



Topics in Numerical Analysis II Computational Inverse Problems

Lecturer: Bangti Jin (b.jin@cuhk.edu.hk)

Chinese University of Hong Kong

October 7, 2023



Outline

1 Kaczmarz method and randomized version

- Kaczmarz method
- Randomized Kaczmarz
- Stochastic gradient descent



Review

linear inverse problems

$$Ax = y$$

- The Landweber method

$$x_{k+1} = x_k - \beta A^*(Ax_k - y)$$

- the method is convergent for exact data: $x_k \rightarrow x^\dagger$
- the method is stable for fixed k

$$\|x_k^\delta - x_k\| \leq \sqrt{k}\delta$$

- convergence rate under source condition $x^\dagger = A^*w$

$$\|x_k - x^\dagger\| \leq (2k + 2)^{-1/2} \|w\|$$

- the discrepancy principle can be applied and is optimal



Review

Landweber method can be very slow ...

- How to accelerate the computation ...
 - it takes many iterations
 - each iteration requires A, A^*
- What are the remedies ?
 - simple: Anderson acceleration
 - simple: Kaczmarz / stochastic gradient descent
 - complex: conjugate gradient, MINRES
 - ...



Anderson acceleration for fixed point equation: $x_{n+1} = T(x_n)$

- let $g(x) = T(x) - x$, $g_k = g(x_k)$
- set x_0 and $m \geq 1$ (memory parameter)

D. G. Anderson. Iterative Procedures for Nonlinear Integral Equations. J. the ACM. 1965; 12 (4): 547–560

```
 $x_1 = T(x_0)$   
for  $k = 1, 2, \dots$  do  
     $m_k = \min(m, k)$   
     $G_k = [g_{k-m_k} \ \dots \ g_k]$   
     $\alpha_k = \arg \min_{\sum_{i=0}^{m_k} \alpha_i = 1} \|G_k \alpha\|$   
     $x_{k+1} = \sum_{i=0}^{m_k} (\alpha_k)_i T(x_{k-m_k+i})$   
end for
```

very effective, but no analysis of the regularizing property!



Old idea: Kaczmarz method starts from initial guess x_0 ,

$$x_{k+1} = x_k + \frac{y_i - (a_i, x_k)}{\|a_i\|^2} a_i, \quad i = (k \bmod n) + 1,$$

orthogonal projection into hyperplanes ...

- Kaczmarz method was proposed in 1937 by Polish mathematician Stefan Kaczmarz.
- It was re-invented by Gordon et al. (1970) under the name algebraic reconstruction technique (ART), for image reconstruction in computed tomography:
[randomized version works better !](#) Natterer 1986
- randomized version first analyzed by Vershynin-Strohmer 2009



Kaczmarz method

Consider the linear system

$$Ax = y$$

with $A \in \mathbb{R}^{m \times n}$, $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$. We write the system matrix A in the form

$$A = \begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_\ell \end{bmatrix}, \quad A_j \in \mathbb{R}^{k_j \times n}, j = 1, \dots, \ell,$$

with $k_1 + \dots + k_\ell = m$, and each submatrix has k_j linearly indep. row vectors, i.e., $\text{rank}(A_j) = k_j \leq n$, and A_j is surjective from $\mathbb{R}^n$ to $\mathbb{R}^{k_j}$.



Similarly, we decompose $y \in \mathbb{R}^m$ into ℓ subvectors

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_\ell \end{bmatrix}, \quad y_j \in \mathbb{R}^{k_j}, j = 1, \dots, \ell.$$

The original equation can be given as the system

$$A_j x = y_j, \quad j = 1, \dots, \ell$$

The j th problem is composed of $k_j \leq n$ linearly indep. linear eq., and the solution space

$$X_j = \{x \in \mathbb{R}^n : A_j x = y_j\}$$

is an $(n - k_j)$ dimensional hyperplane in $\mathbb{R}^n$. This hyperplane is a subspace iff $y_j = 0$.



- define an orthogonal projection $P_j : \mathbb{R}^n \rightarrow X_j$ by requiring

$$P_j z \in X_j \quad \text{and} \quad (I - P_j)z \perp (w_1 - w_2), \quad \forall z \in \mathbb{R}^n, w_1, w_2 \in X_j$$

$P_j z$ is the point closest to z in X_j .

