OPTICAL VORTICES AND COHERENCE IN NANO-OPTICS

by

Matt Smith

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Nanoscale Science

Charlotte

2019

Approved by:

Dr. Gregory J. Gbur

Dr. Thomas Suleski

Dr. Yong Zhang

Dr. Tom Schmedake

Dr. Matthew Eastin

ABSTRACT

MATT SMITH. Optical vortices and coherence in nano-optics. (Under the direction of DR. GREGORY J. GBUR)

In this dissertation, we use theory and computation to conduct three projects in the area of nano-optics. In the first project, we study optical coherence conversion using a plasmonic hole array. This work led to the discovery of the optical coherence band gap. In the second project, we develop a method of producing sub-wavelength arrangements of optical vortices. We demonstrate the validity of this method and propose an experimental realization. In the third project, we apply the techniques of the second to design superresolution lenses.

ACKNOWLEDGEMENTS

I would like to thank Dr. Greg Gbur for being my advisor, and the other members of my committee for giving their time and feedback. I would also like to acknowledge financial support from GASP and from the Air Force Office of Scientific Research. In the course of my dissertation, I used resources from MOSAIC and from University Research Computing at UNCC.

DEDICATION

To my parents, for their unceasing support, confidence, encouragement, and love.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION TO OPTICAL COHERENCE THEORY

1.1 Introduction

In this chapter, we will present an introduction to optical coherence theory, which is the focus of the project in Chapters 2 and 3. For the purposes of this dissertation, we shall not attempt to provide a comprehensive review of the entire subject. Instead, we will only cover the concepts needed to understand the associated projects in Chapters 2 and 3. We refer the interested reader to References [1, chapters 1-4], [2, chapters 2-4], and [3, chapter 10], from which we take the information in this chapter unless cited otherwise.

Most physicists are likely familiar with the idea that light can be either "coherent," such as a laser, or "incoherent," such as sunlight. Most also probably know some characteristics that distinguish coherent light from incoherent. For example, laser light is highly directional and causes interference patterns in a Young-style double pinhole experiment, whereas sunlight is not very directional and does not typically cause interference patterns in a Young experiment. This knowledge doesn't suffice to explain what coherence actually *is*, nor does it describe the richness of coherence theory or reveal just how central coherence is to the physics of light.

So what is coherence? Summed up in one sentence, "Coherence is essentially a consequence of correlations between some components of the fluctuating electric field at two (or more) points," [1, pp. xii]. Light sources found in nature or in a laboratory are not the well-behaved, monochromatic, mono-directional, perfectly (un)polarized plane waves we generally use as models to describe light physics. Real life is messier than that. In particular, physical fields always have some random fluctuations associated with them. Coherence is a measure of how strongly correlated these randomly vibrating fields are between two points in time or space, known as temporal coherence and spatial coherence, respectively. We shall only be concerned with spatial coherence in this dissertation. At first glance, quantifying this correlation might seem like something with no consequence or practical use. However, it turns out that this statistical similarity has profound consequences for the physics of light. In fact, it has been shown that a field's directionality [4], polarization [5, 6], and even its far-field spectrum [7] are all influenced by the field's spatial coherence.

The remainder of this chapter is organized as follows. In Section 1.2 we will review some mathematical concepts needed to understand the rest of the chapter. Then in Section 1.3, we will introduce the basic quantities that form the foundation of coherence theory. Finally, in Section 1.4 we will give some concluding remarks.

1.2 Mathematical Preliminaries

In this section, we will introduce some of the mathematical tools and concepts necessary for understanding coherence theory. As we said in the Introduction, coherence theory is built around the random vibrations that actual light sources possess. Therefore the techniques and concepts that we will need come from *random process theory.*

1.2.1 Average values of random processes

Suppose we have a real field variable $x(t)$ that represents a Cartesian component of a steady-state[1] electric field at some arbitrary point in space and at time t. Let us also suppose that $x(t)$ has random fluctuations. We can measure $x(t)$ in a carefully-controlled series of similar experiments, with the result of the j^{th} experiment denoted ${}^{j}x(t)$. We refer to this set of results as an *ensemble of realizations* or *ensemble of sample functions* of $x(t)$. We denote the entire ensemble using curly brackets as $\{x(t)\}$. So ${}^{j}x(t)$ is then called the j^{th} realization of the ensemble $\{x(t)\}$.

[1] *Steady-state* has a specific meaning in random process theory that we will address later.

Now, at optical frequencies ($\approx 1 \times 10^{15}\,\mathrm{Hz}$), we could not actually measure the time behavior of ${}^{j}x(t)$, due to the rapidity of light vibrations, but we can measure the time average. For a typical realization ${}^{j}x(t)$ of $x(t)$, the time average, denoted using angle brackets $\langle\ldots\rangle_t$, is defined as

$$\left\langle {}^{j}x(t)\right\rangle_t := \lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T} {}^{j}x(t)\,\mathrm{d}t. \tag{1.1}$$

We note that the time average may be different for each realization ${}^{j}x(t)$, so in general we also have an ensemble of time averages, $\{\langle x(t)\rangle_t\}$.

Another important kind of average in the theory of random processes is the *ensemble average* or *expectation value.* This is the time-dependent average over all of the realizations in the ensemble. In general, the number of realizations of a random optical field will be practically infinite, so by denoting the number of realizations as N we can express the ensemble average as

$$\langle x(t)\rangle_{\mathrm{e}} := \lim_{N\to\infty}\frac{1}{N}\sum_{j=1}^{N} {}^{j}x(t)\,, \tag{1.2}$$

which is the discrete form of the ensemble average, denoted by the subscript e on the angle brackets. The continuous form of the ensemble average is given by

$$\langle x(t)\rangle_{\mathrm{e}} := \int x p_1(x,t)\,\mathrm{d}x, \tag{1.3}$$

where the integration is over the allowed range of values of $x(t)$, and $p_1(x,t)\,\mathrm{d}x$ is the probability that x will have a value in the range $(x, x+\mathrm{d}x)$ at time t. We have used x in the integration instead of $x(t)$ because we are integrating over the values $x(t)$ is allowed to take, not over the time behavior of a particular realization ${}^{j}x(t)$. In coherence theory we are typically concerned with finding ensemble averages of the form $\langle z^{*}(t_1)\,z(t_2)\rangle_{\mathrm{e}}$, where $z(t) := x(t) + \mathrm{i}y(t)$, i is the imaginary unit, and the asterisk denotes complex conjugation. This ensemble average is given by

$$\langle z^*(t_1)\, z(t_2)\rangle_{\mathrm{e}} = \iint z_1^* z_2 p_2(z_1, z_2; t_1, t_2)\, \mathrm{d}^2 z_1 \mathrm{d}^2 z_2. \tag{1.4}$$

The probability $p_2(z_1, z_2; t_1, t_2)\, \mathrm{d}^2 z_1 \mathrm{d}^2 z_2$ is the probability that at time t_1, z will be take a value located within an infinitesimal area $\mathrm{d}^2 z_1 = \mathrm{d}x_1 \mathrm{d}y_1$ about the point z_1 and that at time t_2, z will take a value within a similar region $\mathrm{d}^2 z_2 = \mathrm{d}x_2 \mathrm{d}y_2$ about the point z_2. Figure 1.1 shows the relation of the pair of points z_1 and z_2 with their corresponding $\mathrm{d}^2 z_j$ and $\mathrm{d}x_j$ and $\mathrm{d}y_j$, where $j = (1, 2)$.

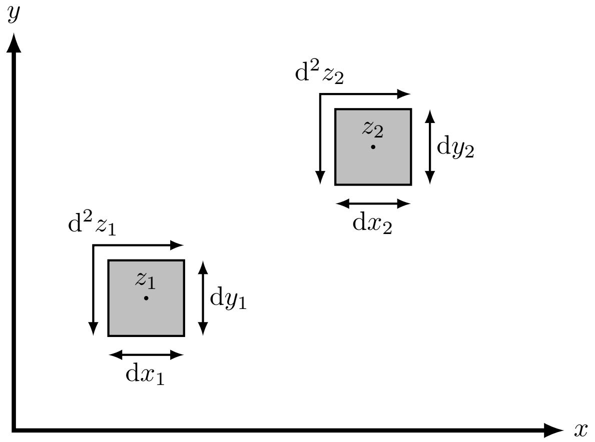

Figure 1.1: Showing the elements $\mathrm{d}^2 z_1$ and $\mathrm{d}^2 z_2$ about the points z_1 and z_2.

Now, we earlier mentioned *steady-state* fields. By that, we mean that the statistical properties of the field do not depend on the origin of time. Such a field is said to be *statistically stationary*; examples include sunlight and continuous-wave lasers. So if $z(t)$ in Eq. (1.4) represents a statistically stationary field, then if the origin of time is shifted by an arbitrary amount τ, the probability $p_2(z_1, z_2; t_1, t_2)\, \mathrm{d}^2 z_1 \mathrm{d}^2 z_2$ will be unchanged. That is,

$$p_2(z_1, z_2; t_1 + \tau, t_2 + \tau)\, \mathrm{d}^2 z_1 \mathrm{d}^2 z_2 = p_2(z_1, z_2; t_1, t_2)\, \mathrm{d}^2 z_1 \mathrm{d}^2 z_2. \tag{1.5}$$

So far we have been considering two types of averages: time averages and ensemble averages. Fortunately, it is often the case that, when the light fluctuations are statically stationary, the two averages are equivalent. That is,

$$\left\langle {}^j x(t) \right\rangle_t = \langle x(t) \rangle_\mathrm{e} . \qquad (1.6)$$

Such fields are called *ergodic.* For the remainder of this chapter, we will assume that the ensembles of fields we deal with are ergodic. Therefore we will drop the t and e subscripts from the angle brackets from now on.

1.2.2 Autocorrelation and cros-correlation functions

The two most important values associated with a real random process are its mean,

$$m(t) := \langle x(t) \rangle \qquad (1.7)$$

and its *autocorrelation function*

$$R(t_1, t_2) := \langle x(t_1)\, x(t_2) \rangle . \qquad (1.8)$$

In this chapter, we will take all of our fields to have a mean of zero. If we take the process to be statistically stationary, then the mean will be independent of time and the autocorrelation function will only depend on the time argument through the difference $\tau = t_2 - t_1$. We can then replace $R(t_1, t_2)$ with $R(\tau)$,

$$R(\tau) := \langle x(t)\, x(t + \tau) \rangle . \qquad (1.9)$$

We will be mostly concerned with complex random variables $z(t)$, in which case the autocorrelation function is defined as

$$R(\tau) = \langle z^*(t)\, z(t + \tau) \rangle . \qquad (1.10)$$

For our purposes in this dissertation, we will primarily be concerned with a version of $R(\tau)$ that can be used for situations with two random variables, $z_1(t)$ and $z_2(t)$, which we will use later to represent field amplitudes at a pair of points. If the two processes are jointly stationary, meaning their joint probability distribution is invariant with respect to translation of the origin of time, the measure of their correlation is the *cross-correlation function*,

$$R_{12}(\tau) := \langle z_1^*(t)\, z_2(t+\tau)\rangle \,. \tag{1.11}$$

It has two important properties:

$$|R_{12}(\tau)| \leq \sqrt{R_{11}(0)\, R_{22}(0)} \tag{1.12a}$$

$$R_{12}(-\tau) = R_{21}^*(\tau)\,. \tag{1.12b}$$

The coherence theory we will describe later in this chapter is built around two quantities, the mutual coherence function and the cross-spectral density, which we will show to be cross-correlation functions.

When the mean of a random process is independent of time and the autocorrelation function depends only on the time difference $\tau = t_2 - t_1$, the process is said to be *stationary in the wide sense.* This is a somewhat less strict version of statistical stationarity, which required that all the process's probability densities remain invariant under translation of the origin of time. In this chapter, we will frequently assume that fields are stationary at least in the wide sense.

1.3 Some Basic Coherence Quantities

Now we can get into the meat of coherence theory. The scalar coherence theory we are considering has two core quantities: the *mutual coherence function*, in the space-time domain, and the *cross-spectral density*, in the space-frequency domain. A third quantity, and the most significant one for this dissertation, is the *spectral degree of coherence*, also in the space-frequency domain. We will develop these now.

1.3.1 Space-time domain

Let us assume that we have a Young-style double pinhole experiment, as shown in Fig. 1.2. We assume that the light is statistically stationary, at least in the wide sense, and that it is *quasi-monochromatic*. Quasi-monochromatic light is light that has a mean radial frequency ω_0 and bandwidth $\Delta\omega$ narrow enough that

$$\frac{\Delta\omega}{\omega_0} \ll 1. \tag{1.13}$$

The light is incident on an opaque screen $\mathcal{A}$ with pinholes at points Q_1 and Q_2. We will be considering the average intensity of the field in the neighborhood of an arbitrary point P on the observation screen $\mathcal{B}$.

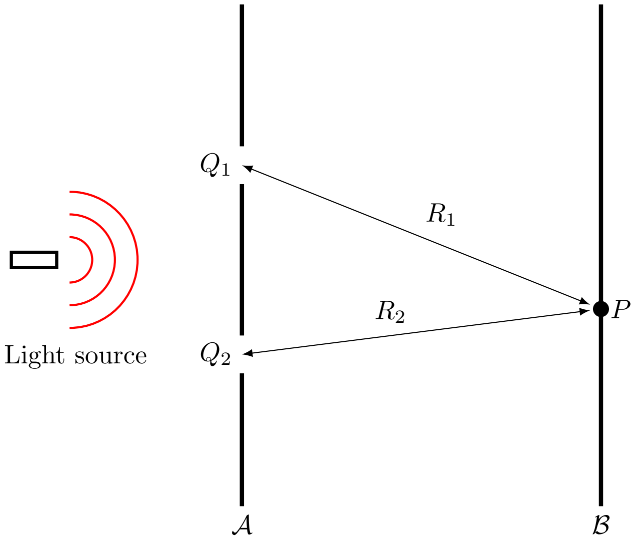

Figure 1.2: Setup and notation of a Young-style double pinhole experiment.

Neglecting polarization, we denote the light vibrations by $V(\mathbf{r}, t)$, a complex scalar. The time it will take for the light to travel the distances R_j $(j = 1, 2)$, is

$$t_j = \frac{R_j}{c} \tag{1.14}$$

where c is the speed of light in vacuum. The field at P will be

$$V(P,t) = K_1 V(Q_1, t - t_1) + K_2 V(Q_2, t - t_2)\,, \tag{1.15}$$

where K_1 and K_2 are to account for diffraction. Here, for small angles of incidence and diffraction,

$$K_j = -\frac{\mathrm{i}}{\lambda_0 R_j} \mathrm{d}A_j, \quad (j = 1, 2), \tag{1.16}$$

where $\mathrm{d}A_j$ are the areas of the pinholes and $\lambda_0 = 2\pi c/\omega_0$ is the mean wavelength of the light. We assume that the pinholes are small enough that the field amplitude is effectively constant over each of them.

As noted earlier, optical fields vibrate at around 1×10^{15} Hz, so the time behavior of the field amplitude in Eq. (1.15) cannot be measured. However, we can measure the average intensity,

$$I(P) := \langle I(P,t) \rangle = \langle V^*(P,t)\, V(P,t) \rangle\,. \tag{1.17}$$

This quantity is independent of the origin of time due to the assumed statistical stationarity. From Eq. (1.15), we have

$$\begin{aligned} I(P) = {} & |K_1|^2 \langle V^*(Q_1, t - t_1)\, V(Q_1, t - t_1) \rangle + |K_2|^2 \langle V^*(Q_2, t - t_2)\, V(Q_2, t - t_2) \rangle \\ & + 2\mathcal{R}\{K_1^* K_2 \langle V^*(Q_1, t - t_1)\, V(Q_2, t - t_2) \rangle\}\,, \end{aligned} \tag{1.18}$$

where $\mathcal{R}\{\}$ denotes the real part. Using the assumed stationarity, we note that $\langle V^*(Q_j, t - t_j)\, V(Q_j, t - t_j) \rangle$ is the average intensity $I(Q_j)$ at pinhole j. We also note that the constants K_j are purely imaginary. Using all of this, we simplify Eq. (1.18) to yield the *interference law for stationary optical fields,*

$$I(P) = |K_1|^2\, I(Q_1) + |K_2|^2\, I(Q_2) + 2\mathcal{R}\{|K_1|\,|K_2|\,\Gamma(Q_1, Q_2, t_1 - t_2)\}\,, \tag{1.19}$$

where

$$\Gamma(Q_1, Q_2, \tau) := \langle V^*(Q_1, t)\, V(Q_2, t + \tau)\rangle\,, \tag{1.20}$$

with $\tau = t_1 - t_2$, is the *mutual coherence function.* This quantity is the cross-correlation function between the two fields at the pinholes, as described in Eqs. (1.11) and (1.12).

The mutual coherence function is a measure of the coherence of the field between the two points Q_1 and Q_2 in the space-time domain. In fact, a normalized version of this quantity, called the *complex degree of coherence*,

$$\gamma(Q_1, Q_2, \tau) := \frac{\Gamma(Q_1, Q_2, \tau)}{\sqrt{\Gamma(Q_1, Q_1, 0)}\sqrt{\Gamma(Q_2, Q_2, 0)}} = \frac{\Gamma(Q_1, Q_2, \tau)}{\sqrt{I(Q_1)}\sqrt{I(Q_2)}} \tag{1.21}$$

can be shown to determine the sharpness of the fringes in Young's experiment with quasi-monochromatic light. First, we note from Eq. (1.12a) that

$$0 \leq |\gamma(Q_1, Q_2, \tau)| \leq 1, \tag{1.22}$$

where 1 corresponds to complete coherence of the light and 0 corresponds to complete incoherence. Next, we define the sharpness of the fringes in the neighborhood of the point P as the *visibility*,

$$\mathcal{V}(P) := \frac{I_{\max}(P) - I_{\min}(P)}{I_{\max}(P) + I_{\min}(P)}, \tag{1.23}$$

where $I_{\max}(P)$ and $I_{\min}(P)$ are the maximum and minimum, respectively, of intensity in the neighborhood of P. With this definition, we can rewrite $\gamma(Q_1, Q_2, \tau)$ as

$\left|\gamma(Q_1, Q_2, \tau)\right| \mathrm{e}^{\mathrm{i}\alpha((Q_1,Q_2,\tau))}$ and use Eqs. (1.19) and (1.21) to show that

$$\mathcal{V}(P) = \left|\gamma(Q_1, Q_2, \tau)\right| . \tag{1.24}$$

Equation (1.24) shows the important result that the sharpness of the interference fringes is controlled by the light's coherence.

One important thing to consider is how $\Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)$ propagates thorugh space, where we are now using $\mathbf{r}$ to denote arbitrary points in space rather than being confined to the screen in Fig. 1.2. Suppose that we have an ensemble $\{V(\mathbf{r}, t)\}$ representing a complex wavefield in free space. Each member of the ensemble satisfies the wave equation,

$$\nabla^2 V(\mathbf{r}, t) = \frac{1}{c^2} \frac{\partial^2}{\partial t^2} V(\mathbf{r}, t) . \tag{1.25}$$

We can then take the complex conjugate of Eq. (1.25), replace $\mathbf{r}$ and t by $\mathbf{r}_1$ and t_1, respectively, and multiply this new equation by $V(\mathbf{r}_2, t_2)$. This yields

$$\nabla_1^2 V^*(\mathbf{r}_1, t_1)\, V(\mathbf{r}_2, t_2) = \frac{1}{c^2} \left[\frac{\partial^2}{\partial {t_1}^2} V^*(\mathbf{r}_1, t_1) \right] V(\mathbf{r}_2, t_2) , \tag{1.26}$$

where ∇_1^2 is the Laplacian operator that only acts on the points in $\mathbf{r}_1$. We can now take the ensemble average of both sides of Eq. (1.26) and interchange the orders of the various operators to yield

$$\nabla_1^2 \left\langle V^*(\mathbf{r}_1, t_1)\, V(\mathbf{r}_2, t_2) \right\rangle = \frac{1}{c^2} \frac{\partial^2}{\partial {t_1}^2} \left\langle V^*(\mathbf{r}_1, t_1)\, V(\mathbf{r}_2, t_2) \right\rangle . \tag{1.27}$$

If the field is statistically stationary, at least in the wide sense, then

$$\begin{aligned} \left\langle V^*(\mathbf{r}_1, t_1)\, V(\mathbf{r}_2, t_2) \right\rangle &= \left\langle V^*(\mathbf{r}_1, t)\, V(\mathbf{r}_2, t + t_2 - t_1) \right\rangle \\ &= \Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) . \end{aligned} \tag{1.28}$$

We note that $\tau = t_2 - t_1$ and that $\frac{\partial^2}{\partial {t_1}^2} = \frac{\partial^2}{\partial \tau^2}$. Thus, we can rewrite Eq. (1.27) as

$$\nabla_1^2 \Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) = \frac{1}{c^2} \frac{\partial^2}{\partial \tau^2} \Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)\,. \tag{1.29}$$

By a similar process we can show that

$$\nabla_2^2 \Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) = \frac{1}{c^2} \frac{\partial^2}{\partial \tau^2} \Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)\,. \tag{1.30}$$

Equations (1.29) and (1.30) are a remarkable result. The statistical correlation between points of a randomly fluctuating quasi-monochromatic wavefield itself propagates as a wave! Rather famously, when Emil Wolf first discovered this result and presented it to Max Born, Born responded, “Wolf, you have always been such a sensible fellow, but now you have become completely crazy!” [8, pp. 287]. This result makes one wonder: what *other* wave-like properties might coherence have?

For this dissertation, we shall be primarily concerned with the space-frequency domain, so although there is much more that could be said about $\Gamma(Q_1, Q_2, \tau)$ and $\gamma(Q_1, Q_2, \tau)$, we will move on. The interested reader may consult Refs. [1, chapter 3] and [2, chapter 4]

1.3.2 Space-frequency domain

The space-frequency domain of coherence theory has parallels to the space-time domain, as one might expect. One parallel is that it also has two main quantities used to define coherence. The first is the *cross-spectral density*, which can be found via Fourier transform of the mutual coherence function, i.e.,

$$W(\mathbf{r}_1, \mathbf{r}_2, \omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)\, \mathrm{e}^{\mathrm{i}\omega\tau} \mathrm{d}\tau. \tag{1.31}$$

From the cross-spectral density, we can derive the most significant quantity for the purposes of this dissertation: the *spectral degree of coherence*,

$$\mu(\mathbf{r}_1, \mathbf{r}_2, \omega) := \frac{W(\mathbf{r}_1, \mathbf{r}_2, \omega)}{\sqrt{S(\mathbf{r}_1, \omega)}\sqrt{S(\mathbf{r}_2, \omega)}}, \tag{1.32}$$

where $S(\mathbf{r}_j, \omega)$ is the *spectral density*, defined as

$$S(\mathbf{r}_j, \omega) := W(\mathbf{r}_j, \mathbf{r}_j, \omega)\,. \tag{1.33}$$

To understand the physical significance of these two quantities, it will help to first consider an alternate representation of the cross-spectral density.

Under the conditions of Hermiticity, non-negative definiteness, and square-integrability over its domain D in free-space, the cross-spectral density can be expressed, for three-dimensional D, as

$$W(\mathbf{r}_1, \mathbf{r}_2, \omega) = \sum_i \sum_j \sum_k \lambda_{ijk}(\omega)\, \phi^*_{ijk}(\mathbf{r}_1, \omega)\, \phi_{ijk}(\mathbf{r}_2, \omega)\,. \tag{1.34}$$

If the domain is two (one) dimensional, then the triple sum would be replaced by a double (single) sum and the triplet of integers ijk would be replaced by a pair (single integer). For notational simplicity, we express all these possibilities as a single symbolic sum and integer n,

$$W(\mathbf{r}_1, \mathbf{r}_2, \omega) = \sum_n \lambda_n(\omega)\, \phi^*_n(\mathbf{r}_1, \omega)\, \phi_n(\mathbf{r}_2, \omega)\,, \tag{1.35}$$

it being understood that what Σ and n stand for depends on the dimensionality of D. Equation (1.35) is known as the *coherent-mode representation* of the cross-spectral density. It is so named because each of the modes $\phi_n(\mathbf{r}, \omega)$ is fully coherent [1, pp. 69]. The functions $\phi_n(\mathbf{r}_j, \omega)$ and $\lambda_n(\omega)$ are the eigenfunctions and eigenvalues, respectively, of the integral equation

$$\int_D W(\mathbf{r}_1, \mathbf{r}_2, \omega)\, \phi_n(\mathbf{r}_1, \omega)\, \mathrm{d}^3 r_1 = \lambda_n(\omega)\, \phi_n(\mathbf{r}_2, \omega)\,. \tag{1.36}$$

The eigenfunctions $\phi_n(\mathbf{r}_1)$ form an orthonormal set over D, and the eigenvalues $\lambda_n(\omega)$ are positive.

Using the coherent-mode representation, we can show that the cross-spectral density can be expressed as a cross-correlation function of sample functions $U(\mathbf{r}, \omega)$. We consider $U(\mathbf{r}, \omega)$ to be of the form

$$U(\mathbf{r}, \omega) = \sum_n a_n(\omega)\, \phi_n(\mathbf{r}, \omega)\,, \tag{1.37}$$

where $a_n(\omega)$ are random coefficients satisfying

$$\langle a_n^*(\omega)\, a_m(\omega) \rangle_\omega = \lambda_n(\omega)\, \delta_{nm}. \tag{1.38}$$

Here, the ω on the angle brackets denotes an average over frequency and δ_{nm} is the Kronecker delta. Note that Eq. (1.38) implies that the eigenfunctions are mutually incoherent; that is, they do not interfere with each other. Using Eq. (1.37), we can consider the cross-correlation function,

$$\langle U^*(\mathbf{r}_1, \omega)\, U(\mathbf{r}_2, \omega) \rangle_\omega = \sum_n \sum_m \langle a_n^*(\omega)\, a_m(\omega) \rangle_\omega\, \phi_n^*(\mathbf{r}_1, \omega)\, \phi_m(\mathbf{r}_2, \omega)\,, \tag{1.39}$$

where we have interchanged the order of summation and ensemble averaging. Using Eq. (1.38), Eq. (1.39) reduces to

$$\langle U^*(\mathbf{r}_1, \omega)\, U(\mathbf{r}_2, \omega) \rangle_\omega = \sum_n \lambda_n(\omega)\, \phi_n^*(\mathbf{r}_1, \omega)\, \phi_n(\mathbf{r}_2, \omega)\,, \tag{1.40}$$

Note that the right-hand side of Eq. (1.37) is the same as the right-hand side of

Eq. (1.35), which means that

$$W(\mathbf{r}_1, \mathbf{r}_2, \omega) = \langle U^*(\mathbf{r}_1, \omega)\, U(\mathbf{r}_2, \omega)\rangle_\omega\,. \tag{1.41}$$

Similarly,

$$S(\mathbf{r}, \omega) = \langle U^*(\mathbf{r}, \omega)\, U(\mathbf{r}, \omega)\rangle_\omega\,. \tag{1.42}$$

Equation (1.41) shows that, for all pairs of points in its domain D, the cross spectral density of a statistically stationary field may be expressed as a cross-correlation function of an ensemble $\{U(\mathbf{r}, \omega)\}$ of space-frequency realizations $U(\mathbf{r}, \omega)$, where cross-correlation functions were defined in Eq. (1.11). This is a somewhat surprising result. Although $W(\mathbf{r}_1, \mathbf{r}_2, \omega)$ was originally given as the Fourier transform of $\Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)$, which is a correlation function, it is not obvious at first glance that $W(\mathbf{r}_1, \mathbf{r}_2, \omega)$ should itself be a correlation function. Equation (1.41) allows us to model $W(\mathbf{r}_1, \mathbf{r}_2, \omega)$ as an average of random monochromatic fields in a Young-style experiment, as we will now show.

We have defined $U(\mathbf{r}, \omega)$ in Eq. (1.37) as a sum of coherent modes $\phi_n(\omega)$, but that doesn't really explain what $U(\mathbf{r}, \omega)$ is. We will show here that $U(\mathbf{r}, \omega)$ can be regarded as the space-dependent part of a monochromatic wavefield. From Eqs. (1.29) and (1.30), we know that the mutual coherence function propagates according to the wave equation. Since $W(\mathbf{r}_1, \mathbf{r}_2, \omega)$ is the Fourier transform of $\Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)$, it follows that $W(\mathbf{r}_1, \mathbf{r}_2, \omega)$ obeys the Helmholtz equation:

$$\nabla_j^2 W(\mathbf{r}_1, \mathbf{r}_2, \omega) + k^2 W(\mathbf{r}_1, \mathbf{r}_2, \omega) = 0 \qquad (j = 1, 2). \tag{1.43}$$

By plugging Eq. (1.35) into Eq. (1.43), multiplying the result by $\phi_m(\mathbf{r}_1, \omega)$, integrating both sides over $\mathbf{r}_1$, and using the orthonormality of the eigenfunctions ϕ_n, we can see that every coherent mode $\phi_n(\mathbf{r}_1, \omega)$ obeys the Helmholtz equation:

$$\nabla^2 \phi_n(\mathbf{r}, \omega) + k^2 \phi_n(\mathbf{r}, \omega) = 0. \tag{1.44}$$

Since $U(\mathbf{r}, \omega)$ is a linear combination of the modes $\phi_n(\mathbf{r}, \omega)$, it follows that $U(\mathbf{r}, \omega)$ also obeys the Helmholtz equation, so

$$\nabla^2 U(\mathbf{r}, \omega) + k^2 U(\mathbf{r}, \omega) = 0. \tag{1.45}$$

This shows that $U(\mathbf{r}, \omega)$ can be considered as the space-dependent part of a monochromatic wavefield $V(\mathbf{r}, t) = U(\mathbf{r}, \omega)\, \mathrm{e}^{-\mathrm{i}\omega t}$. It is important to note that $U(\mathbf{r}, \omega)$ is *not* a Fourier frequency component of the field, since stationary random fields do not have Fourier spectra. (They don't tend to zero as t tends to infinity, therefore they have no Fourier transform.) Rather, $U(\mathbf{r}, \omega)$ is the space-dependent part of a member of the *statistical ensemble* $\left\{V(\mathbf{r}, t) = U(\mathbf{r}, \omega)\, \mathrm{e}^{-\mathrm{i}\omega t}\right\}$ of *monochromatic realizations* of frequency ω. This distinction is not significant for this dissertation, but it is important enough to mention. The interested reader can learn more in Ref. [1, pp. 63].

Let us now examine the physical meaning of the spectral degree of coherence by again considering the Young experiment in Fig. 1.2. We will again assume that the light is stationary, at least in the wide sense, but now we take the light to be broadband rather than quasi-monochromatic. We again assume that the pinholes are small enough that the field amplitude is effectively constant over each of them and that the angles of incidence and diffraction are small. Then the field on $\mathcal{B}$ is given, to a good approximation, by an ensemble of realizations $\{U(P, \omega)\}$, where

$$U(P, \omega) = K_1 U(Q_1, \omega)\, \mathrm{e}^{\mathrm{i}kR_1} + K_2 U(Q_2, \omega)\, \mathrm{e}^{\mathrm{i}kR_2}. \tag{1.46}$$

Here, the constants K_1 and K_2 are defined as before, except instead of a mean wavelength λ_0 we have a wavelength $\lambda = 2\pi c/\omega$, since we are now assuming broadband light, and $k = 2\pi/\lambda$. Although our light is broadband, recall that the random field $U(\mathbf{r}, \omega)$ does not have a Fourier spectrum, so λ is not a Fourier component. However, the field does have a finite *power spectrum*, and the wavelength λ is a component of this.

Now let's substitute Eq. (1.46) into the definition of $S(\mathbf{r}, \omega)$, Eq. (1.42), and use the Hermiticity relation $W(Q_2, Q_1, \omega) = W^*(Q_1, Q_2, \omega)$. This yields

$$S(P, \omega) = |K_1|^2 \, S(Q_1, \omega) + |K_2|^2 \, S(Q_2, \omega) + 2\mathcal{R}\{K_1^* K_2 W(Q_1, Q_2, \omega) \, \mathrm{e}^{-\mathrm{i}\delta}\}, \quad (1.47)$$

where

$$\delta = \frac{2\pi}{\lambda} (R_1 - R_2) \,. \quad (1.48)$$

We note that $|K_j|^2 \, S(Q_j, \omega)$ is the spectral density due only to the hole at Q_j so we can denote

$$S^{(j)}(P, \omega) := |K_j|^2 \, S(Q_j, \omega) \quad (j = 1, 2). \quad (1.49)$$

Using this, we can rewrite Eq. (1.47) as

$$S(P, \omega) = S^{(1)}(P, \omega) + S^{(2)}(P, \omega) + 2\sqrt{S^{(1)}(P, \omega)}\sqrt{S^{(2)}(P, \omega)}\mathcal{R}\{\mu(Q_1, Q_2, \omega) \, \mathrm{e}^{-\mathrm{i}\delta}\} \,. \quad (1.50)$$

If we rewrite $\mu(Q_1, Q_2, \omega)$ as

$$\mu(Q_1, Q_2, \omega) = |\mu(Q_1, Q_2, \omega)| \, \mathrm{e}^{\mathrm{i}\beta(Q_1, Q_2, \omega)}, \quad (1.51)$$

and take the common assumption that $S^{(2)} \approx S^{(1)}$, then Eq. (1.50) becomes

$$S(P, \omega) = 2S^{(1)}(P, \omega) \, \{1 + |\mu(Q_1, Q_2, \omega)| \cos[\beta(Q_1, Q_2, \omega) - \delta]\} \,. \quad (1.52)$$

Equation (1.52) is called the *spectral interference law*, which shows that coherence can cause changes in the spectrum of (possibly broadband) light via the frequency dependence of $\mu(Q_1, Q_2, \omega)$. These correlation-induced spectral changes are known as *the Wolf effect*.

