2}\right)+2|C(t,\Delta t)|\cos\{\arg[C(t,\Delta t)]+\phi'_t t\} \tag{11.85}$$

assuming that the time gate is of infinitesimal duration. A similar algorithm, as described for the spectral shearing interferogram, may be used to extract the temporal phase of the test pulse in this case. In practice, however, it is very difficult to provide a short enough time gate to enable this method to work. Nonlinear optical interactions that cross-correlate the interferogram with the test pulse will not provide enough temporal resolution to resolve the fringes. Therefore this method is restricted to pulses whose duration is long enough that an externally controlled time gate, such as a temporal modulator or a nonlinear interaction with a short optical pulse, can be used.

As with spectrography, practical implementations of spectral shearing interferometry have been demonstrated with nonlinear optics and with entirely linear setups. In electrooptic spectral shearing interferometry (EOSI), the spectral shear is obtained by linear temporal phase modulation, e.g., with lithium-niobate electrooptic phase modulators.[82,83] A symmetric implementation based on such a modulator is shown in Fig. 11.12*a*. The pulse under test is sent into an interferometer that generates two replicas separated by a delay τ. One of the outputs of the interferometer is sent to a phase modulator driven by a sinusoidal high-frequency modulation. The modulation has a period 2τ and is synchronized so that the two replicas are located at the two zero crossings of the modulation. In this configuration, the replicas are sheared by the same amount in opposite directions. The interferogram measured by an optical spectrum analyzer is

$$S(\omega)=\tilde{I}(\omega+\Omega)+\tilde{I}(\omega-\Omega)+2\sqrt{\tilde{I}(\omega+\Omega)\tilde{I}(\omega-\Omega)}\times\cos[\phi_\omega(\omega+\Omega)-\phi_\omega(\omega-\Omega)+\omega\tau] \tag{11.86}$$

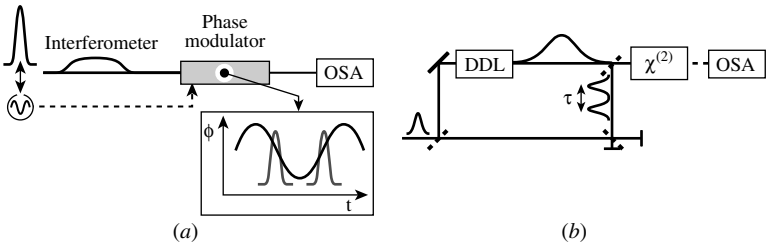


FIGURE 11.12 Schematics of spectral shearing interferometry based on (*a*) a linear temporal phase modulator and (*b*) a nonlinear interaction with a chirped pulse. In part *a*, two replicas of the pulse under test are spectrally sheared in opposite direction by a temporal phase modulator driven by a sinusoidal drive. In part *b*, a chirped pulse is generated by a dispersive delay line (DLL), and nonlinear interaction of two relatively delayed replicas of the test pulse with the chirped pulse induces a relative shear between the two resulting pulses.

where Ω is the shear induced on one of the replicas, which can be as high as several hundreds of gigahertz. Since the relative shear 2Ω must be of the order of a few percents of the bandwidth of the pulse under test to provide adequate sampling in the frequency domain and maintain a measurable finite difference $\phi_\omega(\omega + \Omega) - \phi_\omega(\omega - \Omega)$, EOSI most likely can be implemented for pulses with duration down to 50 fs. Advantageously, the signal intensity does not depend significantly on the amount of chirp present on the pulse, and the technique is accurate as long as one can maintain a linear phase ramp over the temporal support of the pulse under test. Other related techniques measuring the phase difference between the spectral modes of a periodic source have been demonstrated.[84–86]

The nonlinear implementation of spectral shearing interferometry, also known as spectral phase interferometry for direct electric-field reconstruction (SPIDER), can provide large shears by nonlinear interaction of two replicas of the pulse under test with a highly chirped pulse.[87,88] For such a pulse, the instantaneous frequency is a linear function of time and the frequency content in a short time interval is quasi-monochromatic. As seen in Fig. 11.12*b*, part of the input pulse is used to generate a chirped pulse, and two delayed replicas of the input pulse are generated by an interferometer. Nonlinear interaction of one replica of the pulse under test with the chirped pulse leads to a spectral shift given by the instantaneous optical frequency of the chirped pulse during the interaction. The other replica, delayed by τ, interacts with another optical frequency of the chirped pulse. The two converted replicas are relatively sheared by $\Omega = \tau/\phi_2$, where

ϕ_2 is the second-order dispersion of the chirped pulse. For pulses in the hundreds of femtoseconds regime and shorter, spectral shearing interferometry has proved to be a robust and effective method of determining the amplitude and phase of optical, infrared, UX, and indeed EUV pulses.

