THESIS PROPOSAL

JORDAN R. HALL

Abstract. In computational mathematics, it is often of interest to obtain the solution to inverse problems. In this proposal, we consider a noisy map f from a high-dimensional parameter space to a data space of specified dimension, where the gradient of f exists, but may be inaccessible. By solving an inverse problem, we may find a characterization of parameter space based off balancing prior beliefs and observable data; in this way, we might understand the sorts of parameters that correspond to observed data. In particular, using a new approach as in [**?**], we may solve for a probability distribution in parameter space (i.e., a "posterior" density) such that observations in data space are consistent with what we observe from "pushing" the posterior through the map f. In our particular setting, we are interested in exploiting the structure of f to solve an inverse problem, avoiding evaluations of f as much as possible. We may cast the inverse problem as an equivalent regularized, convex minimization problem as in [Butler, Tarantola]. If the gradient of f is accessible, the minimization problem can be solved using gradient-based descent methods. If the gradient of f is inaccessible, we apply Derivative-Free (DF) optimization schemes as in [Chen and Wild]. We may be able to enhance the effectiveness of a DF algorithm by performing dimension reduction on f as in [Russi, Constantine] to explain the behavior of f using as few dimensions in parameter space as possible, which allows for minimization to be performed in fewer dimension, saving computational expense.

Contents

Literature Review and Framework

In this section we provide a literature review focused on inverse problem theory, derivative-free (DF) optimization, and dimension reduction. In the process, we will build a theoretical framework upon which to pose research questions, state initial results, and form a research plan in the later sections of this paper.

Data-Consistent Inversion We begin our discussion of inverse problems by defining a parameter space Λ of dimension N, a map or "model" f, and a data space $\mathcal{D}$, which in our setting may be known values of $f(\cdot)$. The data space will almost always have dimension less than N, which is due to our assumption that evaluations of the model, f, are expensive, which in practice implies that gathering data is also expensive. (?)
In the following, we closely follow [Butler] to formulate the inverse problem. First, we assume prior knowledge of the parameter space, given by the so-called "prior distribution," $\pi_\Lambda(\lambda)$. In practical applications, a prior distribution may be given by experts, but may also be a state of knowledge on Λ obtained by other mathematical or statistical processes. Second, we assume that we have a so-called "observed" density, $\pi_{\mathcal{D}}$, which represents our state of knowledge of observed data, which is uncertain due to noise in f and potential measurement error. Finally, we may form a "push-forward," which is the density obtained by solving a *forward problem*, $f\left(\pi_\Lambda\right)$. We denote the solution to the forward problem with $\pi_\Lambda^{\mathcal{D}}$.

Jordan: take a stab at writing up the approach in Tarantola, summarize

Optimization Methods for Solving Inverse Problems

Varis: I'll write an intro here connecting DF algorithms to noisy simulations due to turbulence/chaos to

J: Put Citation Here

We first consider the Derivative Free (DF) algorithm suited for additive and multiplicative noise outlined by Chen and Wild. This technique requires nothing more than evaluations of the noisy model and random draws from a normal distribution. Briefly, this method finds a new iterate by randomly perturbing the previous iterate in Λ; iterates are not allowed to stray much, though, due to relatively small smoothing factors and step sizes. The smoothing factor and step size in the DF algorithms are of great importance to their convergence and termination. As in [Chen and Wild], both the smoothing factor and step size will depend on a scale factor of the L_1 Lipschitz constant of f. As such, it will be of interest to obtain estimates of L_1, which is not straightforward in a gradient-free setting. We refer to [Others] for Lipschitz constant learning in this setting.

Dimension Reduction In our setting, it may be advantageous to perform dimension reduction on f. In particular, we shall consider Active Subspace methods described by Paul Constantine in [1] and an equivalent method by T.M. Russi in []. These techniques seek to explain outputs $f(x)$ in a subspace $A := A(f; x)$ for which the $\dim(A) < N$.
Assuming that we lack the analytic ∇f, one initializes the method by performing M random draws of $x_i \in \Lambda$. We then compute $f(x_i)$ for all $i = 1, \ldots, M$ samples, which we note will require, at the very least, M evaluations of f; in a realistic setting, this would require M solves of a model (e.g., M solves of a PDE constrained system). In our notations, we

