Robust inversion, dimensionality reduction, and randomized sampling

Aleksandr Aravkin
Michael P. Friedlander
Felix J. Herrmann
Tristan van Leeuwen

November 16, 2011

Abstract We consider a class of inverse problems in which the forward model is the solution operator to linear ODEs or PDEs. This class admits several dimensionality-reduction techniques based on data averaging or sampling, which are especially useful for large-scale problems. We survey these approaches and their connection to stochastic optimization. The data-averaging approach is only viable, however, for a least-squares misfit, which is sensitive to outliers in the data and artifacts unexplained by the forward model. This motivates us to propose a robust formulation based on the Student's t-distribution of the error. We demonstrate how the corresponding penalty function, together with the sampling approach, can obtain good results for a large-scale seismic inverse problem with 50% corrupted data.

Keywords inverse problems · seismic inversion · stochastic optimization · robust estimation

1 Introduction

Consider the generic parameter-estimation scheme in which we conduct m experiments, recording the corresponding experimental input vectors $\{q_1, q_2, \ldots, q_m\}$ and observation vectors $\{d_1, d_2, \ldots, d_m\}$. We model the data for given parameters $x \in \mathbb{R}^n$ by

$$d_i = F_i(x) q_i + \epsilon_i \quad \text{for} \quad i = 1, \ldots, m, \tag{1.1}$$

This work was in part financially supported by the Natural Sciences and Engineering Research Council of Canada Discovery Grant (22R81254) and the Collaborative Research and Development Grant DNOISE II (375142-08). This research was carried out as part of the SINBAD II project with support from the following organizations: BG Group, BPG, BP, Chevron, Conoco Phillips, Petrobras, PGS, Total SA, and WesternGeco.

A. Aravkin, F. J. Herrmann, and T. van Leeuwen
Dept. of Earth and Ocean Sciences, University of British Columbia, Vancouver, BC, Canada
E-mail: {saravkin,fherrmann,tleeuwen}@eos.ubc.ca

M. P. Friedlander
Dept. of Computer Science, University of British Columbia, Vancouver, BC, Canada
E-mail: mpf@cs.ubc.ca

where observation d_i is obtained by the linear action of the forward model $F_i(x)$ on known source parameters q_i, and ϵ_i captures the discrepancy between d_i and prediction $F_i(x)q_i$. The class of models captured by this representation includes solution operators to any linear (partial) differential equation with boundary conditions, where the q_i are the right-hand sides of the equations. A special case arises when $F_i \equiv F$, i.e., the forward model is the same for each experiment.

Inverse problems based on these forward models arise in a variety of applications, including medical imaging and seismic exploration, in which the parameters x usually represent particular physical properties of a material. We are particularly motivated by the full-waveform inversion (FWI) application in seismology, which is used to image the earth's subsurface [36]. In full-waveform inversion, the forward model F is the solution operator of the wave equation composed with a restriction of the full solution to the observation points (receivers); x represents sound-velocity parameters for a (spatial) 2- or 3-dimensional mesh; the vectors q_i encode the location and signature of the ith source experiment; and the vectors d_i contain the corresponding measurements at each receiver. A typical survey in exploration seismology may contain thousands of experiments (shots), and global seismology relies on natural experiments provided by measuring thousands of earthquakes detected at seismic stations around the world. Standard data-fitting algorithms may require months of CPU time on large computing clusters to process this volume of data and yield coherent geological information.

Inverse problems based on the forward models that satisfy (1.1) are typically solved by minimizing some measure of misfit, and have the general form

$$\operatorname*{minimize}_{x} \quad \phi(x) := \frac{1}{m}\sum_{i=1}^{m}\phi_i(x), \tag{1.2}$$

where each $\phi_i(x)$ is some measure of the residual

$$r_i(x) := d_i - F_i(x)q_i \tag{1.3}$$

between the observation and prediction of the ith experiment. The classical approach is based on the least-squares penalty

$$\phi_i(x) = \|r_i(x)\|^2. \tag{1.4}$$

This choice can be interpreted as finding the *maximum a-posteriori* (MAP) likelihood estimate of x, given the assumptions that the errors ϵ_i are independent and follow a Gaussian distribution.

Formulation (1.2) is general enough to capture a variety of models, including many familiar examples. If the d_i and q_i are scalars, and the forward model is linear, then standard least-squares

$$\phi_i(x) = \tfrac{1}{2}(a_i^T x - d_i)^2$$

easily fits into our general formulation. More generally, maximum-likelihood and MAP estimation are based on the form

$$\phi_i(x) = -\log p_i\big(r_i(x)\big),$$

where p_i is a particular probability density function of ϵ_i.

1.1 Dimensionality reduction

Full-waveform inversion is a prime example of an application in which the cost of evaluating each element in the sum of ϕ is very costly: every residual vector $r_i(x)$—required to evaluate one element in the sum of (1.2)—entails solving a partial differential equation on a 2D or 3D mesh with thousands of grid points in each dimension. The scale of such problems is a motivation for using dimensionality reduction techniques that address small portions of the data at a time.

The least-squares objective (1.4) allows for a powerful form of data aggregation that is based on randomly fusing groups of experiments into "meta" experiments, with the effect of reducing the overall problem size. The aggregation scheme is based on Haber et al.'s [17] observation that for this choice of penalty, the objective is connected to the trace of a residual matrix. That is, we can represent the objective of (1.2) by

$$\phi(x) = \frac{1}{m}\sum_{i=1}^{m} \|r_i(x)\|^2 \equiv \frac{1}{m}\,\mathrm{trace}\,\big(R(x)^T R(x)\big), \tag{1.5}$$

where

$$R(x) := [r_1(x), r_2(x), \ldots, r_m(x)]$$

collects the residual vectors (1.3). Now consider a small sample of s weighted averages of the data, i.e.,

$$\widetilde{d}_j = \sum_{i=1}^{m} w_{ij} d_i \quad \text{and} \quad \widetilde{q}_j = \sum_{i=1}^{m} w_{ij} q_i, \quad j = 1, \ldots, s,$$

where $s \ll m$ and w_{ij} are random variables, and collect the corresponding s residuals $\widetilde{r}_j(x) = \widetilde{d}_j - F_j(x)\widetilde{q}_j$ into the matrix $R_W(x) := [\widetilde{r}_1(x), \widetilde{r}_2(x), \ldots, \widetilde{r}_s(x)]$. Because the residuals are linear in the data, we can write compactly

$$R_W(x) := R(x)W \quad \text{where} \quad W := (w_{ij}).$$

Thus, we may consider the sample function

$$\phi_W(x) = \frac{1}{s}\sum_{j=1}^{s} \|\widetilde{r}_j(x)\|^2 \equiv \frac{1}{s}\,\mathrm{trace}\,\big(R_W(x)^T R_W(x)\big) \tag{1.6}$$

based on the s averaged residuals. Proposition 1.1 then follows directly from Hutchinson's [22, §2] work on stochastic trace estimation.

Proposition 1.1. *If* $\mathbb{E}[WW^T] = I$, *then*

$$\mathbb{E}\big[\phi_W(x)\big] = \phi(x) \quad \textit{and} \quad \mathbb{E}[\nabla\phi_W(x)] = \nabla\phi(x).$$

Hutchinson proves that if the weights w_{ij} are drawn independently from a Rademacher distribution, which takes the values ±1 with equal probability, then the stochastic-trace estimate has minimum variance. Avron and Toledo [4] compare the quality of stochastic estimators obtained from other distributions. Golub and von Matt [15] report the surprising result that the estimate obtained with even a single sample ($s = 1$) is often of high quality. Experiments that use the approach in FWI give evidence that good estimates of the true parameters can be obtained at a fraction of the computational cost required by the full approach [19, 24, 39].

1.2 Approach

Although the least-squares approach enjoys widespread use, and naturally accommodates the dimensionality-reduction technique just described, it is known to be unsuitable for very noisy or corrupted data, often encountered in practice. The least-squares formulation also breaks down in the face of systematic features of the data that are unexplained by the model F_i.

Our aim is to characterize the benefits of robust inversion and to describe randomized sampling schemes and optimization algorithms suitable for large-scale applications in which even a single evaluation of the forward model and its action on q_i is computationally expensive. (In practice, the product $F_i(x)q_i$ is evaluated as a single unit.) We interpret these sampling schemes, which include the well-known incremental-gradient algorithm [28], as dimensionality-reduction techniques, because they allow algorithms to make progress using only a portion of the data.

