# Monte Carlo Linear Algebra: A Review and Recent Results

Dimitri P. Bertsekas

Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology

Caradache, France, 2012

# Monte Carlo Linear Algebra

An emerging field combining Monte Carlo simulation and algorithmic linear algebra

Plays a central role in approximate DP (policy iteration, projected equation and aggregation methods)

### Advantage of Monte Carlo

Can be used to approximate sums of huge number of terms such as high-dimensional inner products

### A very broad scope of applications

- Linear systems of equations
- Least squares/regression problems
- Eigenvalue problems
- Linear and quadratic programming problems
- Linear variational inequalities
- Other quasi-linear structures

# Monte Carlo Estimation Approach for Linear Systems

## We focus on solution of $Cx = d$

- Use simulation to compute $C_k \to C$ and $d_k \to d$
- Estimate the solution by matrix inversion $C_k^{-1} d_k \approx C^{-1} d$ (assuming $C$ is invertible)
- Alternatively, solve $C_k x = d_k$ iteratively

## Why simulation?

$C$ may be of small dimension, but may be defined in terms of matrix-vector products of huge dimension

## What are the main issues?

- Efficient simulation design that matches the structure of $C$ and $d$
- Efficient and reliable algorithm design
- What to do when $C$ is singular or nearly singular

# References

## Collaborators: Huizhen (Janey) Yu, Mengdi Wang


- D. P. Bertsekas and H. Yu, "Projected Equation Methods for Approximate Solution of Large Linear Systems," Journal of Computational and Applied Mathematics, Vol. 227, 2009, pp. 27-50.
- H. Yu and D. P. Bertsekas, "Error Bounds for Approximations from Projected Linear Equations," Mathematics of Operations Research, Vol. 35, 2010, pp. 306-329.
- D. P. Bertsekas, "Temporal Difference Methods for General Projected Equations," IEEE Trans. on Aut. Control, Vol. 56, pp. 2128 - 2139, 2011.
- M. Wang and D. P. Bertsekas, "Stabilization of Stochastic Iterative Methods for Singular and Nearly Singular Linear Systems", Lab. for Information and Decision Systems Report LIDS-P-2878, MIT, December 2011 (revised March 2012).
- M. Wang and D. P. Bertsekas, "Convergence of Iterative Simulation-Based Methods for Singular Linear Systems", Lab. for Information and Decision Systems Report LIDS-P-2879, MIT, December 2011 (revised April 2012).
- D. P. Bertsekas, Dynamic Programming and Optimal Control, Vol. II: Approximate Dyn. Programming, Athena Scientific, Belmont, MA, 2012.

# Outline

# Low-Dimensional Approximation

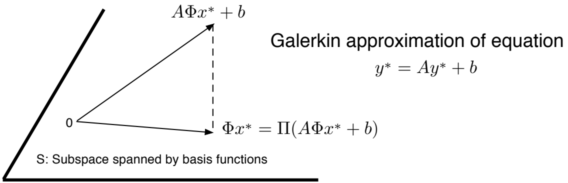
- Start from a high-dimensional equation $y = Ay + b$
- Approximate its solution within a subspace $S = \{\Phi x \mid x \in \Re^s\}$
- Columns of $\Phi$ are basis functions

## Equation approximation approach

Approximate solution $y^*$ with the solution $\Phi x^*$ of an equation defined on $S$

## Important example: Projection/Galerkin approximation

$$\Phi x = \Pi(A\Phi x + b)$$



# Matrix Form of Projected Equation

Let $\Pi$ be projection with respect to a weighted Euclidean norm $\|y\|_\Xi = \sqrt{y'\Xi y}$

The Galerkin solution is obtained from the orthogonality condition

$$\Phi x^* - (A\Phi x^* + b) \perp (\text{Columns of } \Phi)$$

or

$$Cx = d$$

where

$$C = \Phi'\Xi(I - A)\Phi, \qquad d = \Phi'\Xi b$$

## Motivation for simulation

If $y$ is high-dimensional, $C$ and $d$ involve high-dimensional matrix-vector operations

# Another Example: Aggregation

Let $D$ and $\Phi$ be matrices whose rows are probability distributions.

