Technische Universität Braunschweig

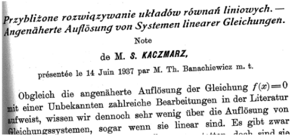
Randomized sparse Kaczmarz methods

Dirk Lorenz, joint with Frank Schöpfer, Feb 9, 2018

Inverse Problems and Machine Learning, Caltech 2018

- **The Kaczmarz method**
- Randomization
- Sparsity
- Split feasibility problems
- Convergence rates

Just solving systems of linear equations


Przybliżone rozwiązywanie układów równań liniowych. — Angenäherte Auflösung von Systemen linearer Gleichungen.

Note

de M. S. KACZMARZ,

présentée le 14 Juin 1937 par M. Th. Banachiewicz m. t.

Obgleich die angenäherte Auflösung der Gleichung $f(x)=0$ mit einer Unbekannten zahlreiche Bearbeitungen in der Literatur aufweist, wissen wir dennoch sehr wenig über die Auflösung von Gleichungssystemen, sogar wenn sie linear sind. Es gibt zwar

- $Ax = b$ pretty arbitrary (but consistent), m rows, n columns

Just solving systems of linear equations

Przybliżone rozwiązywanie układów równań liniowych. — Angenäherte Auflösung von Systemen linearer Gleichungen.

Note

de M. S. KACZMARZ,

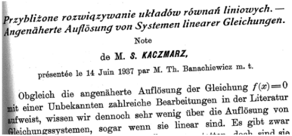
présentée le 14 Juin 1937 par M. Th. Banachiewicz m. t.

Obgleich die angenäherte Auflösung der Gleichung $f(x)=0$ mit einer Unbekannten zahlreiche Bearbeitungen in der Literatur aufweist, wissen wir dennoch sehr wenig über die Auflösung von Gleichungssystemen, sogar wenn sie linear sind. Es gibt zwar

- $Ax = b$ pretty arbitrary (but consistent), m rows, n columns
- Solve only one row $\langle a_i, x\rangle = b$ by projecting onto the hyperplane of solutions:

$$x^{k+1} = x^k - \frac{\langle x^k, a_i\rangle - b_i}{\|a_i\|^2} a_i$$

Just solving systems of linear equations


Przybliżone rozwiązywanie układów równań liniowych. —
Angenäherte Auflösung von Systemen linearer Gleichungen.

Note

de M. S. KACZMARZ,

présentée le 14 Juin 1937 par M. Th. Banachiewicz m. t.

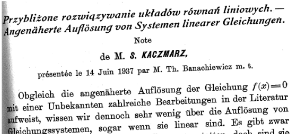
Obgleich die angenäherte Auflösung der Gleichung $f(x)=0$ mit einer Unbekannten zahlreiche Bearbeitungen in der Literatur aufweist, wissen wir dennoch sehr wenig über die Auflösung von Gleichungssystemen, sogar wenn sie linear sind. Es gibt zwar

- $Ax = b$ pretty arbitrary (but consistent), m rows, n columns
- Solve only one row $\langle a_i, x\rangle = b$ by projecting onto the hyperplane of solutions:

$$x^{k+1} = x^k - \frac{\langle x^k, a_i\rangle - b_i}{\|a_i\|^2} a_i$$

- Each projection just needs $O(n)$ operations

Just solving systems of linear equations


Przybliżone rozwiązywanie układów równań liniowych. —
Angenäherte Auflösung von Systemen linearer Gleichungen.

Note

de M. S. KACZMARZ,

présentée le 14 Juin 1937 par M. Th. Banachiewicz m. t.

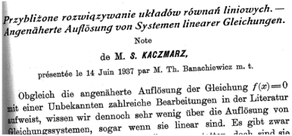
Obgleich die angenäherte Auflösung der Gleichung $f(x)=0$ mit einer Unbekannten zahlreiche Bearbeitungen in der Literatur aufweist, wissen wir dennoch sehr wenig über die Auflösung von Gleichungssystemen, sogar wenn sie linear sind. Es gibt zwar

- $Ax = b$ pretty arbitrary (but consistent), m rows, n columns
- Solve only one row $\langle a_i, x\rangle = b$ by projecting onto the hyperplane of solutions:

$$x^{k+1} = x^k - \frac{\langle x^k, a_i\rangle - b_i}{\|a_i\|^2} a_i$$

- Each projection just needs $O(n)$ operations
- Amount for one pass through all columns same as applying A

Just solving systems of linear equations


Przybliżone rozwiązywanie układów równań liniowych. — Angenäherte Auflösung von Systemen linearer Gleichungen.

Note

de M. S. KACZMARZ,

présentée le 14 Juin 1937 par M. Th. Banachiewicz m. t.

