Approximate Simulation-Based Solution of Large-Scale Least Squares Problems [*]

Mengdi Wang, Nicolas Polydorides, Dimitri P. Bertsekas [†]

September 2009

Abstract

We consider linear least squares problems of very large dimension, such as those arising for example in inverse problems. We introduce an associated approximate problem, within a subspace spanned by a relatively small number of basis functions, and solution methods that use simulation, importance sampling, and low-dimensional calculations. The main components of this methodology are a regression/regularization approach that can deal with nearly singular problems, and an importance sampling design approach that exploits existing continuity structures in the underlying models, and allows the solution of very large problems.

1 Introduction

We consider linear least squares problems of the form

$$\min_{x \in \Re^n} \|Ax - b\|_\zeta^2, \tag{1}$$

where A is an $m \times n$ matrix, b is a vector in $\Re^m$, ζ is a known probability distribution vector with positive components, and $\|\cdot\|_\zeta$ denotes the corresponding weighted Euclidean norm (throughout the paper, all vectors are viewed as column vectors, a prime denotes transposition, and $\|\cdot\|$ denotes the standard unweighted Euclidean norm). We focus on the case where n is

[*] Research supported by the LANL Information Science and Technology Institute, and by NSF Grant ECCS-0801549. Thanks are due to Janey Yu for helpful discussions.

[†] Laboratory for Information and Decision Systems, M.I.T, Cambridge, Mass. 02139.

very large, and we consider an approximation to problem (1), defined over a subspace

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

where Φ is an $n \times s$ matrix whose columns can be viewed as basis functions. A companion paper [PWB09] discusses a similar methodology for the case of overdetermined problems, where m is very large, and n is comparatively small; in this case S is the entire space $\Re^n$ ($\Phi = I$). An important special case where n is large is the solution of a square linear system $Ax = b$ resulting from fine discretization of a continuous operator equation such as a partial differential or integral equation. There is effectively no limit on the dimension of such a system, so the solution by conventional methods may not be possible. Problems of this type are often additionally complicated by near-singularity of the matrix A, which in turn implies near singularity of the corresponding least squares problem (1).

Our approach is based on Monte Carlo simulation. We note that there is a large body of work on the solution of linear systems of equations by using Monte Carlo methods, starting with a suggestion by von Neumann and Ulam, as recounted by Forsythe and Leibler [FL50], and Wasow [Was52] (see also Curtiss [Cur53], [Cur54], and the survey by Halton [Hal70]). We also note recent work on simulation methods that use low-order calculations for solving overdetermined least squares problems [SV09].

Our work 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 and the Petrov-Galerkin method (see e.g., [KZ72]). Our approach is also 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 [BT96], and by Sutton and Barto [SB98]). This methodology was recently extended to apply to general square systems of linear equations and regression problems in a paper by Bertsekas and Yu [BY09], which served as a starting point for the present paper.

The first step in our approach is to substitute Φr in place of x in problem (1) and consider the problem

$$\min_{r \in \Re^s} \|A\Phi r - b\|_\zeta^2. \tag{2}$$

If the solution is unique, it is given by

$$r^* = G^{-1}c, \tag{3}$$

where

$$G = \Phi' A' Z A \Phi, \qquad c = \Phi' A' Z b, \tag{4}$$

Z is the diagonal $m \times m$ matrix having the components of ζ along the diagonal. The vector Φr^* is viewed as an approximation to an exact solution x^* of the least squares problem (1). The paper by Yu and Bertsekas [YB08] and the report by Bertsekas and Yu [BY07] provide bounds on the error $\Phi r^* - x^*$, which involve the weighted Euclidean distance of x^* from S.

The expressions in Eq. (4) involve the formation of sums of a large number of terms (multiple summations of inner products of dimension n), so when n is very large, the direct calculation of G and c is prohibitively expensive. This motivates a simulation-based approach, analogous to Monte Carlo integration, which aims at a running time that is independent of n, but instead depends on the variance of the simulated random variables. The idea is that by using any positive probabilities ξ_i, a sum of a large number of terms $\sum_i v_i$ can be written as the expected value $\sum_i \xi_i (v_i/\xi_i)$, which can be estimated by sampling the values v_i/ξ_i according to the probabilities ξ_i.

In particular, to estimate the entries of G and c by simulation, we write

$$G = \sum_{i=1}^n \sum_{j=1}^n \sum_{\bar{j}=1}^n \zeta_i a_{ij} a_{i\bar{j}} \phi(j) \phi(\bar{j})', \qquad c = \sum_{i=1}^n \sum_{j=1}^n \zeta_i a_{ij} b_j \phi(j), \tag{5}$$

where ζ_i are the diagonal components of Z, a_{ij} are the components of A, and $\phi(j)'$ is the jth s-dimensional row of Φ:

$$\phi(j)' = \big(\Phi_{j1} \cdots \Phi_{js}\big)$$

where $\Phi_{j\ell}$ are the corresponding scalar components of Φ. As suggested in [BY09], to estimate a single scalar component of G, we may generate a sequence of index triples $\{(i_1, j_1, \bar{j}_1), \ldots, (i_T, j_T, \bar{j}_T)\}$ by independently sampling according to some distribution ξ from the set of triples of indices $(i, j, \bar{j}) \in \{1, \ldots, n\}^3$. We may then estimate the ℓth-row-qth-column component of G,

$$G_{\ell q} = \sum_{i=1}^n \sum_{j=1}^n \sum_{\bar{j}=1}^n \zeta_i a_{ij} a_{i\bar{j}} \Phi_{j\ell} \Phi_{\bar{j}q} \tag{6}$$

[cf. Eq. (5)], with $\hat{G}_{\ell q}$ given by

$$\hat{G}_{\ell q} = \frac{1}{T} \sum_{t=1}^T \frac{\zeta_{i_t} a_{i_t j_t} a_{i_t \bar{j}_t} \Phi_{j_t \ell} \Phi_{\bar{j}_t q}}{\xi_{i_t, j_t, \bar{j}_t}}, \tag{7}$$

where $\xi_{i,j,\bar{j}}$ denotes the probability of the index triple $(i, j, \bar{j})$. Similarly, to estimate a single scalar entry of c, we may generate a sequence of index pairs $\{(i_1, j_1), \ldots, (i_T, j_T)\}$ by independently sampling according to some distribution ξ from the set of pairs of indices $(i, j) \in \{1, \ldots, n\}^2$, and then estimate the ℓth component of c with $\hat{c}_\ell$ given by

$$\hat{c}_\ell = \frac{1}{T} \sum_{t=1}^{T} \frac{\zeta_{i_t} a_{i_t j_t} b_{i_t} \Phi_{j_t \ell}}{\xi_{i_t, j_t}}, \tag{8}$$

where $\xi_{i,j}$ denotes the probability of the index pair (i, j). We then approximate r^* by $\hat{r} = \hat{G}^{-1}\hat{c}$, where $\hat{G}$ and $\hat{c}$ are the matrix and vector with components $\hat{G}_{\ell q}$ and $\hat{c}_\ell$. Note that all the above calculations are low-dimensional. Furthermore, a comparison of Eq. (5) and Eqs. (7)-(8), and a simple law of large numbers argument shows that $\hat{G} \to G$, $\hat{c} \to c$, and $\hat{G}^{-1}\hat{c} \to r^*$ as $t \to \infty$, with probability 1.

Note that there are several options for estimation of components of G and c. At one extreme we may generate a single sequence of index triples $\{(i_1, j_1, \bar{j}_1), \ldots, (i_T, j_T, \bar{j}_T)\}$, which can be used to simultaneously estimate all components of G and c, by using Eqs. (7) and (8). At the opposite extreme we may generate a special sequence of index triples to estimate each component of G and c separately, by using Eqs. (7) and (8). There are also intermediate possibilities whereby blocks of components of G are c are simultaneously estimated with a single sequence of index triples. The tradeoff involved is that grouping components into small blocks costs more in simulation overhead, but may result in variance reduction (and smaller number of required samples for a given degree of accuracy) by tailoring the sampling distribution to the structure of the block and the sparsity structure of Φ, based on importance sampling principles (see Section 3). For example, when estimating a component $G_{\ell q}$ using Eq. (7), it is inefficient to generate sample triples $(i, j, \bar{j})$ for which $\Phi_{j\ell}\Phi_{\bar{j}q} = 0$.

