Immersed Interface/Boundary Method

Kazufumi Ito and Zhilin Li
Center for Research in Scientific Computation and Department of Mathematics, North Carolina State University, Raleigh, NC, USA

Introduction

The immersed interface method (IIM) is a numerical method for solving interface problems or problems on irregular domains. Interface problems are considered as partial differential equations (PDEs) with discontinuous coefficients, multi-physics, and/or singular sources along a co-dimensional space. The IIM was originally introduced by LeVeque and Li [7] and Li [8] and further developed in [1, 11]. A monograph of IIM has been published by SIAM in 2006 [12].

The original motivation of the immersed interface method is to improve accuracy of Peskin's immersed boundary (IB) method and to develop a higher-order method for PDEs with discontinuous coefficients. The IIM method is based on uniform or adaptive Cartesian/polar/spherical grids or triangulations. Standard finite difference or finite element methods are used away from interfaces or boundaries. A higher-order finite difference or finite element schemes are developed near or on the interfaces or boundaries according to the interface conditions, and it results in a higher accuracy in the entire domain. The method employs continuation of the solution from the one side to the other side of the domain separated by the interface. The continuation procedure uses the multivariable Taylor's expansion of the solution at selected interface points. The Taylor coefficients are then determined by incorporating the interface conditions and the equation. The necessary interface conditions are derived from the physical interface conditions.

Since interfaces or irregular boundaries are one dimensional lower than the solution domain, the extra costs in dealing with interfaces or irregular boundaries are generally insignificant. Furthermore, many available software packages based on uniform Cartesian/polar/spherical grids, such as FFT and fast Poisson solvers, can be applied easily with the immersed interface method. Therefore, the immersed interface method is simple enough to be implemented by researchers and graduate students who have reasonable background in finite difference or finite element methods, but it is powerful enough to solve complicated problems with a high-order accuracy.

Immersed Boundary Method and Interface Modeling

The immersed boundary (IB) method was originally introduced by Peskin [22,23] for simulating flow patterns around heart valves and for studying blood flows in a heart [24]. First of all, the immersed boundary method is a *mathematical model* that describes elastic structures (or membranes) interacting with fluid flows. For instance, the blood flows in a heart can be considered as a Newtonian fluid governed by the Navier-Stokes equations

$$\rho\left(\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u}\cdot\nabla)\mathbf{u}\right) + \nabla p = \mu\Delta\mathbf{u} + \mathbf{F}, \qquad (1)$$

with the incompressibility condition $\nabla\cdot\mathbf{u} = 0$, where ρ is fluid density, $\mathbf{u}$ fluid velocity, p pressure, and μ fluid

B. Engquist (ed.), *Encyclopedia of Applied and Computational Mathematics*,
DOI 10.1007/978-3-540-70529-1

viscosity. The geometry of the heart is complicated and is moving with time, so are the heart valves, which makes it difficult to simulate the flow patterns around the heart valves. In the immersed boundary model, the flow equations are extended to a rectangular box (domain) with a periodic boundary condition; the heart boundary and valves are modeled as elastic band that exerts force on the fluid. The immersed structure is typically represented by a collection of interacting particles X_k with a prescribed force law. Let $\delta(\mathbf{x})$ be the Dirac delta function. In Peskin's original immersed boundary model, the force is considered as source distribution along the boundary of the heart and thus can be written as

$$\mathbf{F}(\mathbf{x},t) = \int_{\Gamma(\mathbf{s},t)} \mathbf{f}(\mathbf{s},t)\delta(\mathbf{x}-\mathbf{X}(\mathbf{s},t))d\mathbf{s}, \qquad (2)$$

where $\Gamma(\mathbf{s},t)$ is the surface parameterized by $\mathbf{s}$ which is one dimensional in 2D and two dimensional in 3D, say a heart boundary, $\mathbf{f}(\mathbf{s},t)$ is the force density. Since the boundary now is immersed in the entire domain, it is called the *immersed boundary*. The system is closed by requiring that the elastic immersed boundary moves at the local fluid velocity:

$$\begin{aligned}\frac{d\mathbf{X}(\mathbf{s},t)}{dt} &= \mathbf{u}(\mathbf{X}(\mathbf{s},t),t)\\ &= \int \mathbf{u}(\mathbf{x},t)\,\delta\,(\mathbf{x}-\mathbf{X}(\mathbf{s},\mathbf{t}),\mathbf{t})\,d\mathbf{x}, \qquad (3)\end{aligned}$$

here the integration is over the entire domain.

For an elastic material, as first considered by Peskin, the force density is given by

$$\mathbf{f}(\mathbf{s},t) = \frac{\partial \mathbf{T}}{\partial \mathbf{s}}\boldsymbol{\tau}, \qquad \mathbf{T}(\mathbf{s},t) = \sigma\left(\left|\frac{\partial \mathbf{X}}{\partial \mathbf{s}}\right| - 1\right), \qquad (4)$$

the unit tangent vector $\boldsymbol{\tau}(\mathbf{s},t)$ is given by $\boldsymbol{\tau}(\mathbf{s},t) = \dfrac{\partial \mathbf{X}/\partial \mathbf{s}}{|\partial \mathbf{X}/\partial \mathbf{s}|}$. The tension $\mathbf{T}$ assumes that elastic fiber band obeys a linear Hooke's law with stiffness constant σ. For different applications, the key of the immersed boundary method is to derive the force density.

In Peskin's original IB method, the blood flow in a heart is embedded in a rectangular box with a periodic boundary condition. In numerical simulations, a uniform Cartesian grid (x_i, y_j, z_k) can be used.

An important feature of the IB method is to use a discrete delta function $\delta_h(\mathbf{x})$ to approximate the Dirac delta function $\delta(\mathbf{x})$. There are quite a few discrete delta functions $\delta_h(\mathbf{x})$ that have been developed in the literature. In three dimensions, often a discrete delta function $\delta_h(\mathbf{x})$ is a product of one-dimensional ones,

$$\delta_h(\mathbf{x}) = \delta_h(x)\delta_h(y)\delta_h(z). \qquad (5)$$

A traditional form for $\delta_h(x)$ was introduced in [24]:

$$\delta_h(x) = \begin{cases} \dfrac{1}{4h}\left(1+\cos\left(\pi x/2h\right)\right), & \text{if } |x| < 2h,\\ 0, & \text{if } |x| \geq 2h. \end{cases} \qquad (6)$$

Another commonly used one is the hat function:

$$\delta_h(x) = \begin{cases} (h-|x|)/h^2, & \text{if } |x| < h,\\ 0, & \text{if } |x| \geq h. \end{cases} \qquad (7)$$

With Peskin's discrete delta function approach, one can discretize a source distribution on a surface Γ as

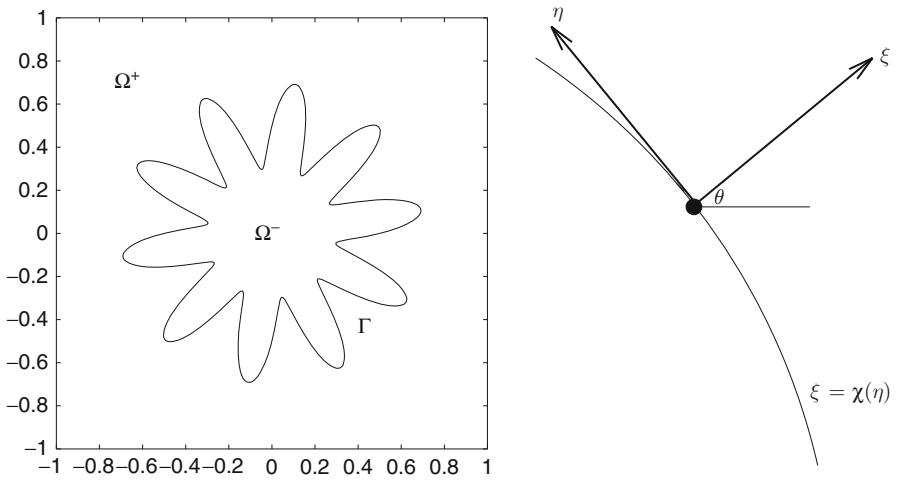
$$\mathbf{F}_{ijk} = \sum_{l=1}^{N_b} \mathbf{f}\,(\mathbf{s}_l)\;\delta_h(x_i - X_\ell)\,\delta_h(y_j - Y_\ell)\,\delta_h(z_k - Z_\ell)\Delta\mathbf{s}_l, \qquad (8)$$

where N_b is the number of discrete points $\{(X_\ell, Y_\ell, Z_\ell)\}$ on the surface $\Gamma(\mathbf{s},t)$. In this way, the singular source is distributed to the nearby grid points in a neighborhood of the immersed boundary $\Gamma(\mathbf{s},t)$. The discrete delta function approach cannot achieve second-order or higher accuracy except when the interface is aligned with a grid line.

In the immersed boundary method, we also need to interpolate the velocity at grid points to the immersed boundary corresponding to (3). This is done again through the discrete delta function

$$\begin{aligned}u(X,Y,Z) = \sum_{ijk} u(x_i, y_j, z_k)\delta_h(x_i - X)\\ \delta_h(y_j - Y)\delta_h(z_k - Z)h_x h_y h_z \qquad (9)\end{aligned}$$

assume (X,Y,Z) is a point on the immersed boundary $\Gamma(\mathbf{s},t)$, h_x, h_y, h_z are mesh sizes in each coordinate direction. Once the velocity is computed, the new location of the immersed boundary is updated through (3). Since the flow equation is defined on a rectangular



Immersed Interface/Boundary Method, Fig. 1 *Left diagram*: a *rectangular* domain $\Omega = \Omega^+ \cup \Omega^-$ with an interface Γ. The coefficients such as $\beta(\mathbf{x})$ have a jump across the interface. *Right diagram*: the local coordinates in the normal and tangential directions, where θ is the angle between the x-axis and the normal direction

domain, standard numerical methods can be applied. For many application problems in mathematical biology, the projection method is used for small to modest Reynolds numbers.

The immersed boundary method is simple and robust. It has been combined with and with adaptive mesh refinement [26, 27]. A few IB packages are available [3]. The IB method has been applied to many problems in mathematical biology and computational fluid mechanics. There are a few review articles on IB method given. Among them are the one given by Peskin in [25] and Mittal and Iaccarina [19] that highlighted the applications of IB method on computational fluid dynamics problems. The immersed boundary method is considered as a regularized method, and it is believed to be first-order accurate for the velocity, which has been confirmed by many numerical simulations and been partially proved [20].

The Immersed Interface Method

We describe the second immersed interface method for a scalar elliptic equation in two-dimensional domain, and we refer to [17] and references therein for general equations, fourth-order method, and the three-dimensional case. A simplified Peskin's model can be rewritten as a Poisson equation of the form:

$$\nabla \cdot (\beta(\mathbf{x}) \nabla u) - \sigma(\mathbf{x})\, u = f(\mathbf{x}), \qquad \mathbf{x} \in \Omega - \Gamma,$$
$$[u]\Big|_\Gamma = 0, \qquad [\beta u_n]\Big|_\Gamma = v(s) \tag{10}$$

where $v(\mathbf{s}) \in C^2(\Gamma)$, $f(\mathbf{x}) \in C(\Omega)$, Γ is a smooth interface, and β is a piecewise constant. Here $u_n = \frac{\partial u}{\partial \mathbf{n}} = \nabla u \cdot \mathbf{n}$ is the normal derivative, and $\mathbf{n}$ is the unit normal direction, and $[u]$ is the difference of the limiting values from different side of the interface Γ, so is $[u_n]$; see Fig. 1 (Left diagram) for an illustration.

Given a Cartesian mesh $\{(x_i, y_j\}$; $x_i = i\, h_x$, $0 \le i \le M$ $y_j = j\, h_y$, $0 \le j \le N$ with the mesh size h_x, h_y, the node (x_i, y_j) is irregular if the central five-point finite difference stencil at (x_i, y_j) has grid points from both side of the interface Γ, otherwise is regular. The IIM uses the standard five-point finite difference scheme at regular grid:

$$\frac{\beta_{i+\frac{1}{2},j} u_{i+1,j} + \beta_{i-\frac{1}{2},j} u_{i-1,j} - (\beta_{i+\frac{1}{2},j} + \beta_{i-\frac{1}{2},j}) u_{ij}}{(h_x)^2}$$
$$+ \frac{\beta_{i,j+\frac{1}{2}} u_{i,j+1} + \beta_{i,j-\frac{1}{2}} u_{i,j-1} - (\beta_{i,j+\frac{1}{2}} + \beta_{i,j-\frac{1}{2}}) u_{ij}}{(h_y)^2}$$
$$-\sigma u_{ij} = f_{ij}. \tag{11}$$

The local truncation error at regular grid points is $O(h^2)$, where $h = \max\{h_x, h_y\}$.

If (x_i, y_j) is an irregular grid point, then the method of undetermined coefficients

$$\sum_{k=1}^{n_s} \gamma_k\, U_{i+i_k, j+j_k} - \sigma_{ij}\, U_{ij} = f_{ij} + C_{ij} \tag{12}$$

is used to determine γ_k's and C_{ij}, where n_s is the number of grid points in the finite difference stencil. We usually take $n_s = 9$. We determine the coefficients in such a way that the local truncation error

$$T_{ij} = \sum_{k=1}^{n_s} \gamma_k \, u\left(x_{i+i_k}, y_{j+j_k}\right) - \sigma_{ij} \, u(x_i, y_j) - f(x_i, y_j) - C_{ij}, \tag{13}$$

is as small as possible in the magnitude.

We choose a projected point $\mathbf{x}_{ij}^* = (x_i^*, y_j^*)$ on the interface Γ of irregular point (x_i, y_j). We use the Taylor expansion at $\mathbf{x}_{ij}^*$ in the local coordinates (ξ, η) so that (12) matches (10) up to second derivatives at $\mathbf{x}_{ij}^*$ from a particular side of the interface, say the $-$ side. This will guarantee the consistency of the finite difference scheme. The local coordinates in the normal and tangential directions is

$$\begin{aligned} \xi &= (x - x^*)\cos\theta + (y - y^*)\sin\theta, \\ \eta &= -(x - x^*)\sin\theta + (y - y^*)\cos\theta, \end{aligned} \tag{14}$$

where θ is the angle between the x-axis and the normal direction, pointing to the direction of a specified side. In the neighborhood of (x^*, y^*), the interface Γ can be parameterized as

$$\xi = \chi(\eta), \quad \text{with} \quad \chi(0) = 0, \quad \chi'(0) = 0. \tag{15}$$

The interface conditions are given

$$[u(\chi(\eta), \eta)] = 0, \quad [\beta(u_\xi(\chi(\eta), \eta) - \chi'(\eta)\, u_\eta(\chi(\eta), \eta))] = \sqrt{1 + |\chi'(\eta)|^2} v(\eta)$$

and the curvature of the interface at (x^*, y^*) is $\chi''(0)$. The Taylor expansion of each $u(x_{i+i_k}, y_{j+j_k})$ at $\mathbf{x}_{ij}^*$ can be written as

$$\begin{aligned} u(x_{i+i_k}, y_{j+j_k}) = u(\xi_k, \eta_k) = u^\pm + \xi_k u_\xi^\pm + \eta_k u_\eta^\pm \\ + \frac{1}{2}\xi_k^2 u_{\xi\xi}^\pm + \xi_k \eta_k u_{\xi\eta}^\pm + \frac{1}{2}\eta_k^2 u_{\eta\eta}^\pm \\ + O(h^3), \end{aligned} \tag{16}$$

where the $+$ or $-$ superscript depends on whether (ξ_k, η_k) lies on the $+$ or $-$ side of Ω. Therefore the local truncation error T_{ij} can be expressed as a linear combination of the values $u^\pm, u_\xi^\pm, u_\eta^\pm, u_{\xi\xi}^\pm, u_{\xi\eta}^\pm, u_{\eta\eta}^\pm$

$$\begin{aligned} T_{ij} = a_1 u^- + a_2 u^+ + a_3 u_\xi^- + a_4 u_\xi^+ + a_5 u_\eta^- \\ + a_6 u_\eta^+ + a_7 u_{\xi\xi}^- + a_8 u_{\xi\xi}^+ + a_9 u_{\eta\eta}^- \\ + a_{10} u_{\eta\eta}^+ + a_{11} u_{\xi\eta}^- + a_{12} u_{\xi\eta}^+ \\ - \sigma u^- - f^- - C_{ij} + O(\max_k |\gamma_k| h^3), \end{aligned} \tag{17}$$

where $h = \max\{h_x, h_y\}$. We drive additional interface conditions [7,8,12] by taking the derivative of the jump conditions with respect to η at $\eta = 0$, and then we can express the quantities from one side in terms of the other side in the local coordinates (ξ, η) as

$$\begin{aligned} u^+ &= u^-, \quad u_\xi^+ = \rho\, u_\xi^- + \frac{v}{\beta^+}, \quad u_\eta^+ = u_\eta^-, \\ u_{\xi\xi}^+ &= -\chi'' u_\xi^- + \chi'' u_\xi^+ + (\rho - 1)\, u_{\eta\eta}^- + \rho u_{\xi\xi}^-, \\ u_{\eta\eta}^+ &= u_{\eta\eta}^- + (u_\xi^- - u_\xi^+)\, \chi'', \\ u_{\xi\eta}^+ &= \left(u_\eta^+ - \rho u_\eta^-\right) \chi'' + \rho\, u_{\xi\eta}^- + \frac{v'}{\beta^+}, \end{aligned} \tag{18}$$

where $\rho = \dfrac{\beta^-}{\beta^+}$. An alternative is to use a collocation method, That is, we equate the interface conditions

$$u^+(\xi_k, \eta_k) = u^-(\xi_k, \eta_k), \quad \beta^+ \frac{\partial u^+}{\partial \nu}(\xi_k, \eta_k) - \beta^- \frac{\partial u^-}{\partial \nu}(\xi_k, \eta_k) = v(\xi_k, \eta_k),$$

where (ξ_k, η_k) is the local coordinates of the three closest projection points to (x_i, y_j) along the equation at (x_i^*, y_j^*); $[\beta(u_{\xi\xi} + u_{\eta\eta})] = 0$. In this way one can avoid the tangential derivative the data v, especially useful for the three-dimensional case.

If we define the index sets $K^\pm = \{k : (\xi_k, \eta_k)$ is on the $\pm$ side of $\Gamma\}$, then a_{2j-1} terms are defined by

$$\begin{aligned} a_1 &= \sum_{k \in K^-} \gamma_k, & a_3 &= \sum_{k \in K^-} \xi_k \gamma_k, \\ a_5 &= \sum_{k \in K^-} \eta_k \gamma_k, & a_7 &= \frac{1}{2} \sum_{k \in K^-} \xi_k^2 \gamma_k, \\ a_9 &= \frac{1}{2} \sum_{k \in K^-} \eta_k^2 \gamma_k, & a_{11} &= \sum_{k \in K^-} \xi_k \eta_k \gamma_k. \end{aligned} \tag{19}$$

The a_{2j} terms have the same expressions as a_{2j-1} except the summation is taken over K^+. From (18) equating the terms in (13) for $(u^-, u^-_\xi, u^-_\eta, u^-_{\xi\xi}, u^-_{\eta\eta}, u^-_{\xi\eta})$, we obtain the linear system of equations for γ_k's:

$$\begin{aligned} a_1 + a_2 &= 0 \\ a_3 + \rho a_4 - a_8 \frac{[\beta]\chi''}{\beta^+} + a_{10} \frac{[\beta]\chi''}{\beta^+} &= 0 \\ a_5 + a_6 + a_{12}(1-\rho)\,\chi'' &= 0 \\ a_7 + a_8\,\rho &= \beta^- \\ a_9 + a_{10} + a_8\,(\rho - 1) &= \beta^- \\ a_{11} + a_{12}\,\rho &= 0. \end{aligned} \tag{20}$$

Once the γ_k's are obtained, we set $C_{ij} = a_{12}\,\dfrac{v'}{\beta^+} + \dfrac{1}{\beta^+}\left(a_4 + (a_8 - a_{10})\chi''\right) v$.

Remark 1

- If $[\beta] = 0$, then the finite difference scheme is the standard one. Only correction terms need to be added at irregular grid points. The correction terms can be regarded as second-order accurate discrete delta functions.
- If $v \equiv 0$, then the correction terms are zero.
- If we use a six-point stencil and (20) has a solution, then this leads to the original IIM [7].
- For more general cases, say both σ and f are discontinuous, we refer the reader to [7, 8, 12] for the derivation.

Enforcing the Maximum Principle Using an Optimization Approach

The stability of the finite difference equations is guaranteed by enforcing the sign constraint of the discrete maximum principle; see, for example, Morton and Mayers [21]. The sign restriction on the coefficients γ_k's in (12) are

$$\begin{aligned} \gamma_k &\geq 0 \quad \text{if} \quad (i_k, j_k) \neq (0,0), \\ \gamma_k &< 0 \quad \text{if} \quad (i_k, j_k) = (0,0). \end{aligned} \tag{21}$$

We form the following constrained quadratic optimization problem whose solution is the coefficients of the finite difference equation at the irregular grid point $\mathbf{x}_{ij}$:

$$\begin{aligned} &\min_\gamma \left\{ \frac{1}{2} \|, -g\|_2^2 \right\}, \quad \text{subject to} \quad A\gamma = b, \\ &\gamma_k \geq 0, \quad \text{if} \quad (i_k, j_k) \neq (0,0); \quad \gamma_k < 0, \text{if} \\ &(i_k, j_k) = (0,0), \end{aligned} \tag{22}$$

where $\gamma = [\,\gamma_1, \gamma_2, \cdots, \gamma_{n_s}\,]^T$ is the vector composed of the coefficients of the finite difference equation; $A\gamma = b$ is the system of linear equations (20); and $g \in R^{n_s}$ has the following components: $g \in R^{n_s}$,

$$\begin{aligned} g_k &= \frac{\beta_{i+i_k, j+j_k}}{h^2}, \quad \text{if} \quad (i_k, j_k) \in \{(-1,0), (1,0), \\ &\quad (0,-1), (0,1)\,\}; \\ g_k &= -\frac{4\beta_{i,j}}{h^2}, \quad \text{if} \quad (i_k, j_k) = (0,0); \\ g_k &= 0, \quad \text{otherwise.} \end{aligned} \tag{23}$$

With the maximum principle, the second-order convergence of the IIM has been proved in [11].

Augmented Immersed Interface Method

The original idea of the augmented strategy for interface problems was proposed in [9] for elliptic interface problems with a piecewise constant but discontinuous coefficient. With a few modifications, the augmented method developed in [9] was applied to generalized Helmholtz equations including Poisson equations on irregular domains in [14]. The augmented approach for the incompressible Stokes equations with a piecewise constant but discontinuous viscosity was proposed in [18], for slip boundary condition to deal with pressure boundary condition in [17], and for the Navier-Stokes equations on irregular domains in [6].

There are at least two motivations to use augmented strategies. The first one is to get a faster algorithm compared to a direct discretization, particularly to take advantages of existing fast solvers. The second reason is that, for some interface problems, an augmented approach may be the only way to derive an accurate algorithm. This is illustrated in the augmented immersed interface method [18] for the incompressible Stokes equations with discontinuous viscosity in which the jump conditions for the pressure and the velocity are coupled together. The augmented techniques enable

us to decouple the jump conditions so that the idea of the immersed interface method can be applied.

While augmented methods have some similarities to boundary integral methods or the integral equation approach to find a source strength, the augmented methods have a few special features: (1) no Green function is needed, and therefore there is no need to evaluate singular integrals; (2) there is no need to set up the system of equations for the augmented variable explicitly; (3) they are applicable to general PDEs with or without source terms; and (4) the method can be applied to general boundary conditions. On the other hand, we may need estimate the condition number of the Schur complement system and develop preconditioning techniques.

Procedure of the Augmented IIM

We explain the procedure of the augmented IIM using the fast Poisson solver on an interior domain as an illustration.

Assume we have linear partial differential equations with a linear interface or boundary condition. The the Poisson equation on an irregular domain Ω, as an example,

$$\Delta u = f(\mathbf{x}), \quad \mathbf{x} \in \Omega, \quad q(u, u_n) = 0, \quad \mathbf{x} \in \partial\Omega, \tag{24}$$

where $q(u, u_n) = 0$ is either a Dirichlet or Neumann boundary condition along the boundary $\partial\Omega$. To use an augmented approach, the domain Ω is embedded into a rectangle $\Omega \subset R$; the PDE and the source term are extended to the entire rectangle R:

$$\Delta u = \begin{cases} f, \text{if} & \mathbf{x} \in \Omega, \\ 0, \text{if} & \mathbf{x} \in R \setminus \Omega, \end{cases} \quad \begin{cases} [u] &= g, \text{ on } \partial\Omega, \\ [u_n] &= 0, \text{ on } \partial\Omega, \\ u &= 0, \text{ on } \partial R. \end{cases} \tag{25}$$

and

$$q(u, u_n) = 0 \quad \text{on } \partial\Omega.$$

The solution u to (25) is a functional $u(g)$ of g. We determine g such that the solution $u(g)$ satisfies the boundary condition $q(u, u_n) = 0$. Note that, given g, we can solve (25) using the immersed interface method with a single call to a fast Poisson solver.

On a Cartesian mesh (x_i, y_j), $i = 0, 1, \cdots, M$, $j = 0, 1, \cdots, N$, $M \sim N$, we use U and G to represent the discrete solution to (25). Note that the dimension of U is $O(N^2)$ while that of G is of $O(N)$. The augmented IIM can be written as

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} U \\ G \end{bmatrix} = \begin{bmatrix} F \\ Q \end{bmatrix}, \tag{26}$$

where A is the matrix formed from the discrete five-point Laplacian; BG are correction terms due to the jump in u, and the boundary condition is discretized by an interpolation scheme $CU + DG = Q$, corresponding to the boundary condition $q(u, u_n) = 0$. The main reason to use an augmented approach is to take advantage of fast Poisson solvers. Eliminating U from (26) gives a linear system for G, the *Schur complement system*,

$$(D - CA^{-1}B)G = Q - CA^{-1}F \stackrel{\text{def}}{=} F_2. \tag{27}$$

This is an $N_b \times N_b$ system for $\mathbf{G}$, a much smaller linear system compared to the one for U, where N_b is the dimension of G. If we can solve the Schur complement system efficiently, then we obtain the solution of the original problem with one call to the fast Poisson solver. There are two approaches to solve the Schur complement system. One is the GMRES iterative method; the other one is a direct method such as the LU decomposition. In either of the cases, we need to know how to find the matrix vector multiplication without forming the sub-matrices A^{-1}, B, C, D explicitly. That is, first we set $G = 0$ and solve the first equation of the (26), to get $U(0) = A^{-1}F$. For a given G the residual vector of the boundary condition is then given by

$$R(G) = C(U(0) - U(G)) + DG - Q.$$

Remark 2 For different applications, the augmented variable(s) can be chosen differently but the above procedure is the same. For some problems, if we need to use the same Schur complement at every time step, it is then more efficient to use the LU decomposition just once. If the Schur complement is varying or only used a few times, then the GMRES iterative method may be a better option. One may need to develop efficient preconditioners for the Schur complement.

Immersed Finite Element Method (IFEM)

The IIM has also been developed using finite element formulation as well, which is preferred sometimes because there is rich theoretical foundation based on Sobolev space, and finite element approach may lead to a better conditioned system of equations. Finite element methods have less regularity requirements for the coefficients, the source term, and the solution than finite difference methods do. In fact, the weak form for one-dimensional elliptic interface problem $(\beta u')' - \sigma u = f(x) + v\delta(x-\alpha)$, $0 < x < 1$ with homogeneous Dirichlet boundary condition is

$$\int_0^1 \left(\beta u'\phi' - \sigma uv\right) dx = -\int_0^1 f\phi dx + v\phi(\alpha), \quad \forall \phi \in H_0^1(0,1). \tag{28}$$

For two-dimensional elliptic interface problems (10), the weak form is

$$\iint_\Omega (\beta \nabla u \nabla \phi - \sigma uv)\, d\mathbf{x} = -\iint_\Omega f\,\phi\, d\mathbf{x} - \int_\Gamma v\phi ds, \quad \forall \phi(\mathbf{x}) \in H_0^1(\Omega). \tag{29}$$

Unless a body-fitted mesh is used, the solution obtained from the standard finite element method using the linear basis functions is only first-order accurate in the maximum norm. In [10], a new immersed finite element for the one-dimensional case is constructed using modified basis functions that satisfy homogeneous jump conditions. The modified basis functions satisfy

$$\phi_i(x_k) = \begin{cases} 1, \text{ if } k = i, \\ 0, \text{ otherwise} \end{cases} \quad \text{and} \quad [\phi_i] = 0, \quad [\beta\phi_i'] = 0. \tag{30}$$

Obviously, if $x_j < \alpha < x_{j+1}$, then only ϕ_j and ϕ_{j+1} need to be changed to satisfy the second jump condition. Using the method of undetermined coefficients, we can conclude that

$$\phi_j(x) = \begin{cases} 0, & 0 \le x < x_{j-1}, \\ \dfrac{x - x_{j-1}}{h}, & x_{j-1} \le x < x_j, \\ \dfrac{x_j - x}{D} + 1, & x_j \le x < \alpha, \\ \dfrac{\rho\,(x_{j+1} - x)}{D}, & \alpha \le x < x_{j+1}, \\ 0, & x_{j+1} \le x \le 1, \end{cases}$$

$$\phi_{j+1}(x) = \begin{cases} 0, & 0 \le x < x_j, \\ \dfrac{x - x_j}{D}, & x_j \le x < \alpha, \\ \dfrac{\rho\,(x - x_{j+1})}{D} + 1, & \alpha \le x < x_{j+1}, \\ \dfrac{x_{j+2} - x}{h}, & x_{j+1} \le x \le x_{j+2}, \\ 0, & x_{j+2} \le x \le 1. \end{cases}$$

where

$$\rho = \frac{\beta^-}{\beta^+}, \quad D = h - \frac{\beta^+ - \beta^-}{\beta^+}(x_{j+1} - \alpha).$$

Using the modified basis function, it has been shown in [10] that the Galerkin method is second-order accurate in the maximum norm. For 1D interface problems, the FD and FE methods discussed here are not very much different. The FE method likely perform better for self-adjoint problems, while the FD method is more flexible for general elliptic interface problems.

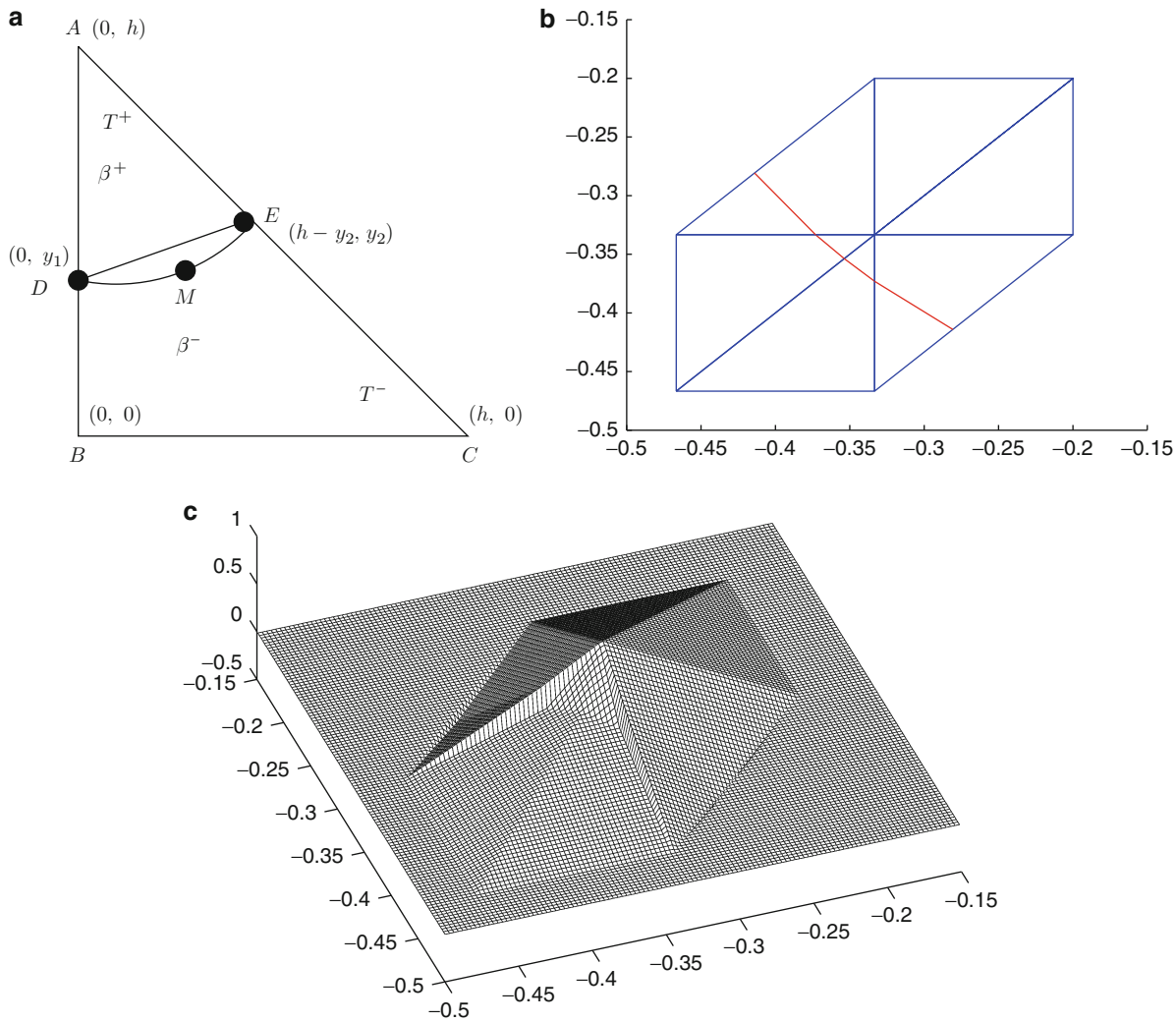
Modified Basis Functions for Two-Dimensional Problems

A similar idea above has been applied to two-dimensional problems with a uniform Cartesian triangulation [15]. The piecewise linear basis function centered at a node is defined as:

$$\phi_i(\mathbf{x}_j) = \begin{cases} 1, \text{ if } i = j \\ 0, \text{ otherwise,} \end{cases}$$

$$[u]|_\Gamma = 0, \quad \left[\beta \frac{\partial \phi_i}{\partial \mathbf{n}}\right]\Big|_\Gamma = 0, \quad \phi_i|_{\partial\Omega} = 0. \tag{31}$$

We call the space formed by all the basis function $\phi_i(\mathbf{x})$ as the immersed finite element space (IFE).



Immersed Interface/Boundary Method, Fig. 2 (**a**) A typical *triangle* element with an interface cutting through. The curve between D and E is part of the interface curve Γ which is approximated by the *line* segment $\overline{DE}$. In this diagram, T is the *triangle* $\triangle ABC$, $T^+ = \triangle ADE$, $T^- = T - T^+$, and T_r is the region enclosed by the $\overline{DE}$ and the arc DME. (**b**) A standard domain of six *triangles* with an interface cutting through. (**c**) A global basis function on its support in the *nonconforming* immersed finite element space. The basis function has small jump across some edges

We consider a reference interface element T whose geometric configuration is given in Fig. 2a in which the curve between points D and E is a part of the interface. We assume that the coordinates at A, B, C, D, and E are

$$(0, h), \quad (0, 0), \quad (h, 0), \quad (0, y_1), \quad (h - y_2, y_2), \tag{32}$$

with the restriction $0 \le y_1 \le h, \quad 0 \le y_2 < h$.

Once the values at vertices A, B, and C of the element T are specified, we construct the following piecewise linear function:

$$u(\mathbf{x}) = \begin{cases} u^+(\mathbf{x}) = a_0 + a_1 x + a_2 (y - h), \\ \text{if } \mathbf{x} = (x, y) \in T^+, \\ u^-(\mathbf{x}) = b_0 + b_1 x + b_2 y, \\ \text{if } \mathbf{x} = (x, y) \in T^-, \end{cases} \tag{33a}$$

$$u^+(D) = u^-(D), \quad u^+(E) = u^-(E),$$

$$\beta^+ \frac{\partial u^+}{\partial n} = \beta^- \frac{\partial u^-}{\partial n}, \tag{33b}$$

where $\mathbf{n}$ is the unit normal direction of the line segment $\overline{DE}$. This is a piecewise linear function in T that

satisfies the natural jump conditions along $\overline{DE}$. The existence and uniqueness of the basis functions and error estimates are given in [15].

It is easy to show that the linear basis function defined at a nodal point exists and it is unique. It has also been proved in [15] that for the solution of the interface problem (10), there is an interpolation function $u_I(\mathbf{x})$ in the IFE space that approximates $u(x)$ to second-order accuracy in the maximum norm.

However, as we can see from Fig. 2c, a linear basis function may be discontinuous along some edges. Therefore such IFE space is a nonconforming finite element space. Theoretically, it is easy to prove the corresponding Galerkin finite element method is at least first-order accurate; see [15]. In practice, its behaviors are much better than the standard finite element without any modifications. Numerically, the computed solution has super linear convergence. More theoretical analysis can be found in [2, 16].

The nonconforming immersed finite element space is also constructed for elasticity problems with interfaces in [4, 13, 29]. There are six coupled unknowns in one interface triangle for elasticity problems with interfaces.

A *conforming* IFE space is also proposed in [15]. The basis functions are still piecewise linear. The idea is to extend the support of the basis function along interface to one more triangle to keep the continuity. The conforming immersed finite element method is indeed second-order accurate. The trade-off is the increased complexity of the implementation. We refer the readers to [15] for the details. The conforming immersed finite element space is also constructed for elasticity problems with interfaces in [4].

Finally, one can construct the quadratic nonconforming element using the quadratic Taylor expansion (16) at the midpoint of the interface. The relation of coefficients of both sides is determined by the interface conditions (17). Then the quadratic element on the triangle is uniquely by the values of the basis six points of the triangle.

Hyperbolic Equations

We consider an advection equation as a model equation

$$u_t + (c(x)u)_x = 0, \quad t > 0, \ x \in R, \quad u(0,x) = u_0(x), \tag{34}$$

where $c = c(x) > 0$ is piecewise smooth. The second-order immersed interface method has been developed in [30]. We describe the higher-order method closely related to CIP methods [28]. CIP is one of the numerical methods that provides an accurate, less-dispersive and less-dissipative numerical solution. The method uses the exact integration in time by the characteristic method and uses the solution u and its derivative $v = u_x$ as unknowns. The piecewise cubic Hermite interpolation for each computational cell in each cell $[x_{j-1}, x_j]$ based on solution values and its derivatives at two endpoints x_{j-1}, x_j. In this way the method allows us to take an arbitrary time step (no CFL limitation) without losing the stability and accuracy. That is, we use the exact simultaneous update formula for the solution u:

$$u(x_k, t + \Delta t) = \frac{c(y_k)}{c(x_k)} u(y_k, t) \tag{35}$$

and for its derivative v:

$$v(x_k, t + \Delta t) = \left(\frac{c'(y_k)}{c(x_k)} - \frac{c'(x_k)}{c(x_k)} \right) \frac{c(y_k)}{c(x_k)} u(y_k, t) + \left(\frac{c(y_k)}{c(x_k)} \right)^2 v(y_k, t). \tag{36}$$

For the piecewise constant equation $u_t + c(x)\, u_x = 0$, we use the piecewise cubic interpolation: $F^-(x)$ in $[x_{j-1}, \alpha]$ and $F^+(x)$ in $[\alpha, x_j]$ of the form $F^{\pm}(x) = \sum_{k=0}^{3} a_k^{\pm} (x - \alpha)^k$. The eight unknowns are uniquely determined via the interface relations and the interpolation conditions at the interface $\alpha \in (x_{j-1}, x_j)$:

$$[u] = 0, \quad [cu_x] = 0, \quad [c^2 u_{xx}] = 0, \quad [c^3 u_{xxx}] = 0, \tag{37}$$

$$F^-(x_{j-1}) = u_{j-1}^n, \quad F_x^-(x_{j-1}) = v_{j-1}^n, \quad F^+(x_j) = u_j^n, \quad F_x^+(x_j) = v_j^n, \tag{38}$$

Thus, we update solution (u^n, v^n) at node x_j by

$$u_j^{n+1} = F^+(x_j - c^+ \Delta t), \quad v_j^{n+1} = F_x^+(x_j - c^+ \Delta t). \tag{39}$$

Similarly, for (34) we have the method based on the interface conditions $[cu] = [c^2 u_x] = [c^3 u_{xx}] = [c^4 u_{xxx}] = 0$ and the updates (35)–(36).

The d'Alembert-based method for the Maxwell equation that extends our characteristic-based method to Maxwell system is developed for the piecewise constant media and then applied to Maxwell system with piecewise constant coefficients. Also, one can extend the exact time integration CIP method for equations in discontinuous media in R^2 and R^3 and the Hamilton Jacobi equation [5].

References

1. Deng, S., Ito, K., Li, Z.: Three dimensional elliptic solvers for interface problems and applications. J. Comput. Phys. **184**, 215–243 (2003)
2. Ewing, R., Li, Z., Lin, T., Lin, Y.: The immersed finite volume element method for the elliptic interface problems. Math. Comput. Simul. **50**, 63–76 (1999)
3. Eyre, D., Fogelson, A.: IBIS: immersed boundary and interface software package. http://www.math.utah.edu/IBIS (1997)
4. Gong, Y.: Immersed-interface finite-element methods for elliptic and elasticity interface problems. North Carolina State University (2007)
5. Ito, K., Takeuchi, T.: Exact time integration CIP methods for scaler hyperbolic equations with variable and discontinuous coefficients. SIAM J. Numer. Anal. pp. 20 (2013)
6. Ito, K., Li, Z., Lai, M.-C.: An augmented method for the Navier-Stokes equations on irregular domains. J. Comput. Phys. **228**, 2616–2628 (2009)
7. LeVeque, R.J., Li, Z.: The immersed interface method for elliptic equations with discontinuous coefficients and singular sources. SIAM J. Numer. Anal. **31**, 1019–1044 (1994)
8. Li, Z.: The immersed interface method – a numerical approach for partial differential equations with interfaces. PhD thesis, University of Washington (1994)
9. Li, Z.: A fast iterative algorithm for elliptic interface problems. SIAM J. Numer. Anal. **35**, 230–254 (1998)
10. Li, Z.: The immersed interface method using a finite element formulation. Appl. Numer. Math. **27**, 253–267 (1998)
11. Li, Z., Ito, K.: Maximum principle preserving schemes for interface problems with discontinuous coefficients. SIAM J. Sci. Comput. **23**, 1225–1242 (2001)
12. Li, Z., Ito, K.: The Immersed Interface Method: Numerical Solutions of PDEs Involving Interfaces and Irregular Domains. SIAM Frontier Series in Applied mathematics, FR33. Society for Industrial and Applied Mathematics, Philadelphia (2006)
13. Li, Z., Yang, X.: An immersed finite element method for elasticity equations with interfaces. In: Shi, Z.-C., et al. (eds.) Proceedings of Symposia in Applied Mathematics. AMS Comput. Phys. Commun. **12**(2), 595–612 (2012)
14. Li, Z., Zhao, H., Gao, H.: A numerical study of electromigration voiding by evolving level set functions on a fixed cartesian grid. J. Comput. Phys. **152**, 281–304 (1999)
15. Li, Z., Lin, T., Wu, X.: New Cartesian grid methods for interface problem using finite element formulation. Numer. Math. **96**, 61–98 (2003)
16. Li, Z., Lin, T., Lin, Y., Rogers, R.C.: Error estimates of an immersed finite element method for interface problems. Numer. PDEs **12**, 338–367 (2004)
17. Li, Z., Wan, X., Ito, K., Lubkin, S.: An augmented pressure boundary condition for a Stokes flow with a non-slip boundary condition. Commun. Comput. Phys. **1**, 874–885 (2006)
18. Li, Z., Ito, K., Lai, M.-C.: An augmented approach for Stokes equations with a discontinuous viscosity and singular forces. Comput. Fluids **36**, 622–635 (2007)
19. Mittal, R., Iaccarino, G.: Immersed boundary methods. Annu. Rev. Fluid Mech. **37**, 239–261 (2005)
20. Mori, Y.: Convergence proof of the velocity field for a Stokes flow immersed boundary method. Commun. Pure Appl. Math. **61**, 1213–1263 (2008)
21. Morton, K.W., Mayers, D.F.: Numerical Solution of Partial Differential Equations. Cambridge University Press (1995)
22. Peskin, C.S.: Flow patterns around heart valves: a digital computer method for solving the equations of motion. PhD thesis, Physiology, Albert Einsten College of Medicine, University Microfilms 72–30 (1972)
23. Peskin, C.S.: Flow patterns around heart valves: a numerical method. J. Comput. Phys. **10**, 252–271 (1972)
24. Peskin, C.S.: Numerical analysis of blood flow in the heart. J. Comput. Phys. **25**, 220–252 (1977)
25. Peskin, C.S.: The immersed boundary method. Acta Numer. **11**, 479–517 (2002)
26. Roma, A.: A multi-level self adaptive version of the immersed boundary method. PhD thesis, New York University (1996)
27. Roma, A., Peskin, C.S., Berger, M.: An adaptive version of the immersed boundary method. J. Comput. Phys. **153**, 509–534 (1999)
28. Yabe, T., Aoki, T.: A universal solver for hyperbolic equations by cubic-polynomial interpolation. I. One-dimensional solver. Comput. Phys. Commun. **66**, 219–232 (1991)
29. Yang, X., Li, B., Li, Z.: The immersed interface method for elasticity problems with interface. Dyn. Contin. Discret. Impuls. Syst. **10**, 783–808 (2003)
30. Zhang, C., LeVeque, R.J.: The immersed interface method for acoustic wave equations with discontinuous coefficients. Wave Motion **25**, 237–263 (1997)

Index Concepts for Differential-Algebraic Equations

Volker Mehrmann
Institut für Mathematik, MA 4-5 TU, Berlin, Germany

Introduction

Differential-algebraic equations (DAEs) present today the state of the art in mathematical modeling of dynamical systems in almost all areas of science and engineering. Modeling is done in a modularized

way by combining standardized sub-models in a hierarchically built network. The topic is well studied from an analytical, numerical, and control theoretical point of view, and several monographs are available that cover different aspects of the subject [1, 2, 9, 14–16, 21, 28, 29, 34].

The mathematical model can usually be written in the form

$$F(t, x, \dot{x}) = 0, \tag{1}$$

where $\dot{x}$ denotes the (typically time) derivative of x. Denoting by $C^k(\mathbb{I}, \mathbb{R}^n)$ the set of k times continuously differentiable functions from $\mathbb{I} = [\underline{t}, \overline{t}] \subset \mathbb{R}$ to $\mathbb{R}^n$, one usually assumes that $F \in C^0(\mathbb{I} \times \mathbb{D}_x \times \mathbb{D}_{\dot{x}}, \mathbb{R}^m)$ is sufficiently smooth and that $\mathbb{D}_x, \mathbb{D}_{\dot{x}} \subseteq \mathbb{R}^n$ are open sets. The model equations are usually completed with initial conditions

$$x(\underline{t}) = \underline{x}. \tag{2}$$

Linear DAEs

$$E\dot{x} - Ax - f = 0, \tag{3}$$

with $E, A \in C^0(\mathbb{I}, \mathbb{R}^{m,n})$, $f \in C^0(\mathbb{I}, \mathbb{R}^m)$ often arise after linearization along trajectories (see [4]) with constant coefficients in the case of linearization around an equilibrium solution. DAE models are also studied in the case when x is infinite dimensional (see, e.g., [7, 37]), but here we only discuss the finite-dimensional case.

Studying the literature for DAEs, one quickly realizes an almost Babylonian confusion in the notation, in the solution concepts, in the numerical simulation techniques, and in control and optimization methods. These differences partially result from the fact that the subject was developed by different groups in mathematics, computer science, and engineering. Another reason is that it is almost impossible to treat automatically generated DAE models directly with standard numerical methods, since the solution of a DAE may depend on derivatives of the model equations or input functions and since the algebraic equations restrict the dynamics of the system to certain manifolds, some of which are only implicitly contained in the model. This has the effect that numerical methods may have a loss in convergence order, are hard to initialize, or fail to preserve the underlying constraints and thus yield physically meaningless results (see, e.g., [2, 21] for illustrative examples). Furthermore, inconsistent initial conditions or violated smoothness requirements can give rise to distributional or other classes of solutions [8, 21, 27, 35] as well as multiple solutions [21]. Here we only discuss *classical solutions*, $x \in C^1(\mathbb{I}, \mathbb{C}^n)$ that satisfy (1) pointwise.

Different approaches of classifying the difficulties that arise in DAEs have led to different so-called *index* concepts, where the index is a "measure of difficulty" in the analytical or numerical treatment of the DAE. In this contribution, the major index concepts will be surveyed and put in perspective with each other as far as this is possible. For a detailed analysis and a comparison of various index concepts with the differentiation index (see [5, 12, 14, 21, 22, 24, 31]). Since most index concepts are only defined for uniquely solvable square systems with $m = n$, here only this case is studied (see [21] for the general case).

Index Concepts for DAEs

The starting point for all index concepts is the linear systems with constant coefficients. In this case, the smoothness requirements can be determined from the Kronecker canonical form [11] of the matrix pair (E, A) under equivalence transformations $E_2 = PE_1Q$, $A_2 = PA_1Q$, with invertible matrices P, Q (see e.g., [21]). The size of the largest Kronecker block associated with an infinite eigenvalue of (E, A) is called *Kronecker index*, and it defines the smoothness requirements for the inhomogeneity f. For the linear variable coefficient case, it was first tried to define a Kronecker index (see [13]). However, it was soon realized that this is not a reasonable concept [5, 17], since for the variable coefficient case, the equivalence transformation is $E_2 = PE_1Q$, $A_2 = PA_1Q - PE_1\dot{Q}$, and it locally does not reduce to the classical equivalence for matrix pencils. Canonical forms under this equivalence transformation have been derived in [17] and existence and uniqueness of solutions of DAEs has been characterized via global equivalence transformations and differentiations.

Since the differentiation of computed quantities is usually difficult, it was suggested in [3] to differentiate first the original DAE (3) and then carry out equivalence transformations. For this, we gather the original equation and its derivatives up to order ℓ into a so-called derivative array:

$$F_\ell(t,x,\dots,x^{(\ell+1)}) = \begin{bmatrix} F(t,x,\dot{x}) \\ \frac{d}{dt}F(t,x,\dot{x}) \\ \vdots \\ (\frac{d}{dt})^\ell F(t,x,\dot{x}) \end{bmatrix}. \quad (4)$$

We require solvability of (4) in an open set and define

$$\begin{aligned} M_\ell(t,x,\dot{x},\dots,x^{(\ell+1)}) &= F_{\ell;\dot{x},\dots,x^{(\ell+1)}}(t,x,\dot{x},\dots,x^{(\ell+1)}), \\ N_\ell(t,x,\dot{x},\dots,x^{(\ell+1)}) &= -(F_{\ell;x}(t,x,\dot{x},\dots,x^{(\ell+1)}), \\ &\quad 0,\dots,0), \\ g_\ell(t) &= F_{\ell;t}, \end{aligned}$$

where $F_{\ell;z}$ denotes the Jacobian of F_ℓ with respect to the variables in z.

The Differentiation Index

The most common index definition is that of the *differentiation index* (see [5]).

Definition 1 Suppose that (1) is solvable. The smallest integer ν (if it exists) such that the solution x is uniquely defined by $F_\nu(t,x,\dot{x},\dots,x^{(\nu+1)}) = 0$ for all consistent initial values is called the *differentiation index* of (1).

Over the years, the definition of the differentiation index has been slightly modified to adjust from the linear to the nonlinear case [3,5,6] and to deal with slightly different smoothness assumptions. In the linear case, it has been shown in [21] that the differentiation index ν is invariant under (global) equivalence transformations, and if it is well defined, then there exists a smooth, pointwise nonsingular $R \in C(\mathbb{I}, \mathbb{C}^{(\nu+1)n,(\nu+1)n})$ such that $RM_\nu = \begin{bmatrix} I_n & 0 \\ 0 & H \end{bmatrix}$. Then from the derivative array $M_\nu(t)\dot{z} = N_\nu(t)z + g_\nu(t)$, one obtains an ordinary differential equation (ODE):

$$\dot{x} = [\, I_n \;\; 0\,]R(t)M_\nu(t)\dot{z} = [\, I_n \;\; 0\,]R(t)N_\nu(t)\begin{bmatrix} I_n \\ 0 \end{bmatrix} x + [\, I_n \;\; 0\,]R(t)g_\nu(t),$$

which is called *underlying ODE*. Any solution of the DAE is also a solution of this ODE. This motivates the interpretation that the differentiation index is the number of differentiations needed to transform the DAE into an ODE.

The Strangeness Index

An index concept that is closely related to the differentiation index and extends to over- and under determined systems is based on the following hypothesis.

Hypothesis 1 *Consider the DAE (1) and suppose that there exist integers μ, a, and d such that the set $\mathbb{L}_\mu = \{z \in \mathbb{R}^{(\mu+2)n+1} \mid F_\mu(z) = 0\}$ associated with F is nonempty and such that for every point $z_0 = (t_0, x_0, \dot{x}_0, \dots, x_0^{(\mu+1)}) \in \mathbb{L}_\mu$, there exists a (sufficiently small) neighborhood in which the following properties hold:*

1. *We have* rank $M_\mu(z) = (\mu+1)n - a$ *on* $\mathbb{L}_\mu$ *such that there exists a smooth matrix function Z_2 of size $(\mu+1)n \times a$ and pointwise maximal rank, satisfying $Z_2^T M_\mu = 0$ on $\mathbb{L}_\mu$.*
2. *We have* rank $\hat{A}_2(z) = a$, *where $\hat{A}_2 = Z_2^T N_\mu [I_n \; 0 \; \cdots \; 0]^T$ such that there exists a smooth matrix function T_2 of size $n \times d$, $d = n - a$, and pointwise maximal rank, satisfying $\hat{A}_2 T_2 = 0$.*
3. *We have* rank $F_{\dot{x}}(t,x,\dot{x})T_2(z) = d$ *such that there exists a smooth matrix function Z_1 of size $n \times d$ and pointwise maximal rank, satisfying* rank $\hat{E}_1 T_2 = d$, *where $\hat{E}_1 = Z_1^T F_{\dot{x}}$.*

Definition 2 Given a DAE as in (1), the smallest value of μ such that F satisfies Hypothesis 1 is called the *strangeness index* of (1).