Electric field reconstruction with spectral shearing interferometry relies primarily on the steps described in Fig. 11.11. Once the spectral phase of the two-time correlation function $\phi_\omega(\omega+\Omega)-\phi_\omega(\omega-\Omega)$ has been extracted, it can be integrated into the spectral phase of the pulse under test $\phi_\omega(\omega)$. The reconstruction algorithm is therefore direct and algebraic, even when the time-nonstationary filters are synthesized using nonlinear optics. Spectral shearing interferometry can also be used without a delay between the two interfering pulses, in which case the phase of the interferometric component can be retrieved by scanning the relative phase of the interfering pulses.[89]

11.4 Conclusions

A phase-space representation of ultrashort optical pulses is useful for three reasons. First, it provides a useful tool for visualizing pulsed fields and enables an intuitive way to understand central concepts such as chirp, group delay, and instantaneous frequency. Second, it enables representation of pulse ensembles in terms of the lowest-order correlation function of the ensemble. Third, it provides a simple framework for understanding measurement methods that are based on square-law detectors, which are universal in optics. In this chapter, we set out the basic definitions required to define the chronocyclic phase space, and its space-time extension, as well as developed a catalog of modern measurement techniques in terms of phase space (or correlation space) distributions. This catalog can be shown to be complete and can be understood in terms of manipulation and sectioning of the phase-space distributions by means of linear filters and photodetectors. This clarifies an important misconception about ultrafast measurements—that they require a nonlinear response somewhere in the apparatus. In fact, a linear filter with a nonstationary response is sufficient. Indeed the minimum necessary conditions for obtaining a signal that encodes sufficient information to invert the electric field of the pulse must contain at least one time-stationary and one time-nonstationary filter. Further, we have shown how all the currently most popular methods can be incorporated into this framework. Because it is necessary to synthesize a nonstationary filter by means of nonlinear optics when one is working with ultrashort pulses (i.e., with durations below 100 fs), inverting the data from the measured signal sometimes requires iterative algorithms, though in some cases it is

possible to find direct inversions that are robust. In both cases, there are well-established practical methods by which the ultrashort pulse may be completely characterized.

The work of Christophe Dorrer was partially supported by the U.S. DOE Office of Inertial Confinement Fusion under Cooperative Agreement No. DE-FC52-08NA28302, the University of Rochester, and the New York State Energy Research and Development Authority. The support of DOE does not constitute an endorsement by DOE of the views expressed in this chapter.

Ian Wamsley was supported by EPSRC (ER/S24015/01 and EP/D503248/1) and the Royal Society, through the Wolfson Research Merit Award scheme.

