Inference in Infinite Dimensional Inverse Problems: Discretization and Duality

By

Philip B. Stark
Department of Statistics
University of California
Berkeley, California 94720

Technical Report No. 261
July 1990

Department of Statistics
University of California
Berkeley, California 94720

Inference in Infinite Dimensional Inverse Problems: Discretization and Duality

Philip B. Stark
Department of Statistics
University of California
Berkeley, CA 94720

July 3, 1990

Abstract

Many techniques for solving inverse problems involve approximating the unknown model, a function, by a finite-dimensional "discretization" or parametric representation. The uncertainty in the computed solution is sometimes taken to be the uncertainty within the parametrization; this can result in unwarranted confidence. This paper presents a method for overcoming the limitations of discretization within the "strict bounds" formalism, a technique for constructing confidence intervals for functionals of the unknown model which can incorporate certain types of prior information. The usual computational approach to strict bounds approximates the "primal" problem in a way that the the resulting confidence intervals are at most long enough to have the nominal coverage probability. There is another approach based on "dual" optimization problems which gives confidence intervals with at least the nominal coverage probability. The pair of intervals derived by the two approaches bracket the correct confidence interval. Gravimetric, seismic, and geomagnetic illustrations are given, and a numerical example in seismology.

Keywords. Bounds on functionals, strict bounds, extremal bounds, confidence set inference, inverse theory, inference, discretization error, conjugate duality.

Acknowledgements. I am grateful to George E. Backus, Peter J. Bickel, Mikhail A. Brodsky, David L. Donoho, Douglas O. Gough, Niklaus Hengartner, Robert L. Parker, and Naum Z. Shor for helpful and encouraging discussions. Michael Stewart gave valuable computational assistance on the Berkeley Cray X-MP/14. This work was supported by National Science Foundation Fellowship DMS-8705843, NSF Presidential Young Investigator award DMS-8957573, NSF Grant DMS-8810192, and NASA Grant NAG 5-883.

0 Preamble

Figure 1 presents two confidence regions for seismic velocity in Earth's core, based on travel-time data reduced by Johnson and Lee [30] and the assumption that velocity increases monotonically in the core (see [64]). Both regions have a nominal 99.9% confidence level. The inner (dashed) envelope was computed by Stark and Parker [61] using a technique that approximates the space of velocity models by a finite-dimensional subspace (discretization of the model space). The envelope is therefore *optimistic*: its coverage probability may be less than the nominal level, since many elements of the space are not available to the algorithm. The wider (solid) envelope was computed by a technique described below in detail, especially in section 7.3.2 and appendix A. That technique, based on a dual problem, does not involve discretizing an infinite-dimensional space. The second envelope is *conservative*: its coverage probability is at least the nominal value, provided the assumptions about data errors and monotonicity constraints hold.

Stark [59] proves that as the number of functions used to approximate the space of velocity models increases (under mild conditions) the results based on discretization converge to the correct bounds. Stark and Parker [61] tested for convergence numerically by doubling the number of approximating functions: the computed bounds did not change visibly. Apparently, the bounds had converged. The set of solid bounds here limit rigorously the extent to which the approach based on discretization might be misleading. The two envelopes together bracket a true 99.9% confidence region for the inverse problem of inferring seismic velocity in the core from these data and prior information.

This paper gives a systematic treatment of this bracketing procedure for making inferences about functionals of the unknown model in infinite-dimensional constrained linear inverse problems.

1 Introduction

A fundamental difficulty in inverse problems is that generally an infinite number of models adequately account for our measurements (see, for example, Backus [2]). Strict bounds are a way to make rigorous inferences in the face of this nonuniqueness. Loosely, strict bounds determine shared properties of the set of models that satisfy our assumptions and fit the observations adequately. Here a "property" is the value of a functional, and for a property to be "shared" means that the range of values the functional takes on the set satisfy an inequality. More precisely, strict bounds are confidence intervals for functionals of the unknown model. Backus [7] calls this idea "Confidence Set Inference." Sabatier [53, 50] and Cuer [16] call the approach "well-posed questions." While Bayesian inversion, stochastic inversion, maximum entropy and other regularizing methods all solve the problem of *construction*, finding a plausible model that

fits the data, Backus provides persuasive reasons for preferring strict bounds or confidence set inference for many problems of *inference*, characterizing the nonuniqueness in inverse problems. This is especially true for problems in whole-Earth geophysics, where there is no observational basis for establishing a prior distribution on the (infinite-dimensional) model space.

We compute strict bounds by using our prior knowledge about the model and the distribution of the data errors to define a confidence region for the true model in model space—a set of "acceptable" models. We then pose optimization problems: find the smallest and largest values some scientifically interesting functional takes on that set. The solutions to these optimization problems give a confidence interval for the functional of the true model. Many different functionals may be considered, corresponding to different properties of the model. If the range of a functional on the set of acceptable models is small, then the data and our prior information allow us to determine that parameter well. If the confidence interval is wide, the available information is insufficient to constrain the property we have chosen. Strict bounds techniques are also useful for experimental design. By computing strict bounds with different prior constraints and hypothetical data, we can determine what measurements, signal-to-noise ratios and prior information would be sufficient to constrain properties we want to know.

Not every functional is bounded above or below on the set of acceptable models, but often even one-sided bounds are informative. On the basis of linear observations without special prior information, norms and seminorms can be bounded only from below. With suitable norm definitions, Shure *et al.* [56] use geomagnetic observations to find lower bounds on the energy stored in the magnetic field of the Earth's core, and on the Ohmic dissipation within the core. Parker and others [46, 43] use norm and seminorm minimization to construct confidence regions for the paleopole position from measurements of seamount magnetism.

The strict bounds approach has been used in many branches of geophysics, including:

- Gravimetry: Parker [42, 44]; Safon *et al.* [54]; Huestis and Ander [28]
- Core magnetism: Shure *et al.* [56, 57] Backus [7]
- Crustal magnetism: Parker and others [46, 43] Huestis and Parker [29]; Bayer and Cuer [9]
- Magnetotellurics: Oldenburg [38]
- Seismic Tomography: Grasso *et al.* [24]
- Travel-time seismology: Bessonova and others [10, 11] Garmany *et al.* [22]; Orcutt [39]; Stark and others [64, 63, 62] Wiggins *et al.* [65]

- Thermal History of the Crust: McNutt and Royden [37]

Strict bounds have also been used to estimate power spectral density (Lang and Marzetta [32, 33, 36]). Lang [31], Oldenburg [38], Rust, Burris and others [49, 15, 47], and Sabatier [51] discuss formulations of the strict bounds problem for bounds on linear functionals when the prior information about the unknown is a pointwise linear inequality. Backus [7] discusses confidence intervals for linear functionals when the prior information is a bound on a quadratic form of the model. His treatment covers systematic errors and the effect of approximating the model by a finite-dimensional discretization, by means of a clever reduction of the infinite-dimensional problem to the standard statistical problem of finding confidence intervals in finite-dimensional problems. Lang's paper [31] is a particularly clear exposition of strict bounds, with a careful treatment of errors introduced in approximating infinite-dimensional models, and a discussion of the dual problem and approximations to it. Donoho [18] treats the problem of estimating a linear functional of a model in l_2 from linear data when the model is constrained to lie in a convex subset of l_2 and the data errors are Gaussian. His results include as a special case optimally short confidence intervals for the problem addressed by Backus [7], in the case the data errors are Gaussian. See section 9.2.

An advantage of the Backus-Gilbert approach to inference [2, 3, 4] and of minimum Hilbert-space norm problems is that the only models entering the computations are linear combinations of a finite number of known functions—the problems are really finite-dimensional. This is also the case with the approach of Donoho [18], at least when the prior information is a ball in l_2. When there is prior information about the model in the form of linear inequalities, when one wishes to look at other norms, or when the space of acceptable models is particularly nasty, the Hilbert-space setting is often unnatural, leading one to formulate the problem in a more general Banach space (however, *c.f.* Donoho [18], who gives a theoretical treatment of arbitrary convex sets in a Hilbert space setting). The analogue of the projection theorem, which so readily allows the Hilbert-space minimum-norm problem to be reduced to a finite number of dimensions, can lead to nonlinear equations in these spaces. It can also be somewhat awkward to incorporate pointwise linear inequality constraints on the model in the Hilbert space framework: the point-evaluator can not be represented as an inner product except in reproducing kernel Hilbert spaces [1], which demand a tremendous amount of smoothness of the model. Hilbert space is ideal when the information about the unknown is an upper bound on a quadratic form or a linear equality, but not a more general "ball" nor a cone (again, compare with [18]). The problems we must solve to make inferences if we wish to use, say, pointwise linear inequalities, have appeared to be relentlessly infinite-dimensional. In order to compute with real data, we seem to be faced with the problem of representing an infinite-dimensional model on a computer using only a finite number of parameters. It is clear that no matter what scheme

we choose to represent the model space with only a finite number of parameters, we can always find some element of the space that is not approximated well. This may result in optimistic bounds on the functionals: the computed bounds may be narrower than the true bounds, giving a false impression of confidence. In some problems, it is possible to show that as the discretization improves, the bounds approach the correct value (*e.g.* Lang [31] ; Stark [59]). There is always a practical limit to the number of parameters we can store and compute. How good are our answers at that practical limit?

Lang [31] studies a finite-dimensional *dual problem* for inferences about linear functionals of a model in the space of continuous functions, with *a priori* pointwise upper and lower bounds on the unknown and a 2-norm measure of misfit to the data. He shows that a discretized and smoothed version of the dual can be used to bracket rigorously the strict bounds. (He also shows that the solution to the discretized primal with a larger value of the acceptable misfit can be used to bracket the true primal value. That approach does not seem helpful in more general settings—see section 8.) This paper owes much to his work, and to the excellent book by Luenberger [35].

It turns out that there are finite-dimensional dual problems for strict bounds in linear inverse problems, even in quite general linear vector spaces. In the dual problem, the objects we encounter are linear combinations of the data functionals, but the functionals are no longer necessarily elements of the model space, as they are in the Hilbert-space setting, since the model space is not in general its own dual. The dual problem has another significant advantage beyond the fact that it is finite-dimensional: its constraints, which may be infinite-dimensional, involve *known* functions, often much nicer than many elements of the set of models we must consider in the primal problem. This can make it much easier to determine the effect of discretization on the dual problem—we shall see below that the discretized dual for many strict bounds problems can be implemented in a way that gives bounds guaranteed to be pessimistic (too wide). Since the usual approach to discretizing the primal problem gives bounds that are too narrow (as the entire model space is not considered), the discretized primal and dual can be used in complementary ways to bracket the correct values of the optimization problems.

Several of the papers mentioned above use a dual approach, for example, Lang [31], Parker [42, 44], and Bessonova *et al.* [10, 11]. The aim of this paper is to develop dual problems for strict bounds in linear inverse problems with a variety of prior constraints in a more inclusive formulation, and to show how to implement the dual so that it can be used in conjunction with the primal to bracket rigorously the correct values of the strict bounds optimization problems. The paper is organized as follows: Section 2 sets up the strict bounds primal problem and touches very briefly on statistical and computational considerations for the primal approach. Section 3 introduces the dual in an intuitive fashion; section 4 gives an algebraic treatment. Section 5 develops the dual further for strict bounds on linear functionals; section 6 for norms and seminorms. Section

7 gives examples of bounding discretization error in some problems in gravimetry, seismology and geomagnetism, including a numerical example (Figure 1) in seismology, via dual optimization problems. Section 8 discusses primal approaches to bounding discretization error, including Lang's [31] and Backus' [7] techniques. Section 9 discusses confidence intervals for a single linear functional, including the methods of Backus [7] and Donoho [18]. Section 10 discusses the incorporation of systematic errors into the computations, with an illustration from helioseismology. Section 11 briefly discusses duality gaps and Lagrangian duality. It shows that for strict bounds problems, Lagrangian duality provides a more satisfactory set of sufficient conditions for the absence of a duality gap than Fenchel's duality Theorem, which is the usual tool for conjugate duality. Section 12 summarizes and indicates some unanswered questions and directions for further research. Appendix A develops the computations needed for the numerical example in seismology.

2 Strict Bounds Primal Problems

2.1 Notation

Lowercase Greek letters, $\alpha,\beta,\ldots$ are elements of the real line $\mathbf{R}$, as are the letters s and t, which will be used occasionally as independent variables. Boldface lowercase Greek letters, $\boldsymbol{\delta}$, $\boldsymbol{\lambda}$, ... are finite-dimensional vectors (elements of $\mathbf{R}^n$). Lowercase Greek letters may also denote functions into the reals, or into $\mathbf{R}^n$ (boldface). Subsets of $\mathbf{R}$ and $\mathbf{R}^n$ are denoted by uppercase and bold uppercase Greek letters (Δ, $\boldsymbol{\Xi}$). Matrices will also be denoted by bold uppercase Greek letters ($\boldsymbol{\Gamma}$); elements of a matrix are denoted by the same letter, doubly indexed (Γ_{ij}). The lowercase italic Roman letters i through q are reserved for integers (except in sections 7.2, 7.3.2, 10.2 and appendix A as noted). The unknown model is x_0, which is assumed to be an element of the real linear vector space X. Lowercase italic Roman letters from the end of the alphabet, such as x_1, x', y, are other elements of X. Subsets of X are denoted by sans-serif uppercase Roman letters (*e.g.* C). The dual space of X, the set of all linear functionals on X, is denoted X^*. Elements of X^* are denoted by starred lowercase Roman letters such as f^*, x^*. Functionals may also be recognized by the fact that their arguments are enclosed by square brackets [ ]. Linearity of x^* of course means that if $\{x_i\}_{i=1}^n$ are elements of X, and $\{\alpha_i\}_{i=1}^n$ are any n real numbers, then

$$x^*\left[\sum_{i=1}^n \alpha_i x_i\right] = \sum_{i=1}^n \alpha_i x^*[x_i]. \tag{1}$$

Subsets of X^* are denoted by starred sans-serif uppercase Roman letters (C^*). Boldface starred lowercase Roman letters ($\mathbf{f}^*$) denote ordered n-tuples of linear functionals on X:

$$\begin{aligned} \mathbf{f}^* \quad &: \quad \mathsf{X} \to \mathbf{R}^n \\ &\qquad x \mapsto \mathbf{f}^*[x] = (f^*{}_1[x], \cdots, f^*{}_n[x]). \end{aligned}$$

General, possibly nonlinear functionals on X (or subsets of X) are denoted by italic uppercase Roman letters ($H[x]$); and general, possibly nonlinear functionals on X^* or subsets of it are denoted by starred italic uppercase Roman letters ($C^*[x^*]$). Primal and dual optimization problems (defined below) are denoted $\mathcal{P}$ and $\mathcal{D}$. $\wp\{\cdot\}$ is the probability that the event in braces occurs. The notation $C[a,b]$ and $C^0[a,b]$ will be used to denote the usual spaces of continuous functions (C is normed with the "sup" norm; C^0 is not).

Our prior information may or may not endow X with a natural norm. When X is normed, it will be assumed that X is complete with respect to the norm topology, so X is a Banach space. In that case X^* will be taken to be the normed dual of X with the induced norm

$$||x^*|| \equiv \sup_{||x|| \le 1} x^*[x]. \tag{2}$$

It is well known that X^* with this norm is also a Banach space [35]. No notational distinction will be made between the norm on X and the norm on X^*.

We possess n real data δ_j:

$$\delta_j = f_j^*[x_0] + \epsilon_j, \quad j = 1, \cdots, n, \tag{3}$$

where the f_j^* are linear functionals on X. If X is normed, the f_j^* will be assumed to lie in the normed dual X^*. The ϵ_j are observational and modeling errors associated with the jth measurement. The data relations will usually be written in vector form:

$$\boldsymbol{\delta} = \mathbf{f}^*[x_0] + \boldsymbol{\epsilon}. \tag{4}$$

We will frequently take the "dot" product of $\mathbf{f}^*$with an n-vector $\boldsymbol{\lambda}$ of real numbers:

$$(\boldsymbol{\lambda} \cdot \mathbf{f}^*)[x] \equiv (\sum_{j=1}^{n} \lambda_j f_j^*)[x] \equiv \sum_{j=1}^{n} \lambda_j f_j^*[x] \ , \quad x \in \mathsf{X}. \tag{5}$$

2.2 The Primal Problem $\mathcal{P}$

Let $P_{\boldsymbol{\epsilon}}$ be the joint distribution of the data errors $\boldsymbol{\epsilon}$. All we need assume formally about the errors is that we can find a $1-\alpha$ confidence set $\boldsymbol{\Xi} \subset \mathbf{R}^n$ for the errors $\boldsymbol{\epsilon}$; *i.e.* any set such that

$$\wp\{\boldsymbol{\epsilon} \in \boldsymbol{\Xi}\} \geq 1 - \alpha, \tag{6}$$

or equivalently

$$\int_{\boldsymbol{\Xi}} dP_{\boldsymbol{\epsilon}} \geq 1 - \alpha.$$

The set $\boldsymbol{\Xi}$ can be used to define a confidence region for x_0: let

$$\begin{aligned} \mathsf{D} &\equiv \{x \in \mathsf{X} : \boldsymbol{\delta} - \mathbf{f}^*[x] \in \boldsymbol{\Xi}\} \\ &= \{x \in \mathsf{X} : \mathbf{f}^*[x] \in \boldsymbol{\delta} - \boldsymbol{\Xi}\}, \end{aligned} \tag{7}$$

where the set

$$\boldsymbol{\delta} - \boldsymbol{\Xi} \equiv \{\boldsymbol{\gamma} \in \mathbf{R}^n : \boldsymbol{\gamma} = \boldsymbol{\delta} - \boldsymbol{\xi} \text{ for some } \boldsymbol{\xi} \in \boldsymbol{\Xi}\}. \tag{8}$$

Let $\boldsymbol{\delta}_0$ denote the noise-free predictions of the true model, *i.e.* $\boldsymbol{\delta}_0 = \mathbf{f}^*[x_0]$. By 4 and the definition of $\boldsymbol{\Xi}$, the probability is at least $1 - \alpha$ that $\boldsymbol{\delta} - \boldsymbol{\delta}_0 \in \boldsymbol{\Xi}$. Thus the probability is at least $1 - \alpha$ that D, the preimage of $\boldsymbol{\delta} - \boldsymbol{\Xi}$, contains the true model x_0:

$$\wp\{\mathsf{D} \ni x_0\} \geq 1 - \alpha. \tag{9}$$

The set D is a $1 - \alpha$ confidence region in model space for x_0.

Let C be a subset of X (possibly all of X) in which we are sure *a priori* that x_0 lies. The set C represents our prior information; for example, we may believe that surely $||x_0|| \leq 1$ (see, for example, the geomagnetic problem in

Backus [7], where this constraint captures the physical requirement that the energy stored in Earth's main magnetic field is less than Earth's mass). Since there is no chance that $x_0 \notin \mathsf{C}$, the probability that $\mathsf{C} \cap \mathsf{D} \ni x_0$ is the same as the probability that $\mathsf{D} \ni x_0$: at least $1 - \alpha$:

$$\begin{aligned} \wp\{\mathsf{C} \cap \mathsf{D} \ni x_0\} &= \wp\{\mathsf{D} \ni x_0\} - \wp\{\mathsf{C}^c \cap \mathsf{D} \ni x_0\} \\ &= \wp\{\mathsf{D} \ni x_0\} \\ &\geq 1 - \alpha, \end{aligned} \tag{10}$$

where C^c is the complement of C in X.

The set $\mathsf{C} \cap \mathsf{D}$ is a $1-\alpha$ confidence set for x_0. It contains, in general, infinitely many models that are acceptable on the basis of their fit to the data, and their satisfaction of the prior constraints available. Any element in $\mathsf{C} \cap \mathsf{D}$ *might* be the truth. The problem of *construction* in inverse problems is to find an element $\hat{x} \in \mathsf{C} \cap \mathsf{D}$. The problem of *inference* in inverse problems is to study the entire set. The set $\mathsf{C} \cap \mathsf{D}$ contains many models—are they similar? In what respect? With the strict bounds technique, we focus on one functional (property of the set) at a time, and measure the "width" of the set with respect to that property by the range of values that property may have over all $\mathsf{C} \cap \mathsf{D}$.

For example, suppose we are interested in the functional

$$\begin{aligned} H \quad &: \quad \mathsf{X} \to \mathbf{R} \\ &\qquad x \mapsto H[x]. \end{aligned}$$

H might be the value of x at a point, a local average of x over some region, a norm of x, or any other functional.

From 10 we find

$$\wp\{H[\mathsf{C} \cap \mathsf{D}] \ni H[x_0]\} \geq 1 - \alpha, \tag{11}$$

where

$$H[\mathsf{C} \cap \mathsf{D}] \equiv \{\beta \in \mathbf{R} : \beta = H[x] \text{ for some } x \in \mathsf{C} \cap \mathsf{D}\}.$$

Define $\gamma^- \equiv \inf_{x \in \mathsf{C} \cap \mathsf{D}} H[x]$ and $\gamma^+ \equiv \inf_{x \in \mathsf{C} \cap \mathsf{D}} H[x]$. Then 11 implies (there may be slack in this inequality)

$$\wp\{\gamma^- \leq H[x_0] \leq \gamma^+\} \geq 1 - \alpha; \tag{12}$$

that is, the interval $[\gamma^-, \gamma^+]$ is a $1 - \alpha$ confidence interval for $H[x_0]$. Figure 2 sketches the relationships defined so far.

Depending on $\boldsymbol{\Xi}$, C, $\mathbf{f}^*$, and H, it can happen that the infimum γ^-, the supremum γ^+, or both fail to be finite. In these cases, only one-sided confidence intervals, or trivial confidence intervals result. See sections 2.3 and 9 for short discussions of the choice of $\boldsymbol{\Xi}$. Note that confidence intervals for any number of functionals may be found this way, with simultaneous $1-\alpha$ coverage probability. The bounds are in general conservative, *i.e.*

$$\wp\{v(\mathcal{P}) \leq H[x_0]\} > 1 - \alpha,$$

since there is some chance that x_0 could be anywhere in C, and $H[x]$ may be between α^- and α^+ in much of C that is not included in D.

We are now faced with a pair of infinite-dimensional optimization problems. Note that $-\inf_{\mathsf{C}\cap\mathsf{D}}\{-H\} = \sup_{\mathsf{C}\cap\mathsf{D}} H = \gamma^+$, so it suffices to consider the minimization of a functional H over $\mathsf{C}\cap\mathsf{D}$. The strict bounds primal problem $\mathcal{P}$ is to find the *value of* $\mathcal{P}$,

$$v(\mathcal{P}) \equiv \inf_{x\in\mathsf{C}\cap\mathsf{D}} H[x] \tag{13}$$

where H is a functional defined on $\mathsf{C}\cap\mathsf{D}$. A *feasible point* of $\mathcal{P}$ is a point $x \in \mathsf{C}\cap\mathsf{D}$.

I know of only two approaches to finding $v(\mathcal{P})$: discretization and duality, the subject of this paper. *Discretization* is the approximation of the primal problem by a problem in a finite-dimensional space. *Duality* leads to certain finite-dimensional problems in the dual space X^*, with the same value as the primal.

In this paper, the only functionals H considered in detail are linear functionals, norms and seminorms. The sorts of prior information treated here are

1. (1) none
2. (2) bounds on the the distance of the true model x_0 from a preferred model x', and
3. (3) the constraint that x lies in a cone.

I hope to treat other nonlinear functionals and different prior information in a future paper.

The current set-up covers quite a range of problems. For example, for geomagnetic problems it can incorporate a "hard quadratic bound" on the energy stored in the Earth's magnetic field or the Ohmic dissipation in the core (see Backus [6, 7]); it can incorporate pointwise upper and lower bounds (such as in Lang [31]), upper and lower bounds on the derivative (as in Stark *et al.* [64]), and positivity or monotonicity constraints (as in Stark and Parker [62]). All of the problems listed in the introduction (within the approximations customarily made for their solution) are instances of problems addressed here.

Deterministic Interpretation of Strict Bounds. As noted by Lang [31], strict bounds can be interpreted deterministically as well as stochastically. Let ϵ be a deterministic rather than stochastic error. We believe that $\epsilon \in \Xi$. The computed strict bounds on $H[x_0]$ are valid provided that supposition is correct. If $\epsilon \notin \Xi$, the computed bounds might not be valid (although, depending on Ξ, H, C, and $\mathbf{f}^*$, they might still be valid). This relates to the connection between statistical estimation and optimal recovery developed by Donoho [18].

