ICES REPORT 10-28

July 2010

Gaussian Beam Decomposition of High Frequency Wave Fields Using Expectation-Maximization

by

G. Ariel, B. Engquist, N. Tanushev, and R. Tsai

The Institute for Computational Engineering and Sciences
The University of Texas at Austin
Austin, Texas 78712

Reference: G. Ariel, B. Engquist, N. Tanushev, and R. Tsai, "Gaussian Beam Decomposition of High Frequency Wave Fields Using Expectation-Maximization ", ICES REPORT 10-28, The Institute for Computational Engineering and Sciences, The University of Texas at Austin, July 2010.

Gaussian Beam Decomposition of High Frequency Wave Fields Using Expectation-Maximization

Gil Ariel[*] Bjorn Engquist[†] Nicolay M. Tanushev[‡] Richard Tsai[§]

June 4, 2010

Abstract

A new numerical method for approximating highly oscillatory wave fields as a superposition of Gaussian beams is presented. The method estimates the number of beams and their parameters automatically. This is achieved by an expectation-maximization algorithm that fits real, positive Gaussians to the energy of the highly oscillatory wave fields and its Fourier transform. Beam parameters are further refined by an optimization procedure that minimizes the difference between the Gaussian beam superposition and the highly oscillatory wave field in the energy norm.

1 Introduction

Numerical simulation of high frequency waves is an active field of computational mathematics with applications in seismic migration [8], computational electro-magnetics [3], semiclassical approximations in quantum mechanics [9] and more. As the term "high frequency" suggests, such applications involve many wave oscillations in the domain of interest and thus, direct numerical simulation methods of the wave propagation are prohibitively computationally costly. The standard approach to surmounting this difficulty is to use an approximate model for the wave propagation that converges to the exact model as the frequency increases. Examples of such asymptotic high frequency methods include geometric optics [3], Gaussian beam methods (GBs) [1, 11], wavefront tracking methods [14] and others. For a comprehensive review, we refer the reader to [6].

[*] Department of Mathematics, Bar Ilan University, Ramat Gan, Israel (arielg@math.biu.ac.il).

[†] Department of Mathematics, The University of Texas at Austin, Austin, TX, 78712, USA (engquist@math.utexas.edu).

[‡] Department of Mathematics, The University of Texas at Austin, Austin, TX, 78712, USA (nicktan@math.utexas.edu).

[§] Department of Mathematics, The University of Texas at Austin, Austin, TX, 78712, USA (ytsai@math.utexas.edu).

In a simple formulation of geometric optics and Gaussian beams, the solution of the hyperbolic partial differential equation (PDE) that models the wave propagation is assumed to be of the form

$$u(x,t) = a(x,t)e^{ik\phi(x,t)} \ , \tag{1.1}$$

where k is the large parameter, $a(x,t)$ is the amplitude and $\phi(x,t)$ is the phase function. For geometric optics ϕ is a real valued function, while for Gaussian beams the phase is complex valued with an imaginary part that concentrates $u(x,t)$ near a certain curve in space time (the central ray). Using the PDE, we find equations for the phase and amplitude and central ray. These equations are independent of k, and consequently, the phase and amplitude are independent of k, thus they can be represented accurately in the domain of interest with far fewer grid points than the original wave field $u(x,t)$. However, even though the original PDE for $u(x,t)$ is linear, the equations determining the phase are usually not. In geometric optics, this non-linearity leads to the breakdown of classical solutions at caustics [6]. The additional assumptions on the phase and amplitude guarantee that Gaussian beams are global asymptotic solutions that are valid even at caustics [11, 12]. Furthermore, as the PDE is linear, superpositions of such Gaussian beams will also be global asymptotic solutions. This idea is the basis of all Gaussian beam methods.

The ultimate goal is to build an approximate solution that is close to the true solution in an appropriate norm for the given problem. Errors of such approximations have two components: how closely the initial data is approximated and how well the PDE is satisfied. In this paper, we will focus on answering the question of how to take general high frequency initial data and approximate it by a linear superposition of a few functions of the form (1.1) that are suitable for providing initial conditions for a Gaussian beam based asymptotic solution method. However, we point out that the question of approximating the initial data and satisfying the PDE are not independent in the sense of the accuracy of the approximate solution. The initial conditions for a Gaussian beam also affect how well the Gaussian beam satisfies the PDE. We will only focus on the initial data, since this is the dominating error at least for short time.

In the geometric optics setting, this question has been addressed in [2]. In that paper, the authors present a method for determining a small number of plane waves, $a_j e^{ik\xi_j\cdot(x-y)}$, that locally approximate the high frequency initial data near a fixed point y. At all points, using the Fourier transform one can always rewrite the initial data as a linear superposition of plane waves. Similarly, in the case of Gaussian beams, one can use the Fourier-Bros-Iaglonitzer (FBI) transform to rewrite the initial data as a linear superposition of Gaussian beams [4]. However, such transform methods generate beams which are Gaussian modulated plane waves. Thus, while these methods may provide an efficient approximation for initial wave fields that consist of superpositions of only a few plane waves, for more general wave fields, which for example contain waves with curved wavefronts, such transform methods will not provide an optimal representation.

As described in section 2 and A, the phase and amplitude of a Gaussian beam are given as Taylor polynomials around the central ray. Thus, the initial data for a Gaussian beam are the parameters (base point and coefficients) that define the Taylor polynomials for the phase and amplitude. We will refer to these initial values as the initial beam parameters or simply the initial beams. The wave fields generated by Gaussian beam parameters can be viewed as a redundant basis for representing general wave fields. In this sense, we can think of the problem of approximating (or decomposing) an initial high frequency data as finding the parameters of a small number of beams such that the wave field generated by the superposition of these beams is a good approximation to the initial data for the PDE. That is, given the initial data u and an error tolerance ε, the objective is to approximate the initial data in the energy norm $||\cdot||_E$ (see equation (2.2)) using the superposition of as few Gaussian beams as possible. One can formally consider a basis-pursuit style formulation

$$\min_a |a|_{\ell_0} \quad \text{s.t.} \quad ||Sa - u||_E^2 \leq \varepsilon^2 \ ,$$

where S is the discretized basis matrix, a are the associated weights and $|a|_{\ell_0}$ is the number of non-zero elements in a. One can then envision using the latest fast algorithms for constrained L^1 minimization for finding a sparse approximation, for example [15]. However, the set of all Gaussian beams forms a high dimensional space which make these algorithms prohibitively inefficient.

In [13], the authors propose a practical method for decomposing a general wave field into a a superposition of Gaussian beams. Their method can be described as a greedy bottom-up approach. At the $(N+1)$ iteration of "the greedy outer loop", a new set of parameters is found for a single beam that approximates the difference between the initial data for the PDE and the wave field generated by previous (N) Gaussian beams. This new set of beam parameters is directly estimated from the residual wave field. Then, the parameters are locally optimized using the Nelder-Mead method [10]. The procedure is repeated until a desired tolerance or number of beams is reached.

The basic assumption underlying this strategy is that the sets of parameters that give the optimal $(N+1)$-beam minimum will be closely related (in parameters space) to the parameters that give the optimal (N)-beam minimum. Thus, the method of sequentially adding beams is highly advantageous if the different beams are close to orthogonal in the energy inner product as the minimum can be reached by optimizing each set of beam parameters independently of the others. This assumption holds in the case of waves that are spatially separated and also the case of crossing wave, in which waves arrive to the same point, but have different directions (see the crossing waves example in section 4). However, waves do not always exhibit this type of behavior. Near a caustic region, waves are traveling in similar directions and are close together in space. In this situation, the individual beams will not be orthogonal and the sets of parameters that give the $(N+1)$-beam minimum may be quite different from the sets of parameters that give the (N)-beam minimum. Therefore, while adding Gaussians decreases the error in the approximation of the initial data for the PDE, the required number of beams to reach a given tolerance may be suboptimal.

In this paper, following the greedy approach of [13], we propose a different decomposition algorithm which can be more efficient in handling wave fields containing many different beams with similar centers and wave directions. Inside each iteration of the greedy outer loop, an additional set of *several* Gaussian beams is constructed simultaneously to fit the difference between the initial wave field and the Gaussian beam approximation from previous iteration. In contrast, we note that in [13], a single beam is added in each iteration. To obtain this set of several beams, we exploit the fact that the energy of a single beam is non-negative and nearly Gaussian shaped. The same holds for the energy of the Fourier transform of a single beam as shown in section 2.1. Accordingly, our method is based on fitting Gaussians to a smoothed version of the energy of the initial data for the PDE and its Fourier transform. This is done using the expectation maximization (EM) algorithm [5], which is reviewed in B. The advantage of the EM algorithm is its efficiency in simultaneously optimizing a large number of parameters that define a superposition of Gaussians.

After the new set of Gaussian beams is identified, all the beam parameters, including the newly constructed and the ones from the previous iteration, are optimized to minimize the error in the energy norm. This approach may bypass some of the local minima, for example near a caustic point, that may be encountered in the approach of [13], in which beams are added sequentially. Of course, our approach is also suboptimal as there is no general methodology for finding the global minimum of a highly multi-dimensional function that is not computationally prohibitive.

Let $u_0(x)$ and $\partial_t u_0(x)$ denote the initial wave field and its derivative with respect to time, respectively. In addition, let u_{GB}^n denote the Gaussian beam approximation after n iterations of the greedy outer loop, where $u_{\mathrm{GB}}^0 \equiv 0$. The structure of a each iteration of the outer loop can be summarized as follows.

- EM-based approximation: Construct a Gaussian beam approximation for the residual wave field, $u_0 - u_{GB}^n$ and $\partial_t u_0 - \partial_t u_{\mathrm{GB}}^n$ using the EM method, as described below. We denote the approximation by v_{GB}.

- Local optimization (section 3.5): Update the sets of beam parameters for u_{GB}^n and v_{GB} constructed so far to minimize the difference between the wave field generated by these beams and the initial data for the PDE. Eliminate beams whose contribution to the overall error is smaller than a prescribed threshold. Let u_{GB}^{n+1} be the beams defined by the optimized parameters.

The EM-based approximation is summarized in the following:

- Pre-processing (section 3.1): Calculate the energy function of the initial data for the PDE and the energy function of its scaled Fourier transform. Mollify these energies by a Gaussian kernel.

- EM (section 3.2): Fit a linear superposition of Gaussians to the mollified energies using the EM method.
- Reconstruction (section 3.3): Reconstruct sets of beam parameters by pairing the Gaussian coefficients obtained by EM from the physical and Fourier domains. All such pairs are tested by projections on the initial data for the PDE in the energy norm. Candidate pairs with small projections are discarded.
- Corrections (section 3.4): Improve the accuracy of the fit by extrapolation.