References

1. G. Steinmeyer, "A review of ultrafast optics and optoelectronics," *J. Opt. A: Pure Appl. Opt.* **5**: 1–15 (2003).
2. C. Spielmann, P. F. Curley, T. Brabec, and F. Krausz, "Ultrabroadband femtosecond lasers," *IEEE J. Quantum Electron.* **30**: 1100–1114 (1994).
3. S. Backus, C. G. Durfee, III, M. M. Murnane, and H. C. Kapteyn, "High power ultrafast lasers," *Rev. Sci. Instrum.* **69**: 1207–1223 (1998).
4. G. Cerullo and S. De Silvestri, "Ultrafast optical parametric amplifiers," *Rev. Sci. Instrum.* **74**: 1–18 (2003).
5. J. Ye and S. T. Cundiff, *Femtosecond Optical Frequency Comb Technology*, Springer, New York, 2005.
6. A. Scrinzi, M. Yu Ivanov, R. Kienberger, and D. M. Villeneuve, "Attosecond physics," *J. Phys. B: At. Mol. Opt. Phys.* **39**: 1–37 (2006).
7. L. Mandel, "Interpretation of instantaneous frequencies," *Am. J. Phys.* **42**: 840–846 (1974).
8. P. J. Loughlin and B. Tacer, "Comments on the interpretation of instantaneous frequency," *IEEE Signal Process. Lett.* **4**: 123–125 (1997).
9. J. W. Goodman, *Introduction to Fourier Optics*, McGraw-Hill, New York, 1996.
10. H. R. Telle, G. Steinmeyer, A. E. Dunlop, J. Stenger, D. H. Sutter, and U. Keller, "Carrier-envelope offset phase control: A novel concept for absolute optical frequency measurement and ultrashort pulse generation," *Appl. Phys. B* **69**: 327–332 (1999).
11. M. Born and E. Wolf, *Principles of Optics*, Pergamon, New York, 1980.
12. S. A. Ponomarenko, G. P. Agrawal, and E. Wolf, "Energy spectrum of a nonstationary ensemble of pulses," *Opt. Lett.* **29**: 394–396 (2004).
13. K. H. Brenner and J. Ojeda-Castañeda, "Ambiguity function and Wigner distribution function applied to partially coherent imagery," *Opt. Acta* **31**: 213–223 (1984).
14. K. H. Brenner and K. Wodkiewicz, "The time-dependent physical spectrum of light and the Wigner distribution function," *Opt. Comm.* **43**: 103–106 (1982).
15. C. A. Hirlimann and J.-F. Morhange, "Wavelet analysis of short light pulses," *Appl. Opt.* **31**: 3263 (1992).
16. J. Paye, "The chronocyclic representation of ultrashort light pulses," *IEEE J. Quantum Electron.* **28**: 2262–2273 (1992).
17. S. Mukamel, C. Ciordas-Ciurdariu, and V. Khidekel, "Wigner spectrograms for femtosecond pulse-shaped heterodyne and autocorrelation measurements," *IEEE J. Quantum Electron.* **32**: 1278–1288 (1996).

18. V. J. Pinto-Robledo and T. A. Hall, "Chronocyclic description of laser pulse compression," *Opt. Comm.* **125**: 348 (1996).
19. J.-P. Likforman, M. Joffre, and V. Thierry Mieg, "Measurement of photon echoes by use of femtosecond Fourier-transform spectral interferometry," *Opt. Lett.* **22**: 1104–1106 (1997).
20. D. Dragoman and M. Dragoman, "Phase space characterization of solitons with the Wigner transform," *Opt. Comm.* **137**: 437–444 (1997).
21. J. Azaña and M. A. Muriel, "Reconstruction of fiber grating period profiles by use of Wigner–Ville distributions and spectrograms," *J. Opt. Soc. Am. A* **17**: 2496–2505 (2000).
22. J.-H. Kim, D. G. Lee, H. J. Shin, and C. H. Nam, "Wigner time-frequency distribution of high-order harmonics," *Phys. Rev. A* **63**: 063403 (2001).
23. M. B. Gaarde, "Time-frequency representations of high order harmonics," *Opt. Express* **8**: 529–536 (2001).
24. D. Dragoman and M. Dragoman, "Time-frequency signal processing of terahertz pulses," *Appl. Opt.* **43**: 3848–3853 (2004).
25. J. Azaña, "Time-frequency (Wigner) analysis of linear and nonlinear pulse propagation in optical fibers," *EURASIP J. Appl. Signal Process.* **10**: 1554–1565 (2005).
26. R. N. Graf and A. Wax, "Temporal coherence and time-frequency distributions in spectroscopic optical coherence tomography," *J. Opt. Soc. Am. A* **24**: 2186–2195 (2007).
27. J. Ojeda-Castañeda, J. Lancis, C. M. Gómez-Sarabia, V. Torres-Company, and P. Andrés, "Ambiguity function analysis of pulse train propagation: Applications to temporal Lau filtering," *J. Opt. Soc. Am. A* **24**: 2268–2273 (2007).
28. S. Fechner, F. Dimler, T. Brixner, G. Gerber, and D. J. Tannor, "The von Neumann picture: A new representation for ultrashort laser pulses," *Opt. Express* **15**: 15387–15401 (2007).
29. A. Rodenberg, S. Fechner, F. Dimler, and D. J. Tannor, "Experimental implementation of ultrashort laser pulses in the von Neumann picture," *Appl. Phys. B* **93**: 763–772 (2008).
30. L. Cohen, *Time-Frequency Analysis*, Prentice-Hall, Englewood Cliffs, N.J., 1995.
31. R. Gase, "Time-dependent spectrum of linear optical systems," *J. Opt. Soc. Am. B* **8**: 850–859 (1991).
32. U. Leonhardt, *Measuring the Quantum State of Light*, Cambridge University Press, Cambridge, 1997.
33. M. E. Casida, J. E. Harriman, and J. L. Anchell, "The Husimi function for electron distributions," *Int. J. Quantum Chem.* **32**: 435–456 (2004).
34. H. M. Ozatkas, Z. Zalevsky, and M. A. Kutay, *The Fractional Fourier Transform with Applications in Optics and Signal Processing*, Wiley, Chichester, 2001.
35. D. Dragoman, "Applications of the Wigner distribution function in signal processing," *EURASIP J. Appl. Signal Process.* **2005**: 1520–1534 (2005).
36. P. Tournois, J.-L. Vernet, and G. Bienvenu, "Sur l'analogie optique de certains montages électroniques: formation d'images temporelles de signaux électriques," *C. R. Acad. Sci. Paris* **267**: 375–378 (1968).
37. B. H. Kolner, "Space-time duality and the theory of temporal imaging," *IEEE J. Quantum Electron.* **30**: 1951–1963 (1994).
38. A. V. Gitin, "Applications of the Wigner function and matrix optics to describe variations in the shape of ultrashort laser pulses propagating through linear optical systems," *IEEE J. Quantum Electron.* **36**: 376–382 (2006).
39. M. T. Kaufman, W. C. Banyai, A. A. Godil, and D. M. Bloom, "Time-to-frequency converter for measuring picosecond optical pulses," *Appl. Phys. Lett.* **64**: 270–272 (1994).
40. I. P. Christov, "Theory of a time telescope," *Opt. Quantum Electron.* **22**: 473–480 (1990).
41. C. V. Bennett and B. H. Kolner, "Upconversion time microscope demonstrating 103X magnification of femtosecond waveforms," *Opt. Lett.* **24**: 783–785 (1999).

