Simulation-Based Approximate Solution of Large-Scale Linear Least Squares Problems and Applications

by

Mengdi Wang

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2010

Author .....
Department of Electrical Engineering and Computer Science
December 15, 2009

Certified by .....
Dimitri P. Bertsekas
Professor
Thesis Supervisor

Accepted by .....
Terry P. Orlando
Chairman, Department Committee on Graduate Students

@

Simulation-Based Approximate Solution of Large-Scale Linear Least Squares Problems and Applications

by

Mengdi Wang

Submitted to the Department of Electrical Engineering and Computer Science
on December 15, 2009, in partial fulfillment of the
requirements for the degree of
Master of Science in Electrical Engineering and Computer Science

Abstract

We consider linear least squares problems, or linear systems that can be formulated into least squares problems, of very large dimension, such as those arising for example in dynamic programming (DP) and 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. We also investigate the use of our regression/regularization approach in temporal difference-type methods in the context of approximate DP. Finally we demonstrate the application of our methodology in a series of practical large-scale examples arising from Fredholm integral equations of the first kind.

Thesis Supervisor: Dimitri P. Bertsekas
Title: Professor

@

Acknowledgments

First and foremost I want to express my sincerest gratitude to my supervisor, Prof. Dimitri P. Bertsekas, who has supported me throughout my first two years at MIT with his patience and knowledge. This thesis would not have been possible without his encouragement, guidance and endless efforts in helping me become a mature researcher and good academic writer. Prof. Bertsekas has been a wonderful mentor in both aspects of academic and life. He is like a father to me, and I simply cannot wish for a better person to be my supervisor.

I am grateful to the fine academic environment provided by the LIDS community. I would like to thank the many faculty members of LIDS, who have inspired me with their brilliance and impressed me with their rigor. I am also indebted to my student colleagues at LIDS for their friendship and support. They have provided a stimulating and fun environment, from which I have benefited greatly. I am especially grateful to my colleague, Nick Polydorides, who has helped me with my thesis in numerous ways.

I wish to thank all my friends, old and new, locally available and geographically separated, for always being there for me. Even if we may have been separated, I would still cherish all the friendship we have had and appreciate the sharing of empathy between us.

Finally, I wish to thank my grandparents, Xianliang Lv and Hongying Jin. They have raised me, loved me, trusted me, and provided a warm home for me. I cannot overstate my gratitude to their unconditional love to me and everlasting faith in me. To them I dedicate this thesis.

@

Contents

Chapter 1

Introduction

1.1 Problem Formulation and Background

We consider the solution of large-scale and complex systems that can be formulated into linear systems of equations, which take the form of

$$Ax = b, \tag{1.1}$$

where A is an $m \times n$ matrix, b is a vector in $\Re^m$, and the solution x is a vector in $\Re^n$. The system can be either underdetermined or overdetermined. A closely related problem is the fixed point solution of linear mappings that take the form of

$$T(x) = Ax + b,$$

where A is an $n \times n$ square matrix and b is a vector in $\Re^n$. These problems typically arise in fields such as dynamic programming (or reinforcement learning), inverse problems, inference, scientific computation, operations research, etc. As the dimension of the system grows, solving the problem becomes increasingly expensive in terms of storage requirements and computational cost, ultimately presenting a formidable challenge in addressing problems of realistic complexity.

In the context of dynamic programming, Eq. (1.1) is known as the Bellman equa-

tion corresponding to a single policy, with x being the cost vector of that policy. Solving the Bellman equation is known as policy evaluation, which is a critical step of the policy iteration method for solution of dynamic programming problems (see the book by Bertsekas [Ber07]). For the case of a decision process with a huge state space, evaluating a single policy may be computationally expensive, let alone evaluating a series of policies.

In the context of linear inverse problems, such systems like Eq. (1.1) typically arise from integral or differential equations which are not necessarily linear, but can be discretized into linear equations of very large but finite dimension [PWB09], such as those encountered in finite element analysis, image processing, geophysical prospecting, and astronomy [Gro07], [BB98]. When such a system results from discretization, there is effectively no limit on the dimension. Problems of these types are often additionally complicated by near-singularity of the matrix A when A is square. Hence the solution by conventional methods may not be possible, even if A is a square matrix.

We consider the least squares formulation of Eq. (1.1), which takes the form of

$$\min_{x \in \Re^n} \|Ax - b\|_\zeta^2, \tag{1.2}$$

where ζ is a known probability distribution vector with positive components, and $\|\cdot\|_\zeta$ denotes the corresponding weighted Euclidean norm.[1] We focus on the case where n is very large, and m is either large or small. We consider an approximation to problem (1.2), 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 or features. The choice of basis functions is a likely problem-dependent issue, and needs to be addressed with care. Typical choices of basis functions include triangular functions, Fourier functions, functions of the exponential family, etc. A special case is to use

[1]Regarding notation: all vectors are viewed as column vectors, a prime denotes transposition, and $\|\cdot\|$ denotes the standard unweighted Euclidean norm.

basis functions involving powers of A, e.g. $\Phi = [b, (I - A)b, \ldots, (I - A)^{s-1}b]$, which are the basis functions implicitly used in the context of Krylov subspace methods (see [Saa03]). This is a natural choice because the solution of $Ax = b$ lies in the space spanned by $(b, (I - A)b, \ldots)$, assuming that A is contractive. Although these basis functions may not be readily available, they can be approximated in the simulation process [BY09]. We also note the recent developments on basis function adaptation methods, which optimize the basis functions over a parameterized family using a cross-entropy method or gradient method (see the papers by Menache [MMS05] and by Yu and Bertsekas [YB09]). However, a detailed analysis and comparison of choices of basis functions is beyond our scope. Throughout this thesis we will assume that Φ is known.

Our approach is an approximation approach based on Monte Carlo simulation. We note that there is a large body of work on the exact or approximate solution of large-scale linear systems of equations. One of the earliest attempts is by Lanczos [Lan58], which successively approximates the solution without explicit matrix inversion. Since then a number of iterative methods have been studied, such as the Landweber iteration, the conjugate gradient method, the LSQR algorithm (see the survey [HH93] for a comprehensive review of these methods). A projection-regularization approach, proposed by O'Leary and Simmons [OS81], approximates the solution within a subspace in which the projected coefficient matrix is bidiagonalizable. A related approach proposed by Calveti and Zhang [DCZ99] suggests the use of Lanczos bidiagonalization with Gauss quadrature. Later, a trust-region formulation was proposed by Rojas and Sorensen [RS02], which poses the regularized problem as an inequality constrained least squares problem.

Also, there have been several proposals in the literature relating to 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 [Cur57], [Cur54], and the survey by Halton [Hal70]). We also note recent work on simulation methods that use low-order calculations for solving severely 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 work. The papers [WPB09] and [PWB09] extended further this methodology in a number of ways and are the basis for this thesis.

1.2 Approximation and Simulation Framework

The first step in our approach is to substitute Φr in place of x in problem (1.2) and consider the approximate problem

$$\min_{r \in \Re^s} \|A\Phi r - b\|_\zeta^2. \tag{1.3}$$

If the solution is unique, it is given by

$$r^* = G^{-1}c, \tag{1.4}$$

where

$$G = \Phi' A' Z A \Phi, \qquad c = \Phi' A' Z b, \tag{1.5}$$

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 (1.2). The paper by Yu and Bertsekas [YB08] and the report [BY07] provide bounds on the error $\Phi r^* - x^*$, which involve the weighted Euclidean distance of x^* from S.

The expressions in Eq. (1.5) 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

$$G = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{\bar{j}=1}^{n} \zeta_i a_{ij} a_{i\bar{j}} \phi_j \phi'_{\bar{j}}, \qquad c = \sum_{i=1}^{n}\sum_{j=1}^{n} \zeta_i a_{ij} b_j \phi_j, \tag{1.6}$$

where ζ_i are the diagonal components of Z, a_{ij} are the components of A, and ϕ'_j is the jth s-dimensional row of Φ:

$$\phi'_j = \begin{pmatrix} \phi_{j1} \cdots \phi_{js} \end{pmatrix}$$

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 $\{(i_0, j_0, \bar{j}_0), \ldots, (i_t, j_t, \bar{j}_t)\}$ 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,

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

[cf. Eq. (1.6)], with $\hat{G}_{\ell q}$ given by

$$\hat{G}_{\ell q} = \frac{1}{t+1} \sum_{k=0}^{t} \frac{\zeta_{i_k} a_{i_k j_k} a_{i_k \bar{j}_k} \phi_{j_k \ell} \phi_{\bar{j}_k q}}{\xi_{i_k, j_k, \bar{j}_k}}, \tag{1.7}$$

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 $\{(i_0, j_0), \ldots, (i_t, j_t)\}$ 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

$$\hat{c}_\ell = \frac{1}{t+1} \sum_{k=0}^{t} \frac{\zeta_{i_k} a_{i_k j_k} b_{i_t} \phi_{j_t \ell}}{\xi_{i_t, j_t}}, \tag{1.8}$$

where $\xi_{i,j}$ denotes the probability of the index pair (i, j). One possibility is to 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. (1.6) and Eqs. (1.7)-(1.8), 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 $\{(i_0, j_0, \bar{j}_0), \ldots, (i_t, j_t, \bar{j}_t)\}$ by sampling independently according to some distribution ξ from the set of index triples $(i, j, \bar{j}) \in \{1, \ldots, n\}^3$. Then we may estimate G and c with $\hat{G}$ and $\hat{c}$ given by

$$\hat{G} = \frac{1}{t+1} \sum_{k=0}^{t} \frac{\zeta_{i_k} a_{i_k j_k} a_{i_k \bar{j}_k}}{\xi_{i_k j_k \bar{j}_k}} \phi_{j_k} \phi'_{\bar{j}_k}, \quad \hat{c} = \frac{1}{t+1} \sum_{k=0}^{t} \frac{\zeta_{i_k} a_{i_k j_k} b_{i_k}}{\xi_{i_k j_k}} \phi_{j_k}. \tag{1.9}$$

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. (1.7) and (1.8). 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 Chapter 3). For example, when estimating a component $G_{\ell q}$ using Eq. (1.7), it is inefficient to generate sample triples $(i, j, \bar{j})$ for which $\phi_{j\ell}\phi_{\bar{j}q} = 0$.

In addition, we remark that there are many ways to generate sample triples $(i, j, \bar{j})$.

The simplest scheme is to sample i_k, j_k, and $\bar{j}_k$ independently from one another, according to distributions μ_1, μ_2, and μ_3, respectively. Then the marginal probabilities for pairs (i_k, j_k) and triples $(i_k, j_k, \bar{j}_k)$ are

$$\xi_{i_k j_k} = \mu_{1,i_k}\mu_{2,j_k}, \qquad \xi_{i_k j_k \bar{j}_k} = \mu_{1,i_k}\mu_{2,j_k}\mu_{3,\bar{j}_k}.$$

An alternative is to generate an independent sequence of indices $\{i_0, i_1, \ldots\}$ according to a distribution μ, and then generate j_k and $\bar{j}_k$ conditioned on each i_k, according to transition probabilities $q_{i_k j_k}$ and $\tilde{q}_{i_k \bar{j}_k}$. In this case, the marginal probabilities are

$$\xi_{i_k j_k} = \mu_{i_k} q_{i_k j_k}, \qquad \xi_{i_k j_k \bar{j}_k} = \mu_{i_k} q_{i_k j_k} \tilde{q}_{i_k \bar{j}_k}.$$

A somewhat more complex scheme is to generate a sequence of state transitions $\{i_0, i_1, \ldots\}$ using an irreducible Markov chain with transition probability matrix P and initial distribution ξ_0. Sampling j_k and $\bar{j}_k$ according to some transition probabilities $q_{i_k j_k}$ and $\tilde{q}_{i_k \bar{j}_k}$ yields marginal probabilities for pairs (i_k, j_k) and triples $(i_k, j_k, \bar{j}_k)$:

$$\xi_{i_k j_k} = (\xi_0' P^k)_{i_k} q_{i_k j_k}, \qquad \xi_{i_k j_k \bar{j}_k} = (\xi_0' P^k)_{i_k} q_{i_k j_k} \tilde{q}_{i_k \bar{j}_k}.$$

Here the choice of P should ensure that all row indices are sampled infinitely often, so that $\hat{G} \to G$ and $\hat{c} \to c$ (and hence also $\hat{G}^{-1}\hat{c} \to r^*$) as $t \to \infty$, with probability 1. In particular, if P is an irreducible Markov chain, we can use as ξ the distribution of long-term frequencies of state visits corresponding to P.

1.3 Overview of this Work

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 approx-

imations $\hat{G}$ 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

$$\hat{c} = \hat{G}r + e, \tag{1.10}$$

where e is the vector

$$e = (G - \hat{G})r + \hat{c} - c, \tag{1.11}$$

which we view as "simulation noise." We then estimate the solution r^* based on Eq. (1.10) 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 $\hat{c}$. [2] In Chapter 2, we discuss the corresponding methods and analysis, then we derive confidence regions and large deviation bounds for the regression solution, so as to quantify the effect of near-singularity of G, and the sample covariances of $\hat{G}$ and $\hat{c}$.

Second, to reduce the effect of the components $(G-\hat{G})$ and $(c-\hat{c})$ of the simulation noise e [cf. Eq. (1.11)], we employ variance reduction techniques, based on importance sampling ideas. The corresponding methods and variance analysis is discussed in Chapter 3.

In summary, the contributions of this work 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:

[2] 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.

- A regression 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 $\hat{G}$ and $\hat{c}$, and by reducing the effect of near-singularity of G through regularization (Chapter 2).
- Nearly optimal importance sampling schemes that can effectively reduce the variances of the components of $\hat{G}$ and $\hat{c}$ (Chapter 3).
- Extensions of the approximation-simulation methodology and analysis to a broader class of related problems, particularly the problems of constrained least squares and the problems arising from approximate dynamic programming (Chapters 4 and 5).

- The derivation of confidence regions and a large deviation bound that quantify the effect of near-singularity of G on the error $(\hat{r} - r^*)$ (Chapter 2).
- 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 (Chapter 3).

The regression and variance reduction ideas of Chapters 1-4 are brought together in an algorithmic methodology that is successfully applied in Chapter 6 to some standard examples of inverse problems of very large dimension $(n \geq 10^9)$[3]. The regression approach, together with the confidence region and large deviation analysis of Chapter 2, also apply to the more general system $\hat{G}r = \hat{c}$, where $\hat{G}$ and $\hat{c}$ are simulation-based approximations to an $s \times s$ matrix G that is not necessarily positive definite or symmetric, and a vector c.

[3]We acknowledge the efforts made by Nick Polydorides for development of these computational experiments.

Chapter 2

Regression Methodology

According to the approximation framework introduced in Chapter 1, we want to estimate the solution of the equation

$$G\, r = c,$$

based on the estimates $\hat{G}$ and $\hat{c}$ obtained by sampling. Direct solution of the estimated system using $\hat{r} = \hat{G}^{-1}\hat{c}$ must contend with two potential difficulties: near-singularity of G and the simulation error induced in estimating G and c. If G is close to be singular, a small simulation error, i.e. $G - \hat{G}$ and $c - \hat{c}$, may result in a large solution error $\hat{r} - r^*$.

This motivates us to consider a regularization/regression approach in solving the system of equations with simulation error. We note that there exists a large body of literature on linear regression methods and related topics, including the generalized least squares method, the iteratively reweighted least squares method, the principal component regression, the total least squares method, the Tikhonov regularization, etc (see the books by Draper and Smith [DS66] and by Jolliffe [Jol02]). The iteratively reweighted least squares method (IRLS) is used for solving optimization problems that can be formulated into least squares problems with covariance matrix depending on the unknown variables (see [BP79], [Gre84], [BL97]). This is an iterative method that solves a series of least squares problems while updating the covariance matrix

accordingly, and has been proved to converge when the cost function is strictly convex and satisfy certain additional conditions (see [BP79]).

Our approach bears similarity with the iteratively reweighted least squares method, and can be extended to a more general context of solving systems of equations with simulation error. We will later show that, our approach of iterative regression converges locally under a relatively mild condition (Section 2.1). Furthermore, we will analyze the error bound for the approximate solution $\hat{r}$ yielded by the regression. In conjunction with the simulation framework introduced in Chapter 1, we will derive a confidence region for $\hat{r}$ (Section 2.2), and an upper bound on the large deviation rate for the convergence of $\hat{r}$ as the number of samples increases (Section 2.3).

2.1 Regression Using Sample Covariance

Let us consider the estimation of $r^* = G^{-1}c$ [cf. Eq. (1.4)] using the model

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

[cf. Eq. (1.10)], where

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

[cf. Eq. (1.11)]. The standard least squares/regression approach yields the estimate

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

where $\bar{r}$ is an *a priori* estimate (for example $\bar{r} = \hat{G}^{-1}\hat{c}$ or a singular valued-based estimate of $\hat{G}^{-1}\hat{c}$), and Σ and Γ are some positive definite symmetric matrices. Equivalently,

$$\hat{r} = (\hat{G}'\Sigma^{-1}\hat{G} + \Gamma^{-1})^{-1}(\hat{G}'\Sigma^{-1}\hat{c} + \Gamma^{-1}\bar{r}). \tag{2.1}$$

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 $\{(G_k, c_k) \mid k = 0, \ldots, t\}$, where the components of the matrix G_k

and the vector c_k are the terms appearing in the summations of Eqs. (1.7) and (1.8), respectively:

$$G_{k,\ell q} = \frac{\zeta_{i_k} a_{i_k j_k} a_{i_k \bar{j}_k} \phi_{j_k \ell} \phi_{\bar{j}_k q}}{\xi_{i_k, j_k, \bar{j}_k}}, \qquad c_{k,\ell} = \frac{\zeta_{i_k} a_{i_k j_k} b_{i_k} \phi_{j_k \ell}}{\xi_{i_k, j_k}}.$$

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

$$e_k = (G_k - \hat{G})\tilde{r} + (\hat{c} - c_k), \qquad k = 0, \ldots, t,$$

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

$$\Sigma = \frac{1}{t+1} \sum_{k=0}^{t} e_k e_k' = \frac{1}{t+1} \sum_{k=0}^{t} \big((G_k - \hat{G})\tilde{r} + (\hat{c} - c_k)\big)\big((G_k - \hat{G})\tilde{r} + (\hat{c} - c_k)\big)'. \quad (2.2)$$

In our experiments (see Chapter 6) 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

$$\Sigma = \Sigma_c + \begin{bmatrix} r' & 0 & \ldots & 0 \\ 0 & r' & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 0 & r' \end{bmatrix}_{s \times s^2} \Sigma_G \begin{bmatrix} r & 0 & \ldots & 0 \\ 0 & r & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 0 & r \end{bmatrix}_{s^2 \times s} \quad (2.3)$$

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

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

where $\operatorname{cov}(\hat{g}_i', \hat{g}_j')$ 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. (2.2) or Eq. (2.3), is to use an *iterative regression* approach: iterate using Eq. (2.1), and estimate r repeatedly with intermediate correction of the matrix Σ. This is the iteration

$$r_{k+1} = \big(\hat{G}'\Sigma(r_k)^{-1}\hat{G} + \Gamma^{-1}\big)^{-1}\big(\hat{G}'\Sigma(r_k)^{-1}\hat{c} + \Gamma^{-1}\bar{r}\big), \tag{2.4}$$

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

$$\Sigma(r) = \frac{1}{t+1}\sum_{k=0}^{t}\big((G_k - \hat{G})r + (\hat{c} - c_k)\big)\big((G_k - \hat{G})r + (\hat{c} - c_k)\big)';$$

[cf. Eq. (2.2)].

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 $\hat{G}$ and $\hat{c}$ 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. (2.3), $\Sigma(r)$

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. (2.3)], so the mapping of Eq. (2.4),

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

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.