The outline of the paper is as follows. Section 2 gives a precise statement of the problem considered and briefly reviews geometric optics and Gaussian beam solutions to the wave equation. Section 3 explains our numerical method with examples in section 4. We summarize our results in section 5. Several technical aspects of the calculations involved are detailed in the appendices.

2 Gaussian beam solutions

Consider the isotropic wave equation with variable coefficients in $\mathbb{R}^d$

$$\begin{aligned} &\Box u = u_{tt}(x,t) - c^2(x)\Delta u(x,t) = 0, \quad t > 0 \\ &u(x,0) = f(x) \\ &u_t(x,0) = g(x), \end{aligned} \tag{2.1}$$

where subscripts denote partial differentiation and Δ is the Laplacian. We seek solutions in which the ratio between the wave length and the scale on which c varies (assumed to be of order one) is large. This ratio, denoted k, satisfies $k \gg 1$. The wave equation is well posed in the energy (semi-)norm

$$||u||_E^2 = k^2 \int_{\mathbb{R}^d} e(x,t)dx, \tag{2.2}$$

where $e(x,t)$ is the energy function weighted by k,

$$e(x,t) = k^{-2}\left[\frac{1}{c^2(x)}|u_t|^2 + |\nabla u(x,t)|^2\right]. \tag{2.3}$$

We will also use the scalar product underlying the energy norm

$$\langle u, v\rangle = k^{-2}\int_{\mathbb{R}^d}\left[\frac{1}{c^2(x)}u_t v_t^* + \nabla u \cdot \nabla v^*\right]dx,$$

where $[\cdot]^*$ denotes complex conjugation. In order to obtain the high-frequency geometric optics approximation, we make the standard ansatz

$$u(x,t) = a(x,t)e^{ik\phi(x,t)}. \tag{2.4}$$

In geometric optics, one finds solutions for $a(x,t)$ and $\phi(x,t)$ in the form of rays, which are the characteristics of an eikonal equation for ϕ. The GB methods goes further and approximates solutions to the wave equations in the form of expansions around a specific ray. For completeness, the derivation of the ray and GB solutions are reviewed in A.

For example, with a constant speed of propagation $c(x) = c$, rays are straight lines. Denote the source point of the ray by ξ, the initial direction by η and the initial Hessian by $M(0) = i\beta$. In one-dimension (1D), a GB has the form

$$u(x,t) = Ae^{ik\eta(x \pm ct - \xi)}e^{-k\beta(x \pm ct-\xi)^2/2}, \tag{2.5}$$

where β is a complex number with a positive real part, $\mathrm{Re}\beta > 0$. In two-dimensions (2D), a single GB has the form

$$u(x,t) = a(t)e^{ik\eta\cdot(x\pm\hat{\eta}ct-\xi)}e^{ik(x\pm\hat{\eta}ct-\xi)^T M(t)(x\pm\hat{\eta}ct-\xi)/2}, \tag{2.6}$$

where $\hat{\eta} = \eta/|\eta|$, $[\cdot]^T$ denotes transposition and β is a complex 2×2 matrix with a positive definite real part, $\mathrm{Re}\beta > 0$. The amplitude and the Hessian matrix are given by

$$\begin{aligned} a(t) &= A\sqrt{\frac{|\eta|^3}{|\eta|^3 \mp i(\det\beta)(\eta^T\beta\eta)ct}} \\ M(t) &= \frac{\pm i|\eta|^3\beta + (\det\beta)(\eta\eta^T)ct}{\pm|\eta|^3 + i(\det\beta)(\eta^T\beta\eta)ct}, \end{aligned} \tag{2.7}$$

where $A = a(0)$.

2.1 Single beam

One of the important observations is that to leading order in k the energy function of a single Gaussian, $e(x,t)$, is a real valued Gaussian. In particular, the initial energy function of a single beams is

$$e(x,0) = 2|A|^2|\eta|^2 e^{-k(x-\xi)^T(\mathrm{Re}\beta)(\mathrm{x}-\xi)} + O(1/\sqrt{k}), \tag{2.8}$$

which is a real and positive Gaussian centered at ξ with covariance $\Sigma_x = (\mathrm{Re}\beta)^{-1}/2$. Note that, by assumption, $|\eta| > 0$ and is of order one (in k). A similar version of (2.8) as well as all expressions in this section hold in the general case of a variable propagation speed $c(x)$ and in any dimension d. Furthermore, due to symmetry with respect to the Gaussian center ξ, the contribution of $O(1/\sqrt{k})$ terms to the total energy $||u||_E$ is of order $O(1/k)$. We refer the reader to A for details.

A similar energy function can be found in Fourier space. To this end we define a weighted Fourier transform

$$\tilde{f}(p,t) = \mathcal{F}f(x,t) = \int f(x,t)e^{ikp\cdot x}dx. \tag{2.9}$$

At $t = 0$, a single transformed GB takes the form

$$\tilde{u}(p,0) = k^{-1/2} A e^{ikp\cdot\xi} e^{-k(p+\eta)^T\beta^{-1}(p+\eta)/2}, \tag{2.10}$$

and the Fourier energy function,

$$\tilde{e}(p,t) = k^{-1}\left[|\tilde{u}_t|^2 + |\nabla_p \tilde{u}(p,t)|^2\right], \tag{2.11}$$

becomes, at $t = 0$,

$$\tilde{e}(p,0) = |A|^2(|\xi|^2 + 2c^2(\xi)|\eta|^2) e^{-k(p+\eta)^T(\mathrm{Re}[\beta^{-1}])(\mathrm{p}+\eta)} + O(1/\sqrt{k}). \tag{2.12}$$

As in position space, the contribution of $O(1/\sqrt{k})$ terms to the total energy $||\tilde{u}||_E$ is of order $O(1/k)$. To leading order in k, $\tilde{e}(p,0)$ is a real, positive Gaussian centered at $p = -\eta$ with covariance $\Sigma_p = (\mathrm{Re}[\beta^{-1}])^{-1}/2$.

The above observations suggest that the two energy functions can be used to reconstruct all the parameters that make up a beam. In position space, $e(x,0)$ given by (2.8) can be used to obtain ξ and $\mathrm{Re}[\beta]$. In Fourier space $\tilde{e}(x,0)$ given by (2.12) can be used to obtain η and $\mathrm{Re}[\beta^{-1}]$. In D we show that this is sufficient to derive $\mathrm{Im}[\beta]$ as well. The amplitude can be obtained by projecting a normalized beam with parameters ξ, η and β on the initial field.

2.2 Superposition of beams

The analysis is more complicated for wave fields consisting of a superposition of several GBs. In 2D,

$$u(x,t) = \sum_{n=1}^{N} a_n(t) e^{ik\eta_n\cdot(x+s_n\hat{\eta}_n t-\xi_n)} e^{ik(x+s_n\hat{\eta}_n t-\xi_n)^T M_n(t)(x+s_n\hat{\eta}_n t-\xi_n)/2}, \tag{2.13}$$

where $s_n = +1$ or -1 is the sign appearing in (2.5) or (2.6). The time dependent amplitude, $a_n(t)$, and Hessian, $M_n(t)$, are given by (2.7) with parameters $\eta = \eta_n$, respectively. At time $t = 0$, the initial field can be written as

$$u(x,0) = \sum_{n=1}^{N} A_n e^{ik\eta_n\cdot(x-\xi_n)} e^{-k(x-\xi_n)^T\beta_n(x-\xi_n)/2}, \tag{2.14}$$

where $A_n = a_n(0)$ and $\beta_n = -iM_n(0)$. The initial energy function is

$$\begin{aligned} e(x,0) = \sum_{n,j=1}^{N} & A_n A_j^* \left(\eta_n\cdot\eta_j + s_n s_j|\eta_n||\eta_j|\right) e^{ik[\eta_n\cdot(x-\xi_n)-\eta_j\cdot(x-\xi_j)]} \\ & \times e^{-k[(x-\xi_n)^T\beta_n(x-\xi_n)+(x-\xi_j)^T\beta_j^*(x-\xi_j)]/2} + O(1/\sqrt{k}). \end{aligned}$$

Since $\{\xi_n\}$ are independent of k, the magnitude of the leading order quadratic term is small unless $\xi_j = \xi_n$. Hence,

$$\begin{aligned} e(x,0) = & \sum_{n,j=1,\xi_n=\xi_j}^{N} A_n A_j^* \left(\eta_n \cdot \eta_j + s_n s_j |\eta_n||\eta_j|\right) \\ & \times e^{ik[\eta_n\cdot(x-\xi_n)-\eta_j\cdot(x-\xi_n)]} e^{-k(x-\xi_n)^T(\beta_n+\beta_j^*)(x-\xi_n)/2} + O(1/\sqrt{k}). \end{aligned}$$

Terms in the double sum oscillate with a frequency of order k unless $\eta_n = \eta_j$ and $\mathrm{Im}\beta_{\mathrm{n}} = \mathrm{Im}\beta_{\mathrm{j}}$. Hence, $e(x,0)$ has the form

$$\begin{aligned} e(x,0) =& 2\sum_{n=1}^{N} |A_n|^2 |\eta_n|^2 e^{-k(x-\xi_n)^T \mathrm{Re}[\beta_{\mathrm{n}}](\mathrm{x}-\xi_{\mathrm{n}})} \\ & + 4 \sum_{n<j,\ \mathrm{non-osc}} \mathrm{Re}[\mathrm{A_n A_j^*}]|\eta_{\mathrm{n}}|^2 \mathrm{e}^{-\mathrm{k(x-\xi_n)^T Re[\beta_n+\beta_j](x-\xi_n)}} \\ & + O(1/\sqrt{k}) + \text{highly oscillatory terms.} \end{aligned}$$

where $\sum_{n<j,\ \mathrm{non-osc}}$ denotes summation over all $n < j$ such that $\eta_n = \eta_j$, $\xi_n = \xi_j$, $s_n = s_j$ and $\mathrm{Im}\beta_{\mathrm{n}} = \mathrm{Im}\beta_{\mathrm{j}}$. The first sum, that corresponds to terms in which $n = j$, describes the energies of each separate beam. The second sum includes some spurious Gaussians that may appear in the energy function due to interference between beams. In order to suppress the oscillatory terms we convolve $e(x,0)$ with a smoothing kernel of the form

$$\chi(x) = e^{-kx^2/(2l)}, \tag{2.15}$$

where $l > 0$ is independent of k. The convolution attenuates oscillations with frequency p by a factor of $e^{-Cp^2/k}$, where $C > 0$ is a constant that depends only on l and the eigenvalues of $\mathrm{Re}\beta$. Hence, high frequencies of the order of k are suppressed by a factor of e^{-Ck}. The convolved energy has the form