2.3 Choice of the Data Confidence Region

See section 10 for a discussion of systematic errors. The strict bounds we calculate depend on the way we choose the set $\boldsymbol{\Xi}$. For a particular collection of functionals H we may wish to make inferences about, we can try to choose $\boldsymbol{\Xi}$ optimally to minimize the maximum of some weighted sum of the lengths of the intervals over all models in C. This problem appears to be unsolved, even for bounds on a single linear functional (see section 9). In this paper we will consider two types of data confidence regions $\boldsymbol{\Xi}$: balls in $\mathbf{R}^n$ in a weighted p-norm (discussed in this section); and "slabs," the region between two parallel hyperplanes (balls in a projection of the data space onto $\mathbf{R}^1$—these degenerate balls are discussed in section 9).

The restriction to balls is not as severe as it first appears, because by defining new data $\boldsymbol{\delta}' \equiv \boldsymbol{\Gamma} \cdot \boldsymbol{\delta}$, where $\boldsymbol{\Gamma}$ is an m by n matrix, we can get as confidence regions families of hypercylinders with various orientations, depending on $\boldsymbol{\Gamma}$, and various shapes, depending on the norm used. The slab discussed by Backus [7] and in section 9 is the case $m = 1$.

Define the p norms, $p \geq 1$, on $\mathbf{R}^n$ as follows:

$$||\epsilon||_p \equiv \left(\sum_{j=1}^{n} |\epsilon_j|^p \right)^{1/p}, \quad 1 \leq p < \infty, \tag{14}$$

and

$$||\epsilon||_\infty \equiv \max_j |\epsilon_j|. \tag{15}$$

Let $\boldsymbol{\Sigma}$ be a symmetric positive-definite n by n matrix. We define the $\boldsymbol{\Sigma}$-weighted p-norms to be

$$||\epsilon||_p^{\boldsymbol{\Sigma}} \equiv ||\boldsymbol{\Sigma} \cdot \epsilon||_p \tag{16}$$

where $\boldsymbol{\Sigma} \cdot \epsilon$ is the matrix-vector dot product

$$(\boldsymbol{\Sigma} \cdot \epsilon)_j \equiv \sum_{k=1}^{n} \Sigma_{jk} \epsilon_k. \tag{17}$$

The norms $||\epsilon||_p^{\boldsymbol{\Sigma}^{-1}}$ are defined analogously, with the matrix $\boldsymbol{\Sigma}^{-1}$ in place of $\boldsymbol{\Sigma}$.

We will need a weighted form of Hölder's inequality. Let $q = p/(p-1)$, $1 \leq p < \infty$, or $q = 1$ if $p = \infty$. Then for any two n-vectors $\boldsymbol{\beta}, \boldsymbol{\gamma}$, and symmetric, positive definite $\boldsymbol{\Sigma}$,

$$\begin{aligned} |\boldsymbol{\beta} \cdot \boldsymbol{\gamma}| &= |\boldsymbol{\beta} \cdot \boldsymbol{\Sigma} \cdot \boldsymbol{\Sigma}^{-1} \cdot \boldsymbol{\gamma}| \\ &\leq ||\boldsymbol{\Sigma} \cdot \boldsymbol{\beta}||_q ||\boldsymbol{\Sigma}^{-1} \cdot \boldsymbol{\gamma}||_p \\ &= ||\boldsymbol{\beta}||_q^{\boldsymbol{\Sigma}} ||\boldsymbol{\gamma}||_p^{\boldsymbol{\Sigma}^{-1}}. \end{aligned} \tag{18}$$

We assume enough is known about the joint distribution of the ϵ_j's so that for fixed p, Σ and $\alpha \in [0,1]$ we may find χ such that

$$\wp\left\{||\epsilon||_p^{\Sigma^{-1}} \leq \chi\right\} \geq 1-\alpha. \tag{19}$$

In general, finding χ for parametric distributions is a nontrivial calculation. If the joint distribution of the ϵ_j is Gaussian with nonsingular covariance matrix Σ^2, then $\left(||\epsilon||_2^{\Sigma^{-1}}\right)^2$ has the chi-squared distribution with n degrees of freedom. Percentage points for the 1-norm of independent Gaussians are tabulated in Parker and McNutt [45]. Percentage points for the infinity-norm of independent, identically distributed Gaussians are easy to calculate, but other than the two-norm, none of these is much fun even for Gaussian random variables if the ϵ_j are correlated. (Of course, if one is not attached to a particular coordinate system, in the Gaussian case one may always rotate the data to form uncorrelated linear combinations provided the covariance matrix is nonsingular.) For other distributions, I do not know what results are extant. When there are sufficiently many repeated observations, it may be possible to find χ by simulation from the empirical distribution of the data.

For most of the paper we take Ξ to be the Σ^{-1}-weighted p-norm ball

$$\Xi = \{\gamma \in \mathbf{R}^n : ||\gamma||_p^{\Sigma^{-1}} \leq \chi\}, \tag{20}$$

where χ is chosen so that $\wp\{\epsilon \in \Xi\} = 1-\alpha$. (Implicitly, we are assuming that the ϵ_j have mean zero; otherwise it would make sense to center the ball at the mean. Backus [6, 7] discusses removing the effect of known nonzero means from the data.)

Then D is the set of models that satisfy the data within χ in the Σ^{-1} weighted p-norm:

$$\begin{aligned} \mathsf{D} &\equiv \{x \in \mathsf{X} : \delta - \mathbf{f}^*[x] \in \Xi\} \\ &= \{x \in \mathsf{X} : ||\mathbf{f}^*[x] - \delta||_p^{\Sigma^{-1}} \leq \chi\}. \end{aligned} \tag{21}$$

There are both computational and statistical considerations for the choice of norms in defining Ξ. The 1,2 and infinity norms often lead to discretized primal problems that can be solved by linear or quadratic programming, rather than general nonlinear optimization—this is a clear advantage. The 2-norm is often considered the most natural when the errors are thought to be Gaussian, but it has several difficulties. First, for small n, the percentage points of the resulting distribution are very sensitive to departures from Gaussianity: if the true distribution is long-tailed (*e.g.* if there are outliers), the 2-norm ball with diameter chosen on the basis of the chi-squared distribution may have an actual coverage probability much smaller than the nominal one. Second, for each linear functional we wish to bound, solving the discretized primal optimization

problem for the two-norm measure of misfit involves a sequence of quadratic programs, each of which may consist of many least-squares problems. Both the one-norm and infinity-norm measures of misfit lead to a single linear programming problem for a bound on a linear functional, which can result in a substantial computational savings. The one-norm measure of misfit leads to a smaller tableau than the infinity-norm, and is more robust in the presence of outliers than the two-norm or infinity-norm. Since we are often just making an educated guess about the distribution of errors, it is desirable to decrease the sensitivity of our results to the accuracy of our guess as much as possible. Finally, there are some cases where the error is thought to be an absolute tolerance rather than a stochastic disturbance (for example instrumental limits or rounding errors in digitization—see also section 10). In such problems, the infinity-norm measure of misfit may be natural.

The geometry of the unit balls $\{\gamma \in \mathbf{R}^n : \|\gamma\|_p^{\mathbf{\Sigma}^{-1}} \leq 1\}$ and the radii of the smallest balls that contain $1-\alpha$ of the mass of the distribution of ϵ are different in these various norms, so the choice of norms will affect the bounds we obtain (although to a smaller extent than the choice of a different sort of region, such as a slab). The details depend on the data functionals, the measured data, the prior information, and the functionals of interest.

2.4 Discretizing the Primal Problem

The most common approach to solving the optimization problem $\mathcal{P}$ is to approximate X by some finite-dimensional subspace $\mathsf{S} \subset \mathsf{X}$, *i.e.* to pick a set of m linearly independent vectors $x_i \in \mathsf{X}$, and solve the discretized primal problem $\tilde{\mathcal{P}}$: Find

$$v(\tilde{\mathcal{P}}) = \inf_{\{\boldsymbol{\alpha} \in \mathbf{R}^m : \sum_{i=1}^m \alpha_i x_i \in \mathsf{C} \cap \mathsf{D}\}} H\left[\sum \alpha_i x_i\right]. \tag{22}$$

There are several goals in choosing S:

1. the possibility of computing $\mathbf{f}^*[x]$ and $H[x]$ analytically for $x \in \mathsf{S}$
2. simplicity of the description of $\mathsf{C} \cap \mathsf{S}$ in terms of the coefficients $\boldsymbol{\alpha}$
3. sufficient flexibility in S that $v(\tilde{\mathcal{P}})$ is close to $v(\mathcal{P})$.

Typical choices of S are piecewise constant functions, and the first m elements of a basis. For example, in whole-earth tomography, most researchers either divide Earth into a set of spherical "boxes," and assume that seismic velocity is constant in these boxes (the first sort of approximation); or they use a truncated spherical harmonic expansion in the angular variables, tensored with some sort of polynomial in radius (the second sort of approximation). In geomagnetism, it is customary to use the first m spherical harmonics (the second sort of discretization). Both sorts of finite-dimensional approximation will be called discretization.

For many choices of C, H and p, the problem $\tilde{\mathcal{P}}$ can be solved on a computer by standard optimization techniques. For example, if C is a cone or a set of models satisfying an infinity-norm bound, H is linear, and p is 1 or infinity, then the discretized primal can be solved by linear programming (*e.g.* [54, 29, 22, 64]). If $p = 2$ and the rest remain the same, the problem can be solved by a series of quadratic programs [49, 63]. If H is the one or infinity norm on a bounded domain, C is a cone or a ball in the infinity norm, and p is 1 or ∞, $\tilde{\mathcal{P}}$ can also be solved by linear programming. If H is a quadratic norm or seminorm, C is a cone, and p is 1 or ∞, $\tilde{\mathcal{P}}$ can be solved by quadratic programming (*e.g.* [62]). It is clear that

$$v(\tilde{\mathcal{P}}) \geq v(\mathcal{P}) \tag{23}$$

since the infimum in $\tilde{\mathcal{P}}$ is over $\mathsf{S} \cap \mathsf{C} \cap \mathsf{D}$, rather than $\mathsf{C} \cap \mathsf{D}$. How much smaller than the approximation could the true value be? We shall see that there is another *finite dimensional* optimization problem, the *dual problem* $\mathcal{D}$, with the property that

$$v(\tilde{\mathcal{P}}) \geq v(\mathcal{P}) \geq v(\mathcal{D}).$$

The pair of finite-dimensional optimization problems $\tilde{\mathcal{P}}$ and $\mathcal{D}$ rigorously bracket the number we wish to know, $v(\mathcal{P})$.

3 Naive Duality

The idea of the dual approach is to get information about $H[x_0]$ directly from the data and prior constraints. Even the fact that a datum has a finite value tells us *something* about the model—its component in the direction "aligned" (see [35] for a precise definition) with the data kernel is (almost surely) finite.

In this section we take X to be the space of continuous real-valued functions on the interval $[\alpha, \beta]$, we take H to be a linear functional, and we assume that $H[x]$ and $\mathbf{f}^*[x]$ can be written as integrals against "kernels:"

$$H[x] = \int_\alpha^\beta h(\tau)x(\tau)d\tau$$

$$\mathbf{f}^*[x] = \int_\alpha^\beta \mathbf{f}(\tau)x(\tau)d\tau,$$

where $\mathbf{f}(\tau)$ is an n-tuple of functions on $[\alpha, \beta]$.

3.1 No Prior Information

Suppose we can find a vector $\boldsymbol{\lambda} \in \mathbf{R}^n$ such that

$$h(\tau) = \boldsymbol{\lambda} \cdot \mathbf{f}(\tau) = \sum_{j=1}^n \lambda_j f_j(\tau)$$

for all $\tau \in [\alpha, \beta]$. Consider the minimization problem $\inf_{x \in \mathsf{D}} H[x]$. We have

$$\begin{aligned} H[x] &= \int \boldsymbol{\lambda} \cdot \mathbf{f}(\tau)x(\tau)d\tau \\ &= \boldsymbol{\lambda} \cdot (\boldsymbol{\delta} + \boldsymbol{\gamma}). \end{aligned}$$

(This observation is basic to Backus-Gilbert resolution theory [2, 3, 4, 5, 8].) Now for any $x \in \mathsf{D}$, $||\gamma||_p^{\boldsymbol{\Sigma}^{-1}} \leq \chi$. It follows from 18, the weighted form of Hölder's inequality, that $|\boldsymbol{\lambda} \cdot \boldsymbol{\gamma}| \leq \chi ||\boldsymbol{\lambda}||_q^{\boldsymbol{\Sigma}}$ for all $x \in \mathsf{D}$, so for all $\boldsymbol{\lambda} \in \mathbf{R}^n$ such that $h(\tau) = \boldsymbol{\lambda} \cdot \tau$,

$$H[x] \geq \boldsymbol{\lambda} \cdot \boldsymbol{\delta} - \chi ||\boldsymbol{\lambda}||_q^{\boldsymbol{\Sigma}}. \tag{24}$$

If the functions $f_j(\tau)$ are linearly independent, then there is at most one $\boldsymbol{\lambda}$ satisfying $\boldsymbol{\lambda} \cdot \mathbf{f}(\tau) = h(\tau)$, and we shall see below that for that $\boldsymbol{\lambda}$,

$$\inf_{x \in \mathsf{D}} H[x] = \boldsymbol{\lambda} \cdot \boldsymbol{\delta} - \chi ||\boldsymbol{\lambda}||_q^{\boldsymbol{\Sigma}}.$$

In any case, we have the "dual" problem $\mathcal{D}$:

$$v(\mathcal{D}) = \sup_{\boldsymbol{\lambda} \in \mathbf{R}^n : \boldsymbol{\lambda} \cdot \mathbf{f}(\tau) = h(\tau)} \boldsymbol{\lambda} \cdot \boldsymbol{\delta} - \chi ||\boldsymbol{\lambda}||_q^{\boldsymbol{\Sigma}}, \tag{25}$$

and $v(\mathcal{D}) \leq v(\mathcal{P}) = \inf_{x \in \mathsf{D}} H[x]$.

3.2 Positivity Constraints

Suppose now we are given the prior information that x_0 lies in the *positive cone* $\mathsf{C} = \{x(\tau) : x(\tau) \geq 0, \quad \tau \in [\alpha, \beta]\}$. This information allows us to relax the constraint that $\boldsymbol{\lambda} \cdot \mathbf{f}(\tau) = h(\tau)$ to the constraint $\boldsymbol{\lambda} \cdot \mathbf{f}(\tau) \leq h(\tau)$, because we can still bound the integral of $(h(\tau) - \boldsymbol{\lambda} \cdot \mathbf{f}(\tau))x(\tau)$:

$$\begin{aligned} H[x] &= \boldsymbol{\lambda} \cdot \mathbf{f}^*[x] + (H[x] - \boldsymbol{\lambda} \cdot \mathbf{f}^*[x]) \\ &= \boldsymbol{\lambda} \cdot (\boldsymbol{\delta} + \boldsymbol{\gamma}) + \int_{\{h(\tau) > \boldsymbol{\lambda} \cdot \mathbf{f}(\tau)\}} (h - \boldsymbol{\lambda} \cdot \mathbf{f}) x \\ &\geq \boldsymbol{\lambda} \cdot \boldsymbol{\delta} - \chi ||\boldsymbol{\lambda}||_q^{\boldsymbol{\Sigma}} + \int_{\{h(\tau) > \boldsymbol{\lambda} \cdot \mathbf{f}(\tau)\}} (h - \boldsymbol{\lambda} \cdot \mathbf{f}) 0 \\ &= \boldsymbol{\lambda} \cdot \boldsymbol{\delta} - \chi ||\boldsymbol{\lambda}||_q^{\boldsymbol{\Sigma}}, \end{aligned}$$

for $x \in \mathsf{C} \cap \mathsf{D}$. Since this relation holds for all $\boldsymbol{\lambda}$ such that $\boldsymbol{\lambda} \cdot \mathbf{f}(\tau) \leq h(\tau)$, this inequality leads to the dual problem

$$v(\mathcal{D}) = \sup_{\boldsymbol{\lambda} : \boldsymbol{\lambda} \cdot \mathbf{f}(\tau) \leq h(\tau)} \boldsymbol{\lambda} \cdot \boldsymbol{\delta} - \chi ||\boldsymbol{\lambda}||_q^{\boldsymbol{\Sigma}}, \tag{26}$$

and we have $v(\mathcal{D}) \leq v(\mathcal{P}) = \inf_{x \in \mathsf{C} \cap \mathsf{D}} H[x]$. This case and the one that follows are discussed in Lang [31]

3.3 Positivity Constraints and Upper Bounds

Suppose we have in addition an upper bound on $x_0(\tau)$, *i.e.*

$$x_0 \in \mathsf{C} \equiv \{x \in \mathsf{X} : 0 \leq x(\tau) \leq u(\tau)\}.$$

We may then completely relax the constraint that $\boldsymbol{\lambda} \cdot \mathbf{f}(\tau) = h(\tau)$, since

$$\begin{aligned} H[x] &= \boldsymbol{\lambda} \cdot \mathbf{f}^*[x] + (H[x] - \boldsymbol{\lambda} \cdot \mathbf{f}^*[x]) \\ &\geq \boldsymbol{\lambda} \cdot \boldsymbol{\delta} - \chi ||\boldsymbol{\lambda}||_q^{\boldsymbol{\Sigma}} + \int_{\{h(\tau) > \boldsymbol{\lambda} \cdot \mathbf{f}(\tau)\}} (h - \boldsymbol{\lambda} \cdot \mathbf{f}) \cdot 0 + \\ &\quad \int_{\{h(\tau) < \boldsymbol{\lambda} \cdot \mathbf{f}(\tau)\}} (h - \boldsymbol{\lambda} \cdot \mathbf{f}) u, \end{aligned}$$

for $x \in \mathsf{C} \cap \mathsf{D}$. We thus arrive at a third dual problem

$$v(\mathcal{D}) = \sup_{\boldsymbol{\lambda} \in \mathbf{R}^n} \boldsymbol{\lambda} \cdot \boldsymbol{\delta} - \chi ||\boldsymbol{\lambda}||_q^{\boldsymbol{\Sigma}} + \int_{\{h(\tau) < \boldsymbol{\lambda} \cdot \mathbf{f}(\tau)\}} (h - \boldsymbol{\lambda} \cdot \mathbf{f}) u. \tag{27}$$

Of course as before

$$v(\mathcal{D}) \leq v(\mathcal{P}) = \inf_{x \in \mathsf{C} \cap \mathsf{D}} H[x],$$

where C is now the set of positive, continuous functions on $[\alpha, \beta]$ that are everywhere less than or equal to $u(\tau)$.

The three dual problems 25,26,27 are representative of the dual problems we will encounter for bounds on linear functionals over abstract vector spaces with no prior information, prior cones, and prior balls. The dual problems all involve maximizing a concave functional of $\boldsymbol{\lambda} \in \mathbf{R}^n$ over convex regions of $\mathbf{R}^n$, and so are instances of general convex programming problems. Note that no particular properties of the space of continuous functions on $[\alpha, \beta]$ were used in deriving the first dual problem (with no prior information); to derive the second dual problem for a prior cone we needed only a lower bound on $(H - \boldsymbol{\lambda} \cdot \mathbf{f}^*)[x]$ for $x \in \mathsf{C}$; and for the prior infinity-norm ball $0 \leq x(\tau) \leq u(\tau)$, we used upper and lower bounds on $(H - \boldsymbol{\lambda} \cdot \mathbf{f}^*)[x]$ for $x \in \mathsf{C}$. Such bounds may be available even in much more abstract settings, allowing us, for example, to avoid the assumption that we are in a space of ***continuous*** functions, unless our prior knowledge supports it. We turn now to a more abstract framework.

4 Algebraic Duality

Let us proceed in a little more generality to derive a lower bound on the value of the primal problem $\mathcal{P}$. Let X be a real linear vector space, $\mathbf{f}^*$ an n-tuple of linear functionals on X, H a functional on X, and let $\mathsf{C} \subset \mathsf{X}$. Let x^* be any element of X^*.

$$\begin{aligned} v(\mathcal{P}) &\equiv \inf_{x \in \mathsf{C} \cap \mathsf{D}} H[x] \\ &= \inf_{x \in \mathsf{C} \cap \mathsf{D}} \{x^*[x] + (H - x^*)[x]\} \\ &\geq \inf_{x \in \mathsf{C} \cap \mathsf{D}} x^*[x] + \inf_{x \in \mathsf{C} \cap \mathsf{D}} \{(H - x^*)[x]\} \\ &\geq \inf_{x \in \mathsf{D}} x^*[x] + \inf_{x \in \mathsf{C}} \{(H - x^*)[x]\}. \end{aligned}$$

But x^* was arbitrary, so

$$\begin{aligned} \inf_{x \in \mathsf{C} \cap \mathsf{D}} H[x] &\geq \sup_{x^* \in \mathsf{X}^*} \left\{ \inf_{x \in \mathsf{D}} x^*[x] + \inf_{x \in \mathsf{C}} \{(H - x^*)[x]\} \right\} \\ &= \sup_{x^* \in \mathsf{X}^*} \{D^*[x^*] + C^*[x^*]\}, \end{aligned} \tag{28}$$

where the functionals C^* and D^* on X^* are defined by

$$C^*[x^*] \equiv \inf_{x \in \mathsf{C}} (H - x^*)[x]. \tag{29}$$

$$D^*[x^*] \equiv \inf_{x \in \mathsf{D}} x^*[x]. \tag{30}$$

Remarkably, equality is possible in 28, even though we seem to have thrown away a great deal of information by breaking the optimization problem into pieces and removing constraints from each of the pieces. When equality does not hold, there is said to be a "duality gap." Fenchel's Duality Theorem and Lagrangian Duality [35] give sufficient conditions for equality in 28; these are discussed in section 11 below.

The only functionals x^* it behooves us to consider are those for which $C^*[x^*]$ and $D^*[x^*]$ are greater than minus infinity. Define the *conjugate sets*

$$\mathsf{C}^* \equiv \{x^* \in \mathsf{X}^* : C^*[x^*] > -\infty\} \tag{31}$$

and

$$\mathsf{D}^* \equiv \{x^* \in \mathsf{X}^* : D^*[x^*] > -\infty\}. \tag{32}$$

The functional x^* provides a nontrivial bound via 28 if and only if $x^* \in \mathsf{C}^* \cap \mathsf{D}^*$. The set D^* is easily characterized. By an argument implicit in Backus [2], the

only linear functionals bounded below on D are linear combinations of the data functionals f_j^*; *i.e.* functionals of the form

$$x^* = \boldsymbol{\lambda} \cdot \mathbf{f}^*, \quad \boldsymbol{\lambda} \in \mathbf{R}^n.$$

We lose nothing by restricting the supremum in 28 to this (at most) n-dimensional subspace of X^*.

We may now compute $\mathsf{D}^*[x^*]$. For $x \in \mathsf{D}$, $\mathbf{f}^*[x] = \boldsymbol{\delta} + \boldsymbol{\gamma}$, where $||\boldsymbol{\gamma}||_p^{\boldsymbol{\Sigma}^{-1}} \leq \chi$. Thus

$$\begin{aligned} \boldsymbol{\delta} \cdot \mathbf{f}^*[x] &= \boldsymbol{\lambda} \cdot \boldsymbol{\delta} + \boldsymbol{\lambda} \cdot \boldsymbol{\gamma} \\ &\geq \boldsymbol{\lambda} \cdot \boldsymbol{\delta} - |\boldsymbol{\lambda} \cdot \boldsymbol{\gamma}| \\ &\geq \boldsymbol{\lambda} \cdot \boldsymbol{\delta} - ||\boldsymbol{\lambda}||_q^{\boldsymbol{\Sigma}} ||\boldsymbol{\gamma}||_p^{\boldsymbol{\Sigma}^{-1}}, \end{aligned}$$

by Hölder's inequality. Since $||\boldsymbol{\gamma}||_p^{\boldsymbol{\Sigma}^{-1}} \leq \chi$,

$$\boldsymbol{\delta} \cdot \mathbf{f}^*[x] \geq \boldsymbol{\lambda} \cdot \boldsymbol{\delta} - \chi ||\boldsymbol{\lambda}||_q^{\boldsymbol{\Sigma}}. \tag{33}$$

If the f_j^* are linearly independent, equality is possible: By an argument not reproduced here (see Backus [6]), if the f_j^* are linearly independent, there exist $\{x_i\}_{i=1}^n$ in X such that

$$f_j^*[x_i] = \begin{cases} 1, & i = j \\ 0, & \text{otherwise.} \end{cases}$$

Define the scalars α_j as follows:

$$\alpha_j = \delta_j - \chi \left(\frac{|\lambda_j|}{||\boldsymbol{\lambda}||_q^{\boldsymbol{\Sigma}}} \right)^{q/p} \frac{\lambda_j}{|\lambda_j|},$$

and let $x' \equiv \sum_j \alpha_j x_j$. It is easily verified that

$$||\mathbf{f}^*[x'] - \boldsymbol{\delta}||_p^{\boldsymbol{\Sigma}^{-1}} = \chi,$$

so $x' \in \mathsf{D}$, and

$$H[x'] = \boldsymbol{\lambda} \cdot \mathbf{f}^*[x'] = \boldsymbol{\lambda} \cdot \boldsymbol{\delta} - \chi ||\boldsymbol{\lambda}||_q^{\boldsymbol{\Sigma}}. \tag{34}$$

Thus the bound in equation 33 is sharp at least when the f_j^* are linearly independent. Substituting 33 into 28, we arrive at the fundamental inequality of the paper:

$$\inf_{x \in \mathsf{C} \cap \mathsf{D}} H[x] \geq v(\mathcal{D}) \equiv \sup_{\boldsymbol{\lambda} \in \mathbf{R}^n} \left\{ \boldsymbol{\lambda} \cdot \boldsymbol{\delta} - \chi ||\boldsymbol{\lambda}||_q^{\boldsymbol{\Sigma}} + C^*[\boldsymbol{\lambda} \cdot \mathbf{f}^*] \right\}. \tag{35}$$

Note that the infimum in the primal problem has been replaced by a supremum (over $\mathbf{R}^n$) in the dual. However, there is still a hidden infinite-dimensional infimum on the right hand side, in C^*.