- define **sequential** proj. $P : \mathbb{R}^n \rightarrow \mathbb{R}^n$ by

$$P = P_\ell P_{\ell-1} \dots P_2 P_1$$

- The Kaczmarz sequence $\{x_k\}_{k=0}^\infty$ is defined by Stephan Kaczmarz 1937

$$x_{k+1} = P x_k, \quad x_0 = 0$$



Theorem

If $X = \cap_{j=1}^{\ell} X_j \neq \emptyset$, then the Kaczmarz sequence $\{x_k\}_{k=0}^{\infty}$ converges to the minimum norm solution $x^{\dagger} = A^{\dagger}y$ as $k \rightarrow \infty$, i.e.

$$\lim_{k \rightarrow \infty} x_k = x^{\dagger}.$$

The proof is quite lengthy



Algebraic reconstruction technique (ART)

Consider the special case where the original problem $Ax = y$, $A \in \mathbb{R}^{m \times n}$, is partitioned into m subproblems, i.e.,

$$A_j x = a_j^\top x = y_j, \quad j = 1, \dots, m,$$

- $a_j^\top$: the j th row of A .
- $A_j : \mathbb{R}^n \rightarrow \mathbb{R}$ is a surjection $\forall j \Leftrightarrow A$ has no empty rows

This version of Kaczmarz method is called algebraic reconstruction technique (ART). ART is used extensively in X-ray tomography.

Richard Gordon, Robert Bender, Gabor Herman 1970



ART iteration

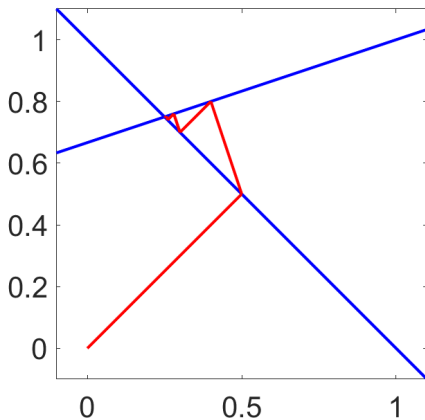
$$A = \begin{bmatrix} 1 & 1 \\ -1 & 3 \end{bmatrix}, \quad y = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

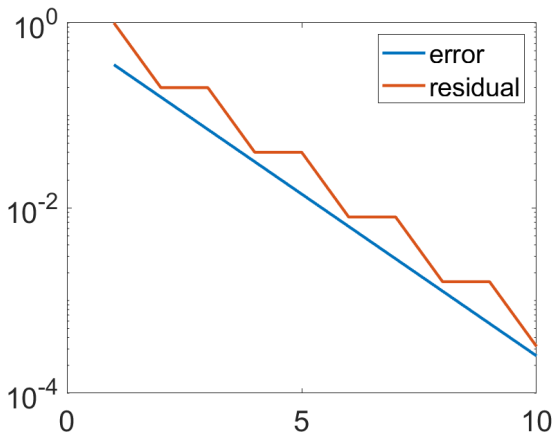
A is invertible and the hyperplanes in $\mathbb{R}^2$ are given by

$$X_1 = \{(x_1, x_2) : x_1 + x_2 = 1\}$$

$$X_2 = \{(x_1, x_2) : -x_1 + 3x_2 = 2\}$$

ART converges to the unique solution $x = (\frac{1}{4}, \frac{3}{4})$. We visualize each projection P_j , $j = 1, 2$, not just the sequential projection P .