Now we can understand the physical meaning of $\mu(Q_1, Q_2, \omega)$ by discussing how Eq. (1.52) can be used to experimentally determine its value. Referring to Eq. (1.52), we note that the spectral density at ω will have a maximum or minimum when $\cos(\beta(Q_1, Q_2, \omega) - \delta)$ is equal to 1 or -1, respectively. We denote these as

$$S_{\max}(P, \omega) := 2S^{(1)}\omega \left[1 + |\mu(Q_1, Q_2, \omega)|\right] \tag{1.53a}$$

$$S_{\min}(P, \omega) := 2S^{(1)}\omega \left[1 - |\mu(Q_1, Q_2, \omega)|\right] \tag{1.53b}$$

respectively. We can use these to define the *spectral visibility*,

$$\mathcal{V}(P, \omega) := \frac{S_{\max}(P, \omega) - S_{\min}(P, \omega)}{S_{\max}(P, \omega) + S_{\min}(P, \omega)}. \tag{1.54}$$

Substituting Eq. (1.53) into Eq. (1.54) shows that

$$\mathcal{V}(P, \omega) = |\mu(Q_1, Q_2, \omega)|\,. \tag{1.55}$$

So the absolute value of the spectral degree of coherence determines the visibility of the fringes. The spectral degree of coherence can be shown using Eq. (1.12a) to have its absolute value bounded by $0 \leq |\mu(Q_1, Q_2, \omega)| \leq 1$, where 0 means complete incoherence (so no spectral fringes) and 1 means complete coherence (so maximal spectral fringes). Suppose we now place narrow-band filters on the pinholes with identical mean frequency ω_0 and bandwidth $\Delta\omega$, so that we now have quasi-monochromatic light again. Then the spectral density at ω_0 will be approximately the intensity of the field, and the spectral visibility will be approximately the same quantity as the regular visibility, defined in Eq. (1.23). This also allows us to see the importance of coherence – it determines how strongly a wavefield interferes with itself. Notice also that $|\mu(Q_1, Q_2, \omega)|$ can take values *between* 0 and 1. These values are called *partial coherence*. Interference fringes in these cases will be present, but will not be maximally contrasted. See Fig. 1.3 for an example.

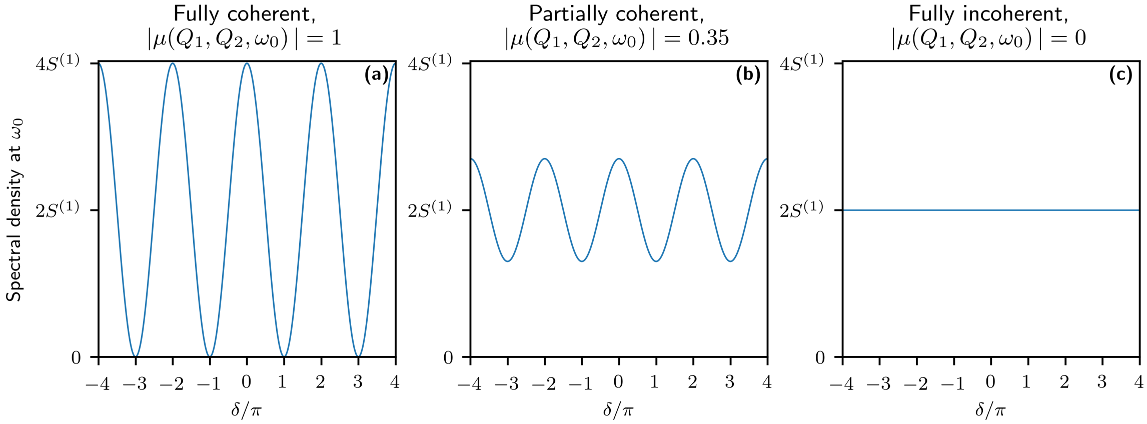

Figure 1.3: Examples of fringes for fully coherent, partially coherent, and fully incoherent light, narrow-band filtered at ω_0.

1.4 Conclusion

In this chapter, we have given the foundations of optical spatial coherence theory. We have shown that it is based around a time-domain cross-correlation function $\Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)$ and a frequency-domain cross-correlation function $W(\mathbf{r}_1, \mathbf{r}_2, \omega)$. We have given the definition and explained the physical significance of the spectral degree of coherence, $\mu(\mathbf{r}_1, \mathbf{r}_2, \omega)$, which will be important for the coherence project given in Chapters 2 and 3. We have also shown the difference between fully coherent, partially coherent, and fully incoherent light by demonstrating their control over light interference patterns in a Young-style double pinhole experiment. As mentioned at the beginning, a field's coherence influences a number of its properties, so being able to control that coherence is desirable for a number of applications.

CHAPTER 2: COHERENCE RESONANCES AND BAND GAPS IN PLASMONIC HOLE ARRAYS

2.1 Introduction

Surface plasmon polaritons, oscillations of electric charge density that propagate as waves confined to the surface of a suitable material, have a number of interesting applications. Their ability to confine light to subwavelength regions enables them to be used for nano-focusing of light beyond the diffraction limit [9]. Their extreme sensitivity to local refractive index makes them useful for sensing applications, such as disease diagnosis [10], blood type identification [11], and solution concentration sensing [12]. Arrays of sub-wavelength holes in metal plates have used surface plasmon polaritons to enhance the photon-to-electron efficiency of semiconductors [13] and as a form of e-paper [14].

Subwavelength holes in metal plates have also been shown to greatly enhance optical transmittance [15] and to modulate optical spatial coherence [16]. This ability to modulate coherence led to the proposal of a device [17] to convert the coherence of a beam from one spectral degree of coherence to another using a subwavelength-thickness metal plate perforated with an array of subwavelength holes. These holes can convert some of the incoming light into surface plasmon waves which can then decouple and combine with light at other holes. This process introduces a correlation between the light at different holes, which changes the degree of spatial coherence. This device was proposed because partially coherent light propagates through atmospheric turbulence more favorably than does fully coherent light [18, 19], so such a device could possibly have point-to-point optical communications applications. However, the optimum degree of coherence depends on the amount of turbulence, so it would be desirable to be able to tune the degree of coherence as needed. A device such as the one in Ref. [17] should permit that. However, Ref. [17] did not study the physical details of the multi-hole process, making it hard to design plates to achieve desired degrees of coherence at desired wavelengths. We will examine some of these details in Chapter 3. In the course of using the simulation method of Ref. [17] to examine these physical details, we stumbled on a previously undiscovered phenomenon – the optical coherence band gap, which can be thought of as a classical analogue of a quantum phenomenon called superradiance/subradiance. That is the main result of this chapter.

The remainder of this chapter is organized as follows. We begin with a brief overview of the model used for this work. Then we go into detail about the optical coherence band gap. Finally, we present some miscellaneous results related to our band gap work.

2.2 Description of Model

We use a simple scalar cylindrical wave model to describe the effects of plasmonic scattering from the holes; this model was first introduced in Ref. [17]. We use this, rather than a full electromagnetic simulation method such as Finite-Difference Time-Domain, for a few reasons. First, a simple model usually makes it easier to glean generalizable physical insights from the results. Second, simple models have much reduced complexity, which saves on computation time and on code production time. In this section, we will review the details of the model and discuss our new additions to it.

We consider a gold plate of subwavelength thickness lying in the $z = 0$ plane perforated by an array of holes with subwavelength diameter. From the $z < 0$ side, the plate is illuminated by a partially coherent quasi-monochromatic scalar field $U_0(\mathbf{r})$, with central wavelength λ_0 and spectral degree of coherence $\mu_0(\mathbf{r}_1, \mathbf{r}_2)$. Upon hitting the plate, some fraction of the field will transmit directly through the incident holes, and some will scatter off the holes into surface plasmon waves[1]. Upon hitting another hole, the plasmon waves can either rescatter or be emitted as light again. The new field being emitted from each hole will be correlated with the light from the other holes, which will generally result in a new degree of coherence, $\mu_f(\mathbf{r}_1, \mathbf{r}_2)$.

Figure 2.1 summarizes the complete process with a cut-away view of a plate with a 2×2 array of holes. The light incident on, and scattered from, each hole is depicted with color-coded arrows. Because of surface plasmons, the light emitted from each hole is a combination of the light from all other holes.

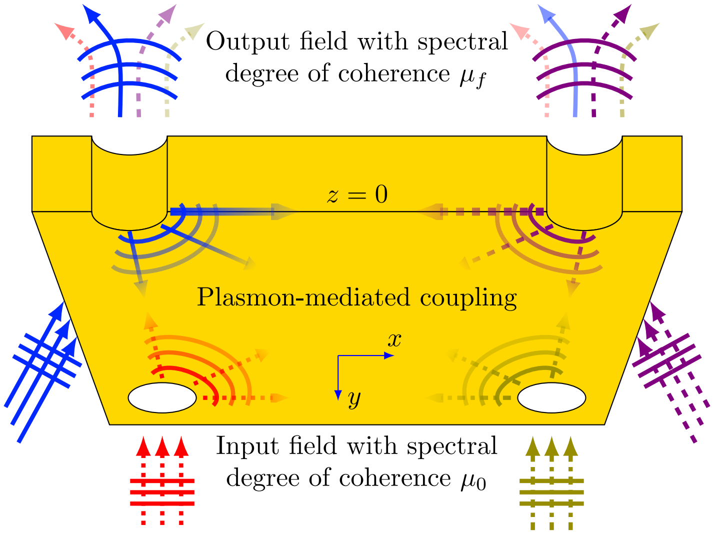
Output field with spectral degree of coherence μ_f
$z = 0$
Plasmon-mediated coupling
x
y
Input field with spectral degree of coherence μ_0

Figure 2.1: A cut-away sketch of a 2×2 plasmonic hole array showing the plasmon scattering process, viewed from below the $z = 0$ side of the plate. Note that the input field is the same everywhere in the plane; the arrows in the figure are color-coded only to show which hole they are incident on.

Each of the following subsections describe a particular element of the model: the spatial coherence, the scalar plasmonic wave scattering, and finally our additions to the model.

[1] Note that there are two kinds of surface plasmon waves: *Localized surface plasmons*, which are confined to a small region such as a nanosphere, and *surface plasmon polaritons*, which are propagating waves which can last for "long" distances before dissipating. In this chapter, when we discuss surface plasmons, we are exclusively referring to surface plasmon polaritons.

2.2.1 Spatial coherence model

We take our incident field to be of Gaussian-Schell form and express its cross-spectral density as an incoherent superposition of coherent modes. As we are concerned only with the behavior of the field right at the plate, we restrict our attention to the $z = 0$ plane. A Schell model field is defined such that the spectral degree of coherence depends only on the distance $|\boldsymbol{\rho}_m - \boldsymbol{\rho}_n|$ between points $\boldsymbol{\rho}_n$ and $\boldsymbol{\rho}_m$ [1, section 5.3.1], where $\boldsymbol{\rho}$ denotes (x, y) coordinates in the $z = 0$ plane. If the incident spectral density S_0 is constant across the $z = 0$ plane, then the incident cross spectral density W_0 is

$$W_0(\boldsymbol{\rho}_n, \boldsymbol{\rho}_m) = S_0 \mu_0(|\boldsymbol{\rho}_m - \boldsymbol{\rho}_n|)\,, \tag{2.1}$$

where μ_0 is the incident spectral degree of coherence. A Gaussian-Schell model field is a Schell model field with a Gaussian dependence on the distance, such that $\mu_0(|\boldsymbol{\rho}_m - \boldsymbol{\rho}_m|)$ is given by

$$\mu_0(|\boldsymbol{\rho}_m - \boldsymbol{\rho}_m|) = \exp\left(\frac{-\left|\boldsymbol{\rho}_m - \boldsymbol{\rho}_n\right|^2}{2\delta^2}\right), \tag{2.2}$$

where δ is the transverse correlation length. We express Eq. (2.1) as an incoherent superposition of modes by writing Eq. (2.2) in terms of its Fourier transform, $\tilde{\mu}_0(\mathbf{k})$,

$$\mu_0(|\boldsymbol{\rho}_m - \boldsymbol{\rho}_m|) = \iint_{-\infty}^{\infty} \tilde{\mu}_0(\mathbf{k})\, \mathrm{e}^{\mathbf{ik}\cdot(\boldsymbol{\rho}_m - \boldsymbol{\rho}_n)} \mathrm{d}^2\mathbf{k}, \tag{2.3}$$

where

$$\tilde{\mu}_0(\mathbf{k}) = \frac{\delta^2}{2\pi} \exp\left(-\frac{1}{2}\delta^2 \left|\mathbf{k}\right|^2\right). \tag{2.4}$$

Combining Eqs. (2.1) to (2.3), we may express cross-spectral density as a superposition of modes in the form

$$W_0(\boldsymbol{\rho}_n, \boldsymbol{\rho}_m) = \iint_{-\infty}^{\infty} \tilde{\mu}_0(\mathbf{k})\, \phi^*_{\mathbf{k}}(\boldsymbol{\rho}_n)\, \phi_{\mathbf{k}}(\boldsymbol{\rho}_m)\, \mathrm{d}^2\mathbf{k}, \tag{2.5}$$

where

$$\phi_{\mathbf{k}}(\boldsymbol{\rho}) := \sqrt{S_0}\mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\boldsymbol{\rho}} \tag{2.6}$$

are plane waves which shall be the coherent modes for our cross-spectral density. Since the plane waves are coherent, Eq. (2.5) is a coherent-mode representation of $W_0(\boldsymbol{\rho}_n, \boldsymbol{\rho}_m)$, which we defined in Eq. (1.35).

Now we consider propagating the field through the plate. For any linear coupling mechanism, the modes will remain mutually incoherent as they traverse the system. Therefore the cross-spectral density on the dark side of the plate $W_f(\boldsymbol{\rho}_n, \boldsymbol{\rho}_m)$ is

$$W_f(\boldsymbol{\rho}_n, \boldsymbol{\rho}_m) = \iint_{-\infty}^{\infty} \tilde{\mu}_0(\mathbf{k})\, \psi^*_{\mathbf{k}}(\boldsymbol{\rho}_n)\, \psi_{\mathbf{k}}(\boldsymbol{\rho}_m)\, \mathrm{d}^2\mathbf{k}, \tag{2.7}$$

where $\psi_{\mathbf{k}}(\boldsymbol{\rho})$ is the $\mathbf{k}^{\mathrm{th}}$ mode of the field on the dark side of the plate. Thus we can evaluate $W_f(\boldsymbol{\rho}_n, \boldsymbol{\rho}_m)$ by propagating individual input modes $\phi_{\mathbf{k}}(\boldsymbol{\rho})$ through the plate to obtain the output modes, denoted $\psi_{\mathbf{k}}(\boldsymbol{\rho})$. Then, using Eq. (2.7), we can calculate the output spectral degree of coherence between pairs of points using the definition of $\mu(\mathbf{r}_1, \mathbf{r}_2)$, Eq. (1.32).

2.2.2 Scalar model for plasmonic field

We first consider our incident field in the $z = 0$ plane to be a coherent mode $\phi_{\mathbf{k}}(\boldsymbol{\rho})$. Upon striking a hole at location $\boldsymbol{\rho}_n$, some fraction α of the mode will transmit directly through the hole, and some of the mode will couple to the plate surface as a surface plasmon wave and scatter to other holes. Assuming that the holes are smaller than the wavelength and that the distance between holes is multiple wavelengths, the holes can be modeled as point scatterers. The output mode $\psi_{\mathbf{k}}(\boldsymbol{\rho})$ is thus

$$\psi_{\mathbf{k}}(\boldsymbol{\rho}_n) = \alpha\phi_{\mathbf{k}}(\boldsymbol{\rho}_n) + \Psi_{\mathbf{k}}(\boldsymbol{\rho}_n)\,, \tag{2.8}$$

where $\Psi_{\mathbf{k}}$ is the plasmonic field from the other holes. The plasmonic field is defined as

$$\Psi_{\mathbf{k}}(\boldsymbol{\rho}_n) := \beta \sum_{m=1,m\neq j}^{N} G(\boldsymbol{\rho}_n, \boldsymbol{\rho}_m)\, \psi_{\mathbf{k}}(\boldsymbol{\rho}_m)\,, \tag{2.9}$$

where N is the number of holes in the system, $G(\boldsymbol{\rho}_n, \boldsymbol{\rho}_m)$ is the scalar plasmonic wave propagating from a hole at position $\boldsymbol{\rho}_m$ to position $\boldsymbol{\rho}_n$, and β is the scattering strength of each scatterer, which we will define shortly. The scalar plasmonic wave is defined as [20]

$$G(\boldsymbol{\rho}_n, \boldsymbol{\rho}_m) := \frac{\mathrm{i}}{4} H_0^{(1)}\left(k_{\mathrm{sp}}\left|\boldsymbol{\rho}_n - \boldsymbol{\rho}_m\right|\right), \tag{2.10}$$

where $H_0^{(1)}$ is the zeroth-order Hankel function of the first kind, k_{sp} is the surface plasmon wavenumber,

$$k_{\mathrm{sp}} = k_0\sqrt{\frac{\epsilon_0\epsilon_m}{\epsilon_0 + \epsilon_m}}, \tag{2.11}$$

ϵ_m is the dielectric constant of the metal, ϵ_0 is the dielectric constant of free space, and $k_0 = 2\pi/\lambda_0$. Equation (2.10) is the Green's function for a point source in 2D space [21, pp. 679]. Combining Eqs. (2.8) and (2.9), the output mode is

$$\boxed{\psi_{\mathbf{k}}(\boldsymbol{\rho}_n) = \alpha\phi_{\mathbf{k}}(\boldsymbol{\rho}_n) + \beta \sum_{m=1,m\neq n}^{N} G(\boldsymbol{\rho}_n, \boldsymbol{\rho}_m)\, \psi_{\mathbf{k}}(\boldsymbol{\rho}_m)\,.} \tag{2.12}$$

Equation (2.12) can be converted to matrix form and solved to obtain the output modes for all holes. Before describing that, let us first give the derivation for the expression for β.

The derivation of β relies on approximating our cylindrical holes in a metal plate as spherical cavities in a metal background and using the method of Ref. [22, section.

5.2]. Consider the electric field radiated by a monochromatic electric dipole with radial frequency ω_0. In the far field, its spatial dependence is expressed as

$$\mathbf{E}_s = \frac{k_0^2}{4\pi\epsilon_m}\left(\hat{\mathbf{r}} \times \mathbf{p}\right) \times \hat{\mathbf{r}} \frac{\exp(\mathrm{i}k_0 r)}{r}, \tag{2.13}$$

where $k_0 = 2\pi/\lambda_0$ is the wavenumber, ϵ_m is the dielectric constant of the background medium, $\mathbf{p}$ is the electric dipole moment, and $\hat{\mathbf{r}}$ is a unit vector in the direction of $\mathbf{r}$, which is the position vector of the point of observation as measured from the dipole, with $r = |\mathbf{r}|$. To model the scattering of the sphere, the dipole moment is set to

$$\mathbf{p} = \epsilon_m \alpha_p E_0 \exp(\mathrm{i}\omega_0 t)\,, \tag{2.14}$$

where E_0 is the amplitude of the incident field and α_p is the polarizability of the sphere. Taking the sphere to be vacuum, the polarizability is given by [23, section 4.4]

$$\alpha_p := 4\pi a^3 \frac{1 - \epsilon_m/\epsilon_0}{1 + 2\epsilon_m/\epsilon_0}, \tag{2.15}$$

where a is the radius of the sphere. Plugging the Eq. (2.14) into Eq. (2.13) and rearranging terms, we have

$$\mathbf{E}_s = E_0 \mathbf{X} \frac{\exp(\mathrm{i}k_0 r)}{r}, \tag{2.16}$$

where

$$\mathbf{X} := \frac{\mathrm{i}k_0^3}{4\pi} \alpha_p \left(\hat{\mathbf{r}} \times \hat{\mathbf{p}}\right) \times \hat{\mathbf{r}} \tag{2.17}$$

is the unitless vector scattering amplitude. In the model of Ref. [17], the plasmon scattering parameter is treated as a scalar analog to the vector $\mathbf{X}$, with $|\beta| \approx |\mathbf{X}|$. Therefore, we set

$$\beta \approx |\mathbf{X}| = \frac{k_0^3}{4\pi}\left|\alpha_p\right|, \tag{2.18}$$

and our final expression for β is

$$\boxed{\beta \approx \left(\frac{2\pi a}{\lambda_0}\right)^3 \left|\frac{1-\epsilon_m/\epsilon_0}{1+2\epsilon_m/\epsilon_0}\right|.} \tag{2.19}$$

Note that a now stands for the radius of our cylindrical holes, rather than the radius of a spherical cavity. Note also that the assumption of spherical cavities in the derivation means we are also assuming that our metal plate is about as thick as the diameter of our holes.

Given Eq. (2.12), we now have a system of N equations with N unknowns. This is a Foldy-Lax system of equations [24, 25] which can be solved by converting Eq. (2.12) to matrix form, as follows. First we express $\phi_{\mathbf{k}}(\boldsymbol{\rho}_{\mathbf{k}})$ and $\psi_{\mathbf{k}}(\boldsymbol{\rho}_{\mathbf{k}})$ as the column vectors

$$\boldsymbol{U}^{(\mathbf{0})} := \left[\phi_{\mathbf{k}}(\boldsymbol{\rho}_1),\ \phi_{\mathbf{k}}(\boldsymbol{\rho}_2),\ \ldots,\ \phi_{\mathbf{k}}(\boldsymbol{\rho}_N)\right]^{\mathrm{T}}, \tag{2.20}$$

$$\boldsymbol{U} := \left[\psi_{\mathbf{k}}(\boldsymbol{\rho}_1),\ \psi_{\mathbf{k}}(\boldsymbol{\rho}_2),\ \ldots,\ \psi_{\mathbf{k}}(\boldsymbol{\rho}_N)\right]^{\mathrm{T}}, \tag{2.21}$$

where the superscript T denotes matrix transposition. Next, we express $G(\boldsymbol{\rho}_n, \boldsymbol{\rho}_m)$ as an $N \times N$ matrix $\boldsymbol{G}$, whose elements are defined by Eq. (2.10), except that the diagonal elements are set to zero since the plasmon waves from holes do not self-interact. That is,

$$\boldsymbol{G} := \begin{bmatrix} 0 & G(\boldsymbol{\rho}_1, \boldsymbol{\rho}_2) & \ldots & G(\boldsymbol{\rho}_1, \boldsymbol{\rho}_N) \\ G(\boldsymbol{\rho}_2, \boldsymbol{\rho}_1) & 0 & \ldots & G(\boldsymbol{\rho}_2, \boldsymbol{\rho}_N) \\ \vdots & \vdots & \ddots & \vdots \\ G(\boldsymbol{\rho}_N, \boldsymbol{\rho}_1) & G(\boldsymbol{\rho}_N, \boldsymbol{\rho}_2) & \ldots & 0 \end{bmatrix}. \tag{2.22}$$

With these definitions, we note that the summation in Eq. (2.12) will be replaced by the matrix multiplication $\boldsymbol{GU}$. Therefore, Eq. (2.12) in matrix form is written as

$$\boldsymbol{U} = \alpha \boldsymbol{U}^{(\mathbf{0})} + \beta \boldsymbol{GU}. \tag{2.23}$$

We can now use Eq. (2.23) to solve for $\boldsymbol{U}$ and thus solve for $\psi_{\mathbf{k}}(\boldsymbol{\rho}_{\mathbf{k}})$. We do this by first moving the product $\beta\boldsymbol{G}\boldsymbol{U}$ to the left-hand side of the equation,

$$\boldsymbol{U} - \beta\boldsymbol{G}\boldsymbol{U} = \alpha\boldsymbol{U}^{(\mathbf{0})}. \tag{2.24}$$

Noting that $\boldsymbol{U} = \boldsymbol{I}\boldsymbol{U}$, where $\boldsymbol{I}$ is the identity matrix, we have

$$\left[\boldsymbol{I} - \beta\boldsymbol{G}\right]\boldsymbol{U} = \alpha\boldsymbol{U}^{(\mathbf{0})}. \tag{2.25}$$

Finally, by inverting the term in brackets, we have solved for $\boldsymbol{U}$:

$$\boxed{\boldsymbol{U} = \alpha\left[\boldsymbol{I} - \beta\boldsymbol{G}\right]^{-1}\boldsymbol{U}^{(\mathbf{0})}.} \tag{2.26}$$

Evaluating Eq. (2.26) gives the output mode $\psi_{\mathbf{k}}(\boldsymbol{\rho})$ at each hole.

Now that we have Eq. (2.26), this is a good time to show that the exact value of α does not matter when calculating the spectral degree of coherence, as long as it is nonzero. Recall that the definition of the spectral degree of coherence is

$$\mu(\boldsymbol{\rho}_n, \boldsymbol{\rho}_m) := \frac{W(\boldsymbol{\rho}_n, \boldsymbol{\rho}_m)}{\sqrt{W(\boldsymbol{\rho}_n, \boldsymbol{\rho}_n)\, W(\boldsymbol{\rho}_m, \boldsymbol{\rho}_m)}} \tag{2.27}$$

where we have used $S(\boldsymbol{\rho}) = W(\boldsymbol{\rho}, \boldsymbol{\rho})$. Let us plug Eq. (2.7) into this definition:

$$\mu(\boldsymbol{\rho}_n, \boldsymbol{\rho}_m) = \frac{\iint_{-\infty}^{\infty} \tilde{\mu}_0(\mathbf{k})\, \psi_{\mathbf{k}}^*(\boldsymbol{\rho}_n)\, \psi_{\mathbf{k}}(\boldsymbol{\rho}_m)\, \mathrm{d}^2\mathbf{k}}{\left(\iint_{-\infty}^{\infty} \tilde{\mu}_0(\mathbf{k})\, \psi_{\mathbf{k}}^*(\boldsymbol{\rho}_n)\, \psi_{\mathbf{k}}(\boldsymbol{\rho}_n)\, \mathrm{d}^2\mathbf{k}\right)^{1/2} \left(\iint_{-\infty}^{\infty} \tilde{\mu}_0(\mathbf{k})\, \psi_{\mathbf{k}}^*(\boldsymbol{\rho}_m)\, \psi_{\mathbf{k}}(\boldsymbol{\rho}_m)\, \mathrm{d}^2\mathbf{k}\right)^{1/2}} \tag{2.28}$$

Now, recall from the definition of the vector $\boldsymbol{U}$ in Eq. (2.21) that each element of $\boldsymbol{U}$, denoted U_j, is the corresponding output mode $\psi_{\mathbf{k}}(\boldsymbol{\rho}_j)$; that is, $U_j = \psi_{\mathbf{k}}(\boldsymbol{\rho}_j)$. Let us define a new vector $\boldsymbol{A}$,

$$\boldsymbol{A} := \left[\boldsymbol{I} - \beta\boldsymbol{G}\right]^{-1}\boldsymbol{U}^{(\mathbf{0})}, \tag{2.29}$$

which means

$$\boldsymbol{U} = \alpha \boldsymbol{A}. \tag{2.30}$$

Since $U_j = \psi_{\mathbf{k}}\left(\boldsymbol{\rho}_j\right)$ and $U_j = \alpha A_j$, where A_j is the j^{th} element of $\boldsymbol{A}$, then

$$\psi_{\mathbf{k}}\left(\boldsymbol{\rho}_j\right) = \alpha A_j. \tag{2.31}$$

Plugging Eq. (2.31) into Eq. (2.28), we have

$$\mu(\boldsymbol{\rho}_n, \boldsymbol{\rho}_m) = \frac{\iint_{-\infty}^{\infty} \tilde{\mu}_0(\mathbf{k})\,(\alpha A_n)^*\,(\alpha A_m)\,\mathrm{d}^2\mathbf{k}}{\left(\iint_{-\infty}^{\infty} \tilde{\mu}_0(\mathbf{k})\,(\alpha A_n)^*\,(\alpha A_n)\,\mathrm{d}^2\mathbf{k}\right)^{1/2}\left(\iint_{-\infty}^{\infty} \tilde{\mu}_0(\mathbf{k})\,(\alpha A_m)^*\,(\alpha A_m)\,\mathrm{d}^2\mathbf{k}\right)^{1/2}}, \tag{2.32}$$

and we can see that all of the α will cancel out. So the specific value of α does not matter, as long as it is nonzero, when calculating the spectral degree of coherence. In our code we set it to 0.5 as a moderate value, but again, it doesn't matter for calculation of the spectral degree of coherence. (It would matter for calculating the absolute throughput, however.)

2.2.3 Our additions to the model

The portions of our model that we have described so far originate in Ref. [17]. In this work, we have added two components to the model.

Our first addition is that we are considering a a range of input wavelengths rather than just one. The work in Ref. [17] used a fixed wavelength and varied the hole spacing. This is not easy to do in experiments, as it would require preparing a large number of hole arrays to use with light of a single wavelength. An experimenter would more naturally want to prepare a single array and vary the wavelength of their light source. As such, we require a wavelength-dependent model for the dielectric constant ϵ_m. We use the critical points model described in Refs. [26, 27], which describes the metal's dielectric function as

$$\begin{aligned}\epsilon_m(\lambda_0) =& \epsilon_\infty - \frac{1}{\lambda_p^2(1/\lambda_0^2 + \mathrm{i}/\gamma_p\lambda_0)} \\ &+ \sum_{n=1}^{2} \frac{A_n}{\lambda_n}\left[\frac{\mathrm{e}^{\mathrm{i}\phi_n}}{(1/\lambda_n - 1/\lambda_0 - \mathrm{i}/\gamma_n)} + \frac{\mathrm{e}^{-\mathrm{i}\phi_n}}{(1/\lambda_n + 1/\lambda_0 + \mathrm{i}/\gamma_n)}\right],\end{aligned} \tag{2.33}$$

where λ_p is the plasma wavelength, λ_n are the interband transition wavelengths, γ_p and γ_n are damping terms, and A_n is an amplitude. Values used in Eq. (2.33) for gold can be found in Table 2.1.

Table 2.1: Values for gold for Eq. (2.33).

Parameter	Value [27]
ϵ_∞	1.54
λ_p	143 nm
γ_p	14 500 nm
A_1	1.27
ϕ_1	$-\pi/4$ rad
λ_1	470 nm
γ_1	1900 nm
A_2	1.1
ϕ_2	$-\pi/4$ rad
λ_2	325 nm
γ_2	1060 nm

This is a good time to discuss how we selected what wavelength range to use for λ_0. Our model relies on several assumptions; one of which is a multi-wavelength spacing between holes. This means we require at least that $d > \lambda_\mathrm{sp}$, where d is the smallest spacing between holes in the array and λ_sp is the surface plasmon wavelength,

$$\lambda_\mathrm{sp} := \frac{2\pi}{\mathcal{R}\{k_\mathrm{sp}\}}, \tag{2.34}$$

where $\mathcal{R}\{\}$ denotes the real part. Another assumption is that the hole diameter is subwavelength; so we require that $2a < \lambda_0$. In addition to these assumptions, we add the constraint that holes not be so far apart that plasmon effects are negligibly small. That is, we require $d < L_{\text{sp}}$, where L_{sp} is the plasmon propagation distance, defined as

$$L_{\text{sp}} := \frac{1}{2\mathcal{I}\{k_{\text{sp}}\}}, \tag{2.35}$$

where $\mathcal{I}\{\}$ denotes the imaginary part. If plasmons travel a distance of L_{sp}, their intensity will be reduced to 1/e of the original value, so we do not want d to be greater than that, or else the plasmonic contribution to the field will be very small. With these three constraints, we chose to set $a = 200\,\text{nm}$ and to keep d near or exceeding $1000\,\text{nm}$. Based on all this, we settled on using a wavelength range $\lambda_0 = [550\,\text{nm}, 850\,\text{nm}]$, which is mostly in the visible.