The preceding approach must contend with two main difficulties:

(a) *The approximation error* associated with restricting the solution to lie in the subspace S. This has to do with the choice of the matrix Φ, and is an important, likely problem-dependent issue, which however we do not discuss in this paper.

(b) *The simulation error* associated with replacing G and c with sampling approximations $\hat{G}$ and $\hat{c}$. For an accurate solution, the amount of sampling required may be excessive, and this difficulty is exacerbated in the common case where G is nearly singular.

We focus on the second difficulty, and we address it in two ways. First, rather than approximating r^* with $\hat{r} = \hat{G}^{-1}\hat{c}$, we use a regression/regularization approach. We write the equation $c = Gr$ as

$$\hat{c} = \hat{G}r + e, \tag{9}$$

where e is the vector

$$e = (G - \hat{G})r + \hat{c} - c, \tag{10}$$

which we view as "simulation noise." We then estimate the solution r^* based on Eq. (9) by using regression, and an approximate sample covariance for e, which is available at essentially no cost as a by-product of the simulation used to obtain $\hat{G}$ and $\hat{c}$.[1] This methodology is discussed in Section 2.

Second, to reduce the effect of the components $(G-\hat{G})$ and $(c-\hat{c})$ of the simulation noise e [cf. Eq. (10)], we employ variance reduction techniques, based on importance sampling ideas. We discuss the corresponding methods and analysis in Sections 3, and in Section 4 we derive confidence regions that quantify the effect of near-singularity of G, and the sample covariances of G and c.

In summary, the contributions of Sections 2-4 are three-fold:

- The development of the necessary ingredients for a simulation-based solution methodology that can address very large least squares problems. These include:
  - A regression/regularization approach that can reduce the solution error $(r^* - \hat{r})$ by reducing the effect of the simulation noises $(G - \hat{G})$ and $(c-\hat{c})$ through the use of the sample covariances of G and c, and by reducing the effect of near singularity of G (Section 2).
  - Nearly optimal importance sampling schemes that can effectively reduce the variances of the components of $\hat{G}$ and $\hat{c}$ (Section 3).
- The development of analytical tools that motivate efficient sampling schemes. In particular, we propose a normalized measure of quality of a sampling distribution, called *divergence factor*, which is used for the design of near-optimal distributions (Section 3).

[1] Given independent samples $v_1, \ldots, v_T$ of a random variable v, by "sample variance of v" we mean the scalar

$$\frac{1}{T}\sum_{t=1}^{T}(v_t - \hat{v})^2,$$

where $\hat{v}$ is the sample mean $\hat{v} = (1/T)\sum_{t=1}^{T} v_t$. The sample covariance of a random vector is defined analogously.

- The derivation of confidence regions that quantify the effect of near-singularity of G on the variance of the error $(\hat{r} - r^*)$ (Section 4).

The regression and variance reduction ideas of Sections 2-4 are brought together in an algorithmic methodology that is successfully applied in Section 5 to some standard examples of inverse problems of very large dimension ($n \geq 10^9$). The regression approach of Section 2 and the confidence region analysis of Section 4 also apply to the more general system $\hat{G}r = \hat{c}$, where $\hat{c}$ and $\hat{G}$ are simulation-based approximations to a vector c and an $s \times s$ matrix G that is not necessarily positive definite or symmetric.

2 Regression Methodology

Let us consider the estimation of r^* using the model

$$\hat{c} = \hat{G}r + e,$$

[cf. Eq. (9)], where

$$e = (G - \hat{G})r + \hat{c} - c$$

[cf. Eq. (10)]. The standard least squares/regression approach yields the estimate

$$\hat{r} = \arg\min_{r} \big\{ (\hat{G}r - \hat{c})'\Sigma^{-1}(\hat{G}r - \hat{c}) + (r - \bar{r})'\Gamma^{-1}(r - \bar{r}) \big\},$$

where $\bar{r}$ is an *a priori* estimate (for example some simple least squares-based approximation of a solution of $\hat{G}r = \hat{c}$), and Σ and Γ are some positive definite symmetric matrices. Equivalently,

$$\hat{r} = (\hat{G}'\Sigma^{-1}\hat{G} + \Gamma^{-1})^{-1}(\hat{G}'\Sigma^{-1}\hat{c} + \Gamma^{-1}\bar{r}). \tag{11}$$

We propose to use as Σ an estimate of the covariance of e, which we can obtain as a byproduct of the simulation. In particular, at the end of the simulation, we have the samples $\big\{(G_t, c_t) \mid t = 1, \ldots, T\big\}$, where the components of the matrix G_t and the vector c_t are the terms appearing in the summations of Eqs. (7) and (8), respectively:

$$G_{t,\ell q} = \frac{\zeta_{i_t} a_{i_t j_t} a_{i_t \bar{j}_t} \Phi_{j_t \ell} \Phi_{\bar{j}_t q}}{\xi_{i_t, j_t, \bar{j}_t}}, \qquad c_{t,\ell} = \frac{\zeta_{i_t} a_{i_t j_t} b_{i_t} \Phi_{j_t \ell}}{\xi_{i_t, j_t}}.$$

We make a choice $\tilde{r}$ of a fixed nominal value/guess of r (for example $\tilde{r} = \hat{G}^{-1}\hat{c}$) and we view the vectors

$$e_t = (G_t - \hat{G})\tilde{r} + (\hat{c} - c_t), \qquad t = 1, \ldots, T,$$

as samples of e, with sample mean equal to 0 (by the definition of $\hat{G}$ and $\hat{c}$), and we use as estimate of the covariance of e the corresponding sample covariance matrix

$$\Sigma = \frac{1}{T}\sum_{t=1}^{T} e_t e_t' = \frac{1}{T}\sum_{t=1}^{T}\big((G_t - \hat{G})\tilde{r} + (\hat{c} - c_t)\big)\big((G_t - \hat{G})\tilde{r} + (\hat{c} - c_t)\big)'. \quad (12)$$

In our experiments we have estimated all the components of G and all the components of c independently. For this case, we view samples of G as vectors in $\Re^{s^2}$ that are independent of the samples of c, since G and c are estimated separately. We then calculate Σ using the sample covariances of G and c, and a nominal value of r. In particular, we have

$$\Sigma = \Sigma_c + \begin{bmatrix} r' & 0 & \dots & 0 \\ 0 & r' & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & r' \end{bmatrix}_{s\times s^2} \Sigma_G \begin{bmatrix} r & 0 & \dots & 0 \\ 0 & r & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & r \end{bmatrix}_{s^2\times s} \quad (13)$$

where Σ_c is the sample covariance of c, and Σ_G is the $s^2 \times s^2$ sample covariance of G, given by

$$\Sigma_G = \begin{bmatrix} \mathrm{cov}(\hat{g}_1', \hat{g}_1') & \mathrm{cov}(\hat{g}_1', \hat{g}_2') & \dots & \mathrm{cov}(\hat{g}_1', \hat{g}_s') \\ \mathrm{cov}(\hat{g}_2', \hat{g}_1') & \mathrm{cov}(\hat{g}_2', \hat{g}_2') & \dots & \mathrm{cov}(\hat{g}_2', \hat{g}_s') \\ \dots & \dots & \dots & \dots \\ \mathrm{cov}(\hat{g}_s', \hat{g}_1') & \mathrm{cov}(\hat{g}_s', \hat{g}_2') & \dots & \mathrm{cov}(\hat{g}_s', \hat{g}_s') \end{bmatrix}, \quad (14)$$

where $\mathrm{cov}(\hat{g}_i', \hat{g}_j')$ is the sample covariance between the i^{th} and j^{th} rows of $\hat{G}$. Note that the sample covariances Σ_c and Σ_G are available as a byproduct of the simulation used to calculate $\hat{G}$ and $\hat{c}$. Moreover, the size of these covariances can be controlled and can be made arbitrarily small by taking a sufficiently large number of samples.

An alternative to using a guess $\tilde{r}$ of r and calculating Σ according to Eq. (12), is to use an *iterative regression* approach: iterate using Eq. (11), and estimate r repeatedly with intermediate correction of the matrix Σ. This is the iteration

$$r_{k+1} = \big(\hat{G}'\Sigma(r_k)^{-1}\hat{G} + \Gamma^{-1}\big)^{-1}\big(\hat{G}'\Sigma(r_k)^{-1}\hat{c} + \Gamma^{-1}\bar{r}\big), \quad (15)$$

where for any r, the matrix $\Sigma(r)$ is given by

$$\Sigma(r) = \frac{1}{T}\sum_{t=1}^{T}\big((G_t - \hat{G})r + (\hat{c} - c_t)\big)\big((G_t - \hat{G})r + (\hat{c} - c_t)\big)';$$

[cf. Eq. (12)].