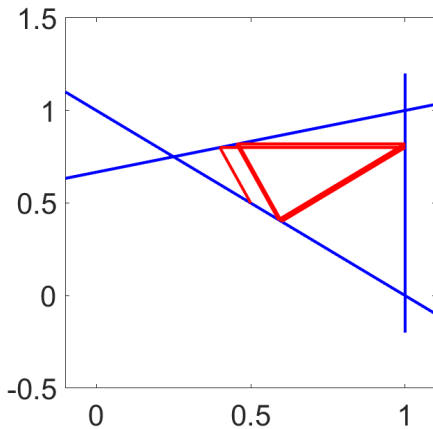


ART iteration

$$A = \begin{bmatrix} 1 & 1 \\ -1 & 3 \\ 1 & 0 \end{bmatrix}, \quad y = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}$$

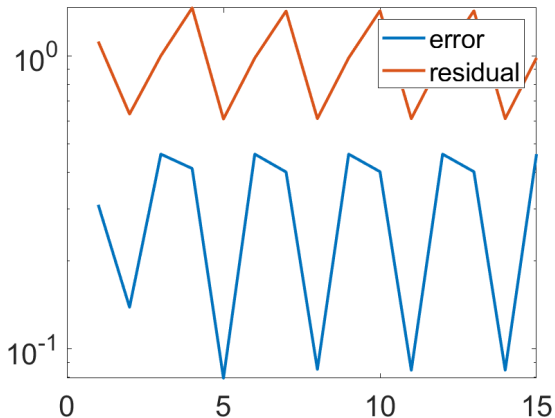
added the third hyperplane $X_3 = \{x_1 = 1\}$

the linear system is inconsistent and does not have a solution. ART appears convergent to a solution.





error with respect to the least-squares solution ...





ART iteration

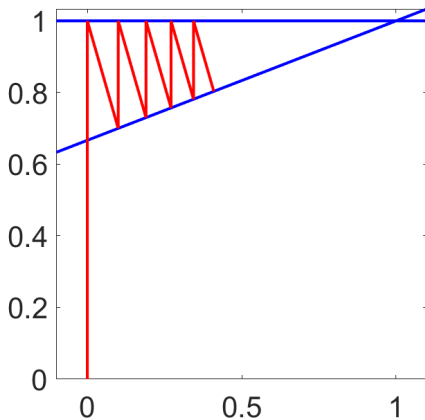
$$A = \begin{bmatrix} 0 & 1 \\ -1 & 3 \end{bmatrix}, \quad y = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

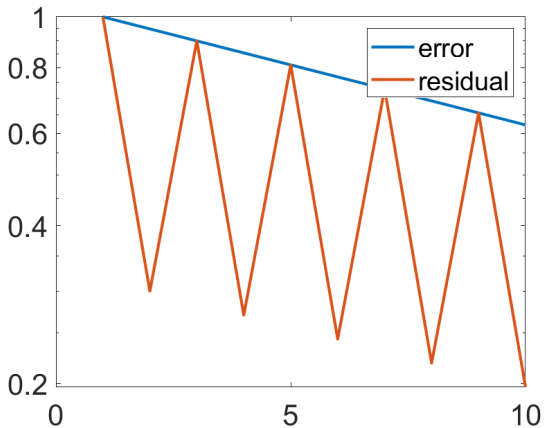
A is invertible and the hyperplanes in $\mathbb{R}^2$ are given by

$$X_1 = \{(x_1, x_2) : x_2 = 1\}$$

$$X_2 = \{(x_1, x_2) : -x_1 + 3x_2 = 2\}$$

ART converges to the unique solution $x = (1, 1)$, but extremely slowly







The computation of the projection P_j :

$$A = \begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_\ell \end{bmatrix}, \quad A_j \in \mathbb{R}^{k_j \times n}, \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_\ell \end{bmatrix}, \quad y_j \in \mathbb{R}_{k_j}, \quad j = 1, \dots, \ell,$$

with each A_j being surjective

- X_j : the hyperplane composed of the solutions to $A_j x = y_j$
- $P_j : \mathbb{R}^n \rightarrow X_j$: orthogonal projection onto X_j
- $Q_j : \mathbb{R}^n \rightarrow \ker(A_j)$: orthogonal projection onto the kernel of A_j
 $\Rightarrow I - Q_j : \mathbb{R}^n \rightarrow \ker(A_j)^\perp = \text{range}(A_j^\top)$: orthogonal projection
- the identity

