Randomized Iterative Methods for Linear Systems

Robert Mansel Gower & Peter Richtárik
University of Edinburgh

LMS Inverse Day on Large-scale and Nonlinear Inverse Problems
September 21, 2015
Robert M Gower
(Edinburgh)

Robert M Gower and P.R.
Randomized Iterative Methods for Linear Systems
The Problem
The Problem

\[ Ax = b \]

Assumption: The system is consistent (i.e., has a solution)

We can also think of this as \( m \) linear equations, where the \( i^{th} \) equation looks as follows:

\[
\sum_{j=1}^{n} A_{ij} x_j = b_i
\]

\[ A_i: x = b_i \]
Minimizing Convex Quadratics

\[
\min_{x \in \mathbb{R}^n} \left[ f(x) = \frac{1}{2} \|Ax - b\|^2 \right] \Rightarrow \nabla f(x) = 0 \Rightarrow A^T Ax = A^T b
\]

This system is consistent.

\[
\min_{x \in \mathbb{R}^n} \left[ f(x) = \frac{1}{2} x^T Ax + b^T x + c \right] \Rightarrow \nabla f(x) = 0 \Rightarrow Ax = b
\]

A = positive definite

This system is consistent.
The Solution
(6 Ways to Skin a Cat)
skin the cat

Term refers to a task which has several ways by which it can be completed. Often used in the expression "there are many ways to skin the cat" or by using "skin this cat" in place of "skin the cat."

My friends and I are going to start a business, but we don't even know where to begin because there are so many ways to skin the cat.

by CRubio April 15, 2007
1. Relaxation Viewpoint

“Sketch and Project”

\[
x^{t+1} = \arg \min_{x \in \mathbb{R}^n} \|x - x^t\|_B^2
\]

subject to \( S^T A x = S^T b \)

One Step Method: \( S = m \times m \) invertible (with probability 1)
2. Optimization Viewpoint
“Constrain and Approximate”

\[
x^{t+1} = \arg \min_{x \in \mathbb{R}^n} \| x - x^* \|^2_B
\]

subject to \( x = x^t + B^{-1} A^T S y \)

\( y \) is free
3. Geometric Viewpoint
“Random Intersect”

Lemma  \( \text{Null}(S^T A) \) and \( \text{Range}(B^{-1}A^T S) \) are \( B \)-orthogonal complements

Proof \[ h \in \text{Null}(S^T A) \quad \Rightarrow \quad \langle B^{-1}A^T Sy, h \rangle_B = (y^T S^T AB^{-1}) Bh = y^T S^T Ah = 0 \]

\[
\{x^{t+1}\} = \left(x^* + \text{Null}(S^T A)\right) \cap \left(x^t + \text{Range}(B^{-1}A^T S)\right)
\]
4. Algebraic Viewpoint
“Random Linear Solve”

\[ x^{t+1} = \text{solution in } x \text{ of the linear system} \]

\[ S^T A x = S^T b \]

\[ x = x^t + B^{-1} A^T S y \]
5. Algebraic Viewpoint

“Random Update”

$x^{t+1} = x^t - B^{-1} A^T S (S^T A B^{-1} A^T S)^\dagger S^T (A x^t - b)$

**Fact:** Every (not necessarily square) real matrix $M$ has a real pseudo-inverse $M^\dagger$.

**Some properties:**

1. $MM^\dagger M = M$
2. $M^\dagger MM^\dagger = M^\dagger$
3. $(M^T M)^\dagger M^T = M^\dagger$
4. $(M^T)^\dagger = (M^\dagger)^T$
5. $(MM^T)^\dagger = (M^\dagger)^T M^\dagger$
6. Analytic Viewpoint “Random Fixed Point”

\[
x^{t+1} - x^* = (I - B^{-1}Z)(x^t - x^*)
\]

\[
Z := A^T S (S^T A B^{-1} A^T S)^\dagger S^T A
\]

\[
(B^{-1}Z)^2 = B^{-1}Z
\]
\[
(I - B^{-1}Z)^2 = I - B^{-1}Z
\]