It has been shown in [21] that if F as in (1) satisfies Hypothesis 1 with characteristic values μ, a, and d, then the set $\mathbb{L}_\mu \subseteq \mathbb{R}^{(\mu+2)n+1}$ forms a (smooth) manifold of dimension $n+1$. Setting

$$\begin{aligned} \hat{F}_1(t,x,\dot{x}) &= Z_1^T F(t,x,\dot{x}), \\ \hat{F}_2(t,x) &= Z_2^T F_\mu(t,x,\hat{z}), \end{aligned}$$

where $\hat{z} = (x^{(1)},\dots,x^{(\mu+1)})$, and considering the *reduced DAE*

$$\hat{F}(t,x,\dot{x}) = \begin{bmatrix} \hat{F}_1(t,x,\dot{x}) \\ \hat{F}_2(t,x) \end{bmatrix} = 0, \quad (5)$$

one has the following (local) relation between the solutions of (1) and (5).

Theorem 1 ([19,21]) *Let F as in (1) satisfy Hypothesis 1 with values μ, a, and d. Then every sufficiently smooth solution of (1) also solves (5).*

It also has been shown in [21] that if $x^* \in C^1(\mathbb{I}, \mathbb{R}^n)$ is a sufficiently smooth solution of (1), then there exist an operator $\hat{\mathcal{F}}: \mathbb{D} \to \mathbb{Y}$, $\mathbb{D} \subseteq \mathbb{X}$ open, given by

$$\hat{\mathcal{F}}(x)(t) = \begin{bmatrix} \dot{x}_1(t) - \mathcal{L}(t, x_1(t)) \\ x_2(t) - \mathcal{R}(t, x_1(t)) \end{bmatrix}, \tag{6}$$

with $\mathbb{X} = \{x \in C(\mathbb{I}, \mathbb{R}^n) \mid x_1 \in C^1(\mathbb{I}, \mathbb{R}^d), x_1(\underline{t}) = 0\}$ and $\mathbb{Y} = C(\mathbb{I}, \mathbb{R}^n)$. Then x^* is a *regular solution* of (6), i.e., there exist neighborhoods $\mathbb{U} \subseteq \mathbb{X}$ of x^*, and $\mathbb{V} \subseteq \mathbb{Y}$ of the origin such that for every $b \in \mathbb{V}$, the equation $\hat{\mathcal{F}}(x) = b$ has a unique solution $x \in \mathbb{U}$ that depends continuously on f.

The requirements of Hypothesis 1 and that of a well-defined differentiation index are equivalent up to some (technical) smoothness requirements (see [18,21]). For uniquely solvable systems, however, the differentiation index aims at a reformulation of the given problem as an ODE, whereas Hypothesis 1 aims at a reformulation as a DAE with two parts, one part which states all constraints and another part which describes the dynamical behavior. If the appropriate smoothness conditions hold, then $\nu = 0$ if $\mu = a = 0$ and $\nu = \mu + 1$ otherwise.

The Perturbation Index

Motivated by the desire to classify the difficulties arising in the numerical solution of DAEs, the *perturbation index* introduced in [16] studies the effect of a perturbation η in

$$F(t, \hat{x}, \dot{\hat{x}}) = \eta, \tag{7}$$

with sufficiently smooth η and initial condition $\hat{x}(\underline{t}) = \underline{\hat{x}}$.

Definition 3 If $x \in C^1(\mathbb{I}, \mathbb{C}^n)$ is a solution, then (1) is said to have *perturbation index* $\kappa \in \mathbb{N}$ along x, if κ is the smallest number such that for all sufficiently smooth $\hat{x}$ satisfying (7) the estimate (with appropriate norms in the relevant spaces)

$$\|\hat{x} - x\| \le C(\|\underline{\hat{x}} - \underline{x}\| + \|\eta\|_\infty + \|\dot{\eta}\|_\infty + \cdots + \|\eta^{(\kappa-1)}\|_\infty) \tag{8}$$

holds with a constant C independent of $\hat{x}$, provided that the expression on the right-hand side in (8) is sufficiently small. It is said to have *perturbation index* $\kappa = 0$ if the estimate

$$\|\hat{x} - x\| \le C(\|\underline{\hat{x}} - \underline{x}\| + \max_{t \in \mathbb{I}} \| \textstyle\int_{\underline{t}}^{t} \eta(s)\, ds \|_\infty) \tag{9}$$

holds.

For the linear variable coefficient case, the following relation holds.

Theorem 2 ([21]) *Let the strangeness index μ of (3) be well defined and let x be a solution of (3). Then the perturbation index κ of (3) along x is well defined with $\kappa = 0$ if $\mu = a = 0$ and $\kappa = \mu + 1$ otherwise.*

The reason for the two cases in the definition of the perturbation index is that in this way, the perturbation index equals the differentiation index if defined. Counting in the way of the strangeness index according to the estimate (8), there would be no need in the extension (9).

It has been shown in [21] that the concept of the perturbation index can also be extended to the non-square case.

The Tractability Index

A different index concept [14, 23, 24] is formulated in its current form for DAEs *with properly stated leading term*:

$$F \frac{\mathrm{d}}{\mathrm{d}t}(Dx) = f(x, t), \ t \in \mathbb{I} \tag{10}$$

with $F \in C(\mathbb{I}, \mathbb{R}^{n,l})$, $D \in C(\mathbb{I}, \mathbb{R}^{l,n})$, $f \in C(\mathbb{I} \times \mathbb{D}_x, \mathbb{R}^n)$, sufficiently smooth such that kernel $F(t) \oplus$ range $D(t) = \mathbb{R}^l$ for all $t \in \mathbb{I}$ and such that there exists a projector $R \in C^1(\mathbb{I}, \mathbb{R}^{l,l})$ with range $R(t) =$ range $D(t)$ and kernel $R(t) =$ kernel $F(t)$ for all $t \in \mathbb{I}$. One introduces the chain of matrix functions:

$$\begin{aligned} \mathcal{G}_0 &= FD, \ \mathcal{G}_1 = \mathcal{G}_0 + \mathcal{B}_0 \mathcal{Q}_0, \ \mathcal{G}_{i+1} \\ &= \mathcal{G}_i + \mathcal{B}_i \mathcal{Q}_i, \ i = 1, 2, \ldots, \end{aligned} \tag{11}$$

where $\mathcal{Q}_i$ is a projector onto $\mathcal{N}_i = \text{kernel}\, \mathcal{G}_i$, with $\mathcal{Q}_i \mathcal{Q}_j = 0$ for $j = 0, \ldots, i-1$, $\mathcal{P}_i = I - \mathcal{Q}_i$, $\mathcal{B}_0 = f_x$, and $\mathcal{B}_i = \mathcal{B}_{i-1}\mathcal{P}_{i-1} - \mathcal{G}_i D^- \frac{\mathrm{d}}{\mathrm{d}t}(D\mathcal{P}_1 \ldots \mathcal{P}_i D^-) D\mathcal{P}_{i-1}$, where D^- is the reflexive generalized inverse of D satisfying $(DD^-) = R$ and $(D^-D) = \mathcal{P}_0$.

Definition 4 ([23]) A DAE of the form (10) with properly stated leading term is said to be *regular with tractability index τ on the interval* $\mathbb{I}$, if there exist a sequence of continuous matrix functions (11) such that

1. $\mathcal{G}_i$ is singular and has constant rank $\bar{r}_i$ on $\mathbb{I}$ for $i = 0, \ldots, \tau - 1$.
2. $\mathcal{Q}_i$ is continuous and $D\mathcal{P}_1 \ldots \mathcal{P}_i D^-$ is continuously differentiable on $\mathbb{I}$ for $i = 0, \ldots, \tau - 1$.
3. $\mathcal{Q}_i \mathcal{Q}_j = 0$ holds on $\mathbb{I}$ for all $i = 1, \ldots, \tau - 1$ and $j = 1, \ldots, i - 1$.
4. $\mathcal{G}_\mu$ is nonsingular on $\mathbb{I}$.

The chain of projectors and spaces allows to filter out an ODE for the differential part of the solution $u = D\mathcal{P}_1 \ldots \mathcal{P}_{\tau-1} D^- Dx$ of the linear version of (10) with $f(x,t) = A(t)x(t) + q(t)$ (see [23]) which is given by

$$\dot{u} - \frac{d}{dt}(D\mathcal{P}_1 \ldots \mathcal{P}_{\tau-1} D^-)u - D\mathcal{P}_1 \ldots \mathcal{P}_{\tau-1} \mathcal{G}_\mu^{-1} A D^-$$
$$u = D\mathcal{P}_1 \ldots \mathcal{P}_{\tau-1} \mathcal{G}_\mu^{-1} q.$$

Instead of using derivative arrays here, derivatives of projectors are used. The advantage is that the smoothness requirements for the inhomogeneity can be explicitly specified and in this form the tractability index can be extended to infinite-dimensional systems. However, if the projectors have to be computed numerically, then difficulties in obtaining the derivatives can be anticipated.

It is still a partially open problem to characterize the exact relationship between the tractability index and the other indices. Partial results have been obtained in [5, 6, 22, 24], showing that (except again for different smoothness requirements) the tractability index is equal to the differentiation index and thus by setting $\tau = 0$ if $\mu = a = 0$ one has $\tau = \mu + 1$ if $\tau > 0$.

The Geometric Index

The geometric theory to study DAEs as differential equations on manifolds was developed first in [30, 32, 33]. One constructs a sequence of sub-manifolds and their parameterizations via local charts (corresponding to the different constraints on different levels of differentiation). The largest number of differentiations needed to identify the DAE as a differential equation on a manifold is then called the *geometric index* of the DAE. It has been shown in [21] that any solvable regular DAE with strangeness index $\mu = 0$ can be locally (near a given solution) rewritten as a differential equation on a manifold and vice versa. If one considers the reduced system (5), then starting with a solution $x^* \in C^1(\mathbb{I}, \mathbb{R}^n)$ of (1), the set $\mathbb{M} = \hat{F}_2^{-1}(\{0\})$ is nonempty and forms the desired sub-manifold of dimension d of $\mathbb{R}^n$, where the differential equation evolves and contains the consistent initial values. The ODE case trivially is a differential equation on the manifold $\mathbb{R}^n$. Except for differences in the smoothness requirements, the geometric index is equal to the differentiation index [5]. This then also defines the relationship to the other indices.

The Structural Index

A combinatorially oriented index was first defined for the linear constant coefficient case. Let $(E(p), A(p))$ be the parameter dependent pencil that is obtained from (E, A) by substituting the nonzero elements of E and A by independent parameters p_j. Then the unique integer that equals the Kronecker index of $(E(p), A(p))$ for all p from some open and dense subset of the parameter set is called the *structural index* (see [25] and in a more general way [26]). For the nonlinear case, a local linearization is employed.

Although it has been shown in [31] that the differentiation index and the structural index can be arbitrarily different, the algorithm of [25] to determine the structural index is used heavily in applications (see, e.g., [38]) by employing combinatorial information to analyze which equations should be differentiated and to introduce extra variables for index reduction [36]. A sound analysis when this approach is fully justified has, however, only been given in special cases [10, 20, 36].

Conclusions

Different index concepts for systems of differential-algebraic equations have been discussed. Except for different technical smoothness assumptions (and in the case of the strangeness index, different counting) for regular and uniquely solvable systems, these concepts are essentially equivalent to the differentiation index. However, all have advantages and disadvantages when it comes to generalizations, numerical methods, or control techniques. The strangeness index and the perturbation index also extend to non-square systems, while the tractability index allows a direct generalization to infinite-dimensional systems.

References

1. Ascher, U.M., Petzold, L.R.: Computer Methods for Ordinary Differential and Differential-Algebraic Equations. SIAM Publications, Philadelphia (1998)
2. Brenan, K.E., Campbell, S.L., Petzold, L.R.: Numerical Solution of Initial-Value Problems in Differential Algebraic Equations, 2nd edn. SIAM Publications, Philadelphia (1996)
3. Campbell, S.L.: A general form for solvable linear time varying singular systems of differential equations. SIAM J. Math. Anal. **18**, 1101–1115 (1987)
4. Campbell, S.L.: Linearization of DAE's along trajectories. Z. Angew. Math. Phys. **46**, 70–84 (1995)
5. Campbell, S.L., Gear, C.W.: The index of general nonlinear DAEs. Numer. Math. **72**, 173–196 (1995)
6. Campbell, S.L., Griepentrog, E.: Solvability of general differential algebraic equations. SIAM J. Sci. Comput. **16**, 257–270 (1995)
7. Campbell, S.L., Marszalek, W.: Index of infinite dimensional differential algebraic equations. Math. Comput. Model. Dyn. Syst. **5**, 18–42 (1999)
8. Cobb, J.D.: On the solutions of linear differential equations with singular coefficients. J. Differ. Equ. **46**, 310–323 (1982)
9. Eich-Soellner, E., Führer, C.: Numerical Methods in Multibody Systems. Teubner Verlag, Stuttgart (1998)
10. Estévez-Schwarz, D., Tischendorf, C.: Structural analysis for electrical circuits and consequences for MNA. Int. J. Circuit Theor. Appl. **28**, 131–162 (2000)
11. Gantmacher, F.R.: The Theory of Matrices I. Chelsea Publishing Company, New York (1959)
12. Gear, C.W.: Differential-algebraic equation index transformations. SIAM J. Sci. Stat. Comput. **9**, 39–47 (1988)
13. Gear, C.W., Petzold, L.R.: Differential/algebraic systems and matrix pencils. In: Kågström, B., Ruhe, A. (eds.) Matrix Pencils, pp. 75–89. Springer, Berlin (1983)
14. Griepentrog, E., März, R.: Differential-Algebraic Equations and Their Numerical Treatment. Teubner Verlag, Leipzig (1986)
15. Hairer, E., Wanner, G.: Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems, 2nd edn. Springer, Berlin (1996)
16. Hairer, E., Lubich, C., Roche, M.: The Numerical Solution of Differential-Algebraic Systems by Runge–Kutta Methods. Springer, Berlin (1989)
17. Kunkel, P., Mehrmann, V.: Canonical forms for linear differential-algebraic equations with variable coefficients. J. Comput. Appl. Math. **56**, 225–259 (1994)
18. Kunkel, P., Mehrmann, V.: Local and global invariants of linear differential-algebraic equations and their relation. Electron. Trans. Numer. Anal. **4**, 138–157 (1996)
19. Kunkel, P., Mehrmann, V.: A new class of discretization methods for the solution of linear differential algebraic equations with variable coefficients. SIAM J. Numer. Anal. **33**, 1941–1961 (1996)
20. Kunkel, P., Mehrmann, V.: Index reduction for differential-algebraic equations by minimal extension. Z. Angew. Math. Mech. **84**, 579–597 (2004)
21. Kunkel, P., Mehrmann, V.: Differential-Algebraic Equations: Analysis and Numerical Solution. EMS Publishing House, Zürich (2006)
22. Lamour, R.: A projector based representation of the strangeness index concept. Preprint 07-03, Humboldt Universität zu Berlin, Berlin, Germany, (2007)
23. März, R.: The index of linear differential algebraic equations with properly stated leading terms. Results Math. **42**, 308–338 (2002)
24. März, R.: Characterizing differential algebraic equations without the use of derivative arrays. Comput. Math. Appl. **50**, 1141–1156 (2005)
25. Pantelides, C.C.: The consistent initialization of differential-algebraic systems. SIAM J. Sci. Stat. Comput. **9**, 213–231 (1988)
26. Pryce, J.: A simple structural analysis method for DAEs. BIT **41**, 364–394 (2001)
27. Rabier, P.J., Rheinboldt, W.C.: Classical and generalized solutions of time-dependent linear differential-algebraic equations. Linear Algebra Appl. **245**, 259–293 (1996)
28. Rabier, P.J., Rheinboldt, W.C.: Nonholonomic motion of rigid mechanical systems from a DAE viewpoint. SIAM Publications, Philadelphia (2000)
29. Rabier, P.J., Rheinboldt, W.C.: Theoretical and Numerical Analysis of Differential-Algebraic Equations. Handbook of Numerical Analysis, vol. VIII. Elsevier Publications, Amsterdam (2002)
30. Reich, S.: On a geometric interpretation of differential-algebraic equations. Circuits Syst. Signal Process. **9**, 367–382 (1990)
31. Reißig, G., Martinson, W.S., Barton, P.I.: Differential-algebraic equations of index 1 may have an arbitrarily high structural index. SIAM J. Sci. Comput. **21**, 1987–1990 (2000)
32. Rheinboldt, W.C.: Differential-algebraic systems as differential equations on manifolds. Math. Comput. **43**, 473–482 (1984)
33. Rheinboldt, W.C.: On the existence and uniqueness of solutions of nonlinear semi-implicit differential algebraic equations. Nonlinear Anal. **16**, 647–661 (1991)
34. Riaza, R.: Differential-Algebraic Systems: Analytical Aspects and Circuit Applications. World Scientific Publishing Co. Pte. Ltd., Hackensack (2008)
35. Seiler, W.M.: Involution-the formal theory of differential equations and its applications in computer algebra and numerical analysis. Habilitation thesis, Fak. f. Mathematik, University of Mannheim, Mannheim, Germany (2002)
36. Söderlind, G.: Remarks on the stability of high-index DAE's with respect to parametric perturbations. Computing **49**, 303–314 (1992)
37. Tischendorf, C.: Coupled systems of differential algebraic and partial differential equations in circuit and device simulation. Habilitation thesis, Inst. für Math., Humboldt-Universität zu Berlin, Berlin, Germany (2004)
38. Unger, J., Kröner, A., Marquardt, W.: Structural analysis of differential-algebraic equation systems: theory and applications. Comput. Chem. Eng. **19**, 867–882 (1995)

Information Theory for Climate Change and Prediction

Michal Branicki
School of Mathematics, The University of Edinburgh, Edinburgh, UK

Keywords

Climate change; Information theory; Kulback-Leibler divergence; Relative entropy; Reduced-order predictions

Mathematics Subject Classification

Primary: 94A15, 60H30, 35Q86, 35Q93; Secondary: 35Q94, 62B10, 35Q84, 60H15

Description

The Earth's climate is an extremely complex system coupling physical processes for the atmosphere, ocean, and land over a wide range of spatial and temporal scales (e.g., [5]). In contrast to predicting the small-scale, short-term behavior of the atmosphere (i.e., the "weather"), climate change science aims to predict the planetary-scale, long-time response in the "climate system" induced either by changes in external forcing or by internal variability such as the impact of increased greenhouse gases or massive volcanic eruptions [14]. Climate change predictions pose a formidable challenge for a number of intertwined reasons. First, while the dynamical equations for the actual climate system are unknown, one might reasonably assume that the dynamics are nonlinear and turbulent with, at best, intermittent energy fluxes from small scales to much larger and longer spatiotemporal scales. Moreover, all that is available from the true climate dynamics are coarse, empirical estimates of low-order statistics (e.g., mean and variance) of the large-scale horizontal winds, temperature, concentration of greenhouse gases, etc., obtained from sparse observations. Thus, a fundamental difficulty in estimating sensitivity of the climate system to perturbations lies in predicting the coarse-grained response of an extremely complex system from sparse observations of its past and present dynamics combined with a suite of imperfect, reduced-order models.

For several decades, the weather forecasts and the climate change predictions have been carried out through comprehensive numerical models [5, 14]. However, such models contain various errors which are introduced through lack of resolution and a myriad of parameterizations which aim to compensate for the effects of the unresolved dynamical features such as clouds, ocean eddies, sea ice cover, etc. Due to the highly nonlinear, multi-scale nature of this extremely high-dimensional problem, it is quite clear that – despite the ever increasing computer power – no model of the climate system will be able to resolve all the dynamically important and interacting scales.

Recently, a stochastic-statistical framework rooted in information theory was developed in [1, 10–12] for a systematic mitigation of error in reduced-order models and improvement of imperfect coarse-grained predictions. This newly emerging approach blends physics-constrained dynamical modeling, stochastic parameterization, and linear response theory, and it has at least two mathematically desirable features: (i) The approach is based on a skill measure given by the relative entropy which, unlike other metrics for uncertainty quantification in atmospheric sciences, is invariant under the general change of variables [9, 13]; this property is very important for unbiased model calibration especially in high-dimensional problems. (ii) Minimizing the loss of information in the imperfect predictions via the relative entropy implies simultaneous tuning of all considered statistical moments; this is particularly important for improving predictions of nonlinear, non-Gaussian dynamics where the statistical moments are interdependent.

Improving Imperfect Predictions

Assume that the reduced-order model(s) used to approximate the truth resolve the dynamics within a finite-dimensional domain, Ω, $\dim(\Omega) < \infty$, of the full phase space. The variables, $\boldsymbol{u} \in \Omega$, resolved by the model can represent, for example, the first N Fourier modes of the velocity and temperature fields. We are interested in improving imperfect probabilistic predictions of the true dynamics on the resolved variables $\boldsymbol{u} \in \Omega$ given the the time-dependent probability den-

sity, $\pi_t^{\mathrm{M}}(\boldsymbol{u})$, of the model for $t \in \mathcal{I}$ which approximates the marginal probability density, $\pi_t(\boldsymbol{u})$, of the truth.

The lack of information in the probability density π relative to the density π^{M} can be measured through the relative entropy, $\mathcal{P}(\pi, \pi^{\mathrm{M}})$, given by [8,9]

$$\mathcal{P}(\pi, \pi^{\mathrm{M}}) = \int_{\Omega} \pi \ln \frac{\pi}{\pi^{\mathrm{M}}}, \tag{1}$$

where we skipped the explicit dependence on time and space in the probability densities. The relative entropy $\mathcal{P}(\pi, \pi^{\mathrm{M}})$ originates from Shannon's information theory (e.g., [3]), and it provides a useful measure of *model error* in imperfect probabilistic predictions (e.g., [10]) due to its two metric-like properties: (i) $\mathcal{P}(\pi, \pi^{\mathrm{M}})$ is nonnegative and zero only when $\pi = \pi^{\mathrm{M}}$, and (ii) $\mathcal{P}(\pi, \pi^{\mathrm{M}})$ is invariant under any invertible change of variables which follows from the independence of $\mathcal{P}$ of the dominating measure in π and π^{M}. These properties can be easily understood in the Gaussian framework when $\pi^{\mathrm{G}} = \mathcal{N}(\bar{\boldsymbol{u}}, R)$ and $\pi^{\mathrm{M,G}} = \mathcal{N}(\bar{\boldsymbol{u}}^{\mathrm{M}}, R^{\mathrm{M}})$, and the relative entropy is simply expressed by

$$\mathcal{P}(\pi^{\mathrm{G}}, \pi^{\mathrm{M,G}}) = \left[\tfrac{1}{2}(\bar{\boldsymbol{u}} - \bar{\boldsymbol{u}}^{\mathrm{M}}) R_{\mathrm{M}}^{-1} (\bar{\boldsymbol{u}} - \bar{\boldsymbol{u}}^{\mathrm{M}})\right] + \tfrac{1}{2}\left[\mathrm{tr}[R R_{\mathrm{M}}^{-1}] - \ln \det[R\, R_{\mathrm{M}}^{-1}] - \dim[\bar{\boldsymbol{u}}]\right], \tag{2}$$

which also highlights the fact that minimizing $\mathcal{P}$ requires simultaneous tuning of both the model mean and covariance.

Given a class $\mathcal{M}$ of reduced-order models for the resolved dynamics on $\boldsymbol{u} \in \Omega$, the best model $\mathrm{M}_{\mathcal{I}}^{*} \in \mathcal{M}$ for making predictions over the time interval, $\mathcal{I} \equiv [t \;\; t+T]$, is given by

$$\mathscr{P}_{\mathcal{I}}(\pi, \pi^{\mathrm{M}_{\mathcal{I}}^{*}}) = \min_{\mathrm{M} \in \mathcal{M}} \mathscr{P}_{\mathcal{I}}(\pi, \pi^{\mathrm{M}}),$$

$$\mathscr{P}_{\mathcal{I}}(\pi, \pi^{\mathrm{M}}) \equiv \frac{1}{T} \int_{t}^{t+T} \mathcal{P}(\pi_s, \pi_s^{\mathrm{M}}) \mathrm{d}s, \tag{3}$$

where $\mathscr{P}_{\mathcal{I}}(\pi, \pi^{\mathrm{M}})$ measures the total lack of information in π^{M} relative to the truth density π within $\mathcal{I}$; note that for $T \to 0$ the best model, $\mathrm{M}_{\mathcal{I}}^{*} \in \mathcal{M}$, is simply the one minimizing the relative entropy (1) at time t. The utility of the relative entropy for quantifying the model error extends beyond the formal definition in (3) with the unknown truth density, π, and it stems from the fact that (1) can be written as [13]

$$\mathcal{P}(\pi, \pi^{\mathrm{M,L}}) = \mathcal{P}(\pi, \pi^{\mathrm{L}}) + \mathcal{P}(\pi^{\mathrm{L}}, \pi^{\mathrm{M,L}}), \tag{4}$$

where

$$\pi^{\mathrm{L}} = C^{-1} \exp\left(-\textstyle\sum_{i=1}^{\mathrm{L}} \theta_i\, E_i(\boldsymbol{u})\right),$$

$$C = \int_{\Omega} \exp\left(-\sum_{i=1}^{\mathrm{L}} \theta_i\, E_i(\boldsymbol{u})\right), \tag{5}$$

is the *least-biased estimate* of π based on L moment constraints

$$\int_{\Omega} \pi^{\mathrm{L}}(\boldsymbol{u}) E_i(\boldsymbol{u}) \mathrm{d}\boldsymbol{u} = \int_{\Omega} \pi(\boldsymbol{u}) E_i(\boldsymbol{u}) \mathrm{d}\boldsymbol{u}, \quad i = 1, \ldots, \mathrm{L}, \tag{6}$$

for the set of functionals $\boldsymbol{E} \equiv (E_1, \ldots, E_{\mathrm{L}})$ on the space Ω of the variables resolved by the imperfect models. Such densities were shown by Jaynes [7] to be least-biased in terms of information content and are obtained by maximizing the Shannon entropy, $\mathcal{S} = -\int \pi \ln \pi$, subject to the constraints in (6). Here, we assume that the functionals $\boldsymbol{E}$ are given by tensor powers of the resolved variables, $\boldsymbol{u} \in \Omega$, so that $E_i(\boldsymbol{u}) = \boldsymbol{u}^{\otimes i}$ and the expectations $\bar{E}_i$ yield the first L uncentered statistical moments of π; note that in this case, π^{L} for $\mathrm{L} = 2$ is a Gaussian density. In fact, the Gaussian framework when both the measurements of the truth dynamics and its model involve only the mean and covariance presents the most practical setup for utilizing the framework of information theory in climate change applications; note that considering only $\mathrm{L} = 2$ in (5) does not imply assuming that the underlying dynamics is Gaussian but merely focuses on tuning to the available second-order statistics of the truth dynamics.

In weather or climate change prediction, the complex numerical models for the climate system are calibrated (often in an ad hoc fashion) by comparing the spatiotemporal model statistics with the available coarse statistics obtained from various historical observations [5, 14]; we refer to this procedure as the *calibration phase* on the time interval $\mathcal{I}_{\mathrm{c}}$. The model optimization (3) carried out in the calibration phase can be represented, using the relationship (4), as

$$\mathscr{P}_{\mathcal{I}_c}(\pi, \pi^{\mathrm{M}^*_{\mathcal{I}_c}}) = \mathscr{P}_{\mathcal{I}_c}(\pi, \pi^{\mathrm{L}}) + \min_{\mathrm{M}\in\mathcal{M}} \mathscr{P}_{\mathcal{I}_c}(\pi^{\mathrm{L}}, \pi^{\mathrm{M,L}}), \tag{7}$$

where $\mathrm{M}^*_{\mathcal{I}_c} \in \mathcal{M}$ is the model with the smallest lack of information within $\mathcal{I}_c$. The first term, $\mathscr{P}_{\mathcal{I}_c}(\pi, \pi^{\mathrm{L}})$, in (7) represents an *intrinsic information barrier* [1,10] which cannot be overcome unless more measurements L of the truth are incorporated. The second term in (7) can be minimized directly since the least-biased estimates, π^{L}, of the truth which are known within $\mathcal{I}_c$; note that if $\mathscr{P}_{\mathcal{I}_c}(\pi^{\mathrm{L}}, \pi^{\mathrm{M}^*_{\mathcal{I}_c}}) \neq 0$, the corresponding information barrier can be reduced by enlarging the class of models $\mathcal{M}$.

The utility of the information-theoretic optimization principle (7) for improving climate change projections is best illustrated by linking the statistical model fidelity on the unperturbed attractor/climate and improved probabilistic predictions of the perturbed dynamics. Assume that the truth dynamics are perturbed so that the corresponding least-biased density, $\pi^{\mathrm{L},\delta}$, is perturbed smoothly to

$$\pi^{\mathrm{L},\delta} = \pi^{\mathrm{L}} + \delta\pi^{\mathrm{L}}, \quad \int_\Omega \delta\pi^{\mathrm{L}} = 0, \tag{8}$$

where π^{L} denotes the unperturbed least-biased density (5), and we skipped the explicit dependence on time and space. For stochastic dynamical systems with time-independent, invariant measure on the attractor, rigorous theorems guarantee this smooth dependence under minimal hypothesis [6]; for more general dynamics, this property remains as an empirical conjecture. Now, the lack of information in the perturbed least-biased model density, $\pi^{\mathrm{M},\delta}$, relative to the perturbed least-biased estimate of the truth, $\pi^{\mathrm{L},\delta}$, can be expressed as (see, e.g., [2,10])

$$\mathcal{P}\left(\pi^{\mathrm{L},\delta}, \pi^{\mathrm{M},\delta}\right) = \ln\left(C^{\mathrm{M},\delta}/C^{\delta}\right) + \left(\boldsymbol{\theta}^{\mathrm{M},\delta} - \boldsymbol{\theta}^{\delta}\right)\cdot \bar{\boldsymbol{E}}^{\delta}, \tag{9}$$

where $\bar{\boldsymbol{E}}^{\delta} = \bar{\boldsymbol{E}} + \delta\bar{\boldsymbol{E}}$ denotes the vector of L statistical moments with respect to the perturbed truth density, π^{δ}, and we suppressed the time dependence for simplicity. For smooth perturbations of the truth density $\delta\pi^{\mathrm{L}}$ in (8), the moment perturbations $\delta\bar{\boldsymbol{E}}$ remain small so that the leading-order Taylor expansion of (9) combined with the Cauchy-Schwarz inequality leads to the following link between the error in the perturbed and unperturbed truth and model densities:

$$\mathscr{P}_{\mathcal{I}}(\pi^{\mathrm{L},\delta}, \pi^{\mathrm{M},\delta}) \leqslant \left\|\boldsymbol{\theta}^{\mathrm{M}} - \boldsymbol{\theta}\right\|^{1/2}_{L^2(\mathcal{I})} \left\|\bar{\boldsymbol{E}}^{\delta}\right\|^{1/2}_{L^2(\mathcal{I})} + \mathcal{O}\big((\delta\bar{\boldsymbol{E}})^2\big), \tag{10}$$

where $\boldsymbol{\theta}^{\mathrm{M}}$, $\boldsymbol{\theta}$ are the Lagrange multipliers of the unperturbed densities π^{L}, π^{M} assumed in the form (5) and determined in the calibration phase $\mathcal{I}_c$ on the unperturbed attractor/climate. Thus, the result in (10) implies that optimizing the statistical model fidelity on the unperturbed attractor via (7) implies improved predictions of the perturbed dynamics. Illustration of the utility of the principle in (7) on a model of turbulent tracer dynamics, can be found in [11].

Multi-model Ensemble Predictions and Information Theory

Multi-model ensemble (MME) predictions are a popular technique for improving predictions in weather forecasting and climate change science (e.g., [4]). The heuristic idea behind MME prediction framework is simple: given a collection of imperfect models, consider predictions obtained through the convex superposition of the individual forecasts in the hope of mitigating model error. However, it is not obvious which models, and with what weights, should be included in the MME forecast in order to achieve the best predictive performance. Consequently, virtually all existing operational MME prediction systems are based on equal-weight ensembles which are likely to be far from optimal [4]. The information-theoretic framework allows for deriving a sufficient condition which guarantees prediction improvement via the MME approach relative to the single model forecasts [2].

The probabilistic predictions of the multi-model ensemble are represented in the present framework by the mixture density

$$\pi^{\mathrm{MME}}_{\boldsymbol{\alpha},t}(\boldsymbol{u}) \equiv \sum_i \alpha_i \pi^{\mathrm{M}_i}_t(\boldsymbol{u}), \quad \boldsymbol{u}\in\Omega, \tag{11}$$

where $\sum \alpha_i = 1$, $\alpha_i \geqslant 1$, and $\pi^{\mathrm{M}_i}_t$ represent probability densities associated with the imperfect models M_i in the class $\mathcal{M}$ of available models. Given the MME density $\pi^{\mathrm{MME}}_{\boldsymbol{\alpha},t}$, the optimization principle (7) over the time interval $\mathcal{I}$ can be expressed in terms of the weight vector $\boldsymbol{\alpha}$ as

$$\mathscr{P}_{\mathcal{I}}\left(\pi, \pi^{\mathrm{MME}}_{\boldsymbol{\alpha}^*_{\mathcal{I}}}\right) = \min_{\boldsymbol{\alpha}} \mathscr{P}_{\mathcal{I}}\left(\pi, \pi^{\mathrm{MME}}_{\boldsymbol{\alpha}}\right). \tag{12}$$

Clearly, MME prediction with the ensemble of models $\mathrm{M}_i \in \mathcal{M}$ is more skilful in terms of information content than the single model prediction with $\mathrm{M}_\diamond$ when

$$\mathscr{P}_{\mathcal{I}}\left(\pi, \pi^{\mathrm{MME}}_{\boldsymbol{\alpha}}\right) - \mathscr{P}_{\mathcal{I}}\left(\pi, \pi^{\mathrm{M}_\diamond}\right) < 0. \tag{13}$$

It turns out [2] that by exploiting the convexity of the relative entropy (1) in the second argument, i.e., $\mathcal{P}\big(\pi, \sum_{i=1} \alpha_i \pi^{\mathrm{M}_i}\big) \leqslant \sum_{i=1} \alpha_i \mathcal{P}\big(\pi, \pi^{\mathrm{M}_i}\big)$, it is possible to obtain a sufficient condition for improving imperfect predictions via the MME approach with $\pi^{\mathrm{MME}}_{\boldsymbol{\alpha}}$ relative to the single model predictions with $\mathrm{M}_\diamond$ in the form

$$\mathscr{P}_{\mathcal{I}}\left(\pi^{\mathrm{L}}, \pi^{\mathrm{M}_\diamond}\right) > \sum_{i \neq \diamond} \beta_i \, \mathscr{P}_{\mathcal{I}}\left(\pi^{\mathrm{L}}, \pi^{\mathrm{M}_i}\right),$$
$$\beta_i = \frac{\alpha_i}{1-\alpha_\diamond}, \quad \sum_{i \neq \diamond} \beta_i = 1, \tag{14}$$

where $\mathrm{M}_\diamond, \mathrm{M}_i \in \mathcal{M}$, and π^{L} is the least-biased density (5) based on L moment constraints which is practically measurable in the calibration phase. Further variants of this condition expressed via the statistical moments $\bar{\boldsymbol{E}}$, $\bar{\boldsymbol{E}}^{\mathrm{M}}$ are discussed in [2]. Here, we only highlight one important fact concerning the improvement of climate change predictions via the MME approach; using analogous arguments to those leading to (10) in the single model framework and the convexity of the relative entropy, the following holds in the MME framework:

$$\mathscr{P}_{\mathcal{I}}\left(\pi^{\mathrm{L},\delta}, \pi^{\mathrm{MME},\delta}_{\boldsymbol{\alpha}}\right) \leqslant \Big\| \sum_i \alpha_i \, \boldsymbol{\theta}^{\mathrm{M}_i} - \boldsymbol{\theta} \Big\|^{1/2}_{L^2(\mathcal{I})} \, \big\| \bar{\boldsymbol{E}}^{\delta} \big\|^{1/2}_{L^2(\mathcal{I})}$$
$$+ \mathcal{O}\big((\delta \bar{\boldsymbol{E}})^2\big), \tag{15}$$

where, for simplicity in exposition, the models M_i in MME are assumed to be in the least-biased form (5) and $\boldsymbol{\theta}$, $\boldsymbol{\theta}^{\mathrm{M}_i}$ are the Lagrange multipliers of the unperturbed truth and model densities determined in the calibration phase (see [2] for a general formulation). Thus, for sufficiently small perturbations, optimizing the weights $\boldsymbol{\alpha}$ in the density $\pi^{\mathrm{MME}}_{\boldsymbol{\alpha}}$ on the unperturbed attractor via (12) implies improved MME predictions of the perturbed truth dynamics. The potential advantage of MME predictions lies in the fact [2] that for optimal-weight MME in the training phase $\mathscr{P}_{\mathcal{I}}(\pi^{\mathrm{L}}, \pi^{\mathrm{MME}}_{\boldsymbol{\alpha}^*}) \leqslant \mathscr{P}_{\mathcal{I}}(\pi^{\mathrm{L}}, \pi^{\mathrm{M}^*_{\mathcal{I}}})$, where $\mathrm{M}^*_{\mathcal{I}}$ is the best single model within the training phase in terms of information content. However, MME predictions are inferior to the single model predictions when the MME weights $\boldsymbol{\alpha}$ are such that the condition (14) with $\mathrm{M}_\diamond = \mathrm{M}^*_{\mathcal{I}}$ is not satisfied. In summary, while the MME predictions can be superior to the single model predictions, the model ensemble has to be constructed with a sufficient care, and the information-theoretic framework provides means for accomplishing this task in a systematic fashion.

References

1. Branicki, M., Majda, A.J.: Quantifying uncertainty for long range forecasting scenarios with model errors in non-Guassian models with intermittency. Nonlinearity **25**, 2543–2578 (2012)
2. Branicki, M., Majda, A.J.: An information-theoretic framework for improving multi model ensemble forecasts. J. Nonlinear Sci. (2013, submitted) doi:10.1007/s00332-015-9233-1
3. Cover, T.A., Thomas, J.A.: Elements of Information Theory. Wiley-Interscience, Hoboken (2006)
4. Doblas-Reyes, F.J., Hagedorn, R., Palmer, T.N.: The rationale behind the success of multi-model ensembles in seasonal forecasting. II: calibration and combination. Tellus **57**, 234–252 (2005)
5. Emanuel, K.A., Wyngaard, J.C., McWilliams, J.C., Randall, D.A., Yung, Y.L.: Improving the Scientific Foundation for Atmosphere-Land Ocean Simulations. National Academic Press, Washington, DC (2005)
6. Hairer, M., Majda, A.J.: A simple framework to justify linear response theory. Nonlinearity **12**, 909–922 (2010)
7. Jaynes, E.T.: Information theory and statistical mechanics. Phys. Rev. **106**(10), 620–630 (1957)
8. Kleeman, R.: Measuring dynamical prediction utility using relative entropy. J. Atmos. Sci. **59**(13), 2057–2072 (2002)
9. Kullback, S., Leibler, R.: On information and sufficiency. Ann. Math. Stat. **22**, 79–86 (1951)
10. Majda, A.J., Gershgorin, B.: Quantifying uncertainty in climage change science through empirical information theory. Proc. Natl. Acad. Sci. **107**(34), 14958–14963 (2010)
11. Majda, A.J., Gershgorin, B.: Link between statistical equilibrium fidelity and forecasting skill for complex systems with model error. Proc. Natl. Acad. Sci. **108**(31), 12599–12604 (2011)
12. Majda, A.J., Gershgorin, B.: Improving model fidelity and sensitivity for complex systems through empirical information theory. Proc. Natl. Acad. Sci. **108**(31), 10044–10049 (2011)
13. Majda, A.J., Abramov, R.V., Grote, M.J.: Information Theory and Stochastics for Multiscale Nonlinear Systems. CRM Monograph Series, vol. 25. AMS, Providence (2005)

I

14. Randall, D.A.: Climate models and their evaluation. In: Climate Change 2007: The Physical Science Basis, Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, pp. 589–662. Cambridge University Press, Cambridge/New York (2007)

Inhomogeneous Media Identification

Fioralba Cakoni
Department of Mathematics, Rutgers University, New Brunswick, NJ, USA

Definition

Inhomogeneous media identification is the problem of determining the physical properties of an unknown inhomogeneity from its response to various interrogating modalities. This response, recorded in measured data, comes as a result of the interaction of the inhomogeneity with an exciting physical field. Inhomogeneous media identification is mathematically modeled as the problem of determining the coefficients of some partial differential equations with initial or boundary data from a knowledge of the solution on the measurement domain.

Formulation of the Problem

This survey discusses only the problem of *inhomogeneous media identification in inverse scattering theory*. Scattering theory is concerned with the effects that inhomogeneities have on the propagation of waves and in particular time-harmonic waves. In the context of this presentation, scattering theory provides the mathematical tools for imaging of inhomogeneous media via acoustic, electromagnetic, or elastic waves with applications to such fields as radar, sonar, geophysics, medical imaging, and nondestructive testing. For reasons of brevity, we focus our attention on the case of acoustic waves and refer the reader to Cakoni-Colton-Monk [5] for a comprehensive reading on media identification using electromagnetic waves. Since the literature in the area is enormous, we have only referenced a limited number of papers and monographs and hope that the reader can use these as starting point for further investigations.

We begin by considering the propagation of sound waves of small amplitude in R^3 viewed as a problem in fluid dynamics. Let $p(x,t)$ denote the pressure of the fluid which is a small perturbation of the static case, i.e., $p(x,t) = p_0 + \epsilon P_1(x,t) + \cdots$ where $p_0 > 0$ is a constant. Assuming that $p_1(x,t)$ is time harmonic, $p_1(x,t) = \Re\left\{u(x)e^{-i\omega t}\right\}$, we have that u satisfies (Colton-Kress 1998 [8])

$$\Delta u + \frac{\omega^2}{c^2(x)}u = 0 \tag{1}$$

where ω is the frequency and $c(x)$ is the sound speed. Equation (1) governs the propagation of time-harmonic acoustic waves of small amplitude in a slowly varying inhomogeneous medium. We still must prescribe how the wave motion is initiated and what is the boundary of the region contained in the fluid. We shall only consider the simplest case when the inhomogeneity is of compact support denoted by D, the region of consideration is all of R^3, and the wave motion is caused by an incident field u^i satisfying the unperturbed linearized equations being scattered by the inhomogeneous medium. Assuming that $c(x) = c_0 =$ constant for $x \in R^3 \setminus \overline{D}$, the total field $u = u^i + u^s$ satisfies

$$\Delta u + k^2 n(x) u = 0 \qquad \text{in } R^3 \tag{2}$$

and the scattered field u^s fulfills the Sommerfeld radiation condition

$$\lim_{|x|\to\infty} |x| \left(\frac{\partial u^s}{\partial |x|} - iku^s \right) = 0 \tag{3}$$

which holds uniformly in all directions $x/|x|$ where $k = \omega/c_0$ is the wave number and $n = c_0^2/c^2$ is the refractive index in the case of non-absorbing media. An absorbing medium is modeled by adding an absorption term which leads to a refractive index with a positive imaginary part of the form

$$n(x) = \frac{c_0^2}{c^2(x)} + i\frac{\gamma(x)}{k}$$

in terms of an absorption coefficient $\gamma > 0$ in $\overline{D}$. In the sequel, the *refractive index* n is assumed to be a piecewise continuous complex-valued function such

that $n(x) = 1$ for $x \notin D$ and $\Re(n) > 0$ and $\Im(n) \geq 0$. For a vector $d \in R^3$, with $|d| = 1$, the function $e^{ikx\cdot d}$ satisfies the Helmholtz equations in R^3, and it is called a *plane wave*, since $e^{i(kx\cdot d-\omega t)}$ is constant on the planes $kx \cdot d - \omega t$ =const. Summarizing, given the incident field u^i and the physical properties of the inhomogeneity, the *direct scattering problem* is to find the scattered wave and in particular its behavior at large distances from the scattering object, i.e., its far-field behavior. The *inverse scattering problem* takes this answer to the direct scattering problem as its starting point and asks what is the nature of the scatterer that gave rise to such far-field behavior?

Identification of Inhomogeneities from Far-Field Data

It can be shown that radiating solutions u^s to the Helmholtz equation (i.e., solutions that satisfy the Sommerfeld radiation condition (3)) assume the asymptotic behavior

$$u^s(x) = \frac{e^{ik|x|}}{|x|}\left\{u_\infty(\hat{x}) + O\left(\frac{1}{|x|}\right)\right\}, \quad |x| \to +\infty \tag{4}$$

uniformly for all directions $\hat{x}$ where the function u_∞ defined on the unit sphere S^2 is known as the far-field pattern of the scattered wave. For plane wave incidence $u^i(x, d) = e^{ikx\cdot d}$, we indicate the dependence of the far-field pattern on the incident direction d and the observation direction $\hat{x}$ by writing $u_\infty = u_\infty(\hat{x}, d)$. The *inverse scattering problem* or in other words *inhomogeneous media identification problem* can now be formulated as the problem of determining the index of refraction n (and hence also its support D) from a knowledge of the far-field pattern $u_\infty(\hat{x}, d)$ for $\hat{x}$ and d on the unit sphere S^2 (or a subset of S^2). All the results presented here are valid in R^2 as well. Also, it is possible to extend our discussion to the case of point source incidence and near-field measurements (see [3]).

Uniqueness

The first question to approach the problem is whether the inhomogeneous media is identifiable from the exact data, which in mathematical terms is known as the uniqueness problem. The uniqueness problem for inverse scattering by an inhomogeneous medium in R^3 was solved by Nachman [13], Novikov [14], and Ramm [16] who based their analysis on the fundamental work of Sylvester and Uhlmann [17]. Their uniqueness proof was considerably simplified by Hähner [9] (see [2, 13, 17] and the references in [8] and [10]). The uniqueness problem for an inhomogeneous media in R^2, which is a formerly determined problem, was recently solved by Bukhgeim [2]. In particular, under the assumptions on the refractive index stated in the Introduction, the following uniqueness result holds.

Theorem 1 *The refractive index n in (2) is uniquely determined from $u_\infty(\hat{x}, d)$ for $\hat{x}, d \in S^2$ and a fixed value of the wave number k.*

It is important to notice that owning to fact that u_∞ is real analytic in $S^2 \times S^2$, for the uniqueness problem, it suffices to know $u_\infty(\hat{x}, d)$ for $\hat{x}, d$ on subsets of S^2 having an accumulation point.

The identifiability problem for the matrix index of refraction of an anisotropic media is more complicated. In the mathematical model of the scattering by anisotropic media, Eq. (2) is replaced by

$$\nabla \cdot A\nabla u + k^2 n(x)u = 0 \qquad \text{in } R^3 \tag{5}$$

where n satisfies the same assumptions as in Introduction and A is a 3×3 piecewise continuous matrix-valued function with a positive definite real part, i.e., $\xi \cdot \Re(A)\xi > \alpha|\xi|^2, \alpha > 0$ in $\overline{D}$, non-positive imaginary part, i.e., $\xi \cdot \Im(A)\xi \leq 0$ in $\overline{D}$ and $A = I$ in $R^3 \setminus \overline{D}$. In general, it is known that $u_\infty(\hat{x}, d)$ for $\hat{x}, d \in S^2$ does not uniquely determine the matrix A even it is known for all wave numbers $k > 0$, and hence without further a priori assumptions, the determination of D is the most that can be hoped. To this end, Hähner (2000) proved that the support D of an anisotropic inhomogeneity is uniquely determined from $u_\infty(\hat{x}, d)$ for $\hat{x}, d \in S^2$ and a fixed value of the wave number k provided that either $\xi \cdot \Re(A)\xi > \beta|\xi|^2$ or $\xi \cdot \Re(A^{-1})\xi > \beta|\xi|^2$ for some constant $\beta > 1$.

Reconstruction Methods

Recall the scattering problem described by (2)–(3) for the total field $u = u^i + u^s$ with plane wave incident field $u^i := e^{ikx\cdot d}$. The total field satisfies the *Lippmann-Schwinger equation* (see [8])

$$u(x) = e^{ikx\cdot d} - \frac{k^2}{4\pi}\int_{R^3}\frac{e^{ik|x-y|}}{|x-y|}m(y)u(y)\,dy,\quad x\in R^3, \tag{6}$$

and the corresponding far-field pattern is given by

$$u_\infty(\hat{x},d) = -\frac{k^2}{4\pi}\int_{R^3} e^{-ik\hat{x}\cdot y}m(y)u(y)\,dy,\quad \hat{x},d\in S^2. \tag{7}$$

Since the function $m := 1-n$ has support D, the integrals in (6) and (7) can in fact be written over a bounded domain containing D. The goal is to reconstruct $m(x)$ from a knowledge of (the measured) far-field pattern $u_\infty(\hat{x},d)$ based on (7). The dependence of (7) on the unknown m is in a nonlinear fashion; thus, the inverse medium problem is genuinely a nonlinear problem. The reconstruction methods can, roughly speaking, be classified into three groups, Born or weak scattering approximation, nonlinear optimization techniques, and qualitative methods (we remark that this classification is not inclusive).

Born Approximation

Born approximation, known otherwise as weak scattering approximation, turns the inverse medium scattering problem into a linear problem and therefore is often employed in practical applications. This process is justified under restrictive assumption that the scattered field due to the inhomogeneous media is only a small perturbation of incident field, which at a given frequency is valid if either the corresponding contrast $n-1$ is small or the support D is small. Hence, assuming that $k^2\|m\|_\infty$ is sufficiently small, one can replace u in (7) by the plane wave incident field $e^{ikx\cdot d}$, thus obtaining the linear integral equation for m

$$u_\infty(\hat{x},d) = -\frac{k^2}{4\pi}\int_{R^3} e^{-ik(\hat{x}-d)\cdot y}m(y)\,dy,\quad \hat{x},d\in S^2. \tag{8}$$

Solving (8) for the unknown m corresponds to inverting the Fourier transform of m restricted to the ball of radius $2k$ centered at the origin, i.e., only incomplete data is available. This causes uniqueness ambiguities and leads to severe ill-posedness of the inversion. For details we refer the reader to Langenberg [12].

Nonlinear Optimization Techniques

These methods avoid incorrect model assumptions inherent in weak scattering approximation and consider the full nonlinear inverse medium problem. To write a nonlinear optimization setup, note that the inverse medium problem is equivalent to solving the system of equations composed by (6) and (7) for u and m where u_∞ is in practice the (noisy) measured data u_∞^δ with $\delta>0$ being the noise level. Thus, a simple least square approach looks for minimizing the cost functional

$$\mu(u,m) := \frac{\|u^i + Tmu - u\|^2_{L^2(B\times S^2)}}{\|u^i\|^2_{L^2(B\times S^2)}} + \frac{\|u_\infty^\delta - Fmu - u\|^2_{L^2(S^2\times S^2)}}{\|u_\infty^\delta\|^2_{L^2(B\times S^2)}}$$

for u and m over admissible sets, where Tmu denotes the integral in (6) and Fmu denotes the integral in (7). The discrete versions of this optimization problem suffer from a large number of unknowns and thus is expensive. Regularization techniques are needed to handle instability due to ill-posedness.

A more rigorous mathematical approach to deal with nonlinearity in (6) and (7) is the Newton-type iterative method. To this end, it is possible to reformulate the inverse medium problem as a nonlinear operator equation by introducing the operator $\mathcal{F}: m\to u_\infty$ that maps $m := 1-n$ to the far-field pattern $u_\infty(\cdot,d)$ for plane incidence $u^i(x) = e^{ikx\cdot d}$. In view of uniqueness theorem, $\mathcal{F}$ can be interpreted as an injective operator from $\mathcal{B}(B)$ (the space of bounded functions defined on a ball B containing the support D of m) into $L^2(S^2\times S^2)$ (the space of square integrable function on $S^2\times S^2$). From (7) we can write

$$(\mathcal{F}(m))(\hat{x},d) = -\frac{k^2}{4\pi}\int_B e^{-ik\hat{x}\cdot y}m(y)u(y)\,dy,\quad \hat{x},d\in S^2 \tag{9}$$

where $u(\cdot,d)$ is the unique solution of (6). Note that $\mathcal{F}$ is a compact operator, owning this to its analytic kernel; thus, (9) is severely ill-posed. From the latter it can be seen that the Fréchet derivative v_q of u with respect to m (in direction q) satisfies the Lippmann-Schwinger equation

$$v_q(x,d) + \frac{k^2}{4\pi}\int_B \frac{e^{ik|x-y|}}{|x-y|}\left[m(y)v_q(y,d) + q(y)u(y,d)\right]dy,\quad x\in B$$

which implies the following expression for the Fréchet derivative of $\mathcal{F}$

$$(\mathcal{F}'(m)q)(\hat{x},d) = -\frac{k^2}{4\pi}\int_B e^{-ik(\hat{x}-d)\cdot y} \left[m(y)v_q(y,d) + q(y)u(y,d)\right] dy, \quad \hat{x},d \in S^2.$$

Observe that $\mathcal{F}'(m)q = v_{q,\infty}$ where $v_{q,\infty}$ is the far-field pattern of the radiating solution to $\Delta v + k^2 n v = -k^2 uq$. It can be shown that $\mathcal{F}'(m)$ is injective (see [8, 10]). With the help of Fréchet derivative, it is now possible to replace (7) by its linearized version

$$\mathcal{F}(m) + \mathcal{F}'(m)q = u_\infty \tag{10}$$

which, given an initial guess m, it is solved for q to obtain an update $m+q$. Then as in the classical Newton iterations, this linearization procedure is iterated until some stopping criteria are satisfied. Of course the linearized equation inherits the ill-posedness of the nonlinear equation, and therefore regularization is required. If u_∞^δ is again the noisy far-field measurements, Tikhonov regularization replaces (10) by

$$\alpha q + [\mathcal{F}'(m)]^* \mathcal{F}'(m)q = [\mathcal{F}'(m)]^* \left\{u_\infty^\delta - \mathcal{F}(m)\right\}$$

with some positive regularization parameter α and the L^2 adjoint $[\mathcal{F}'(m)]^*$ of $\mathcal{F}'(m)$. Of course for the Newton method to work, one needs to start with a good initial guess incorporating available a priori information, but in principle the method can be formulated for one or few incident directions.