Using the sample covariances of $\hat{G}$ and $\hat{c}$ 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 Chapter 6, as well as with small test problems; see Appendix B.1. Let us also note that in our tests, the iterative regression scheme (2.4), when it converged, typically converged within few iterations and 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.

2.2 Confidence Region

Up till now, we have introduced the essential elements of our approach to estimate r^* [cf. Eq. (1.4)] by using regression, based on the estimates $\hat{G}$ and $\hat{c}$ obtained using 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 2.1, in conjunction with the simulation formulas of Eqs. (1.7)-(1.8) given in Chapter 1.

We will derive a $(1-\theta)$-confidence region for the approximate solution $\hat{r}$, where θ is a given small positive scalar. We consider the case where regularization of the form $\Gamma^{-1} = \beta I$ is used, for some $\beta > 0$. Then the solution $\hat{r}$ of Eq. (2.1) can be rewritten as

$$\hat{r} = \left(\hat{G}'\Sigma^{-1}\hat{G} + \beta I\right)^{-1}\left(\hat{G}'\Sigma^{-1}\hat{c} + \beta\bar{r}\right), \tag{2.5}$$

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 region for the error $\hat{r} - r^*$. Let us denote

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

so from Eq. (2.5), the error can be written as

$$\hat{r} - r^* = \left(\hat{G}'\Sigma^{-1}\hat{G} + \beta I\right)^{-1}\left(\hat{G}'\Sigma^{-1/2}d + \beta(\bar{r} - r^*)\right). \tag{2.6}$$

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

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

For a large number of samples, we may assume 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

$$\|\hat{d}\| \leq \sqrt{P^{-1}(1-\theta; s)} \tag{2.7}$$

with probability $(1-\theta)$, where $P(\,\cdot\,; s)$ is the regularized Gamma function of degree s and $P^{-1}(\,\cdot\,; s)$ denotes its inverse function. 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 2.1. 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 region for the error $\hat{r} - r^*$ (assuming that $\hat{d}$ can be treated as a Gaussian random variable).

Proposition 1 *We have*

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

where

$$\begin{aligned}\sigma(\Sigma, \beta) = &\max_{i=1,\ldots,s} \left\{ \frac{\lambda_i}{\lambda_i^2 + \beta} \right\} \left\| \Sigma^{-1/2} \hat{\Sigma}^{1/2} \right\| \sqrt{P^{-1}(1-\theta; s)} \\ &+ \max_{i=1,\ldots,s} \left\{ \frac{\beta}{\lambda_i^2 + \beta} \right\} \|\bar{r} - r^*\|,\end{aligned} \tag{2.8}$$

and $\lambda_1, \ldots, \lambda_s$ 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 $\Lambda = \mathrm{diag}\{\lambda_1, \ldots, \lambda_s\}$, and U, V are unitary matrices ($UU' = VV' = I$). Then, Eq. (2.6) becomes

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

where the third equality follows from Eq. (2.9). 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 $(\bar{r} - r^*)$

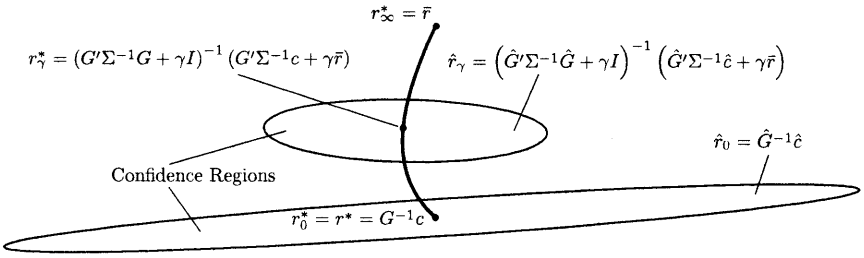


Figure 2-1: Illustration of confidence regions for different values of the regularization parameter β. For different values of $\beta \in [0, \infty]$, the figure shows the estimates $\hat{r}_\beta$, corresponding to a finite number of samples, and the exact values r^*_β, corresponding to an infinite number of samples. By Eq. (2.6), we may view $r^*_\beta - \hat{r}$ as the sum of a "simulation error" whose norm is bounded by the first term in the estimate (2.8), and a "regularization error" whose norm is bounded by the second term in the estimate (2.8).

has singular values $\beta/(\lambda_i^2 + \beta)$. Taking the norm of both sides and using the triangle inequality, it follows that

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

Since Eq. (2.7) holds with probability $(1-\theta)$, the desired result follows. ■

Note that the constant $\sigma(\Sigma, \beta)$ of Eq. (2.8) is 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. 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. 2-1).

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

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

with probability $(1-\theta)$. 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 region involving the singular values of G rather than $\hat{G}$; see Appendix A.1. While the derivation is more complicated, it supports the qualitative conclusion that the radius of the confidence region is proportional to $\|G^{-1}\|$.

2.3 Large Deviation Bound

In this section, we attempt to analyze convergence of the estimate $\hat{r}$ [cf. Eq. (2.1)] as the sampling effort grows. In particular, we focus on the large deviation behavior of $\hat{r}$ as the number of samples t increases, in conjunction with the estimation formulas Eqs. (1.7)-(1.8).

Let $\hat{\Sigma}_1$ be the covariance of $G_1 r^* - c_1$ where G_1 and c_1 are estimates of G and c using only 1 sample, so that

$$\hat{\Sigma}_1 = t\,\hat{\Sigma}.$$

Let also

$$\hat{d}_t = \hat{\Sigma}_1^{-1/2}(\hat{c} - \hat{G}r^*) = \hat{\Sigma}_1^{-1/2}\Sigma^{1/2}d, \tag{2.9}$$

hence $\hat{d}_t$ is the mean of t independent unit gaussian random variables.

We now derive the following upper bound on the large deviation rate of the error $\hat{r} - r^*$ (assuming that $\hat{d}_t$ can be treated as the mean of t independent Gaussian random variables):

Proposition 2 *For any $y > 0$, we have*

$$\begin{aligned}\limsup_{t\to\infty} \frac{1}{t}\log \mathbf{P}\left(\|\hat{r} - r^*\| \geq \max_{i=1,\ldots,s}\left\{\frac{\beta}{\lambda_i^2+\beta}\right\}\|\bar{r} - r^*\| + y\right)\\ \leq -\frac{1}{2}\min_{i=1,\ldots,s}\left\{\left(\lambda_i + \frac{\beta}{\lambda_i}\right)^2\right\}\|\Sigma^{-1}\hat{\Sigma}_1\|^{-1},\end{aligned} \tag{2.10}$$

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

Proof. In analogy with Eq. (2.6), we may write

$$\hat{r} - r^* = W\hat{d}_t + \left(\hat{G}'\Sigma^{-1}\hat{G} + \beta I\right)^{-1} \beta(\bar{r} - r^*), \tag{2.11}$$

where we define $W = \left(\hat{G}'\Sigma^{-1}\hat{G} + \beta I\right)^{-1} \hat{G}'\Sigma^{-1}\hat{\Sigma}_1^{1/2}$. According to the proof of Prop. 1, the second term of the righthand-side of Eq. (2.11) satisfies that

$$\left\| \left(\hat{G}'\Sigma^{-1}\hat{G} + \beta I\right)^{-1} \beta(\bar{r} - r^*) \right\| \leq \max_{i=1,\ldots,s} \left\{ \frac{\beta}{\lambda_i^2 + \beta} \right\} \|\bar{r} - r^*\|.$$

By comparing the above relation with Eq. (2.11), we may obtain, for any $y > 0$, that

$$\mathbf{P}\left(\|\hat{r} - r^*\| \geq \max_{i=1,\ldots,s} \left\{ \frac{\beta}{\lambda_i^2 + \beta} \right\} \|\bar{r} - r^*\| + y \right) \leq \mathbf{P}(\|W\hat{d}_t\| \geq y). \tag{2.12}$$

Note $\hat{d}_t = \hat{\Sigma}_1^{-1/2}(\hat{G}r^* - \hat{c})'$ is the mean of t independent unit gaussian random variables. So $W\hat{d}_t$ is the sum of t independent gaussian random variables with covariance WW'. According to the large deviation theory for multi-dimensional gaussians, we have

$$\lim_{t\to\infty} \frac{1}{t} \log \mathbf{P}(\|W\hat{d}_t\| \geq y) = -\frac{1}{2}\|WW'\|^{-1}y^2 = -\frac{1}{2}\|W\|^{-2}y^2. \tag{2.13}$$

Using the matrix norm inequality and the singular value decomposition given in the proof of Prop. 1, we obtain

$$\begin{aligned} \|W\| &= \left\| \left(\hat{G}'\Sigma^{-1}\hat{G} + \beta I\right)^{-1} (\hat{G}'\Sigma^{-1}\hat{\Sigma}_1^{1/2}) \right\| \\ &\leq \left\| \left(\hat{G}'\Sigma^{-1}\hat{G} + \beta I\right)^{-1} (\hat{G}'\Sigma^{-1/2}) \right\| \left\| \Sigma^{-1/2}\hat{\Sigma}_1^{1/2} \right\| \\ &\leq \max_{i=1,\ldots,s} \left\{ \frac{\lambda_i}{\lambda_i^2 + \beta} \right\} \left\| \Sigma^{-1/2}\hat{\Sigma}_1^{1/2} \right\|. \end{aligned}$$

Hence (2.13) becomes

$$\limsup_{t\to\infty} \frac{1}{t} \log \mathbf{P}(\|W\hat{d}_t\| \geq y) \leq -\frac{1}{2} y^2 \left(\max_{i=1,\ldots,s} \left\{ \frac{\lambda_i}{\lambda_i^2 + \beta} \right\} \left\| \Sigma^{-1/2} \hat{\Sigma}_1^{1/2} \right\| \right)^{-2}$$
$$= -\frac{1}{2} \min_{i=1,\ldots,s} \left\{ \left(\lambda_i + \frac{\beta}{\lambda_i} \right)^2 \right\} \|\Sigma^{-1}\hat{\Sigma}_1\|^{-1}.$$

Finally, by combining the above relation with (2.12), we complete the proof. ■

Let us consider again the limiting case where $\beta = 0$ and $\Sigma = \hat{\Sigma}_1$, assuming that $\hat{G}$ is invertible. Now the approximate solution is $\hat{r} = \hat{G}^{-1}\hat{c}$, and the preceding proof can be used to show that

$$\limsup_{t\to\infty} \frac{1}{t} \log \mathbf{P}\left(\|\hat{G}^{-1}\hat{c} - r^*\| \geq y \right) \leq -\frac{1}{2} \min_{i=1,\ldots,s} \{\lambda_i^2\}\, y^2,$$

where $\lambda_1, \ldots, \lambda_s$ are singular values of $\hat{\Sigma}_1^{-1/2}\hat{G}$. This implies that the rate of exponential convergence of $\hat{r}$ is determined by near-singularity of the matrix $\hat{G}$, and hence by near-singularity of the matrix G (since $\hat{G}$ is close to G). On the other hand, by taking β to be a reasonable positive value, we may tune up the rate of convergence of $\hat{r}$ (corresponding to the righthand-side of Eq. (2.10)), while enduring a small constant bias (corresponding to the term $\max_{i=1,\ldots,s} \left\{ \frac{\beta}{\lambda_i^2 + \beta} \right\} \|\bar{r} - r^*\|$ on the lefthand-side of Eq. (2.10)).

Remarks:

Our regression approach has been applied successfully to several temporal difference-type methods used in solving approximate DP problems (given in Chapter 5), and to several large-scale practical inverse problems (given in Chapter 6). The analysis of Sections 2.2 and 2.3 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.

We finally note that the ability of our regularization/regression 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 (2.1). The theoretical results on confidence regions and the large deviation bound of Sections 2.2 and 2.3 also apply to this general case, provided that the estimates $\hat{G}$ and $\hat{c}$ have known statistics.

Chapter 3

Variance Reduction by Importance Sampling

In this chapter, we return to the simulation framework introduced in Chapter 1 and focus on the issue of variance reduction of simulation. In particular, we will consider the use of importance sampling and some related theoretical aspects (Sections 3.1-3.2). Then we will propose a few practical approaches for designing near-optimal sampling distributions, in the context of our approximation-simulation framework for solution of large least squares problems (Section 3.3).

Recall from Chapter 1 that the simulation that generates the estimates $\hat{G}$ and $\hat{c}$ using Eqs. (1.7)-(1.8) can be carried out in several ways. For example, we may generate a single sequence of independent index triples $\{(i_0, j_0, \bar{j}_0), \ldots, (i_t, j_t, \bar{j}_t)\}$ 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 and 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 multidi-

mensional sums or integrals (see the survey paper by Halton [Hal70] and the book by Evans and Swartz [ES00]). Recent developments on importance sampling 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 importance sampling that we have used. In particular, we will derive estimates of the covariances of the estimation errors $\hat{G} - G$ and $\hat{c} - c$, 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 and c using Eqs. (1.7)-(1.8) 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. (1.2)]. 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

$$z = \sum_{\omega \in \Omega} v(\omega), \tag{3.1}$$

where Ω is a finite set and $v : \Omega \mapsto \Re^d$ is a function of $\omega \in \Omega$. 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 importance sampling 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

$$\hat{z} = \frac{1}{T}\sum_{t=1}^{T}\frac{v(\omega_t)}{\xi(\omega_t)}. \tag{3.2}$$

Clearly $\hat{z}$ is unbiased:

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

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

$$\mathrm{cov}(\hat{z}) = \frac{1}{T}\left(\sum_{\omega\in\Omega}\frac{v(\omega)v(\omega)'}{\xi(\omega)} - zz'\right). \tag{3.3}$$

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. (3.3) becomes

$$\mathrm{var}(\hat{z}) = \frac{z^2}{T}\left(\sum_{\omega\in\Omega}\frac{\left(v(\omega)/z\right)^2}{\xi(\omega)} - 1\right). \tag{3.4}$$

Assuming that $v(\omega) \geq 0$ for all $\omega \in \Omega$,[1] the optimal distribution is $\xi^* = v/z$ and the corresponding minimum variance value is 0. However, ξ^* cannot be computed without knowledge of z.

[1]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 $z_2 = \sum_{\omega\in\Omega}v^-(\omega)$.

(ii) $d > 1$: In this case, the covariance of $\hat{z}$ [cf. Eq. (3.3)] is a matrix that cannot be minimized directly. One possibility is to minimize instead an estimate of a norm of the matrix $\sum_{\omega\in\Omega}\left(v(\omega)v(\omega)'/\xi(\omega)\right)$. We have

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

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

$$\xi^*(\omega) = C\cdot\left\|v(\omega)v(\omega)'\right\|^{\frac{1}{2}}, \qquad \omega\in\Omega, \tag{3.5}$$

where C is a normalizing constant.

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

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

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

$$\xi^*(\omega) = \frac{y'v(\omega)}{y'z}, \qquad \omega\in\Omega. \tag{3.6}$$

Note that calculating exactly ξ^* is impractical with both formulas (3.5) and (3.6).

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 $\nu = y'v$ if we are interested in the uncertainty of $\hat{z}$ along a specific direction y. The probability distribution ξ^* minimizes the cost function

$$F_\xi = \sum_{\omega\in\Omega}\frac{\nu(\omega)^2}{\xi(\omega)} \tag{3.7}$$

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

$$\xi^*(\omega) = C \cdot \nu(\omega), \qquad \omega \in \Omega, \tag{3.8}$$

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. (3.7) 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

$$\hat{\xi}(\omega) = \hat{C} \cdot \hat{\nu}(\omega), \qquad \omega \in \Omega, \tag{3.9}$$

where $\hat{\nu}$ is an approximation to ν and $\hat{C}^{-1} = \sum_{\omega \in \Omega} \hat{\nu}(\omega)$, 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 pairs of trial points and corresponding function values $\{(\bar{\omega}_1, \nu(\bar{\omega}_1)), (\bar{\omega}_2, \nu(\bar{\omega}_2)), \dots\}$.

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

(iv) Generate the sample sequence $\{(\omega_1, v(\omega_1)), \dots, (\omega_T, v(\omega_T))\}$ according to $\hat{\xi}$ and compute the estimate $\hat{z}$ using Eq. (3.2).

For further insight into the preceding procedure, it is useful to introduce the following normalized version of the cost function F_ξ of Eq. (3.7):

$$D_\xi = \frac{1}{\left(\sum_{\omega\in\Omega}\nu(\omega)\right)^2}\sum_{\omega\in\Omega}\frac{\nu(\omega)^2}{\xi(\omega)}, \tag{3.10}$$

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

$$\begin{aligned} &\text{minimize} \quad D_\xi \\ &\text{s.t.} \quad \sum_{\omega\in\Omega}\xi(\omega) = 1, \quad \xi \geq 0. \end{aligned} \tag{3.11}$$

Using Eqs. (3.9)-(3.10), we can express $D_{\hat{\xi}}$ as

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

We have

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

and

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

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

$$D_{\hat{\xi}} \leq \max_{\omega\in\Omega}\frac{\hat{\nu}(\omega)}{\nu(\omega)}\cdot\max_{\omega\in\Omega}\frac{\nu(\omega)}{\hat{\nu}(\omega)}. \tag{3.12}$$

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 (3.12) to D_ξ small, and hence also the cost function F_ξ small.

As a final remark of this section, there are situations where ξ^* has a complicated form but can be written as the product of a finite number of relatively simple

distributions, i.e.,

$$\xi^* = \xi_1^* \cdots \xi_K^*.$$

This is often the case when ξ^* is multidimensional and the simulation is done sequentially. Then it is convenient to approximate the simple distributions $\xi_1^*, \ldots, \xi_K^*$ with $\hat{\xi}_1, \ldots, \hat{\xi}_K$ respectively, and then to calculate $\hat{\xi}$ by

$$\hat{\xi} = \hat{\xi}_1 \cdots \hat{\xi}_K.$$

Let $D_{\hat{\xi}}$ be the divergence factor of $\hat{\xi}$ with respect to ξ^*, and let $D_{\hat{\xi}_k}$ be the divergence factor of $\hat{\xi}_k$ with respect to ξ_k^*. We can show that

$$D_{\hat{\xi}} = \sum_{\omega \in \Omega} \frac{\xi_1^*(\omega)^2 \cdots \xi_K^*(\omega)^2}{\hat{\xi}_1(\omega) \cdots \hat{\xi}_K(\omega)} \leq \left(\sum_{\omega \in \Omega} \frac{\xi_1^*(\omega^2)}{\hat{\xi}_1(\omega)} \right) \cdots \left(\sum_{\omega \in \Omega} \frac{\xi_K^*(\omega)^2}{\hat{\xi}_K(\omega)} \right) = D_{\hat{\xi}_1} \cdots D_{\hat{\xi}_K}.$$

Hence by using estimates of divergence factors for the simple distributions, we may obtain an upper bound for the divergence factor of the product distribution.

3.2 Variance Analysis of IS Distribution Obtained by Piecewise Approximation

Let us consider approximation of the optimal sampling distribution $\xi^* = C \cdot \nu$ [cf. Eq. (3.8)] 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

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

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. (3.12), 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. (3.12) 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

$$\hat{\nu} = \sum_{k=1}^{K} \nu(\omega_k) \cdot \mathbf{1}_{\Omega_k}, \tag{3.13}$$

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. (3.12), under some reasonable smoothness conditions.

Proposition 3 *If* $\log \nu$ *exists and is Lipschitz continuous with Lipschitz constant* $\eta > 0$, *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. (3.12), yields the desired result. ■

Proposition 4 *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. (3.12), 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 5 *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

$$\begin{aligned} \nu(\omega) &= \nu(\omega_k) + \int_0^1 \nabla\nu\big(\omega_k + t(\omega - \omega_k)\big)\mathrm{d}t \\ &= \nu(\omega_k) + \nabla\nu(\omega_k)'(\omega - \omega_k) \\ &\quad + \int_0^1 \Big(\nabla\nu\big(\omega_k + t(\omega - \omega_k)\big) - \nabla\nu(\omega_k)\Big)\,\mathrm{d}t. \end{aligned} \tag{3.14}$$

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. (3.14) 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. (3.12), 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. 4 (which involves ρ) with the one of Prop. 5 (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. (3.12). 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. 3-5.