$B^{-1}Z$ projects orthogonally onto Range$(B^{-1}A^T S)$

$I - B^{-1}Z$ projects orthogonally onto Null$(S^T A)$
Verifying that $B^{-1}Z$ is a Projection

\[
(B^{-1}Z)^2 = B^{-1}A^TS(S^T AB^{-1}A^TS)^{\dagger}S^TAB^{-1}A^TS(S^T AB^{-1}A^TS)^{\dagger}S^TA
\]

\[
= B^{-1}A^TS(S^T AB^{-1}A^TS)^{\dagger}S^TA
\]

\[
= B^{-1}Z
\]

\[
M^\dagger MM^\dagger = M^\dagger
\]

Eigenvalues of $B^{-1}Z$ are in \{0,1\}
Theory
Complexity / Convergence

**Theorem [RG’15]** For every solution $x^*$ of $Ax = b$ we have

$$
\mathbf{E} \left[ x^{t+1} - x^* \right] = \left( I - B^{-1} \mathbf{E}[Z] \right) \mathbf{E} \left[ x^t - x^* \right]
$$

Moreover,

$$
\| \mathbf{E} \left[ x^t - x^* \right] \|_B \leq \rho^t \| x^0 - x^* \|_B
$$

where

$$
\rho := \| I - B^{-1} \mathbf{E}[Z] \|_B
$$

$$
\| M \|_B := \max_{\| x \|_B = 1} \| Mx \|_B
$$

1. $\mathbf{E}[Z] \succ 0$
2. $\| M \|_B := \max_{\| x \|_B = 1} \| Mx \|_B$

$$
\mathbf{E} \left[ \| x^t - x^* \|_B^2 \right] \leq \rho^t \| x^0 - x^* \|_B^2
$$
Proof of \( 1 \)

\[
x^{t+1} - x^* = (I - B^{-1}Z)(x^t - x^*)\]

Taking expectations conditioned on \( x^t \), we get

\[
E[x^{t+1} - x^* | x^t] = (I - B^{-1}E[Z])(x^t - x^*).
\]

Taking expectation again gives

\[
\begin{align*}
E[x^{t+1} - x^*] &= E\left[ E\left[ x^{t+1} - x^* | x^t \right] \right] \\
&= E \left[ (I - B^{-1}E[Z])(x^t - x^*) \right] \\
&= (I - B^{-1}E[Z])E[x^t - x^*].
\end{align*}
\]

Applying the norms to both sides we obtain the estimate

\[
\|E[x^{t+1} - x^*]\|_B \leq \|I - B^{-1}E[Z]\|_B \|E[x^t - x^*]\|_B \cdot \rho.
\]
The Rate: Lower and Upper Bounds

\[ d \triangleq \text{Rank}(S^T A) = \text{dim}(\text{Range}(B^{-1} A^T S)) = \text{Tr}(B^{-1} Z) \]

**Theorem [RG‘15]**

\[ 0 \leq 1 - \frac{\mathbf{E}[d]}{n} \leq \rho \leq 1 \]

**Insight:** The method is a contraction (without any assumptions on \( S \) whatsoever). That is, things can not get worse.

**Insight:** The lower bound on the rate improves as the dimension of the search space in the “constrain and approximate” viewpoint grows.
Proof

\[ \rho = \|I - B^{-1} \mathbf{E}[Z]\|_B \]

\[ = \lambda_{\text{max}}(I - B^{-1/2} \mathbf{E}[Z]B^{-1/2}) \]

\[ = 1 - \lambda_{\text{min}}(B^{-1/2} \mathbf{E}[Z]B^{-1/2}) \]

\[ = 1 - \lambda_{\text{min}}(\mathbf{E}[B^{-1/2} Z B^{-1/2}]) \]

\[ \geq 1 - \frac{\text{Tr} (\mathbf{E} [B^{-1} Z])}{n} \]

\[ = 1 - \frac{\mathbf{E} [\text{Tr}(B^{-1} Z)]}{n} \]
The Rate: Sufficient Condition for Convergence

