A Quasi Monte Carlo Method for Large-Scale Inverse Problems

Nick Polydorides, Mengdi Wang, and Dimitri P. Bertsekas

Abstract We consider large-scale linear inverse problems with a simulation-based algorithm that approximates the solution within a low-dimensional subspace. The algorithm uses Tikhonov regularization, regression, and low-dimensional linear algebra calculations and storage. For sampling efficiency, we implement importance sampling schemes, specially tailored to the structure of inverse problems. We emphasize various alternative methods for approximating the optimal sampling distribution and we demonstrate their impact on the reduction of simulation noise. The performance of our algorithm is tested on a practical inverse problem arising from Fredholm integral equations of the first kind.

1 Introduction

Many problems in computational science and engineering are characterized by experimental design, measurement acquisition, and parameter estimation or prediction. This process involves mathematical modeling of the physical systems pertinent to the observations, as well as estimation of unknown model parameters from the acquired measurements by formulating and solving an inverse problem. Quite often solving the inverse problem subject to measurement errors and model uncertainties becomes computationally prohibitive, particularly for high-dimensional parameter spaces and precise forward models [5].

In this paper we consider ill-posed inverse problems that upon discretization yield large systems of linear equations. Such problems formulated as Fredholm integral

N. Polydorides
EEWRC, The Cyprus Institute, Nicosia, Cyprus
e-mail: nickpld@cyi.ac.cy

M. Wang · D. P. Bertsekas
LIDS, MIT, Cambridge, MA, USA
e-mail: mwang@mit.edu; dimitrib@mit.edu

equations of the first kind typically arise in several areas of engineering and natural science including image processing, geophysical prospecting and wave scattering [11]. The main characteristic of these problems is that the integral operator that maps the model parameters to the observed data does not have a continuous inverse and thus a small amount of noise in the data may trigger an arbitrarily large variation in the estimated parameters. This inherent property of ill-posed problems is reflected also in the discrete problem setting causing the coefficients matrix of the respective linear system to be ill-conditioned or singular. Consider for example the integral equation

$$b(y) = \int_{x_1}^{x_2} dx \, \alpha(x, y) f(x) + \eta(y)$$

to which we associate, through a numerical integration rule, the linear model

$$b = Af + \eta \tag{1}$$

where $A \in \Re^{m \times n}$ is a dense ill-conditioned matrix, $b \in \Re^m$ is the data vector, $f \in \Re^n$ is the discretization of the unknown function and $\eta \in \Re^m$ is some additive noise. In order to enforce stability in estimating f from noisy data b one may apply Tikhonov regularization, expressed as a penalized least-squares problem

$$\min_{f \in \Re^n} \|b - Af\|_{\zeta}^2 + \lambda \|f\|^2, \tag{2}$$

where $\zeta \in \Re^m$ is a known probability distribution with positive components and $\lambda \in \Re$ is a positive regularization parameter. This problem is shown to have a unique regularized solution f_t, obtained by solving the linear system

$$(A'ZA + \lambda I) f_t = A'Zb,$$

where $Z \in \Re^{m \times m}$ is the diagonal matrix based on ζ, I is the identity matrix and prime denotes transposition. The value of λ is chosen such that $(A'ZA + \lambda I)$ is full rank and well-conditioned for inversion [2]. When n or m is very large, computing f_t becomes challenging, hence we propose to approximate f_t within a low-dimensional subspace

$$S = \{ \Phi r \mid r \in \Re^s \},$$

where $\Phi \in \Re^{n \times s}$ is a matrix whose columns represent the s discrete basis functions spanning S. The type of basis functions can be arbitrary but we assume throughout that Φ has rank s. Our proposed methodology involves subspace approximation, Monte-Carlo simulation, regression, and most significantly, *only* low-dimensional vector operations, e.g. of order s. Let $\Pi : \Re^n \mapsto S$ be an orthogonal projection operator. By decomposing f to its orthogonal components, $f = \Pi f + (I - \Pi) f$, we have

$$b = A(\Pi f + (I - \Pi) f) + \eta = A\Pi f + \epsilon, \tag{3}$$

where the error term $\epsilon = A(I - \Pi)f + \eta$ encompasses the impact of subspace approximation and the additive noise. By representing Πf as Φr, and applying a Galerkin projection to S weighted by ζ, we obtain

$$c = Gr + z \tag{4}$$

where

$$c = \Phi' A' Z b, \qquad G = \Phi' A' Z A \Phi \qquad z = \Phi' A' Z \epsilon.$$

The new projected operator $G \in \Re^{s \times s}$ is now of moderate dimension but is typically still ill-conditioned and may not be invertible. Suppose that instead of evaluating G and c by performing the high-dimensional matrix products, we use estimators $\hat{G}$ and $\hat{c}$ obtained by stochastic simulation. In such case we formulate the linear model

$$\hat{c} = \hat{G}r + w, \quad \text{where} \quad w = z + (\hat{c} - c) + (G - \hat{G})r. \tag{5}$$