## Aggregation equation

By forming convex combinations of variables (i.e., $y \approx \Phi x$) and equations (using $D$), we obtain an aggregate form of the fixed point problem $y = Ay + b$:

$$x = D(A\Phi x + b)$$

or $Cx = d$ with

$$C = DA\Phi, \qquad d = Db$$

## Connection with projection/Galerkin approximation

The aggregation equation yields

$$\Phi x = \Phi D(A\Phi x + b)$$

$\Phi D$ is an oblique projection in some of the most interesting types of aggregation [if $D\Phi = I$ so that $(\Phi D)^2 = \Phi D$] - also a seminorm projection.

# Another Example: Large-Scale Regression

## Weighted least squares problem

Consider

$$\min_{y \in \Re^n} \|Wy - h\|_\Xi^2,$$

where $W$ and $h$ are given, $\|\cdot\|_\Xi$ is a weighted Euclidean norm, and $y$ is high-dimensional.

We approximate $y$ within the subspace $S = \{\Phi x \mid x \in \Re^s\}$, to obtain

$$\min_{x \in \Re^s} \|W\Phi x - h\|_\Xi^2.$$

## Equivalent linear system $Cx = d$

$$C = \Phi' W' \Xi W \Phi, \qquad d = \Phi' W' \Xi h$$

# Key Idea for Simulation

## Critical Problem

Compute sums $\sum_{i=1}^{n} a_i$ for very large $n$ (or $n = \infty$)

## Convert Sum to an Expected Value

Introduce a sampling distribution $\xi$ and write

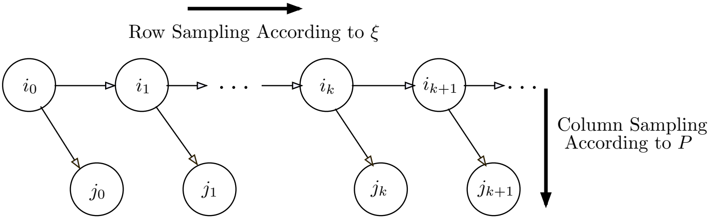
$$\sum_{i=1}^{n} a_i = \sum_{i=1}^{n} \xi_i \left( \frac{a_i}{\xi_i} \right) = E_\xi\{\hat{a}\}$$

where the random variable $\hat{a}$ has distribution

$$P\left\{ \hat{a} = \frac{a_i}{\xi_i} \right\} = \xi_i, \qquad i = 1, \ldots, n$$

- We “invent" $\xi$ to convert a “deterministic" problem to a “stochastic" problem that can be solved by simulation.
- Complexity advantage: Running time is independent of the number $n$ of terms in the sum, only the distribution of $\hat{a}$.
- Importance sampling idea: Use a sampling distribution that matches the problem for efficiency (e.g., make the variance of $\hat{a}$ small) .

# Row and Column Sampling for System $Cx = d$




- Row sampling: Generate sequence $\{i_0, i_1, \dots\}$ according to $\xi$ (the diagonal of $\Xi$), i.e., relative frequency of each row $i$ is $\xi_i$
- Column sampling: Generate sequence $\{(i_0, j_0), (i_1, j_1), \dots\}$ according to some transition probability matrix $P$ with

$$p_{ij} > 0 \quad \text{if} \quad a_{ij} \neq 0,$$

  i.e., for each $i$, the relative frequency of $(i, j)$ is $p_{ij}$
- Row sampling may be done using a Markov chain with transition matrix $Q$ (unrelated to $P$)
- Row sampling may also be done without a Markov chain - just sample rows according to some known distribution $\xi$ (e.g., a uniform)

# Simulation Formulas for Matrix Form of Projected Equation

- Approximation of $C$ and $d$ by simulation:

$$C = \Phi'\Xi(I - A)\Phi \quad \sim \quad C_k = \frac{1}{k+1}\sum_{t=0}^{k} \phi(i_t)\left(\phi(i_t) - \frac{a_{i_t j_t}}{p_{i_t j_t}}\phi(j_t)\right)',$$

$$d = \Phi'\Xi b \quad \sim \quad d_k = \frac{1}{k+1}\sum_{t=0}^{k} \phi(i_t) b_{i_t}$$

- We have by law of large numbers $C_k \to C$, $d_k \to d$.
- Equation approximation: Solve the equation $C_k x = d_k$ in place of $Cx = d$.

