SGD and Randomized projection algorithms for overdetermined linear systems

Deanna Needell

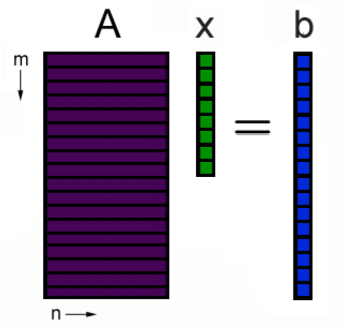
Claremont McKenna College

IPAM, Feb. 25, 2014

Includes joint work with Eldar, Ward, Tropp, Srebro-Ward

Setup

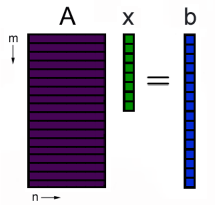
Let $Ax = b$ be an *overdetermined*, full rank system of equations.



Setup

Setup

Let $Ax = b$ be an *overdetermined*, full rank system of equations.



Goal

From A and b we wish to recover unknown x. Assume $m \gg n$.

This talk:

- Accelerate Kaczmarz method via dimension reduction [Eldar-N, 2011]
- Accelerate via optimal relaxation [N-Ward, 2013]
- Accelerate via blocking and pavings [N-Tropp, 2013]
- Partially weighted sampling via SGD analysis [N-Sbrero-Ward, 2014]

This talk:

- Accelerate Kaczmarz method via dimension reduction [Eldar-N, 2011]
- Accelerate via optimal relaxation [N-Ward, 2013]
- Accelerate via blocking and pavings [N-Tropp, 2013]
- Partially weighted sampling via SGD analysis [N-Sbrero-Ward, 2014]

This talk:

- Accelerate Kaczmarz method via dimension reduction [Eldar-N, 2011]
- Accelerate via optimal relaxation [N-Ward, 2013]
- Accelerate via blocking and pavings [N-Tropp, 2013]
- Partially weighted sampling via SGD analysis [N-Sbrero-Ward, 2014]

This talk:

- Accelerate Kaczmarz method via dimension reduction [Eldar-N, 2011]
- Accelerate via optimal relaxation [N-Ward, 2013]
- Accelerate via blocking and pavings [N-Tropp, 2013]
- Partially weighted sampling via SGD analysis [N-Sbrero-Ward, 2014]

Method

$$\begin{bmatrix} \text{———} \; a_1 \; \text{———} \\ \text{———} \; a_2 \; \text{———} \\ \vdots \quad \vdots \quad \ddots \quad \vdots \\ \text{———} \; a_m \; \text{———} \end{bmatrix} \cdot \begin{bmatrix} x \\ \\ \\ \end{bmatrix} = \begin{bmatrix} b[1] \\ b[2] \\ \cdot \\ \cdot \\ \cdot \\ b[m] \end{bmatrix}$$

Kaczmarz

1. Start with initial guess x_0
2. $x_{k+1} = x_k + (b[i] - \langle a_i, x_k \rangle) a_i$ where $i = (k \mod m) + 1$
3. Repeat (2)

Method

$$\begin{bmatrix} \text{——} & a_1 & \text{——} \\ \text{——} & a_2 & \text{——} \\ \vdots & \vdots & \ddots & \vdots \\ \text{——} & a_m & \text{——} \end{bmatrix} \cdot \begin{bmatrix} x \\ \\ \end{bmatrix} = \begin{bmatrix} b[1] \\ b[2] \\ \cdot \\ \cdot \\ \cdot \\ b[m] \end{bmatrix}$$

Kaczmarz

1. Start with initial guess x_0
2. $x_{k+1} = x_k + (b[i] - \langle a_i, x_k \rangle) a_i$ where $i = (k \mod m) + 1$
3. Repeat (2)

Method

$$\begin{bmatrix} \text{———} \; a_1 \; \text{———} \\ \text{———} \; a_2 \; \text{———} \\ \vdots \quad \vdots \quad \ddots \quad \vdots \\ \text{———} \; a_m \; \text{———} \end{bmatrix} \cdot \begin{bmatrix} x \\ \\ \\ \end{bmatrix} = \begin{bmatrix} b[1] \\ b[2] \\ \cdot \\ \cdot \\ \cdot \\ b[m] \end{bmatrix}$$

Kaczmarz

1. Start with initial guess x_0
2. $x_{k+1} = x_k + (b[i] - \langle a_i, x_k \rangle) a_i$ where $i = (k \bmod m) + 1$
3. Repeat (2)

Method

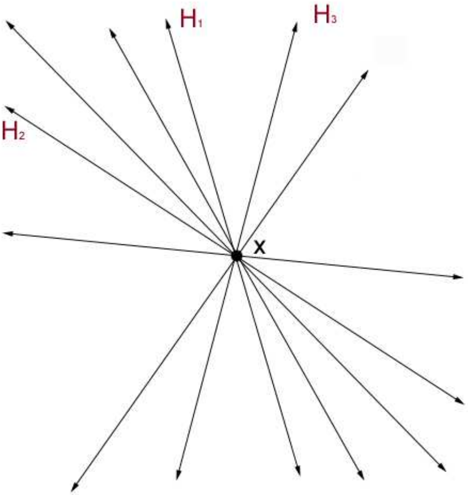
$$\begin{bmatrix} \text{——} \; a_1 \; \text{——} \\ \text{——} \; a_2 \; \text{——} \\ \vdots \quad \vdots \quad \ddots \quad \vdots \\ \text{——} \; a_m \; \text{——} \end{bmatrix} \cdot \begin{bmatrix} x \\ \\ \end{bmatrix} = \begin{bmatrix} b[1] \\ b[2] \\ \cdot \\ \cdot \\ \cdot \\ b[m] \end{bmatrix}$$

Kaczmarz

1. Start with initial guess x_0
2. $x_{k+1} = x_k + (b[i] - \langle a_i, x_k \rangle) a_i$ where $i = (k \bmod m) + 1$
3. Repeat (2)

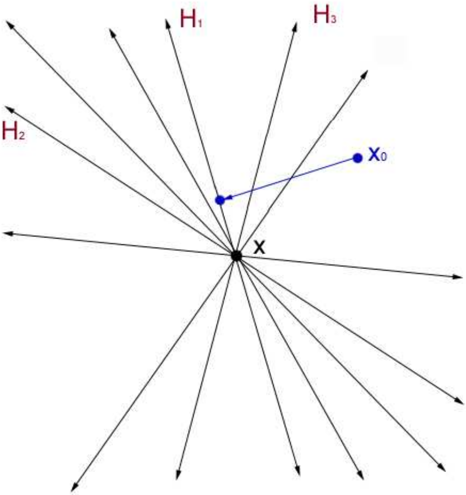
Geometrically

Denote $H_i = \{w : \langle a_i, w \rangle = b[i]\}$.



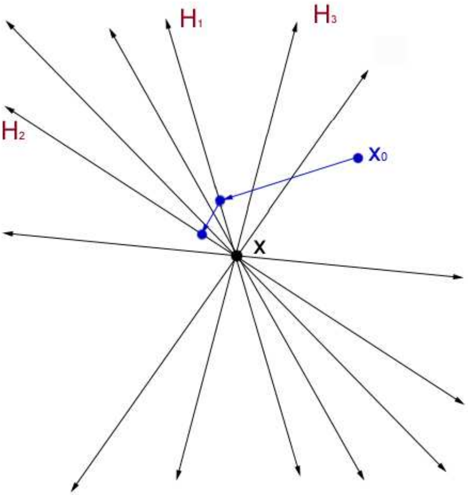
Geometrically

Denote $H_i = \{w : \langle a_i, w\rangle = b[i]\}$.



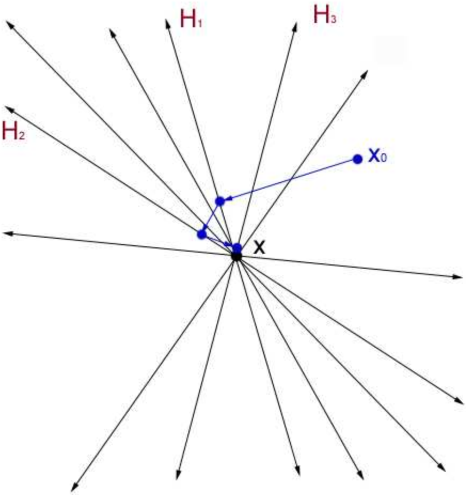
Geometrically

Denote $H_i = \{w : \langle a_i, w \rangle = b[i]\}$.