3.3 Designing Near-Optimal Sampling Schemes

3.3.1 An Importance Sampling Scheme for Estimating $G_{\ell q}$

As an illustration of the preceding importance sampling ideas, let us focus on the estimation of the component $G_{\ell q}$ by generating a sequence of index triples and using the following equation [cf. Eq. (1.7)]:

$$\hat{G}_{\ell q} = \frac{1}{t+1} \sum_{k=0}^{t} \frac{\zeta_{i_k} a_{i_k j_k} a_{i_k \bar{j}_k} \phi_{j_k \ell} \phi_{\bar{j}_k q}}{\xi_{i_k, j_k, \bar{j}_k}}.$$

In this case the sample space Ω and the function ν are

$$\Omega = \{1, \ldots, n\}^3, \qquad \nu(i, j, \bar{j}) = \zeta_i a_{ij} a_{i\bar{j}} \phi_{j\ell} \phi_{\bar{j}q}.$$

We want to design the sampling distribution ξ so that it is close to ξ^* and belongs to some family of relatively simple distribution functions.

We have used a scheme that generates the indices i, j, and $\bar{j}$ sequentially. The

optimal distribution satisfies

$$\xi^*_{ij\bar{j}} \propto (\phi_{j\ell}\|a_j\|_1)\,(\phi_{\bar{j}q}\|a_{\bar{j}}\|_1)\frac{\zeta_i a_{ij} a_{i\bar{j}}}{\|a_j\|_1\|a_{\bar{j}}\|_1},$$

where we denote by a_j the jth column of A, and denote by $\|a_j\|_1$ the L_1 norm of a_j (i.e., $\|a_j\|_1 = \sum_{i=1}^n |a_{ij}|$). We approximate ξ^* by approximating the functions

$$\phi_{j\ell}\|a_j\|_1, \qquad \phi_{\bar{j}q}\|a_{\bar{j}}\|_1, \qquad \frac{\zeta_i a_{ij} a_{i\bar{j}}}{\|a_j\|_1\|a_{\bar{j}}\|_1}$$

with distributions

$$\xi_j, \qquad \xi_{\bar{j}}, \qquad \xi(i \mid j, \bar{j}),$$

respectively, so $\xi^*_{ij\bar{j}}$ is approximated with $\xi_{ij\bar{j}} = \xi_j\,\xi_{\bar{j}}\,\xi(i \mid j, \bar{j})$.

Let us denote by $\mathrm{T}(\cdot)$ the approximation operator that maps a set of trial values of ν to an approximation of the entire function ν. For instance, we can take $\mathrm{T}(\cdot)$ to be a *piecewise constant approximation*: for any $y \in \Re^n$ and $I = \{i_1, \ldots, i_K\} \subset \{1, \ldots, n\}$,

$$\mathrm{T}\left([y_i]_{i\in I}\right) = \sum_{k=1}^{K} y_{i_k}\,\mathbf{1}\left([(i_{k-1} + i_k)/2, (i_k + i_{k+1})/2]\right),$$

where $\mathbf{1}$ denotes the function that is equal to 1 within the corresponding interval, and 0 otherwise, and we define $i_0 = -i_1$ and $i_{K+1} = 2n - i_K$. For another example, we may take T to be the *piecewise linear approximation*,

$$\mathrm{T}\left([y_i]_{i\in I}\right) = \sum_{k=1}^{K-1} \mathbf{L}\left((i_k, y_{i_k}), (i_{k+1}, y_{i_{k+1}})\right)\mathbf{1}\left([i_k, i_{k+1}]\right),$$

where we denote by $\mathbf{L}$ the linear function that takes value y_{i_k} at i_k and value $y_{i_{k+1}}$ at i_{k+1}, and assume without loss of generality that $i_1 = 0$ and $i_K = n$. The resulting importance sampling algorithm is summarized in what follows.

An Importance Sampling Scheme for Estimating $G_{\ell q}$:

1. Select a small set $[a_{ij}]_{i,j\in I}$ of components of A, and a corresponding small set of rows $[\phi'_j]_{j\in I}$ of Φ.
2. Generate the sample triple $(i_k, j_k, \bar{j}_k)$ by
   (a) sampling j_k according to $\xi_{j_k} \propto \mathrm{T}_{\mathrm{j_k}}\left(\left[\phi_{\mathrm{j}\ell}\sum_{\mathrm{i}\in\mathrm{I}} \mathrm{a_{ij}}\right]_{\mathrm{j}\in\mathrm{I}}\right)$,
   (b) sampling $\bar{j}_k$ according to $\xi_{\bar{j}_k} \propto \mathrm{T}_{\bar{\mathrm{j}}_\mathrm{k}}\left(\left[\phi_{\mathrm{jq}}\sum_{\mathrm{i}\in\mathrm{I}} \mathrm{a_{ij}}\right]_{\mathrm{j}\in\mathrm{I}}\right)$,
   (c) sampling i_k conditioned on j_k and $\bar{j}_k$ according to

$$\xi(i_k \mid j_k, \bar{j}_k) \propto \mathrm{T}_{\mathrm{i_k}}\left(\left[\zeta_{\mathrm{i}}\mathrm{a_{ij_k}}\mathrm{a_{i\bar{j}_k}}\right]_{\mathrm{i}\in\mathrm{I}}\right)$$

   where we denote by $\mathrm{T}_j(\cdot)$ the jth component of $\mathrm{T}(\cdot)$.

Figure 3-1 illustrates the last step of this importance sampling scheme, where A is an 1000×1000 matrix, $Z = I$ and $\mathrm{T}(\cdot)$ is taken to be the operator of piecewise constant/linear approximation. We start with a low-dimensional representation of A, namely $[a_{ij}]_{i,j\in I}$, which can be implemented using a uniformly spaced discretization as illustrated in Fig. 3-1(a). The resulting distributions $\xi(i_k \mid j_k, \bar{j}_k)$ are plotted in Fig. 3-1(b)-(c), and compared with the exact optimal conditional distribution.

3.3.2 Variations

The importance sampling scheme given in the preceding section is only one possibility for generating samples to estimate $G_{\ell q}$. An alternative is to replace the distributions in steps 2(a) and 2(b) with

$$\xi_{j_k} \propto \phi_{j_k \ell}, \qquad \xi_{\bar{j}_k} \propto \phi_{j_k q},$$

or with approximations of the above functions. This simplified version is easier to implement, and may reduce the computational complexity greatly if Φ is known to have a simple analytical form.

We may also change the order of generating i_k, j_k, and $\bar{j}_k$. For instance, we can generate i_k first, and then j_k and $\bar{j}_k$ conditioned on i_k, according to the distributions

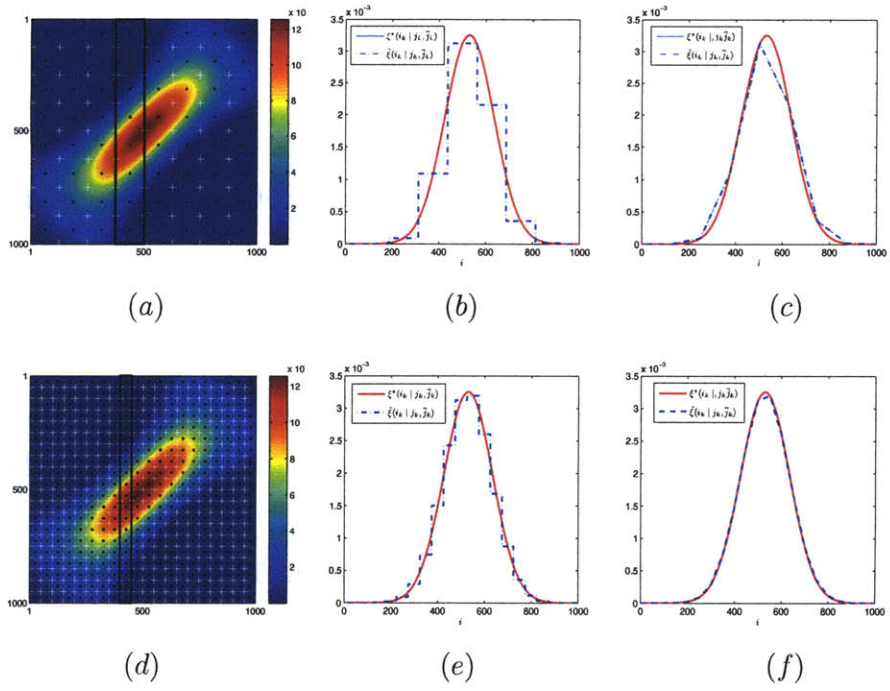


Figure 3-1: Illustration of step 2(c) of the proposed IS scheme for estimation of $G_{\ell q}$. In (a), the color field represents the 1000×1000 matrix A; the two vertical lines represent the columns a_{j_k} and $a_{\bar{j}_k}$; and the grid represents $[a_{ij}]_{i,j\in I}$, which is an 8×8 discretization of A. In (b)/(c) the conditional distribution $\xi(i_k \mid j_k, \bar{j}_k)$ (obtained by piecewise constant/linear approximation using $[a_{ij}]_{i,j\in I}$) is plotted against the optimal distribution $\xi^*(i_k \mid j_k, \bar{j}_k)$. In (d)-(f) the same process is repeated with a finer 20×20 discretization of A.

$$\xi_{i_k} \propto \|a_{i_k}\|_1, \qquad \xi(j_k \mid i_k) \propto \phi_{j_k \ell} a_{i_k j_k}, \qquad \xi(\bar{j}_k \mid i_k) \propto \phi_{j_k q} a_{i_k j_k}.$$

If A and Φ have complicated forms, we may first replace them with coarse approximations, and then introduce a step of function approximation when computing the above distributions. When A has relatively sparse rows, by sampling the row index first, we may greatly improve the efficiency of sampling.

The most straightforward scheme is to approximate the three-dimensional function $\nu = \zeta_i a_{ij} a_{i\bar{j}} \phi_{j\ell} \phi_{\bar{j}q}$ directly: first take trial samples from the sample space $\Omega = \{1, \ldots, n\}^3$ and approximate ν by fitting some function (e.g., a piecewise constant/linear function) based on the trial samples. More specifically, we may take $I \subset \{1, \ldots, n\}$

and obtain $[\nu(i,j,\bar{j})]_{i,j,\bar{j}\in I}$. Then we can compute the approximate function by

$$\hat{\nu} = \mathrm{T}\left([\nu(i,j,\bar{j})]_{i,j,\bar{j}\in I}\right),$$

where we maintain the notation $\mathrm{T}(\cdot)$ for the operator of function approximation, and finally normalize $\hat{\nu}$ to obtain a distribution function. However, this scheme may be computationally expensive, because it involves selecting trial samples from a three-dimensional space and then sampling according to a three-dimensional distribution. A critical choice in all the schemes mentioned above is the function approximation operator $\mathrm{T}(\cdot)$, and a good choice may depend on the characteristics of the problem at hand, i.e., A and Φ.

Also, as an alternative to the piecewise constant/linear approximation used in Figure 3-1, we may consider an approximation approach based on a least squares fit from some family of parameterized functions. In particular, we may approximate the function $\nu : \Omega \mapsto \Re$ by introducing a parametric family, which we denote by

$$\mathfrak{F} = \left\{ f_\eta \;\middle|\; \eta \in \Re^d \right\},$$

where $f_\eta : \Omega \mapsto \Re$ is a function parameterized by η; or we may consider the family of a finite sum of parameterized functions, in which case

$$\mathfrak{F} = \left\{ \sum_{k=1}^{M} f_{\eta_k} \;\middle|\; \eta_k \in \Re^d, k = 1, \ldots, M \right\},$$

where M is a positive integer. Given the trial samples and corresponding function values $\{\nu(i,j,\bar{j})\}_{i,j,\bar{j}\in I}$, we can approximate ν with $\hat{\nu}$ by minimizing the total squared error corresponding to the trial samples, i.e.,

$$\mathrm{T}\left([\nu(i,j,\bar{j})]_{i,j,\bar{j}\in I}\right) = \operatorname*{argmin}_{\hat{\nu}\in\mathfrak{F}} \sum_{i,j,\bar{j}\in I} \|\nu(i,j,\bar{j}) - \hat{\nu}(i,j,\bar{j})\|^2.$$

This approach may be preferable under circumstances where ν is known to have some special structure, in which case we may choose $\mathfrak{F}$ accordingly and improve the

approximation accuracy. As a final remark, although we have focused on designing importance sampling schemes for estimation of the component $G_{\ell q}$, the preceding algorithmic approach can be easily extended to estimating components of c, and more generally to estimating other quantities that take the form of a large sum.

Chapter 4

Extensions and Related Methods

In the preceding chapters we have focused on the least squares problem

$$\min_{x \in \Re^n} \|Ax - b\|_\zeta^2 \tag{4.1}$$

[cf. Eq. (1.2)]. Now we want to extend the preceding methodology and analysis to a broader class of problems. In particular, we will consider the projected equation approach for square problems (Section 4.1), a special least squares approach for underdetermined problems (Section 4.2), and the constrained least squares problems (Section 4.3).

4.1 Projected Equation Approach for Square Problems

For square linear systems $Ax = b$ [cf. Eq. (1.1)], we consider an alternative approximation approach to the least squares-based approach [cf. Section 1.2]. Equation (1.1) may be written as a fixed point equation

$$x = T(x) = (I - A)x + b.$$

As earlier, we introduce an $n \times s$ matrix Φ, and assume that the columns of Φ are linearly independent. Letting S be the subspace spanned by columns of Φ, we want to approximate the fixed point solution with $x = \Phi r$. We consider an approximate version of the fixed point equation,

$$\Phi r = \Pi T(\Phi r), \tag{4.2}$$

where we denote by Π the projection to S with respect to some weighted Euclidean norm. This is known as the *projected equation.*

We assume that ΠA is invertible, so there exists a unique solution r^* of Eq. (4.2). We have

$$r^* = \operatorname*{argmin}_{r \in \Re^s} \|\Phi r - ((I - A)\Phi r^* + b))\|_\zeta, \tag{4.3}$$

where $\|\cdot\|_\zeta$ denotes the weighted Euclidean norm and ζ is a probability distribution vector with positive components. Equation (4.3) is equivalent to

$$r^* = C^{-1}d,$$

where we define

$$C = \Phi' Z A \Phi, \qquad d = \Phi' Z b.$$

Note that this r^* is different from the least squares solution of Eq. (1.4).

Now we may generate a sequence of indexes pairs $\{(i_0, j_0), \ldots, (i_t, j_t)\}$ by independently sampling according to the distribution ξ, and may estimate C and d with

$$\hat{C} = \frac{1}{t+1} \sum_{k=0}^{t} \frac{\zeta_{i_k} a_{i_k j_k}}{\xi_{i_k j_k}} \phi_{i_k} \phi'_{j_k}, \qquad \hat{d} = \frac{1}{t+1} \sum_{k=0}^{t} \frac{\zeta_{i_k} b_{i_k}}{\xi_{i_k}} \phi_{i_k}. \tag{4.4}$$

Based on the estimates $\hat{C}$ and $\hat{d}$, we may estimate the solution r^* of the projected equation using the regression approach of Chapter 2.

This projected equation approach may be less sensitive to simulation error, compared with our least squares-based approach. Here estimating C and d using Eq. (4.4) involves sampling from a two-dimensional space, while estimating G and c [cf. Eq.

(1.5)] using Eqs. (1.7)-(1.8) involves sampling from a three-dimensional space. As a result, designing the importance sampling distributions for C and d is relatively simple. Therefore the estimation of C and d is likely to involve less simulation error, compared with the estimation of G and c.

In the context of a policy evaluation problem in DP, we take x to be the corresponding cost vector, and take $A = I - \alpha P$, where P is the probability transition matrix corresponding to a single policy and α is a scalar satisfying $0 < \alpha < 1$. Solving the simulation-based approximation to the projected equation is known as the LSTD (*least squares temporal difference*) method, which we will discuss in more detail in the subsequent Chapter 5.

4.2 A Special Case for Underdetermined Problems

In dealing with severely underdetermined problems (see [KS04] for examples of inverse problems of this type), we can estimate the components of the high-dimensional solution x^* of problem (4.1) directly, without subspace approximation. Assuming that m is reasonably small, we propose to take $\Phi = I$ and adapt the preceding regression and simulation methodology as follows.

Let $\Sigma^{-1} = I$, $\Gamma^{-1} = \beta I$, and $Z = I$ in Eq. (2.1) for some $\beta > 0$, and let $\hat{x} = \Phi\hat{r} = \hat{r}$ and $\bar{x} = \Phi\bar{r} = \bar{r}$. Equation (2.1) can be rewritten as

$$\hat{x} = \bar{x} + (A'A + \beta I)^{-1}A'(b - A\bar{x}). \tag{4.5}$$

We now note that

$$A'(AA' + \beta I) = A'AA' + \beta A' = (A'A + \beta I)A',$$

and that both matrices $(AA' + \beta I)$ and $(A'A + \beta I)$ are positive definite and thus invertible. Hence we have

$$(A'A + \beta I)^{-1}A' = A'(AA' + \beta I)^{-1}.$$

Thus Eq. (4.5) is equivalent with

$$\hat{x} = \bar{x} + A'(F + \beta I)^{-1} d, \tag{4.6}$$

where we define the $m \times m$ matrix F and the m-dimensional vector d by

$$F = AA', \qquad d = b - A\bar{x}.$$

In analogy with the estimation of G and c by using Eqs. (1.7)-(1.8), we may generate one sample sequence $\{(i_0, j_0), \ldots, (i_t, j_t)\}$ per component of F and d, and estimate $F_{\ell q}$ and d_ℓ respectively with

$$\hat{F}_{\ell q} = \frac{1}{t+1} \sum_{k=0}^{t} \frac{a_{\ell i_k} a_{q j_k}}{\xi_{i_k j_k}}, \qquad \hat{d}_\ell = b - \frac{1}{t+1} \sum_{k=0}^{t} \frac{a_{\ell i_k} \bar{x}_{i_k}}{\xi_{i_k}}. \tag{4.7}$$

Alternatively, in analogy with the estimation of G and c by using Eq. (1.9), we may generate a single sample sequence $\{(i_0, j_0), \ldots, (i_t, j_t)\}$ according to a distribution ξ, and estimate F and d with

$$\hat{F} = \frac{1}{t+1} \sum_{k=0}^{t} \frac{1}{\xi_{i_k j_k}} a_{i_k} a'_{j_k}, \quad \hat{d} = b - \frac{1}{t+1} \sum_{k=0}^{t} \frac{\bar{x}_{i_k}}{\xi_{i_k}} a_{i_k},$$

where a_{i_k} denotes the i_kth column of A. We now obtain the approximate solution $\hat{x}$ whose ith entry is computed as

$$\hat{x}_i = \bar{x}_i + a'_i (\hat{F} + \beta I)^{-1} \hat{d}.$$

In this way, we are able to estimate components of x^* directly, using only low-dimensional vector operations. We have successfully used this approach in the solution of some severely underdetermined inverse problems; see Sections 6.1-6.2 in Chapter 6.