42. C. V. Bennett and B. H. Kolner, "Principles of parametric temporal imaging—part I: System configurations," *IEEE J. Quantum Electron*. **36**: 430–437 (2000).
43. C. V. Bennett and B. H. Kolner, "Principles of parametric temporal imaging-part II: System performance," *IEEE J. Quantum Electron*. **36**: 649–655 (2000).
44. A. G. Kostenbauder, "Ray-pulse matrices: A rational treatment for dispersive optical systems," *IEEE J. Quantum Electron*. **26**: 1148–1157 (1990).
45. J. Paye and A. Migus, "Space-time Wigner functions and their application to the analysis of a pulse shaper," *J. Opt. Soc. Am. B* **12**: 1480–1490 (1995).
46. Y. Sutoh, Y. Yasuno, K. Harada, M. Itoh, and T. Yatagai, "Analysis of spatiotemporal coupling in a femtosecond pulse shaper by the Wigner distribution function," *Opt. Eng*. **40**: 1717–1723 (2001).
47. I. A. Walmsley and V. Wong, "Characterization of the electric field of ultrashort optical pulses," *J. Opt. Soc. Am. B* **13**: 2453–2463 (1996).
48. E. B. Treacy, "Measurement and interpretation of dynamic spectrograms of picosecond light pulses," *J. Appl. Phys*. **42**: 3848–3858 (1971).
49. J. Paye, M. Ramaswamy, J. G. Fujimoto, and E. P. Ippen, "Measurement of the amplitude and phase of ultrashort light pulses from spectrally resolved autocorrelation," *Opt. Lett*. **18**: 1946–1948 (1993).
50. R. Trebino, K. W. Delong, D. N. Fittinghoff, J. N. Sweetser, M. A. Krumbügel, B. A. Richman, and D. J. Kane, "Measuring ultrashort laser pulses in the time-frequency domain using frequency-resolved optical gating," *Rev. Sci. Instrum*. **68**: 3277–3295 (1997).
51. S. Linden, J. Kuhl, and H. Giessen, "Amplitude and phase characterization of weak blue ultrashort pulses by downconversion," *Opt. Lett*. **24**: 569–571 (1999).
52. A. Efimov and A. J. Taylor, "Supercontinuum generation and solition timing jitter in SF6 soft glass photonic crystal fibers," *Opt. Express* **16**: 5942–5953 (2008).
53. J. M. Dudley, X. Gu, L. Xu, M. Kimmel, E. Zeek, P. O'Shea, R. Trebino, et al., "Cross-correlation frequency resolved optical gating analysis of broadband continuum generation in photonic crystal fiber: Simulations and experiments," *Opt. Express* **10**: 1215–1221 (2002).
54. H. R. Lange, M. A. Franco, J.-F. Ripoche, B. S. Prade, P. Rousseau, and A. Mysyrowicz, "Reconstruction of the time profile of femtosecond laser pulses through cross-phase modulation," *IEEE J. Selected Topics Quantum Electron*. **4**: 295–300 (1998).
55. C. Dorrer and I. Kang, "Simultaneous temporal characterization of telecommunication optical pulses and modulators using spectrograms," *Opt. Lett*. **27**: 1315–1317 (2002).
56. A. S. L. Gomes, A. S. Gouveia-Neto, and J. R. Taylor, "Direct measurement of chirped optical pulses with picosecond resolution," *Electron. Lett*. **22**: 41–42 (1986).
57. Y. Ozeki, Y. Takushima, H. Yoshimi, K. Kikuchi, H. Yamauchi, and H. Taga, "Complete characterization of picosecond optical pulses in long-haul dispersion-managed transmission systems," *IEEE Photon. Technol. Lett*. **17**: 648–650 (2005).
58. J. L. A. Chilla and O. E. Martinez, "Analysis of a method of phase measurement of ultrashort pulses in the frequency domain," *IEEE J. Quantum Electron*. **27**: 1228–1235 (1991).
59. V. Wong and I. A. Walmsley, "Ultrashort-pulse characterization from dynamic spectrograms by iterative phase retrieval," *J. Opt. Soc. Am. B* **14**: 944–949 (1997).
60. D. T. Reid, "Algorithm for complete and rapid retrieval of ultrashort pulse amplitude and phase from a sonogram," *IEEE J. Quantum Electron*. **35**: 1584–1589 (1999).
61. D. J. Kane, "Principal components generalized projections: A review," *J. Opt. Soc. Am. B* **25**: 120–132 (2008).
62. A. C. Kak and M. Slaney, *Principles of Computerized Tomographic Imaging*, IEEE Press, New York, 1987.