It is easily verified that C^* is a concave functional: suppose $\alpha, \beta \geq 0$ and $\alpha + \beta = 1$. We have

$$\begin{aligned} C^*[\alpha x^* + \beta y^*] &= \inf_{x \in \mathsf{C}} \{(H - \alpha x^* - \beta y^*)[x]\} \\ &= \inf_{x \in \mathsf{C}} \{\alpha(H - x^*)[x] + \beta(H - y^*)[x]\} \\ &\geq \alpha \inf_{x \in \mathsf{C}} \{(H - x^*)[x]\} + \beta \inf_{x \in \mathsf{C}} \{(H - y^*)[x]\} \\ &= \alpha C^*[x^*] + \beta C^*[y^*] \quad \square \end{aligned} \tag{36}$$

One may show similarly that D^* is concave. Since $\boldsymbol{\lambda} \cdot \mathbf{f}^*$ is linear in $\boldsymbol{\lambda}$, it follows that $C^*[\boldsymbol{\lambda} \cdot \mathbf{f}^*]$ is concave in $\boldsymbol{\lambda}$. As $\boldsymbol{\lambda} \cdot \boldsymbol{\delta}$ and $-\|\boldsymbol{\lambda}\|_q^{\boldsymbol{\Sigma}}$ are also concave in $\boldsymbol{\lambda}$, the dual functional (the right hand side of 35) is concave in $\boldsymbol{\lambda}$. The concavity of C^* and D^* yield immediately the convexity of C^*, D^* and $\mathsf{C}^* \cap \mathsf{D}^*$. Thus the dual problem is always the maximization of a concave functional over a convex set.

The following sections explore some functionals H and sets C where C^* and $C^*[\boldsymbol{\lambda} \cdot \mathbf{f}^*]$ can be characterized usefully, as D^* and D^* were. In the next section we take H to be a linear functional, and look at some sets C where C^* and C^* are simple.

5 Bounds on Linear Functionals

Within this section $H[x]$ is always a linear functional. We will consider three kinds of prior information: $\mathsf{C} = \mathsf{X}$, C is a cone, and C is a ball.

5.1 No Prior Information

Here $\mathsf{C} = \mathsf{X}$. Since H is linear, by the standard argument,

$$C^*[x^*] \equiv \inf_{x \in \mathsf{X}} (H - x^*)[x] = -\infty,$$

unless $H - x^* = 0$. Thus $\mathsf{C}^* = H$ and for all $x^* \in \mathsf{C}^*$, $C^*[x^*] = 0$. If there is no $\boldsymbol{\lambda}$ such that $\boldsymbol{\lambda} \cdot \mathbf{f}^* = H$, the value of the dual problem is defined to be $-\infty$. Otherwise, it is

$$v(\mathcal{D}) = \sup_{\{\boldsymbol{\lambda} \in \mathbf{R}^n : \boldsymbol{\lambda} \cdot \mathbf{f}^* = H\}} \{\boldsymbol{\lambda} \cdot \boldsymbol{\delta} - \chi ||\boldsymbol{\lambda}||_q^{\boldsymbol{\Sigma}}\}.$$

This is the maximization of a convex functional subject to an infinite set of linear equality constraints. For $q = 1$ or ∞, the dual problem can be written as a semi-infinite equality constrained least-squares problem. If $q = 2$, it is a semi-infinite least-squares problem. If the f_j^* are linearly independent, there is at most one $\boldsymbol{\lambda}$ satisfying the constraints, so there is no maximization to perform, and by the argument leading to equation 34, there is no duality gap. This is analogous to section 3.1.

5.2 Prior Cones

A set $\mathsf{P} \subset \mathsf{X}$ is a *cone with vertex at the origin* if $x \in \mathsf{P}$ implies that $\gamma x \in \mathsf{P}$ for all $\gamma \geq 0$. A set C is a *cone with vertex* x' if

$$\mathsf{C} = x' + \mathsf{P} \equiv \{x' + y : y \in \mathsf{P}\},$$

for some point $x' \in \mathsf{X}$ and some cone P with vertex at the origin. C is *convex cone* if C is a cone and C is convex. We now consider prior information of the form $x \in \mathsf{C}$, where C is a convex cone. Examples of information of this kind are pointwise upper or lower bounds on the model, monotonicity constraints, and so on. Such constraints arise in seismology, gravimetry, medical tomography, and magnetotellurics, diffraction-limited optics and radio astronomy; as seismic velocity, mass density, electrical conductivity, and the number of photons coming from an object are all nonnegative.

It is easily verified that C^* is a convex cone in X^*. The convexity of C^* follows from the more general inequality 36. That C^* is a cone can be shown using the following characterization of C^*: recall that

$$\mathsf{C}^* \equiv \{x^* \in \mathsf{X}^* : \inf_{x \in \mathsf{C}} (H - x^*)[x] > -\infty\}$$

We show that

$$\mathsf{C}^* = \tilde{\mathsf{C}}^* \equiv \{x^* \in \mathsf{X}^* : \inf_{x \in \mathsf{C}} (H - x^*)[x] = (H - x^*)[x']\}. \tag{37}$$

It is obvious that $\tilde{\mathsf{C}}^* \subset \mathsf{C}^*$; we need to show that $\mathsf{C}^* \subset \tilde{\mathsf{C}}^*$. Suppose $x^* \notin \tilde{\mathsf{C}}^*$, so for some $y \in \mathsf{C}$, $(H - x^*)[y] < (H - x^*)[x']$. Since C is a cone, $y_\alpha \equiv x' + \alpha(y - x') \in \mathsf{C}$ for all $\alpha > 0$. Define $x^*_H \equiv H - x^*$. By the linearity of x^*_H,

$$x^*_H[y_\alpha] = x^*_H[x'] + \alpha(x^*_H[y] - x^*_H[x']).$$

As $\alpha \to \infty$,

$$x^*_H[y_\alpha] = (H - x^*)[y_\alpha] \to -\infty,$$

so $x^* \notin \mathsf{C}^*$. Now if $x^* \in C^*$, $x^*[x] \geq x^*[x']$, $\forall x \in \mathsf{C}^*$. Then $\alpha x^*[x] \geq \alpha x^*[x']$ $\forall x \in \mathsf{C}^*$ and all $\alpha \geq 0$. Therefore by the argument supporting 37, $\alpha x^* \in \mathsf{C}^*$, and thus C^* is a cone with vertex at the origin. □

Now if C is a cone with vertex x', we have

$$v(\mathcal{D}) \equiv \sup_{\boldsymbol{\lambda} \in \mathbf{R}^n} \left\{ \boldsymbol{\lambda} \cdot \boldsymbol{\delta} - \chi ||\boldsymbol{\lambda}||_q^{\boldsymbol{\Sigma}} + \inf_{x \in \mathsf{C}} (H - \boldsymbol{\lambda} \cdot \mathbf{f}^*)[x] \right\}$$

$$= \sup_{\{\boldsymbol{\lambda}: \boldsymbol{\lambda} \cdot \mathbf{f}^* \in \mathsf{C}^*\}} \left\{ \boldsymbol{\lambda} \cdot \boldsymbol{\delta} - \chi ||\boldsymbol{\lambda}||_q^{\boldsymbol{\Sigma}} + \inf_{x \in \mathsf{C}} (H - \boldsymbol{\lambda} \cdot \mathbf{f}^*)[x] \right\}$$

$$= \sup_{\{\boldsymbol{\lambda}: \boldsymbol{\lambda} \cdot \mathbf{f}^* \in \mathsf{C}^*\}} \left\{ \boldsymbol{\lambda} \cdot \boldsymbol{\delta} - \chi ||\boldsymbol{\lambda}||_q^{\boldsymbol{\Sigma}} + (H - \boldsymbol{\lambda} \cdot \mathbf{f}^*)[x'] \right\}, \tag{38}$$

by 37. Equation 38 becomes useful if there is an easy way to characterize the set of $\boldsymbol{\lambda} \in \mathbf{R}^n$ such that $\boldsymbol{\lambda} \cdot \mathbf{f}^* \in \mathsf{C}^*$.

Example. Let $\mathsf{X} = C^0[0, 1]$, the space of continuous functions on $[0, 1]$ and let C be the positive cone:

$$\mathsf{C} \equiv \{x \in \mathsf{X} : x(\tau) \geq 0, \quad \tau \in [0, 1]\}.$$

Then X^* is the space of measures on $[0, 1]$ with

$$x^*[x] = \int_0^1 x(\tau) dx^*(\tau),$$

and C^* is the set of positive measures on $[0, 1]$. If H and f_j^* are measures that are continuous with respect to Lebesgue measure, *i.e.* if

$$H[x] = \int_0^1 h(\tau) x(\tau) d\tau$$

and

$$f_j^*[x] = \int_0^1 f_j(\tau)x(\tau)d\tau$$

for some real-valued functions $h(\tau)$ and $f_j(\tau)$, then $\boldsymbol{\lambda} \cdot \mathbf{f}^* \in \mathsf{C}^*$ if and only if

$$h(\tau) - \sum_j \lambda_j f_j(\tau) \geq 0, \quad \tau \in [0,1].$$

This infinite set of constraints on $\boldsymbol{\lambda}$ can be discretized by imposing them at a finite set of points in $[0,1]$. If something is known about the continuity of $h(\tau)$ and $f_j(\tau)$, it may be possible (as we shall see below in section 7) to impose constraints in the discretized dual problem so that the set of admissible $\boldsymbol{\lambda}$ is a subset of the set in the original infinitely constrained problem.

5.3 Prior Balls

Here we take C to be the set $\{x \in \mathsf{X} : ||x - x'|| \leq v\}$. A quadratic norm might be induced by a "hard quadratic bound," such as a constraint on the amount of energy stored in Earth's magnetic field (Backus [6, 7]). An infinity norm bound could be induced by pointwise upper and lower bounds on the unknown, such as arise in gravimetry, travel-time tomography, magnetotellurics, spectroscopy, *etc.* In this section we assume that X is complete with respect to the norm (so it is a Banach space) and take X^* to be its normed dual. We assume that H and f_j^* are in X^*; since X^* is a linear space, $\boldsymbol{\lambda} \cdot \mathbf{f}^* \in \mathsf{X}^*$ for all $\boldsymbol{\lambda} \in \mathbf{R}^n$. From the definition of the norm on X^* and the linearity of x^*, it is immediate that

$$|x^*[x]| \leq ||x^*||\,||x||. \tag{39}$$

Consider the conjugate set C^* for this problem:

$$\mathsf{C}^* \equiv \{x^* \in \mathsf{X}^* : \inf_{x \in \mathsf{C}} (H - \boldsymbol{\lambda} \cdot \mathbf{f}^*)[x] > -\infty\}.$$

Now

$$\begin{aligned}
\inf_{x \in \mathsf{C}} (H - \boldsymbol{\lambda} \cdot \mathbf{f}^*)[x] &= \inf_{||x - x'|| \leq v} (H - \boldsymbol{\lambda} \cdot \mathbf{f}^*)[x] \\
&= \inf_{||y|| \leq v} \{(H - \boldsymbol{\lambda} \cdot \mathbf{f}^*)[y] + (H - \boldsymbol{\lambda} \cdot \mathbf{f}^*)[x']\} \\
&\geq -v||(H - \boldsymbol{\lambda} \cdot \mathbf{f}^*)|| + (H - \boldsymbol{\lambda} \cdot \mathbf{f}^*)[x'] > -\infty
\end{aligned} \tag{40}$$

by equation 39. Thus $\mathsf{C}^* = \mathsf{X}^*$, the entire normed dual space. This leads to the dual problem: find

$$\begin{aligned}
v(\mathcal{D}) &\equiv \sup_{\boldsymbol{\lambda} \in \mathbf{R}^n} \left\{ \boldsymbol{\lambda} \cdot \boldsymbol{\delta} - \chi||\boldsymbol{\lambda}||_q^{\boldsymbol{\Sigma}} - v||H - \boldsymbol{\lambda} \cdot \mathbf{f}^*|| + (H - \boldsymbol{\lambda} \cdot \mathbf{f}^*)[x'] \right\} \\
&= H[x'] + \sup_{\boldsymbol{\lambda} \in \mathbf{R}^n} \left\{ \boldsymbol{\lambda} \cdot (\boldsymbol{\delta} - \mathbf{f}^*[x']) - \chi||\boldsymbol{\lambda}||_q^{\boldsymbol{\Sigma}} - \nu||H - \boldsymbol{\lambda} \cdot \mathbf{f}^*|| \right\},
\end{aligned} \tag{41}$$

which is the unconstrained maximization of a concave functional of $\boldsymbol{\lambda}$.

Example 1. Let $\mathsf{X} \equiv l_2^{\omega}$ be the Hilbert space of sequences $x = (x_j)_{j=1}^{\infty}$ that are square-summable with the positive weight sequence ω:

$$\mathsf{X} = \left\{ x = (x_j)_{j=1}^{\infty} : \sum_{1}^{\infty} \omega_j x_j^2 < \infty \right\}$$

Let f_j^* and H denote both the functionals and the elements of l_2^{ω} that "represent" them. Suppose we can compute analytically the infinite sums $f_j^*[f_k^*]$, $f_j^*[H]$, $f_j^*[x']$, $H[H]$ and $H[x']$. Let $\boldsymbol{\Gamma}$ be the Gram matrix with elements $\Gamma_{jk} = f_j^*[f_k^*]$, $j, k \in \{1, \cdots, n\}$. Then equation 41 reduces to

$$\begin{aligned} v(\mathcal{D}) \quad = \quad & H[x'] + \sup_{\boldsymbol{\lambda} \in \mathbf{R}^n} \left\{ \boldsymbol{\lambda} \cdot (\boldsymbol{\delta} - \mathbf{f}^*[x']) - \chi ||\boldsymbol{\lambda}||_q^{\boldsymbol{\Sigma}} - \right. \\ & \left. v \left(\boldsymbol{\lambda} \cdot \boldsymbol{\Gamma} \cdot \boldsymbol{\lambda} - 2\boldsymbol{\lambda} \cdot (\mathbf{f}^*[H]) + H[H] \right)^{1/2} \right\}, \end{aligned}$$

and there is no need to discretize the dual problem. This example relates to the geomagnetic problem addressed by Backus [7], where the model x is a sequence of coefficients in the spherical harmonic expansion of the scalar magnetic potential of Earth's main magnetic field.

Example 2. Suppose our prior information is that the unknown is a continuous function $x_0(\tau)$ and that

$$0 \leq x_0(\tau) \leq u(\tau). \tag{42}$$

It is then natural to take X to be the Banach space of continuous functions with finite weighted sup-norm

$$||x|| \equiv \sup_{\tau} \left| \frac{2}{u(\tau)} x(\tau) \right|.$$

With the norm defined this way, 42 can be written

$$x_0 \in \mathsf{C} \equiv \{ x \in \mathsf{X} : ||x - x'|| \leq 1 \},$$

where $x'(\tau) = u(\tau)/2$. The normed dual of X is the set of finite signed measures; i.e. $x^* \in \mathsf{X}^*$ can be written

$$x^*[x] = \int x(\tau) dx^*(\tau),$$

and the norm on X^* is

$$||x^*|| \equiv \int \frac{u(\tau)}{2} |dx^*(\tau)|.$$

The dual problem 41 becomes

$$v(\mathcal{D}) = H[x'] + \sup_{\boldsymbol{\lambda} \in \mathbf{R}^n} \left\{ \boldsymbol{\lambda} \cdot (\boldsymbol{\delta} - \mathbf{f}^*[x']) - \chi ||\boldsymbol{\lambda}||_q^{\boldsymbol{\Sigma}} - \int \frac{u}{2} |d(H - \boldsymbol{\lambda} \cdot \mathbf{f}^*)| \right\}.$$

6 Bounds on Norms and Seminorms

In this section we construct dual problems for lower bounds on norms and seminorms of the difference between the set of acceptable models and some "preferred" model x'. We shall not consider any special prior information: in this section $\mathsf{C} = \mathsf{X}$. Two classes of problems that are thus not treated are lower bounds on norms or seminorms when the unknown x_0 is known to lie in some cone, and bounds on norms or seminorms when x_0 is known to lie in a ball in a different norm.

6.1 Norms

Here we take X to be a Banach space, X^* to be its normed dual, and assume $f_j^* \in \mathsf{X}^*$, $j = 1, \cdots, m$. The primal problem is to find $v(\mathcal{P}) = \inf_{x \in \mathsf{C} \cap \mathsf{D}} ||x - x'||$. The conjugate functional $C^*[x^*] = \inf_{x \in \mathsf{X}} \{||x - x'|| - x^*[x]\}$. The conjugate set

$$\mathsf{C}^* = \{x^* \in \mathsf{X}^* : \inf_x \{||x - x'|| - x^*[x]\} > -\infty\}.$$

If we translate the origin to x',

$$\begin{aligned} \mathsf{C}^* &= \{x^* \in \mathsf{X}^* : \inf_x \{||x|| - x^*[x] + x^*[x']\} > -\infty\} \\ &= \{x^* \in \mathsf{X}^* : ||x^*|| \leq 1\}, \end{aligned}$$

as we may easily see: Suppose $||x^*|| > 1$, and let $0 < \epsilon < ||x^*|| - 1$. Pick any $y_0 \in \mathsf{X}$ such that

$$\frac{|x^*[y_0]|}{||y_0||} \geq ||x^*|| - \epsilon > 1,$$

and let

$$y = \frac{y_0}{||y_0||} \frac{x^*[y_0]}{|x^*[y_0]|}.$$

Then for $\gamma > 0$, $||\gamma y|| = \gamma$ and

$$x^*[\gamma y] \geq \gamma ||x^*|| - \gamma\epsilon = \gamma(||x^*|| - \epsilon).$$

Thus

$$\begin{aligned} C^*[x^*] &\leq \inf_{\gamma > 0} \{||\gamma y|| - x^*[\gamma y] + x^*[x']\} \\ &\leq \inf_{\gamma > 0} \{\gamma(1 - ||x^*|| + \epsilon)\} + x^*[x'] = -\infty. \end{aligned}$$

So if $||x^*|| > 1$, $x^* \notin \mathsf{C}^*$. Suppose $||x^*|| \leq 1$. Then for all $x \in \mathsf{X}$,

$$\begin{aligned} ||x|| - x^*[x] + x^*[x'] &\geq ||x|| - ||x^*||||x|| + x^*[x'] \\ &= ||x||(1 - ||x^*||) + x^*[x'] \\ &\geq x^*[x'], \end{aligned} \tag{43}$$

so $x^* \in \mathsf{C}^*$ □

By 43 we see immediately that $C^*[x^*] = x^*[x']$ for all $x^* \in \mathsf{C}^*$. Thus the dual problem is to find

$$v(\mathcal{D}) = \sup_{\boldsymbol{\lambda} \in \mathbf{R}^m : ||\boldsymbol{\lambda} \cdot \mathbf{f}^*|| \leq 1} \boldsymbol{\lambda} \cdot (\boldsymbol{\delta} - \mathbf{f}^*[x']) - \chi ||\boldsymbol{\lambda}||_q^{\boldsymbol{\Sigma}}. \tag{44}$$

Example: One-sided confidence interval for Ohmic dissipation in the core.

As shown by the work of Shure, Parker and Backus [56] (**SPB** in the following), the Ohmic dissipation in the core can be lower-bounded by a norm of the magnetic scalar potential at the core-mantle boundary; thus measurements of components of Earth's magnetic field can be used to infer a one-sided confidence interval for the dissipation. **SPB** focus on constructing models achieving the minimum; here we are concerned primarily with the confidence interval. They use a two-norm measure of misfit to the data; for contrast let's use the infinity norm. We assume that the observational errors are due to crustal magnetization, the magnitude of which can be bounded using knowledge of rock properties, so that

$$\left| \frac{\epsilon_j}{\sigma_j} \right| \leq \chi.$$

(That is, we treat the errors as systematic rather than random—see also section 10.) **SPB** represent the main field in $l_2^{\boldsymbol{\omega}}$ by the sequence of coefficients in the spherical harmonic expansion (referred to Earth's core) of the magnetic scalar potential. Here

$$l_2^{\boldsymbol{\omega}} \equiv \left\{ (x_0, x_1, x_2, \cdots) : \sum_{j=0}^{\infty} x_j^2 \omega_j < \infty \right\},$$

where the positive weight sequence $(\omega_j)_{j=1}^{\infty}$ is determined by the norm that measures Ohmic dissipation. We have

$$v(\mathcal{D}) = \sup_{\boldsymbol{\lambda} : ||\boldsymbol{\lambda} \cdot \mathbf{f}^*|| \leq 1} \boldsymbol{\lambda} \cdot \boldsymbol{\delta} - \chi ||\boldsymbol{\lambda}||_q^{\boldsymbol{\Sigma}}.$$

Now for the norms in **SPB**, one may evaluate the Gram matrix $\boldsymbol{\Gamma}$:

$$\Gamma_{jk} \equiv < f_j^*, f_k^* >,$$

and the constraint $||\boldsymbol{\lambda} \cdot \mathbf{f}^*|| \leq 1$ becomes

$$\boldsymbol{\lambda} \cdot \boldsymbol{\Gamma} \cdot \boldsymbol{\lambda} \leq 1.$$

Thus

$$v(\mathcal{D}) = \sup_{\boldsymbol{\lambda} : \boldsymbol{\lambda} \cdot \boldsymbol{\Gamma} \cdot \boldsymbol{\lambda} \leq 1} \{ \boldsymbol{\lambda} \cdot \boldsymbol{\delta} - \chi ||\boldsymbol{\lambda}||_1^{\Sigma} \},$$

which can be solved as a sequence of quadratic programming problems.

6.2 Seminorms

We write $|x|$ for the seminorm of x. The primal problem is to find

$$\inf_{x \in \mathsf{C} \cap \mathsf{D}} |x - x'|.$$

We shall consider only the case $\mathsf{C} = \mathsf{X}$, no prior information. Then $C^*[x^*] = \inf_{x \in \mathsf{X}} \{|x - x'| - x^*[x]\}$. Recall (*e.g.* [35]) that the elements of a real linear vector space that have zero seminorm form a linear subspace

$$\mathsf{M} \equiv \{x \in \mathsf{X} : |x| = 0\}.$$

We define the quotient space X/M, X modulo M, by identifying vectors x and y if there exists $m \in \mathsf{M}$ such that $y = x + m$. On the space of equivalence classes set up this way, $|\cdot|$ is a norm—the entire subspace M is identified with the zero element. We now have a minimum-norm problem on the space X/M, and the results of the previous section can be used.