4.3 Constrained Least Squares Problems

4.3.1 Equality Constrained Problems

As a variation of problem (4.1), consider the following equality constrained least squares problem

$$\begin{aligned} &\min_{x\in\Re^n} \|Ax-b\|_\zeta^2 \\ &\text{s.t.} \quad Lx = 0, \end{aligned} \tag{4.8}$$

where L is an $l \times n$ matrix. Following a similar approximation approach, we restrict the solution of problem (4.8) to lie within the subspace S. Now the constraint $Lx = 0$ becomes $L\Phi r = 0$ or equivalently $r'\Phi' L' L\Phi r = 0$, which is also equivalent with $\Phi' L' L\Phi r = 0$. Thus, we may write the approximate problem as

$$\begin{aligned} &\min_{r\in\Re^s} \|A\Phi r-b\|_\zeta^2, \\ &\text{s.t.} \quad \Phi' L' L\Phi r = 0. \end{aligned} \tag{4.9}$$

We assume that there exists at least one feasible solution for this problem. Introducing a Lagrange multiplier vector $\lambda \in \Re^l$ and using standard duality arguments, we obtain the following necessary and sufficient condition for (r^*, λ^*) to be an optimal solution-Lagrange multiplier pair for problem (4.9):

$$H \begin{pmatrix} r^* \\ \lambda^* \end{pmatrix} = f, \tag{4.10}$$

where we define the $2s \times 2s$ matrix H and $2s$-vector f as

$$H = \begin{pmatrix} \Phi' A' Z A\Phi & \Phi' L' L\Phi' \\ \Phi' L' L\Phi & 0 \end{pmatrix}, \qquad f = \begin{pmatrix} \Phi' A' Z b \\ 0 \end{pmatrix}.$$

We may now apply our simulation and regression approach of the preceding chapters to solution of the system (4.10) (which is always low-dimensional, even if L has

a large row dimension). In particular, similar to Eq. (1.9), we may generate a sample sequence

$$\{(i_0, j_0, \bar{j}_0), \ldots, (i_t, j_t, \bar{j}_t)\}$$

according to the distribution ξ, and estimate H and f with $\hat{H}$ and $\hat{f}$ given by

$$\hat{H} = \frac{1}{t+1} \sum_{k=0}^{t} \frac{1}{\xi_{i_k j_k \bar{j}_k}} \begin{pmatrix} \zeta_{i_k} a_{i_k j_k} a_{i_k \bar{j}_k} \phi_{j_k} \phi'_{\bar{j}_k} & l_{i_k j_k} l_{i_k \bar{j}_k} \phi_{j_k} \phi'_{\bar{j}_k} \\ l_{i_k j_k} l_{i_k \bar{j}_k} \phi_{j_k} \phi'_{\bar{j}_k} & 0 \end{pmatrix},$$

and

$$\hat{f} = \frac{1}{t+1} \sum_{k=0}^{t} \frac{\zeta_{i_k} a_{i_k j_k} b_{i_k}}{\xi_{i_k}} \begin{pmatrix} \phi_{j_k} \\ 0 \end{pmatrix}.$$

Alternatively, we may generate one sample sequence per component of H and f, and estimate the components with formulas that are similar to Eqs. (1.7)-(1.8).

4.3.2 Inequality Constrained Problems

Another variation of problem (4.1) is the inequality constrained least squares problem

$$\begin{aligned} &\min_{x \in \Re^n} \|Ax - b\|_\zeta^2 \\ &\text{s.t.} \quad Lx \leq g, \end{aligned} \tag{4.11}$$

where L is an $l \times n$ matrix and *the row dimension l is assumed to be small.* We consider an approximation of this problem restricted within the subspace S, given by

$$\begin{aligned} &\min_{r \in \Re^s} \|A\Phi r - b\|_\zeta^2 \\ &\text{s.t.} \quad L\Phi r \leq g, \end{aligned}$$

or equivalently

$$\begin{aligned} &\min_{r \in \Re^s} r'Gr - 2c'r, \\ &\text{s.t.} \quad Mr \leq g, \end{aligned}$$

where G and c are defined in Eq. (1.5), and $M = L\Phi$. We may now apply the simulation approach of preceding chapters. For example, we may generate one single sample sequence, then estimate G and c with $\hat{G}$ and $\hat{c}$ using Eq. (1.9), and estimate M with $\hat{M}$ given by

$$\hat{M} = \frac{1}{t+1} \sum_{k=0}^{t} \frac{1}{\xi_{i_k}} l_{i_k} \phi'_{i_k},$$

where we denote by l_i the ith column of L. Alternatively, we may generate one sample sequence per component of M, and estimate $M_{\ell q}$ with

$$\hat{M}_{\ell q} = \frac{1}{t+1} \sum_{k=0}^{t} \frac{1}{\xi_{i_k}} l_{\ell i_k} \phi_{i_k q}.$$

The resulting approximate problem,

$$\min_{r \in \Re^s} r'\hat{G}r - 2\hat{c}'r,$$
$$\text{s.t.} \quad \hat{M}r \leq g,$$

is low-dimensional in both the cost function and the inequality constraints. Now we can apply standard quadratic programming techniques to solve this problem. Note that it is essential to assume that L has a small row dimension, so that $\hat{M}$ has low dimension.

Chapter 5

Application to Approximate Dynamic Programming

In this chapter, we consider the application of our methodology to the class of policy evaluation algorithms for approximate solution of DP problems that are computationally intensive. These algorithms come from the field of approximate dynamic programming, or neuro-dynamic programming, or reinforcement learning (see the books by Bertsekas [Ber07], by Bertsekas and Tsitsiklis [BT96], and by Sutton and Barto [SB98], and see also the more recent books [Bor08], [CFHM07], [Liu01], [Mey07], [Pow07]). In particular, we will focus on two popular methods for solution of projected Bellman's equation based on least squares: the LSTD (*least squares temporal difference*) method and the LSPE (*least squares projected equation*) method.

5.1 Simulation-Based Projected Bellman's Equation

We focus on the evaluation of a given policy by solving the Bellman equation

$$J = g + \alpha P J,$$

where J is the cost vector, P is the state transition probability matrix specified by the given policy, g is the transition cost vector, and α is a discount factor. We approximate the cost vector using the linear architecture $J = \Phi r$, where Φ is the basis matrix and its columns span the subspace S.

We introduce the Euclidean norm $\|\cdot\|_\xi$ where ξ is the invariant distribution of P, and we consider the projected Bellman's equation to S with respect to $\|\cdot\|_\xi$. According to Section 4.1, the projected Bellman's equation can be written as

$$Cr^* = d,$$

where

$$C = \Phi'\Xi(I - \alpha P)\Phi, \qquad d = \Phi'\Xi g,$$

and Ξ is the matrix with components of ξ along its diagonal.

In the LSTD (*least squares temporal difference*) method, proposed in [BB96], we may generate a infinitely long sequence of state transitions $\{(i_0, i_1), \ldots, (i_k, i_{k+1}), \ldots\}$ according to P, and estimate C and d with

$$C_k = \frac{1}{k+1}\sum_{t=0}^{k} \phi_{i_t}\left(\phi'_{i_t} - \alpha\phi'_{i_{t+1}}\right), \qquad d_k = \frac{1}{k+1}\sum_{t=0}^{k} g_{i_t}\phi_{i_t}. \tag{5.1}$$

Another possibility, proposed by [BY09], is to generate the sequence $\{i_0, \ldots, i_k, \ldots\}$ according to P, and to generate the transitions $\{(i_0, j_0), \ldots, (i_k, j_k), \ldots\}$ according to transition probabilities Q. Then we may estimate C and d with

$$C_k = \frac{1}{k+1}\sum_{t=0}^{k} \phi_{i_t}\left(\phi'_{i_t} - \frac{\alpha p_{i_t j_t}}{q_{i_t j_t}}\phi'_{j_t}\right), \qquad d_k = \frac{1}{k+1}\sum_{t=0}^{k} g_{i_t}\phi_{i_t}, \tag{5.2}$$

where $p_{i_t j_t}$ are the components of P and $q_{i_t j_t}$ are the transition probabilities of Q. Note the similarity between with Eq. (5.1), Eq. (5.2) and the estimation formula Eq. (4.4) in Section 4.1. Under certain irreducibility condition, it can be shown that $C_k \to C$ and $d_k \to d$ with probability 1 (see [BY07] and [BY09]).

Based on the estimates C_k and d_k, the LSTD method evaluates the approximate

cost vector by

$$\hat{r}_k = C_k^{-1} d_k. \tag{5.3}$$

Since $C_k \to C$ and $d_k \to d$, we have $\hat{r}_k \to C^{-1}d$. So the LSTD iteration is guaranteed to converge to the exact solution r^* of the projected Bellman's equation.

The LSPE (*least squares projected equation*) iteration, proposed in [BI96] (see also [NB03], [BBN04]), takes the form of

$$r_k = r_{k-1} - \gamma D_k^{-1}(C_k r_{k-1} - d_k), \tag{5.4}$$

where γ is a positive scalar and D_k is an estimate of the matrix $D = \Phi' \Xi \Phi$ given by

$$D_k = \frac{1}{k+1} \sum_{t=0}^{k} \phi_{i_t} \phi_{i_t}'. \tag{5.5}$$

The convergence of LSPE has been proved in [BBN04]; and as suggested in [Ber09], this analysis is through for a more general context where D_k is not of the form of Eq. (5.5) , provided that the mapping of iteration (5.4) be asymptotically contractive .

5.2 Regression-Based LSTD

The LSTD method must also contend with the difficulty induced by near-singularity of C (and also C_k) and the simulation errors $C_k - C$ and $d_k - d$. These two factors together may amplify the solution error:

$$\hat{r}_k - r^* = C_k^{-1} d_k - C^{-1} d.$$

As an alternative to the standard LSTD iteration [cf. Eq. (5.3)], we consider the estimation of r^* using the regularized regression approach given in Chapter 2. This yields the *regression-based LSTD* (reg-LSTD) iteration:

$$\hat{r}_k = \left(C_k' \Sigma^{-1} C_k + \Gamma^{-1}\right)^{-1} \left(C_k' \Sigma^{-1} d_k + \Gamma^{-1} \bar{r}\right) \tag{5.6}$$

where Σ and Γ are some positive definite symmetric matrices, and $\bar{r}$ is an prior guess or nominal value of r^*. For example, one simple choice is to let

$$\bar{r} = 0, \qquad \Sigma^{-1} = I, \qquad \Gamma^{-1} = \beta I,$$

so Eq. (5.6) becomes

$$\hat{r}_k = \left(C_k' C_k + \beta I\right)^{-1} C_k' d_k.$$

Due to the regularization, this regression-based version of LSTD is less susceptible to near-singularity of C and the simulation error of $C_k - C$ and $d_k - d$, as justified in Chapter 2. However, when passing the limit to $k \to \infty$, we have

$$\begin{aligned}\lim_{k\to\infty} \hat{r}_k &= \lim_{k\to\infty} \left(C_k' \Sigma^{-1} C_k + \Gamma^{-1}\right)^{-1} \left(C_k' \Sigma^{-1} d_k + \Gamma^{-1} \bar{r}\right) \\ &= \left(C' \Sigma^{-1} C + \Gamma^{-1}\right)^{-1} \left(C' \Sigma^{-1} d + \Gamma^{-1} \bar{r}\right).\end{aligned}$$

Therefore we have shown that reg-LSTD incurs a constant bias, and converges to a limit point other than r^*.

Let us also mention the limiting case when β decreases to zero. We can show that, as $\beta \to 0$, Eq. (5.6) converges to

$$\hat{r}_k = C_k^\dagger d_k, \tag{5.7}$$

where $\dagger$ denotes the Moore-Penrose pseudoinverse; [1] see Prop. 8 in Appendix A.2. Hence Eq. (5.7) can be viewed as a special form of reg-LSTD, in which the pseudoinverse solution $\hat{r}_k$ happens to be the minimal norm solution of $C_k r = d_k$. Moreover, Eq. (5.7) implies the potential possibility of extending the traditional LSTD [cf. Eq. (5.3)] to the case where Φ may have linearly dependent columns. It can be shown that this version of LSTD still converges to a solution of the projected Bellman's equation. A stronger conclusion is that, the high-dimensional behavior of iteration (5.7) is invariant under different representations of the subspace S; see Prop. 9 in Appendix A.3.

[1] The MoorePenrose pseudoinverse $A^\dagger$ of an $m \times n$ real matrix A is defined as the unique $n \times m$ matrix satisfying $AA^\dagger A = A, A^\dagger A A^\dagger = A^\dagger, (AA^\dagger)' = AA^+$, and $(A^\dagger A)' = A^\dagger A$; [Moo20], [Pen55].

5.3 Iterative Regression-Based LSTD

Based on the regression formula Eq. (5.6), a particular approach of interest is to let

$$\Sigma = \Sigma_k, \qquad \bar{r} = \hat{r}_{k-1},$$

where Σ_k is an estimate of the sample covariance of $C_k r^* - d_k$ and the prior estimate $\bar{r}$ is taken to be the latest iterate. In this case, Eq. (5.6) becomes

$$\hat{r}_k = \left(C_k' \Sigma_k^{-1} C_k + \Gamma^{-1}\right)^{-1} \left(C_k' \Sigma_k^{-1} d_k + \Gamma^{-1} \hat{r}_{k-1}\right), \tag{5.8}$$

or

$$\hat{r}_k = \hat{r}_{k-1} - \left(C_k' \Sigma_k^{-1} C_k + \Gamma^{-1}\right)^{-1} C_k' \Sigma_k^{-1} \left(C_k \hat{r}_{k-1} - d_k\right). \tag{5.9}$$

This type of reg-LSTD becomes an iterative algorithm. We will now establish its convergence.

We note the similarity between iteration (5.9) and LSPE [cf. Eq. (5.4)]. This similarity may imply a proof of convergence for this iterative reg-LSTD that is similar to the proof of convergence for LSPE. We may view Eq. (5.8) as a *scaled LSPE* iteration:

$$\hat{r}_k = \hat{r}_{k-1} - D_k^{-1}(C_k \hat{r}_{k-1} - d_k), \tag{5.10}$$

where D_k^{-1} is the matrix

$$D_k^{-1} = \left(C_k' \Sigma_k^{-1} C_k + \Gamma^{-1}\right)^{-1} C_k' \Sigma_k^{-1}.$$

As suggested in [Ber09], the line of analysis of the convergence proof of LSPE is through in a more general context, where D_k is not necessarily symmetric. In particular, it can be shown that, the scaled LSPE converges asymptotically to r^* with probability 1 if the matrix $I - D_k^{-1} C_k$ eventually becomes contractive.

Now let us show that the mapping of iteration (5.9) will become strictly contractive eventually. First we give the following lemma:

Lemma 1 *Any $s \times s$ positive definite symmetric matrices A and B satisfy that*

$$\|(A+B)^{-1}B\| < 1.$$

Proof. Since A is positive definite, there exists an $s \times s$ invertible matrix C such that $A = C'C$. We have

$$(A+B)^{-1}B = (C'C+B)^{-1}B = C^{-1}\left(I + (C^{-1})'BC^{-1}\right)^{-1}(C^{-1})'BC^{-1}C.$$

Since B is positive definite and C is invertible, the matrix $(C^{-1})'BC^{-1}$ is also positive definite and symmetric. So we may write its eigenvalue decomposition as $(C^{-1})'BC^{-1} = V'\Lambda V$, where V is an unitary matrix, $\Lambda = \text{diag}\{\lambda_1, \ldots, \lambda_s\}$ and $\lambda_1, \ldots, \lambda_s$ are the positive eigenvalues.

We have

$$\begin{aligned}
(A+B)^{-1}B &= C^{-1}(I + (C^{-1})'BC^{-1})^{-1}(C^{-1})'BC^{-1}C \\
&= C^{-1}(I + V'\Lambda V)^{-1}V'\Lambda VC \\
&= C^{-1}(V'V + V'\Lambda V)^{-1}V'\Lambda VC \\
&= C^{-1}V'(I+\Lambda)^{-1}\Lambda VC \\
&= (VC)^{-1}\text{diag}\left\{\frac{\lambda_1}{1+\lambda_1}, \ldots, \frac{\lambda_s}{1+\lambda_s}\right\}VC.
\end{aligned}$$

Hence we have obtained the eigenvalue decomposition of $(A+B)^{-1}B$. Noting the fact that $\frac{\lambda_i}{1+\lambda_i} < 1$ for all λ_i, we obtain that eigenvalues of $(A+B)^{-1}B$ are strictly contained in the unit circle. This implies that $\|(A+B)^{-1}B\| < 1$. ■

Now we are ready to establish the convergence results for iteration (5.9):

Proposition 6 *Assuming that $C_k \to C$, $d_k \to d$ and $\Sigma_k \to \Sigma$, where Σ is a positive definite matrix, the iterative reg-LSTD iteration* (5.9) *converges to $r^* = C^{-1}d$ almost surely.*

Proof. Consider an arbitrary sample path, along which $C_k \to C$, $d_k \to d$ and $\Sigma_k \to \Sigma$.

We have

$$I - D_k^{-1}C_k \to \left(C'\Sigma^{-1}C + \Gamma^{-1}\right)^{-1}\Gamma^{-1}$$

where $C'\Sigma^{-1}C$ and Γ^{-1} are positive definite and symmetric matrices. Now we may use a similar line of analysis as used in [YB06] and [BBN04], i.e., the convergence of iteration (5.9) hinges on its deterministic portion.

Using Lemma 1 we have

$$\| \left(C'\Sigma^{-1}C + \Gamma^{-1}\right)^{-1}\Gamma^{-1}\| < 1.$$

Since $\| \cdot \|$ is a continuous mapping, there exists $\epsilon > 0$ such that

$$\|I - D_k^{-1}C_k\| \leq \| \left(C'\Sigma^{-1}C + \Gamma^{-1}\right)^{-1}\Gamma^{-1}\| + \epsilon < 1,$$

for all k sufficiently large. This implies that (5.9) eventually becomes strictly contractive, so that $\hat{r}_k$ will converge to the fixed point r^* of the limiting mapping. Noting that the set of all convergent sample paths has a probability measure equal to 1, therefore we obtain that $\hat{r}_k \to r^*$ with probability 1. ■

Here we have proposed to use Σ_k, which is equal to an estimate of the sample covariance of $C_k r^* - d_k$, following the idea of Section 2.1. For estimation of this sample covariance, one possibility is to calculate Σ_k to be the sample covariance of

$$\{\hat{C}_1\tilde{r} - \hat{d}_1, \ldots, \hat{C}_k\tilde{r} - \hat{d}_k\},$$

where $\tilde{r}$ is a prior guess or nominal value of r^*, and $\hat{C}_k$ and $\hat{d}_k$ are the sample matrices and vectors summed up for calculating C_k and d_k. Using a law of large numbers argument, we can show that this Σ_k converges to a positive definite matrix, i.e. the covariance matrix of the corresponding probability space. Another possibility is to use the sample covariance of

$$\{\hat{C}_1\hat{r}_0 - \hat{d}_1, \ldots, \hat{C}_k\hat{r}_{k-1} - \hat{d}_k\}.$$

In this way, we avoid the step of guessing the value of r^*. For this case, the asymptotic behavior of Σ_k and the convergence of the corresponding reg-LSTD iteration is not yet understood, and is a direction for further research.

5.4 Optimistic Versions

The standard LSTD/LSPE methods focus on evaluating a single policy by generating an infinite number of samples. Now we consider their optimistic versions, in which case the underlying control policy is updated every few iterations. For estimation of C_k and d_k, these optimistic algorithms make use of either only those samples generated by the current control, or all samples with uniform weights, using the formula Eq. (5.1) or Eq. (5.2); see [Ber07] and [BT96].

We consider an alternative way of reusing the historical samples, which is to use discounted weights for historical samples instead of uniform weights. One possibility is to update the policy upon taking each sample and induce a discount factor in estimating C and d, given by

$$C_k = (1-\eta)C_{k-1} + \eta\phi_{i_k}\left(\phi'_{i_k} - \frac{\alpha p_{i_k j_k}}{q_{i_k j_k}}\phi'_{j_k}\right), \qquad d_k = (1-\eta)d_{k-1} + \eta g_{i_k}\phi'_{i_k},$$

where η is a positive scalar such that $0 < \eta \leq 1$.