$$\begin{aligned} e_l(x) =& \chi(x) * e(x,0) \\ =& 2\sum_{n=1}^{N} |A_n|^2 |\eta_n|^2 e^{-k(x-\xi_n)^T \Sigma(\mathrm{Re}[\beta_{\mathrm{n}}])(\mathrm{x}-\xi_{\mathrm{n}})} \\ & + 4 \sum_{n<j,\ \mathrm{non-osc}} \mathrm{Re}[\mathrm{A_n A_j^*}]|\eta_{\mathrm{n}}|^2 \mathrm{e}^{-\mathrm{k(x-\xi_n)^T \Sigma(Re[\beta_n+\beta_j])(x-\xi_n)/2}} \\ & + O(1/\sqrt{k}), \end{aligned} \tag{2.16}$$

where, $\Sigma^{-1}(\cdot)$ is the new variance which is changed due to the convolution. In 2D it is given by

$$\Sigma(B) = \frac{B + 2l(\det B)\mathcal{I}}{1 + 2l\mathrm{Tr}B + 4l^2 \det B}, \tag{2.17}$$

where Tr denotes the trace and $\mathcal{I}$ is the identity matrix. The important conclusion of this derivation is that (2.16) is a linear mixture of real, positive Gaussians. Following the same procedure in Fourier space, the convolved Fourier initial energy is

$$
\begin{aligned}
\tilde{e}_l(p) &= \chi(p) * \tilde{e}(p,0) \\
&= \sum_{n=1}^{N} |A_n|^2(|\xi_n|^2 + 2c^2(\xi_n)|\eta_n|^2)e^{-k(p+\eta_n)^T\Sigma(\mathrm{Re}[\beta_{\mathrm{n}}^{-1}])(\mathrm{p}+\eta_{\mathrm{n}})} \\
&+ 2 \sum_{\substack{n<j \\ \mathrm{non-osc-p}}} \mathrm{Re}[\mathrm{A_n A_j^*}](|\xi_{\mathrm{n}}|^2 + 2\mathrm{s_n s_j}\mathrm{c}^2(\xi_{\mathrm{n}})|\eta_{\mathrm{n}}|^2)\mathrm{e}^{-\mathrm{k}(\mathrm{p}+\eta_{\mathrm{n}})^{\mathrm{T}}\Sigma(\mathrm{Re}[\beta_{\mathrm{n}}^{-1}+(\beta_{\mathrm{j}}^*)^{-1}])(\mathrm{p}+\eta_{\mathrm{n}})/2} \\
&+ O\left(1/\sqrt{k}\right),
\end{aligned}
\tag{2.18}
$$

where $\sum_{n<j,\ \mathrm{non-osc-p}}$ denotes a sum over the non-oscillatory terms in $\tilde{e}(p,0)$, obtained by substituting the weighted Fourier transform of (2.13) into (2.11). Terms in $\tilde{e}(p,0)$ oscillate with a frequency of order k unless $\eta_n = \eta_j$, $\xi_n = \xi_j$ and $\mathrm{Im}\beta_{\mathrm{n}}^{-1} = \mathrm{Im}\beta_{\mathrm{j}}^{-1}$. We readily see that, to leading order in k, the convolved Fourier energy function $\tilde{e}_l(p)$ is a mixture of real, positive Gaussians. However, there are some degenerate situations in which the $O(1)$ terms vanish due to exact cancelations between the first and second sums. For example, take $N = 2$, $A_1 = A_2 \neq 0$, $\xi_1 = \xi_2 = 0$, $\eta_1 = \eta_2 \neq 0$, $s_1 = -s_2$ and $\beta_1 = \beta_2$. We do not pursue this situation further since it can be eliminated using the smoothing kernel.

3 Numerical method

In this section, we detail the different stages making out a single iteration of the greedy outer loop in our numerical algorithm: pre-processing, EM, reconstruction, corrections and parameter optimization.

3.1 Pre-processing

The purpose of the pre-processing stage is to change the initial condition into a form that can be approximated by a linear combination of real and positive Gaussians. The steps, described in section 2.2, consists of convolving the initial energy function, $e(x,0)$ given by (2.3), with the smoothing kernel (2.15). Then, the initial field $u(x,0)$ is Fourier transformed using FFT. The initial Fourier energy $\tilde{e}(p,0)$ is calculated and convolved with a similar kernel. The process yields e_l and $\tilde{e}_l$ given by (2.16) and (2.18). Smoothing the initial data removes high frequency oscillations which allows using a coarser grid than required for a numerically accurate description of a high frequency wave.

3.2 Expectation-Maximization

We now explain the method used to approximate the real and positive smoothed energy functions $e_l(x)$ and $\tilde{e}_l(p)$ using a linear combination of Gaussians. For brevity, we refer only to the energy in position space, $e_l(x)$. The same process is applied to approximate $\tilde{e}_l(p)$. In section 2.2, we show that if the solution is indeed a superposition of GBs, then e_l and $\tilde{e}_l$ are, to leading order in k, a linear combination of real and positive Gaussians. Let,

$$e_l(x) = \sum_{j=1}^{N} D_j G_j(x) \ ; \quad G_j(x) = z_i^{-1} e^{-(x-\mu_j)\cdot\sigma_j^{-1}(x-\mu_j)/2}. \tag{3.1}$$

The values of e_l are given on a grid with M points $X = \{x_i\}_{i=1}^{M}$. Here, z_i are normalization constants such that $\sum_{i=1}^{M} G_j(x_i) = 1$ for all $j = 1 \ldots N$. The energy function, e_l is normalized so that $\Sigma_i e_l(x_i) = 1$. This implies that $\sum_j D_j = 1$. The parameters for fitting the normalized $e_l(x)$ using N Gaussians are found using the EM algorithm. The procedure consists of picking an initial random guess of parameters $\{A_j, \mu_j, \sigma_j\}_{j=1}^{N}$ and iterating the following calculations:

$$\begin{aligned} D'_j &= \sum_i e_l(x_i) p_{ij} \\ \mu'_j &= \sum_i e_l(x_i) \frac{p_{ij}}{D'_j} x_i \\ \sigma'_j &= \sum_i e_l(x_i) \frac{p_{ij}}{D'_j} x_i x_i^T - \mu'_j (\mu'_j)^T, \end{aligned} \tag{3.2}$$

where

$$p_{ij} = \frac{A_j G_j(x_i)}{\sum_{j=1}^{N} D_j G_j(x_i)}.$$

The algorithm is explained and motivated in B.

The EM algorithm [5] is an iterative process that converges to a local extremum of the likelihood for obtaining the observation e_l from a random sample of Gaussians with coefficients $\{A_j, \mu_j, \sigma_j\}_{j=1}^{N}$. Local minima are stable while local maxima are unstable. Therefore, in general, the algorithm may not converge to the global minimum or the set of parameters which will give the smallest final fit error (in the energy norm). However, it can be shown that with a single Gaussian EM converges in a single iteration. In addition, if the initial field is composed of a sum of Gaussian beams which are well separated in both position and Fourier space, then the likelihood which EM minimizes is unique and iteration of (3.2) will converge for all reasonable initial guesses. By a reasonable guess we exclude some degenerate cases and initial guesses that are so far from the global minimum that the coefficients p_{ij} in (3.2) are smaller than the round-off error. The computational complexity of each iteration is $O(N)$.

The number of Gaussians required for the fit is not known in advance. In order for the GB approximation to be consistent, all beam parameters should be of order one (in k). This condition can be manifested in a requirements for some maximal and minimal beam width that can be used to evaluate the number of beams. Furthermore, an EM fit with particularly thin or wide Gaussian can respectively suggest the need to reduce or increase the number of beams. In practice, the number of Gaussians, N, was increased gradually until the error (in the L_1 norm) was below a given threshold (usually $5-10\%$). Note that the algorithm converges quickly even for large N and several different random initial guesses for EM can be checked.

3.3 Reconstruction

The EM fit provides a list of N_1 and N_2 Gaussians fitted to the smoothed position and Fourier energy functions, respectively. Pairing up each Gaussian in position space with a Gaussian in Fourier space yields a list of $N_1 N_2$ pairs with four parameters: position center, μ, position covariance, σ, Fourier center $\tilde{\mu}$ and Fourier variance $\tilde{\sigma}$. The amplitudes are discarded. These four multi-dimensional parameters will be used to construct a candidate Gaussian beam. Comparing with (2.16) and (2.18) we find that

$$\begin{aligned}
&\xi = \mu \\
&\eta = -\tilde{\mu} \\
&\Sigma(\mathrm{Re}[\beta]) = \sigma^{-1} \\
&\Sigma(\mathrm{Re}[\beta^{-1}]) = \tilde{\sigma}^{-1},
\end{aligned}$$

where $\Sigma(\cdot)$ is the widening of variances due to the smoothing convolution, given, for example in two dimensions, by (2.17). C describes a simple iterative method for inverting this matrix-valued function. Hence, one can derive candidate values for ξ, η, $\mathrm{Re}[\beta]$ and $\mathrm{Re}[\beta^{-1}]$. In D, we describe a method for using the real part of the inverse, $\mathrm{Re}[\beta^{-1}]$, in order to reconstruct the imaginary part of β. The solution for $\mathrm{Im}[\beta]$ is not unique. In general, for fixed $A = \mathrm{Re}[\beta]$, there are 2^d possible real and symmetric matrices B that give the same $\mathrm{Re}[(\mathrm{A}+\mathrm{iB})^{-1}]$ in d dimensions. The method described in D is numerical, however, we also give an analytic formula for one and two dimensions. The formula provides all solutions. For robustness, we also use the case in which β is purely real with its real part derived from the fit in position space alone. Finally, each parameter combination could occur with either $s = +1$ or $s = -1$ as this information is not manifested directly in the EM fit.

To summarize, the EM stage provides us with $N = 2(1 + 2^d)N_1 N_2$ candidate Gaussians. Each candidate is projected on the initial field to find its amplitude. Candidates with a poor projection are discarded.

3.4 Corrections

The EM fit provides a good approximation to the parameters making up the initial beams. However, the process has several sources of errors such as neglected terms of order k^{-1} (c.f. A.2) and the inversion of $\Sigma(\cdot)$ (c.f. C). These errors and others can be compensated for using the following extrapolation procedure.