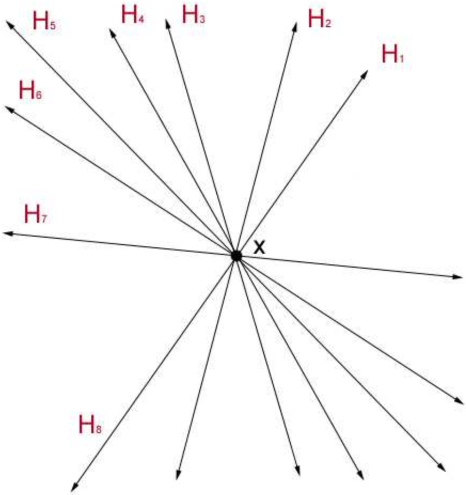
Geometrically

Denote $H_i = \{w : \langle a_i, w \rangle = b[i]\}$.



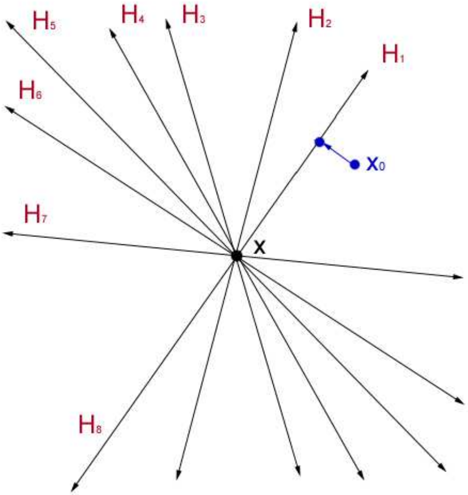
But what if...

Denote $H_i = \{w : \langle a_i, w \rangle = b[i]\}$.



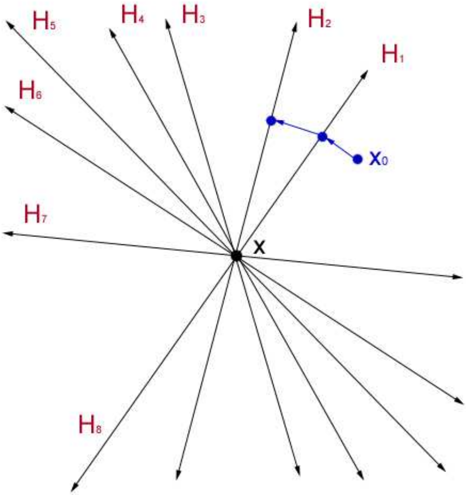
But what if...

Denote $H_i = \{w : \langle a_i, w \rangle = b[i]\}$.



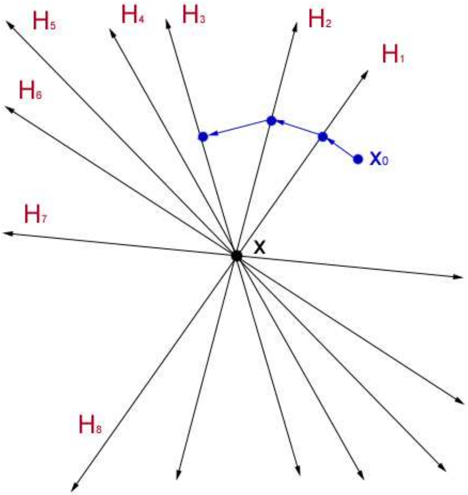
But what if...

Denote $H_i = \{w : \langle a_i, w \rangle = b[i]\}$.



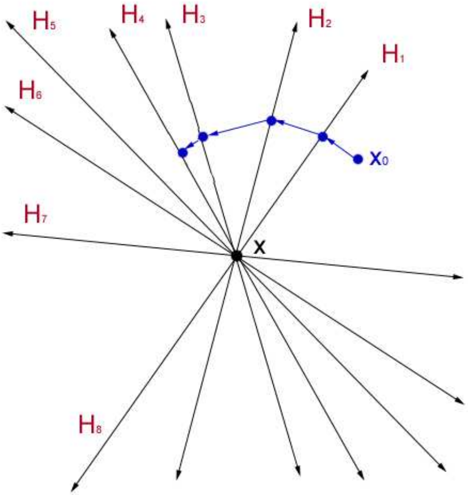
But what if...

Denote $H_i = \{w : \langle a_i, w \rangle = b[i]\}$.



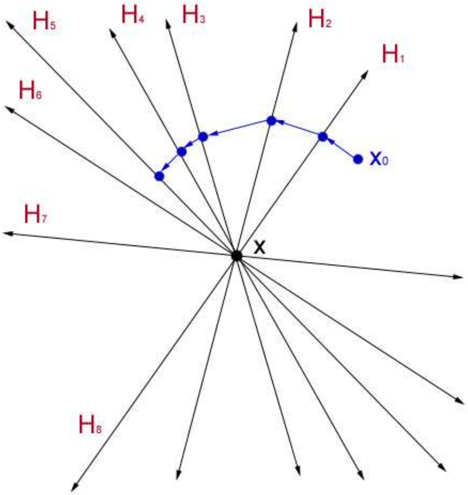
But what if...

Denote $H_i = \{w : \langle a_i, w\rangle = b[i]\}$.



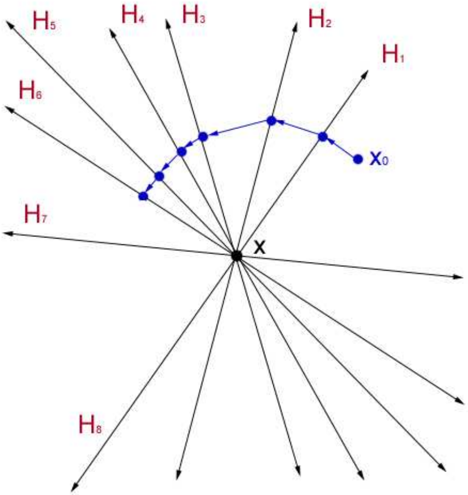
But what if...

Denote $H_i = \{w : \langle a_i, w\rangle = b[i]\}$.



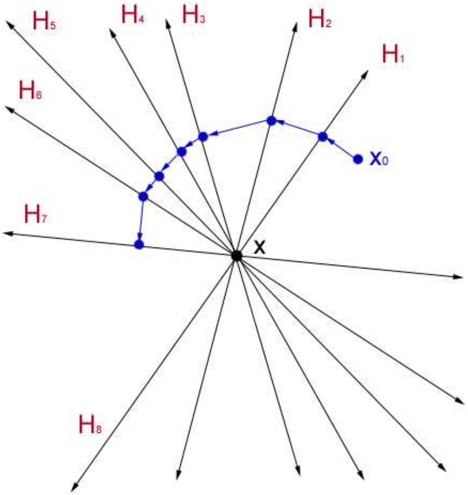
But what if...

Denote $H_i = \{w : \langle a_i, w \rangle = b[i]\}$.



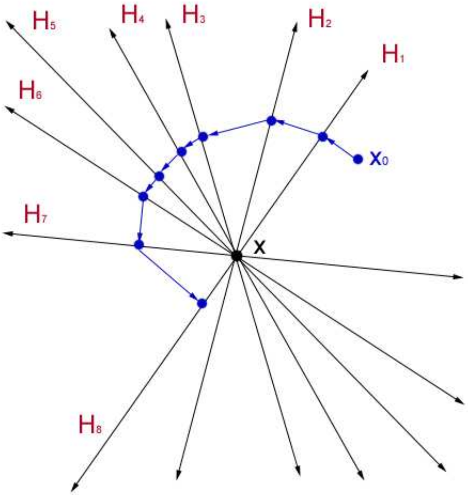
But what if...

Denote $H_i = \{w : \langle a_i, w \rangle = b[i]\}$.



But what if...

Denote $H_i = \{w : \langle a_i, w \rangle = b[i]\}$.