This paper is organized into the following components:

Robust statistics (§2). We survey robust approaches from a statistical perspective, and present a robust approach based on the heavy-tailed Student's t-distribution. We show that all log-concave error models share statistical properties that differentiate them from heavy-tailed densities (such as the Student's t) and limit their ability to work in regimes with large outliers or significant systematic corruption of the data. We demonstrate that densities outside the log-concave family allow extremely robust formulations that yield reasonable inversion results even in the face of major data contamination.

Sample average approximations (§3). We propose a dimensionality-reduction technique based on sampling the available data, and characterize the statistical properties that make it suitable as the basis for an optimization algorithm to solve the general inversion problem (1.2). These techniques can be used for the general robust formulation described in §2, and for formulations in which forward models F_i vary with i.

Stochastic optimization (§4) We review stochastic-gradient, randomized incremental-gradient, and sample-average methods. We show how the assumptions required by each method fit with the class of inverse problems of interest, and can be satisfied by the sampling schemes discussed in §3.

Seismic inversion (§5) We test the proposed sample-average approach on the robust formulation of the FWI problem. We compare the inversion results obtained with the new heavy-tailed approach to those obtained using robust log-concave models and conventional methods, and demonstrate that a useful synthetic velocity model can be recovered by the heavy-tailed robust method in an extreme case with 50% missing data. We also compare the performance of stochastic algorithms and deterministic approaches, and show that the robust result can be obtained using only 30% of the effort required by a deterministic approach.

2 Robust Statistics

A popular approach in robust regression is to replace the least-squares penalty (1.4) on the residual with a penalty that increases more slowly than the 2-norm. (Virieux and Operto [40] discuss the difficulties with least-squares regression, which are especially egregious in seismic inversion.)

One way to derive a robust approach of this form is to assume that the noise ϵ_i comes from a particular non-Gaussian probability density, p_i, and then find the maximum likelihood (ML) or maximum a posteriori (MAP) estimate of the parameters x that maximizes the likelihood that the residual vectors $r_i(x)$ are realizations of the random variable ϵ_i, given the observations d_i. Because the negative logarithm is monotone decreasing, it is natural to minimize the negative log of the likelihood function rather than maximizing the likelihood itself. In fact, when the distribution of the errors ϵ_i is modeled using a log-concave density

$$p(r) \propto \exp\big(-\rho(r)\big),$$

with a convex loss function ρ, the MAP estimation problem is equivalent to the formulation (1.2), with

$$\phi_i(x) = \rho(r_i(x)) \quad \text{for} \quad i = 1, \dots, m. \tag{2.1}$$

One could also simply start with a penalty ρ on $r_i(x)$, without explicitly modelling the noise density; estimates obtained this way are generally known as M-estimates [20]. A popular choice that follows this approach is the Huber penalty [20, 21, 27].

Robust formulations are typically based on convex penalties ρ—or equivalently, on log-concave densities for ϵ_i— that look quadratic near 0 and increase linearly far from 0. In the seismic context, the Huber penalty is considered by Guitton and Symes [16], who cite many previous examples of the use of 1-norm penalty in the geophysical context. Huber and 1-norm penalties are further compared on large-scale seismic problems by Brossier et al. [8], and a Huber-like (but strictly convex) hyperbolic penalty is described by Bube and Nemeth [10], with the aim of avoiding possible non-uniqueness associated with the Huber penalty.

Clearly, practitioners have a preference for convex formulations. However, it is important to note that

- for nonlinear forward models F_i, the optimization problem (1.2) is typically nonconvex even for convex penalties ρ (it is difficult to satisfy the compositional requirements for convexity in that case);
- even for linear forward models F_i, it may be beneficial to choose a nonconvex penalty in order to guard against outliers in the data.

We will justify the second point from a statistical perspective. Before we proceed with the argument, we introduce the Student's t-density, which we use in designing our robust method for FWI.

2.1 Heavy-tailed distribution: Student's t

Robust formulations using the Student's t-distribution have been shown to outperform log-concave formulations in various applications [1]. In this section, we introduce the Student's t-density, explain its properties, and establish a result that underscores how different heavy-tailed distributions are from those in the log-concave family.

The scalar Student's t-density function with mean μ and positive degrees-of-freedom parameter ν is given by

$$p(\, r \mid \mu, \nu \,) \propto \left(1 + (r-\mu)^2/\nu\right)^{-(1+\nu)/2}. \tag{2.2}$$

The density is depicted in Figure 1(a). The parameter ν can be understood by recalling the origins of the Student's t-distribution. Given n i.i.d. Gaussian variables x_i with mean μ, the normalized sample mean

$$\frac{\bar{x}-\mu}{S/\sqrt{n}} \tag{2.3}$$

follows the Student's t-distribution with $\nu = n-1$, where the sample variance $S^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$ is distributed as a χ^2 random variable with $n-1$ degrees of freedom. As $\nu \to \infty$, the characterization (2.3) immediately implies that the Student's t-density converges pointwise to the density of $N(0,1)$. Thus, ν can be interpreted as a tuning parameter: for low values one expects a high degree of non-normality, but as ν increases, the distribution behaves more like a Gaussian distribution. This interpretation is highlighted in [25].

For a zero-mean Student's t-distribution ($\mu = 0$), the log-likelihood of the density (2.2) gives rise to the nonconvex penalty function

$$\rho(r) = \log(1 + r^2/\nu), \tag{2.4}$$

which is depicted in Figure 1(b). The nonconvexity of this penalty is equivalent to the sub-exponential decrease of the tail of the Student's t-distribution, which goes to 0 at the rate $1/r^{\nu+1}$ as $r \to \infty$.

The significance of these so-called *heavy tails* in outlier removal becomes clear when we consider the following question: Given that a scalar residual deviates from the mean by more than t, what is the probability that it actually deviates by more than $2t$?

The 1-norm is the slowest-growing convex penalty, and is induced by the Laplace distribution, which is proportional to $\exp(-\|\cdot\|_1)$. A basic property of the scalar Laplace distribution is that it is memory free. That is, given a Laplace distribution with mean $1/\alpha$, then the probability relationship

$$\Pr(|r| > t_2 \mid |r| > t_1) = \Pr(|r| > t_2 - t_1) = \exp(-\alpha[t_2 - t_1]) \tag{2.5}$$

holds for all $t_2 > t_1$. Hence, the probability that a scalar residual is at least $2t$ away from the mean, given that it is at least t away from the mean, decays exponentially fast with t. For large t, it is unintuitive to make such a strong claim for a residual already known to correspond to an outlier.

Contrast this behavior with that of the Student's t-distribution. When $\nu = 1$, the Student's t-distribution is simply the Cauchy distribution, with a density proportional to $1/(1+r^2)$. Then we have that

$$\lim_{t\to\infty} \Pr(|r| > 2t \mid |r| > t) = \lim_{t\to\infty} \frac{\frac{\pi}{2} - \arctan(2t)}{\frac{\pi}{2} - \arctan(t)} = \frac{1}{2}.$$

Remarkably, the conditional probability is independent of t for large residuals. This cannot be achieved with any probability density arising from a convex penalty, because (2.5) provides a lower bound for this family of densities, as is shown in the following theorem.