Qualitative Methods

In recent years alternative methods for imaging of inhomogeneous media have emerged which avoid incorrect model assumptions of weak approximations but, as opposed to nonlinear optimization techniques, require essentially no a priori information on the scattering media. Nevertheless, they seek limited information about scattering object and need multistatic data, i.e, several incident fields each measured at several observation directions. Such methods come under the general title of qualitative methods in inverse scattering theory. Most popular examples of such approaches are linear sampling method (Cakoni-Colton [3]), factorization method (Kirsch-Grinberg [11]), and singular sources method (Potthast [15]). Typically, these methods seek to determine an approximation to the support of the inhomogeneity by constructing a support indicator function and in some cases provide limited information on material properties of inhomogeneous media. We provide here a brief exposé of the linear sampling method. To this end let us define the *far-field operator* $F : L^2(S^2) \to L^2(S^2)$ by

$$(Fg)(\hat{x}) := \int_{S^2} u_\infty(\hat{x};d,k)g(d)ds(d) \tag{11}$$

We note that by linearity $(Fg)(\hat{x})$ is the far-field pattern corresponding to (1) where the incident field u^i is a *Herglotz wave function* $v_g(x) := \int_{S^2} e^{ikx\cdot d} g(d)ds(d)$. For given $k > 0$ the far-field operator is injective with dense range if and only if there does not exist a nontrivial solution $v, w \in L^2(D)$, $v - w \in H^2(D)$ of the transmission eigenvalue problem

$$\Delta w + k^2 n(x) w = 0 \text{ and } \Delta v + k^2 v = 0 \text{ in } D \tag{12}$$

$$w = v \text{ and } \frac{\partial w}{\partial \nu} = \frac{\partial v}{\partial \nu} \text{ on } \partial D \tag{13}$$

such that v is a Herglotz wave function. Values of $k > 0$ for which (12)–(13) has nontrivial solutions are called *transmission eigenvalues*. If $\Im(n) = 0$, there exits an infinite discrete set of transmission eigenvalues accumulating only at $+\infty$, [7]. Consider now the *far-field equation* $(Fg)(\hat{x}) = \Phi_\infty(\hat{x},z,k)$ where $\Phi_\infty(x,z,k) := \frac{1}{4\pi} e^{-ik\hat{x}\cdot z}$ (is the far-field pattern of the fundamental solution $\frac{e^{ik|x-y|}}{4\pi|x-y|}$ to the Helmholtz equation). The far-field equation is severely ill-posed owing to the compactness of the far-field operator which is an integral operator with analytic kernel.

Theorem 2 *Assume that k is not a transmission eigenvalue. Then: (1) If $z \in D$ for given $\epsilon > 0$ there exists $g_{z,\epsilon,k} \in L^2(S^2)$ such that $\|Fg_{z,\epsilon,k} - \Phi_\infty(\cdot,z,k)\|_{L^2(S^2)} < \epsilon$ and the corresponding Herglotz function satisfies $\lim_{\epsilon\to 0} \|v_{g_{z,\epsilon,k}}\|_{L^2(D)}$ exists finitely, and for a fixed $\epsilon > 0$, $\lim_{z\to\partial D} \|v_{g_{z,\epsilon,k}}\|_{L^2(D)} = +\infty$. (2) If $z \in R^3 \setminus \overline{D}$ and $\epsilon > 0$, every $g_{z,\epsilon,k} \in L^2(S^2)$ satisfying $\|Fg_{z,\epsilon,k} - \Phi_\infty(\cdot,z,k)\|_{L^2(S^2)} < \epsilon$ is such that $\lim_{\epsilon\to 0} \|v_{g_{z,\epsilon,k}}\|_{L^2(D)} = +\infty$.*

The *linear sampling method* is based on attempting to compute the function $g_{z,\epsilon,k}$ in the above theorem by using Tikhonov regularization as the unique minimizer of the *Tikhonov functional* (see [8])

$$\|F^{\delta}g - \Phi(\cdot, z)\|^2_{L^2(\Omega)} + \alpha\|g\|^2_{L^2(S^2)} \qquad (14)$$

where the positive number $\alpha := \alpha(\delta)$ is known as the *Tikhonov regularization parameter* and $F^{\delta}g$ is the noisy far-field operator where u_∞ in (7) is replaced by the noisy far-field data u^{δ}_∞ with $\delta > 0$ being the noise level (note that $\alpha_\delta \to 0$ as $\delta \to 0$). In particular, one expects that this regularized solution will be relatively smaller for $z \in D$ than $z \in R^3 \setminus \overline{D}$, and this behavior can be visualized by color coding the values of the regularized solution on a grid over some domain containing the support D of the inhomogeneity and thus providing a reconstruction of D. A precise mathematical statement on the described behavior of the regularized solution to the far-field equation is based on factorization method which instead of the far-field operator F considers $(F^*F)^{1/4}$ where F^* is the L^2 adjoint of F (see [11]). For numerical examples using linear sampling method, we refer the reader to [3].

Having reconstructed the support of the inhomogeneity D, we then obtain information on $n(x)$ for non-absorbing media, i.e., if $\Im(n) = 0$. Assume to this end that $n(x) > 1$ (similar results hold for $0 < n(x) < 1$), fix a $z \in D$, and consider a range of wave number $k > 0$. If $g_{\delta,z,k}$ is now the Tikhonov-regularized solution of the far-field equation (14), then we have that: (1) for $k > 0$ *not* a transmission eigenvalue $\lim_{\delta\to 0} \|v_{g_{\delta,z,k}}\|_{L^2(D)}$ exists finitely [1] and (2) for $k > 0$ a transmission eigenvalue $\lim_{\delta\to 0} \|v_{g_{\delta,z,k}}\|_{L^2(D)} = +\infty$ (for almost all $z \in D$) [4]. In practice, this means if $\|g_{\delta,z,k}\|_{L^2(S^2)}$ is plotted against k, the transmission eigenvalues will appear as sharp picks and thus providing a way to compute transmission eigenvalues from far-field measured data. A detailed study of transmission eigenvalue problem [7] reveals that the first transmission is related to the index of refraction n. More specifically, letting $n_* = \inf_D n(x)$ and $n^* = \sup_D n(x)$, the following Faber-Krahn type inequalities hold:

$$k^2_{1,n(x),D} \geq \frac{\lambda_1(D)}{n^*} \qquad (15)$$

where $k_{1,n(x),D}$ is the first transmission eigenvalue corresponding to d and $n(x)$ and $\lambda_1(D)$ is the first Dirichlet eigenvalue for $-\Delta$ in D, and

$$0 < k_{1,D,n^*} \leq k_{1,D,n(x)} \leq k_{1,D,n_*} \qquad (16)$$

which is clearly seen to be isoperimetric for $n(x)$ equal to a constant. In particular, (16) shows that for n constant, the first transmission eigenvalue is monotonic decreasing function of n, and moreover, this dependence can be shown to be continuous and strictly monotonic. Using (16), for a measured first transmission eigenvalue $k_{1,D,n(x)}$, we can determine a unique constant n_0 that satisfies $0 < n_* \leq n_0 \leq n^*$, where this constant is such that $k_{1,D,n_0} = k_{1,D,n(x)}$. This n_0 is an integrated average of $n(x)$ over D.

A more interesting question is what does the first transmission eigenvalue say about the matrix index of refraction A for the scattering problem for anisotropic media (5). Assuming $n = 1$, $\overline{\xi} \cdot \Re(A)\xi > |\xi|^2$ and $\overline{\xi} \cdot \Im(A)\xi = 0$ in (5), similar analysis for the corresponding transmission eigenvalue problem leads to the isoperimetric inequality $0 < k_{1,D,a^*} \leq k_{1,D,A(x)} \leq k_{1,D,a_*}$ [6]. Hence, it is possible to compute a constant a_0 such that k_{1,D,a_0} equals the (measured) first transmission eigenvalue $k_{1,D,A(x)}$, and this constant satisfies $0 < a_* \leq a_0 \leq a^*$, where $a_* = \inf_D a_1(x)$, $a^* = \sup a_3(x)$, and $a_1(x)$ and $a_3(x)$ are the smallest and the largest eigenvalues of the matrix $A^{-1}(x)$, respectively. The latter inequality is of particular interest since $A(x)$ is not uniquely determined from the far field, and to our knowledge this is the only information obtainable to date about $A(x)$ that can be determined from far-field data (see [6] for numerical examples).

Cross-References

▸ Optical Tomography: Applications

References

1. Arens, T.: Why linear sampling works. Inverse Probl. **20** (2004)
2. Bukhgeim, A.: J. Recovering a potential from Cauchy data in the two dimensional case. Inverse Ill-Posed Probl. **16** (2008)
3. Cakoni, F., Colton, D.: Qualitative Methods in Inverse Scattering Theory. Springer, Berlin (2006)

4. Cakoni, F., Colton, D., Haddar, H.: On the determination of Dirichlet and transmission eigenvalues from far field data. Comptes Rendus Math. **348** (2010)
5. Cakoni, F., Colton, D., Monk, P.: The Linear Sampling Method in Inverse Electromagnetic Scattering CBMS-NSF, vol. 80. SIAM, Philadelphia (2011)
6. Cakoni, F., Colton, D., Monk, P., Sun, J.: The inverse electromagnetic scattering problem for anisotropic media. Inverse Probl. **26** (2010)
7. Cakoni, F., Gintides, D., Haddar, H.: The existence of an infinite discrete set of transmission eigenvalues. SIAM J. Math. Anal. **42** (2010)
8. Colton, D., Kress, R.: Inverse Acoustic and Electromagnetic Scattering Theory, 2nd edn. Springer, Berlin (1998)
9. Hähner, P.: A periodic Faddeev-type solution operator. J. Differ. Equ. **128**, 300–308 (1996)
10. Hähner, P.: Electromagnetic wave scattering. In: Pike, R., Sabatier, P. (eds.) Scattering. Academic, New York (2002)
11. Kirsch, A., Grinberg, N.: The Factorization Method for Inverse Problems. Oxford University Press, Oxford (2008)
12. Langenberg, K.: Applied inverse problems for acoustic, electromagnetic and elastic wave scattering. In: Sabatier (ed.) Basic Methods of Tomography and Inverse Problems. Adam Hilger, Bristol/Philadelphia (1987)
13. Nachman, A.: Reconstructions from boundary measurements. Ann. Math. **128** (1988)
14. Novikov, R.: Multidimensional inverse spectral problems for the equation $-\Delta\psi + (v(x) - Eu(x))\psi = 0$. Transl. Funct. Anal. Appl. **22**, 263–272 (1888)
15. Potthast, R.: Point Sourse and Multipoles in Inverse Scattering Theory. Research Notes in Mathematics, vol. 427. Chapman and Hall/CRC, Boca Raton (2001)
16. Ramm, A.G.: Recovery of the potential from fixed energy scattering data. Inverse Probl. **4**, 877–886 (1988)
17. Sylvester, J., Uhlmann, G.: A global uniqueness theorem for an inverse boundary value problem. Ann. Math. **125** (1987)

Initial Value Problems

Ernst Hairer and Gerhard Wanner
Section de Mathématiques, Université de Genève, Genève, Switzerland

We describe initial value problems for ordinary differential equations and dynamical systems, which have a tremendous range of applications in all branches of science. We also explain differential equations on manifolds and systems with constraints.

Ordinary Differential Equations

An ordinary differential equation is a formula

$$\dot{y} = f(t, y),$$

which relates the time derivative of a function $y(t)$ to its function value. Any function $y(t)$ defined on an interval $I \subset \mathbb{R}$ and satisfying $\dot{y}(t) = f(t, y(t))$ for all $t \in I$ is called a solution of the differential equation. If the value $y(t)$ is prescribed at some point t_0, we call the problem

$$\dot{y} = f(t, y), \qquad y(t_0) = y_0$$

an initial value problem. In most situations of practical interest the function $y(t)$ is vector-valued, so that we are in fact concerned with a system

$$\begin{aligned} \dot{y}_1 &= f_1(t, y_1, \ldots, y_n), \;\; y_1(t_0) = y_{10}, \\ &\vdots \qquad\qquad\qquad\quad \vdots \\ \dot{y}_n &= f_n(t, y_1, \ldots, y_n), \;\; y_n(t_0) = y_{n0}. \end{aligned}$$

The differential equation is called autonomous if the vector field f does not explicitly depend on time t.

An equation of the form

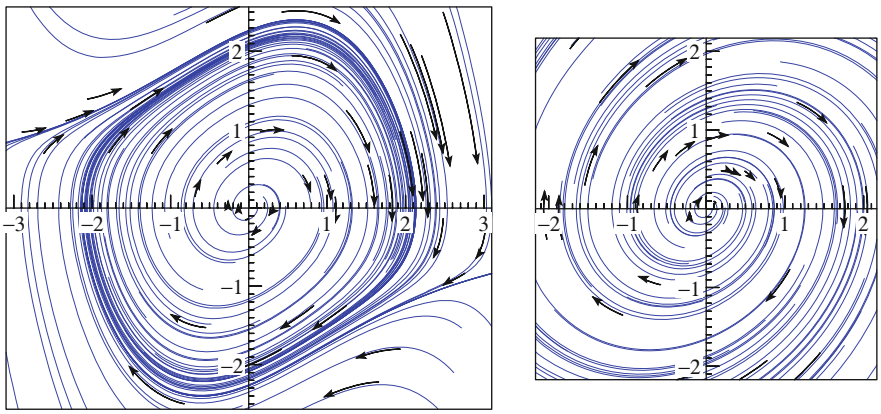
$$y^{(k)} = f(t, y^{(k-1)}, \ldots, \dot{y}, y)$$

is a differential equation of order k. By introducing the variables $y_1 = y$, $y_2 = \dot{y}$, $\ldots$, $y_k = y^{(k-1)}$, and adding the equations $\dot{y}_j = y_{j+1}$ for $j = 1, \ldots, k-1$, such a problem is transformed into a system of first-order equations.

Example 1 The Van der Pol oscillator is an autonomous second-order differential equation. Written as a first-order system the equations are given by

$$\begin{aligned} \dot{y}_1 &= y_2 \\ \dot{y}_2 &= \mu\,(1 - y_1^2)\,y_2 - y_1. \end{aligned}$$

Since the problem is autonomous, solutions can conveniently be plotted as paths in the phase space (y_1, y_2). Several of them can be seen in Fig. 1 (left). Arrows indicate the direction of the flow. We observe that all solutions tend for a large time to a periodic solution (limit cycle).



Initial Value Problems, Fig. 1 Solutions in the phase space of the Van der Pol oscillator for $\mu = 0.4$ (*left*); solutions of the linearized equation (*right*)

Linear Systems with Constant Coefficients

Systems of differential equations can be solved analytically only in very special situations. One of them are linear equations with constant coefficients,

$$\dot{y} = Ay, \qquad y(0) = y_0,$$

where $y(t) \in \mathbb{R}^n$, and A is a constant matrix of dimension n. A linear change of coordinates $y = Tz$ transforms the system into $\dot{z} = \Lambda z$ with $\Lambda = T^{-1}AT$. If T can be chosen such that $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ is diagonal, we obtain $z_j(t) = \mathrm{e}^{\lambda_j t} c_j$, and the solution $y(t)$ via the relation $y(t) = Tz(t)$. The free parameters $c_1, \ldots, c_n$ can be chosen to match the initial condition $y(0) = y_0$.

If the matrix A cannot be diagonalized, it can be transformed to upper triangular form (Schur or Jordan canonical form). Starting with $z_n(t)$, the functions $z_j(t)$ can be obtained successively by solving scalar, inhomogeneous linear equations with constant coefficients.

An explicit formula for the solution of the linear system $\dot{y} = Ay$ is obtained by using the matrix exponential

$$y(t) = \exp(At)\, y_0, \qquad \exp(At) = \sum_{k=0}^{\infty} A^k \frac{t^k}{k!}.$$

Example 2 If we neglect in the Van der Pol equation, for y_1 small, the cubic term $y_1^2 y_2$, we obtain the system

$$\dot{y}_1 = y_2$$
$$\dot{y}_2 = \mu\, y_2 - y_1,$$

which leads, for $0 < \mu < 2$, to complex eigenvalues $\lambda_{1,2} = \nu \pm \mathrm{i}\omega$ with $\nu = \frac{\mu}{2}$ and $\omega = \sqrt{1 - \nu^2}$. The solutions are thus linear combinations of $\mathrm{e}^{\nu t} \cos \omega t$ and $\mathrm{e}^{\nu t} \sin \omega t$. Some of these outward spiraling solutions are displayed in Fig. 1 (right) and mimic those of the Van der Pol equation close to the origin.

Existence, Uniqueness, and Differentiability of the Solutions

Whenever it is not possible to find the solution of a differential equation in analytic form, it is still of interest to study its existence, uniqueness, and qualitative properties.

Existence and Uniqueness

Consider a differential equation $\dot{y} = f(t, y)$ with a continuously differentiable function $f : U \to \mathbb{R}^n$, where $U \subset \mathbb{R} \times \mathbb{R}^n$ is an open set, and let $(t_0, y_0) \in U$. Then, there exists a unique function $y : I \to \mathbb{R}^n$ on a (maximal) open interval $I = I(t_0, y_0)$ such that

- $\dot{y}(t) = f(t, y(t))$ for $t \in I$ and $y(t_0) = y_0$.
- $(t, y(t))$ approaches the border of U whenever t tends to the left (or right) end of the interval I.
- If $z : J \to \mathbb{R}^n$ is a solution of $\dot{y} = f(t, y)$ satisfying $z(t_0) = y_0$, then $J \subset I$ and $z(t) = y(t)$ for $t \in J$.

The statement is still true if the differentiability assumption is weakened to a "local Lipschitz condition." In this case the local existence and uniqueness result is known as the theorem of Picard–Lindelöf.

Variational Equation

If the dependence on the initial condition is of interest, one denotes the solution by $y(t, t_0, y_0)$. It is defined on the set

$$D = \{(t, t_0, y_0)\,;\, (t_0, y_0) \in U,\, t \in I(t_0, y_0)\}.$$

This set is open, and the solution $y(t, t_0, y_0)$ is continuously differentiable with respect to all variables. Its derivative with respect to the initial value y_0 is the solution of the variational equation

$$\dot{\Psi}(t) = \frac{\partial f}{\partial y}\big(t, y(t, t_0, y_0)\big)\,\Psi(t), \qquad \Psi(t_0) = I.$$

Stability

The stability of a solution tells us how sensible it is with respect to perturbations in the initial value.

Stability of Linear Problems

The analytic solution of a problem $\dot{y} = Ay$ is a linear combination of expressions $p(t)\mathrm{e}^{\lambda t}$, where λ is an eigenvalue of A and $p(t)$ is a polynomial of degree $k-1$, where k is the dimension of the Jordan block corresponding to λ. As a consequence we have for solutions of $\dot{y} = Ay$:

- If all eigenvalues of A satisfy $\Re\lambda < 0$, then $y(t) \to 0$ for $t \to \infty$; the solution is called asymptotically stable.
- If all eigenvalues of A satisfy $\Re\lambda \le 0$ and the Jordan block of eigenvalues with $\Re\lambda = 0$ is of dimension one, then $y(t)$ is bounded for $t \to \infty$; the solution is called stable.
- If there exists an eigenvalue with $\Re\lambda > 0$ or an eigenvalue with $\Re\lambda = 0$ whose Jordan block is larger than one, then most solutions are unbounded for $t \to \infty$; the problem is called unstable.

The same is true for the difference between two solutions, because the problem is linear.

Stability for Nonlinear Problems

The stability investigation of solutions for nonlinear problems is much more involved. However, there are simple criteria for stationary solutions (i.e., $y(t) = y_0$, where $f(y_0) = 0$) of autonomous differential equations $\dot{y} = f(y)$:

- If all eigenvalues of the matrix $f'(y_0)$ satisfy $\Re\lambda < 0$, then the stationary solution is asymptotically stable. This means that it is stable (i.e., for every $\varepsilon > 0$ there exists a $\delta > 0$ such that, if $\|z_0\| < \delta$, we have $\|y(t, 0, y_0 + z_0) - y_0\| < \varepsilon$ for all $t \ge 0$), and that for sufficiently small $\|z_0\|$ one has $y(t, 0, y_0 + z_0) \to y_0$ for $t \to \infty$.
- If there exists an eigenvalue of $f'(y_0)$ satisfying $\Re\lambda > 0$, then the stationary solution is unstable. This means that there exist arbitrarily small perturbations z_0 for which the solution $y(t, 0, y_0 + z_0)$ moves away from y_0 (example: the origin for Van der Pol's equation in Fig. 1 is unstable).

Contractivity

If the vector field satisfies a one-sided Lipschitz condition, i.e., there exists a number ν such that

$$\langle f(t, y) - f(t, z), y - z\rangle \le \nu\|y - z\|^2$$

for all y and z, then the difference of any two solutions can be estimated as

$$\|y(t) - z(t)\|^2 \le \mathrm{e}^{\nu(t-t_0)}\|y(t_0) - z(t_0)\|^2 \quad \text{for} \quad t \ge t_0.$$

Differential Equations on Manifolds

There are problems where solutions of a differential equation evolve on a submanifold of $\mathbb{R}^n$. The manifold is typically given by algebraic constraints (preservation of energy and momentum, first integrals, holonomic constraints for mechanical systems). Much of the theory of differential equations (existence, uniqueness, etc.) carries over to this situation.

Closely related to differential equations on manifolds are so-called differential-algebraic equations. They can be written in the form

$$M\dot{y} = f(t, y), \qquad y(t_0) = y_0$$

with a constant but possibly singular matrix M. We cannot expect that such a problem has always a (local) solution, even if $f(t, y)$ is sufficiently smooth. One immediately sees that $f(t_0, y_0)$ has to be in the range of M, but this is not sufficient in general.

Problems of Index 1

Consider problems of the form

$$\begin{aligned} \dot{y} &= f(t, y, z), \quad & y(t_0) &= y_0 \\ 0 &= g(t, y, z), \quad & z(t_0) &= z_0, \end{aligned}$$

where the Jacobian matrix $\frac{\partial g}{\partial z}$ is invertible in a neighborhood of (t_0, y_0, z_0) (index 1 condition). Obviously, the initial values have to satisfy $g(t_0, y_0, z_0) = 0$. This permits to apply the implicit function theorem and to express $z = \zeta(t, y)$ from the algebraic relation. As a consequence the problem is equivalent to the ordinary differential equation $\dot{y} = f(t, y, \zeta(t, y))$ and the standard theory can be applied.

Problems of Index 2

Problems from control theory often have the form

$$\begin{aligned} \dot{y} &= f(y, z), \quad & y(0) &= y_0 \\ 0 &= g(y), \quad & z(0) &= z_0, \end{aligned}$$

where for notational convenience we suppress the dependence of t. The index 1 condition is violated, because g does not depend on z. Differentiating the algebraic relation with respect to time yields

$$g_y(y) f(y, z) = 0.$$

If $(g_y f_z)(y_0, z_0)$ is invertible (index 2 condition), the implicit function theorem implies that $z = \zeta(y)$ close to the initial value. We thus get a differential equation $\dot{y} = f(y, \zeta(y))$ on the manifold $\mathcal{M} = \{y\,;\, g(y) = 0\}$. Consistent initial values have to satisfy both constraints, $g(y_0) = 0$ and $(g_y f)(y_0, z_0) = 0$.

Problems of Index 3

Mechanical systems with holonomic constraints are problems of the form

$$\begin{aligned} \dot{y} &= f(y, z), \quad & y(0) &= y_0 \\ \dot{z} &= h(y, z, u), \quad & z(0) &= z_0 \\ 0 &= g(y), \quad & u(0) &= u_0. \end{aligned}$$

One has to differentiate twice the algebraic relation to be able to write $u = v(y, z)$. If this is possible, we get a differential equation for (y, z) on the manifold $\mathcal{M} = \{(y, z)\,;\, g(y) = 0,\, (g_y f)(y, z) = 0\}$. Consistent initial values have to satisfy $(y_0, z_0) \in \mathcal{M}$, and $u_0 = v(y_0 z_0)$.

Notes

There are many excellent books on the theory of ordinary differential equations. Let us just mention the classical monographs by Arnold [1], Hartman [5], and Chapter I of [3]. Concerning the theory of differential-algebraic equations we refer to [2] and to Chapters VI and VII of [4].

References

1. Arnold, V.I.: Ordinary Differential Equations. Universitext, Springer-Verlag, Berlin (2006), Translated from the Russian, Second printing of the 1992 edition.
2. Brenan, K.E., Campbell, S.L., Petzold, L.R.: Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations. Classics in Appl. Math. SIAM, Philadelphia (1996)
3. Hairer, E., Nørsett, S.P., Wanner, G.: Solving Ordinary Differential Equations I. Nonstiff Problems, Springer Series in Computational Mathematics, vol. 8, 2nd edn. Springer, Berlin (1993)
4. Hairer, E., Wanner, G.: Solving Ordinary Differential Equations II. Stiff and Differential-Algebraic Problems, Springer Series in Computational Mathematics, vol. 14, 2nd edn. Springer, Berlin (1996)
5. Hartman, P.: Ordinary Differential Equations, Classics in Applied Mathematics, vol. 38. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (2002). Corrected reprint of the second (1982) edition (Birkhäuser, Boston, MA)

Integro-Differential Equations: Computation

Hermann Brunner
Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong SAR, China
Department of Mathematics and Statistics, Memorial University of Newfoundland, St. John's, NL, Canada

Mathematics Subject Classification

65R99; 65M60

Funding: Hong Kong Research Grants Council (HKBU 200207)

Synonyms

Integro-differential equation (IDE)

Integro Differential Equations

The standard form of a first-order, nonlinear Volterra IDE for an unknown function $u = u(t)$ is

$$u'(t) = f(t, u(t)) + \int_0^t k(t, s, u(s))\, ds, \quad t \in [0, T], \tag{1}$$

complemented by an initial condition $u(0) = u_0$. In applications, k has often the *Hammerstein* form $k(t, s, u) = K(t, s)G(s, u)$, where G is smooth and K is either bounded (or even smooth) or weakly singular (integrable), e.g., $K(t, s) = (t - s)^{\alpha-1}$ $(0 < \alpha < 1)$ or $K(t, s) = \log(t - s)$.

Many Volterra-type IDEs arising in mathematical modelling processes (Volterra [18], Brunner [2, Sects. 3.6, 4.8, and 7.8], Janno and von Wolfersdorf [9], Shakourifar and Enright [16]) are of *nonstandard* form (in each equation, the nonstandard part is underlined):

$$u'(t) = f(t, u(t)) + \int_0^t k(t, s, \underline{u(t)}, u(s))\, ds \tag{2}$$

$$u'(t) = f(t, u(t)) + \int_0^t k(t, s, u(s), \underline{u'(s)})\, ds \tag{3}$$

$$u'(t) = f(t, u(t)) + \int_0^t K(t, s)\underline{u(t - s)}u(s)\, ds \tag{4}$$

$$u'(t) = f(t, u(t), u(\theta(t))) + \int_{\underline{\theta(t)}}^t k(t, s, u(s), u'(s))\, ds \tag{5}$$

In (5), θ denotes a *delay function* satisfying $\theta(t) < t$ (e.g., $\theta(t) = t - \tau$, $\tau > 0$: constant delay).

In an IDE of *Fredholm* type, the limits of integration are fixed, and the order of the IDE is usually even. Thus, a typical Fredholm IDE has the (Hammerstein) form

$$u^{(2m)}(t) + \sum_{j=0}^{2m-1} a_j(t)u^{(j)}(t) = \int_0^T \sum_{j=0}^{2m} K_j(t, s)$$
$$G_j(s, u^{(j)}(s))\, ds = f(t), t \in [0, T], \tag{6}$$

where u is subject to boundary conditions at $t = 0$ and $t = T$ (Ganesh and Sloan [8] and references).

The solution $u = u(t, x)$ of a *partial* Volterra or Fredholm IDE depends also on the spatial variable $x \in \Omega \subset \mathbb{R}^N$ (where Ω is bounded or unbounded). The following are representative examples of such equations arising in a variety of applications where memory effects play a role, for example, in heat conduction or viscoelasticity in materials with memory, and in stochastic processes of financial mathematics (Renardy et al. [15], Prüss [14], Matache et al. [12]; see also Souplet [17] and Appell et al. [1], especially for partial Volterra-Fredholm IDEs and Fredholm IDEs):

$$u_t + Au = \int_0^t h(t - s)Bu(s, \cdot)\, ds + f(t, x),$$
$$x \in \Omega,\ t \geq 0 \tag{7}$$

$$u_{tt} + Au = \int_0^t h(t - s)Bu(s, \cdot)\, ds + f(t, x),$$
$$x \in \Omega,\ t \geq 0 \tag{8}$$

$$u_t + \int_0^t h(t - s)Au(s, \cdot)\, ds = f(t, x),$$
$$x \in \Omega,\ t \geq 0 \tag{9}$$

$$u_t + Au = \int_\Omega G(t - s, x, \xi)Bu(\cdot, \xi)\, d\xi + f(t, x),$$
$$x \in \Omega,\ t \geq 0 \tag{10}$$

Here, A denotes a linear or nonlinear elliptic spatial partial differential operator (typically: $A = -\Delta$), while B is a spatial partial differential operator of order not exceeding two. The convolution kernel h either is bounded (and smooth) or has the form $h(z) = z^{\alpha-1}$ $(0 < \alpha < 1)$.

Fractional diffusion and wave equations represent another class of partial IDEs whose numerical solution is presently receiving considerable attention. Representative examples of such IDEs are

$$\frac{1}{\Gamma(\alpha)} \int_0^t (t - s)^{\alpha-1} \frac{\partial u(s, \cdot)}{\partial s}\, ds - \Delta u = f(t, x)$$
$$(0 < \alpha < 1) \tag{11}$$

and

$$u_t - \frac{1}{\Gamma(\alpha)} \int_0^t (t - s)^{\alpha-1} \Delta u(s, \cdot)\, ds = F(t, x, u, \nabla u) \tag{12}$$

(see Cuesta et al. [7] and Brunner et al. [5], also for references). Note that these IDEs are intermediate between the diffusion equation ($\alpha = 0$) and the wave equation ($\alpha = 1$).

Computational Solution of IDEs

Collocation methods (Brunner [2], also for higher-order IDEs) and *discontinuous Galerkin* (DG) methods (Brunner and Schötzau [3]) based on piecewise polynomials with respect to suitable meshes $I_h := \{t_n : 0 = t_0 < t_1 < \cdots < t_N = T\}$ are the methods of choice for the computational solution of general Volterra IDEs (1)–(5). If the kernel function k contains an integrable singularity like $(t-s)^{\alpha-1}$ $(0 < \alpha < 1)$, the solution $u(t)$ has an unbounded second derivative at $t = 0$; in order to obtain high-order collocation or DG solutions, meshes I_h that are suitably refined (graded) near $t = 0$ have to be employed (Brunner et al. [4] and Brunner and Schötzau [3]). The same is true if these methods are used as time-stepping methods in spatially semi-discretized partial Volterra IDEs with weakly singular kernels (Mustapha et al. [13]; see section "Computational Solution of Partial IDEs" below).

If the IDE (1) contains a *Hammerstein* kernel of *convolution* type, $k(t,s,u) = h(t-s)G(s,u)$, the computationally most efficient methods are the ones based on *convolution quadrature* techniques, combined with adaptive step-size control (López-Fernández et al. [11]).

While the efficient computational solution of *autoconvolution* IDEs (4) remains to be studied, it is well understood for *delay* IDEs (5) (Brunner [2]). An effective algorithm is presented in Shakourifar and Enright [16]: the underlying numerical method is based on explicit continuous *Runge–Kutta* methods with adaptive step-size control.

Turning to boundary-value problems for *Fredholm* IDEs (6), it is shown in Ganesh and Sloan [8] that *orthogonal collocation* yields an efficient and highly accurate computational scheme for solving such IDEs.

Computational Solution of Partial IDEs

The spatial discretization of time-dependent partial IDEs (approximation of the spatial partial differential operators A and B in (7)–(9)) – based on finite element/Galerkin techniques (cf. Chen and Shih [6] and references) – leads to *high-dimensional* systems of (linear or nonlinear) IDEs of the forms (1). The *time-stepping* methods for discretizing these systems of IDEs are usually adaptations of the computational methods for IDEs described in section "Computational Solution of IDEs". Typical examples of *one-point collocation* schemes are the *backward Euler* method and the implicit *Crank-Nicolson* method (Chen and Shih [6]). Time stepping by means of the *DG* method and its *hp* implementation is described in Larsson et al. [10] and Mustapha et al. [13], respectively.

In the case of *convolution kernels*, convolution quadrature time-stepping schemes, for example, those based on the second-order backward differentiation formula, yield fast and efficient computational methods (Cuesta et al. [7] and López-Fernández et al. [11]; the latter paper also contains a pseudocode of the algorithm).

A major problem in the computational solution of partial IDEs (especially IDEs of Fredholm type (10)) is that the matrices arising in the spatial semi-discretization of (7)–(10) are densely populated, owing to the *nonlocal* integral terms. Thus, in 2D and 3D spatial environments, the design of an efficient and fast time-stepping scheme will have to employ (wavelet-based) *matrix compression* techniques applied to the system of IDEs resulting from the spatial semi-discretization. In the case of parabolic Fredholm IDEs of the form (10), an efficient such algorithm is presented in Matache et al. [12] for Fokker-Planck IDEs modelling Markov processes with jumps.

There is a rapidly increasing number of papers on the computational solution of *fractional diffusion and wave equations* (11) and (12), as shown for example in Cuesta et al. [7] and the references in Brunner et al. [5]. Compare also López-Fernández et al. [11], Sect. 5.

References

1. Appell, J.M., Kalitvin, A.S., Zabrejko, P.P.: Partial Integral Operators and Integro-Differential Equations. Marcel Dekker Inc, New York (2000)
2. Brunner, H.: Collocation Methods for Volterra Integral and Related Functional Differential Equations. Cambridge University Press, Cambridge (2004)
3. Brunner, H., Schötzau, D.: *hp*-discontinuous Galerkin time-stepping for Volterra integro-differential equations. SIAM J. Numer. Anal. **44**, 224–245 (2006)

4. Brunner, H., Pedas, A., Vainikko, G.: Piecewise polynomial collocation methods for linear Volterra integro-differential equations with weakly singular kernels. SIAM J. Numer. Anal. **39**, 957–982 (2001)
5. Brunner, H., Ling, L., Yamamoto, M.: Numerical simulation of 2D fractional subdiffusion problems. J. Comput. Phys. **229**, 6613–6622 (2010)
6. Chen, C., Shih, T.: Finite Element Methods for Integrodifferential Equations. World Scientific, River Edge (1998)
7. Cuesta, E., Lubich, C., Palencia, C.: Convolution quadrature time discretization of fractional diffusion-wave equations. Math. Comput. **75**, 673–696 (2006)
8. Ganesh, M., Sloan, I.H.: Optimal order spline methods for nonlinear differential and integro-differential equations. Appl. Numer. Math. **29**, 445–478 (1999)
9. Janno, J., von Wolfersdorf, L.: Integro-differential equations of first order with autoconvolution integral. J. Integral Equ. Appl. **21**, 39–75 (2009)
10. Larsson, S., Thomée, V., Wahlbin, L.B.: Numerical solution of parabolic integro-differential equations by the discontinuous Galerkin method. Math. Comput. **67**, 45–71 (1998)
11. López-Fernández, M., Lubich, C., Schädle, A.: Adaptive, fast, and oblivious convolution in evolution equations with memory. SIAM J. Sci. Comput. **30**, 1015–1037 (2008)
12. Matache, A.-M., Schwab, C., Wihler, T.: Fast numerical solution of parabolic integro-differential equations with applications in finance. SIAM J. Sci. Comput. **27**, 369–393 (2005)
13. Mustapha, K., Brunner, H., Mustapha, H., Schötzau, D.: An hp-version discontinuous Galerkin method for integro-differential equations of parabolic type. SIAM J. Numer. Anal. **49**, 1369–1396 (2011)
14. Prüss, J.: Evolutionary Integral Equations and Applications. Birkhäuser, Basel (1993)
15. Renardy, M., Hrusa, W.J., Nohel, J.: Mathematical Problems in Viscoelasticity. Wiley, New York (1987)
16. Shakourifar, M., Enright, W.H.: Reliable approximate solution of systems of Volterra integro-differential equations with time-dependent delays. SIAM J. Sci. Comput. **33**, 1134–1158 (2011)
17. Souplet, P.: Blow-up in nonlocal reaction-diffusion equations. SIAM J. Math. Anal. **29**, 1301–1334 (1998)
18. Volterra, V.: Lessons in the Mathematical Theory of the Struggle for Survival (French, reprint of the 1931 Gauthier-Villars edition). Éditions Jacques Gabay, Sceaux (1990)

Interferometric Imaging and Time Reversal in Random Media

Liliana Borcea
Department of Mathematics, University of Michigan, Ann Arbor, MI, USA

Mathematics Subject Classification

35Q60; 35Q86; 60G99; 78A48

Definition Terms

Array collection of sensors (wave sources and receivers) located close together so they behave as an entity, the array.

Imaging process of creating a map of large scale variations of the wave speed in a medium from measurements of the wave field at an array of sensors.

Random media mathematical models of heterogeneous media with uncertain microstructure.

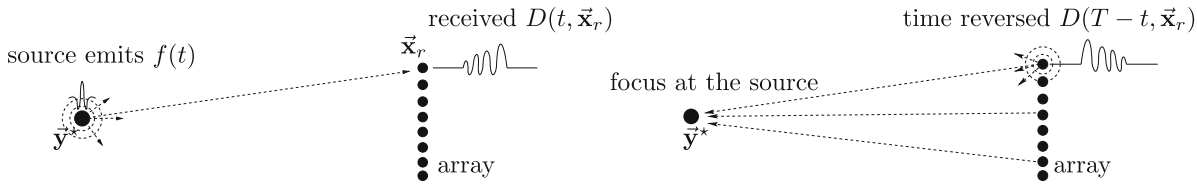
Time reversal process of reversing in time the waves measured at an array, and reemitting them in the medium where they came from, so that they can propagate an refocus at the source.

Short Description

We present a comparative study of time reversal and array imaging in random media. We explain that the time reversal process is fundamentally different than imaging, and it cannot be used for imaging purposes. We also describe briefly the resolution of time reversal and imaging. Since they occur in random media, the resolution theory is augmented with the important concept of statistical stability. It refers to robustness of the processes with respect to different realizations of the random medium.

Description

Time reversal is a physical experiment that uses special arrays of transducers, called time reversal mirrors (TRM) [12]. The transducers in a TRM operate as both receivers and sources, as illustrated in Fig. 1. First, they record the signals emitted by a remote localized source. Then, they time reverse these signals and reemit them into the medium. The waves propagate back toward the source and focus near it. In *passive array imaging*, the transducers are only receivers that record the array data, the signals from the localized source. Then, the data are processed numerically to obtain an imaging function evaluated at points $\vec{\mathbf{y}}$ in a search domain. The peaks of this function are the estimates of the source location.



Interferometric Imaging and Time Reversal in Random Media, Fig. 1 Schematic of the time reversal experiment. On the *left*, we illustrate a localized source that emits a signal $f(t)$. The transducers at locations $\vec{\mathbf{x}}_r$ in the array record the signal $D(t,\vec{\mathbf{x}}_r)$. On the *right*, we illustrate how the transducers emit the time-reversed signal, and how the waves travel back to the source, where they focus

It is often said that any imaging process involves some form of time reversal. This is true in some sense *if imaging occurs in media that are known in detail*. Then, numerical propagation of the waves in our model of the medium resembles closely the physical wave propagation in the true medium. We consider here heterogeneous media, cluttered by inhomogeneities that scatter the waves. They arise in applications like ground or foliage penetrating radar, seismic exploration, shallow water acoustics, nondestructive evaluation of heterogeneous materials like aging concrete, and so on. When imaging in clutter, we know at best the large-scale, smooth features of the medium. If we do not know them, it may be feasible to estimate them using a process called *velocity estimation* that requires additional data. See, for example, the semblance velocity estimation approach described in [10] or the travel time tomography approach [14]. However, we cannot know in detail and it is not feasible to estimate the small-scale structure of cluttered media, the inhomogeneities. That is to say, there is uncertainty about the clutter, which is why we model it as a random spatial process and speak of imaging in random media.

The time reversal experiment can be carried out without any knowledge of the medium, and surprisingly at first, clutter may improve the wave focusing at the source [12]. Time reversal requires however that we *observe the field at the time of refocus, and in the vicinity of the source*, which is of course not possible in imaging applications. That is to say, *time reversal cannot be used for imaging*. In what follows, we describe in detail the fundamental differences between time reversal and imaging in clutter, using the mathematical model of the scalar wave equation with randomly fluctuating wave speed.

Mathematical Model

The acoustic pressure $p(t,\vec{\mathbf{x}})$ solves the wave equation

$$\frac{1}{c^2(\vec{\mathbf{x}})}\frac{\partial^2 p(t,\vec{\mathbf{x}})}{\partial t^2} - \Delta p(t,\vec{\mathbf{x}}) = F(t,\vec{\mathbf{x}}),\ \vec{\mathbf{x}} \in \mathbb{R}^n, \quad t > 0, \tag{1}$$

in a medium with wave velocity $c(\vec{\mathbf{x}})$, satisfying $c(\vec{\mathbf{x}}) = c_o(\vec{\mathbf{x}})[1 + \gamma\mu(\vec{\mathbf{x}})]$. Here $c_o(\vec{\mathbf{x}})$ is the smooth, mean speed that describes the large-scales feature of the medium, and $\mu(\vec{\mathbf{x}})$ is a random function that models the inhomogeneities. We assume it to be stationary, with mean $\mathbb{E}\{\mu(\vec{\mathbf{x}})\} = 0$ and with autocorrelation $\mathcal{R}(\vec{\mathbf{x}}) = \mathbb{E}\{\mu(\vec{\mathbf{x}}' + \vec{\mathbf{x}})\mu(\vec{\mathbf{x}}')\}$ normalized by $\mathcal{R}(0) = 1$. The amplitude of the random fluctuations is modeled by the dimensionless parameter γ.

We neglect any boundaries in the problem and suppose that the waves propagate in the whole space $\mathbb{R}^n$, with $n = 2$ or 3. The medium is assumed quiescent $p(t,\vec{\mathbf{x}}) = 0$ before the source excitation, modeled by $F(t,\vec{\mathbf{x}}) = f(t)\rho(\vec{\mathbf{x}})$. Here, $f(t)$ is the emitted signal, a short pulse, and $\rho(\vec{\mathbf{x}}) \geq 0$ is the source density, compactly supported in a small ball centered at $\vec{\mathbf{y}}^\star$ and normalized to integrate to one.

The Array and System of Coordinates

There are N transducers at locations $\vec{\mathbf{x}}_r$, in a compact set $\mathcal{A}$ on an $n-1$-dimensional surface. They are closely spaced so that they behave as a collective entity, the array. In the analysis, it is usually assumed for simplicity that $\vec{\mathbf{x}}_r$ are uniformly spaced on a mesh of small size h, to allow the continuum approximation

$$h^{n-1}\sum_{r=1}^{N}\varphi(\vec{\mathbf{x}}_r) \approx \int_{\mathcal{A}} ds(\vec{\mathbf{x}})\,\varphi(\vec{\mathbf{x}}). \tag{2}$$

Here, $ds(\vec{\mathbf{x}})$ is the infinitesimal area of the surface and φ is an arbitrary integrable function. We take for

simplicity a planar array, with $\mathcal{A}$ a square of side a for $n = 3$ and $\mathcal{A}$ a line segment of length a for $n = 2$. We call a the *array aperture*.

The system of coordinates has origin at the center of the array and range axis z orthogonal to it. Then, the transducer locations are $\vec{\mathbf{x}}_r = (\mathbf{x}_r, 0)$, with cross-range $\mathbf{x}_r \in \mathcal{A}$ satisfying $|\mathbf{x}_r| \leq a/2$, for $r = 1, \ldots N$. For convenience, we assume that the center $\vec{\mathbf{y}}^\star$ of the source is on the range axis, at distance L from the array, $\vec{\mathbf{y}}^\star = (\mathbf{0}, L)$. The points $\vec{\mathbf{y}}$ in the search domain $\mathcal{Y}$, where we either observe the time-reversed field or we compute the image, are offset from $\vec{\mathbf{y}}^\star$ by $\boldsymbol{\xi}$ in cross-range and by η in range, $\vec{\mathbf{y}} = (\boldsymbol{\xi}, L + \eta)$.

Model of the Array Data

With $G(t, \vec{\mathbf{x}}, \vec{\mathbf{y}})$ the Green's function of the wave equation, we get

$$p(t, \vec{\mathbf{x}}_r) = f(t) \star_t \int_{\mathbb{R}^n} d\vec{\mathbf{y}}\, \rho(\vec{\mathbf{y}})\, G(t, \vec{\mathbf{x}}_r, \vec{\mathbf{y}}), \quad (3)$$

where $\star_t$ denotes convolution in time. Since it is easier to deal with convolutions in the frequency domain, we use the Fourier transform to write

$$\begin{aligned} p(t, \vec{\mathbf{x}}_r) &= \int_{\infty}^{\infty} \frac{d\omega}{2\pi}\, \hat{p}(\omega, \vec{\mathbf{x}}_r) e^{-i\omega t}, \\ \hat{p}(\omega, \vec{\mathbf{x}}_r) &= \hat{f}(\omega) \int_{\mathbb{R}^d} d\vec{\mathbf{y}}\, \rho(\vec{\mathbf{y}})\, \hat{G}(\omega, \vec{\mathbf{x}}_r, \vec{\mathbf{y}}), \end{aligned} \quad (4)$$

with $\hat{G}(\omega, \vec{\mathbf{x}}, \vec{\mathbf{y}})$ the outgoing Green's function of the Helmholtz equation. The source signal is modeled by

$$\begin{aligned} f(t) &= \cos(\omega_o t) f_B(t), \\ \hat{f}(\omega) &= \int_{-\infty}^{\infty} d\omega f(t) e^{i\omega t} \\ &= \frac{1}{2}\left[\hat{f}_B(\omega - \omega_o) + \hat{f}_B(\omega + \omega_o)\right], \end{aligned} \quad (5)$$

where $f_B(t)$ is a real-valued base-band pulse, with Fourier transform $\hat{f}_B(\omega)$ supported at $\omega \in [-B/2, B/2]$. We call B the *bandwidth* and ω_o the *central frequency*.

The transducers record over a time window $\chi_T(t)$ of duration T. We model it by $\chi_T(t) = T^{-1}\chi(t/T)$, with the function $\chi(u)$ of dimensionless argument u, compactly supported in the unit interval $[0, 1]$. For example, we may take $\chi(u) = 1_{[0,1]}(u)$, the indicator function equal to one when $u \in [0, 1]$ and zero otherwise.

The model of the array data is $D(t, \vec{\mathbf{x}}_r) = \chi_T(t) p(t, \vec{\mathbf{x}}_r)$, for $r = 1, \ldots, N$, with Fourier transform

$$\begin{aligned} \hat{D}(\omega, \vec{\mathbf{x}}_r) &= \int_{-\infty}^{\infty} d\omega' \, \frac{\hat{\chi}\,[(\omega - \omega')T]}{2\pi} \hat{p}(\omega', \vec{\mathbf{x}}_r) \\ &= \int_{-\infty}^{\infty} d\omega' \, \frac{\hat{\chi}\,[(\omega - \omega')T]}{2\pi} \hat{f}(\omega') \\ &\quad \int_{\mathbb{R}^n} d\vec{\mathbf{y}}\, \rho(\vec{\mathbf{y}})\, \hat{G}(\omega', \vec{\mathbf{x}}_r, \vec{\mathbf{y}}). \end{aligned} \quad (6)$$

We often call the signals $D(t, \vec{\mathbf{x}}_r)$ *data time traces*, to emphasize that they are functions of time.

Model of the Time Reversal Function

Each transducer in the array reverses the received signal

$$\begin{aligned} F(t, \vec{\mathbf{x}}_r) &= D(T - t, \vec{\mathbf{x}}_r), \\ \hat{F}(\omega, \vec{\mathbf{x}}_r) &= \int_{-\infty}^{\infty} dt\, e^{i\omega t} D(T - t, \vec{\mathbf{x}}_r) = \overline{\hat{D}(\omega, \vec{\mathbf{x}}_r)} e^{i\omega T}, \end{aligned} \quad (7)$$

and reemits it in the medium. The acoustic pressure observed at points $\vec{\mathbf{y}} \in \mathcal{Y}$ is

$$\begin{aligned} p^{\mathrm{TR}}(t, \vec{\mathbf{y}}) &= \int_{-\infty}^{\infty} \frac{d\omega}{2\pi}\, e^{-i\omega t} \sum_{r=1}^{N} \hat{F}(\omega, \vec{\mathbf{x}}_r) \hat{G}(\omega, \vec{\mathbf{x}}_r, \vec{\mathbf{y}}) \\ &= \int_{-\infty}^{\infty} \frac{d\omega}{2\pi}\, e^{i\omega(T-t)} \sum_{r=1}^{N} \overline{\hat{D}(\omega, \vec{\mathbf{x}}_r)} \hat{G}(\omega, \vec{\mathbf{x}}_r, \vec{\mathbf{y}}), \end{aligned} \quad (8)$$

where the bar denotes complex conjugate. It is expected to focus back at the source, at time $t = T$, so we define the time reversal function

$$\begin{aligned} \mathcal{J}^{\mathrm{TR}}_{\rho,\chi}(\vec{\mathbf{y}}) &= p^{\mathrm{TR}}(t = T, \vec{\mathbf{y}}) \\ &= \int_{-\infty}^{\infty} \frac{d\omega}{2\pi} \sum_{r=1}^{N} \overline{\hat{D}(\omega, \vec{\mathbf{x}}_r)} \hat{G}(\omega, \vec{\mathbf{x}}_r, \vec{\mathbf{y}}). \end{aligned} \quad (9)$$

The indexes ρ, χ indicate its dependence on the source density ρ and the recording window χ. In the analysis, it is usual to assume an ideal point

source $\rho(\vec{\mathbf{y}}) = \delta(\vec{\mathbf{y}} - \vec{\mathbf{y}}^\star)$ and an infinite time window $\hat{\chi}(\omega T) = 2\pi\delta(\omega)$, where $\delta(\cdot)$ is the Dirac delta distribution. It is also usual to make the continuum array aperture approximation (2) and forget the scaling factor h^{n-1}. The time reversal function becomes, under these simplifications,

$$\mathcal{J}^{\mathrm{TR}}(\vec{\mathbf{y}}) = \int_{-\infty}^{\infty} \frac{d\omega}{2\pi} \overline{\hat{f}(\omega)} \int_{\mathcal{A}} d\mathbf{x}\, \overline{\hat{G}(\omega, \vec{\mathbf{x}}, \vec{\mathbf{y}}^\star)} \hat{G}(\omega, \vec{\mathbf{x}}, \vec{\mathbf{y}}), \quad \vec{\mathbf{x}} = (\mathbf{x}, 0). \tag{10}$$

Reverse Time Migration and the Least Squares Approach to Imaging

The least squares estimate $\rho^{\mathrm{LS}}(\vec{\mathbf{x}})$ of the source density is the minimizer of the array data misfit

$$\min_{\rho \in L^2(\mathbb{R}^n)} \mathcal{O}(\rho), \quad \mathcal{O}(\rho) = \langle \mathcal{M}\rho - D, \mathcal{M}\rho - D \rangle = \int_{-\infty}^{\infty} \frac{d\omega}{2\pi} \sum_{r=1}^{N} \left| [\mathcal{M}\rho](\omega, \vec{\mathbf{x}}_r) - \hat{D}(\omega, \vec{\mathbf{x}}_r) \right|^2. \tag{11}$$

Here, we assume a square integrable ρ, and let $\mathcal{M}$ be the forward operator that takes ρ to the Hilbert space of the data, with inner product denoted by $\langle \cdot, \cdot \rangle$. We have, similar to (6),

$$[\mathcal{M}\rho](\omega, \vec{\mathbf{x}}_r) = \int_{-\infty}^{\infty} d\omega' \, \frac{\hat{\chi}\left[(\omega - \omega')T\right]}{2\pi} \hat{f}(\omega') \int_{\mathbb{R}^n} d\vec{\mathbf{y}}\, \rho(\vec{\mathbf{y}})\, \hat{G}_o(\omega', \vec{\mathbf{x}}_r, \vec{\mathbf{y}}), \tag{12}$$

where $\hat{G}_o$ is the outgoing Green's function of the Helmholtz equation in the medium with wave speed $c_o(\vec{\mathbf{x}})$, our estimate of the true wave speed $c(\vec{\mathbf{x}})$. We assume henceforth, for simplicity, $\hat{\chi}(\omega T) = 2\pi\delta(\omega)$.