$$P_j x = z + Q_j(x - z), \quad \forall x \in \mathbb{R}^n, z \in X_j$$

This formula is indep. of the choice of z



the proof of the identity

$$P_j x = z + Q_j(x - z), \quad \forall x \in \mathbb{R}^n, z \in X_j$$

- for any $z \in X_j$, $P_j x \in X_j$

$$A_j P_j x = A_j z + A_j Q_j(x - z) = y_j$$

- for any $z_1, z_2 \in X_j$, $z_1 - z_2 \in \ker(A_j) + (I - Q_j) : \mathbb{R}^n \rightarrow \ker(A_j)^\perp$

$$(x - P_j x, z_1 - z_2) = ((I - Q_j)(x - z), z_1 - z_2) = 0$$



Lemma

The projection P_j can be written explicitly as

$$P_j x = x + A_j^\top (A_j A_j^\top)^{-1} (y_j - A_j x), \quad \forall x \in \mathbb{R}^n.$$



- $A_j A_j^\top \in \mathbb{R}^{k_j \times k_j}$ is invertible: $A_j : \mathbb{R}^n \rightarrow \mathbb{R}^{k_j}$ surjective $\Rightarrow A_j^\top$ injective $\Rightarrow A_j A_j^\top$ injective:

$$A_j A_j^\top z = 0 \Rightarrow z^\top A_j A_j^\top z = 0 \Rightarrow \|A_j^\top z\|^2 = 0 \Rightarrow z = 0$$

$\Rightarrow$ injective square matrix $A_j A_j^\top$ is invertible.

- Fix any $x \in \mathbb{R}^n$, let

$$P_j x = z + Q_j(x - z), \quad \text{with } z \in X_j$$

$\Rightarrow$

$$\begin{aligned} x - P_j x &= (x - z) - Q_j(x - z) \\ &= (I - Q_j)(x - z) \in \ker(A_j)^\perp = \text{range}(A_j^\top) \end{aligned}$$



- there exists $w \in \mathbb{R}^{k_j}$ s.t.

$$A_j^\top w = x - P_j x$$

- $+ P_j x \in X_j \Rightarrow$

$$A_j A_j^\top w = A_j x - A P_j x = A_j x - y_j$$

- Solving for w and substitution:

$$A_j^\top (A_j A_j^\top)^{-1} (A_j x - y_j) = x - P_j x.$$



implementation of ART:

- when the submatrices $A_j = a_j^\top$, $j = 1, \dots, m$, are rows of the original system matrix A , y_j , $j = 1, \dots, m$, are the components of y , the inverse is just real numbers

$$(A_j A_j^\top)^{-1} = (a_j^\top a_j)^{-1} = 1 / \|a_j\|^2$$

- update

$$P_j x = x + (y_j - (a_j, x)) \frac{a_j}{\|a_j\|^2}$$



discrepancy principle for Kaczmarz method

- the measurement y^δ is a noisy version of exact data $y^\dagger$ with

$$\|y^\delta - y^\dagger\| = \delta > 0$$

- The Morozov's discrepancy principle works for the Kaczmarz iteration as follows: choose the smallest k s.t.

$$\|y^\delta - Ax_k^\delta\| \leq \delta$$

if such k exists.



remark

- Unlike truncated SVD and Landweber method, the condition

$$\delta > \|y^\delta - Py^\delta\|$$

where P is the projection onto the range of A , is insufficient to guarantee the existence of a stopping index k

- The evaluation of the full residual is costly ...

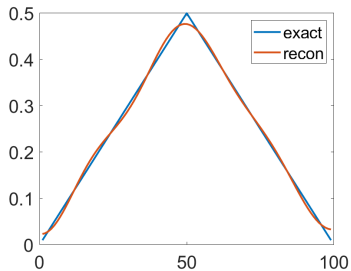
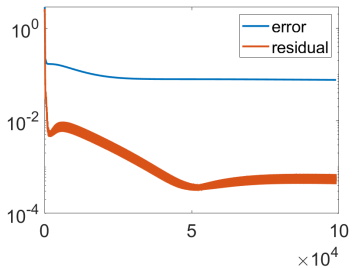


Example: discretized inverse heat conduction

- simulate the data, add some noise, same value δ
- use discrepancy principle for early stopping
-