In Fig. 2.2 we compare the critical points model with experimental data [28]; the agreement is excellent, particularly over the wavelength range we have selected for our simulations. We also note that for a flat surface to support surface plasmon waves we require [29, pp. 5]

$$\mathcal{R}\{\epsilon_m\} < 0 \tag{2.36a}$$

$$|\mathcal{R}\{\epsilon_m\}| > \epsilon_0 \tag{2.36b}$$

$$\mathcal{I}\{\epsilon_m\} < |\mathcal{R}\{\epsilon_m\}|\,. \tag{2.36c}$$

Figure 2.2 shows that these conditions are met over our chosen wavelength range.

Our second addition to the model of Ref. [17] addresses one of the challenges of analyzing the results of this model. Namely, the spectral degree of coherence is defined in terms of a pair of points, call them $\boldsymbol{\rho}_n$ and $\boldsymbol{\rho}_m$. For N holes, this results in many possible combinations of pairs, scaling roughly $\mathcal{O}(N^2)$, and thus many values of $\mu(\boldsymbol{\rho}_n, \boldsymbol{\rho}_m)$. Specifically, for N holes there are N^2 pairs. Now, two of the properties of the spectral degree of coherence are that $\mu(\boldsymbol{\rho}_n, \boldsymbol{\rho}_m) = \mu(\boldsymbol{\rho}_m, \boldsymbol{\rho}_n)$ and that $\mu(\boldsymbol{\rho}_n, \boldsymbol{\rho}_n) = 1$; these reduce the number of pairs we need to consider. The end

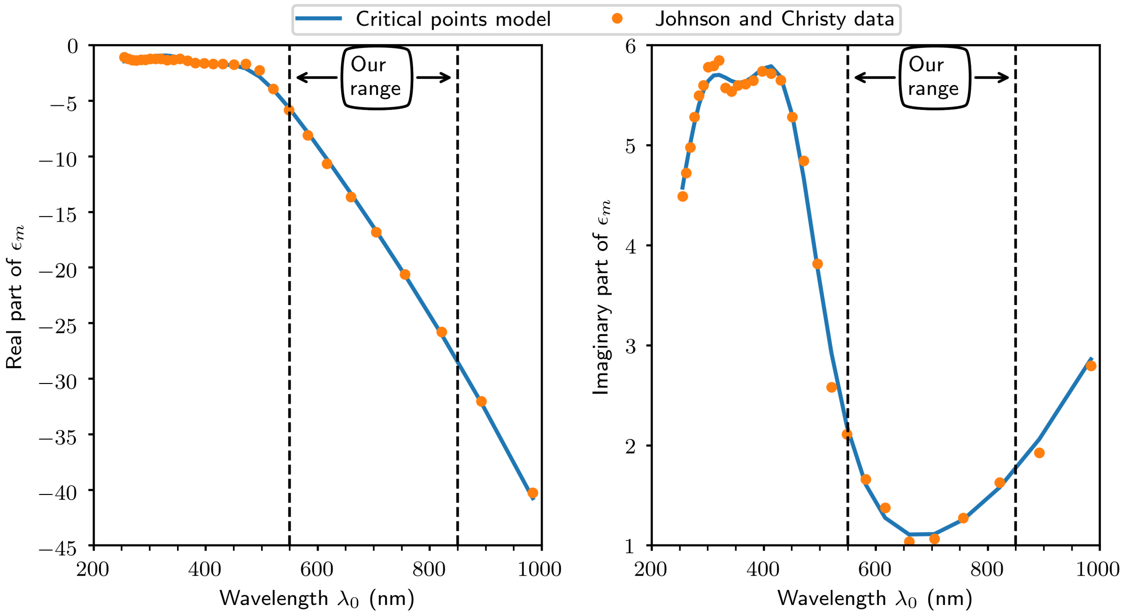

Figure 2.2: Comparing the critical points model of Eq. (2.33) with experimental data by Johnson and Christy [28]. The dashed lines indicate the wavelength range we use for most of the calculations in this chapter.

result is that for N holes, the number of pairs that would need to be considered is $\Delta(N-1)$, where $\Delta(n)$ is the n^{th} triangular number, $\Delta(n) = \sum_{j=1}^{n} j$, which quickly becomes very large. For example, $\Delta(9) = 45$, $\Delta(16) = 136$, and $\Delta(25) = 325$. This was a challenge in the work in Ref. [17]; only a few combinations could be plotted, which made it difficult to discern overall trends in the coherence. Here, we simplify the analysis by considering the average coherence over all hole pairs. For the output coherence, we denote this average as M_f and define it as

$$M_f := \frac{1}{N}\sum_{n=1}^{N}\frac{1}{N-1}\sum_{m=1,m\neq n}^{N}\left|\mu_f(\boldsymbol{\rho}_n,\boldsymbol{\rho}_m)\right|. \tag{2.37}$$

We define the input average M_0 similarly in terms of $\mu_0(\boldsymbol{\rho}_n,\boldsymbol{\rho}_m)$. We note that we use the absolute value of $\mu_f(\boldsymbol{\rho}_n,\boldsymbol{\rho}_m)$ in Eq. (2.37) to avoid low correlations caused by different $\mu_f(\boldsymbol{\rho}_n,\boldsymbol{\rho}_m)$ being out of phase. We also note one limitation of using this value is that it will, by definition, tend to get very low as the array size increases.

Some typical results obtained using this method are shown on the right side Fig. 2.3. On the left is a depiction of the hole geometry used: a square array of holes periodic in two dimensions with period d. The data in Fig. 2.3 are somewhat chaotic, with many narrow peaks, but there are noticable trends. For example, one can note that the coherence is generally higher at the lower and higher wavelengths than it is in the middle. Particularly interesting are the sudden dips at a few wavelengths, specifically $\lambda_0 \approx 610\,\text{nm}$, $650\,\text{nm}$, and $750\,\text{nm}$. This one plot is certainly simpler to interpret than would be the coherence between the 79 800 hole pairs one would otherwise have to analyze individually!

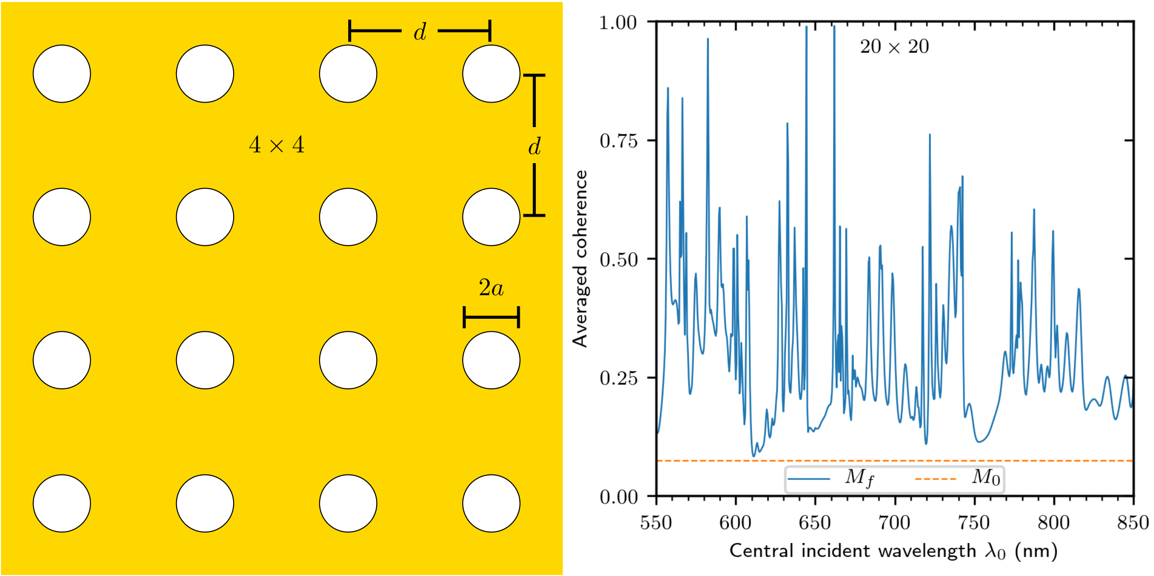

Figure 2.3: Left: geometry of a square array of holes, in this case 4×4, with period d and radius a. Right: M_f and M_0 of a 20×20 array, with $d = 1000\,\text{nm}$, $a = 200\,\text{nm}$, and $\delta = 1000\,\text{nm}$.

2.3 Optical Coherence Band Gap

Looking back at those prominent dips we mentioned in Fig. 2.3, we notice that they have similar shape to that of band gaps. Band gaps are a well-known phenomenon in optics, in which a periodic structure causes certain spectral regions of light to have a transmittance of zero. A common example is a periodic structure of evenly-spaced layers of materials with alternating refractive index, known as a *Bragg reflector*.

In this project, published in Ref. [30], we demonstrate a similar band gap phenomenon for optical coherence. This idea gains some prior plausibility if we recall from Eqs. (1.29) and (1.30) that the mutual coherence function obeys the wave equation and from Eq. (1.43) that the cross-spectral density obeys the Helmholtz equation. Since coherence propagates a wave, it is plausible to think that it might exhibit band gap behavior given a suitable periodic medium of propagation. Now, there is some difficulty in analyzing the results from periodic 2D arrays like the one shown in Fig. 2.3. Specifically, they have a large number of peaks in M_f and they have are large number of distances between holes. Therefore, for this project, we decided to use *linear* arrays of periodically spaced holes, as shown in Fig. 2.4, with holes of radius a spaced periodically with lattice constant d. This provides us a structure similar to that of a Bragg reflector. We constrain our holes to lie along the x axis, and we also take the polarization of the incident beam to lie along the x axis. So the vectors in our model now have a y component of zero, meaning that in our equations the vectors $\boldsymbol{\rho}$ and $\mathbf{k}$ can be replaced with scalars x and k, respectively. We note that this structure means that band gaps will depend on a periodicity *transverse* to the direction of propagation, as opposed to intensity band gaps which depend on periodicity parallel to the direction of propagation. That is, we have periodicity (of holes) along the x direction, while the incoming light is propagating along the z direction; for intensity band gaps, the periodicity (of refractive index layers) would also be along the z direction.

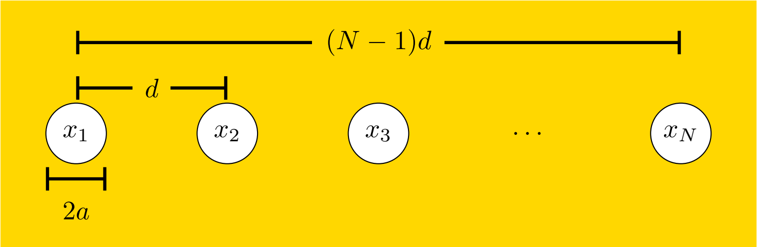

Figure 2.4: Geometry and notation of linear hole arrays.

To verify that what we have observed is indeed a band gap, we test two important properties characteristic to band gaps. The first is that they require a certain number of periodic layers in order to appear, and they remain relatively unchanged by adding more. In Fig. 2.5 we show M_f calculation results for a series of array sizes from 1×2 to 1×100. As the array size increases, the broad peak on the left, centered at about $\lambda_0 = 570\,\mathrm{nm}$, significantly narrows and slightly blueshifts. The valley on the right, centered at about $\lambda_0 = 720\,\mathrm{nm}$, flattens and then, at 1×10, becomes a peak. This peak broadens and turns into a cluster of peaks while redshifting. Between these two peaks, beginning with the 1×10 array, is a mostly flat region centered at about $\lambda_0 = 670\,\mathrm{nm}$ where M_f is almost exactly equal to M_0. This flat region is what we are identifying as a coherence band gap. We note that the band gap region takes a few holes to appear, and after a certain number (10 in this case) it is relatively unchanged as more holes are added. This behavior matches the band gap behavior we were examining.

The second important property of band gaps we wanted to verify is that they only appear in a periodic medium. We test our band gap's dependence on periodicity by randomly moving the holes within the array. If the flat region is truly a band gap, destroying the periodicity should destroy the band-gap-like behavior. So, let us consider the n^{th} hole in the array, originally at location x_n. We assign it a randomized coordinate X_n via

$$X_n := x_n + Z_n \sigma, \tag{2.38}$$

where Z_n is a number drawn from the standard normal distribution and σ is a chosen standard deviation. After obtaining a configuration of randomized coordinates for all holes, we checked that all hole pairs satisfied the condition

$$|x_n - x_m| > 2a, \tag{2.39}$$

to prevent the holes from overlapping. If any pair in the configuration failed that test,

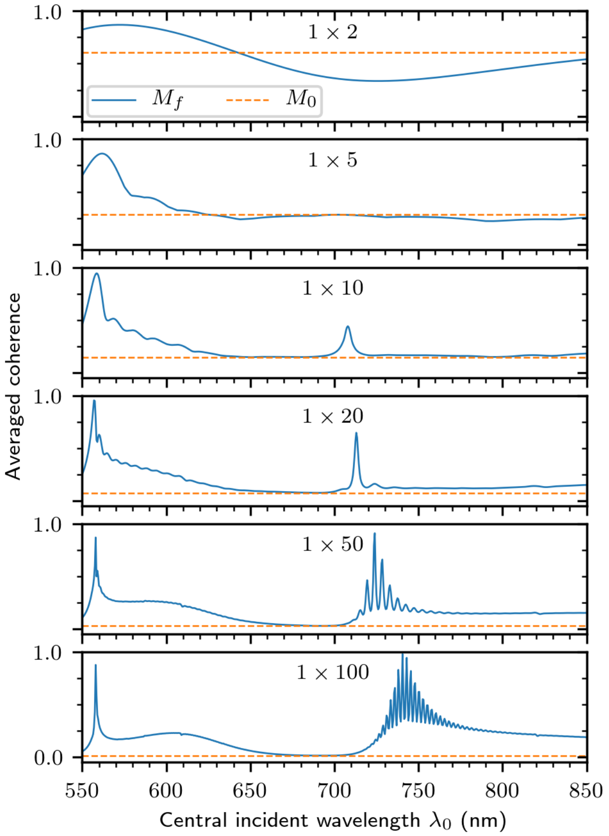

Figure 2.5: Averaged output and input coherence for increasingly large hole arrays, with $d = 1000\,\text{nm}$ and $\delta = 1000\,\text{nm}$. There is a flat valley between $\lambda_0 \approx 660\,\text{nm}$ and $\lambda_0 \approx 700\,\text{nm}$; this is the coherence band gap we will be examining. It is noticeable with as few as 10 holes in the array. *Figure and caption from Ref. [30], ©American Physical Society, used with permission.*

that configuration was discarded and a new one obtained until Eq. (2.39) was satisfied for all pairs. In Fig. 2.6, we show results for three different standard deviations, plus the unrandomized case ($\sigma = 0$). The results in Fig. 2.6 are the averaged M_f from 100 configurations for each nonzero σ. We can see that even at small σ, the band gap nature is already destroyed. Increasing σ further makes the coherence more of a large, broad peak with no flatness. Figures 2.5 and 2.6 together provide strong evidence that the flat region we are describing is a band gap.

We now turn to investigating the origin of the band gap structure. We first simplify our discussion by only considering a 1×2 array, as shown in Fig. 2.7. Light incident on the hole at x_1 will create a SPP propagating to x_2, denoted G_1. Upon hitting the

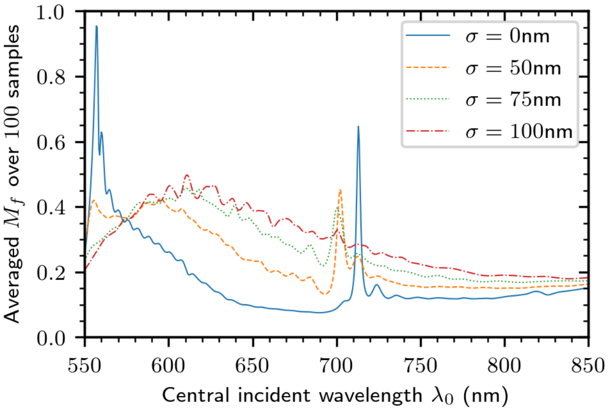

Figure 2.6: Coherence of randomized 1×20 hole arrays, with $d = 1000\,\text{nm}$ and $\delta = 1000\,\text{nm}$. For each value of σ (except zero), 100 randomizations were done, and their averaged coherence M_f calculated. These M_f were then averaged together to produce this figure. *Figure and caption from Ref. [30], ©American Physical Society, used with permission.*

hole at x_2, part of G_1 will reflect back to x_1; we denote this wave G_{1r}. Additionally, there will be a SPP wave propagating from x_2 to x_1, denoted G_2. There would also be a reflected G_{2r} mode, and since the pair of holes acts similarly to a Fabry-Perot cavity, there can be a large number of other reflections, but these three are sufficient to consider for our explanation. Now, at x_1 the G_2 mode will have acquired a phase of $k_{\text{sp}}d$. Thus, when

$$\mathcal{R}\{k_{\text{sp}}\}\, d = \nu_2 \pi \tag{2.40}$$

and ν_2 is an even (odd) integer, G_2 will constructively (destructively) interfere with G_1, which has a phase of 0 at x_1. We call these "ν_2 modes." Similarly, G_{1r} will have acquired a phase of $k_{\text{sp}}2d$ at x_1, and when

$$\mathcal{R}\{k_{\text{sp}}\}\, 2d = \nu_{1r} \pi, \tag{2.41}$$

where ν_{1r} is an even (odd) integer, G_{1r} will have constructive (destructive) interference with G_1 at x_1. We call these "ν_{1r} modes." Combining Eqs. (2.40) and (2.41), we can see that these two conditions will coincide when

$$\nu_{1r} = 2\nu_2. \tag{2.42}$$

This means that ν_2 modes will only ever coincide with constructively interfering ν_{1r} modes. Now, in a larger array, every hole will have the same relationship characterized by Eqs. (2.40) to (2.42) with its neighboring holes, all at the same wavelength. This means that the entire array will have identical ν_2 modes and identical ν_{1r} modes.

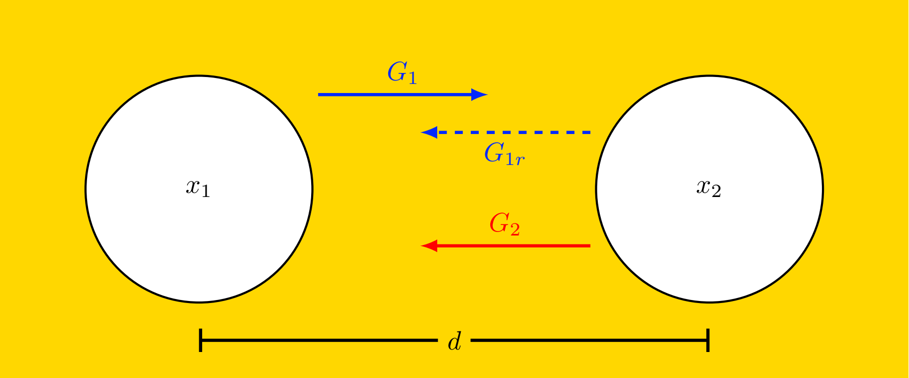

Figure 2.7: Notation of plasmon modes used to derive the band gap condition.

To show this behavior, in Fig. 2.8 we show M_f of a 1×50 array as a function of $\mathcal{R}\{k_{\text{sp}}\}\, d$. In order to show many cycles, we neglect the wavelength-dependent behavior of β and ϵ_m, setting them to constant values of $\beta = 5$ and $\epsilon_m \approx -10.6488 + \text{i}1.3734$. These values were used in Ref. [17] for a wavelength of $\lambda_0 \approx 600\,\text{nm}$. We can see that there are band gaps around $\nu_2 = 3$ and $\nu_2 = 5$, while there are large peaks near $\nu_2 = 4$ and $\nu_2 = 6$. This is consistent with with our model in Eqs. (2.40) to (2.42). At even ν_2, both G_2 and G_{1r} are in phase with G_1, which increases coherence. In the region between odd ν_2 and odd ν_{1r}, G_2 and G_{1r} will mostly destructively interfere with G_1. This minimizes the overall plasmonic contribution to the field at the holes, which reduces the output coherence to near that of the input coherence. This destructive interference is the cause of the coherence band gaps we have observed. We note that in our testing, we have not *always* observed band gaps appear near odd ν_2, but they have appeared *only* near odd ν_2.

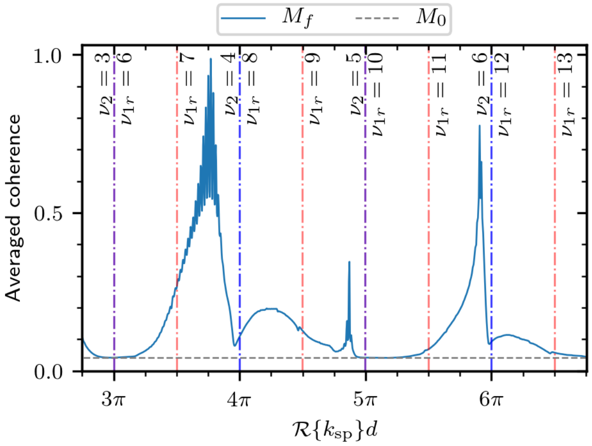

Figure 2.8: M_f for a 1×50 array, with $d = 1200\,\mathrm{nm}$ and $\delta = 1500\,\mathrm{nm}$. Here, the dielectric constant and scattering parameter were set to constant values of $\epsilon_m \approx -10.6488 + \mathrm{i}1.3734$ and $\beta = 5$, independent of wavelength. Red lines are wavelengths of destructive interference, blue of constructive interference, and purple is where the two coincide. *Figure and caption from Ref. [30], ©American Physical Society, used with permission.*

We note that there is an alternative interpretation of the band gap results in this chapter. Namely, surface plasmons scattering from holes can be seen as a classical form of *quantum superradiance and subradiance*, first examined by Dicke [31]. These phenomena have to do with how a collection of atoms (or ions or quantum dots, etc.) interact with incident light and with the field due to stimulated emission from neighboring atoms. Specifically, superradiance is when the atoms radiate coherently due to constructive interaction with the neighboring atoms; subradiance is when the radiation is suppressed by destructive interference with the neighboring atoms. The analogy between this quantum phenomenon and our plasmon multiple-scattering system is readily apparent. The holes take the place of atoms, and generated SPPs take the place of stimulated emission. This analogous relationship was first noted by Ropers et al. [32]. This interpretation can be confirmed by considering the coherence of the arrays in Fig. 2.5. The peak at around $\lambda_0 = 560\,\mathrm{nm}$ is a superradiant peak, since at that wavelength d is about twice the plasmon wavelength ($\lambda_{\mathrm{sp}} \approx 519\,\mathrm{nm}$), which corresponds to constructive interference. The valley that turns into a peak at about $\lambda_0 = 710\,\mathrm{nm}$ is a subradiant peak because then d is about 3/2 the plasmon wavelength ($\lambda_{\mathrm{sp}} \approx 689\,\mathrm{nm}$), which corresponds to destructive interference.

2.4 Other Results

2.4.1 Fano shape

Looking at these M_f curves, particularly in Figs. 2.5 and 2.8, we can observe that some of them have what appears to be a Fano resonance shape. To test this, we decided to try to fit a Fano-style curve against some of the resonances in M_f. These fits are shown in Fig. 2.9. Each of the resonances corresponds to a value of ν_2, so there is a fit for each value of ν_2 shown. Each fit, denoted y_{ν_2}, is of the form of the Fano resonance equation [33],

$$y_{\nu_2} = \frac{\left(F_{\nu_2}\gamma_{\nu_2} + \lambda_0 - \lambda_0^{(\nu_2)}\right)^2}{\left(\lambda_0 - \lambda_0^{(\nu_2)}\right)^2 + \gamma_{\nu_2}^2}, \tag{2.43}$$

where $\lambda_0^{(\nu_2)}$ is the wavelength at which the resonance occurs, γ_{ν_2} is the width, and F_{ν_2} is the Fano parameter, which describes the degree of asymmetry. Each y_{ν_2} was then normalized to the value of its corresponding peak. Table 2.2 shows the values used in Eq. (2.43) to produce the fits in Fig. 2.9, which were determined by trial and error. Figure 2.9 shows fits for resonances with ν_2 from 3 to 7 for a 1×20 array with $\delta = 1000\,\mathrm{nm}$. We can see that the Fano lines can make a good fit for these coherence resonances, which gives evidence that there might be some underlying Fano-style mechanism.

There are a couple of problems with this idea. First, while the Fano curve fits individual peaks well, the sum of all the fits does not. Call the sum in Fig. 2.9(a)

Table 2.2: Values of parameters used in Eq. (2.43) to produce Fig. 2.9.

	F_{ν_2}	γ_{ν_2} (nm)	$\lambda_0^{(\nu_2)}$ (nm)
$\nu_2 = 7$	2.5	2.0	621.0
$\nu_2 = 6$	2.5	2.0	709.5
$\nu_2 = 5$	3.5	3.5	840.5
$\nu_2 = 4$	2.5	8.0	642.0
$\nu_2 = 3$	8.5	4.5	829.0

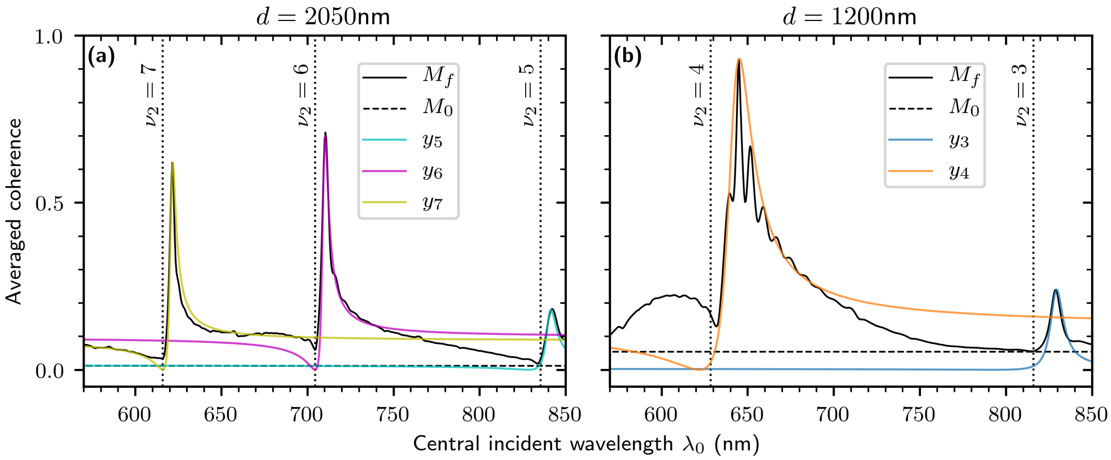

Figure 2.9: M_f, M_0, and Fano fits y_{ν_2} for a 1×20 array, with $a = 200$ nm and $\delta = 1000$ nm.

$Y = y_5 + y_6 + y_7$. In our testing, Y would fit well on the $\nu_2 = 6$ and $\nu_2 = 7$ peaks, but not the $\nu_2 = 5$ peak. This is because M_f is linearly decreasing $\nu_2 = 6$ to $\nu_2 = 5$, where Y is flat. This prevented Y from overlapping M_f on that peak. The second big problem can bee seen in Fig. 2.9(b), which shows M_f for $\nu_2 = 3$ and 4, with $d = 1200$ nm. The Fano fit is not as good as in Fig. 2.9(a). The $\nu_2 = 3$ M_f is mostly symmetrical, so while y_3 fits the peak, it's not very insightful to put a Fano fit on it. (This same issue applies to the $\nu_2 = 5$ peak in Fig. 2.9(a), so this may just be a high-wavelength effect.) The real problem is at $\nu_2 = 4$. This resonance has many narrow peaks on a broad superstructure and is severely rounded to the left of the dip. This is evidence against the Fano idea. However, the model works so well for the higher ν_2 resonances that we think this would be worth some more investigation in the future. In fact, it almost seems like the Fano model works better as ν_2 increases.

2.4.2 Aid for future designs

In this subsection, we give some miscellaneous results that may be useful for future design of these coherence conversion systems with linear hole arrays.

In Fig. 2.10 we show the location of ν_2 modes as a function of λ_0 and d. In addition, we also indicate two regions where our model breaks down. In the green region, d is less than λ_{sp}, so our assumption of multiple wavelengths between holes is invalid in that region. In the red region, d is greater than L_{sp}, the plasmon propagation distance, so plasmonic effects become greatly diminished.

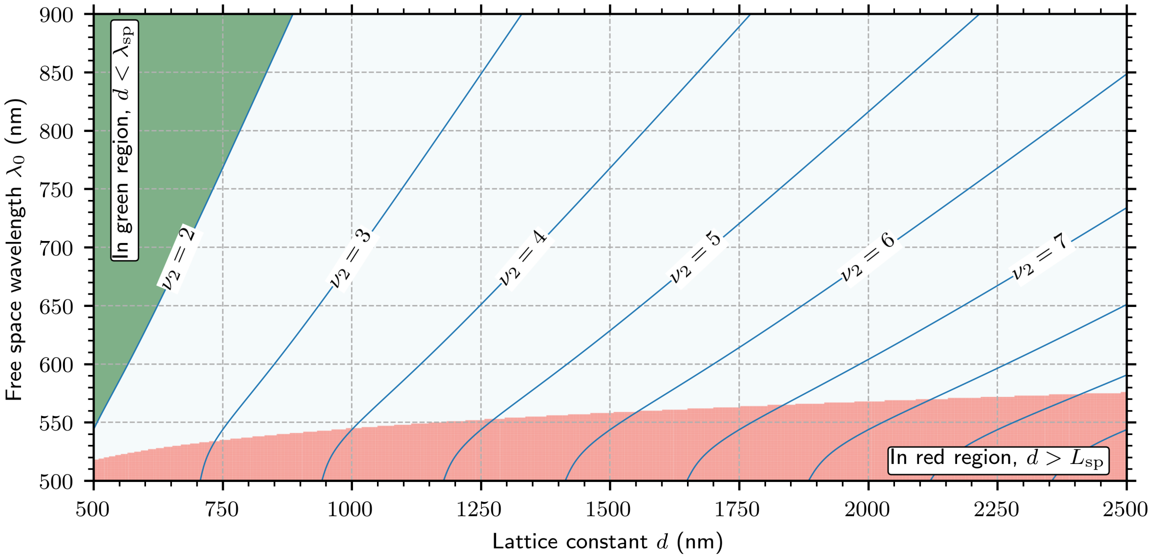

Figure 2.10: Location of ν_2 modes as a function of λ_0 and d. Our model breaks down in the shaded regions for the reasons specified in the text boxes.

In Fig. 2.11, we show the averaged coherence of a 1 × 20 array, showing how M_f changes as a function of d and a, along with λ_0. The ν_2 modes are also indicated. Figure 2.11(a) shows M_f as a function of d and λ_0. We can see that there is a high coherence "band" next to the ν_2 modes. In Fig. 2.11(b), we show M_f as a function of a and λ_0. The behavior here is intriguing. The coherence is near zero for small a. At value of a over about 170 nm, there are are "islands" of high coherence. These islands follow curves that seem to asymptotically approach the $\nu_2 = 3$ line in opposite directions. Examining the cause of this behavior, and what role a plays analogous to more well-known examples of band theory, could be an interesting new project.

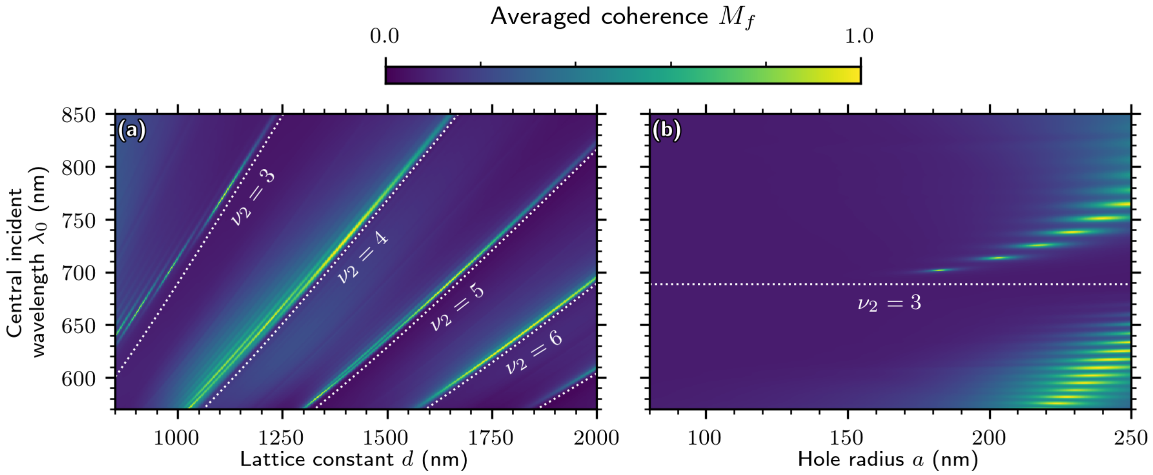

Figure 2.11: Averaged coherence of a 1×20 array, with $\delta = 1000\,\text{nm}$. (a) M_f as a function of d and λ_0, with $a = 200\,\text{nm}$. (b) M_f as a function of a and λ_0, with $d = 1000\,\text{nm}$.