Theorem 2.1. *Consider any scalar density p arising from a symmetric convex and differentiable penalty ρ via $p(t) = \exp(-\rho(t))$, and take any point t_0 with $\rho'(t_0) = \alpha_0 > 0$. Then for all $t_2 > t_1 \geq t_0$, the conditional tail distribution induced by $p(r)$ satisfies*

$$\Pr(|r| > t_2 \mid |r| > t_1) \leq \exp(-\alpha_0[t_2 - t_1])\ .$$

Proof. Define $l(t) = \rho(t_1) + \alpha_1(t - t_1)$, with $\alpha_1 = \rho'(t_1)$, to be the (global) linear under-estimate for ρ at t_1. Define $F(t) = \int_t^\infty p(r)\,dr$. Because $p(t)$ is log-concave and differentiable, it follows from [5, Corollary 3] that the ratio $p(t)/F(t)$ (known as the failure rate) is monotonically increasing, so in particular

$$\frac{p(t_1)}{F(t_1)} \leq \frac{p(t_2)}{F(t_2)}\ , \qquad \text{or equivalently,} \qquad \frac{F(t_2)}{F(t_1)} \leq \frac{p(t_2)}{p(t_1)}\ .$$

By assumption on the functions ℓ and ρ,

$$\rho(t_2) - \ell(t_2) \geq \rho(t_1) - \ell(t_1) = 0,$$

which implies that

$$\begin{aligned}\Pr(|r| > t_2 \mid |r| > t_1) = \frac{F(t_2)}{F(t_1)} &\leq \frac{\exp(-\rho(t_2))}{\exp(-\rho(t_1))} \\ &= \exp(-[\rho(t_2) - \ell(t_1)]) \\ &\leq \exp(-[\ell(t_2) - \ell(t_1)]) \\ &= \exp(-\alpha_1[t_2 - t_1])\ .\end{aligned}$$

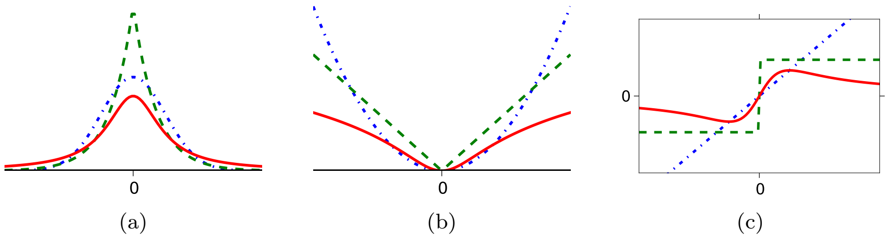
To complete the proof, note that the derivative ρ' is monotonic in t, by the following characterization of convexity for ρ:

$$\frac{\rho'(t_1) - \rho'(t_0)}{t_1 - t_0} \geq 0 \quad \text{for all} \quad t_0, t_1.$$

Then we have $\alpha_0 \leq \alpha_1$ for $t_0 \leq t_1$. □

In order to apply this theorem to the Laplace distribution with the 1-norm penalty $\rho(r) = |r|$, we need to add the condition $t_0 > 0$, which excludes the point of non-differentiability. (We have been unable to relax the differentiability assumption that was needed to assert that the ratio $p(t)/F(t)$ is monotonic.)

For log-concave densities in Theorem 2.1, the *influence function* is defined to be $\rho'(t)$, and for a general distribution it is the derivative of the negative log of the density. These functions provide further insight into the difference between the behaviors of log-concave densities and heavy-tailed densities such as the Student's. In particular, they measure the effect of the size of a residual on the negative log likelihood. The Student's t-density has a so-called *redescending* influence function: as residuals grow larger, they are effectively ignored by the model. Figure 1 shows



(a) (b) (c)

Fig. 1: The Gaussian (·−), Laplace (−−), and Student's t- (—) distributions: (a) densities, (b) penalties, and (c) influence functions.

the relationships among densities, penalties, and influence functions of two log-concave distributions (Gaussian and Laplacian) and those of the Student's t, which is not log-concave. If we examine the derivative

$$\rho'(r) = \frac{2r}{\nu + r^2}$$

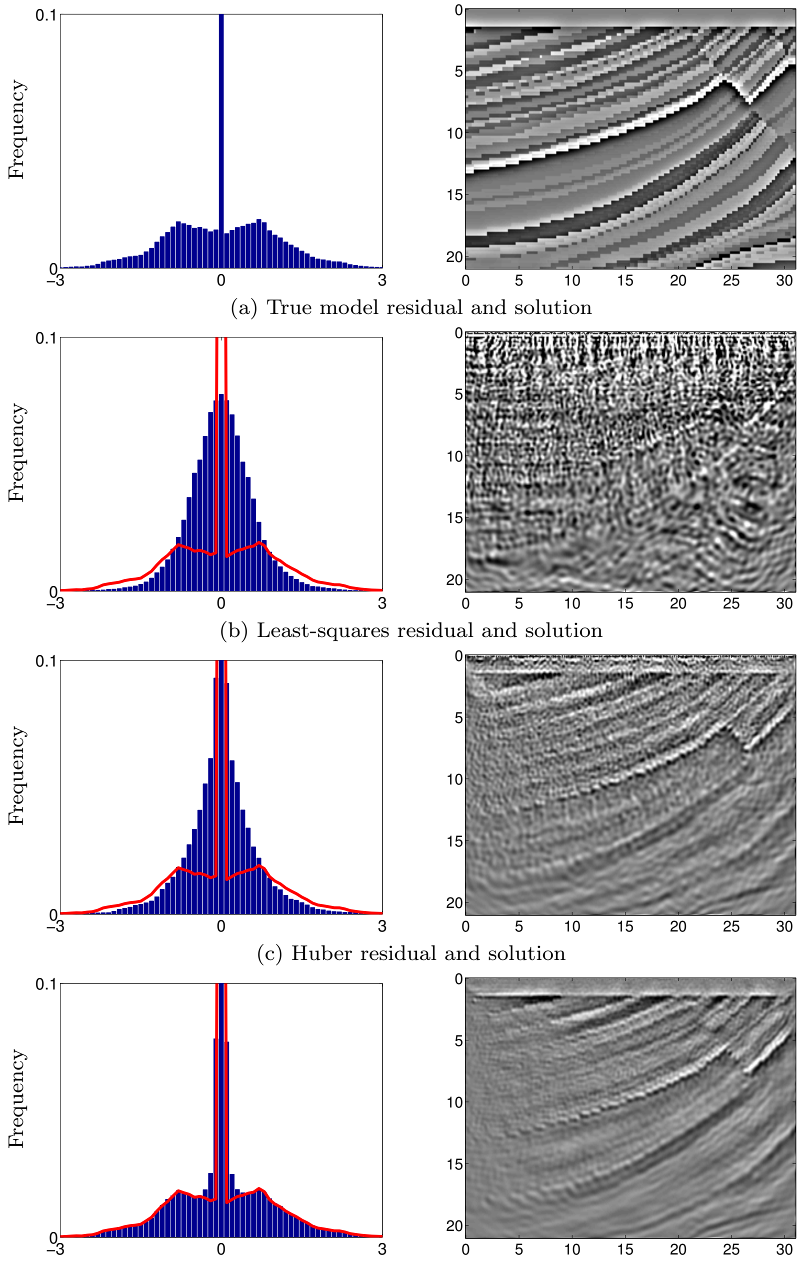
of the Student's t-penalty (2.4), it is clear that large residuals have a small influence when $r^2 \gg \nu$. For small r, on the other hand, the derivative resembles that of the least-squares penalty. See Hampel et al. [18] for a discussion of influence-function approaches to robust statistics, and redescending influence functions in particular, and Shevlyakov et al. [33] for further connections.

There is an implicit tradeoff between convex and non-convex penalties (and their log-concave and non-log-concave counterparts). Convex models are easier to characterize and solve, but may be wrong in a situation in which large outliers are expected. Nonconvex penalties are particularly useful with large outliers.

2.2 The Student's t in practice

Figure 2 compares the reconstruction obtained using the Student's t-penalty, with those obtained using least-squares and Huber penalties, on an FWI experiment (described more fully in §5). These panels show histograms of the residuals (1.3) that are obtained at different solutions, including the true solution, and the solutions recovered by solving (1.2) where the subfunctions ϕ_i in (2.1) are defined by the least-squares, Huber, and Student's t- penalties.