The least squares solution solves the normal equations $[\mathcal{M}^\star \mathcal{M} \rho^{\mathrm{LS}}](\vec{\mathbf{y}}) = [\mathcal{M}^\star D](\vec{\mathbf{y}})$, where $\mathcal{M}^\star$ is the adjoint operator that takes the data to the Hilbert space $L^2(\mathbb{R}^n)$,

$$[\mathcal{M}^\star D](\vec{\mathbf{y}}) = \int_{-\infty}^{\infty} \frac{d\omega}{2\pi}\, \hat{f}(\omega) \sum_{r=1}^{N} \overline{\hat{D}(\omega, \vec{\mathbf{x}}_r)} \hat{G}_o(\omega, \vec{\mathbf{x}}_r, \vec{\mathbf{y}}). \tag{13}$$

The normal operator $\mathcal{M}^\star\mathcal{M} : L^2(\mathbb{R}^n) \to L^2(\mathbb{R}^n)$ is given by $[\mathcal{M}^\star\mathcal{M}\rho](\vec{\mathbf{y}}) = \int_{\mathbb{R}^n} d\vec{\mathbf{y}}'\, \rho(\vec{\mathbf{y}}') \mathcal{K}(\vec{\mathbf{y}}, \vec{\mathbf{y}}')$, with kernel

$$\mathcal{K}(\vec{\mathbf{y}}, \vec{\mathbf{y}}') = \int_{-\infty}^{\infty} \frac{d\omega}{2\pi} \left| \hat{f}(\omega) \right|^2 \sum_{r=1}^{N} \overline{\hat{G}_o(\omega, \vec{\mathbf{x}}_r, \vec{\mathbf{y}}')} \hat{G}_o(\omega, \vec{\mathbf{x}}_r, \vec{\mathbf{y}}). \tag{14}$$

Note that $\mathcal{K}(\vec{\mathbf{y}}, \vec{\mathbf{y}}')$ is the time reversal function for a point source at $\vec{\mathbf{y}}'$ that emits a signal with Fourier transform $|\hat{f}(\omega)|^2$, in the *smooth, fictitious medium* with wave speed $c_o(\vec{\mathbf{x}})$. It peaks at $\vec{\mathbf{y}} = \vec{\mathbf{y}}'$, and it is large in a vicinity of $\vec{\mathbf{y}}'$, as described by the resolution limits given in section "Resolution and Robustness of Time Reversal and Imaging in Random Media." This implies that the right-hand side in the normal equations is large around the support of $\rho^{\mathrm{LS}}(\vec{\mathbf{x}})$, and thus, it defines an imaging function

$$\mathcal{J}^{\mathrm{M}}(\vec{\mathbf{y}}) = [\mathcal{M}^\star D](\vec{\mathbf{y}}) = \int_{-\infty}^{\infty} \frac{d\omega}{2\pi}\, \hat{f}(\omega) \sum_{r=1}^{N} \overline{\hat{D}(\omega, \vec{\mathbf{x}}_r)} \hat{G}_o(\omega, \vec{\mathbf{x}}_r, \vec{\mathbf{y}}), \tag{15}$$

known as *reverse time migration*. Often, the factor $\hat{f}(\omega)$ is neglected, because it does not play a big role when the signal $f(t)$ is a pulse. However, for long signals like chirps [11, Section 3.1.2], the factor is important. Explicitly, the convolution of $f(t)$ with $f(-t)$ compresses these signals as if the sources emitted a pulse. The Fourier transform of $f(-t) \star_t f(t)$ is $|\hat{f}(\omega)|^2$, as it appears in (14).

Reverse time migration is common in geophysics [2], radar [11], and elsewhere, but most often it is replaced by its simplified version known as *Kirchhoff migration*

$$\mathcal{J}^{\mathrm{KM}}(\vec{\mathbf{y}}) = \int_{-\infty}^{\infty} \frac{d\omega}{2\pi} \sum_{r=1}^{N} \hat{D}(\omega, \vec{\mathbf{x}}_r) e^{-i\omega\tau(\vec{\mathbf{x}}_r, \vec{\mathbf{y}})} = \sum_{r=1}^{N} D(\tau(\vec{\mathbf{x}}_r, \vec{\mathbf{y}}), \vec{\mathbf{x}}_r). \tag{16}$$

The simplification uses the high frequency, geometrical optics approximation of the Green's function

$$\hat{G}_o(\omega, \vec{\mathbf{x}}, \vec{\mathbf{y}}) \approx \alpha(\omega_o, L) e^{i\omega\tau(\vec{\mathbf{x}}, \vec{\mathbf{y}})}, \tag{17}$$

with approximately constant amplitude α, under the assumptions $|\xi|, a, \eta \ll L$. The travel time $\tau(\vec{\mathbf{x}}, \vec{\mathbf{y}})$ is given by Fermat's principle, $\tau(\vec{\mathbf{x}}, \vec{\mathbf{y}}) = \min \int dl\, c^{-1}(\vec{\mathbf{r}}(l))$, where the minimum is over all paths $\vec{\mathbf{r}}(l)$ parametrized by $l \in \mathbb{R}$ that start at $\vec{\mathbf{y}}$ and end at $\vec{\mathbf{x}}$.

Coherent Interferometric Imaging

Equations (15) and (16) show how migration forms images by superposing the data traces $D(t, \vec{\mathbf{x}}_r)$ back-propagated to $\vec{\mathbf{y}} \in \mathcal{Y}$, either with the Green's function $\hat{G}_o$ or with the travel time τ. The *coherent interferometric* (CINT) imaging approach introduced in [6, 7] back-propagates to $\vec{\mathbf{y}} \in \mathcal{Y}$ local cross-correlations of the data traces at nearby receivers, instead of the traces themselves. The *local cross-correlations* are defined by

$$\mathcal{C}(t, \Delta t, \vec{\mathbf{x}}_r, \vec{\mathbf{x}}_{r'}; T_c) = \int_{\infty}^{\infty} dt'\, \phi_c(t') D\left(t + \frac{\Delta t}{2} - t', \vec{\mathbf{x}}_r\right) D\left(t - \frac{\Delta t}{2} - t', \vec{\mathbf{x}}_{r'}\right)$$
$$= \int_{-\infty}^{\infty} \frac{d\omega}{2\pi} e^{-i\omega\Delta t} \int_{-\infty}^{\infty} \frac{d\tilde{\omega}}{2\pi} e^{-i\tilde{\omega}t}\, \hat{\phi}(\tilde{\omega} T_c)\, \hat{D}\left(\omega + \frac{\tilde{\omega}}{2}, \vec{\mathbf{x}}_r\right) \overline{\hat{D}\left(\omega - \frac{\tilde{\omega}}{2}, \vec{\mathbf{x}}_{r'}\right)}. \quad (18)$$

They are computed over a time window $\phi_c(t)$ of width T_c, modeled by $\phi_c(t) = T_c^{-1} \phi\,(t/T_c)$, using the function $\phi(u)$ of dimensionless argument u, and compactly supported at $|u| \le 1/2$.

Let us assume for simplicity that the high frequency, geometrical optics approximation (17) applies. The mathematical model of the CINT imaging function is

$$\mathcal{J}^{\mathrm{CINT}}(\vec{\mathbf{y}}; T_c, X_c) = \int_{-\infty}^{\infty} \frac{d\omega}{2\pi} \int_{-\infty}^{\infty} \frac{d\tilde{\omega}}{2\pi}\, \hat{\phi}(\tilde{\omega} T_c) \sum_{r,r'=1}^{N} \psi\left(\frac{|\vec{\mathbf{x}}_r - \vec{\mathbf{x}}_{r'}|}{X_c(\omega)}\right) \hat{D}\left(\omega + \frac{\tilde{\omega}}{2}, \vec{\mathbf{x}}_r\right) \overline{\hat{D}\left(\omega - \frac{\tilde{\omega}}{2}, \vec{\mathbf{x}}_{r'}\right)} \times$$
$$\exp\left\{-i\omega\left[\tau(\vec{\mathbf{x}}_r, \vec{\mathbf{y}}) - \tau(\vec{\mathbf{x}}_{r'}, \vec{\mathbf{y}})\right] - i\tilde{\omega}\, \frac{\left[\tau(\vec{\mathbf{x}}_r, \vec{\mathbf{y}}) + \tau(\vec{\mathbf{x}}_{r'}, \vec{\mathbf{y}})\right]}{2}\right\}. \quad (19)$$

Note how it superposes the local cross-correlations (18) back-propagated to $\vec{\mathbf{y}}$ by evaluating them at the mean travel time $t = \left[\tau(\vec{\mathbf{x}}_r, \vec{\mathbf{y}}) + \tau(\vec{\mathbf{x}}_{r'}, \vec{\mathbf{y}})\right]/2$, and at the difference travel time $\Delta t = \tau(\vec{\mathbf{x}}_r, \vec{\mathbf{y}}) - \tau(\vec{\mathbf{x}}_{r'}, \vec{\mathbf{y}})$. Note also that we introduced another window function $\psi(u)$, supported at $|u| \le 1/2$. Its purpose is to restrict the superposition in (19) to the receivers that are not further than the distance $X_c(\omega)$ apart. In general, this distance may vary in the bandwidth.

Resolution and Robustness of Time Reversal and Imaging in Random Media

The performance of the time reversal and imaging processes is assessed by their *resolution* and *robustness*. The *resolution* quantifies the ability of the process to distinguish between two localized sources. We analyze it by estimating the support of the point-spread function, the model of the process for a point-like source. The models derived above are random, because the waves travel from the source to the array in a random medium. Therefore, we quantify the resolution using the mean (statistical expectation) of the models.

A *robust* process gives a high signal-to-noise (SNR) ratio. Recall that we look for the peaks of the random functions that model time reversal and imaging. By high SNR, we mean that these peaks are insensitive to the noise and are clearly distinguishable. Usually, one considers additive, uncorrelated, instrument noise in the data. Here, we consider *clutter noise* due to scattering of the waves in the medium. It is not additive, it has a complex structure, it exhibits correlations across the array and over frequencies, and it is much harder

to mitigate than instrument noise. The high SNR of imaging (or time reversal) in random media means that the random fluctuation of the images (or wave field) induced by the clutter noise is small, and therefore, the results are insensitive to the particular realization of the random medium. Such robustness is called *statistical stability* [3, 8, 13], and it is an essential quality of any useful method in random media.

Resolution

For simplicity, we use the continuum array approximation (2) and assume that the background medium is homogeneous, with constant wave speed c_o. The Green's function is approximated by (17), with $\tau(\vec{\mathbf{x}}, \vec{\mathbf{y}}) = |\vec{\mathbf{x}} - \vec{\mathbf{y}}|/c_o$. We have a point source at $\vec{\mathbf{y}}^\star$.

To quantify the resolution, we estimate the support of the mean point spread functions oftime reversal, KM and CINT. We need the first and second statistical moments of the random Greens' function $\hat{G}(\omega, \vec{\mathbf{x}}, \vec{\mathbf{y}})$. The details of the calculation of these moments depend on the particular model of the fluctuations $\mu(\vec{\mathbf{x}})$. For *mixing, isotropic* fluctuations, that is, fluctuations with integrable correlation function $\mathbb{R}(\vec{\mathbf{x}}) = \mathbb{R}(|\vec{\mathbf{x}}|)$, the moments have the generic form

$$\begin{aligned}\mathbb{E}\left\{\hat{G}(\omega, \vec{\mathbf{x}}, \vec{\mathbf{y}}^\star)\right\} &\approx \hat{G}_o(\omega, \vec{\mathbf{x}}, \vec{\mathbf{y}}^\star) \exp\left[-\frac{\omega^2}{2\Omega_d^2}\right] \\ &\approx \alpha(\omega_o, L) \exp\left[i\omega\tau(\vec{\mathbf{x}}, \vec{\mathbf{y}}^\star) - \frac{\omega^2}{2\Omega_d^2}\right],\end{aligned} \tag{20}$$

$$\begin{aligned}\mathbb{E}\left\{\hat{G}\left(\omega + \frac{\tilde{\omega}}{2}, \left(\mathbf{x} + \frac{\tilde{\mathbf{x}}}{2}, 0\right), \vec{\mathbf{y}}^\star\right) \overline{\hat{G}\left(\omega - \frac{\tilde{\omega}}{2}, \left(\mathbf{x} - \frac{\tilde{\mathbf{x}}}{2}, 0\right), \vec{\mathbf{y}}^\star\right)}\right\} &\approx |\alpha(\omega_o, L)|^2 \times \\ \exp\left[i\omega\Delta\tau(\mathbf{x}, \tilde{\mathbf{x}}, \vec{\mathbf{y}}^\star) + i\tilde{\omega}\bar{\tau}(\mathbf{x}, \tilde{\mathbf{x}}, \vec{\mathbf{y}}^\star) - \frac{\tilde{\omega}^2}{2\Omega_d^2} - \frac{|\tilde{\mathbf{x}}|^2}{2X_d^2(\omega)}\right],\end{aligned} \tag{21}$$

where we let

$$\bar{\tau}(\mathbf{x}, \tilde{\mathbf{x}}, \vec{\mathbf{y}}) = \frac{\tau\left[\left(\mathbf{x} + \frac{\tilde{\mathbf{x}}}{2}, 0\right), \vec{\mathbf{y}}\right] + \tau\left[\left(\mathbf{x} - \frac{\tilde{\mathbf{x}}}{2}, 0\right)), \vec{\mathbf{y}}\right]}{2},$$

$$\Delta\tau(\mathbf{x}, \tilde{\mathbf{x}}, \vec{\mathbf{y}}) = \tau\left[\left(\mathbf{x} + \frac{\tilde{\mathbf{x}}}{2}, 0\right), \vec{\mathbf{y}}\right] - \tau\left[\left(\mathbf{x} - \frac{\tilde{\mathbf{x}}}{2}, 0\right), \vec{\mathbf{y}}\right].$$

We refer the reader to [6, Appendix B] for the derivation of these formulas in the random paraxial (forward scattering) regime, in which $\lambda_o \ll a \ll L$ and the random fluctuations are small $\gamma \ll 1$, with correlation length ℓ (typical size of the inhomogeneities) satisfying $\ell \ll L$. See also [5, Lemma 3.2] for the derivation of the same moment formulas, under a much simpler model of the random fluctuations that gives only random wave front distortions.

The first moment formula (20) says that the mean field is exponentially damped. There is no absorption in our model. The damping means that the wave field loses coherence because of scattering in the medium, and the incoherent field $\hat{G} - E\{\hat{G}\}$ becomes the dominant part of $\hat{G}$. The second moment formula (22) says that the wave fields are statistically correlated over frequency offsets satisfying $|\tilde{\omega}| \lesssim \Omega_d$ and over transducer offsets $\tilde{\mathbf{x}}$ satisfying $|\tilde{\mathbf{x}}| \lesssim X_d(\omega)$. We call Ω_d the *decoherence frequency* and $X_d(\omega)$ the *decoherence length*. Their precise expressions are model dependent, but they are in general determined by the correlation function $\mathbb{R}(|\vec{\mathbf{x}}|)$, and they decrease with range L. The decoherence length is also proportional to the wavelength $\lambda = 2\pi c_o/\omega$, and we write it in the form $X_d(\omega) = \frac{\lambda L}{a_e(L)}$, with $a_e(L)$ having units of length and increasing with range. It is called in [3, 6] the *effective aperture* for the reasons explained below.

The resolution study is simpler in the Fraunhofer diffraction regime [9], where $a \ll L$ and the Fresnel number $a^2/(\lambda L)$ is small. It allows us to linearize phases in the models of time reversal and imaging and obtain simpler expressions that can be interpreted as decompositions in plane waves.

Cross-Range Resolution

Consider search points $\vec{\mathbf{y}}$ that are offset from the source location only in cross-range: $\vec{\mathbf{y}} = (\boldsymbol{\xi}, L)$. The expectation of the time reversal function is

$$\mathbb{E}\{\mathcal{J}^{\mathrm{TR}}(\boldsymbol{\xi},L)\} \approx |\alpha(\omega_o)L|^2 \int_{-\infty}^{\infty} \overline{\hat{f}(\omega)} \int_{\mathcal{A}} d\mathbf{x}\, \exp\left\{ i\omega\left[\tau(\vec{\mathbf{x}},\vec{\mathbf{y}}) - \tau(\vec{\mathbf{x}},\vec{\mathbf{y}}^\star)\right] - \frac{|\boldsymbol{\xi}|^2}{2X_d^2(\omega)} \right\} \tag{22}$$

and we obtain with the approximation $\tau(\vec{\mathbf{x}},\vec{\mathbf{y}}) - \tau(\vec{\mathbf{x}},\vec{\mathbf{y}}^\star) \approx \boldsymbol{\xi}\cdot\nabla_{\mathbf{y}}\tau(\vec{\mathbf{x}},\vec{\mathbf{y}}^\star) \approx -\frac{\boldsymbol{\xi}\cdot\mathbf{x}}{c_o L}$ that

$$\mathbb{E}\{\mathcal{J}^{\mathrm{TR}}(\boldsymbol{\xi},L)\} \approx |\alpha(\omega_o)L|^2 a^{n-1} \int_{-\infty}^{\infty} \overline{\hat{f}(\omega)} e^{-\frac{|\boldsymbol{\xi}|^2}{2X_d^2(\omega)}} \prod_{j=1}^{n-1} \mathrm{sinc}\left(\frac{\pi a \xi_j}{\lambda L}\right). \tag{23}$$

Here, ξ_j are the components of vector $\boldsymbol{\xi}$ and $\mathrm{sinc}(u) = \sin(u)/u$. The expectation of the KM function is

$$\mathbb{E}\{\mathcal{J}^{\mathrm{KM}}(\boldsymbol{\xi},L)\} = \int_{-\infty}^{\infty} \frac{d\omega}{2\pi}\, \hat{f}(\omega) \int_{\mathcal{A}} d\mathbf{x}\, \mathbb{E}\{\hat{G}(\omega,\vec{\mathbf{x}},\vec{\mathbf{y}}^\star)\} e^{-i\omega\tau(\vec{\mathbf{x}},\vec{\mathbf{y}})}$$
$$\approx \alpha(\omega_o,L) a^{n-1} \int_{-\infty}^{\infty} \hat{f}(\omega) e^{-\frac{\omega^2}{2\Omega_d^2}} \prod_{j=1}^{n-1} \mathrm{sinc}\left(\frac{\pi a \xi_j}{\lambda L}\right). \tag{24}$$

The expectation of the CINT function is more complicated

$$\mathbb{E}\{\mathcal{J}^{\mathrm{CINT}}(\boldsymbol{\xi},L)\} \approx |\alpha(\omega_o,L)|^2 \int_{-\infty}^{\infty} \frac{d\omega}{2\pi} \int_{-\infty}^{\infty} \frac{d\tilde{\omega}}{2\pi} \overline{\hat{f}\left(\omega+\frac{\tilde{\omega}}{2}\right)} \hat{f}\left(\omega-\frac{\tilde{\omega}}{2}\right) \hat{\phi}(\tilde{\omega}T_c) \int_{\mathcal{A}} d\mathbf{x} \int_{\mathbb{R}^{n-1}} d\tilde{\mathbf{x}}\, \psi\left(\frac{|\tilde{\mathbf{x}}|}{X_c(\omega)}\right) \times$$
$$\exp\left\{ i\tilde{\omega}\left[\bar{\tau}(\mathbf{x},\tilde{\mathbf{x}},\vec{\mathbf{y}}^\star) - \bar{\tau}(\mathbf{x},\tilde{\mathbf{x}},\vec{\mathbf{y}})\right] + i\omega\left[\Delta\tau(\mathbf{x},\tilde{\mathbf{x}},\vec{\mathbf{y}}^\star) - \Delta\tau(\mathbf{x},\tilde{\mathbf{x}},\vec{\mathbf{y}})\right] - \frac{\tilde{\omega}^2}{2\Omega_d^2} - \frac{|\tilde{\mathbf{x}}|^2}{2X_d^2(\omega)} \right\}. \tag{25}$$

We can simplify it by assuming:

1. A small X_d (i.e., a small $|\tilde{\mathbf{x}}|$), so that $\bar{\tau}(\mathbf{x},\tilde{\mathbf{x}},\vec{\mathbf{y}}) \approx \tau(\vec{\mathbf{x}},\vec{\mathbf{y}})$ and $\Delta\tau(\mathbf{x},\tilde{\mathbf{x}},\vec{\mathbf{y}}^\star) \approx \tilde{\mathbf{x}}\cdot\nabla_{\mathbf{x}}\tau(\vec{\mathbf{x}},\vec{\mathbf{y}}) \approx \frac{\tilde{\mathbf{x}}\cdot(\mathbf{x}-\boldsymbol{\xi})}{L}$.
2. A small Ω_d (i.e., a small $|\tilde{\omega}|$) and a smooth pulse, so that $\hat{f}(\omega \pm \tilde{\omega}/2) \approx \hat{f}(\omega)$.
3. The windows $\hat{\phi}(\tilde{\omega}T_c)$ and $\psi(|\tilde{\mathbf{x}}|/X_c)$ are one in the essential support of the Gaussians in $\tilde{\omega}$ and $\tilde{\mathbf{x}}$ in (25) and zero outside. We obtain after some straightforward calculations

$$\mathbb{E}\{\mathcal{J}^{\mathrm{CINT}}(\boldsymbol{\xi},L)\} \approx (2\pi)^{\frac{n}{2}-1} \Omega_d\, |\alpha(\omega_o,L)|^2 \left[\frac{a\,L}{a_e(L)}\right]^{n-1} \int_{-\infty}^{\infty} \frac{d\omega}{2\pi} |\hat{f}(\omega)|^2 \lambda^{n-1} \exp\left[-\frac{2\pi^2|\boldsymbol{\xi}|^2}{a_e^2(L)}\right]. \tag{26}$$

Conclusions Equations (23)–(26) show that the mean time reversal and imaging functions peak at the true source location, i.e., at $\boldsymbol{\xi} = \mathbf{0}$. However, they have different resolution. The resolution of KM is the same as that in the homogeneous medium. It is defined as the distance between the peak of the sinc function and its first zero, and it is given by the Rayleigh resolution formula [9]

$$\frac{\lambda_o L}{a}\left[1 + O\left(\frac{B}{\omega_o}\right)\right] \approx \frac{\lambda_o L}{a}, \qquad \text{if } B \ll \omega_o. \tag{27}$$

The resolution of time reversal is *better*, assuming that $a_e(L) > a$,

$$|\boldsymbol{\xi}| \lesssim X_d(\omega) = \frac{\lambda_o L}{a_e(L)}\left[1 + O\left(\frac{B}{\omega_o}\right)\right] \approx \frac{\lambda_o L}{a_e(L)}. \tag{28}$$

This happens when the cumulative wave scattering in the random medium is strong and causes the waves to decorrelate over small distances X_d. The improved cross-range focusing is called *super-resolution*. It was discovered and demonstrated experimentally in [12] and has been explained theoretically in terms of the enhanced effective aperture $a_e(L)$ in [3, 6, 13]. The resolution of CINT is proportional to the effective aperture $|\boldsymbol{\xi}| \lesssim \frac{a_e(L)}{2\pi}$, and thus, it *deteriorates as wave scattering becomes stronger*.

Range Resolution

When the search points $\vec{\mathbf{y}} = (\mathbf{0}, L + \eta)$ are offset only in range from $\vec{\mathbf{y}}^\star$,

$$\mathbb{E}\{\mathcal{J}^{\mathrm{TR}}(\mathbf{0}, L+\eta)\} \approx |\alpha(\omega_o, L)|^2 a^{n-1} \int_{-\infty}^{\infty} \frac{d\omega}{2\pi} \overline{\hat{f}(\omega)} \exp\left(-\frac{\omega^2}{2\Omega_d^2}\frac{\eta}{L} + i\frac{\omega}{c_o}\eta\right). \quad (29)$$

Here, we used the moment formula [6, Appendix B]

$$\mathbb{E}\left\{\overline{\hat{G}\left(\omega, \vec{\mathbf{x}}, \vec{\mathbf{y}}^\star\right)}\hat{G}\left(\omega, \vec{\mathbf{x}}, \vec{\mathbf{y}}\right)\right\} \approx \overline{\hat{G}_o\left(\omega, \vec{\mathbf{x}}, \vec{\mathbf{y}}^\star\right)}\hat{G}_o\left(\omega, \vec{\mathbf{x}}, \vec{\mathbf{y}}\right) \exp\left(-\frac{\omega^2}{2\Omega_d^2}\frac{\eta}{L}\right), \quad (30)$$

and the approximation $\tau(\vec{\mathbf{x}}, \vec{\mathbf{y}}) - \tau(\vec{\mathbf{x}}, \vec{\mathbf{y}}^\star) \approx \eta/c_o$. For the KM function, we get

$$\mathbb{E}\{\mathcal{J}^{\mathrm{KM}}(\mathbf{0}, L+\eta)\} \approx \alpha(\omega_o, L) a^{n-1} \int_{-\infty}^{\infty} \frac{d\omega}{2\pi} \hat{f}(\omega) \exp\left(-\frac{\omega^2}{2\Omega_d^2} - i\frac{\omega}{c_o}\eta\right), \quad (31)$$

and for CINT,

$$\mathbb{E}\{\mathcal{J}^{\mathrm{CINT}}(\mathbf{0}, L+\eta)\} \approx (2\pi)^{\frac{n}{2}-1}\Omega_d |\alpha(\omega_o, L)|^2 \left[\frac{aL}{a_e(L)}\right]^{n-1} \int_{-\infty}^{\infty} \frac{d\omega}{2\pi} |\hat{f}(\omega)|^2 \lambda^{n-1} \exp\left[-\frac{\eta^2}{2(c_o/\Omega_d)^2}\right]. \quad (32)$$

Conclusions All the mean functions peak at the source location, where $\eta = 0$, but they have different resolution. The range resolution of time reversal is

$$|\eta| \lesssim \min\left\{\frac{c_o}{B}, L\left(\frac{\Omega_d}{\omega_o}\right)^2\right\}. \quad (33)$$

In most regimes, it is comparable to that of the mean KM function, $|\eta| \lesssim c_o/B$, determined by the pulse bandwidth. However, the range resolution of CINT is worse in random media, where cumulative wave scattering causes wave decorrelation over frequency offsets $\Omega_d < B$. We have $|\eta| \lesssim c_o/\Omega_d$.

Statistical Stability

There is another fundamental difference between time reversal, KM, and CINT imaging. Note how the mean KM function at the peak is exponentially damped because of the factor $\exp[-\omega^2/(2\Omega_d^2)]$. In random media, where $\Omega_d \ll \omega_o$, this is typically almost zero. The magnitude of the random fluctuations of $\mathcal{J}^{\mathrm{KM}}(\vec{\mathbf{y}})$ are determined by its standard deviation $\sigma^{\mathrm{KM}}(\vec{\mathbf{y}})$. Its calculation involves the second moments (22) of the Green's function, and it is similar to that of computing $\mathbb{E}\{\mathcal{J}^{\mathrm{CINT}}\}$. The SNR is the ratio $\mathbb{E}\{\mathcal{J}^{\mathrm{KM}}(\vec{\mathbf{y}}^\star)\}/\sigma^{\mathrm{KM}}(\vec{\mathbf{y}}^\star)$. It is exponentially small, of the order $\exp[-\omega_o^2/(2\Omega_d^2)]$, no matter how large the array aperture is. If we had uncorrelated, additive noise, the SNR would improve for larger arrays, because the noise would be averaged out by the superposition over the many sensors. The random medium noise is much more complex, and in general, it cannot be removed by simply increasing the array aperture. The KM method is not useful in imaging in random media, because the signal, the value of the function at the expected peak $\vec{\mathbf{y}}^\star$, is faint and not distinguishable from the noise, the random fluctuations of the image.

The mean time reversal and CINT functions are not exponentially damped as KM is. This is key to their robustness. Examples of proofs of the statistical stability of time reversal and CINT imaging are in [13] and in [8], respectively. They assume a paraxial, forward scattering regime, and certain asymptotic limits, and show that $\mathcal{J}^{\mathrm{TR}}(\vec{\mathbf{y}})$ and $\mathcal{J}^{\mathrm{CINT}}(\vec{\mathbf{y}})$ converge in probability to a deterministic limit. A more quantitative statistical stability study requires the calculation of the SNR, which is much more difficult than for KM, because it involves fourth-order moments of the Green's function. The SNR of CINT has been calculated only recently in [5], for a simple model of the random medium that gives only random wave front distortions, but does not account for multiple wave scattering. The result in [5] shows that the SNR of CINT is large and it can be improved by increasing the array aperture.

Note that statistical stability of time reversal typically holds only in broadband [3]. The stability of CINT is also in broadband and subject to choosing the proper time and transducer offset thresholds T_c and X_c in (19). In section "Resolution and Robustness of Time Reversal and Imaging in Random Media," we made the optimal choice with thresholds given by the decoherence frequency and length, $1/T_c = \Omega_d$ and $X_c = X_d$. If we chose $1/T_c > \Omega_d$ and $X_c > X_d$

instead, the resolution analysis would have stayed the same, but the stability result would not hold. It turns out the thresholding by T_c and X_c has a statistical smoothing effect [8] and it is essential for a robust CINT imaging process. The smoothing comes at the expense of loss of resolution. If we chose $1/T_c < \Omega_d$ and $X_c < X_d$, the resolution of CINT would be worse, by a factor X_d/X_c in cross-range and $T_c\Omega_d$ in range. This trade-off between resolution and stability in CINT can be used to determine the optimal thresholding parameters T_c, X_c, without apriori knowledge about the statistics of the medium, that is, about Ω_d and X_d. This is the idea of the adaptive CINT algorithm introduced and studied in [7].

Summary

We have described the fundamental differences between the time reversal process and imaging in random media. Wave scattering may lead to *super-resolution* of time reversal [12], but this is not useful in imaging. Traditional imaging methods, like reverse time migration cannot be used for robust imaging in random media. Coherent interferometry can give robust results, but its resolution deteriorates as the cumulative wave scattering effects increase. CINT by itself will not work in strong scattering media, but in some cases, it can be complemented with additional data preprocessing designed to filter out clutter effects [1,4]. We discussed only imaging with passive arrays, because it is the natural setting for comparison with time reversal. We refer to [5,7] for studies of CINT imaging of scatterers with active arrays.

References

1. Alonso, R., Borcea, L., Papanicolaou, G., Tsogka, C.: Detection and imaging in strongly backscattering randomly layered media. Probl. **27**, 025004 (2011)
2. Biondi, B.: 3D seismic imaging. Society of Exploration Geophysicists, Tulsa (2006)
3. Blomgren, P., Papanicolaou, G., Zhao, H.: Super-resolution in time-reversal acoustics. J. Acoust. Soc. Am. **111**, 230 (2002)
4. Borcea, L., del Cueto, F., Papanicolaou, G., Tsogka, C.: Filtering random layering effects in imaging. SIAM Multiscale Model. Simul. **8**, 751–781 (2010)
5. Borcea, L., Garnier, J., Papanicolaou, G., Tsogka, C.: Enhanced statistical stability in coherent interferometric imaging. Inverse Probl. Inverse Problems, 27(8), 2011, p. 085003.
6. Borcea, L., Papanicolaou, G., Tsogka, C.: Interferometric array imaging in clutter. Inverse Probl. **21**, 1419–1460 (2005)
7. Borcea, L., Papanicolaou, G., Tsogka, C.: Adaptive interferometric imaging in clutter and optimal illumination. Inverse Probl. **22**, 1405–1436 (2006)
8. Borcea, L., Papanicolaou, G., Tsogka, C.: Asymptotics for the space-time Wigner transform with applications to imaging. In: Rozovskii, B.L., Baxendale, P.H., Lototsky S.V. (eds.) Stochastic Differential Equations: Theory and Applications. Interdisciplinary Mathematical Sciences, vol. 2. World Scientific, Singapore/Hackensack (2007)
9. Born, M., Wolf, E.: Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light, 7th edn . Cambridge University Press, Cambridge (1999)
10. Carazzone, J., Symes, W.: Velocity inversion by differential semblance optimization. Geophysics **56**, 654–663 (1991)
11. Curlander, J., McDonough, R.: Synthetic Aperture Radar – Systems and Signal Processing (Book). Wiley, New York (1991)
12. Fink, M.: Time reversed acoustics. Phys. Today **50**, 34 (1997)
13. Papanicolaou, G., Ryzhik, L., Sølna, K.: Statistical stability in time reversal. SIAM J. Appl. Math. **64**(4), 1133–1155 (2004)
14. Uhlmann, G.: Travel time tomography. J. Korean Math. Soc. **38**(4), 711–722 (2001)

Interior Point Methods

Osman Güler
Department of Mathematics and Statistics, University of Maryland Baltimore County, Baltimore, MD, USA

Description

Interior point methods (IPMs) are a class of algorithms for solving convex optimization problems which are efficient in theory (they have polynomial-time worst complexity) and in practice. They caused a true revolution in optimization and are widely considered to be one of the most important, if not the most important, developments in optimization within the last 30 years. They influenced nearly all existing areas of continuous optimization (convex and nonconvex) and discrete optimization and opened new areas of investigation in optimization, such as semidefinite programming, symmetric cone programming, and semialgebraic programming. Their influence continues today.

The revolution started in 1984 when Narendra Karmarkar announced his famous projective algorithm [10] for linear programming (LP). This algorithm has polynomial-time complexity, and more importantly, Karmarkar claimed that it was much faster than the simplex method on large, sparse linear programs. Although this dramatic claim did not quite materialize, IPMs are competitive today with the simplex method, and most LP software today (such as CPLEX) have both simplex and IPM options, although Karmarkar's original algorithm is now obsolete.

Most of the early attention in IPMs was directed toward LP and its close relatives, such as (convex) quadratic programming (QP) and (monotone) linear complementarity problems (LPC). At first, Karmarkar's algorithm did not fit any paradigm within optimization, but within a couple of years, connections were established with the logarithmic barrier methods of the 1950s and 1960s in which Newton's method is used at each iteration. Once this connection was understood, progress came quickly. By the late 1980s, duality theory of linear programming was incorporated into IPM, and in the early 1990s, the problem of finding an initial feasible point was elegantly answered with the invention of self-dual embedding techniques. By the mid-1990s, this part of IPM theory matured. Much more information on interior point methods for LP, QP, and LCP can be found in the books [20, 24, 25].

Around 1988, Nesterov and Nemirovski [15] dramatically expanded the scope of interior point methods to include *all* of convex programming. Their inspiration came from the earlier work of Renegar [18], who had devised a polynomial-time path-following logarithmic center method that uses Newton's method at each iterate. By a careful analysis of the logarithmic barrier function, they showed that only three properties of it are essential to obtain polynomial-time algorithms, calling any function satisfying them a *self-concordant barrier (s.c.b.) function*. Moreover, they showed that one can find such a s.c.b. function on any (regular) closed convex set, fittingly calling it the *universal barrier function*. Finally, Nesterov and Nemirovski showed that the Fenchel dual of the universal barrier for a convex cone is a s.c.b. for the dual cone. This means that the duality theory of convex programming works very well with IPM, making it possible to devise natural primal-dual IPM.

It was now possible, at least in theory, to devise IPM to solve any convex program in polynomial time. However, it is notoriously hard to compute the universal barrier function (or any other s.c.b.) for a general convex set. We know how to compute a suitable barrier function for some structured classes of problems such as LP, QP, semidefinite programming (SDP), symmetric and homogeneous cone programming, and hyperbolic programming. Nesterov and Nemirovski developed a kind of calculus to construct more s.c.b. out of known ones, such as for direct products and intersections of convex sets. Thus, in practice, IPM today is restricted to problems for which we can construct a *computable* s.c.b. using these techniques.

After LP, the next success story (perhaps its greatest) for IPM was the emergence, in the early 1990s, of semidefinite programming (SDP) as a major paradigm in convex programming. This is the problem of minimizing a linear function over the intersection of the cone of symmetric, positive semidefinite matrices (semidefinite cone) with an affine subset. We will discuss this and other exciting developments after we develop some terminology.

Due to space considerations, our treatment will be concise. Fortunately, a reader who wishes to learn more about IPMs can find much more detailed information in the two excellent survey articles [13, 14].

Self-Concordant Barrier Functions

Let C be a *regular* convex set (a closed convex set with nonempty interior and containing no entire lines) in a finite-dimensional inner product space E. A self-concordant barrier function is a C^3 function $F : \mathrm{int}(C) \to \mathbb{R}$ which is strongly convex (the Hessian $D^2F(x)$ is positive definite at any $x \in \mathrm{int}(C)$) and satisfies the following properties, for all $x \in \mathrm{int}(C)$ and for all $h \in E$:

$$|D^3F(x)[h,h,h]| \le 2(D^2F(x)[h,h])^{3/2}, \qquad \text{(self-concordance)}$$

$$|DF(x)[h]|^2 \le \vartheta D^2F(x)[h,h],$$

$$F(x) \to \infty \text{ as } x \to \partial C. \qquad \text{(barrier property)}$$

Here, $D^kF(x)[h,\ldots,h]$ is the kth directional of F at x along the direction h. The second property is satisfied if F is logarithmically homogeneous, $F(tx) = F(x) - \vartheta \log t$.

Let $C \subset \mathbb{R}^n$ be a regular convex set. Nesterov and Nemirovski's *universal barrier function* on C is given by

$$u(x) = c \log \mathrm{vol}(C^\circ(x)),$$

where $C^\circ(x)$ is the polar set $C^\circ(x) = \{y \in \mathbb{R}^n : \langle z - x, y\rangle \le 1 \text{ for all } z \in C\}$ and c is an absolute constant. It is shown in [6] that if $C = K$ is a regular convex cone, then

$$u(x) = c \log \int_{K^*} e^{-\langle x,y\rangle} dy,$$

where $K^* := \{s \in E : \langle x, s\rangle \ge 0 \text{ for all } x \in K\}$ is the dual cone of K.

Another universal barrier function was announced in 2012 by Hildebrand [8], who calls his function the *Einstein-Hessian self-concordant barrier*. It is the (unique) convex solution to the Monge-Ampère partial differential equation

$$u(x) = \frac{1}{2} \log \det D^2 u(x), \quad u|_{\partial K} = \infty.$$

It has slightly better theoretical properties than the original universal barrier function in the sense that its parameter value is exactly n ($\vartheta = n$) and it is *symmetric* under duality, that is, the Fenchel dual function u^* is the Einstein-Hessian barrier function for K^*. As mentioned before, both universal functions are hard to compute in general.

Conic Optimization

In principle, IPMs can be applied to any convex optimization problem, but the theory is simpler when applied to a problem in *conic* form, and the duality theory becomes more symmetric. Since there is no essential loss in generality (and most software deal with this kind of format), we limit our discussion to conic form.

A primal-dual pair of problems (P) and (D) in conic form is given by

$$\begin{array}{ll} \min \langle c, x\rangle & \max \langle b, y\rangle \\ \text{s.t. } Ax = b \ (P) & s.t. \ A^* y + s = c \ (D) \\ \quad x \in K, & \quad s \in K^*, \end{array}$$

where $A : E \to F$ is a linear operator between two finite-dimensional Euclidean spaces E and F, $A^* : F \to E$ its adjoint, $c \in E$, $b \in F$, $K \subset E$ is a regular convex cone in E, and $K^* := \{s \in E : \langle x, s\rangle \ge 0 \text{ for all } x \in K\}$ is the dual cone of K. It is well known in convex analysis that if one of the problems, say (P), has an interior feasible point $x \in \mathrm{int}(K)$ and $\inf(P) > -\infty$, then (D) has an optimal solution, and the strong duality theorem holds, that is, $\inf(P) = \max(D)$. It follows that if both programs (P) and (D) have interior feasible solutions, then both programs have optimal solutions and $\min(P) = \max(D)$.

The convex cones corresponding to LP, SDP, and QCP are the nonnegative orthant, the semidefinite cone, and the Lorentz cone given by $\{(x, t) \in \mathbb{R}^n \times \mathbb{R} : \|x\| \le t\}$, respectively.

The traditional interior penalty function method dating back to the 1960s [4] (p. 42) proceeds as follows in trying to solve our problem (P): under mild conditions on the barrier function F, the "path"

$$x(t) := \arg\min \{\langle c, x\rangle + tF(x) : Ax = b\}, \quad t > 0$$

exists and converges to the optimal solution set of (P) as $t \downarrow 0$. Suppose that x_k is "close to $x(t_k)$" in some measure. We then set the parameter t to a smaller value $t_{k+1} < t_k$ in some fashion and try to minimize the affine constrained penalty function $P_{k+1}(x) = \langle c, x\rangle + t_{k+1} F(x)$ subject to $Ax = b$ using a minimization method, say Newton's method, until we find a point x_{k+1} which is "close to $x(t_{k+1})$" and start all over again.

The computational complexity issues were not considered in the 1960s – they came later. One of the main contributions of Nesterov and Nemirovski was to show that if F is a s.c.b., then the (damped) Newton method performs very well in minimizing the penalty function $P_{k+1}(x)$. For example, if we choose $t_k / t_{k+1} = 1 + c\vartheta^{-1/2}$, only one Newton iteration is needed to go from x_k to x_{k+1}. This is a "short-step" path-following method which follows the central path closely. These are slow in practice. Path-following methods that are implemented choose t_{k+1} much more aggressively, leading to "long-step" methods. There exist dual and primal-dual variants of path-following algorithms. The interested reader should consult the survey articles [13, 14] and the books [15, 19] for more details.

Semidefinite Programming

We recall that a semidefinite program is a conic program (P) in which K is the semidefinite cone, that is, the cone of symmetric positive semidefinite matrices. By the late 1980s, it was already established by Nesterov and Nemirovski that the function $F(x) = -\log\det X$ is a s.c.b. for the semidefinite cone; hence, their IPMs could solve it in polynomial time. Several events in the early 1990s catapulted SDP into a major paradigm in convex optimization. First of all, Alizadeh [1] introduced a polynomial-time primal-dual IPM for SDP and showed that several eigenvalue problems can be formulated as SDP problems and some combinatorial optimization can be approximated as SDP problems. Secondly, Vandenberghe and Boyd [21] gave many examples of problems from engineering and elsewhere that can be formulated as SDP. Finally, and most dramatically, Goemans and Williamson [5] demonstrated that the SDP relaxation of the *maximum cut* problem from the graph theory delivers a solution whose expected value is at least 0.87856 times the optimal value, an improvement of about 38 % over previously known methods.

SDP has much greater modeling capabilities than LP, and since mid-1990s, much research effort has gone into finding out what classes of problems can be expressed as SDP. This effort is continuing today. The books [2, 23] and the article [13] contain a wealth of information on SDP.

Symmetric Cone Programming

In the 1990s, the theory of IPM expanded and deepened in several directions. The emergence of symmetric cone programming was one of them. Nesterov and Todd [16] identified a class of convex cones, which they called *self-scaled*, for which it is possible to devise long-step IPMs. They showed that this theory applies to the important classes of convex programming such as LP, QCP, and SDP. At about the same time, the author's article [6] brought the concepts of *symmetric cones, Euclidean Jordan algebras, and homogeneous convex cones* into IPM. We recall that a convex cone K is called *homogeneous* if the linear automorphisms of K are transitive, that is, given any two points $x, y \in K$, there is an automorphism T such that $T(K) = K$ and $T(x) = y$. A homogeneous cone is called *symmetric* if its dual cone (with respect to some Euclidean inner product) is equal to itself.

It turns out that the function $\int_{K^*} e^{-\langle x,y\rangle} dy$ that appears in the formula for the universal barrier had a substantial role in the classification of both symmetric cones (by Koecher) and homogeneous cones (by Vinberg). Moreover, symmetric cones are *exactly* the cones of squares of *Euclidean Jordan algebras*. These Jordan algebras were classified in 1930s by Jordan, von Neumann, and Wigner [9] in their quest for using Euclidean Jordan algebras as a basis for quantum mechanics. They were unsuccessful, however, because they found that there exist only five classes of elementary Jordan algebras. The cone of squares of these algebras correspond to the following five classes of convex cones: semidefinite cone over the real numbers, complex numbers and quaternions, the Lorentz (quadratic or ice-cream) cone, and a single exceptional cone, namely, the 3×3 semidefinite cone over the octonions. The book by Faraut and Korányi [3] is an excellent source for the theories of symmetric cones and Euclidean Jordan algebras. The author completed the cycle of correspondences by showing that self-scaled cones are exactly the symmetric cones.

Thus, the long-step primal-dual IPM methods of Nesterov and Todd are limited to a few, yet very important classes of convex optimization problems. Several software packages exist for symmetric cone programming including SeDuMi and SDPT3.

Hyperbolic Polynomials

A homogeneous polynomial $p : R^n \to R$ is called *hyperbolic* in direction d if $p(d) > 0$ and the map $t \mapsto p(x + td)$ has all *real* roots. The hyperbolicity cone $K(p, d)$ of p is the connected component of $\{x : p(x) \neq 0\}$ containing d or equivalently the set $K(p, d) = \{x \in \mathbb{R}^n : \text{all roots of } t \mapsto p(x + td) \text{ are negative}\}$. These polynomials originally appeared in partial differential equations, but they are also useful in IPMs. The theory of hyperbolic polynomials is currently active and has been found useful in optimization, combinatorics, and many other areas.

The basic facts about hyperbolic polynomials are: (i) the hyperbolicity cone $K(p,d)$ is convex *(Gårding 1950)*, (ii) the function $F(x) = -\log p(x)$ is s.c.b. barrier on $K(p,d)$ (thus alleviating the notorious problem of finding a computable s.c.b.), and (iii) more inequalities hold among the directional derivatives; see [7]. This last fact implies that it is possible to implement polynomial-time "long-step" IPMs for hyperbolic programming.

A conjecture of Peter Lax (1958) states that a homogeneous polynomial p of three variables is hyperbolic of degree m in the direction $e = (1,0,0)$ and satisfies $p(e) = 1$ if and only if there exist $m \times m$ real, symmetric matrices A_1 and A_2 such that $p(t_1,t_2,t_3) = \det(t_1 I + t_2 A_1 + t_3 A_2)$. Lewis, Parillo, and Ramana [12], using a deep result of Vinnikov, showed in 2003 that the Lax conjecture is true. This inspired another conjecture, called *generalized Lax conjecture*, which is still open. It claims that every hyperbolicity cone is a slice of the semidefinite cone, that is, the intersection of a semidefinite cone and a linear subspace; see [22].

Semialgebraic Programming

In the 2000s, yet another major class of optimization problems, this time optimization problems involving polynomial equations and inequalities (semialgebraic programming), was linked to SDP through the sum of squares approximation [17] and through moment problems [11]. This is a current area of intensive research. Software packages such as GloptiPoly and SOSTools are dedicated to semialgebraic programming.

References

1. Alizadeh, F.: Interior point methods in semidefinite programming with applications to combinatorial optimization. SIAM J. Optim. **5**, 13–51 (1995)
2. Ben-Tal, A., Nemirovski, A.S.: Lectures on Modern Convex Optimization. SIAM, Philadelphia (2001)
3. Faraut, J., Korányi, A.: Analysis on Symmetric Cones. Oxford University Press, New York (1994)
4. Fiacco, A.V., McCormick, G.P.: Nonlinear Programming, 2nd edn. SIAM, Philadelphia (1990)
5. Goemans, M.X., Williamson, D.P.: Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. J. Assoc. Comput. Mach. **42**, 1115–1145 (1995)
6. Güler, O.: Barrier functions in interior point methods. Math. Oper. Res. **21**(4), 860–885 (1996)
7. Güler, O.: Hyperbolic polynomials and interior point methods for convex programming. Math. Oper. Res. **22**, 350–377 (1997)
8. Hildebrand, R.: Einstein-Hessian barriers on convex cones. Optimization Online e-prints (2012)
9. Jordan, P., von Neumann, J., Wigner, E.: On an algebraic generalization of the quantum mechanical formalism. Ann. Math. (2) **35**, 29–64 (1934)
10. Karmarkar, N.: A new polynomial-time algorithm for linear programming. Combinatorica **4**, 373–395 (1984)
11. Lasserre, J.B.: Global optimization with polynomials and the problem of moments. SIAM J. Optim. **11**, 796–817 (2000/2001)
12. Lewis, A.S., Parrilo, P.A., Ramana, M.V.: The Lax conjecture is true. Proc. Am. Math. Soc. **133**, 2495–2499 (electronic) (2005)
13. Nemirovski, A.S.: Advances in convex optimization: conic programming. In: International Congress of Mathematicians, vol. I, pp. 413–444. European Mathematical Society, Zürich (2007)
14. Nemirovski, A.S., Todd, M.J.: Interior-point methods for optimization. Acta Numer. **17**, 191–234 (2008)
15. Nesterov, Y.E., Nemirovski, A.S.: Interior–Point Polynomial Algorithms in Convex Programming. SIAM, Philadelphia (1994)
16. Nesterov, Y.E., Todd, M.J.: Self–scaled barriers and interior–point methods for convex programming. Math. Oper. Res. **22**, 1–42 (1997)
17. Parrilo, P.A.: Semidefinite programming relaxations for semialgebraic problems. Math. Program. **96**(2, Ser. B), 293–320 (2003)
18. Renegar, J.: A polynomial-time algorithm, based on Newton's method, for linear programming. Math. Program. **40**(1, Ser. A), 59–93 (1988)
19. Renegar, J.: A Mathematical View of Interior-Point Methods in Convex Optimization. MPS/SIAM Series on Optimization. SIAM, Philadelphia (2001)
20. Roos, C., Terlaky, T., Vial, J.-P.: Interior Point Methods for Linear Optimization. Springer, New York (2006)
21. Vandenberghe, L., Boyd, S.: Semidefinite programming. SIAM Rev. **38**, 49–95 (1996)
22. Vinnikov, V.: LMI representations of convex semialgebraic sets and determinantal representations of algebraic hypersurfaces: past, present, and future. In: Mathematical Methods in Systems, Optimization, and Control, pp. 325–349. Birkhäuser/Springer Basel AG, Basel (2012)
23. Wolkowicz, H., Saigal, R., Vandenberghe, L. (eds.): Handbook of Semidefinite Programming. Kluwer Academic, Boston (2000)
24. Wright, S.J.: Primal-Dual Interior-Point Methods. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (1997)
25. Ye, Y.: Interior Point Algorithms. Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley, New York (1997)

Interpolation

Jean–Paul Berrut
Département de Mathématiques, Université de Fribourg, Fribourg/Pérolles, Switzerland

Mathematics Subject Classification

41A05; 65D05

Short Definition

In one-dimensional numerical analysis, *interpolation* is a solution of the problem of determining a function from a finite number of its values: it constructs a curve which exactly takes on given values at a finite number of points.

The Taylor Series and Newton's Interpolation Formula

In calculus classes, one learns the nth *Taylor polynomial* of a function f sufficiently smooth about a point x_0:

$$P_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k.$$

This approximation is extremely useful for theoretical purposes; however, it has several drawbacks in numerical practice: for instance, it requires the knowledge of the derivatives of f at x_0 and, since the information is concentrated in one point, it rapidly becomes ill conditioned (unstable) as x moves away from x_0.

These difficulties disappear by going over to *interpolation*: when approximating real functions, one takes as input instead of the $f^{(k)}(x_0)$ the values of f at $n+1$ distinct abscissas (nodes) $x_0, x_1, \ldots, x_n$ on some interval $[a, b]$ and replaces the derivatives by *divided differences*

$$\frac{f'(x_0)}{1!} \approx \frac{f(x_1) - f(x_0)}{x_1 - x_0} =: f[x_0, x_1]$$
$$\frac{f''(x_0)}{2!} \approx \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0} =: f[x_0, x_1, x_2]$$

and the powers of $x - x_0$ by products of $x - x_j$: with $f[x_0] := f(x_0)$, this yields the *Newton interpolation polynomial* of degree at most n

$$\begin{aligned} p_n(x) = f[x_0] + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2] \\ (x - x_0)(x - x_1) + \ldots \\ \ldots + f[x_0, x_1, \ldots, x_n](x - x_0)(x - x_1) \\ \ldots (x - x_{n-1}). \end{aligned} \tag{1}$$

Interpolation is the property that the approximation goes through the values of f at the given abscissas:

$$p_n(x_j) = f_j := f(x_j), \qquad j = 0, \ldots, n.$$

In effect, (1) merely is the Newton *form* of the interpolating polynomial; several other representations exist, but there is only one interpolating polynomial of degree at most n: would there be two, their difference would have the $n+1$ zeros x_j, which is impossible.

Newton's form has some favorable features: once the divided differences have been computed, which requires $\mathcal{O}(n^2)$ arithmetic operations, merely $\mathcal{O}(n)$ operations are necessary for evaluating p_n at a point x; adding a new abscissa x_{n+1} is immediate, as it just requires extension of (1) with the next term $f[x_0, \ldots, x_{n+1}](x - x_0) \ldots (x - x_n)$; interpolation of a matrix function is straightforward.

However, it also has some drawbacks: two of the severe ones are the facts that the formula, and unfortunately also numerical values of $p_n(x)$ in usual arithmetic, strongly depends on the ordering of the nodes and that the divided differences depend on f.

There fortunately are several other forms of the polynomial: an important one is *Neville's*, a cousin of Newton mostly used for extrapolation to a limit, yet another is *Lagrange's*, which has many decisive advantages over Newton's.

Lagrange Interpolation Formula

Waring and Euler independently had the following constructive idea for deriving p_n: to every x_j they considered the polynomial ℓ_j of degree n that takes the value 1 at x_j and vanishes at all other nodes:

$$\ell_j(x) = \lambda_j \prod_{\substack{i=0 \\ i\neq j}}^{n} (x - x_i), \quad \text{with}$$

$$\lambda_j := 1 \Big/ \prod_{\substack{i=0 \\ i\neq j}}^{n} (x_j - x_i), \quad j = 0, \ldots, n.$$

The unicity then warrants the validity of the *Lagrange interpolation formula*

$$p_n(x) = \sum_{j=0}^{n} f_j \ell_j(x), \tag{2}$$

which expresses p_n as sort of a linear combination of the interpolated values f_j with coefficient functions ℓ_j *which do not depend on* f; it is therefore efficiently used as ansatz in all kinds of solution methods, e.g., in pseudospectral methods for differential equations.

Unfortunately, evaluating (2) requires $\mathcal{O}(n^2)$ operations for every x and is unstable; this led most authors to discard the Lagrange form in favor of Newton's for most of the twentieth century. However, it may easily be modified into

$$p_n(x) = \ell(x) \sum_{j=0}^{n} \frac{\lambda_i}{x - x_j} f_j, \quad \ell(x) := \prod_{j=0}^{n} (x - x_j). \tag{3}$$

N. Higham [6] has shown backward as well as forward stability of this formula, which makes it the most suited of all, at least as far as accuracy is concerned.

As with the Newton form, $\mathcal{O}(n)$ operations are necessary for evaluating p_n at some x when the *weights* λ_j have been determined. Updating the λ_j when a new node x_{n+1} is added is a $\mathcal{O}(n)$ process as well [2, 3], so that Lagrange asymptotically is as fast as Newton in this respect. There even exist closed, $\mathcal{O}(1)$ formulas for λ_j for some of the most important sets of points, i.e., Chebyshev points of the four kinds, and equidistant points on the interval and on the complex unit circle [2]. No expensive computations are then needed for the weights λ_j, and thus only $\mathcal{O}(n)$ operations are required for evaluating p_n, something no other formula seems to achieve. A fast numerical formula has recently been found even for Legendre points by Wang et al; see [9].