2.5 Conclusion

In this chapter, we have theoretically demonstrated the existance of an optical coherence band gap using simulations of a gold plate with a linear array of subwavelength holes. We described the model that we used and gave examples and explanations to support our conclusions. We also discussed the possibility of a Fano-style mechanism contributing to the coherence resonance shape and provided plots to aid future designers using or simulating this device.

CHAPTER 3: COHERENCE CONVERSION AND TRANSMITTANCE WITH SQUARE ARRAYS

3.1 Introduction

In the previous chapter, we used simulations of linear arrays of holes in a gold sheet to show the existence of an optical coherence band gap. However, a real coherence conversion deviece would more likely use a 2D square arrays of holes like the one that was shown in Fig. 2.3. In this chapter, we will study these sorts of arrays using the same simulation method as in Chapter 2. This work had two aims. The first was to characterize the behavior of the array in response to changing parameters of the array, specifically, the number of holes, the hole spacing (or lattice constant) d, and the hole radius a. The second was to examine whether high intensity transmittance correlates with high coherence. This second aim is important because our system not only needs to change the degree of coherence, but also to have enough throughput to be measurable at the detector.

3.2 Response of Coherence and Transmittance to Changing Parameters

3.2.1 Simulation details

In all our simulations, we consider the coherence and transmittance of gold plates with square hole arrays, as shown in Fig. 3.1, over a wavelength range from $\lambda_0 = 550\,\mathrm{nm}$ to $\lambda_0 = 850\,\mathrm{nm}$. In the previous chapter, we introduced an averaged coherence M_f in order to analyze the coherence results. However, in this work we are interested in comparing the transmittance through individual holes to the spectral degree of coherence between holes, so we will generally not use the averaged coherence in this chapter. We restrict our attention to three hole pairs: two holes in the center (holes

A and B in Fig. 3.1), two of the outermost corners (C and D), and two holes in the lower left corner (E and F). We take it that the behavior of these hole pairs will give a sense of the behavior of the entire array. We examine the effect of changing three array properties: the lattice constant d (see Fig. 3.1), the hole radius a, and the array size. When increasing array size, we only use configurations with an even number of holes on each side in order to have two holes in the center of the array. For all of these simulations, the transverse correlation length is $\delta = 1000\,\mathrm{nm}$.

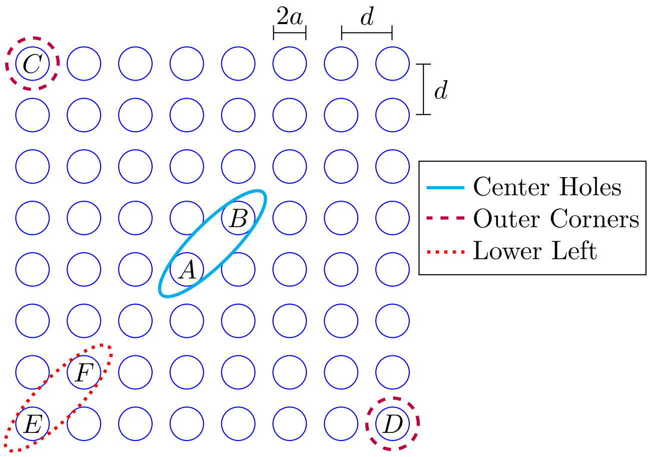

Figure 3.1: An example 8 × 8 square hole array, with the three hole pairs we're considering indicated as shown in the legend. The individual holes are also lettered.

In this work, we are interested in seeing how the coherence of a pair of holes coincide with the transmittance at those holes. This is because the coherence conversion device would need to both produce the desired spectral degree of coherence and also have enough intensity to be measurable at the detector. The transmittance $T(\boldsymbol{\rho}_n)$ at the n^{th} hole is defined as

$$\boxed{T(\boldsymbol{\rho}_n) := \frac{S_f(\boldsymbol{\rho}_n)}{S_0},} \tag{3.1}$$

where S_0 is the input spectral density and $S_f(\boldsymbol{\rho}_n)$ is the output spectral density,

$$S_f(\boldsymbol{\rho}_n) = \iint_{-\infty}^{\infty} \tilde{\mu}_0(\mathbf{k}) \psi^*_{\boldsymbol{k}}(\boldsymbol{\rho}_n)\, \psi_{\boldsymbol{k}}(\boldsymbol{\rho}_n)\, \mathrm{d}^2\mathbf{k}. \tag{3.2}$$

Recall that we defined S_0 to be a constant across the plane of the array. Without loss of generality, we assign $S_0 = 1$. Recall also that the spectral density of a field, as a function of wavelength, is the field's power spectrum, so the definition of transmittance in Eq. (3.1) is the ratio of the output and input intensities. It is to be noted that the definition of $T(\boldsymbol{\rho}_n)$ depends on α, the fraction of light directly transmitted through the holes, via the output intensity in Eq. (3.2). Furthermore, α will in general depend on the wavelength of the incoming light, as transmittance will depend on the ratio a/λ_0. However, we do not think that the specific value of α at any specific wavelength will have much effect on whether there is a peak at that wavelength, as it would be a slow function of λ_0, whereas transmittance is a fast function of λ_0, as we will soon see. We therefore choose to scale out the effects of α by defining

$$T_{\mathrm{s}}(\boldsymbol{\rho}_n) := \frac{T(\boldsymbol{\rho}_n)}{\alpha^2}. \tag{3.3}$$

This is the transmittance we shall use for the remainder of the paper.

It is worth mentioning that, due to the symmetry of our system, both holes in the center hole pair (holes A and B in Fig. 3.1) will always have identical transmittance. Similarly, the outer corner holes (C and D) will always have identical transmittance. The lower left holes (E and F), however, will not. But, hole E will have the same transmittance as holes C and D (the outer corners). Because of this, we will exclude the transmittance of hole E from consideration, and whenever we discuss the transmittance of the lower left pair, we mean only the transmittance of hole F.

3.2.2 Changing array size

A realistic hole array for a coherence conversion would likely have at least dozens of holes in it. Thus, we chose to examine the response of increasingly large numbers of holes in the array in order to get a sense of the limiting behavior. We would expect the behavior of the system to converge at large arrays since, as the number of holes in the array increases,the effects of the boundary will be insignificant compared to the center. We examined hole arrays from 2 × 2 (4 holes) to 20 × 20 (400 holes). We give sample results for changing array size in Fig. 3.2, which shows the input and output spectral degrees of coherence and the transmittance for all three hole pairs for array sizes 6 × 6 (36 holes), 8 × 8 (64 holes), 12 × 12 (144 holes), and 16 × 16 (256 holes). Here $d = 1000\,\mathrm{nm}$ and $a = 200\,\mathrm{nm}$.

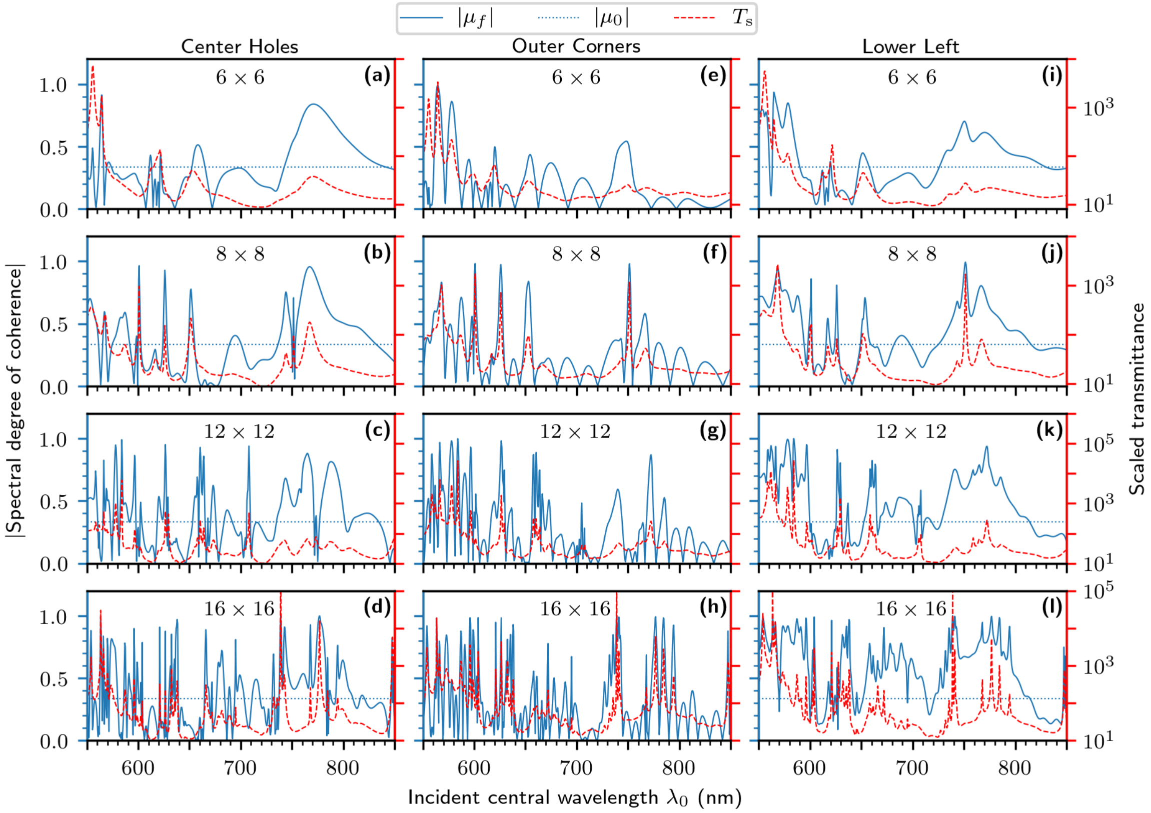

Figure 3.2: Effect of increasing array size on coherence and transmittance. Here, $a = 200\,\mathrm{nm}$ and $d = 1000\,\mathrm{nm}$. The left column, subplots (a-d), shows the coherence and transmittance for the center holes. The center column, subplots (e-h), shows the coherence and transmittance for the outer corner holes. The right column, subplots (i-l), shows the coherence and transmittance for the lower left holes. Note that all transmittances are plotted on a logarithmic scale.

Before examining the effect of array size, it is worth comparing the general output coherence behavior of the hole pairs relative to each other, as these trends will hold for all results that follow. The outer corners pair's $|\mu_f(\boldsymbol{\rho}_n, \boldsymbol{\rho}_m)|$ oscillates more rapidly than the other two pairs, generally having far more peaks and zeros than they do. The lower left pair oscillates the least strongly, rarely ever getting near zero. Instead, it tends to have narrow sub-peaks superimposed on a few broad, shallow peaks. The central hole pair's behavior is between that of the other two pairs.

Now let's consider what happens as array size increases by examining Fig. 3.2 in detail. First, it should be noted that, for all pairs, as the array size increases, the coherence and transmittance peaks generally increase in number and have decreasing width. Next, the center hole pair, Fig. 3.2(a-d), has a rather prominent coherence peak at about $\lambda_0 = 770\,\text{nm}$ and low array sizes. As the array size increases, this feature gradually splits into two, until the array size is 16×16, at which point it is now two distinct peaks with a notable gap in between where $|\mu_f(\boldsymbol{\rho}_n, \boldsymbol{\rho}_m)|$ is not much greater than $|\mu_0(\boldsymbol{\rho}_n, \boldsymbol{\rho}_m)|$. Next, the outer corners pair, Fig. 3.2(e-h). Unlike the center hole pair, this pair does not have any particular prominent peak that is consistent for different array sizes until the array size exceeds 8×8, at which point there is a small but notable peak at about $\lambda_0 = 740\,\text{nm}$. As the array size increases, this peak remains but has several taller sub-peaks come out of it, with nearby peaks remaining low. Finally, the lower left pair, Fig. 3.2(i-l), mostly follows the same pattern as the center hole pair. There is a broad peak with a center roughly between $\lambda_0 = 750\,\text{nm}$ and $\lambda_0 = 770\,\text{nm}$. As the array size increases, this peak has sub-peaks grow up out of it, with more sub-peaks forming as the array size increases. When the array size gets over 12×12, these sub-peaks separate into two clusters with a peakless gap in between. In addition, several other peaks sprout sub-peaks and either split or merge as the array size increases.

For all hole pairs, the 18×18 and 20×20 array coherence and transmittance results (not shown) do not differ significantly from those of the 16×16 array, indicating that adding additional holes will likely not cause results qualitatively much different from those presented here, although the precise location, height, and number of peaks does change.

Finally, it can be seen that the maxima of transmittance and coherence often coincide exactly, and that the transmittance spectrum usually follows the general trend of the coherence spectrum. Altogether, Fig. 3.2 suggests that it would be difficult to make a coherence conversion device work by *reducing* coherence. This is because wavelengths where $|\mu_f(\boldsymbol{\rho}_n, \boldsymbol{\rho}_m)| < |\mu_0(\boldsymbol{\rho}_n, \boldsymbol{\rho}_m)|$ tend to have lower transmittance.

3.2.3 Changing lattice constant

For tuning coherence spectra to a desired wavelength, the lattice constant d is likely to be one of the most important properties to control, since it can tune the holes to lie where surface plasmons constructively or destructively interfere, whichever is desired. We show results for changing lattice constant d in Figs. 3.3 and 3.4 for a 6×6 hole array with 200 nm hole radius.

Figure 3.3 shows the output spectral degree of coherence and the scaled transmittance for all three hole pairs for values of d from 850 nm to 2000 nm. We see that coherence peaks tend to form "bands" of high coherence or low coherence which follow a linear relationship between d and λ_0. The bands do not all have the same slope; those in the upper left corner have a steeper slope than those in the lower right corner. All of the peaks redshift, which is what would be expected, since larger hole separation would resonate with longer plasmon wavelengths, which have an almost, but not quite, linear dependence on λ_0. Additionally, it can be seen that the transmittance behavior often matches that of $|\mu_f(\boldsymbol{\rho}_n, \boldsymbol{\rho}_m)|$, as expected.

Figure 3.4 shows the input and output spectral degrees of coherence of the center hole pair for three different values of d. This is to see the relationship between transmittance and coherence more clearly. As we have seen from Figs. 3.2 and 3.3, the transmittance generally follows the same trend as the output coherence, and wavelengths where $|\mu_f(\boldsymbol{\rho}_n, \boldsymbol{\rho}_m)| < |\mu_0(\boldsymbol{\rho}_n, \boldsymbol{\rho}_m)|$ tends to have lower transmittance.

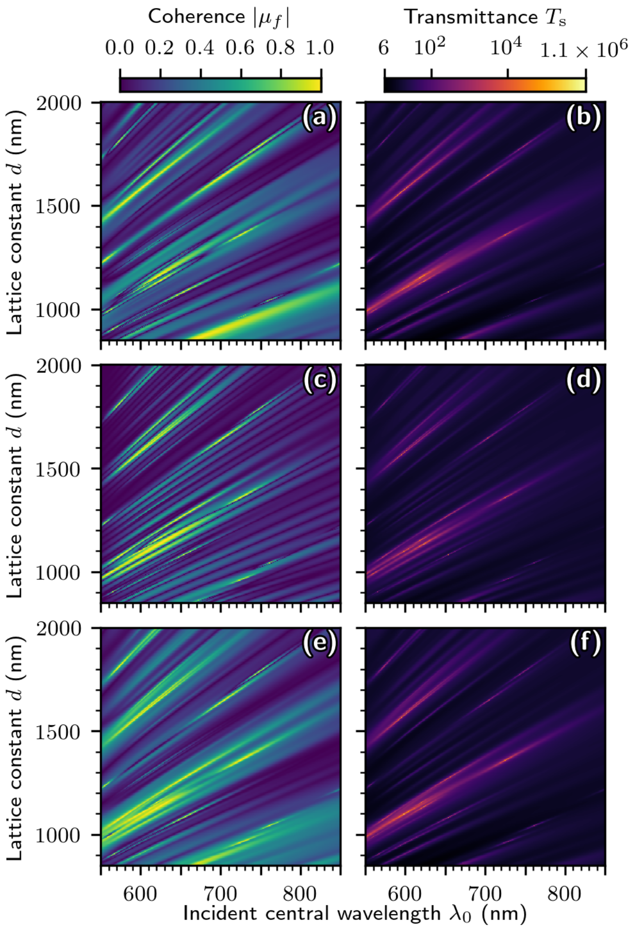

Figure 3.3: Coherence and transmittance of a 6 × 6 configuration, varying d, with $a = 200\,\text{nm}$. (a) Center holes coherence. (b) Center holes transmittance. (c) Outer corners coherence. (d) Outer corners transmittance. (e) Lower left corner coherence. (f) Lower left corner transmittance. Note that the transmittances are on a logarithmic scale.

3.2.4 Changing hole radius

The hole radius directly affects the scattering strength β of the holes, as shown in Eq. (2.19). We would thus expect that increasing a would cause the output coherence to increasingly differ from the input coherence. Figure 3.5 shows the effect of changing hole radius on coherence and transmittance for all three hole pairs for values of a ranging from 100 nm to 250 nm. It can be seen that increasing the hole radius increases the oscillatory nature of the coherence. That is, for small a, $|\mu_f(\boldsymbol{\rho}_n, \boldsymbol{\rho}_m)|$ has a value that is almost constant, with small gradual oscillations. Additionally, $|\mu_f(\boldsymbol{\rho}_n, \boldsymbol{\rho}_m)|$

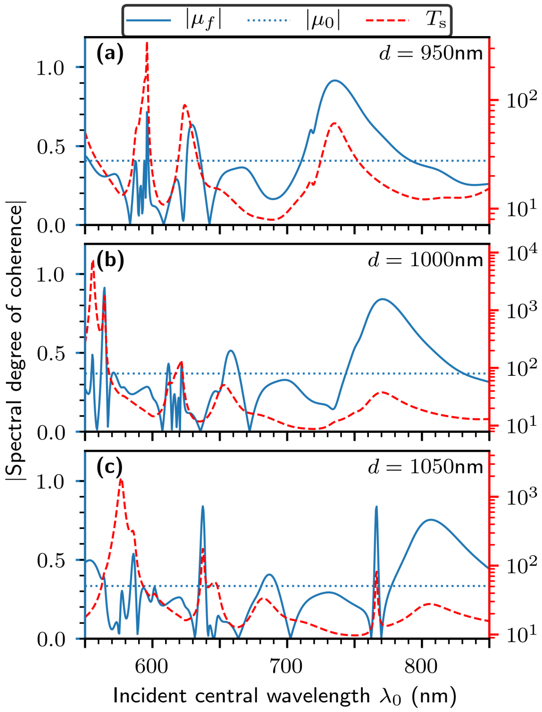

Figure 3.4: Center hole output and input coherence and scaled transmittance for a 6×6 hole array showing the effect of changing lattice constant d. Here, $a = 200\,\text{nm}$. (a) $d = 950\,\text{nm}$. (b) $d = 1000\,\text{nm}$. (c) $d = 1050\,\text{nm}$.

does not differ significantly from $|\mu_0(\boldsymbol{\rho}_n, \boldsymbol{\rho}_m)|$, which is about 0.33 for the center hole and lower corner pairs and is about 5×10^{-24} for the outer corner pair. As a increases, $|\mu_f(\boldsymbol{\rho}_n, \boldsymbol{\rho}_m)|$ differs more from $|\mu_0(\boldsymbol{\rho}_n, \boldsymbol{\rho}_m)|$ and oscillates more strongly and rapidly. And again, transmittance maxima tend to coincide with coherence maxima.

It is worth noting that in Fig. 3.5 there many coherence peaks which are at the same location for all three hole pairs. Perhaps most prominently, there is a string of narrow peaks beginning at $a = 150\,\text{nm}$, $\lambda_0 = 698\,\text{nm}$ and continuing up and to the right which is present for all three hole pairs. This is a good indication that the coherence conversion effects at these locations are affecting the global state of coherence in roughly the same way, rather than just affecting isolated hole pairs.

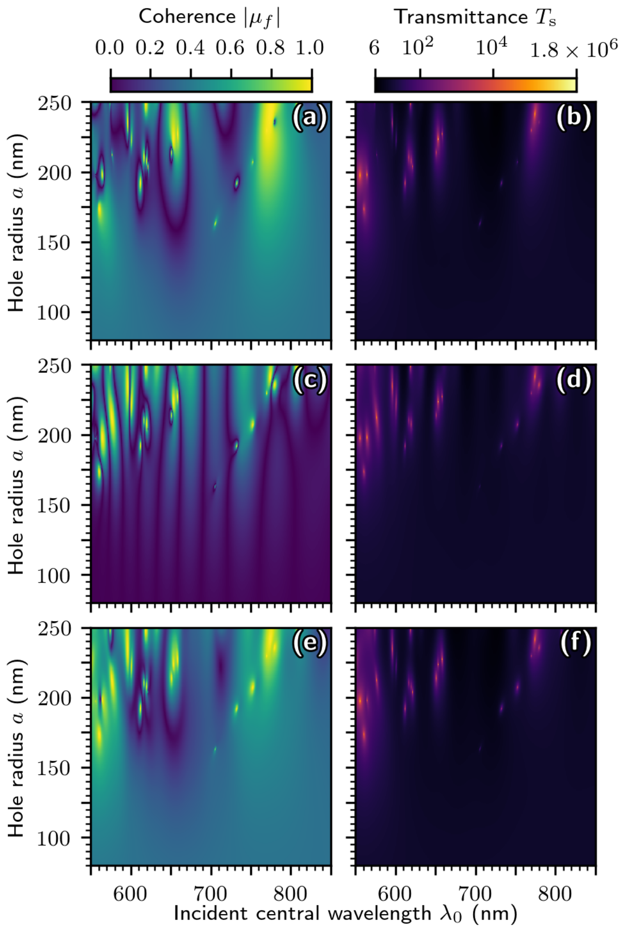

Figure 3.5: Coherence and transmittance of an 6×6 configuration, varying a, with $d = 1000\,\text{nm}$. (a) Center holes coherence. (b) Center holes transmittance. (c) Outer corners coherence. (d) Outer corners transmittance. (e) Lower left corner coherence. (f) Lower left corner transmittance. Note that the transmittances are on a logarithmic scale.

3.2.5 Relationship between coherence and transmittance

It would be good to look at the relationship between transmittance and coherence more quantitatively than we have up until now. We do this by seeing how frequently the maxima of $|\mu_f(\boldsymbol{\rho}_n, \boldsymbol{\rho}_m)|$ coincide with the maxima of $T_\text{s}(\boldsymbol{\rho}_n)$, to within $\pm 1\,\text{nm}$ (e.g., what occurs in Fig. 3.2(a) at $\lambda_0 \approx 770\,\text{nm}$). We use $1\,\text{nm}$ because the wavelength resolution we are using is $0.5\,\text{nm}$, and with discretization there is some ambiguity in where a maximum truly is. It also allows us to count maxima that may be very close, but not perfectly aligned.

Figure 3.6 shows the fraction of coinciding maxima as a function of increasing array size. The total number of maxima of $|\mu_f(\boldsymbol{\rho}_n, \boldsymbol{\rho}_m)|$ over the wavelength range we're considering is denoted N_μ. Similarly, the number of transmittance maxima is denoted N_T. The number of these maxima that coincide within ±1 nm over the wavelength range is denoted N. It can be seen that N/N_μ and N/N_T generally increase with increasing array size. This suggests that arrays with very many holes, as may be used for a real coherence conversion device, will have high correlation between transmittance maxima and coherence maxima.

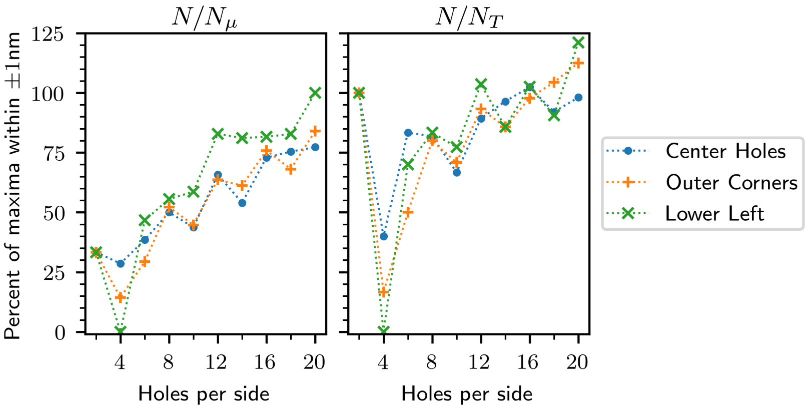

Figure 3.6: Comparing the location of coherence and transmittance maxima that coincide within ±1 nm. Note that for higher array sizes, a single coherence/transmittance maximum may be within ±1 nm of multiple transmittance/coherence maxima, so the percentage may exceed 100%.

As there is no significant difference in behavior between N/N_μ and N/N_T, to reduce the number of plots we will only consider N/N_T from here on out.

Figure 3.7 shows N/N_T for changing lattice constant d, with one subfigure for each hole pair. For the center holes, N/N_T is fairly uniform: it is mostly between 50% and 80%, occassionally hitting 100%. For the outer corners, N/N_T begins near 20% and increases steadily to around 80% at $d \approx 1850\,\text{nm}$ before declining again. For the lower left pair, N/N_T varies almost sinusoidally between about 40% and 100%. The insets show histograms of the data from their respective plots. For the central hole pair, N/N_T is centered on about 50%, with a spread mostly from 40% to 90%. The outer corner pair has a similar distribution as the center hole pair. The lower left pair's distribution is centered at roughly 70%. Taken together, these results suggest what we already guessed intuitively from the figures: that the maxima of $|\mu_f(\boldsymbol{\rho}_n, \boldsymbol{\rho}_m)|$ and $T_\mathrm{s}(\boldsymbol{\rho}_n)$ coincide more often than not. That said, d does not seem to have a strong effect on this, as N/N_T does not significantly increase or decrease as d increases for all three pairs.

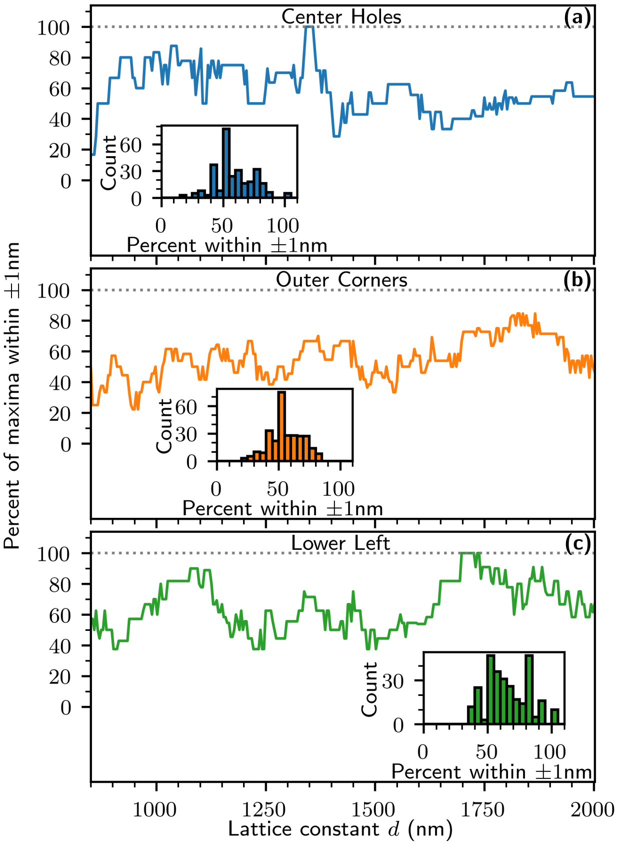

Figure 3.7: Counting coinciding peaks while changing d on a 6×6 array, with $a = 200\,\mathrm{nm}$. (a) N/N_T as function of d for the center hole pair. (b) Outer corners pair. (c) Lower left pair. The insets show a histogram of the data in their respective plots.

Figure 3.8 shows the N/N_T for changing hole radius a for the three hole pairs. They all follow a similar trend. Recall from Fig. 3.5 that, for low a, there are not many maxima of coherence or transmittance, so all the percentages are around $^1/_2$, $^1/_3$, $^1/_4$, and $^1/_5$. This means that, for low a, typically only one or zero maxima coincide. After about $a = 150\,\mathrm{nm}$, N/N_T gradually increases, maxing out between 70% and 100%.

The insets shows histograms of the data. Aside from the large number at 0%, the distributions are widely spread. From all this, can see that a does seem to have some effect on the coincidence of coherence and transmittance maxima. Specifically, for a less than about 150 nm, the few maxima present do not coincide much.

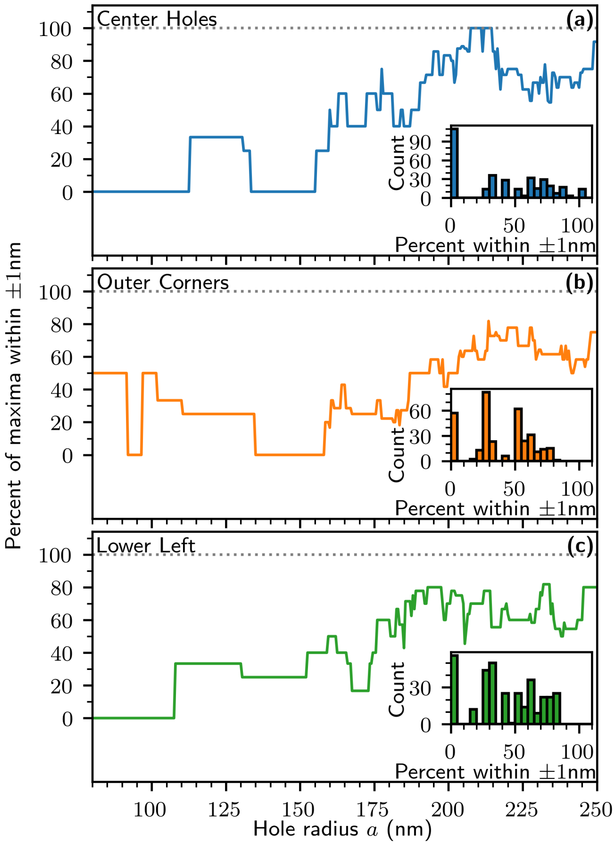

Figure 3.8: ounting coinciding peaks while changing a on a 6×6 array, with $d = 1000$ nm. (a) N/N_T as function of d for the center hole pair. (b) Outer corners pair. (c) Lower left pair. The insets show a histogram of the data in their respective plots.

3.3 Averaged Coherence

As an aid to anyone who might continue work on this project, show the averaged coherence of a square array as a function of λ_0, d, and a in Fig. 3.9. The ν_2 modes are indicated as they were in Fig. 2.11. We can see in Fig. 3.9(a) that we seem to still have band-gap-like behavior near the ν_2 modes. In Fig. 3.9(b), we see that, above the $\nu_2 = 3$ line, we have a line of isolated coherence peaks approaching the line, similar to that of Fig. 2.11(b). However, unlike Fig. 2.11(b), there is not another line of peaks asymptotically approaching the line from below.

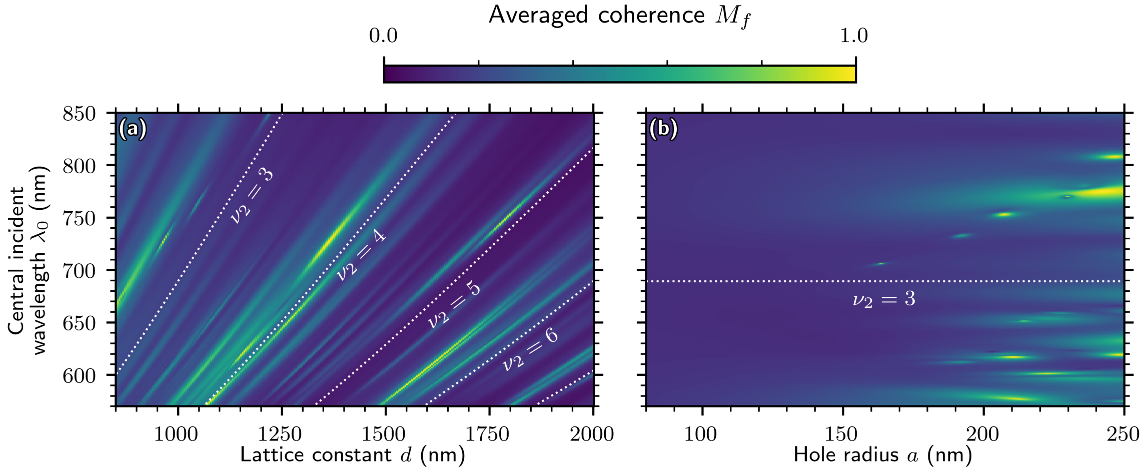

Figure 3.9: Averaged coherence M_f of a 6×6 array, with $\delta = 1000\,\mathrm{nm}$. (a) M_f as a function of λ_0 and d, with $a = 200\,\mathrm{nm}$. (b) M_f as a function of λ_0 and a, with $d = 1000\,\mathrm{nm}$.