The experiment simulates 50% missing data using a random mask that zeros out half of the data obtained via a forward model at the true value of x. A residual histogram at the true x therefore contains a large spike at 0, corresponding to the residuals for correct data, and a multimodal distribution of residuals for the erased data. The least-squares recovery yields a residual histogram that resembles a Gaussian distribution. The corresponding inversion result is useless, which is not surprising, because the residuals at the true solution are very far from Guassian. The reconstruction using the Huber penalty is a significant improvement over the conventional least-squares approach, and the residual has a shape that resembles the Laplace distribution, which is closer to the shape of the true residual. The Student's t approach yields the best reconstruction, and, remarkably, produces



(a) True model residual and solution

(b) Least-squares residual and solution

(c) Huber residual and solution

(d) Student's t residual and solution

Fig. 2: Residual histograms (normalized) and solutions for an FWI problem. The histogram at (a) the true solution shows that the errors follow a tri-modal distribution (superimposed on the other histogram panels for reference). The residuals for (b) least-squares and (c) Huber reconstructions follow the model error densities (i.e., Gaussian and Laplace). The residuals for (d) the Student t reconstruction, however, closely match the distribution of the actual errors.

a residual distribution that matches the multi-modal shape of the true residual histogram. This is surprising because the Student's t-distribution is unimodal, but the residual shape obtained using the inversion formulation is not. It appears that the statistical prior implied by the Student's t-distribution is weak enough to allow the model to converge to a solution that is almost fully consistent with the good data, and completely ignors the bad data.

Despite several successful applications in statistics and control theory [13, 25], Student's t-formulations do not enjoy widespread use, especially in the context of nonlinear regression and large-scale inverse problems. Recently, however, they were shown to work very well for robust recovery in nonlinear inverse problems such as Kalman smoothing and bundle adjustment [1], and to outperform the Huber penalty when inverting large synthetic models [2, 3]. Moreover, because the corresponding penalty function is smooth, it is usually possible to adapt existing algorithms and workflows to work with a robust formulation.

In order for algorithms to be useful with industrial-scale problems, it is essential that they be designed for conventional and robust formulations that use a relatively small portion of the data in any computational kernel. We lay the groundwork for these algorithms in the next section.

3 Sample average approximations

The data-averaging approach used to derive the approximation (1.6) may not be appropriate when the misfit functions ϕ_i are something other than the 2-norm. In particular, a result such as Proposition 1.1, which reassures us that the approximations are unbiased estimates of the true functions, relies on the special structure of the 2-norm, and is not available to us in the more general case. In this section, we describe sampling strategies—analogous to the stochastic-trace estimation procedure of §1.1—that allow for more general misfit measures ϕ_i. In particular, we are interested in a sampling approach that allows for differential treatment across experiments i, and for robust functions.

We adopt the useful perspective that each of the constituent functions ϕ_i and the gradients $\nabla\phi_i$ are members of a fixed population of size m. The aggregate objective function and its gradient,

$$\phi(x) = \frac{1}{m}\sum_{i=1}^{m}\phi_i(x) \quad \text{and} \quad \nabla\phi(x) = \frac{1}{m}\sum_{i=1}^{m}\nabla\phi_i(x),$$

can then simply be considered to be population averages of the individual objectives and gradients, as reflected in the scaling factors $1/m$. A common method for estimating the mean of a population is to sample only a small subset $\mathcal{S} \subseteq \{\, 1,\dots,m \,\}$ to derive the sample averages

$$\phi_{\mathcal{S}}(x) = \frac{1}{s}\sum_{i\in\mathcal{S}}\phi_i(x) \quad \text{and} \quad \nabla\phi_{\mathcal{S}}(x) = \frac{1}{s}\sum_{i\in\mathcal{S}}\nabla\phi_i(x), \tag{3.1}$$

where $s = |\mathcal{S}|$ is the sample size. We build the subset $\mathcal{S}$ as a uniform random sampling of the full population, and in that case the sample averages are unbiased:

$$\mathbb{E}[\phi_{\mathcal{S}}(x)] = \phi(x) \quad \text{and} \quad \mathbb{E}[\nabla\phi_{\mathcal{S}}(x)] = \nabla\phi(x). \tag{3.2}$$

The cost of evaluating these sample-average approximations is about s/m times that for the true function and gradient. (Non-uniform schemes, such as importance and stratified sampling, are also possible, but require prior knowledge about the relative importance of the ϕ_i.) We use these quantities to drive the optimization procedure.

This approach constitutes a kind of dimensionality-reduction scheme, and it is widely used by census takers to avoid the expense of measuring the entire population. In our case, measuring each element of the population means an evaluation of a function ϕ_i and its gradient $\nabla\phi_i$. The goal of probability sampling is to design randomized sampling schemes that estimate statistics—such as these sample averages—with quantifiable error; see, for example, Lohr's introductory text [26].

The stochastic-optimization methods that we describe in §4 allow for approximate gradients, and thus can take advantage of these sampling schemes. The error analysis of the sample-average method described in §4.3 relies on the second moment of the error

$$e = \nabla\phi_{\mathcal{S}} - \nabla\phi \tag{3.3}$$

in the gradient. Because the sample averages are unbiased, the expected value of the squared error of the approximation reduces to the variance of the norm of the sample average:

$$\mathbb{E}\big[\|e\|^2\big] = \mathbb{V}\big[\|\nabla\phi_{\mathcal{S}}\|\big]. \tag{3.4}$$

This error is key to the optimization process, because the accuracy of the gradient estimate ultimately determines the quality of the search directions available to the underlying optimization algorithm.

3.1 Sampling with and without replacement

Intuitively, the size s of the random sample influences the norm of the error e in the gradient estimate. The difference between uniform sampling schemes with or without replacement greatly affects how the variance of the sample average decreases as the sample size increases. In both cases, the variance of the estimator is proportional to the sample variance

$$\sigma_g := \frac{1}{m-1}\sum_{i=1}^{m} \|\nabla\phi_i - \nabla\phi\|^2 \tag{3.5}$$

of the population of gradients $\{\, \nabla\phi_1, \dots, \nabla\phi_m \,\}$ evaluated at x. This quantity is inherent to the problem and independent of the chosen sampling scheme.

When sampling from a finite population without replacement (i.e., every element in $\mathcal{S}$ occurs only once), then the error e_n of the sample average gradient satisfies

$$\mathbb{E}[\|e_n\|^2] = \frac{1}{s}\left(1 - \frac{s}{m}\right)\sigma_g\,; \tag{3.6}$$

for example, see Cochran [12] or Lohr [26, §2.7]. Note that the expected error decreases with s, and—importantly—is exactly 0 when $s = m$. On the other hand, in a sample average gradient built by uniform sampling with replacement, every sample draw of the population is independent of the others, so that the error e_r of this sample average gradient satisfies

$$\mathbb{E}[\|e_r\|^2] = \frac{1}{s}\sigma_g. \tag{3.7}$$

This error goes to 0 as $1/s$, and is never 0 when sampling over a finite population.

Comparing the expected error between sampling with and without replacement for finite populations, we note that

$$\mathbb{E}[\|e_n\|^2] = \Big(1 - \frac{s}{m}\Big)\,\mathbb{E}[\|e_r\|^2],$$

and so sampling without replacement yields a uniformly lower expected error than independent finite sampling.

3.2 Data averaging

The data-averaging approach discussed in §1.1 for the objective (1.5) does not immediately fit into the sample-average framework just presented, even though the function ϕ_W defined in (1.6) is a sample average. Nevertheless, for all sampling schemes described by Proposition 1.1, the sample average

$$\phi_W(x) = \frac{1}{s}\sum_{j=1}^{s}\widetilde{\phi}_i(x), \quad \text{with} \quad \widetilde{\phi}_i(x) := \|R(x)w_i\|^2,$$

is in some sense a sample average of an infinite population. If the random vectors are uncorrelated—as required by Proposition 1.1—than, as with (3.7), the error

$$e_w = \nabla\phi_W - \phi$$

of the sample average gradient is proportional to the sample variance of the population of gradients of ϕ_W. That is,

$$\mathbb{E}[\|e_w\|^2] = \frac{1}{s}\widetilde{\sigma}_g,$$

where $\widetilde{\sigma}_g$ is the sample variance of the population of gradients $\{\,\nabla\widetilde{\phi}_1,\dots,\nabla\widetilde{\phi}_m\,\}$.

The particular value of $\widetilde{\sigma}_g$ will depend on the distribution from which the weights w_i are drawn; for some distributions of w_i this quantity may even be infinite, as is shown by the following results.

The sample variance (3.5) is always finite, and the analogous sample variance $\widetilde{\sigma}_g$ of the implicit functions $\nabla\widetilde{\phi}_i$ is finite under general conditions on w.