A few words about the condition (stability) of the problem: polynomial interpolation may only be used in practice for arbitrary (large) n when the points are distributed on the interval so as to accumulate at the extremities. To be more precise, assume that the problem has been scaled so that the interval of interpolation is $[-1, 1]$. Then every node x_j may be mapped to two vertically aligned points on the unit circle E by the application $\phi(x) = \arccos x$ to yield a node distribution on E. For good conditioning, these nodes should be about evenly distributed on E. This is the case, e.g., for Chebyshev and Legendre points, but not for equidistant ones.

Polynomial interpolation with good nodes such as Chebyshev's and Legendre's is unbeaten for very smooth functions if one may increase n as well. For Chebyshev points of the first kind, for instance, the interpolation error may be bounded as

$$|P_n(x) - f(x)| \le 2^{-n} \frac{M_{n+1}}{(n+1)!}, \quad x \in [-1, 1].$$

$$M_{n+1} := \max_{\xi \in [-1,1]} \left| f^{(n+1)}(\xi) \right|,$$

Thus, when M_{n+1} does not grow much faster with n than $(n+1)!$, the error decreases exponentially with n. Results are very similar with Chebyshev points of the second kind, which are more important in practice as they contain the extremities of the interval; notice that to experiment with their fantastic efficiency, also in applications, there is no need to write programs any longer: one may just download the public domain software Chebfun [9].

When the nodes cannot be chosen, one usually turns to piecewise polynomial interpolants called *splines*, which we do not elaborate on here [4].

The Barycentric Formula

Formula (3) may still be improved for actual computation. One of the difficulties is the growth which may occur in the various factors $\ell(x)$ and λ_j for large n and requires adjustments such as the use of logarithms.One may get rid of common factors by the following manipulations: one considers besides the interpolant of f that of the function identically 1, which by the unicity equals 1, divides each side of (3) by that of the corresponding formula and cancels $\ell(x)$ to obtain

$$p_n(x) = \sum_{j=0}^{n} \frac{\lambda_j}{x - x_j} f_j \Bigg/ \sum_{j=0}^{n} \frac{\lambda_j}{x - x_j}. \qquad (4)$$

Equation (4) is the *barycentric formula* for p_n. Higham [6] has proved that it is merely forward stable and given a particular example for which (3) and (4) yield different results; this does not happen in actual practice, however.

As the λ_j appear in the numerator and the denominator, any common factor independent of j may be cancelled to yield very elegant simple formulas for the corresponding *simplified weights* λ_j^*. One has $\lambda_j^* = (-1)^i \binom{n}{j}$ for equidistant nodes and $\lambda_j^* = (-1)^i \delta_i \eta_j$ with

$$\eta_j := \begin{cases} \sin\phi_j, & \text{1st kind,} \\ 1, & \text{2nd kind,} \\ \sin(\phi_j/2), & \text{3rd kind,} \\ \cos(\phi_j/2), & \text{4th kind,} \end{cases} \qquad \delta_j := \begin{cases} 1/2, & x_j = 1 \text{ or } x_j = -1, \\ 1, & \text{otherwise} \end{cases}$$

for Chebyshev points $x_j = \cos\phi_j$ [1]. For the complex roots of unity, $x_j := e^{j2\pi \mathrm{i}/n}$, $\lambda_j^* = x_j$. Another advantage of (4) is guaranteed interpolation even when the λ_j are in error (as long as none of them vanishes).

Interpolation is a vast subject, of which we have just touched the simple polynomial version. We note, in particular, that the so efficient polynomial interpolation between Chebyshev nodes is a special case of trigonometric interpolation between equidistant nodes [1] and that the latter is itself the restriction to periodic functions of sinc interpolation on the infinite line [8]. Hermite–Birkhoff interpolation considers the case in which derivatives are prescribed on top of the function values at the nodes. Another extension is rational interpolation, in which the interpolant is a quotient of two polynomials: see [7] for the classical nonlinear version and [5] for the linear case.

The literature on interpolation is huge, as a chapter of about every numerical analysis book is devoted to it. We have limited ourselves to a few of the most recent citations, from which the reader will be able to access the classic literature.

References

1. Berrut, J.-P.: Baryzentrische Formeln zur trigonometrischen Interpolation (I). Z. Angew. Math. Phys. **35**, 91–105 (1984)
2. Berrut, J.-P., Trefethen, L.N.: Barycentric Lagrange interpolation. SIAM Rev. **46**, 501–517 (2004)
3. Dahlquist, G., Björck, Å.: Numerical Methods in Scientific Computing, vol. 1. SIAM, Philadelphia (2008)
4. de Boor, C.: A Practical Guide to Splines, Revised Edition. Applied Mathematical Sciences, vol. 27. Springer, New York (2001)
5. Floater, M.S., Hormann, K.: Barycentric rational interpolation with no poles and high rates of approximation. Numer. Math. **107**, 315–331 (2007)
6. Higham, N.: The numerical stability of barycentric Lagrange interpolation. IMA J. Numer. Anal. **24**, 547–556 (2004)
7. Pachón, R., Gonnet, P., van Deun, J.: Fast and stable rational interpolation in roots of unity and Chebyshev points. SIAM J. Numer. Anal. **50**, 1713–1734 (2012)
8. Stenger, F.: Handbook of Sinc Numerical Methods. Chapman and Hall, Boca Raton (2010)
9. Trefethen, L.N.: Approximation Theory and Approximation Practice. SIAM, Philadelphia, (2013)

Interval Arithmetics

Siegfried M. Rump
Institute for Reliable Computing, Hamburg University of Technology, Hamburg, Germany
Faculty of Science and Engineering, Waseda University, Tokyo, Japan

Synonyms

Automatic error analysis; Interval analysis; Reliable computing; Rigorous error bounds

Definition

The *raison d'être* of interval arithmetic is to obtain rigorous error bounds for computational results. The worst case error estimates for arithmetical operations are used in *verification methods* to solve many numerical problems with full rigor and in a reasonable

computing time, not far from that of a traditional approximate numerical method.

Historical Background

Intervals are well known in mathematics. Archimedes' inclusion $[\frac{223}{71}, \frac{22}{7}]$ of π using the 96-sided polygon is one of the oldest examples. Numbers afflicted with a tolerance (*Ungenaue Zahlen*) such as 3.14 ± 0.01 and operations over those were used by Gauss; see also [4]. Higher order terms were sometimes neglected.

In the nineteenth and early twentieth centuries, sequences of (nested) intervals were introduced as one way to formalize real numbers. Apparently, this was known to Bolzano in 1817 and was formalized by Bachmann [1]. Also [27] was in this spirit.

The challenge is to compute error bounds for numerical problems; arithmetical operations are helpful, but by no means sufficient (see below). In February 1956, Sunaga [25] is the first to use interval arithmetic to compute error bounds for the solution of numerical problems. This seminal paper, handwritten in Japanese and much ahead of its time, introduces and investigates real and complex interval arithmetic (with floating-point bounds), inf-sup and mid-rad representation, the natural interval extension of functions, the interval Newton procedure, Simpson's rule with verification, error bounds for the solution of initial value problems, and more. It remained completely unrecognized.

In the late 1950s, with the rise of digital computers, interval operations with floating-point endpoints seem to be common knowledge, cf. [2, 6, 16]. In the sequel, undoubtedly Moore popularized interval arithmetic [17, 18].

Standard Intervals

If a quantity is not precisely known and/or there is no simple characterization of it, it may be represented by an interval. For example, $\pi \in [3.14, 3.15]$ is a true statement and may be used to obtain error bounds for functions involving π, such as $\sqrt{\pi} \in [\sqrt{3.14}, \sqrt{3.15}] \subseteq [1.772, 1.775]$.

The result of an operation such as $a+b$ for $a, b \in \mathbb{R}$ with $a \in [a_1, a_2]$ and $b \in [b_1, b_2]$ satisfies $a + b \in [a_1 + b_1, a_2 + b_2]$. More general, denote by $\mathbb{IR}$ the set of nonempty closed real intervals, and let $\boldsymbol{a} := [a_1, a_2], \boldsymbol{b} := [b_1, b_2] \in \mathbb{IR}$ be given. An interval operation $\circ \in \{+, -, \cdot, /\}$ is defined by (provided $0 \notin \boldsymbol{b}$ in case of division)

$$\boldsymbol{a} \circ \boldsymbol{b} := \boldsymbol{c} = [c_1, c_2] \tag{1}$$
$$\text{with } c_1 := \min_{i,j}\{a_i \circ b_j\} \quad \text{and} \quad c_2 := \max_{i,j}\{a_i \circ b_j\}.$$

Obviously, $a \in \boldsymbol{a}$ and $b \in \boldsymbol{b}$ implies $a \circ b \in \boldsymbol{a} \circ \boldsymbol{b}$, and the result is optimal. For computational purposes, the general definition (1) can be improved by using case distinctions. For example [20],

$$a_1 \geq 0 \text{ and } b_2 < 0 \text{ implies } \boldsymbol{a}/\boldsymbol{b} = [\, a_2/b_2 \,,\, a_1/b_1 \,]\,. \tag{2}$$

For a function $f : \mathbb{R} \to \mathbb{R}$ composed of arithmetic operations, the *natural interval extension* $F : \mathbb{IR} \to \mathbb{IR} \cup \{\texttt{NaI}\}$ is defined by replacing each arithmetic operation by the corresponding interval operation, where `NaI` (`Not an Interval`) is the result of an invalid operation. For $F(\boldsymbol{x}) \neq \texttt{NaI}$, it follows the remarkable property $x \in \boldsymbol{x} \Rightarrow f(x) \in F(\boldsymbol{x})$, so that $F(\boldsymbol{x})$ encloses the range of f over the interval $\boldsymbol{x} \in \mathbb{IR}$.

This inclusion property can be maintained for standard functions, as previously noted for the square root. Moreover, also for non-monotonic functions, the range can be enclosed. For example, for all $\boldsymbol{x} = [x_1, x_2] \in \mathbb{IR}$ and $|\boldsymbol{x}| := \max\{|x_1|, |x_2|\}$, it follows

$$\sin(\boldsymbol{x}) \subseteq \big((\boldsymbol{x}^2/20 - 1)\boldsymbol{x}^2/6 + 1\big)\boldsymbol{x} + [-e, e],$$
$$\text{where } e := |\boldsymbol{x}|^7/7!\,. \tag{3}$$

The inclusion is correct but broad for larger $|\boldsymbol{x}|$. With some effort, narrow inclusions for arbitrary $\boldsymbol{x}$ can be computed as in INTLAB [23], the Matlab toolbox for reliable computing, and interval extensions for all elementary standard functions and operations between those are obtained. This leads to the inclusion of the range of nonelementary and other functions, such as a definite integral. An example of a crude inclusion is

$$\int_a^b f(x)\,dx \in h \sum_{i=1}^{n} f(\boldsymbol{x}^{(i)}) \quad \text{with}$$
$$\boldsymbol{x}^{(i)} := [\, a + (i-1)h \,,\, a + ih \,] \tag{4}$$

for an integrable function $f : [a, b] \to \mathbb{R}$, $1 \leq n \in \mathbb{N}$ and $h := \frac{b-a}{n}$. As an example, consider $f(x) := \sin\sqrt{x + \pi}$ with $\sqrt{[x_1, x_2]} = [\sqrt{x_1}, \sqrt{x_2}\,]$ with $x_1 \geq 0$, $\pi \in [3.14, 3.15]$ and (3) to include the sine function. Using $n = 4$ and $n = 64$ in (4) proves

$$\int_{-2}^{1} f(x)\,dx \subseteq [2.34, 3.51] \quad \text{and}$$
$$\int_{-2}^{1} f(x)\,dx \subseteq [2.84, 2.98]\,, \tag{5}$$

respectively. By using a better inclusion of π, a better inclusion function of the sine, and, of course, a better quadrature formula, an accurate inclusion of the integral can be computed. For example, the executable Matlab/INTLAB code

```
f='sin(x+exp(x))';app=quad(f,0,8),
   incl=verifyquad(f,0,8)
```

uses the Matlab quadrature routine `quad` and the INTLAB routine `verifyquad` [23]. It computes in double precision (corresponding to 16 decimal places) the approximation `app` $= 0.25110272$, without warning; the inclusion $[0.34740016, 0.34740018]$ needs about 1.5 times the computing time but shows that no digit of the approximation is correct.

Overestimation and the Dependency Problem

The result of a sequence of interval operations is *either* `NaI` *or* a completely rigorous inclusion of the true result. This ease of use comes at a price. Identical interval quantities occurring more than once cannot be recognized as such and are treated as independent data. For instance,

$$[3.14, 3.15] - [3.14, 3.15] = [-0.01, 0.01] \tag{6}$$

is best possible: The first interval might be an inclusion of 3.14, but the second of 3.15, say. The potential information that both intervals represent π is lost.

The natural interval extension of an arithmetic expression yields the exact range if each variable occurs only once [18]; otherwise, the overestimation may be arbitrarily large. As an example, consider $f(x) := e^{x^2-4x}$ on $\boldsymbol{x} := [2, 4]$. The natural interval extension yields a true inclusion but gross overestimation

$$f(\boldsymbol{x}) \subseteq \exp([4, 16] - [8, 16]) = \exp([-12, 8])$$
$$= [6.14 \cdot 10^{-6}, 2980.96]\,. \tag{7}$$

The reformulation $f(x) = e^{(x-2)^2-4}$ of the *original* function contains the variable x only once, and the natural interval extension produces the exact range

$$f(\boldsymbol{x}) \subseteq \exp([0, 2]^2 - 4) = \exp([-4, 0]) = [0.0183, 1]\,. \tag{8}$$

Based on interval operations, so-called verification methods compute verified error bounds for the solution of a numerical problem. The challenge is to utilize interval operations in a way that potential overestimation is diminished (see below).

Interval Vectors and Matrices

A matrix (vector) with interval entries forms an interval matrix (vector). Interval operations are the natural extension of the real operations [20]. For example, for an interval matrix $\mathbf{A} = (\mathbf{a}_{ij})$ and an interval vector $\mathbf{x} = (\mathbf{x}_j)$, the entries of $\mathbf{y} = \mathbf{A} \cdot \mathbf{x}$ are

$$\mathbf{y}_i = \sum_j \mathbf{a}_{ij} \cdot \mathbf{x}_j \tag{9}$$

using scalar interval sums and products in the right-hand side. Note that the scalar interval operations in (1) are identical to the power set operation, i.e.,

$$\boldsymbol{a} \circ \boldsymbol{b} = \{a \circ b : a \in \boldsymbol{a}, b \in \boldsymbol{b}\} \quad \text{for } \circ \in \{+, -, \cdot, /\} \tag{10}$$

(with $0 \notin \boldsymbol{b}$ in case of division), but for interval matrix and vector operations only the inclusion principle holds true. In (9), $\boldsymbol{y}$ is the narrowest interval vector including the power set operation, i.e., $\{Ax : A \in \boldsymbol{A}, x \in \boldsymbol{x}\} \subseteq \boldsymbol{A} \cdot \boldsymbol{x}$ is best possible.

Alternatives to Intervals: Other Representations of Sets

Interval arithmetic is *one* (elegant) possibility to estimate the error of numerical operations. A generalized interval arithmetic including intervals $[-\infty, a_2] \cup [a_1, \infty]$ is introduced in [10] and [11].

Generally, any subset $\mathbb{S} \subseteq \mathbb{PR}$ of the power set of the real numbers with computable operations $\circ : \mathbb{S} \times \mathbb{S} \to \mathbb{S}$ can be used to estimate numerical errors. Similarly, other sets of vectors $\mathbb{S} \subseteq \mathbb{PR}^n$ may be used.

A natural candidate is a set $\mathbb{S}$ of polytopes, such as the set of standard simplices. More generally, parallelepipeds are introduced as "affine arithmetic" in [3] and successfully used to solve initial value problems [5, 15]. Moreover, hyperellipsoids were considered in [8] and arithmetical operations defined in [21].

Convex conic representable sets and relaxation techniques based on semi-definite programming have

been used by Jansson [9] to solve large optimization problems.

Implementation on Digital Computers

For the computation of rigorous error bounds on digital computers, intervals with floating-point bounds have to be used. Denote by $\mathbb{F}$ a set of floating-point numbers, for example, according to the IEEE 754 arithmetic standard [7]. In general, real operations between floating-point numbers are not in $\mathbb{F}$, such as $r := 1/10$. But there are unique $f_1, f_2 \in \mathbb{F}$ with $f_1 \le r \le f_2$ and a minimal distance $f_2 - f_1$. Those can be computed in [7] using directed rounding, i.e., the quotient $1/10$ computed in rounding downwards yields f_1, the largest floating-point number f with $f \le 1/10$, and when rounding upwards, the result is f_2, the smallest $f \in \mathbb{F}$ with $1/10 \le f$.

The vast majority of today's computers adhere to the IEEE 754 standard, so that all four basic arithmetic operations are available in rounding downwards and upwards (and, of course, in rounding to nearest). For $\boldsymbol{a} = [a_1, a_2]$ and $\boldsymbol{b} = [b_1, b_2]$ with $a_1, a_2, b_1, b_2 \in \mathbb{F}$, interval operations are thus defined by

$$\boldsymbol{c} = \boldsymbol{a} \circ \boldsymbol{b} := [c_1, c_2]$$
$$\text{with } c_1 := \min_{i,j} a_i \circ_\nabla b_j \text{ and } c_2 := \max_{i,j} a_i \circ_\Delta b_j, \tag{11}$$

where $\circ_\nabla$ and $\circ_\Delta$ denote the result in rounding downwards and upwards, respectively. Thus the bounds of $\boldsymbol{c}$ are *computed floating-point numbers*, and it follows $a \circ b \in \boldsymbol{c}$ for all *real* $a \in \boldsymbol{a}$, $b \in \boldsymbol{b}$. Again, simplifications of (11) such as $[a_1, a_2] - [b_1, b_2] = [a_1 -_\nabla b_2, a_2 -_\Delta b_1]$, and by case distinctions for multiplication and division are obvious.

Operations for interval vectors and matrices with floating-point bounds are defined similar to (9) using directed rounding.

Note that the range of a function defined by a sequence of arithmetic operations and (elementary) standard functions is rigorously enclosed *solely using floating-point operations*. A value $f(\pi)$ can be bounded as well by replacing π by an enclosing interval with floating-point endpoints, etc.

Verification Methods

The appealing inclusion of the range of a function by its natural interval extension would stand to reason to replace in an algorithm each operation by its corresponding interval operation. Gaussian elimination modified this way either produces `NaIs` or delivers rigorous error bounds for the solution of a linear system.

However, such an approach is almost certainly bound to fail [24, Sect. 10.1]. Even for toy problems the discussed dependency problem leads to wide intervals, eventually causing premature program termination by a denominator interval containing zero. Here is a major difference to numerical methods, where replacing real operations by floating-point operations usually produces satisfactory results.

In contrast, a *verification method* is based on a mathematical theorem and uses interval arithmetic to verify the assumptions. As a simple example, let matrices $A, R \in \mathbb{F}^{n\times n}$, a vector $b \in \mathbb{F}^n$, and a potential inclusion $\boldsymbol{x} \in \mathbb{IF}^n$ of $A^{-1}b$ be given. If

$$Rb \in \boldsymbol{z},\ I - RA \in \boldsymbol{C},\ \|\,|\boldsymbol{C}|\,\|_\infty < 1 \quad \text{and}$$
$$\boldsymbol{z} + \boldsymbol{C}\boldsymbol{x} \subseteq \boldsymbol{x} \tag{12}$$

for I denoting the identity matrix and $|\boldsymbol{C}| := (|\boldsymbol{C}_{ij}|)$, then A is non-singular and $A^{-1}b \in \boldsymbol{x}$. Basically this is already proved (for nonlinear functions) in [10] by using fixed-point theorems; an explicit formulation as an existence test is given in [19]. The quantities $\boldsymbol{z}$ and $\boldsymbol{C}$ are calculated in interval arithmetic with floating-point endpoints.

Note that there are no assumptions on A, R, b, or $\boldsymbol{x}$ other than (12), in particular not on the condition number of A. This principle is elaborated in verification methods on a much higher level together with the construction of suitable test sets $\boldsymbol{x}$. Applying (12) to interval data $\boldsymbol{A}, \boldsymbol{b}$, the "solution set" $\Sigma(\boldsymbol{A}, \boldsymbol{b}) := \{x \in \mathbb{R}^n : \exists\, A \in \boldsymbol{A}\ \exists\, b \in \boldsymbol{b} \text{ with } Ax = b\}$ is included by $\boldsymbol{x}$. The exact computation of $\Sigma(\boldsymbol{A}, \boldsymbol{b})$ is NP-hard [22].

Based on the above, verification methods for various standard problems in numerical analysis have been developed from systems of nonlinear equations, eigenproblems, and general, constrained, and semidefinite programming problems to ordinary and partial differential equations. For an overview, see [24].

Among the many references for verification methods are [20, 24]; libraries for interval operations in C++ include C-XSC [12] and Profil/BIAS [13]. The examples in this article are computed in INTLAB [23], the widely used Matlab toolbox for reliable computing. It is completely written in Matlab and covers interval arithmetic, standard functions, automatic differentiation and various verification methods and demos.

Nontrivial problems have been solved using verification methods by so-called computer-assisted proofs. For example, Tucker [26] received the 2004 EMS prize awarded by the European Mathematical Society for "giving a rigorous proof that the Lorenz attractor exists for the parameter values provided by Lorenz. This was a long standing challenge to the dynamical system community, and was included by Smale in his list of problems for the new millennium. The proof uses computer estimates with rigorous bounds based on higher dimensional interval arithmetics."

References

1. Bachmann, P.: Vorlesungen über die Natur der Irrationalzahlen, Theorie der Irrationalzahlen, B.G. Teubner, Leipzig (1892)
2. Collins, G.E.: Interval arithmetic for automatic error analysis. Technical report, IBM, Mathematics and Applications Department, New York (1960)
3. Comba, J.L.D., Stolfi, J.: Affine arithmetic and its applications to computer graphics. Presented at SIBGRAPI'93, Recife, 20–22 Oct 1993
4. Dwyer, P.S.: Linear Computations. Wiley, New York/London (1951)
5. Eijgenraam, P.: The Solution of Initial Value Problems Using Interval Arithmetic. Mathematisch Centrum, Amsterdam (1981)
6. Fischer, P.C.: Automatic propagated and round-off error analysis. In: Proceedings of the ACM National Meeting, Urbana, 11–13 June, pp. 39.1–2 (1958)
7. IEEE Standard for Floating-Point Arithmetic, In IEEE Std 754-2008 (29 August 2008), pp. 1–58.
8. Jackson, L.W.: A comparison of ellipsoidal and interval arithmetic error bounds, numerical solutions of nonlinear problems (notice). SIAM Rev. **11**, 114 (1969)
9. Jansson, C.: On verified numerical computations in convex programming. Jpn. J. Ind. Appl. Math. **26**, 337–363 (2009)
10. Kahan, W.M.: A More Complete Interval Arithmetic. Lecture Notes for a Summer Course at the University of Michigan (1968)
11. Kaucher, E.: Interval analysis in the extended interval space $\mathbb{IR}$. Comput. Suppl. **2**, 33–49 (1980)
12. Klatte, R., Kulisch, U., Wiethoff, A., Lawo, C., Rauch, M.: C-XSC: A C++ Class Library for Extended Scientific Computing. Springer, Berlin (1993)
13. Knüppel, O.: PROFIL/BIAS – a fast interval library. Computing **53**, 277–287 (1994)
14. Krawczyk, R.: Newton-Algorithmen zur Bestimmung von Nullstellen mit Fehlerschranken. Computing **4**, 187–201 (1969)
15. Lohner, R.: Einschließung der Lösung gewöhnlicher Anfangs- und Randwertaufgaben und Anordnungen. PhD thesis, University of Karlsruhe (1988)
16. Moore, R.E.: Automatic error analysis in digital computation. Technical report LMSD-48421, Lockheed Missiles and Space Division, Sunnyvale (1959)
17. Moore, R.E.: Interval arithmetic and automatic error analysis in digital computing. Dissertation, Stanford University (1963)
18. Moore, R.E.: Interval Analysis. Prentice-Hall, Englewood Cliffs (1966)
19. Moore, R.E.: A test for existence of solutions for non-linear systems. SIAM J. Numer. Anal. **4**, 611–615 (1977)
20. Neumaier, A.: Interval Methods for Systems of Equations. Encyclopedia of Mathematics and Its Applications. Cambridge University Press, Cambridge (1990)
21. Neumaier, A.: The wrapping effect, ellipsoid arithmetic, stability and confidence regions. Comput. Suppl. **9**, 175–190 (1993)
22. Poljak, S., Rohn, J.: Checking robust nonsingularity is NP-hard. Math. Control Signals Syst. **6**, 1–9 (1993)
23. Rump, S.M.: INTLAB – INTerval LABoratory, version 7.1. http://www.ti3.tuhh.de/rump (1998–2013)
24. Rump, S.M.: Verification methods: rigorous results using floating-point arithmetic. Acta Numer. **19**, 287–449 (2010)
25. Sunaga, T.: Geometry of numerals. Master's thesis, University of Tokyo (1956)
26. Tucker, W.: The Lorenz attractor exists. C. R. Acad. Sci. Paris Sér. I Math. **328**(12), 1197–1202 (1999)
27. Young, R.C.: The algebra of many-valued quantities. Math. Ann. **104**, 260–290 (1931)

Inverse Boundary Problems for Electromagnetic Waves

Gunther Uhlmann[1] and Ting Zhou[2]
[1]Department of Mathematics, University of Washington, Seattle, WA, USA
[2]Department of Mathematics, Northeastern University, Boston, MA, USA

Introduction

In this chapter we consider inverse boundary problems for electromagnetic waves. The goal is to determine the electromagnetic parameters of a medium by making measurements at the boundary of the medium.

We concentrate on fixed energy problems. We first discuss the case of electrostatics, which is called Electrical Impedance Tomography (EIT). This is also called Calderón problem since the mathematical formulation of the problem and the first results in the multidimensional case were due to A.P. Calderón [11]. In this case the electromagnetic parameter is the conductivity of the medium, and the equation modelling the problem is the conductivity equation. Then we discuss the more general case of recovering all the electromagnetic parameters of the medium, electric permittivity, magnetic permeability, and electrical conductivity of the medium by making boundary measurements, and the equation modeling the problem is the full system of Maxwell's equations. Finally we consider the problem of determining electromagnetic inclusions and obstacles from electromagnetic boundary measurements. A common feature of the problems we study is that they are fixed energy problems. The type of electromagnetic waves that we use to probe the medium are complex geometrical optics solutions to Maxwell's equations.

Electrical Impedance Tomography

The problem that Calderón proposed was whether one can determine the electrical conductivity of a medium by making voltage and current measurements at the boundary of the medium. Calderón was motivated by oil prospection. In the 1940s he worked as an engineer for Yacimientos Petrolíferos Fiscales (YPF), the state oil company of Argentina, and he thought about this problem then although he did not publish his results until many years later. For applications of electrical methods in geophysics, see [52]. EIT also arises in medical imaging given that human organs and tissues have quite different conductivities. One potential application is the early diagnosis of breast cancer [54]. The conductivity of a malignant breast tumor is typically 0.2 mho which is significantly higher than normal tissue which has been typically measured at 0.03 mho. For other medical imaging applications, see [22].

We now describe more precisely the mathematical problem. Let $\Omega \subseteq \mathbb{R}^n$ be a bounded domain with smooth boundary (many of the results we will describe are valid for domains with Lipschitz boundaries). The isotropic electrical conductivity of Ω is represented by a bounded and positive function $\gamma(x)$. In the absence of sinks or sources of current and given a voltage potential on the boundary $f \in H^{\frac{1}{2}}(\partial\Omega)$, the induced potential $u \in H^1(\Omega)$ solves the Dirichlet problem

$$\nabla \cdot (\gamma \nabla u) = 0 \text{ in } \Omega, \quad u\big|_{\partial\Omega} = f. \tag{1}$$

The Dirichlet-to-Neumann map, or voltage to current map, is given by

$$\Lambda_\gamma(f) = \left(\gamma \frac{\partial u}{\partial \nu}\right)\Big|_{\partial\Omega} \tag{2}$$

where ν denotes the unit outer normal to $\partial\Omega$.

The inverse problem of EIT is to determine γ knowing Λ_γ. It is difficult to find a systematic way of prescribing voltage measurements at the boundary to be able to find the conductivity. Calderón took instead a different route. Using the divergence theorem we have

$$Q_\gamma(f) := \int_\Omega \gamma |\nabla u|^2 dx = \int_{\partial\Omega} \Lambda_\gamma(f) f \, dS \tag{3}$$

where dS denotes surface measure and u is the solution of (1). In other words $Q_\gamma(f)$ is the quadratic form associated to the linear map $\Lambda_\gamma(f)$, and to know $\Lambda_\gamma(f)$ or $Q_\gamma(f)$ for all $f \in H^{\frac{1}{2}}(\partial\Omega)$ is equivalent. $Q_\gamma(f)$ measures the energy needed to maintain the potential f at the boundary. Calderón's point of view is that if one looks at $Q_\gamma(f)$, the problem is changed to finding enough solutions $u \in H^1(\Omega)$ of the conductivity equation in order to find γ in the interior. He carried out this approach for the linearized EIT problem at constant conductivity. He used the harmonic functions $e^{x\cdot\rho}$ with $\rho \in \mathbb{C}^n$, $\rho \cdot \rho = 0$.

Complex Geometrical Optics Solutions with a Linear Phase

Sylvester and Uhlmann [46, 47] constructed in dimension $n \geq 2$ complex geometrical optics (CGO) solutions of the conductivity equation for C^2 conductivities that behave like Calderón exponential solutions for large frequencies. This can be reduced to constructing solutions in the whole space (by extending $\gamma = 1$ outside a large ball containing Ω) for the Schrödinger equation with potential.

Let $\gamma \in C^2(\mathbb{R}^n)$, γ strictly positive in $\mathbb{R}^n$, and $\gamma = 1$ for $|x| \geq R$, $R > 0$. Let $L_\gamma u = \nabla \cdot \gamma \nabla u$. Then we have

$$\gamma^{-\frac{1}{2}} L_\gamma(\gamma^{-\frac{1}{2}}) = \Delta - q, \quad q = \frac{\Delta\sqrt{\gamma}}{\sqrt{\gamma}}. \tag{4}$$

Therefore, to construct solutions of $L_\gamma u = 0$ in $\mathbb{R}^n$, it is enough to construct solutions of the Schrödinger equation $(\Delta - q)u = 0$ with q of the form (4). The next result states the existence of complex geometrical optics solutions for the Schrödinger equation associated to any bounded and compactly supported potential.

Theorem 1 ([46, 47]) *Let $q \in L^\infty(\mathbb{R}^n)$, $n \geq 2$, with $q(x) = 0$ for $|x| \geq R > 0$. Let $-1 < \delta < 0$. There exists $\epsilon(\delta)$ and such that for every $\rho \in \mathbb{C}^n$ satisfying $\rho \cdot \rho = 0$ and $\frac{\|(1+|x|^2)^{1/2} q\|_{L^\infty(\mathbb{R}^n)} + 1}{|\rho|} \leq \epsilon$, there exists a unique solution to*

$$(\Delta - q)u = 0$$

of the form

$$u = e^{x\cdot\rho}(1 + \psi_q(x, \rho)) \tag{5}$$

with $\psi_q(\cdot, \rho) \in L^2_\delta(\mathbb{R}^n)$. Moreover $\psi_q(\cdot, \rho) \in H^2_\delta(\mathbb{R}^n)$, and for $0 \leq s \leq 2$ there exists $C = C(n, s, \delta) > 0$ such that $\|\psi_q(\cdot, \rho)\|_{H^s_\delta} \leq \frac{C}{|\rho|^{1-s}}$.

Here $L^2_\delta(\mathbb{R}^n) = \{f; \int(1+|x|^2)^\delta |f(x)|^2 dx < \infty\}$ with the norm given by $\|f\|^2_{L^2_\delta} = \int(1 + |x|^2)^\delta |f(x)|^2 dx$, and $H^m_\delta(\mathbb{R}^n)$ denotes the corresponding Sobolev space. Note that for large $|\rho|$ these solutions behave like Calderón's exponential solutions. If 0 is not a Dirichlet eigenvalue for the Schrödinger equation, we can also define the DN map

$$\Lambda_q(f) = \frac{\partial u}{\partial \nu}|_{\partial\Omega}$$

where u solves

$$(\Delta - q)u = 0; \quad u|_{\partial\Omega} = f.$$

More generally we can define the set of Cauchy data for the Schrödinger equation as the set

$$\mathbb{C}_q = \left\{ \left(u\Big|_{\partial\Omega}, \frac{\partial u}{\partial \nu}\Big|_{\partial\Omega} \right) \right\}, \tag{6}$$

where $u \in H^1(\Omega)$ is a solution of

$$(\Delta - q)u = 0 \text{ in } \Omega. \tag{7}$$

We have $\mathbb{C}_q \subseteq H^{\frac{1}{2}}(\partial\Omega) \times H^{-\frac{1}{2}}(\partial\Omega)$. If 0 is not a Dirichlet eigenvalue of $\Delta - q$, then $\mathbb{C}_q$ is the graph of the DN map.

The Calderón Problem in Dimension $n \geq 3$

The identifiability question in EIT was resolved for smooth enough isotropic conductivities. The result is

Theorem 2 ([47]) *Let $\gamma_i \in C^2(\overline{\Omega})$, γ_i strictly positive, $i = 1, 2$. If $\Lambda_{\gamma_1} = \Lambda_{\gamma_2}$, then $\gamma_1 = \gamma_2$ in $\overline{\Omega}$.*

In dimension $n \geq 3$ this result is a consequence of a more general result. Let $q \in L^\infty(\Omega)$.

Theorem 3 ([47]) *Let $q_i \in L^\infty(\Omega)$, $i = 1, 2$. Assume $\mathbb{C}_{q_1} = \mathbb{C}_{q_2}$, and then $q_1 = q_2$.*

Theorem 2 has been extended to conductivities having $3/2$ derivatives in some sense in [7,42]. Uniqueness for conormal conductivities in $C^{1+\epsilon}$ was shown in [18]. It is an open problem whether uniqueness holds in dimension $n \geq 3$ for Lipschitz or less regular conductivities. For conormal potentials with singularities including almost a delta function of a hypersurface, uniqueness was shown in [18]. The regularity condition on the conductivity was improved recently to C^1 conductivities in [20], to conductivities in $W^{1,n}, n = 3, 4, 5$ in [19] and Lipschitz conductivities in [13] in all dimensions larger than 3. The case of piecewise analytic conductivities has been settled earlier in [31]. Stability for EIT using CGO solutions was shown by Alessandrini [1], and a reconstruction method was proposed by Nachman [36].

Other Applications

We give a short list of other applications to inverse problems using the CGO solutions described above for the Schrödinger equation.

Quantum Scattering

In dimension $n \geq 3$ and in the case of a compactly supported electric potential, uniqueness for the fixed energy scattering problem was proven in [36, 39, 43]. For compactly supported potentials knowledge of the scattering amplitude at fixed energy is equivalent to knowing the Dirichlet-to-Neumann map for the Schrödinger equation measured on the boundary of a large ball containing the support of the potential (see [48] for an account). Then Theorem 3 implies the result. Melrose [35] suggested a related proof that

uses the density of products of scattering solutions. Applications of CGO solutions to the 3-body problem were given in [49].

Optics

The DN map associated to the Helmholtz equation $-\Delta + k^2 n(x)$ with an isotropic index of refraction n determines uniquely a bounded index of refraction in dimension $n \geq 3$.

Optical Tomography in the Diffusion Approximation

In this case we have $\nabla \cdot D(x)\nabla u - \sigma_a(x)u - i\omega u = 0$ in Ω where u represents the density of photons, D the diffusion coefficient, and σ_a the optical absorption. Using Theorem 2 one can show in dimension three or higher that if $\omega \neq 0$, one can recover both D and σ_a from the corresponding DN map. If $\omega = 0$, then one can recover one of the two parameters.

Photoacoustic Tomography

Applications of CGO solutions to quantitative photoacoustic tomography were given in [4, 5].

The Partial Data Problem in Dimension $n \geq 3$

In several applications in EIT, one can only measure currents and voltages on part of the boundary. Substantial progress has been made recently on the problem of whether one can determine the conductivity in the interior by measuring the DN map on part of the boundary.

The paper [10] used the method of Carleman estimates with a linear weight to prove that, roughly speaking, knowledge of the DN map in "half" of the boundary is enough to determine uniquely a C^2 conductivity. The regularity assumption on the conductivity was relaxed to $C^{1+\epsilon}, \epsilon > 0$ in [30]. Stability estimates for the uniqueness result of [10] were given in [21].

The result [10] was substantially improved in [29]. The latter paper contains a global identifiability result where it is assumed that the DN map is measured on any open subset of the boundary of a strictly convex domain for all functions supported, roughly, on the complement. The key new ingredient is the construction of a larger class of CGO solutions than the ones considered in the previous sections. These have the form

$$u = e^{\tau(\phi + i\psi)}(a + r), \tag{8}$$

where $\nabla\phi \cdot \nabla\psi = 0, |\nabla\phi|^2 = |\nabla\psi|^2$, and ϕ are limiting Carleman weights (LCW). Moreover a is smooth and nonvanishing and $\|r\|_{L^2(\Omega)} = O(\frac{1}{\tau})$, $\|r\|_{H^1(\Omega)} = O(1)$. Examples of LCW are the linear phase $\phi(x) = x \cdot \omega, \omega \in S^{n-1}$, used previously, and the nonlinear phase $\phi(x) = \ln|x - x_0|$, where $x_0 \in \mathbb{R}^n \setminus \overline{\text{ch}(\Omega)}$ which was used in [29]. Here ch (Ω) denotes the convex hull of Ω. All the LCW in $\mathbb{R}^n$ were characterized in [17]. In two dimensions any harmonic function is an LCW.

The CGO solutions used in [29] are of the form

$$u(x, \tau) = e^{1/n|x-x_0| + id(\frac{x-x_0}{|x-x_0|}, \omega)}(a + r) \tag{9}$$

where x_0 is a point outside the convex hull of Ω, ω is a unit vector, and $d(\frac{x-x_0}{|x-x_0|}, \omega)$ denotes distance. We take directions ω so that the distance function is smooth for $x \in \overline{\Omega}$. These are called *complex spherical waves* since the level sets of the real part of the phase are spheres centered at x_0. Further applications of these type of waves are given below. A reconstruction method based on the uniqueness proof of [29] was proposed in [38].

The Two-Dimensional Case

In EIT Astala and Päivärinta [2], in a seminal contribution, have extended significantly the uniqueness result of [37] for conductivities having two derivatives in an appropriate sense and the result of [8] for conductivities having one derivative in appropriate sense, by proving that any L^∞ conductivity in two dimensions can be determined uniquely from the DN map. The proof of [2] relies also on the construction of CGO solutions for the conductivity equation with L^∞ coefficients and the $\overline{\partial}$ method. This is done by transforming the conductivity equation to a quasi-regular map.

For the partial data problem, it is shown in [26] that for a two-dimensional bounded domain, the Cauchy data for the Schrödinger equation measured on an arbitrary open subset of the boundary determines uniquely the potential. This implies, for the conductivity equation, that if one measures the current fluxes at the boundary on an arbitrary open subset of the boundary produced by voltage potentials supported in the same subset, one can determine uniquely the conductivity. The paper [26] uses Carleman estimates with weights which are harmonic functions with nondegenerate critical points to construct appropriate complex geometrical optics solutions to prove the result.

For the Schrödinger equation Bukhgeim in a breakthrough [9] proved that a potential in $L^p(\Omega)$, $p > 2$ can be uniquely determined from the set of Cauchy data as defined in (6). Assume now that $0 \in \Omega$. Bukhgeim constructs CGO solutions of the form

$$u_1(z,k) = e^{z^2 k}(1 + \psi_1(z,k)),$$
$$u_2(z,k) = e^{-\bar{z}^2 k}(1 + \psi_2(z,k)) \qquad (10)$$

where $z, k \in \mathbb{C}$, and we have used the complex notation $z = x_1 + i x_2$. Moreover ψ_1 and ψ_2 decay uniformly in Ω, in an appropriate sense, for $|k|$ large. Note that the weight $z^2 k$ in the exponential is a limiting Carleman weight since it is a harmonic function but it has a nondegenerate critical point at 0.

Anisotropic Conductivities

Anisotropic conductivities depend on direction. The muscle tissue in the human body is an important example of an anisotropic conductor. For instance, cardiac muscle has a conductivity of 2.3 mho in the transverse direction and 6.3 in the longitudinal direction. The conductivity in this case is represented by a positive definite, smooth, symmetric matrix $\gamma = (\gamma^{ij}(x))$ on Ω.

Under the assumption of no sources or sinks of current in Ω, the potential u in Ω, given a voltage potential f on $\partial\Omega$, solves the Dirichlet problem

$$\sum_{i,j=1}^{n} \frac{\partial}{\partial x_i}\left(\gamma^{ij}\frac{\partial u}{\partial x_j}\right) = 0 \text{ on } \Omega, \quad u|_{\partial\Omega} = f. \qquad (11)$$

The DN map is defined by

$$\Lambda_\gamma(f) = \sum_{i,j=1}^{n} \nu^i \gamma^{ij} \frac{\partial u}{\partial x_j}\Big|_{\partial\Omega} \qquad (12)$$

where $\nu = (\nu^1, \ldots, \nu^n)$ denotes the unit outer normal to $\partial\Omega$ and u is the solution of (11). The inverse problem is whether one can determine the matrix γ by knowing Λ_γ. Unfortunately, Λ_γ does not determine γ uniquely. Let $\psi : \overline{\Omega} \to \overline{\Omega}$ be a C^∞ diffeomorphism with $\psi|_{\partial\Omega} = \mathrm{Id}$ where Id denotes the identity map. We have

$$\Lambda_{\tilde{\gamma}} = \Lambda_\gamma \qquad (13)$$

where

$$\tilde{\gamma} = \left(\frac{(D\psi)^T \circ \gamma \circ (D\psi)}{|\det D\psi|}\right) \circ \psi^{-1}. \qquad (14)$$

Here $D\psi$ denotes the (matrix) differential of ψ, $(D\psi)^T$ its transpose, and the composition in (14) is to be interpreted as multiplication of matrices.

We have then a large number of conductivities with the same DN map: any change of variables of Ω that leaves the boundary fixed gives rise to a new conductivity with the same electrostatic boundary measurements. The question is then whether this is the only obstruction to unique identifiability of the conductivity.

In two dimensions this has been shown for $L^\infty(\Omega)$ conductivities in [3]. This is done by reducing the anisotropic problem to the isotropic one by using isothermal coordinates and using Astala and Päivärinta's result in the isotropic case [2]. Earlier results were for C^3 conductivities using the result of Nachman [37], for Lipschitz conductivities in [44] using the techniques of [8], and [45] for anisotropic conductivities close to constant.

In three or more dimensions, this has been shown for real-analytic conductivity ion domains with real-analytic boundary. In fact this problem admits a geometric formulation on manifolds [34], and it has been proven for real-analytic manifolds with boundary [32]. New CGO solutions were constructed in [17] for anisotropic conductivities or metrics for which roughly speaking the metric or conductivity is Euclidean in one direction.

Full Maxwell's Equations

Inverse Boundary Value Problems

In the present section, we consider the inverse boundary value problems for the full time-harmonic Maxwell's equations in a bounded domain, that is, to reconstruct three key electromagnetic parameters: electric permittivity $\varepsilon(x)$, conductivity $\sigma(x)$, and magnetic permeability $\mu(x)$, as functions of the spatial variables, from a specified set of electromagnetic field measurements taken on the boundary. To be more specific, let $E(x)$ and $H(x)$ denote the time-harmonic electric and magnetic fields inside the domain $\Omega \subset \mathbb{R}^3$. At the frequency $\omega > 0$, E and H satisfy the time-harmonic Maxwell's equations

$$\nabla \times E = i\omega\mu H, \quad \nabla \times H = -i\omega\gamma E \qquad (15)$$

where $\gamma(x) = \varepsilon(x) + i\sigma(x)$. Assume that the parameters are L^∞ functions in Ω and, for some positive constants $\varepsilon_m, \varepsilon_M, \mu_m, \mu_M$, and σ_M,

$$\varepsilon_m \le \varepsilon(x) \le \varepsilon_M, \quad \mu_m \le \mu(x) \le \mu_M,$$
$$0 \le \sigma(x) \le \sigma_M \quad \text{for } x \in \overline{\Omega}. \tag{16}$$

To introduce the solution space, we define

$$H^1_{\mathrm{Div}}(\Omega) := \left\{ u \in \left(H^1(\Omega)\right)^3 \;\middle|\; \mathrm{Div}(\nu \times u|_{\partial\Omega}) \in H^{1/2}(\partial\Omega) \right\}$$

where on the boundary $\partial\Omega$, ν is the outer normal unit vector and Div denotes the surface divergence. Let $TH^{1/2}_{\mathrm{Div}}(\partial\Omega)$ denote the Sobolev space obtained by taking natural tangential traces of functions in $H^1_{\mathrm{Div}}(\Omega)$ on the boundary. It is well-known that (15) admits a unique solution $(E, H) \in H^1_{\mathrm{Div}}(\Omega) \times H^1_{\mathrm{Div}}(\Omega)$ with imposed boundary electric (or magnetic) condition $\nu \times E = f \in TH^{1/2}_{\mathrm{Div}}(\partial\Omega)$ (or $\nu \times H = g \in TH^{1/2}_{\mathrm{Div}}(\partial\Omega)$), except for a discrete set of resonant frequencies $\{\omega_n\}$ in the dissipative case, namely, $\sigma = 0$.

Then the inverse boundary value problem is to recover ε, σ, and μ from the boundary measurements encoded as the well-defined impedance map

$$\Lambda^\omega : TH^{1/2}_{\mathrm{Div}}(\partial\Omega) \to TH^{1/2}_{\mathrm{Div}}(\partial\Omega)$$
$$f = \nu \times E|_{\partial\Omega} \mapsto \nu \times H|_{\partial\Omega}.$$

We remark that the impedance map Λ^ω is a natural analogue of the Dirichlet-to-Neumann map for EIT, since it carries enough information of the electromagnetic energy associated to the system.

The underlying problem was first formulated in [15], and a local uniqueness result was obtained based on Calderón's linearization idea, that is, the parameters that are slightly perturbed from constants can be uniquely determined by the impedance map. For the global uniqueness and reconstruction of the parameters, the following result was proved in [41], and the proof was simplified later in [40] by introducing the so-called generalized Sommerfeld potentials.

Theorem 4 ([40, 41]) *Let $\Omega \subset \mathbb{R}^3$ be an open bounded domain with a $C^{1,1}$-boundary and a connected complement $\mathbb{R}^3\backslash\overline{\Omega}$. Assume that ε, σ, and μ are in $C^3(\mathbb{R}^3)$ satisfying the condition (16) in Ω and $\varepsilon(x) = \varepsilon_0$, $\mu(x) = \mu_0$, and $\sigma(x) = 0$ when $x \in \mathbb{R}^3\backslash\overline{\Omega}$ for some constants ε_0 and μ_0. Assume that $\omega > 0$ is not a resonant frequency. Then the knowledge of Λ^ω determines the functions ε, σ, and μ uniquely. Recently, the regularity assumed in this result for the electromagnetic parameters has been improved to C^1 [14].*

A closely related problem to the one considered here is the inverse scattering problem of electromagnetism, that is, to reconstruct the unknown parameters from the far-field pattern of the scattered electromagnetic fields. It is shown in [16] that the refractive index $n(x)$ (corresponding to, e.g., known constant μ but unknown $\varepsilon(x)$ and $\sigma(x)$) can be uniquely determined by the far-field patterns of scattered electric fields satisfying

$$\nabla \times \nabla \times E - k^2 n(x) E = 0.$$

The approach is based on the ideas in [47] of constructing CGO type of solutions of the form $E = e^{ix\cdot\zeta}(\eta + R_\zeta)$ where $\zeta, \eta \in \mathbb{C}^3$, $\zeta \cdot \zeta = k^2$, and $\zeta \cdot \eta = 0$.

For Maxwell's equations (15), more generalized solutions of such type were constructed in [41] as follows.

Proposition 1 ([41]) *Suppose the parameters ε, σ, and μ satisfy the condition in Theorem 4. Let $\eta, \theta, and \zeta \in \mathbb{C}^3$ satisfy $\zeta \cdot \zeta = \omega^2$, $\zeta \times \eta = \omega\mu_0\theta$, and $\zeta \times \theta = -\omega\mu_0\eta$. Then for $|\zeta|$ large enough, the Maxwell's equation (15) admits a unique global solution (E, H) of the form*

$$E = e^{ix\cdot\zeta}(\eta + R_\zeta) \quad H = e^{ix\cdot\zeta}(\theta + Q_\zeta) \tag{17}$$

where $R_\zeta(x)$ and $Q_\zeta(x)$ belong to $(L^2_{-\delta}(\mathbb{R}^3))^3$ for $\delta \in [\frac{1}{2}, 1]$.

However, such vector CGO type solutions for both [16] and [41] do not have the property that R_ζ decays like $O(|\zeta|^{-1})$, which was a key ingredient in the proof of the uniqueness in the scalar case. The nature of this difficulty is that the vector-valued analogue of Faddeev's fundamental solution (for the scalar Schrödinger equation), used in the construction of (17), does not share the decaying property of it. In [16], this is tackled by constructing R_ζ that decays to zero in certain distinguished directions as $|\zeta|$ tends to infinity. By rotations,

such special set of solutions are enough to determine the refractive index.

In [41], the approach to the final proof of uniqueness starts with the following identity obtained integrating by parts

$$\int_{\partial\Omega} \nu \times E \cdot \overline{H_0} + \Lambda^{\omega}(\nu \times E|_{\partial\Omega}) \cdot \overline{E_0} dS = i\omega \int_{\Omega} (\mu - \mu_0) H \cdot \overline{H_0} - (\gamma - \varepsilon_0) E \cdot \overline{E_0} dx \quad (18)$$

where (E, H) is an arbitrary solution of (15), while (E_0, H_0) is a solution in the free space where $\varepsilon = \varepsilon_0$, $\sigma = 0$, and $\mu = \mu_0$. It is shown that if one let $|\zeta|$ tend to infinity along a certain manifold (similar to the choices of directions and by rotations in [16]), the right-hand side of (18) has the asymptotic to be a nonlinear functional of unknown parameters ε, σ, and μ. It results in a semilinear elliptic equation of the parameters, and their uniqueness is a direct corollary of the unique continuation principle.

On the other hand, the article [40] reduces significantly the asymptotic estimates used in [41] by an augmenting technique, in which the Maxwell's equations are transformed into a matrix Schrödinger equation. To be more specific, denoting scalar functions $\Phi = \frac{i}{\omega}\nabla \cdot \gamma E$ and $\Psi = \frac{i}{\omega}\nabla \cdot \mu H$, we consider the following rescalization

$$X := \left(\frac{1}{\omega\gamma\mu^{1/2}}\Phi, \gamma^{1/2}E, \mu^{1/2}H, \frac{1}{\omega\mu\gamma^{1/2}}\Psi \right)^T \in (\mathcal{D}')^8. \quad (19)$$

Such rescalization is particularly chosen so that one has, under conditions on Φ and Ψ, the equivalence between Maxwell's equations (15) and a Dirac system about X

$$(P(i\nabla) - k + V)\, X = 0, \qquad P(i\nabla) := i \begin{pmatrix} 0 & \nabla\cdot & 0 & 0 \\ \nabla & 0 & \nabla\times & 0 \\ 0 & -\nabla\times & 0 & \nabla\cdot \\ 0 & 0 & \nabla\cdot & 0 \end{pmatrix} \quad (20)$$

where $k = \omega(\varepsilon_0\mu_0)^{1/2}$ and $V \in (C^{\infty}(\mathbb{R}^3))^8$ (Here we assume the unknown parameters are C^{∞}). For a more detailed argument on the rescalization, we refer the readers to [12, 28]. Moreover the operator $(P(i\nabla) - k + V)$ is related to the matrix Schrödinger operator by

$$(P(i\nabla) - k + V)\left(P(i\nabla) + k - V^T\right) = -(\Delta + k^2)\mathbf{1}_8 + Q \quad (21)$$

where $\mathbf{1}_8$ is the identity matrix and the potential $Q \in \left(C^{\infty}(\mathbb{R}^3)\right)^{8\times 8}$ is compactly supported. Therefore, the generalized Sommerfeld potential Y defined by $X = \left(P(i\nabla) + k - V^T\right) Y$ satisfies the Schrödinger equation

$$-(\Delta + k^2)Y + QY = 0, \quad (22)$$

for which we can construct the CGO solution for some constant vector $y_{0,\zeta}$

$$Y_{\zeta} = e^{ix\cdot\zeta}(y_{0,\zeta} + v_{\zeta}) \quad (23)$$

where v_{ζ} decays to zero as $O(|\zeta|^{-1})$. The rest of the proof is based on the identity

$$-i \int_{\partial\Omega} Y_0^* \cdot P(\nu) X dS = \int_{\Omega} Y_0^* \cdot QY dx \quad (24)$$

where Y_0^* annihilates $P(i\nabla) + k$ and $P(\nu)$ is the matrix with $i\nabla$ replaced by ν in $P(i\nabla)$. Then substitute the CGO solution Y_{ζ} into the identity, and let Y_0^* depend on ζ in an appropriate way. Taking $|\zeta|$ to infinity, the left-hand side of (24) can be computed from the impedance map Λ^{ω}, and the right-hand side converges to functionals of Q. Such functionals carry the information of the unknown parameters, and the reconstruction of each of them is possible when proper directions, along which ζ diverges, are chosen.

For the partial data problem, namely, to determine the parameters from the impedance map only made on part of the boundary, there are not as many results as in the scalar case. It is shown in [12] that if the measurements $\Lambda^{\omega}(f)$ are taken only on a nonempty open subset Γ of $\partial\Omega$ for $f = \nu \times E|_{\partial\Omega}$ supported in γ, where the inaccessible part $\overline{\partial\Omega \setminus \Gamma}$ is part of a plane or a sphere, the electromagnetic parameters can still be uniquely determined. Combined with the augmenting

argument in [40], the proof in [12] generalized the reflection technique used in [27], where the restriction on the shape of the inaccessible part comes from. As for another well-known method in dealing with partial data problems based on the Carleman estimates [10,29], there are however significant difficulties in generalizing the method to the full system of Maxwell's equations, e.g., the CGO solutions constructed using Carleman estimates.