Another possibility is to generate a number of samples after updating the policy, and estimate C and d given by

$$C_k = (1-\eta)C_{\bar{k}} + \frac{\eta}{k-\bar{k}}\sum_{t=\bar{k}+1}^{k}\phi_{i_t}\left(\phi'_{i_t} - \frac{\alpha p_{i_t j_t}}{q_{i_t j_t}}\phi'_{j_t}\right), \tag{5.11}$$

and

$$d_k = (1-\eta)d_{\bar{k}} + \frac{\eta}{k-\bar{k}}\sum_{t=\bar{k}+1}^{k} g_{i_t}\phi'_{i_t}, \tag{5.12}$$

where $\bar{k}$ corresponds to the latest iteration in which the policy has been updated, and $p_{i_t j_t}$ are components of the transition probability matrix under the current policy.

For these cases, the asymptotic behavior of C_k and d_k is not clear. To ensure the

convergence of these optimistic versions of LSTD and LSPE, we may add a decreasing stepsize, e.g., to modify Eq. (5.9) into

$$\hat{r}_k = \hat{r}_{k-1} - \gamma_k \left(C_k' \Sigma^{-1} C_k + \Gamma^{-1}\right)^{-1} C_k' \Sigma^{-1} \left(C_k \hat{r}_{k-1} - d_k\right), \tag{5.13}$$

where γ_k is a sequence of positive scalars decreasing to zero. Theoretical results on the convergence of these optimistic algorithms are not yet established, and the effect of regression is not fully understood. These are interesting directions that worth both theoretical and experimental investigation.

5.5 Example: Parking Problem

In this section we experiment with a toy Markov decision process to demonstrate the use of regression on LSTD/LSPE and their optimistic versions. For more computational experiments on these methods we refer to the paper by Yu and Bertsekas [YB06], and the books by Bertsekas [Ber07] and by Bertsekas and Tsitsiklis [BT96].

We consider the well-known example of parking problem (see [BT96]), in which a driver is searching for parking space through an area of N consecutive spaces. The driver starts at state 1, and from state i he either parks and returns to state 1 or goes next to state $i+1$. Each space is vacant with probability p, independent of whether other spaces are vacant or not. The driver can only observe whether a space is vacant after arriving at that space. If the driver parks in space i he incurs a cost of $c(i)$. If the driver eventually reaches the space N, he must park and incurs a large cost $c(N)$.

This problem can be formulated as a stochastic shortest path problem, and the Bellman equation can be written as

$$J^*(i) = p \min\{c(i) + \alpha J^*(1),\ \alpha J^*(i+1)\} + (1-p)\alpha J^*(i+1), \qquad i = 1, \ldots, N-1,$$
$$J^*(N) = c(N),$$

where J^* denotes the optimal cost vector and α denotes the discount factor. At an arbitrary state $i < N$, the optimal policy is to park and return to state 1 if space i is

vacant and $c(i) + \alpha J^*(1) < \alpha J^*(i+1)$, or to continue driving otherwise.

We consider the case where

$$p = 0.05, \qquad N = 100, \qquad c(i) = i, \qquad c(N) = 1000.$$

For approximation of J^*, we introduce a quadratic architecture and approximate J^* within the subspace

$$S = \left\{ \Phi r \mid r \in \Re^3 \right\},$$

where Φ is the matrix whose ith row is $\phi_i' = [1 \ i \ i^2]$.

Example 1 (Comparison of LSTD, LSPE and Reg-LSTD) *For comparison of LSTD, LSPE and their regression-based versions, we focus on evaluating a single policy of the parking problem. We generate a trajectory of state transitions* $\{(i_0, i_1), \ldots, (i_k, i_{k+1}), \ldots\}$ *under this policy and calculate* (C_k, d_k) *by using Eq.* (5.1)*. We use the following iterations:*

LSTD: $\hat{r}_k = C_k^\dagger d_k$

LSPE: $r_k = r_{k-1} - D_k^\dagger (C_k r_{k-1} - d_k)$,

Reg-LSTD1: Eq. (5.6) *with* $\Sigma = I$, $\beta = 10^{-2}$, $\bar{r} = \hat{r}_{k-1}$,

Reg-LSTD2: Eq. (5.6) *with* $\Sigma = \Sigma_k$, $\beta = 10$, $\bar{r} = \hat{r}_{k-1}$,

Reg-LSTD3: Eq. (5.6) *with* $\Sigma = I$, $\beta = 10^{-4}$, $\bar{r} = 0$,

Reg-LSTD4: Eq. (5.6) *with* $\Sigma = \Sigma_k$, $\beta = 10^{-1}$, $\bar{r} = 0$,

where D_k *is calculated by using Eq.* (5.5)*. Here* Σ_k *is the sample covariances of*

$$\{(\hat{C}_1 r^* - \hat{d}_1), \ldots, (\hat{C}_k r^* - \hat{d}_k)\},$$

and the values of β *have been tailored to reach a near-optimal performance for each algorithm. The computational results are illustrated in Fig. 5-1.*

A first observation from Fig. 5-1 is that, the regression-based versions of LSTD perform in a more stable manner compared with standard LSTD. Another qualitative observation is that, the iterative versions of reg-LSTD using $\bar{r} = \hat{r}_{k-1}$ (reg-LSTD1/2) yield sequences of updates similar to that of LSPE.

Regarding convergence, the LSTD, LSPE and reg-LSTD with $\bar{r} = r_{k-1}$ (reg-LSTD1/2) converge to one another and eventually converge to r^*, which is consistent with the analysis of [YB06]. On the other hand, the regression-based LSTD with $\bar{r} = 0$ (reg-LSTD3/4) incur a constant bias and converge to some limit other than r^*, which validates our analysis.

Example 2 (More on the Iterative Reg-LSTD) *Continuing with the settings of Example 1, let us focus on the iteration:*

$$\hat{r}_k = \left(C_k'\Sigma_k^{-1}C_k + \beta I\right)^{-1}\left(C_k'\Sigma_k^{-1}d_k + \beta\hat{r}_{k-1}\right)$$

[cf. reg-LSTD2 of Example 1], where Σ_k is the sample covariance of the samples

$$\{(\hat{C}_1 r^* - \hat{d}_1), \ldots, (\hat{C}_k r^* - \hat{d}_k)\}.$$

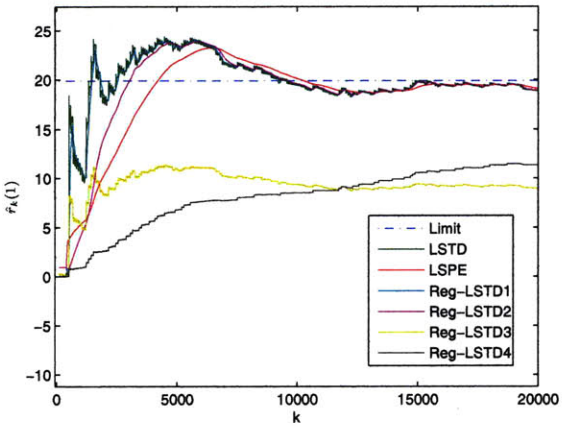
We generate a single trajectory $\{(i_0, i_1), \ldots, (i_k, i_{k+1}), \ldots\}$ and experiment with alternative values of β. We illustrate the corresponding results in Fig. 5-2.

As can be seen in Fig. 5-2, as β decreases to zero, the sequence of updates of iterative reg-LSTD gets closer to the sequence of updates of LSTD, and becomes more susceptible to simulation error. On the other hand, the iterative reg-LSTD converges slowly for large values of β; and may converge more slowly than LSPE does, for β sufficiently large.

We have also experimented this algorithm with Σ_k being the sample covariance of

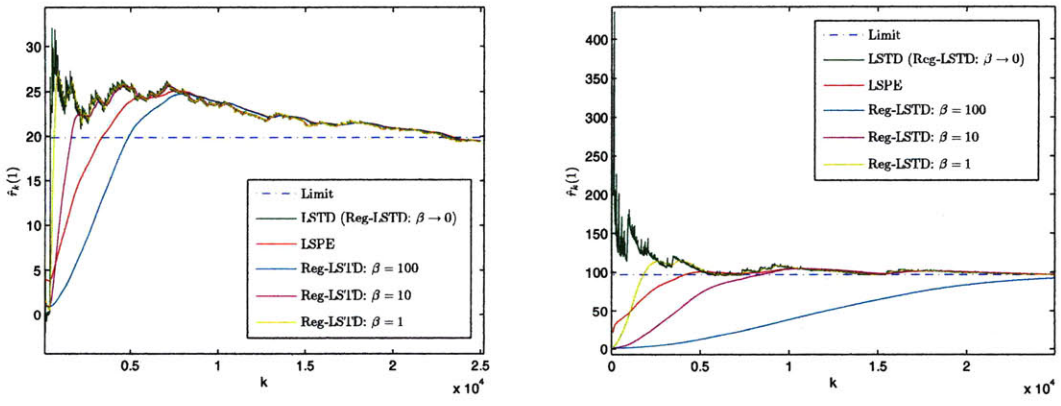
$$\{(\hat{C}_1 \hat{r}_0 - \hat{d}_1), \ldots, (\hat{C}_k \hat{r}_{k-1} - \hat{d}_k)\}.$$

The resulted sequences of updates are similar with those plotted in Fig. 5-2.



$\alpha = 0.95$

Figure 5-1: Sequences of updates of Example 1 based on a single trajectory. Plotted are sequences of one component of the updates of LSTD, LSPE, reg-LSTD1/2/3/4, based on a single trajectory.



$\alpha = 0.95$ $\alpha = 0.99$

Figure 5-2: Sequences of updates of Example 2 based on a single trajectory. The green curve corresponds the standard LSTD iteration, which is also the limiting case when $\beta \to 0$; others curves correspond to iteration (2) with β taking alternative values.

Example 3 (Optimistic Iterative Reg-LSTD using Uniformly Weighted History) *This example tests the optimistic version of iterative reg-LSTD [cf. Eq.* (5.8)*]. We experiment with several values of* β *and generate* 10 *independent trajectories for each* β*. Each trajectory* $\{(i_0, i_1), \ldots, (i_k, i_{k+1}), \ldots\}$ *is generated according to a policy which is updated every* 50 *samples; and based on this trajectory, we calculate* C_k *and* d_k *by using Eq.* (5.1) *and calculate the updates of reg-LSTD by using Eq.* (5.8)*. For comparison, we also test the optimistic versions of LSTD and LSPE. The corresponding results are illustrated in Fig. 5-3.*

We observe that, at one extreme, if β is very small, the reg-LSTD is largely affected by the simulation noise and resembles the LSTD; at the other extreme, if β is too large, the iteration of reg-LSTD has a modulus of contractiveness close to 1 and converges very slowly.

Example 4 (Optimistic LSTD/LSPE/Iterative Reg-LSTD using Discounted History) *In this example we test the optimistic versions of LSTD/LSPE/iterative reg-LSTD using discounted history. We generate 5 independent sample trajectories for each method. For each trajectory, the policy is updated every* $K = 100$ *samples; and* C_k *and* d_k *are calculated as the discounted weighted average of all samples [cf. Eqs.* (5.11)-(5.12)*] with* $\eta = 0.8$*. We illustrate the corresponding results in Fig. 5-4.*

As illustrated in Fig. 5-4, the optimistic LSTD/LSPE, which does not use regulariztion, does not converge under the scheme using discounted history. On the other hand, the reg-LSTD converges well under this scheme. We have also experimented with other values of K, η and β, but have only obtained convergent results in a few cases. Further experiments and analysis is needed before the convergence of these optimistic versions is better understood.

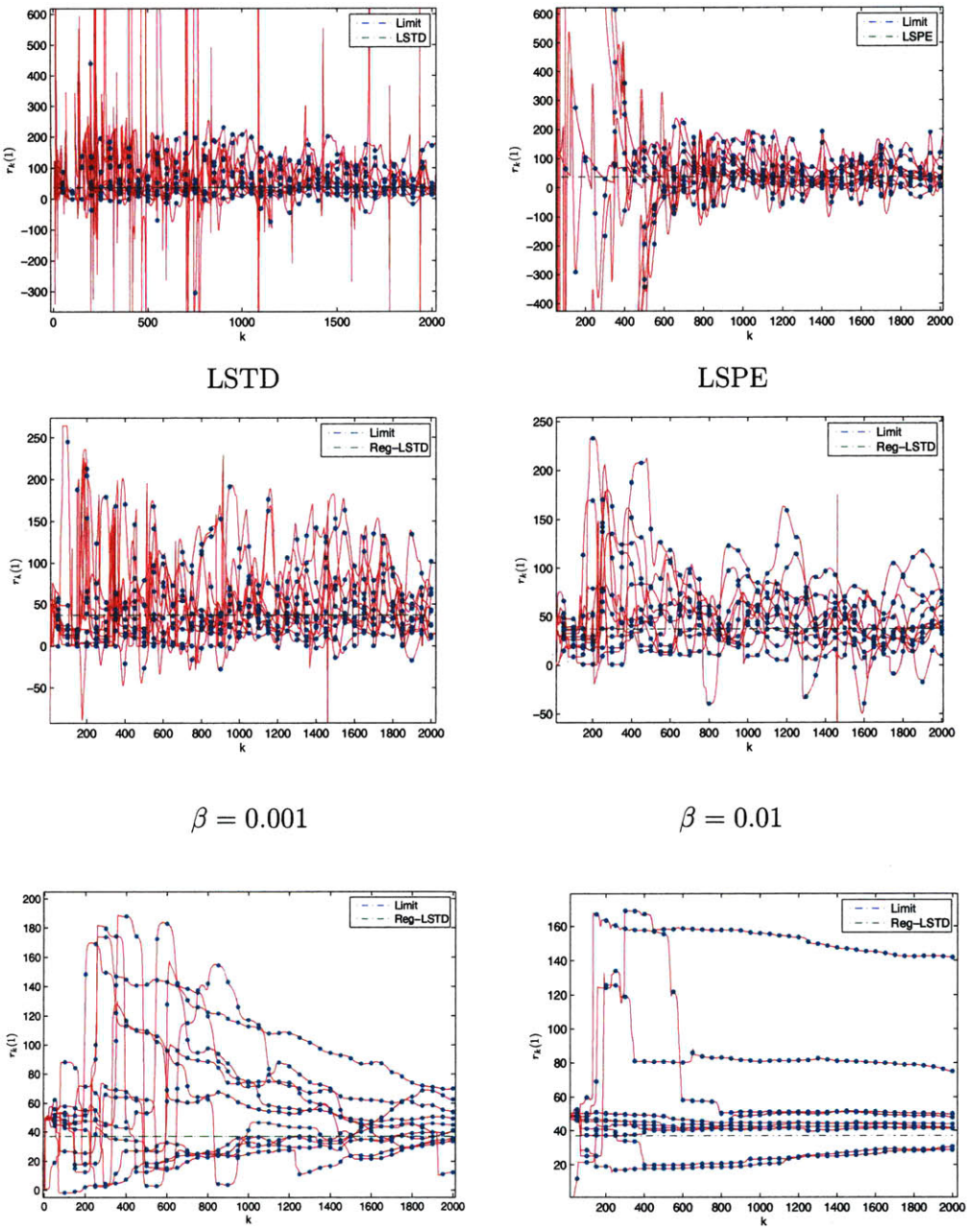


Figure 5-3: Sequences of updates of LSTD/LSPE/iterative reg-LSTD using uniformly weighted history of Example 3, with $K = 50$ and $\alpha = 0.98$. Each figure plots the updates of one component of $\hat{r}_k$ based on 10 independent trajectories. The dots correspond to the iterations of policy updation.

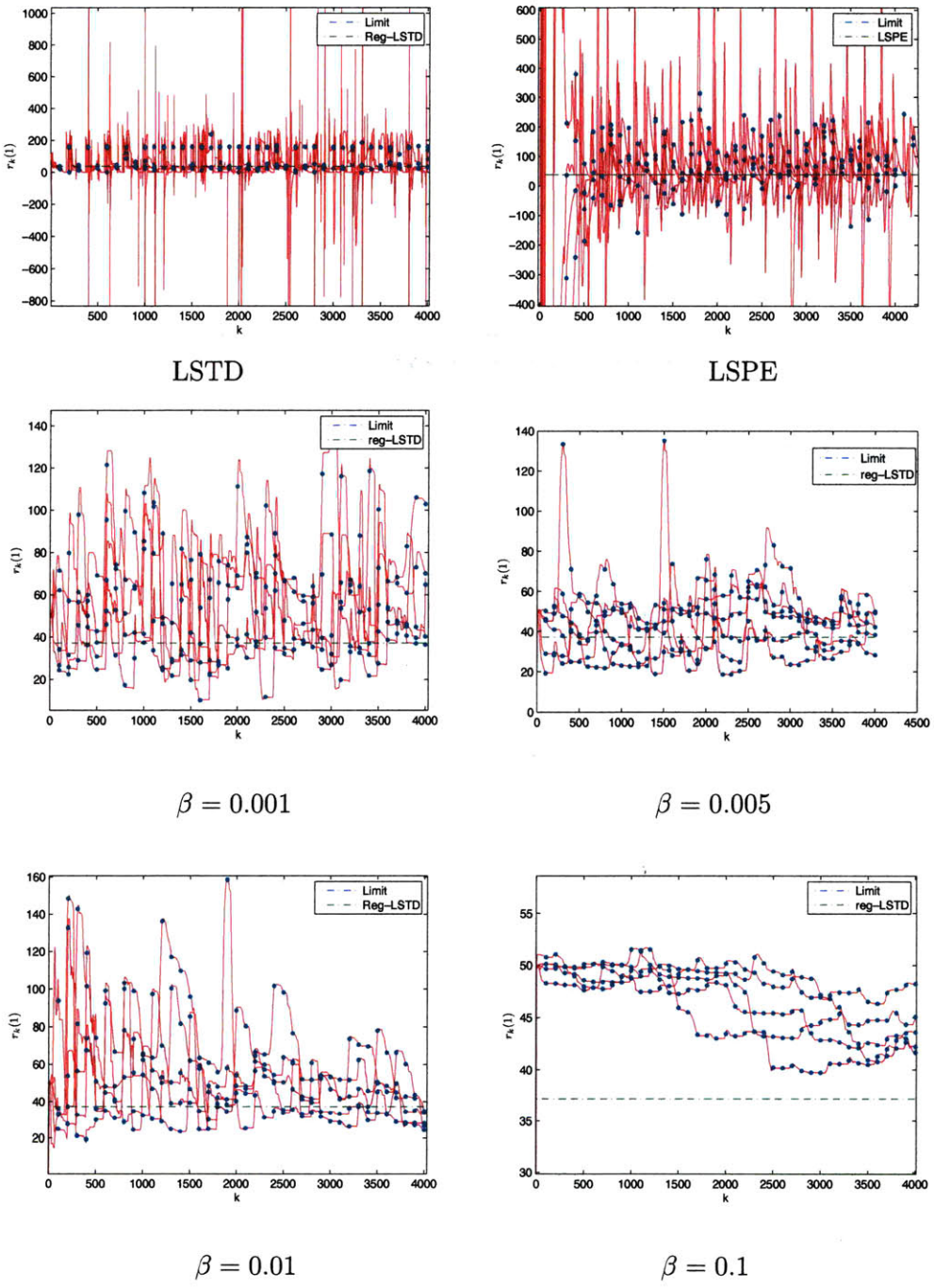


Figure 5-4: Sequences of updates of optimistic LSTD/LSPE/iterative reg-LSTD using discounted history of Example 4, with $K = 100$, $\alpha = 0.98$ and $\eta = 0.8$. Each figure plots the updates of one component of $\hat{r}_k$ based on 5 independent trajectories. The dots correspond to the iterations of policy updation.