Lemma

If $E[Z]$ is invertible, then

(i) $\rho < 1$,

(ii) $A$ has full column rank, and

(iii) $x^*$ is unique

\[ \rho = 1 - \lambda_{\text{min}}(B^{-1}E[Z]) \]
Special Case: Randomized Kaczmarz Method
Randomized Kaczmarz (RK) Method


RK arises as a special case for parameters $B, S$ set as follows:

$B = I$

$S = e^i = (0, \ldots, 0, 1, 0, \ldots, 0)$ with probability $p_i$

$x^{t+1} = x^t - \frac{A_i : x^t - b_i}{\|A_i :\|^2} (A_i :)^T$

RK was analyzed for $p_i = \frac{\|A_i :\|^2}{\|A\|^2_F}$
**RK: Derivation and Rate**

**General Method**

\[
x^{t+1} = x^t - B^{-1} A^T S \left( S^T A B^{-1} A^T S \right)^\dagger S^T (A x^t - b)
\]

**Special Choice of Parameters**

\[
P(S = e^i) = p_i \quad B = I \quad S = e^i
\]

**Complexity Rate**

\[
p_i = \frac{\|A_i\|^2}{\|A\|^2_F}
\]

\[
E \left[ \|x^t - x^*\|_2^2 \right] \leq \left( 1 - \frac{\lambda_{\min} (A^T A)}{\|A\|^2_F} \right)^t \|x^0 - x^*\|_2^2
\]
RK = SGD with a “smart” stepsize

\[ Ax = b \text{ vs } \min_x \frac{1}{2} \| Ax - b \|^2 \]

Apply RK

\[ x^{t+1} = x^t - \frac{A_i: x^t - b_i}{\| A_i : \|^2} (A_i:)^T \]

Apply SGD

\[ x^{t+1} = x^t - h^t \nabla f_i(x^t) = x^t - \frac{h^t}{p_i} (A_i: x^t - b_i)(A_i:)^T \]

RK is equivalent to applying SGD with a specific (smart!) constant stepsize!

\[ x^{t+1} = \arg \min_{x \in \mathbb{R}^n} \| x - x^* \|^2_2 \text{ s.t. } x = x^t + y (A_i:)^T, \quad y \in \mathbb{R} \]
RK: Further Reading


Special Case: Randomized Coordinate Descent
Randomized Coordinate Descent in 2D
Randomized Coordinate Descent in 2D
Randomized Coordinate Descent in 2D
Randomized Coordinate Descent in 2D
Randomized Coordinate Descent in 2D
Randomized Coordinate Descent in 2D
Randomized Coordinate Descent in 2D
Randomized Coordinate Descent in 2D

SOLVED!
Randomized Coordinate Descent (RCD)

\[\min_{x \in \mathbb{R}^n} \left[ f(x) = \frac{1}{2} x^T A x - b^T x \right] \]
\[x^* = A^{-1} b\]

Assume: Positive definite

RCD arises as a special case for parameters \(B, S\) set as follows:
\[B = A \quad \quad S = e^i = (0, \ldots, 0, 1, 0, \ldots, 0)\] with probability \(p_i\)

Recall: In RK we had \(B = I\)

RCD was analyzed for \(p_i = \frac{A_{ii}}{\text{Tr}(A)}\)
RCD: Derivation and Rate

**General Method**

\[ x^{t+1} = x^t - B^{-1}A^T S (S^T AB^{-1} A^T S)^\dagger S^T (A x^t - b) \]

**Special Choice of Parameters**

- \( B = A \)
- \( S = e^i \)
- \( p_i = \frac{A_{ii}}{\text{Tr}(A)} \)

**Complexity Rate**

\[ E \left[ \| x^t - x^* \|^2_A \right] \leq \left( 1 - \frac{\lambda_{\text{min}}(A)}{\text{Tr}(A)} \right)^t \| x^0 - x^* \|^2_A \]
RCD uses “Exact Line Search”