Consider a superposition of beams given by a set of parameters θ_0 that generate a field u_0. Applying our decomposition method (pre-processing, EM and reconstruction) to this field, yields an approximated set of beams with parameters θ_1 which is close, but not identical to θ_0. Let u_1 denote the field generated by GBs with parameters θ_1. Applying our fitting procedure to u_1 yields a new parameter set θ_2, which is again similar, but not exactly the same as θ_1. The difference between θ_2 and θ_1 can be used to evaluate the unknown error of θ_1 compared to θ_0.

We formally denote the error in the fitted parameters as a function of the initial beam parameters as $\epsilon E(\theta)$, where ϵ is a small parameter, for example of order $1/k$, such that the error function itself is of order unity. The main assumption here is that $E(\theta)$ is continuously differentiable in θ for some range of parameters. With θ_0 unknown, one can devise an extrapolation algorithm as follows. Let $\theta_1 = \theta_0 + \epsilon E(\theta_0)$ and $\theta_2 = \theta_1 + \epsilon E(\theta_1)$. Expanding $E(\theta_1)$ around θ_0,

$$\theta_2 - \theta_1 = \epsilon E(\theta_1) = \epsilon E(\theta_0) + \epsilon O(\theta_1 - \theta_0) = \epsilon E(\theta_0) + \epsilon^2 O(E(\theta_0)).$$

We conclude that $2\theta_1 - \theta_2$ is an improved approximation (of order ϵ^2) for the unknown θ_0.

If the difference $|\theta_1 - \theta_2|$ is small enough, then the new β will have a positive definite real part. In practice, we verify that $2\theta_1 - \theta_2$ are admissible GB parameters and that the error in using $2\theta_1 - \theta_2$ is indeed smaller than the error in using θ_1. We found that one or two iterations of this procedure can considerably improve results. The correction is done in two steps: first for the reconstruction step stage alone and then for the three stages together (pre-processing, EM and reconstruction). The computational cost is low as one can use the previous EM fit as initial conditions for the new one. See section 4.1 for an example.

3.5 Local optimization

The final stage of the process is to gradually update beam parameters to decrease the overall error in the energy norm. The beams obtained from the previous stages using smoothing, EM, reconstruction and corrections serve as initial conditions. In 2D, each beam involves 13 real parameters. Since amplitudes and the sign s_n can be found using least squares [13], this implies minimization in a $10N$-dimensional parameter space. This high dimensional minimization should be carried over all of the parameters simultaneously. However, to accomplish it in practice, we iterate over all parameters, holding all but one fixed and optimizing over it using steps of fixed size until a local minima (as a function of the single parameter being changed) is reached. Then, beams whose contribution to the error is lower than some threshold are removed, which is determined by looking at the error in approximating the initial field with each one of the beams removed. This is an important step that can eliminate beams whose parameters are similar. The optimization steps are repeated with decreasingly smaller step sizes to a prescribed tolerance.

One of the fundamental assumptions underlying GBs is that all beam parameters are appropriately scaled in terms of k. Violating this requirement leads to a poor approximation of the beams evolution in time. To this end, we add a penalty to the fit error in order to enforce the scaling.

4 Examples

In this section we describe several numerical experiments. In the first three examples, the initial field is generated from a superposition of beams. Hence, the purpose of the example is to demonstrate that the algorithm can successfully reconstruct the generating beams. The last two examples describe a field which cannot be written as a finite sum of GBs. These raise two questions: what is the optimal way to approximate the field with beams to a given tolerance and how well can our algorithm approximate the optimal combination. All example are constructed with $k = 50$, which is a relatively modest scale separation. This is a more challenging scenario as multi-scale algorithms tend to improve with larger scale separation, i.e. larger k.

Note that the number of beams used for the final fit is not a parameter in our algorithm. Instead, the method automatically adjusts the number of beams according to fitness criteria such as the required precision of the EM fit and other optimization parameters. As explained in the introduction, our goal is to find a representation of the initial wave field using a small number of GBs.

In the following, we define the EM-error as the L_1 norm of the difference between the smoothed energy function and the linear combination of Gaussians found by EM. By error, or fit-error, we refer to the energy norm of the difference between a superposition of GBs and the initial wave field, $||u_0 - u^n_{\text{GB}}||_E$. Relative errors are relative to the norm of the initial fields (L_1 for EM and energy otherwise).

4.1 A single beam

We approximate the initial field generated by a single GB with the following coefficients

$$\begin{aligned}
&A = 1 + i \\
&\xi = (0, 0.5) \\
&\eta = (0.5, 0.5) \\
&\beta = \begin{pmatrix} 1 & 0.2 + i \\ 0.2 + i & 1 \end{pmatrix} \\
&s = +1.
\end{aligned} \tag{4.1}$$

The initial field and energy function are depicted in figure 1. With a single Gaussian the energy function is, to leading order in k, Gaussian. The EM algorithm converges in a single iteration with a relative EM-error of about 1.3% in both position and Fourier spaces. The reconstruction stage yields a single GB whose coefficients ξ, η and β are about 5% off:

$$\begin{aligned}
&A = 1.1 + 0.8i \\
&\xi = (0.00, 0.498) \\
&\eta = (0.51, 0.52) \\
&\beta = \begin{pmatrix} 0.97 - 0.05i & 0.23 + 1.00i \\ 0.23 + 1.00i & 0.48 + 0.13i \end{pmatrix} \\
&s = +1.
\end{aligned}$$

Despite the close match in coefficients, the relative fit error is about 21%. A single correction iteration (c.f. section 3.4) yields a new set of beam parameters which are about 0.1% away from (4.1). The relative fit error is 0.7%. A second correction iteration yields a beam with a negligible 0.002% error. Note that this Gaussian beam is obtained without any non-linear optimization in parameters space to reduce the error in the energy norm (section 3.5).

4.2 A focus point

We approximate the following field generated by eight beams (eight combinations of ±) focused at the origin, as in [13]:

$$\begin{aligned}
&A = 1 \\
&\xi = (0, 0) \\
&\eta = (\pm 0.7, 0) \quad \text{and} \quad (0, \pm 0.7) \\
&\beta = \mathcal{I} \\
&s = \pm 1.
\end{aligned}$$

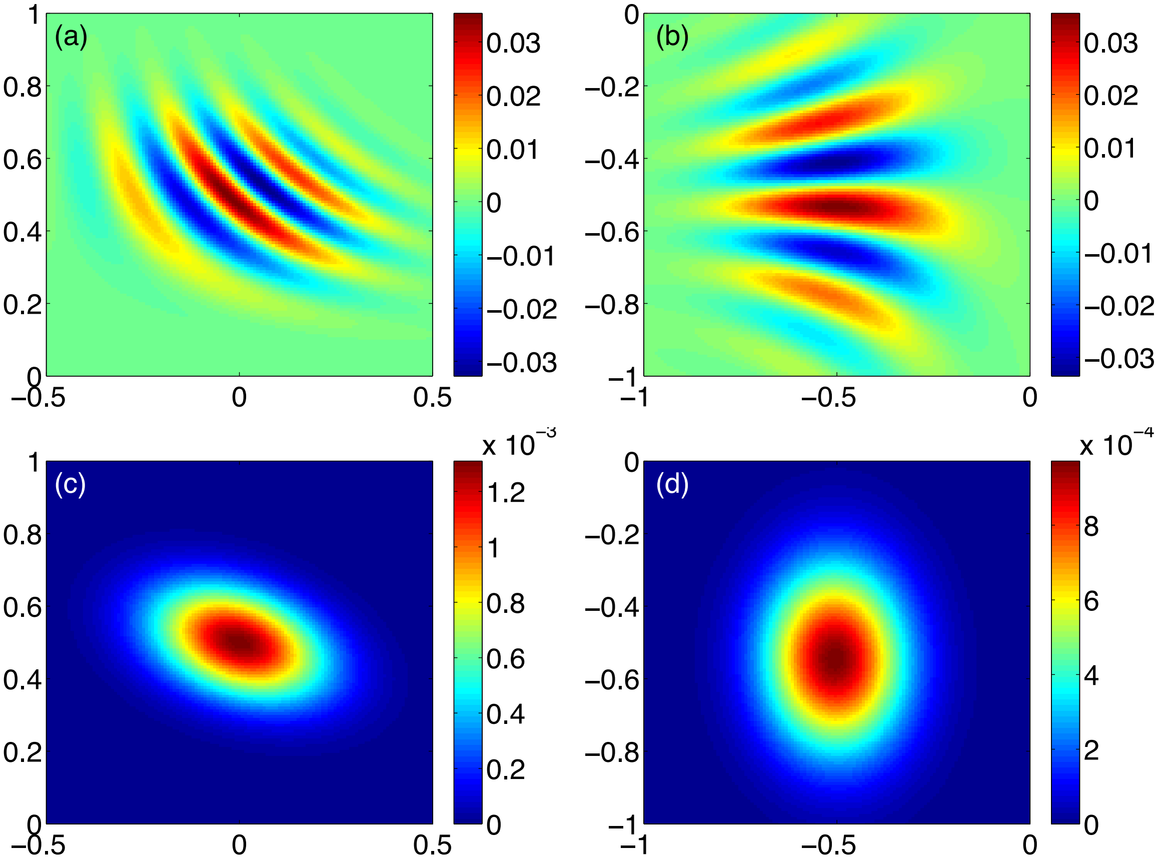

Figure 1: A single Gaussian beams. (a) the real part of the field, (b) the real part of the weighted Fourier transform, (c) the energy function in position space, and (d) the energy function in Fourier space.

The field, energy function and smoothed energy function are depicted in figure 2. In position space (left), the energy is oscillatory due to interference between the beams. Following a convolution with a smoothing kernel, the energy appears Gaussian and suggests using a single Gaussian in position space and four in Fourier space. The reconstruction step yields $2*(1+2^2)*1*4 = 40$ candidate beams, out of which only 8 have a significant projection on the initial field. The relative fit error of the 8 beams is 27%. Corrections (section 3.4) and parameter optimization reduces the error to 1.8%. A 200×200 grid was used.