Proposition 3.1. *The sample variance $\widetilde{\sigma}_g$ of the population $\{\,\nabla\widetilde{\phi}_1,\dots,\nabla\widetilde{\phi}_m\,\}$ of gradients is finite when the distribution for w_i has finite fourth moments.*

Proof. The claim follows from a few simple bounds (all sums run from 1 to m):

$$\begin{aligned}
\widetilde{\sigma}_g &\leq \mathbb{E}\left[\|\nabla\phi_{w_i}\|^2\right] \\
&= 4\mathbb{E}\left[\left\|\left(\sum_i \nabla r_i(x)w_i\right)\left(\sum_i r_i(x)w_i\right)\right\|^2\right] \\
&\leq 4\mathbb{E}\left[\left\|\sum_i \nabla r_i(x)w_i\right\|^2\left\|\sum_i r_i(x)w_i\right\|^2\right] \\
&\leq 4\mathbb{E}\left[\left(\sum_i \|\nabla r_i(x)\|^2\|w_i\|^2\right)\left(\sum_i \|r_i(x)\|^2\|w_i\|^2\right)\right] \\
&\leq 4\mathbb{E}\left[\left(m\max_i \|\nabla r_i(x)\|^2\sum_i \|w_i\|^2\right)\left(m\max_i \|r_i(x)\|^2\sum_i \|w_i\|^2\right)\right] \\
&\leq 4m^2\max_i \|\nabla r_i(x)\|^2\cdot\max_i \|r_i(x)\|^2\mathbb{E}\Big[\sum_{ij}\|w_i\|^2\|w_j\|^2\Big]\,.
\end{aligned}$$

The quantity $\mathbb{E}\big[\sum_{ij}\|w_i\|^2\|w_j\|^2\big] < \infty$ when the fourth moments are finite. □

As long as $\widetilde{\sigma}_g$ is nonzero, the expected error of uniform sampling without replacement is asymptotically better than the expected error that results from data averaging. That is,

$$\mathbb{E}[\|e_n\|^2] < \mathbb{E}[\|e_w\|^2] \quad \text{for all } s \text{ large enough.}$$

At least as measured by the second moment of the error in the gradient, the simple random sampling without replacement has the benefit of yielding a good estimate when compared to these other sampling schemes.

4 Stochastic optimization

Stochastic optimization, which naturally allows for inexact gradient calculations, meshes well with the various sampling and averaging strategies described in §3. We review several approaches that fall under the stochastic optimization umbrella, and describe their relative benefits.

4.1 Stochastic gradient methods

Stochastic gradient methods for minimizing a differentiable function ϕ, not necessarily of the form defined in (1.2), can be generically expressed by the iteration

$$x_{k+1} = x_k - \alpha_k d_k \quad \text{with} \quad d_k := s_k + e_k, \tag{4.1}$$

where α_k is a positive stepsize, s_k is a descent direction for ϕ, and e_k is a random noise term. Bertsekas and Tsitsiklis [7, Prop. 3] give general conditions under which the iterates converge to a stationary point, i.e.,

$$\lim_{k\to\infty} \nabla\phi(x_k) = 0.$$

Note that unless the minimizer is unique, this does not imply that the sequence of iterates $\{x_k\}$ converges. Chief among the required conditions are that $\nabla\phi$ is globally Lipschitz, i.e., for some positive L,

$$\|\nabla\phi(x) - \nabla\phi(y)\| \le L\|x-y\| \quad \text{for all } x \text{ and } y;$$

that for all k,

$$s_k^T\nabla\phi(x_k) \le -\mu_1\|\nabla\phi(x_k)\|^2, \tag{4.2a}$$

$$\|s_k\| \le \mu_2(1+\|\nabla\phi(x_k)\|), \tag{4.2b}$$

$$\mathbb{E}[e_k] = 0 \quad \text{and} \quad \mathbb{E}\big[\|e_k\|^2\big] < \mu_3, \tag{4.2c}$$

for some positive constants μ_1, μ_2, and μ_3; and that the steplengths satisfy the infinite travel and summable conditions

$$\sum_{k=0}^{\infty}\alpha_k = \infty \quad \text{and} \quad \sum_{k=0}^{\infty}\alpha_k^2 < \infty. \tag{4.3}$$

Many authors have worked on similar stochastic-gradient methods, but the Bertsekas and Tsitsiklis [7] is particularly general; see their paper for further references.

Note that the randomized sample average schemes (with or without replacement) from §3 can be immediately used to design a stochastic gradient that satisfies (4.2b). It suffices to choose the sample average of the gradient (3.1) as the search direction:

$$d_k = \nabla\phi_S(x_k).$$

Because the sample average $\nabla\phi_S$ is unbiased—cf. (3.2)—this direction is on average simply the steepest descent, and can be interpreted as having been generated from the choices

$$s_k = \nabla\phi(x_k) \quad \text{and} \quad e_k = \nabla\phi_S(x_k) - \nabla\phi(x_k).$$

Moreover, the sample average has finite variance—cf. (3.6)–(3.7)—and so the direction s_k and the error e_k clearly satisfy conditions (4.2).

The same argument holds for the data-averaging scheme outlined in §1.1, as long as the distribution of the mixing vector admits an unbiased sample average with a finite variance. Propositions 1.1 and 3.1 establish conditions under which these requirements hold.

Suppose that ϕ is strongly convex with parameter μ, which implies that

$$\frac{\mu}{2}\|x_k - x^*\|^2 \le \phi(x_k) - \phi(x^*),$$

where x^* is the unique minimizer of ϕ. Under this additional assumption, further statements can be made about the rate of convergence. In particular, the iteration (4.1), with $s_k = \nabla\phi(x_k)$, converges sublinearly, i.e.,

$$\mathbb{E}[\|x_k - x^*\|] = \mathcal{O}(1/k). \tag{4.4}$$

where the steplengths $\alpha_k = \mathcal{O}(1/k)$ are decreasing [30, §2.1]. This is in fact the optimal rate among all first-order stochastic methods [29, §14.1].

A strength of the stochastic algorithm (4.1) is that it applies so generally. All of the sampling approaches that we have discussed so far, and no doubt others, easily fit into this framework. The convergence guarantees are relatively weak for our purposes, however, because they do not provide guidance on how a sampling strategy might influence the speed of convergence. This analysis is crucial within the context of the sampling schemes that we consider, because we want to gain an understanding of how the sample size influences the speed of the algorithm.

4.2 Incremental-gradient methods

Incremental-gradient methods, in their randomized form, can be considered a special case of stochastic gradient methods that are especially suited to optimizing sums of functions such as (1.2). They can be described by the iteration scheme

$$x_{k+1} = x_k - \alpha_k \nabla\phi_{i_k}(x_k), \tag{4.5}$$

for some positive steplengths α_k, where the index i_k selects among the m constituent functions of ϕ. In the deterministic version of the algorithm, the ordering of the subfunctions ϕ_i is predetermined, and the counter $i_k = (k \bmod m) + 1$ makes a full sweep through all the functions every m iterations. In the randomized version, i_k is at each iteration randomly selected with equal probability from the indices $1, \ldots, m$. (The Kaczmarz method for linear system [23] is closely related, and a randomized version of it is analyzed by Strohmer and Vershynin [35].)

In the context of the sampling discussion in §3, the incremental-gradient algorithm can be viewed as an extreme sampling strategy that at each iteration uses only a single function ϕ_i (i.e., a sample of size $s = 1$) in order to form a sample average ϕ_S of the gradient. For the data-averaging case of §1.1, this corresponds to generating the approximation ϕ_W from a single weighted average of the data (i.e., using a single random vector w_i to form $R(x)w_i$).