It can be shown that this iteration converges locally in the following sense: given any initial estimate r_0, it generates a sequence $\{r_k\}$ that converges to a fixed point $\hat{r}$ satisfying

$$\hat{r} = \big(\hat{G}'\Sigma(\hat{r})^{-1}\hat{G} + \Gamma^{-1}\big)^{-1}\big(\hat{G}'\Sigma(\hat{r})^{-1}\hat{c} + \Gamma^{-1}\bar{r}\big)$$

provided that the sample covariances of the entries of G and c are below a sufficiently small threshold.

A precise statement and a detailed proof of this local convergence property is outside our scope, so we just provide a heuristic argument. If $\eta \geq 0$ is an upper bound to the sample covariances of the components of G, then from Eq. (13), $\Sigma(r)$ is written as

$$\Sigma_c + \Psi_{r,\eta},$$

where $\Psi_{r,\eta}$ is a matrix satisfying $\|\Psi_{r,\eta}\| \leq q\|r\|^2\eta$ for some constant q. When $\eta = 0$, $\Sigma(r)$ is the constant Σ_c [cf. Eq. (13)], so the mapping of Eq. (15),

$$r \mapsto \big(\hat{G}'\Sigma(r)^{-1}\hat{G} + \Gamma^{-1}\big)^{-1}\big(\hat{G}'\Sigma(r)^{-1}\hat{c} + \Gamma^{-1}\bar{r}\big),$$

is a constant mapping (independent of r), and hence it is a contraction of modulus 0. It follows that for small η, this mapping is also a contraction for r within a given bounded region. This essentially guarantees the local convergence property stated earlier.

Our regression approach was used successfully in large-scale practical inverse problems, some of which are given in Section 5 and some others are discussed in the companion paper [PWB09]. The analysis of Section 4 provides an analytical justification. It shows that the error $(r^* - \hat{G}^{-1}\hat{c})$ is strongly affected by the norm of G^{-1}, so if G is near-singular the error can be very large even when the number of samples used is large; this is consistent with long-standing experience in solving linear equations. In this case, by using a regularization term, we can greatly reduce the error variance at the expense of a relatively small bias in the estimates $\hat{G}$ and $\hat{c}$. While the choice of the regularization matrix Γ is not clear a priori, this should typically not be a major problem, because trial-and-error experimentation with different values of Γ involves low-dimensional linear algebra calculations once $\hat{G}$ and $\hat{c}$ become available.

Also, using the sample covariances of G and c in place of some other positive definite matrices makes sense on intuitive grounds, and has resulted in substantial benefits in terms of solution error variance. This was empirically verified with the examples of Section 5, as well as with small test problems. Let us also note that in our tests, the iterative regression scheme (15), when it converged, gave on the average a small improvement in the quality of the estimate $\hat{r}$. However, the scheme is not guaranteed to converge, and indeed it diverged in many instances where the simulation noise was substantial. It may be argued that divergence of the iterative regression scheme is an indication that the number of samples used is insufficient for a high quality estimate, and that more sampling is required. However, this is only a conjecture at this point, and further experimentation is needed to arrive at a reliable conclusion regarding the potential advantages of iterating within our regression scheme.

We finally note that the ability of our regularization approach to deal with near-singular problems suggests that it should be successful in dealing with general square linear systems of the form $Gr = c$, where G is not necessarily symmetric and positive definite, but may be nearly singular. If the components of G and c are computed by simulation together with corresponding sample covariances, reliable estimates of r may be obtained using the regression/regularization formula (11).

3 Importance Sampling, Design of Sampling Distribution, and Error Variance Bounds

The simulation that generates the estimates $\hat{G}$ and $\hat{c}$ using Eqs. (7)-(8) can be carried out in several ways. For example, we may generate a single sequence of independent index triples $\{(i_1, j_1, \bar{j}_1), \ldots, (i_T, j_T, \bar{j}_T)\}$ according to a distribution ξ, and estimate all entries of G and c simultaneously; or at the other extreme, we may generate a separate sequence of independent index triples (or pairs) with a separate sampling distribution for each scalar component of G (or c). The motivation for this is that we may tailor the sampling distribution to the component with the aim of reducing the variance of the corresponding estimation error, based on ideas from importance sampling. In general, we may specify a partition of G and c into blocks of components, and generate a separate sequence of index triples per block.

Importance sampling (IS) is a basic simulation technique for estimating multidimensional sums or integrals [Hal70], [ES00]. Recent developments on IS have focused on changing the sampling distribution adaptively, in order to obtain estimates with nice asymptotic behavior [OB92], [LC98], [DW05]. We will next provide a variance analysis of a nonadaptive type of IS that we have used. In particular, we will derive estimates of the covariances of the estimation errors G and c, and a normalized measure of quality of the sampling distribution ξ, called *divergence factor*, which will in turn motivate various suboptimal but practically implementable choices of ξ.

3.1 Variance Analysis for Importance Sampling

The estimation of components of G (or c) using Eqs. (7) [or (8)] amounts to estimation of a sum of a large number of terms (as many as n^3 for components of G and as many as n^2 for components of c). When a single component of G or c is estimated, this is a sum of scalars [cf. Eq. (6)]. When a block of components of G or c is estimated, this is a sum of multidimensional vectors. To cover all cases, we will consider the problem of estimating sums of the more abstract form

$$z = \sum_{\omega \in \Omega} v(\omega), \tag{16}$$

where Ω is a finite set and $v : \Omega \mapsto \Re^d$ is a function of ω. In the case of estimation of components of G (or c), ω is a triple $(i, j, \bar{j})$ [or pair (i, j), respectively].

According to the IS technique, we introduce a distribution ξ that assigns positive probability $\xi(\omega)$ to every nonzero element $\omega \in \Omega$, and we generate a sequence

$$\{\omega_1, \ldots, \omega_T\}$$

of independent samples from Ω according to ξ. We estimate z with

$$\hat{z} = \frac{1}{T} \sum_{t=1}^{T} \frac{v(\omega_t)}{\xi(\omega_t)}. \tag{17}$$

Clearly $\hat{z}$ is unbiased:

$$E[\hat{z}] = \frac{1}{T} \sum_{t=1}^{T} \sum_{\omega \in \Omega} \xi(\omega) \frac{v(\omega)}{\xi(\omega)} = \sum_{\omega \in \Omega} v(\omega) = z.$$

Furthermore, by using the independence of the samples, the covariance of $\hat{z}$ is given by

$$\mathrm{cov}(\hat{z}) = \frac{1}{T^2} \sum_{t=1}^{T} \sum_{\omega \in \Omega} \xi(\omega) \left(\frac{v(\omega)}{\xi(\omega)} - z \right) \left(\frac{v(\omega)}{\xi(\omega)} - z \right)',$$

which can be written as

$$\mathrm{cov}(\hat{z}) = \frac{1}{T} \left(\sum_{\omega \in \Omega} \frac{v(\omega) v(\omega)'}{\xi(\omega)} - z z' \right). \tag{18}$$

A natural question is to find the sampling distribution ξ that minimizes a measure of this error covariance for a fixed number of samples T. We will consider separately the two cases where z is one-dimensional ($d = 1$), and where z is multi-dimensional ($d > 1$).

(i) $d = 1$**:** Then Eq. (18) becomes

$$\mathrm{var}(\hat{z}) = \frac{z^2}{T}\left(\sum_{\omega\in\Omega}\frac{\big(v(\omega)/z\big)^2}{\xi(\omega)} - 1\right). \tag{19}$$

Assuming that $v(\omega) \geq 0$ for all $\omega \in \Omega$,[2] the optimal distribution is $\xi^* = v/z$ and the corresponding minimum variance value is 0. However, ξ^* cannot be computed without knowledge of z.

(ii) $d > 1$**:** In this case, the covariance for $\hat{z}$ [cf. Eq. (18)] is a matrix that cannot be minimized directly. One possibility is to minimize instead an estimate of a norm of the matrix $\sum_{\omega\in\Omega}\big(v(\omega)v(\omega)'/\xi(\omega)\big)$. We have

$$\left\|\sum_{\omega\in\Omega}\frac{v(\omega)v(\omega)'}{\xi(\omega)}\right\| \leq \sum_{\omega\in\Omega}\frac{\big\|v(\omega)v(\omega)'\big\|}{\xi(\omega)}.$$

Minimizing this upper bound yields a near-optimal sampling distribution:

$$\xi^*(\omega) = C\cdot\big\|v(\omega)v(\omega)'\big\|^{\frac{1}{2}}, \qquad \omega\in\Omega, \tag{20}$$

where C is a normalizing constant.