In the anisotropic setting, where the electromagnetic parameters depend on direction and are regarded as matrix-valued functions, one of the uniqueness results was obtained in [28] for Maxwell's equations on certain admissible Riemannian manifolds. Such manifold has a product structure and includes compact manifolds in Euclidean space, hyperbolic space and $\mathbb{S}^3$ minus a point, and also sufficiently small sub-manifolds of conformally flat manifolds as examples. A construction of CGO solutions based on direct Fourier arguments was provided with a suitable uniqueness result.

Identifying Electromagnetic Obstacles by the Enclosure Method

As another application of the important CGO solutions for scalar conductivity equations and Helmholtz equations, in [24], the enclosure method was introduced to determine the shape of an obstacle or inclusion embedded in a bounded domain with known background parameters like conductivity or sound speed, from the boundary measurements of electric currents or sound waves. The fundamental idea of this method is to implement the low penetrating ability of CGO plane waves due to its rapidly decaying property away from the key planes. The energies associated with such waves show little evidence of the existence of the inclusion unless the key planes have intersection with it. These planes will enclose the inclusion from each direction, and the convex hull can be reconstructed. The method was improved in [23] by the complex spherical waves constructed in [29] to enclose some non-convex part of the shape of electrostatic inclusions. For the application on more generalized systems of two variables, in which case more choices of CGO solutions are available, we refer the article [51]. Numerical simulations of the approach were done in [23,25].

For the full time-harmonic system of Maxwell's equations, the enclosure method is generalized in [53] to identify the electromagnetic obstacles embedded in lossless background media. Suppose the obstacle D satisfies $\overline{D} \subset \Omega$ and $\Omega \backslash \overline{D}$ is connected. It is embedded in a lossless electromagnetic medium, and therefore the EM fields in $\Omega \backslash \overline{D}$ satisfy

$$\nabla \times E = i\omega\mu H, \quad \nabla \times H = -i\omega\varepsilon E, \tag{25}$$

with perfect magnetic obstacle condition $\nu \times H|_{\partial D} = 0$. With well-defined boundary impedance map denoted by Λ_D^ω on $\partial\Omega$ for nonresonant frequency ω, the inverse problem aims to recover the convex hull of D. The candidates of the probing waves are among the CGO solutions for the background medium, of the form

$$\begin{aligned} E_0 &= \varepsilon^{1/2} e^{\tau(x\cdot\rho - t) + i\sqrt{\tau^2+\omega^2}\, x\cdot\rho^\perp}(\eta + R_\tau), \\ H_0 &= \mu^{1/2} e^{\tau(x\cdot\rho - t) + i\sqrt{\tau^2+\omega^2}\, x\cdot\rho^\perp}(\theta + Q_\tau) \end{aligned} \tag{26}$$

where the planes used to enclose the obstacle are level sets $\{x \cdot \rho = t\}$. It is possible to compute, from the impedance map Λ_D^ω, an energy difference between two systems: the domain with obstacle and the background domain without an obstacle, for the same boundary CGO inputs. This is denoted as an indicator function given by

$$I_\rho(\tau, t) := i\omega \int_{\partial\Omega} (\nu \times E_0) \cdot \overline{(\Lambda_D^\omega - \Lambda_\emptyset^\omega)(\nu \times E_0) \times \nu} dS. \tag{27}$$

Since that as $\tau \to \infty$, the CGO EM fields (26) decay to zero exponentially on the half space $\{x \cdot \rho < t\}$ and grow exponentially on the other half, and one would expect $\lim_{\tau\to\infty} I_\rho(\tau, t) = 0$, i.e., no energy detection, as long as D stays in $\{x \cdot \rho < t\}$. On the other hand, if D has any intersection with the opposite closed half space $\{x \cdot \rho \geq 0\}$, the limit should not any longer be small. This provides a way by testing different $\rho \in \mathbb{S}^2$ and $t > 0$ to detect where the boundary of D lies. However, for the full system of Maxwell's equation, a difficulty arises when showing the nonvanishing property of the indicator function in the latter case. This is again mainly because that the CGO solutions' remainder terms R_τ and Q_τ do not decay. To address this, one can choose the relatively free incoming constant fields $\eta = \eta_\tau$ and $\theta = \theta_\tau$ that share different asymptotic speeds as τ tends to infinity.

In this way, one can prove that the lower bound of the indicator function is dominated by the CGO magnetic energy in D, which is never vanishing. Hence the enclosure method is developed. We would like to point out that in [53], the construction of CGO solutions for the system is based on the augmenting technique in [40] and the choice of constant fields η_τ and θ_τ is similar to that in [16, 40, 41].

A natural improvement of the enclosure method as in the scalar case is to examine the reconstruction of the non-convex part of the shape of D. The complex spherical waves constructed in [29] using Carleman estimates are CGO solutions with nonlinear phase $\ln|x - x_0|$ where $x_0 \in \mathbb{R}^3\backslash\overline{\Omega}$, with spherical level sets. When replacing the linear-phase-CGO solutions in the enclosure method by complex spherical waves, the obstacle or the inclusion is enclosed by the exterior of spheres. However, for Maxwell's equations, the Carleman estimate argument has not been carried out yet. Instead, it is shown, in [53], that one can implement the Kelvin transformation

$$T : x \mapsto R^2 \frac{x - x_0}{|x - x_0|^2} + x_0, \quad x_0 \in \mathbb{R}^3\backslash\overline{\Omega}, \ R > 0,$$

which maps spheres passing x_0 to planes. The invariance of Maxwell's equations under T makes it possible to compute the impedance map associated to the image domain $T(\Omega)$ and apply the enclosure method there with linear-phase-CGO solutions. This is equivalent to enclosing in the original domain with spheres, which are pre-images of the planes. We notice that the pullbacks of the linear-phase-CGO fields in the image space are complex spherical fields in the original space with LCW

$$\varphi(x) = R^2 \frac{(x - x_0) \cdot \rho}{|x - x_0|^2} + x_0 \cdot \rho.$$

References

1. Alessandrini, G.: Stable determination of conductivity by boundary measurements. Appl. Anal. **27**, 153–172 (1988)
2. Astala, K., Päivärinta, L.: Calderón inverse conductivity problem in the plane. Ann. Math. **163**, 265–299 (2006)
3. Astala, K., Lassas, M., Päiväirinta, L.: Calderón inverse problem for anisotropic conductivity in the plane. Commun. Partial Diff. Eqn. **30**, 207–224 (2005)
4. Bal, G., Uhlmann, G.: Inverse diffusion theory of photoacoustics. Inverse Probl. **26**, 085010 (2010)
5. Bal, G., Ren, K., Uhlmann, G., Zhou, T.: Quantitative thermo-acoustics and related problems. Inverse Probl. **27** (2011), 055007
6. Barceló, J.A., Faraco, D., Ruiz, A.: Stability of Calderón inverse problem in the plane. J. des Math. Pures Appl. **88**(6), 522–556 (2007)
7. Brown, R., Torres, R.: Uniqueness in the inverse conductivity problem for conductivities with $3/2$ derivatives in L^p, $p > 2n$. J. Fourier Anal. Appl. **9**, 1049–1056 (2003)
8. Brown, R., Uhlmann, G.: Uniqueness in the inverse conductivity problem with less regular conductivities in two dimensions. Commun. PDE **22**, 1009–10027 (1997)
9. Bukhgeim, A.: Recovering the potential from Cauchy data in two dimensions. J. Inverse Ill-Posed Probl. **16**, 19–34 (2008)
10. Bukhgeim, L., Uhlmann, G.: Recovering a potential from partial Cauchy data. Commun. PDE **27**, 653–668 (2002)
11. Calderón, A.P.: On an inverse boundary value problem. In: Seminar on Numerical Analysis and Its Applications to Continuum Physics, Río de Janeiro, pp. 65–73. Sociedade Brasileira de Matematica, Rio de Janeiro (1980)
12. Caro, P., Ola, P., Salo, M.: Inverse boundary value problem for Maxwell equations with local data. Commun. PDE **34**, 1425–1464 (2009)
13. Caro, P., Rogers, K.: Global uniqueness for the Calderón problem with Lipschitz conductivities. arXiv: 1411.8001
14. Caro, P., Zhou, T.: On global uniqueness for an IBVP for the time-harmonic Maxwell equations. Analysis & PDE, **7**(2), 375–405 (2014)
15. Cheney, M., Isaacson, D., Somersalo, E.: A linearized inverse boundary value problem for Maxwell's equations. J. Comput. Appl. Math. **42**, 123–136 (1992)
16. Colton, D., Päivärinta, L.: The uniqueness of a solution to an inverse scattering problem for electromagnetic waves. Arch. Ration. Mech. Anal. **119**, 59–70 (1992)
17. Dos Santos Ferreira, D., Kenig, C.E., Salo, M., Uhlmann, G.: Limiting Carleman weights and anisotropic inverse problems. Invent. Math. **178**, 119–171 (2009)
18. Greenleaf, A., Lassas, M., Uhlmann, G.: The Calderón problem for conormal potentials, I: global uniqueness and reconstruction. Commun. Pure Appl. Math. **56**, 328–352 (2003)
19. Haberman, B.: Uniqueness in Calderón's problem for conductivities with unbounded gradient, arXiv:1410.2201
20. Haberman, B., Tataru, D.: Uniqueness in Calderón's problem with Lipschitz conductivities. Duke Math J., **162**, 497–516 (2013)
21. Heck, H., Wang, J.-N.: Stability estimates for the inverse boundary value problem by partial Cauchy data. Inverse Probl. **22**, 1787–1796 (2006)
22. Holder, D.: Electrical Impedance Tomography. Institute of Physics Publishing, Bristol/Philadelphia (2005)
23. Ide, T., Isozaki, H., Nakata, S., Siltanen, S., Uhlmann, G.: Probing for electrical inclusions with complex spherical waves. Commun. Pure. Appl. Math. **60**, 1415–1442 (2007)
24. Ikehata, M.: How to draw a picture of an unknown inclusion from boundary measurements: two mathematical inversion algorithms. J. Inverse Ill-Posed Probl. **7**, 255–271 (1999)
25. Ikehata, M., Siltanen, S.: Numerical method for finding the convex hull of an inclusion in conductivity from boundary measurements. Inverse Probl. **16**, 1043–1052 (2000)

26. Imanuvilov, O., Uhlmann, G., Yamamoto, M.: The Calderón problem with partial data in two dimensions. J. Am. Math. Soc. **23**, 655–691 (2010)
27. Isakov, V.: On uniqueness in the inverse conductivity problem with local data. Inverse Probl. Imaging **1**, 95–105 (2007)
28. Kenig, C., Salo, M., Uhlmann, G.: Inverse problem for the anisotropic Maxwell equations. Duke Math. J. **157**(2), 369–419 (2011)
29. Kenig, C., Sjöstrand, J., Uhlmann, G.: The Calderón problem with partial data. Ann. Math. **165**, 567–591 (2007)
30. Knudsen, K.: The Calderón problem with partial data for less smooth conductivities. Commun. Partial Differ. Equ. **31**, 57–71 (2006)
31. Kohn R., and Vogelius M.: Determining conductivity by boundary measurements II. Interior results. Comm. Pure Appl. Math. **38**, 643–667 (1985)
32. Lassas, M., Uhlmann, G.: Determining a Riemannian manifold from boundary measurements. Ann. Sci. École Norm. Sup. **34**, 771–787 (2001)
33. Lassas, M., Taylor, M., Uhlmann, G.: The Dirichlet-to-Neumann map for complete Riemannian manifolds with boundary. Commun. Geom. Anal. **11**, 207–222 (2003)
34. Lee, J., Uhlmann, G.: Determining anisotropic real-analytic conductivities by boundary measurements. Commun. Pure Appl. Math. **42**, 1097–1112 (1989)
35. Melrose, R.B.: Geometric Scattering Theory. Cambridge University Press, Cambridge/New York (1995)
36. Nachman, A.: Reconstructions from boundary measurements. Ann. Math. **128**, 531–576 (1988)
37. Nachman, A.: Global uniqueness for a two-dimensional inverse boundary value problem. Ann. Math. **143**, 71–96 (1996)
38. Nachman, A., Street, B.: Reconstruction in the Calderón problem with partial data. Commun. PDE **35**, 375–390 (preprint)
39. Novikov, R.G.: Multidimensional inverse spectral problems for the equation $-\Delta\psi + (v(x) - Eu(x))\psi = 0$. Funktsionalny Analizi Ego Prilozheniya **22**, 11–12, Translation in Funct. Anal. Appl. **22**, 263–272 (1988)
40. Ola, P., Someralo, E.: Electromagnetic inverse problems and generalized Sommerfeld potential. SIAM J. Appl. Math. **56**, 1129–1145 (1996)
41. Ola, P., Päivärinta, L., Somersalo, E.: An inverse boundary value problem in electrodynamics. Duke Math. J. **70**, 617–653 (1993)
42. Päivärinta, L., Panchenko, A., Uhlmann, G.: Complex geometrical optics for Lipschitz conductivities. Rev. Mat. Iberoam. **19**, 57–72 (2003)
43. Ramm, A.: Recovery of the potential from fixed energy scattering data. Inverse Probl. **4**, 877–886 (1988)
44. Sun, Z., Uhlmann, G.: Anisotropic inverse problems in two dimensions. Inverse Probl. **19**, 1001–1010 (2003)
45. Sylvester, J.: An anisotropic inverse boundary value problem. Commun. Pure Appl. Math. **43**, 201–232 (1990)
46. Sylvester, J., Uhlmann, G.: A uniqueness theorem for an inverse boundary value problem in electrical prospection. Commun. Pure Appl. Math. **39**, 92–112 (1986)
47. Sylvester, J., Uhlmann, G.: A global uniqueness theorem for an inverse boundary value problem. Ann. Math. **125**, 153–169 (1987)
48. Uhlmann, G.: Inverse boundary value problems and applications. Astérisque **207**, 153–211 (1992)
49. Uhlmann, G., Vasy, A.: Low-energy inverse problems in three-body scattering. Inverse Probl. **18**, 719–736 (2002)
50. Uhlmann, G., Wang, J.-N.: Complex spherical waves for the elasticity system and probing of inclusions. SIAM J. Math. Anal. **38**, 1967–1980 (2007)
51. Uhlmann, G., Wang, J.-N.: Reconstruction of discontinuities using complex geometrical optics solutions. SIAM J. Appl. Math. **68**, 1026–1044 (2008)
52. Zhdanov, M.S., Keller, G.V.: The Geoelectrical Methods in Geophysical Exploration. Methods in Geochemistry and Geophysics, vol. 31. Elsevier, Amsterdam/New York (1994)
53. Zhou, T.: Reconstructing electromagnetic obstacles by the enclosure method. Inverse Probl. Imaging **4**, 547–569 (2010)
54. Zou, Y., Guo, Z.: A review of electrical impedance techniques for breast cancer detection. Med. Eng. Phys. **25**, 79–90 (2003)

Inverse Nodal Problems: 1-D

Chun-Kong Law
Department of Applied Mathematics, National Sun Yat-sen University, Kaohsiung, Taiwan

Mathematics Subject Classification

34A55; 34B24

Synonyms

Inverse Nodal Problems 2-D; Inverse Spectral Problems 1-D Theoretical Results; Inverse Spectral Problems 2-D: Theoretical Results; Inverse Spectral Problems 1-D Algorithms; Multidimensional Inverse Spectral Problems; Regularization of Inverse Problems

Glossary

Nodal data A set of nodal points (zeros) of all eigenfunctions.

Nodal length The distance between two consecutive nodal points of one eigenfunction.

Quasinodal set A double sequence $\{x_k^n\}$ that satisfies the asymptotic behavior as given in (2).

Short Definition

This is the inverse problem of recovering parameters in a Sturm-Liouville-type equation using the nodal data.

Description

Consider the Sturm-Liouville operator H:

$$Hy = -y'' + q(x)y \,, \tag{1}$$

with boundary conditions

$$\begin{cases} y(0)\cos\alpha + y'(0)\sin\alpha = 0 \\ y(1)\cos\beta + y'(1)\sin\beta = 0 \end{cases}.$$

Here $q \in L^1(0,1)$ and $\alpha, \beta \in [0,\pi)$. Let λ be the nth eigenvalue of the operator H and $0 < x_1^{(n)} < x_2^{(n)} < \cdots < x_{n-1}^{(n)} < 1$ be the $(n-1)$ nodal points of the nth eigenfunction. The double sequence $\{x_k^{(n)}\}$ is called the *nodal set* associated with H. Also, let $l_k^{(n)} = x_{k+1}^{(n)} - x_k^{(n)}$ be the associated *nodal length*. We define the function $j_n(x)$ on $(0,1)$ by $j_n(x) = \max\{k : x_k^{(n)} \le x\}$. Hence, if x and n are fixed, then $j = j_n(x)$ implies $x \in [x_j^{(n)}, x_{j+1}^{(n)})$.

In many applications, certain nodal set associated with a potential can be measured. Hence, it is desirable to recover the potential with this nodal set. The inverse nodal problem was first defined by McLaughlin [19]. She showed that knowledge of the nodal points alone can determine the potential function in $L^2(0,1)$ up to a constant. Up till now, the issues of uniqueness, reconstruction, smoothness, and stability are all solved for $q \in L^1(0,1)$ [14, 16, 19, 25].

Reconstruction Formula, Smoothness, and Stability

For simplicity, we consider the Dirichlet boundary conditions $\alpha = \beta = 0$. We can turn the Sturm-Liouville equation into the integral equation

$$y(x) = \frac{\sin sx}{s} + \frac{1}{s}\int_0^x \sin[s(x-t)]q(t)y(t)\,dt,$$

for a solution y, satisfying $y(0) = 0$, $y'(0) = 1$. After an iteration and some trigonometric calculations, when $y(x) = 0$ and $\cos(sx)$ is not close to 0,

$$\tan(sx) = \frac{1}{2s}\int_0^x (1 - \cos(2st))q(t)\,dt + o\left(\frac{1}{s^2}\right).$$

From this, one can easily derive asymptotic estimates of the parameters $s_n = \sqrt{\lambda_n}$ and $x_k^{(n)}$, by letting $x = 1$ and $x = x_k^{(n)}$, respectively:

$$\begin{aligned} s_n &= n\pi + \frac{1}{2s_n}\int_0^1 (1 - \cos(2s_n t))q(t)dt \\ &\quad + o\left(\frac{1}{s_n^2}\right) \\ x_k^{(n)} &= \frac{k\pi}{s_n} + \frac{1}{2s_n^2}\int_0^{x_k^{(n)}} (1 - \cos(2s_n t))\,q(t)\,dt \\ &\quad + o\left(\frac{1}{s_n^3}\right). \end{aligned} \tag{2}$$

Hence, the nodal length is given by

$$l_k^{(n)} = \frac{\pi}{s_n} + \frac{1}{2s_n^2}\int_{x_k^{(n)}}^{x_{k+1}^{(n)}} (1 - \cos(2s_n t))q(t)\,dt + o\left(\frac{1}{s_n^3}\right),$$

from which, one arrives at

$$\begin{aligned} &2s_n^2\left(\frac{s_n l_{j_n(x)}^{(n)}}{\pi} - 1\right) \\ &= \frac{s_n}{\pi}\int_{x_k^{(n)}}^{x_{k+1}^{(n)}} (1 - \cos(2s_n t))q(t)\,dt + o(1) \\ &\to q(x) \end{aligned}$$

where the convergence is pointwise a.e. as well as L^1. If we put in the asymptotic expression of s_n, we obtain that pointwisely a.e. and in L^1 [7, 16],

$$q(x) = \lim_{n\to\infty} 2n^2\pi^2\left(n l_{j_n(x)}^{(n)} - 1 + \frac{l_{j_n(x)}^{(n)}}{2n\pi^2}\int_0^1 q\right). \tag{3}$$

Thus, given the nodal set plus the constant $\int_0^1 q$, one can recover the potential function. Unlike most inverse spectral problems, the reconstruction formula here is direct and explicit. However, the problem

is overdetermined, as the limit does not require starting terms.

On the other hand, we define the difference quotient operator δ as follows:

$$\delta a_i^{(n)} = \frac{a_{i+1}^{(n)} - a_i^{(n)}}{x_{i+1}^{(n)} - x_i^{(n)}} = \frac{\Delta a_i^{(n)}}{l_i^{(n)}}\,; \quad \text{and}$$

$$\delta^k a_i^{(n)} = \frac{\delta^{k-1} a_{i+1}^{(n)} - \delta^{k-1} a_i^{(n)}}{l_i^{(n)}}.$$

Hence, the above reconstruction formula, whose term is a step function, can be linked up to a continuous function

$$F_n^{(0)}(x) = 2n^2\pi^2 \left\{ \left(n + \frac{1}{2n\pi^2} \int_0^1 q \right) \left(l_{j_n(x)}^{(n)} + \delta l_{j_n(x)}^{(n)} \cdot \left(x - x_{j_n(x)}^{(n)} \right) \right) - 1 \right\}.$$

Furthermore, we let

$$F_n^{(k)}(x) = 2n^3\pi^2 \left\{ \delta^k l_j^{(n)} + \delta^{k+1} l_j^{(n)} \cdot (x - x_j^{(n)}) \right\}.$$

With these definitions, one can show the following theorem [15, 16].

Theorem 1 *Suppose q is C^{N+1} on $[0,1]$ $(N \geq 1)$. Then for each $x \in (0,1)$ and $k = 0, \ldots, N$, as $n \to \infty$,*

$$q^{(k)}(x) = F_n^{(k)}(x) + O\left(\frac{1}{n}\right).$$

Conversely, if $F_n^{(k)}$ is uniformly convergent on compact subsets of $(0,1)$, for each $k = 1, \ldots, N$, then q is C^N on $(0,1)$, and $F_n^{(k)}$ is uniformly convergent to $q^{(k)}$ on compact subsets of $(0,1)$.

The proof depends on the definition of $\delta^k a_i$. Let $G^{(1)} = \lim_{n\to\infty} F_n^{(1)}$. Then

$$\begin{aligned}
&\int_0^x G^{(1)}(t)\,dt = \lim_{n\to\infty} \int_0^x F_n^{(1)}(t)\,dt \\
&= \lim_{n\to\infty} 2n^3\pi^2 \int_0^x \left[\delta l_{j_n(t)}^{(n)} + \delta^2 l_{j_n(t)}^{(n)} \cdot \left(t - x_{j_n(t)} \right) \right] dt \\
&= \lim_{n\to\infty} 2n^3\pi^2 \sum_{k=1}^{j-1} \left[l_k^{(n)} \delta l_k^{(n)} + \frac{1}{2} \left(x_{k+1}^{(n)} - x_k^{(n)} \right)^2 \delta^2 l_k^{(n)} \right] \\
&= \lim_{n\to\infty} n^3\pi^2 \sum_{i=1}^{j-1} l_k^{(n)} \left(\delta l_k^{(n)} + \delta l_{k+1}^{(n)} \right) \\
&= \lim_{n\to\infty} n^3\pi^2 \sum_{i=1}^{j-1} \left(l_{k+2}^{(n)} - l_k^{(n)} \right) \\
&= \lim_{n\to\infty} 2n^3\pi^2 \left(l_j^{(n)} - l_1^{(n)} \right) \\
&= q(x) - q(0),
\end{aligned}$$

using the facts such as $l_k^{(n)} \delta l_k^{(n)} = l_{k+1}^{(n)} - l_k^{(n)}$,

$$l_{k+1}^{(n)} - l_k^{(n)} = o\left(\frac{1}{n^3}\right), \qquad \text{and} \qquad \delta l_k^{(n)} = o\left(\frac{1}{n^2}\right).$$

The rest of the proof is similar.

Next, we would like to add that this inverse nodal problem is also stable [14]. Let X and $\overline{X}$ be the nodal sets associated with the potential function q and $\overline{q}$ respectively. Define

$$S_n(X, \overline{X}) := n^2\pi^2 \sum_{k=0}^{n-1} |l_k^{(n)} - \overline{l}_k^{(n)}|$$

$$d_0(X, \overline{X}) := \overline{\lim}_{n\to\infty} S_n(X, \overline{X}), \quad \text{and}$$

$$d(X, \overline{X}) := \overline{\lim}_{n\to\infty} \frac{S_n(X, \overline{X})}{1 + S_n(X, \overline{X})}.$$

Note that it is easy to show that $d(X, \overline{X}) \leq d_0(X, \overline{X})$. If $d_0(X, \overline{X}) < \infty$, then

$$d_0(X, \overline{X}) \leq \frac{d(X, \overline{X})}{1 - d(X, \overline{X})}.$$

That means, $d_0(X, \overline{X})$ is close to 0 if and only if $d(X, \overline{X})$ is close to 0. We shall state the following theorem without proof:

Theorem 2 $\|q - \overline{q}\|_{L^1} = 2d_0(X, \overline{X})$.

Numerical Aspects

In [12], Hald and McLaughlin give two numerical algorithms for the reconstruction of q. One of the algorithms can be induced from (3) above, while the other needs the information about the eigenvalues as well. Some other algorithms are also given for the other coefficient functions such as elastic modulus and density function. In [9], a Tikhonov regularization

approach is taken instead, on the foundation that the problem is overdetermined and ill-posed. Let $Q = \{p \in H^1((0,1)) : \int_0^1 p(x)dx = 0\}$. Also let $X(n) \subset \mathbf{R}^{n+1}$ such that $\mathbf{x} \in X(n)$ implies $\mathbf{x} = \{0, x_1, \ldots, x_{n-1}, 1)$. Letting $\mathbf{z}(n, p)$ be the zero set of the nth eigenfunction, we define a Tikhonov functional on $X(n) \times Q$

$$E(n, \epsilon, \mathbf{x}, p) = |\mathbf{x} - \mathbf{z}(n, p)|^2 + \epsilon \int_0^1 p'(x)^2 \, dx.$$

Let p_ϵ be the minimizer, which exists. When n is large enough and $\mathbf{x} = \mathbf{z}(n, q)$, then

$$\|p_\epsilon - q\|_{L^2} \le C \left(\frac{1}{n^2} + \epsilon n^5 \right) \int_0^1 q'^2.$$

Further Remarks

In fact, boundary data can also be reconstructed with nodal data [4]. Hill's operator was also tackled with successfully [5]. In [10], some of the arguments above are refined and made more compact. C.L. Shen solved the inverse nodal problem for the density function [20, 21,23], while Hald and McLaughlin [13] weakened the condition to bounded variations. Shen, together with Shieh, further investigated the 2×2 vectorial Sturm-Liouville system with certain nodal sets [22]. Buterin and Shieh [2] gave some reconstruction formulas for the two coefficient functions p and q in the diffusion operator $-y'' + (2\lambda p + q)y = \lambda^2 y$. Law, Lian, and Wang [17] solved the inverse nodal problem for the one-dimensional p-Laplacian eigenvalue problem, which is a nonlinear analogue of the Sturm-Liouville operator. Finally, the problem for Dirac operators was also studied by C.F. Yang [27].

More studies on other equations or systems are encouraged. Right now, most methods here make use of asymptotics of eigenvalues and nodal points. It would be desirable to explore other methods that can avoid the overdetermination of data. We add here that X.F. Yang used the nodal data on a subinterval $I = (0, b)$, where $b > 1/2$, to determine the potential function in $L^1(0, 1)$ uniquely [6, 26]. Recently, it has been shown that an arbitrarily short interval $I = (a_1, a_2)$ containing the point 1/2 suffices to determine uniquely [1]. Furthermore, there is the issue of existence. Is there any condition, no matter how strong, that can guarantee that some sequence is the nodal set of some potential function?

References

1. Browne, P.J., Sleeman, B.D.: Inverse nodal problems for Sturm-Liouville equations with eigenparameter dependent boundary conditions. Inverse Probl. **12**, 377–381 (1996)
2. Buterin, S.A., Shieh, C.T.: Inverse nodal problems for differential pencils. Appl. Math. Lett. **22**, 1240–1247 (2009)
3. Cheng, Y.H.: Reconstruction of the Sturm-Liouville operator on a p-star graph with nodal data. Rocky Mt. J. Math. **42**, 1431–1446 (2011)
4. Cheng, Y.H., Law, C.K.: On the quasinodal map for the Sturm-Liouville problem. Proc. R. Soc. Edinb. **136A**, 71–86 (2006)
5. Cheng, Y.H., Law, C.K.: The inverse nodal problem for Hill's equation. Inverse Probl. **22**, 891–901 (2006)
6. Cheng, Y.H., Law, C.K., Tsay, J.: Remarks on a new inverse nodal problem. J. Math. Anal. Appl. **248**, 145–155 (2000)
7. Chen, Y.T., Cheng, Y.H., Law, C.K., Tsay, J.: L^1 convergence of the reconstruction formula for the potential function. Proc. Am. Math. Soc. **130**, 2319–2324 (2002)
8. Cheng, Y.H., Shieh, C.T., Law, C.K.: A vectorial inverse nodal problem. Proc. Am. Math. Soc. **133**(5), 1475–1484 (2005)
9. Chen, X., Cheng, Y.H., Law, C.K.: Reconstructing potentials from zeros of one eigenfunction. Trans. Am. Math. Soc. **363**, 4831–4851 (2011)
10. Currie, S., Watson, B.A.: Inverse nodal problems for Sturm-Liouville equations on graphs. Inverse Probl. **23**, 2029–2040 (2007)
11. Guo, Y., Wei, G.: Inverse problems: Dense nodal subset on an interior subinterval, J. Differ. Equ. **255**, 2002–2017 (2013)
12. Hald, O.H., McLaughlin, J.R.: Solutions of inverse nodal problems. Inverse Probl. **5**, 307–347 (1989)
13. Hald, O.H., Mclaughlin, J.R.: Inverse problems: recovery of BV coefficients from nodes. Inverse Probl. **14**, 245–273 (1998)
14. Law, C.K., Tsay, J.: On the well-posedness of the inverse nodal problem. Inverse Probl. **17**, 1493–1512 (2001)
15. Law, C.K., Yang, C.F.: Reconstructing the potential function and its derivatives using nodal data. Inverse Probl. **14**, 299–312 (1998)
16. Law, C.K., Shen, C.L., Yang, C.F.: The inverse nodal problem on the smoothness of the potential function. Inverse Probl. **15**, 253–263 (1999); Erratum **17**, 361–364 (2001)
17. Law, C.K., Lian, W.C., Wang, W.C.: Inverse nodal problem and Ambarzumyan problem for the p-Laplacian. Proc. R. Soc. Edinb. **139A**, 1261–1273 (2009)
18. Lee, C.J., McLaughlin, J.R.: Finding the density for a membrane from nodal lines. In: Chavent, G., et al. (eds.) Inverse Problems in Wave Propagation, pp. 325–345. Springer, New York (1997)
19. McLaughlin, J.R.: Inverse spectral theory using nodal points as data-a uniqueness result. J. Differ. Equ. **73**, 354–362 (1988)
20. Shen, C.L.: On the nodal sets of the eigenfunctions of the string equation. SIAM J. Math. Anal. **6**, 1419–1424 (1988)
21. Shen, C.L.: On the nodal sets of the eigenfunctions of certain homogeneous and nonhomgeneous membranes. SIAM J. Math. Anal. **24**, 1277–1282 (1993)

22. Shen, C.L., Shieh, C.T.: An inverse nodal problem for vectorial Sturm-Liouville equations. Inverse Probl. **16**, 349–356 (2000)
23. Shen, C.L., Tsai, T.M.: On a uniform approximation of the density function of a string equation using eigenvalues and nodal points and some related inverse nodal problems. Inverse Probl. **11**, 1113–1123 (1995)
24. Shieh, C.T., Yurko, V.A.: Inverse nodal and inverse spectral problems for discontinuous boundary value problems. J. Math. Anal. Appl. **374**, 266–272 (2008)
25. Yang, X.F.: A solution of the inverse nodal problem. Inverse Probl. **13**, 203–213 (1997)
26. Yang, X.F.: A new inverse nodal problem. J. Differ. Equ. **169**, 633–653 (2001)
27. Yang, C.F., Huang, Z.Y.: Reconstruction of the Dirac operator from nodal data. Integral Equ. Oper. Theory **66**, 539–551 (2010)

Inverse Optical Design

Owen D. Miller and Eli Yablonovitch
Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA, USA

Synonyms

Electromagnetic shape optimization; Electromagnetic topology optimization; Inverse electromagnetic design

Definition

Inverse optical design requires finding a dielectric structure, if it exists, that produces a desired optical response. Such a problem is the inverse of the more common problem of finding the optical response for a given dielectric structure.

Overview

Inverse design represents an important new paradigm in electromagnetics. Over the past few decades, substantial progress has been made in computing the electromagnetic response of a given structure with sources, to the point where several commercial programs provide computational tools for a wide array of problems. Electromagnetic design, however, remains primarily restricted to heuristic methods in which scientists intuit structures that (hopefully) have the characteristics they desire. Inverse design promises to overtake such methods and provide an efficient approach for achieving nonintuitive, superior designs.

The inverse design problem cannot be solved by simply choosing a desired electric field and numerically computing the dielectric structure. It is generally unknown whether such a field can exist and, if so, whether the dielectric structure producing it has a simple physical realization. Instead, the inverse problem needs to be approached through iteration: given an initial structure, how should one iterate such that the final structure most closely achieves the desired functionality? From this viewpoint, it is clear that inverse design problems can be treated as optimization problems, in which the "merit function" to be optimized represents the desired functionality. The merit function is subject to the constraint that all fields, frequencies, etc., must be solutions of Maxwell's equations; consequently, inverse design is also sometimes referred to as PDE-constrained optimization.

This entry primarily focuses on the methodology for finding the optimal design of an electromagnetic structure. We describe the physical mechanism underpinning adjoint-based optimization, in which two simulations for each iteration provide information about how to update the structure. We then discuss some applications of the method and key research results in the literature.

Electromagnetic Optimization

As previously discussed, successful inverse design finds a structure through iterative optimization: an initial design is created, computations are done to find a new design, the design is updated, and the loop continues. Whether an optimization is successful is defined by the efficiency and effectiveness of the computations for finding a new design. In some fields of optimization, stochastic methods such as genetic algorithms or simulated annealing provide the computations. The inefficiency of completing the many electromagnetics simulations required, however, renders such methods generally ineffective in the optical regime. Instead, the so-called "adjoint" approach provides quicker computations while

exploiting the fact that the fields must be solutions of Maxwell's equations.

Adjoint-based optimization is well known in mathematics and engineering. As applied to PDE-constrained problems, [4] provides a general introduction while [1] and [10] work out the optimization equations for elasticity and electromagnetic systems, respectively. Instead of taking a purely equation-based approach, however, we will present the adjoint-based optimization technique from a more intuitive viewpoint, to understand the physical origins of adjoint fields [9,11].

For concreteness, we will consider a simplified problem in two dimensions with a specific merit function. The dimensionality will allow us to treat the electric field as a scalar field. The picture will be clearer with these simplifications, and generalizing to three dimensions and a larger group of merit functions does not change the underlying optimization mechanism.

The crux of the optimization routine is the decision of how to update the structure from one iteration to the next. Consider, for example, a problem in which the electric field intensity at a single point, x_0, is to be maximized. The merit function J would take the form

$$J = \frac{1}{2}|E(x_0)|^2 \tag{1}$$

If the change in structure is small between the two iterations, the change in the fields is also relatively small. The change in merit function can then be approximated as

$$\begin{aligned}\delta J &\approx \frac{1}{2}\left[E^*(x_0)\delta E(x_0) + E(x_0)\delta E^*(x_0)\right] \\ &= \mathrm{Re}\left[E^*(x_0)\delta E(x_0)\right]\end{aligned} \tag{2}$$

Eq. 2 is very important: it states that the change in merit function is simply the product of the (conjugated) original field $E(x_0)$ with the change in field incurred by the change in geometry, $\delta E(x_0)$. The question becomes whether a change in geometry can be chosen to ensure that $\delta J > 0$ (or $\delta J < 0$), so that the merit function increases (decreases) each iteration.

The simplest method for choosing a new structure would be brute force. One could add or subtract a small piece of dielectric at every allowable point in the domain, run a simulation to check whether the merit function has increased, and then choose the structure that most increased the merit function. However, this would take thousands or millions of simulations per iteration and is clearly unfeasible. The adjoint method, however, gleans the same information from only two simulations. This is accomplished by exploiting symmetry properties.

The first step is to recognize that a small piece of dielectric acts like an electric dipole. If a small sphere of radius a and dielectric constant ϵ_2 is added to a background with dielectric ϵ_1, the scattering from the sphere will be approximately equivalent to the fields radiated by an electric dipole with dipole moment [5]:

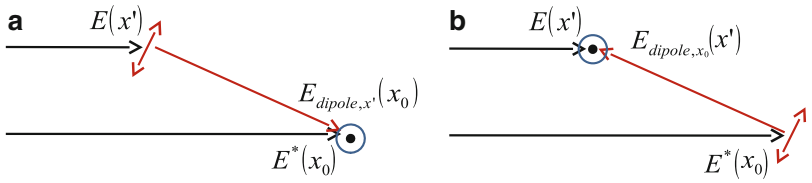
$$p = 4\pi\epsilon_0\left(\frac{\epsilon_2 - \epsilon_1}{\epsilon_2 + 2\epsilon_1}\right)a^3 E_{\mathrm{inc}} \tag{3}$$

where E_{inc} is the value of the incident field at the location of the dielectric. Although Eq. 3 assumes a three-dimensional sphere, the two-dimensional case differs only by numerical pre-factors. The addition of dielectric at a point x', then, is equivalent to the addition of an electric dipole driven by $E_{\mathrm{inc}} = E(x')$. The change in field at x_0 can be expressed as $\delta E(x_0) = E(x')E_{\mathrm{dipole},x'}(x_0)$, where $E_{\mathrm{dipole},x'}(x_0)$ is the normalized electric field at x_0 from a dipole at x' and $E(x')$ provides the driving term. δJ can be rewritten:

$$\begin{aligned}\delta J &= \mathrm{Re}\left[E^*(x_0)\delta E(x_0)\right] \\ &= \mathrm{Re}\left[E^*(x_0)E(x')E_{\mathrm{dipole},x'}(x_0)\right] \\ &= \mathrm{Re}\left[E^*(x_0)E_{\mathrm{dipole},x_0}(x')E(x')\right] \\ &= \mathrm{Re}\left[W(x')E(x')\right]\end{aligned} \tag{4}$$

The first step in Eq. 4 is the replacement of $\delta E(x_0)$. The next step is the realization that placing a dipole at x' and measuring at x_0 is equivalent to placing a dipole at x_0 and measuring at x'. This can be proved by the symmetry of the Green's function or by the recognition that the optical paths are identical and the fields must therefore be equivalent. The final step is to define the adjoint field $W(x') = E^*(x_0)E_{\mathrm{dipole},x_0}(x')$. By analogy with the definition of δE, it is clear that $W(x')$ is the field of a dipole at x_0 with $E_{\mathrm{inc}} = E^*(x_0)$. Figure 1 motivates this particular sequence of operations.

With the original form of δJ, as in Eq. 2, one would need a simulation to find $E(x_0)$ and then countless simulations to find $\delta E(x_0)$ for every possible x' at which to add dielectric, as the dipole location would



Inverse Optical Design, Fig. 1 Illustration of how adjoint-based electromagnetic optimization exploits symmetry properties. (**a**) shows a simple but inefficient method for testing whether to add dielectric at x'. A first simulation (*black*) finds $E(x_0)$ and $E(x')$. Then a second simulation (*red*) is run with an electric dipole at x'. Finally, $\delta J = \mathrm{Re}\left[E^*(x_0)E(x')E_{\mathrm{dipole},x'}(x_0)\right]$. In order to decide the location of optimal δJ, many simulations with dipoles at each different x' would have to be run. In (**b**) the equivalent calculation is more efficiently completed. The second simulation (*red*) is of an electric dipole at x_0, instead of x', and the fields are multiplied at x'. In this way, only a single extra simulation is required, with a dipole at x_0, and δJ is known for all possible values of x'

change every simulation. By switching the measurement and dipole locations, however, the countless simulations have been reduced to a single one. Placing the dipole at x_0 and measuring the resulting field at x', one can calculate $W(x')$ *everywhere* with a single simulation. This is why adjoint optimization requires only two simulations. Importantly, if the merit function had been the sum or integral of the field intensities on a larger set of points, still only two simulations would be required. In that case, dipoles would be simultaneously placed at each of the points. For a mathematically rigorous derivation of the adjoint field in a more general electromagnetics setting, consult [10].

Applications

Adjoint-based optimization has been used as a design tool in numerous electromagnetics applications. The authors of [12], for example, designed scattering cylinders such that the radiation from a terminated waveguide was highly directional and asymmetric. Their design was nonintuitive and would have been almost impossible to achieve through heuristic methods. Although nominally designed in the rf regime, the design would work at optical frequencies in a scaled-down configuration.

Similar research has shown other ways in which optimized designs can mold the flow of light in desirable ways. Using photonic crystals for waveguiding and routing is a promising technology, but achieving optimal designs is difficult in practice. In [6] and [7], the authors used adjoint-based optimization techniques to design a high-bandwidth T-junction and an efficient 90° waveguide bend, respectively, in photonic crystal platforms.

Plasmonics represents another field in which optimal design may prove particularly useful. In a recent paper, Andkjær et al. [2] designed a grating coupler to efficiently couple surface plasmons to incoming and outgoing waves. The coupler was superior to previous designs achieved by other methods and demonstrates how one might couple into or out of future plasmon-based technologies.

Although the examples above and the previous discussion focused on optimizing merit functions in which the fields are the primary variables, the technique extends to eigenfrequencies and other variables. Bandgap optimization, in which the gap between two eigenfrequencies is maximized, is actually a self-adjoint problem for which only a single simulation per iteration is required. Optimal structures with large bandgaps were designed in [3] and [8].

Discussion

Inverse design has been an invaluable tool in fields such as aerodynamic design and mechanical optimization. It seems clear that it can provide the same function in optical design, especially as computational power continues to improve. Through a more intuitive understanding of the optimization mechanism, the technique may become more accessible to a wider audience of researchers. By treating dielectrics through their dipole moments and iterating through small changes in structure, simple initial structures can morph into nonintuitive, superior designs. Whereas

the current forefront of electromagnetic computation is the quick solution of the response to a given structure, the inverse problem of computing the structure for a given response may prove much more powerful in the future.

Cross-References

▸ Adjoint Methods as Applied to Inverse Problems

References

1. Allaire, G., De Gournay, F., Jouve, F., Toader, A.: Structural optimization using topological and shape sensitivity via a level set method. Control Cybern. **34**(1), 59 (2005)
2. Andkjær, J., Nishiwaki, S., Nomura, T., Sigmund, O.: Topology optimization of grating couplers for the efficient excitation of surface plasmons. J. Opt. Soc. Am. B **27**(9), 1828–1832 (2010)
3. Cox, S., Dobson, D.: Band structure optimization of two-dimensional photonic crystals in H-polarization. J. Comput. Phys. **158**(2), 214–224 (2000)
4. Giles, M., Pierce, N.: An introduction to the adjoint approach to design. Flow Turbul. Combust. **65**, 393–415 (2000)
5. Jackson, J.: Classical Electrodynamics, 3rd edn. Wiley, New York (1999)
6. Jensen, J., Sigmund, O.: Topology optimization of photonic crystal structures: a high-bandwidth low-loss T-junction waveguide. J. Opt. Soc. Am. B **22**(6), 1191–1198 (2005)
7. Jensen, J., Sigmund, O., Frandsen, L., Borel, P., Harpoth, A., Kristensen, M.: Topology design and fabrication of an efficient double 90 degree photonic crystal waveguide bend. IEEE Photon. Technol. Lett. **17**(6), 1202 (2005)
8. Kao, C., Osher, S., Yablonovitch, E.: Maximizing band gaps in two-dimensional photonic crystals by using level set methods. Appl. Phys. B Lasers Opt. **81**(2), 235–244 (2005)
9. Lalau-Keraly, C.M., Bhargava, S., Miller, O.D., Yablonovitch, E.: Adjoint shape optimization applied to electromagnetic design. Opt. Exp. **21**(18), 21,693–21,701 (2013)
10. Masmoudi, M., Pommier, J., Samet, B.: The topological asymptotic expansion for the Maxwell equations and some applications. Inverse Probl. **21**, 547 (2005)
11. Miller, O.D.: Photonic design: from fundamental solar cell physics to computational inverse design. Phd thesis, University of California, Berkeley (2012). http://arxiv.org/abs/1308.0212
12. Seliger, P., Mahvash, M., Wang, C., Levi, A.: Optimization of aperiodic dielectric structures. J. Appl. Phys. **100**, 34,310 (2006)

Inverse Problems: Numerical Methods

Martin Burger
Institute for Computational and Applied Mathematics, Westfälische Wilhelms-Universität (WWU) Münster, Münster, Germany

Synonyms

Discretization; Ill-posed problems; Inverse problems; Numerical methods; Regularization

Definition

Inverse problems are problems where one looks for a cause of an observed or desired effect via mathematical model, usually by inverting a forward problem. The forward problem such as solving partial differential equations is usually well posed in the sense of Hadamard, whereas the inverse problem such as determining unknown parameter functions or initial values is ill posed in most cases and requires special care in numerical approaches.

Introduction

Inverse problem approaches (often called inverse modeling in engineering) have become a key technique to recover quantitative information in many branches of science. Prominent examples include medical image reconstruction, nondestructive material testing, seismic imaging, and remote sensing. The common abstract approach to inverse problems is to use a forward model that links the unknown u to the available data f, which very often comes as a system of partial differential equations (or integral formulas derived from partial differential equations, cf., e.g., [5, 13]). Solving the forward model given u is translated to evaluating a (possibly nonlinear) operator F. The inverse problem then amounts to solving the operator equation

$$F(u) = f. \tag{1}$$

Particular complications arise due to the fact that in typical situations the solution of (1) is not well

posed, in particular u does not depend continuously on the data and the fact that practical data always contain measurement and modeling errors. Therefore any computational approach is based on a regularization method, which is a well-posed approximation to (1) parameterized by a regularization parameter α, which tunes the degree of approximation. In usual convention, the original problem is recovered in the limit $\alpha \to 0$ and no noise, and in presence of noise, α needs to be tuned to obtain optimal reconstructions. We will here take a deterministic perspective and denote by δ the noise level, i.e., the maximal norm difference between f and the exact data Ku^* (u^* being the unknown exact solution).

The need for regularization makes numerical methods such as discretization and iterative schemes quite peculiar in the case of inverse problems; they are always interwoven with the regularization approach. There are two possible issues appearing:

- The numerical methods can serve themselves as regularizations, in which case classical questions of numerical analysis have to be reconsidered. For Example, if regularization is achieved by discretization, it is not the key question how to obtain a high order of convergence as the discretization fineness decreases to zero, but it is at least equally important how much robustness is achieved with respect to the noise and how the discretization fineness is chosen optimally in dependence of the noise level.
- The regularization is carried out by a different approach, e.g., a variational penalty. In this case one has to considered numerical methods for a parametric problem, and robustness is desirable in the case of small regularization parameter and decreasing noise level, which indeed yields similarities to the first case, often also to singular perturbation problems in differential equations.

Iterative Regularization Methods

Iterative methods, which we generally write as

$$u_{k+1} = G(u_k, F(u_k) - f, \beta_k), \tag{2}$$

yield a first instance of numerical methods for regularization. Simple examples are the Landweber iteration

$$u_{k+1} = u_k - \beta_k F'(u_k)^*(F(u_k) - f) \tag{3}$$

and the Levenberg-Marquardt method for nonlinear problems

$$u_{k+1} = u_k - (F'(u_k)^* F'(u_k) + \beta_k I)^{-1} F'(u_k)^* (F(u_k) - f). \tag{4}$$

In this case the regularization is the maximal number of iterations k_*, which are carried out, i.e., $\alpha = \frac{1}{k_*}$. Instead of convergence, one speaks of *semiconvergence* in this respect (cf., [9]):

- In the case of exact data $f = Ku^*$, one seeks classical convergence $u_k \to u^*$.
- In the case of noisy data $f = Ku^* + n_\delta$ with $\|n_\delta\| \le \delta$, one seeks to choose a maximal number of iterates $k_*(\delta)$, such that $u^\delta_{k_*(\delta)}$ converges to u^* as $\delta \to 0$, where u^δ_k denotes the sequence of iterates obtained with data $f = Ku^* + n_\delta$. Note that in this case the convergence concerns the stopped iterates of different iteration sequences.

Major recent challenges are iterative methods in reflexive and nonreflexive Banach spaces (cf., [8, 14]).

Regularization by Discretization

Discretization of infinite-dimensional inverse problems needs to be understood as well as a regularization technique, and thus again convergence as discretization fineness tends to zero differs from classical aspects of numerical analysis. If the data are taken from an m-dimensional subspace and the unknown is approximated in an n-dimensional subspace, one usually ends up with a problem of the form

$$Q_m F(P_n u) = f, \tag{5}$$

typically with Q_m and P_n being projection operators. Again a semiconvergence behavior appears, where the regularization parameter is related to $\frac{1}{n}$ respectively and $\frac{1}{m}$.

In the case of linear inverse problems, it is well understood that the discretization leads to an ill-conditioned linear system, and the choice of basis functions is crucial for the conditioning. In particular Galerkin-type discretization with basis functions in the range of the adjoint operator are efficient ways to discretize inverse problems (cf., [5, 12]). The role of adaptivity in inverse problems has been explored

recently (cf., [1]), again with necessary modifications compared to the case of well-posed problems due to the fact that a posteriori error estimates without assumptions on the solution are impossible.

Variational Regularization Methods

The most frequently used approach for the stable solution of inverse problems are variational regularization techniques, which consist in minimizing a functional of the form (cf., [3,5])

$$E_\alpha(u) = D(F(u), f) + \alpha J(u), \tag{6}$$

where J is a regularization functional and D is an appropriate distance measure, frequently a square norm in a Hilbert space (related to classical least-squares methods)

$$D(F(u), f) = \frac{1}{2}\|F(u) - f\|^2. \tag{7}$$

The role of the regularization functional from a theoretical point of view is to enforce well posedness in the minimization of E_α, typically by enforcing compactness of sublevel sets of E_α in an appropriate topology. Having in mind the Banach-Alaoglu theorem, it is not surprisingly that the most frequent choice of regularization functionals are powers of norms in appropriate Banach spaces, whose boundedness implies weak compactness. From a practical point of view, the role of the regularization functional is to introduce a priori knowledge by highly penalizing unexpected or unfavorable solutions. In particular in underdetermined cases, the minimization of J needs to determine appropriate solutions, a paradigm which is heavily used in the adjacent field of compressed sensing (cf., [4]).

Besides discretization issues as mentioned above, a key challenge is the construction of efficient optimization methods to minimize E_α. In the past squared norms or seminorms in Hilbert spaces (e.g., in L^2 or H^1) have been used frequently, so that rather standard algorithms for differentiable optimization have been used. The main challenge when using Newton-type methods is efficient solution of the arising large linear systems; several preconditioning approaches have been proposed, some at the interface to optimal control and PDE-constrained optimization (cf., e.g., [2]). In particular in the twenty-first century, nonsmooth regularization functionals such as total variation and ℓ^1-type norms became more and more popular, since they can introduce prior knowledge more effectively. A variety of numerical optimization methods has been proposed in such cases, in particular Augmented Lagrangian methods have become popular (cf., [6]).

Bayesian Inversion

The use of Bayesian approaches for inverse problems has received growing attention in the recent years (cf., e.g., [7]) due to frequent availability of prior knowledge as well as increasing detail in the statistical characterization of noise and other uncertainties in inverse problems, which can be handled naturally. The basis of Bayesian inversion in a finite-dimensional setup is Bayes' formula for the posterior probability density

$$p(u|f) = \frac{p(f|u)\ p(u)}{p(f)}. \tag{8}$$

Here $p(f|u)$ is the data likelihood, into which the forward model and the noise are incorporated, and $p(u)$ respectively $p(f)$ are a priori probability densities for the unknown and the data, respectively. Since $p(f)$ is just a scaling factor when f is fixed, it is usually neglected. Most effort is used to model the prior probability density, which is often related to regularization functionals in variational methods via

$$p(u) \sim e^{-\alpha J(u)}. \tag{9}$$

A standard approach to compute estimates is *maximum a posteriori probability (MAP)* estimation, which amounts to maximize $p(u|f)$ subject to u. By the equivalent minimization of the negative log likelihood, MAP estimation can be translated into variational regularization; the role of the statistical approach boils down to selecting appropriate regularization functionals and data terms based on noise models. In order to quantify uncertainty, also conditional mean (CM) estimates

$$\hat{u} = \int u\ p(u|f)\ du, \tag{10}$$

variances, and other quantifying numbers of the posterior distribution are used. These are all based on integration of the posterior in very high dimensions,

since clearly the infinite-dimensional limit should be approximated. The vast majority of approaches is based on Markov chain Monte Carlo (MCMC) methods (cf., e.g., [7]); see also [15] for a deterministic approach. The construction of efficient sampling schemes for posterior distribution with complicated priors is a future computational challenge of central importance.

A program related to classical numerical analysis is the convergence of posteriors and different estimates as the dimension of the space for the unknown (possibly also for the data) tends to infinity, an issue that has been investigated under the keyword of *discretization invariance* in several instances recently (cf., [10, 11]).