Recall Viewpoint 2 ("Constrain and Approximate"):

\[ x^{t+1} = \arg \min_{x \in \mathbb{R}^n} \| x - x^* \|^2_B \]
subject to \( x = x^t + B^{-1} A^T S y \)
\( y \) is free

Observation: \[ \| x - x^* \|^2_A = (x - x^*)^T A(x - x^*) \]
\[ = x^T A x - 2(x^*)^T A x + (x^*)^T A x^* \]
\[ = x^T A x - 2b^T x + b^T x^* \]
\[ = 2f(x) + b^T x^* \]

In RCD we have: \( B = A \quad S = e^i \)

\[ x^* = A^{-1} b \]

Insight:
RCD exactly minimizes \( f \) along a random coordinate direction!
Nesterov considered the problem:

\[
\min_{x \in \mathbb{R}^n} f(x)
\]

Nesterov assumed that the following inequality holds for all \( x, h \) and \( i \):

\[
f(x + he^i) \leq f(x) + \nabla_i f(x)h + \frac{L_i}{2}h^2
\]

Given a current iterate \( x \), choosing \( h \) by minimizing the RHS gives:

**Nesterov’s RCD method:**

\[
x^{t+1} = x^t - \frac{1}{L_i} \nabla_i f(x^t)e^i
\]

We recover RCD as we have seen it:

\[
x^{t+1} = x^t - \frac{(A_{i:})^T x^t - b_i}{A_{ii}} e^i
\]
Special Case: Randomized Newton Method
Randomized Newton (RN)


\[
\min_{x \in \mathbb{R}^n} \left[ f(x) = \frac{1}{2} x^T Ax - b^T x \right]
\]
\[x^* = A^{-1} b\]

Assume: Positive definite

RN arises as a special case for parameters \( B, S \) set as follows:

\[B = A\]
\[S = I_{|C|} \text{ with probability } p_C\]

\[p_C \geq 0 \quad \forall C \subseteq \{1, \ldots, n\}\]
\[\sum_{C \subseteq \{1, \ldots, n\}} p_C = 1\]

RCD is special case with \( p_C = 0 \) whenever \(|C| \neq 1\)
**RN: Derivation**

**General Method**

\[
x^{t+1} = x^t - B^{-1} A^T S (S^T A B^{-1} A^T S)^\dagger S^T (A x^t - b)
\]

**Special Choice of Parameters**

\[
B = A \\
S = I_C \text{ with probability } p_C
\]

\[
x^{t+1} = x^t - I_C ( (I_C)^T A I_C )^{-1} (I_C)^T (A x^t - b)
\]

This method minimizes \( f \) exactly in a random subspace spanned by the coordinates belonging to \( C \)

**Complexity Rate**

Will talk about this more later in the “curvature” part
$n = 2$

$C = \{1, 2\}$
Special Case: Gaussian Descent
Gaussian Descent

General Method

\[ x^{t+1} = x^t - B^{-1} A^T S (S^T A B^{-1} A^T S)^\dagger S^T (A x^t - b) \]

Special Choice of Parameters

\[ S \sim N(0, \Sigma) \]

Positive definite covariance matrix

Complexity Rate

\[ \mathbf{E} \left[ \| x^t - x^* \|^2_B \right] \leq \rho^t \| x^0 - x^* \|^2_B \]


\[
x^{t+1} = x^t - h^t B^{-1/2} \xi
\]

\[
\xi := B^{-1/2} A^T S
\]

\[
\xi \sim N(0, \Omega)
\]

\[
\Omega := B^{-1/2} A^T \Sigma A B^{-1/2}
\]
Gaussian Descent: The Rate

\[ \rho = 1 - \lambda_{\min}(B^{-1}E[Z]) \]
\[ = 1 - \lambda_{\min}(B^{-1/2}E[Z]B^{-1/2}) \]
\[ = 1 - \lambda_{\min}(E[A^T S (S^T A B^{-1} A^T S)^\dagger S^T A] B^{-1/2}) \]
\[ = 1 - \lambda_{\min}(E \left[ B^{-1/2} A^T S (S^T A B^{-1} A^T S)^\dagger S^T A B^{-1/2} \right] \]
\[ = 1 - \lambda_{\min}(E \left[ \frac{\xi \xi^T}{\|\xi\|_2^2} \right] ) \]