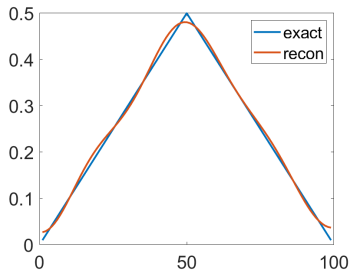
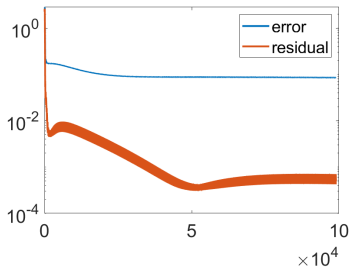


heat example with exact data





heat example with 1% noise





Randomized Kaczmarz method

T. Strohmer, R. Vershynin J. Fourier Anal. Appl. 2009: randomized Kaczmarz method

- choose the i th equation with probability prop. to $\|a_i\|^2$
- the method is a version of stochastic gradient descent



Let $x^\dagger$ be a solution to $Ax = y$. Then randomized Kaczmarz converges to x in expectation:

$$\mathbb{E}[\|x_k - x^\dagger\|^2] \leq (1 - \kappa(A)^{-2})^k \|x_0 - x^\dagger\|^2$$

with

$$\kappa(A) = \|A\|_F \|A^{-1}\|$$

The method **converges exponentially fast** to $x^\dagger$, and the convergence rate depends only on the scaled condition number $\kappa(A)$



- there holds the inequality

$$\sum_{j=1}^m |(z, a_j)|^2 \geq \frac{\|z\|^2}{\|A^{-1}\|^2}, \quad \forall z \in \mathbb{R}^n$$

- the identity $\|A\|_F^2 = \sum_{j=1}^m \|a_j\|^2 + \kappa(A) = \|A\|_F \|A^{-1}\| \Rightarrow$

$$\sum_{j=1}^m \frac{\|a_j\|^2}{\|A\|_F^2} \left(z, \frac{a_j}{\|a_j\|} \right)^2 \geq \kappa(A)^{-2} \|z\|^2$$

LHS is the expectation of some random variable.

- The solution space of j th equation is $\{x : (x, a_j) = y_j\}$, with normal $\frac{a_j}{\|a_j\|}$



- define a random variable z whose values are these normals:

$$Z = \frac{a_j}{\|a_j\|} \quad \text{with probability } \frac{\|a_j\|^2}{\|A\|_F^2}, \quad j = 1, \dots, m$$

- there holds the **fundamental inequality**

$$\mathbb{E}[(z, Z)^2] \geq \kappa(A)^{-2} \|z\|^2, \quad \forall z \in \mathbb{R}^n$$

- The orthogonal projection P onto the sol. space of a random eq. of $Ax = y$ is given by

$$Pz = z - (z - x^\dagger, Z)Z$$



- the approx. x_k is computed from x_{k-1} via $x_k = P_k x_{k-1}$, with $P_1, P_2, \dots$ are indep. realization of the random projection P
- The vector $x_k - x_{k-1}$ in the kernel of P_{i_k} : orthogonal to the sol. space of $(a_{i_k}, x) = y_{i_k}$ onto which projects $\Rightarrow$

$$\|x_k - x^\dagger\|^2 = \|x_{k-1} - x^\dagger\|^2 - \|x_{k-1} - x_k\|^2$$

(Pythagorean theorem)

- by the definition of x_k

$$x_{k-1} - x_k = (x_{k-1} - x^\dagger, Z_k) Z_k$$

Z_k : indep. realizations of the random vector Z

- Z_k unit norm $\Rightarrow$ error reduction

$$\|x_k - x^\dagger\|^2 = \left(1 - \left(\frac{x_{k-1} - x^\dagger}{\|x_{k-1} - x^\dagger\|}, Z_k\right)^2\right) \|x_{k-1} - x^\dagger\|^2$$



- taking conditional expectation on both sides

$$\mathbb{E}_k[\|x_k - x^\dagger\|^2] = (1 - \mathbb{E}\left(\frac{x_{k-1} - x^\dagger}{\|x_{k-1} - x^\dagger\|}, Z_k\right)^2) \|x_{k-1} - x^\dagger\|^2$$