If we are only interested in the uncertainty of $\hat{z}$ along a particular direction d, we may minimize $d'\mathrm{cov}(\hat{z})d$, which is determined by the term

$$d'\left(\sum_{\omega\in\Omega}\frac{v(\omega)v(\omega)'}{\xi(\omega)}\right)d = \sum_{\omega\in\Omega}\frac{\big(d'v(\omega)\big)^2}{\xi(\omega)}.$$

In this way, we map the uncertainty of $\hat{z}$ to a one-dimensional subspace. Under the assumption that $d'v(\omega) \geq 0$ for all $\omega \in \Omega$, the corresponding "optimal" sampling distribution is

$$\xi^*(\omega) = \frac{d'v(\omega)}{d'z}, \qquad \omega\in\Omega. \tag{21}$$

[2] This may be assumed without loss of generality. When v takes negative values, we may decompose v as

$$v = v^+ - v^-,$$

so that both v^+ and v^- are positive functions, and then estimate separately $z_1 = \sum_{\omega\in\Omega} v^+(\omega)$ and $z_2 = \sum_{\omega\in\Omega} v^-(\omega)$.

Note that calculating exactly ξ^* is impractical with both formulas (20) and (21).

In both cases (i) and (ii), we see that ξ should be designed to fit some function, which we generically denote by ν. In the one-dimensional case, $\nu = v$. In the multi-dimensional case, $\nu = \|vv'\|^{1/2}$, if we want to minimize the upper bound for some norm of $\mathrm{cov}\{\hat{z}\}$, or $\nu = d'v$ if we are interested in the uncertainty of $\hat{z}$ along a specific direction d. The probability distribution ξ^* minimizes the cost function

$$F_\xi = \sum_{\omega \in \Omega} \frac{\nu(\omega)^2}{\xi(\omega)} \tag{22}$$

over all distributions ξ, and is of the form $\xi^*(\omega) = C \cdot |\nu(\omega)|$, where C is a positive normalization constant $[C^{-1} = \sum_{\omega \in \Omega} \nu(\omega)]$. In our subsequent analysis, we assume without loss of generality that $\nu(\omega) \geq 0$ for all $\omega \in \Omega$, so we may write

$$\xi^*(\omega) = C \cdot \nu(\omega), \qquad \omega \in \Omega, \tag{23}$$

with

$$C^{-1} = \sum_{\omega \in \Omega} \nu(\omega).$$

Since computing ξ^* is impractical (C is as hard to compute as the sum z that we wish to estimate), we are motivated to use a suboptimal sampling distribution. One possibility is to introduce a restricted class of distributions Ξ, and try to optimize the cost F_ξ of Eq. (22) over all $\xi \in \Xi$. For example, Ξ may be a class of piecewise constant or piecewise linear distributions over Ω. We have adopted a related approach, whereby instead of $\xi^* = C \cdot \nu$, we use a suboptimal distribution $\hat{\xi}$ of the form

$$\hat{\xi}(\omega) = \hat{C} \cdot \hat{\nu}(\omega), \qquad \omega \in \Omega, \tag{24}$$

with

$$\hat{C}^{-1} = \sum_{\omega \in \Omega} \hat{\nu}(\omega),$$

such that for all $\omega \in \Omega$, we have $\hat{\nu}(\omega) > 0$ if $\nu(\omega) > 0$. We select $\hat{\nu}$ by "fitting" ν from some restricted class of functions, using the values of ν at a relatively small subset of "trial" points.

The overall estimation procedure is as follows:

(i) Choose the target/desired function ν.

(ii) Generate trial pairs $\big\{(\bar{\omega}_1, \nu(\bar{\omega}_1)), (\bar{\omega}_2, \nu(\bar{\omega}_2)), \dots\big\}$.

(iii) Approximate ν with a function $\hat{\nu}$ from a restricted class, based on the trial pairs, and obtain the corresponding sampling distribution $\hat{\xi} = \hat{C} \cdot \hat{\nu}$.

(iv) Generate the sample sequence $\big\{(\omega_1, v(\omega_1)), \dots, (\omega_T, v(\omega_T))\big\}$ according to $\hat{\xi}$ and compute the estimate $\hat{z}$ using Eq. (17).

For further insight into the preceding procedure, it is useful to introduce the following normalized version of the cost function F_ξ of Eq. (22):

$$D_\xi = \frac{1}{\big(\sum_{\omega\in\Omega} \nu(\omega)\big)^2} \sum_{\omega\in\Omega} \frac{\nu(\omega)^2}{\xi(\omega)}, \tag{25}$$

which we call the *divergence factor*. The minimization of F_ξ can equivalently be written as

$$\begin{aligned} &\text{minimize} \quad D_\xi \\ &\text{s.t.} \quad \sum_{\omega\in\Omega} \xi(\omega) = 1, \quad \xi \geq 0. \end{aligned} \tag{26}$$

Using Eqs. (24)-(25), we can express $D_{\hat{\xi}}$ as

$$D_{\hat{\xi}} = \frac{\hat{C}^{-1}}{\big(\sum_{\omega\in\Omega} \nu(\omega)\big)^2} \sum_{\omega\in\Omega} \frac{\nu(\omega)^2}{\hat{\nu}(\omega)}.$$

We have

$$\hat{C}^{-1} = \sum_{\omega\in\Omega} \hat{\nu}(\omega) \leq \sum_{\omega\in\Omega} \nu(\omega) \cdot \max_{\omega\in\Omega} \frac{\hat{\nu}(\omega)}{\nu(\omega)},$$

and

$$\sum_{\omega\in\Omega} \frac{\nu(\omega)^2}{\hat{\nu}(\omega)} \leq \sum_{\omega\in\Omega} \nu(\omega) \cdot \max_{\omega\in\Omega} \frac{\nu(\omega)}{\hat{\nu}(\omega)},$$

so by combining the preceding relations, we obtain the following bound:

$$D_{\hat{\xi}} \leq \max_{\omega\in\Omega} \frac{\hat{\nu}(\omega)}{\nu(\omega)} \cdot \max_{\omega\in\Omega} \frac{\nu(\omega)}{\hat{\nu}(\omega)}. \tag{27}$$

This provides an intuitive interpretation of our approach: by fitting ν with $\hat{\nu}$, we keep the ratios $\hat{\nu}/\nu$ and $\nu/\hat{\nu}$ near the unit function. This keeps the upper bound (27) to D_ξ small, and hence also the cost function F_ξ small. We will next focus on how to approximate ν. We will consider a few methods, such as piecewise constant and piecewise linear approximations, and analyze the resulting divergence factor.

3.2 Designing the Sampling Distribution by Piecewise Approximation

Let us consider approximation of the optimal sampling distribution $\xi^* = C \cdot \nu$ [cf. Eq. (23)] by piecewise approximation of ν. Given a partition $\{\Omega_k\}_{k=1}^K$ for Ω, we approximate separately ν on each Ω_k with some function $\hat{\nu}_k$. Then we approximate ν and ξ^* by

$$\hat{\nu} = \sum_{k=1}^{K} \hat{\nu}_k \cdot \mathbf{1}_{\Omega_k}, \qquad \hat{\xi}(\omega) = \hat{C} \cdot \hat{\nu}(\omega), \quad \forall\ \omega \in \Omega,$$

where $\mathbf{1}_{\Omega_k}$ denotes the function that is equal to 1 within Ω_k and 0 otherwise, and $\hat{C}$ is the normalizing constant.

We select a special point ω_k within each set Ω_k, at which the approximation is "anchored" in the sense that $\hat{\nu}_k(\omega_k) = \nu(\omega_k)$. We assume that Ω is a subset of a Euclidean space, and we introduce the scalar

$$\rho = \max_{k=1,\ldots,K} \sup_{\omega \in \Omega_k} \|\omega - \omega_k\|,$$

which is a measure of how fine the partition is. In the following analysis, we will view $\nu(\omega)$, $\omega \in \Omega$, as the values of a continuous function (also denoted ν for convenience), which is defined over the convex hull of Ω. From the estimate of Eq. (27), we see that under reasonable assumptions on ν, the deviation of $\hat{\nu}/\nu$ from the unit function decreases as ρ decreases. As a result we can control $D_{\hat{\xi}}$ and thus the corresponding simulation error covariance, and make them as small as desired by using a sufficiently fine partition.