Let

$$\mathsf{M}_\perp^* \equiv \{x^* \in \mathsf{X}^* : x^*[x] = 0 \ \ \forall x \in \mathsf{M}\}.$$

The set $\mathsf{M}_\perp^*$, the orthogonal complement of M in X^*, is a linear subspace of X^*. It can be thought of as the dual space to the space of equivalence classes we have constructed: it consists of all linear functionals on the original space that ignore "components" in M. A linear functional $x^* \in \mathsf{M}_\perp^*$ assigns the same value to all vectors that are equivalent modulo M (assigning the value 0 to all elements of M). It is clear that unless $x^*[x]$ vanishes for all $x \in \mathsf{M}$, $C^*[x^*] = -\infty$: x^* provides a nontrivial bound only if $x^* \in \mathsf{M}_\perp^*$. Define the functional $|\cdot|$ on $\mathsf{M}_\perp^*$ as follows:

$$|x^*| \equiv \sup_{x \in \mathsf{X}} \frac{|x^*[x]|}{|x|}.$$

Then $|x^*|$ is a norm on $\mathsf{M}_\perp^*$, dual to the norm $|x|$ on X/M. The same argument as before gives $\mathsf{C}^* = \{x^* \in \mathsf{M}_\perp^* : |x^*| \leq 1\}$, and $C^*[x^*] = x^*[x']$. Thus the dual problem is

$$v(\mathcal{D}) = \sup_{\{\boldsymbol{\lambda} \in \mathbf{R}^n : \boldsymbol{\lambda} \cdot \mathbf{f}^* \in \mathsf{M}_\perp^* \text{ and } |\boldsymbol{\lambda} \cdot \mathbf{f}^*| \leq 1\}} \boldsymbol{\lambda} \cdot (\boldsymbol{\delta} - \mathbf{f}^*[x']) - \chi ||\boldsymbol{\lambda}||_q^{\boldsymbol{\Sigma}}. \qquad (45)$$

Example: Distance from an m-dimensional subspace.

Here we define

$$|x| \equiv \inf_{\boldsymbol{\gamma} \in \mathbf{R}^m} ||x - \boldsymbol{\gamma} \cdot \mathbf{x}_0||,$$

where $\mathbf{x}_0$ is a fixed m-tuple $(x_1, x_2, \cdots, x_m)$ of elements of X, and the dot product is defined in the obvious way. We then have the dual problem

$$v(\mathcal{P}) \equiv \inf_{x \in \mathsf{D}} |x|$$

$$\geq \quad v(\mathcal{D}) \equiv \sup_{\{\boldsymbol{\lambda}:\boldsymbol{\lambda}\cdot\mathbf{f}^*[x_i]=0, i=1,\cdots m, \text{ and } |\boldsymbol{\lambda}\cdot\mathbf{f}^*|\leq 1\}} \boldsymbol{\lambda}\cdot\boldsymbol{\delta} - \chi||\boldsymbol{\lambda}||_q^{\boldsymbol{\Sigma}}.$$

7 Discretization Error and the Dual

Lang [31] gives a treatment of discretization error for the dual when $\mathsf{X} = C[a, b]$, the Banach space of continuous functions on $[a, b]$ with the "sup" norm; $\mathsf{C} = \{x : x_l(\tau) \leq x(\tau) \leq x_u(\tau)\}$, where x_l and x_u are known piecewise constant functions; H is linear and has a piecewise constant kernel; and the kernels for the data functionals $f_j(\tau)$ are differentiable with bounded derivative. Here we develop a more general treatment, illustrated with geophysical examples.

There are two places where discretization is an issue in the dual problem: in the dual functional, and in the dual constraints ($\boldsymbol{\lambda} \cdot \mathbf{f}^* \in C^*$). It is fairly straightforward to treat errors in the dual functional from discretization in concrete cases where the primal functionals are integrals against known functions (or sums with known weights). The integrability or moduli of continuity of the functions involved (or expressions for the weights in the sums) let one control the maximum possible error. If that is subtracted from the value of the discretized dual, the adjusted value of the discretized dual can not exceed the value of the exact dual. Discretized constraints are just as easy to treat, but conceptually are quite different. The constraints in the dual, when there are any, are infinite-dimensional. Discretizing the constraint set generally admits $\boldsymbol{\lambda}$'s that are not in the continuous constraint set. In order to ensure that the value of the discretized dual is conservative, we must impose the discrete constraints in such a way that the feasible set for the discretized dual is a subset of the feasible set for the continuous problem (or at least that the optimal $\boldsymbol{\lambda}$ in the discretized problem is an element of the original feasible set).

Note that any feasible point of the properly discretized dual gives a rigorous bound on the value of the primal problem. Optimizing the dual functional gives tighter and tighter bounds, but it may be that the computational expense involved in optimizing the dual functional is not worthwhile, if a reasonable (useful) feasible point can be obtained inexpensively and a complete optimization is onerous. This is particularly likely to be the case when the dual functional is not differentiable, unless the dual problem can be written as a linear programming problem. The numerical example in section 7.3.2 below illustrates this. It is also the case that when there are many data, and hence the dimension of the dual problem is large, it can be computationally advantageous to use only a subset of the data: the results remain conservative.

The sections that follow describe in greater detail the approach advocated in this paper: "squeezing" the value of the primal between the discretized primal, and a discretized dual adjusted for the two sources of error just mentioned. The approach for minimum norm problems, bounds on linear functionals of a model in a cone, and linear functionals of a model in a ball are illustrated with examples from gravimetry, seismology, and geomagnetism, respectively. A numerical example is given for linear functionals of a model in an infinity-norm ball in a seismic problem in section 7.3.2.

7.1 Ideal Bodies for Gravity Interpretation—a Minimum Infinity-Norm Problem

This section illustrates the approach with the "ideal bodies" technique for the interpretation of gravity anomalies (Parker [42, 44]). The problem is as follows: we make n gravity anomaly measurements δ_j at points $\{\tau_j\}_{j=1}^n$, $\tau_j \in \mathbf{R}^3$, at a height β above the surface ($\zeta = 0$) of the Earth $\mathbf{E}$, a bounded subset of $\mathbf{R}^3$. Let $\hat{\boldsymbol{\zeta}}$ be the vertical unit vector. The measured gravity anomalies are assumed to be due to a buried anomalous mass $x_0(\tau)$, plus observational noise. The data relations are

$$\delta_j = \int_{\mathbf{E}} \frac{\mathcal{G}\hat{\boldsymbol{\zeta}} \cdot (\tau_j - \tau)}{\|\tau_j - \tau\|_2^3} x_0(\tau) d\tau + \epsilon_j \tag{46}$$

where $\mathcal{G}$ is the gravitational constant and $\|\tau\|_2 = \left(\sum_{k=1}^3 \tau_k^2\right)^{1/2}$ is the Euclidean 2-norm on $\mathbf{R}^3$. We wish to use the data and the (assumed) known distribution of ϵ to get a one-sided confidence interval for the maximum density contrast of the anomalous body x_0. This problem is the basis for inferences about the depth of burial and other interesting properties of the anomalous mass—see Parker [42, 44] for details.

In the discretized primal, we minimize the norm of the solution over a subspace of the entire model space. In the dual problem, we wish to discretize the constraint $\|\boldsymbol{\lambda} \cdot \mathbf{f}^*\| \leq 1$; in order to do so, we need some control over the discrepancy between the functionals $\mathbf{f}^*$ and their discrete approximations.

Here X is the space of functions on the bounded domain $\mathbf{E} \subset \mathbf{R}^3$ with finite "sup" norm; the normed dual space of X is the linear space of finitely additive signed measures with finite (unsigned) mass:

$$\mathsf{X}^* = \{\mu : \int_{\mathbf{E}} |d\mu| < \infty\}.$$

The measures representing f_j^* are absolutely continuous with respect to Lebesgue measure $d\tau$, and $\mathbf{f}^*[x]$ may be written

$$\mathbf{f}^*[x] = \int_{\mathbf{E}} \mathbf{f}(\tau) x(\tau) d\tau,$$

where $\mathbf{f}$ is the n-vector of continuous real-valued functions

$$f_j(\tau) \equiv \frac{\mathcal{G}\hat{\boldsymbol{\zeta}} \cdot (\tau_j - \tau)}{\|\tau_j - \tau\|_2^3},$$

and $d\tau$ is Lebesgue measure on $\mathbf{R}^3$.

7.1.1 Discretizing the Primal

We may discretize the primal problem by partitioning the region $\mathbf{E}$ into m "boxes" Δ_i, and defining the indicator functions

$$x_i(\tau) = \begin{cases} 1, & \tau \in \Delta_i, \\ 0, & \text{otherwise;} \end{cases}$$

and the matrix Φ with elements

$$\Phi_{ji} = f_j^*[x_i].$$

Let $\alpha \in \mathbf{R}^m$. The discrete primal problem is then to find

$$\min\{\sup_i |\alpha_i|\} \text{ such that } \|\Phi \cdot \alpha - \delta\|_1 \leq \chi.$$

This may be written as a linear programming problem:

$$\min \sigma$$

such that

$$\begin{aligned} \sigma &\geq 0 \quad (\mathbf{R}) \\ \mathbf{o} \leq \pi, \nu &\leq \sigma \mathbf{1} \quad (\mathbf{R}^m) \\ \Phi \cdot (\pi - \nu) - \delta + \eta - \zeta &= \mathbf{o} \quad (\mathbf{R}^m) \\ \eta, \zeta &\geq \mathbf{o} \quad (\mathbf{R}^n) \\ \mathbf{1} \cdot (\eta + \zeta) &\leq \chi \quad (\mathbf{R}) \end{aligned}$$

where $\mathbf{1} = (1, 1, \ldots, 1)^T$ and $\mathbf{o} = (0, 0, \ldots, 0)^T$.

See Huestis and Ander [28] and Huestis [27] for techniques when the fit to the data is measured in the infinity-norm.

7.1.2 Discretizing the Dual

Recall that the dual problem is to find

$$\sup_{\{\lambda : \|\lambda \cdot \mathbf{f}^*\|_1 \leq 1\}} \lambda \cdot \delta - \chi \|\lambda\|_\infty.$$

The norm of the functional $\lambda \cdot \mathbf{f}^*$ is

$$\|\lambda \cdot \mathbf{f}^*\|_1 = \int_{\mathbf{E}} |\lambda \cdot \mathbf{f}(\tau)| d\tau.$$

We may discretize the dual problem by dividing $\mathbf{E}$ into m boxes Δ_i as in the discretized primal. Let $\mathbf{f}^i$ be the n-vector with components

$$f_j^i \equiv \int_{\Delta_i} f_j(\tau),$$

let $\tilde{\mathbf{f}}(\tau)$ be the corresponding piecewise constant function

$$\tilde{\mathbf{f}}(\tau) \equiv \mathbf{f}^i, \quad \tau \in \Delta_i,$$

and let $\tilde{\mathbf{f}}^*$ be the corresponding vector of functionals:

$$\tilde{\mathbf{f}}^*[x] \equiv \int_{\mathbf{E}} \tilde{\mathbf{f}}(\tau) x(\tau) d\tau.$$

We wish to compare

$$\|\boldsymbol{\lambda} \cdot \mathbf{f}^*\| \equiv \int_{\mathbf{E}} |\boldsymbol{\lambda} \cdot \mathbf{f}(\tau)|$$

with its discrete approximation

$$\begin{aligned} \|\boldsymbol{\lambda} \cdot \tilde{\mathbf{f}}^*\| &\equiv \int_{\mathbf{E}} |\boldsymbol{\lambda} \cdot b\tilde{f} f(\tau)| \\ &= \sum_i \Delta_i |\boldsymbol{\lambda} \cdot \mathbf{f}^i|. \end{aligned}$$

Denote the magnitude of their difference by $\mathcal{E}$:

$$\begin{aligned} \mathcal{E} &\equiv \left| \|\boldsymbol{\lambda} \cdot \mathbf{f}^*\| - \|\boldsymbol{\lambda} \cdot \tilde{\mathbf{f}}^*\| \right| \\ &\leq \|\boldsymbol{\lambda} \cdot (\mathbf{f}^* - \tilde{\mathbf{f}}^*\| \\ &\leq |\boldsymbol{\lambda}| \cdot \left(\|f_j^* - \tilde{f}_j^*\| \right)_{j=1}^n \end{aligned} \tag{47}$$

Now the rightmost term is

$$\begin{aligned} \varrho &\equiv \left(\|f_j^* - \tilde{f}_j^*\| \right)_{j=1}^n \\ &= \left(\sum_i \int_{\Delta_i} |f_j^* - \tilde{f}_j^*| \right)_{j=1}^n . \end{aligned} \tag{48}$$

For some sorts of prismatic Δ_i, this can be computed analytically. (See section 7.3.2 for a similar result.) For others, it will be necessary to bound the components of ϱ, which can be achieved using the modulus of continuity: Define $\boldsymbol{\rho}^i(v)$ to be the n-vector with components:

$$\rho_j^i \equiv \sup_{\tau, \sigma \in \Delta_i} |f_j(\tau) - f_j(\sigma)|.$$

Clearly

$$\varrho_j \leq \sum_i \Delta_i \rho_j^i,$$

where Δ_i is the volume of $\boldsymbol{\Delta}_i$. The monotonicity of the functions $f_j(\boldsymbol{\tau})$ with distance from $\boldsymbol{\tau}_j$ makes computation of ρ_j^i straightforward.

Since the $f_j(\boldsymbol{\tau})$ are differentiable, there is another inequality we may use as a bound:

$$\rho_j^i \leq \Delta_i \sum_{k=1}^{3} \sup_{\boldsymbol{\tau} \in \boldsymbol{\Delta}_i} |\partial_{\tau_k} f_j(\boldsymbol{\tau})|.$$

These two final techniques rely on the fact that the observations are made above the surface, at height at least $\beta > 0$; otherwise, the f_j have singularities in $\mathbf{E}$ and the moduli of continuity are infinite.

By one of these methods or some other, we can find a vector $\boldsymbol{\varrho}$ satisfying

$$\mathcal{E} \leq |\boldsymbol{\lambda}| \cdot \boldsymbol{\varrho}.$$

Thus we have

$$\|\boldsymbol{\lambda} \cdot \mathbf{f}^*\|_1 \leq \sum_i \Delta_i |\boldsymbol{\lambda} \cdot \mathbf{f}^i| + |\boldsymbol{\lambda}| \cdot \boldsymbol{\varrho}. \tag{49}$$

Let $\boldsymbol{\Phi}$ be the n by m matrix whose elements are given by

$$\Phi_{ji} \equiv \Delta_i f_j^i.$$

Write $\boldsymbol{\lambda}$ as the difference between two positive n-vectors $\boldsymbol{\pi}$ and $\boldsymbol{\nu}$, and let $\boldsymbol{\eta}$ and $\boldsymbol{\zeta}$ be two positive m-vectors. Equation 49 shows that the set

$$\{\boldsymbol{\lambda} : \|\boldsymbol{\lambda} \cdot \mathbf{f}^*\|_1 \leq 1\}$$

contains the set

$$\{\boldsymbol{\lambda} = \boldsymbol{\pi} - \boldsymbol{\nu} : \boldsymbol{\pi}, \boldsymbol{\nu} \geq 0 \text{ and } \boldsymbol{\Phi} \cdot (\boldsymbol{\pi} - \boldsymbol{\nu}) + \boldsymbol{\eta} - \boldsymbol{\zeta} = 0 \text{ and } \mathbf{1} \cdot (\boldsymbol{\eta} + \boldsymbol{\zeta}) + \boldsymbol{\varrho} \cdot (\boldsymbol{\pi} + \boldsymbol{\nu}) \leq 1\},$$

where $\mathbf{1}$ is the vector $(1, \cdots, 1)^T$. The conservatively discretized dual problem is thus the (large) linear programming problem

$$\min(-\boldsymbol{\delta} + \mathbf{f}^*[x^\Gamma]) \cdot (\boldsymbol{\pi} - \boldsymbol{\nu}) + \sigma\chi$$

such that

$$\begin{aligned} \sigma &\geq 0 \quad (\mathbf{R}) \\ 0 \leq \boldsymbol{\pi}, \boldsymbol{\nu} &\leq \sigma \mathbf{1} \quad (\mathbf{R}^n) \\ \boldsymbol{\eta}, \boldsymbol{\zeta} &\geq 0 \quad (\mathbf{R}^m) \\ \boldsymbol{\Phi} \cdot (\boldsymbol{\pi} - \boldsymbol{\nu}) + \boldsymbol{\eta} - \boldsymbol{\zeta} &= 0 \quad (\mathbf{R}^m) \\ \mathbf{1} \cdot (\boldsymbol{\eta} + \boldsymbol{\zeta}) + \boldsymbol{\varrho} \cdot (\boldsymbol{\pi} + \boldsymbol{\nu}) &\leq 1. \end{aligned}$$

The value of this discretized dual is less than the value of the exact dual

$$\max_{\boldsymbol{\lambda}:\|\boldsymbol{\lambda}\cdot\mathbf{f}^*\|_1\leq 1} \{\boldsymbol{\lambda}\cdot(\boldsymbol{\delta}-\mathbf{f}^*[x'])-\chi\|\boldsymbol{\lambda}\|_\infty\}.$$

which is less than or equal to the value of the primal

$$\inf_{x\in\mathsf{C}\cap\mathsf{D}} \{\sup_{\tau}|x(\tau)|\},$$

where

$$\mathsf{C}\equiv\mathsf{X}=\{\text{functions on } \mathbf{E} \text{ with finite sup norm}\},$$

and

$$\mathsf{D}\equiv\{x\in\mathsf{X}:\|\mathbf{f}^*[x]-\boldsymbol{\delta}\|_1\leq\chi\}$$

7.2 Linear Functionals of a Model in a Cone: Seismic Velocity Bounds from $\tau(p)$ and $X(p)$

This section illustrates bounding discretization error in the dual for bounds on linear functionals of a model in a cone, with the seismic inverse problem of inferring bounds on velocity as a function of depth from $\tau(p)$ (vertical delay time as a function of ray parameter) and $X(p)$ (horizontal distance as a function of ray parameter) data from a halfspace in which velocity varies only with depth, and increases monotonically with depth. The section 7.3.2 covers an extension, derived from a problem in a spherical geometry, in which an upper bound on the rate of increase of velocity with respect to depth is available. Section 10.2 covers an extension to the case where it is not known precisely which data functionals were measured. The approach taken here originated with Garmany [21] and is discussed and developed in Garmany *et al.* [22], Orcutt [39], and Stark and others [64, 63, 61, 59, 60].

In this section, section 7.3.2, and section 10.2, p stands for ray parameter and takes on real, rather than integer values. We model the Earth in terms of depth x as a function of velocity v, rather than velocity as a function of depth. The data mapping functionals for this problem are

$$f_j^*[x] = \int_a^b 2Re\{(v^{-2}-p_j^2)^{1/2}\}dx(v), \quad j=1,\cdots n_\tau \tag{50}$$

$$f_j^*[x] = \int_a^b 2p_jRe\{(v^{-2}-p_j^2)^{-1/2}\}dx(v), \quad j=n_\tau+1,\cdots,n \tag{51}$$

where a is the velocity at the surface of the Earth, $b>\max_j\{p_j^{-1}\}$, and x is a measure on $[a,b]$. The equations 50 are for the measurements of $\tau(p)$; the equations 51 are for $X(p)$ measurements. To ensure that $x(v)$ assigns a single velocity to each depth (excluding a countable number of discontinuities) we

must ask that x be a *positive* measure on $[a, b]$. Perhaps the most interesting linear functional to consider is

$$H_v[x] \equiv \int_a^v dx(v)$$

which gives the depth to the velocity v. It is the equivalent of the point evaluator for this problem: by finding strict bounds on H_v for many values of v, we can construct a "corridor" containing the models consistent with the data (in the sense that $\|\mathbf{f}^*[x] - \delta\|_2^{\Sigma} \leq \chi$) and the prior constraint $x(dv) \geq 0$.

7.2.1 Discretizing the Primal

There are numerous references for discretizing the problem of finding bounds on linear functionals in positivity-constrained problems using numerical programming, including Garmany, Orcutt, and Parker [22], Lang [31], Oldenburg [38], Rust and Burris [49], Sabatier [51, 52], Safon, Vasseur and Cuer [54], and Stark and others [59, 64, 63, 62]. Of these, Lang [31], Rust and Burris [49], Stark and Parker [63], and Stark [59] address the two-norm measure of misfit to the data.

As usual, the primal problem may be discretized by parametrizing the depth-velocity model in terms of a linear combination of a known finite set of functions. In this problem, piecewise constant, piecewise linear, piecewise quadratic, and piecewise reciprocal quadratic representations have been used. Once the approximating functions have been chosen it is straightforward to discretize the primal problem; the reader is referred to the references above.

7.2.2 Discretizing the Dual

It seems to be considerably more difficult to discretize the dual to this problem in a sharp way than it is for prior balls—even positivity constraints seem to lead to mixed-integer linear programming, a combinatorial optimization problem. It is still possible to guarantee conservative answers to the discretized dual, but in general the answers will be *very* conservative. The fundamental difficulty is that positivity constraints in the primal lead to the constraint in the discretized dual that the absolute value of some quantities be *greater* than a constant. Linear programming can readily impose the constraint $|x_i| < c$, but not $|x_i| > c$. The constraint $|x_i|$ can be written $x_i < c$ *and* $-x_i < c$, whereas $|x_i| > c$ involves a disjunction: $x_i > c$ *or* $-x_i > c$. Such disjunctions can, however, be treated with mixed-integer linear programming (see, for example, Papadimitriou and Steiglitz [41]). It is also possible to use a stronger constraint instead so that linear programming can solve the discretized dual at the cost of a less sharp, but still conservative result.

Finally, there are other methods for solving semi-infinite linear programming problems, such as imposing the positivity constraints at a finite grid *without* the "fudge factors" mentioned above, then *a posteriori* testing to see if the constraint

is violated between grid points. If so, more points are added to the grid, *etc.* To do this rigorously again requires the use of the modulus of continuity of the functions to determine how fine a grid one must use when performing the test for violations of the positivity constraint. See Hettich [25] and Hu [26] for more details and other algorithms.

Here is a sketch of what is involved in following through for positivity constraints. Take X to be functions on $[a,b]$, C the set of functions pointwise positive on $[a,b]$. Assume as before that the functionals $\mathbf{f}^*[x]$ can be written $\mathbf{f}^*[x] = \int_a^b \mathbf{f}(t)x(t)dt$ for n continuous functions $f_j(t)$. We may readily see that $\mathsf{C}^* \cap \mathsf{D}^*$ is the set of linear functionals formed by linear combinations of the kernels $f_j(t)$ that are everywhere less than or equal to $h(t)$:

$$\mathsf{C}^* \cap \mathsf{D}^* = \{x^* \in \mathsf{X}^* : x^*[x] = \int_a^b \boldsymbol{\lambda} \cdot \mathbf{f}(t)x(t)dt \text{ and } h(t) - \boldsymbol{\lambda} \cdot \mathbf{f}(t) \geq 0,\ t \in [a,b].$$

The difficulty in discretizing this problem is somehow to characterize $\mathsf{C}^* \cap \mathsf{D}^*$ with only a finite number of constraints. Again, we resort to the modulus of continuity of the functions $\mathbf{f}(t)$ and $h(t)$, and express the modulus of continuity of $h(t) - \boldsymbol{\lambda} \cdot \mathbf{f}(t)$ in terms of them. We want to be sure that the discrete samples of $h - \boldsymbol{\lambda} \cdot \mathbf{f}$ are sufficiently large that between samples the function can not become negative. There are two factors that "use up" the available modulus of continuity: the function can't go negative if its values at both endpoints are sufficiently large, nor if the difference between the values at the endpoints is sufficiently large (the function "consumes" its modulus climbing from one endpoint to the other). It is straightforward but tedious to turn this idea into an expression for how large the linear combination must be at its sample points to ensure it does not go negative between samples.

7.3 Linear Functionals of a Model in a Ball

In this section we consider discretizing the problem of finding strict bounds on a linear functional of a model where the prior information is that the model lies in a ball in some norm around a known point. We have

$$\begin{aligned} v(\mathcal{P}) &= \inf_{\{x : \|\mathbf{f}^*[x] - \boldsymbol{\delta}\|_p^{\boldsymbol{\Sigma}^{-1}} \leq \chi \text{ and } \|x - x'\| \leq v\}} h^*[x] \\ &\geq v(\mathcal{D}) = \max_{\boldsymbol{\lambda} \in \mathbf{R}^n} \left\{ \boldsymbol{\lambda} \cdot (\boldsymbol{\delta} - \mathbf{f}^*[x']) - \chi \|\boldsymbol{\lambda}\|_q^{\boldsymbol{\Sigma}} - v\|\boldsymbol{\lambda} \cdot \mathbf{f}^* - h^*\| + h^*[x'] \right\}. \end{aligned}$$

See Safon *et al.* [54], Parker [42, 44], Stark and others [64, 63, 59], Lang and Marzetta [31, 32, 33] for examples where the ball is in the infinity norm. See Backus [7, 6] for a two-norm example.