3.4 Conclusions

In this chapter, we have examined the coherence and transmittance response of square hole arrays to changing array size, hole spacing, and hole radius. We have also shown that coherence maxima tend to coincide with transmittance maxima, though not always.

CHAPTER 4: CONSTRUCTION OF ARBITRARY VORTEX AND SUPEROSCILLATORY FIELDS

4.1 Introduction

In this project, we have developed a method to mathematically construct *superoscillatory fields* (fields which are "faster than Fourier" [34]) in the transverse plane of a beam which also involves arbitrary placement of *optical vortices* (zero-intensity lines about which the phase rotates). This research was published in Ref. [35]. In this chapter, we will first describe some of the basic physics of superoscillations and of optical vortices. Then we will describe our mathematical method for creating fields containing vortices that can also be superoscillatory. After that, we will give some demonstrations of our method and then discuss some error-checking we did to make sure our method is sound. We end with a few concluding thoughts.

4.2 Superoscillations

For a long time, it was thought due to Fourier analysis that no band limited signal $f(x)$ could oscillate faster than its highest frequency component. For example, if the function is band limited such that its Fourier transform $\tilde{f}(k_x)$ is zero outside the interval $[-k_L, k_L]$, then we would expect $f(x)$ to have no oscillations of higher frequency than k_L. It has since been shown that this is not always the case. As first popularized by Berry [34], a band limited signal can be made to oscillate *arbitrarily* fast in the presence of closely-spaced zeros. These functions are what we call superoscillatory.

We give an example of a superoscillatory function in Fig. 4.1. Figure 4.1(a) shows the reciprocal-space function $\tilde{g}(k)$, which is band limited at $k = \pm k_L$. With these band limits, standard Fourier theory would predict that the real-space function $g(x)$

would not have much oscillation within a period $\lambda_{\min} = 2\pi/k_L$. In Fig. 4.1(b), we show the normalized absolute value of $g(x)$, the inverse Fourier transform of $\tilde{g}(k)$ with $k_L = 1$. With this value of k_L, we would typically not expect to have a full period of oscillations within a range less than 2π. One of the easiest ways of identifying a period of oscillations is to count zeros: when the function crosses zero three times, that is a period. So in Fig. 4.1(b), we would typically not expect to see three zeros within the period 2π, and for the most part we do not. However, looking at the inset, we can see that the function has three zeros within a period from -1 to 1, corresponding to a period of 2. This is obviously less than 2π, so these oscillations are faster than would be expected. These oscillations are called *superoscillations*, and the region of a function where they occur (here, from about -1 to 1) is called a *superoscillatory region* of the function. One important feature of superoscillatory function that we can see here is large sidelobes. Superoscillatory functions always have sidelobes next to the superoscillatory region that are much larger than the superoscillations are – usually by several orders of magnitude [34]. We can see here that the sidelobes are roughly two orders of magnitude greater than the superoscillations. When it comes to applications, these sidelobes are usually the greatest challenge to overcome. In Fig. 4.1(c), we show the normalized magnitude of $g(x)$ again, along with a sinusoid of frequency k_L. Here, we have used a logarithmic scale. Logarithmic scales are often used when plotting superoscillatory functions for two reasons. First, it helps make the superoscillatory region visible by not letting the sidelobes dominate the plot. Second, it makes the zeros easier to see; the very narrow downward dips in Fig. 4.1(c) are zeros of the functions[1]. We can easily see that $g(x)$ has the three zeros spaced more closely together than does the sinusoid, again confirming the superoscillatory nature of $g(x)$. In the rest of this chapter, we will usually use logarithmic scales to examine

[1] Technically, zero on a logarithmic scale would be negative infinity. However, that is not typically visible in logarithmic plots, either due to discretization or to aesthetic choice (as was done here, since vertical lines going all the way to the bottom of the plot can clutter the figure and be distracting).

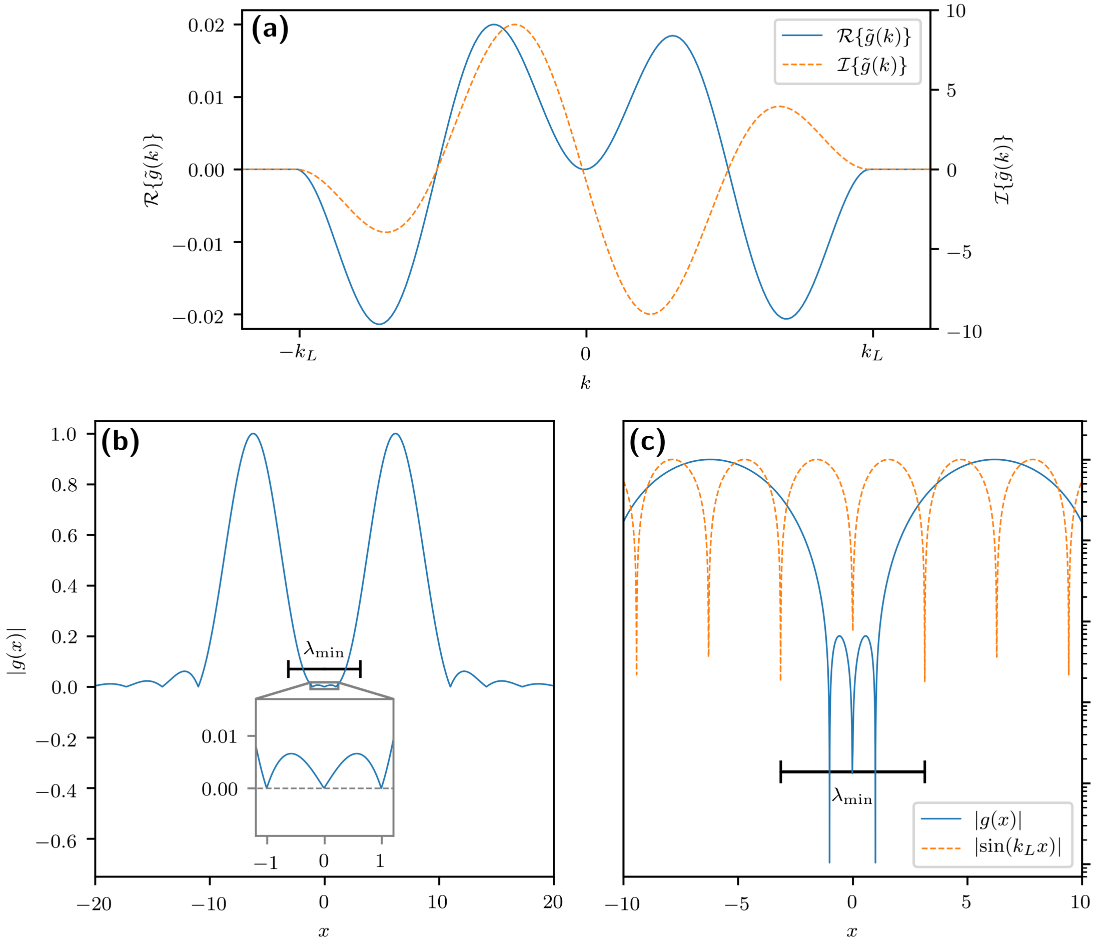

Figure 4.1: Example of a superoscillatory function. (a) Real and imaginary parts of a reciprocal-space function $\tilde{g}(k)$, band limited by $\pm k_L$. (b) Normalized absolute value of the Fourier transform $g(x)$ of $\tilde{g}(k)$, with $k_L = 1$. The inset shows oscillations exceeding the maximum frequency of the signal. (c) Normalized absolute value of $g(x)$ along with a sinusoid of frequency k_L, the maximum frequency of $\tilde{g}(k)$. Here, $k_L = 1$. Note the logarithmic scale to help show the location of zeros.

the magnitude in superoscillatory regions.

Superoscillations have been investigated in a number of fields; here we will give a few examples. In quantum mechanics, it has been shown that a quantum particle with bounded momentum (a band limit) can have its momentum increase after passing a superoscillatory part of its wavefunction through a neutral slit [36]. It was also shown that superoscillations in quantum wavefunctions can persist for a longer time than might be expected [37]. In signal processing, superoscillations are part of why it is difficult to define a time-varying bandwidth: their existence means that the oscillation frequency about some point in time is not a reliable indicator of the signal's bandwidth [38]. That same paper showed that "while a frequency limit does not pose a limit to how quickly a function can vary, a frequency limit does pose a limit to how much a function's Nyquist rate samples can be peaked" [38]. In addition, it has been shown that, to achieve a certain number N of superoscillations, the energy required increases polynomially with $1/k_L$ and exponentially with N [39]. In optics, it has been shown that, in the speckle pattern of random waves with disc-shaped band limits, $1/5$ of the area of the speckle pattern is superoscillatory, so superoscillations are more common that may have been intuitively thought [40]. One important application of superoscillations in optics is superresolution – beating the diffraction limit in light focusing through a lens, which is important for nano-optics applications. We discuss such an application in Chapter 5.

4.3 Optical Vortices

Optical vortices are phase structures that occur in monochromatic light fields as lines in 3D space where the intensity of the light is zero [41, pp. 223]. If we consider the transverse plane of a beam, the field is typically modeled by use of complex numbers, where the field $U(x, y)$ is of the form

$$U(x,y) = U_r(x,y) + \mathrm{i}U_i(x,y) = U_m(x,y)\,\mathrm{e}^{\mathrm{i}\theta(x,y)}, \tag{4.1}$$

where $U_r(x, y)$ is the real part of $U(x, y)$, $U_i(x, y)$ is the imaginary part, $U_m(x, y)$ is the amplitude (note that the beam's intensity is the square of its amplitude, in appropriate units), and $\theta(x, y)$ is the phase. It follows that $\theta = \tan^{-1}(U_i/U_r)$; if the amplitude U_m is zero at some point in the plane, then at that point the phase will be $\theta = \tan^{-1}(0/0)$, which is undefined, or *singular*. About these points, the beam's phase will circulate, undergoing some integer multiple of 2π cycles about the zero point [41, pp. 228]. These singularities in phase are what create the vortex behavior we will be examining in this section. Because of this, the study of optical vortices is often called *singular optics* [41].

4.3.1 Basics of optical vortices

Many of the basic properties of optical vortices can be seen by looking at the phase of *Laguerre-Gauss beams*. In the waist plane (z=0), Laguerre-Gauss beams are defined, without normalization, in polar coordinates as [21, pp. 649]

$$E_{\mathrm{LG}}(\rho,\phi) := E_0\left(\frac{\sqrt{2}}{w_0}\right) L_p^{|l|}\left(\frac{2\rho^2}{w_0^2}\right)\exp\left(\frac{-\rho^2}{w_0^2}\right)\rho^{|l|}\exp(\mathrm{i}l\phi)\,, \tag{4.2}$$

where w_0 is the beam width in the waist plane, E_0 is a constant,and $L_p^{|l|}(2\rho^2/w_0^2)$ are the associated Laguerre polynomials. Althought they are often expressed in polar coordinates, for our purposes it will be more useful to express the Laguerre-Gauss beam in Cartesian coordinates:

$$E_{\mathrm{LG}}(x,y) = E_0\left(\frac{\sqrt{2}}{w_0}\right) L_p^{|l|}\left(\frac{2\rho^2}{w_0^2}\right)\exp\left(\frac{-\rho^2}{w_0^2}\right)(x\pm \mathrm{i}y)^{|l|}\,, \tag{4.3}$$

where the plus/minus sign on y corresponds to the sign of l in Eq. (4.2). In Fig. 4.2, we plot the normalized amplitude (top row) and the phase (bottom row) of a few of these beams.

From Fig. 4.2, we can see several properties of optical vortices. The simplest optical vortex, an $l = 1$, $p = 0$ mode, is shown in Fig. 4.2(a) and (b). We can see that the amplitude has a bright ring with a dark hole in the center where the intensity is zero. Looking at the phase, we can see that the phase undergoes a full 2π rotation, from $-\pi$ to π, in a counter-clockwise direction centered on the origin. That point at the origin, which the phase rotates about and where the amplitude is zero, is the phase singularity; this singularity and the surrounding phase structure together are

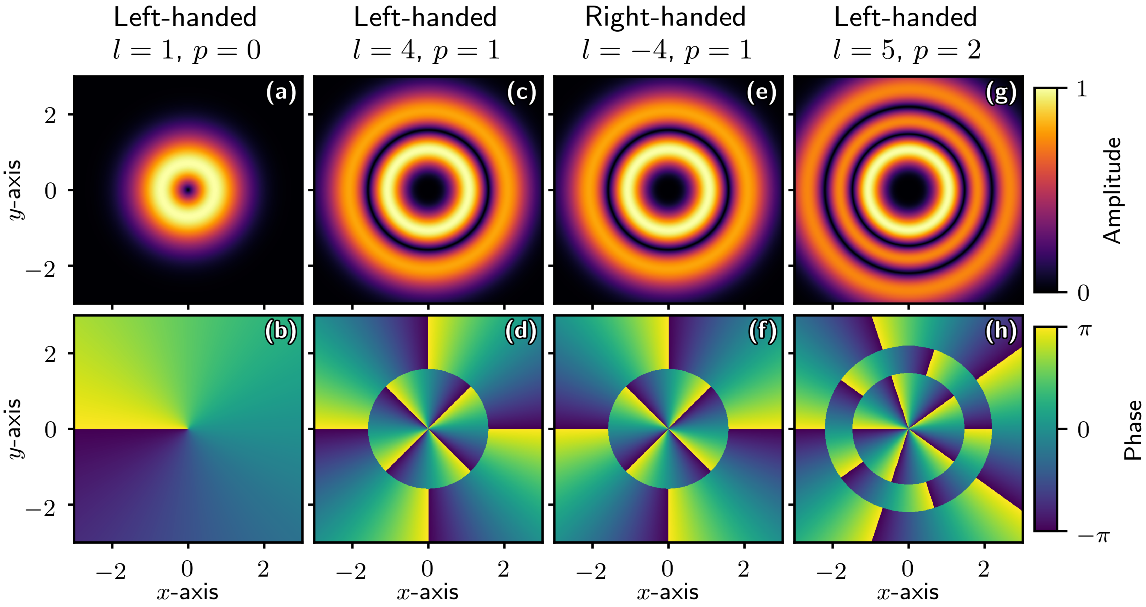

Figure 4.2: Examples of Laguerre-Gauss beams.

the optical vortex. The counter-clockwise direction is denoted *left-handed.* It is so named because if your left thumb is on the vortex, pointing into the page, your fingers will curl in the direction of rotation, from $-\pi$ to π. Optical vortices can be either left-handed or right-handed, depending on whether we take the sign on $x \pm iy$ in Eq. (4.3) to be positive or negative, respectively. In the next column, figures (c) and (d), we see the amplitude and phase of a $l = 4$, $p = 1$ beam. Incrementing p from 0 to 1 has added a new ring about the center where the intensity is zero. Additionally, increasing l to 4 has caused the dark region in the center to widen. Looking now at the phase, we can notice a few differences from the $l = 1$, $p = 0$ phase. First the phase now undergoes four full rotations about the vortex. The number of rotations the phase completes about a vortex is called the vortex's *topological charge* or its *order* [42, pp. 17]. So this is a 4th order vortex, whereas the the vortex in (b) was a 1st order vortex. The order of a vortex will always be an integer because the field is assumed to be *analytic* [41, pp. 228], a term from the calculus of complex variables which indicates that the field has no discontinuities. At the zero-intensity ring, we can see that the phase abruptly changes by π. In the third column, (e) and (f), we show the amplitude and phase of an $l = -4$, $p = 1$ beam, which is right-handed. Comparing this phase with the phase in (d), we can see that the phase does indeed rotate in the opposite direction. It is important to note that the amplitudes in (c) and (e) are exactly the same – changing the handedness of the vortex produced no change in the intensity pattern of the beam. This can be a problem when trying to detect optical vortices. Finally, in (g) and (h) we see the amplitude and phase of an $l = 5$, $p = 2$ beam. Incrementing p to 2 has added another zero ring, so we can conclude that the number of zero rings is equal to p. We will refer to this phase and amplitude plot later in this chapter.

One feature of optical vortices not shown by the Laguerre-Gauss beams in Fig. 4.2 is that they tend to exist in pairs of opposite handedness [43], and the lines of constant phase tend to "originate" on one member of the pair and "terminate" on the other. If there is not an opposite-handed vortex to pair with, we may formally say that there is one at infinity [44]. This is suggested by Fig. 4.2, where lines of constant phase (most noticeably the line where the phase transitions from π to $-\pi$) are lines going to infinity.

Another feature of optical vortices not seen in Fig. 4.2 is their *dislocation type*. This terminology comes from the language of crystal lattices and defects in such lattices. Adding a new plane of atoms to a crystal lattice dislocates other atoms in the lattice generally in one of two ways: either as an *edge dislocation* or a *screw dislocation*. As it is not a large component of this chapter's project, we will not delve much into this topic in this dissertation; the interested reader can learn more in Refs. [41, sections 3-5] and [42, chapter 3]. Briefly, a screw dislocation is defined as the line of zero intensity being parallel to the axis of propagation. This causes the phase to have a helical shape along the direction of propagation, like the threads on a screw. The vortices at the center of the Laguerre-Gauss beams in Fig. 4.2 are screw dislocations, as can be seen by their direction and the phase rotation about them. An edge dislocation is defined as the zero-intensity line being perpendicular to the axis of propagation. When viewed from along this axis, this dislocation looks line an "edge" in the phase plot across which the phase abruptly changes by a value of π. The phase does circulate around an edge dislocation, but the circulation isn't visible because we are viewing it perpendicularly. A circular edge dislocation would look like the phase of the zero-intensity rings in Fig. 4.2 (d), (f), and (h). Note that, although we said that the zero-intensity rings in the $p > 0$ beams look like edge dislocation, they are *not* edge dislocations. This is because the rings in the plot are not actually rings in the beam. They extend along the length of the beam, so in 3D space they are actually "cylinders" of zero-intensity[2], not lines or rings, so they are not optical vortices at all. It is also possible for a vortex to be neither parallel nor perpendicular to the axis of propagation; such vortices are called *mixed edge/screw dislocations.* Also note that the zero-intensity lines are usually only *locally* straight; they can make shapes like knots, braids, or twisted loops [45], as can be seen in laser speckle [46].

Note that there is not actually a sharp discontinuity in the phase when the phase crosses from π to $-\pi$, as it might seem from the phase plots in Fig. 4.2. The lines formed when the phase crosses that $\pi/-\pi$ transition are an artifact of the color map used in the plots. A more realistic picture of the phase could be obtained using a color map that has the same color for both the maximum and the minimum. However, those color maps tend to make it more difficult to identify vortex location, handedness, and order, which are the vortex properties we are most concerned with for this project.

4.3.2 Applications of optical vortices

Optical vortex beams have a number of interesting applications. Here, we will briefly give a few, just to give a flavor of what is possible.

One intriguing application of optical vortices is their possible use in micromachines.

[2] If you want to be technical, Laguerre-Gauss beams expand in either direction from the waist plane, so the shape is less of a cylinder and more like a tube of paper with a rubber band in the middle.

Optical vortices possess orbital angular momentum, and they can impart this momentum to objects, offering many applications for rotating devices at the microscale. The rotation can be imparted by simple absorption [47] or by birefringence [48]. These rotations have been used to produce gear-like systems [49, 50]. Small particles not in the central axis of the beam can orbit the center of the beam [51, 52]; rows of such beams have led to the production of microoptomechanical pumps [53].

Optical vortices have been used to show the existence of the *rotational Doppler shift*. As the name implies, this is a version of the well-known translational Doppler shift that manifests with rotations. Imagine we have a vortex propagating in a straight line, and we consider a reference frame which is rotating about the same axis as the vortex. The speed at which the reference frame is rotating relative to the vortex will change the perceived angular speed of the vortex. Since the vortex rotates with the same frequency as the wave, this results in a changed perception in the wave's frequency as well – this is the rotational Doppler shift. This new Doppler shift has been used for both translating and rotating optically trapped particles [54] and measuring the rotation of an object [55].

One promising application of optical vortices is for point-to-point free-space optical communications. Referring back to Laguerre-Gauss beams, each mode is orthogonal to the others, so in principle some number M of vortices of different order m could be used to transmit information simultaneously. In addition, vortices are fairly resistant to atmospheric turbulence, reducing their distortion on propagation. There are some challenges to overcome, though. For example, after some amount of propagation, individual vortices can wander enough to miss the detector, and higher-order vortices tend to break up into lower-order ones (i.e., a vortex of order m will break into m 1st order vortices) [56].

4.3.3 Creation and detection of optical vortices

Here, we describe a few methods for creating optical vortices and detecting them; this information is by no means exhaustive.

There are several ways of creating beams with optical vortices. One of the simplest, conceptually, is to create a Laguerre-Gauss beam by passing a Gaussian laser mode through a spiral phase plate [57]. However, this does not create a pure mode, and such a phase plate can be difficult to fabricate, since the height of the plate needs to be on the order of the wavelength [42, pp. 64-65]. A laser beam can be converted to a vortex beam using computer-generated holograms [58] or liquid-crystal displays [59, 60, 61]. One group created optical vortices using surface plasmons excited in spiral grooves on a thin gold film with a central cylindrical aperture. The plasmon modes would impart orbital angular momentum to the light, thus causing the output to be a vortex mode [62].

Detecting vortices is a difficult challenge, since they are inherently a phase structure, while optical measurements are usually of intensity. Simply detecting zeros of intensity is not enough, as a pure zero cannot be separated from a very low signal that happens to be below the experimental uncertainty. Also, as can be seen from Fig. 4.2(c-f), vortices with opposite handedness can produce identical intensity patterns. One detection method has been to interfere a vortex beam with its mirror image [63]. Several methods exist based on diffraction, where vortices of different handedness/order will produce different diffraction patterns. This is often done with a triangular aperture [64, 65], though other shapes have been used [66, 67, 68]. Computer-generated holograms can be used to separate different modes to different parts of the observation plane: a bright spot in a certain location means a specific type of vortex was present [69, 70, 71]. A similar idea separates vortex types to different parts of the observation plane by using geometric techniques [72].

4.4 Mathematical Method for Making Superoscillatory Optical Vortex Fields

One of the challenges of superoscillatory functions has been that they require somewhat complicated mathematics in order to produce them, such as asymptotics [34] or Tschebyscheff polynomials [73]. A relatively simple Fourier method for producing superoscillations in 1D functions was published by Chremmos and Fikioris [74]. Our method for constructing arbitrary and superoscillatory optical vortex fields is initially based on that method, but extends the work to 2D fields of complex variables and allows for the creation of optical vortices of arbitrary location, handedness, and order.

We begin with a two-dimensional, band limited function $\tilde{f}(k_x, k_y)$ in reciprocal space. The corresponding real-space function $f(x, y)$ is given by the inverse Fourier transform (IFT)

$$f(x,y) = \frac{1}{(2\pi)^2} \iint_{-k_L}^{k_L} \tilde{f}(k_x,k_y)\, \mathrm{e}^{\mathrm{i}(k_x x + k_y y)} \mathrm{d}k_x \mathrm{d}k_y, \tag{4.4}$$

where ±k_L are the band limits of the function. We now multiply this by the N^{th}-order polynomial $h(\bar{z})$,

$$h(\bar{z}) := \sum_{n=0}^{N} a_n \bar{z}^n, \tag{4.5}$$

where $\bar{z} := x + \mathrm{i}y$ and a_n are real constants. Let us call the result of this multiplication $g(x, y)$,

$$g(x,y) := h(\bar{z})\, f(x,y)\,. \tag{4.6}$$

We note that this new function $g(x, y)$ will have the same band limits as $f(x, y)$. We can see this by looking at its Fourier transform (FT),

$$\tilde{g}(k_x,k_y) = \sum_{n=0}^{N} a_n \mathrm{i}^n \left[\frac{\partial}{\partial k_x} + \mathrm{i}\frac{\partial}{\partial k_x}\right]^n \tilde{f}(k_x,k_y)\,, \tag{4.7}$$

where we have used the property that

$$\mathcal{F}\{x^n f(x)\} = \mathrm{i}^n \frac{\partial^n}{\partial k_x^n} \tilde{f}(k_x, k_y)\,, \tag{4.8}$$

and similarly for y and k_y, where $\mathcal{F}\{\}$ denotes the FT operation. Recall that $\tilde{f}(k_x, k_y)$ is zero outside of its band limit. Because of this, the constants a_n and the differentiation do not cause $\tilde{g}(k_x, k_y)$ to have a greater bandwidth than $\tilde{f}(k_x, k_y)$. However, Eq. (4.7) does impose a limitation on $\tilde{f}(k_x, k_y)$. Namely, the first $N-1$ derivatives of $\tilde{f}(k_x, k_y)$ must be continuous, or else $\tilde{g}(k_x, k_y)$ will have Dirac-delta singularities [74].

Now we can see how this method produces superoscillations and optical vortices. For $g(x, y)$ to be superoscillatory, it must have oscillations that are more rapid than its highest band limit. We can create these oscillations by placing zeros in $g(x, y)$ using the polynomial $h(\bar{z})$. We can see how to do this by rewriting $h(\bar{z})$ in terms of its roots as

$$h(\bar{z}) = \prod_{n=0}^{N} (\bar{z} - \bar{z}_n)\,, \tag{4.9}$$

where $\bar{z}_n$ is the n^{th} root of $h(\bar{z})$. We can choose any $\bar{z}_n$ we want; at those points in the plane, $h(\bar{z})$ will be zero, and so $g(x, y)$ will be zero. If we place these zeros close enough together, then $g(x, y)$ will be oscillating faster than the band limits k_L would normally permit and thus it would be superoscillatory. (Obviously, if the zeros are placed far apart from each other, they will not necessarily make the field superoscillatory.) And, since at the zeros $g(x, y)$ has the form $(x - x_n) + \mathrm{i}(y - y_n)$, like the Laguerre-Gauss beam in Eq. (4.3), $g(x, y)$ will have optical vortices at those locations as well.

Note that there is no restriction on which points can be zeros: this allows us to place zeros in *any arbitrary* arrangement. Furthermore, this method gives arbitrary control over both the handedness and the order of the vortices. The polynomial in Eq. (4.9) produces left-handed vortices by default. To produce right-handed vortices, simply take the complex conjugate of the term in parentheses. That is, replace $(\bar{z} - \bar{z}_n)$ with $(\bar{z} - \bar{z}_n)^*$. To have a vortex of order m, simply use the desired $\bar{z}_n$ as a root m times. A word of caution: you cannot simply replace $(\bar{z} - \bar{z}_n)$ in Eq. (4.9) with $(\bar{z} - \bar{z}_n)^m$. That is because, although the m^{th}-order vortex is a single vortex, it requires m derivatives to avoid singularities, so with this method it is best to think of an m^{th}-order vortex as m vortices of order 1 occupying the same point. To produce a m^{th}-order vortex which is right-handed, apply the conjugation at all m roots. We give the algorithm for our method in Algorithm 1 below.

Algorithm 1 Superoscillatory optical vortex field algorithm

```
Choose $N$ points in the plane where you want zeros.
Define a coordinate system $\bar{z} := x + \mathrm{i}y$ and assign these points their coordinates, where the $n^{\mathrm{th}}$ zero has coordinate $\bar{z}_n$.
Create the polynomial $h(\bar{z})$ using Eq. (4.9), applying conjugation where a right-handed vortex is desired.
Define a band limited function $\tilde{f}(k_x, k_y)$ in reciprocal space which has at least $N - 1$ continuous derivatives.
Take the IFT of $\tilde{f}(k_x, k_y)$ to obtain the real-space function $f(x, y)$.
Obtain $g(x, y)$ using Eq. (4.6). The field $g(x, y)$ will have optical vortices at the points $(x_n, y_n)$ and will be superoscillatory if these zeros are close enough together.
```

This method could be implemented by a spatial light modulator (SLM) and a thin lens in a $2f$ configuration, where f is the focal length of the lens, as shown in Fig. 4.3. In this configuration, the image in the rear focal plane is the FT of the object in the front focal plane [75, pp. 87]. So, the SLM could display the reciprocal space field $\tilde{g}(k_x, k_y)$, which would yield the superoscillatory vortex field $g(x, y)$ in the rear focal plane.

We note that, as we have described it, this method will produce vortices with screw dislocations. It is also possible to make vortices with mixed edge/screw dislocations. To do so, simply change the definition of the coordinate system $\bar{z}$ from $\bar{z} := x + \mathrm{i}y$ to $\bar{z} := \alpha x + \mathrm{i}\beta y$, where α and β are real constants. The relative value of α and β will determine how much the zero line is tilted along each axis.

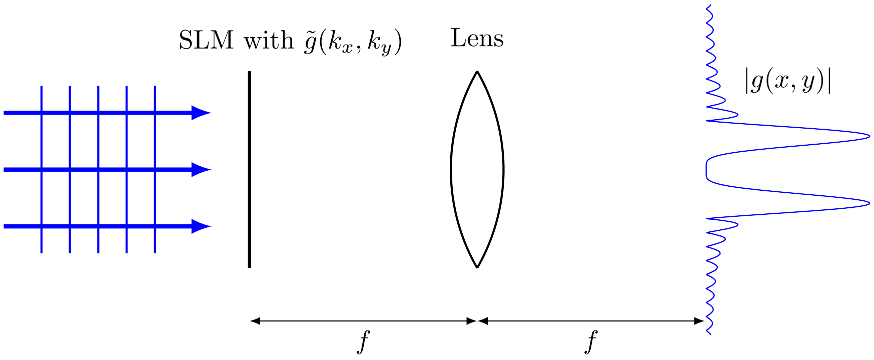

Figure 4.3: A spatial light modulator (SLM) and a thin lens in a $2f$ configuration.

When we submitted the first version of our paper [35], a reviewer pointed out that our resulting vortex arrangements appeared somewhat similar to those that could be obtained by a perturbation method [76]. However, our method has four distinct advantages. First, the perturbation method can only produces vortices along straight lines or at the vertices of regular polygons, whereas our method allows for completely arbitrary placement of each vortex. Second, our method allows for arbitrary handedness for every vortex; the perturbation method can only make vortices all with the same handedness. Third, our method allows for higher-order vortices, while the perturbation method only yields first-order vortices. Finally, our method allows for the possibility of making mixed edge/screw vortices.

4.5 Demonstrations of Our Method

We will examine three test functions to demonstrate our superoscillatory/optical vortex field creation method. The first test function will be primarily aimed at demonstrating the superoscillatory nature of the resulting fields. The second test function will mainly look at our control over the handedness of the resulting optical vortices. The third will demonstrate our ability to place vortices at arbitrary locations and with arbitrary order.

For our first test case, we will begin with the band limited function

$$\tilde{f}_{\text{circ}}(k_x, k_y) := \cos\left(\frac{\pi}{2k_L}k\right)^5, \tag{4.10}$$

where $k := \sqrt{k_x^2 + k_y^2}$ and $\tilde{f}_{\text{circ}}(k_x, k_y)$ is band limited by setting to zero where $|k| \geq k_L$. This function, denoted $\tilde{f}_{\text{circ}}$ because of its circular symmetry, is plotted in Fig. 4.4. Since the cosine term is raised to a power of 5, it can have at most 5 derivatives without discontinuities (note that while $\cos^n(x)$ is infinitely differentiable, a *band limited* $\cos^n(x)$, when band limited to its first zeros, is only differentiable n times). With this function we will use the polynomial

$$h_{\text{circ}}(\bar{z}) := \bar{z}^5 - \frac{5}{4}\bar{z}^3 + \frac{1}{4}\bar{z}, \tag{4.11}$$

which has roots along the x axis at $\bar{z} = 0$, ± 0.5, and ± 1. Our final field is

$$g_{\text{circ}}(x, y) := h_{\text{circ}}(\bar{z})\, f_{\text{circ}}(x, y)\,, \tag{4.12}$$

where $f_{\text{circ}}(x, y)$ is the IFT of $\tilde{f}_{\text{circ}}(k_x, k_y)$, which we calculated numerically.