We will now discuss the cases of piecewise constant and piecewise linear approximation, as examples of the broader class of polynomial approximation methods. Other types of approximating functions may be used, such as Fourier series up to some order, and weighted sums of Gaussian functions. Their analysis may follow a similar line, based on the bound of Eq. (27) for the divergence factor D_{ξ}.

3.2.1 Piecewise constant approximation

Given a partition $\{\Omega_k\}_{k=1}^K$ of Ω and the point $\omega_k \in \Omega_k$ for each k, consider the piecewise constant approximation

$$\hat{\nu}_k(\omega) = \nu(\omega_k), \qquad \forall\ \omega \in \Omega_k.$$

Then

$$\hat{\nu} = \sum_{k=1}^{K} \nu(\omega_k) \cdot \mathbf{1}_{\Omega_k}, \tag{28}$$

and the corresponding sampling distribution is

$$\hat{\xi} = \hat{C} \cdot \sum_{k=1}^{K} \nu(\omega_k) \cdot \mathbf{1}_{\Omega_k},$$

where

$$\hat{C}^{-1} = \sum_{k=1}^{K} n_k \, \nu(\omega_k),$$

and n_k is the number of points in the set Ω_k.

The following propositions provide upper bounds for the divergence factor $D_{\hat{\xi}}$ based on Eq. (27), under some reasonable smoothness conditions.

Proposition 1 *If* $\log \nu$ *exists and is Lipschitz continuous with Lipschitz constant* η*, then*

$$D_{\hat{\xi}} \leq e^{2\eta\rho}.$$

Proof. By the Lipschitz continuity assumption we have $|\log \nu(x) - \log \nu(y)| \leq \eta \|x - y\|$ for any $x, y \in \Omega$, which implies that

$$\max_{x,y \in \Omega_k} \left\{ \frac{\nu(x)}{\nu(y)}, \frac{\nu(y)}{\nu(x)} \right\} \leq e^{\eta \max_{x,y\in\Omega_k} \|x-y\|} \leq e^{2\eta\rho}.$$

This together with Eq. (27), yields the desired result. ■

Proposition 2 *If* ν *is Lipschitz continuous with Lipschitz constant* $\eta > 0$*, and for some* $\beta > 0$ *we have* $\nu(\omega) \geq \beta$ *for all* $\omega \in \Omega$*, then*

$$D_{\hat{\xi}} \leq \left(1 + \frac{\eta\rho}{\beta}\right)^2.$$

Proof. For any $\omega \in \Omega_k$, by the Lipschitz continuity of ν we have

$$|\nu(\omega_k) - \nu(\omega)| \leq \eta \|\omega_k - \omega\| \leq \eta\rho.$$

Using the assumption $\nu \geq \beta$, we obtain

$$\frac{\nu(\omega_k)}{\nu(\omega)} \leq 1 + \frac{|\nu(\omega) - \nu(\omega_k)|}{\nu(\omega)} \leq 1 + \frac{\eta\rho}{\beta},$$

and by symmetry, the same bound holds for $\nu(\omega)/\nu(\omega_k)$. This together with Eq. (27), yields the desired result. ■

3.2.2 Piecewise linear approximation

Let us assume that ν is differentiable, with gradient at ω denoted by $\nabla\nu(\omega)$. Given a partition $\{\Omega_k\}_{k=1}^K$ of Ω and the point $\omega_k \in \Omega_k$ for each k, we consider a piecewise linear approximation whereby the function ν is approximated within Ω_k by the linear function

$$\hat{\nu}_k(\omega) = \nu(\omega_k) + \nabla\nu(\omega_k)'(\omega - \omega_k), \qquad \omega \in \Omega_k.$$

The following proposition gives a corresponding upper bound for D_ξ.

Proposition 3 *Assume that $\nabla\nu$ is Lipschitz continuous with Lipschitz constant $\eta > 0$ and that for some $\beta > 0$ we have $\nu(\omega) \geq \beta$ for all $\omega \in \Omega$. Then*

$$D_{\hat{\xi}} \leq \left(1 + \frac{\eta\rho^2}{2\beta}\right)^2.$$

Proof. For any k and any $\omega \in \Omega_k$ we have

$$\begin{aligned} \nu(\omega) &= \nu(\omega_k) + \int_0^1 \nabla\nu\big(\omega_k + t(\omega - \omega_k)\big)\mathrm{d}t \\ &= \nu(\omega_k) + \nabla\nu(\omega_k)'(\omega - \omega_k) \qquad (29) \\ &\quad + \int_0^1 \Big(\nabla\nu\big(\omega_k + t(\omega - \omega_k)\big) - \nabla\nu(\omega_k)\Big)\,\mathrm{d}t. \end{aligned}$$

Using the Lipschitz continuity of $\nabla\nu$, we have for all $t \in [0, 1]$,

$$\big\|\nabla\nu\big(\omega_k + t(\omega - \omega_k)\big) - \nabla\nu(\omega_k)\big\| \leq t\eta\|\omega - \omega_k\| \leq t\eta\rho.$$

Hence the third term in Eq. (29) can be bounded by $\eta\rho^2/2$, which implies that

$$|\nu(\omega) - \hat{\nu}(\omega)| \leq \frac{\eta\rho^2}{2}, \qquad \forall\, \omega \in \Omega.$$

Since $\nu \geq \beta$, we see that an upper bound for both $\max_\omega\{\hat{\nu}(\omega)/\nu(\omega)\}$ and $\max_\omega\{\nu(\omega)/\hat{\nu}(\omega)\}$ is

$$\max\left\{1 + \frac{|\nu(\omega) - \hat{\nu}(\omega)|}{\nu(\omega)}, 1 + \frac{|\nu(\omega) - \hat{\nu}(\omega)|}{\hat{\nu}(\omega)}\right\} \leq 1 + \frac{\eta\rho^2}{2\beta}.$$

This together with Eq. (27), yields the desired result. ■

The qualitative advantage of piecewise linear versus piecewise constant approximation for small ρ can be seen by comparing the bound of Prop. 2 (which involves ρ) with the one of Prop. 3 (which involves ρ^2).

Let us finally mention that in the case of a piecewise approximation, there is a bound for the divergence factor that is slightly sharper than the one of Eq. (27). It is given by

$$D_{\hat{\xi}} \leq \left(\sum_{k=1}^{K} \eta_k \max_{\omega \in \Omega_k} \frac{\hat{\nu}_k(\omega)}{\nu(\omega)} \right) \cdot \left(\sum_{k=1}^{K} \eta_k \max_{\omega \in \Omega_k} \frac{\nu(\omega)}{\hat{\nu}_k(\omega)} \right),$$

where

$$\eta_k = \frac{\sum_{\omega \in \Omega_k} \nu(\omega)}{\sum_{\omega \in \Omega} \nu(\omega)}.$$

Using this bound one may obtain slightly sharper but qualitatively similar estimates to the ones of Props. 1-3.

4 Confidence Regions

The preceding sections have focused on the essential elements of our approach to obtain low-variance estimates of the components of the matrix G and the vector c for a fixed number of samples. In this section, we will focus on quantifying the effect of the number of samples on the quality of the estimate $\hat{r}$ produced by the regression methodology of Section 2, in conjunction with the simulation formulas of Eqs. (7)-(8).

We will derive a $(1-\theta)$-confidence region for the approximate solution $\hat{r}$, where θ is a given small positive number. We consider the case where regularization of the form $\Gamma^{-1} = \beta I$ is used, for some $\beta > 0$. Then assuming that the inverses below exist, the approximate solution $\hat{r}$ of Eq. (11) can be rewritten as

$$\hat{r} = \left(\hat{G}'\Sigma^{-1}\hat{G} + \beta I \right)^{-1} \left(\hat{G}'\Sigma^{-1}\hat{c} + \beta\bar{r} \right), \tag{30}$$

where Σ is some positive definite symmetric matrix.