Obgleich die angenäherte Auflösung der Gleichung $f(x)=0$ mit einer Unbekannten zahlreiche Bearbeitungen in der Literatur aufweist, wissen wir dennoch sehr wenig über die Auflösung von Gleichungssystemen, sogar wenn sie linear sind. Es gibt zwar

- $Ax = b$ pretty arbitrary (but consistent), m rows, n columns
- Solve only one row $\langle a_i, x\rangle = b$ by projecting onto the hyperplane of solutions:

$$x^{k+1} = x^k - \frac{\langle x^k, a_i\rangle - b_i}{\|a_i\|^2} a_i$$

- Each projection just needs $O(n)$ operations
- Amount for one pass through all columns same as applying A
- Stefan Kaczmarz [1937]: Convergent to some solution for all consistent systems

Learning with Kaczmarz

- Unknown distibution ρ on $X \times Y = \mathbf{R}^d \times \mathbf{R}$, regression function $f_\rho(a) = \int y \, d\rho(y \mid a)$
- Hypothesis space $\mathcal{H} = \{f_x \in L^2_{\rho_X}, x \in \mathbf{R}^d\}, f_x(a) = \langle a, x\rangle$
- Learning: Obtain samples $a \in X'$, $b \in Y$ sequentially and try to learn x
- Kaczmarz: Update x^k by

$$x^{k+1} = x^k - \frac{\langle x^k, a\rangle - b}{\|a\|^2} a$$

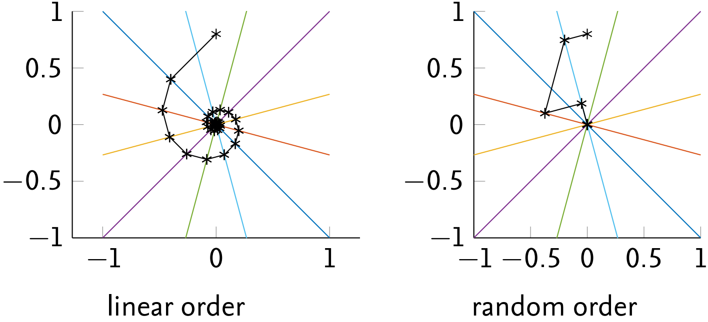
- Goal: Show that x^k converges to some x^* such that

$$f_{x^*} = \operatorname*{argmin}_{f \in \mathcal{H}} \mathbf{E}(f) = \operatorname*{argmin}_{f \in \mathcal{H}} \int_{X \times Y} (b - f(a))^2 d\rho$$

  [Lin, Zhou 2015]
- Here focus on Kaczmarz as an algorithm for solving systems

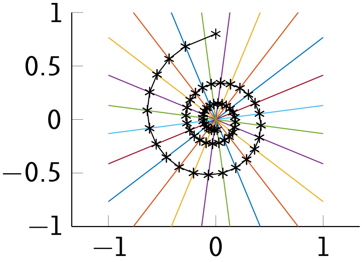
Convergence speed?

$m = 6$ rows, $n = 2$ columns:

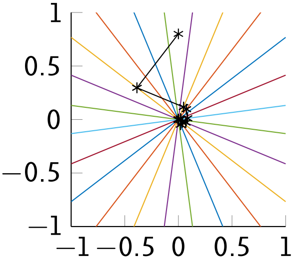


Convergence speed?

$m = 12$ rows, $n = 2$ columns:



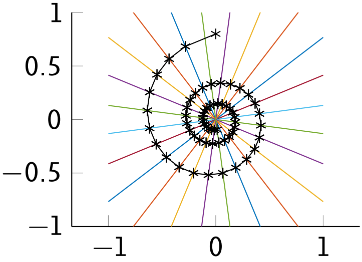
linear order



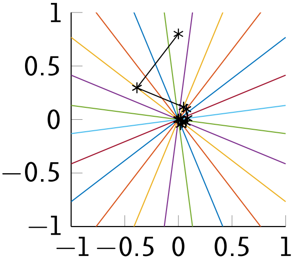
random order

Convergence speed?

rows, $n = 2$ columns:



linear order



random order

- Btw: Randomized Kaczmarz is stochastic gradient descent for $\sum_i (\langle a_i, x\rangle - b_i)^2$

- The Kaczmarz method
- **Randomization**
- Sparsity
- Split feasibility problems
- Convergence rates

Randomization leads to linear convergence

- In each iteration, choose index i with probability p_i.
- If $\hat{x}$ solves (i.e. $\langle \hat{x}, a_i\rangle = b_i$), then

$$\|x^{k+1} - \hat{x}\|^2 = \|x^k - \hat{x}\|^2 - \frac{(\langle x^k - \hat{x}, a_i\rangle)^2}{\|a_i\|^2}.$$