4.3 A tight superposition of beams

We generate ten GBs with random coefficients. To make the decomposition challenging, the centers of all the beams are crowded in a small area in both position and Fourier

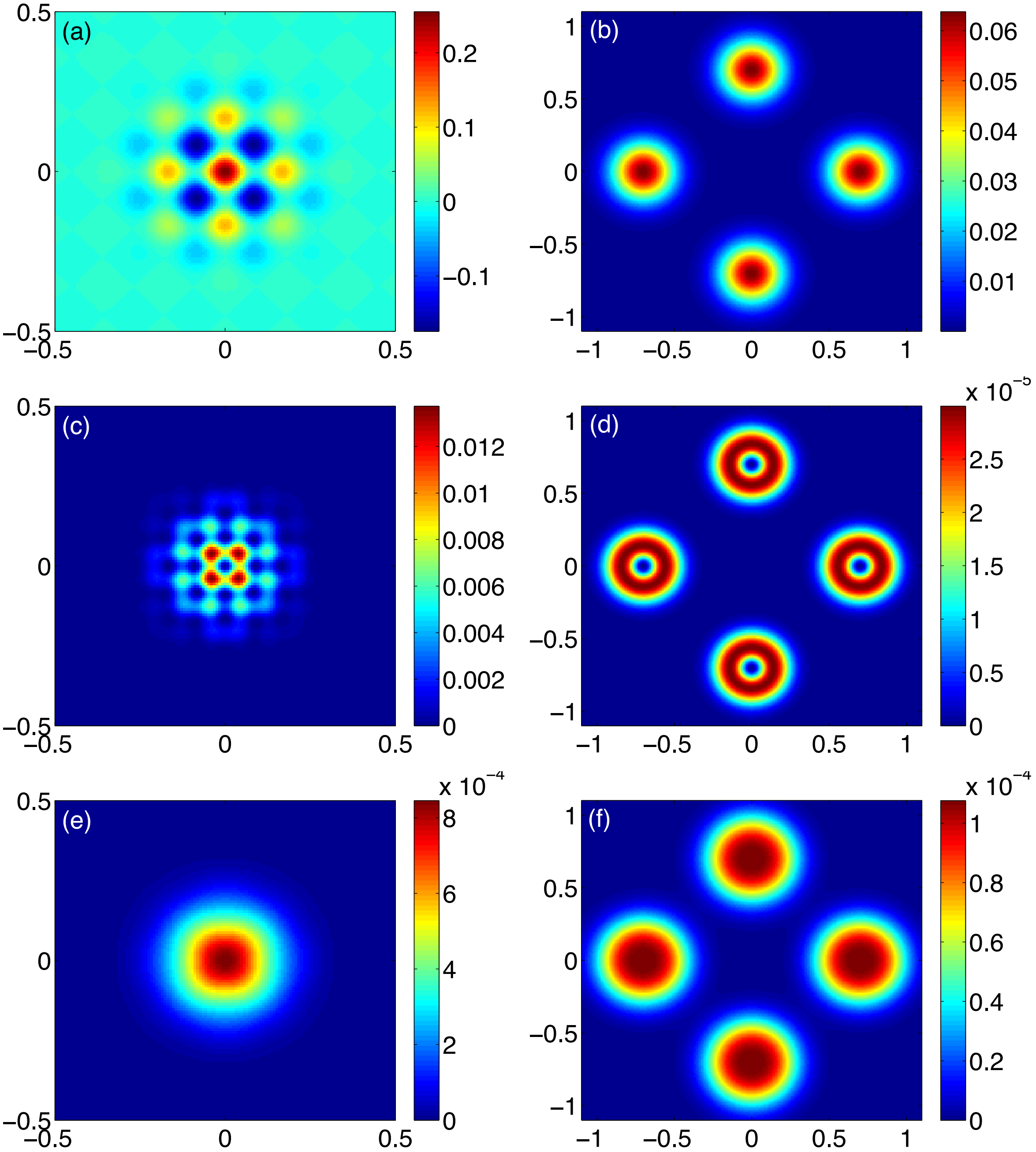

Figure 2: A superposition of eight Gaussian beams at a focus point. (a) the real part of the field, (b) the real part of the weighted Fourier transform, (c) the energy function in position space, (d) the energy function in Fourier space, (e) the smoothed energy function in position space, and (f) the smoothed energy function in Fourier space.

domains. Figure 3a-b shows the real part of the initial field in both domains. Figure 3c-d shows the associated energy functions. Even though the energy function of a single beam is a Gaussian, the energy function of the superposition shows oscillations of the order of $1/k$ due to interference between different beams. As the figure shows, the Gaussian structure of the energy function is not evident. Figure 3e-f shows the energy function after convolution with the smoothing kernel (2.15).

The algorithm was implemented on the domain depicted in figure 3 on a coarse 50×50 rectangular grid. In addition, points in which the energy was below a given threshold were ignored. This left fewer that 1000 points to consider in each domain ($M < 1000$). The EM fit resulted in 12 Gaussians, the smallest number which gave an EM-error smaller than 10%. The first iteration of the greedy outer loop yielded 11 GBs that approximate the initial field with a 22% error. Closer inspection of the result showed that the algorithm captured 8 out of the original 10 GBs correctly and compensated for the error with three other beams which were misplaced. The result demonstrates that the final optimization algorithm did not converge to the correct result even though we used enough Gaussians for the approximation. This is either because the minimization algorithm got trapped in a local minimum or that convergence was too slow. After the first iteration of the outer loop, the field generated by the 11 beams was subtracted from the initial one and the EM process was repeated. Using 5 Gaussians in the EM fit the algorithm approximates the difference field with an EM-error of 6%. Reconstructing GBs from the EM data yielded 4 GBs which were combined with the previous fit. Repeating the optimization step, which includes parameter optimization and removal non-contributing Gaussians, yielded 11 GBs with an 18% error. A Closer inspection showed that the 11 Gaussian include 9 of the original ones. The third iteration of the greedy outer loop yielded 10 GBs with a relative error of 8%.

4.4 A modulated plane wave

We fit a wave field given by a plane wave in 2D, modulated by a Gaussian in one dimension, as depicted in figure 4a. The EM fit resulted in 5 Gaussians in position space and a single one in Fourier space. Also, as explained in section 3.5, in order to prevent the optimization process from converging toward beams that are exceedingly wide, we add a penalty if the smallest eigenvalue of $\mathrm{Re}[\beta]$ is smaller than 0.3. This threshold corresponds to a beam whose standard deviation is about 0.5.

The first iteration of the greedy outer loop yielded two beams that fit the initial field with a 40% relative error. A second iteration yielded five additional beams which reduce the relative error to about 4%. The approximating beams are depicted in figure 4b-h.

4.5 The double slit experiment

To test our method with data that does not have an underlying Gaussian beam superposition and also exhibits many of the typical wave phenomena, including crossing waves

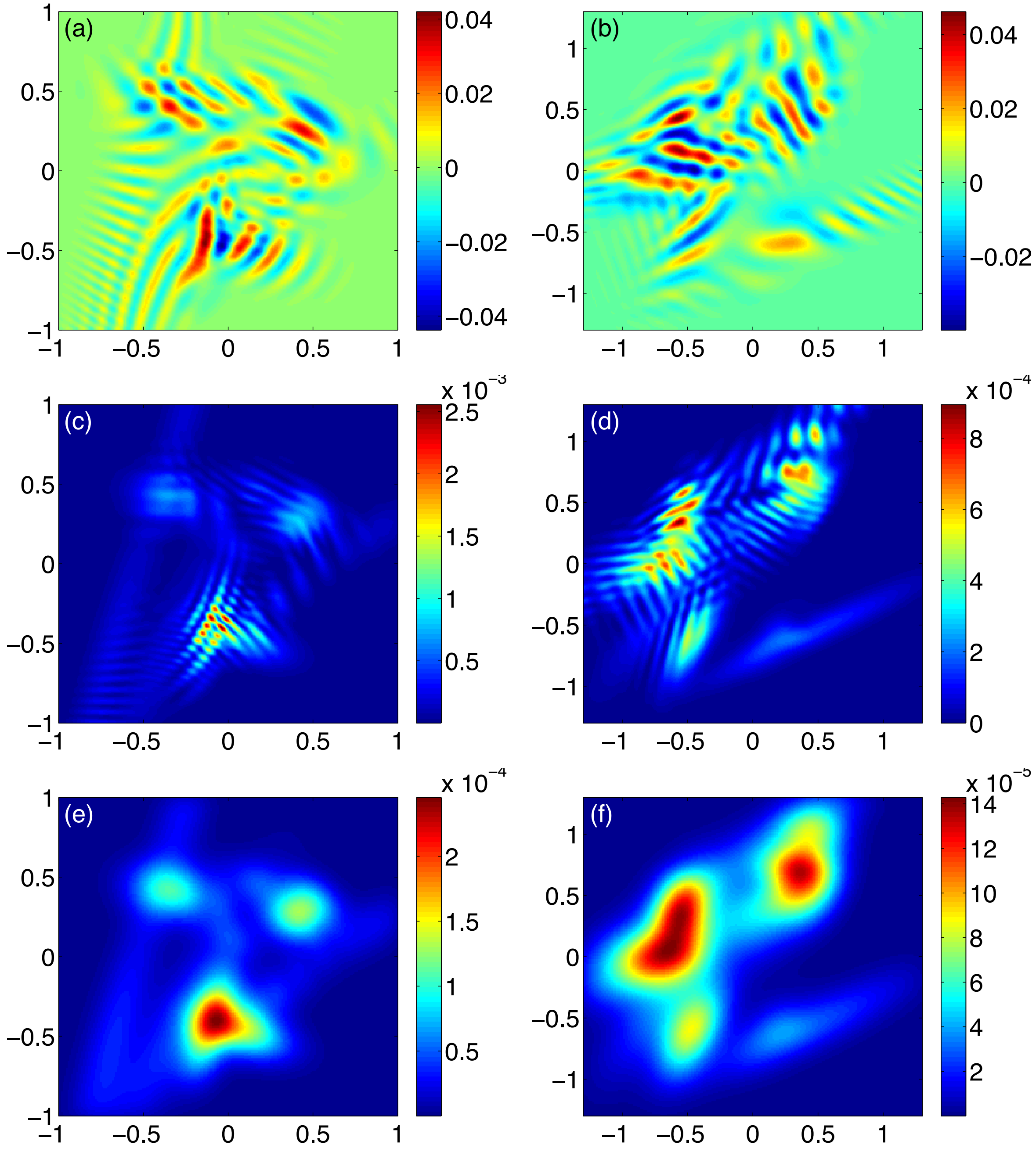

Figure 3: A superposition of ten Gaussian beams with random coefficients. (a) the real part of the field, (b) the real part of the weighted Fourier transform, (c) the energy function in position space, (d) the energy function in Fourier space, (e) the smoothed energy function in position space, and (f) the smoothed energy function in Fourier space.