- independence + the fundamental inequality

$$\mathbb{E}_k[\|x_k - x^\dagger\|^2] \leq (1 - \kappa(A)^{-2}) \|x_{k-1} - x^\dagger\|^2$$

- taking full conditional yields

$$\mathbb{E}[\|x_k - x^\dagger\|^2] \leq (1 - \kappa(A)^{-2}) \mathbb{E}[\|x_{k-1} - x^\dagger\|^2] \leq (1 - \kappa(A)^{-2})^k \|x_0 - x^\dagger\|^2$$



convergence analysis via SVD

■ error $e_k = x_k - x^*$

■ error recursion

$$e_{k+1} = \left(I - \frac{a_{i_k} a_{i_k}^t}{\|a_{i_k}\|^2} \right) e_k. \quad (1)$$

$I - \frac{a_{i_k} a_{i_k}^t}{\|a_{i_k}\|^2}$ is an orthogonal projection operator

■ $\sigma_{\min} \|e\| \leq \|Ae\| \leq \sigma_1 \|e\|$

$$\begin{aligned}\|e_{k+1}\|^2 &= \|e_k\|^2 - \frac{2}{\|a_{i_k}\|^2} \langle a_{i_k}, e_k \rangle \langle a_{i_k}, e_k \rangle + \langle a_{i_k}, e_k \rangle^2 \frac{\|a_{i_k}\|^2}{\|a_{i_k}\|^4} \\ &= \|e_k\|^2 - \frac{2}{\|a_{i_k}\|^2} \langle a_{i_k}, e_k \rangle \langle a_{i_k}, e_k \rangle + \langle a_{i_k}, e_k \rangle^2 \frac{1}{\|a_{i_k}\|^2}\end{aligned}$$

Upon noting the identity $\sum_{i=1}^n a_i a_i^t = A^t A$, taking expectation

$$\begin{aligned}\mathbb{E}[\|e_{k+1}\|^2 | e_k] &\leq \|e_k\|^2 - \frac{2}{\|A\|_F^2} \langle e_k, A^t A e_k \rangle + \frac{\|A e_k\|^2}{\|A\|_F^2} = \|e_k\|^2 - \frac{\|A e_k\|^2}{\|A\|_F^2} \\ &+ \|A e\| \geq \sigma_{\min} \|e\| \Rightarrow\end{aligned}$$

$$\mathbb{E}[\|e_{k+1}\|^2 | e_k] \leq \|e_k\|^2 - \frac{\sigma_{\min}^2 \|e_k\|^2}{\|A\|_F^2} = (1 - \kappa(A)^{-2}) \|e_k\|^2$$



Theorem

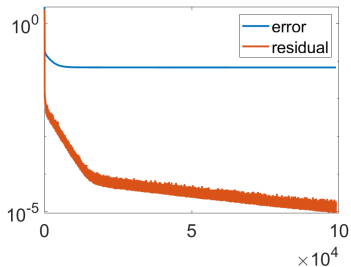
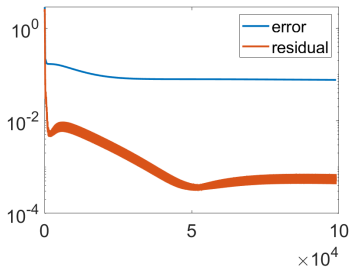
Let $c_1 = \frac{\sigma_L^2}{\|A\|_F^2}$ and $c_2 = \frac{\sum_{i=L+1}^r \sigma_i^2}{\|A\|_F^2}$. Then there hold

$$\mathbb{E}[\|P_L e_{k+1}\|^2 | e_k] \leq (1 - c_1) \|P_L e_k\|^2 + c_2 \|P_H e_k\|^2,$$

$$\mathbb{E}[\|P_H e_{k+1}\|^2 | e_k] \leq c_2 \|P_L e_k\|^2 + (1 + c_2) \|P_H e_k\|^2.$$