References

1. Benameur, H., Kaltenbacher, B.: Regularization of parameter estimation by adaptive discretization using refinement and coarsening indicators. J. Inverse Ill-Posed Probl. **10**, 561–584 (2002)
2. Biegler, L., Ghattas, O., Heinkenschloss, M., van Bloemen Waanders, V. (eds.) Large-Scale PDE-Constrained Optimization, Springer, New York (2003)
3. Burger, M., Osher, S.: Convergence rates of convex variational regularization. Inverse Probl. **20**, 1411–1421 (2004)
4. Donoho, D.L.: Compressed sensing. IEEE Trans. Inform. Theory **52**, 1289–1306 (2006)
5. Engl, H., Hanke, M., Neubauer, A.: Regularization of Inverse Problems. Kluwer, Dordrecht (1996)
6. Goldstein, T., Osher, S.: The split Bregman method for L1-regularized problems. SIAM J. Imagin. Sci. **2**, 323–343 (2009)
7. Kaipio, J., Somersalo, A.: Statistical and Computational Inverse Problems. Springer, Heidelberg (2005)
8. Kaltenbacher, B., Schoepfer, F., Schuster, T.: Convergence of some iterative methods for the regularization of nonlinear ill-posed problems in Banach spaces. Inverse Probl. **25**, 065003 (2009)
9. Kaltenbacher, B., Neubauer, A., Scherzer, O.: Iterative Regularization Methods for Nonlinear Ill-Posed Problems. De Gruyter, Berlin (2008)
10. Lassas, M., Saksman, S., Siltanen, S.: Discretization invariant Bayesian inversion and Besov space priors. Inverse Probl. Imaging **3**, 87–122 (2009)
11. Lehtinen, M.S., Päivärinta, L., Somersalo, E.: Linear inverse problems for generalised random variables. Inverse Probl. **5**, 599–612 (1989)
12. Natterer, F.: Numerical methods in tomography. Acta Numer. **8**, 107–141 (1999)
13. Natterer, F.: Imaging and inverse problems of partial differential equations. Jahresber. Dtsch. Math. Ver. **109**, 31–48 (2007)
14. Osher, S., Burger, M., Goldfarb, D., Xu, J., Yin, W.: An iterative regularization method for total variation based image restoration. Multiscale Model. Simul. **4**, 460–489 (2005)
15. Schwab, C., Stuart, A.M.: Sparse deterministic approximation of Bayesian inverse problems. Inverse Probl. to appear. (2012)

Inverse Spectral Problems: 1-D, Algorithms

Paul E. Sacks
Department of Mathematics, Iowa State University, Ames, IA, USA

Synonyms

Inverse eigenvalue problems; Inverse Sturm-Liouville problems; Numerical methods

Introduction

In this entry we will describe techniques which have been developed for numerical solution of inverse spectral problems for differential operators in one space dimension, for which the model is the inverse Sturm-Liouville problem. Let $V = V(x)$ be a given real valued potential on the interval $[0, 1]$ and consider the eigenvalue problem

$$\phi'' + (\lambda - V(x))\phi = 0 \quad 0 < x < 1 \quad \phi(0) = \phi(1) = 0 \tag{1}$$

As is well known, there exists an infinite sequence of real eigenvalues

$$\lambda_1 < \lambda_2 < \ldots \lambda_n \to +\infty \tag{2}$$

We will always assume at least that $V \in L^2(0, 1)$, although much of what is said below is valid in larger spaces. The inverse spectral problem of interest is to recover $V(x)$ from spectral data – there are many different versions of this, depending on exactly what is meant by "spectral data." In the simplest case this would simply mean the eigenvalues, but one quickly sees that this is not enough information, unless the class of V's is considerably restricted.

We therefore define some additional quantities. Let $\phi_n(x)$ be an eigenfunction corresponding to λ_n normalized by $||\phi_n||_{L^2(0,1)} = 1$, and set

$$\rho_n = \frac{1}{\phi_n'(0)^2} \tag{3}$$

$$\kappa_n = \log\left(|\phi_n'(1)|/|\phi_n'(0)|\right) \tag{4}$$

Also, let μ_n denote the nth eigenvalue of (1) when the boundary condition at $x = 1$ is replaced by $\phi'(1) + H\phi(1) = 0$ for some fixed $H \in \mathbb{R}$. The following asymptotic expressions are known.

$$\lambda_n = (n\pi)^2 + \int_0^1 V(s)\,ds + a_n \qquad \sum_{n=1}^{\infty} a_n^2 < \infty \tag{5}$$

$$\rho_n = \frac{1}{2(n\pi)^2}\left(1 + \frac{b_n}{n}\right) \qquad \sum_{n=1}^{\infty} b_n^2 < \infty \tag{6}$$

$$\kappa_n = \frac{c_n}{n} \qquad \sum_{n=1}^{\infty} c_n^2 < \infty \tag{7}$$

$$\mu_n = ((n-\frac{1}{2})\pi)^2 + \int_0^1 V(s)\,ds + 2H + d_n \quad \sum_{n=1}^{\infty} d_n^2 < \infty \tag{8}$$

We may then formulate three corresponding inverse spectral problems:

Problem 1 Determine V given $\{\lambda_n\}_{n=1}^{\infty}, \{\rho_n\}_{n=1}^{\infty}$

Problem 2 Determine V given $\{\lambda_n\}_{n=1}^{\infty}, \{\kappa_n\}_{n=1}^{\infty}$

Problem 3 Determine V given $\{\lambda_n\}_{n=1}^{\infty}, \{\mu_n\}_{n=1}^{\infty}$

It is known that each of the above problems has at most one solution in an appropriate function space, such as $L^2(0,1)$. From a computational point of view we are always dealing with a finite subset of the data, such as the first N terms of each sequence, so that careful consideration should be given to how to compensate for the missing data.

There are many obvious and not so obvious variants of these problems which have been studied, but due to limited space we will focus only on these three. We mention, however, that one widely studied special case, when V is symmetric with respect to the midpoint $x = 1/2$, may be viewed as a special case of Problem 2, since $\kappa_n = 0$ for any n automatically. In this case we may cite [4,6,9,12] as general references for the theory of inverse Sturm-Liouville problems.

Whichever problem is being solved and whichever of the methods described in the following sections is to be used, it is almost always useful to do a preliminary reduction to the case when the mean value of the potential $\int_0^1 V(s)\,ds$ is zero. This may be done by first making an estimate of $\int_0^1 V(s)\,ds$ based on the asymptotic behavior of the eigenvalues, such as

$$\int_0^1 V(s)\,ds = \lim_{n\to\infty} \lambda_n - (n\pi)^2 \tag{9}$$

which follows from (5), and then taking into account the obvious fact that $\lambda_n - \int_0^1 V(s)\,ds$ is the nth eigenvalue for the shifted potential $V(x) - \int_0^1 V(s)\,ds$.

Computational Methods

In this section we give details of several widely applicable and representative methods.

Integral Equation Method

The seminal paper Gelfand and Levitan [5], one of the very earliest substantial works on the theory of the inverse spectral problem, also supplies, in principle, a practical computational method for Problem 1. Assuming as above that V has mean value zero, the algorithm is as follows:

- Set

$$g(t) = \sum_{n=1}^{\infty}\left(2n\pi \sin n\pi t - \frac{1}{\sqrt{\lambda_n}\rho_n}\sin\sqrt{\lambda_n}t\right). \tag{10}$$

- Set $f(x,t) = \frac{1}{2}\left(G(|x-t|) - G(x+t)\right)$ where $G(t) = \int_0^t g(s)\,ds$.
- Solve the integral equation

$$f(x,t) + \int_0^x K(x,z) f(z,t)\,dz + K(x,t) = 0 \qquad 0 \le t \le x \le 1 \tag{11}$$

  for $K(x,t)$.
- Obtain the potential from $V(x) = 2\frac{d}{dx}K(x,x)$.

The asymptotic behaviors (5), (6) guarantee that $g \in L^2(0,2)$, but some care should be taken with numerical evaluation of g since it is the difference of two divergent series. If the available data consists of λ_n, ρ_n for $n \le N$, then using the Nth partial sum of the series (10) as an approximation to g amounts to specifying that $\lambda_n = (n\pi)^2$, $\rho_n = \frac{1}{2(n\pi)^2}$ for $n > N$, which are the exact values these quantities would have when

$V = 0$. The integral equation (11) may be numerically solved, for example, by a collocation method or by seeking the solution as a linear combination of suitable basis functions.

Method of Overdetermined Hyperbolic Problems

The next method was introduced in [14], in which the inverse spectral problem was shown to be equivalent to a certain overdetermined boundary value problem for a hyperbolic partial differential equation, which may be solved by an iteration technique. The method is easily adaptable to any of Problems 1–3, as well as many other variants – we will focus on Problem 2 for definiteness.

The kernel $K(x,t)$ appearing in (11) is known to have a number of other interesting properties (see [4, 9]), the first of which we need is that it serves as the kernel of an integral operator which transforms a solution of

$$\phi'' + \lambda\phi = 0 \qquad \phi(0) = 0 \tag{12}$$

into a corresponding solution of

$$\phi'' + (\lambda - V(x))\phi = 0 \qquad \phi(0) = 0 \tag{13}$$

Specifically, if we denote by $\theta(x,\lambda)$ the solution of (13) normalized by $\theta'(0,\lambda) = \sqrt{\lambda}$, then

$$\theta(x,\lambda) = \sin\sqrt{\lambda}x + \int_0^x K(x,t)\sin\sqrt{\lambda}t\,dt \tag{14}$$

The key point here is that K does not depend on λ. The second property of K we will use here is that if defined for $t < 0$ by odd extension, it satisfies the Goursat problem

$$K_{tt} - K_{xx} + V(x)K = 0 \qquad 0 < |t| < x < 1 \tag{15}$$

$$K(x,\pm x) = \pm\frac{1}{2}\int_0^x V(s)\,ds \qquad 0 < x < 1 \tag{16}$$

Observing that $\theta(1,\lambda_n) = 0$ and $\theta'(1,\lambda_n) = \sqrt{\lambda_n}(-1)^n e^{\kappa_n}$, we obtain (recall we still assume V has zero mean)

$$\int_0^1 K(1,t)\sin\sqrt{\lambda_n}t\,dt = -\sin\sqrt{\lambda_n} \tag{17}$$

$$\int_0^1 K_x(1,t)\sin\sqrt{\lambda_n}t\,dt = \sqrt{\lambda_n}((-1)^n e^{\kappa_n} - \cos\sqrt{\lambda_n}) \tag{18}$$

With spectral data λ_n, κ_n known, these systems of equations can be shown to uniquely determine (since $K(1,t)$, $K_x(1,t)$ are both odd)

$$K(1,t) := G_0(t) \quad K_x(1,t) := G_1(t) \qquad -1 < t < 1 \tag{19}$$

which we think of as Cauchy data for $K(x,t)$ on the segment $\{(1,t) : -1 < t < 1\}$. Equations (15), (16), and (19) now constitute an overdetermined hyperbolic boundary value problem for $K(x,t)$, if V were known, and the inverse spectral problem may be regarded as that of determining the pair $\{V(x), K(x,t)\}$ so that (15), (16), and (19) hold. One numerical method proposed in [14] for obtaining the solution in this way is the fixed point iteration scheme

$$V_{n+1}(x) = 2\frac{d}{dx}u(x,x;V_n) \qquad V_0(x) \equiv 0 \tag{20}$$

where $u(x,t;V)$ denotes the solution of (15), (19) in the domain $\{(x,t) : |t| < x < 1\}$, which is a well-posed Cauchy problem. A convergence theorem for $V \in L^\infty(0,1)$ is given in [14]. The algorithm may be summarized as:

- Solve the systems (17), (18) for $G_0(t) = K(1,t)$, $G_1(t) = K_x(1,t)$ and extend both to be odd functions on $(-1,1)$.
- Carry out the iteration step (20) until a suitable stopping criterion is satisfied.

Numerical solution of (17), (18) is most conveniently achieved by looking for $K(1,t), K_x(1,t)$ as linear combinations of suitable basis functions, chosen to match the expected boundary behavior of $K(1,t), K_x(1,t)$ as well as possible. For example, since $G_0(0) = G_0(1) = G_1(0) = 0$, but $G_1(1) \neq 0$ in general, the choices

$$K(1,t) = \sum a_j \sin j\pi t$$
$$K_x(1,t) = \sum b_j \sin(j - \frac{1}{2})\pi t \tag{21}$$

seem to work best. The numerical evaluation of $u(x,t;V)$ may be conveniently carried out by means of a finite difference scheme in characteristic coordinates.

Optimization Method

In [13] an optimization technique is proposed, which we describe in the case of Problem 3 with $H = 0$ for simplicity. Denote by $\lambda(n, V), \mu(n, V)$ respectively the nth eigenvalue of (1) and the corresponding problem when the right-hand boundary condition is replaced by $\phi'(1) = 0$. Let $\omega_n > 0$ be weights to be specified later and define the objective functional

$$J(V) = \sum_{n=1}^{\infty} \omega_n \left[(\lambda(n, V) - \lambda_n)^2 + (\mu(n, V) - \mu_n)^2 \right] \tag{22}$$

We assume at least that $\sum_{n=1}^{\infty} \omega_n < \infty$, from which it follows, taking into account the known asymptotic behavior of the eigenvalues, that $J(V)$ is well defined for $V \in L^1(0, 1)$. With the stronger restriction $\sum_{n=1}^{\infty} n\omega_n < \infty$, it is shown in [13] that the unique solution of Problem 3 is also the one and only critical point of J.

An explicit expression for the gradient of J may also be derived, namely,

$$\begin{aligned} DJ(V)\delta V = & 2 \sum_{n=1}^{\infty} \int_0^1 \omega_n \left[(\lambda(n, V) - \lambda_n) g_{1n}^2(x) \right. \\ & \left. + (\mu(n, V) - \mu_n) g_{2n}^2(x)\, dx \right] \delta V(x)\, dx \end{aligned} \tag{23}$$

where g_{1n}, g_{2n} denote respectively $L^2(0, 1)$ normalized eigenfunctions corresponding to $\lambda(n, V)$ and $\mu(n, V)$. One may now use some kind of standard unconstrained smooth optimization method to locate the unique global minimum of J.

The approach of [13] may be summarized as:

- Apply some gradient-based unconstrained minimization algorithm to the functional J defined in (22).

Further Discussion

The use of this algorithm is relatively costly, due to the need to accurately solve the two direct eigenvalue problems for the eigenvalues and normalized eigenfunctions at each step of the iteration process. The numerical examples in [13] are carried out using the Polack-Ribiere variant of the conjugate gradient algorithm.

Matrix Methods

A final class of methods uses a finite difference approximation to reduce the inverse Sturm-Liouville problem to a corresponding matrix inverse eigenvalue problem. The following approach to numerical solution of Problem 2 is taken from [3], and a number of refinements have been made by later authors (see, e.g., [1]). Assume that the available data is λ_j, κ_j for $j = 1, \ldots M$. Using an obvious central differencing with a uniform grid $x_j = hj$, $j = 1 \ldots 2M$, $h = 1/(2M + 1)$, we obtain in place of (1) a $2M \times 2M$ matrix equation in the form

$$(h^{-2}A + Q)y = \lambda y \tag{24}$$

where A is the symmetric tridiagonal matrix with $A_{jj} = 2$, $A_{j,j+1} = -1$ and $Q = \text{diag}\{V(x_1), \ldots V(x_{2M})\}$.

For fixed h and arbitrary diagonal matrix Q, let $\nu_j(Q)$ denote the jth eigenvalue of $h^{-2}A + Q$, $j = 1, \ldots M$, and let

$$\tau_j(Q) = \log \left| \frac{y_{j,M}}{y_{j,1}} \right| \tag{25}$$

where $y_{j,k}$ denotes the kth component of an eigenvector corresponding to $\nu_j(Q)$. If Q is the discretization of the exact potential V, then it will be true that $\nu_j(Q)$ tends to λ_j as $h \to 0$ for fixed j, but it is well known to be highly nonuniform with respect to j. On the other hand, the leading asymptotics of the discrepancy can be computed explicitly and introduced as a correction term – similar considerations hold for the approximation of κ_j by $\tau_j(Q)$. Thus, we define mappings $\alpha(Q) = [\alpha_1(Q), \ldots \alpha_M(Q)]^T$ and $\beta(Q) = [\beta_1(Q), \ldots \beta_M(Q)]^T$ where

$$\alpha_j(Q) = \nu_j(Q) + (j\pi)^2 - \frac{4}{h^2} \sin^2 \frac{j\pi h}{2} - \lambda_j \tag{26}$$

$$\beta_j(Q) = \frac{2 \sin j\pi h}{h\pi} \tau_j(Q) - 2j\kappa_j \tag{27}$$

If we then let

$$F(Q) = \begin{bmatrix} \alpha(Q) \\ \beta(Q) \end{bmatrix} \tag{28}$$

then the approximate solution is sought as a solution of the $2M \times 2M$ nonlinear system $F(Q) = 0$. A modified Newton scheme is used in [3]:

$$Q_{n+1} = Q_n - DF(0)^{-1} F(Q_n) \qquad Q_0 = 0 \tag{29}$$

where $DF(0)$ denotes the Jacobian of F at $Q = 0$. An explicit expression for the entries of $DF(0)$ may be calculated, namely,

$$DF(0)_{jk} = \begin{cases} 2h \sin^2 jkh\pi & j = 1, \dots M \\ \pi h \sin 2(j-M)kh\pi & j = M+1, \dots 2M \end{cases} \tag{30}$$

It follows from this that $DF(0)$ is nonsingular, and in fact the condition number with respect to the Euclidean norm may be shown to be $\sqrt{2M+1}$. Thus, the successive iterates are well defined and that the scheme is convergent at least provided that the solution Q is sufficiently small.

The algorithm may be summarized as follows:

- For a given guess Q_n, solve the direct matrix eigenvalue problem for $(h^{-2}A + Q_n)$ to obtain $\nu_j(Q_n), \tau_j(Q_n)$ for $j = 1, \dots M$.
- Compute $F(Q_n)$ using (26)–(28).
- Compute Q_{n+1} using (29) and the explicit form of $DF(0)$ for $n = 1, 2, \dots$ until a suitable stopping criterion is satisfied.

Further Discussion

This method requires the solution of the direct eigenvalue problem for a potentially large matrix at each step of the iteration process. Convergence is only guaranteed for sufficiently small Q, although in practice it seems quite robust. The way the algorithm is stated here, the stepsize h and the number of spectral data $2M$ are tied together, but more recent variants of this approach have loosened such restriction. The use of matrices arising from higher order discretizations of the ODE has also been investigated. In particular the Numerov discretization scheme has received special attention because it allows for higher order accuracy with respect to h while still only using a 3-point stencil.

Related Problems

We conclude by mentioning several other classes of inverse spectral problems to which some or all of the above methods may be adapted:

- *Inverse spectral problems associated with other forms of second-order differential operators:* Some important examples are $(\eta(x)\phi')' + \lambda\eta(x)\phi = 0$ or $\phi'' + \lambda\rho(x)\phi = 0$. The different types are all equivalent, via the Liouville transform if the coefficients are smooth enough, and this leads to certain equivalences among the various inverse spectral problems which can be formulated. But from a computational point of view, it may be more appropriate to treat each form directly.
- *Inverse spectral problems for second-order differential operators with singular points:* An important special case is

$$\phi'' + \left(\lambda - \frac{\ell(\ell+1)}{x^2} - V(x)\right)\phi = 0 \quad 0 < x < 1 \tag{31}$$

for $\ell = 0, 1, \dots$ which arises from the corresponding 3-D problem after separation of variables. The strong singularity at the origin generally prevents any straightforward use of the methods described above.

- *Inverse spectral problems with partially known coefficients:* One well-studied case of this is the problem posed in [7] of determining $V(x)$ from the eigenvalues $\{\lambda_n\}_{n=1}^{\infty}$, assuming that $V(x)$ is known on one half of the interval.
- *Inverse spectral problems with eigenparameter-dependent boundary conditions:* A number of interesting direct and inverse spectral problems may be stated in the form of problem (1) with the boundary condition at $x = 1$ replaced by

$$\phi'(1) = f(\lambda)\phi(1) \tag{32}$$

for some choice of f. For example, the interior transmission eigenvalue problem introduced in [2] leads to the case $f(\lambda) = \sqrt{\lambda}\cot\sqrt{\lambda}a$ for a certain parameter a. Numerical methods for the corresponding inverse spectral problem are studied in [11]. Another interesting example which may be viewed in this framework is the inverse resonance problem for a compactly supported potential, which may be viewed as the case $f(\lambda) = i\sqrt{\lambda}$ (see, e.g., [8]). The case of a linear fractional transformation $f(\lambda) = (a\lambda + b)/(c\lambda + d)$ is studied in [10].

References

1. Andrew, A.L.: Computing Sturm-Liouville potentials from two spectra. Inverse Probl. **22**(6), 2069–2081 (2006). doi:10.1088/0266-5611/22/6/010. http://dx.doi.org/10.1088/0266-5611/22/6/010
2. Colton, D., Monk, P.: The inverse scattering problem for time-harmonic acoustic waves in an inhomogeneous medium. Q. J. Mech. Appl. Math. **41**(1), 97–125

(1988). doi:10.1093/qjmam/41.1.97. http://dx.doi.org/10.1093/qjmam/41.1.97
3. Fabiano, R.H., Knobel, R., Lowe, B.D.: A finite-difference algorithm for an inverse Sturm-Liouville problem. IMA J. Numer. Anal. **15**(1), 75–88 (1995). doi:10.1093/imanum/15.1.75. http://dx.doi.org/10.1093/imanum/15.1.75
4. Freiling, G., Yurko, V.: Inverse Sturm-Liouville Problems and Their Applications. Nova Science Publishers, Huntington (2001)
5. Gel'fand, I.M., Levitan, B.M.: On the determination of a differential equation from its spectral function. Am. Math. Soc. Trans. (2) **1**, 253–304 (1955)
6. Gladwell, G.M.L.: Inverse problems in vibration. In: Solid Mechanics and Its Applications, vol. 119, 2nd edn. Kluwer Academic, Dordrecht (2004)
7. Hochstadt, H., Lieberman, B.: An inverse Sturm-Liouville problem with mixed given data. SIAM J. Appl. Math. **34**(4), 676–680 (1978)
8. Korotyaev, E.: Inverse resonance scattering on the half line. Asymptot. Anal. **37**(3–4), 215–226 (2004)
9. Levitan, B.M.: Inverse Sturm-Liouville Problems. VSP, Zeist (1987). Translated from the Russian by O. Efimov
10. McCarthy, C.M., Rundell, W.: Eigenparameter dependent inverse Sturm-Liouville problems. Numer. Funct. Anal. Optim. **24**(1–2), 85–105 (2003). doi:10.1081/NFA-120020248. http://dx.doi.org/10.1081/NFA-120020248
11. McLaughlin, J.R., Polyakov, P.L., Sacks, P.E.: Reconstruction of a spherically symmetric speed of sound. SIAM J. Appl. Math. **54**(5), 1203–1223 (1994). doi:10.1137/S0036139992238218. http://dx.doi.org/10.1137/S0036139992238218
12. Pöschel, J., Trubowitz, E.: Inverse Spectral Theory. Volume 130 of Pure and Applied Mathematics. Academic, Boston (1987)
13. Röhrl, N.: A least-squares functional for solving inverse Sturm-Liouville problems. Inverse Probl. **21**(6), 2009–2017 (2005). doi:10.1088/0266-5611/21/6/013. http://dx.doi.org/10.1088/0266-5611/21/6/013
14. Rundell, W., Sacks, P.E.: Reconstruction techniques for classical inverse Sturm-Liouville problems. Math. Comput. **58**(197), 161–183 (1992). doi:10.2307/2153026. http://dx.doi.org/10.2307/2153026

Inverse Spectral Problems: 1-D, Theoretical Results

Mourad Sini
Johann Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Sciences, Linz, Austria

Introduction

Let p, q, and ρ be real-valued, bounded, and measurable functions defined on the interval $(0, 1)$ with p and ρ positive. We denote by λ_i and ϕ_i, $i \in \mathbb{N}$, the associated eigenvalues and the $L^2_\rho(0,1)$-orthonormal eigenfunctions of the Sturm-Liouville problem:

$$\begin{cases} -(pu')' + qu - \lambda\rho u = 0, \text{ in } (0,1), \\ u(0) = u(1) = 0, \end{cases} \tag{1}$$

where $L^2_\rho(0,1)$ is the $L^2(0,1)$-space with the scalar product $(f,g) := \int_0^1 f(x)g(x)\rho(x)dx$. If we replace in (1), $u(1) = 0$ by $(pu')(1) = 0$, then we have another sequence of eigenvalues and eigenfunctions, which we denote by $(\mu_i)_{i=1}^\infty$ and $(e_i)_{i=1}^\infty$, respectively. In the next sections, we will discuss the following two types of inverse spectral problems:

1. *The Borg-Levinson inverse spectral problem.* It consists of the reconstruction of some of the three coefficients p, q, and ρ from the spectral data $(\lambda_i, \mu_j)_{i=1}^\infty$.
2. *The Gelfand inverse spectral problem.* It consists of the reconstruction of some of the three coefficients p, q, and ρ from the spectral data $(\lambda_i, |(p\phi_i')(0)|)_{i=1}^\infty$.

Different boundary conditions rather than the one in (1) can be taken. In addition, other types of inverse spectral problems have been also considered in the literature. We can cite among others, for the case $p = \rho = 1$, for instance, the one related to the spectral data $(\lambda_n, \log\frac{|\phi_n'(1)|}{|\phi_n'(0)|})_{n\in\mathbb{N}}$ or to the mixed data, i.e., given $(\lambda_n)_{n\in\mathbb{N}}$ and the a priori information that q is symmetric with respect to the middle point $x_0 := \frac{1}{2}$. We cite also the spectral data consisting of the sequence $(\lambda_n)_{n\in\mathbb{N}}$ and the nodal points (or the zeros of the corresponding eigenfunctions ϕ_n), called the inverse nodal problem. More information on these cases can be found in the following references [3, 8–10], for instance. In this paper, we focus only on the Gelfand and the Borg-Levinson spectral problems. An observation we can make is that we cannot obtain more than one of the three coefficients p, q, and ρ. To see this, assume in addition that p and ρ are of class $C^2(0,1)$. We define the Liouville transformation $y(x) := \frac{1}{L}\int_0^x \sqrt{\frac{\rho}{p}}(t)dt$, $x \in [0,1]$, with $L := \int_0^1 \sqrt{\frac{\rho}{p}}(t)dt$. Using this transformation as a coordinate transformation, we can verify that the Gelfand as well as the Borg-Levinson spectral data related to the *general* form Sturm-Liouville equation $-(pu')' + qu - \lambda\rho u = 0$, in $(0,1)$ are equal to the ones of the *normal* form Sturm-Liouville equation

$-u'' + Vu - \lambda u = 0$, in $(0,1)$, where V, which is a bounded and measurable real valued function, is given as a combination of the three coefficients p, q, and ρ the dependent variable is scaled by the fourth root of $p\rho$. Note that this transformation is not valid for discontinuous coefficients p and ρ.

The literature on these $1-D$ inverse spectral problems is huge. So, instead of reviewing the known results, we chose to review some of the popular ideas for solving the Gelfand and the Borg-Levinson problem considering the Sturm-Liouville equation of the normal form. (We assumed the potential V to be bounded, but, of course, this is not optimal, and many of the results stated here are known for potentials belonging to larger spaces.) Indeed, in section "The Asymptotic Expansion Technique", we mention the asymptotic expansion technique used for the first time by Borg and then simplified by Levinson at the end of the 1940s, see [2, 7]. In section "The Integral Equation Technique", we explain briefly the integral equation method by Gelfand and Levitan introduced in the 1950s for solving the Gelfand inverse spectral problems; see [4]. During the period from the 1950s till the 1980s, these two approaches have been extensively studied by many authors; see the references [8–10, 12] for more information on these methods and the related results till mid-1980s. In section "The C-Property", we consider the method of the C-property by Ramm (see [11]) and in section "The Boundary Control Method" the so-called boundary control method by Belichev and Kurylev both introduced in the mid-1980s; see [1] for the original version and [5] for a different presentation. We describe these methods for proving the uniqueness results. However, we warn the reader that two of them (the Gelfand-Levitan and the boundary control methods) are reconstructive. In addition, it is worth mentioning that the boundary control method has been also stated for the multidimensional problems; see [1]. Our goal in this paper is to explain the ideas by highlighting, with details, the link between the spectral data and the main mathematical tool proposed in each of the mentioned approaches. Regarding the step from the main mathematical tool to the final result, either we give some details, when it is possible, or we provide an appropriate reference.

The starting point for solving these problems is the following asymptotic formulas for the eigenmodes (λ_n, ϕ_n) in terms of n; see [2] for the original proof or [6] for a more simplified proof using Volterra-type integral equations in addition to some complex analysis techniques.

Lemma 1 *The sequence of eigenvalues $(\lambda_n)_{n\in\mathbb{N}}$ has the following asymptotic expression:*

$$\lambda_n = n^2\pi^2 + \int_0^1 V(t)dt + O\left(\frac{1}{n}\right) \tag{2}$$

and the sequence of normalized eigenfunctions $(\phi_n(x))_{n\in\mathbb{N}}$ behaves as follows:

$$\begin{aligned} \phi_n(x) &= \sqrt{2}\sin(n\pi x) + O\left(\frac{1}{n}\right) \quad \textit{and} \\ \phi_n'(x) &= \sqrt{2}n\pi\cos(n\pi x) + O(1) \end{aligned} \tag{3}$$

for $n \to \infty$, uniformly for $x \in [0,1]$.

The Asymptotic Expansion Technique

The original idea of Borg and as simplified by Levinson for solving the Borg-Levinson inverse spectral problem goes as follows. We introduce the Cauchy problem satisfied by $u := u(x,\lambda)$

$$\begin{cases} -u'' + Vu - \lambda u = 0, & \text{in } (0,1), \\ u(0,\lambda) = 0, \text{ and } u'(0,\lambda) = 1 \end{cases} \tag{4}$$

and the one satisfied by $v := v(x,\lambda)$

$$\begin{cases} -v'' + Vv - \lambda v = 0, & \text{in } (0,1), \\ v(1,\lambda) = 0, \text{ and } v'(1,\lambda) = 1. \end{cases} \tag{5}$$

Similar to the asymptotic expansion (3), we have

$$\begin{cases} u(x,\lambda) = \dfrac{\sin(\sqrt{\Re\lambda}x)}{\sqrt{\Re\lambda}} + O\left(\dfrac{e^{|\Im\lambda|x}}{|\Re\lambda|^2}\right) \\ u'(x,\lambda) = \sin(\sqrt{\Re\lambda}x) + O\left(\dfrac{e^{|\Im\lambda|x}}{|\Re\lambda|}\right) \end{cases} \tag{6}$$

for $|\lambda| \to \infty$, uniformly in $[0,1]$. Note that $\frac{\phi_n}{\phi_n'(0)}$ satisfies (4) and $\frac{\phi_n}{e_n'(1)}$ satisfies (5) with $\lambda := \lambda_n$. We define the characteristic function $w(\lambda) := u(1,\lambda)$. It is an entire function and has as zeros the eigenvalues λ_n, $n = 1, 2, \ldots$ The expansion (6) implies that $w(\lambda)$ is entire of order $1/2$, and hence, by the Hadamard's

factorization theorem, it is completely characterized by its zeros, λ_n, $n = 1, 2, \ldots$, i.e., $w(\lambda) = C(V)\Pi_{n=1}^{\infty}\left(1 - \frac{\lambda}{\lambda_n}\right)$ with some constant $C(V)$. Let now V_1 and V_2 have the same Borg-Levinson spectral data. We define $u_j(x,\lambda)$ as the solution of (4) for potential V_j. Hence the corresponding characteristic functions $w_j(\lambda)$ satisfy $w_1 = w_2 =: w$ as functions since $C(V_1) = C(V_2)$ from the first property in (6). We define also $v_j(x,\lambda)$ to be the solution of (5) for the potential V_j. From (4) and (5), we deduce that

$$u_j(x,\lambda_n) = C_n v_j(x,\lambda_n) \tag{7}$$

where C_n is independent of $j, j = 1, 2$. Indeed, it is clear that $u_j(x,\lambda_n) = C_n^j v_j(x,\lambda_n)$ (since $u_j(x,\lambda_n) = \frac{\phi_n^j(x)}{(\phi_n^j)'(0)}$ and $v_j(x,\lambda_n) = \frac{\phi_n^j(x)}{(\phi_n^j)'(1)}$). But $u_j'(1,\lambda)$ is the characteristic function associated to the mixed boundary conditions $u(0) = u'(1) = 0$. Hence, similar to $w(\lambda)$, it is characterized by its eigenvalues $\mu_n, n = 1, 2, \ldots$. Since $u_j'(1,\lambda_n) = C_n^j$, then $C_n^1 = C_n^2 =: C_n$, for n in $\mathbb{N}$. Let us mention that, for this approach, the equality of the Borg-Levinson spectral data is used only to prove (7) and $u_1(1,\lambda) = u_2(1,\lambda) =: w(\lambda)$.

The main new argument of Levinson (see [7]) starts from here. He defines the following function

$$H(x,\lambda) := \frac{1}{w(\lambda)} v_2(x,\lambda) \int_0^x u_1(\xi,\lambda) f(\xi) d\xi, \quad \forall \lambda \neq \lambda_n, \tag{8}$$

where $f \in C_0^1[0,1]$. Using the property (6) and an appropriate contour of integration, he shows that

$$\int_{\Gamma_N} H(x,\lambda) d\lambda - \pi i f(x) \to 0, \ N \to \infty \tag{9}$$

where Γ_N is a circle of center $\lambda = 0$ and radius between $\lambda_N^{1/2}$ and $\lambda_N^{3/2}$. Applying the residue theorem to the left-hand side of (9) and using the identity (7), for $j = 2$, we obtain $f(x) = 2\sum_{n=1}^{\infty} \frac{u_2(x,\lambda_n)\int_0^x u_1(t,\lambda_n) f(t) dt}{C_{nw}'(\lambda_n)}$. Applying the same calculations to $\frac{1}{w(\lambda)} u_2(x,\lambda)\int_x^1 v_1(\xi,\lambda) f(\xi) d\xi$ instead of $H(x,\lambda)$, we obtain $f(x) = 2\sum_{n=1}^{\infty} \frac{u_2(x,\lambda_n)\int_x^1 u_1(t,\lambda_n) f(t) dt}{C_{nw}'(\lambda_n)}$. Summing up these last two identities, we get the first expansion of f: $f(x) = \sum_{n=1}^{\infty} \frac{u_2(x,\lambda_n)\int_0^1 u_1(t,\lambda_n) f(t) dt}{C_{nw}'(\lambda_n)}$. Now exchanging the roles of u_1 and v_1 by u_2 and v_2, we obtain the second expansion of f: $f(x) = \sum_{n=1}^{\infty} \frac{u_2(x,\lambda_n)\int_0^1 u_2(t,\lambda_n) f(t) dt}{C_{nw}'(\lambda_n)}$. As a conclusion of these two expansions, we have

$$\sum_{n=1}^{\infty} \frac{u_2(x,\lambda_n)\int_0^1 [u_2(t,\lambda_n) - u_1(t,\lambda_n)] f(t) dt}{C_n w'(\lambda_n)} = 0.$$

Finally, using orthogonality properties of $u_2(x,\lambda_n)$, $n \in \mathbb{N}$, and choosing $f(t) := \sin(n\pi x)$, for instance, we deduce that $u_1(x,\lambda_n) = u_2(x,\lambda_n)$, in $[0,1]$, which implies that $V_1 = V_2$.

The Integral Equation Technique

We give here the main idea of the approach by Gelfand and Levitan in their seminal paper [4], for solving the Gelfand inverse spectral problem. We start by stating the following key theorem which relates solutions of the Cauchy problems for the equations $-u'' + V_j u - \lambda u = 0$, $j = 1, 2$ via a Volterra integral operator of which kernel is the solution of a Goursat problem with potential $V_1 - V_2$; see [6].

Theorem 1 *Let $u_j(\cdot,\lambda) \in C^2[0,1]$ be the solutions of the hyperbolic problem:*

$$\begin{cases} -u_j'' + V_j u_j = \lambda u_j, \ in \ (0,1), \ u_j(0,\lambda) = 0, \\ j = 1, 2 \ u_1'(0,\lambda) = u_2'(0,\lambda). \end{cases} \tag{10}$$

Let also $K \in C(\overline{\Delta})$ be the solution of the Goursat type problem

$$\begin{cases} \dfrac{\partial^2}{\partial x^2} K(x,t) - \dfrac{\partial^2}{\partial t^2} K(x,t) + (V_1(t) - V_2(x)) K(x,t) = 0, \ \ in \ \Delta \\ K(x,0) = 0, \ \ in \ [0,1] \\ K(x,x) = \dfrac{1}{2}\int_0^x (V_1(t) - V_2(t)) dt, \ \ in \ [0,1] \end{cases} \tag{11}$$

where $\Delta := \{(x,t) \in \mathbb{R}^2,\ 0 < t < x < 1\}$. *Then we have*

$$u_1(x,\lambda) = u_2(x,\lambda) + \int_0^x K(x,t)u_2(t,\lambda)dt, \quad \text{in } [0,1],\ \lambda \in \mathbb{C}. \tag{12}$$

Let us now explain how this theorem can be used to prove the uniqueness property. Assume that both the potentials V_1 and V_2 have the same eigenvalues $\lambda_n^1 = \lambda_n^2$ and the same traces of eigenfunctions $(\phi_n^1)'(0) = \pm(\phi_n^2)'(0)$, $n \in \mathbb{N}$. First, recall that $u_j(x,\lambda_n) = \frac{\phi_n^j(x)}{(\phi_n^j)'(0)}$, $j = 1,2$, satisfies (10); hence they also satisfy (12). Since $u_j(1,\lambda_n) = 0$, then $\int_0^1 K(1,t)\phi_n^2(t)dt = 0, \forall n \in \mathbb{N}$, which implies from the denseness of the eigenfunctions $\phi_n^2, n \in \mathbb{N}$, in $L^2(0,1)$ that

$$K(1,t) = 0,\ t \in [0,1]. \tag{13}$$

Second, it is shown (see [13], for instance) that from the equality of the Gelfand spectral, we have $(\phi_n^1)'(1) = \pm C(\phi_n^2)'(1)$ where C is constant. By the asymptotic expansion in (6), we deduce that $C = 1$. Using the representation (12) applied for λ_n and taking the derivative and then the trace on the point $x_0 = 1$, we obtain $\int_0^1 \frac{\partial}{\partial x}K(1,t)\phi_n^2(t)dt = 0,\ \forall n \in \mathbb{N}$, from which we deduce that

$$\frac{\partial}{\partial x}K(1,t) = 0,\ t \in [0,1]. \tag{14}$$

Resuming, we have shown that $K(x,t)$ satisfies the Cauchy problem in Δ given by the first equation in (11) and the initial conditions (13) and (14). From the uniqueness of the solutions of this Cauchy problem (see [6]), we deduce that K is identically zero. As a conclusion, we obtain from the last equation of (11) that $\int_0^x (V_1(x) - V_2(x))dx = 0$, in $[0,1]$, and hence $V_1 = V_2$.

The C-Property

A. Ramm introduced a method for proving the uniqueness property for one-dimensional inverse spectral and inverse scattering problems; see [11] for more details. It is based on the following property which he called the C-property. Let $u_j(x,\lambda)$ be the solution of (4) for $V = V_j$, $j = 1,2$, and then we have the following property; see [11] for the proof.

Theorem 2 *The set of products* $(u_1(\cdot,\lambda)u_2(\cdot,\lambda))_{\lambda>0}$ *is dense in* $L^1(0,1)$, *i.e., let* $h \in L^1(0,1)$ *such that* $\int_0^1 h(x)u_1(x,\lambda)u_2(x,\lambda)dx = 0$ *for every* $\lambda > 0$ *then* $h = 0$.

Let us explain how this result answers the uniqueness question of the Gelfand inverse spectral problem. Multiplying the first equation of (4) corresponding to $j = 1$ by $u_2(\cdot,\lambda)$ and conversely the one corresponding to $j = 2$ by $u_1(\cdot,\lambda)$, integrating by parts and taking the difference, we obtain

$$\int_0^1 (V_1 - V_2)(x)u_1(x,\lambda)u_2(x,\lambda)dx = u_1'(1,\lambda)u_2(1,\lambda) - u_2'(1,\lambda)u_1(1,\lambda),\ \forall\lambda \in \mathbb{C}. \tag{15}$$

As a next step, we show that the equality of the Gelfand spectral data implies that

$$u_1'(1,\lambda)u_2(1,\lambda) - u_2'(1,\lambda)u_1(1,\lambda) = 0,\ \forall\lambda \in \mathbb{C}. \tag{16}$$

Hence the C-property, i.e., Theorem 2, implies that $V_1 = V_2$.

In the following lines, we give a very short justification of (16). As we explained in section "The Asymptotic Expansion Technique", $u_j(1,\lambda)$ is completely characterized by its eigenvalues. Hence $u_1(1,\lambda) = u_2(1,\lambda)$. Remark that we need only the equality of the eigenvalues to obtain this equality. If in addition we have the equality of the traces of the eigenfunctions, then we have the equality of the derivatives. We state this in the following lemma.

Lemma 2 *If the Gelfand spectral data are equal for* $j = 1,2$, *then we have the identity*

$$u_1'(1,\lambda) = u_2'(1,\lambda),\ \forall\lambda \in \mathbb{C}. \tag{17}$$

Proof We introduce the function $\overline{u}_j$ satisfying the problem:

$$\begin{cases} -\overline{u}_j'' + V_j\overline{u}_j = 0, \ \text{ in } (0,1),\ j = 1,2 \\ \overline{u}_j(0,\lambda) = 0,\ \overline{u}_j(1,\lambda) = u_j(1,\lambda). \end{cases} \tag{18}$$

We set $w_j := u_j - \overline{u}_j$, and then it satisfies

$$\begin{cases} -w_j'' + V_j w_j = \lambda u_j, \ \text{ in } (0,1), \ j = 1,2 \\ w_j(0,\lambda) = w_j(1,\lambda) = 0. \end{cases} \tag{19}$$

Multiplying (19) by ϕ_n^j and integrating by parts, we obtain

$$\int_0^1 w_j(x)\phi_n^j(x)dx = -\frac{\lambda}{(\lambda-\lambda_n)\lambda_n}(\phi_n^j)'(1)u_j(1,\lambda). \tag{20}$$

Using (20), we write $w_j = \sum_1^\infty \left[\int_0^1 w_j(x)\phi_n^j(x)dx\right] \phi_n^j = -\sum_1^\infty \frac{\lambda(\phi_n^j)'(1)u_j(1,\lambda)}{(\lambda-\lambda_n)\lambda_n}\phi_n^j$. Taking the derivative and the trace on the point $x = 1$, we have $w_j'(1,\lambda) = -\sum_{n=1}^\infty \frac{\lambda|(\phi_n^j)'(1)|^2 u_j(1,\lambda)}{(\lambda-\lambda_n)\lambda_n}$. From the equality of the Gelfand spectral data(We explained in the previous section how the Gelfand spectral data imply that $|(\phi_n^1)'(1)| = |(\phi_n^2)'(1)|$, $\forall n \in \mathbb{N}$.) and the equality $u_1(1,\lambda) = u_2(1,\lambda)$, shown before, we see that $w_1'(1,\lambda) = w_2'(1,\lambda)$. Hence

$$u_1'(1,\lambda) - u_2'(1,\lambda) = \overline{u}_1'(1,\lambda) - \overline{u}_1'(1,\lambda). \tag{21}$$

Now, remark that $\overline{u}_j(x,\lambda) = \overline{\overline{v}}_j(x)u_j(1,\lambda)$ where $\overline{\overline{v}}_j$ is the solution of (18) replacing $u(1,\lambda)$ by 1. Hence $\overline{u}_j'(1,\lambda) = (\overline{\overline{v}}_j)'(1)u_j(1,\lambda)$. Recalling that $u_1(1,\lambda) = u_2(1,\lambda)$, the identity (21) becomes

$$u_1'(1,\lambda) - u_2'(1,\lambda) = ((\overline{\overline{v}}_1)'(1) - (\overline{\overline{v}}_2)'(1))u_1(1,\lambda). \tag{22}$$

From the identities in (6) taken for λ real and positive, the identity (22) can be written as

$$O\left(\frac{1}{\lambda}\right) = \left[(\overline{\overline{v}}_1)'(1) - (\overline{\overline{v}}_2)'(1)\right]\left[\frac{\sin(\sqrt{\lambda})}{\sqrt{\lambda}} + O\left(\frac{1}{\lambda^2}\right)\right] \tag{23}$$

which implies that $(\overline{\overline{v}}_1)'(1) - (\overline{\overline{v}}_2)'(1) = 0$ and hence $u_1'(1,\lambda) - u_2'(1,\lambda) = 0$. This ends the proof of Lemma 2.

The Boundary Control Method

The boundary control method introduced by Belishev (see [1]) is based on a combination of properties of the solutions of dynamical problems with the control theory of partial differential equations. Comparing it to the previous methods, it has the potential to be applied to the higher dimension inverse spectral and dynamical problems. The reader can refer to the review works [1] and [5] for more details. In this section, we show the main ideas of this theory needed to solve the $1-D$ Gelfand inverse spectral problem. For this, we state first the following hyperbolic problem related to our Sturm-Liouville model:

$$\begin{cases} \frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} + Vu = 0, \ \text{ in } \ (0,T)\times(0,1), \\ u(t,0) = f(t), \ u(t,1) = 0, \ t \in (0,T) \\ u(0,x) = \frac{\partial u}{\partial t}(0,x) = 0, \ x \in (0,1) \end{cases} \tag{24}$$

where $f \in H^1(0,T)$ such that $f(0) = 0$ and T is a positive constant. This problem is well posed. We set u^f its solution. The justification of the boundary control method for the $1-D$ problems is based on the following arguments:

1. **Domain of influence of the waves**. The support of $u^f(t,x)$ is given explicitly by the speed of propagation(For the general form Sturm-Liouville model (1), the speed of propagation is $\int_0^x \sqrt{\frac{\rho}{p}}(t)dt$; hence $\Gamma_t = \left\{x \in (0,1), \ \int_0^x \sqrt{\frac{\rho}{p}}(t)dt < t\right\}$.) (in our case it equals 1), i.e., $\{(t,x) \in (0,T)\times(0,1), \ x < t\}$. For $t > 0$ fixed, we set $\Gamma_t := \{x \in (0,1), \ x < t\} = (0,t)$.
2. **Fourier expansion of the waves**. We use the sequence $(\phi_n, \lambda_n)_{n\in\mathbb{N}}$ of the eigenvalues and eigenfunctions of the corresponding Sturm-Liouville equation with Dirichlet boundary conditions to represent u^f as follows:

$$u^f(t,x) = \sum_{i=1}^\infty u_i^f(t)\phi_i(x) \tag{25}$$

where the Fourier coefficients $u_i^f(t) := \int_0^1 u^f(t,x)\phi_i(x)dx$ are completely characterized by the spectral data $(\phi_n'(0), \lambda_n)_{n\in\mathbb{N}}$, i.e., $u_i^f(t) = (\phi_n)'(0)\int_0^1 f(s)\frac{\sin(\sqrt{\lambda_i}(t-s))}{\sqrt{\lambda_i}}ds$, since it is the solution of the Cauchy problem (Replace $\sqrt{\lambda_i}$ by $\sqrt{|\lambda_i|}$ for possible negative eigenvalues λ_i or $\frac{\sin(\sqrt{\lambda_i}(t-s))}{\sqrt{\lambda_i}}$ by $t-s$ if $\lambda_i = 0$.)

$$\begin{cases} \frac{\partial^2 u^f}{\partial t^2} - \lambda_i u^f = (\phi_n)'(0)f(t), \ \text{ in } \ (0,T), \\ u_i^f(0) = \frac{d}{dt}u_i^f(t) = 0, \ t \in (0,T). \end{cases} \tag{26}$$

3. **Boundary controllability**. There are two types of boundary controllability. First, the exact boundary controllability for the problem (24) is to find, for every fixed $t \in (0,T)$, for every $z(x)$ in $L^2(\Gamma_t)$ a function (i.e., a control) $f \in L^2(0,T)$ such that $u^f(t,x) = z(x)$. Second, we have the approximate boundary controllability where we replace the equality $u^f(T,x) = z(x)$ by an approximation; see [5], for instance. The second property is enough for our purpose.

Based on these three arguments, we prove the following theorem which characterizes fully the eigenfunctions ϕ_n in Γ_t, for every $t \leq 1$ using only the Gelfand spectral data.

Theorem 3 *Let V_j, $j = 1,2$ be two potentials such that the corresponding Gelfand spectral data $(\lambda_i^j, |(\phi_i^j)'(0)|)_{i\in\mathbb{N}}$, $j = 1,2$, are equal. Then, we have*

$$\int_0^t (\phi_i^1)^2(x)dx = \int_0^t (\phi_i^2)^2(x)dx, \ \forall i \in \mathbb{N}, \ \forall t \in (0,T). \tag{27}$$

From (27), we have $(\phi_i^1)^2(x) = (\phi_i^2)^2(x)$, for $x \in [0,T]$ (or for $x \in [0,1]$ if $T \geq 1$) and $i \in \mathbb{N}$. Taking $i = 1$, we have $V_1 = \frac{(\phi_1^1)''-\lambda_1\phi_1^1}{\phi_1^1} = \frac{(\phi_1^2)''-\lambda_1\phi_1^2}{\phi_1^2} = V_2$ in $(0,1)$ knowing that the eigenfunction ϕ_1^j never vanish in $(0,1)$.

The proof of Theorem 3 goes as follows. Let $(f_k)_{k\in\mathbb{N}}$ be a dense set in $H_0^1(0,1)$. From the well-posedness of the problem (24) and the approximate boundary controllability, we deduce that finite combinations of the functions $u^{f_k}(t,x)$ is dense in $L^2(0,t)$. From the second argument we know that the Fourier coefficients of u^{f_k} can be reconstructed from the Gelfand spectral data. By a Gram-Schmidt orthonormalization procedure, we can find an orthogonal basis of $L^2(0,t)$ given by combinations of u^{f_k}, i.e., $v_s := \sum_1^{n(s)} d_s u^{f_i}$. By linearity, we have $v_s = u^{g_s}$ where $g_s := \sum_1^{n(s)} d_s f_i$. Now, we write $\phi_j = \sum_{k=1}^{\infty} \left[\int_0^t \phi_j(x)v_k(t,x)dx\right] v_k(t,x)$ in $(0,t)$, and hence

$$\int_0^t \phi_i(x)\phi_j(x)dx = \sum_{k=1}^{\infty} \int_0^t \phi_j(x)v_k(t,x)dx \int_0^t v_k(t,x)\phi_i(x)dx. \tag{28}$$

Again from the second argument above, we know that

$$\int_0^t \phi_j(x)v_k(t,x)dx = (\phi_j)'(0) \int_0^t g_k \frac{\sin\sqrt{\lambda_j}(t-s)}{\sqrt{\lambda_j}}ds. \tag{29}$$

Taking $i = j$ in (28) and using (29), we see that the Gelfand spectral data completely characterize the quantities $\int_0^t (\phi_i(x))^2dx$, $i \in \mathbb{N}$. This ends the proof of Theorem 3.

References

1. Belishev, M.I.: Recent progress in the boundary control method. Inverse Probl. **23**(5), R1–R67 (2007)
2. Borg, G.: Eine Umkerung der Sturm–Liouville Eigenwertarfgabe. Acta Math. **78**, 1–96 (1946)
3. Hald, O.H., McLaughlin, J.R.: Solutions of inverse nodal problems. Inverse Probl. **5**(3), 307–347 (1989)
4. Gel'fand, I.M., Levitan, B.M.: On the determination of a differential equation from its special function. Izv. Akad. Nauk SSR. Ser. Mat. **15**, 309–360 (1951) (Russian); English transl. in Am. Math. Soc. Trans. Ser. **2**(1), 253–304 (1955)
5. Katchalov, A., Kurylev, Y., Lassas, M.: Inverse Boundary Spectral Problems. Chapman/Hall/CRC Monographs and Surveys in Pure and Applied Mathematics, vol. 123, p. xx+290. Chapman/Hall/CRC, Boca Raton (2001)
6. Kirsch, A.: An introduction to the mathematical theory of inverse problems. Applied Mathematical Sciences, vol. 120, p. x+282. Springer, New York (1996)
7. Levinson, N.: The inverse Sturm-Liouville problem. Math. Tidsskr. B **25**, 25–30 (1949)
8. Levitan, B.M.: Inverse Strum-Liouville Problems. VNU Science, Ultrecht (1987)
9. McLaughlin, J.R.: Analytical methods for recovering coefficients in differential equations from spectral data. SIAM Rev. **28**(1), 53–72 (1986)
10. Poschel, J., Trubowitz, E.: Inverse Spectral Theory. Pure and Applied Mathematics, vol. 130, p. x+192. Academic, Boston (1987)
11. Ramm, A.: Inverse Problems. Mathematical and Analytical Techniques with Applications to Engineering. Springer, New York (2005)
12. Rundell, W., Sacks, P.: Reconstruction techniques for classical inverse Sturm-Liouville problems. Math. Comput. **58**(197), 161–183 (1992)
13. Sini, M.: Some uniqueness results of discontinuous coefficients for the one-dimensional inverse spectral problem. Inverse Probl. **19**(4), 871–894 (2003)

Inversion Formulas in Inverse Scattering

Clifford J. Nolan
Department of Mathematics and Statistics, University of Limerick, Limerick, Ireland

Abstract

We survey some important inversion formulas in inverse scattering with a particular emphasis on those having their roots in the Radon transform. The history of the latter transform and its inversion spans approximately a century. While the Radon transform had a modest beginning, it now forms a cornerstone of modern-day medical imaging, nondestructive testing of materials, etc. It is therefore fitting that we collect inversion formulas from diverse sources together in this article.

Synonyms

Artefacts; Asymptotic; Backprojection; Image; Inversion; Microlocal

Introduction

Inverse scattering is a term that is widely used in both the mathematics and physics. Due in part to the maturity of the subject, *inverse scattering* has come to mean quite different things to different research communities. For example, it can mean nondestructive testing of materials using ultrasound, or it might mean the applications of semigroups connected with the wave equation, as in Lax and Philips' seminal work [1].

Physics and mathematics have common ground when it comes to approximating scattered waves in the guise of the *Born approximation*. From the mathematical perspective, this shows up as a linearization of the wave equation. However, there are situations where a rigorous justification of this "approximation" is still lacking, and therefore, one should be guided by physical principles and experiments. At the same time, research continues into a mathematical justification.

Because of the diversity of the meaning of the subject matter, we have chosen examples of inverse scattering which are united by a common theme: the Radon transform. This is because many situations arise in practice where scattered waves can be approximated as an integral transform of wave equation coefficients over lines, curves, surfaces, etc. Therefore, when one measures such scattered waves, one is measuring a Radon or generalized Radon transform (GRT) of the coefficients. The goal is to recover these coefficients from the measurements.

The Radon Transform

Since our unifying theme is the Radon transform and its inversion, we begin our discussion with a brief description of what the Radon transform actually is and then proceed to discuss some of the more common ways in which it may be inverted.