Randomized Kaczmarz

$$\begin{bmatrix} \text{———} \; a_1 \; \text{———} \\ \text{———} \; a_2 \; \text{———} \\ \vdots \quad \vdots \quad \ddots \quad \vdots \\ \text{———} \; a_m \; \text{———} \end{bmatrix} \cdot \begin{bmatrix} x \\ \\ \\ \end{bmatrix} = \begin{bmatrix} b[1] \\ b[2] \\ \cdot \\ \cdot \\ \cdot \\ b[m] \end{bmatrix}$$

Kaczmarz

1. Start with initial guess x_0
2. $x_{k+1} = x_k + (b[i] - \langle a_i, x_k \rangle) a_i$ where i is chosen *randomly*
3. Repeat (2)

Randomized Kaczmarz

$$\begin{bmatrix} \text{———} & a_1 & \text{———} \\ \text{———} & a_2 & \text{———} \\ \vdots & \vdots & \ddots & \vdots \\ \text{———} & a_m & \text{———} \end{bmatrix} \cdot \begin{bmatrix} x \\ \\ \\ \end{bmatrix} = \begin{bmatrix} b[1] \\ b[2] \\ \cdot \\ \cdot \\ \cdot \\ b[m] \end{bmatrix}$$

Kaczmarz

1. Start with initial guess x_0
2. $x_{k+1} = x_k + (b[i] - \langle a_i, x_k \rangle) a_i$ where i is chosen *randomly*
3. Repeat (2)

Randomized Kaczmarz

Theorem [Strohmer-Vershynin]: Consistent case $Ax = b$

1. Start with initial guess x_0
2. $x_{k+1} = x_k + (b_p - \langle a_p, x_k \rangle) a_p$ where $\mathbb{P}(p = i) = \frac{\|a_i\|_2^2}{\|A\|_F^2}$
3. Repeat (2)

Randomized Kaczmarz

Theorem [Strohmer-Vershynin]: Consistent case $Ax = b$

1. Start with initial guess x_0
2. $x_{k+1} = x_k + (b_p - \langle a_p, x_k \rangle) a_p$ where $\mathbb{P}(p = i) = \frac{\|a_i\|_2^2}{\|A\|_F^2}$
3. Repeat (2)

Randomized Kaczmarz (RK)

Theorem [Strohmer-Vershynin]

- Let $R = \|A\|_f^2 \|A^{-1}\|^2$
- Then $\mathbb{E}\|x_k - x\|_2^2 \leq \left(1 - \frac{1}{R}\right)^k \|x_0 - x\|_2^2$
- Well conditioned $A \rightarrow$ Convergence in $\mathrm{O}(n)$ iterations $\rightarrow$ $\mathrm{O}(n^2)$ total runtime.
- Better than $\mathrm{O}(mn^2)$ runtime for Gaussian elimination and empirically often faster than Conjugate Gradient.

Randomized Kaczmarz (RK)

Theorem [Strohmer-Vershynin]

- Let $R = \|A\|_f^2 \|A^{-1}\|^2$
- Then $\mathbb{E}\|x_k - x\|_2^2 \leq \left(1 - \frac{1}{R}\right)^k \|x_0 - x\|_2^2$
- Well conditioned $A \rightarrow$ Convergence in $\mathrm{O}(n)$ iterations $\rightarrow$ $\mathrm{O}(n^2)$ total runtime.
- Better than $\mathrm{O}(mn^2)$ runtime for Gaussian elimination and empirically often faster than Conjugate Gradient.

Theorem [Strohmer-Vershynin]

- Let $R = \|A\|_f^2 \|A^{-1}\|^2$
- Then $\mathbb{E}\|x_k - x\|_2^2 \leq \left(1 - \frac{1}{R}\right)^k \|x_0 - x\|_2^2$
- Well conditioned $A \rightarrow$ Convergence in $\mathrm{O}(n)$ iterations $\rightarrow$ $\mathrm{O}(n^2)$ total runtime.
- Better than $\mathrm{O}(mn^2)$ runtime for Gaussian elimination and empirically often faster than Conjugate Gradient.

Randomized Kaczmarz (RK)

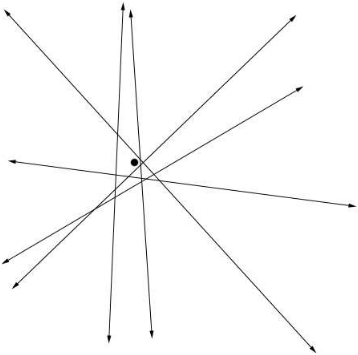
Theorem [Strohmer-Vershynin]

- Let $R = \|A\|_f^2 \|A^{-1}\|^2$
- Then $\mathbb{E}\|x_k - x\|_2^2 \leq \left(1 - \frac{1}{R}\right)^k \|x_0 - x\|_2^2$
- Well conditioned $A \to$ Convergence in $\mathrm{O}(n)$ iterations $\to$ $\mathrm{O}(n^2)$ total runtime.
- Better than $\mathrm{O}(mn^2)$ runtime for Gaussian elimination and empirically often faster than Conjugate Gradient.

Randomized Kaczmarz (RK) with noise

Inconsistent systems

We now consider the system $Ax = b + e$.



Randomized Kaczmarz (RK) with noise

Theorem [N]

- Let $Ax = b + e$. Then

$$\mathbb{E}\|x_k - x\|_2 \leq \left(1 - \frac{1}{R}\right)^{k/2} \|x_0 - x\|_2 + \sqrt{R}\|e\|_\infty$$

.

- This bound is sharp and attained in simple examples.
- Note can set $e = Ax^\star - b$ where $x^\star$ is LS solution.

Randomized Kaczmarz (RK) with noise

Theorem [N]

- Let $Ax = b + e$. Then

$$\mathbb{E}\|x_k - x\|_2 \le \left(1 - \frac{1}{R}\right)^{k/2} \|x_0 - x\|_2 + \sqrt{R}\|e\|_\infty$$

.

- This bound is sharp and attained in simple examples.
- Note can set $e = Ax^\star - b$ where $x^\star$ is LS solution.

Randomized Kaczmarz (RK) with noise

Theorem [N]

- Let $Ax = b + e$. Then

$$\mathbb{E}\|x_k - x\|_2 \leq \left(1 - \frac{1}{R}\right)^{k/2} \|x_0 - x\|_2 + \sqrt{R}\|e\|_\infty$$

.

- This bound is sharp and attained in simple examples.
- Note can set $e = Ax^\star - b$ where $x^\star$ is LS solution.

Randomized Kaczmarz (RK) with noise

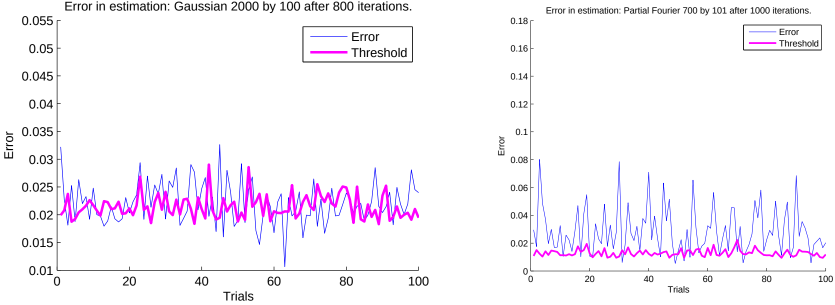


Figure Comparison between actual error (blue) and predicted threshold (pink). Scatter plot shows exponential convergence over several trials.

Even better convergence?

- Recall $x_{k+1} = x_k + (b[i] - \langle a_i, x_k \rangle) a_i$
- Since these projections are orthogonal, the optimal projection is one that maximizes $\|x_{k+1} - x_k\|_2$.
- Equivalently, one which maximizes: $\frac{|b[i] - \langle a_i, x_k \rangle|}{\|a_i\|_2}$.
- We should pick the row which maximizes this. But – can only afford to search through a constant number.
- Idea: Use dimension reduction (ala Johnson-Lindenstrauss) to approximate these terms and search through a large number of them.

Even better convergence?

- Recall $x_{k+1} = x_k + (b[i] - \langle a_i, x_k \rangle) a_i$
- Since these projections are orthogonal, the optimal projection is one that maximizes $\|x_{k+1} - x_k\|_2$.
- Equivalently, one which maximizes: $\frac{|b[i] - \langle a_i, x_k \rangle|}{\|a_i\|_2}$.
- We should pick the row which maximizes this. But – can only afford to search through a constant number.
- Idea: Use dimension reduction (ala Johnson-Lindenstrauss) to approximate these terms and search through a large number of them.

Even better convergence?

- Recall $x_{k+1} = x_k + (b[i] - \langle a_i, x_k \rangle) a_i$
- Since these projections are orthogonal, the optimal projection is one that maximizes $\|x_{k+1} - x_k\|_2$.
- Equivalently, one which maximizes: $\frac{|b[i] - \langle a_i, x_k \rangle|}{\|a_i\|_2}$.
- We should pick the row which maximizes this. But – can only afford to search through a constant number.
- Idea: Use dimension reduction (ala Johnson-Lindenstrauss) to approximate these terms and search through a large number of them.

Even better convergence?