Bertsekas and Tsitsiklis [6, Prop. 3.8] describe conditions for convergence of the incremental-gradient algorithm for functions with globally Lipschitz continuous gradients, when the steplengths $\alpha_k \to 0$ as specified by (4.3). Note that it is necessary for the steplengths $\alpha_k \to 0$ in order for the iterates x_k produced by (4.5) to ensure stationarity of the limit points. Unless we assume that $\nabla\phi(\bar{x}) = 0$ implies that $\phi_i(\bar{x}) = 0$ for all i, a stationary point of ϕ is not a fixed point of the iteration process; Solodov [34] and Tseng [37] study this case. Solodov [34] further describes how bounding the steplengths away from zero yields limit points $\bar{x}$ that satisfy the approximate stationarity condition

$$\|\nabla\phi(\bar{x})\| = \mathcal{O}\big(\inf_k \alpha_k\big).$$

With the additional assumption of strong convexity of ϕ, it follows from Nedić and Bertsekas [28] that the randomized incremental-gradient algorithm with a decreasing stepsize $\alpha_k = \mathcal{O}(1/k)$ converges sublinearly accordingly to (4.4). They also show that keeping the stepsize constant as $\alpha_k \equiv m/L$ implies that

$$\mathbb{E}[\|x_k - x^*\|^2] \le \mathcal{O}([1 - \mu/L]^k) + \mathcal{O}(m/L).$$

This expression is interesting because the first term on the right-hand side decreases at a linear rate, and depends on the condition number μ/L of ϕ; this term is present for any deterministic first-order method with constant stepsize. Thus, we can see that with the strong-convexity assumption and a constant stepsize, the incremental-gradient algorithm has the same convergence characteristics as steepest descent, but with an additional constant error term.

4.3 Sampling methods

The incremental-gradient method described in §4.2 has the benefit that each iteration costs essentially the same as evaluating only a single gradient element $\nabla\phi_i$. The downside is that they achieve only a sublinear convergence to the exact solution, or a linear convergence to an approximate solution. The sampling approach described in Friedlander and Schmidt [14] allows us to interpolate between the one-at-a-time incremental-gradient method at one extreme, and a full gradient method at the other.

The sampling method is based on the iteration update

$$x_{k+1} = x_k - \alpha g_k, \quad \alpha = 1/L, \tag{4.6}$$

where L is the Lipschitz constant for the gradient, and the search direction

$$g_k = \nabla\phi(x_k) + e_k \tag{4.7}$$

is an approximation of the gradient; the term e_k absorbes the discrepancy between the approximation and the true gradient. We define the direction g_k in terms of the sample average gradient (3.1), and then e_k corresponds to the error defined in (3.3).

When the function ϕ is strongly convex and has a globally Lipschitz continuous gradient, than the following theorem links the convergence of the iterates to the error in the gradient.

Theorem 4.1. *Suppose that* $\mathbb{E}[\|e_k\|^2] \le B_k$, *where* $\lim_{k\to\infty} B_{k+1}/B_k \le 1$. *Then each iteration of algorithm* (4.6) *satisfies for each* $k = 0, 1, 2, \ldots,$

$$\mathbb{E}[\|x_k - x^*\|^2] \le \mathcal{O}([1-\mu/L]^k) + \mathcal{O}(C_k), \tag{4.8}$$

where $C_k = \max\{B_k, (1-\mu/L+\epsilon)^k\}$ *for any positive* ϵ.

It is also possible to replace g_k in (4.6) with a search direction d_k that is the solution of the system

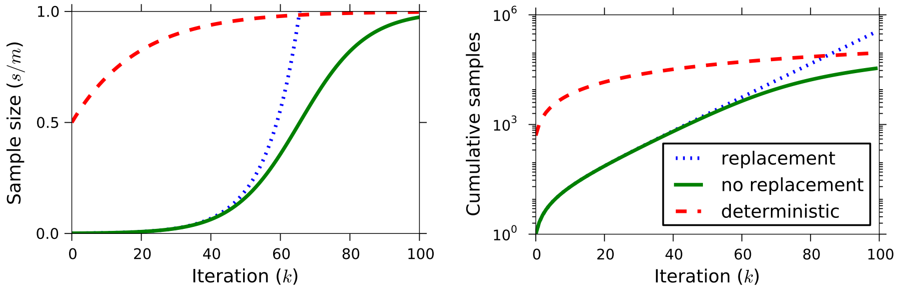
$$H_k d = g_k, \tag{4.9}$$

for any sequence of Hessian approximations H_k that are uniformly positive definite and bounded in norm, as can be enforced in practice. Theorem 4.1 continues to hold in this case, but with different constants μ and L that reflect the conditioning of the "preconditioned" function; see [14, §1.2].

It is useful to compare (4.4) and (4.8), which are remarkably similar. The distance to the solution, for both the incremental-gradient method (4.5) and the gradient-with-errors method (4.6), is bounded by the same linearly convergent term. The second terms in their bounds, however, are crucially different: the accuracy of the incremental-gradient method is bounded by a multiple of the fixed steplength; the accuracy of the gradient-with-errors method is bounded by the norm of the error in the gradient.

Theorem 4.1 is significant because it furnishes a guide for refining the sample $\mathcal{S}_k$ that defines the average approximation

$$g_k = \frac{1}{s_k}\sum_{i\in\mathcal{S}_k}\phi_i(x_k)$$



(a) (b)

Fig. 3: Comparing the difference between the errors in the sample averages, randomized with replacement, randomized without replacement, and deterministic. (a) The sample size (fraction of the total population of $m = 1000$) required to reduce the error linearly with error constant 0.9. (b) The corresponding cumulative number of samples used.

of the gradient of ϕ, where s_k is the size of the sample $\mathcal{S}_k$; cf. (3.1). In particular, (3.6) and (3.7) give the second moment of the errors of these sample averages, which correspond precisely to the gradient error defined by (4.7). If we wish to design a sampling strategy that gives a linear decrease with a certain rate, then a policy for the sample size s_k needs to ensure that it grows fast enough to induce $\mathbb{E}[\|e_k\|^2]$ to decrease with at least that rate.

It is instructive to compare how the sample average error decreases in the randomized (with and without replacement) and deterministic cases. We can more easily compare the randomized and deterministic variants by following Bertsekas and Tsitsiklis [6, §4.2], and assuming that

$$\|\nabla\phi_i(x)\|^2 \le \beta_1 + \beta_2\|\nabla\phi(x)\|^2 \quad \text{for all } x \text{ and } i = 1,\dots,m,$$

for some constants $\beta_1 \ge 0$ and $\beta_2 \ge 1$. Together with the Lipschitz continuity of ϕ, we can provide the following bounds:

$$\text{randomized, without replacement} \qquad \mathbb{E}[\|e_k\|^2] \le \frac{1}{s_k}\left[1-\frac{s_k}{m}\right]\left[\frac{m}{m-1}\right]\beta_k \tag{4.10a}$$

$$\text{randomized, with replacement} \qquad \mathbb{E}[\|e_k\|^2] \le \frac{1}{s_k}\left[\frac{m}{m-1}\right]\beta_k \tag{4.10b}$$

$$\text{deterministic} \qquad \|e_k\|^2 \le 4\left[\frac{m-s_k}{m}\right]^2\beta_k, \tag{4.10c}$$

where $\beta_k = \beta_1 + 2\beta_2 L[\phi(x_k) - \phi(x^*)]$. These bounds follow readily from the derivation in [14, §§3.1–3.2]. Figure 3 illustrates the difference between these bounds on an example problem with $m = 1000$. The panel on the left shows how the sample size needs to be increased in order for the right-hand-side bounds in (4.10) to decrease linearly at a rate of 0.9. The panel on the right shows the cumulative sample size, i.e., $\sum_{i=0}^{k} s_i$. Uniform sampling without replacement yields a uniformly and significantly better bound than the other sampling strategies. Both types of sampling are admissible, but sampling without replacement requires a much slower rate of growth of s to guarantee a linear rate.

The strong convexity assumption needed to derive the error bounds used in this section is especially strong because the inverse problem we use to motivate the sampling approach is not a convex problem. In fact, it is virtually impossible to guarantee convexity of a composite function such as (2.1) unless the penalty function $\rho(\cdot)$ is convex and each $r_i(\cdot)$ is affine. This is not the case for many interesting inverse problems, such as full waveform inversion, and for nonconvex loss functions corresponding to distributions with heavy tails, such as Student's t.

Even relaxing the assumption on ϕ from strong convexity to just convexity makes it difficult to design a sampling strategy with a certain convergence rate. The full-gradient method for convex (but not strongly) functions has a sublinear convergence rate of $\mathcal{O}(1/k)$. Thus, all that is possible for a sampling-type approach that introduces errors into the gradient is to simply maintain that sublinear rate. For example, if $\|e_k\|^2 \leq B_k$, and $\sum_{k=1}^{\infty} B_k < \infty$, then the iteration (4.6) maintains the sublinear rate of the gradient method [14, Theorem 2.6]. The theory for the strongly convex case is also supported by empirical evidence, where sampling strategies tend to outperform basic incremental-gradient methods.