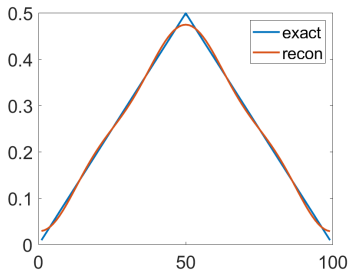
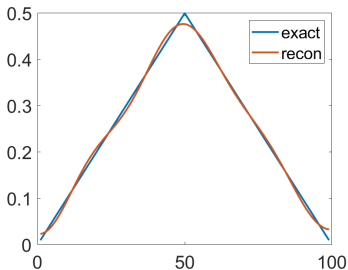
heat example with exact data



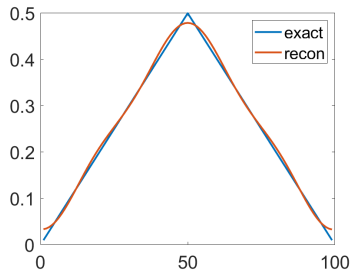
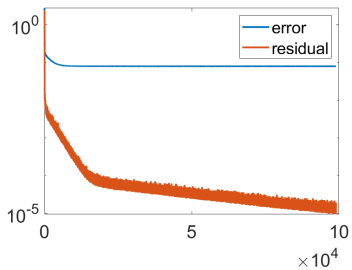
left Kaczmarz, and right randomized Kaczmarz



heat example with exact data



left Kaczmarz, and right randomized Kaczmarz





randomized Kaczmarz as SGD

solving the problem by approximately minimizing

$$J(x) = (2n)^{-1} \|Ax - y\|^2 = \frac{1}{n} \sum_{j=1}^n f_j(x), \quad f_j(x) = \frac{1}{2} ((a_j, x) - y_j)^2.$$

many different approaches:

- gradient descent (a.k.a. Landweber iteration)

$$x_{k+1} = x_k - \eta_k n^{-1} A^t (Ax_k - y)$$

acceleration by step-size, Nesterov trick, momentum, ...

- **stochastic** gradient descent Robbins-Monro 1950s $f_j = \frac{1}{2} ((a_{i_k}, x) - b_{i_k})^2$

$$x_{k+1} = x_k - \eta_k ((a_{i_k}, x_k) - b_{i_k}) a_{i_k}$$

index i_k uniformly over the set $\{1, \dots, n\}$



RKM and SGD: the linear equation $Ax = b$ in least squares form:

$$\min_{x \in \mathbb{R}^n} \frac{1}{2n} \sum_{i=1}^n |(a_i, x) - y_i|^2 \triangleq f(x) = \sum_{i=1}^n \frac{\|a_i\|^2}{\|A\|_F^2} f_i(x),$$

with

$$f_i(x) = \frac{\|A\|_F^2}{2n\|a_i\|^2} |(a_i, x) - y_i|^2$$

and sampling probability of drawing i th row being

$$p_i = \frac{\|a_i\|^2}{\|A\|_F^2}.$$

this is a valid probability distribution

$$\sum_{i=1}^n p_i = \sum_{i=1}^n \frac{\|a_i\|^2}{\|A\|_F^2} = 1$$



Needell et al 2016; Jiao-Jin-Lu 2017 Inverse Problems

RKM is a (weighted) SGD with a constant stepsize $\beta = n/\|A\|_F^2$.

$$\begin{aligned}x_{k+1} &= x_k - \beta \partial_x f_{i_k}(x_k) \\&= x_k - \frac{n}{\|A\|_F^2} \cdot \frac{\|A\|_F^2}{n\|a_{i_k}\|^2} ((a_{i_k}, x) - y_{i_k}) a_{i_k} \\&= x_k - ((a_{i_k}, x) - y_{i_k}) \frac{a_{i_k}}{\|a_{i_k}\|^2}\end{aligned}$$



The SGD has a long history in statistics

- Robbins-Monro 1950s, a few monographs Kushner-Yin 2003
various asymptotic distributional results ...
- revived interest in machine learning and signal processing
- accelerated variants represent the state of art solvers for large-scale problems: machine learning, inverse problems etc.



Asymptotic convergence

Question: Is SGD consistent (as $\delta \rightarrow 0$) ?