Then an approximate solution r^* can be computed from the regularized regression

$$\min_{r \in \Re^s} \|\hat{G}r - \hat{c}\|_{\Sigma^{-1}}^2 + \lambda \|r - \bar{r}\|^2, \tag{6}$$

where $\Sigma \in \Re^{m \times m}$ is the noise covariance matrix of w and $\bar{r}$ is an initial guess on the solution. With minimal loss of generality we assume that η and ϵ are random variables with zero mean. The simulation-based regularized problem (6) admits the unique solution

$$\hat{r} = (\hat{G}' \Sigma^{-1} \hat{G} + \lambda I)^{-1} (\hat{G}' \Sigma^{-1} \hat{c} + \lambda \bar{r}), \tag{7}$$

although, because w is a function of r (cf. (5)), the noise covariance Σ is a function of the required solution. To overcome this problem one option is to evaluate a constant covariance based on a nominal r, such as the prior for example, yielding $\Sigma = \Sigma(\bar{r})$. Another possibility is a form of iterative regularized regression, whereby we iteratively estimate the optimal solution using an intermediate correction of $\Sigma(r)$ as

$$\hat{r}_{k+1} = \left(\hat{G}' \Sigma(\hat{r}_k)^{-1} \hat{G} + \lambda I\right)^{-1} \left(\hat{G}' \Sigma(\hat{r}_k)^{-1} \hat{c} + \lambda \bar{r}\right), \tag{8}$$

for $k \geq 0$ and $r_0 = \bar{r}$. The iteration was shown to converge locally to a fixed point of (8), provided that a matrix Euclidean norm of $\Sigma(r)$ is sufficiently small [19]. The estimation of $\hat{G}$, $\hat{c}$ and $\Sigma(\hat{r}_k)$ using stochastic simulation is addressed next.

2 Approximation Based on Simulation and Regression

Our approach is based on stochastic simulation. We note that there is a large body of work on the solution of linear systems using Monte Carlo methods, starting with a suggestion by von Neumann and Ulam, as recounted by Forsythe and Leibler [10], (see also Curtiss [6] and the survey by Halton [12]). For a thorough review of the

methods including some important recent developments we refer the readers to the books by Asmussen et al. [1] and Lemieux [15].

Our approach differs from the works just mentioned in that it involves not only simulation, but also approximation of the solution within a low-dimensional subspace in the spirit of Galerkin approximation (see e.g. [5]). We also like to draw the distinction from Markov chain Monte Carlo methods used in the context of linear Bayesian estimation, where the *a posteriori* probability distribution is sampled using, for example, the Metropolis-Hastings or the Gibbs algorithms [14]. Our work is related to the approximate dynamic programming methodology that aims to solve forms of Bellman's equation of very large dimension by using simulation (see the books by Bertsekas and Tsitsiklis [3], and by Sutton and Barto [18]). This methodology has recently been extended to general square systems of linear equations and regression problems in a paper by Bertsekas and Yu [4], which served as a starting point for the present paper.

The use of simulation for linear algebra operations has also been adopted by Drineas et al. in a series of papers [7–9] in the context of randomized algorithms for massive dataset analysis. The authors propose sampling the entries of large matrices, in order to construct new sparser or smaller matrices that behave like the high-dimensional ones. In their analysis they consider products of several matrices where they randomly take samples according to an importance sampling distribution that relates to the Euclidean norms of the columns. In their work they make no assumptions on the matrices, as opposed to our methodology, which applies primarily to matrices of smooth structure like those arising from discretization of Fredholm kernels.

2.1 Markov Chain Monte Carlo Framework

In [4] the authors suggest generating a long finite sequence of indices $\{i_0, \ldots, i_t\}$ according to a nominal probability distribution ξ and two sequences of transitions $\{(i_0, j_0), \ldots, (i_t, j_t)\}$ and $\{(i_0, h_0), \ldots, (i_t, h_t)\}$ according to some transition probabilities ρ_{ij} and ρ_{ih} respectively. This yields estimates of the low-dimensional G and c as

$$\hat{G} = \frac{1}{t+1} \sum_{p=0}^{t} \frac{\zeta_{i_p} a_{i_p j_p} a_{i_p h_p}}{\xi_{i_p} \rho_{i_p j_p} \rho_{i_p h_p}} \phi_{j_p} \phi'_{h_p}, \qquad \hat{c} = \frac{1}{t+1} \sum_{p=0}^{t} \frac{\zeta_{i_p} a_{i_p j_p} b_{i_p}}{\xi_{i_p} \rho_{i_p j_p}} \phi_{j_p}, \quad (9)$$

where a_{ij} denotes the (i, j)th component of A, and ϕ'_j is the jth row of Φ, assuming that $\rho_{ij} > 0$ whenever $a_{ij} > 0$. Apart from the estimators one obtains a sample-based estimator of the covariance given by

$$\Sigma(\hat{r}_k) = \frac{1}{t+1} \sum_{p=0}^{t} w_p w'_p = \frac{1}{t+1} \sum_{p=0}^{t} \big((G_p - \hat{G})\hat{r}_k + (\hat{c} - c_p)\big)\big((G_p - \hat{G})\hat{r}_k + (\hat{c} - c_p)\big)',$$

$$(10)$$

where each w_p can be viewed as a sample of w, while G_p and c_p denote the corresponding sample terms averaged to yield $\hat{G}$ and $\hat{c}$. For further discussion and a derivation of a confidence region for $\hat{r}_k$ obtained by introducing (9) and (10) into (8) we refer to [19]. For the needs of this work, we borrow an important result from [4], in the form of the following theorem.

Theorem 1. *As $t \to \infty$ we have $\hat{G} \to G$, and $\hat{c} \to c$ with probability 1, where $\hat{G}$ and $\hat{c}$ are given by (9).*

Proof. The proof is in [4].