We shall assume that the vector $\mathbf{f}^*[x']$ and the scalar $h^*[x']$ can be evaluated analytically, but that $\|\boldsymbol{\lambda} \cdot \mathbf{f}^* - h^*\|$ can not in general. We will discretize this term; we wish to bound the error committed. In examples, this amounts to bounding

the error in a quadrature scheme, or in a truncated infinite sum. Denote by $\tilde{\mathbf{f}}^*$ and $\tilde{h}^*$ the discrete approximations to $\mathbf{f}^*$ and h^*. In analogy to equation 47, let

$$\begin{aligned}
\mathcal{E} &\equiv \left| \|\boldsymbol{\lambda}\cdot\mathbf{f}^* - h^*\| - \|\boldsymbol{\lambda}\cdot\tilde{\mathbf{f}}^* - \tilde{h}^*\| \right| \\
&\leq \|\boldsymbol{\lambda}\cdot\mathbf{f}^* - h^* - (\boldsymbol{\lambda}\cdot\tilde{\mathbf{f}}^* - \tilde{h}^*)\| \\
&= \|\boldsymbol{\lambda}\cdot(\mathbf{f}^* - \tilde{\mathbf{f}}^*) - (h^* - \tilde{h}^*)\| \\
&\leq \|\boldsymbol{\lambda}\cdot(\mathbf{f}^* - \tilde{\mathbf{f}}^*)\| + \|h^* - \tilde{h}^*\| \\
&\leq |\boldsymbol{\lambda}|\cdot(\|f_j^* - \tilde{f}_j^*\|)_{j=1}^n + \|h^* - \tilde{h}^*\| \\
&\leq \|\boldsymbol{\lambda}\|_q^{\boldsymbol{\Sigma}} \left\|(\|f_j^* - \tilde{f}_j^*\|)_{j=1}^n\right\|_p^{\boldsymbol{\Sigma}^{-1}} + \|h^* - \tilde{h}^*\|, \qquad (52)
\end{aligned}$$

where in the penultimate equation, the absolute value of $\boldsymbol{\lambda}$ is taken componentwise. If we can find numbers ϱ_f and ϱ_h such that

$$\varrho_f \geq \left\|(\|f_j^* - \tilde{f}_j^*\|)_{j=1}^n\right\|_p^{\boldsymbol{\Sigma}^{-1}} \qquad (53)$$

and

$$\varrho_h \geq \|h^* - \tilde{h}^*\|, \qquad (54)$$

then

$$\mathcal{E} \leq \varrho_f\|\boldsymbol{\lambda}\|_q^{\boldsymbol{\Sigma}} + \varrho_h,$$

and hence

$$v(\mathcal{D}) \geq \max_{\boldsymbol{\lambda}}\left\{\boldsymbol{\lambda}\cdot(\boldsymbol{\delta} - \mathbf{f}^*[x']) - (\chi + v\varrho_f)\|\boldsymbol{\lambda}\|_q^{\boldsymbol{\Sigma}} - v\|\boldsymbol{\lambda}\cdot\tilde{\mathbf{f}}^* - \tilde{h}^*\| + h^*[x'] - v\varrho_h\right\}. \qquad (55)$$

Note that this amounts to solving the dual with an increased value of χ, plus a constant factor $-v\varrho_h$.

We now set out to find numbers ϱ_f and ϱ_h satisfying 53 and 54 using different discrete approximations to $\mathbf{f}^*$ and h^*.

7.3.1 Discretizing the Dual in L_p Spaces

We specialize to $\mathsf{X} = L_{p'}[a,b]$ $(1 \leq p' < \infty)$, $\mathsf{X}^* = L_{q'}[a,b]$, $\mathbf{f}^*[x] = \int_a^b \mathbf{f}(t)x(t)dt$, and $h^*[x] = \int_a^b h(t)x(t)dt$. An example where $\mathsf{X} = L_\infty$ appears in section 7.3.2. Divide the interval $[a,b]$ as before into m subintervals

$$\Delta_i \equiv \begin{cases} [t_i, t_{i+1}), & 1 \leq i \leq m \\ [t_m, t_{m+1}], & i = m \end{cases}$$

where $t_1 = a$ and $t_{m+1} = b$, and let $\Delta_i = t_{i+1} - t_i$.

We will discretize f_j and h by functions constant in each Δ_i. Let these approximations have the values f_j^i and h^i in Δ_i. Define

$$\mathcal{I}_j^i(f) \equiv \int_{\Delta_i} |f_j(t) - f_j^i|^{q'} dt \tag{56}$$

$$\mathcal{I}_j^i(h) \equiv \int_{\Delta_i} |h(t) - h^i|^{q'} dt. \tag{57}$$

We can then write

$$\|f_j^* - \tilde{f}_j^*\| = \left(\sum_{i=1}^m \mathcal{I}_j^i(f) \right)^{1/q'}$$

$$\|h^* - \tilde{h}^*\| = \left(\sum_{i=1}^m \mathcal{I}_j^i(h) \right)^{1/q'}.$$

These lead obviously to values of ϱ_f and ϱ_h that work in equation 55. The fact that $f_j, h \in L_{q'}$ implies that with any sensible choice of $\tilde{f}_j^*$ and $\tilde{h}^*$ these can be made as small as desired by choosing Δ_i sufficiently short.

We will look at two different schemes for assigning f_j^i and h^i and finding the corresponding ϱ_f and ϱ_h. The first is based directly on the integrability of f_j and h; the second uses continuity properties of those functions that might hold in some problems. The first approach will be illustrated numerically in section 7.3.2 for a seismic problem.

Prototype 1. Define

$$f_j^i \equiv \frac{1}{\Delta_i} \int_{\Delta_i} f_j(t) dt \tag{58}$$

$$h^i \equiv \frac{1}{\Delta_i} \int_{\Delta_i} h(t) dt \tag{59}$$

Since $f_j, h \in L_{q'}$, and $L_q[a,b] \subset L_r[a,b]$ for $r \leq q$, the integrals exist. (If the original $L_{p'}$ space was weighted, weights can be introduced as needed.) This is in some sense the "correct" way to discretize problems in L_p spaces since it assumes only that the kernels $f_j(t)$ and $h(t)$ are integrable.

Assuming we can bound $\mathcal{I}_j^i(f)$ and $\mathcal{I}_j^i(h)$, we are done: let ρ_f be the n-vector with components

$$(\rho_f)_j \geq \left(\sum_{i=1}^m \mathcal{I}_j^i(f) \right)^{1/q'}.$$

Then $\varrho_f = \|\rho_f\|_p^{\Sigma^{-1}}$ and $\varrho_h = \left(\sum_{i=1}^m \mathcal{I}_j^i(h)\right)^{1/q'}$ work. See section 7.3.2.

Prototype 2. This approach requires continuity of $f_j(t)$ and $h(t)$. We discretize

$\mathbf{f}$ by sampling at arbitrary points within the intervals Δ_i. Let $\bar{t}_i \in \Delta_i$, $\mathbf{f}^i \equiv \mathbf{f}(\bar{t}_i)$, and $h^i \equiv h(\bar{t}_i)$. Define $\rho_f^i(\Delta)$ to be the n-vector whose j^{th} component is

$$(\rho_f^i)_j(\Delta) \equiv \sup_{|t-\bar{t}_i| \leq \Delta} |f_j(t) - f_j(\bar{t}_i)|,$$

and define

$$\rho_h^i(\Delta) \equiv \sup_{|t-\bar{t}_i| \leq \Delta} |h(t) - h(\bar{t}_i)|.$$

Then clearly

$$\begin{aligned} \|f_j^* - \tilde{f}_j^*\| &\leq \left(\sum_i \int_{\Delta_i} |f_j(t) - f_j^i|^{q'} dt \right)^{1/q'} \\ &\leq \left(\sum_i \Delta_i \left[(\rho_f^i)_j(\Delta_i) \right]^{q'} \right)^{1/q'}, \end{aligned}$$

and similarly

$$\|h^* - \tilde{h}^*\| \leq \left(\sum_i \Delta_i \rho_h^i(\Delta_i)^{q'} \right)^{1/q'};$$

from which values of ϱ_f and ϱ_h follow directly by inspection of equations 53 and 54. As in the previous subsections, it may be easier to use different bounds on the change of the $\mathbf{f}(t)$ than to use ρ_f^i, or to use Prototype 1.

7.3.2 Numerical Example: Seismic Velocity in Earth's Core

See section 7.2 for the problem set-up; the difference here is that we will incorporate the constraint, developed in [64], that $0 \leq x_0(dv) \leq \frac{R}{v} dv$, where R is the radius of Earth's core. The addition of the upper bound on the measure forces x_0 to be continuous with respect to Lebesgue measure; thus dx/dv exists and is less than $\frac{R}{v}$. This means that our model x_0 is in the space of continuous functions possessing finite first derivatives. In a slight stretch of notation, we will denote this space by $C^{1-}[a,b]$. This is to draw a distinction from the usual notation, in which $C^1[a,b]$ is the space of *continuously differentiable* functions on $[a,b]$. Thus x_0 lies in the ball

$$\mathsf{C} = \left\{ x \in C^{1-}[a,b] : 0 \leq \frac{dx}{dv} \leq \frac{R}{v} \right\}.$$

Now since we are interpreting x as a measure, the following functional is really a norm:

$$\|x\| \equiv \sup_{v \in [a,b]} \frac{2v}{R} \left| \frac{dx(v)}{dv} \right|. \tag{60}$$

Now endow $C^{1-}[a,b]$ with that norm. Note that it is unimportant for the subsequent analysis whether $C^{-1}[a,b]$ is complete with respect to the norm 60. We may write $\mathsf{C} = \{x \in C^{1-}[a,b] : \|x - x'\| \leq 1\}$, where x' is defined by $\frac{dx'}{dv} \equiv \frac{R}{2v}$. Now the norm on the dual space X^* is by definition

$$\|x^*\| \equiv \sup_{\|x\| \leq 1} x^*[x].$$

If $x^*[x]$ can be written $x^*[x] = \int_a^b x^*(v)dx(v)$, then it is easy to see that

$$\|x^*\| = \int_a^b \frac{R}{2v}|x^*(v)|dv, \tag{61}$$

which is a weighted 1-norm of the function $x^*(v)$. We are interested only in norms of $\boldsymbol{\lambda} \cdot \mathbf{f}^* - h^*$, which *can* be written in that form, as integrals of dx/dv against ordinary Lebesgue integrable functions. This is important, as the dual of L_∞ is rather nastier than L_1, which is not the dual of any normed space.

This example is particularly interesting since some of the data functionals (those corresponding to $X(p)$ data—see equation 63) have integrable singularities. Recall from section 7.2 that the functionals h^* of interest in this problem are of the form

$$h^*[x] = \int_a^b 1_{\{v \leq v_h\}} dx(v),$$

where 1_Ω is the indicator function of the set Ω. The integrals $\mathbf{f}^*[x']$ and $h^*[x']$ can be performed analytically; the results are in appendix A. Thus as indicated above in section 7.3.1, we are concerned only with the discrete approximation of $\|\boldsymbol{\lambda} \cdot \mathbf{f}^* - h^*\|$. For our choice of h^* in this problem, the second term in 55 can be made to equal zero by including v_h among the points in the discretization of $[a,b]$. Hence the only approximation we need to control is $\|\boldsymbol{\lambda} \cdot (\mathbf{f}^* - \tilde{\mathbf{f}}^*)\|$. The kernels $f_j(v)$ are as follows (see equations 50 and 51):

$$f_j(v) = 2Re\{(v^{-2} - p_j^2)^{1/2}\}, \quad j = 1, \cdots n_\tau \tag{62}$$

$$f_j(v) = 2p_j Re\{(v^{-2} - p_j^2)^{-1/2}\}, \quad j = n_\tau + 1, \cdots, n \tag{63}$$

As the $X(p)$ kernels 63 are singular at p_j^{-1}, their moduli of continuity are infinite, and Prototype 2 in section 7.3.1 fails. We use the approach in Prototype 1, which is requires only integrability of the $\tau(p)$ and $X(p)$ kernels.

Define f_j^i as in equation 58. In analogy to equation 56, but with weights, let

$$\begin{aligned} \mathcal{I}_j^i &\equiv \int_{\boldsymbol{\Delta}_i} \frac{R}{2v}|f_j(v) - f_j^i|dv \\ &= \int_{\boldsymbol{\Delta}_i} \frac{R}{2v}\left|f_j(v) - \frac{1}{\Delta_i}\int_{\boldsymbol{\Delta}_i} f_j(t)dt\right|dv. \end{aligned} \tag{64}$$

Then

$$\|f_j^* - \tilde{f}_j^*\| = \sum_{i=1}^{m} \mathcal{I}_j^i, \tag{65}$$

which can be made as small as desired by making Δ_i sufficiently short. Let ρ_f^i be the n-vector whose j^{th} component is

$$(\rho_f^i)_j \equiv \|f_j^* - \tilde{f}_j^*\|. \tag{66}$$

If we

$$\varrho_f \equiv \|\rho_f^i\|_p^{\Sigma^{-1}}, \tag{67}$$

equation 55 will hold. Appendix A shows how to compute ϱ numerically.

Figure 3 shows the $\tau(p)$ and $X(p)$ kernels for the data we will analyze. The data are from the work of Johnson and Lee [30]; strict bounds in the discretized primal were found for this data set by Stark and others [64, 63].

Johnson and Lee analyze the data using the strict bounds technique of Bessonova and others [10, 11], which requires interpolation of the finite set of data available. Stark *et al.* [64] demonstrate that the numerical programming approach followed here is superior; see also the argument in Orcutt *et al.* [40].

It is clear from Figure 3 that for ϱ_f to be small, the discretization must be very fine near the origin, where the $\tau(p)$ kernels change rapidly, and in the range 25-35 $km\ s^{-1}$, near the singularities of the $X(p)$ kernels.

Once we have computed ϱ_f, we need to solve the discretized dual optimization problem. We will take $p = 2$ and Σ diagonal following Stark and Parker [63, 61]. Let ω be the m-vector of integrals of $\frac{R}{2v}$ over the intervals of discretization:

$$\begin{aligned} \omega_i &\equiv \int_{\Delta_i} \frac{R}{2v} dv \\ &= \frac{R}{2} \ln \frac{v_{i+1}}{v_i}. \end{aligned}$$

Let Φ be the matrix of the discretized values of $\mathbf{f}$, weighted by ω:

$$\Phi_{ji} \equiv f_j^i \omega_i; \tag{68}$$

let γ be the vector of the discretized values of h weighted by ω:

$$\gamma_i \equiv h^i \omega_i; \tag{69}$$

and let

$$\delta' \equiv \delta - \mathbf{f}^*[x'] \tag{70}$$

$$\chi' \equiv \chi + \varrho_f. \tag{71}$$

Then, provided we include v_h in the discretization of $[a, b]$, the adjusted discrete dual is

$$\begin{aligned} v(\tilde{\mathcal{D}}) &= \max_{\boldsymbol{\lambda} \in \mathbf{R}^n} \left\{ \boldsymbol{\lambda} \cdot \boldsymbol{\delta}' - \chi' \|\boldsymbol{\lambda}\|_2^{\boldsymbol{\Sigma}} - \|\boldsymbol{\lambda} \cdot \boldsymbol{\Phi} - \boldsymbol{\gamma}\|_1 + h^*[x'] \right\} \\ &= h^*[x'] - \min_{\boldsymbol{\lambda}} \left\{ -\boldsymbol{\lambda} \cdot \boldsymbol{\delta}' + \chi' \|\boldsymbol{\lambda}\|_2^{\boldsymbol{\Sigma}} + \|\boldsymbol{\lambda} \cdot \boldsymbol{\Phi} - \boldsymbol{\gamma}\|_1 \right\} \qquad (72) \\ &= h^*[x'] - \min_{\substack{\{\boldsymbol{\lambda}, \boldsymbol{\pi}, \boldsymbol{\nu} \in \mathbf{R}^n : \\ \boldsymbol{\pi}, \boldsymbol{\nu} \geq 0 \\ \boldsymbol{\lambda} \cdot \boldsymbol{\Phi} + \boldsymbol{\pi} - \boldsymbol{\nu} = \boldsymbol{\gamma}\}}} \left\{ \chi' \|\boldsymbol{\lambda}\|_2^{\boldsymbol{\Sigma}} + \mathbf{1} \cdot (\boldsymbol{\pi} + \boldsymbol{\nu}) - \boldsymbol{\lambda} \cdot \boldsymbol{\delta}' \right\}, \end{aligned}$$

where $\mathbf{1}$ is a $2m$-vector of ones. This final optimization is a quadratic programming problem in $n + 2m$ variables, with positivity constraints on $2m$ of those variables, and m linear equality constraints. Unfortunately, the value of m needed to get ϱ down to 2.29, which is about 4% of $\chi = 59.7$ in this problem, is about 19,000. Since n is only 30, it is substantially more efficient to use unconstrained nonlinear optimization and equation 72. Although it is continuous, the objective functional is not differentiable everywhere. However, since the functional is concave, subgradients of the negative of the objective functional does exist; appendix A gives such a subgradient. The objective functionals were minimized using a two step procedure involving the Nelder-Mead simplex method (see [48, 20]) and Stanford Systems Optimization Laboratory algorithm NPSOL. See section A.4 for details.

Figure 1 compares the solutions to this problem for 89 values of v_h (178 optimization problems, solid lines) with the discretized primal results of Stark and Parker [61] (dashed lines), using the data from Johnson and Lee [30], with uncertainties assigned to the $X(p)$ data by Stark and Parker [63], and 19,118 points in the discretized dual. The primal results were obtained using 100 basis functions, equally spaced in $1/v$, plus extra basis functions either side of v_h. The results were essentially unchanged when 200 basis functions were used. The computational algorithm used solves the problem as a sequence of linearly constrained, quadratic programming problems with upper and lower bounds on the variables. The central algorithm of the method in [61] was modeled after the phenomenally robust NNLS algorithm of Lawson and Hanson [34].

The bounds in figure 1 from the dual are extra-conservative, as the algorithmic procedure used to solve the dual may not find the global optimal values. Note, however, that any value of the discretized dual is a rigorous bound on the value of the primal: by spending more computer money to find better values of the discrete dual functional, we get tighter results, but no extra rigor.

Note that there is no duality gap for this problem, by conditions given in section 11: the feasibility of the primal problem guarantees that the dual is finite, and solutions in C exist with misfit less than χ. In cases where there is known to be no duality gap, there is another way to bracket the true confidence

interval: the value of the discretized, corrected dual is less than the value of the true dual (which is equal to the true primal), and the value of the true dual is less than the value of the discrete dual without correction terms. Thus by solving the dual problem with two different values of χ' ($\chi + \varrho_f$ and χ), we can bracket a confidence interval with at least its nominal coverage probability. For this procedure to work, one must have a computational algorithm for the dual that is guaranteed to converge to the global extremum. In spite of the convexity of the problem, the procedure used here may not achieve the global optimal value; thus using a smaller χ' would not necessarily bracket the correct results from the "inside."

7.3.3 Separable Hilbert Spaces: Core Magnetism

Backus [7] shows that the problem of inferring the value of a single coefficient in the spherical harmonic expansion of the core field, or of inferring the average magnetization of a disc on the core, using observations of Earth's magnetic field from satellites and ground observatories and the prior information that the equivalent mass of the energy stored in the field is less than the mass of Earth, can be posed in the infinite-dimensional space of coefficients in the spherical harmonic expansion of the magnetic potential on the core. The constraint that the energy in the field is bounded by a particular constant constrains the model to lie in a (weighted) two-norm ball. The functionals that evaluate a spherical harmonic component or the average field in a disc are obviously linear.

All separable Hilbert spaces are isomorphic to l_2, so without loss of generality, we assume some countable orthonormal basis has been chosen, and consider only $\mathsf{X} = l_2$, the space of coefficients in that basis. We have assumed that h^* and f_j^* are bounded, so by the Riesz representation theorem there exist sequences $(h_k)_{k=1}^{\infty}$ and $(f^*_{j\,k})_{k=1}^{\infty}$, $j = 1, 2, \cdots, n$ such that the linear functionals

$$h^*[x] = \sum_{k=1}^{\infty} h_k x_k,$$

$$f_j^*[x] = \sum_{k} f^*_{j\,k} x_k,$$

and we have $\|x\| = (\sum x_k^2)^{\frac{1}{2}}$. As Hilbert space is self-dual, $\|f_j^*\| = (\sum_k f^{*\,2}_{j\,k})^{\frac{1}{2}}$.

Note that, as in example 1 of section 5.3, if the inner products $f_j^*[f_k^*]$ and $f_j^*[h^*]$ can be evaluated, there is no need to discretize the dual problem. In fact, there is no need to discretize the primal problem, as we shall see: It is intuitively clear on geometric grounds (the projection theorem) that the relevant subset of X for this problem is the subspace M spanned by the set $\{h^*, f_1^*, \cdots f_n^*\}$. Any component not in M makes $\|x\|$ larger, and does not affect $\|\mathbf{f}^*[x] - \delta\|$ nor $h^*[x]$.

Or consider this more rigorous justification: M is closed since it is a finite-dimensional subspace; therefore X can be written as the direct sum $\mathsf{M} + \mathsf{M}^{\perp}$,

where $\mathsf{M}^\perp$ is the orthogonal complement of M (see for example [35]). That is, any $x \in \mathsf{X}$ can be written uniquely as $x = y + z$, where $y \in \mathsf{M}$ and $z \in \mathsf{M}^\perp$. Consider any sequence $(x_j)_{j=1}^\infty$ of feasible points for the primal problem for which $h^*[x_j] \to v(\mathcal{P})$. By direct substitution one may verify that the sequence $(y_j)_{j=1}^\infty$ of components of the x_j in M is also feasible, and that $h^*[x_j] = h^*[y_j]$. Thus the value of the primal restricted to M is the same as for all X.)

Therefore, it suffices to consider $x \in \mathsf{M}$. Let's assume that the set $\{h^*, f_1^*, \cdots f_n^*\}$ is linearly independent, *i.e.* that M is $n+1$-dimensional. Rename $f_{n+1}^* \equiv h^*$. We may write $x \in \mathsf{M}$ uniquely as $x = \sum_{\jmath=1}^{n+1} \beta_j f_j^*$. We have

$$||x||^2 = \boldsymbol{\beta} \cdot \boldsymbol{\Gamma} \cdot \boldsymbol{\beta}$$

where $\boldsymbol{\Gamma}$ is the "augmented" Gram matrix with elements

$$\Gamma_{jk} \equiv f_j^*[f_k^*], \quad j,k = 1, \cdots, n+1.$$

Similarly, for $x \in \mathsf{M}$,

$$||\mathbf{f}^*[x] - \boldsymbol{\delta}||^2 = \boldsymbol{\beta} \cdot \boldsymbol{\Lambda} \cdot \boldsymbol{\beta} - 2\boldsymbol{\delta} \cdot {}^J\boldsymbol{\Gamma} \cdot \boldsymbol{\beta} + ||\boldsymbol{\delta}||^2,$$

where $\boldsymbol{\Lambda}$ is the $n+1$ by $n+1$ matrix

$$\boldsymbol{\Lambda} = {}^J\boldsymbol{\Gamma}^T {}^J\boldsymbol{\Gamma},$$

and ${}^J\boldsymbol{\Gamma}$ is the n by $n+1$ submatrix of $\boldsymbol{\Gamma}$ consisting of its first n rows (those corresponding to $\mathbf{f}^*$). (This can be generalized easily to $\boldsymbol{\Sigma} \neq I$ by the introduction of a weight matrix; the inner products needed are the same.) For $x \in \mathsf{M}$ we also have $h^*[x] = \boldsymbol{\gamma} \cdot \boldsymbol{\beta}$ where $\boldsymbol{\gamma}$ is the $n+1^{st}$ row of $\boldsymbol{\Gamma}$.

Thus the primal is equivalent to minimizing

$$\min_{\boldsymbol{\beta}} \boldsymbol{\gamma} \cdot \boldsymbol{\beta} : \boldsymbol{\beta} \cdot \boldsymbol{\Gamma} \cdot \boldsymbol{\beta} \leq v \text{ and } \boldsymbol{\beta} \cdot \boldsymbol{\Lambda} \cdot \boldsymbol{\beta} - 2\boldsymbol{\delta} \cdot {}^J\boldsymbol{\Gamma} \cdot \boldsymbol{\beta} + ||\boldsymbol{\delta}||^2 \leq \chi.$$

This can be achieved by means of two Lagrange multipliers, and implemented computationally by imposing $\boldsymbol{\gamma} \cdot \boldsymbol{\beta} = \zeta$ as a constraint, and minimizing $\mu||x||^2 + \lambda||\mathbf{f}^*[x] - \boldsymbol{\delta}||^2$. The argument of Stark and Parker [63] can be adapted to show that the minimum value of $\mu||x||^2 + \lambda||\mathbf{f}^*[x] - \boldsymbol{\delta}||^2$ is quasiconvex in ζ. One may show then that there exist ζ, $\mu \geq 0$, and $\lambda \geq 0$ such that

$$||\boldsymbol{\delta}||^2 + \min_{\boldsymbol{\beta}:\boldsymbol{\gamma}\cdot\boldsymbol{\beta}=\zeta} \lambda\left(\boldsymbol{\beta} \cdot \boldsymbol{\Lambda} \cdot \boldsymbol{\beta} - 2\boldsymbol{\delta} \cdot {}^J\boldsymbol{\Gamma} \cdot \boldsymbol{\beta}\right) + \mu\left(\boldsymbol{\beta} \cdot \boldsymbol{\Gamma} \cdot \boldsymbol{\beta}\right) = v(\mathcal{P})$$

i.e. the problem can be solved as a sequence of linearly constrained least-squares problems. See Lawson and Hanson [34] for details.