- Recall $x_{k+1} = x_k + (b[i] - \langle a_i, x_k \rangle) a_i$
- Since these projections are orthogonal, the optimal projection is one that maximizes $\|x_{k+1} - x_k\|_2$.
- Equivalently, one which maximizes: $\frac{|b[i] - \langle a_i, x_k \rangle|}{\|a_i\|_2}$.
- We should pick the row which maximizes this. But – can only afford to search through a constant number.
- Idea: Use dimension reduction (ala Johnson-Lindenstrauss) to approximate these terms and search through a large number of them.

Even better convergence?

- Recall $x_{k+1} = x_k + (b[i] - \langle a_i, x_k \rangle) a_i$
- Since these projections are orthogonal, the optimal projection is one that maximizes $\|x_{k+1} - x_k\|_2$.
- Equivalently, one which maximizes: $\frac{|b[i] - \langle a_i, x_k \rangle|}{\|a_i\|_2}$.
- We should pick the row which maximizes this. But – can only afford to search through a constant number.
- Idea: Use dimension reduction (ala Johnson-Lindenstrauss) to approximate these terms and search through a large number of them.

Initialize Set $k = 0$, create a $d \times n$ Gaussian matrix Φ and set $\alpha_i = \Phi a_i$. Repeat the following $\mathrm{O}(n)$ times:

Select Select n rows with same prob. dist. Calculate

$$\gamma_i = \frac{|b[i] - \langle \alpha_i, \Phi x_k \rangle|}{\|\alpha_i\|_2},$$

and set $j = \mathrm{argmax}_i\, \gamma_i$.

Test For a_j and the first row a_l selected out of the n, explicitly calculate

$$\gamma_j^* = \frac{|b[j] - \langle a_j, x_k \rangle|}{\|a_j\|_2} \quad \text{and} \quad \gamma_l^* = \frac{|b[l] - \langle a_l, x_k \rangle|}{\|a_l\|_2}.$$

If $\gamma_l^* > \gamma_j^*$, set $j = l$.

Project Set

$$x_{k+1} = x_k + \frac{b[j] - \langle a_j, x_k \rangle}{\|a_j\|_2^2} a_j.$$

Update Set $k = k + 1$.

[Eldar-N]

Fix an estimation x_k and denote by x_{k+1} and x^*_{k+1} the next estimations using the RKJL and the standard RK method, respectively. Set $\gamma^*_j = |\langle a_j, x_k\rangle|^2$ and reorder so that $\gamma^*_1 \geq \gamma^*_2 \geq \ldots \geq \gamma^*_m$. Then when $d = C\delta^{-2}\log n$,

$$\mathbb{E}\|x_{k+1}-x\|_2^2 \leq \min\left[\mathbb{E}\|x^*_{k+1} - x\|_2^2 - \sum_{j=1}^{m}\left(p_j - \frac{1}{m}\right)\gamma^*_j + 2\delta, \quad \mathbb{E}\|x^*_{k+1} -$$

where p_j are non-negative values satisfying $\sum_{j=1}^{m} p_j = 1$ and $p_1 \geq p_2 \geq \ldots \geq p_m = 0$.

Large initial computation but accelerated convergence.

Accelerated RK via JL

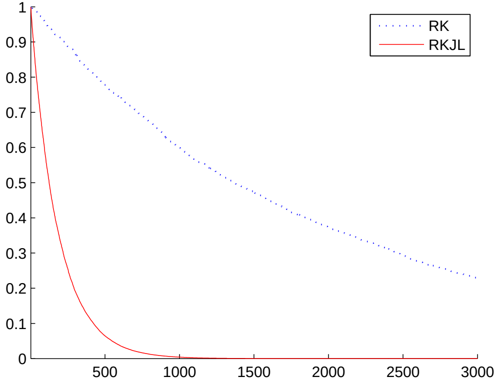
[Eldar-N]

Fix an estimation x_k and denote by x_{k+1} and x^*_{k+1} the next estimations using the RKJL and the standard RK method, respectively. Set $\gamma^*_j = |\langle a_j, x_k \rangle|^2$ and reorder so that $\gamma^*_1 \geq \gamma^*_2 \geq \ldots \geq \gamma^*_m$. Then when $d = C\delta^{-2} \log n$,

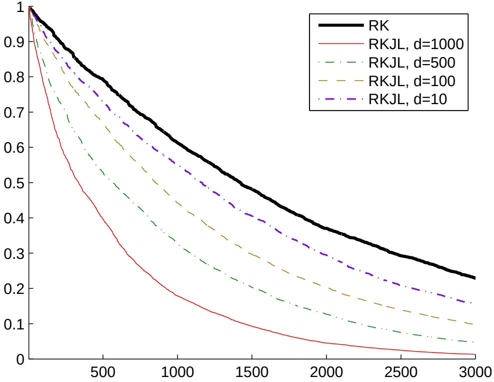
$$\mathbb{E}\|x_{k+1} - x\|_2^2 \leq \min \left[\mathbb{E}\|x^*_{k+1} - x\|_2^2 - \sum_{j=1}^{m} \left(p_j - \frac{1}{m}\right)\gamma^*_j + 2\delta, \quad \mathbb{E}\|x^*_{k+1} -$$

where p_j are non-negative values satisfying $\sum_{j=1}^m p_j = 1$ and $p_1 \geq p_2 \geq \ldots \geq p_m = 0$.

Large initial computation but accelerated convergence.



ℓ_2-Error (y-axis) as a function of the iterations (x-axis). The dashed line is standard Randomized Kaczmarz, and the solid line is the modified one, without a Johnson-Lindenstrauss projection. Instead, the best move out of the randomly chosen n rows is used. Note that we cannot afford to do this computationally.



ℓ_2-Error (y-axis) as a function of the iterations (x-axis) for various values of d with $m = 60000$ and $n = 1000$.

Even better (cheaper) convergence?

- Recall $x_{k+1} = x_k + (b[i] - \langle a_i, x_k \rangle) a_i$
- Since these projections are orthogonal, the optimal projection is one that maximizes $\|x_{k+1} - x_k\|_2$.
- What if we relax: $x_{k+1} = x_k + \gamma(b[i] - \langle a_i, x_k \rangle) a_i$
- Can we choose γ optimally?
- Idea: In each "iteration," project once with relaxation optimally and then project normally.

Even better (cheaper) convergence?

- Recall $x_{k+1} = x_k + (b[i] - \langle a_i, x_k \rangle) a_i$
- Since these projections are orthogonal, the optimal projection is one that maximizes $\|x_{k+1} - x_k\|_2$.
- What if we relax: $x_{k+1} = x_k + \gamma (b[i] - \langle a_i, x_k \rangle) a_i$
- Can we choose γ optimally?
- Idea: In each "iteration," project once with relaxation optimally and then project normally.

Even better (cheaper) convergence?

- Recall $x_{k+1} = x_k + (b[i] - \langle a_i, x_k \rangle) a_i$
- Since these projections are orthogonal, the optimal projection is one that maximizes $\|x_{k+1} - x_k\|_2$.
- What if we relax: $x_{k+1} = x_k + \gamma(b[i] - \langle a_i, x_k \rangle) a_i$
- Can we choose γ optimally?
- Idea: In each "iteration," project once with relaxation optimally and then project normally.

Even better (cheaper) convergence?

- Recall $x_{k+1} = x_k + (b[i] - \langle a_i, x_k \rangle) a_i$
- Since these projections are orthogonal, the optimal projection is one that maximizes $\|x_{k+1} - x_k\|_2$.
- What if we relax: $x_{k+1} = x_k + \gamma(b[i] - \langle a_i, x_k \rangle) a_i$
- Can we choose γ optimally?
- Idea: In each "iteration," project once with relaxation optimally and then project normally.

Even better convergence?

Two-subspace Kaczmarz

- Randomly select two rows, a_s and a_r
- Perform initial projection: $y = x_k + \gamma(b[i] - \langle a_i, x_k \rangle)a_i$ with γ optimal
- Peform second projection: $x_{k+1} = y + (b[i] - \langle a_i, y \rangle)a_i$
- Repeat

Even better convergence?

Two-subspace Kaczmarz

- Randomly select two rows, a_s and a_r
- Perform initial projection: $y = x_k + \gamma(b[i] - \langle a_i, x_k \rangle)a_i$ with γ optimal
- Peform second projection: $x_{k+1} = y + (b[i] - \langle a_i, y \rangle)a_i$
- Repeat

Even better convergence?