Remark 1. Under the conditions of Theorem 1, if the eigenvalues of the sample-based covariance $\Sigma(\hat{r}_k)$ are bounded below by a positive scalar, then iteration (8) yields $\hat{r}_k \to r^*$ with probability 1, where Φr^* is the target high-dimensional regularized solution.

2.2 Variance Reduction by Importance Sampling

The central idea of our simulation method is to evaluate G and c as weighted averages of samples generated by a probabilistic mechanism. In this context, a critical issue is the reduction of the variance of the estimation errors $\hat{G} - G$ and $\hat{c} - c$. To achieve this goal we use importance sampling, which can be shown to yield estimators of minimal variance when an *optimal* probability distribution is used for generating the samples [12]. Let Ω be a discrete sample space, $\nu : \Omega \mapsto \Re$ be a function and $\{i_0, i_1, \ldots, i_t\}$ be the sequence of samples generated from Ω independently according to distribution ξ. Then consider estimating the large sum $u = \sum_{i \in \Omega} \nu_i$ as

$$\hat{u} = \frac{1}{t+1} \sum_{p=0}^{t} \frac{\nu_{i_p}}{\xi_{i_p}},$$

and designing ξ so that the variance of $\hat{u}$ is minimized. If ν is nonnegative, the variance is

$$\text{var}\{\hat{u}\} = \frac{u^2}{t+1} \left(\sum_{\omega \in \Omega} \frac{(\nu(\omega)/u)^2}{\xi(\omega)} - 1 \right),$$

from where it is now apparent that the choice $\xi^* = \nu u^{-1}$ is the optimal zero-variance sampling distribution. Note that the non-negativity of ν is not critical, for if ν admits negative values, it is trivial to decompose as $\nu = \nu^+ - \nu^-$ so that both ν^+ and ν^- are positive functions. In such a situation $\hat{u}$ is computed by estimating separately $u^+ = \sum_{i \in \Omega} \nu_i^+$ and $u^- = \sum_{i \in \Omega} \nu_i^-$. As is well known, calculating the optimal ξ^* is impractical since it requires knowledge of the unknown sum. However, designing a computationally tractable approximation $\hat{\xi}$ that nearly minimizes the L_1 norm $\|\xi - \nu u^{-1}\|_1$ can be shown to reduce the variance of $\hat{u}$. In the remaining part of

this section we discuss some schemes for designing sampling distributions tailored to the data of the linear ill-posed inverse problems, so to achieve variance reduction.

2.2.1 Designing Importance Sampling Distributions with Polynomial Bases

We focus on estimating the (l, q)th entry of the symmetric, $s \times s$ matrix G and the lth element of vector c independently in an element by element fashion. Noticing that these can be expressed as high-dimensional sums (of dimensions n^3 and n^2 respectively)

$$G_{lq} = \phi_l' A' Z A \phi_q = \sum_{i=1}^{n} \zeta_i \left(\sum_{j=1}^{n} a_{ij} \phi_{jl} \right) \left(\sum_{h=1}^{n} a_{ih} \phi_{hq} \right), \tag{11}$$

$$c_l = \phi_l' A' Z b = \sum_{i=1}^{n} \zeta_i \left(\sum_{j=1}^{n} a_{ij} \phi_{jl} \right) b_i. \tag{12}$$

One may consider a sequence of independent *uniformly* distributed sample indices $\{(i_p, j_p, h_p)\}_{p=0}^{t}$ and $\{(i_p, j_p)\}_{p=0}^{t}$ from the spaces $[1, n]^3$ and $[1, n]^2$, and compute the Monte Carlo estimators

$$\hat{G}_{lq} = \frac{1}{t+1} \sum_{p=0}^{t} \frac{\zeta_{i_p} a_{i_p j_p} a_{i_p h_p} \phi_{j_p l} \phi_{h_p q}}{n^{-3}}, \quad \hat{c}_l = \frac{1}{t+1} \sum_{p=0}^{t} \frac{\zeta_{i_p} a_{i_p j_p} \phi_{j_p l} b_{i_p}}{n^{-2}}.$$

Alternatively, one may design an importance sampling distribution customized for G_{lq} as in (11).[1] In this case let the sample space be $\Omega = [1, n]^3$ and consider the function

$$v(i, j, h) = \zeta_i a_{ij} a_{ih} \phi_{jl} \phi_{hq},$$

assuming for simplicity that $v(i, j, h)$ is nonnegative. The aim here is to construct, in a computationally efficient manner, a sampling distribution $\hat{\xi}$ that approximates the optimal

$$\xi_{G_{lq}}^*(i, j, h) = \frac{v(i, j, h)}{G_{lq}}, \quad \text{where } G_{lq} = \sum_{i,j,h=1}^{n} v(i, j, h) = \sum_{i=1}^{n} \zeta_i \|a_i \phi_l\|_1 \|a_i \phi_q\|_1$$

and belongs to some family of relatively simple distribution functions. In the above a_i is the ith row of A and $\|a_i \phi_l\|_1$ is the L_1 norm of the Hadamard product of a_i and ϕ_l. As it now becomes apparent, ξ^* is not only high-dimensional and impractical

---

[1] Unless otherwise stated, from now on we deal exclusively with G_{lq}. A simplified analysis applies to c_l.