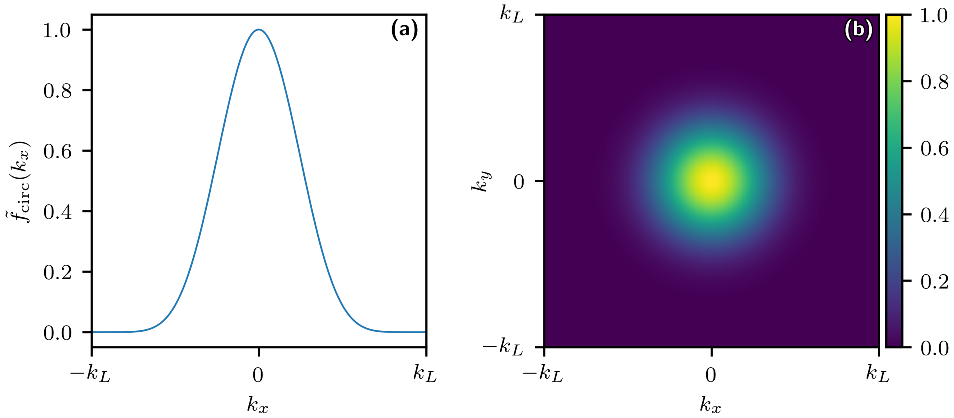

Figure 4.4: (a) x-axis slice of $\tilde{f}_{\text{circ}}(k_x, k_y)$. (b) Plot of $\tilde{f}_{\text{circ}}(k_x, k_y)$.

In Fig. 4.5 we plot the normalized amplitude and the phase of $g_{\text{circ}}(x, y)$. In

Fig. 4.5(a), we can see that the amplitude has a circular shape similar to that of the $l = 5$ Laguerre-Gauss beam in Fig. 4.2(g). In Fig. 4.5(b) (note the logarithmic scale), we see a close-up of the dark spot at the center of Fig. 4.5(a). The zeros are where they should be according to Eq. (4.11). Now, with $k_L = 1$, we would typically not expect to see more than three zeros within a period $\lambda_{\text{min}} = 2\pi$, but we have five zeros in a line less than half of that, so the field is superoscillatory. Note that the amplitude of the field in the superoscillatory region is five or six orders of magnitude less that of the surrounding bright ring in Fig. 4.5(a), consistent with the behavior of superoscillations discussed in Section 4.2. In Fig. 4.5(c) we can see that the field has a phase structure similar to that of the $l = 5$, $p = 2$ Laguerre-Gauss beam in Fig. 4.2(h), with five full 2π rotations about the center and phase jumps of π at the zero rings. In Fig. 4.5(d) we see a close-up of the phase in Fig. 4.5(c). We can compare this with the amplitude in Fig. 4.5(b) to confirm that our zeros are in fact optical vortices, which are left-handed and of order 1 as expected. Note also that since all the vortices are left-handed, their right-handed counterparts are at infinity; this is why those lines at the $\pi/-\pi$ transition tend toward infinity.

To further examine the superoscillatory nature of these zeros, in Fig. 4.6 we plot the x axis of the normalized amplitude of $g_{\text{circ}}(x, y)$ along with a sinusoid of frequency k_L. It can be seen that, over the period λ_{min}, the sinusioid has three zeros, whereas $g_{\text{circ}}(x, y)$ has five. This confirms the superoscillatory nature of these zeros.

Our second test function will show our control over the handedness of vortices. We use a new $\tilde{f}(k_x, k_y)$ which has rectangular symmetry,

$$\tilde{f}_{\text{rect}}(k_x, k_y) := \cos\left(\frac{\pi}{2k_L}k_x\right)^6 \cos\left(\frac{\pi}{2k_L}k_y\right)^6 . \tag{4.13}$$

This function is band limited by setting it nonzero only where $k_x < k_L$ and $k_y < k_L$. See Fig. 4.7 for a visualization. Since the cosine terms in $\tilde{f}_{\text{rect}}(k_x, k_y)$ are raised to

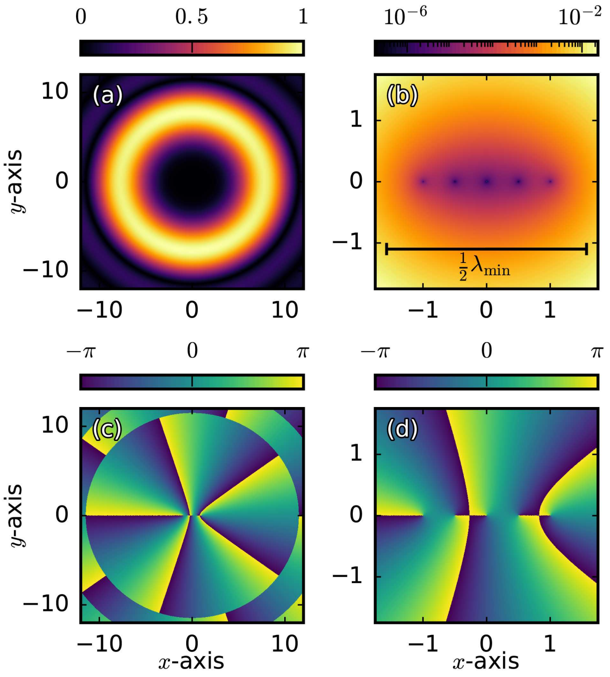

Figure 4.5: Magnitude and phase plots for the circular case, with $k_L = 1$. (a) Normalized magnitude of $g_{\text{circ}}(x, y)$. (b) Magnitude of $g_{\text{circ}}(x, y)$, zoomed in to show the zeros at the roots of Eq. (4.11) and normalized to (a). Note the logarithmic scale. (c) Phase of the field in (a). (d) Phase of the field in (b). *Reprinted with permission from ref [35], OSA.*

the power 6, there can be up to six zeros added to the field. With this function, the polynomial we will use is

$$h_{\text{rect}}(\bar{z}) := \prod_{n=0}^{5} \mathcal{C}^n \left\{ \bar{z} - \mathrm{i}\mathrm{e}^{\mathrm{i}\pi(2n+1)/6} \right\}, \tag{4.14}$$

where $\mathcal{C}$ is a complex conjugation operator. This polynomial will produce a regular hexagon of vortices on the unit circle with alternating handedness. Our final field is

$$g_{\text{rect}}(x, y) := h_{\text{rect}}(\bar{z})\, f_{\text{rect}}(x, y)\,, \tag{4.15}$$

where $f_{\text{rect}}(x, y)$ is the IFT of $\tilde{f}_{\text{rect}}(k_x, k_y)$, which we calculated numerically.

In Fig. 4.8 we plot the normalized amplitude and the phase of $g_{\text{rect}}(x, y)$. In

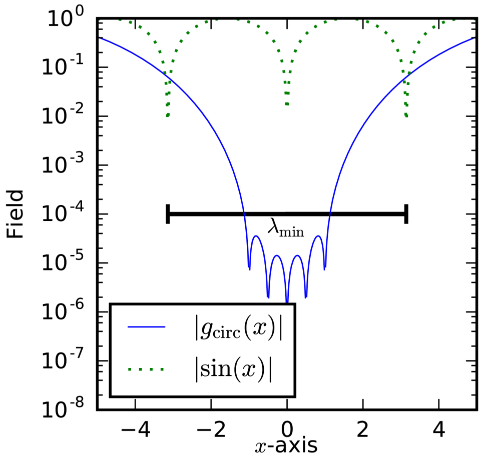

Figure 4.6: x-axis of $|g_{\mathrm{circ}}(x, y)|$, with $k_L = 1$, plotted with a sinusoid of the minimum wavelength associated with this bandwidth. The thick black line is the wavelength of the sinusoid. Note the logarithmic scale. *Reprinted with permission from ref [35], OSA.*

Fig. 4.8(a) we can see that the amplitude has a rectangular shape, which we would expect from the definition of $\tilde{f}_{\mathrm{rect}}(k_x, k_y)$ in Eq. (4.13). Its center has a dark circle, similar to the one in Fig. 4.5(a), where the added zeros are. In Fig. 4.8(b) we can see that the zeros do form a regular hexagon on the unit circle as expected. In Fig. 4.8(c) we can see that the phase of the field in Fig. 4.8(a). In Fig. 4.8(d) we can see that the vortices alternate handedness as planned. we can also see that, now that each vortex is paired with an opposite-handed vortex, we no longer have the $\pi/-\pi$ transition

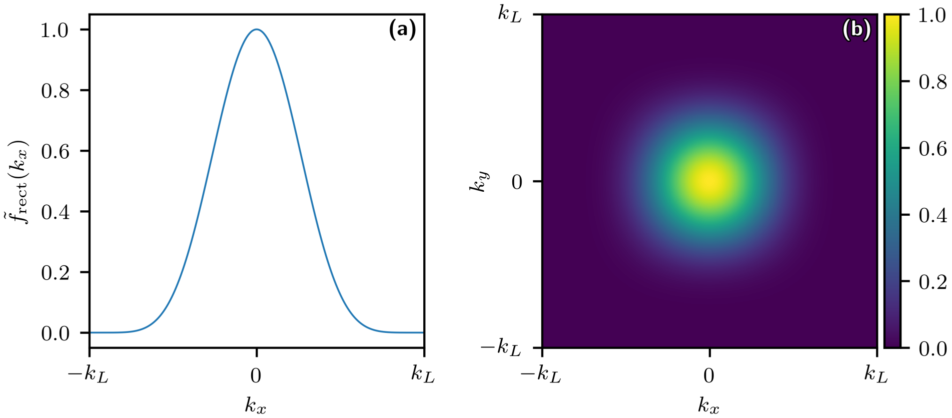

Figure 4.7: (a) x-axis slice of $\tilde{f}_{\mathrm{rect}}(k_x, k_y)$ (b) Plot of $\tilde{f}_{\mathrm{rect}}(k_x, k_y)$.

lines stretching to infinity. Rather, they connect to an opposite-handed vortex.

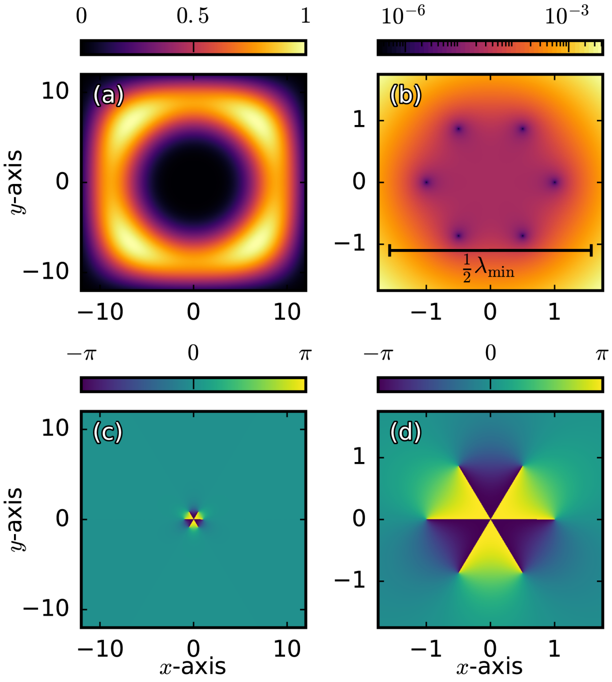

Figure 4.8: Magnitude and phase plots for the rectangular case, with $k_L = 1$. (a) Normalized magnitude of $g_{\text{rect}}(x, y)$. (b) Magnitude of $g_{\text{rect}}(x, y)$, zoomed in to show the zeros at the roots of Eq. (4.14) and normalized to (a). Note the logarithmic scale. (c) Phase of the field in (a). (d) Phase of the field in (b). *Reprinted with permission from ref [35], OSA.*

Our final test case is shown in Fig. 4.9, demonstrating our ability to make arbitrary arrangements of vortices of arbitrary order. Here we have used our method to produce a field with 76 vortices arranged to spell “UNCC.” The vortices alternate handedness within each letter, and in the second “C” they are all second order, except for the endpoints. For the reciprocal-space function $\tilde{f}(k_x, k_y)$, we used $\tilde{f}_{\text{rect}}(k_x, k_y)$ except with a power of 120 instead of 6, which provides more than enough derivatives. The results in Fig. 4.9 prove that our method can be used to produce arbitrary arrangements of vortices of arbitrary handedness and order.

4.6 Error-Checking Our Method

With the rectangular function, we can do a couple of checks on the accuracy of our method. First, the IFT of $\tilde{f}_{\text{rect}}(k_x, k_y)$ has an analytical form we can compare our

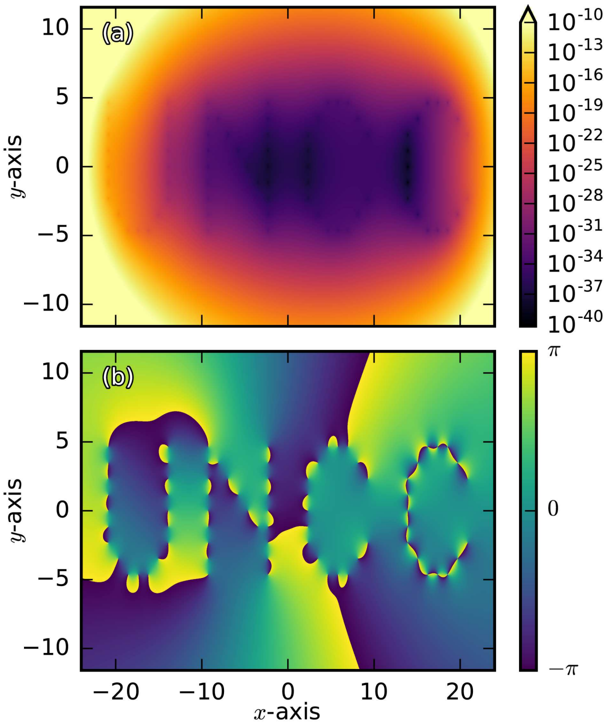

Figure 4.9: Magnitude and phase plots for a field with vortices arranged to spell "UNCC." Each of the letters has alternating right-handed/left-handed vortices. The vortices in the second "C" are second order, with the exception of the end points. (a) Normalized magnitude of the field. To help make the vortices visible, the color map was capped at a maximum of 1×10^{-10}. (b) Phase of the field. *Reprinted with permission from ref [35], OSA.*

numerically-obtained IFT against:

$$f^{(\mathrm{an})}_{\mathrm{rect}}(x, y) = \frac{1}{(2\pi)^2} \sum_{m=0}^{6} \sum_{j=0}^{6} \binom{6}{m} \binom{6}{j} \operatorname{sinc}[\beta_m(x)] \operatorname{sinc}[\beta_j(y)]\,, \tag{4.16}$$

where

$$\beta_m(x) := (6 - 2m)\,\frac{\pi}{2} + x k_L \tag{4.17a}$$

$$\beta_j(y) := (6 - 2j)\,\frac{\pi}{2} + y k_L. \tag{4.17b}$$

Comparing our numerically computed $|g_{\mathrm{rect}}(x, y)|$ with $\left|g^{(\mathrm{an})}_{\mathrm{rect}}(x, y)\right|$, where $g^{(\mathrm{an})}_{\mathrm{rect}}(x, y) = h(\bar{z})\, f^{(\mathrm{an})}_{\mathrm{rect}}(x, y)$, yielded a root-mean-square difference less that $1.1 \times 10^{-10}\,\%$ of the mean value of $|g_{\text{rect}}(x,y)|$.

One possible issue with our SLM-based system is discretization. That is, while we have so far made the image-plane $g(x,y)$ using $h(\bar{z})$ and the IFT of $\tilde{f}(k_x,k_y)$, the real system would work by having the illuminated SLM display a pixelated $\tilde{g}(k_x,k_y)$, the FT of $g(x,y)$. It is possible that the discrete pixels of the SLM could cause the superoscillatory nature of the field to be lost. As a way of checking this, we analytically derived the FT of $g(x,y)$, and then took the numerical IFT of the resulting $\tilde{g}(k_x,k_y)$. This numerical IFT is taken to be a simple approximation of the effect of having a discrete screen. Since this Fourier transform means computing a lot of derivatives, per Eq. (4.7), we used a relatively simple setup, with only three zeros. We used $f_{\text{rect}}(x,y)$ as a starting point, and multiplied it by a polynomial

$$h^{(\text{test})}_{\text{rect}}(\bar{z}) = \bar{z}\,(\bar{z}-1)\,(\bar{z}+1) \tag{4.18}$$

which has roots at 0 and ±1. The FT of the resulting $g^{(\text{test})}_{\text{rect}}(x,y)$ is

$$\begin{aligned}\tilde{g}^{(\text{test})}_{\text{rect}}(k_x,k_y) = 6b\mathrm{c}_x^3\mathrm{c}_y^3\bigg\{&\frac{1}{2}\mathrm{i}\mathrm{c}_x^2\mathrm{c}_y\left[1-88b^2+\left(1+92b^2\right)\cos(2bk_y)\right]\mathrm{s}_x+20\mathrm{i}b^2\mathrm{c}_y^3\mathrm{s}_x^3\\&-\frac{1}{2}\mathrm{c}_x^3\left[1-22b^2+\left(1+18b^2\right)\cos(2bk_y)\right]\mathrm{s}_y-90b^2\mathrm{c}_x\mathrm{c}_y^2\mathrm{s}_x^2\mathrm{s}_y\bigg\},\end{aligned} \tag{4.19}$$

where

$$b := \frac{\pi}{2k_L} \tag{4.20a}$$

$$\mathrm{c}_x := \cos(bk_x) \tag{4.20b}$$

$$\mathrm{c}_y := \cos(bk_y) \tag{4.20c}$$

$$\mathrm{s}_x := \sin(bk_x) \tag{4.20d}$$

$$\mathrm{s}_y := \sin(bk_y)\,. \tag{4.20e}$$

We then compared taking the discrete FT of this analytically-derived $\tilde{g}_{\text{rect}}^{(\text{test})}(k_x, k_y)$ against the $g_{\text{rect}}^{(\text{test})}(x, y)$ obtained using our normal method (taking the IFT of $\tilde{f}_{\text{rect}}(k_x, k_y)$ then multiplying by $h_{\text{rect}}^{(\text{test})}(\bar{z})$). Both cases had the superoscillatory zeros, and the root-mean-square difference between their normalized amplitudes was $1.8 \times 10^{-8}\,\%$ of the mean of $\left|g_{\text{rect}}^{(\text{test})}(x, y)\right|$ obtained via our normal method.

4.7 Conclusions

In this chapter, we have used simulations to demonstrate the validity of a technique for mathematically producing optical vortices at arbitrary locations in the transverse plane of a beam which can make the field be superoscillatory. Additionally, this technique allows for arbitrary control over each vortex's handedness and order.

CHAPTER 5: SUPEROSCILLATORY LENS

5.1 Introduction

In Chapter 4 we introduced a method for creating arbitrary arrangements of optical vortices which can also be superoscillatory. In this chapter, we use a modification of this method to propose a *superresolution* lens, a lens with a resolution higher than would be expected by the conventional diffraction limit. A simple example of this idea was first demonstrated by Gbur [77]; here, we examine it in more depth.

5.2 Superresolution and Superoscillatory Lenses

"Resolution" is a somewhat ambiguous concept, as there is no unique way to mathematically define the resolution of a lens. That said, resolution is related to the lens's *point spread function* (PSF), which is the lens's response to a point source. To borrow the language of linear system theory, the PSF is the lens's impulse response. The PSF of a typical circular lens is the *Airy disk* [78, pp. 469-472], shown in Fig. 5.1(a). The Airy disk represents the intensity of light in the image plane and is defined, in terms of an arbitrary variable x, as

$$I(x) := I_0 \left[\frac{2J_1(x)}{x}\right]^2, \tag{5.1}$$

where I_0 is the maximum intensity and $J_1(x)$ is the Bessel function of the first kind of order one. A pair of incoherent point sources, whose combined image is the sum of their individual intensities, would produce a pair of overlapping Airy discs like the ones shown in Fig. 5.1(b-c). Resolution, then, has to do with the distance the two point sources must be apart in order for their central lobes to be distinguishable: the shorter this distance, the better the resolution. As we said, this is ambiguous, and somewhat arbitrary, as the two lobes can overlap partially while still being distinguishable in some sense. In any case, the resolution of a lens depends on the width of the central lobe in its PSF. For example, one common way to define resolution is the *Rayleigh criterion*, which says that two point sources are just resolved when the center of one source's Airy disk coincides with the first zero ring of the other source's Airy disk [78, pp. 472]. In other words, the resolvable distance in the Rayleigh criterion is the distance from the center of the PSF's central lobe to its first zero ring. The two objects in Fig. 5.1(b) satisfy the Rayleigh criterion and are resolvable – their central lobes overlap but can be reasonably distinguished. On the other hand, the two objects in Fig. 5.1(c) are too close together and are not resolvable; the two objects' central lobes combine into a single lobe in the image. In this chapter, the Rayleigh criterion is the definition of resolution distance that we will use. This definition obviously describes the width of the central lobe and is conceptually linked with our method from Chapter 4, which also depends on zeros.

In recent years, superoscillation has begun to be used as a method for achieving a superresolution lens. Here, we will give a few examples. The idea for a superoscillatory lens (SOL) was first introduced by Huang and Zheludev in 2009 [79]. They gave an example of a 1D lens consisting of an optical mask to modulate the intensity and phase of light to produce a superoscillatory region in the image plane. The lens was designed using prolate spheroidal wavefunctions, and was shown to be theoretically robust against manufacturing errors. Another SOL was designed by Rogers et al. [80] using a mask made of alternating rings of unity and zero transmittance of varying width. Using a scanning confocal setup with an incident wavelength of 640 nm, they demonstrated imaging superior to a conventional lens with a numerical aperture of 1.4. Another kind of SOL was made using a strategy called *optical eigenmodes*, which used SLMs to generate superoscillatory spots [81, 82, 83]. Yuan, Rogers, and Zheludev used an optical mask to make achromatic SOLs [84].

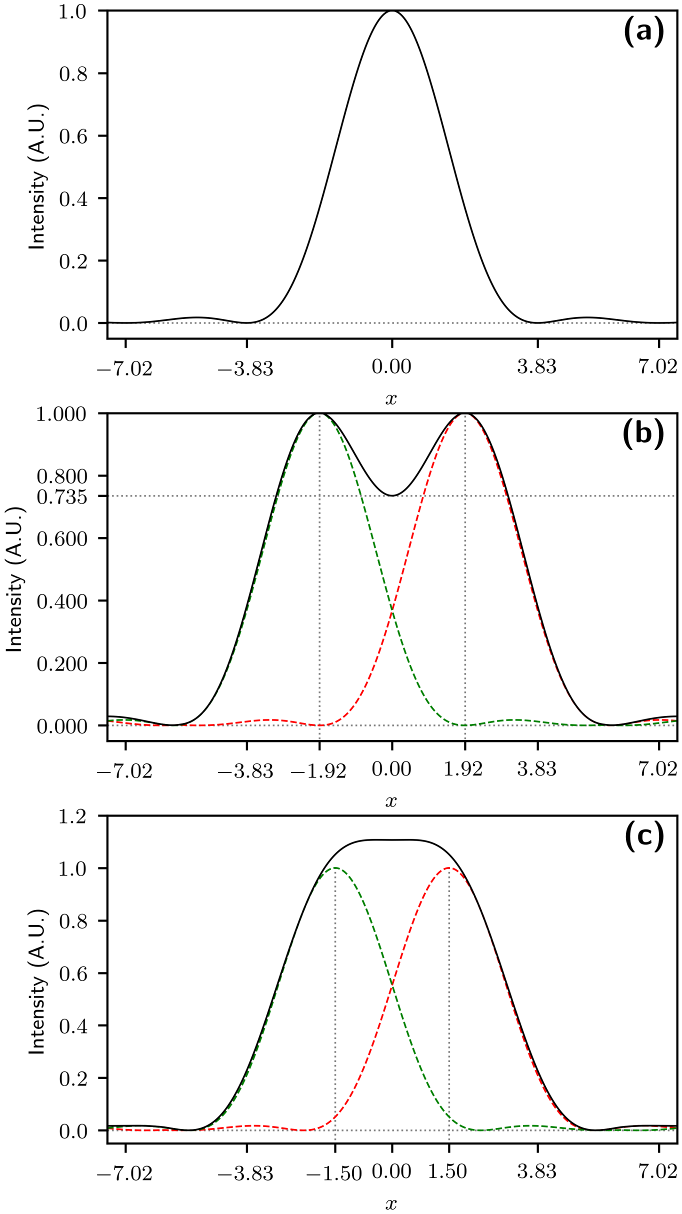

Figure 5.1: (a) Cross-section of an Airy disk. The ticks on x axis (other than zero) are the locations of zero rings. (b) A pair of Airy disks, resolvable according to the Rayleigh criterion. Dashed lines are the disks; the solid line is their combined image. The dotted lines are guides to the eye. (c) An unresolvable pair of Airy disks. The line styles have the same meaning as in (b).

As might be expected, one of the main challenges with using superoscillations for imaging is the sidelobes – if a pair of point sources is far enough apart, the central lobe of one will be overwhelmed by the sidelobe of the other, and vice versa. This means that, in practice, a SOL can only image a small area at a time, typically using a scanning confocal setup [80, 83]. So the two goals for designing a superresolution lens using superoscillations are to have good resolution (narrow central lobe) and to keep the sidelobes as far from the central lobe as possible. We will show that our method can address both of those goals.

5.3 Our Superoscillatory Lens Method

A sketch of our SOL setup is shown in Fig. 5.2. We have an ordinary thin lens contained within an aperture, with a phase mask on the object plane side to modulate the incoming light. The modulation from the phase mask is what will produce the superoscillation effect in the image plane. In principle, the lens and the mask can be combined. The image plane is at a distance d_I from the lens plane, and the object plane is at a distance d_o from the lens plane. The image distance, object distance, and focal length f of the lens obey the lens law:

$$\frac{1}{f} = \frac{1}{d_o} + \frac{1}{d_I}. \tag{5.2}$$

For the moment, we will only consider a point object located at the origin of the object plane, in order to obtain the SOL's point-spread function.

Our point object is illuminated by monochromatic light of wavelength λ_0. Light from the object propagates along the z axis through the lens and into the image plane. By Fresnel diffraction [21, pp. 351], the field in the image plane will be

$$U(\mathbf{r}_I, d_I) = \mathrm{e}^{\mathrm{i}k_0 d_I} \frac{\mathrm{i}}{\lambda_0 d_I} \iint_A U_0(\mathbf{r}_L)\, l(\mathbf{r}_L)\, t(\mathbf{r}_L)\, \mathrm{e}^{\mathrm{i}\frac{k_0}{2d_I}|\mathbf{r}_I - \mathbf{r}_L|^2} \mathrm{d}^2\mathbf{r}_L \tag{5.3}$$

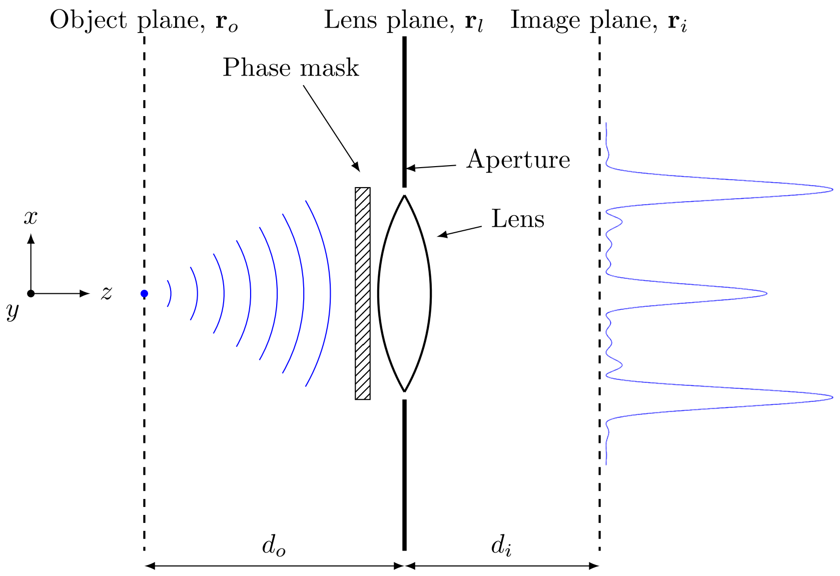
Object plane, $\mathbf{r}_o$
Lens plane, $\mathbf{r}_l$
Image plane, $\mathbf{r}_i$
Phase mask
Aperture
Lens
x
y
z
d_o
d_i

Figure 5.2: Superoscillatory lens setup with a point object.

where $\mathbf{r}_I$ is the x and y coordinates in the image plane, $k_0 = 2\pi/\lambda_0$ is the wavenumber of the light, A is the area of the lens, and $\mathbf{r}_L$ is the x and y coordinates in the lens plane. The function $U_0(\mathbf{r}_L)$ is the field from the object in the image plane,

$$U_0(\mathbf{r}_L) := U_0 \mathrm{e}^{\mathrm{i}\frac{k_0}{2d_o}|\mathbf{r}_o - \mathbf{r}_L|^2}, \tag{5.4}$$

where $\mathbf{r}_o$ is location of the point object in the object plane and U_0 is the field amplitude. The function $l(\mathbf{r}_L)$ is the lens function, modeling the effect the lens has on the wave, defined as [75, pp. 91]

$$l(\mathbf{r}_L) := \mathrm{e}^{-\mathrm{i}\frac{k_0}{2f}|\mathbf{r}_L|^2}. \tag{5.5}$$

The function $t(\mathbf{r}_L)$ is the transmission function of the phase mask, which we will define later. Putting this all together, in the image plane we have

$$U(\mathbf{r}_I, d_I) = \mathrm{e}^{\mathrm{i}k_0 d_I} \frac{\mathrm{i}}{\lambda_0 d_I} U_0 \iint_A \mathrm{e}^{\mathrm{i}\frac{k_0}{2d_o}|\mathbf{r}_o - \mathbf{r}_L|^2} t(\mathbf{r}_L)\, \mathrm{e}^{-\mathrm{i}\frac{k_0}{2f}|\mathbf{r}_L|^2} \mathrm{e}^{\mathrm{i}\frac{k_0}{2d_I}|\mathbf{r}_I - \mathbf{r}_L|^2} \mathrm{d}^2\mathbf{r}_L. \tag{5.6}$$

We will now show that, in the image plane, Eq. (5.6) is an aperture-limited Fourier transform operation on $t(\mathbf{r}_L)$.

Rearranging the exponents and using the vector property $|\mathbf{a}-\mathbf{b}|^2 = |\mathbf{a}|^2 + |\mathbf{b}|^2 - 2\mathbf{a}\cdot\mathbf{b}$, we obtain

$$U(\mathbf{r}_I, d_I) = \mathrm{e}^{\mathrm{i}k_0 d_I} \frac{\mathrm{i}}{\lambda_0 d_I} U_0 \iint_A t(\mathbf{r}_L)\, \mathrm{e}^{\mathrm{i}\frac{k_0}{2}\left(\frac{|\mathbf{r}_o|^2+|\mathbf{r}_L|^2-2\mathbf{r}_o\cdot\mathbf{r}_L}{d_o} - \frac{|\mathbf{r}_L|^2}{f} + \frac{|\mathbf{r}_I|^2+|\mathbf{r}_L|^2-2\mathbf{r}_I\cdot\mathbf{r}_L}{d_I}\right)} \mathrm{d}^2\mathbf{r}_L. \tag{5.7}$$

Using the lens law, Eq. (5.2), yields

$$U(\mathbf{r}_I, d_I) = \mathrm{e}^{\mathrm{i}k_0 d_I} \frac{\mathrm{i}}{\lambda_0 d_I} U_0 \iint_A t(\mathbf{r}_L)\, \mathrm{e}^{\mathrm{i}\frac{k_0}{2}\left(\frac{|\mathbf{r}_o|^2-2\mathbf{r}_o\cdot\mathbf{r}_L}{d_o} + \frac{|\mathbf{r}_I|^2-2\mathbf{r}_I\cdot\mathbf{r}_L}{d_I}\right)} \mathrm{d}^2\mathbf{r}_L. \tag{5.8}$$

Rearranging terms, we now have

$$U(\mathbf{r}_I, d_I) = \mathrm{e}^{\mathrm{i}k_0 d_I} \frac{\mathrm{i}}{\lambda_0 d_I} U_0 \mathrm{e}^{\mathrm{i}\frac{k_0}{2}\left(\frac{|\mathbf{r}_o|^2}{d_o} + \frac{|\mathbf{r}_I|^2}{d_I}\right)} \iint_A t(\mathbf{r}_L)\, \mathrm{e}^{-\mathrm{i}k_0\left(\frac{\mathbf{r}_o}{d_o}+\frac{\mathbf{r}_I}{d_I}\right)\cdot\mathbf{r}_L} \mathrm{d}^2\mathbf{r}_L, \tag{5.9}$$

We can now see that Eq. (5.9) is a Fourier transform of $t(\mathbf{r}_L)$. This can be made even more explicit by defining

$$\mathbf{k} := k_0 \left(\frac{\mathbf{r}_o}{d_o} + \frac{\mathbf{r}_I}{d_I}\right), \tag{5.10}$$

and we have

$$U(\mathbf{r}_I, d_I) = \mathrm{e}^{\mathrm{i}k_0 d_I} \frac{\mathrm{i}}{\lambda_0 d_I} U_0 \mathrm{e}^{\mathrm{i}\frac{k_0}{2}\left(\frac{|\mathbf{r}_o|^2}{d_o} + \frac{|\mathbf{r}_I|^2}{d_I}\right)} \iint_A t(\mathbf{r}_L)\, \mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{r}_L} \mathrm{d}^2\mathbf{r}_L, \tag{5.11}$$

which finally means that

$$U(\mathbf{r}_I, d_I) \propto \mathcal{F}\{t(\mathbf{r}_L)\}. \tag{5.12}$$

where $\mathcal{F}\{\}$ denotes the Fourier transform operation. We can thus define

$$U_I(\mathbf{r}_I) := \mathcal{F}\{t(\mathbf{r}_L)\} \tag{5.13a}$$

$$t(\mathbf{r}_L) = \mathcal{F}^{-1}\{U_I(\mathbf{r}_I)\}. \tag{5.13b}$$

where $\mathcal{F}^{-1}$ denotes the inverse Fourier transform and $U_I(\mathbf{r}_I)$ is equal to $U(\mathbf{r}_I, d_I)$, except with the constants in front of the integral omitted because they make no qualitative difference to the shape of the intensity in the image plane.