We denote by r_β^* the solution that would be obtained if $\hat{G} = G$ and $\hat{c} = c$:

$$r_\beta^* = \left(G'\Sigma^{-1}G + \beta I \right)^{-1} \left(G'\Sigma^{-1}c + \beta\bar{r} \right)$$

which differs from r^* since $\beta \neq 0$. We will now derive a confidence interval for the error $\hat{r} - r^*$. Let us denote

$$d = \Sigma^{-1/2}(\hat{c} - \hat{G}r^*),$$

so from Eq. (30), the error can be written as

$$\hat{r} - r^* = \left(\hat{G}'\Sigma^{-1}\hat{G} + \beta I \right)^{-1} \left(\hat{G}'\Sigma^{-1/2}d + \beta(\bar{r} - r^*) \right). \tag{31}$$

Let also $\hat{\Sigma}$ be the covariance of $(\hat{c} - \hat{G}r^*)$, and let

$$\hat{d} = \hat{\Sigma}^{-1/2}(\hat{c} - \hat{G}r^*) = \hat{\Sigma}^{-1/2}\Sigma^{1/2}d. \tag{32}$$

For a large number of samples, we may assume (by the central limit theorem) that $(\hat{c}-\hat{G}r^*)$ is a zero mean Gaussian random s-dimensional vector, so that the scalar

$$\|\hat{d}\|^2 = (\hat{c} - \hat{G}r^*)'\hat{\Sigma}^{-1}(\hat{c} - \hat{G}r^*)$$

can be treated as a chi-square random variable with s degrees of freedom. Assuming this, we have

$$\|\hat{d}\| \leq \sqrt{P^{-1}(1-\theta; s)} \tag{33}$$

with probability $(1-\theta)$, where $P^{-1}(1-\theta; s)$ is the threshold value v at which the probability that a chi square random variable with s degrees of freedom takes value greater than v is $(1-\theta)$. In our algorithm, Σ can be any positive definite matrix, but we have focused on the case where Σ is the sample covariance of $(\hat{c} - \hat{G}\tilde{r})$, where $\tilde{r}$ is only a guess of r^*, such as $\hat{G}^{-1}\hat{c}$; cf. Section 2. If $\tilde{r}$ is close to r^*, then Σ is close to $\hat{\Sigma}$ and d is close to $\hat{d}$.

We now derive the following confidence interval for the error $\hat{r} - r^*$ (assuming that $\hat{d}$ can be treated as a Gaussian random variable).

Proposition 4 *We have*

$$\mathbf{P}\big(\|\hat{r} - r^*\| \leq \sigma(\Sigma, \beta)\big) \geq 1 - \theta,$$

where

$$\begin{aligned}\sigma(\Sigma, \beta) = &\max_{i=1,\dots,s}\left\{\frac{\lambda_i}{\lambda_i^2 + \beta}\right\}\left\|\Sigma^{1/2}\hat{\Sigma}^{-1/2}\right\|\sqrt{P^{-1}(1-\theta; s)}\\ &+ \max_{i=1,\dots,s}\left\{\frac{\beta}{\lambda_i^2 + \beta}\right\}\|\bar{r} - r^*\|,\end{aligned} \tag{34}$$

and $\lambda_1, \dots, \lambda_s$ are the singular values of $\Sigma^{-1/2}\hat{G}$.

Proof. Let $\Sigma^{-1/2}\hat{G} = U\Lambda V'$ be the singular value decomposition of $\Sigma^{-1/2}\hat{G}$, where $\Lambda = \text{diag}\{\lambda_1, \dots, \lambda_s\}$, and U, V are unitary matrices ($UU' = VV' = I$). Then, Eq. (31) becomes

$$\begin{aligned}\hat{r} - r^* &= \big(V\Lambda U'U\Lambda V' + \beta I\big)^{-1}\big(V\Lambda U'd + \beta(\bar{r} - r^*)\big)\\ &= V(\Lambda^2 + \beta I)^{-1}\Lambda U'd + \beta\, V(\Lambda^2 + \beta I)^{-1}V'(\bar{r} - r^*)\\ &= V(\Lambda^2 + \beta I)^{-1}\Lambda U'\,\Sigma^{1/2}\hat{\Sigma}^{-1/2}\hat{d} + \beta\, V(\Lambda^2 + \beta I)^{-1}V'(\bar{r} - r^*),\end{aligned}$$

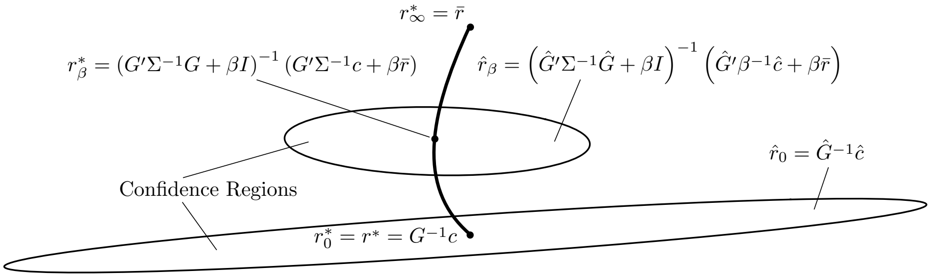
$r^*_\infty = \bar{r}$
$r^*_\beta = (G'\Sigma^{-1}G + \beta I)^{-1}\,(G'\Sigma^{-1}c + \beta\bar{r})$
$\hat{r}_\beta = \left(\hat{G}'\Sigma^{-1}\hat{G} + \beta I\right)^{-1}\left(\hat{G}'\beta^{-1}\hat{c} + \beta\bar{r}\right)$
$\hat{r}_0 = \hat{G}^{-1}\hat{c}$
Confidence Regions
$r^*_0 = r^* = G^{-1}c$

Figure 1: Illustration of confidence regions for different values of the regularization parameter β. For different values of $\beta \in [0, \infty]$, the figure shows the estimates $\hat{r}_\beta$, corresponding to a finite number of samples, and the exact values r^*_β, corresponding to an infinite number of samples. By Eq. (31), we may view $\hat{r}_\beta - r^*$ as the sum of a "simulation error" whose norm is bounded by the first term in the estimate (34), and a "regularization error" whose norm is bounded by the second term in the estimate (34).

where the third equality follows from Eq. (32). The matrix $V(\Lambda^2 + \beta I)^{-1}\Lambda U'$ in the above equality has singular values $\lambda_i/(\lambda_i^2 + \beta)$, while the matrix multiplying $(\bar{r} - r^*)$ has singular values $\beta/(\lambda_i^2 + \beta)$. Taking the norm of both sides and using the triangle inequality, it follows that

$$\|\hat{r} - r^*\| \leq \max_{i=1,\ldots,s}\left\{\frac{\lambda_i}{\lambda_i^2 + \beta}\right\}\left\|\Sigma^{1/2}\hat{\Sigma}^{-1/2}\right\|\|\hat{d}\| + \max_{i=1,\ldots,s}\left\{\frac{\beta}{\lambda_i^2 + \beta}\right\}\|\bar{r} - r^*\|. \tag{35}$$

Since Eq. (33) holds with probability $(1 - \theta)$, the desired result follows. ■

Note from Eq. (35) that the error $\|\hat{r}_k - r^*\|$ is bounded by the sum of two terms. The first term, reflects the *simulation error*, and depends on $\|\hat{d}\|$, which can be made arbitrarily small by using a sufficiently large number of samples [cf. Eq. (32)]. The second term reflects the *regularization error* (the bias introduced by the quadratic $\beta\|r - \bar{r}\|^2$ in the regularized cost function) and diminishes with β, but it cannot be made arbitrarily small by using more samples (see Fig. 1).

Now consider the limiting case of the preceding proposition where $\beta = 0$ and $\Sigma = \hat{\Sigma}$, assuming that $\hat{G}$ is invertible. In this case $\hat{r} = \hat{G}^{-1}\hat{c}$, and the preceding proof can be used to show that

$$\mathbf{P}\left(\|\hat{G}^{-1}\hat{c} - r^*\| \leq \max_{i=1,\ldots,s}\left\{\frac{1}{\lambda_i}\right\}\sqrt{P^{-1}(1 - \theta; s)}\right) \geq 1 - \theta.$$

This shows that the level of confidence is adversely affected by near singularity of the matrix $\hat{G}$, and hence by near singularity of the matrix G (since $\hat{G}$ is close to G). It is also possible to derive a confidence interval involving the singular values of G rather than $\hat{G}$. While the derivation is more complicated and will not be given in this paper, it supports the qual-

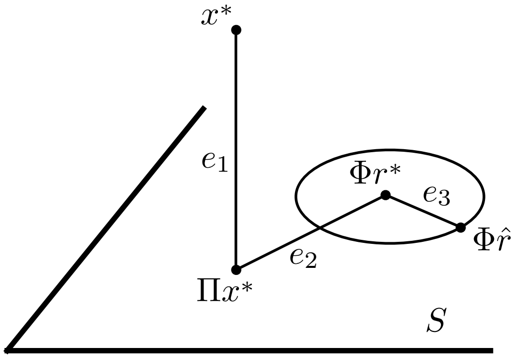

Figure 2: Comparison of the projection error e_1, the subspace approximation error e_2 and the simulation plus regularization error e_3, where x^* is the exact solution to $Ax = b$, Πx^* is the projection of x^* on the subspace S, Φr^* is the exact solution to the approximate low-dimensional system [cf. Eq. (3)] and $\Phi\hat{r}$ is the approximate solution obtained by the proposed algorithm.

itative conclusion that the radius of the confidence interval is proportional to $\|G^{-1}\|$.