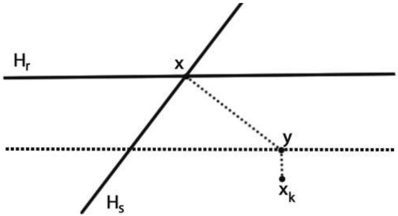
Two-subspace Kaczmarz

- Randomly select two rows, a_s and a_r
- Perform initial projection: $y = x_k + \gamma(b[i] - \langle a_i, x_k \rangle)a_i$ with γ optimal
- Peform second projection: $x_{k+1} = y + (b[i] - \langle a_i, y \rangle)a_i$
- Repeat

Two-subspace Kaczmarz

- Randomly select two rows, a_s and a_r
- Perform initial projection: $y = x_k + \gamma(b[i] - \langle a_i, x_k \rangle)a_i$ with γ optimal
- Peform second projection: $x_{k+1} = y + (b[i] - \langle a_i, y \rangle)a_i$
- Repeat

Geometrically, we choose γ in such a way:



Two-subspace Kaczmarz

The optimal choice of γ in a single iteration is

$$\gamma = \frac{-\langle a_r - \langle a_s, a_r\rangle a_s, x_k - x + (b_s - \langle x_k, a_s\rangle)a_s\rangle}{(b_r - \langle x_k, a_r\rangle)\|a_r - \langle a_s, a_r\rangle a_s\|_2^2}.$$

Two-Subspace Kaczmarz method

- Select two distinct rows of A uniformly at random
- $\mu_k \leftarrow \langle a_r, a_s\rangle$
- $y_k \leftarrow x_{k-1} + (b_s - \langle x_{k-1}, a_s\rangle)a_s$
- $v_k \leftarrow \frac{a_r - \mu_k a_s}{\sqrt{1-|\mu_k|^2}}$
- $\beta_k \leftarrow \frac{b_r - b_s\mu_k}{\sqrt{1-|\mu_k|^2}}$
- $x_k \leftarrow y_k + (\beta_k - \langle y_k, v_k\rangle)v_k$

Two-subspace Kaczmarz

The optimal choice of γ in a single iteration is

$$\gamma = \frac{-\langle a_r - \langle a_s, a_r\rangle a_s, x_k - x + (b_s - \langle x_k, a_s\rangle)a_s\rangle}{(b_r - \langle x_k, a_r\rangle)\|a_r - \langle a_s, a_r\rangle a_s\|_2^2}.$$

Two-Subspace Kaczmarz method

- Select two distinct rows of A uniformly at random
- $\mu_k \leftarrow \langle a_r, a_s\rangle$
- $y_k \leftarrow x_{k-1} + (b_s - \langle x_{k-1}, a_s\rangle)a_s$
- $v_k \leftarrow \frac{a_r - \mu_k a_s}{\sqrt{1-|\mu_k|^2}}$
- $\beta_k \leftarrow \frac{b_r - b_s\mu_k}{\sqrt{1-|\mu_k|^2}}$
- $x_k \leftarrow y_k + (\beta_k - \langle y_k, v_k\rangle)v_k$

Two-Subspace Kaczmarz

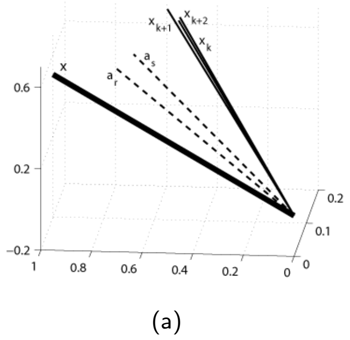


Figure For coherent systems, the one-subspace randomized Kaczmarz algorithm (a) converges more slowly than the two-subspace Kaczmarz algorithm (b).

Define the coherence parameters:

$$\Delta = \Delta(A) = \max_{j \neq k} |\langle a_j, a_k \rangle| \quad \textit{and} \quad \delta = \delta(A) = \min_{j \neq k} |\langle a_j, a_k \rangle|. \quad (1)$$

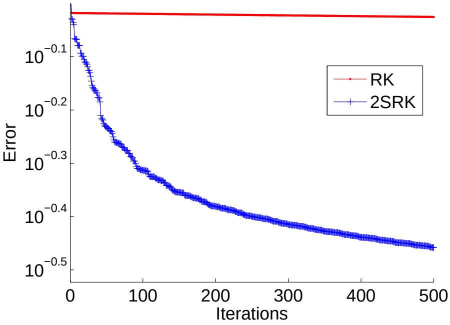


Figure Randomized Kaczmarz (RK) versus two-subspace RK (2SRK). A has highly coherent rows with $\delta = 0.992$ and $\Delta = 0.998$.

Two-Subspace Kaczmarz

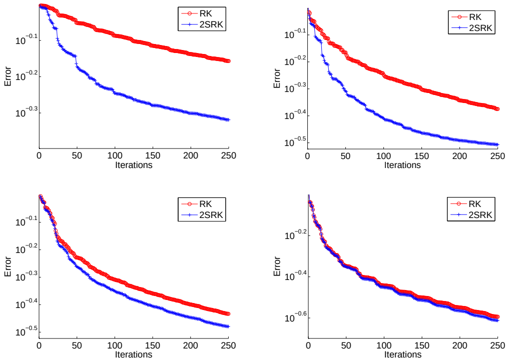


Figure Randomized Kaczmarz (RK) versus two-subspace RK (2SRK). A has highly coherent rows with coherence parameters (a) $\delta = 0.837$ and $\Delta = 0.967$, (b) $\delta = 0.534$ and $\Delta = 0.904$, (c) $\delta = 0.018$ and $\Delta = 0.819$, and (d) $\delta = 0$ and $\Delta = 0.610$.

Results

Recall the coherence parameters:

$$\Delta = \Delta(A) = \max_{j \neq k} |\langle a_j, a_k \rangle| \quad \textit{and} \quad \delta = \delta(A) = \min_{j \neq k} |\langle a_j, a_k \rangle|. \quad (2)$$

Theorem [N-Ward]

Let $b = Ax + e$, then the two-subspace Kaczmarz method yields

$$\mathbb{E}\|x - x_k\|_2 \leq \eta^{k/2}\|x - x_0\|_2 + \frac{3}{1-\sqrt{\eta}} \cdot \frac{\|e\|_\infty}{\sqrt{1-\Delta^2}},$$

where $D = \min\left\{\frac{\delta^2(1-\delta)}{1+\delta}, \frac{\Delta^2(1-\Delta)}{1+\Delta}\right\}$, $R = m\|A^{-1}\|^2$ denotes the scaled condition number, and $\eta = \left(1 - \frac{1}{R}\right)^2 - \frac{D}{R}$.

Results

Recall the coherence parameters:

$$\Delta = \Delta(A) = \max_{j \neq k} |\langle a_j, a_k \rangle| \quad \textit{and} \quad \delta = \delta(A) = \min_{j \neq k} |\langle a_j, a_k \rangle|. \quad (2)$$

Theorem [N-Ward]

Let $b = Ax + e$, then the two-subspace Kaczmarz method yields

$$\mathbb{E}\|x - x_k\|_2 \leq \eta^{k/2}\|x - x_0\|_2 + \frac{3}{1 - \sqrt{\eta}} \cdot \frac{\|e\|_\infty}{\sqrt{1 - \Delta^2}},$$

where $D = \min\left\{\frac{\delta^2(1-\delta)}{1+\delta}, \frac{\Delta^2(1-\Delta)}{1+\Delta}\right\}$, $R = m\|A^{-1}\|^2$ denotes the scaled condition number, and $\eta = \left(1 - \frac{1}{R}\right)^2 - \frac{D}{R}$.

Results

Remarks

1. When $\Delta = 1$ or $\delta = 0$ we recover the same convergence rate as provided for the standard Kaczmarz method since the two-subspace method utilizes two projections per iteration.

2. The bound presented in the theorem is a pessimistic bound. Even when $\Delta = 1$ or $\delta = 0$, the two-subspace method improves on the standard method if any rows of A are highly correlated (but not equal).

Results

Remarks

1. When $\Delta = 1$ or $\delta = 0$ we recover the same convergence rate as provided for the standard Kaczmarz method since the two-subspace method utilizes two projections per iteration.
2. The bound presented in the theorem is a pessimistic bound. Even when $\Delta = 1$ or $\delta = 0$, the two-subspace method improves on the standard method if any rows of A are highly correlated (but not equal).