to compute, store and sample, but it also requires n-dimensional vector products and sums. Using Bayes' theorem and the conditional probability law the optimal distribution can be reformulated in a product form as

$$\xi^*(i, j, h) = \xi(h|i, j)\xi(i, j) = \xi(h|i, j)\xi(j|i)\xi(i), \qquad (13)$$

where the marginal distributions are $\xi(i, j) = \sum_{h=1}^{n} \xi(i, j, h)$ and $\xi(i) = \sum_{j=1}^{n} \xi(i, j)$. We propose to approximate ξ^* by approximating the constituent sampling distributions

$$\xi(i) = \frac{v_{hj}(i)}{G_{lq}}, \quad \xi(j|i) = \frac{v_h(i|j)}{\sum_{i=1}^{n} v_h(i, j)}, \quad \xi(h|i, j) = \frac{v(i, j, h)}{\sum_{i,j=1}^{n} v(i, j, h)}, \qquad (14)$$

corresponding to the functions

$$v_{hj}(i) = \sum_{j=1}^{n} v_h(i, j), \quad v_h(i, j) = \sum_{h=1}^{n} v(i, j, h), \quad v(i, j, h) = \zeta_i a_{ij} a_{ih} \phi_{jl} \phi_{hq}. \qquad (15)$$

To accomplish this assume a low-dimensional discretization of the sampling space, for example a uniform cubical grid. For instance let $\Omega = \Omega^k \times \Omega^k \times \Omega^k$, where $\Omega^k = \cup_{i=1}^{K} \Theta_i$, and $\Theta_1, \ldots, \Theta_K$ are K connected disjoint subsets of $[1, n]$. Moreover, let $\psi_i : \Theta_i \to \Re$ be a polynomial function with support over Θ_i and let I_{Θ_i} denote a small nonempty set of points in Θ_i, for $i = 1, \ldots, K$. Then one can approximate v by $\tilde{v}$ using ψ_i and samples of v at I_{Θ_i}. If ψ_i is a constant function then I_{Θ_i} requires only one point, whereas if it is linear then two sample points are needed in each Θ_i and so on with higher degree polynomials. The advantage of using polynomial bases is that the approximate functions in (15) can be summed up to yield the probability distributions in (14) without element-wise summation, since the sums of discrete polynomial functions can be evaluated analytically. It is now easy to see that as the grid dimension grows, i.e. $K \to n$, then $\tilde{v} \to v$, so that the approximate ξ will converge to the optimum ξ^*, albeit with an increase of computational complexity. The suitability of the proposed importance sampling scheme relies predominantly on the ease of forming $\tilde{v}$ using a relatively small K so that $\|\tilde{v} - v\|_1$ is small and therefore so is $\|\hat{\xi} - \xi^*\|_1$. This is fundamentally due to the smooth structure of v, a property that stems from the smooth structure of the model matrix A in (1) (resp. the Fourier series of the Fredholm kernel $\alpha : f \mapsto b$), which *always* holds true in linear ill-posed inverse problems [11]. Once $\hat{G}$ and $\hat{c}$ are estimated, the low-dimensional solution can be computed by (7) or (8). Moreover, since the components of G and c are estimated independently, one may view the samples of G as vectors in $\Re^{s^2}$ that are independent of the samples of c. Thus we can estimate the simulation error covariance by

$$\Sigma(\hat{r}) = \Sigma_c + \begin{bmatrix} \hat{r}' & 0 & \dots & 0 \\ 0 & \hat{r}' & \dots & 0 \\ \ddots & \ddots & \ddots & \ddots \\ 0 & \dots & 0 & \hat{r}' \end{bmatrix} \Sigma_G \begin{bmatrix} \hat{r} & 0 & \dots & 0 \\ 0 & \hat{r} & \dots & 0 \\ \ddots & \ddots & \ddots & \ddots \\ 0 & \dots & 0 & \hat{r} \end{bmatrix}, \qquad (16)$$

where $\Sigma_c \in \Re^{s \times s}$ is the sample-based covariance of c and $\Sigma_G \in \Re^{s^2 \times s^2}$ is the sample-based covariance of G, which is given by

$$\Sigma_G = \begin{bmatrix} \text{cov}(\hat{g}_1', \hat{g}_1') & \text{cov}(\hat{g}_1', \hat{g}_2') & \dots & \text{cov}(\hat{g}_1', \hat{g}_s') \\ \text{cov}(\hat{g}_2', \hat{g}_1') & \text{cov}(\hat{g}_2', \hat{g}_2') & \dots & \text{cov}(\hat{g}_2', \hat{g}_s') \\ \ddots & \ddots & \dots & \ddots \\ \text{cov}(\hat{g}_s', \hat{g}_1') & \text{cov}(\hat{g}_s', \hat{g}_2') & \dots & \text{cov}(\hat{g}_s', \hat{g}_s') \end{bmatrix},$$

where $\text{cov}(\hat{g}_i', \hat{g}_j')$ is the sample covariance between the ith and jth rows of $\hat{G}$.