have $\mathcal{D} = \{f(x_i)\}_{i=1}^{M}$. Next, we need $\nabla_\Lambda f$ evaluated at x_i for all $i = 1, \ldots, M$, which we assume that we do not have in closed analytic form. Hence, we generally need some gradient approximation method, and typically a locally linear approximation to the gradient is a fair balance between reasonably estimating the gradient and not pushing computational expenses to an unreasonable regime. With this approximation formed, we denote each estimation to $\nabla f(x_i)$ with $\widehat{\nabla f}(x_i)$ and we define the $M \times n$ matrix

$$W := \begin{bmatrix} \widehat{\nabla f}(x_1)^T \\ \cdots \\ \widehat{\nabla f}(x_i)^T \\ \cdots \\ \widehat{\nabla f}(x_M)^T \end{bmatrix}, \tag{1}$$

which defines a covariance structure of f over Λ. This interpretation of (1) leads us to the idea of computing the Singular Value Decomposition of W,

$$W = U\Sigma V^*, \tag{2}$$

where U is $M \times M$ unitary, Σ is $M \times n$ diagonal with the singular values of W along its diagonal, and V^* is $n \times n$ unitary. With the singular values of W in hand, we search for a drop-off in the spectrum of W. In detail, we plot the singular values, $\{\sigma_i\}_{i=1}^{n}$ and seek a drop-off in magnitude between some pair of singular values, σ_j and σ_{j+1}. The active subspace is the span of $u_1, \ldots, u_j$, which are the first j columns of W, the so-called "left singular vectors" of W. We let $A\left(f; \mathcal{D}\right) := \text{span}\{u_1, \ldots, u_j\}$ to denote the active subspace of f with respect to the data $\mathcal{D}$. We choose to use a notation with $\mathcal{D}$ included to emphasize the dependence of the active subspace on the random draws made in Λ, which led to our particular $\mathcal{D}$.

The fact that $u_1, \ldots, u_j$ correspond to the nontrivial singular values is exactly why they account for the most amount of variance in function values. In fact, one can view AS analysis as an artful choice of principal components after a *full* Principal Component Analysis (PCA) is performed in the gradient space W; for more details on this viewpoint, we refer the interested reader to Section 6.4 in T.M. Russi [3].

For a point $\tilde{x} \in \Lambda$, we may write $\displaystyle\sum_{i=1}^{j-1} \left(u_i^T \tilde{x}\right) u_i \in A(f; \mathcal{D})$, which is a projection of the point $\tilde{x}$ in the "active directions" of f. We call this projection an "active variable," which is a point in the active subspace A, which we will use to abbreviate $A(f; \mathcal{D})$ when it is clear to which f and data $\mathcal{D}$ we are referring.

We have arrived at the property that

$$f\left(\sum_{i=1}^{j-1}\left(u_i^T x\right) u_i\right) \approx f(x). \tag{3}$$

In practice, we can check the extent to which the active subspace accounts for functions values $f(x)$ by checking for resolution in a so-called “sufficient summary plot” [1], where we plot active variables against function values. In these plots, we hope to see a pattern between the active variables versus the actual function values. For example, if f is quadratic in its active variables, then we expect to see quadratically-resolved sufficient summary plots. We will revisit sufficient summary plots in the Methods and Algorithms section. For now, we continue describing the AS minimization method with the understanding that sufficient summary plots allow us to check the “fit” of the active subspace.

Varis: I’ll write this, covering Grad-Shafonov Equations, and why these MHD equilibria are important to kinetic simulations.

Application: Magnetic Equilibria in Tokamaks

Research Questions

References

[1] Constantine, Paul G. “Active Subspaces: Emerging Ideas for Dimension Reduction in Parameter Studies.” SIAM, 2015.
The Active Subspaces software library and interactive Jupyter notebooks can be found at `https://github.com/paulcon/active_subspaces`.

[2] Chen and Wild. “Randomized Derivative-Free Optimization of Noisy Convex Functions.” Funded by the Department of Energy. 2015.

[3] Russi, Trent M. “Uncertainty Quantification with Experimental Data and Complex System Models.” Dissertation, University of California Berkeley. 2010.

[4] Smith, Ralph.“Uncertainty Quantification: Theory, Implementation, and Applications.” SIAM, 2013.