\( \xi := B^{-1/2} A^T S \)
\( \xi \sim \mathcal{N}(0, \Omega) \)
\( \Omega := B^{-1/2} A^T \Sigma A B^{-1/2} \)

XY and YX have the same spectrum
Gaussian Descent: The Rate

Lemma [GR’15]

\[
\mathbf{E} \left[ \frac{\xi \xi^T}{\|\xi\|_2^2} \right] \preceq \frac{2}{\pi} \frac{\Omega}{\text{Tr}(\Omega)}
\]

\[
\rho \leq 1 - \frac{2 \lambda_{\min}(\Omega)}{\pi \text{Tr}(\Omega)}
\]

This follows from the general lower bound \( 1 - \frac{\mathbf{E}[d]}{n} \leq \rho \) since \( d = 1 \)
Gaussian Descent: Further Reading


Experiments
Data

\( m = 1,000; \ n = 500 \)

(a) rand

(b) sprandn
Real data (Matrix Market)

(a) illc1033 \((m = 1,850; n = 750)\)

(b) well1033 \((m = 1,033; n = 320)\)
Importance Sampling
Importance Sampling

Assume that $S$ is discrete:

$$S = S_i \quad \text{with probability} \quad p_i \quad (i = 1, \ldots, r)$$

**Question**

Consider $S_1, \ldots, S_r$ fixed. How to choose the probabilities $p_1, \ldots, p_r$ which optimize the convergence rate $\rho = 1 - \lambda_{\min}(B^{-1}E[Z])$?

$$\max_p \left\{ \lambda_{\min}(B^{-1}E[Z]) \quad \text{subject to} \quad \sum_{i=1}^{r} p_i = 1, \; p \geq 0 \right\}$$

- Can be reformulated as an SDP (Semidefinite Program)
- Leads to different probabilities than those proposed for RK and RCD!

$$V_i = B^{-1/2}A^T S_i$$
RCD: Optimal Probabilities Can Lead to a Remarkable Improvement

<table>
<thead>
<tr>
<th>data set</th>
<th>$\rho_c$</th>
<th>$\rho^*$</th>
<th>$1 - 1/n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>rand(50,50)</td>
<td>$1 - 2 \cdot 10^{-6}$</td>
<td>$1 - 3.05 \cdot 10^{-6}$</td>
<td>$1 - 2 \cdot 10^{-2}$</td>
</tr>
<tr>
<td>mushrooms-ridge</td>
<td>$1 - 5.86 \cdot 10^{-6}$</td>
<td>$1 - 7.15 \cdot 10^{-6}$</td>
<td>$1 - 8.93 \cdot 10^{-3}$</td>
</tr>
<tr>
<td>aloi-ridge</td>
<td>$1 - 2.17 \cdot 10^{-7}$</td>
<td>$1 - 1.26 \cdot 10^{-4}$</td>
<td>$1 - 7.81 \cdot 10^{-3}$</td>
</tr>
<tr>
<td>liver-disorders-ridge</td>
<td>$1 - 5.16 \cdot 10^{-4}$</td>
<td>$1 - 8.25 \cdot 10^{-3}$</td>
<td>$1 - 1.67 \cdot 10^{-1}$</td>
</tr>
<tr>
<td>covtype.binary-ridge</td>
<td>$1 - 7.57 \cdot 10^{-14}$</td>
<td>$1 - 1.48 \cdot 10^{-6}$</td>
<td>$1 - 1.85 \cdot 10^{-2}$</td>
</tr>
</tbody>
</table>
RK: Convenient vs Optimal

(a) liver-disorders-popt-k
(b) rand(500,100)
RCD: Convenient vs Optimal

(a) aloi

(b) covtype.libsvm.binary

(c) liver-disorders-ridge

(d) mushrooms-ridge-opt
THE END