63. M. Beck, M. G. Raymer, I. A. Walmsley, and V. Wong, "Chronocyclic tomography for measuring the amplitude and phase structure of optical pulses," *Opt. Lett.* **18**: 2041–2043 (1993).
64. C. Dorrer and I. Kang, "Complete temporal characterization of short optical pulses by simplified chronocyclic tomography," *Opt. Lett.* **28**: 1481–1483 (2003).
65. T. Alieva, M. J. Bastiaans, and L. Stankovic, "Signal reconstruction from two close fractional Fourier power spectra," *IEEE Trans. Signal Process.* **51**: 112–123 (2003).
66. C. Dorrer, "Characterization of nonlinear phase shifts by use of the temporal transport-of-intensity equation," *Opt. Lett.* **30**: 3237–3239 (2005).
67. S. C. Woods and A. H. Greenaway, "Wavefront sensing by use of a Green's function solution to the intensity transport equation," *J. Opt. Soc. Am. A* **20**: 508–512 (2003).
68. L. Lepetit, G. Chériaux, and M. Joffre, "Linear techniques of phase measurement by femtosecond spectral interferometry for applications in spectroscopy," *J. Opt. Soc. Am. B* **12**: 2467–2474 (1995).
69. D. N. Fittinghoff, J. L. Bowie, J. N. Sweetser, R. T. Jennings, M. A. Krumbügel, K. W. Delong, R. Trebino, et al., "Measurement of the intensity and phase of ultraweak, ultrashort laser pulses," *Opt. Lett.* **21**: 884–886 (1996).
70. C. Dorrer, "Complete characterization of periodic optical sources by use of sampled test-plus-reference interferometry," *Opt. Lett.* **30**: 2022–2024 (2005).
71. K. C. Chu, J. P. Heritage, R. S. Grant, K. X. Liu, A. Dienes, W. E. White, and A. Sullivan, "Direct measurement of the spectral phase of femtosecond pulses," *Opt. Lett.* **20**: 906 (1995).
72. K. C. Chu, J. P. Heritage, R. S. Grant, and W. E. White, "Temporal interferometric measurement of femtosecond spectral phase," *Opt. Lett.* **21**: 1842–1844 (1996).
73. R. M. Fortenberry and V. Wayne, "Apparatus for characterizing short optical pulses," U.S. Patent 5,684,586, 1997.
74. C. Dorrer, "Chromatic dispersion measurement using direct instantaneous frequency measurement," *Opt. Lett.* **29**: 204–206 (2004).
75. S. Prein, S. A. Diddams, and J.-C. Diels, "Complete characterization of femtosecond pulses using an all-electronic detector," *Opt. Comm.* **123**: 567–573 (1996).
76. V. Messager, F. Louradour, C. Froehly, and A. Barthélémy, "Coherent measurement of short laser pulses based on spectral interferometry resolved in time," *Opt. Lett.* **28**: 743–745 (2003).
77. P. Kockaert, M. Peeters, S. Coen, P. Emplit, M. Haelterman, and O. Deparis, "Simple amplitude and phase measuring technique for ultrahigh-repetition-rate lasers," *IEEE Photon. Technol. Lett.* **12**: 187–189 (2000).
78. J. E. Rothenberg and D. Grischkowsky, "Measurement of optical phase with subpicosecond resolution by time-domain interferometry," *Opt. Lett.* **12**: 99–101 (1987).
79. V. A. Zubov and T. I. Kuznetsova, "Solution of the phase problem for time-dependent optical signals by an interference system," *Sov. J. Quantum Electron.* **21**: 1285–1286 (1991).
80. V. Wong and I. A. Walmsley, "Analysis of ultrashort pulse-shape measurement using linear interferometers," *Opt. Lett.* **19**: 287–289 (1994).
81. M. Takeda, H. Ina, and S. Kobayashi, "Fourier-transform method of fringe-pattern analysis for computer-based topography and interferometry," *J. Opt. Soc. Am. A* **72**: 156–160 (1982).
82. C. Dorrer and I. Kang, "Highly sensitive direct femtosecond pulse characterization using electro-optic spectral shearing interferometry," *Opt. Lett.* **28**: 477–479 (2003).
83. J. Bromage, C. Dorrer, I. A. Begishev, N. G. Usechak, and J. D. Zuegel, "Highly sensitive, single-shot characterization for pulse widths from 0.4 to 85 ps using electro-optic shearing interferometry," *Opt. Lett.* **31**: 3523–3525 (2006).