The Discrete Primal. Problems in separable Hilbert spaces are usually discretized by *truncation*—representing x, f_j^*, h^* by their first m components

$(x_1, x_2, \cdots, x_m)$, *etc.*. Let X_m denote the m-dimensional subspace of X where all components in the chosen basis other than the first m equal zero. Let P_m denote the operator that projects an element of X onto X_m:

$$P_m x \equiv (x_1, x_2, \cdots, x_m, 0, 0, \cdots),$$

and let P_{R^m} map X onto $\mathbf{R}^m$ similarly:

$$P_{R^m} x \equiv (x_1, x_2, \cdots, x_m).$$

Let P_m and P_{R^m} act componentwise on n-tuples of elements of X; *e.g.* $P_m \mathbf{f}^* = (P_m {f^*}_1, P_m {f^*}_2, \cdots, P_m {f^*}_n)$. Note that $f_j^*[P_m x] = (P_m f_j^*)[P_m x]$. Thus

$$\begin{aligned} \mathsf{D} \cap \mathsf{X}_m &= \{x \in \mathsf{X}_m : \mathbf{f}^*[x] - \delta \in \Xi\} \\ &= \{x \in \mathsf{X}_m : (P_m \mathbf{f}^*)[x] - \delta \in \Xi\} \end{aligned} \tag{73}$$

Let $\Phi = P_{R^m} \mathbf{f}^*$ be the n by m matrix with elements

$$\Phi_{jk} = {f^*}_{j\,k}, \quad j = 1, \cdots, n; k = 1, \cdots, m,$$

and let

$$\gamma = P_{R^m} h = ({h^*}_1, {h^*}_2, \cdots, {h^*}_m).$$

The discretized primal is to find

$$\begin{gathered} \inf_{\alpha \in \mathbf{R}^m} \quad \gamma \cdot \alpha, \\ \|\Phi \cdot \alpha - \delta\|_p^{\Sigma^{-1}} \leq \chi \\ \|\alpha\|_2 \leq \nu \end{gathered}$$

which for $p = 1, 2$, or ∞ can be solved using quadratic programming iteratively.

The Discrete Dual. We discretize the dual by truncation as well, again assuming that $\mathbf{f}^*[x']$ and $h^*[x']$ can be found analytically. If necessary, it is straightforward to account for their approximate computation too. As in section 7.3, the error $\mathcal{E}$ in the approximation of $\|\lambda \cdot \mathbf{f}^* - h^*\|$ by $\|\lambda \cdot \tilde{\mathbf{f}^*} - \tilde{h^*}\|$ is bounded by:

$$\mathcal{E} \leq \|\lambda\|_q^{\Sigma} \left\| (\|f_j^* - P_m f_j^*\|)_{j=1}^n \right\|_p^{\Sigma^{-1}} + \|h^* - P_m h^*\|.$$

Here $\mathbf{f}^* - P_m \mathbf{f}^*$ and $h^* - P_m h^*$ are *known*. Let $\varrho_f \equiv \left\| (\|f_j^* - P_m f_j^*\|)_{j=1}^n \right\|_p^{\Sigma^{-1}}$ and $\varrho_h \equiv \|h^* - P_m h^*\|$. Then 55 holds, and the value of the exact dual is at least

$$\max_{\lambda \in \mathbf{R}^n} \{\lambda \cdot (\delta - \mathbf{f}^*[x']) - (\chi + v\varrho_f)\|\lambda\|_q^{\Sigma} - \nu\|\lambda \cdot \Phi - \gamma\| - v\varrho_h\},$$

which is the unconstrained maximization of a concave functional and is a completely finite-dimensional problem.

8 Primal Approaches to Discretization Error

Backus [7] gives an intriguing treatment of discretization error for confidence intervals on a single linear functional of a model in a two-norm Hilbert-space ball, based upon adding a "systematic error" term that can be bounded using the Cauchy-Schwarz inequality and the prior constraint $||x_0|| \leq 1$. Backus advocates using as the discrete basis a subset of the singular functions of the mapping $\mathbf{f}^*$ restricted to the span of the data functionals ($\{x \in \mathsf{X} : x = \boldsymbol{\lambda} \cdot \mathbf{f}^*\}$). His technique for constructing confidence intervals can produce intervals that are longer when more basis functions are used; he proposes a method of choosing the number of functions optimally using the singular values in that decomposition.

Lang [31] proposes a very appealing method of bounding discretization error: solve the discrete *primal* problem again with an increased value of χ (analogous to the term $v\varrho_f$ in section 7.3). One of the attractions of this approach is that the same software can be used to bound $v(\mathcal{P})$ above and below, with the addition of a small amount of code to determine the necessary increase to χ. Lang works out how much χ must be increased for bounds on linear functionals in a model problem where X is the space of continuous functions on an interval, C is an infinity-norm ball ($0 \leq x(t) \leq u(t)$, where u is piecewise constant); and the linear functional is given by integration against a piecewise constant weight function. His basic idea can be extended to deal with prior balls in other norms, and to linear functionals not given by piecewise constant weights (this necessitates the introduction of another correction term analogous to $v\varrho_h$ in section 7.3). However, it does not appear to extend to cases where there is less prior information—*e.g.* if x_0 is not known to lie in a ball—and it does not work if the linear functionals $\mathbf{f}^*$ and h^* are not in the *normed* dual space.

The basic idea is to use the bound on the norm of x_0 to determine an increase to χ that guarantees the set $\tilde{\mathsf{C}} \cap \tilde{\mathsf{D}}$ is at least as "wide" in the direction normal to the hyperplane that represents h^* as is the set $\mathsf{C} \cap \mathsf{D}$. One wants to incorporate the errors induced by the approximations to x and $\mathbf{f}^*$ into the acceptable misfit. This can be done by considering the norm of the operator that is the difference $f^*_{j,\Delta}$ between f^*_j and its discrete approximation $\tilde{f}^*_j$: the error is bounded by

$$||f^*_{j,\Delta}||\,||x|| = \nu||f^*_{j,\Delta}||.$$

By making the discretization sufficiently fine, one can ensure that $||f^*_{j,\Delta}||$ is small. Then the contribution of the discretization to the misfit is bounded by $\nu||\boldsymbol{\varrho}_f||_p^{\boldsymbol{\Sigma}^{-1}}$, where the n-vector $\boldsymbol{\varrho}_f \equiv (||f^*_{j,\Delta}||)_{j=1}^n$.

Similarly, one needs to allow for possible errors in $h^*[x]$, which can be bounded in the same way using the norm of the difference between h^* and its discrete approximation. Denote the norm of this operator by ϱ_{h^*}. Then the value of the primal is bounded below by

$$\inf_{\boldsymbol{\alpha} \in \tilde{\mathsf{C}} \cap \tilde{\mathsf{D}}'} \tilde{h}^*[\boldsymbol{\alpha} \cdot \boldsymbol{x}] - \nu\varrho_{h^*}$$

where

$$\tilde{\mathsf{D}}' = \{\alpha : ||\delta - \tilde{\mathbf{f}}^*[\alpha \cdot \boldsymbol{x}]||_p^{\Sigma^{-1}} \leq \chi + \nu||\varrho_f||_p^{\Sigma^{-1}}\}.$$

In the one practical case the author tried to use this approach, the seismic problem in section 7.3.2, $\nu||\varrho_f||_p^{\Sigma^{-1}}$ was greater than the misfit of the zero model; *i.e.*, greater than the norm of the data vector δ. This is an indication that the discretization in the problem was too crude, but figure 1 shows that the discretization problem is not as severe as indicated by such a large value of $\nu||\varrho_f||_p^{\Sigma^{-1}}$. Thus this approach may not be sharp enough to give practical results in all problems where a bound on the norm of x_0 is available. A reason for this is that (for example, in L_p spaces with $p \neq \infty$) a model that achieves its norm bound in each little interval of discretization in general does not fit the data.

9 Bounds on a Single Linear Functional

So far, we have chosen the data space confidence region Ξ without regard to the functional H or the prior information C. Our procedure allows simultaneous confidence intervals for any number of functionals to be computed, but in principle, Ξ could be chosen optimally; for example, to give the smallest expected weighted sum of the lengths of the confidence intervals for a fixed finite collection of functionals. Let $\mathrm{S}(\alpha)$ denote the set of $1-\alpha$ confidence regions for ϵ:

$$\mathrm{S}(\alpha) \equiv \{\Xi \subset \mathbf{R}^n : P\{\epsilon \in \Xi\} \geq 1-\alpha\}.$$

Using a ball in the data space as the confidence region leads to wider confidence intervals than necessary to guarantee the nominal coverage probability if one is interested in only a finite number of linear functionals. Since we can never compute bounds on more than a finite number of functionals, it may be of primary interest to get the bounds on those functionals as tight as possible while maintaining their (simultaneous) coverage probability. Backus [7] argues that one should stick to a single functional at a time; I believe there are cases where it is natural to consider more than one. For example, one might be interested in a confidence region for the paleopole position from paleomagnetic data (R.L. Parker, personal communication). Since two numbers (a three-dimensional unit vector) are needed to specify the paleopole position, we need a two-dimensional confidence set. In that case, it makes more sense to look at the two numbers as a vector and find, say, the region on the surface of the sphere with smallest area, rather than to put separate confidence intervals on the components.

If one wished to construct a simultaneously valid envelope such as that in Figure 1, using one functional at a time and a set Ξ tailored to each functional, the coverage probability decreases to something like $1 - N\alpha_0$, where N is the number of functionals and α_0 is the nominal coverage probability for each functional. Clearly as N grows, α_0 must approach zero to guarantee simultaneous coverage probability $1-\alpha$; it is easier and probably more efficient to use the approach here, where a single Ξ is used for all the functionals. To get the tightest bounds on any particular collection of functionals, one needs to tailor the data confidence set Ξ to the functionals and the available prior information C.

As far as I know, the optimal choice of Ξ is an unsolved problem, even for a single linear functional. For example, even the following simpler statistical problem appears to be unsolved: Suppose we observe $\delta = \mu + \epsilon$, where $\epsilon \sim N(0,1)$, and $\mu \in [-\tau, \tau]$. For a fixed coverage probability $1-\alpha$, what procedure for assigning a confidence interval to μ from δ produces, for the least favorable $\mu \in [-\tau, \tau]$, the shortest confidence interval in expectation? One must be able to answer such a question before tackling the problem of optimal confidence regions for bounds on functionals. The approach taken implicitly in this paper is to use the confidence interval $[\delta - 1.96, \delta + 1.96] \cap [-\tau, \tau]$. It is easy to convince oneself that this is suboptimal, but further analysis is needed to determine just

how bad it is.

9.1 Confidence Slabs

In this section we examine a particular choice of Ξ that can give a tighter confidence interval for a single linear functional than do balls. The choice made here is the same as that used by Backus [7] and by Donoho [18], but the confidence intervals we get differ due to the way the information $x_0 \in \mathsf{C}$ is used. The set we shall pick is a "slab," the region between two parallel hyperplanes in data space. See also Stark [58]. One way to think of this region is in terms of projections of the data space $\mathbf{R}^n$ into $\mathbf{R}$. This amounts to taking a particular linear combination of the data functionals, say $\gamma \cdot \mathbf{f}^*$, and the corresponding datum $\gamma \cdot \delta$ and solving the resulting single-datum inverse problem. Thus slabs are a special case ($n = 1$) of the previous results. For $\mathbf{R}$, all the l_p norms are the same, $||\gamma \cdot \epsilon|| = |\gamma \cdot \epsilon|$. Using the knowledge of the distribution of ϵ, it is easy to compute the distribution of $\gamma \cdot \epsilon$. Let us sketch a motivation for this choice of data confidence regions, and develop the dual problem directly.

Suppose for the moment that X^* is normed. Define

$$\gamma \equiv \arg \inf_{\beta \in \mathbf{R}^n} ||H - \beta \cdot \mathbf{f}^*||,$$

$$\hat{\gamma} \equiv \frac{\gamma}{||\gamma||}, \tag{74}$$

$$g^* \equiv \hat{\gamma} \cdot \mathbf{f}^*, \tag{75}$$

and pick χ so that

$$P\{|\epsilon \cdot \hat{\gamma}| \leq \chi\} \geq 1 - \alpha.$$

(If X^* is not normed, pick any $\gamma \in \mathbf{R}^n$.) Let

$$\mathsf{D} \equiv \{x \in \mathsf{X} : |\delta \cdot \hat{\gamma} - g^*[x]| \leq \chi\}.$$

Then

$$P\{\mathsf{C} \cap \mathsf{D} \ni x_0\} = P\{\mathsf{D} \ni x_0\} \geq 1 - \alpha.$$

The resulting primal optimization problem to get the lower endpoint of a $1 - \alpha$ confidence interval for $H[x_0]$ is to find

$$v(\mathcal{P}) = \inf_{x \in \mathsf{C} \cap \mathsf{D}} H[x].$$

The dual functionals C^* and D^* for this problem are

$$\mathsf{D}^*[x^*] \equiv \inf_{x \in \mathsf{D}} x^*[x] = \begin{cases} \beta\delta \cdot \hat{\gamma} - \chi|\beta|, & x^* = \beta g^* \text{ for some } \beta \in \mathbf{R} \\ -\infty, & \text{otherwise,} \end{cases}$$

and

$$\mathsf{C}^*[x^*] \equiv \inf_{x \in \mathsf{C}} (H[x] - x^*[x]) = -v||H - x^*||,$$

so

$$v(\mathcal{D}) = \sup_{\beta \in \mathbf{R}} \{\beta\delta \cdot \hat{\gamma} - \chi|\beta| - v||H - \beta g^*||\}.$$

Note that this can be derived directly from the n-dimensional case by taking linear combinations as mentioned above, and setting $n = 1$. Section 11 discusses conditions for equality $v(\mathcal{P}) = c(\mathcal{D})$.

9.2 Hilbert-Spaces: Backus and Donoho

Suppose that X is a Hilbert space so that $\mathsf{X}^* = \mathsf{X}$. Then it is easy to see that $\gamma = \Gamma^{-1} \cdot \mathbf{f}^*[H]$, where Γ is the Gram matrix with elements

$$\Gamma_{ij} = < f_i^*, f_j^* > .$$

A bit of algebra yields

$$v(\mathcal{D}) = \sup_{\beta} \left\{ \beta\delta \cdot \Gamma^{-1} \cdot \mathbf{f}^*[H] - \chi|\beta| - v\left(||H||^2 + \beta(\beta - 2)\mathbf{f}^*[H] \cdot \Gamma^{-1} \cdot \mathbf{f}^*[H]\right)^{\frac{1}{2}} \right\}.$$

The expression can be arranged so that evaluating the dual functional does not involve inverting the Gram matrix, just solving the system of equations

$$\Gamma \cdot \gamma = \mathbf{f}^*[H]. \tag{76}$$

Now if the distribution of errors ϵ is symmetric, the value of χ needed for $1 - \alpha$ coverage probability is independent of the direction $\hat{\gamma}$—any unit vector gives the same value. Thus if we choose to solve 9.1 only approximately (say by a singular value truncation procedure, as advocated by Backus [7]) we still get a valid confidence interval for $H[x_0]$.

It is always possible to get a bound on the improvement in going from the ball in $\mathbf{R}^n$ to the slab by computing bounds with a $\mathbf{R}^n$ ball whose diameter is chosen on the basis of the χ_1^2 distribution rather than the χ_n^2 distribution. The proof is obvious: the chi-squared 1 ball is a subset of the slab. Let $H^{\|}$ denote the approximation to H in the linear span of the data functionals.

In our notation, the endpoints of Backus' confidence intervals are given by

$$\pm \inf_{x \in \mathsf{D}} \pm H^{\|}[x] \pm \inf_{x \in \mathsf{C}} \pm (H - H^{\|})[x].$$

Examine the lower endpoint (plus signs throughout):

$$\begin{aligned} \inf_{x \in \mathsf{D}} H^{\|}[x] + \inf_{x \in \mathsf{C}} (H - H^{\|})[x] &\leq \inf_{x \in \mathsf{C} \cap \mathsf{D}} H^{\|}[x] + \inf_{x \in \mathsf{C} \cap \mathsf{D}} (H - H^{\|})[x] \\ &\leq \inf_{x \in \mathsf{C} \cap \mathsf{D}} (H^{\|} + H - H^{\|})[x] \\ &= \inf_{x \in \mathsf{C} \cap \mathsf{D}} H[x] \end{aligned} \tag{77}$$

which is the lower endpoint our procedure gives. Similarly, Backus' upper endpoint is always larger, and so, at the same level of approximation used to find λ, his intervals are always longer, to obtain the same nominal coverage probability. However, Backus performs a discrete optimization over linear combinations of the singular functions of the Gram matrix to determine the approximation to λ that yields the shortest interval. With our approach, that optimization is impermissible, since for each λ, the width of the computed interval depends on the measured data. However, it *is* permissible to use Backus' or Donoho's (see below) *a priori* computation of the approximation that is optimal, then to use strict bounds to find the confidence interval with that approximation to λ. This will yield shorter intervals than either Backus' or Donoho's technique used by themselves.

By a different approach, based on the modulus of continuity of H, Donoho [18] shows how to construct optimally short (minimax) fixed-length confidence intervals based on affine estimates for linear functionals of models in l_2 constrained to lie in arbitrary convex subsets C of X when the errors ϵ are Gaussian. These intervals are shorter than those computed by Backus: they are optimal in a much larger class. Donoho also shows that these intervals based on affine estimates are nearly optimal among all measurable estimates, including nonlinear ones. For the present problem, where $\mathsf{C} = \{x \in l_2 : \|x\| \leq 1\}$, and H is linear, Donoho's approach leads to the computation:

$$\begin{aligned} \omega(\nu) &\equiv \sup\{|H[x_1] - H[x_{-1}| : \|\mathbf{f}^*[x_1] - \mathbf{f}^*[x_{-1}]\|_2 \leq \nu \text{ and } x_i \in \mathsf{C}\} \\ &= \sup\{H[x] : \|\mathbf{f}^*[x]\|_2 \leq \nu \text{ and } \|x\| \leq 2\} \end{aligned} \tag{78}$$

as one may show without difficulty. The optimization problem 78 is in fact a strict bounds problem with data misfit measured in the two-norm, a data vector identically equal to zero, and prior information that the unknown lies in a Hilbertian ball. Thus the development in section 7.3.3 and example 1 of section 5.3 is applicable to the computation; only $(n+1)(n+2)/2$ infinite-dimensional computations, $f_j^*[f_k^*]$ and $H[f_j^*]$, and $\|H\|^2$, are required.

Donoho's minimax affine confidence interval has length

$$\sup_{\nu \geq 0} \left(\frac{\omega(\nu)}{\nu} \right) \chi_\alpha(\nu/2, \sigma),$$

where $\chi_\alpha(\cdot, \cdot)$ is a universal function for such problems and must be evaluated numerically, and σ is the standard deviation of the (Gaussian) observational noise. Once the function χ_α is tabulated once and for all, Donoho's approach is computationally less taxing than Backus', as it requires no secondary optimization over truncation levels. Donoho and Stark [19] give a numerical comparison of Backus' and Donoho's confidence intervals for the geomagnetic problem, and show that the optimal truncation-based confidence interval in a problem of this sort is no more than twice as long as the minimax affine-based fixed-length confidence interval.

10 Systematic Errors

It is straightforward to incorporate the effects of known systematic errors into the confidence intervals constructed by strict bounds. In order to end up with solvable optimization problems, however, it is necessary that the systematic error sets be convex; otherwise the optimization problems are over nonconvex sets, local optima may not be global optima, and there may be duality gaps. Of course, the convex hull of the systematic error set may be used, but this can enlarge the confidence intervals substantially. The following treatment of systematic errors in the data, or of modelling errors, is similar to the treatment by Backus [7]. Backus' treatment of systematic errors in the evaluation of H carries over to our approach unchanged: one merely "tacks on" the maximum systematic error in the obvious way after evaluating the confidence interval.

We model the data relations with systematic errors as follows:

$$\boldsymbol{\delta} = \mathbf{f}^*[x_0] + \boldsymbol{\epsilon} + \boldsymbol{\omega}, \tag{79}$$

where $\boldsymbol{\epsilon}$ is random error as before, and $\boldsymbol{\omega} \in \boldsymbol{\Omega}$ is a systematic error drawn from the known set $\boldsymbol{\Omega}$. Clearly, we may get a $1-\alpha$ confidence region for $\boldsymbol{\delta}_0 \equiv \mathbf{f}^*[x_0]$ by translating any of our previous regions $\boldsymbol{\Xi}$ by the set $\boldsymbol{\Omega}$:

$$\wp\{\boldsymbol{\delta} - \boldsymbol{\Xi} - \boldsymbol{\Omega} \ni \mathbf{f}^*[x_0]\} \geq 1-\alpha\},$$

where

$$\boldsymbol{\delta} - \boldsymbol{\Xi} - \boldsymbol{\Omega} \equiv \{\boldsymbol{\gamma} \in \mathbf{R}^n : \boldsymbol{\gamma} = \boldsymbol{\delta} - \boldsymbol{\xi} - \boldsymbol{\omega} \text{ for some } \boldsymbol{\xi} \in \boldsymbol{\Xi} \text{ and some } \boldsymbol{\omega} \in \boldsymbol{\Omega}\}.$$

We may thus proceed as before if we define D to be

$$\mathsf{D} \equiv \{x \in \mathsf{X} : \mathbf{f}^*[x] \in \boldsymbol{\delta} - \boldsymbol{\Xi} - \boldsymbol{\Omega}\}.$$

If the set $\boldsymbol{\Omega}$ is a ball in the norm we are using to measure data misfit, the resulting optimization problem can be solved by the same techniques as before.

10.1 Example: Ball-Type Constraints.

Suppose

$$\boldsymbol{\Omega} = \{\boldsymbol{\omega} \in \mathbf{R}^n : \|\boldsymbol{\omega} - \boldsymbol{\omega}_0\|_{p'}^{\boldsymbol{\Sigma}'^{-1}} \leq \zeta\}$$

for some $\boldsymbol{\omega}_0 \in \mathbf{R}^n$ and $\zeta \geq 0$. For $p = p'$ and $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}'$ it is obvious that this additional uncertainty can be incorporated by setting

$$\mathsf{D} = \{x \in \mathsf{X} : \|\mathbf{f}^*[x] - \boldsymbol{\delta} - \boldsymbol{\omega}\|_p^{\boldsymbol{\Sigma}^{-1}} \leq \chi + \zeta,$$

and proceeding as before. If $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}'$ are both diagonal and $p = p'$, results are easily obtained.

Suppose $p \neq p'$. The following "quick and dirty" approach is straightforward and conservative, but often not sharp. Let $p'' = \max(p, p')$. Since the unit ball in l_p is contained in the unit ball in $l_{p''}$ for $p'' \geq p$, one may conservatively incorporate systematic errors by expanding $\boldsymbol{\Omega}$ or $\boldsymbol{\Xi}$ to the p''-norm, then using the same technique as before for $p = p'$.

Obviously, the situation is very simple if a slab is used as the confidence region, provided the systematic error set is an interval. When it is not, it is sometimes possible to do something with mixed integer linear programming, which can treat disjunctive constraints (see Papadimitriou and Steiglitz, [41]).

10.2 Uncertain Data Functionals: an example from Helioseismology

The mathematical model of an inverse problem we adopted in equation 4, where the data functionals f_j^* were assumed to be known, is too restrictive for some inverse problems. In many cases, including notably curved-ray travel-time tomography, we do not know precisely what functionals we have measured: we merely know that $\mathbf{f}^* \in \mathcal{F}$, with some probability, for a known class $\mathcal{F}$ of functionals. In curved-ray tomography this comes about because we do not know the raypaths. In other problems it can come about because we do not know precisely the characteristics of our measuring instruments. In some circumstances, this uncertainty in what was measured can be cast as a systematic error.

The following is an illustration from the asymptotic inversion of free-oscillation data; there is an analogous problem for the inversion of $\tau(p)$ and $X(p)$ data for Earth as well (see sections 7.2 and 7.3.2), although it has been ignored in the literature so far. The asymptotic inversion of free-oscillation data via Abel's equation was first done for Earth by Brodsky and Levshin (the first paper in English is [12]). The technique has also been used for the inversion of solar free-oscillation data by Christensen-Dalsgaard and others [14], and by Brodsky and Vorontsov [13]. The following development will implicitly be for Sun, because the additional constraint of monotonicity of soundspeed with depth does not hold for Earth, except arguably in the core [64].