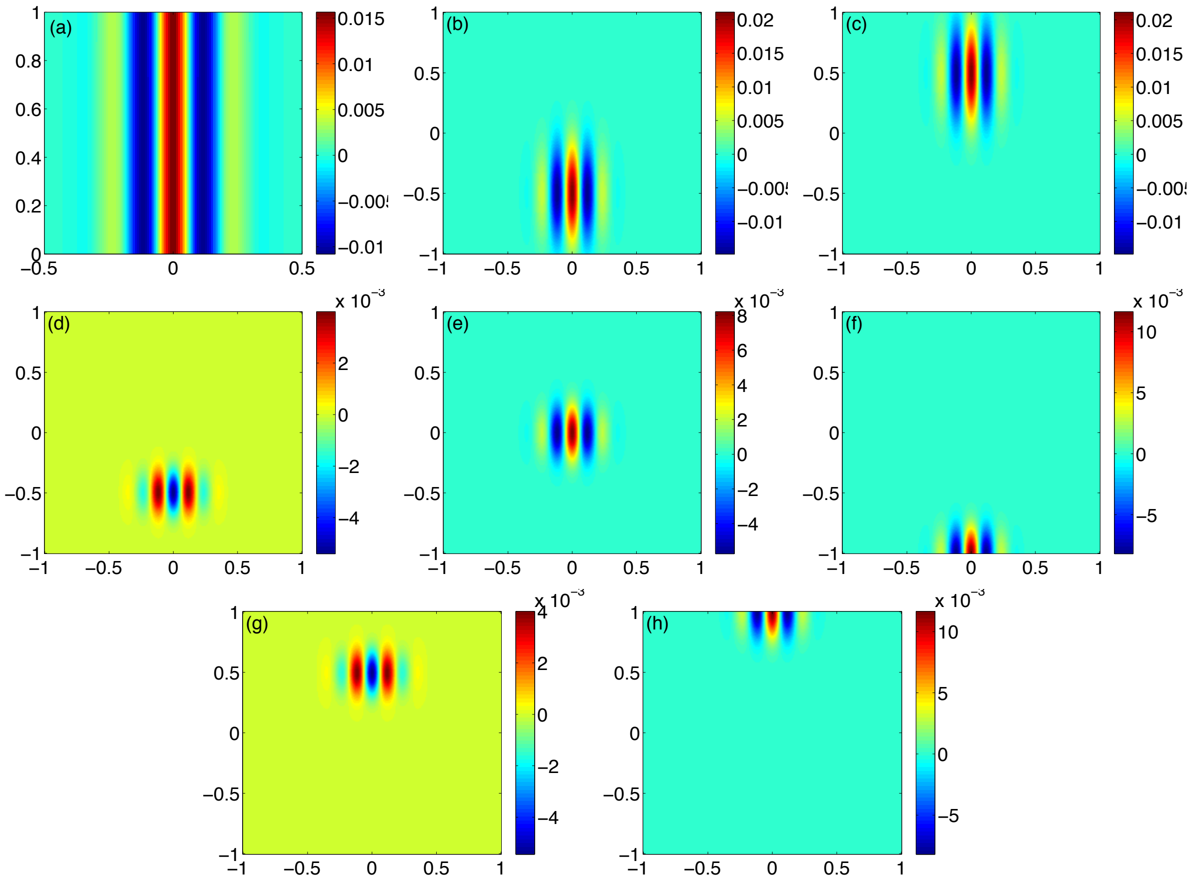

Figure 4: (a) A plane wave modulated by a Gaussian. (b)-(h) Seven beams approximating (a) with a 4% error.

and spreading, we look at the classical double slit experiment. To generate the data, we simulate coherent waves as they pass though two slits, using a standard second order finite difference method with absorbing boundary conditions [7]. The slits are closely spaced together and their width is similar to the wave length. The wave field after the waves have passed though the two slits in shown in figure 5. We will decompose this field into a sum of a few Gaussian beams.

Our method was applied using the smoothing kernel with width $l = 0.2$ in position space and $l = 0.3$ in Fourier space. With a cutoff EM-error at 5%, EM found four Gaussians in position space and five Gaussians in Fourier space. The first iteration of the greedy outer loop yielded 8 GBs with a 30% error. A second iteration yielded a total of 14 GBs with a 10% error.

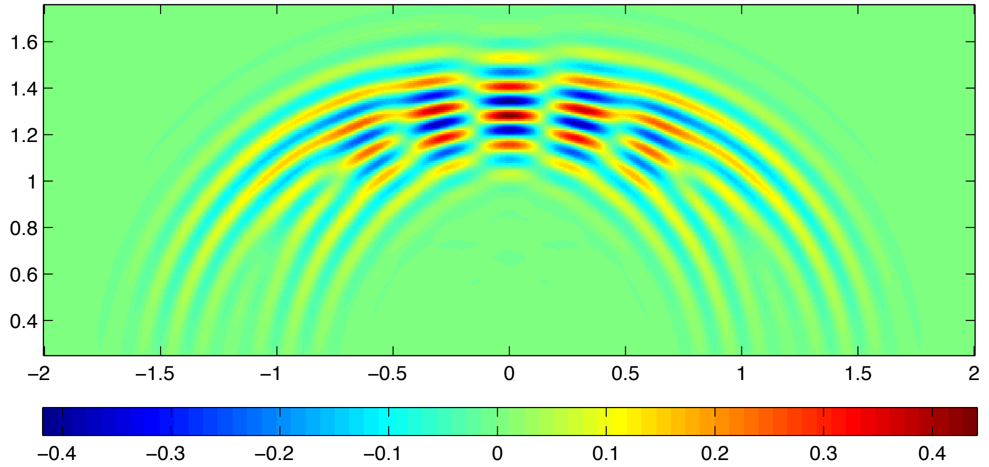

Figure 5: Real part of the wave field to be decomposed for the double slit experiment.

5 Summary

We presented a numerical method for approximating a high frequency wave field using Gaussian beams and applied it to decompose wave fields consisting of Gaussian beams and to more general wave fields in two dimensions.

Our approach approximates the energy functions of the wave equation in both the position space and the fourier space. By considering both spaces simultaneously, our strategy has an advantage of decomposing waves which may not be easily distinguished in one space but are separated in the other. We apply the well-established Expectation-Maximization algorithm which allows for efficient search of multiple Gaussians approximating the mollified energy functions. The EM fit is then processed into a superposition of Gaussian beams which approximates the high frequency wave field.

We demonstrate that our algorithm provides an efficient way of approximating high frequency wave fields by superposition of a relatively small number of Gaussian beams. We suggest that generalizations of our algorithm to other types of highly oscillatory fields, for example, fields generated by solutions to the Schrödinger equation, may be advantageous.

Acknowledgments

The authors were partially supported by the NSF under grant No. DMS-0714612. NT was also partially supported by the NSF under grant No. DMS-0636586 (UT Austin

RTG).

A The Gaussian beams approximation

In this appendix we review the derivation of rays and GBs in the variable coefficient wave equation in $\mathbb{R}^d$ (2.1). The GB equations are solved exactly for the simple case of constant propagation speed in 2D.

A.1 Geometric Optics

In order to obtain the high-frequency geometric optics approximation, one makes the ansatz

$$u(x,t) = a(x,t)e^{ik\phi(x,t)}, \tag{A.1}$$

where $k \gg 1$ is a large parameter characterizing the ratio between the wave length and the scale on which c varies (assumed to be of order one). Substituting (A.1) into the wave equation (2.1) and equating equal powers of k yields the eikonal equation for the phase ϕ and a transport equation for the amplitude a. To leading order in k,

$$\begin{aligned} &|\phi_t|^2 - c^2(x)|\nabla\phi(x)|^2 = 0 \\ &2\phi_t a_t - 2c^2(x)\nabla\phi\cdot\nabla a = -a\Box\phi. \end{aligned} \tag{A.2}$$

Without loss of generality, we assume that $\phi_t \leq 0$, otherwise, take $t \mapsto -t$. The eikonal equation has the form of a Hamilton-Jacobi equation

$$\phi_t + H(x,\nabla\phi) = 0, \tag{A.3}$$

with

$$H(x,p) = c(x)|p|. \tag{A.4}$$

In geometric optics, one describes waves using rays, which are the characteristics of (A.3). Parameterizing the characteristics by s, we look for a solution $\phi = \phi(t(s),x(s))$ and a trajectory $x(s)$ such that $z(s) = \phi(t(s),x(s))$ satisfies an ODE. We denote $p = \nabla_x\phi$, where the subscript is added in order to emphasize that the gradient is with respect to x, $\nabla_x u = (\partial_{x_1},\ldots,\partial_{x_d})^T$ and $[\cdot]^T$ denotes the transpose. Differentiating with respect to s yields

$$\begin{aligned} &\frac{dz}{ds} = \phi_t\frac{dt}{ds} + p\frac{dx}{ds} \\ &\frac{dp}{ds} = \frac{d}{ds}\nabla\phi(t(s),x(s)) = \nabla\phi_t\frac{dt}{ds} + \nabla_x\nabla_x^T\phi\frac{dx}{ds}. \end{aligned} \tag{A.5}$$

Note that $\nabla_x\nabla_x^T$ is the Hessian. In addition, differentiating the PDE (A.3) with respect to x yields

$$\nabla_x\phi_t + \nabla_x H(x,p) + \nabla_p H(x,p)\nabla_x\nabla_x^T\phi = 0. \tag{A.6}$$

We now see that if $dt/ds = 1$ and $dx/ds = \nabla_p H(x,p)$, then one could eliminate the second order term $(\nabla_x p^T)$ from (A.5). Substituting (A.6) into (A.5) and taking $s = t$ yields Hamilton's equations of motion

$$\begin{aligned} \dot{x} &= \nabla_p H(x,p) \\ \dot{p} &= -\nabla_x H(x,p), \end{aligned} \tag{A.7}$$

where dot denotes differentiation with respect to time t. For the case at hand, $H = c(x)|p|$, and the characteristics are given by

$$\begin{aligned} \dot{x} &= c(x)\hat{p} \\ \dot{p} &= -|p|\nabla c(x). \end{aligned} \tag{A.8}$$

with some initial conditions $x(0) = \xi$ and $p(0) = \nabla\phi(0) = \eta$. Here, $\hat{p} = p/|p|$. Thus, the Hamiltonian $H(x,p)$ is conserved under the dynamics, $H(x,p) = H(x_0,p_0) = H_0$. Since H is constant, (A.3) implies that along the rays the phase is linear in time

$$\phi(t) = \phi_0 - H_0 t. \tag{A.9}$$

Without loss of generality we take $\phi_0 = 0$ since the phase $e^{ik\phi_0}$ is just a multiplicative factor that does not change further derivations.