2.2.2 The Simulation Algorithm

The resulting importance sampling (IS) algorithm for estimating G_{lq} is summarized as follows:

1. Divide the sampling space $[1, n]$ into K disjoint intervals $\Theta_1, \dots, \Theta_K$.
2. Fix $d + 1$ points I_{Θ_i} in $\Theta_i, i = 1, \dots, K$, with $d \geq 0$.
3. Choose the bases of dth order polynomial functions $\psi_i : \Theta_i \mapsto \Re$, $i = 1, \dots, K$.
4. Form the weights matrix $N \in \Re^{(d+1)K \times (d+1)K \times (d+1)K}$ by evaluating $\nu(i, j, h)$ at $I_{\Theta_i} \times I_{\Theta_j} \times I_{\Theta_h}$, for $i, j, h = 1, \dots, K$.
5. Sum N over the h-dimension to get $N_h \in \Re^{(d+1)K \times (d+1)K}$.
6. Sum N_h over the j-dimension to get $N_{hj} \in \Re^{(d+1)K}$.
7. For $p = 0, \dots, t$:

   a. Evaluate the sum $Q = \sum_{i=1}^{(d+1)K} |N_{hj}(i)|$, construct distribution $q(i) = |N_{hj}(i)|/Q$ and take sample s_i from $\cup_{i=1}^{K} I_{\Theta_i}$ according to distribution q.
   b. Let I_{Θ_l} be the set containing s_i, construct the distribution q_l over Θ_l by interpolating with the bases ψ. Sample i_p from Θ_l according to distribution q_l with probability P_{i_p}.
   c. Evaluate the sum $Q = \sum_{j=1}^{(d+1)K} |N_h(s_i, j)|$, and construct $q(j) = |N_h(s_i, j)|/Q$ and take sample s_j from $\cup_{i=1}^{K} I_{\Theta_i}$ according to q.
   d. Let I_{Θ_m} be the set containing s_j, and construct distribution q_m over Θ_m by interpolating. Sample j_p from Θ_m according to q_m with probability P_{j_p}.
   e. Evaluate the sum $Q = \sum_{h=1}^{(d+1)K} |N(s_i, s_j, h)|$, and construct $q(h) = |N(s_i, s_j, h)|/Q$ and take sample s_h from $\cup_{i=1}^{K} I_{\Theta_i}$ according to q.

    f. Let I_{Θ_n} be the set containing s_h, and construct distribution q_n over Θ_n by interpolating. Sample h_p from Θ_n according to q_n with probability P_{h_p}.

    g. Register sample (i_p, j_p, h_p) with probability $\xi(i_p, j_p, h_p) = P_{i_p} P_{j_p} P_{h_p}$ and evaluate $v_p = \zeta_{i_p} a_{i_p j_p} a_{i_p h_p} \phi_{j_p l} \phi_{h_p q}$.

    h. Evaluate pth sample mean:

        i. $\hat{G}_{lq} = v_p / \xi(i_p, j_p, h_p)$ if $p = 0$.

        ii. $\hat{G}_{lq} = \frac{p}{p+1}\hat{G}_{lq} + \frac{1}{p+1}v_p / \xi(i_p, j_p, h_p)$ if $p > 0$.

8. End sampling.
9. Evaluate t-sample variance $\mathrm{var}(\hat{G}_{lq})$.
10. Evaluate the total error covariance using (16).
11. Compute the solution approximation from (7) or (8).

3 Discrete Linear Inverse Problems

Linear ill-posed inverse problems typically occur in applications of image processing, emission tomography, wave diffraction, palaeo-climatology, and heat transfer, and are usually expressed in Fredholm integral equations. Discretizing these equations yields linear systems with ill-conditioned coefficient matrices. This is an inherent characteristic of ill-posed problems and has been analyzed in various publications, including [11] and [2] which emphasize its implications to the existence, uniqueness and stability of the solution. In particular, the condition number of the coefficient matrix obtained by discretization can be shown to increase with the dimension n, sometimes at an exponential rate in which case the problem is said to be heavily ill-posed.

    In our development we have assumed the structure of the matrix A to be smooth, implying that neighboring entries have almost identical values. This property is due to the spectral properties of the Fredholm operators in consideration. Figure 1 illustrates this effect on a moderately sized discretized kernel $A \in \Re^{n \times n}$ with $n = 10^3$ for a problem arising from geophysics. A large-scale numerical study based on this model problem is investigated next.

3.1 Test Example: Gravitational Prospecting

Gravitational prospecting is a problem typically encountered in hydrocarbon exploration. Suppose a mass of density $f(\theta)$ is distributed on a circular ring O_i of radius r_i centered at the origin, where $0 \leq \theta \leq 2\pi$. Allow also a concentric circle O_o of radius r_o, with $r_o \ll r_i$ lying on the same plane, where the centrally directed component of gravitational force $b(\phi)$ is measured, for $0 \leq \phi \leq 2\pi$. According to the law of cosines the squared distance between a mass element situated on O_i at an angle θ and a point of O_o at ϕ is
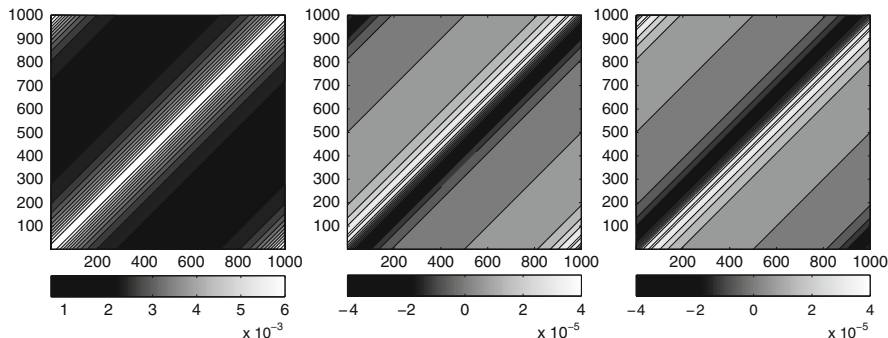
Fig. 1 Contour plots of the elements of A (*left*) for a smaller scale problem with $n = 10^3$ and its gradients in row index i (*middle*) and column index j (*right*) to indicate the smooth structure of the model matrix as manifested by the flat regions in the gradient plots. At dimension $n = 10^3$ the condition number of the A is 2.2×10^{20}.