The parameter D

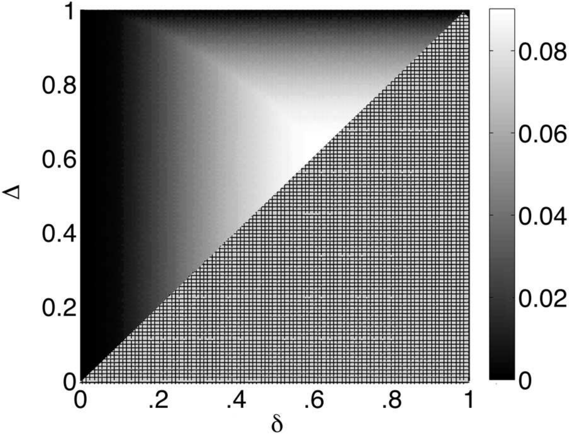


Figure A plot of the improved convergence factor D as a function of the coherence parameters δ and $\Delta \geq \delta$.

Generalization to more than two rows?

Block Kaczmarz [N-Tropp]

Randomized Block Kaczmarz method

Given a partition of the rows, T:

- Select a block τ of the partition at random
- $x_k \leftarrow x_{k-1} + A_\tau{}^\dagger(b_\tau - A_\tau x_{k-1})$

The convergence rate heavily depends on the conditioning of the blocks A_τ $\rightarrow$ need to control geometric properties of the partition.

Block Kaczmarz [N-Tropp]

Randomized Block Kaczmarz method

Given a partition of the rows, T:

- Select a block τ of the partition at random
- $x_k \leftarrow x_{k-1} + {A_\tau}^\dagger (b_\tau - A_\tau x_{k-1})$

The convergence rate heavily depends on the conditioning of the blocks A_τ $\rightarrow$ need to control geometric properties of the partition.

Block Kaczmarz [N-Tropp]

Randomized Block Kaczmarz method

Given a partition of the rows, T:

- Select a block τ of the partition at random
- $x_k \leftarrow x_{k-1} + A_\tau{}^\dagger(b_\tau - A_\tau x_{k-1})$

The convergence rate heavily depends on the conditioning of the blocks $A_\tau \rightarrow$ need to control geometric properties of the partition.

Block Kaczmarz [N-Tropp]

Randomized Block Kaczmarz method

Given a partition of the rows, T:

- Select a block τ of the partition at random
- $x_k \leftarrow x_{k-1} + A_\tau{}^\dagger (b_\tau - A_\tau x_{k-1})$

The convergence rate heavily depends on the conditioning of the blocks A_τ $\rightarrow$ need to control geometric properties of the partition.

Block Kaczmarz

Row paving

A (d, α, β) *row paving* of a matrix A is a partition $T = \{\tau_1, \dots, \tau_d\}$ of the row indices that verifies

$$\alpha \le \lambda_{\min}(A_\tau A_\tau^*) \quad \text{and} \quad \lambda_{\max}(A_\tau A_\tau^*) \le \beta \quad \text{for each } \tau \in T.$$

Theorem [N-Tropp]

Suppose A admits an (d, α, β) row paving T and that $b = Ax + e$. The convergence of the block Kaczmarz method satisfies

$$\mathbb{E}\|x_k - x\|_2^2 \le \left[1 - \frac{\sigma_{\min}^2(A)}{\beta d}\right]^k \|x_0 - x\|_2^2 + \frac{\beta}{\alpha} \cdot \frac{\|e\|_2^2}{\sigma_{\min}^2(A)}. \quad (3)$$

Block Kaczmarz

Row paving

A (d, α, β) *row paving* of a matrix A is a partition $T = \{\tau_1, \dots, \tau_d\}$ of the row indices that verifies

$$\alpha \leq \lambda_{\min}(A_\tau A_\tau^*) \quad \text{and} \quad \lambda_{\max}(A_\tau A_\tau^*) \leq \beta \quad \text{for each } \tau \in T.$$

Theorem [N-Tropp]

Suppose A admits an (d, α, β) row paving T and that $b = Ax + e$. The convergence of the block Kaczmarz method satisfies

$$\mathbb{E}\|x_k - x\|_2^2 \leq \left[1 - \frac{\sigma_{\min}^2(A)}{\beta d}\right]^k \|x_0 - x\|_2^2 + \frac{\beta}{\alpha} \cdot \frac{\|e\|_2^2}{\sigma_{\min}^2(A)}. \quad (3)$$

Good row pavings [Bougain-Tzafriri]

For any $\delta \in (0,1)$, A admits a row paving with

$$d \leq C \cdot \delta^{-2} \|A\|^2 \log(1+n) \quad \text{and} \quad 1-\delta \leq \alpha \leq \beta \leq 1+\delta.$$

Theorem [N-Tropp]

Let A have row paving above with $\delta = 1/2$. The block Kaczmarz method yields

$$\mathbb{E}\|x_k - x\|_2^2 \leq \left[1 - \frac{1}{C\kappa^2(A)\log(1+n)}\right]^k \|x_0 - x\|_2^2 + \frac{3\|e\|_2^2}{\sigma_{\min}^2(A)}.$$

Block Kaczmarz

Good row pavings [Bougain-Tzafriri]

For any $\delta \in (0,1)$, A admits a row paving with

$$d \leq C \cdot \delta^{-2} \|A\|^2 \log(1+n) \quad \text{and} \quad 1-\delta \leq \alpha \leq \beta \leq 1+\delta.$$

Theorem [N-Tropp]

Let A have row paving above with $\delta = 1/2$. The block Kaczmarz method yields

$$\mathbb{E}\|x_k - x\|_2^2 \leq \left[1 - \frac{1}{C\kappa^2(A)\log(1+n)}\right]^k \|x_0 - x\|_2^2 + \frac{3\|e\|_2^2}{\sigma_{\min}^2(A)}.$$

Block Kaczmarz

Theorem [Bougain-Tzafriri, Vershynin, Tropp]

A random partition of the row indices with $m \geq \|A\|^2$ blocks is a row paving with upper bound $\beta \leq 6 \log(1 + n)$, with probability at least $1 - n^{-1}$.

Theorem [Bourgain-Tzafriri, Vershynin, Tropp]

Suppose that A is incoherent. A random partition of the row indices into m blocks where $m \geq C \cdot \delta^{-2} \|A\|^2 \log(1 + n)$ is a row paving of A whose paving bounds satisfy $1 - \delta \leq \alpha \leq \beta \leq 1 + \delta$, with probability at least $1 - n^{-1}$.

Block Kaczmarz

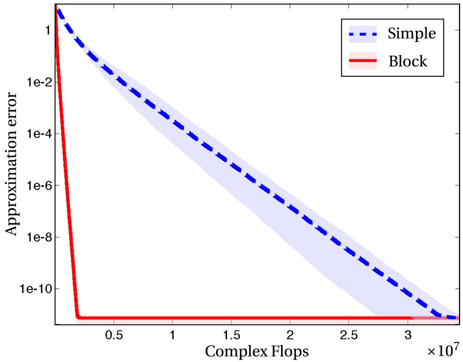


Figure The matrix A is a fixed 300×100 matrix consisting of 15 partial circulant blocks. Error $\|x_k - x\|_2$ per flop count.

Block Kaczmarz

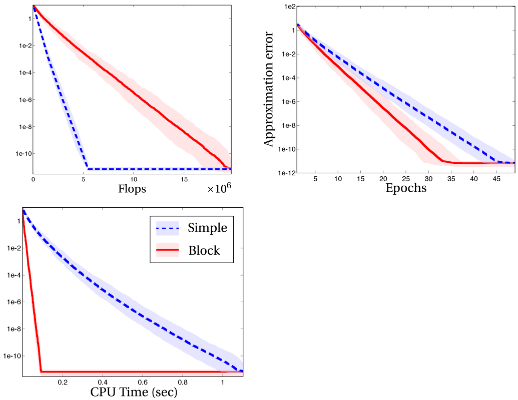


Figure The matrix A is a fixed 300×100 matrix with rows drawn randomly from the unit sphere, with $d = 10$ blocks. Error $\|x_k - x\|_2$ over various computational resources.

Block Kaczmarz

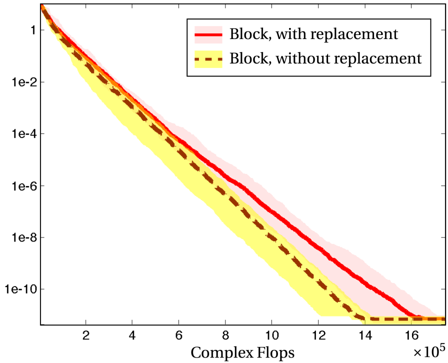


Figure Shout out to going Hogwild – with versus without replacement for circulant matrix.