The GB approximation also requires the value of the Hessian, $\nabla\nabla^T\phi$ along the ray. Similar to the derivation of p, we write $M(s) = \nabla\nabla^T\phi(x(s))$ and differentiate with respect to s. The chain rule yields a three dimensional tensor involving all third order derivatives of ϕ with respect to x. Differentiating (A.6) with respect to x involves the same tensor. Thus, all third order derivatives can be eliminated. We obtain

$$\dot{M} = -M(\nabla_p\nabla_p^T H)M - M(\nabla_p\nabla_x^T H) - (\nabla_p^T\nabla_x H)M - (\nabla_x\nabla_x^T H). \tag{A.10}$$

Using (A.4)

$$\begin{aligned} \nabla_p\nabla_p^T H &= \frac{c(x)}{|p|}\left[\mathcal{I} - \hat{p}\hat{p}^T\right] \\ \nabla_p\nabla_x^T H &= \hat{p}(\nabla_x c(x))^T \\ \nabla_x\nabla_x^T H &= |p|\nabla_x\nabla_x^T c(x), \end{aligned} \tag{A.11}$$

where $\mathcal{I}$ is the identity matrix. Initial conditions are $M(0) = i\beta$ where $\mathrm{Re}\beta > 0$.

Similarly, the characteristics of the linear transport equation for the amplitudes (A.2) are found as follows. Denoting $X = (x,t) \in \mathbb{R}^d \times \mathbb{R}$, the characteristics for X, parameterized by s, satisfy $dX/ds = (2\phi_t, 2c^2(x)\nabla\phi)$. This can be written as,

$$\begin{aligned} t &= 2H_0 s = 2c(x)|p|s \\ \frac{dx}{dt} &= \frac{dx}{ds}\frac{ds}{dt} = 2c^2(x)p\frac{1}{2c(x)|p|} = c(x)\hat{p}. \end{aligned} \tag{A.12}$$

Hence, the characteristics of a are the same as those of ϕ (compare with (A.8)). As a result, the derivative of a along the ray is given by

$$\begin{aligned} \frac{da}{ds} &= a_t \frac{dt}{ds} + \nabla_x a \cdot \frac{dx}{ds} \\ &= 2H_0 a_t - 2c^2 p \cdot \nabla_x a = -(2a_t\phi_t - 2c^2(x)\nabla_x a \cdot \nabla_x \phi) = a\Box\phi, \end{aligned} \tag{A.13}$$

where we used the transport equation (A.2). Re-parameterizing with respect to time and using (A.4) and (A.9) yields

$$\dot{a} = 2H_0 a\Box\phi = -2c^3(x)|p|\mathrm{Tr}[M]a, \tag{A.14}$$

where $\mathrm{Tr}[\cdot]$ denotes the trace.

For example, with a constant speed $c(x) = c$,

$$\begin{aligned} p &= \eta \\ x &= \xi + c\eta t \\ \dot{M} &= -\frac{c}{|\eta|}M(\mathcal{I} - \hat{\eta}\hat{\eta}^T)M \ \ ; \ M(0) = i\beta \\ \dot{a} &= -2c^3|p|\mathrm{Tr}[M]a \ \ ; \ a(0) = A. \end{aligned} \tag{A.15}$$

Solving the equations for $M(t)$ and $a(t)$ yields in 2D (2.7).

A.2 Gaussian beams

The main idea underling the GB approximation is to expand the solution of the phase around a particular beam [11]. Let $X(t)$ denote the characteristics of a ray originating at $t = 0$ from a point ξ and with an initial direction η. The phase, its gradient and Hessian are denoted $\Phi(t)$, $P(t)$ and $M(t)$, respectively. In addition, denote the amplitude obtained by integrating (A.14) as $A(t)$. Note that we have chosen to parameterized the ray with respect to time. The GB approximation for $\phi(x,t)$ is a second order Taylor polynomial for ϕ around the point $(X(t),t)$, i.e.,

$$\begin{aligned} \phi(x,t) &= \Phi(t) + P(t)\cdot[x - X(t)] + \frac{1}{2}[x - X(t)]^T M(t)[x - X(t)] \\ a(x,t) &= A(t). \end{aligned} \tag{A.16}$$

Therefore,

$$\begin{aligned} u(x,0) =& Ae^{ik\eta\cdot(x-\xi)}e^{-k(x-\xi)^T\beta(x-\xi)/2} \\ \nabla_x u(x,0) =& ku(x,0)\left[i\eta - \beta(x-\xi)\right] \\ u_t(x,0) =& ku(x,0)\big[k^{-1}\dot{a}(t) + ic(x)|\eta| - i|\eta|\nabla c(x)\cdot(x-\xi) - \\ & c(x)\hat{\eta}^T\beta(x-\xi) + i(x-\xi)^T M'(0)(x-\xi)/2\big] \end{aligned} \tag{A.17}$$

where $A = a(0)$. The derivation above used the characteristic equation (A.8). Substituting into (2.3) yields the energy function. However, since the exponent is small for $|x-\xi| > 1/\sqrt{k}$, terms that are of order $|x-\xi|^2$ are comparable with $1/k$ and can be neglected. Similarly, $c(x)$ and $\nabla c(x)$ can also be expanded around ξ. The expansion yields

$$e(x,0) = |A|^2|\eta|^2 e^{-k(x-\xi)^T(\mathrm{Re}\beta)(\mathrm{x}-\xi)}\left[1 - c^{-2}(\xi)\nabla c(\xi)\cdot(x-\xi)\right] + O(1/k). \tag{A.18}$$

Furthermore, the leading in the expression above are symmetric with respect to the Gaussian center, ξ. Hence, terms that are proportional to $x-\xi$ cancel upon integration of $e(x,0)$. To keep notation simple, we write

$$e(x,0) = 2|A|^2|\eta|^2 e^{-k(x-\xi)^T(\mathrm{Re}\beta)(\mathrm{x}-\xi)} + O(1/\sqrt{k}). \tag{A.19}$$

and remember that the contribution of the last term to the overall energy is smaller. In the case of constant propagation speed, $c(x) = 1$, and the energy function reduces to (2.8).

Hence, the contribution of exponential terms that are multiplied by polynomials to the total energy is of order $1/k$.

Similarly, in the frequency domain, substituting (A.16) into (2.9) yields

$$\tilde{u}(p,t) = k^{-1/2}a(t)e^{ik\Phi(t)}e^{ikp\cdot X(t)}e^{-i[p+P(t)]^T M^{-1}(t)[p+P(t)]}. \tag{A.20}$$

At $t = 0$ this expression becomes (2.10). Substituting into the Fourier energy function (2.11) and using the characteristic equation (A.8) yields (2.12).

B Expectation-Maximization

In this section we describe how to apply the expectation-maximization (EM) algorithm [5] to approximate a probability distribution given on a set of points using Gaussian random variables.

Let $f(x)$ denote a non-negative density function on $\mathbb{R}^d$. Let $X = \{x_i\}_{i=1}^M$ denote a list of M points in $\mathbb{R}^d$ and denote $f_i = f(x_i)$. Without loss of generality we assume that $f(x)$ induces a probability measure on X, i.e., $\sum_{i=1}^M f_i = 1$. For example, in the numerical examples described in section 4, $\{x_i\}_{i=1}^M$ are the points on a rectangular two-dimensional grid with $f(x)$ above some fixed threshold.

The purpose of this section is to use the sample of $f(x)$ at the points x_i in order to find a linear combination of N that approximates $f(x)$ in a probabilistic sense Gaussians

$$g(x) = \sum_{j=1}^N A_j G_j(x)\ ; \quad G_j(x) = z_i^{-1} e^{-(x-\mu_j)\cdot\Sigma_j^{-1}(x-\mu_j)/2}. \tag{B.1}$$

Here, z_i are normalization constants such that $\sum_{i=1}^{M} G_j(x_i) = 1$ for all $j = 1 \dots N$ and $\sum_{i=1}^{M} A_i = 1$. Hence, $G_j(x)$ and $g(x)$ are probability measures on X. To this end, let θ denote the set of parameters defining the N Gaussians, i.e., $\theta = \{A_j, \mu_j, \Sigma_j\}_{j=1}^{N}$, where, for all $j = 1 \dots N$, $A_j \in \mathbb{R}$, $A_j \geq 0$, $\mu_j \in \mathbb{R}^d$ and Σ_j are positive definite $d \times d$ matrices. In addition, we require that $\sum_{i=1}^{M} A_i = 1$. An intuitive approach for constructing an EM algorithm is to think of a random game that generates points from X with probabilities $g_i = g(x_i)$. First choose a Gaussian j out of $\{1, \dots, N\}$ with probabilities $A_1, \dots, A_N$, respectively. Then, draw a point x_i with probability $G_j(x_i)$. Thus, for fixed θ, the probability for getting x_i is g_i. Indeed, the probability that point x_i was generated from the Gaussian G_j, denoted p_{ij}, is

$$p_{ij} = A_j G_j(x_i)/g(x_i). \tag{B.2}$$

Therefore, A_j is the average number of points chosen from Gaussian j, weighted by g_i:

$$A_j = \sum_i g_i p_{ij}, \tag{B.3}$$

μ_j is the weighted average position of points drawn from Gaussian j,

$$\mu_j = \sum_i g_i \frac{p_{ij}}{A_j} x_i, \tag{B.4}$$

and Σ_j is the associated covariance matrix

$$\Sigma_j = \sum_i g_i \frac{p_{ij}}{A_j} x_i x_i^T - \mu_j \mu_j^T. \tag{B.5}$$

We see that, if $g_i = f_i$ for all i, then θ is a fixed point of the map $\theta \to \theta' = \{A'_j, \mu'_j, \Sigma'_j\}_{j=1}^{N}$ given by:

$$\begin{aligned} A'_j &= \sum_i f_i p_{ij} \\ \mu'_j &= \sum_i f_i \frac{p_{ij}}{A'_j} x_i \\ \Sigma'_j &= \sum_i f_i \frac{p_{ij}}{A'_j} x_i x_i^T - \mu'_j (\mu')_j^T. \end{aligned} \tag{B.6}$$

Following [5] one can show that, for general non-negative and normalized $f(x)$, (B.6) defines a contraction and that fixed points are local minima for the likelihood of obtaining a distribution f_i over X from a set of parameters θ.

Summarizing, for the case at hand the EM algorithm can be applied as follows: start with an initial guess θ_0 and iterate (B.6) until the process converges within a given tolerance. The resulting parameters describe a linear combination of Gaussians of the form of B.1 that approximate the distribution $f(x)$ on X.