5 Numerical experiments in seismic inversion

A good candidate for the sampling approach we have discussed is the full waveform inversion problem from exploration geophysics, which we address using a robust formulation. The goal is to obtain an estimate of subsurface properties of the earth using seismic data. To collect the data, explosive charges are detonated just below the surface, and the energy that reflects back is recorded at the surface by a large array of geophones. The resulting data consist of a time-series collection for thousands of source positions.

The estimate of the medium parameters is based on fitting the recorded and predicted data. Typically, the predicted data are generated by solving a PDE whose coefficients are the features of interest. The resulting PDE-constrained optimization problem can be formulated in either the time [36] or the frequency [32] domain. It is common practice to use a simple scalar wave equation to predict the data, effectively assuming that the earth behaves like a fluid—in this case, sound speed is the parameter we seek.

Raw data are processed to remove any unwanted artifacts; this requires significant time and effort. One source of unwanted artifacts in the data is equipment malfunction. If some of the receivers are not working properly, the resulting data can be either zero or contaminated with an unusual amount of noise. And even if we were to have a perfect estimate of the sound speed, we still would not expect to be able to fit our model perfectly to the data. The presence of these outliers in the data motivates us (and many other authors, e.g., [8, 9, 16]) to use robust methods for this application. We compare the results of robust Student's t-based inversion to those obtained using least-squares and Huber robust penalties, and we compare the performance of deterministic, incremental-gradient, and sampling methods in this setting.

5.1 Modelling and gradient computation for full waveform inversion

The forward model for frequency-domain acoustic FWI, for a single source function q, assumes that wave propagation in the earth is described by the scalar Helmholtz equation

$$A_\omega(x)u = [\omega^2 x + \nabla^2]u = q,$$

where ω is the angular frequency, x is the squared-slowness (seconds/meter)^2, and u represents the wavefield. The discretization of the Helmholtz operator includes absorbing boundary conditions, so that $A_\omega(x)$ and u are complex-valued. The data are measurements of the wavefield obtained at the receiver locations $d = Pu$. The forward modelling operator $F(x)$ is then given by

$$F(x) = PA^{-1}(x),$$

where A is a sparse block-diagonal matrix, with blocks A_ω indexed by the frequencies ω. Multiple sources q_i are typically modeled as discretized delta functions with a frequency-dependent weight. The resulting data are then modeled by the equation $d_i = F(x)q_i$, and the corresponding residual equals $r_i(x) = d_i - F(x)q_i$ (cf. (1.3)).

For a given loss function ρ, the misfit function and its gradient are defined as

$$\phi(x) = \sum_{i=1}^{m} \rho(r_i(x)) \quad \text{and} \quad \nabla\phi(x) = \sum_{i=1}^{m} \nabla F(x)^* \nabla\rho(r_i(x)),$$

where $\nabla F(x)$ is the Jacobian of F. The action of the adjoint of the Jacobian on a vector y can be efficiently computed via the adjoint-state method [36] as follows:

$$\nabla F(x)^* y = G(x,u)^* v,$$

where $G(x,u)$ is the (sparse) Jacobian of $A(x)u$ with respect to x, and u and v are solutions of the linear systems

$$A(x)u = q \quad \text{and} \quad A(x)^* v = y.$$

The Huber penalty function for vectors y is

$$\rho(y) = \sum_i \zeta_i, \quad \text{where} \quad \zeta_i = \begin{cases} y_i^2/2\mu & \text{if } |y_i| \le \mu \\ |y_i| - \mu/2 & \text{otherwise.} \end{cases}$$

The Student's t penalty function (2.4) for vectors y is defined by

$$\rho(y) = \sum_i \log(1 + y_i^2/\nu).$$

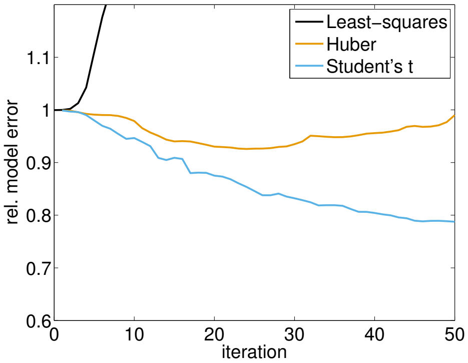


Fig. 4: Relative error between the true and reconstructed models for least-squares, Huber, and Student t penalties. In the least-squares case, the model error is not reduced at all. Slightly better results are obtained with the Huber penalty, although the model error starts to increase after about 20 iterations. The Students t penalty gives the best result.

5.2 Experimental setup and results

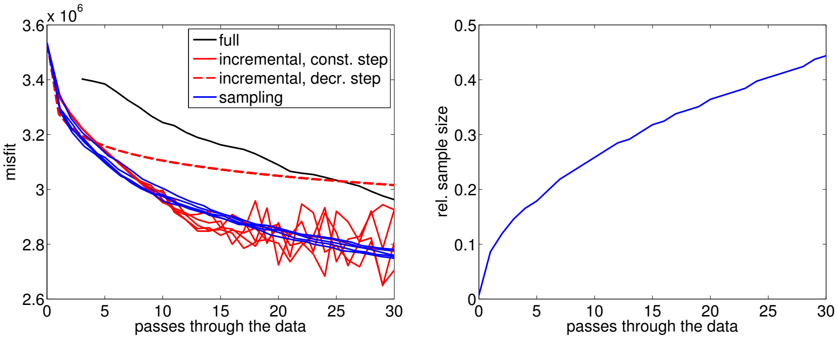
Observed data are generated for the model x^*, depicted in Figure 2(a), for 6 frequencies, and 151 point sources located at the surface. To simulate a scenario in which half of the receivers at unknown locations have failed, we multiply the data with a mask that zeroes out 50% of the data at random locations. The resulting data thus differ from the prediction $F(x^*)$ given by the true solution x^*. A spike in the histogram of the residuals $r_i(x^*)$ evaluated at the true solution x^*, shown in Figure 2a, shows these outliers. The noise does not fit well with any simple prior distribution that one might like to use. We solve the resulting optimization problem with the least-squares, Huber, and Student t- penalties using a limited-memory BFGS method. Figure 4 tracks across iterations the relative model error $||x_k - x^*||/||x^*||$ for all three approaches. Histograms of the residuals after 50 iterations are plotted in Figures 2(c)–(e). The residuals for the least-squares and Huber approaches resemble Gaussian and Laplace distributions respectively. This fits well with the prior assumption on the noise, but does not fit the true residual at all. The residual for the Student's t approach does *not* resemble the prior distribution at all. The slowly increasing penalty function allows for enough freedom to let the residual evolve into the true distribution.

Next, we compare the performance of the incremental-gradient (§4.2) and sampling (§4.3) algorithms against the full-gradient method. For the incremental-gradient algorithm (4.5), at each iteration we randomly choose i uniformly over the set $\{ 1, 2, \ldots, m \}$, and use either a fixed stepsize $\alpha_k \equiv \alpha$ or a decreasing stepsize $\alpha_k = \alpha/\lfloor k/m \rfloor$. The sampling method is implemented via the iteration

$$x_{k+1} = x_k - \alpha_k d_k,$$

where d_k solves the system (4.9), and H_k is a limited-memory BFGS Hessian approximation updated using the pairs $(\Delta x_k, \Delta g_k)$, where

$$\Delta x_k := x_{k+1} - x_k \quad \text{and} \quad \Delta g_k := g_{k+1} - g_k;$$



(a) (b)

Fig. 5: (a) Convergence of different optimization strategies on the Students t penalty: Limited-memory BFGS using the full gradient ("full"), incremental gradient with constant and decreasing step sizes, and the sampling approach. Different lines of the same color indicate independent runs with different random number streams. (b) The evolution of the amount of data used by the sampling method.

the limited-memory Hessian is based on a history of length 4. Nocedal and Wright [31, §7.2] describe the recursive procedure for updating H_k. The batch size is increased at each iteration by only a single element, i.e.,

$$s_{k+1} = \min\{\, m,\; s_k + 1 \,\}.$$

The members of the batch are redrawn at every iteration, and we use an Armijo backtracking linesearch based on the sampled function $(1/s_k)\sum_{i\in\mathcal{S}_k}\phi_i(x)$.