## Algorithms

- Matrix inversion approach: $x^* \approx C_k^{-1} d_k$ (if $C_k$ is invertible for large $k$)
- Iterative approach: $x_{k+1} = x_k - \gamma G_k (C_k x_k - d_k)$

# Multistep Methods - TD($\lambda$)-Type

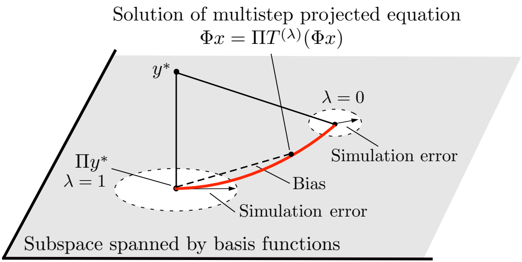
Instead of solving (approximately) the equation $y = T(y) = Ay + b$, consider the multistep equivalent

$$y = T^{(\lambda)}(y)$$

where for $\lambda \in [0, 1)$

$$T^{(\lambda)} = (1 - \lambda) \sum_{\ell=0}^{\infty} \lambda^{\ell} T^{\ell+1}$$

- Special multistep sampling methods
- Bias-variance tradeoff

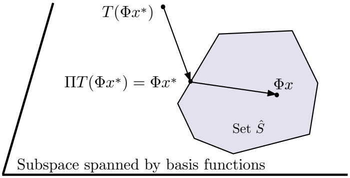


# Constrained Projected Equations

- Consider

$$\Phi x = \Pi T(\Phi x) = \Pi(A\Phi x + b)$$

where $\Pi$ is the projection operation onto a closed convex subset $\hat{S}$ of the subspace $S$ (w/ respect to weighted norm $\|\cdot\|_\Xi$; $\Xi$: positive definite).




- From the properties of projection,

$$\big(\Phi x^* - T(\Phi x^*)\big)'\Xi(y - \Phi x^*) \geq 0, \qquad \forall\, y \in \hat{S}$$

- This is a linear variational inequality: Find $x^*$ such that

$$f(\Phi x^*)'(y - \Phi x^*) \geq 0, \qquad \forall\, y \in \hat{S},$$

where $f(y) = \Xi\big(y - T(y)\big) = \Xi\big(y - (Ay + b)\big)$.

# Equivalence Conclusion

## Two equivalent problems

- The projected equation

$$\Phi x = \Pi T(\Phi x)$$

where $\Pi$ is projection with respect to $\|\cdot\|_\Xi$ on convex set $\hat{S} \subset S$

- The special-form VI

$$f(\Phi x^*)'\Phi(x - x^*) \geq 0, \qquad \forall\, x \in X,$$

where

$$f(y) = \Xi\big(y - T(y)\big), \qquad X = \{x \mid \Phi x \in \hat{S}\}$$

## Special linear cases: $T(y) = Ay + b$

- $\hat{S} = \Re^n$: VI <==> $f(\Phi x^*) = \Xi\big(\Phi x^* - T(\Phi x^*)\big) = 0$ (linear equation)
- $\hat{S} =$ subspace: VI <==> $f(\Phi x^*) \perp \hat{S}$ (e.g., projected linear equation)
- $f(y)$ the gradient of a quadratic, $\hat{S}$: polyhedral (e.g., approx. LP and QP)
- Linear VI case (e.g., cooperative and zero-sum games with approximation)

# Deterministic Solution Methods - Invertible Case of $Cx = d$

## Matrix Inversion Method

$$x^* = C^{-1}d$$

## Generic Linear Iterative Method

$$x_{k+1} = x_k - \gamma G(Cx_k - d)$$

where:

- $G$ is a scaling matrix, $\gamma > 0$ is a stepsize
- Eigenvalues of $I - \gamma GC$ within the unit circle (for convergence)