In its *simplest setting*, the Radon transform takes a function of two variables $f(x, y)$ (having suitable decay properties) and evaluates line integrals of this function. Therefore, the Radon transform is a function on the space of lines. We parametrize a line by specifying its distance (s) from the origin and the direction ($\theta \in S^1$) to which it is perpendicular. For example, such a line is described by the following set of points:

$$L(\theta, s) = \{x \in \mathbb{R}^2 | \, x \cdot \theta = s\} \tag{1}$$

The Radon transform Rf of f is defined as the following line integral:

$$Rf(\theta, s) = \int_{L(\theta,s)} f \, dl \tag{2}$$

The latter definition has obvious extensions to higher dimensions (where the integration takes place over $n-1$ dimensions and $\theta \in S^{n-1}$), e.g., integrals of functions over hyperplanes in three or higher dimensions.

One can also consider integrals of functions over a family of submanifolds, i.e., a GRT. As pointed out in a survey article by Strauss [2], Radon was somewhat fortunate to have his name attached to such integral transforms. Strauss supports his claim by point-

ing out that only 3 years earlier, Funk [3] investigated a similar integral transform, pertaining to integrals of functions over great circles on a sphere (now called the "Funk transform"). Funk obtained inversion formulas for his transform, and it seems clear that Radon was influenced by (and refers to) this work.

Inversion of the Radon Transform

A simple way to obtain an inversion formula for (2) is to perform an elementary calculation [4] that shows that

$$\widehat{Rf}(\theta,\sigma) = \hat{f}(\sigma\theta) \tag{3}$$

where σ is the Fourier variable dual to s, so that the left-hand side refers to one-dimensional Fourier transform and the right-hand side refers to a regular two-dimensional Fourier transform. Formula (3) is referred to as the "Projection Slice Theorem" and illustrates a connection between the Fourier and Radon transforms. It immediately gives an inversion formula for the Radon transform: inverse Fourier transform the one-dimensional Fourier transform of a Radon transform. The same formula is valid in higher dimensions, and the same comment often applies to related transforms which we discuss below. There are many inversion techniques based on variants of this formula.

While the relationship with the Fourier transform can be useful, it is perhaps not as instructive as another common constructive inversion technique which is based on the adjoint of the Radon transform. A straightforward calculation of the formal L^2-adjoint R^* of R is seen to be the operation of integrating over all lines that go through the point of evaluation:

$$R^*g(x) = \int_{S^1} g(\theta, x\cdot\theta)\, d\theta \tag{4}$$

and in higher dimensions, S^1 is replaced by S^{n-1}.

Clearly, the lines going through any particular point x influence the Radon transform of f, and it seems natural that if we were to evaluate $R^*g(x)$ with $g = Rf$, then these lines would contribute to $R^*g(x)$ in (4), while lines that do not go through x don't carry information about $f(x)$. It is not surprising then to learn of the following inversion formula:

$$f = (4\pi)^{-1} I_1 R^* Rf \tag{5}$$

where I_n is the Riesz potential, defined as follows:

$$\widehat{I_m g}(\xi) = |\xi|^{-m}\, \hat{g}(\xi) \tag{6}$$

valid for $0 < m < n$. Let's take stock of Formula (5) for a moment. It provides an inversion formula which is given by application of the adjoint R^* followed by application of a filter.

The analogue of the above calculation in three and higher dimensions ($n \in \mathbb{N}$) leads to the following general inversion formula ([4], p.10) valid, for example, when f belongs to the class of Schwartz functions $\mathcal{S}(\mathbb{R}^2)$:

$$f = \begin{cases} c_n R^* H \frac{d^{(n-1)}}{ds^{(n-1)}} Rf, & n \text{ even} \\ \\ c_n R^* \frac{d^{(n-1)}}{ds^{(n-1)}} Rf, & n \text{ odd} \end{cases} \tag{7}$$

where

$$c_n = \begin{cases} 2^{-1}(2\pi)^{1-n}(-1)^{(n-2)/2}, & n \text{ even} \\ \\ 2^{-1}(2\pi)^{1-n}(-1)^{(n-1)/2}, & n \text{ odd} \end{cases} \tag{8}$$

and H denotes the Hilbert transform with respect to the variable s.

Remark 1 Note the difference between the inversion formulae for even and odd dimensions; the former leads to a nonlocal inversion formula while the latter to a local formula. This is reminiscent of the qualitative difference between solutions of the wave equation in even and odd dimensions, and indeed, Helgason ([5], p. 1) refers to the fact that (apart from the filtering process) the Radon inversion formula is a decomposition into plane waves, as seen in Formula (4) above.

Scattering in the Context of the GRT

Geophysical Applications

In the mid-1980s, Gregory Beylkin published a paper [6] which revolutionized geophysical subsurface imaging, using high-frequency scattered seismic waves. The paper demonstrated how scattered seismic waves in the earth's subsurface could be modeled as the output of a GRT of the earth's *reflectivity function*. The latter function is

$$v(x) := c_0^{-2}(x) - c^{-2}(x) \tag{9}$$

where the speed of (acoustic) wave propagation is viewed as a superposition of a smooth (background) component c_0 and a highly oscillatory component δc, i.e.,

$$c(x) = c_0(x) + \delta c(x) \tag{10}$$

with δc encoding discontinuities in wave speed across interfaces of different materials, for example. The reflectivity can also model point inclusions, among other scatterers.

The basic assumptions made in [6] were as follows. Waves scattered just once from the time when they leave the source (usually a buried explosive) to when they are recorded back on the earth's surface (by a buried geophone). All waves which arrive at the geophone without scattering (usually strong signals, known as *first arrivals*) are filtered out. Beylkin also made an essential simplifying assumption that no caustics develop either in the incident or scattered waves.

Under the above assumptions, Beylkin showed that the scattered pressure field $\delta p(r, t)$ measured at receiver location r at time $t > 0$ could be written as the output of a GRT, which integrates the reflectivity function over a family space-time *move-out surfaces*, parametrized by (r, t). More precisely, the latter integral transform is a Fourier integral operator (FIO); see [7–10] for information on these operators which are studied in *microlocal analysis*. If we denote by δp the scattered acoustic pressure field due to high-frequency perturbations δc in the sound speed, Beylkin's result can be written symbolically as

$$\delta p = F \delta c \tag{11}$$

where F is a FIO. The latter FIO is asymptotically equivalent to the GRT mentioned above. In view of the previous section, it should not be a surprise that inversion of F involves the formal L^2-adjoint F^* of F. In fact, if we form an "image"

$$I = F^* \delta p \equiv F^* F \delta c \tag{12}$$

we see that this is effectively the result of applying F^*F to the unknown δc that we wish to recover. The latter image is referred to as the *migrated section* in geophysics literature. The method upon which geophysicists derive such a procedure is quite similar to our discussion on the Radon transform. In fact, they argue that for a scatterer to contribute to the data collected by each receiver, the scatterer must lie on an associated move-out surface, and by superimposing all such contributions (effected by integrating data over receiver locations at suitable time off-sets), the main contribution comes from constructive interference at the true scatterer location.

Beylkin showed that under the above assumptions, F^*F is a pseudodifferential operator (ΨDO). Furthermore, F^*F is elliptic when restricted to reflectivity functions whose singularities are visible in the data δp. This means that it's possible to follow-up F^* with a microlocal "filter" G (the analogue of I_m from the previous section):

$$GF^* p \sim \delta c \tag{13}$$

where $\sim$ is an asymptotic approximation which means that the left-hand side recovers δc except for where it is not visible using the scattering data δp. This is all that one can reasonably expect in any case.

The similarity between Formulae (5) and (13) is almost self-evident. Indeed, the only substantive difference is that the operator G cancels the geometrical spreading amplitude that is built into F. Also (13) is an asymptotic inversion formula in the sense that it only inverts for high frequencies that are contained in δc. This is not the only example where the inversion formula that emerges from a microlocal treatment matches very closely an exact inversion formula, as applied to the common-or-garden test functions like the Schwartz functions in the previous section.

In 1988, Rakesh [11] showed that even if one allows caustics to be present in either the incident or reflected waves, then F is still a FIO. And in 1997, Nolan and Symes [12] gave geometrical conditions, related to ray-geometry and source–receiver configurations that guarantee F^*F is a ΨDO. Therefore, when these geometrical conditions are satisfied, a similar inversion formula to (13) applies. In the case of sources and receivers varying independently over a codimension 1 submanifold of the earth's surface, various authors [13, 14] examined the effect of relaxing some of the latter assumptions.

Artifacts

Leading on from the last section, we comment on what can be done when an inversion formula is not available. Even if the above conditions for F^*F to be a ΨDO are not satisfied (so that an inversion formula like (13) is no longer available), it is common practice to *backproject* the data anyway. That is, one applies F^* to the data in the hope of gleaning certain information about the reflectivity function. At this stage, an examination of the wavefront relation Λ of the FIO F is necessary to understand the content of the image $F^*\delta p$. We point out that even though this discussion is in the context of a geophysical example, the conclusions apply in numerous other contexts as well.

The wavefront relation $\Lambda \subset T^*(Y \times X)$ is a *Lagrangian submanifold* of the *phase/cotangent space* $T^*(Y \times X)$ and is derived from the kinematical properties of the incident and scattered ray fields. The manifold X is the earth's subsurface. The manifold Y is the Cartesian cross-product of (i) the manifold where the sources and receivers are placed and (ii) the interval of time, over which the scattered waves are recorded. The main thing that needs to happen in order for F^*F to be a ΨDO (and thus obtain the usual inversion formula) is that the natural projection

$$\Lambda \longrightarrow T^*Y \tag{14}$$

is an embedding. This condition is known as the *Bölker* condition and is often not satisfied except under the assumptions like those given by Beylkin [6] and Nolan and Symes [12], for example. When the conditions are not satisfied, then the backprojected data will contain artifacts (e.g., see [12, 15]). In the final section, we give a brief explanation as to why such artifacts appear.

The Cone Beam and Attenuated Ray Transforms

If one considers weighted integrals of functions over lines, we arrive at a model for X-ray images. At the start of this century, Novikov [16] derived an inversion formula for this transform which obviously has important consequences for medical imaging. We briefly describe the model and the associated inversion formula here.

The attenuated ray transform can be defined (in two dimensions for ease of exposition, with generalization to higher dimensions possible) via the cone beam transform as follows. Let $f \in \mathcal{S}(\mathbb{R}^2)$ and for $a \in \mathbb{R}^2$, $\theta \in S^2$. The cone beam transform is defined by

$$Df(a,\theta) = \int_0^\infty f(a + t\theta)dt \tag{15}$$

Grangeat's Ph.D. thesis developed the following formula (see [4, 17]):

$$\frac{\partial}{\partial s} Rf(\theta, a \cdot \theta) = \int_{\omega \in \theta^\perp \cap S^2} \frac{\partial}{\partial \theta} Df(a,\omega)d\omega \tag{16}$$

The latter formula gives an inversion formula for the cone beam transform, given that we know how to invert the Radon transform. Note that $\theta = (\cos(\theta), \sin(\theta))$, $\theta^\perp = (-\sin(\theta), \cos(\theta))$.

Related to the cone beam transform is the attenuated ray transform, described by

$$R_\mu f(\theta, x) = \int f(x + t\theta)e^{-D\mu(x,\theta^\perp)} \tag{17}$$

for some $\mu \in \mathcal{S}(\mathbb{R}^n)$.

An efficient inversion formula for the attenuated ray transform can be obtained by following Novikov's formula [16], summarized by the following [4]. Let $h = (I + iH)R\mu/2$. Then for μ, $f \in \mathcal{S}(\mathbb{R}^2)$,

$$f(x) = (4\pi)^{-1} Re\ \nabla \cdot R^*_{-\mu}\left(\theta e^{-h} H e^h R_\mu f\right) \tag{18}$$

There have been many other papers since then, extending Novikov's work [18], in the quest for more efficient reconstructions.

Tensor Tomography

In the previous sections, the unknown quantity to be recovered or imaged was a scalar field, such as the reflectivity function, or density of a material. We now consider the situation that one encounters, for example, in elasticity or electromagnetism. Here, one is interested in recovering a tensor, such as the stress tensor, the electrical permittivity, etc. A perfect example involves both these areas at once, namely, photoelasticity. Since we don't have the space to develop the details of this example here, we will discuss the problem in the abstract and follow some results in Sharafutdinov's book [19]. The integral transform that

arises in these kinds of problems may be written as

$$If(\gamma) = \int_\gamma f_{i_1 i_2 \ldots i_m}(x(t))\, \dot{x}_1(t)\dot{x}_2(t)\ldots\dot{x}_m(t)\, dt \tag{19}$$

where $f \in C^\infty(S^m\tau'_M)$, the space of smooth covariant tensor fields of degree m. Here, γ is a geodesic of a manifold M, τ'_M is the cotangent bundle of M, and S^m denotes a section over M. Sharafutdinov refers to the *data* measured by these integrals as a *hodograph*, since it is initially motivated by *tensor tomography*, where the integral transforms measure sojourn times of rays between pairs of boundary points of the manifold M.

In such problems, it often happens that I in (19) has a kernel. This was also the case earlier when trying to invert for a scalar field; e.g., the Funk transform obviously vanishes on functions which are antisymmetric on great circles. However, here one can see even more scope for the existence of a kernel, so this is another obstruction to inversion without imposing additional assumptions.

When trying to invert (19), it is a good idea to decompose the tensor f into its potential and solenoidal parts (in analogue to Helmholtz's decomposition for a vector field):

$$f = {}^s f + dv,\ \delta\, {}^s f = 0 \tag{20}$$

where δ is the *divergence* and $-d$ is its dual with respect to the L^2 inner product. Then, following [19], we can write down the inversion formula

$${}^s f = (-\triangle)^{1/2}\left(\sum_{k=0}^{[m/2]} c_k i^k j^k\right)\mu^m If + du \tag{21}$$

where the operator i is defined by the property that $iu = u\delta$ and j is the dual of i. $[m/2]$ is the integral part of $m/2$. The coefficients c_k and the potential u are explicitly described in [19]. Also the integral moment operator μ^m is defined by

$$(\mu^m\phi)_{i_1 i_2 \ldots i_m}(x) = \omega_n^{-1}\int_\Omega \xi_{i_1}\ldots\xi_{i_m}\, \phi(x,\xi)\, d\omega(\xi) \tag{22}$$

where ω_n is the volume of the unit sphere Ω in dimension n.

It turns out that If only determines f up to an arbitrary summand dv. Also, If determines a system of local linear functionals WIf acting on If, where W is called the *Saint-Venant* operator. This is all the information that can be recovered from the hodograph.

The author of this article has noticed that there is a stark difference between the kernel of I in the current setting and its analogue in the microlocal setting. That is, when trying to recover the components of, say, the stress tensor, Sharafutdinov's work tells us what kind of kernel to expect. Aside from Sharafutdinov's general results, it is well known in the literature that usually only certain linear combinations of the stress tensor can be recovered (due to the nontrivial kernel). However, such kernels may become trivial when one is only inverting for the high-frequency components (e.g., see [20]).

The Unifying Theme: Backprojection

In the context of scattering, it should be obvious at this stage that whether we are looking to invert the integral transforms exactly or asymptotically, the adjoint of the transform is present as part of the inversion formula at some stage.

We already remarked why it was plausible to see R^* appearing in inversion of the Radon transform and it is reasonable to extend this to the GRTs that we have seen above too. Perhaps the microlocal point of view gives the clearest picture as to why one should expect application of the adjoint (i.e., backprojection) to appear in the inversion formulas. The wavefront relation Λ describes ordered pairs of singularities, or more precisely ordered pairs of *wavefront sets* $((y,\eta),(x,\xi))$ where $(x,\xi) \in WF(\delta c), (y,\eta) \in WF(\delta p)$. The integral transform F maps these wavefront sets or singularities (x,ξ) into (y,η). So Λ relates singularities in the model (δc) to their corresponding singularities in the data (δp). The adjoint maps singularities in the reverse direction, so its wavefront relation consists of ordered pairs $((x,\xi),(y,\eta))$. Therefore, provided the Bölker condition is satisfied, wavefront set elements (x,ξ) will be imaged at the correct location (due to the injectivity of the projection $\Lambda \to T^*Y$). Therefore, it is very natural to backproject the data in the hope of reconstructing an image without artifacts.

Finally, when the Bölker condition is not satisfied and one goes ahead and backprojects the data anyway,

one can now see a mechanism that explains how artifacts arise in an image – they are due to the fact that now Λ is a many-to-one relation.

References

1. Lax, P.D., Phillips, R.S.: Scattering Theory, Rev. edn. Academic Press, Boston (1989)
2. Strauss, R.S.: Radon Inversion - Variations on a Theme. Am. Math. Mon. **89**(6), 377–384 (1982)
3. Radon, J.: Über Flächen mit lauter geschlossenen geodätischen Linien. Math. Ann. **74**, 278–300 (1914)
4. Natterer, F.: Mathematical Methods in Image Reconstruction. SIAM Monographs on Mathematical Modeling and Computation. Society for Industrial and Applied Mathematics, Philadelphia (2001)
5. Helgason, S.: The Radon Transform. Birkhauser, Boston (1999)
6. Beylkin, G.: Imaging of discontinuities in the inverse scattering problem by inversion of a causal generalized Radon transform. J. Math. Phys. **26**, 99–108 (1985)
7. Duistermaat, J.J.: Fourier Integral Operators. Birkhäuser, Boston (1996)
8. Tréves, F.: Introduction to Pseudodifferential and Fourier Integral Operators, vols. 1–2. Plenum Press, New York (1982)
9. Duistermaat, J.J., Guillemin, V.W., Hörmander, L.: Matematics Past and Present, Fourier Integral Operators. Springer, Berlin Heidelberg (1991)
10. Grigis, A., Sjöstrand, J.: Microlocal Analysis for Differential Operators: An Introduction. London Mathematical Society. Cambridge University Press, Cambridge/New York (1994)
11. Rakesh, B.: A linearised inverse problem for the wave equation. Commun. PDE **13**, 573–601 (1988)
12. Nolan, C., Symes, W.: Global solution of a linearized inverse problem for the wave equation. Commun. Partial Differ. Equ. **22**, 919–952 (1997)
13. Ten Kroode, A.P.E., Smit, D.J., Verdel, A.R.: A microlocal analysis of migration. Wave Motion **28**, 149–172 (1998)
14. Stolk, C.: Microlocal analysis of a seismic linearized inverse problem. Wave Motion **32**, 267–290 (2000)
15. Nolan, C.J., Cheney, M.: Synthetic aperture inversion. Inverse Probl. **18**(1), 221–235 (2002)
16. Novikov, R.G.: An inversion formula for the attenuated X-ray transformation. Ark. Mat. **40**, 145–167 (2002)
17. Grangeat, P.: Mathematical framework of cone beam 3D reconstruction via the first derivative of the radon transform. In: Herman, G.T., Luis, A.K., Natterer, F. (eds.) Mathematical Methods in Tomography. Lecture Notes in Mathematics. Springer, Berlin (1991)
18. Boman, J., Strömberg, J.: Novikov's inversion formula for the attenuated radon transform – a new approach. J. Geom. Anal. **14**(2), 185–198 (2004)
19. Sharafutdinov, V.A.: Integral Geometry of Tensor Fields. VSP, Utrecht (1994)
20. Ryan, N.: High-frequency elastic wave inversion, Ph.D. thesis, University of Limerick (2010)

Invisibility Cloaking

Matti Lassas[1] and Graeme Milton[2]
[1]Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland
[2]Department of Mathematics, The University of Utah, Salt Lake City, UT, USA

Synonyms

Cloaking due to Anomalous Localized Resonance (CALR); Plasmonic cloaking

Glossary

Active cloaking Cloaking which uses one or more active sources, such as antennas, to generate appropriate fields to cloak the incoming field.

Broadband cloaking Cloaking over an entire interval of frequencies.

Exterior cloaking Cloaking where the cloaking region is outside the cloaking device.

Metamaterial An artificially structured composite material, often locally periodic, which has effective properties outside those usually found in nature.

Neutral inclusion An inclusion which is invisible to one or more applied fields.

Passive cloaking Cloaking where the cloak is composed only of passive materials which respond causally to applied fields, but which do not in themselves radiate energy.

Transformation optics Using the principle of invariance of electromagnetic equations (at fixed frequency) to map from one solution of Maxwell's equations in one geometry, to another solution of Maxwell's equations in another possibly more interesting geometry. The same principle applies to many other equations, including the conductivity equations and acoustic equations.

Abstract

We discuss recent mathematical theory of cloaking, that is, on making objects invisible. We describe three different approaches for this. The first approach, the

I

use of transformation optics, is based on the fact that the equations that govern a variety of wave phenomena, including electrostatics, electromagnetism, and acoustics, have transformation laws under changes of variables which allow one to design material parameters that steer waves around a hidden region, returning them to their original path on the far side. In the second one, cloaking by anomalous resonance, the field generated by a discrete collection of polarizable dipoles resonates with the cloaking device in such a way to almost cancel the field acting on the polarizable dipoles rendering them and the cloaking device almost invisible. In the third one, active exterior cloaking, active sources generate almost localized fields which cancel the incident field in a region to create a quiet zone, in which an object may be hidden.

Cloaking and Transformation Optics

There have been many scientific prescriptions for invisibility in various settings, starting from the first proposals [3, 19] introduced decades ago. However, since 2003 there has been a wave of serious theoretical proposals [1, 15, 16, 25, 30, 33, 37] in the physics and mathematics literature and a widely reported experiment by Schurig et al. [39], for cloaking devices – structures that would not only make an object invisible but also undetectable to electromagnetic waves, thus making it *cloaked*.

There are many alternative proposals for invisibility which we will describe next. The first one, called *transformation optics* [9,38,41], means the design of optical devices with customized effects on wave propagation, made possible by taking advantage of the transformation rules for the material properties of optics, the index of refraction $n(x)$ for geometric optics; the electrical permittivity $\varepsilon(x)$ and magnetic permeability $\mu(x)$ for vector optics, as described by Maxwell's equations; and the conductivity $\sigma(x)$ appearing in the static limit of electromagnetism.

To explain the principle of transformation optics, let us start with the conductivity equation with anisotropic conductivity. An anisotropic conductivity on a domain $\Omega \subset \mathbb{R}^n$ is defined by a symmetric, positive semi-definite matrix-valued function, $\sigma = [\sigma^{ij}(x)]_{i,j=1}^n$. In the absence of sources or sinks, a static electrical potential $u(x)$ in the domain Ω satisfies

$$\nabla\cdot\sigma\nabla u = \sum_{j,k=1}^{n} \frac{\partial}{\partial x^j}\sigma^{jk}(x)\frac{\partial}{\partial x^k}u(x) = 0. \quad (1)$$

The boundary value $u|_{\partial\Omega}$ corresponds to the voltage on the boundary, and the co-normal derivative, $B_\sigma u|_{\partial\Omega}$, corresponds to the current through the boundary. Here, $B_\sigma u = \sum_{j=1}^n \nu_j \sigma^{jk}\frac{\partial}{\partial x^k}u$, and ν is the unit normal vector of $\partial\Omega$. The set of all possible voltage-current pairs which can be observed on $\partial\Omega$ corresponds then to the set of Cauchy data of solutions u of Eq. (1). We denote the set of Cauchy data of solutions, which are the function space X and correspond to the conductivity σ, by

$$\Sigma_X(\sigma) = \{(u|_{\partial\Omega}, B_\sigma u|_{\partial\Omega});\ u \in X, \nabla\cdot\sigma\nabla u = 0\}. \quad (2)$$

For conductivities which are bounded both from below and above by positive constants, one usually considers solutions in the Sobolev space $X = H^1(\Omega)$.

Let us next consider Eq. (1) in different coordinates. Let $F(x) = (F^1(x), \ldots, F^n(x))$ be a diffeomorphism $F : \Omega \to \Omega$ with $F|_{\partial\Omega} =$ Identity. We consider the change of variables $y = F(x)$ and set $v = u\circ F^{-1}$, that is, $u(y) = v(F(x))$. Using the fact that u satisfies the conductivity equation (1) and the chain rule, one sees that v satisfies the conductivity equation $\nabla\cdot\tilde{\sigma}\nabla v = 0$ in Ω, where $\tilde{\sigma} = F_*\sigma$ is the push-forward of the conductivity σ by F given by

$$(F_*\sigma)^{jk}(y) = \frac{1}{\det[\frac{\partial F}{\partial x}(x)]}\sum_{p,q=1}^{n}\frac{\partial F^j}{\partial x^p}(x)\left.\frac{\partial F^k}{\partial x^q}(x)\sigma^{pq}(x)\right|_{x=F^{-1}(y)}. \quad (3)$$

Moreover, $v|_{\partial\Omega} = u|_{\partial\Omega}$ and the chain rule implies that $B_{\tilde{\sigma}}v|_{\partial\Omega} = B_\sigma u|_{\partial\Omega}$. Thus,

$$\Sigma_X(F_*\sigma) = \Sigma_X(\sigma) \quad (4)$$

for $X = H^1(\Omega)$. This implies that all conductivities $F_*\sigma$ with arbitrary boundary preserving diffeomorphism F give rise to the same electrical measurements at the boundary. This was first observed in [22] following a remark by Luc Tartar. For electromagnetism at fixed frequency this same idea is implicit in the work of [9].

The above has two physical interpretations. First one is that if we change coordinates in Ω using the diffeomorphism F and write the conductivity equation in the new coordinates, physical observations on the boundary $\partial\Omega$ do not change. The other interpretation of (4) is that if we keep the coordinates in Ω fixed and change the conductivity according to the formula (3), then the physical observations on the boundary does not change. This interpretation is the basis of the transformation optics.

Cloaking via Transformation Optics for Electrostatics

To obtain cloaking using the above equivalence of observations, i.e., (4), we use a *singular* transformation F stretching (or "blowing up") the origin to the ball $\overline{B}_1$, where $B_R \subset \mathbb{R}^3$ denotes the ball of radius R centered at origin. An example of such transformation is

$$F : B_2 \setminus \{0\} \to B_2 \setminus \overline{B}_1, \quad F(x) = \left(\frac{|x|}{2} + 1\right)\frac{x}{|x|}, \quad 0 < |x| < 2. \tag{5}$$

In the rest of the section, we reserve the notation F to denote the map (5).

In $\mathbb{R}^3$, we define the cloaking conductivity $\tilde{\sigma}$ by

$$\tilde{\sigma}(x) = (F_*\sigma_0)(x), \quad \text{for } 1 < |x| \le 2 \quad \text{and} \quad \tilde{\sigma}(x) = \gamma(x) I, \quad \text{for } |x| \le 1, \tag{6}$$

where $\sigma_0 = I$ and γ in the ball B_1 is non-degenerate, that is, $c_1 \le \gamma(x) \le c_2$, for $x \in B_1$ where $c_1, c_2 > 0$. In spherical coordinates $(r, \phi, \theta) \mapsto (r\sin\theta\cos\phi, r\sin\theta\sin\phi, r\cos\theta)$, $\tilde{\sigma}$ is given by

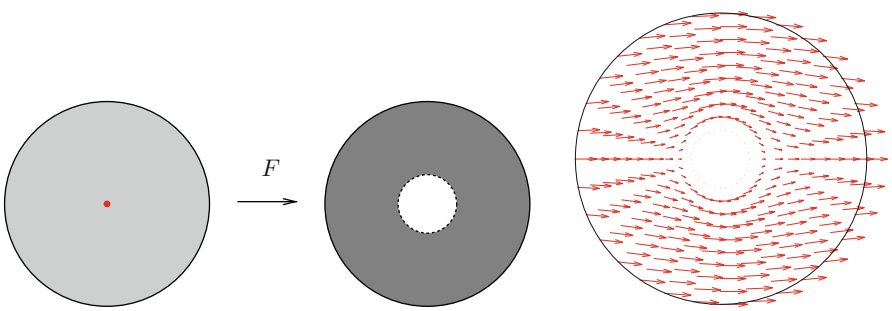
$$\tilde{\sigma}(r,\phi,\theta) = \begin{pmatrix} 2(r-1)^2\sin\theta & 0 & 0 \\ 0 & 2\sin\theta & 0 \\ 0 & 0 & 2(\sin\theta)^{-1} \end{pmatrix}, \quad 1 < r \le 2.$$

Note that $\tilde{\sigma}$ is degenerate on the sphere of radius 1 in the sense that it is not bounded from below by any positive multiple of the identity matrix I. Then, if u satisfies conductivity equation $\nabla\cdot\sigma_0\nabla u = 0$ in B_2 where $\sigma_0 = I$ is the constant isotropic conductivity, one sees that $\tilde{u}(x) = u(F^{-1}(x))$ satisfies in $B_2 \setminus \overline{B}_1$ the conductivity equation $\nabla\cdot\tilde{\sigma}\nabla\tilde{u} = 0$.

The currents associated to this singular conductivity on B_2 are shown in Fig. 1 (right). No currents originating at ∂B_2 have access to the region B_1, so that (heuristically) if the conductivity is changed in B_1, the measurements on the boundary ∂B_2 do not change. Moreover, all voltage-to-current measurements made on ∂B_2 give the same results as the measurements on the surface of a ball filled with homogeneous, isotropic material. The object is said to be *cloaked*, and the structure on $B_2 \setminus \overline{B}_1$ producing this effect is said to be a *cloaking device*. This was proven in dimensions $n \ge 3$ in [15, 16] by showing that $\Sigma_X(\sigma) = \Sigma_X(\tilde{\sigma})$ for $X = H^1(\Omega) \cap L^\infty(\Omega)$. In [11] the analogous identity is shown also for the Hilbert space X defined with the norm $(\tilde{\sigma}\nabla u, \nabla u)^{1/2}_{L^2(\Omega)}$. Similar results in the two-dimensional, or in cylindrical case, are shown in [20].

Cloaking via Transformation Optics for Electromagnetism

In the same 2006 issue of Science, there appeared two papers with transformation optics-based proposals for cloaking. Leonhardt [25] gave a description, based on conformal mapping, of inhomogeneous indices of



Invisibility Cloaking, Fig. 1 *Left*: Map $F : B_2 \setminus \{0\} \to B_2 \setminus \overline{B}_1$. *Right*: Analytic solutions for the currents with conductivity $\tilde{\sigma}$

refraction $n(x)$ in two dimensions that would cause light rays to go around a region and emerge on the other side as if they had passed through empty space (for which $n(x) \equiv 1$). On the other hand, Pendry, Schurig, and Smith [37] gave a prescription for values of permittivity ε and permeability μ yielding a cloaking device for electromagnetic waves, based on the fact that ε and μ transform in the same way as the conductivity σ under changes of variables, cf. (3). In fact, also this construction used the above singular transformation (5). In [25] and [37] the obtained mathematical models were also suggested to be realized physically, at least approximately, using artificially structured materials, *metamaterials*.

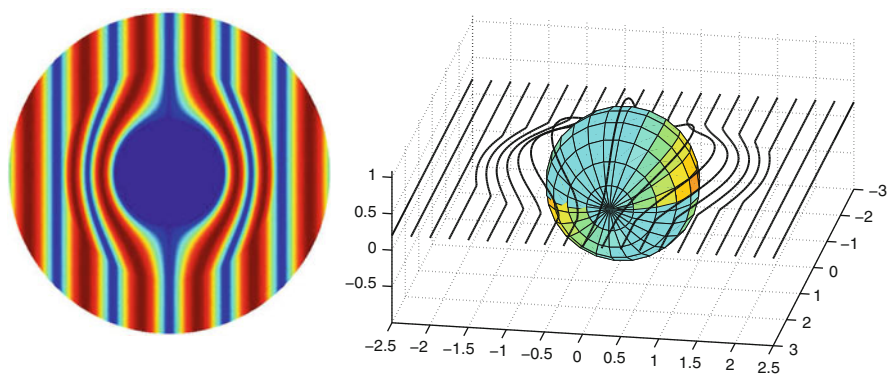
Next we consider the cloaking construction suggested in [37] and consider time-harmonic electric and magnetic fields $\mathbf{E}(x,t) = E(x)e^{i\omega t}$ and $\mathbf{H}(x,t) = H(x)e^{i\omega t}$ with frequency ω. When ε and μ are the permittivity and permeability in the domain $\Omega \subset \mathbb{R}^3$, then E, H satisfy time-harmonic Maxwell's equations,

$$\nabla \times H = -i\omega\varepsilon E \quad \nabla \times E = i\omega\mu H. \tag{7}$$

Let $\varepsilon_0 = I$ and $\mu_0 = I$ denote the constant permittivity and permeability (note that in the mathematical model we consider, all physical units are omitted). Then one defines the cloaking permittivity $\tilde{\varepsilon}$ and the cloaking permeability $\tilde{\mu}$ by setting $\tilde{\varepsilon} = \tilde{\mu} = \tilde{\sigma}$, where $\tilde{\sigma}$ is given in (6) with $\sigma_0 = I$.

Using ray optics to approximate electromagnetic waves, it was deduced in [38] that the light rays in the layer $B_2 \setminus \overline{B}_1$ with material parameters $\tilde{\varepsilon}$ and $\tilde{\mu}$ are images of straight lines in the map F (see Fig. 2 (right)). Thus the light rays go around a region and emerge on the other side as if they had passed through empty space, making the presence of any object being in B_1 undetected. Also, the behavior of the electromagnetic fields E and H as solutions of differential equations has been analyzed using the transformation law under change of coordinates. This was done in [37] in the layer $B_2 \setminus \overline{B}_1$ and in [11] in the whole ball B_2 taking in to account that $\tilde{\varepsilon}$ and $\tilde{\mu}$ are degenerate at the surface $|x| = 1$. Transformation optics-based cloaks can be obtained also for the Helmholtz equation by using the Riemannian metric $\tilde{g} = F_* g_0$, obtained by blowing up the Euclidian metric g_0 with the map (5) (see [8, 11]). Solutions corresponding to such cloaks are shown in Fig. 2, left.

From the practical point of view, one needs to consider what kind of materials are needed to realize an invisibility cloak, working at least with waves with a given frequency. Such materials with customized values of ϵ and μ, referred to as *metamaterials*, have been under extensive study in recent years. The label "metamaterial" usually attaches to macroscopic material structures having a man-made cellular architecture and producing combinations of material parameters not available in nature, due to resonances induced by the geometry of the cells [10]. Using metamaterial cells designed to resonate near the desired frequency, it is possible to obtain a wide range of permittivity and permeability tensors *at a given frequency*, so that they may have very large, very small, or even negative eigenvalues. The use of resonance phenomenon also explains why the material properties of such metamaterials strongly depend on the frequency. Also, Fig. 2 (right) shows why transformation optics-based cloaks only work perfectly for a single frequency: we see in the figure that a light ray traveling around the cloak travels a longer Euclidean distance than a straight line segment. Thus, if ϵ_0 and μ_0 were the vacuum electromagnetic parameters, and one could build a material



Invisibility Cloaking, Fig. 2 *Left*: The real part of the solution of a Helmholtz equation with a cloak at the plane $z = 0$. *Right*: Inside a cloaking device corresponding to $\tilde{\varepsilon}$ and $\tilde{\mu}$, the light rays go around the cloaked object

with permittivity and permeability $\tilde{\epsilon}$ and $\tilde{\mu}$ for all frequencies, then the velocity of the signal propagation would be faster than the speed of light in a vacuum. For a recent development on broadband cloaking in a surrounding medium with refractive index greater than 1, we refer to [27] and on properties of approximate cloaking constructions to [14, 21].

Cloaking via Anomalous Resonance

In contrast to transformation-based cloaking, cloaking due to anomalous resonance [5, 33, 35] is exterior cloaking: it has the intriguing feature that the cloaked region lies outside the cloaking device. First, to understand anomalous resonance [32, 34], consider the dielectric equation $\nabla \cdot \varepsilon \nabla V = 0$ in $\mathbb{R}^2 \setminus \{x_0\}$ in the presence of a dielectric annulus, having the scalar permittivity

$$\begin{aligned} \varepsilon(x) &= 1 \quad \text{for } |x| \geq r_s \\ &= \varepsilon_s, \quad \text{for} \quad r_c < |x| \leq r_s, \\ &= 1, \quad \text{for } |x| \leq r_c, \end{aligned} \tag{8}$$

where ε_s has the unusual value $\varepsilon_s = -1$. At $x_0 = (a, 0)$, we place a dipole of strength k oriented along the x_1-axis (corresponding to adding a source term which is proportional to the x_1 partial derivative of a delta function), and we look for solutions with $V \to 0$ as $x \to \infty$ (corresponding to the absence of a source at infinity). When $a > r_s^2/r_c$, the two-dimensional real potential $V(x_1, x_2) = \Re e\, U(z)$ with $z = x_1 + i x_2$ and $U(z)$ given by

$$\begin{aligned} U(z) &= \frac{k}{z-a}, \quad \text{for } |x| \geq r_s \\ &= -\frac{k}{a} - \frac{k r_s^2/a^2}{z - r_s^2/a} \quad \text{for } r_c < |x| \leq r_s, \\ &= \frac{k r_c^2/r_s^2}{z - a r_c^2/r_s^2} \quad \text{for } |x| \leq r_c \end{aligned} \tag{9}$$

solves the equations. Curiously, the solution in $|x| \geq r_s$ is exactly the same as for a homogeneous medium with $\varepsilon(x) = 1$ everywhere [34]. Thus the presence of the annulus does not influence the fields in $|x| \geq r_s$: the annulus is invisible to dipolar sources or more generally to any sources with support outside $|x| = r_s^2/r_c$ [32]. When $r_s^2/r_c > a > r_s$, the formula (9) does not provide a solution since it is singular at $z = r_s^2/a$, and at $z = a r_c^2/r_s^2$, nor should a solution necessarily exist as the partial differential equation is not elliptic.

However with $\varepsilon_s = -1 + i\eta$, with η real, the complex potential V_η satisfying $\nabla \cdot \varepsilon \nabla V_\eta = 0$ in $\mathbb{R}^2 \setminus \{x_0\}$, with $V_\eta \to 0$ as $x \to \infty$ and with the same dipole source so that $V_\eta \approx \Re e[k/(z-a)]$ near $z = a$, can be found by series expansions, for any $a > r_s$ and $\eta > 0$. Such potentials represent quasistatic solutions to Maxwell's equations, giving the fields in the vicinity of a hollow cylinder (represented by the annulus) with outer radius much smaller than the wavelength, and relative permittivities somewhat close to -1 can be realized using materials such as silver, gold, and silicon carbide at an appropriate frequency. The potential V_η exhibits strikingly unusual behavior as $\eta \to 0$. For $r_s^2/r_c > a$, define D as the union of the two annuli $r_s^2/a > |z| > a r_c^2/r_s^2$ and $r_s^3/(a r_c) > |z| > a r_c/r_s$, and for $a > r_s^2/r_c$, take D to be empty. The region D is the region of anomalous resonance: the L^2 norm of V_η inside any compact set within D diverges to infinity as $\eta \to 0$. The potential V_η develops large oscillations (called surface plasmons in physics) inside D with growing amplitude as $\eta \to 0$. This localized resonance is called anomalous because D depends on the position a of the source. Outside D, V_η converges pointwise as $\eta \to 0$ to the smooth potential $V = \Re e\, U$. For small η and $r_s^2/r_c > a$, it appears from outside D almost as if V_η has singularities at $z = r_s^2/a$, and at $z = a r_c^2/r_s^2$. (Anomalous resonance and the presence of such ghost singularities, discovered in [34], accounts for the superresolution of a superlens [36], which is a slab of material with $\varepsilon = \mu \approx -1$, surrounded by material with $\varepsilon = \mu = 1$.)

Given any source-free region, Ω an important physical quantity is

$$W_\eta(\Omega) = \int_\Omega \Im m(\varepsilon) |\nabla V_\eta|^2 = \Im m \int_{\partial\Omega} \varepsilon (\nabla V_\eta \cdot \nu) V_\eta^* \tag{10}$$

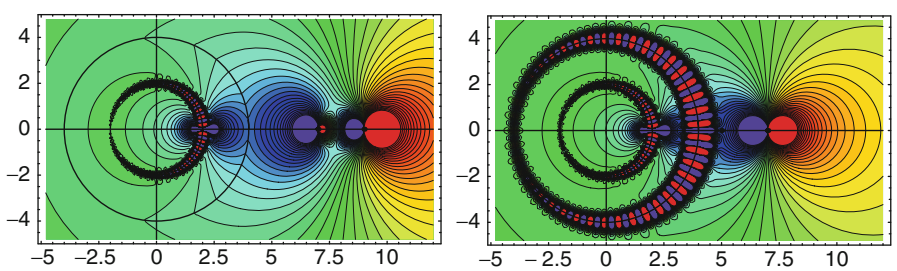
which in quasistatics is proportional to the electrical power dissipated in Ω (the * denotes complex conjugation, and ν denotes the outward unit normal to $\partial\Omega$). When Ω includes the shell region $r_c < |x| < r_s$ and k is fixed, then $W_\eta(\Omega) \to \infty$ as $\eta \to 0$ if the dipole source lies in D, i.e., $r_\# > a > r_s$, where $r_\# = \sqrt{r_s^3/r_c}$. As any realistic source can only produce

bounded power, it follows that k must go to zero as $\eta \to 0$. But then $V_\eta \to 0$ outside D. The dipole source will become essentially cloaked: the energy flowing from it is all channeled to the annulus and virtually does not escape outside the radius $r_\#$. For this reason the annulus $r_\# > |x| > r_s$ is called the cloaking region. In fact any finite collection of dipole sources located at fixed positions in the cloaking region which produce bounded power must all become cloaked as $\eta \to 0$ [33]. If these dipole sources are not active sources, but rather polarizable dipoles whose strength is proportional to the field acting on them, then these must also become cloaked [5, 35]. Somehow the field in D must adjust itself so that the field acting on each polarizable dipole in the collection is almost zero. It is still not exactly clear what is and is not cloaked by the annulus. Although some progress has recently been made: see, for example, [2]. Numerical evidence suggests that dielectric disks within the cloaking region are only partially cloaked [4]. Cloaking also extends to polarizable dipoles near two or three dimensional superlenses. There is also numerical evidence [24] to suggest that an object near a superlens can be cloaked at a fixed frequency if the appropriate "antiobject" is embedded in the superlens (Fig. 3).

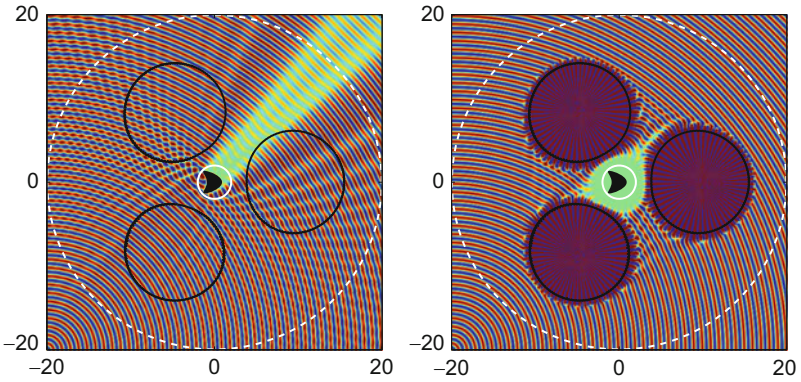
Active Exterior Cloaking

Active cloaking has the advantage of being broadband, but may require advance knowledge of the probing fields. Miller [28] found that active controls rather than passive materials could be used to achieve interior cloaking. Active exterior cloaking is easiest to see in the context of two-dimensional electrostatics, where it reduces to finding a polynomial which is approximately 0 within one disk in $\mathbb{C}$ and approximately 1 within a second disjoint disk [17]. To see this, let $B_r(\xi) \subset \mathbb{R}^2$ denote the disk of radius r centered at $x = (\xi, 0)$. Suppose we are given a potential $V(x)$ which, for simplicity, is harmonic in $\mathbb{R}^2$. The desired cloaking device, located at the origin, produces a potential $V_d(x)$ which is harmonic in $\mathbb{R}^2 \setminus \{0\}$ with $V_d(x)$ almost zero outside a sufficiently large ball $B_\gamma(0)$ so that the cloaking device is hard to detect outside the radius γ. At the same time we desire that the total potential $V(x) + V_d(x)$ (and its gradient) be almost zero in a ball $B_\alpha(\delta) \subset B_\gamma(0)$ not containing the origin, which is the cloaking region: a (non-resonant) object can be placed there with little disturbance to the surrounding fields because the field acting on it is very small. After applying the inverse transformation $z = 1/(x_1 + ix_2)$ and introducing harmonic conjugate potentials to obtain the analytic extensions v and v_d of V and V_d, the problem becomes: find $v_d(z)$ analytic in $\mathbb{C}$ such that $v_d \approx 0$ in $B_{1/\gamma}(0)$ and $v_d \approx -v$ in $B_{\alpha_*}(\delta_*)$, where $B_{\alpha_*}(\delta_*)$ is the image of $B_\alpha(\delta)$ under the inverse transformation. Since the product of two analytic functions is again analytic, this can be reformulated: find $w(z)$ analytic in $\mathbb{C}$ such that $w \approx 0$ in $B_{1/\gamma}(0)$ and $w \approx 1$ in $B_{\alpha_*}(\delta_*)$. To recover v_d one needs to multiply w by a polynomial which approximates $-v$ in $B_{\alpha_*}(\delta_*)$. When $1/\gamma$ and α_* are small enough, one can take $w(z)$ to be the Hermite interpolation polynomial of degree $2n-1 \gg 1$ satisfying

$$w(0) = 0, \quad w(\delta_*) = 1,$$
$$w^j(0) = w^j(\delta_*) = 0 \quad \text{for } j = 1, 2, \ldots, n-1 \quad (11)$$



Invisibility Cloaking, Fig. 3 *Left*: Equipotentials for the real part of the potential with one fixed dipole source on the right and a neighboring polarizable dipole on the left, outside the cylinder with $\varepsilon_s = -1 + 10^{-12}i$. *Right*: The equipotentials when the cylinder is moved to the right so it cloaks the polarizable dipole, leaving the exterior field close to that of the fixed dipole in free space (Taken from [33])



Invisibility Cloaking, Fig. 4 *Left*: Scattering of waves by a kite-shaped object, with the three active cloaking devices turned off. *Right*: The wave pattern with the devices turned on, showing almost no scattering

where $w^j(z)$ is the jth derivative of $w(z)$. As $n \to \infty$, one can show that $w(z)$ converges to 0 (and to 1) in the side of the figure eight $|z^2 - \delta_* z| < \delta_*^2/4$ containing the origin (and containing δ_*, respectively). Outside this figure eight, and excluding the boundary, $w(z)$ diverges to infinity. In practice the cloaking device cannot be a point, but should rather be an extended device encompassing the origin, and the device should produce the required potential V_d. Choosing the boundary of this device to be where V_d is not too large forces the device to partially wrap around the cloaking region, leaving a "throat" connecting the cloaking region to the outside. The width of the throat goes to zero as $n \to \infty$, but it appears to go to zero slowly. Thus one can get good cloaking with throat sizes that are not too small. This active exterior cloaking extends to the Helmholtz equation (see Fig. 4) and in that context works over a broad range of frequencies [18]. Numerical results show that an object can be effectively cloaked from an incoming pulse with a device having throats that are reasonably large.

References

1. Alu, A., Engheta, N.: Achieving transparency with plasmonic and metamaterial coatings. Phys. Rev. E **72**, 016623 (2005)
2. Ammari, H., Ciraolo, G., Kang, H., Lee, H., Milton, G.W.: Spectral theory of a Neumann–Poincaré-type operator and analysis of cloaking due to anomalous localized resonance. Arch. Ration. Mech. Anal. **208**(2), 667–692 (2013)
3. Benveniste, Y., Miloh, T.: Neutral inhomogeneities in conduction phenomenon. J. Mech. Phys. Solids **47**, 1873 (1999)
4. Bruno, O.P., Lintner, S.: Superlens-cloaking of small dielectric bodies in the quasistatic regime. J. Appl. Phys. **102**, 124502 (2007)
5. Bouchitté, G., Schweizer, B.: Cloaking of small objects by anomalous localized resonance. Q. J. Mech. Appl. Math. **63**, 437–463 (2010)
6. Cai, W., Chettiar, U., Kildishev, A., Milton, G., Shalaev, V.: Non-magnetic cloak with minimized scattering. Appl. Phys. Lett. **91**, 111105 (2007)
7. Calderón, A.P.: On an inverse boundary value problem. Seminar on Numerical Analysis and its Applications to Continuum Physics (Rio de Janeiro, 1980), pp. 65–73, Soc. Brasil. Mat., Río de Janeiro (1980)
8. Chen, H., Chan, C.T.: Acoustic cloaking in three dimensions using acoustic metamaterials. Appl. Phys. Lett. **91**, 183518 (2007)
9. Dolin, L.S.: To the possibility of comparison of three-dimensional electromagnetic systems with nonuniform anisotropic filling. Izv. Vyssh. Uchebn. Zaved. Radiofizika **4**(5), 964–967 (1961)
10. Eleftheriades, G., Balmain, K. (eds.): Negative-Refraction Metamaterials. IEEE/Wiley, Hoboken (2005)
11. Greenleaf, A., Kurylev, Y., Lassas, M., Uhlmann, G.: Full-wave invisibility of active devices at all frequencies. Commun. Math. Phys. **275**, 749–789 (2007)
12. Greenleaf, A., Kurylev, Y., Lassas, M., Uhlmann, G.: Electromagnetic wormholes and virtual magnetic monopoles from metamaterials. Phys. Rev. Lett. **99**, 183901 (2007)
13. Greenleaf, A., Kurylev, Y., Lassas, M., Uhlmann, G.: Electromagnetic wormholes via handlebody constructions. Commun. Math. Phys. **281**, 369–385 (2008)
14. Greenleaf, A., Kurylev, Y., Lassas, M., Uhlmann, G.: Approximate quantum and acoustic cloaking. J. Spectr. Theory **1**, 27–80 (2011). doi:10.4171/JST/2. arXiv:0812.1706v1
15. Greenleaf, A., Lassas, M., Uhlmann, G.: Anisotropic conductivities that cannot detected in EIT. Physiolog. Meas. (special issue on Impedance Tomography) **24**, 413–420 (2003)
16. Greenleaf, A., Lassas, M., Uhlmann, G.: On nonuniqueness for Calderón's inverse problem. Math. Res. Lett. **10**(5–6), 685–693 (2003)
17. Guevara Vasquez, F., Milton, G.W., Onofrei, D.: Active exterior cloaking. Phys. Rev. Lett **103**, 073901 (2009)
18. Guevara Vasquez, F., Milton, G.W., Onofrei, D.: Broadband exterior cloaking. Opt. Express **17**, 14800–14805 (2009)

19. Kerker, M.: Invisible bodies. J. Opt. Soc. Am. **65**, 376–379 (1975)
20. Kohn, R., Shen, H., Vogelius, M., Weinstein, M.: Cloaking via change of variables in electrical impedance tomography. Inver. Prob. **24**, 015016 (2008)
21. Kohn, R., Onofrei, D., Vogelius, M., Weinstein, M.: Cloaking via change of variables for the Helmholtz Equation. Commun. Pure Appl. Math. **63**, 1525–1531 (2010)
22. Kohn, R., Vogelius, M.: Identification of an unknown conductivity by means of measurements at the boundary. In: McLaughlin, D. (ed.) Inverse Problems. SIAM-AMS Proceedings vol. 14, pp. 113–123. American Mathematical Society, Providence (1984). ISBN 0-8218-1334-X
23. Lee, J., Uhlmann, G.: Determining anisotropic real-analytic conductivities by boundary measurements. Commun. Pure Appl. Math. **42**, 1097–1112 (1989)
24. Lai, Y., Chen, H., Zhang, Z.-Q., Chan, C.T.: Complementary media invisibility cloak that cloaks objects at a distance outside the cloaking shell. Phys. Rev. Lett. **102**, 093901 (2009)
25. Leonhardt, U.: Optical conformal mapping. Science **312**, 1777–1780 (2006)
26. Leonhardt, U., Philbin, T.: General relativity in electrical engineering. New J. Phys. **8**, 247 (2006)
27. Leonhardt, U., Tyc, T.: Broadband invisibility by non-euclidean cloaking. Science **323**, 110–112 (2009)
28. Miller, D.A.B.: On perfect cloaking. Opt. Express **14**, 12457–12466 (2006)
29. Milton, G.: The Theory of Composites. Cambridge University Press, Cambridge/New York (2002)
30. Milton, G.: New metamaterials with macroscopic behavior outside that of continuum elastodynamics. New J. Phys. **9**, 359 (2007)
31. Milton, G., Briane, M., Willis, J.: On cloaking for elasticity and physical equations with a transformation invariant form. New J. Phys. **8**, 248 (2006)
32. Milton, G.W., Nicorovici, N.-A.P., McPhedran, R.C., Podolskiy, V.A.: A proof of superlensing in the quasistatic regime, and limitations of superlenses in this regime due to anomalous localized resonance. Proc. R. Soc. A **461**, 3999–4034 (2005)
33. Milton, G., Nicorovici, N.-A.: On the cloaking effects associated with anomalous localized resonance. Proc. R. Soc. A **462**, 3027–3059 (2006)
34. Nicorovici, N.-A.P., McPhedran, R.C., Milton, G.W.: Optical and dielectric properties of partially resonant composites. Phys. Rev. B **49**, 8479–8482 (1994)
35. Nicorovici, N.-A.P., Milton, G.W., McPhedran, R.C., Botten, L.C.: Quasistatic cloaking of two-dimensional polarizable discrete systems by anomalous resonance. Opt. Express **15**, 6314–6323 (2007)
36. Pendry, J.B.: Negative refraction makes a perfect lens. Phys. Rev. Lett. **85**, 3966–3969 (2000)
37. Pendry, J.B., Schurig, D., Smith, D.R.: Controlling electromagnetic fields. Science **312**, 1780–1782 (2006)
38. Pendry, J.B., Schurig, D., Smith, D.R.: Calculation of material properties and ray tracing in transformation media. Opt. Express **14**, 9794 (2006)
39. Schurig, D., Mock, J., Justice, B., Cummer, S., Pendry, J., Starr, A., Smith, D.: Metamaterial electromagnetic cloak at microwave frequencies. Science **314**, 977–980 (2006)
40. Sylvester, J., Uhlmann, G.: A global uniqueness theorem for an inverse boundary value problem. Ann. Math. **125**, 153–169 (1987)
41. Ward, A., Pendry, J.: Refraction and geometry in Maxwell's equations. J. Modern Opt. **43**, 773–793 (1996)