Chapter 6

Application to Practical Inverse Problems

In this chapter we apply the proposed algorithmic methodology to a number of practical inverse problems, involving both underdetermined systems (Sections 6.1 and 6.2) and square systems (Sections 6.3-6.6). These problems take the form of Fredholm integral equations of the first kind, and are discretized into linear systems

$$Ax = b$$

[cf. Eq. (1.1)], where A is $m \times n$ or $n \times n$, and n, the dimension of the solution space, is taken to be $n = 10^9$. The matrix A is typically ill-conditioned and dense. The components of A and b are accessible, and can be computed analytically.

We aim for the solution (or the least squares fit solution) x^* of the discretized system $Ax = b$. For square systems, we consider its approximate solution within a subspace spanned by $s = 50$ or $s = 100$ multi-resolution basis functions, which are piecewise constant functions with disjoint local support [KS04]. For underdetermined systems, we use the approach introduced in Section 4.1 and estimate specific components of x^* directly. Note that the computational complexity is completely determined by s (or m for underdetermined systems). Our experiments are run on a dual processor personal computer with 4GB RAM running Matlab. The estimates $\hat{G}$ and $\hat{c}$ (or

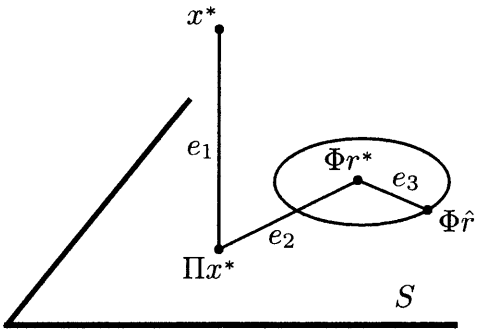


Figure 6-1: Comparison of the projection error e_1, the subspace approximation error e_2 and the simulation plus regularization error e_3, where x^* is the exact solution to $Ax = b$, Πx^* is the projection of x^* on the subspace S, Φr^* is the high-dimensional correspondent of the exact solution to the approximate low-dimensional system [cf. Eq. (1.4)] and $\Phi\hat{r}$ is the high-dimensional correspondent of the approximate solution obtained by the proposed algorithm.

$\hat{F}$ and $\hat{d}$ for underdetermined systems) are obtained component-by-component based on separate sample sequences using Eqs. (1.7)-(1.8) (or Eq. (4.7)). Each sequence is generated by using the importance sampling scheme given in Section 3.3, where we discretize the n-vectors involved (i.e., a_j and ϕ_j) into vectors of dimension 100, and use piecewise linear approximation to compute the sampling distributions. We have estimated each component of G and c (or F and d) with 10^4 samples and each sample takes $50\mu s$ on average.

In presenting the computational results, we will compare three types of error, as illustrated in Fig. 6-1: (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. The performances of our importance sampling schemes are assessed with the total sample covariances of estimated components of $\hat{G}$ and $\hat{c}$ (or $\hat{F}$ and $\hat{d}$ for underdetermined systems).

6.1 The Inverse Contamination Release History Problem

This is an underdetermined problem, whereby we seek to recover the release history of an underground contamination source based on measurements of plume concentration. Let $u(w,\tau)$ be the contaminant concentration at time τ and distance w away from the source, and let $x(\tau)$ be the source release at time τ. The transport of contaminant in the ground is governed by the advection-diffusion model [WU96]

$$\frac{\partial u}{\partial \tau} = D\frac{\partial^2 u}{\partial w^2} - V\frac{\partial u}{\partial w}, \qquad w \geq 0,\ \tau \in [0,\tau_f],$$

subject to Cauchy initial and boundary conditions

$$u(0,\tau) = x(\tau), \quad u(w,0) = 0, \quad \lim_{w\to\infty} u(w,\tau) = 0,$$

where D and V are coefficients for the diffusion and velocity respectively. At a time $T \geq \tau_f$ the plume concentration is distributed as

$$u(w,T) = \int_0^T \mathrm{d}\tau\, \mathbf{A}(w, T-\tau)\, x(\tau),$$

where $\mathbf{A}$ is the transport kernel

$$\mathbf{A}(w,T-\tau) = \frac{w}{\sqrt{4\pi D(T-\tau)^3}} \exp\left\{-\frac{\big(w - V(T-\tau)\big)^2}{4D(T-\tau)}\right\}.$$

In our experiment we take $D = 0.8$, $V = 1$, $T = 300$, and $\tau_f = 250$, and we assume the unknown release history to be

$$x(\tau) = \sum_{i=1}^{5} \kappa_i \exp\left\{-\frac{(\tau-\mu_i)^2}{2\sigma_i^2}\right\},$$

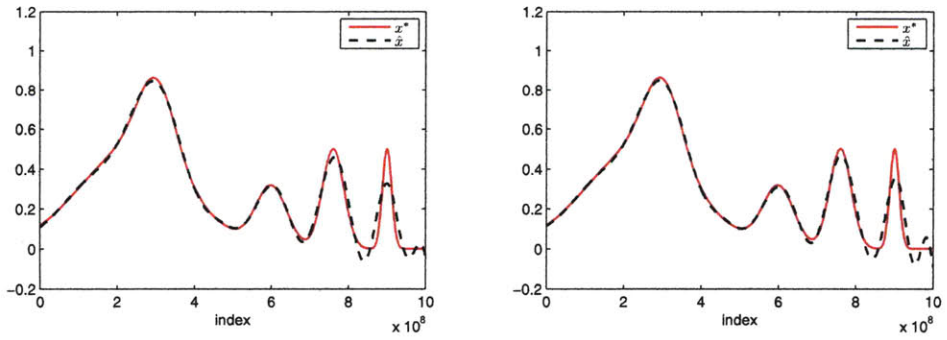


Figure 6-2: The simulation-based approximate solution $\hat{x}$ for the contamination release history reconstruction problem, compared with the exact solution x^*, with $m = 50$ (left) and $m = 100$ (right).

where

$$\kappa = \{0.5, 0.4, 0.3, 0.5, 0.5\},\ \mu = \{60, 75, 150, 190, 225\},\ \sigma = \{35, 12, 10, 7, 3\},$$

and we discretize it into a vector of length 10^9, which is used as the vector x^*. Then we compute m borehole concentration measurements at locations $\{w_i\}_{i=1}^m$, as a discretization of $u(w, T)$ and form the vector b.

In accordance with Section 4.1, we formulate the problem into Eq. (4.6) and estimate F and d using simulation. Then we compute 1000 entries of $\hat{x}$ using the estimates $\hat{F}$ and $\hat{d}$, the regularization matrix $\Gamma^{-1} = 10^{-11}I$ and the initial guess $\bar{r} = 0$. In Fig. 6-2, we compare the resulted entries $\hat{x}_i$ against those of the exact solution x^*.

To analyze the effect of importance sampling, we evaluate the simulation error in terms of the total sample variances for components of $\hat{F}$ and $\hat{d}$. In Fig. 6-3 we compare the reduction of simulation error for alternative importance sampling schemes and alternative ways of function approximation. It can be seen that the proposed importance sampling scheme substantially reduces the simulation error and improves the simulation efficiency. Similar results have been observed in the subsequent problems.

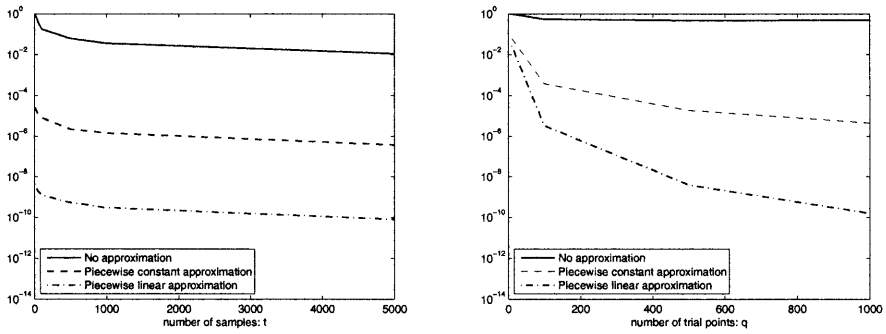


Figure 6-3: The reduction of simulation error for alternative importance sampling schemes. The simulation error is measured in terms of the sum of sample covariances for components of $\hat{F}$ and $\hat{d}$. The solid lines represent the case where no approximation is implemented and a uniform sampling distribution is used; the dotted lines represent the cases where importance sampling is used, with distributions obtained by piecewise constant/linear approximations. The left figure illustrates the reduction of simulation error as the number of samples t varies, while the number of trial points (i.e., the cardinality of I, which is introduced for the purpose of function approximation; see Chapter 3) is fixed at $q = 500$; the right figure plots the results when q varies, with the number of samples fixed at $t = 1000$.

6.2 The Gravitational Prospecting Problem

This is an inverse problem encountered in searching for oil and natural gas resources. We want to estimate the earth density distribution based on measurements of gravitational force at some distance away from the surface.

Here we consider a simplified version of this problem as posed in [Gro07], where the spatial variation of the density is confined within the interior of a ring-shaped domain, and the measurements b are take on a circular trajectory positioned at the same plane but outside the ring. When the unknown density function x and the data are defined on concentric trajectories, we express the problem in polar coordinates as

$$b(\varphi) = \int_0^{2\pi} \mathrm{d}\theta \mathbf{A}(\varphi, \theta) x(\theta), \quad 0 \leq \varphi \leq 2\pi.$$

where

$$\mathbf{A}(\varphi, \theta) = \frac{2 - \cos(\varphi - \theta)}{(5 - 4\cos(\varphi - \theta))^{3/2}}.$$

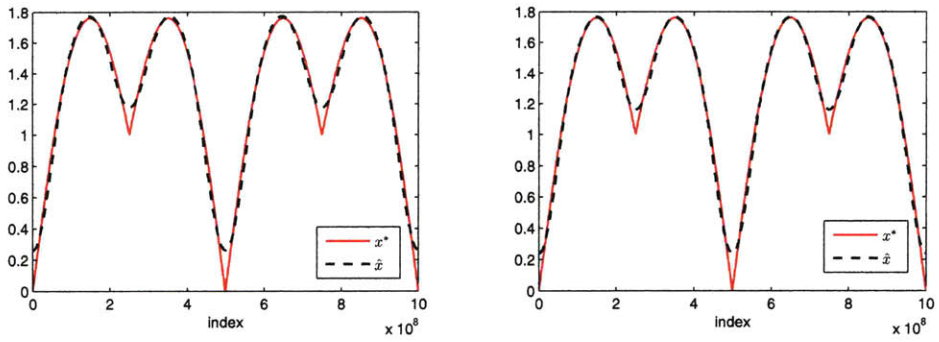


Figure 6-4: The simulation-based approximate solution $\hat{x} = \Phi\hat{r}$ for the gravitational prospecting problem, compared with the exact solution x^*, with $m = 50$ (left) and $m = 100$ (right).

In the experiment, we take the unknown density function to be

$$x(\theta) = |\sin\theta| + |\sin 2\theta|, \quad 0 \leq \theta \leq 2\pi,$$

so the measurement function b can be computed accordingly. We discretize the problem into a system of $m = 50$ and $m = 100$ equations, corresponding to m measurements, with $n = 10^9$ unknowns. For regularization we use $\Gamma^{-1} = 10^{-13}I$ and $\bar{r} = 0$. The approximate solution $\hat{x}$ is illustrated in Fig. 6-4, compared with x^* the exact solution.

6.3 The Second Derivative Problem

This problem refers to differentiating noisy signals that are usually obtained from experimental measurements. This problem has been extensively studied and the solution has been shown to exhibit instability with increasing level of noise [Cul71]. We denote by b the noisy function to be differentiated and denote by x its second derivative. It is given by

$$b(w) = \int_0^1 \mathrm{d}\tau\ \mathbf{A}(w,\tau)x(\tau) \qquad 0 \leq \tau, w \leq 1,$$

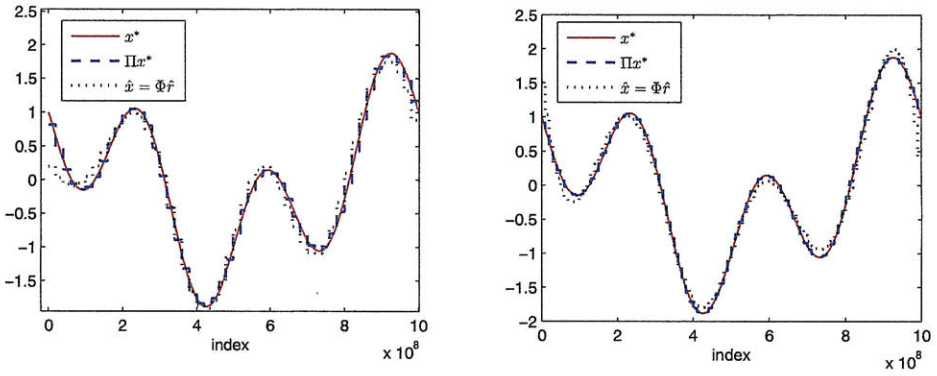


Figure 6-5: The simulation-based approximate solution $\hat{x} = \Phi\hat{r}$ for the second derivative problem, compared with the exact solution x^* and the projected solution Πx^*. The subspace S has dimension $s = 50$ for the left plot and dimension $s = 100$ for the right plot.

where $\mathbf{A}(w, \tau)$ is the Green's function of the second derivative operator

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

In our experiment, we have used

$$x(\tau) = \cos(2\pi\tau) - \sin(6\pi\tau).$$

Following the approach of [Han94] we discretize the integral using the Galerkin method, and obtain a system of n linear equations with n unknowns where $n = 10^9$.

We consider the approximate solution of the system using the preceding methodology, with the initial guess $\bar{r} = 0$ and the regularization matrix $\Gamma = 10^{-5} L_3' L_3$, where L_3 is the $(s-3) \times s$ third-order difference operator. The obtained approximate solution $\Phi\hat{r}$ is presented in Fig. 6-5, and is compared with the exact solution x^* and the projected solution Πx^*.

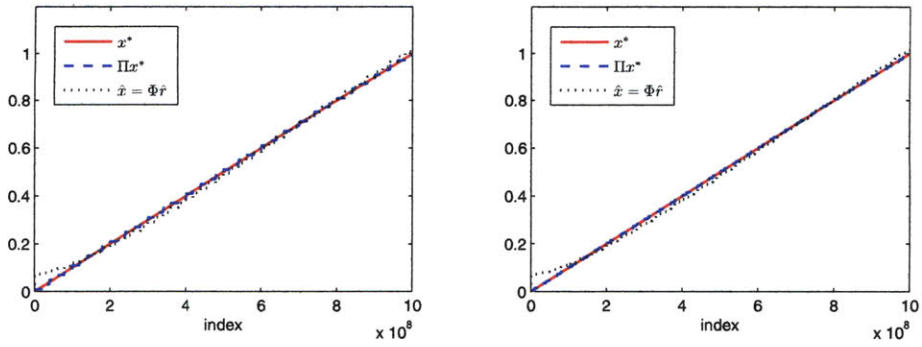


Figure 6-6: The simulation-based approximate solution $\hat{x} = \Phi\hat{r}$ for the Fox-Goodwin problem, compared with the exact solution x^* and the projected solution Πx^*. The subspace S has dimension $s = 50$ for the left plot and dimension $s = 100$ for the right plot.

6.4 The Fox and Goodwin Problem

This problem, introduced by Fox and Goodwin in [FG53], considers the solution of the integral equation

$$b(w) = \int_0^1 \mathrm{d}\tau \sqrt{\mathrm{w}^2 + \tau^2}\, \mathrm{x}(\tau), \quad 0 \leq \mathrm{w} \leq 1.$$

As shown in [FG53], this is a severely ill-posed problem and the condition number of its discretized integral operator increases exponentially with n. In our experiment, we assume the unknown solution to be

$$x(\tau) = \tau, \quad 0 \leq \tau \leq 1,$$

compute b accordingly, and discretize the system into a square linear system of dimension $n = 10^9$.

We consider its approximate solution in the subspace spanned by $s = 50$ or $s = 100$ multi-resolution basis functions, and introduce the regularization matrix $\Gamma^{-1} = 10^{-3}I$ and the initial guess $\bar{r} = 0$. The obtained approximate solution $\Phi\hat{r}$ is presented in Fig. 6-6, plotted against the exact solution x^* and the projected solution Πx^*.

6.5 The Inverse Heat Conduction Problem

This problem seeks to reconstruct the time profile of a heat source by monitoring the temperature at a fixed distance away [Car82]. The one-dimensional heat transfer in a homogeneous quarter plane medium, with known heat conductivity α, is expressed by the elliptic partial differential (heat) equation

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

$$u(w,0) = 0, \quad u(0,\tau) = x(\tau),$$

where $u(w,\tau)$ denotes the temperature at location w and time τ. Let b be the temperature at a fixed location $\bar{w}$ away from the source, and it satisfies

$$b(\tau) = \int_0^T \mathrm{d}\upsilon \mathbf{A}(\upsilon,\tau)\mathrm{x}(\upsilon),$$

where $\mathbf{A}$ is a lower-triangular kernel given by

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

In the experiment we take $T = 1$ and take the unknown target temperature function to be

$$x(\tau) = \sum_{i=1}^{3} \kappa_i \exp\Big\{-\frac{(\tau-\mu_i)^2}{2\sigma_i^2}\Big\}, \quad 0 \leq \upsilon \leq 1,$$

with $\kappa = \{4,3,6\} \times 10^{-4}$, $\mu = \{0.3, 0.6, 0.8\}$ and $\sigma = \{0.1, 0.1, 0.05\}$, so b can be obtained accordingly.

We discretize the integral equation into a linear square system of dimension $n = 10^9$ and consider its approximate solution within the subspace spanned by $s = 50$ or $s = 100$ multi-resolution basis functions. Also we assume an initial guess $\bar{r} = 0$ and the regularization matrix $\Gamma^{-1} = \beta L_1' L_1$, where L_1 is the $(s-1) \times s$ discrete first-order difference operator and $\beta = 10^{-5}$. The computational results are illustrated in Fig. 6-7.

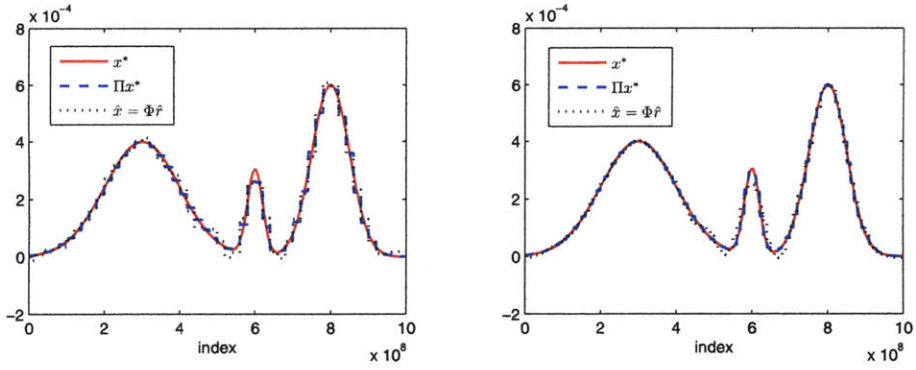


Figure 6-7: The simulation-based approximate solution $\hat{x} = \Phi\hat{r}$ for the inverse heat conduction problem, compared with the exact solution x^* and the projected solution Πx^*. The subspace S has dimension $s = 50$ for the left-hand plot and dimension $s = 100$ for the right-hand plot.