The Fourier relationship in Eq. (5.13) forms the basis of our SOL method. The transmission function $t(\mathbf{r}_L)$ is zero outside of the aperture, so it takes the place of the band limited function $\tilde{f}(k_x, k_y)$ from our vortex method in Chapter 4. The image field $U_I(\mathbf{r}_I)$ thus takes the place of $f(x, y)$, which we will again multiply by a polynomial $h(\mathbf{r}_I)$ to produce zero rings at a desired distance from the center of the image plane. This produces a modified image $U_I{}'(\mathbf{r}_I)$,

$$U_I{}'(\mathbf{r}_I) := h(\mathbf{r}_I)\, U_I(\mathbf{r}_I)\,. \tag{5.14}$$

Finally, we need a modified transmission function $t'(\mathbf{r}_L)$, which is the operation the phase mask needs to perform on the light to produce the desired field $U_I{}'(\mathbf{r}_I)$. We do this by taking the IFT of $U_I{}'(\mathbf{r}_I)$,

$$t'(\mathbf{r}_L) := \mathcal{F}^{-1}\{U_I{}'(\mathbf{r}_I)\}\,. \tag{5.15}$$

There is an important change to the polynomial, however: we want to create zero *rings* about the center of the image plane, rather than zero *points* in the image plane. That is, we want to set all points at a distance

$$|\mathbf{r}_I|^2 = \sqrt{x_I^2 + y_I^2} \tag{5.16}$$

from the center of the image plane to zero. This means our polynomial is now defined as

$$h(\mathbf{r}_I) := \prod_{n=1}^{N} \left(|\mathbf{r}_I|^2 - r_n^2 \right), \tag{5.17}$$

where r_n is the radius of the n^{th} zero ring that we add. Notice that we squared the terms in parentheses. This has to do with the Fourier transform. Recall that our polynomial will Fourier transform into derivatives. If we take the FT of $|\mathbf{r}_I|^2$, we have

$$\mathcal{F}\{|\mathbf{r}_I|^2\} = \mathcal{F}\{x_I^2 + y_I^2\} = \frac{\partial^2}{\partial x_L^2} + \frac{\partial^2}{\partial y_L^2}, \tag{5.18}$$

whereas if we took the FT of $|\mathbf{r}_I|$, we would formally expect, based on Eq. (4.8),

$$\mathcal{F}\{|\mathbf{r}_I|\} = \mathcal{F}\left\{\sqrt{x_I^2 + y_I^2}\right\} = \sqrt{\frac{\partial^2}{\partial x_L^2} + \frac{\partial^2}{\partial y_L^2}}. \tag{5.19}$$

However, this square root of derivative operators has no clear meaning, so we square the terms in Eq. (5.17). This also changes the required number of derivatives. Recall from our discussion of Eq. (4.7) that $\tilde{f}(k_x, k_y)$ must have at least $N - 1$ derivatives in order to avoid singularities, where N is the number of added zeros. Here, since our polynomial is made of squared terms, each added zero ring contributes two derivatives, rather than one. This means that, for N added zero rings, the transmission function $t(\mathbf{r}_L)$ must have at least $2N - 1$ continuous derivatives.

The choice of zero ring locations r_n is what will make our lens a superresolution lens. In this chapter, we are defining the resolution of a circular lens is the distance from the central lobe of the image to its first zero ring; we denote this distance Δr. As long as at least one r_n is less than Δr, the lens will have better resolution than the diffraction limit. In principle this would allow for arbitrarily high resolution, within the limits of Fourier optics, but as r_n gets smaller, we expect that the sidelobes would overwhelm the central lobe. Thus, choosing low r_n must be balanced by managing the distance and relative power of the sidelobes to the central lobe.

The steps of the SOL method are summarized in Algorithm 2.

Algorithm 2 Superoscillatory Lens Algorithm

```
Define a transmission function $t(\mathbf{r}_L)$ in the lens plane, band limited by the aperture.
Fourier transform to yield $U_I(\mathbf{r}_I)$ For a circular aperture, $U_I(\mathbf{r}_I)$ will have zero-rings at distance $\Delta r$ from the center.
Define $h(\mathbf{r}_I) := \prod_{n=1}^{N} \left(|\mathbf{r}_I|^2 - r_n^2\right)$, where $r_n < \Delta r$ for at least one $n$.
Define ${U_I}'(\mathbf{r}_I) := h(\mathbf{r}_I)\, U_I(\mathbf{r}_I)$, which will have sub-diffraction zero-rings wherever $r_n < \Delta r$.
Calculate $t'(\mathbf{r}_L) := \mathcal{F}^{-1}\{{U_I}'(\mathbf{r}_I)\}$, which is the transmission function needed in the lens plane to yield the desired pattern ${U_I}'(\mathbf{r}_I)$ in the image plane.
```

5.4 Algorithm Demonstrations

In this section we will demonstrate the validity of our SOL method by way of a few examples. Our first task is to choose a lens to improve its resolution. The parameters of the lens we chose and the rest of the setup are given in Table 5.1. As we are only aiming to demonstrate the validity of our method, the choice of lens is somewhat arbitrary at this point. We chose this particular lens because of its small value of Δr. The magnification was chosen because that power is often used in nanolithography, which could (eventually) be a possible application for a system like this. The object and image distances were calculated to yield that magnification. The wavelength was chosen because it is small but still part of the visible portion of the spectrum.

Table 5.1: Parameters of the simulated lens setup.

Lens diameter D	1.27 cm
Lens radius a	0.635 cm
Focal length f	13 mm
Magnification m	$^1/_4$
Object distance d_o	65 mm
Image distance d_I	16.25 mm
Light wavelength λ_0	385 nm

Let us examine the PSF of our lens before we begin improving on it. The transmission function $t(\mathbf{r}_L)$ we use for that is

$$t(\mathbf{r}_L) = \begin{cases} 1 & r_L \leq a \\ 0 & r_L > a, \end{cases} \tag{5.20}$$

which is just the transmission function if the phase mask were not there. This can be seen in Fig. 5.3(a-b). The PSF of this lens is shown in Fig. 5.3(c-d), from which we can see that, for this unmodified lens, $\Delta r^{\text{lens}} \approx 600\,\text{nm}$, where Δr^{lens} is the resolution (the Δr) of this lens specifically. This is the resolution we will try to improve upon in this chapter.

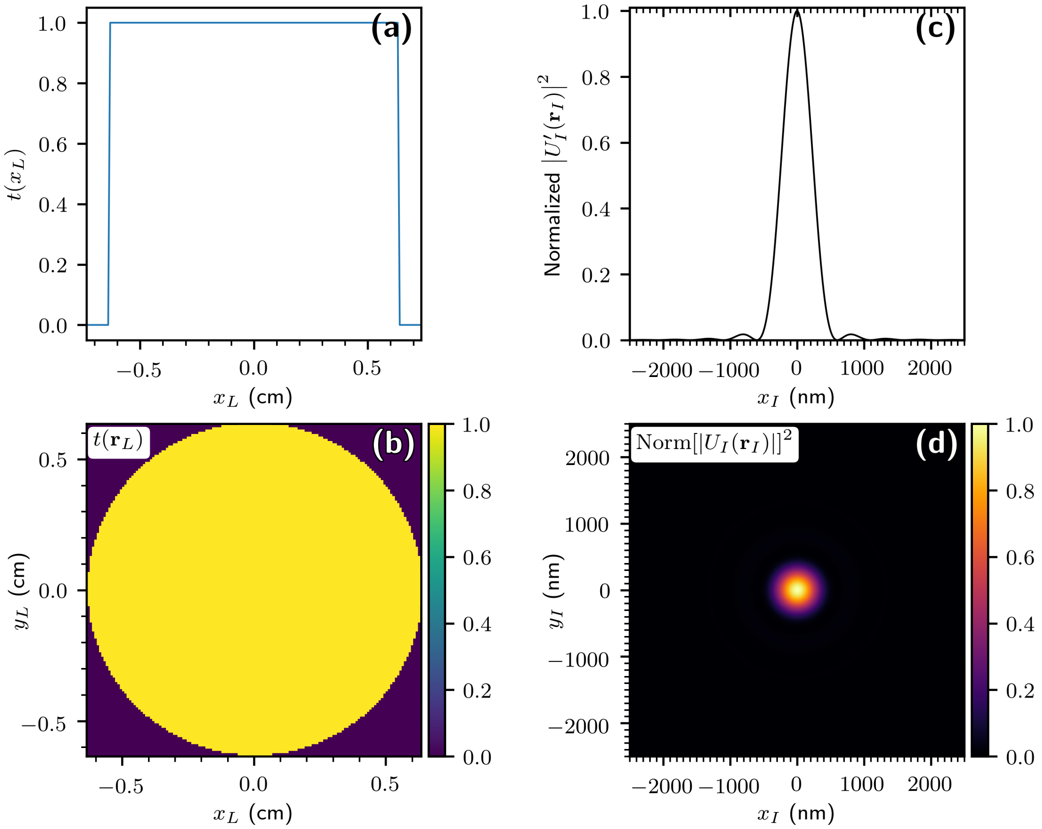

Figure 5.3: Transmission function and PSF of unmodified lens. (a) x axis slice of transmission function $t(\mathbf{r}_L)$. (b) Color map of $t(\mathbf{r}_L)$. (c-d) x axis slice and color map, respectively, of PSF $U_I(\mathbf{r}_I)$.

To help the reader know quickly which plots are in the lens plane and which are in the image plane, from now on our figures will follow the color schemes used in Fig. 5.3. Lens (image) line plots will contain blue (black) lines, and lens (image) color plots will use the blue-green-yellow (black-red-yellow) color map.

Now that we have seen the regular lens's flat transmission function and PSF, we can look at some strategies for making it into a superresolution lens using our method.

5.4.1 Adding zeros to central lobe

Following our algorithm, we will add a zero ring at a distance $r_n < \Delta r^{\text{lens}}$. But first, we need to define a starting transmission function $t(\mathbf{r}_L)$. We cannot use the flat transmission function of the lens, because its hard edges will produce discontinuous derivatives. We decide to use a function similar to $\tilde{f}_{\text{circ}}(k_x, k_y)$ from Eq. (4.10):

$$t(\mathbf{r}_L) = \begin{cases} \cos^{10}\left(\frac{\pi}{2a}\mathbf{r}_L\right) & r_L \leq a \\ 0 & \text{o.w..} \end{cases} \tag{5.21}$$

Since it is raised to the power 10, this transmission function can support up to 5 zero rings. For this transmission function, $\Delta r \approx 3030\,\text{nm}$, determined numerically. We will now use Fig. 5.4 to walk through our method in its entirety.

In Fig. 5.4(a-b) we plot an x axis slice and a color plot, respectively, of the cosine transmission function in Eq. (5.21). In Fig. 5.4(c), we plot an x axis slice of the unmodified PSF, $U_I(\mathbf{r}_I)$, of a point source through this lens (dashed orange line) and the modified PSF, $U_I{}'(\mathbf{r}_I)$, after having added a zero to it. We also plot the PSF from the unmodified lens, $U_I^{\text{lens}}(\mathbf{r}_I)$, for comparison. For the modified PSF, we have added a zero ring at $r_L = 500\,\text{nm}$, which is less than Δr^{lens}. We can see that there is a new zero at that location; we can also see a pair of large sidelobes, as expected. We can also see that the central lobe's intensity is a little under 40% of that of the sidelobes. Figure 5.4(d) shows the modified PSF (note in the text box that "Norm" denotes a normalization operator). In Fig. 5.4(e-f) we take the IFT of $U_I{}'(\mathbf{r}_I)$ to obtain the modified transmission function, $t'(\mathbf{r}_L)$, that would be needed in the lens plane in order to produce our modified PSF with superoscillatory zeros. We can see that it is partially negative. This is okay, as negative $t(\mathbf{r}_L)$ corresponds to a phase shift of π. On a technical note, we manually set all points outside $r_L \leq a$ to zero[1]. In Fig. 5.4(g), we show the PSF of the modified transmission function in (e-f), denoted ${U_I}'_2(\mathbf{r}_I)$ (black line) and the PSF ${U_I}'(\mathbf{r}_I)$ from In Fig. 5.4(c) (orange dashed line). The purpose of this plot is to see whether the transmission function in (e-f) can actually produce the desired image from Fig. 5.4(c). The two lines are in excellent agreement. In Fig. 5.4(h) we show a color plot of ${U_I}'_2(\mathbf{r}_I)$; we can see that it matches the PSF from Fig. 5.4(d).

Altogether, Fig. 5.4 shows that our method can produce a transmission function which will have greater resolution than that of the starting lens. However, the sidelobes in this case are both large and quite close to the central lobe. We will turn to addressing that problem now.

On a side note, we will refrain from using the 2D color plots in the future, as they do not add much information not contained in the line plots.

5.4.2 Adding zeros to sidelobes

We can attempt to mitigate the sidelobes by placing zeros on them. That is, we can place a zero near the peak of a sidelobe to split it into smaller, more manageable sidelobes. We demonstrate this in Fig. 5.5, using the same setup as was used for Fig. 5.4. In Fig. 5.5(a), we added a zero at $r_L = 500\,\mathrm{nm}$, just like before. We can see that the sidelobes are centered at around 1230 nm. In Fig. 5.5(b), we have added a second zero, at 1370 nm. We can see that the resulting sidelobes are now actually smaller than the central lobe. We can attempt to reduce these by adding more zeros

[1] That step technically shouldn't be mathematically necessary, as $t'(\mathbf{r}_L)$ should have the same band limits as $t(\mathbf{r}_L)$, from Eq. (4.7). However, the numerical FFT algorithm can create very small but non-zero values outside the band limit. We set them to zero just to avoid them having any effect in future steps.

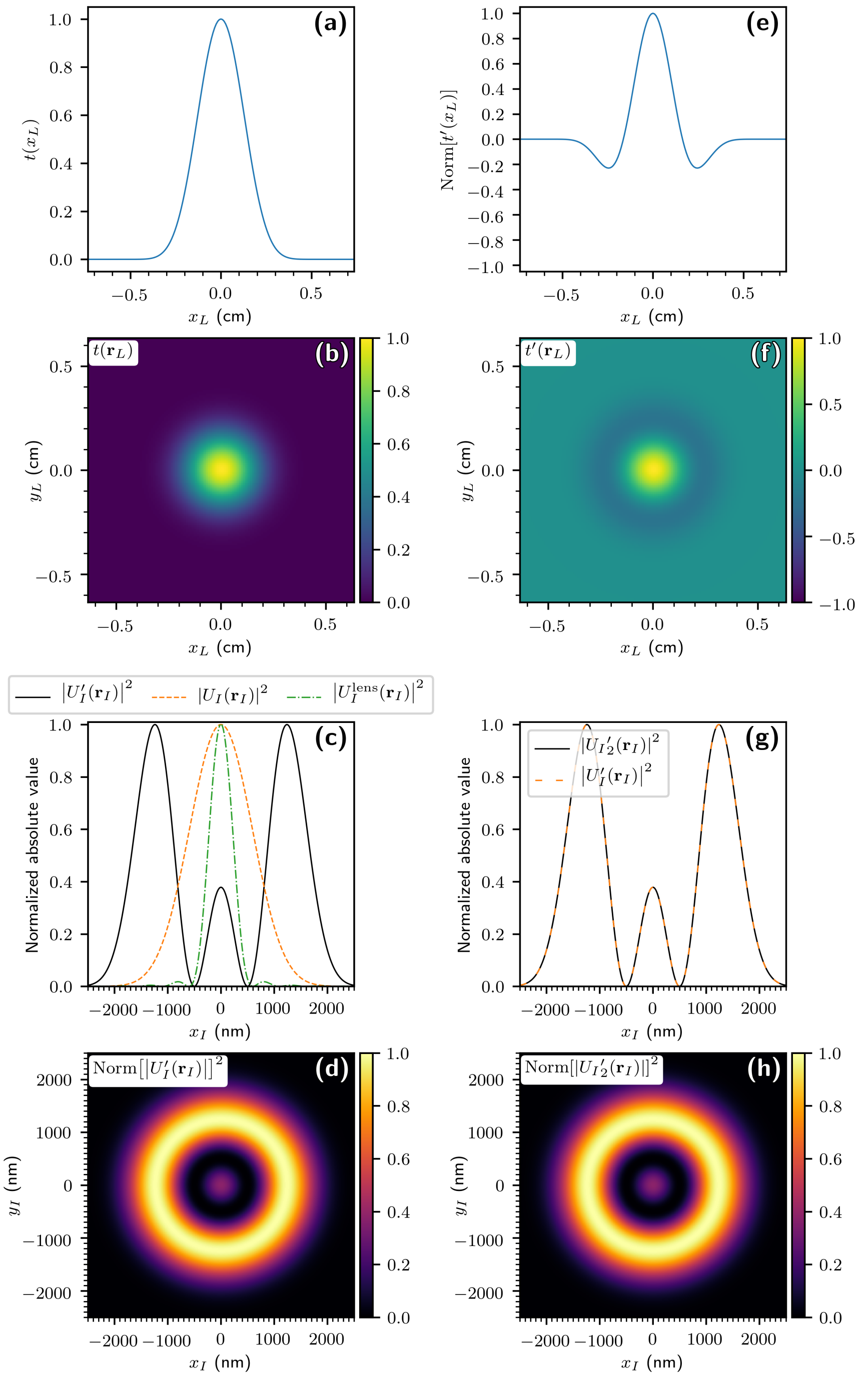

Figure 5.4: Demonstrating the SOL procedure by adding zeros to central lobe of the image produced by a band limited cosine transmission function. See text for a description of each subfigure.

at their peaks. We show this in Fig. 5.5(c), where we have added zeros at 900 nm and 1900 nm. The sidelobes next to the central peak are now greatly reduced. There are also new sidelobes at about 2400 nm, which are again larger than the central lobe. However, our central lobe is now about 70% of their intensity, which is an improvement from Fig. 5.5(a). Furthermore, these large sidelobes are much farther from the central lobe than they were in Fig. 5.5, which should lead to a larger viewing area. Let us now walk through our SOL method to obtain the modified transmission function needed to produce this PSF.

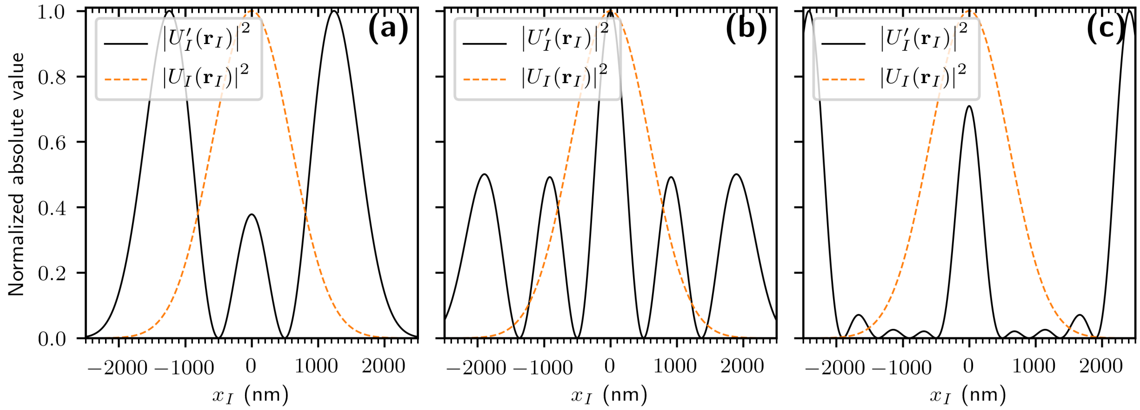

Figure 5.5: A $\cos^{10}(\pi \mathbf{r}_L/2a)$ SOL, gradually adding zeros. For this transmission function, we have $\Delta r \approx 3030$ nm. (a) A zero has been added at $r_I = 500$ nm. (b) A second zero has been added at $r_I = 1370$ nm to suppress the large sidelobes in (a). (c) Two more zeros have been added at $r_I = 900$ nm and 1900 nm to suppress the sidelobes in (b).

In Fig. 5.6 we walk through our SOL method to produce the PSF shown in Fig. 5.5(c). In Fig. 5.6(a) we show the original transmission function. In Fig. 5.6(b) we show the unmodified PSF and the modified PSF with added zeros. We also show the PSF of the original lens. We can see that, not only has the resolution been improved, but the first sidelobes of ${U_I}'(\mathbf{r}_I)$ are of comparable height to the first sidelobes of $U_I^{\text{lens}}(\mathbf{r}_I)$. In Fig. 5.6(c) we show the transmission function needed to generate the modified PSF in Fig. 5.6(b). This a plausibly realizable transmission function. We have again manually set all points of $t'(\mathbf{r}_L)$ outside $r_L \leq a$ to zero. In Fig. 5.6(d), we find the PSF of the transmission function in Fig. 5.6(c) and compare it to PSF in Fig. 5.6(b). We can see that they match, so this transmision function works for making a SOL with resolution of 500 nm, which is better than the 600 nm resolution of the original lens.

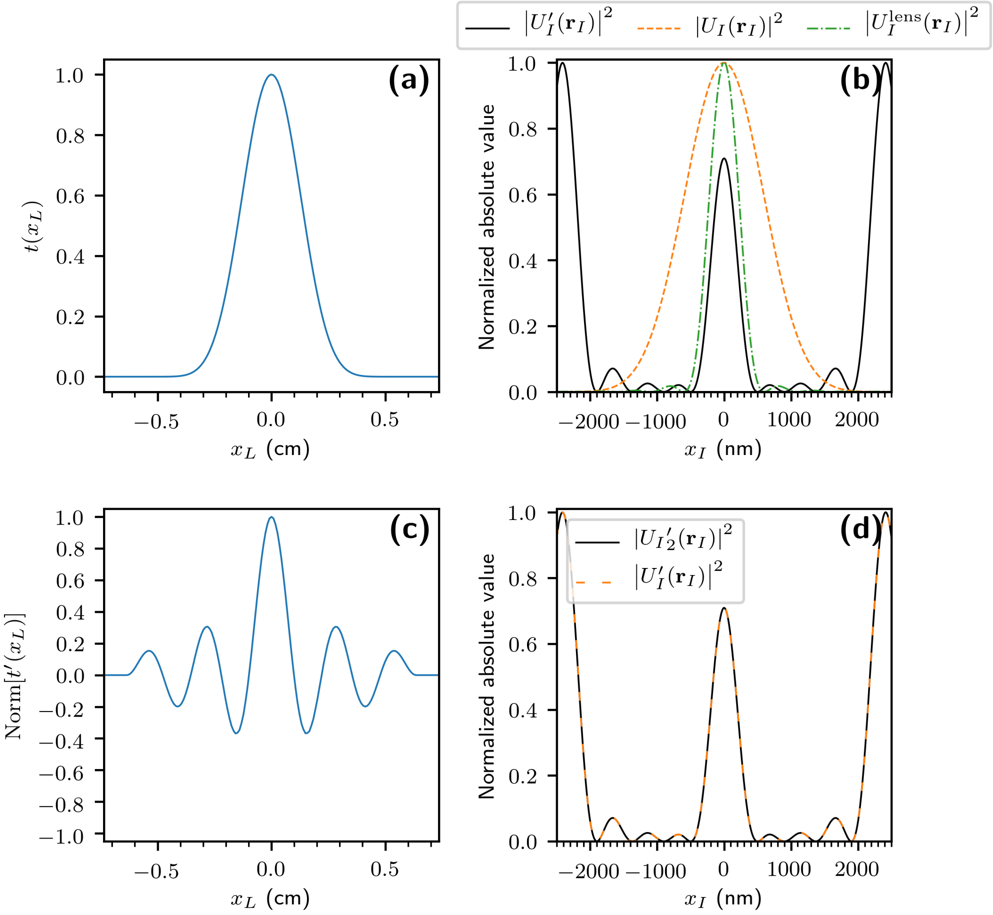

Figure 5.6: A $\cos^{10}(\pi \mathbf{r}_L/2a)$ SOL, with four zeros added. (a) Unmodified lens transmission function. (b) PSF with and without added zeros (black and orange lines, respectively) and PSF of the original lens. (c) Modified transmission function in lens to yield the PSF in (b). (d) Comparing the PSF of the transmission function in (c) (black line) against the PSF from (b) (orange dashed line).

As a further demonstration of the improved resolution of this SOL, in Fig. 5.7 we compare the image formed by two point objects using our SOL against the original lens. The point objects are positioned such that their images are 500 nm apart, which means their central peaks coincide with the other's first zero ring. We can see in

Fig. 5.7(a) that the two are resolvable with the SOL, while in Fig. 5.7(b) with the original lens they barely are, if at all. We also note that, while the image in Fig. 5.7(a) satisfies the Rayleigh criterion, they are less resolvable than the Airy disks back in Fig. 5.1(b). For the two Airy disks in Fig. 5.1(b), the minimum between the two peaks is about 74% of the neighboring maxima; in Fig. 5.7(a) that minimum is about 81% of the neighboring maxima. This suggests that, for smaller-radius zero rings, caution must be taken in simply using the radius as a definition of resolution.

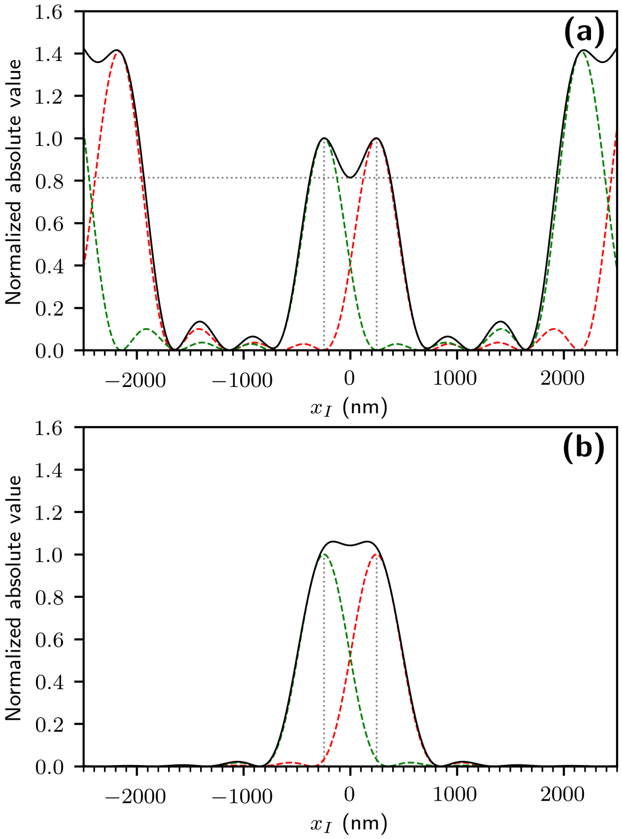

Figure 5.7: Image formed by two point sources using our SOL and the original lens. Intensities are normalized to the central PSF lobe. (a) Image formed by two point objects using our SOL from Fig. 5.6, resolvable according to the Rayleigh criterion. Dashed lines are the individual PSFs; the solid line is their combined image. The dotted lines are guides to the eye. (b) The same point objects viewed through the original lens, which are not as resolvable. The line styles have the same meaning as in (c).

5.5 Adding zeros to Δr

So far we have attempted to achive superresolution by adding zeros to the PSF's central lobe and to resulting sidelobes. There is an alternative strategy that our method can be used for: adding zeros to Δr. The lens with chosen transmission function will naturally have a zero at Δr; by adding more zeros to Δr using our method, we can make this into a higher-order zero. Doing this should narrow the central lobe, thus improving the resolution [2]. One appeal of this method is that it gives us a "free" zero. That is, the original zero does not produce a derivative after Fourier transform, so it doesn't count against our $2N-1$ limit.

We demonstrate this in Fig. 5.8. We again use the cosine transmission function defined in Eq. (5.21), shown in Fig. 5.8(a). In Fig. 5.8(b), we add five zeros to Δr, making that point a sixth-order zero. Figure 5.8(c) and (d) show the modified transmission function $t'(\mathbf{r}_L)$ and that this function successfully reconstructs the desired PSF $U_{I2}'(\mathbf{r}_I)$. We can see from Fig. 5.8(b) that the resolution from this method is far worse than that of the unmodified lens, and only marginally better than the original PSF from before the zeros were added. However, this need not be considered a failure of this method. The cosine transmission function we are using produces a very wide PSF. If a wider transmission function were used, which produced a narrower PSF, then adding zeros to Δr would be more likely to be successful.

5.6 Intensity Considerations

Up to this point, we have only been considering our PSF's shape. It is obviously important to also consider the intensity as we add zeros and how the intensity compares to the unmodified lens.

In Fig. 5.9 we plot the PSFs with added zeros from Fig. 5.5 on the same vertical

[2] Strictly following the definition of resolution that we have been using in this chapter, the resolution won't have improved because we haven't brought any zeros closer to the center of the image plane. It should be clear, however, that narrowing the central lobe will make it more resolvable with a lobe from a second point object, so this is an improvement in resolution.

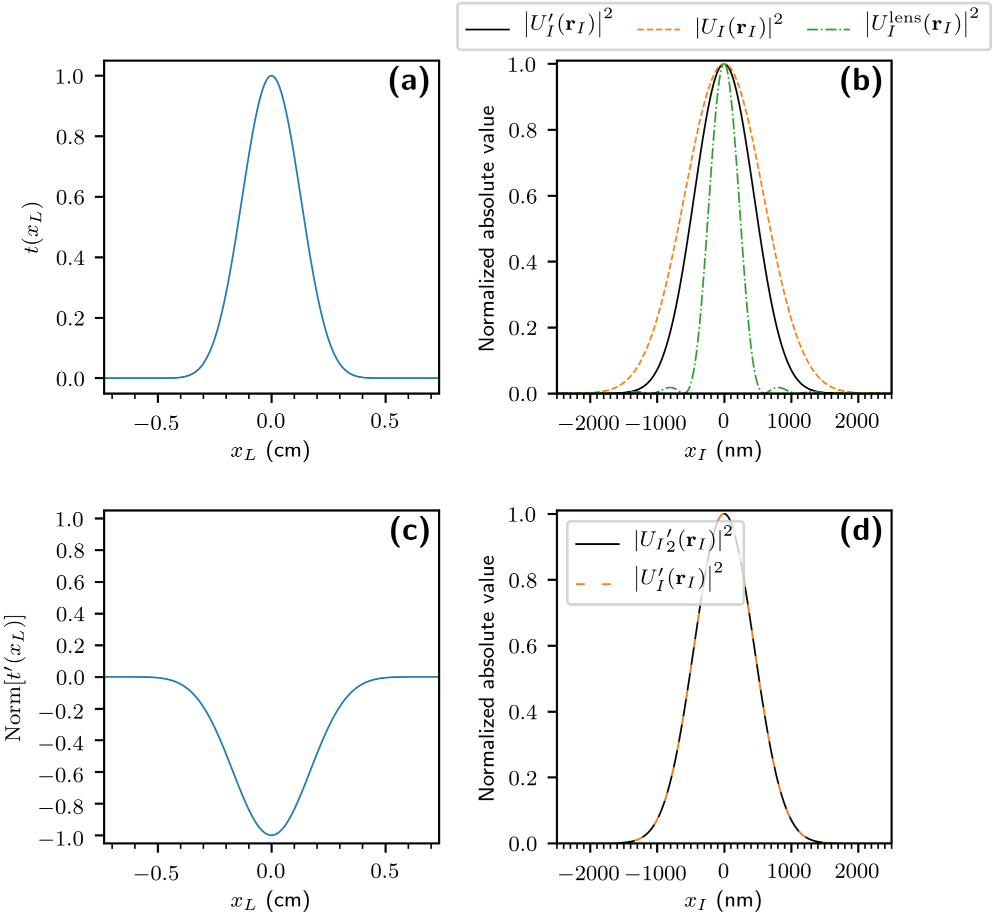

Figure 5.8: A $\cos^{10}(\pi \mathbf{r}_L/2a)$ SOL, with five zeros added at Δr. (a) Unmodified lens transmission function. (b) PSF with and without added zeros (black and orange lines, respectively) and the PSF of the original lens. (c) Modified transmission function in lens to yield the image in (b). (d) Comparing the PSF of the transmission function in (c) (black line) against the PSF from (b) (orange dashed line).

scale. We have normalized them to the maximum intensity of the unmodified lens's PSF's central lobe; we denote this intensity $I_{\text{center}}^{\text{lens}}$. We can see that the SOL intensities are about two orders of magnitude lower than that of the unmodified lens. This is to be expected, since the transmission functions needed to produce the SOL will transmit significantly less light. We can also see that the PSF with two added zeros has an intensity about twice as great as that of the PSF with one or four added zeros. This is because the transmission function associated with this PSF allows more total power through. One might intuitively expect that adding zeros would always decrease total intensity, but this result shows that is not always the case.