## Special cases:

- Projection/Richardson's method: $C$ positive semidefinite, $G$ positive definite symmetric
- Proximal method (quadratic regularization)
- Splitting/Gauss-Seidel method

# Simulation-Based Solution Methods - Invertible Case

Given sequences $C_k \to C$ and $d_k \to d$

### Matrix Inversion Method

$$x_k = C_k^{-1} d_k$$

### Iterative Method

$$x_{k+1} = x_k - \gamma G_k (C_k x_k - d_k)$$

where:

- $G_k$ is a scaling matrix with $G_k \to G$
- $\gamma > 0$ is a stepsize

$x_k \to x^*$ if and only if the deterministic version is convergent

# Solution Methods - Singular Case (Assuming a Solution Exists)

Given sequences $C_k \to C$ and $d_k \to d$. Matrix inversion method does not apply

## Iterative Method

$$x_{k+1} = x_k - \gamma G_k(C_k x_k - d_k)$$

Need not converge to a solution, even if the deterministic version does

## Questions:

- Under what conditions is the stochastic method convergent?
- How to modify the method to restore convergence?

# Simulation-Based Solution Methods - Nearly Singular Case

**The theoretical view**

If $C$ is nearly singular, we are in the nonsingular case

**The practical view**

If $C$ is nearly singular, we are essentially in the singular case (unless the simulation is extremely accurate)

The eigenvalues of the iteration

$$x_{k+1} = x_k - \gamma G_k(C_k x_k - d_k)$$

get in and out of the unit circle for a long time (until the “size" of the simulation noise becomes comparable to the “stability margin" of the iteration)

Think of roundoff error affecting the solution of ill-conditioned systems (simulation noise is far worse)

# Deterministic Iterative Method - Convergence Analysis

Assume that $C$ is invertible or singular (but $Cx = d$ has a solution)

## Generic Linear Iterative Method

$$x_{k+1} = x_k - \gamma G(Cx_k - d)$$

## Standard Convergence Result

Let $C$ be singular and denote by $\mathbf{N}(C)$ the nullspace of $C$. Then:
$\{x_k\}$ is convergent (for all $x_0$ and sufficiently small $\gamma$) to a solution of $Cx = d$ if and only if:

(a) Each eigenvalue of $GC$ either has a positive real part or is equal to 0.

(b) The dimension of $\mathbf{N}(GC)$ is equal to the algebraic multiplicity of the eigenvalue 0 of $GC$.

(c) $\mathbf{N}(C) = \mathbf{N}(GC)$.

# Proof Based on Nullspace Decomposition for Singular Systems

**For any solution $x^*$, rewrite the iteration as**

$$x_{k+1} - x^* = (I - \gamma GC)(x_k - x^*)$$

**Linarly transform the iteration**

Introduce a similarity transformation involving $\mathbf{N}(C)$ and $\mathbf{N}(C)^\perp$

Let $U$ and $V$ be orthonormal bases of $\mathbf{N}(C)$ and $\mathbf{N}(C)^\perp$:

$$\begin{aligned}
[U\ V]'(I - \gamma GC)[U\ V] &= I - \gamma \begin{bmatrix} U'GCU & U'GCV \\ V'GCU & V'GCV \end{bmatrix} \\
&= I - \gamma \begin{bmatrix} 0 & U'GCV \\ 0 & V'GCV \end{bmatrix} \\
&\equiv \begin{bmatrix} I & -\gamma N \\ 0 & I - \gamma H \end{bmatrix},
\end{aligned}$$