Recall SGD to minimize $F(x) = \mathbb{E}f_i(x)$ [N-Srebro-Ward]

SGD

Input:

- Initial estimate $x_0 \in \mathbb{R}^d$
- Degree of nonuniform sampling $\lambda \in [0, 1]$
- Step size $\gamma > 0$
- Tolerance parameter $\delta > 0$
- Access to the source distribution $\mathcal{D}$
- If $\lambda < 1$: bounds on the Lipschitz constants L_i;

$k \leftarrow 0$

Repeat:

$$k \leftarrow k + 1$$

Draw an index $i \sim \mathcal{D}^{(\lambda)}$

$$x_k \leftarrow x_{k-1} - \frac{\gamma}{w_\lambda(i)} \nabla f_i(x_{k-1})$$

Recall SGD to minimize $F(x) = \mathbb{E}f_i(x)$ [N-Srebro-Ward]

Convergence rate for SGD with partially biased sampling

Let f_i be continuously differentiable convex functionals, where each ∇f_i has Lipschitz constant L_i, and let $F(x) = \mathbb{E}_{i\sim\mathcal{D}} f_i(x)$ be μ-strongly convex. Set $\sigma^2 = \mathbb{E}_{i\sim\mathcal{D}}\|\nabla f_i(x_\star)\|_2^2$, where $x_\star$ is the minimizer of

$$x_\star = \operatorname*{argmin}_x F(x).$$

Then the iterate x_k satisfies

$$\mathbb{E}\|x_k - x_\star\|_2^2 \leq \Big[1 - 2\gamma\mu(1-\gamma\alpha)\Big]^k \|x_0 - x_\star\|_2^2 + \frac{\gamma\beta\sigma^2}{\mu(1-\gamma\alpha)},$$

where the expectation is with respect to the random sampling in the Algorithm, $\alpha = \alpha(\lambda) = \min\left(\frac{\overline{L}}{1-\lambda}, \frac{\sup_i L_i}{\lambda}\right)$, and $\beta = \beta(\lambda) = \min\left(\frac{1}{\lambda}, \frac{\overline{L}}{(1-\lambda)\inf_i L_i}\right)$.

SGD and Kaczmarz

Want to minimize:

$$F(x) = \frac{1}{2}\sum_{i=1}^{n}(\langle a_i, x\rangle - b_i)^2 = \frac{1}{2}\|Ax - b\|_2^2$$

which can be formulated as a general problem of minimizing $F(x) = \mathbb{E}f_i(x)$ where

- The components are $f_i = \frac{n}{2}(\langle a_i, x\rangle - b_i)^2$
- The Lipschitz constants are $L_i = n\|a_i\|_2^2$, and the average Lipschitz constant is $\frac{1}{n}\sum_i L_i = \|A\|_F^2$.
- The strong convexity parameter is $\mu = \frac{1}{\|A^{-1}\|^2}$, so that $K(A) := \overline{L}/\mu = \|A\|_F^2\|A^{-1}\|^2$
- The residual is $\sigma^2 = n\sum_i \|a_i\|_2^2 |\langle a_i, x_\star\rangle - b_i|^2$.

SGD and Kaczmarz

Want to minimize:

$$F(x) = \frac{1}{2}\sum_{i=1}^{n}(\langle a_i, x\rangle - b_i)^2 = \frac{1}{2}\|Ax - b\|_2^2$$

which can be formulated as a general problem of minimizing $F(x) = \mathbb{E} f_i(x)$ where

- The components are $f_i = \frac{n}{2}(\langle a_i, x\rangle - b_i)^2$
- The Lipschitz constants are $L_i = n\|a_i\|_2^2$, and the average Lipschitz constant is $\frac{1}{n}\sum_i L_i = \|A\|_F^2$.
- The strong convexity parameter is $\mu = \frac{1}{\|A^{-1}\|^2}$, so that $K(A) := \overline{L}/\mu = \|A\|_F^2\|A^{-1}\|^2$
- The residual is $\sigma^2 = n\sum_i \|a_i\|_2^2 |\langle a_i, x_\star\rangle - b_i|^2$.

Want to minimize:

$$F(x) = \frac{1}{2}\sum_{i=1}^{n}(\langle a_i, x\rangle - b_i)^2 = \frac{1}{2}\|Ax - b\|_2^2$$

which can be formulated as a general problem of minimizing $F(x) = \mathbb{E}f_i(x)$ where

- The components are $f_i = \frac{n}{2}(\langle a_i, x\rangle - b_i)^2$
- The Lipschitz constants are $L_i = n\|a_i\|_2^2$, and the average Lipschitz constant is $\frac{1}{n}\sum_i L_i = \|A\|_F^2$.
- The strong convexity parameter is $\mu = \frac{1}{\|A^{-1}\|^2}$, so that $K(A) := \overline{L}/\mu = \|A\|_F^2\|A^{-1}\|^2$
- The residual is $\sigma^2 = n\sum_i\|a_i\|_2^2 |\langle a_i, x_\star\rangle - b_i|^2$.

SGD and Kaczmarz

Want to minimize:

$$F(x) = \frac{1}{2}\sum_{i=1}^{n}(\langle a_i, x\rangle - b_i)^2 = \frac{1}{2}\|Ax - b\|_2^2$$

which can be formulated as a general problem of minimizing $F(x) = \mathbb{E} f_i(x)$ where

- The components are $f_i = \frac{n}{2}(\langle a_i, x\rangle - b_i)^2$
- The Lipschitz constants are $L_i = n\|a_i\|_2^2$, and the average Lipschitz constant is $\frac{1}{n}\sum_i L_i = \|A\|_F^2$.
- The strong convexity parameter is $\mu = \frac{1}{\|A^{-1}\|^2}$, so that $K(A) := \overline{L}/\mu = \|A\|_F^2\|A^{-1}\|^2$
- The residual is $\sigma^2 = n\sum_i \|a_i\|_2^2 |\langle a_i, x_\star\rangle - b_i|^2$.

SGD and Kaczmarz

Want to minimize:

$$F(x) = \frac{1}{2}\sum_{i=1}^{n}(\langle a_i, x\rangle - b_i)^2 = \frac{1}{2}\|Ax - b\|_2^2$$

which can be formulated as a general problem of minimizing $F(x) = \mathbb{E} f_i(x)$ where

- The components are $f_i = \frac{n}{2}(\langle a_i, x\rangle - b_i)^2$
- The Lipschitz constants are $L_i = n\|a_i\|_2^2$, and the average Lipschitz constant is $\frac{1}{n}\sum_i L_i = \|A\|_F^2$.
- The strong convexity parameter is $\mu = \frac{1}{\|A^{-1}\|^2}$, so that $K(A) := \overline{L}/\mu = \|A\|_F^2\|A^{-1}\|^2$
- The residual is $\sigma^2 = n\sum_i \|a_i\|_2^2 |\langle a_i, x_\star\rangle - b_i|^2$.

SGD and Kaczmarz

Want to minimize:

$$F(x) = \frac{1}{2}\sum_{i=1}^{n}(\langle a_i, x\rangle - b_i)^2 = \frac{1}{2}\|Ax - b\|_2^2$$

which can be formulated as a general problem of minimizing $F(x) = \mathbb{E}f_i(x)$ where

- The components are $f_i = \frac{n}{2}(\langle a_i, x\rangle - b_i)^2$
- The Lipschitz constants are $L_i = n\|a_i\|_2^2$, and the average Lipschitz constant is $\frac{1}{n}\sum_i L_i = \|A\|_F^2$.
- The strong convexity parameter is $\mu = \frac{1}{\|A^{-1}\|^2}$, so that $K(A) := \overline{L}/\mu = \|A\|_F^2\|A^{-1}\|^2$
- The residual is $\sigma^2 = n\sum_i\|a_i\|_2^2|\langle a_i, x_\star\rangle - b_i|^2$.

SGD and Kaczmarz

Consider the relaxed Kaczmarz method:

$$x_{k+1} = x_k + c \cdot \frac{b_i - \langle a_i, x_k \rangle}{\|a_i\|_2^2} a_i \quad \mathbb{P}(i) = \|a_i\|_2^2 / \|A\|_F^2$$

Convergence rate for Kaczmarz with fully biased sampling

Set $e = Ax_\star - b$, $a_{\min}^2 = \inf_i \|a_i\|_2^2$, $a_{\max}^2 = \sup_i \|a_i\|_2^2$, and $e_{\max}^2 = \sup_i e_i^2$. Then

$$\mathbb{E}\|x_k - x_\star\|_2^2 \leq \left[1 - \frac{2c(1-c)}{K(A)}\right]^k \|x_0 - x_\star\|_2^2 + \frac{c}{1-c} K(A) \tilde{r},$$

with $\widetilde{r} = (a_{\max}^2 / a_{\min}^2) \min\left\{e_{\max}^2 / a_{\max}^2,\ \|e\|_2^2 / \|A\|_F^2\right\}$.