$$\rho_t^2 = r_o^2 + r_i^2 - 2r_o r_i \cos(\theta - \phi),$$

while the angle χ formed between the normal component of the gravity force at ϕ and the line connecting that point to the gravitating element $f(\theta)d\theta$ of the inner ring satisfies

$$\rho_t \cos(\chi) = r_o - r_i \cos(\theta - \phi).$$

In effect, the overall gravitational force exerted at the measuring angle ϕ is

$$b(\phi) = \gamma \int_0^{2\pi} d\theta \frac{1}{2\rho_t^2} \cos(\chi) f(\theta),$$

where γ is the universal gravity constant. Taking for simplicity $\gamma = 1$, $r_o = 1$ and $r_i = 0.5$ yields the Fredholm equation

$$b(\phi) = \int_0^{2\pi} d\theta \, \alpha(\phi, \theta) \, f(\theta), \qquad 0 \le \phi \le 2\pi,$$

with kernel

$$\alpha(\phi, \theta) = \frac{2 - \cos(\phi - \theta)}{\left(5 - 4\cos(\phi - \theta)\right)^{3/2}}.$$

The integral equation is discretized using a midpoint quadrature rule on a uniform grid with $n = 10^6$ points $\{\theta_i, \phi_j\}_{i,j=1}^n$ spanning over $[0, 2\pi] \times [0, 2\pi]$. To approximate the solution we choose a subspace spanned by an orthogonal basis of $s = 10^2$ piecewise constant functions, and consider reconstructing a subspace approximation of the regularized density f given data $b \in \Re^n$. To test the performance of the proposed scheme in reducing the simulation noise we run the Algorithm 2.2.2 for various t and K, each time using piecewise constant, linear
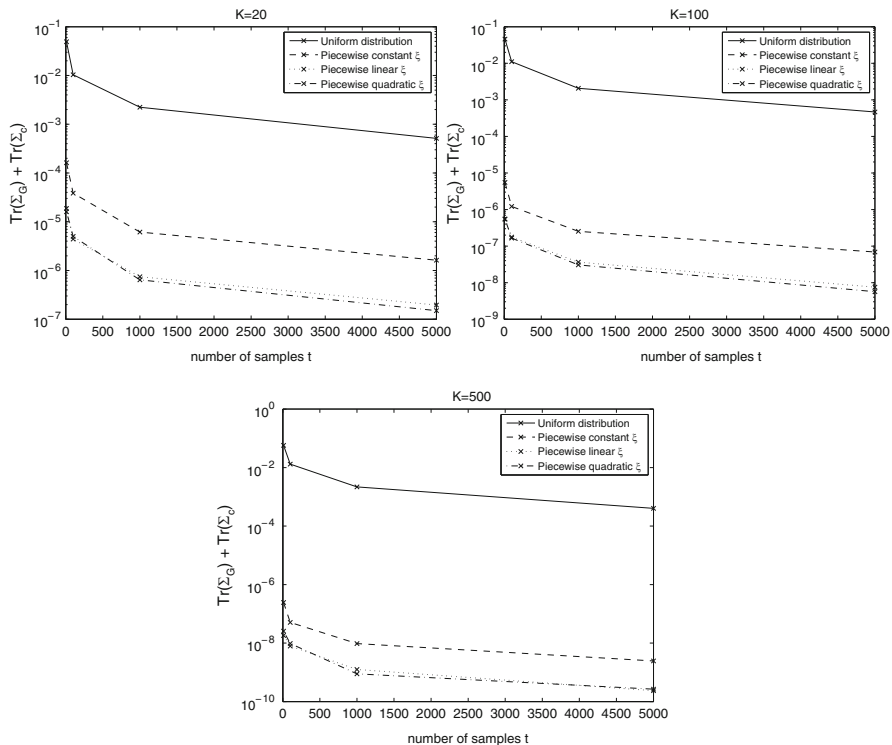
Fig. 2 Reduction in simulation noise $Tr(\Sigma_G) + Tr(\Sigma_c)$ with the number of acquired samples. From the *left* the cases with $K = 20, 100$ and 500 intervals, assuming $s = 10^2$ and $n = 10^6$. In each graph the *solid line* is with the naive Monte Carlo sampling (uniform distribution), the *dashed* for a piecewise constant approximation of the optimum IS distribution ξ^*, the *dotted* for a piecewise linear and the *dash-dotted* for a quadratic approximation of the optimum IS distribution. Notice that the simulation error reduces in increasing K and in implementing a higher-order approximation of the optimal IS distribution. In all cases the proposed scheme outperforms the naive Monte Carlo sampling in reducing the variance of the estimators.

and quadratic basis functions for approximating the optimal importance sampling distribution. The graphs of Fig. 2 illustrate the reduction of the simulation noise, quantified in terms of the sum of the traces of the two sample-based covariances as it is affected by t and K. Notice that $Tr(\Sigma_G) + Tr(\Sigma_c)$ reduces with increasing the number of samples and/or the degree of the polynomials ψ used in approximating the optimal distribution. The corresponding graphs obtained with uniform sampling are plotted for comparison in order to show the superiority of the importance sampling scheme.