5 Computational Results

Our proposed simulation and regression methodology has been tested on a few large inverse problems, two of which will be discussed here: a heat conduction problem [Car82] and the evaluation of the second derivative of a noisy function [Han94]. These systems take the form of Fredholm integral equations of the first kind, and are approximated by the Galerkin or Gauss quadrature methods, which can yield square linear systems of the form $Ax = b$ of arbitrarily large dimension [DW74]. Note that for both test problems, the exact solutions x^* are known and available for comparison. Additional computational results may be found in the companion paper [PWB09].

We consider a subspace S spanned by multi-resolution functions, which are pairwise orthogonal piecewise constant functions with disjoint local supports. The scalar components of G and c are estimated separately, and since G is symmetric, it suffices to estimate its upper-triangular components, yielding a total of $(s^2 + 3s)/2$ scalars to be estimated. Each component $G_{\ell q}$ and c_ℓ is estimated using a specially designed sampling distribution, which involves a piecewise low-order spline approximation to A (we refer to the companion paper [PWB09] for more details). The experiments were computed on a dual processor laptop computer with 4GB RAM running Matlab, using 10^4 samples per entry of G and c, with each sample taking $50\mu s$ on average. Note that the proposed algorithm is well suited for parallel processing.

In presenting the computational results, we will compare three types of error, as illustrated in Fig. 2: (i) the *projection error* $e_1 = \Pi x^* - x^*$, which measures the distance between the exact solution and the subspace S; (ii) the *subspace approximation error* $e_2 = \Phi r^* - \Pi x^*$, which measures the distance of Φr^* and the "best" approximation of x^* within S; and (iii) the *simulation error* $e_3 = \Phi \hat{r} - \Phi r^*$, which can be made arbitrarily small by sufficient sampling. We will also compare the algorithm's performance for alternative importance sampling distributions.

5.1 The inverse heat conduction problem

The problem is to reconstruct the time profile of a heat source by monitoring the temperature at a fixed location away from the source [Car82]. The one-dimensional heat transfer in a homogeneous quarter plane is expressed as an elliptic partial differential (heat) equation,

$$\frac{\partial u}{\partial \tau} = \alpha \frac{\partial^2 u}{\partial \sigma^2}, \qquad \sigma \geq 0,\ \tau \geq 0,$$

$$u(\sigma, 0) = 0, \qquad u(0, \tau) = x(\tau),$$

where $u(\sigma, \tau)$ is the temperature at location σ and time τ, and α is the heat conductivity constant. Let $b(\cdot) = u(\bar{\sigma}, \cdot)$ be the temperature history at a location $\bar{\sigma}$ away from the source. It satisfies the following Volterra integral equation,

$$b(\tau) = \int_0^\tau \mathrm{d}\upsilon \, \frac{\bar{\sigma}/\alpha}{\sqrt{4\pi(\upsilon - \tau)^3}} \exp\Big(-\frac{(\bar{\sigma}/\alpha)^2}{4(\upsilon - \bar{\tau})}\Big) \, x(\upsilon), \quad 0 \leq \tau \leq T, \tag{36}$$

or equivalently

$$b(\tau) = \int_0^T \mathrm{d}\upsilon A(\upsilon, \tau) x(\upsilon), \quad 0 \leq \tau \leq T, \tag{37}$$

where A is a lower-triangular kernel given by

$$\mathbf{A}(\upsilon, \tau) = \begin{cases} \frac{\bar{\sigma}/\alpha}{\sqrt{4\pi(\upsilon - \tau)^3}} \exp(-\frac{(\bar{\sigma}/\alpha)^2}{4(\upsilon - \tau)}), & 0 \leq \tau < \upsilon \leq T, \\ 0, & 0 \leq \upsilon \leq \tau \leq T. \end{cases}$$

The integral equation (37) is discretized into a linear square system of dimension $n = 10^9$. We consider a subspace S spanned by $s = 50$ and $s = 100$

multi-resolution basis functions, and assume an initial guess $\bar{r} = 0$ and a regularization weight matrix $\Gamma^{-1} = \beta L_1' L_1$, where L_1 is the $(s-1) \times s$ discrete first-order difference operator given by,

$$L_1 = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & 0 \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots & 0 \\ 0 & \cdots & 0 & -1 & 1 \end{pmatrix},$$

and β is a small positive scalar. The simulation results are illustrated in Figs. 3-4, which are consistent with our earlier analysis and conjectures.

5.2 The second derivative problem

This classical inverse problem refers to differentiating noisy signals that typically arise from experimental measurements. Let b be the noisy signal function and x be the desired derivative function, so that

$$b(\upsilon) = \int_0^1 \mathrm{d}\tau \, \mathbf{A}(\upsilon, \tau) x(\tau) \qquad \tau, \upsilon \in [0,1] \tag{38}$$

where we denote by $\mathbf{A}(\upsilon, \tau)$ the Green's function of the second derivative defined by

$$\mathbf{A}(\upsilon, \tau) = \begin{cases} \upsilon(\tau - 1) & \upsilon < \tau, \\ \tau(\upsilon - 1) & \upsilon \geq \tau. \end{cases} \tag{39}$$

This problem is known to be mildly ill-posed, exhibiting instabilities with increasing levels of noise in υ [Cul71]. Following the approach of Hansen [Han94] we discretize Eq. (38) using the Galerkin method and obtain a linear square system $Ax = b$ of dimension n. We also impose an initial guess of $\bar{r} = 0$ with $\Gamma^{-1} = \beta L_2' L_2$, where the $(s-2) \times s$ matrix L_2 is the discrete second-order difference operator given by

$$L_2 = \begin{pmatrix} 1 & -2 & 1 & 0 & \dots & 0 \\ 0 & 1 & -2 & 1 & \dots & 0 \\ \vdots & 0 & \ddots & \ddots & \vdots & 0 \\ 0 & 0 & \dots & 1 & -2 & 1 \end{pmatrix},$$

and β is a small positive scalar. We have chosen a subspace S spanned by $s = 50$ and $s = 100$ multi-resolution basis functions. The experimental results are shown in Figs. 5-6. In Fig. 5, the approximate solution $\Phi\hat{r}$ is

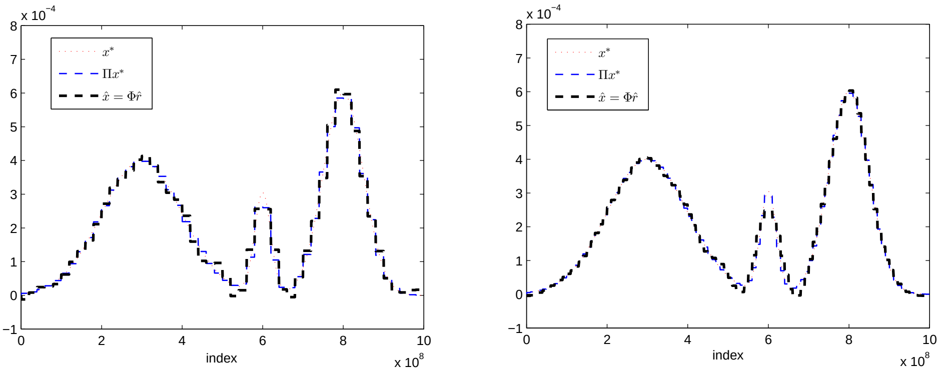

Figure 3: The simulation-based approximate solution $\hat{x} = \Phi\hat{r}$ for the inverse heat conduction problem, compared with the exact solution x^* and the projected solution Πx^*. The dimension is $n = 10^9$, and the subspace S has dimension $s = 50$ for the left-hand plot and dimension $s = 100$ for the right-hand plot. The number of samples used per component of G and c is 10,000.