where $H$ has eigenvalues with positive real parts. Hence for some $\gamma > 0$,

$$\rho(I - \gamma H) < 1,$$

so $I - \gamma H$ is a contraction ... $\blacksquare$

# Nullspace Decomposition of Deterministic Iteration

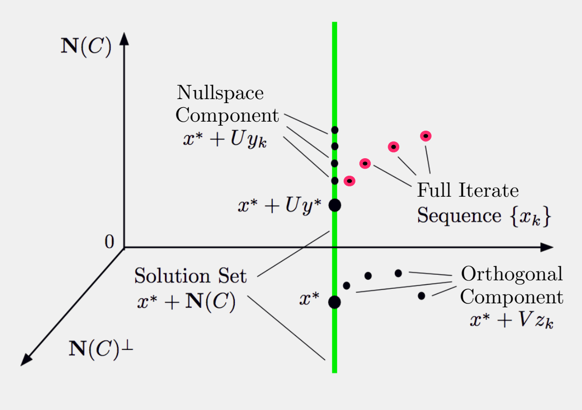



Figure: Iteration decomposition into components on $\mathbf{N}(C)$ and $\mathbf{N}(C)^\perp$.

$$x_k = x^* + Uy_k + Vz_k$$

- Nullspace component: $y_{k+1} = y_k - \gamma N z_k$
- Orthogonal component: $z_{k+1} = z_k - \gamma H z_k$ CONTRACTIVE

# Stochastic Iterative Method May Diverge

The stochastic iteration

$$x_{k+1} = x_k - \gamma G_k (C_k x_k - d_k)$$

approaches the deterministic iteration

$$x_{k+1} = x_k - \gamma G(Cx_k - d), \qquad \text{where} \quad \rho(I - \gamma GC) \leq 1.$$

However, since

$$\rho(I - \gamma G_k C_k) \to 1$$

$\rho(I - \gamma G_k C_k)$ may cross above 1 too frequently, and we can have divergence.

Difficulty is that the orthogonal component is now coupled to the nullspace component with simulation noise

# Divergence of the Stochastic/Singular Iteration

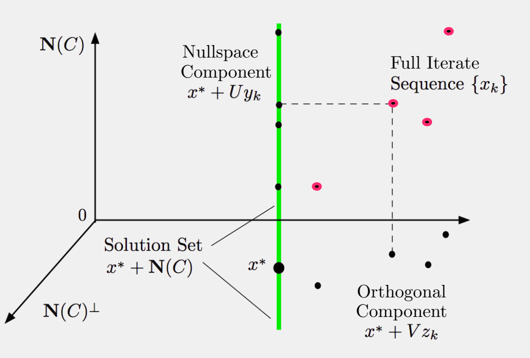



Figure: NOISE LEAKAGE FROM $\mathbf{N}(C)$ to $\mathbf{N}(C)^\perp$

$$x_k = x^* + Uy_k + Vz_k$$

- Nullspace component: $y_{k+1} = y_k - \gamma N z_k +$ Noise$(y_k, z_k)$
- Orthogonal component: $z_{k+1} = z_k - \gamma H z_k +$ Noise$(y_k, z_k)$

# Divergence Example for a Singular Problem

## $2 \times 2$ Example

Let the noise be $\{e_k\}$: MC averages with mean 0 so $e_k \to 0$, and let

$$x_{k+1} = \begin{bmatrix} 1 + e_k & 0 \\ e_k & 1/2 \end{bmatrix} x_k$$

- Nullspace component $y_k = x_k(1)$ diverges:

$$\prod_{t=1}^{k} (1 + e_t) = O(e^{\sqrt{k}}) \to \infty$$

- Orthogonal component $z_k = x_k(2)$ diverges:

$$x_{k+1}(2) = 1/2 x_k(2) + e_k \prod_{t=1}^{k} (1 + e_t),$$

where

$$e_k \prod_{t=1}^{k} (1 + e_t) = O\left(\frac{e^{\sqrt{k}}}{\sqrt{k}}\right) \to \infty.$$

# What Happens in Nearly Singular Problems?

- “Divergence" until Noise $<<$ “Stability Margin" of the iteration
- Compare with roundoff error problems in inversion of nearly singular matrices

## A Simple Example

Consider the inversion of a scalar $c > 0$, with simulation error $\eta$. The absolute and relative errors are

$$E = \frac{1}{c+\eta} - \frac{1}{c}, \qquad E_r = \frac{E}{1/c}.$$

By a Taylor expansion around $\eta = 0$:

$$E \approx \frac{\partial\big(1/(c+\eta)\big)}{\partial\eta}\bigg|_{\eta=0}\eta = -\frac{\eta}{c^2}, \qquad E_r \approx -\frac{\eta}{c}.$$

For the estimate $\frac{1}{c+\eta}$ to be reliable, it is required that

- $|\eta| << |c|$.
- Number of i.i.d. samples needed: $k \gg 1/c^2$.