84. J. Debeau, B. Kowalski, and R. Boittin, "Simple method for the complete characterization of an optical pulse," *Opt. Lett*. **23**: 1784–1786 (1998).
85. M. Kwakernaak, R. Schreieck, and A. Neiger, "Spectral phase measurement of mode-locked diode laser pulses by beating sidebands generated by electrooptical mixing," *IEEE Photon. Technol. Lett*. **12**: 1677–1679 (2000).
86. I. Kang and C. Dorrer, "Method of optical pulse characterization using sinusoidal optical phase modulations," *Opt. Lett*. **32**: 2538–2540 (2007).
87. C. Iaconis and I. A. Walmsley, "Spectral phase interferometry for direct electric-field reconstruction of ultrashort optical pulses," *Opt. Lett*. **23**: 792–794 (1998).
88. C. Iaconis and I. A. Walmsley, "Self-referencing spectral interferometry for measuring ultrashort optical pulses," *IEEE J. Quantum Electron*. **35**: 501–509 (1999).
89. Y. Ozeki, S. Takasaka, and M. Sakano, "Electrooptic spectral shearing interferometry using a Mach-Zehnder modulator with a bias voltage sweeper," *IEEE Photon. Technol. Lett*. **18**: 911–913 (2006).

This page intentionally left blank

Index

N

O

P

Q

R

S

T

V

W

Y

Z