This is one of the basic requirements of a reconstruction scheme ...
since x_k^δ is random, which norm do we use ?

- strong convergence

$$\lim_{\delta \rightarrow 0} \mathbb{E}[\|x_{k(\delta)}^\delta - x^\dagger\|^2] = 0 \quad ???$$

- weak convergence

$$\lim_{\delta \rightarrow 0} \|\mathbb{E}[x_{k(\delta)}^\delta] - x^\dagger\|^2 = 0 \quad ???$$

- high-probability / almost sure ...



Question:

$$\lim_{\delta \rightarrow 0} \mathbb{E}[\|x_{k(\delta)}^\delta - x^\dagger\|^2] = 0 \quad ???$$

Answer: **Yes** for the step size choice:

$$\eta_j = c_0 j^{-\alpha}, \quad \alpha \in (0, 1)$$

if

$$\lim_{\delta \rightarrow 0} k(\delta) = \infty \quad \text{and} \quad \lim_{\delta \rightarrow 0} k(\delta)^{\frac{1-\alpha}{2}} \delta = 0.$$



basic strategy of proof: by bias-variance decomposition

$$\begin{aligned}\mathbb{E}[\|x_k^\delta - x^\dagger\|^2] &= \underbrace{\|\mathbb{E}[x_k^\delta] - x^\dagger\|^2}_{\text{mean}} + \underbrace{\mathbb{E}[\|\mathbb{E}[x_k^\delta] - x_k^\delta\|^2]}_{\text{variance}} \\ &\leq 2 \underbrace{\|\mathbb{E}[x_k^\delta - x_k]\|^2}_{\text{propagation error}} + 2 \underbrace{\|\mathbb{E}[x_k] - x^\dagger\|^2}_{\text{approximation error}} + \underbrace{\mathbb{E}[\|\mathbb{E}[x_k^\delta] - x_k^\delta\|^2]}_{\text{stochastic error}}\end{aligned}$$

- propagation error and approximation error are well understood (i.e., Landweber method)
- stochastic error needs to be controlled ...



how to control the stochastic error

$$\mathbb{E}[\|x_{k+1}^\delta - \mathbb{E}[x_{k+1}^\delta]\|^2] \leq \sum_{j=1}^k \eta_j^2 \|B^{\frac{1}{2}} \Pi_{j+1}^k(B)\|^2 \mathbb{E}[\|Ax_j^\delta - y^\delta\|^2]$$

$\Rightarrow$ bound the expected residual properly: for any $\ell \geq 2$

$$\mathbb{E}[\|Ax_{k+1}^\delta - y^\delta\|^2] \leq ck^{-\min(\ell\alpha, 2(1-\alpha), (2p+1)(1-\alpha))} \ln^\ell k + c'\delta^2 \ln^2 k.$$

with the parameter p (source condition)

$$B^p w = x^\dagger - x_1$$

regularity condition on the initial error: the larger is p , the smoother is the solution ...



limiting cases:

- no solution regularity $p = 0$

$$\mathbb{E}[\|Ax_{k+1}^\delta - b^\delta\|^2] \leq ck^{-\min(\ell\alpha, 1-\alpha)} \ln^\ell k + c'\delta^2 \ln^2 k.$$

- very good regularity

$$\mathbb{E}[\|Ax_{k+1}^\delta - b^\delta\|^2] \leq ck^{-\min(\ell\alpha, 2(1-\alpha))} \ln^\ell k + c'\delta^2 \ln^2 k.$$

$\Rightarrow$ randomness **restricts the best possible decay** of the residual error estimates

$$\mathbb{E}[\|x_k^\delta - x^\dagger\|^2] \leq ck^{-\min(\ell\alpha, 1-\alpha, 2p(1-\alpha))} \ln^\ell k + c\delta^2 k^{1-\alpha}.$$

by choosing proper $k(\delta) \Rightarrow$ error estimates

SGD is consistent with convergence rate (if we know how to stop)!

challenge: Nobody knows how to stop !



what is beyond ?

- SGD / randomized Kaczmarz is regularizing in the mean squares error sense.
- stopping criterion is not well understood
- convergence in other modes (a.s., high-probability) not well understood