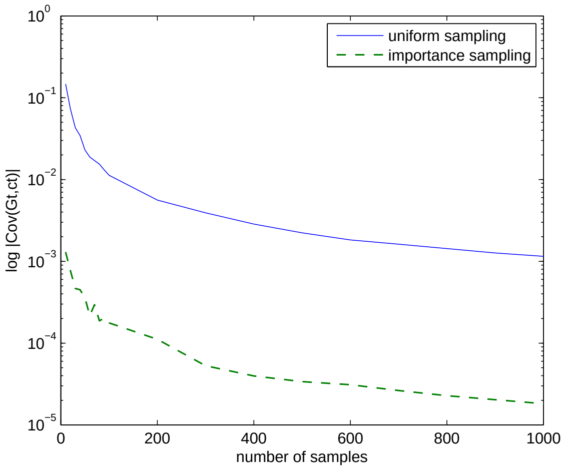

Figure 4: The decrease of the simulation error for the inverse heat conduction problem as a function of the number of samples per component of G and c. The figure compares the norm of the simulation error as a function of the number of samples for two sampling distributions.

compared with the exact solution x^* and the projected solution Πx^*. In Fig. 6, the simulation error are illustrated for alternative choices of sampling distribution when the number of samples and the number of partitions for piecewise approximation vary.

6 Conclusion

We have considered the approximate solution for large-scale least squares problems on a subspace spanned by a given set of features or basis functions. We have proposed a simulation-based regression methodology, that uses low-dimensional calculations. Through the use of importance sampling with near-optimally designed sampling distributions, our methodology can overcome the challenges posed by near-singularity of the problem and excessive variance of simulation noise.

The utility of our methodology will likely be judged on the basis of its ability to solve challenging large-scale problems. Examples of such problems were given in this paper and the companion paper [PWB09]. Additional research, targeted to specific applications, will be very helpful in clarifying the range of potential uses of our methods. Another direction worth investigating is the approximate solution of infinite-dimensional least squares problems using approximation within a low-dimensional subspace and simulation. The main ideas underlying such an approach should be similar to the ones of the present paper, but the corresponding mathematical analysis will likely be more complex.

References

[BT96] D. P. Bertsekas and J. Tsitsiklis. *Neuro-Dynamic Programming.* Athena Scientific, Belmont, MA, 1996.

[BY07] D. P. Bertsekas and H. Yu. Solution of large systems of equations using approximate dynamic programming methods. *Lab. for Information and Decision Systems Report LIDS-P-2754, MIT*, 2007.

[BY09] D. P. Bertsekas and H. Yu. Projected equation methods for approximate solution of large linear systems. *J. of Computational and Applied Mathematics*, 227:27–50, 2009.

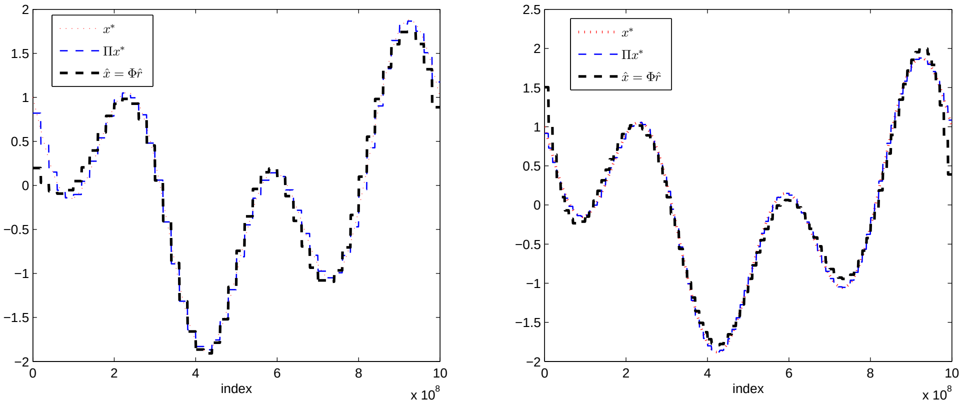

Figure 5: The simulation-based approximate solution $\hat{x} = \Phi\hat{r}$ for the second derivative problem, compared with the exact solution x^* and the projected solution Πx^*. The dimension is $n = 10^9$, and the subspace S has dimension $s = 50$ for the left-hand plot and dimension $s = 100$ for the right-hand plot. The number of samples used per component of G and c is 10,000.

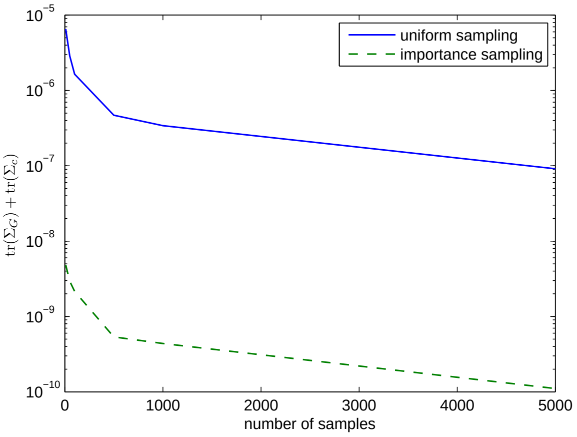

Figure 6: The decrease of the simulation error for the second derivative problem as a function of the number of samples per component of G and c. The figure compares the norm of the simulation error as a function of the number of samples, for two sampling distributions.

[Car82] A. Carasso. Determining surface temperatures from interior observations. *SIAM Journal of Applied Mathematics*, 42(3):558–574, 1982.

[Cul71] J. Cullum. Numerical differentiation and regularization. *SIAM Journal of Numerical Analysis*, 8:254–265, 1971.

[Cur53] J. H. Curtiss. Monte Carlo methods for the iteration of linear operators. *UMN*, 12:149–174, 1953.

[Cur54] J. H. Curtiss. A theoretical comparison of the efficiencies of two classical methods and a Monte Carlo method for computing one component of the solution of a set of linear algebraic equations. *Proc. Symposium on Monte Carlo Methods*, pages 191–233, 1954.

[DW74] L. M. Delves and J. Walsh, editors. *Numerical Solution of Integral Equations*. Oxford University Press, Oxford, 1974.

[DW05] P. Dupuis and H. Wang. Dynamic importance sampling for uniformly recurrent Markov chains. *The Annals of Applied Probability*, 15:1–38, 2005.

[ES00] M. Evans and T. Swartz. Approximating integrals via monte carlo simulation and deterministic methods. *Oxford University Press*, 2000.

[FL50] G. E. Forsythe and R. A. Leibler. Matrix inversion by a Monte Carlo method. *Mathematical Tables and Other Aids to Computation*, 4:127–129, 1950.

[Hal70] J. H. Halton. A retrospective and prospective survey of the Monte Carlo method. *SIAM Review*, 12:1–63, 1970.

[Han94] P. C. Hansen. Regularization tools: A matlab package for analysis and solution of discrete ill-posed problems. *Numerical Algorithms*, 6:1–35, 1994.

[KZ72] M. A. Krasnoselskii and P. P. Zabreiko. *Approximate Solution of Operator Equations*. Wolters-Noordhoff Pub., Groningen, Groningen, 1972.

[LC98] J. S. Liu and R. Chen. Sequential Monte Carlo methods for dynamic systems. *Journal of the American Statistical Association*, 93:1032–1044, 1998.

[OB92] M.-S. Oh and J. O. Berger. Adaptive importance sampling in Monte Carlo integration. *Journal of Statistical Computation and Simulation*, 41:143–168, 1992.

[PWB09] N. Polydorides, M. Wang, and D. P. Bertsekas. Approximate solution of large-scale linear inverse problems with Monte Carlo simulation. *Lab. for Information and Decision Systems Report, MIT*, 2009.

[SB98] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction.* The MIT Press, Cambridge, MA, 1998.

[SV09] T. Strohmer and R. Vershynin. A randomized Kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications*, 15:262–278, 2009.

[Was52] W. R. Wasow. A note on inversion of matrices by random walks. *Mathematical Tables and Other Aids to Computation*, 6:127–129, 1952.

[YB08] H. Yu and D. P. Bertsekas. New error bounds for approximations from projected linear equations. *Lab. for Information and Decision Systems Report LIDS-P-2797, MIT*, 2008.