In our numerical tests we choose not to add any measurement noise, i.e. $\eta = 0$. When solving ill-posed inverse problems with synthetic data the noise free case is considered as a "contrived" example as it allows for processing unrealistically precise data, (see for example Chap. 5 in [14]). On the other hand, there is a large
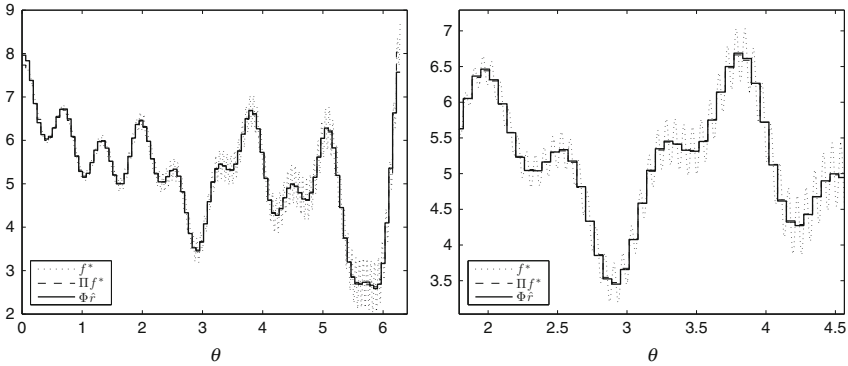
Fig. 3 *Left*, the true image f^*, its projection Πf^* in S spanned by $s = 10^2$ piecewise constant orthogonal basis functions, and the result of the simulated approximation $\Phi \hat{r}$. Angle $\theta \in [0, 2\pi]$ is discretized in $n = 10^6$ elements. To the right a more detailed view of the results for $2 \leq \theta \leq 4.5$. Notice that the curves of Πf^* and $\Phi \hat{r}$ are almost overlapping.

body of literature on how to adjust the regularization parameter λ in (6) so as to counteract the impact of noise and stabilize the solution. An extensive survey of such methods is presented in Chap. 5 of [13]. In particular, notice that the problem under consideration involves a square linear system of manageable dimension, where the data vector $\hat{c}$ and coefficients matrix $\hat{G}$ include simulation errors for which we can estimate their element-wise variance based on samples. Moreover, $\hat{G}$ is symmetric and ill-conditioned and the overall noise includes the subspace approximation error and any additive noise contained in the original data b. In this context, to choose λ we adopt the discrete Picard condition [17], which relies on the singular value decomposition of the low-dimensional $\Sigma(\hat{r}_k)^{-1/2}\hat{G}$, implemented after each iteration (8).

In this study we focus on demonstrating the performance of the Algorithm 2.2.2 in estimating G and c with reduced simulation error. In particular, our claim is that for $\eta = 0$ and a sufficiently small $\lambda > 0$, as the number of samples increases the recursive formula (7) will generate a solution $\hat{r}$ such that $\Phi \hat{r} \rightarrow \Phi r^*$. In turn, this relies on reducing the simulation noise as illustrated by the graphs of Fig. 2. Moreover, notice that in realistic experimental conditions, physical noise and measurement precision are likely to result in $\|\eta\| \gg Tr(\Sigma_G) + Tr(\Sigma_c)$. In this case the covariance of the overall noise in (5) will be predominantly determined by that of η.

The results presented in Fig. 3 have been obtained after implementing Algorithm 2.2.2 with $t = 2 \times 10^3$, $K = 10^2$, $n = 10^6$, $s = 10^2$, ψ piecewise linear, and $\lambda = 10^{-7}$. The figure shows for comparison the true solution f^* used to compute the data b, its subspace projection Πf^* and subspace approximation $\Phi \hat{r}$ as computed by introducing $\hat{G}$, $\hat{c}$, Σ_G, and Σ_c into (8) after only a single iteration. The similarity between Πf^* and $\Phi \hat{r}$ is indicative of the small variance in the estimated $\hat{G}$ and $\hat{c}$. The total computation time, almost exclusively dissipated

in estimating the upper triangular part of $\hat{G}$ (5,150 entries) and the 10^2 entries of c was about 8.5 h on a 2.66 GHz quad processor computer with 4 GB RAM running Matlab [16].

4 Special Case: Underdetermined Problems

Quite often, practical limitations impose a limit to the amount of data that can realistically be measured to estimate a certain set of parameters. By contrast, there is always a quest for increasing the amount of information extracted from an inverse solution, e.g. in terms of its degrees of freedom or resolution. This mismatch in the dimensions of the parameter and data spaces evidently yields underdetermined inverse problems. These problems are addressed in the context of the minimum-norm Backus-Gilbert regularization method [2].

In dealing with severely underdetermined problems, one can implement our algorithm to estimate the components of the high-dimensional solution directly, without the need for subspace approximation. Assuming now that $A \in \Re^{s \times n}$ where s is reasonably small and n is very large by comparison, we may adapt the preceding methodology to estimate f_t from

$$(A'ZA + \lambda I)f_t = A'Zb. \tag{17}$$

Using the matrix inversion lemma [14], the solution can also be expressed as

$$f_t = A'(AA' + \lambda Z^{-1})^{-1}b,$$

which by contrast to (17) requires only the inversion of a low-dimensional matrix. Using this lemma, it is also easy to prove that if $(A'ZA + \lambda I)$ is well conditioned then so is $(AA' + \lambda Z^{-1})$. In such a case the s-dimensional matrix we seek to estimate by simulation is $G = AA'$, whose element G_{lq} we express as a finite sum of functions