C Deconvolving variances

In section 2 we saw that convolving the energy function of a GB with the smoothing kernel function (2.15) changes the variance matrix. In 1D, the relation between the original variance B^{-1} and the convolved Σ^{-1} is:

$$\Sigma(B) = \frac{B}{1+2lB}. \tag{C.1}$$

In 2D it is given by (2.17):

$$\Sigma(B) = \frac{B + 2l(\det B)\mathcal{I}}{1 + 2l\mathrm{Tr}B + 4l^2 \det B}, \tag{C.2}$$

where $\mathcal{I}$ is the identity matrix. As a result, fitting a Gaussian to the convolved energy function yields a biased variance. Hence, we would like to invert the above formulas. This is simple in 1D. In two or more dimensions we use the fact that the shift in B is independent of k, but is of order $l < 1$. We rewrite (C.2) as

$$B = \Sigma + 2l\left[(\mathrm{Tr}B)\Sigma - (\det B)\mathcal{I}\right] + 4l^2(detB)\Sigma. \tag{C.3}$$

Recall that l is known and we are solving for B. For small enough values of l, the solution can be done iteratively by taking

$$\begin{aligned} B_0 &= \Sigma \\ B_{j+1} &= \Sigma + 2l\left[(\mathrm{Tr}B_j)\Sigma - (\det B_j)\mathcal{I}\right] + l^2 4(detB_j)\Sigma. \end{aligned} \tag{C.4}$$

In Fourier space, one is actually looking for B^{-1} rather than B. While it is still possible to use (C.4) and invert, we found that this approach introduced a large numerical error if $|\det B|$ is small. Instead, one can invert (C.3), expand to some order in l and solve iteratively. For example, the order two approximation is

$$\begin{aligned} B_0^{-1} &= \Sigma^{-1} \\ B_{j+1}^{-1} &= \left[1 - lC_j + l^2(C_j^2 - D_j) + O(l^3)\right]\Sigma^{-1}, \end{aligned} \tag{C.5}$$

where

$$\begin{aligned} C_j &= -\frac{2}{\det B_j}\Sigma^{-1} + \frac{\mathrm{Tr}B_j}{\det B_j}\mathcal{I} \\ D_j &= \frac{4}{\det B_j}\mathcal{I}. \end{aligned} \tag{C.6}$$

D Reconstructing β

Let β denote a complex $d \times d$ symmetric matrix with a positive definite real part. Denote $A = \mathrm{Re}[\beta]$ and $C = \mathrm{Re}[\beta^{-1}]$. One can show that C is also positive definite. In this appendix we address the following problem: given A and C, can one determine β?

D.1 General dimension

Denote $B = \mathrm{Im}[\beta]$ and $D = \mathrm{Im}[\beta^{-1}]$, i.e., $\beta = A + iB$ and $\beta^{-1} = C + iD$. Then,

$$\mathcal{I} = \beta\beta^{-1} = (A + iB)(C + iD) = (AC - BD) + i(AD + BC), \tag{D.1}$$

where $\mathcal{I}$ is the identity matrix. Hence

$$\begin{aligned} AC - BD &= \mathcal{I} \\ AD + BC &= 0. \end{aligned} \tag{D.2}$$

Since A is positive definite it is invertible and

$$D = -A^{-1}BC.$$

Substituting into the real part of (D.1) and multiplying by C^{-1} yields

$$BA^{-1}B = C^{-1} - A. \tag{D.3}$$

Equation (D.3) is a quadratic equation for the missing imaginary part, B.

Since A is real and symmetric, its inverse is diagonalizable with an orthonormal matrix, i.e., $A^{-1} = Q\Lambda Q^T$, $\Lambda = \mathrm{diag}\{\lambda_1, \dots, \lambda_d\}$. Multiplying (D.3) by Q^T on the left and Q on the right yields

$$\tilde{B}\Lambda\tilde{B} = H,$$

where $\tilde{B} = Q^T BQ$ and $H = Q^T(C^{-1} - A)Q$ are real and symmetric matrices. Hence, without loss of generality, we need to solve a matrix equation (for B) of the form

$$B\Lambda B = H. \tag{D.4}$$

Denoting the entries of B and H by $\{b_{ij}\}_{i,j=1}^d$ and $\{h_{ij}\}_{i,j=1}^d$, respectively, (D.4) consists of $n = d(d+1)/2$ equations and unknowns:

$$eq(i,j): \quad \sum_{k=1}^{d} \lambda_j b_{ik} b_{jk} = h_{ij}, \tag{D.5}$$

for all $i \leq j$. Arranging the matrix elements $\{b_{ij}\}_{i \leq j}$ in the form of a vector $\mathbf{b}$, (D.5) can be written as n quadratic equations

$$eq(i,j): \quad \mathbf{b}^T L^{ij} \mathbf{b} = h_{ij}, \tag{D.6}$$

where, for each $i \leq j$, L^{ij} is a sparse $n \times n$ symmetric matrix. For example, in two dimensions

$$L^{11} = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & 0 \end{pmatrix}; \; L^{12} = \frac{1}{2}\begin{pmatrix} 0 & \lambda_1 & 0 \\ \lambda_1 & 0 & \lambda_2 \\ 0 & \lambda_2 & 0 \end{pmatrix}; \; L^{22} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \lambda_1 & 0 \\ 0 & 0 & \lambda_2 \end{pmatrix}$$

and in three dimensions

$$L^{11} = \begin{pmatrix} \lambda_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \lambda_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \; ; \; L^{12} = \frac{1}{2} \begin{pmatrix} 0 & \lambda_1 & 0 & 0 & 0 & 0 \\ \lambda_1 & 0 & 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \lambda_3 & 0 \\ 0 & \lambda_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \lambda_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$L^{13} = \frac{1}{2} \begin{pmatrix} 0 & 0 & \lambda_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \lambda_2 & 0 \\ \lambda_1 & 0 & 0 & 0 & 0 & \lambda_3 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \lambda_3 & 0 & 0 & 0 \end{pmatrix} \; ; \; L^{22} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \lambda_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \lambda_3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$L^{23} = \frac{1}{2} \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \lambda_2 & 0 \\ 0 & 0 & 0 & \lambda_2 & 0 & \lambda_3 \\ 0 & 0 & 0 & 0 & \lambda_3 & 0 \end{pmatrix} \; ; \; L^{33} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \lambda_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \lambda_2 & 0 \\ 0 & 0 & 0 & 0 & 0 & \lambda_3 \end{pmatrix} .$$

Since all eigenvalues λ_j are strictly positive, the diagonal equations $e(i,i)$ are elliptic cylinders. The rest are hyperbolic cylinders. Furthermore, since the free axes of the cylinders are orthogonal, there is no degeneracy and the number of solutions is finite, 2^d at most.

Generally, (D.4) should be solved numerically. Since the equations are quadratic, Newton-Raphson in accurate and efficient. The principle axes of the cylinders described in (D.6) divide R^d into regions that correspond to the basins of attractions of the different solutions. Hence, all solutions can be identified by appropriately chosen initial guesses.

D.2 Two dimensions

In 2D equation (D.4) can be solved analytically. First, we note that (D.4) is homogeneous in the sense that, for all $\alpha > 0$,

$$(\sqrt{\alpha}B)\Lambda(\sqrt{\alpha}B) = \alpha H.$$

Hence, without loss of generality we take $\lambda_2 = 1/\lambda_1$ and consider two cases: $\det H = 0$ and $\det H = 1$. Solving using Mathematica and simplifying yields four solutions. With $\det H = 0$,

$$B = \sqrt{\frac{\lambda_1}{h_{22} + h_{11}\lambda_1}} \begin{pmatrix} h_{11} & \pm h_{12} \\ \pm h_{12} & h_{22} \end{pmatrix},$$

and $-B_{\pm}$. With $\det H = 1$, if $h_{12} \neq 0$

$$B_{\pm} = \sqrt{\frac{\lambda_1\lambda_2}{h_{11}\lambda_1 + h_{22}\lambda_2 \pm 2\lambda_1\lambda_2}} \begin{pmatrix} h_{11} \pm \lambda_2 & h_{12} \\ h_{12} & h_{22} \pm \lambda_1 \end{pmatrix},$$

and

$$B_{\pm} = \begin{pmatrix} \sqrt{h_{11}\lambda_1} & 0 \\ 0 & \pm\sqrt{h_{22}\lambda_2} \end{pmatrix},$$

otherwise. The other two solutions are $-B_{\pm}$. It can be shown that in all the expressions above the denominator is strictly positive.

References

[1] V. M. Babich and V. S. Buldyrev. *Asymptotic methods in short wave diffraction problems. Vol. 1.* “Nauka”, Moscow (in Russian), 1972.

[2] J.D. Benamou, F. Collino, and O. Runborg. Numerical microlocal analysis of harmonic wavefields. *J. Comput. Phys.*, 199(2):717–741, 2004.

[3] D. Bouche, F. Molinet, and R. Mittra. *Asymptotic methods in electromagnetics.* Springer-Verlag, Berlin, 1997.

[4] S. Bougacha, J.-L. Akian, and R. Alexandre. Gaussian beams summation for the wave equation in a convex domain. *Commun. Math. Sci.*, 7(4):973–1008, 2009.

[5] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. B (Methodological)*, 39:1–38, 1977.

[6] B. Engquist and O. Runborg. Computational high frequency wave propagation. *Acta Numer.*, 12:181–266, 2003.

[7] Bjorn Engquist and Andrew Majda. Absorbing boundary conditions for the numerical simulation of waves. *Mathematics of Computation*, 31(139):629–651, 1977.

[8] N. R. Hill. Gaussian beam migration. *Geophys.*, 55:1416–1428, 1990.

[9] V. P. Maslov and M. V. Fedoriuk. *Semiclassical approximation in quantum mechanics*, volume 7 of *Mathematical Physics and Applied Mathematics.* D. Reidel Publishing Co., Dordrecht, 1981. Translated from the Russian by J. Niederle and J. Tolar, Contemporary Mathematics, 5.

[10] J. A. Nelder and R. Mead. A simplex method for function minimization. *The Computer Journal*, 7(4):308–313, 1965.

[11] J. Ralston. Gaussian beams and the propagation of singularities. In *Studies in partial differential equations*, volume 23 of *Maa Stud. Math.*, pages 206–248. Math. Assoc. America, Washington, DC, USA, 1982.

[12] N. Tanushev. Superpositions and higher order Gaussian beams. *Commun. Math. Sci.*, 6(2):449–475, 2008.

[13] N. Tanushev, B. Engquist, and R. Tsai. Gaussian beam decomposition of high frequency wave fields. *J. Comp. Phys.*, 228:8856–8871, 2009.

[14] A.K. Tornberg and B. Engquist. The segment projection method for interface tracking. *Commun. Pure Appl. Math.*, 56:47–79, 2003.

[15] W. Yin, S. Osher, D. Goldfarb, and J. Darbon. Bregman iterative algorithm for compressed sensing and related problems. *SIAM J. Imag. Sci.*, 1(1):143–168, 2008.