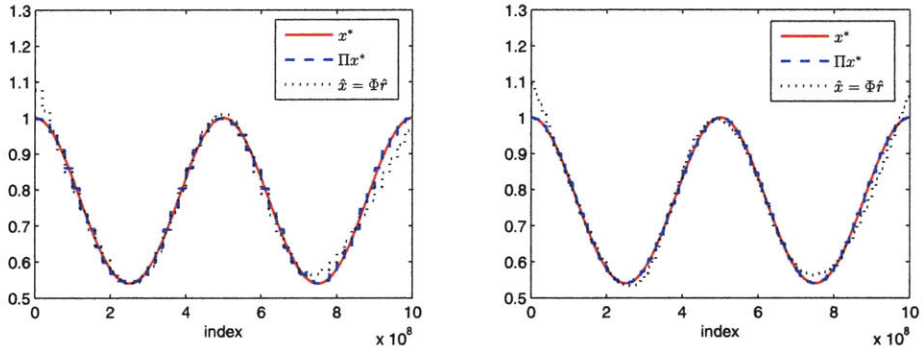


Figure 6-8: The simulation-based approximate solution $\hat{x} = \Phi\hat{r}$ for the optics problem, compared with the exact solution x^* and the projected solution Πx^*. The subspace S has dimension $s = 50$ for the left plot and dimension $s = 100$ for the right plot.

6.6 A Problem in Optical Imaging

Consider light passing through a thin slit, where the intensity of the diffracted light is a function of the outgoing angle and can be measured by some instrument. We wish to reconstruct the light intensity at the incoming side of the slit based on these measurements. Let x be the incoming light intensity as a function of the incoming angle, and let b be the outgoing light intensity as a function of the outgoing angle, so that

$$b(\varphi) = \int_{-\pi/2}^{\pi/2} \mathrm{d}\theta \mathbf{A}(\varphi, \theta)\, \mathbf{x}(\theta), \quad \varphi, \theta \in [-\pi/\mathbf{2}, \pi/\mathbf{2}],$$

where

$$\mathbf{A}(\varphi,\theta) = \Big(\cos\varphi + \cos\theta\Big)^{\mathbf{2}} \left(\frac{\sin(\pi(\sin\varphi + \sin\theta))}{\pi(\sin\varphi + \sin\theta)}\right)^{\mathbf{2}}$$

(we refer to [Jr.72] for further explanation of the physical aspects of this application). We discretize this integral equation into a square system of dimension $n = 10^9$, and consider its approximation within the subspace spanned by $s = 50$ and $s = 100$ multi-resolution functions. The regularization matrix is taken to be $\Gamma^{-1} = \beta L_3' L_3$ and $\bar{r} = 0$, where L_3 is the third-order difference operator and $\beta = 10^{-5}$. The corresponding computational results are plotted in Fig. 6-8.

Chapter 7

Summary and Future Work

In this thesis, we have considered the approximate solution for large-scale linear least squares problems on a subspace spanned by a given set of basis functions or features. We have proposed an approximation-simulation methodology that relies exclusively on low-dimensional calculations, and involves regularized regression, importance sampling and designing near-optimal samplings distributions tailored to the model matrices and basis functions. Our regression approach makes use of the sample covariances obtained as a byproduct of the simulation, and can effectively reduce the effect of near-singularity of the system and simulation error induced by sampling, for which we have provided theoretical justification. This simulation-based regularized regression can also be applied to large-scale problems with equality or inequality constraints, and to approximate policy iteration algorithms in the context of large-scale DP problems.

The performance of our algorithmic methodology has been numerically evaluated on a number of classical inverse problems of dimensions up to 10^9. The computation experiments demonstrate an adequate reduction in simulation noise after a relatively small number of samples and an attendant improvement in quality of the resulted 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 reason-

ably 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.

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 have been given in this thesis and also the related papers [WPB09] and [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 thesis, but the corresponding mathematical analysis will likely be more complex.

Appendix A

Additional Proofs

A.1 A Different Confidence Region of Section 2.2

This is another derivation of an $(1-\theta)$-confidence region for the solution $\hat{r}$ of the regression, as compared with Prop. 1 in Section 2.2. This version of confidence region involves the constant quantities Σ_G, Σ_c, G and c instead of $\hat{G}$. Let us start with the case where there is no regularization, so the exact solution for problem (1.3) is $r^* = G^{-1}c$ and our approximate solution is $\hat{r} = \hat{G}^{-1}\hat{c}$.

Based on the simulation formulas (1.7)-(1.8), the errors $(g - \hat{g})$ [obtained by concatenating columns of $(\hat{G} - G)$] and $(c - \hat{c})$ are multivariate approximately normal random variables, with covariances Σ_G and Σ_c respectively. They will be treated as normal in the following analysis, which is essentially true when the number of samples is large. Accordingly, the term $(g - \hat{g})'\Sigma_G^{-1}(g - \hat{g})$ is a chi-square random variable with s^2 degrees of freedom. Thus we may write the $(1 - \theta/2)$-confidence region for $\hat{g}$ as

$$R_G = \left\{ \hat{g} \in \Re^{s^2} \;\middle|\; (g - \hat{g})'{\Sigma_G}^{-1}(g - \hat{g}) \leq P^{-1}\left(1 - \theta/2; s^2\right) \right\}, \tag{A.1}$$

where $P(\cdot; s^2)$ denotes the regularized Gamma function with s^2 degrees of freedom and $P^{-1}(\cdot; s^2)$ denotes its inverse. Similarly, $(c - \hat{c})'\Sigma_c^{-1}(c - \hat{c})$ is a chi-square random

variable with s degrees of freedom, so the $(1-\theta/2)$-confidence region for $\hat{c}$ is

$$R_c = \left\{ \hat{c} \in \Re^s \;\middle|\; (c-\hat{c})'\Sigma_c^{-1}(c-\hat{c}) \leq P^{-1}(1-\theta/2;s) \right\}, \tag{A.2}$$

where $P(\cdot;s)$ denotes the regularized Gamma function with s degrees of freedom and $P^{-1}(\cdot;s)$ denotes its inverse.

Based on Eqs. (A.1) and (A.2), we can derive an $(1-\theta)-$confidence region for $\hat{r} = \hat{G}^{-1}\hat{c}$ in the following proposition.

Proposition 7 *If*

$$P^{-1}\left(1-\theta/2;s^2\right) > \|\Sigma_G\|\|G^{-1}\|^2 \tag{A.3}$$

then

$$\begin{aligned} \mathbf{Prob}\Big(\|\hat{r}-r^*\| \leq \|G^{-1}\|\|c\|\sqrt{\|\Sigma_G\|P^{-1}(1-\theta/2;s^2)} \\ + \|G^{-1}\|\sqrt{\|\Sigma_c\|P^{-1}(1-\theta/2;s)} + O\left(\|\Sigma_G\| + \sqrt{\|\Sigma_G\|\|\Sigma_c\|}\right)\Big) \geq 1-\theta. \end{aligned} \tag{A.4}$$

Proof. We will use the following matrix norm inequalities, for $s \times s$ symmetric positive definite matrices A, B, and vector $h \in \Re^s$,

$$\|AB\| \leq \|A\|\|B\|, \tag{A.5}$$

$$\|A\| \leq \|A\|_F, \tag{A.6}$$

$$\left\|A^{-1}\right\|^{-1}\|h\|^2 \leq h'Ah, \tag{A.7}$$

where $\|A\|$ denotes the standard Euclidean norm of A (the largest eigenvalue) $\|A\|_F$ denotes the Frobenius norm of the matrix A (the Euclidean norm of the s^2-dimensional vector comprising its components).

Consider any $\hat{G} \in R_G$ and $\hat{c} \in R_c$. Using Eq. (A.7), we have

$$\|G-\hat{G}\|_F^2 = \|g-\hat{g}\|^2 \leq \|\Sigma_G\|(g-\hat{g})'\,{\Sigma_G}^{-1}(g-\hat{g}), \tag{A.8}$$

$$\|c-\hat{c}\|^2 \leq \|\Sigma_c\|\,(c-\hat{c})'\Sigma_c^{-1}(c-\hat{c}). \tag{A.9}$$

Combining Eqs. (A.8)-(A.9) with Eqs. (A.1)-(A.2), and using Eq. (A.6), we obtain

$$\|G - \hat{G}\| \leq \|G - \hat{G}\|_F \leq \sqrt{\|\Sigma_G\| P^{-1}(1 - \theta/2; s^2)}, \tag{A.10}$$

and

$$\|c - \hat{c}\| \leq \sqrt{\|\Sigma_c\| P^{-1}(1 - \theta/2; s)}. \tag{A.11}$$

Then by Eq. (A.5) we have

$$\|G^{-1}(\hat{G} - G)\| \leq \|G^{-1}\| \|G - \hat{G}\|, \tag{A.12}$$

so the smallest singular value of $\left(I + G^{-1}(G - \hat{G})\right)$ satisfies

$$\left\| \left(I + G^{-1}(G - \hat{G})\right)^{-1} \right\|^{-1} \geq 1 - \left\|G^{-1}\right\| \|G - \hat{G}\| > 0, \tag{A.13}$$

where the last inequality follows from Eqs. (A.10) and (A.3).

Using the relation

$$G^{-1} - \hat{G}^{-1} = \left(I + G^{-1}(\hat{G} - G)\right)^{-1} \left(G^{-1}(G - \hat{G})G^{-1}\right), \tag{A.14}$$

we obtain

$$\begin{aligned} \left\|G^{-1} - \hat{G}^{-1}\right\| &\leq \left\| \left(I + G^{-1}(\hat{G} - G)\right)^{-1} \right\| \cdot \left\|G^{-1}(\hat{G} - G)\right\| \\ &\leq \frac{\|G^{-1}\| \left\|\hat{G} - G\right\|}{1 - \|G^{-1}\| \left\|G - \hat{G}\right\|} \\ &= \|G^{-1}\| \left\|G - \hat{G}\right\| + O\left(\left\|G - \hat{G}\right\|^2\right), \end{aligned} \tag{A.15}$$

where the first inequality uses Eq. (A.5), and the second inequality uses relations (A.12)-(A.13). Using Eq. (A.10) we can further bound the right-hand side of Eq. (A.15) by

$$\left\|G^{-1} - \hat{G}^{-1}\right\| \leq \|G^{-1}\| \sqrt{\|\Sigma_G\| P^{-1}(1 - \theta/2; s^2)} + O\left(\|\Sigma_G\|\right).$$

Now we write the difference between $\hat{r}$ and r^* as

$$\begin{aligned}\hat{r} - r^* &= \hat{G}^{-1}\hat{c} - G^{-1}c \\ &= -(G^{-1} - \hat{G}^{-1})c - G^{-1}(c - \hat{c}) + (G^{-1} - \hat{G}^{-1})(c - \hat{c}),\end{aligned} \tag{A.16}$$

whose norm satisfies

$$\begin{aligned}\|\hat{r} - r^*\| &\leq \|G^{-1} - \hat{G}^{-1}\|\|c\| + \|G^{-1}\|\|c - \hat{c}\| + \|G^{-1} - \hat{G}^{-1}\|\|c - \hat{c}\| \\ &\leq \|G^{-1}\|\|c\|\sqrt{\|\Sigma_G\|P^{-1}(1 - \theta/2; s^2)} \\ &\quad + \|G^{-1}\|\sqrt{\|\Sigma_c\|P^{-1}(1 - \theta/2; s)} + O\left(\|\Sigma_G\| + \sqrt{\|\Sigma_G\|\|\Sigma_c\|}\right)\end{aligned} \tag{A.17}$$

Since $\hat{G}$ and $\hat{c}$ are arbitrary, the relation (A.17) is satisfied for all $(\hat{G}, \hat{c})$ in $R_G \times R_c$, whose probability satisfies

$$\mathbf{P}(R_G \times R_c) = \mathbf{P}(R_G)\,\mathbf{P}(R_c) = 1 - \theta + O(\theta^2),$$

since $\hat{G}$ and $\hat{c}$ are independent. Therefore Eq. (A.17) is satisfied with probability greater than $(1 - \theta)$, which proves the desired result. ■

Prop. 7 gives an approximate $(1 - \theta)$-confidence region for $\hat{r}$ given by

$$\begin{aligned}\Big\{\hat{r} \in \Re^s \;\Big|\; &\|r^* - \hat{r}\| \leq \|G^{-1}\|\|c\|\sqrt{\|\Sigma_G\|P^{-1}(1 - \theta/2; s^2)} \\ &+ \|G^{-1}\|\sqrt{\|\Sigma_c\|P^{-1}(1 - \theta/2; s)}\Big\}\end{aligned} \tag{A.18}$$

where the higher-order term has been omitted. This is consistent with the conclusion of Prop. 1 that the level of confidence of $\hat{r}$ is adversely affected by a large value of $\|G^{-1}\|$, which corresponds to a near-singular matrix G. We also note that as expected, by reducing Σ_G and Σ_c through extra simulation, the confidence region can be made arbitrarily small.

Moreover, the condition (A.3) of Prop. 7 indicates the amount of sampling efforts needed for $\hat{r}$ to bear a certain level of confidence. In other words, if the number of samples is not large enough to satisfy (A.3), the regression approach would not yield any reliable estimate. According to the condition (A.3), we also note that, the

number of samples required becomes substantial if G is close to singular.

Let us now consider the case where we use regularization, i.e., $\Gamma^{-1} \neq 0$. Then assuming that the inverses below exist, the approximate solution $\hat{r}$ of Eq. (2.1) can be rewritten as

$$\hat{r} = \left(\hat{G} + \Sigma\hat{G}'^{-1}\Gamma^{-1}\right)^{-1}\left(\hat{c} + \Sigma\hat{G}'^{-1}\Gamma^{-1}\bar{r}\right), \tag{A.19}$$

Now an approximate $(1-\theta)$-confidence region for $\hat{r}$ is

$$\begin{aligned}\Bigg\{\hat{r} \in \Re^s \;\Big|\; \|r^* - \hat{r}\| \leq \|(G+\tilde{\Sigma})^{-1}\|\|c\|\sqrt{\|\Sigma_G\|P^{-1}(1-\theta/2;s^2)} \\ + \|(G+\tilde{\Sigma})^{-1}\|\sqrt{\|\Sigma_c\|P^{-1}(1-\theta/2;s)}\Bigg\},\end{aligned} \tag{A.20}$$

where $\tilde{\Sigma} = \Sigma\hat{G}'^{-1}\Gamma^{-1}$.

A.2 Connection Between Reg-LSTD and Pseudoinverse-Based LSTD

Proposition 8 *Let $\hat{C}$ be an arbitrary $s \times s$ matrix, Σ be an $s \times s$ positive definite symmetric matrix and $\hat{d}$ be a vector in $\Re^s$, then*

$$\lim_{\beta \to 0}\left(\hat{C}'\Sigma^{-1}\hat{C} + \beta I\right)^{-1}\left(\hat{C}'\Sigma^{-1}\hat{d} + \beta\bar{r}\right) = \hat{C}^{\dagger}\hat{d},$$

Proof. Since Σ is positive definite and symmetric, there exists an invertible matrix A such that $\Sigma^{-1} = A'A$. We may use the singular value decomposition and write $A\hat{C}$ as $A\hat{C} = U'\Lambda V$, where U and V are $s \times s$ unitary basis matrices and Λ takes the form of

$$\Lambda = \text{diag}\{\lambda_1, \ldots, \lambda_z, 0, \ldots, 0\},$$

where z is the rank of $A\hat{C}$ and $\lambda_1, \ldots, \lambda_z$ are the positive singular values of $A\hat{C}$.

We derive

$$
\begin{aligned}
&\left(\hat{C}'\Sigma^{-1}\hat{C}+\beta I\right)^{-1}\left(\hat{C}'\Sigma^{-1}\hat{d}\right)\\
&=\left(V'\Lambda U'U\Lambda V+\beta I\right)^{-1}\left(V'\Lambda UA\hat{d}+V'V\beta\bar{r}\right)\\
&=\left(V'(\Lambda^2+\beta I)V\right)^{-1}V'\left(\Lambda UA\hat{d}+V\beta\bar{r}\right)\\
&=V'(\Lambda^2+\beta I)^{-1}\Lambda UA\hat{d}+V'\beta(\Lambda^2+\beta I)^{-1}V\bar{r}\\
&=V'\begin{bmatrix}
\frac{\lambda_1}{\lambda_1^2+\beta} & 0 & 0 & \dots & 0\\
0 & \dots & 0 & \dots & 0\\
0 & \dots & \frac{\lambda_z}{\lambda_z^2+\beta} & \dots & 0\\
\dots & \dots & \dots & 0 & 0\\
0 & \dots & \dots\dots & \dots & 0
\end{bmatrix}UA\hat{d}+V'\begin{bmatrix}
\frac{\beta}{\lambda_1^2+\beta} & 0 & 0 & \dots & 0\\
0 & \dots & 0 & \dots & 0\\
0 & \dots & \frac{\beta}{\lambda_z^2+\beta} & \dots & 0\\
\dots & \dots & \dots & 0 & 0\\
0 & \dots & \dots\dots & \dots & 0
\end{bmatrix}V\bar{r}.
\end{aligned}
\tag{A.21}
$$

Note that

$$
\lim_{\beta\to 0}\frac{\lambda_i}{\lambda_i^2+\beta}=\frac{1}{\lambda_i},\qquad \lim_{\beta\to 0}\frac{\beta}{\lambda_i^2+\beta}=0,\qquad i=1,\dots,z.
$$

Taking $\beta\to 0$ on both sides of Eq. (A.21), we obtain

$$
\lim_{\beta\to 0}\left(\hat{C}'\Sigma^{-1}\hat{C}+\beta I\right)^{-1}\left(\hat{C}'\Sigma^{-1}\hat{d}+\beta\bar{r}\right)=V'\begin{bmatrix}
\frac{1}{\lambda_1} & 0 & 0 & \dots & 0\\
0 & \dots & 0 & \dots & 0\\
0 & \dots & \frac{1}{\lambda_z} & \dots & 0\\
\dots & \dots & \dots & 0 & 0\\
0 & \dots & \dots\dots & \dots & 0
\end{bmatrix}UA\hat{d}=(A\hat{C})^{\dagger}A\hat{d},
$$

where the last equality uses the property of pseudoinverse. Finally we have $(AC)^{\dagger}A\hat{d}=C^{\dagger}A^{\dagger}A\hat{d}=\hat{C}^{\dagger}\hat{d}$. ■

A.3 Feature-Scaling Free Property of Pseudoinverse-Based LSTD

We are interested in the behavior of the pseudoinverse-based LSTD method [cf. Eq. (5.7)] when the same feature space S is represented by alternative sets of basis functions. Consider two sets of basis functions of S, columns of the $n \times s$ matrix Φ and columns of the $n \times \tilde{s}$ matrix Ψ, which satisfy

$$S = \mathrm{Ra}(\Phi) = \mathrm{Ra}(\Psi).$$

Let $\{C_{k,\Phi}, d_{k,\Phi}\}$ be the sequence generated by Eq. (5.1) (or Eq. (5.2)) and $\{C_{k,\Psi}, d_{k,\Psi}\}$ be the corresponding sequence generated when Φ is replaced with Ψ.

We denote by $\{\hat{r}_{\Phi,k}\}$ the iterates generated by the LSTD-type iteration (5.7) using $\{C_{k,\Phi}, d_{k,\Phi}\}$; and we denote by $\{\hat{r}_{\Psi,k}\}$ the corresponding sequence generated by the iteration using $\{C_{k,\Psi}, d_{k,\Psi}\}$ instead.

Proposition 9 *The high-dimensional sequence obtained from the pseudoinverse-based LSTD iteration* (5.7) *and the simulation process of Eq.* (5.1) *(or Eq.* (5.2)*), is feature-scaling free, i.e.,*

$$\Phi\hat{r}_{k,\Phi} = \Psi\hat{r}_{k,\Psi}.$$

Proof. Assuming without loss of generality that $s \geq \tilde{s}$, we first restrain the analysis to the case when Ψ has a full rank equaling $\tilde{s}$. In this case, $C_{k,\Psi}$ is invertible for k sufficiently large (which we assumes implicitly). Moreover, since Φ and Ψ have the same range, there exists some $\tilde{s} \times s$ matrix B such that $\Phi = \Psi B$ and B has full rank $\tilde{s}$. So the $\tilde{s} \times \tilde{s}$ matrix BB' is positive definite and thus invertible.