The convergence plots for several runs of the sampling method and the stochastic gradient method with $\alpha = 10^{-6}$ are shown in Figure 5(a). Figure 5(b) plots the evolution of the amounts of data sampled.

6 Discussion and conclusions

The numerical experiments we have conducted using the Student's t-penalty are encouraging, and indicate that this approach can overcome some of the limitations of convex robust penalties such as the Huber norm. Unlike the least-squares and Huber penalties, the Student t-penalty does not force the residual into a shape prescribed by the corresponding distribution. The sampling method successfully combines the steady convergence rate of the full-gradient method with the inexpensive iterations provided by the incremental-gradient method.

The convergence analysis of the sampling method, based on Theorem 4.1, relies on bounding the second moment of the error in the gradient, and hence the variance of the sample average (see (3.4)). The bound on the second-moment arises because of our reliance on the concept of an *expected* distance to optimality $\mathbb{E}[\|x_k - x^*\|^2]$. However, other probabilistic measures of distance to optimality may be more appropriate; this would influence our criteria for bounding the error in the gradient. For example, Avron and Toledo [4] measure the quality of a sample average using an "epsilon-delta" argument that provides a bound on the sample size needed to achieve a particular accuracy ϵ with probability $1-\delta$.

Other refinements are certainly possible. For example, van den Doel and Ascher [38] advocate an adaptive approach for increasing the sample size. Byrd et al. [11] use a sample average approximation of the Hessian, which may provide better results in practice than the limited-memory BFGS approximation that we use in §5.

References

1. A. Aravkin, *Robust Methods with Applications to Kalman Smoothing and Bundle Adjustment*, PhD thesis, University of Washington, Seattle, WA, June 2010.
2. A. Aravkin, T. van Leeuwen, and M. P. Friedlander, *Robust inversion via semistochastic dimensionality reduction*, in Submitted to ICASSP 2012, arXiv:1110.0895, 2011.
3. A. Aravkin, T. van Leeuwen, and F. Herrmann, *Robust full waveform inversion with students t-distribution*, in Proceedings of the SEG, San Antonio, Texas, 2011, Society for Exploration Geophysics.
4. H. Avron and S. Toledo, *Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix*, J. ACM, 58 (2011), pp. 8:1–8:34.
5. M. Bagnoli and T. Bergstrom, *Log-concave probability and its applications*, Economic Theory, 26 (2005), pp. 445–469.
6. D. Bertsekas and J. Tsitsiklis, *Neuro-dynamic programming*, Athena Scientific, 1996.
7. D. P. Bertsekas and J. N. Tsitsiklis, *Gradient convergence in gradient methods with errors*, SIAM J. Optim., 10 (2000), pp. 627–642.
8. R. Brossier, S. Operto, and J. Virieux, *Which data residual norm for robust elastic frequency-domain full waveform inversion?*, Geophysics, 75 (2010), pp. R37–R46.
9. K. P. Bube and R. T. Langan, *Hybrid ℓ_1/ℓ_2 minimization with applications to tomography*, Geophysics, 62 (1997), pp. 1183–1195.
10. K. P. Bube and T. Nemeth, *Fast line searches for the robust solution of linear systems in the hybrid ℓ_1/ℓ_2 and huber norms*, Geophysics, 72 (2007), pp. A13–A17.
11. R. H. Byrd, G. M. Chin, W. Neveitt, and J. Nocedal, *On the use of stochastic hessian information in optimization methods for machine learning*, SIAM Journal on Optimization, 21 (2011), pp. 977–995.
12. W. G. Cochran, *Sampling Techniques*, Jon Wiley, third ed., 1977.
13. L. Fahrmeir and R. Kunstler, *Penalized likelihood smoothing in robust state space models*, Metrika, 49 (1998), pp. 173–191.
14. M. P. Friedlander and M. Schmidt, *Hybrid deterministic-stochastic methods for data fitting*, tech. rep., Univ. of British Columbia, April 2011. revised September 2011.
15. G. H. Golub and U. von Matt, *Quadratically constrained least squares and quadratic problems*, Numer. Math., 59 (1991), pp. 561–580.
16. A. Guitton and W. W. Symes, *Robust inversion of seismic data using the huber norm*, Geophysics, 68 (2003), pp. 1310–1319.
17. E. Haber, M. Chung, and F. J. Herrmann, *An effective method for parameter estimation with pde constraints with multiple right hand sides*, Tech. Rep. TR-2010-4, UBC-Earth and Ocean Sciences Department, 2010.
18. F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel, *Robust Statistics: The Approach Based on Influence Functions*, Wiley Series in Probability and Statistics, 1986.
19. F. Herrmann, M. P. Friedlander, and O. Yilmaz, *Fighting the curse of dimensionality: compressive sensing in exploration seismology*, tech. rep., University of British Columbia, 2011.
20. P. J. Huber, *Robust Statistics*, John Wiley & Sons, Inc., New York, 1981.
21. P. J. Huber and E. M. Ronchetti, *Robust Statistics*, John Wiley and Sons, 2nd ed., 2009.
22. M. Hutchinson, *A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines*, Communications in Statistics - Simulation and Computation, 19 (1990), pp. 433–450.

23. S. Kaczmarz, *Angenäherte auflösung von systemen linearer gleichungen*, Bull. Int. Acad. Polon. Sci. A, 355 (1937), p. 357.
24. J. R. Krebs, J. E. Anderson, D. Hinkley, R. Neelamani, S. Lee, A. Baumstein, and M.-D. Lacasse, *Fast full-wavefield seismic inversion using encoded sources*, Geophysics, 74 (2009), pp. WCC177–WCC188.
25. K. L. Lange, R. J. A. Little, and J. M. G. Taylor, *Robust statistical modeling using the t distribution*, Journal of the American Statistical Association, 84 (1989), pp. 881–896.
26. S. L. Lohr, *Sampling: design and analysis*, Duxbury Press, Pacific Grove, 1999.
27. R. A. Maronna, D. Martin, and Yohai, *Robust Statistics*, Wiley Series in Probability and Statistics, John Wiley and Sons, 2006.
28. A. Nedic and D. Bertsekas, *Convergence rate of incremental subgradient algorithms*, Stochastic Optimization: Algorithms and Applications, (2000), pp. 263–304.
29. A. Nemirovski, *Efficient methods in convex programming*, Lecture notes, (1994).
30. A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, *Robust stochastic approximation approach to stochastic programming*, SIAM J. Optim., 19 (2009), pp. 1574–1609.
31. J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer Series in Operations Research, Springer, 1999.
32. R. Pratt and M. Worthington, *Inverse theory applied to multi-source cross-hole tomography. part i: Acoustic wave-equation method.*, Geophysical Prospecting, 38 (1990), pp. 287–310.
33. G. Shevlyakov, S. Morgenthaler, and A. Shurygin, *Redescending m-estimators*, Journal of Statistical Planning and Inference, 138 (2008), pp. 2906–2917.
34. M. Solodov, *Incremental gradient algorithms with stepsizes bounded away from zero*, Computational Optimization and Applications, 11 (1998), pp. 23–35.
35. T. Strohmer and R. Vershynin, *A randomized Kaczmarz algorithm with exponential convergence*, Journal of Fourier Analysis and Applications, 15 (2009), pp. 262–278.
36. A. Tarantola, *Inversion of seismic reflection data in the acoustic approximation*, Geophysics, 49 (1984), pp. 1259–1266.
37. P. Tseng, *An incremental gradient(-projection) method with momentum term and adaptive stepsize rule*, SIAM Journal on Optimization, 8 (1998), pp. 506–531.
38. K. van den Doel and U. Ascher, *Adaptive and stochastic algorithms for eit and dc resistivity problems with piecewise constant solutions and many measurements*, Tech. Rep., University of British Columbia, September 2011. `http://www.cs.ubc.ca/~ascher/papers/doas2.pdf`.
39. T. van Leeuwen, A. Aravkin, and F. Herrmann, *Seismic waveform inversion by stochastic optimization*, International Journal of Geophysics, 2011 (2011), p. ID 689041.
40. J. Virieux and S. Operto, *An overview of full-waveform inversion in exploration geophysics*, Geophysics, 74 (2009), pp. 127–152.