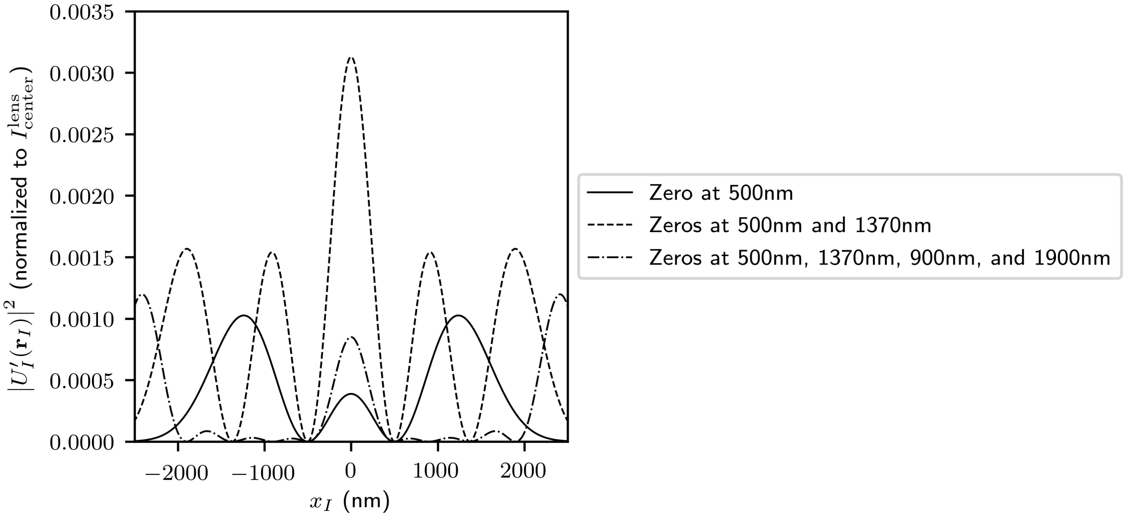

Figure 5.9: Intensities of the PSFs with added zeros from Fig. 5.5, normalized to $I_{\text{center}}^{\text{lens}}$, the maximum intensity of the central lobe of the unmodified lens's PSF.

In Fig. 5.9 we plot the relative intensities of the central lobe and first sidelobe of an SOL ($I_{\text{center}}^{\text{SOL}}$ and $I_{\text{side}}^{\text{SOL}}$, respectively) as a function the radius of a single zero ring. We also plot the ratio of the SOL's central lobe to its sidelobe. We can see in Fig. 5.10(a) and (b) that the relative intensities increase as the zero ring radius increaeses, until they both kink at about 935 nm. After this point, the intensity of the sidelobe decreases and the intensity of the central lobe increases less rapidly. We are not sure what happens at 935 nm to make this happen. In Fig. 5.10(c), we see that, as expected, very small ring radii make the central lobe exponentially smaller than the sidelobe.

5.7 Alternative Lens Transmission Functions

The cosine transmission function defined in Eq. (5.21) works for our SOL, but it has a problem. It has high values only in a small area near the center, and its value is near zero for a large portion of its area. Additionally, it is a fairly narrow shape, which means that it will produce a fairly wide PSF. Ideally, we would like a function that is relatively flat over most of its area and tapers down to zero at the edge, but

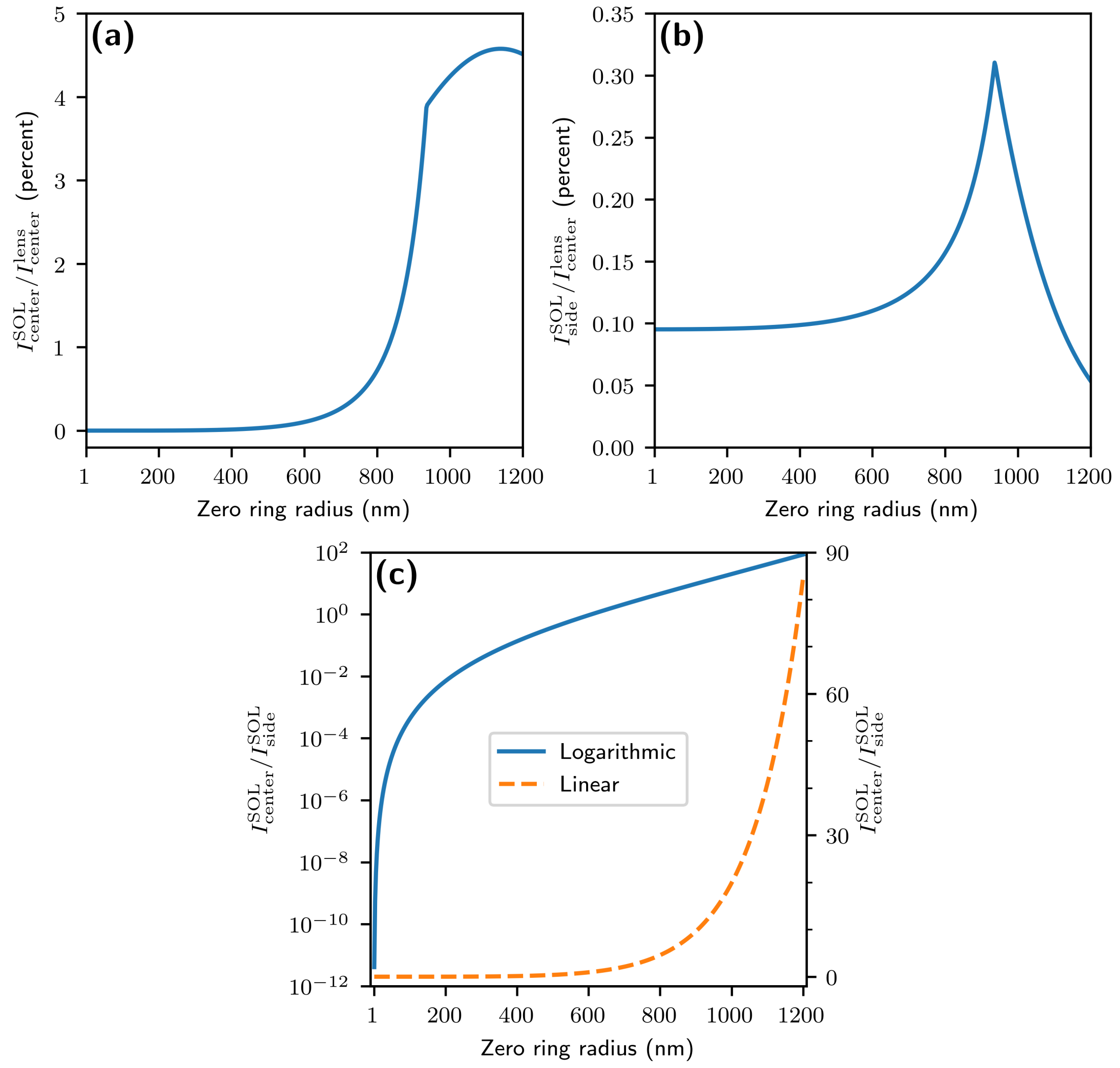

Figure 5.10: SOL lobe intensities as a function of zero ring radius. (a) Ratio of SOL central lobe max intensity, $I^{\text{SOL}}_{\text{center}}$, to unmodified lens max intensity, $I^{\text{lens}}_{\text{center}}$. (b) Ratio of SOL sidelobe max intensity, $I^{\text{SOL}}_{\text{side}}$, to unmodified lens max intensity. (c) Ratio of SOL central lobe max intensity to sidelobe max intensity. For (a) and (b), note that $I^{\text{lens}}_{\text{center}}$ is a constant.

which has sufficient derivatives for us to place new zeros in the image plane. (There is a trade-off, though: a narrow PSF will have sidelobes close to the center lobe, likely reducing the viewing area.) While we haven't found such a function yet, in this section we discuss some special functions that we tried and discuss their problems.

First we tried Riemann theta functions, specifically $\theta_3(x,q)$ [85],

$$\theta_3(x,q) = 1 + 2\sum_{n=1}^{\infty} q^{n^2}\cos(2nx)\,. \tag{5.22}$$

One is plotted in Fig. 5.11 (after having done some rotating, shifting, etc. to get it into that orientation). It failed to support more than one zero. (By that, we mean $t(x_L)$ had singular derivatives and/or that ${U_I}'(x_I)$ could not be reconstructed from $t'(x_l)$.) That's probably because the function is defined as a sum of cosines. As we know, a cosine function raised to the power n only has n derivatives after band limiting. Since this function is a sum of cosines, we effectively only have $n = 1$, so it doesn't work.

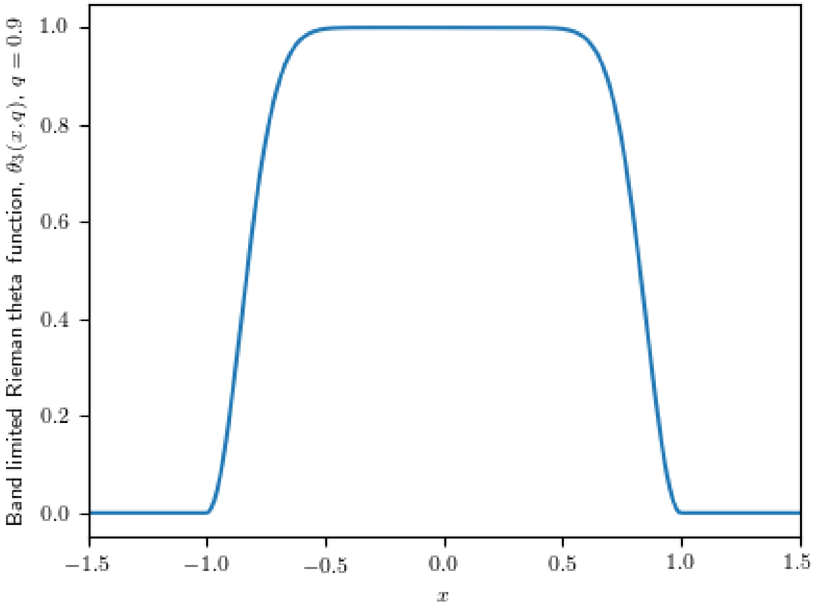

Figure 5.11: Band limited Riemann theta function $\theta_3(x,q)$, with $q = 0.9$, after rotating, shifting, etc.

Next, we tried a Jacobian elliptic function [86],

$$\mathrm{dn}(x,k) = \frac{\theta_4(0,q)}{\theta_3(0,q)}\frac{\theta_3(\zeta,q)}{\theta_4(\zeta,q)} \tag{5.23}$$

where

$$\zeta = \frac{\pi x}{2K(k)} \tag{5.24}$$

$$K(k) = \frac{\pi}{2}\theta_3^2(0,q)\,. \tag{5.25}$$

One is plotted in Fig. 5.12 (after having done some rotating, shifting, etc. to get it into that orientation). It also failed to support more than one zero. This function is defined in terms of Riemann theta functions, so it likely failed for the same reason.

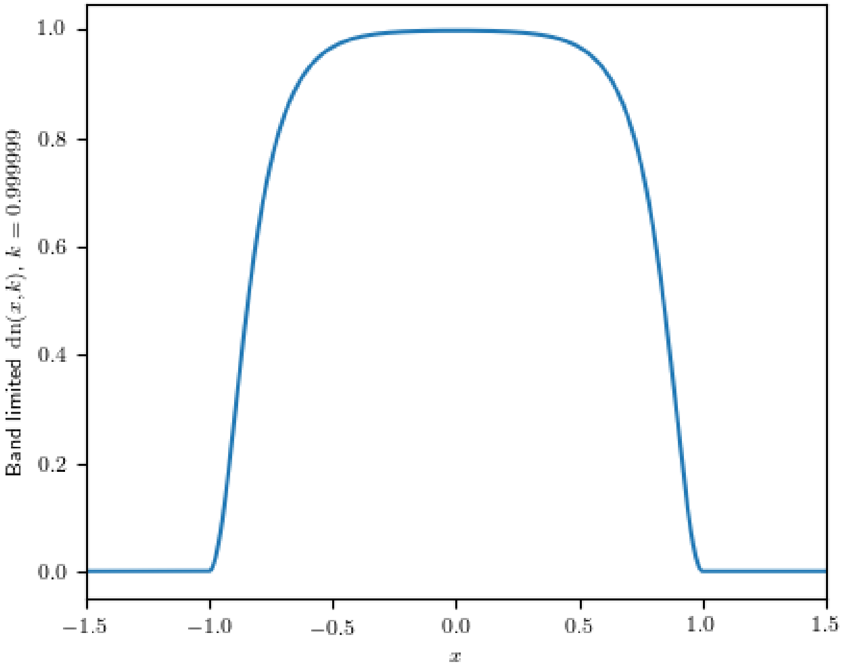

Figure 5.12: Band limited Jacobian elliptic function dn(x,k), with $k = 0.999999$, after rotating, shifting, etc.

5.8 Conclusions

In this chapter, we have described a way to use our zero-creation method from Chapter 4 to design a superoscillatory lens. We have shown by simulated examples that it can produce zeros rings closer to the center of the point spread function than are those of an unmodified lens, beating the diffraction limit. We discuss a superresolution strategy involving adding zeros next to the central lobe, which could possibly work for a different transmission function from the one we used. We also discussed the ideal criteria of a transmission function and our attempts to find such a function.

REFERENCES

[1] E. Wolf, *Introduction to the Theory of Coherence and Polarization of Light.* Cambridge University Press, 2007.

[2] L. Mandel and E. Wolf, *Optical Coherence and Quantum Optics.* Cambridge University Press, 2013.

[3] M. Born and E. Wolf, *Principles of Optics.* Cambridge University Press, 7 ed., 1999.

[4] E. Wolf, "Coherence and radiometry," *J. Opt. Soc. Am.*, vol. 68, pp. 6–17, Jan 1978.

[5] D. F. V. James, "Change of polarization of light beams on propagation in free space," *J. Opt. Soc. Am. A*, vol. 11, pp. 1641–1643, May 1994.

[6] E. Wolf, "Correlation-induced changes in the degree of polarization, the degree of coherence, and the spectrum of random electromagnetic beams on propagation," *Opt. Lett.*, vol. 28, pp. 1078–1080, July 2003.

[7] E. Wolf and D. F. V. James, "Correlation-induced spectral changes," *Rep. Prog. Phys*, vol. 59, no. 6, pp. 771–778, 1996.

[8] E. Wolf, "Recollections of Max Born," *Astrophysics and Space Science*, vol. 227, pp. 277–297, May 1995.

[9] D. K. Gramotnev and S. I. Bozhevolnyi, "Nanofocusing of electromagnetic radiation," *Nature Photonics*, vol. 8, no. 1, pp. 13–22, 2014.

[10] W. R. Wong, O. Krupin, S. D. Sekaran, F. R. Mahamd Adikan, and P. Berini, "Serological diagnosis of dengue infection in blood plasma using long-range surface plasmon waveguides," *Analytical Chemistry*, vol. 86, no. 3, pp. 1735–1743, 2014.

[11] O. Krupin, C. Wang, and P. Berini, "Selective capture of human red blood cells based on blood group using long-range surface plasmon waveguides," *Biosensors and Bioelectronics*, vol. 53, no. Supplement C, pp. 117–122, 2014.

[12] B. Schwarz, P. Reininger, D. Ristanic, H. Detz, A. M. Andrews, W. Schrenk, and G. Strasser, "Monolithically integrated mid-infrared lab-on-a-chip using plasmonics and quantum cascade structures," *Nature Communications*, vol. 5, p. 4085, 06 2014.

[13] J. Li, S. K. Cushing, P. Zheng, F. Meng, D. Chu, and N. Wu, "Plasmon-induced photonic and energy-transfer enhancement of solar water splitting by a hematite nanorod array," *Nature Communications*, vol. 4, p. 2651, 10 2013.

[14] K. Xiong, D. Tordera, G. Emilsson, O. Olsson, U. Linderhed, M. P. Jonsson, and A. B. Dahlin, "Switchable plasmonic metasurfaces with high chromaticity containing only abundant metals," *Nano Letters*, vol. 17, no. 11, pp. 7033–7039, 2017.

[15] K. L. van der Molen, K. J. Klein Koerkamp, S. Enoch, F. B. Segerink, N. F. van Hulst, and L. Kuipers, "Role of shape and localized resonances in extraordinary transmission through periodic arrays of subwavelength holes: Experiment and theory," *Phys. Rev. B*, vol. 72, p. 045421, Jul 2005.

[16] C. H. Gan, G. Gbur, and T. D. Visser, "Surface plasmons modulate the spatial coherence of light in Young's interference experiment," *Phys. Rev. Lett.*, vol. 98, p. 043908, January 2007.

[17] C. H. Gan, Y. Gu, T. D. Visser, and G. Gbur, "Coherence converting plasmonic hole arrays," *Plasmonics*, vol. 7, no. 2, pp. 313–322, 2012.

[18] G. Gbur and O. Korotkova, "Angular spectrum representation for the propagation of arbitrary coherent and partially coherent beams through atmospheric turbulence," *J. Opt. Soc. Am. A*, vol. 24, pp. 745–752, March 2007.

[19] A. Dogariu and S. Amarande, "Propagation of partially coherent beams: turbulence-induced degradation," *Opt. Lett.*, vol. 28, pp. 10–12, January 2003.

[20] S. Bozhevolnyi and V. Coello, "Elastic scattering of surface plasmon polaritons: Modeling and experiment," *Physical review. B, Condensed matter*, vol. 58, pp. 10899–10910, October 1998.

[21] G. J. Gbur, *Mathematical Methods for Optical Physics and Engineering.* Cambridge University Press, 2011.

[22] C. F. Bohren and D. R. Huffman, *Absorption and Scattering of Light by Small Particles.* John Wiley & Sons, 1983.

[23] J. D. Jackson, *CLassical Electrodynamics.* John Wiley & Sons, second ed., 1975.

[24] L. L. Foldy, "The multiple scattering of waves I. general theory of isotropic scattering by randomly distributed scatterers," *Physical Review*, vol. 67, pp. 107–119, February 1945.

[25] M. Lax, "Multiple scattering of waves. II. the effective field in dense systems," *Physical Review*, vol. 85, pp. 621–629, February 1952.

[26] P. G. Etchegoin, E. C. Le Ru, and M. Meyer, "An analytic model for the optical properties of gold," *J. Chem. Phys.*, vol. 125, no. 16, 2006.

[27] P. G. Etchegoin, E. C. Le Ru, and M. Meyer, "Erratum: "an analytic model for the optical properties of gold" [j. chem. phys.125, 164705 (2006)]," *J. Chem. Phys.*, vol. 127, no. 18, 2007.

[28] P. Johnson and R. Christy, "Optical constant of noble metals," *Physical Review B*, vol. 6, no. 12, pp. 4370–4379, 1972.

[29] H. Raether, *Surface Plasmons on Smooth and Rough Surfaces and on Gratings.* Springer-Verlag Berlin Heidelberg GmbH, 1988.

[30] M. Smith and G. Gbur, "Coherence resonances and band gaps in plasmonic hole arrays," *Phys. Rev. A*, vol. 99, p. 023812, Feb 2019.

[31] R. H. Dicke, "Coherence in spontaneous radiation processes," *Phys. Rev.*, vol. 93, pp. 99–110, Jan 1954.

[32] C. Ropers, D. J. Park, G. Stibenz, G. Steinmeyer, J. Kim, D. S. Kim, and C. Lienau, "Femtosecond light transmission and subradiant damping in plasmonic crystals," *Phys. Rev. Lett.*, vol. 94, p. 113901, Mar 2005.

[33] B. Luk'yanchuk, N. I. Zheludev, S. A. Maier, N. J. Halas, P. Nordlander, H. Giessen, and C. T. Chong, "The Fano resonance in plasmonic nanostructures and metamaterials," *Nature Materials*, vol. 9, pp. 707–715, SEP 2010.

[34] M. Berry, "Faster than Fourier," in *Quantum Coherence and Reality; in Celebration of the 60th Birthday of Yakir Aharonov* (J. S. Anandan and J. L. Safko, eds.), pp. 55–65, World Scientific, Singapore, 1994.

[35] M. K. S. Smith and G. J. Gbur, "Construction of arbitrary vortex and superoscillatory fields," *Opt. Lett.*, vol. 41, pp. 4979–4982, Nov 2016.

[36] M. S. Calder and A. Kempf, "Analysis of superoscillatory wave functions," *J. Math. Phys.*, vol. 46, JAN 2005.

[37] M. V. Berry and S. Popescu, "Evolution of quantum superoscillations and optical superresolution without evanescent waves," *J. Phys. A Math. Gen.*, vol. 39, no. 22, pp. 6965–6977, 2006.

[38] A. Kempf, "Black holes, bandwidths and Beethoven," *J. Math. Phys.*, vol. 41, no. 4, pp. 2360–2374, 2000.

[39] P. J. S. G. Ferreira and A. Kempf, "Superoscillations: Faster than the Nyquist rate," *IEEE T. Signal Proces.*, vol. 54, pp. 3732–3740, OCT 2006.

[40] M. R. Dennis, A. C. Hamilton, and J. Courtial, "Superoscillation in speckle patterns," *Opt. Lett.*, vol. 33, pp. 2976–2978, Dec 2008.

[41] M. Soskin and M. Vasnetsov, "Singular optics," in *Progress in Optics* (E. Wolf, ed.), vol. 42, ch. 4, pp. 219–276, Elsevier, 2001.

[42] G. J. Gbur, *Singular Optics.* CRC Press, 2017.

[43] J. F. Nye, J. V. Hajnal, and J. H. Hannay, “Phase saddles and dislocations in two-dimensional waves such as the tides,” *Proceedings of the Royal Society A*, vol. 417, pp. 7–20, May 1988.

[44] M. S. Soskin, V. N. Gorshkov, M. V. Vasnetsov, J. T. Malos, and N. R. Heckenberg, “Topological charge and angular momentum of light beams carrying optical vortices,” *Phys. Rev. A*, vol. 56, pp. 4064–4075, Nov 1997.

[45] M. V. Berry and M. R. Dennis, “Knotted and linked phase singularities in monochromatic waves,” *Proceedings: Mathematical, Physical and Engineering Sciences*, vol. 457, no. 2013, pp. 2251–2263, 2001.

[46] K. O’Holleran, M. R. Dennis, and M. J. Padgett, “Topology of light’s darkness,” *Phys. Rev. Lett.*, vol. 102, p. 143902, Apr 2009.

[47] N. B. Simpson, K. Dholakia, L. Allen, and M. J. Padgett, “Mechanical equivalence of spin and orbital angular momentum of light: an optical spanner,” *Opt. Lett.*, vol. 22, pp. 52–54, Jan 1997.

[48] M. E. J. Friese, T. A. Nieminen, N. R. Heckenberg, and H. Rubinsztein-Dunlop, “Optical alignment and spinning of laser-trapped microscopic particles,” *Nature*, vol. 394, pp. 348–350, July 1998.

[49] M. E. J. Friese, H. Rubinsztein-Dunlop, J. Gold, P. Hagberg, and D. Hanstorp, “Optically driven micromachine elements,” *Applied Physics Letters*, vol. 78, no. 4, pp. 547–549, 2001.

[50] P. Galajda and P. Ormos, “Complex micromachines produced and driven by light,” *Applied Physics Letters*, vol. 78, no. 2, pp. 249–251, 2001.

[51] J. E. Curtis and D. G. Grier, “Structure of optical vortices,” *Phys. Rev. Lett.*, vol. 90, p. 133901, Apr 2003.

[52] V. Garcés-Chávez, D. McGloin, M. J. Padgett, W. Dultz, H. Schmitzer, and K. Dholakia, “Observation of the transfer of the local angular momentum density of a multiringed light beam to an optically trapped particle,” *Phys. Rev. Lett.*, vol. 91, p. 093602, Aug 2003.

[53] K. Ladavac and D. G. Grier, “Microoptomechanical pumps assembled and driven by holographic optical vortex arrays,” *Opt. Express*, vol. 12, pp. 1144–1149, Mar 2004.

[54] J. Arlt, M. MacDonald, L. Paterson, W. Sibbett, K. Dholakia, and K. Volke-Sepulveda, “Moving interference patterns created using the angular Doppler-effect,” *Opt. Express*, vol. 10, pp. 844–852, Aug 2002.

[55] M. P. J. Lavery, F. C. Speirits, S. M. Barnett, and M. J. Padgett, “Detection of a spinning object using light’s orbital angular momentum,” *Science*, vol. 341, pp. 537–540, Aug 2013.

[56] G. Gbur and R. K. Tyson, “Vortex beam propagation through atmospheric turbulence and topological charge conservation,” *J. Opt. Soc. Am. A*, vol. 25, pp. 225–230, Jan 2008.

[57] M. Beijersbergen, R. Coerwinkel, M. Kristensen, and J. Woerdman, “Helical-wavefront laser beams produced with a spiral phaseplate,” *Optics Communications*, vol. 112, pp. 321 – 327, Dec 1994.

[58] N. R. Heckenberg, R. McDuff, C. P. Smith, H. Rubinsztein-Dunlop, and M. J. Wegener, “Laser beams with phase singularities,” *Optical and Quantum Electronics*, vol. 24, pp. S951–S962, Sep 1992.

[59] M. Reicherter, T. Haist, E. U. Wagemann, and H. J. Tiziani, “Optical particle trapping with computer-generated holograms written on a liquid-crystal display,” *Opt. Lett.*, vol. 24, pp. 608–610, May 1999.

[60] D. Ganic, X. Gan, M. Gu, M. Hain, S. Somalingam, S. Stankovic, and T. Tschudi, “Generation of doughnut laser beams by use of a liquid-crystal cell with a conversion efficiency near 100%,” *Opt. Lett.*, vol. 27, pp. 1351–1353, Aug 2002.

[61] E. Brasselet, N. Murazawa, H. Misawa, and S. Juodkazis, “Optical vortices from liquid crystal droplets,” *Phys. Rev. Lett.*, vol. 103, p. 103903, Sep 2009.

[62] Y. Gorodetski, A. Drezet, C. Genet, and T. W. Ebbesen, “Generating far-field orbital angular momenta from near-field optical chirality,” *Phys. Rev. Lett.*, vol. 110, p. 203906, May 2013.

[63] M. Harris, C. A. Hill, P. R. Tapster, and J. M. Vaughan, “Laser modes with helical wave fronts,” *Phys. Rev. A*, vol. 49, pp. 3119–3122, Apr 1994.

[64] J. M. Hickmann, E. J. S. Fonseca, W. C. Soares, and S. Chávez-Cerda, “Unveiling a truncated optical lattice associated with a triangular aperture using light’s orbital angular momentum,” *Phys. Rev. Lett.*, vol. 105, p. 053904, Jul 2010.

[65] L. E. E. de Araujo and M. E. Anderson, “Measuring vortex charge with a triangular aperture,” *Opt. Lett.*, vol. 36, pp. 787–789, Mar 2011.

[66] L. Yongxin, T. Hua, P. Jixiong, and L. Baida, “Detecting the topological charge of vortex beams using an annular triangle aperture,” *Optics & Laser Technology*, vol. 43, pp. 1233–1236, Oct 2011.

[67] H. Tao, Y. Liu, Z. Chen, and J. Pu, “Measuring the topological charge of vortex beams by using an annular ellipse aperture,” *Applied Physics B*, vol. 106, pp. 927–932, Mar 2012.

[68] Y. Liu, S. Sun, J. Pu, and L. Baida, “Propagation of an optical vortex beam through a diamond-shaped aperture,” *Optics & Laser Technology*, vol. 45, pp. 473–479, 2013.

[69] M. Golub, A. Prokhorov, I. Sisakayan, and V. Soĭfer, “Synthesis of spatial filters for investigation of the transverse mode composition of coherent radiation,” *Soviet Journal of Quantum Electronics*, vol. 12, no. 9, pp. 1208–1209, 1982.

[70] M. Golub, S. Karpeev, S. Krivoshlykov, A. Prokhorov, I. Sisakyan, and V. Soĭfer, “Experimental investigation of spatial filters separating transverse modes of optical fields,” *Soviet Journal of Quantum Electronics*, vol. 13, no. 8, pp. 1123–1124, 1983.

[71] M. Golub, S. Karpeev, S. Krivoshlykov, P. A.M, I. Sisakyan, and V. Soĭfer, “Spatial filter investigation of the distribution of power between transverse modes in a fiber waveguide,” *Soviet Journal of Quantum Electronics*, vol. 14, no. 9, pp. 1255–1256, 1984.

[72] G. C. G. Berkhout, M. P. J. Lavery, J. Courtial, M. W. Beijersbergen, and M. J. Padgett, “Efficient sorting of orbital angular momentum states of light,” *Phys. Rev. Lett.*, vol. 105, p. 153601, Oct 2010.

[73] A. M. H. Wong and G. V. Eleftheriades, “Sub-wavelength focusing at the multi-wavelength range using superoscillations: An experimental demonstration,” *IEEE T. Antenn. Propag.*, vol. 59, pp. 4766–4776, Dec 2011.

[74] I. Chremmos and G. Fikioris, “Superoscillations with arbitrary polynomial shape,” *J. Phys. A Math. Theor.*, vol. 48, no. 26, p. 265204, 2015.

[75] J. W. Goodman, *Introduction to Fourier Optics.* McGraw-Hill, 1968.

[76] M. R. Dennis, “Rows of optical vortices from elliptically perturbing a high-order beam,” *Opt. Lett.*, vol. 31, pp. 1325–1327, May 2006.

[77] G. Gbur, “Using superoscillations for superresolved imaging and subwavelength focusing,” *Nanophotonics*, vol. 8, pp. 205–225, Nov 2018.

[78] E. Hecht, *Optics.* Pearson Education, fourth ed., 2002.

[79] F. M. Huang and N. I. Zheludev, “Super-resolution without evanescent waves,” *Nano Letters*, vol. 9, no. 3, pp. 1249–1254, 2009.

[80] E. T. F. Rogers, J. Lindberg, T. Roy, S. Savo, J. E. Chad, M. R. Dennis, and N. I. Zheludev, “A super-oscillatory lens optical microscope for subwavelength imaging,” *NATURE MATERIALS*, vol. 11, pp. 432–435, March 2012.

[81] M. Mazilu, J. Baumgartl, S. Kosmeier, and K. Dholakia, “Optical eigenmodes; exploiting the quadratic nature of the energy flux and of scattering interactions,” *Opt. Express*, vol. 19, pp. 933–945, Jan 2011.

[82] J. Baumgartl, S. Kosmeier, M. Mazilu, E. T. F. Rogers, N. I. Zheludev, and K. Dholakia, “Far field subwavelength focusing using optical eigenmodes,” *Applied Physics Letters*, vol. 98, May 2011.

[83] S. Kosmeier, M. Mazilu, J. Baumgartl, and K. Dholakia, "Enhanced two-point resolution using optical eigenmode optimized pupil functions," *Journal of Optics*, vol. 13, p. 105707, Sep 2011.

[84] G. H. Yuan, E. T. Rogers, and N. I. Zheludev, "Achromatic super-oscillatory lenses with sub-wavelength focusing," *Light: science & applications*, vol. 6, Sep 2017.

[85] National Institute of Standards and Technology, "NIST digital library of mathematical functions: Theta functions: Definitions and periodic properties." https://dlmf.nist.gov/20.2. Last accessed: 03-23-1019.

[86] National Institute of Standards and Technology, "NIST digital library of mathematical functions: Jacobian elliptic functions: Definitions." https://dlmf.nist.gov/22.2. Last accessed: 03-23-1019.