In this section we abandon some of our previous notation for consistency with current usage. Let ω_{kl} be the eigenfrequency with radial number k and angular number l. We shall assume that we observe *triples* (δ_{kl}, k, l), where $\delta_{kl} = \omega_{kl} + \epsilon_{kl}$, and as usual ϵ_{kl} is observational error. It can be shown (see [12]) that

$$\frac{\pi(k+\alpha)}{\omega_{kl}} = \int_{r_t}^{R} \left(\frac{r^2}{c^2(r)} - \left(\frac{l+1/2}{\omega_{kl}} \right)^2 \right)^{1/2} r^{-1} dr + O(\omega_{kl}^{-2}). \qquad (80)$$

Here R is the solar radius, r is distance from the center of Sun, $c(r)$ is the soundspeed at radius r, α is a phase shift factor that captures the effect of the

boundary condition at $r = R$; α is assumed to be known (techniques for its estimation are the subject of work in progress with M. Brodsky, and uncertainties in α are easily incorporated into the subsequent analysis). Define

$$d_{kl} \equiv \frac{\pi(k+\alpha)}{\omega_{kl}} \tag{81}$$

$$p_{kl} \equiv \frac{(l+1/2)c(R)}{\omega_{kl}R} \tag{82}$$

$$x \equiv \frac{R\ln(\frac{R}{r})}{c(R)} \tag{83}$$

$$v(x) \equiv \frac{c(R\exp\{-xc(R)/R\})}{c(R)\exp\{-xc(R)/R\}}, \tag{84}$$

where the last two equations are the Earth-flattening transformations of Gerver and Marcushevitch [23]. If we substitute 81-84 into 80, we find to first order in ω_{kl}^{-1},

$$d_{kl} = \int_0^{x_t} \left(v^{-2} - p_{kl}^2\right)^{1/2} dx. \tag{85}$$

Under the assumed monotonicity of $c(r)/r$ (according to D.O. Gough, personal communication, this assumption is not usually disputed for Sun, and all current solar models satisfy it), we have monotonicity of $v(x)$ and hence can change variables to $x(v)$. (See Stark *et al.* [64] for a more rigorous justification.) Then

$$d_{kl} = \int_1^{p_{kl}^{-1}} \left(v^{-2} - p_{kl}^2\right)^{1/2} dx(v).$$

The reader might notice this is equation 50 with $p_j = p_{kl}$. However, in this case (as, in fact, for the travel-time problem) the kernel in the integral equation and the limit of integration are not known precisely—they depend upon ω_{kl}, while we observe only $\delta_{kl} = \omega_{kl} + \epsilon_{kl}$.

We will address this complication by finding a systematic error term to add to the data uncertainties. The basic idea is to look at how different the integrals

$$\mathcal{I}(p,x) \equiv \int_1^{p^{-1}} (v^{-2} - p^2)^{1/2} dx(v) \tag{86}$$

can be for p ranging throughout confidence intervals for each of the p_{kl} derived from the observed δ_{kl}, and x ranging over the set of *a priori* acceptable solutions. The possibility of bounding this error depends on the following constraint, which is generally accepted by the helioseismological community, but by no means certain (D.O. Gough, personal communication): $c(r)$ increases monotonically with depth in the sun, down to a radius of about $0.2R$. As in the case for $\tau(p)$ and $X(p)$ data mentioned in section 7.3.2, and in detail in [64], monotonicity of $c(r)$ translates into the upper bound $x(dv) \leq \frac{R}{v}dv$. We may then take $\mathsf{X} =$

$C^{1-}[1,b]$, where the space C^{1-} is defined in section 7.3.2, and b satisfies $b \geq \max_{(k,l)} p_{kl}^{-1}$.

Suppose we can find from the statistics of ϵ a set of numbers χ_{kl} such that

$$\wp\{\delta_{kl} - \chi_{kl} \leq \omega_{kl} \leq \delta_{kl} + \chi_{kl}\} \geq 1 - \beta,$$

for fixed $\beta \in [0,1]$. Let

$$\begin{aligned} d_{kl}^{-} &\equiv \frac{\pi(k+\alpha)}{\delta_{kl} + \chi_{kl}}, \\ d_{kl}^{+} &\equiv \frac{\pi(k+\alpha)}{\delta_{kl} - \chi_{kl}} \end{aligned}$$

and

$$\begin{aligned} p_{kl}^{-} &\equiv \frac{(l+1/2)c(R)}{(\delta_{kl} + \chi_{kl})R}, \\ p_{kl}^{+} &\equiv \frac{(l+1/2)c(R)}{(\delta_{kl} - \chi_{kl})R}. \end{aligned}$$

Then

$$\wp\left\{d_{kl}^{-} \leq d_{kl} \leq d_{kl}^{+}\right\} \geq 1 - \beta,$$

and

$$\wp\left\{p_{kl}^{-} \leq p_{kl} \leq p_{kl}^{+}\right\} \geq 1 - \beta.$$

Let us fix k and l and drop them from the notation. Let

$$\begin{aligned} \hat{d} &\equiv \frac{d^{+} + d^{-}}{2} \\ \hat{p} &\equiv \frac{d^{+} + d^{-}}{2} \end{aligned}$$

We wish to bound how bad an approximation $\mathcal{I}(\hat{p}; x_0)$ is to $\mathcal{I}(p; x_0)$, where x_0 is the true model and $p = p_{kl}$ is the true value.

With probability $1 - \beta$, $p \in [p^{-}, p^{+}]$, and hence with probability $1 - \beta$,

$$\begin{aligned} |\mathcal{I}(p; x_0) - \mathcal{I}(\hat{p}; x_0)| &\leq \sup_{q \in [p^{-}, p^{+}]} |\mathcal{I}(q; x_0) - \mathcal{I}(\hat{p}; x_0)| \\ &\leq \sup_{q, r \in [p^{-}, p^{+}]} |\mathcal{I}(q; x_0) - \mathcal{I}(\hat{p}; x_0)| \\ &\leq \sup_{x \in C} \sup_{q, r \in [p^{-}, p^{+}]} |\mathcal{I}(q; x) - \mathcal{I}(\hat{p}; x)| \\ &= R\left[p_{kl}^{+}\{\cos^{-1} p_{kl}^{+} - \tan(\cos^{-1} p_{kl}^{+})\} - \right. \\ &\quad \left. p_{kl}^{-}\{\cos^{-1} p_{kl}^{-} - \tan(\cos^{-1} p_{kl}^{-})\}\right] \\ &\equiv \sigma_{kl}, \end{aligned}$$

provided p^{-1} and $\hat{p}^{-1}$ are velocities that occur shallower than $0.3R$. Thus

$$\wp\{d_{kl}^- - \sigma_{kl} \le \mathcal{I}(\hat{p}_{kl}; x_0) \le d_{kl}^+ + \sigma_{kl}\} \ge 1 - \beta.$$

Now

$$\wp\{d_{kl}^- - \sigma_{kl} \le \mathcal{I}(\hat{p}_{kl}; x_0) \le d_{kl}^+ + \sigma_{kl}, \ \ \forall (k,l) \in K\} \ge 1 - |K|\beta,$$

where K is the set of all pairs (k,l) in the data set, and $|K|$ is the cardinality of K. If the errors ϵ_j are independent, $1 - |K|\beta$ can be replaced by $(1-\beta)^{|K|}$. Define β' to be $|K|\beta$ or $1-(1-\beta)^{|K|}$, as appropriate. Now let

$$\mathsf{D} \equiv \left\{x \in C^{1-}[1,b] : \gamma_{kl}^- \le \mathcal{I}(\hat{p}_{kl}; x) \le \gamma_{kl}^+, \ \forall (k,l) \in K\right\}.$$

Then $\wp\{\mathsf{D} \ni x_0\} \ge 1 - \beta'$. As in section 7.3.2,

$$\mathsf{C} \equiv \left\{x \in C^{1-}[1,b] : 0 \le \frac{dv}{dv} \le \frac{R}{v}\right\}.$$

As usual, $\wp\{\mathsf{C} \cap \mathsf{D} \ni x_0\} \ge 1 - \beta'$, so we may use this set for strict bounds with coverage probability at least $1 - \beta'$: we have succeeded in capturing the uncertainty in the data functionals.

11 Duality Gaps

If there is a duality gap, that is if $v(\mathcal{P}) > v(\mathcal{D})$, then the bound we get by solving the dual problem is still valid, but not tight. We can not "squeeze in" on the value of $\mathcal{P}$ indefinitely using the discretized dual problem—the discretized dual always gives pessimistic results. It is thus of interest to determine when there might be a duality gap, to know if a large difference between the values of the discretized primal and dual problems indicates that the true bound on H may be much smaller than $v(\tilde{\mathcal{P}})$. One heuristic check one may perform is to solve the discretized dual problem without the correction terms for discretization, and see if the discrete primal and dual results then overlap. Here we pursue a set of rigorous sufficient conditions.

The fundamental result for the algebraic approach to conjugate duality we have taken here is Fenchel's Duality Theorem (see Luenberger [35]), which gives the following sufficient conditions for the absence of a duality gap:

1. $\mathsf{C} \cap \mathsf{D}$ has points in the relative interior of C and D,
2. the epigraph of H has nonempty interior, or D has nonempty interior, and
3. $v(\mathcal{P})$ is finite

We obviously need a topology on X to apply the Theorem. We would like to be able to treat cases where no topology is available, for instance when all we know is that x_0 is in a cone. Fortunately, there is another approach through Lagrangian duality (see Luenberger [35] for the general theory, Stark [58] for application to this problem.)

The sufficient conditions for the absence of a duality gap in Lagrangian duality require no topology on X. What makes Lagrangian duality fly here is that we determine membership in D via a mapping into $\mathbf{R}$, with positive cone $\mathbf{R}^+$, such that $x \in \mathsf{D}$ if and only if some convex functional G^* of x is less than or equal to 0. In most of the paper, the functional

$$G^*[x] \equiv ||\mathbf{f}^*[x] - \delta||_p^{\Sigma^{-1}} - \chi.$$

Now Lagrangian duality ([35], Ch. 8.6, Th. 1) says, in our language, that if C is convex, H is convex, $v(\mathcal{P})$ is finite, and there exists $x \in \mathsf{C}$ with $G^*[x] < 0$, then

$$v(\mathcal{P})) = v(\mathcal{D}_L) \equiv \max_{\gamma \geq 0} \inf_{x \in \mathsf{C}} \{H[x] + \gamma G^*[x]\}. \tag{87}$$

We shall see momentarily that the expression on the right hand side of 87 is equivalent to the conjugate dual derived algebraically in section 4. We remark that while Lagrangian duality is useful here to establish the absence of duality gaps, the actual computation of the value of the dual is easier in the original conjugate dual formulation, and the algebraic development of the conjugate dual in the first place is more transparent.

The conjugate dual is (from 35)

$$\begin{aligned} v(\mathcal{D}) &= \sup_{\boldsymbol{\lambda}\in\mathbf{R}^n} \{\boldsymbol{\lambda}\cdot\boldsymbol{\delta} - \chi||\boldsymbol{\lambda}||_q^{\boldsymbol{\Sigma}} + \inf_{x\in\mathsf{C}}(H[x] - \boldsymbol{\lambda}\cdot\mathbf{f}^*[x])\} \\ &= \sup_{\boldsymbol{\lambda}}\left\{-\chi||\boldsymbol{\lambda}||_q^{\boldsymbol{\Sigma}} + \inf_{x\in\mathsf{C}}(H[x] - \boldsymbol{\lambda}\cdot(\mathbf{f}^*[x] - \boldsymbol{\delta}))\right\} \\ &\geq \sup_{\boldsymbol{\lambda}}\left\{-\chi||\boldsymbol{\lambda}||_q^{\boldsymbol{\Sigma}} + \inf_{x\in\mathsf{C}}\left(H[x] - ||\boldsymbol{\lambda}||_q^{\boldsymbol{\Sigma}}||\mathbf{f}^*[x] - \boldsymbol{\delta}||_p^{\boldsymbol{\Sigma}^{-1}}\right)\right\} \\ &= \sup_{\alpha\geq 0}\left\{-\alpha\chi + \inf_{x\in\mathsf{C}}\left(H[x] - \alpha||\mathbf{f}^*[x] - \boldsymbol{\delta}||_p^{\boldsymbol{\Sigma}^{-1}}\right)\right\} \end{aligned} \tag{88}$$

which is the Lagrangian dual 87 defined above. Now since $v(\mathcal{P}) \geq v(\mathcal{D}) \geq v(\mathcal{D}_L)$, where $\mathcal{D}_L$ is the Lagrangian dual, it follows from 88 that if there is no duality gap for the Lagrangian dual, *i.e.* if $v(\mathcal{P}) = v(\mathcal{D}_L)$, there can be no gap for the conjugate dual $\mathcal{D}$. □

Thus the absence of duality gaps depends on

1. the primal $\mathcal{P}$ having finite value $v(\mathcal{P})$, and
2. the existence of an $x_1 \in \mathsf{C}$ with $G^*[x_1] < 0$.

The second condition is easier to discuss, so let's begin with it. If there is a feasible point for the discretized primal, which can be established computationally by numerical programming, then either the second item is satisfied, or an infinitesimal increase to χ will satisfy it. Since the value of χ is estimated from suppositions about the distribution of ϵ, one might well be disconcerted if the particular value of χ chosen happens to be the smallest for which the data can be fit, and might not hesitate to increase χ a smidgen. The second condition can be checked *a priori*, but it depends on the prior information C, and the nature of the functional H. We explore the various cases treated in the paper:

11.1 No Prior Information, H linear

We have seen in the case of bounds on linear functionals with $\mathsf{C} = \mathsf{X}$, no prior information, there is no duality gap provided the data functionals are linearly independent in X^*. This result is purely algebraic.

11.2 Prior Convex Cone, H linear

Let K^* be the *conjugate cone* (also called the *polar cone*) of C:

$$\mathsf{K}^* = \{x^* \in \mathsf{X}^* : \forall x \in \mathsf{C}, x^*[x] \geq 0\},$$

and let F^* be the span of the f_j^*. Then, for $\boldsymbol{\Xi}$ a ball, $v(\mathcal{P})$ is finite provided $H \in \mathsf{F}^* + \mathsf{K}^*$, and so there will be no duality gap provided (2) is satisfied.

11.3 Prior Ball, H linear

Clearly, if H is a bounded linear functional (the only case we consider when C is a ball), $v(\mathcal{P})$ is finite if the problem is feasible at all, since $H[x] \geq \|H\| v + H[x']$ for $x \in \mathsf{C}$.

11.4 No Prior Information, H a norm

For H a norm, $\mathsf{C} = \mathsf{X}$, and linearly independent f_j^* in X^*, $v(\mathcal{P})$ is finite: it is bounded below by zero, and above by a finite value depending on the norms of the individual f_j^*'s and δ.

For H a seminorm, we may pose the problem in X/M, where $|\cdot|$ is a norm and then use the previous result. There is a (surmountable) difficulty if there exists $y \in \mathsf{M}$ with $f_j^*[y] \neq 0$ for some j.

Thus under these reasonable regularity conditions, which often appear to obtain, duality gaps are not in practice a problem, and the discrete-primal/discrete-dual bracketing procedure should allow the true value $v(\mathcal{P})$ to be "squeezed" as accurately as desired.

12 Discussion

The discretized dual with correction terms provides a method for rigorous inferences about certain functionals in some constrained infinite-dimensional linear inverse problems. In conjunction with the discretized dual problem, or by itself without correction terms when there is known to be no duality gap, it enables a true $1-\alpha$ confidence interval for the functional to be bracketed computationally.

The dual approach is probably attractive computationally as long as the number of data is relatively small (compared with the number of model elements) and the data functionals are reasonably well behaved (in contrast to the $X(p)$ functionals in section 7.3.2). In whole Earth travel-time tomography, where the data number millions, current computer technology would require one to chose a subset of the data kernels to work with to limit the dimensionality of $\boldsymbol{\lambda}$ to a reasonable figure. It is probably not too difficult to determine *a priori* a good subset to pick to bound a fixed *linear* functional: one wants kernels that are most nearly aligned with H so that a good approximation to H is possible within their span. One is probably not going to do too badly by picking the subset so that the support of the functions that represent those components of $\mathbf{f}^*$ and H roughly match.

I expect that is possible to treat more general convex functionals via the Hahn-Banach Theorem [35]. The ideas are similar to the development for minimum-norm problems, but the set $\mathbf{C}^*$ has a more complicated description than $||x^*|| \leq 1$. Donoho [17] gives situations in which lower bounds on nonlinear functionals of a probability density can be computed. His work could likely be generalized to inverse problems. It also should be possible to treat more general convex constraints than L_p balls, and to get a bound on problems with nonconvex constraints by using the convex hull of the constraint set.

In the primal problem, it is often expensive to incorporate constraints; when is additional knowledge likely to shorten the confidence intervals significantly? This question is not explored here, but it seems that the dual problem is the natural vehicle for estimating *a priori* the power of *a priori* constraints.

Relation to Backus' work and Donoho's work. The approach to constructing confidence intervals taken here is fundamentally different from that of Backus [7], and Donoho [18]. Both Backus' and Donoho's procedures produce fixed-length confidence intervals—the length of the confidence interval depends on the data functionals, the functional to be bounded, assumptions about the noise, and so on, but not on the measured data values. The measured data merely determine the center of the interval. Backus' procedure does not in general produce the shortest fixed-length confidence interval; among procedures based on affine transformations of the data, Donoho's procedure is optimal. Furthermore, he shows that the resulting intervals are within a small fraction of optimal among *all* procedures for producing fixed-length intervals. With, for example, Gaussian errors, both procedures have positive probability of produc-

ing confidence intervals that do not intersect $H[\mathsf{C}]$, the set of feasible values of H.

The approach taken here gives data-dependent confidence intervals—not only the location, but also the length of the confidence intervals depend on the measured data. With this approach, one can get shorter intervals than with the data-dependent approach (although they can be longer), and the confidence intervals include only values of H that could arise from elements of C. Again, as mentioned in section 9, one way to ensure that they are *always* shorter is to use as the data confidence region Ξ the slab geven by Backus' or Donoho's analysis, then use the techniques here to bound h^*. The problem of finding the shortest (in expectation) data-dependent confidence interval for a restricted parameter appears to be unsolved.

Another fundamental difference is that both of these approaches are geared to finding confidence intervals for a single linear functional. In contrast, the method described here can produce simultaneously valid confidence intervals for any number of functionals, including some nonlinear functionals. Both Backus' and Donoho's techniques are developed only for models in l_2, in contrast to the present work, which does not even require that X be a topological space. However, Donoho's technique allows C to be an arbitrary convex subset of X, while this paper discusses only balls (in arbitrary norms) and convex cones.

A Numerical Computations for Seismic Bounds

This appendix covers the computations needed in section 7.3.2.

A.1 Computation of $h^*[x']$ and $\mathbf{f}^*[x']$

These can be done analytically:

$$h^*[x'] = \int_a^{v_h} \frac{R}{2v} dv = \frac{R}{2} \ln\left(\frac{v_h}{a}\right). \tag{89}$$

Let

$$\begin{aligned} \mathcal{I}_j(v) &\equiv \int f_j(v) \frac{R}{2v} dv \\ &= \begin{cases} -R\left(p_j \sin^{-1}(p_j v) + (v^{-2} - p_j^2)^{1/2}\right), & 1 \le j \le n_\tau \\ R \sin^{-1}(p_j v), & n_\tau < j \le n, \end{cases} \end{aligned} \tag{90}$$

for $v \le p_j^{-1}$, and 0 otherwise; and where the $f_j(v)$ are defined in equations 62 and 63. Then

$$\begin{aligned} f_j^*[x'] &= \mathcal{I}_j(p_j^{-1}) - \mathcal{I}_j(a) \\ &= \begin{cases} R\left[(a^{-2} - p_j^2)^{1/2} + p_j\left(\sin^{-1}(ap_j) - \frac{\pi}{2}\right)\right], & 1 \le j \le n_\tau \\ R\left(\frac{pi}{2} - \sin^{-1}(ap_j)\right), & n_\tau < j \le n \end{cases} \end{aligned} \tag{91}$$

A.2 Computation of f_j^i and h^i

These computations will also be needed below to compute ϱ_f. The functions $f_j(v)$ are defined in equations 62 and 63.

$$\begin{aligned} \mathcal{T}_j(v) &\equiv \int f_j(v) dv \\ &= \begin{cases} 2\left[(1 - v^2 p_j^2)^{1/2} - \ln\left((vp_j)^{-1} + ((vp_j)^{-2} - 1)^{1/2}\right)\right], & 1 \le j \le n_\tau \\ -2(p_j^{-2} - v^2)^{1/2}, & n_\tau < j \le n, \end{cases} \end{aligned} \tag{92}$$

for $v \le p_j^{-1}$, and 0 otherwise. Let

$$v_{i,j} \equiv \min\{v_i, p_j^{-1}\}.$$

Then

$$f_j^i \equiv \frac{1}{\Delta_i} \int_{\Delta_i} f_j(v) dv = \frac{1}{v_{i+1} - v_i} \left(\mathcal{T}_j(v_{i+1,j}) - \mathcal{T}_j(v_{i,j})\right).$$

We have trivially

$$\begin{aligned} h^i &= \frac{1}{\Delta_i} \int_{\Delta_i} 1_{v \le v_h} dv \\ &= \begin{cases} 1, & v_i < v_h \\ 0, & \text{otherwise} \end{cases} \end{aligned}$$

since we have included v_h in the discretization.

A.3 Computation of ϱ_f

As figure 3 shows, the $f_j(v)$ are nonnegative and monotonic, which is helpful: the integrals $\mathcal{I}_j^i$ in equation 64 can each be broken into integrals on two intervals, one where $f_j \le f_j^i$ and one where $f_j > f_j^i$. The endpoint of these intervals in the interior of Δ_i can be found explicitly since the monotonicity of the f_j guarantees that they can be inverted. Let

$$\hat{w}_{i,j} \equiv \begin{cases} \left(\left(\frac{f_j^i}{2}\right)^2 + p_j^2\right)^{-1/2}, & 1 \le j \le n_\tau \\ p_j^{-1}\left(\left(\frac{2}{f_j^i}\right)^2 + 1\right)^{-1/2}, & n_\tau < j \le n. \end{cases} \tag{93}$$

Then $f_j(\hat{w}_{i,j}) = f_j^i$. Define $w_{i,j} = \min(\hat{w}_{i,j}, p_j^{-1})$. Recalling equation 90 we can now find the integrals $\mathcal{I}_j^i$ of equation 64.

For $1 \le j \le n_\tau$, the $\tau(p)$ kernels are monotonic decreasing, so

$$\begin{aligned} \mathcal{I}_j^i &= \int_{\Delta_i} |f_j - f_j^i| \frac{R}{2v} dv \\ &= \int_{v_i}^{w_{i,j}} (f_j - f_j^i) \frac{R}{2v} dv + \int_{w_{i,j}}^{v_{i+1}} (f_j^i - f_j) \frac{R}{2v} dv \\ &= \int_{v_{i,j}}^{w_{i,j}} f_j \frac{R}{2v} dv - f_j^i \int_{v_i}^{w_{i,j}} \frac{R}{2v} dv \\ &\quad + f_j^i \int_{w_{i,j}}^{v_{i+1}} \frac{R}{2v} dv - \int_{w_{i,j}}^{v_{i+1,j}} f_j \frac{R}{2v} dv \\ &= 2\mathcal{I}_j(w_{i,j}) - \mathcal{I}_j(v_{i+1,j}) - \mathcal{I}_j(v_{i,j}) + f_j^i \frac{R}{2} \ln\left(\frac{v_{i+1} v_i}{w_{i,j}^2}\right). \end{aligned} \tag{94}$$

For $n_\tau < j \le n$, the $X(p)$ kernels are monotonic increasing, so

$$\begin{aligned} \mathcal{I}_j^i &= -\int_{v_{i,j}}^{w_{i,j}} f_j \frac{R}{2v} dv + f_j^i \int_{v_i}^{w_{i,j}} \frac{R}{2v} dv \\ &\quad - f_j^i \int_{w_{i,j}}^{v_{i+1}} \frac{R}{2v} dv + \int_{w_{i,j}}^{v_{i+1,j}} f_j \frac{R}{2v} dv \\ &= -2\mathcal{I}_j(w_{i,j}) + \mathcal{I}_j(v_{i+1,j}) + \mathcal{I}_j(v_{i,j}) - f_j^i \frac{R}{2} \ln\left(\frac{v_{i+1} v_i}{w_{i,j}^2}\right). \end{aligned} \tag{95}$$

Then we can compute $\mathcal{I}_j = \sum_{i=1}^m \mathcal{I}_j^i$, and

$$\varrho_f = \|(\mathcal{I}_j)_{j=1}^n\|_p^{\Sigma^{-1}}.$$

A.4 Numerical Solution of the Dual

Define the functional

$$\mu_{\boldsymbol{\gamma}}(\boldsymbol{\lambda}) \equiv -\boldsymbol{\lambda} \cdot \boldsymbol{\delta}' + \chi' \|\boldsymbol{\lambda}\|_2^{\boldsymbol{\Sigma}^{-1}} + \|\boldsymbol{\lambda} \cdot \boldsymbol{\Phi} - \boldsymbol{\gamma}\|_1,$$

where $\boldsymbol{\Phi}$, $\boldsymbol{\gamma}$, $\boldsymbol{\delta}'$ and χ' and are defined in equations 68, 69, 70 and 71. To find the minimum of $h^*[x]$ over $\mathsf{C} \cap \mathsf{D}$ we compute

$$h^*[x'] - \inf_{\boldsymbol{\lambda}} \mu_{\boldsymbol{\gamma}}(\boldsymbol{\lambda}),$$

and to find the maximum of $h^*[x]$ we compute

$$h^*[x'] + \inf_{\boldsymbol{\lambda}} \mu_{-\boldsymbol{\gamma}}(\boldsymbol{\lambda}).$$

There is obviously positive scalar homogeneity for the optimization of linear functionals; to enhance numerical stability in the computations, h^* was normalized by the ratio of its norm to the norm of the largest data functional.