Let $\psi(i)'$ denote the rows of Ψ, so we have $\phi(i)' = \psi(i)'B$ for all i. Comparing Eq. (5.1) (or Eq. (5.2)) and using this relation, we have, for all k, that

$$C_{k,\Phi} = B'C_{k,\Psi}B, \quad d_{k,\Phi} = B'd_{k,\Psi}. \tag{A.22}$$

By definition for the Moore-Penrose pseudoinverse, we write for all k that,

$$C_{k,\Phi}C_{k,\Phi}^{\dagger}C_{k,\Phi} = C_{k,\Phi}. \tag{A.23}$$

Using Eq. (A.22) to replace $C_{k,\Phi}$ in Eq. (A.23), we obtain

$$(B'C_{k,\Psi}B)C_{k,\Phi}^{\dagger}(B'C_{k,\Psi}B) = B'C_{k,\Psi}B. \tag{A.24}$$

Multiplying Eq. (A.24) with B on the left side, and with B' on the right side, we have

$$B(B'C_{k,\Psi}B)C_{k,\Phi}^{\dagger}(B'C_{k,\Psi}B)B' = BB'C_{k,\Psi}BB',$$

and noting that BB' and $C_{k,\Psi}$ are invertible, we can cross-off them on both sides and obtain

$$BC_{k,\Phi}^{\dagger}B' = C_{k,\Psi}^{-1}. \tag{A.25}$$

Also note that $d_{k,\Phi} = B'd_{k,\Psi}$, then Eq. (A.25) implies

$$BC_{k,\Phi}^{\dagger}d_{k,\Phi} = C_{k,\Psi}^{-1}d_{k,\Psi}. \tag{A.26}$$

Finally, we multiply both sides of Eq. (A.26) with Ψ, appeal to the fact $\Phi = \Psi B$, and obtain

$$\Phi C_{k,\Phi}^{\dagger}d_{k,\Phi} = \Psi C_{k,\Psi}^{-1}d_{k,\Psi}.$$

Now consider the general case when neither Φ nor Ψ is necessarily full rank. Let Θ be the $n \times \hat{s}$ matrix consisted of independent column vectors of Ψ, where we denote by $\hat{s}$ the common rank of Φ and Ψ. Consider the sequence $\{C_{k,\Theta}, d_{k,\Theta}\}$ generated by Eq. (5.1) (or Eq. (5.2)) using Θ as the basis matrix. Note here Θ has the same range as that of Φ and Ψ, and Θ is full rank. Hence we can apply results of the previous case and obtain

$$\Psi C_{k,\Phi}^{\dagger}d_{k,\Psi} = \Theta C_{k,\Theta}^{-1}d_{k,\Theta} = \Phi C_{k,\Phi}^{\dagger}d_{k,\Phi},$$

which is the desired equation. ■

Appendix B

Additional Numerical Experiments

B.1 Test Problems of Section 2.1

We test the regularized regression of Section 2.1 in the following examples.

Example 5 (One-time regression using sample covariance) *Let G be a 2×2 positive diagonal matrix, r^* and c be vectors in $\Re^2$, given by*

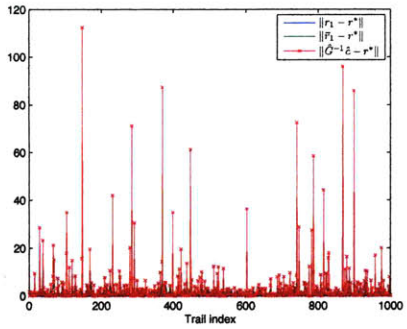
$$G = \begin{bmatrix} \sigma_1 & 0 \\ 0 & 1 \end{bmatrix}, \quad r^* = \begin{bmatrix} 1 \\ \sigma_2 \end{bmatrix}, \quad c = \begin{bmatrix} \sigma_1 \\ \sigma_2 \end{bmatrix},$$

for scalars σ_1 and σ_2. Let the simulation errors $(\hat{G}-G)$ and $(\hat{c}-c)$ be normal random variables, which have means at zero and covariances given by

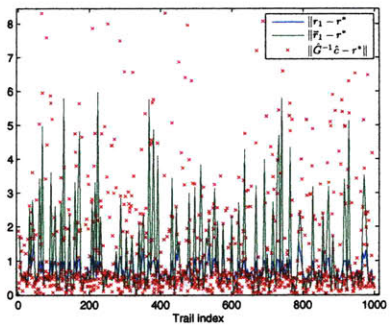
$$\Sigma_G = \begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_3 \end{bmatrix}, \quad \Sigma_c = \begin{bmatrix} 0.01 & 0 \\ 0 & 0.01 \end{bmatrix},$$

for some $\sigma_3 > 0$. The regularization matrix Γ^{-1} [cf. Eq. (2.1)] is given by $\Gamma^{-1} = \beta I$ for some $\beta > 0$, and the prior guess $\bar{r}$ is taken at the origin.

The initial estimate r_0 is taken to be $r_0 = \hat{G}^{-1}\hat{c}$. Then an approximate solution, denoted by r_1, is calculated by iterating Eq. (2.4) only once, starting from r_0. For

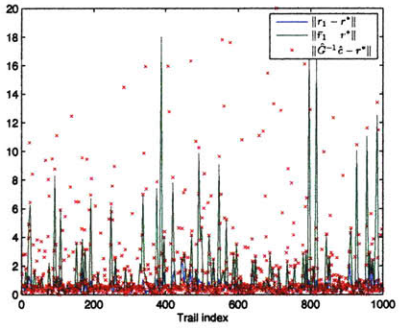


(a) $\beta = 0.01$

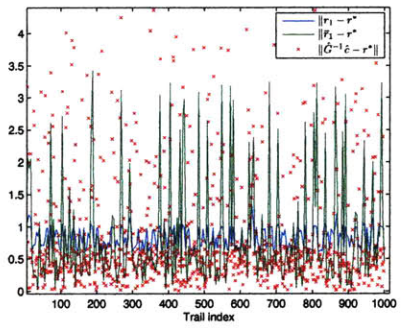


(b) $\beta = 0.01$ (zoomed- in version)

Figure B-1: Illustration of sample errors of Example 5. Figure (a) contains sample errors; and Figure (b) is a zoomed-in version where outlying points are omitted.

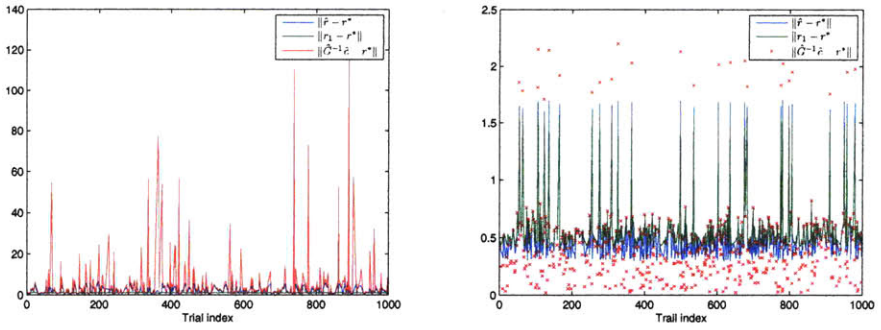


(a) $\beta = 0.005$



(b) $\beta = 0.05$

Figure B-2: Sample errors of Example 5 for alternative values of β. Figures (a) and (b) present the sample errors for $\beta = 0.005$ and $\beta = 0.05$ respectively, where unspecified parameters are identical with those of Fig. B-1. Both (a) and (b) have been zoomed such that outlying points are omitted. A comparison of (a) and (b) reveals the tradeoff between the bias and the variance induced by regularization.



(a) (b)

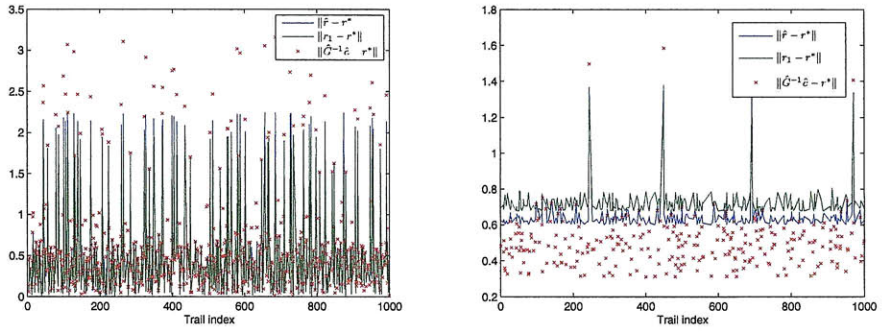
Figure B-3: Sample errors of Example 2 with $\beta = 0.001$. Figure (a) presents the sample errors of all 1000 trials, and figure (b) presents the errors of those trials in which the iterative regression converges.

comparison, a third approximate solution, denoted by $\bar{r}_1$, is calculated using Eq. (2.1) where Σ is taken to be the constant matrix

$$\bar{\Sigma} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}.$$

For given values of parameters $(\sigma_1 = 1, \sigma_2 = 10, \sigma_3 = 1, \beta = 0.01)$, the regression method is tested on 1000 independent trial pairs $(\hat{G}, \hat{c})$ generated according to the given normal distributions. Figure B-1 plots the errors $\|r_1 - r^*\|$, $\|\bar{r}_1 - r^*\|$ and $\|r_0 - r^*\|$ for all 1000 trials, and Figure B-2 compares the results when alternative values of β are used. Note that the plotted errors have been re-scaled by a factor $\|r^*\|^{-1}$ in all numerical results.

Example 6 (Iterative regression using sample covariance) *Under the same settings of Example 5, the iterative regression method is tested against the regression method that iterates once. For given values of σ_1, σ_2, σ_3 and β, 1000 independent trial pairs $(\hat{G}, \hat{c})$ are generated. For each $(\hat{G}, \hat{c})$, Eq. (2.4) is iterated with the initial estimate $r_0 = \hat{G}^{-1}\hat{c}$, yielding r_1 after the first iteration, and yielding the fixed point approximate solution $\hat{r}$ if the corresponding iterates converge.*



(a) $\beta = 0.0001$ (b) $\beta = 0.01$

Figure B-4: Sample errors of converged trials for Example 2, with $\beta = 0.0001$ in (a) and $\beta = 0.01$ in (b), while other parameters are identical as those in Fig. B-3.

Fig. B-3 plots the sample errors of those trials in which the iteration converges, compared with the samples errors of all 1000 trials. Comparing Figs. B-3(a) and B-3(b), we see that the errors of "good" trials, in which the iteration converges, have far smaller variance than general errors. This implies that the convergence of iterative regression is a indicator of the quality of $(\hat{G}, \hat{c})$ thus the confidence of r_0. For given parameters $(\sigma_1 = 1, \sigma_2 = 10, \sigma_3 = 1, \beta = 0.01)$, Figure B-4 plots the sample errors of convergent trials when β varies.

B.2 Validation of Proposition 1

Example 7 (Validation of Confidence Region) *Let G be a 2×2 positive diagonal matrix, r^* and c be vectors in $\Re^2$, given by*

$$G = \begin{bmatrix} 100 & 0 \\ 0 & \sigma \end{bmatrix}, \quad r^* = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad c = \begin{bmatrix} 100 \\ \sigma \end{bmatrix},$$

where σ is a positive scalar. Let the simulation errors $(\hat{G} - G)$ and $(\hat{c} - c)$ be normal random variables, which have means at zero and covariances given by

$$\Sigma_G = \begin{bmatrix} 1000 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & \sigma^2 \end{bmatrix}, \quad \Sigma_c = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

Consider the estimate solution $\hat{r}$ obtained by using (2.1) with

$$\Gamma^{-1} = 0.5I, \quad \bar{r} = r^*,$$

and $\Sigma(r)$ is estimated with $\Sigma(\hat{G}^{-1}\hat{c})$.

We compares the $(1-\theta)$ error bounds $[\sigma(\Sigma(\hat{G}^{-1}\hat{c}), \gamma)]$ with the actual sample errors $\|\hat{r}_1 - r^*\|$ for $(1-\theta) = 0.75, 0.95$ and $\sigma = 1, 10, 100$. For each setting of parameters, 200 independent sample trials $(\hat{G}, \hat{c})$ are generated. The corresponding sample errors and estimated error bounds are presented in the following table:

	$\sigma = 1$	$\sigma = 10$	$\sigma = 100$
Violation rate of 75% bound	11%	16.50%	17.50%
Violation rate of 95% bound	3.4%	7.10%	8.60%

These results demonstrate the soundness of the $(1-\theta)$ error bound given by Prop. 1. We also observe that, the $(1-\theta)$ error bound is more likely to be overestimated, as the system becomes closer to be singular (corresponds to small value of σ).

Bibliography

[BB96] S. J. Bradke and A. G. Barto. Linear least squares methods for temporal difference learning. *Machine Learning*, 22:33–57, 1996.

[BB98] M. Bertero and P. Boccacci. *Introduction to Inverse Problems in Imaging.* Institute of Physics, Bristol, 1998.

[BBN04] D. P. Bertsekas, V. Borkar, and A. Nedic. Improved temporal difference methods with linear function approximation. *Learning and Approximate Dynamic Programming, by A. Barto, W. Powell, J. Si, (Eds.)*, pages 231–255, 2004.

[Ber07] D. P. Bertsekas. *Dynamic Programming and Optimal Control (Volumes I and II). The third edition.* Athena Scientific, Nashua, USA, 2007.

[Ber09] D. P. Bertsekas. Projected equations, variational inequalities, and temporal difference methods. *Lab. for Information and Decision Systems Report LIDS-P-2808, MIT*, 2009.

[BI96] D. P. Bertsekas and S. Ioffe. Temporal difference-based policy iteration and applications in neuro-dynamic programming methods. *Lab. for Info. and Dec. Sys. Report LIDS-P-2349, MIT*, 1996.

[BL97] Kenneth P. Bube and Robert T. Langanz. Hybrid L1/L2 minimization with applications to tomography. *Geophysics*, 1997.

[Bor08] V. S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint.* N. Y., 2008.

[BP79] R. H. Byrd and D. A. Pyne. Convergence of iteratively reweighted least squares algorithm for robust regression. *Technical Report, Department of Mathematical Sciences, The Johns Hopkins University*, page 313, 1979.

[BT96] D. P. Bertsekas and J. Tsitsiklis. *Neuro-Dynamic Programming.* Athena Scientific, Nashua, USA, 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. *Journal of Computational and Applied Mathematics*, 227:27–50, 2009.

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

[CFHM07] H. S. Chang, M. C. Fu, J. Hu, and S. I. Marcus. *Simulation-Based Algorithms for Markov Decision Processes.* N. Y., 2007.

[Cul71] J. Cullum. Numerical differentiation and regularization. *SIAM Journal of Numerical Analysis*, 8(2):254–265, 1971.

[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.

[Cur57] J. H. Curtiss. Monte Carlo methods for the iteration of linear operators. *UMN*, 12(5(77)):149–174, 1957.

[DCZ99] L. Reichel D. Calvetti and Q. Zhang. *Applied and Computational Control, Signals and Systems*, volume 1, chapter Iterative solution methods for large linear discrete ill-posed problems, pages 313–367. Birkhauser, 1999.

[DS66] Norman Richard Draper and Harry Smith. *Applied Regression Analysis.* John Wiley and Sons, Inc, New York, U.S., 1966.

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

[ES00] M. Evans and T. Swartz. *Approximating Integrals via Monte Carlo Simulation and Deterministic Methods.* Oxford University Press, Oxford, 2000.

[FG53] L. Fox and E. T. Goodwin. The numerical solution of non-singular linear integral equations. *Phil. Trans. R. Soc. Lond. A*, 245:501–534, 1953.

[FL50] G. E. Forsythe and R. A. Leibler. Matrix inversion by a Monte Carlo method. *Mathematical Tables and Other Aids to Computation*, 4(31):127–129, 1950.

[Gre84] P. J. Green. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society. Series B (Methodological)*, (2):149, 1984.

[Gro07] C. W. Groetsch. Integral equations of the first kind, inverse problems and regularization: a crash course. *Journal of Physics: Conference Series: Inverse Problems in Applied Sciences, towards breakthrough*, 73:1–32, 2007.

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

[HH93] M. Hanke and P. C. Hansen. Regularization methods for large scale problems. *Surveys on Mathematics for Industry*, 3:253–315, 1993.

[Jol02] I. T. Jolliffe. *Principal Component Analysis.* Springer, New York, U.S., 2002.

[Jr.72] C. B. Shaw Jr. Imrovement of the resolution of an instrument by numerical solution of an integral equation. *Journal of Mathematical Analysis and Applications*, 37:83–112, 1972.

[KS04] J. Kaipio and E. Somersalo. *Statistical and Computational Inverse Problems.* Springer, New York, 2004.

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

[Lan58] C. Lanczos. Iterative solution of large-scale linear systems. *J. Soc. Indust. and Appl. Math.*, 6(91):91–109, 1958.

[LC98] J. S. Liu and R. Chen. Sequential Monte Carlo methods for dynamic systems. *Journal of the American Statistical Association*, 93(443):1032–1044, 1998.

[Liu01] J. S. Liu. *Monte Carlo Strategies in Scientific Computing.* N. Y., 2001.

[Mey07] S. Meyn. *Control Techniques for Complex Networks.* N. Y., 2007.

[MMS05] I Menache, S Mannor, and N Shimkin. Basis function adaptation in temporal difference reinforcement learning. *Annals of Operations Research*, 2005.

[Moo20] E. H. Moore. On the reciprocal of the general algebraic matrix. *Bulletin of the American Mathematical Society*, page 394395, 1920.

[NB03] A. Nedic and D. P. Bertsekas. Least squares policy evaluation algorithms with linear function approximation. *Discrete Event Dynamic Systems: Theory and Applications*, 13:79–110, 2003.

[OB92] M.-S. Oh and J. O. Berger. Adaptive importance sampling in Monte Carlo integration. *Journal of Statistical Computation and Simulation*, 41(3-4):143–168, 1992.

[OS81] D. P. O'Leary and J. A. Simmons. A bidiagonalization-regularization procedure for large scale discretizations of ill-posed problems. *SIAM J. on Scientific and Statistical Computing*, 2(4):474–489, 1981.

[Pen55] R. Penrose. A generalized inverse for matrices. *Proceedings of the Cambridge Philosophical Society*, page 406413, 1955.

[Pow07] W. B. Powell. *Approximate Dynamic Programming: Solving the Curses of Dimensionality.* 2007.

[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.

[RS02] M. Rojas and D. C. Sorensen. A trust region approach to the regularization of large-scale discrete forms of ill-posed problems. *SIAM J. on Scientific Computing*, 23(6):1843–1861, 2002.

[Saa03] Y. Saad. *Iterative Methods for Sparse Linear Systems, 2nd ed.* SIAM, Phila., PA, 2003.

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

[SV09] T. Strohmer and R. Vershynin. A randomized Kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications*, 15(2):262–278, 2009.

[Was52] W. R. Wasow. A note on inversion of matrices by random walks. *Mathematical Tables and Other Aids to Computation*, 6(38):78–81, 1952.

[WPB09] M. Wang, N. Polydorides, and D. P. Bertsekas. Approximate simulation-based solution of large-scale least squares problems. *LIDS Report, MIT*, 2009.

[WU96] A. D. Woodbury and T. J. Ulrych. Minimum relative entropy inversion: Theory and application to recovering the release history of a groundwater contaminant. *Water Resoures Research*, 9(32):2671–2681, 1996.

[YB06] H. Yu and D. P. Bertsekas. Convergence results for some temporal difference methods based on least squares. *Lab. for Information and Decision Systems Report 2697, MIT*, 2006.

[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.

[YB09] H. Yu and D. P. Bertsekas. Basis function adaptation methods for cost approximation in mdp. *Proc. of IEEE International Symposium on Adaptive Dynamic Programming and Reinforcement Learning*, 2009.