- Taking the expectation over the choice of i gives

$$\begin{aligned}\mathbf{E}(\|x^{k+1} - \hat{x}\|^2) &= \|x^k - \hat{x}\|^2 - \sum_i p_i \frac{(\langle x^k - \hat{x}, a_i\rangle)^2}{\|a_i\|^2}\\ &= \|x^k - \hat{x}\|^2 - \langle A(x^k - \hat{x}), DA(x^k - \hat{x})\rangle\end{aligned}$$

  with $D = \operatorname{diag}(p_i/\|a_i\|^2)$.
- Gives uniform improvement

$$\mathbf{E}(\|x^{k+1} - \hat{x}\|^2) \le (1-\lambda)\|x^k - \hat{x}\|^2, \quad \lambda = \lambda_{\min}(A^T DA)$$

Theorem

$A \in \mathbf{R}^{m\times n}$, $m \geq n$ with full rank, $A\hat{x} = b$, then iterates of randomized Kaczmarz fulfill

$$\mathbf{E}(\|x^k - \hat{x}\|^2) \leq (1-\lambda)^k \|x^0 - \hat{x}\|^2$$

with $\lambda = \lambda_{\min}(A^T DA)$, $D = \operatorname{diag}(p_i/\|a_i\|^2)$.

Theorem

$A \in \mathbf{R}^{m\times n}$, $m \geq n$ with full rank, $A\hat{x} = b$, then iterates of randomized Kaczmarz fulfill

$$\mathbf{E}(\|x^k - \hat{x}\|^2) \leq (1-\lambda)^k \|x^0 - \hat{x}\|^2$$

with $\lambda = \lambda_{\min}(A^T DA)$, $D = \operatorname{diag}(p_i/\|a_i\|^2)$.

- Result due to [Stohmer, Vershynin 2009]

Theorem

$A \in \mathbf{R}^{m\times n}$, $m \geq n$ with full rank, $A\hat{x} = b$, then iterates of randomized Kaczmarz fulfill

$$\mathbf{E}(\|x^k - \hat{x}\|^2) \leq (1-\lambda)^k \|x^0 - \hat{x}\|^2$$

with $\lambda = \lambda_{\min}(A^T DA)$, $D = \operatorname{diag}(p_i/\|a_i\|^2)$.

- Result due to [Stohmer, Vershynin 2009]
- Choice $p_i = \frac{\|a_i\|^2}{\|A\|_F^2}$ gives $D = \|A\|_F^{-2} I$, i.e.

$$\lambda = \frac{\lambda_{\min}(A^T A)}{\|A\|_F^2} = \frac{\sigma_{\min}(A)}{\|A\|_F^2} =: \kappa(A)$$

Theorem

$A \in \mathbf{R}^{m\times n}$, $m \geq n$ with full rank, $A\hat{x} = b$, then iterates of randomized Kaczmarz fulfill

$$\mathbf{E}(\|x^k - \hat{x}\|^2) \leq (1-\lambda)^k \|x^0 - \hat{x}\|^2$$

with $\lambda = \lambda_{\min}(A^T DA)$, $D = \operatorname{diag}(p_i / \|a_i\|^2)$.

- Result due to [Stohmer, Vershynin 2009]
- Choice $p_i = \frac{\|a_i\|^2}{\|A\|_F^2}$ gives $D = \|A\|_F^{-2} I$, i.e.

$$\lambda = \frac{\lambda_{\min}(A^T A)}{\|A\|_F^2} = \frac{\sigma_{\min}(A)}{\|A\|_F^2} =: \kappa(A)$$

- Experimentally: above p not optimal, other p give larger λ

Underdetermined systems

- Consider $Ax = b$, underdetermined but consistent

Technische Universität Braunschweig

Underdetermined systems

- Consider $Ax = b$, underdetermined but consistent
- Which solution does Kaczmarz pick?

Underdetermined systems

- Consider $Ax = b$, underdetermined but consistent
- Which solution does Kaczmarz pick?
- Initialization $x^0 = 0$ (or $x^0 \in \operatorname{rg} A^T$), then all iterates $x^k \in \operatorname{rg} A^T$

Underdetermined systems

- Consider $Ax = b$, underdetermined but consistent
- Which solution does Kaczmarz pick?
- Initialization $x^0 = 0$ (or $x^0 \in \operatorname{rg} A^T$), then all iterates $x^k \in \operatorname{rg} A^T$
- Assume $\hat{x}$ solution in $\operatorname{rg} A^T$

Underdetermined systems

- Consider $Ax = b$, underdetermined but consistent
- Which solution does Kaczmarz pick?
- Initialization $x^0 = 0$ (or $x^0 \in \operatorname{rg} A^T$), then all iterates $x^k \in \operatorname{rg} A^T$
- Assume $\hat{x}$ solution in $\operatorname{rg} A^T$
- $Z \in \mathbf{R}^{n\times m}$, columns form ONB of $\operatorname{rg} A^T$, then $x^k = ZZ^T x^k$, $ZZ^T\hat{x} = \hat{x}$.