The actual numerical solution of the optimization problem is nontrivial, as the objective functional, while convex and hence subdifferentiable, is not differentiable everywhere. Both the two-norm piece and the one-norm piece fail to have derivatives on a set of zero measure. Here is a subgradient of $\mu(\boldsymbol{\lambda})$:

$$\partial_{\boldsymbol{\lambda}} \mu_{\boldsymbol{\gamma}}(\boldsymbol{\lambda}) \equiv \begin{cases} -\boldsymbol{\lambda} + \chi' \frac{\boldsymbol{\lambda} \cdot \boldsymbol{\Sigma}}{\|\boldsymbol{\lambda}\|_2^{\boldsymbol{\Sigma}}} + \sum_{i=1}^{m-1} \boldsymbol{\Phi}_i \mathrm{sgn}(\boldsymbol{\lambda} \cdot \boldsymbol{\Phi}_i - \gamma_i), & \|\boldsymbol{\lambda}\|_2^{\boldsymbol{\Sigma}} \neq 0 \\ -\boldsymbol{\lambda} + \sum_{i=1}^{m-1} \boldsymbol{\Phi}_i \mathrm{sgn}(\boldsymbol{\lambda} \cdot \boldsymbol{\Phi}_i - \gamma_i), & \text{otherwise,} \end{cases}$$

where $\boldsymbol{\Phi}_i$ is the i^{th} column of $\boldsymbol{\Phi}$, and

$$\mathrm{sgn}\alpha = \begin{cases} 1, & \alpha > 0 \\ 0, & \alpha = 0 \\ -1, & \alpha < 0 \end{cases}.$$

The lack of smoothness in this problem is sufficient to disrupt algorithms that assume two continuous derivatives, for example Stanford Systems Optimization Laboratories NPSOL. Shor [55] and Fletcher [20] give examples showing that steepest descent methods can converge to nonstationary points for nonsmooth convex functions.

The 178 optimization problems for the 89 upper and lower bounds on velocities were solved by a two step procedure. In the first step, the Nelder-Mead simplex algorithm (see, *e.g.* [20, 48]) was used to find an initial solution. This method is inefficient but robust, requiring from 8,000 to 40,000 function evaluations to find each minimum (*i.e.* terminate); each function evaluation involves a matrix multiply with dimension about 20,000 by 30, and 178 functionals must be optimized, so this is rather expensive. After the Nelder-Mead algorithm

found a reasonably good point, Stanford Systems Optimization Laboratory algorithm NPSOL was used to improve it, supplied with the subgradient in place of the gradient. I found empirically that the algorithm could decrease the objective value when a sufficiently good starting point was used. This may indicate that the objective functional is differentiable in a neighborhood of the optimal solution. Typically, NPSOL required 200-400 major iterations, and about 500 function and gradient evaluations to "converge," *i.e.* terminate. In the event the resulting bounds appeared too pessimistic (for example, if the bounds were non-monotonic), the Nelder-Mead algorithm was restarted from the best point found by NPSOL, and then NPSOL was run again. It is clear on reflection that the actual *bounds themselves* must be monotonic (since each point on the optimal bounds is contained in some monotonic model fitting the data), so this is a diagnostic for sharp solutions.

In addition, I tried the subgradient algorithm (see Shor [55]), with step size $h_k = constant * (.999)^k$. This choice of step length is illegitimate from a theoretical point of view as $\sum_k h_k < \infty$, but substantial numerical experience suggests it works well (N.Z. Shor, personal communication). (For that matter, unless the algorithm finds a point where the subgradient vanishes, no finite number of steps is sufficient.) To try to ensure that the method descends at each step, a bisection was performed up to 31 times in each step to find a point at which the value of the functional was less than the current value. My experience with the generalized gradient algorithm was that it was even slower than the two-stage procedure described above. I am hopeful that the ellipsoid algorithm [41] or the method of space-dilation in the direction of the gradient (SDG) [55] would be more efficient.

The computations were coded in Fortran77 and performed on a set of 5 Sun SPARCstation 1's, and a Cray X-MP/14. The velocity interval was discretized increasingly finely until $\varrho_f \approx 2.29$, about 4% of the value $\chi = 59.7$. That required 19118 points in the interval $[a, b]$; 15000 equally spaced in slowness (reciprocal velocity) from a to the reciprocal of the smallest p_j in the $X(p)$ data, and 4000 equally spaced in slowness from there to b; plus samples at $1/p_j$ for each j, and samples at each velocity v_h. The value $\chi = 59.7$ is appropriate for 99.9% confidence ($\alpha = 0.001$) for the two-norm of 30 independent, zero-mean, unit-variance Gaussian random variables. Stark and Parker [63] show empirically that the primal results are not sensitive to the precise value of χ to within about 10-20% at that level, so a coarser discretization could probably have been used without appreciably changing the results.

References

[1] N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68:337–404, 1950.

[2] G. Backus. Inference from inadequate and inaccurate data, I. *Proc. Natl. Acad. Sci.*, 65:1–7, 1970.

[3] G. Backus. Inference from inadequate and inaccurate data, II. *Proc. Natl. Acad. Sci.*, 65:281–287, 1970.

[4] G. Backus. Inference from inadequate and inaccurate data, III. *Proc. Natl. Acad. Sci.*, 67:282–289, 1970.

[5] G. Backus and F. Gilbert. The resolving power of gross Earth data. *Geophys. J. Roy. astr. Soc.*, 16:169–205, 1968.

[6] G.E. Backus. Comparing hard and soft prior bounds in geophysical inverse problems. *Geophys. J.*, 94:249–261, 1988.

[7] G.E. Backus. Confidence set inference with a prior quadratic bound. *Geophys. J.*, 97:119–150, 1989.

[8] G.E. Backus and F. Gilbert. Uniqueness in the inversion of inaccurate gross Earth data. *Phil. Trans. Roy. Soc. Lon. A.*, 266:123–192, 1970.

[9] R. Bayer and M. Cuer. Inversion tri-dimensionnelle des données aéromagnétiques sur le massif volcanique du Mont-Dore: Implications structurales et géothermiques. *Ann. Géophys.*, 37:347–365, 1981.

[10] E.N. Bessonova, V.M. Fishman, V.Z. Ryaboyi, and G.A. Sitnikova. The tau method for inversion of travel times, I. Deep seismic sounding data. *Geophys. J. Roy. astr. Soc.*, 36:377–398, 1974.

[11] E.N. Bessonova, V.M. Fishman, M.G. Shnirman, G.A. Sitnikova, and L.R. Johnson. The tau method for inversion of travel times, II. Earthquake data. *Geophys. J. Roy. astr. Soc.*, 46:87–108, 1976.

[12] Michael Brodskii and Anatoly Levshin. An asymptotic approach to the inversion of free oscillation data. *Geophys. J. Roy. astr. Soc.*, 58:631–654, 1979.

[13] M.A. Brodsky and S.V. Vorontsov. On the technique of the inversion of helioseismological data. In J. Christensen-Dalsgaard and S. Frandsen, editors, *Advances in Helio- and Asteroseismology*, pages 137–140, IAU, 1988.

[14] J. Christensen-Dalsgaard, T.L. Duvall, D.O. Gough, J.W. Harvey, and E.J. Rhodes Jr. Speed of sound in the solar interior. *Nature*, 378–382, 1985.

[15] J.E. Cope and B.W. Rust. Bounds on solutions of linear systems with inaccurate data. *SIAM J. Numer. Anal.*, 16:950–963, 1979.

[16] M. Cuer. *Des Questions bien posees dans des Problemes Inverses de Gravimetrie et Geomagnetisme: une Nouvelle Application de la Programmation Lineare.* PhD thesis, Académie de Montpellier, Université des Sciences et Techniques du Languedoc, 1984.

[17] D.L. Dohoho. Nonparametric inference about functionals of a density, I. *Technical Report, Department of Statistics, UC Berkeley*, 1984.

[18] D.L. Donoho. Statistical estimation and optimal recovery. *Technical Report No. 214, Department of Statistics, UC Berkeley*, 1989.

[19] D.L. Donoho and P.B. Stark. Minimax confidence intervals in geomagnetism. *Technical Report, Department of Statistics, UC Berkeley*, 1990.

[20] R. Fletcher. *Practical Methods of Optimization.* John Wiley and Sons, 1987.

[21] J.D. Garmany. On the inversion of travel times. *Geophys. Res. Lett.*, 6:277–279, 1979.

[22] J.D. Garmany, J.A. Orcutt, and R.L. Parker. Travel-time inversion: a geometrical approach. *J. Geophys. Res.*, 84:3615–3622, 1979.

[23] M.L. Gerver and V.M. Markushevich. On the characteristic properties of travel-time curves. *Geophys. J. Roy. astr. Soc.*, 13:241–246, 1967.

[24] J.R. Grasso, M. Cuer, and G. Pascal. Use of two inverse techniques. Application to a local structure in the New Hebrides island arc. *Geophys. J. Roy. astr. Soc.*, 75:437–472, 1983.

[25] R. Hettich, editor. *Semi-Infinite Programming*, Springer-Verlag, New York, 1979.

[26] Hui Hu. *Semi-Infinite Programming.* PhD thesis, Stanford University, 1988.

[27] S.P. Huestis. Uniform norm minimization in three dimensions. *Geophysics*, 51:1141–1145, 1986.

[28] S.P. Huestis and M.E. Ander. IDB2—a Fortran program for computing extremal bounds in gravity data interpretation. *Geophysics*, 48:999–1010, 1983.

[29] S.P. Huestis and R.L. Parker. Bounding the thickness of the oceanic magnetized layer. *J. Geophys. Res.*, 82:5293–5303, 1977.

[30] L. R. Johnson and R. C. Lee. Extremal bounds on the p velocity in the earth's core. *Bull. Seismol. Soc. Am.*, 75:115–130, 1985.

[31] S.W. Lang. Bounds from noisy linear measurements. *IEEE Trans. Info. Th.*, IT-31:498–508, 1985.

[32] S.W. Lang and T.L. Marzetta. Bounds from noisy linear measurements. In H.W. Schüssler, editor, *Signal Processing II: Theories and Applications*, pages 491–494, Elsevier Science Publishers B.V. (North-Holland), 1983.

[33] S.W. Lang and T.L. Marzetta. A linear programming approach to bounding spectral power. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 847–850, IEEE, IEEE, April 1983.

[34] C.W. Lawson and R.J. Hanson. *Solving Least Squares Problems.* John Wiley and Sons, Inc., New York, 1974.

[35] D.G. Luenberger. *Optimization by Vector Space Methods.* John Wiley and Sons, Inc., New York, 1969.

[36] T.L. Marzetta and S.W. Lang. Power spectral density bounds. *IEEE Trans. Info. Th.*, IT-30:117–122, 1983.

[37] M. McNutt and L. Royden. Extremal bounds on geotherms in eroding mountain belts from metamorphic pressure-temperature conditions. *Geophys. J. Roy. astr. Soc.*, 88:81–95, 1987.

[38] D.W. Oldenburg. Funnel functions in linear and nonlinear appraisal. *J. Geophys. Res.*, 88:7387–7398, 1983.

[39] J.A. Orcutt. Joint linear, extremal inversion of seismic kinematic data. *J. Geophys. Res.*, 85:2649–2660, 1980.

[40] J.A. Orcutt, R.L. Parker, P.B. Stark, and J.D. Garmany. Comment concerning 'A method of obtaining a velocity-depth envelope from wide-angle seismic data' by R. Mithal and J.B. Diebold. *Geophysical Journal*, 95:209–212, 1988.

[41] C.H. Papadimitriou and K. Steiglitz. *Combinatorial Optimization: Algorithms and Complexity.* Prentice-Hall, Inc., New Jersey, 1982.

[42] R.L. Parker. Best bounds on density and depth from gravity data. *Geophysics*, 39:644–649, 1974.

[43] R.L. Parker. A statistical theory of seamount magnetism. *J. Geophys. Res.*, 93:3105–3115, 1988.

[44] R.L. Parker. The theory of ideal bodies for gravity interpretation. *Geophys. J. Roy. astr. Soc.*, 42:315–334, 1975.

[45] R.L. Parker and M.K. McNutt. Statistics for the one-norm misfit measure. *J. Geophys. Res.*, 85:4429–2230, 1980.

[46] R.L. Parker, L. Shure, and J. Hildebrand. The application of inverse theory to seamount magnetism. *Reviews of Geophysics*, 25:17–40, 1987.

[47] J.E. Pierce and B.W. Rust. Constrained least squares interval estimation. *SIAM J. Sci. Stat. Comput.*, 6:670–683, 1985.

[48] W.H. Press, B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling. *Numerical Recipes in C.* Cambridge University Press, 1988.

[49] B.W. Rust and W.R. Burrus. *Mathematical Programming and the Numerical Solution of Linear Equations.* American Elsevier Pub. Co., Inc., New York, 1972.

[50] P.C. Sabatier. Basic concepts and methods of inverse problems. In P.C. Sabatier, editor, *Basic Methods of Tomography and Inverse Problems*, pages 471–667, Adam Hilger, 1987.

[51] P.C. Sabatier. Positivity constraints in linear inverse problems—I. General theory. *Geophys. J. Roy. astr. Soc.*, 48:415–441, 1977.

[52] P.C. Sabatier. Positivity constraints in linear inverse problems—II. Applications. *Geophys. J. Roy. astr. Soc.*, 48:443–459, 1977.

[53] P.C. Sabatier. Well-posed questions and exploration of the space of parameters in linear and non-linear inversion. In F. Santosa, Y.-H. Pao, W.W. Symes, and C. Holland, editors, *Inverse Problems of Acoustic and Elastic Waves*, pages 82–103, SIAM, 1984.

[54] C. Safon, G. Vasseur, and M. Cuer. Some applications of linear programming to the inverse gravity problem. *Geophysics*, 42:1215–1229, 1977.

[55] N.Z. Shor. *Minimization Methods for Non-Differentiable Functions.* Springer-Verlag, 1985.

[56] L. Shure, R.L. Parker, and G.E. Backus. Harmonic splines for geomagnetic modelling. *Phys. Earth Planet. Int.*, 28:215–229, 1982.

[57] L. Shure, R.L. Parker, and R.A. Langel. A preliminary harmonic spline model from Magsat data. *J. Geophys. Res.*, 90:11,505–11,512, 1985.

[58] P.B. Stark. Rigorous computer solutions for infinite-dimensional inverse problems. In P.C. Sabatier, editor, *Inverse Problems in Action*, Springer-Verlag, 1990.

[59] P.B. Stark. Rigorous velocity bounds from soft $\tau(p)$ and $x(p)$ data. *Geophys. J. Roy. astr. Soc.*, 89:987–996, 1987.

[60] P.B. Stark. Strict bounds and applications. In P.C. Sabatier, editor, *Some Topics on Inverse Problems*, World Scientific, Singapore, 1988.

[61] P.B. Stark and R.L. Parker. Correction to 'Velocity bounds from statistical estimates of $\tau(p)$ and $x(p)$' by Philip B. Stark and Robert L. Parker. *J. Geophys. Res.*, 93:13,821–13,822, 1988.

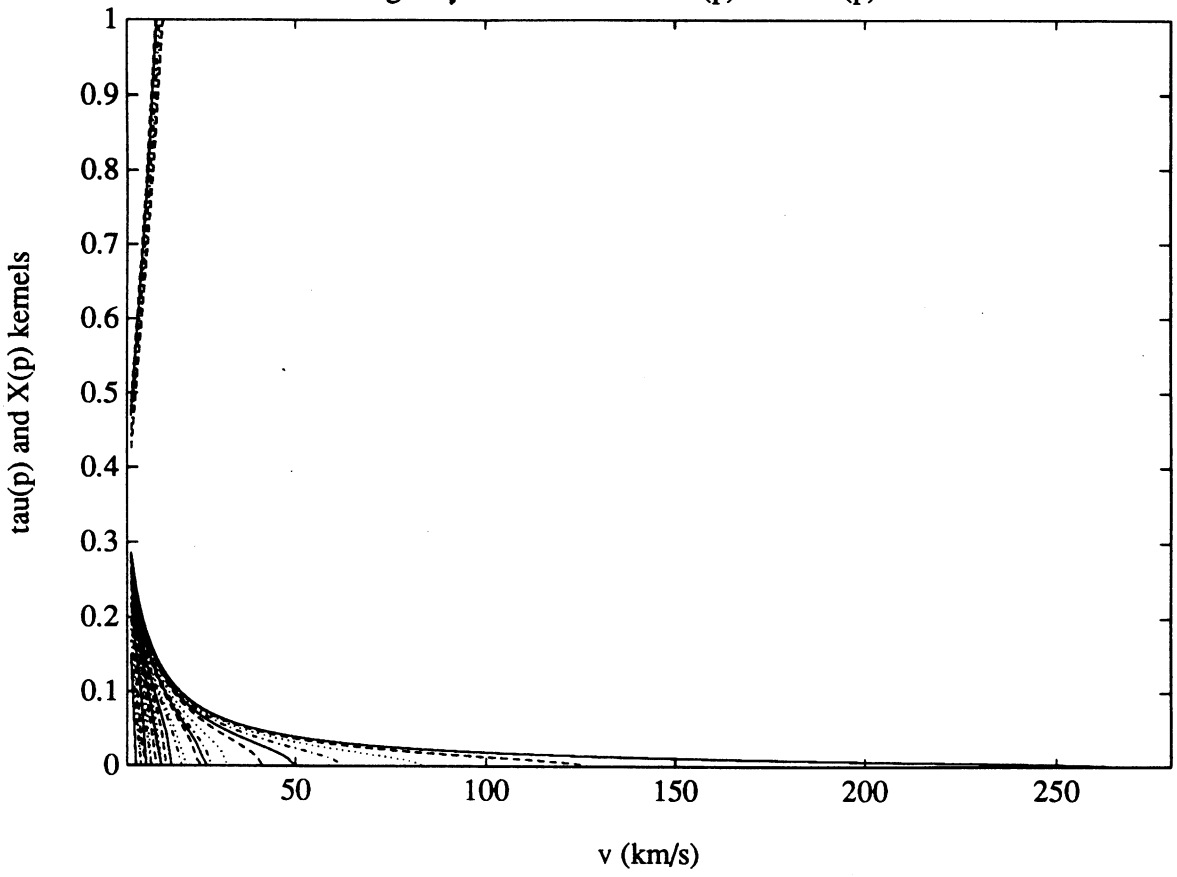
[62] P.B. Stark and R.L. Parker. Smooth profiles from $\tau(p)$ and $X(p)$ data. *Geophys. J. Roy. astr. Soc.*, 89:997–1010, 1987.

[63] P.B. Stark and R.L. Parker. Velocity bounds from statistical estimates of $\tau(p)$ and $X(p)$. *J. Geophys. Res.*, 92:2713–2719, 1987.

[64] P.B. Stark, R.L. Parker, G. Masters, and J.A. Orcutt. Strict bounds on seismic velocity in the spherical Earth. *J. Geophys. Res.*, 91:13,892–13,902, 1986.

[65] R.A. Wiggins, G.A. McMechan, and M.N. Toksöz. Range of Earth structure nonuniqueness implied by body wave observations. *Rev. Geophys.*, 11:87–114, 1973.

Figure 2: Kernels for 25 tau(p) and 5 X(p) data



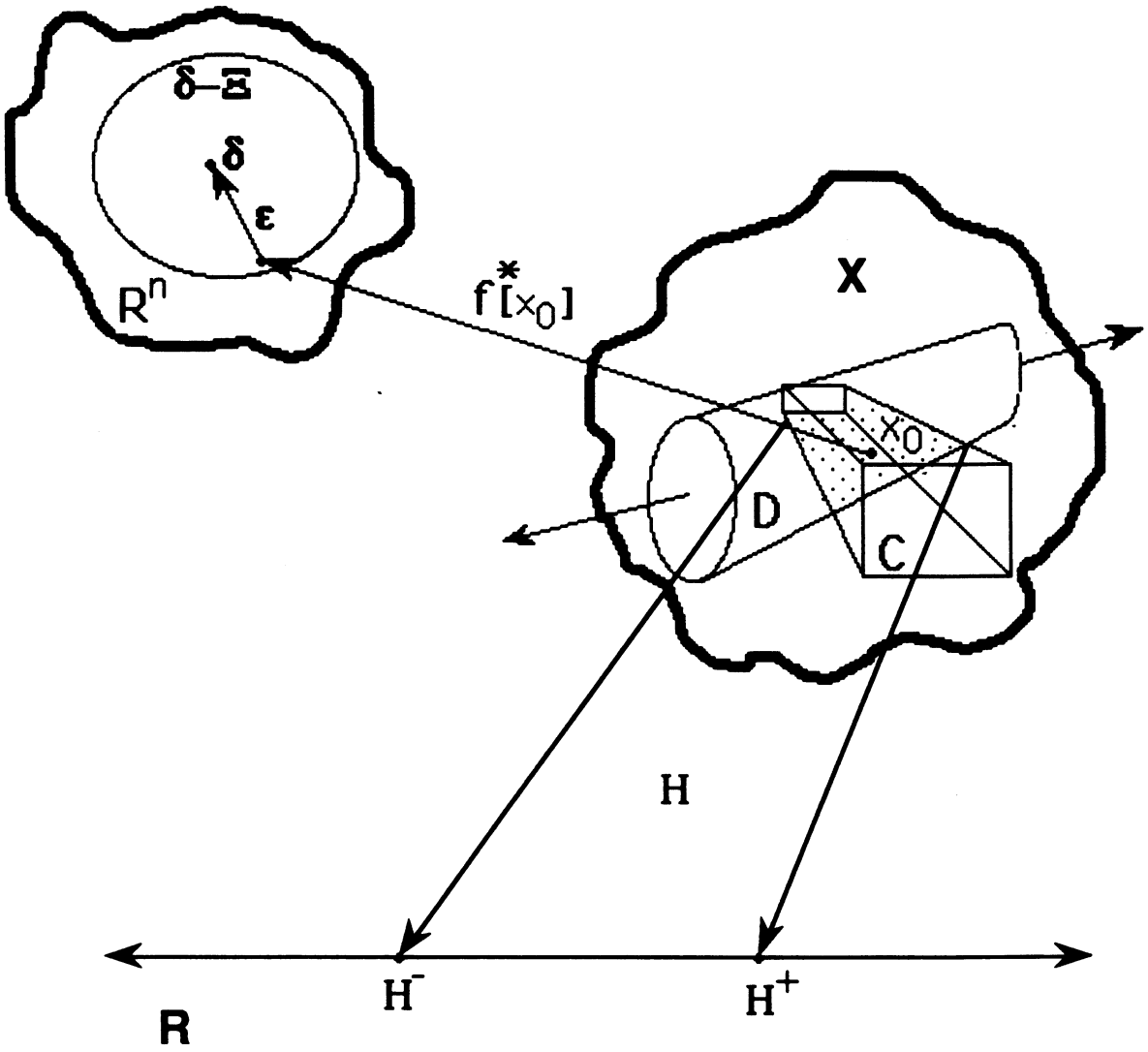


Figure 2: Geometrical Schematic of Strict Bounds; Illustration of the Notation in Section 2.