# Nullspace Consistent Iterations

## Nullspace Consistency and Convergence of Residual

- If $\mathbf{N}(G_k C_k) \equiv \mathbf{N}(C)$, we say that the iteration is nullspace-consistent.
- Nullspace consistent iteration generates convergent residuals ($Cx_k - d \to 0$), iff the deterministic iteration converges.

## Proof Outline:

$$x_k = x^* + Uy_k + Vz_k$$

- Nullspace component: $y_{k+1} = y_k - \gamma N z_k +$ Noise$(y_k, z_k)$
- Orthogonal component: $z_{k+1} = z_k - \gamma H z_k +$ Noise$(z_k)$ DECOUPLED

LEAKAGE FROM $\mathbf{N}(C)$ IS ANIHILATED by $V$ so

$$Cx_k - d = CVz_k \to 0$$

$\blacksquare$

# Interesting Special Cases

## Proximal/Quadratic Regularization Method

$$x_{k+1} = x_k - (C_k' C_k + \beta I)^{-1} C_k' (C_k x_k - d_k)$$

Can diverge even in the nullspace consistent case.

- In the nullspace consistent case, under favorable conditions $x_k \to$ some solution $x^*$.
- In these cases the nullspace component $y_k$ stays constant.

## Approximate DP (projected equation and aggregation)

The estimates often take the form

$$C_k = \Phi' M_k \Phi, \qquad d_k = \Phi' h_k,$$

where $M_k \to M$ for some positive definite $M$.

- If $\Phi$ has dependent columns, the matrix $C = \Phi' M \Phi$ is singular.
- The iteration using such $C_k$ and $d_k$ is nullspace consistent.
- In typical methods (e.g., LSPE) $x_k \to$ some solution $x^*$.

# Stabilization of Divergent Iterations

## A Stabilization Scheme

Shifting the eigenvalues of $I - \gamma G_k C_k$ by $-\delta_k$:

$$x_{k+1} = (1 - \delta_k)x_k - \gamma G_k \left(C_k x_k - d_k\right).$$

## Convergence of Stabilized Iteration

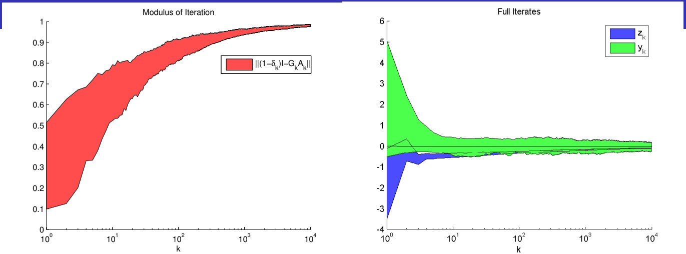
Assume that the eigenvalues are shifted slower than the convergence rate of the simulation:

$$(C_k - C, d_k - d, G_k - G)/\delta_k \to 0, \qquad \sum_{k=0}^{\infty} \delta_k = \infty$$

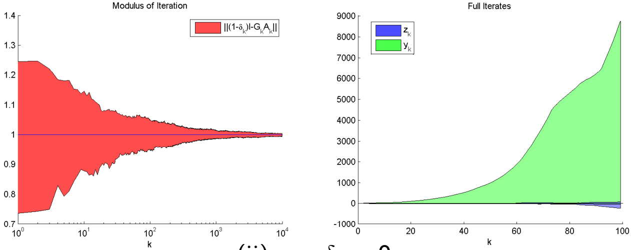
Then the stabilized iteration generates $x_k \to$ some $x^*$ iff the deterministic iteration without $\delta_k$ does.

- Stabilization is interesting even in the nonsingular case
- It provides a form of “regularization"

# Stabilization of the Earlier Divergent Example




(i) $\delta_k = k^{-1/3}$




(ii) $\delta_k = 0$

Thank You!