SGD and Kaczmarz

Consider the relaxed Kaczmarz method:

$$x_{k+1} = x_k + c \cdot \frac{b_i - \langle a_i, x_k \rangle}{\|a_i\|_2^2} a_i \quad \mathbb{P}(i) = \|a_i\|_2^2 / \|A\|_F^2$$

Convergence rate for Kaczmarz with fully biased sampling

Set $e = Ax_\star - b$, $a_{\min}^2 = \inf_i \|a_i\|_2^2$, $a_{\max}^2 = \sup_i \|a_i\|_2^2$, and $e_{\max}^2 = \sup_i e_i^2$. Then

$$\mathbb{E}\|x_k - x_\star\|_2^2 \le \left[1 - \frac{2c(1-c)}{K(A)}\right]^k \|x_0 - x_\star\|_2^2 + \frac{c}{1-c} K(A) \tilde{r},$$

with $\widetilde{r} = (a_{\max}^2 / a_{\min}^2) \min \left\{ e_{\max}^2 / a_{\max}^2, \; \|e\|_2^2 / \|A\|_F^2 \right\}$.

SGD and Kaczmarz

$$x_{k+1} = x_k + c \cdot \frac{b_i - \langle a_i, x_k \rangle}{\|a_i\|_2^2} a_i \quad \mathbb{P}(i) = \|a_i\|_2^2 / \|A\|_F^2$$

Convergence rate for Kaczmarz with fully biased sampling

$$\mathbb{E}\|x_k - x_\star\|_2^2 \leq \left[1 - \frac{2c(1-c)}{K(A)}\right]^k \|x_0 - x_\star\|_2^2 + \frac{c}{1-c} K(A) \tilde{r},$$

- Small step size c diminishes the convergence horizon.
- Tradeoff between convergence horizon and convergence rate.
- Non-uniform sampling.

Convergence rate for randomized Kaczmarz with uniform sampling

Let D be the diagonal matrix with terms $d_{j,j} = \|a_i\|_2$ and set $e_w = D^{-1}(Ax_\star^w - b)$, where

$$x_\star^w = \underset{x}{\operatorname{argmin}} \frac{1}{2}\|D^{-1}(Ax - b)\|_2^2.$$

Then

$$\mathbb{E}\|x_k - x_\star^w\|_2^2 \leq \left[1 - \frac{2c(1-c)}{K(D^{-1}A)}\right)\Big]^k \|x_0 - x_\star^w\|_2^2 + \frac{c}{1-c} K(D^{-1}A) r_w, \quad (4)$$

where $r_w = \|e_w\|_2^2/n$.

SGD and Kaczmarz

Convergence rate for randomized Kaczmarz with uniform sampling

$$\mathbb{E}\|x_k - x_\star^w\|_2^2 \leq \left[1 - \frac{2c(1-c)}{K(D^{-1}A)}\right)\right]^k \|x_0 - x_\star^w\|_2^2 + \frac{c}{1-c} K(D^{-1}A) r_w,$$

- Convergence to pre-conditioned system solution.
- Small step size still diminishes convergence horizon.
- Uniform sampling!

SGD and Kaczmarz

Convergence rate for hybrid randomized Kaczmarz

For any $\lambda \in [0, 1]$,

$$\mathbb{E}\|x_k - x_\star\|_2^2 \le \left(1 - \frac{2\gamma_{\min}(1 - \gamma_{\max}\alpha)}{\|A^{-1}\|^2}\right)\Big)^k \|x_0 - x_\star\|_2^2 + \frac{\gamma_{\max}\beta a_{\max} n \|A^{-1}\|^2 \|e\|_2^2}{(1 - \gamma_{\max}\alpha)},$$

where $a_{\min} = \min_i \|a_i\|_2^2$, $a_{\max} = \max_i \|a_i\|_2^2$,
$\alpha = \min\left(\frac{\|A\|_F^2}{1-\lambda}, \frac{n a_{\max}}{\lambda}\right)$, $\beta = \min\left(\frac{1}{\lambda}, \frac{\|A\|_F^2}{n a_{\min}(1-\lambda)}\right)$,
$\gamma_{\min} = \frac{c\lambda}{n a_{\max}} + \frac{c(1-\lambda)}{\|A\|_F^2}$, and $\gamma_{\max} = \frac{c\lambda}{n a_{\min}} + \frac{c(1-\lambda)}{\|A\|_F^2}$.

Allows for an alternative way to tradeoff.

Convergence rate for hybrid randomized Kaczmarz

For any $\lambda \in [0,1]$,

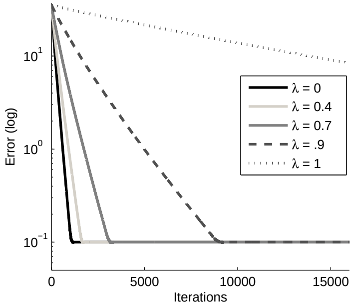
$$\mathbb{E}\|x_k - x_\star\|_2^2 \le \left(1 - \frac{2\gamma_{\min}(1-\gamma_{\max}\alpha)}{\|A^{-1}\|^2}\right)\Bigg)^k \|x_0 - x_\star\|_2^2 + \frac{\gamma_{\max}\beta a_{\max} n \|A^{-1}\|^2 \|e\|_2^2}{(1-\gamma_{\max}\alpha)},$$

where $a_{\min} = \min_i \|a_i\|_2^2$, $a_{\max} = \max_i \|a_i\|_2^2$,
$\alpha = \min\left(\frac{\|A\|_F^2}{1-\lambda}, \frac{n a_{\max}}{\lambda}\right)$, $\beta = \min\left(\frac{1}{\lambda}, \frac{\|A\|_F^2}{n a_{\min}(1-\lambda)}\right)$,
$\gamma_{\min} = \frac{c\lambda}{n a_{\max}} + \frac{c(1-\lambda)}{\|A\|_F^2}$, and $\gamma_{\max} = \frac{c\lambda}{n a_{\min}} + \frac{c(1-\lambda)}{\|A\|_F^2}$.

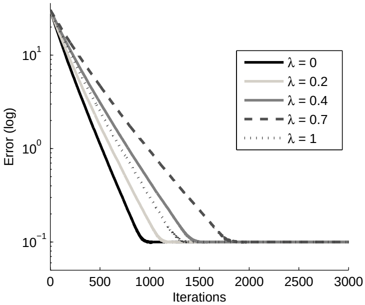
Allows for an alternative way to tradeoff.

SGD and Kaczmarz

Can also consider "variant" of Kaczmarz method, SGD with $f_i(x) = \frac{n}{2}(\langle a_i, x\rangle - b_i)^2$, sampling uniformly $1-\lambda$ proportion of the time.

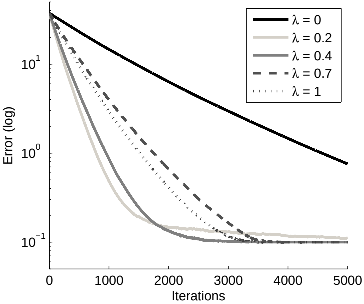


Entries of A N(0,1) but last row N(0,100)

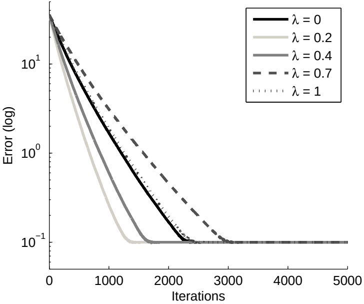


Entries of A N(0,1)

Can also consider "variant" of Kaczmarz method, SGD with $f_i(x) = \frac{n}{2}(\langle a_i, x\rangle - b_i)^2$, sampling uniformly $1 - \lambda$ proportion of the time.



Entries of A_{ik} N(0,j), large residual.



Entries of A_{ik} N(0,j), small residual.

For more information

E-mail:

- dneedell@cmc.edu

Web: www.cmc.edu/pages/faculty/DNeedell

References:

- Strohmer, Vershynin, "A randomized Kaczmarz algorithm with exponential convergence", J. Four. Ana. and App. 2009.
- Needell, "Randomized Kaczmarz solver for noisy linear systems", BIT Num. Math., 2010.
- Needell, Ward, "Two-subspace Projection Method for Coherent Overdetermined Systems", J. Four. Ana. and App., 2013.
- Needell, Tropp , "Paved with Good Intentions: Analysis of a Randomized Block Kaczmarz Method", Lin. Alg. App., 2014.
- Needell, Srebro, Ward, "Stochastic gradient descent and the randomized Kaczmarz algorithm", submitted.