$$G_{lq} = \sum_{i,j=1}^{n} v(i,j) = \sum_{i,j=1}^{n} a_{li}a_{qj}.$$

To obtain an approximation to the optimal importance sampling distribution for v we work similar to the algorithm described above, essentially dividing the sampling space $[1, n]$ into K disjoint intervals $\Theta_1, \ldots, \Theta_K$, where we take a number of arbitrary points I_{Θ_i} and interpolate d-degree polynomial functions in each Θ_i. After t samples, the importance sampling yields the estimator given by

$$\hat{G}_{lq} = \frac{1}{t+1} \sum_{p=0}^{t} \frac{a_{li_p}a_{qi_p}}{\xi(i_p, j_p)}, \qquad l, q = 1, \ldots, s,$$

and its t-sample variance $\text{var}(\hat{G}_{lq})$. Consequently the ith element of the solution can be computed by

$$f_i = a_i'(\hat{G} + \lambda Q^{-1})^{-1}b, \tag{18}$$

where $a_i \in \Re^s$ is the ith column of A and Q is the noise covariance encompassing the additive noise and the simulation error. Notice that since we do not use subspace approximation, the approximation error is essentially zero.

5   Conclusions and Future Directions

In this paper, we have considered the approximate solution of linear inverse problems within a low-dimensional subspace spanned by a given set of basis functions. We have proposed a simulation-based regularized regression approach that involves importance sampling and low-dimensional computation, and that relies on designing sampling distributions customized to the model matrices and basis functions spanning the subspace. We have elaborated on a few approaches for designing near-optimal sampling distributions, which exploit the continuous structure of the underlying models. The performance of our method has been evaluated with a number of numerical tests using a classical inverse problem. The computation experiments demonstrate an adequate reduction of simulation error after a relatively small number of samples and an attendant improvement in quality of the obtained approximate solution.

A central characteristic of our methodology is the use of low-dimensional calculations in solving high-dimensional problems. Two important approximation issues arise within this context: first the solution of the problem should admit a reasonably accurate representation in terms of a relatively small number of basis functions, and second, the problem should possess a reasonably continuous/smooth structure so that effective importance sampling distributions can be designed with relatively small effort. In our computational experiments, simple piecewise polynomial approximations have proved adequate, but other more efficient alternatives may be possible. We finally note that the use of regularized regression based on a sample covariance obtained as a byproduct of the simulation was another critical element for the success of our methodology with nearly singular problems.

References

1. Asmussen, S., and Glynn, P. W.: Stochastic Simulation. Springer, New York (2007)
2. Bertero, M., and Boccacci, P.: Introduction to Inverse Problems in Imaging. IoP, Bristol (2002)
3. Bertsekas, D.P., and Tsitsiklis, J.: Neuro-Dynamic Programming. Athena Scientific (1996)

4. Bertsekas D.P., and Yu, H.: Projected Equation Methods for Approximate Solution of Large Linear Systems. J. Comp. Appl. Math., **227**, 27–50 (2009)
5. Biegler, L., Biros, G., Ghattas, O., Heinkenschloss, M., Keyes, D., Mallick, B., Marzouk, Y., Tenorio, L., van Blomen Waanders, B., and Willcox, K. (eds.): Large-Scale Inverse Problems and Quantification of Uncertainty. Wiley, Chichester (2011)
6. Curtiss, J. H.: Monte Carlo Methods for the Iteration of Linear Operators. J. Math. Phys., **32(4)**, 209–232 (1953)
7. Drineas, P., Kannan, R., and Mahoney, M.W.: Fast Monte Carlo Algorithms for Matrices I: Approximating Matrix Multiplication. SIAM J. Comput. **36**, 132–157 (2006)
8. Drineas, P., Kannan, R., and Mahoney M.W.: Fast Monte Carlo Algorithms for Matrices II: Computing a Low-Rank Approximation to a Matrix. SIAM J. Comput. **36**, 158–183 (2006)
9. Drineas, P., Kannan, R., and Mahoney M.W.:, Fast Monte Carlo Algorithms for Matrices III: Computing a Compressed Approximate Matrix Decomposition. SIAM J. Comput. **36**, 184–206 (2006)
10. Forsythe, G.E., and Leibler, R.A.: Matrix Inversion by a Monte Carlo Method. Math. Tabl. Aids to Comp., **6(38)**, 78–81 (1950)
11. Groetsch, C.W.: The Theory of Tikhonov Regularization for Fredholm Equations of the First Kind. Pitman, London (1984)
12. Halton, J.H.: A Retrospective and Prospective Survey of the Monte Carlo Method. SIAM Review, **12(1)** (1970)
13. Hansen, P.C.: Discrete Inverse Problems: Insight and Algorithms. SIAM, Philadelphia (2010)
14. Kaipio, J., and Somersalo, E.: Statistical and Computational Inverse Problems. Springer, New York (2004)
15. Lemieux, C.: Monte Carlo and Quasi-Monte Carlo Sampling, Springer, New York (2009)
16. Matlab, The Mathworks Ltd
17. O'Leary, D.P.: Near-optimal Parameters for Tikhonov and Other Regularization Methods. SIAM J. on Scientific Computing, **23(4)**, 1161–1171 (2001)
18. Sutton, R.S., and Barto, A.G.: Reinforcement Learning: An Introduction. The MIT Press, Boston (1998)
19. Wang, M., Polydorides, N., and Bertsekas, D.P.: Approximate Simulation-Based Solution of Large-Scale Least Squares Problems, Report LIDS-P-2819, MIT (2009)