Underdetermined systems

- Consider $Ax = b$, underdetermined but consistent
- Which solution does Kaczmarz pick?
- Initialization $x^0 = 0$ (or $x^0 \in \operatorname{rg} A^T$), then all iterates $x^k \in \operatorname{rg} A^T$
- Assume $\hat{x}$ solution in $\operatorname{rg} A^T$
- $Z \in \mathbf{R}^{n\times m}$, columns form ONB of $\operatorname{rg} A^T$, then $x^k = ZZ^T x^k$, $ZZ^T\hat{x} = \hat{x}$.
- As above:

$$\mathbf{E}(\|x^k - \hat{x}\|^2) \leq (1-\lambda)^k \|x^0 - \hat{x}\|^2$$

$\lambda = \lambda_{\min}(Z^T A^T DAZ)$, $D = \operatorname{diag}(p_i / \|a_i\|^2)$

Underdetermined systems

- Consider $Ax = b$, underdetermined but consistent
- Which solution does Kaczmarz pick?
- Initialization $x^0 = 0$ (or $x^0 \in \operatorname{rg} A^T$), then all iterates $x^k \in \operatorname{rg} A^T$
- Assume $\hat{x}$ solution in $\operatorname{rg} A^T$
- $Z \in \mathbf{R}^{n\times m}$, columns form ONB of $\operatorname{rg} A^T$, then $x^k = ZZ^T x^k$, $ZZ^T\hat{x} = \hat{x}$.
- As above:

$$\mathbf{E}(\|x^k - \hat{x}\|^2) \leq (1-\lambda)^k \|x^0 - \hat{x}\|^2$$

  $\lambda = \lambda_{\min}(Z^T A^T DAZ)$, $D = \operatorname{diag}(p_i/\|a_i\|^2)$
- Convergence to minimum-norm solution $\hat{x}$

- The Kaczmarz method
- Randomization
- **Sparsity**
- Split feasibility problems
- Convergence rates

Kaczmarz converging to sparse solutions?

- Kaczmarz converges to (unique) solution in $x^0 + \operatorname{rg} A^T$ (if consistent)

Kaczmarz converging to sparse solutions?

- Kaczmarz converges to (unique) solution in $x^0 + \operatorname{rg} A^T$ (if consistent)
- This is the solution with min $\|x\|_2$

Kaczmarz converging to sparse solutions?

- Kaczmarz converges to (unique) solution in $x^0 + \operatorname{rg} A^T$ (if consistent)
- This is the solution with min $\|x\|_2$
- Convergence to other solutions? (e.g. min $\|x\|_1$)

Kaczmarz converging to sparse solutions?

- Kaczmarz converges to (unique) solution in $x^0 + \operatorname{rg} A^T$ (if consistent)
- This is the solution with min $\|x\|_2$
- Convergence to other solutions? (e.g. min $\|x\|_1$)
- Kaczmarz

$$x^{k+1} = x^k - \frac{a_i^T x_k - b_i}{\|a_i\|_2^2} a_i$$

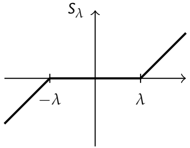
Technische Universität Braunschweig

Kaczmarz converging to sparse solutions?

- Kaczmarz converges to (unique) solution in $x^0 + \operatorname{rg} A^T$ (if consistent)
- This is the solution with min $\|x\|_2$
- Convergence to other solutions? (e.g. min $\|x\|_1$)
- Sparse Kaczmarz

$$z^{k+1} = z^k - \frac{a_i^T x_k - b_i}{\|a_i\|_2^2} a_i$$

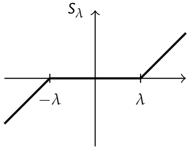
$$x^{k+1} = S_\lambda(z^{k+1})$$



Kaczmarz converging to sparse solutions?

- Kaczmarz converges to (unique) solution in $x^0 + \operatorname{rg} A^T$ (if consistent)
- This is the solution with min $\|x\|_2$
- Convergence to other solutions? (e.g. min $\|x\|_1$)
- Sparse Kaczmarz

$$z^{k+1} = z^k - \frac{a_i^T x_k - b_i}{\|a_i\|_2^2} a_i$$
$$x^{k+1} = S_\lambda(z^{k+1})$$



- **Theorem** [L, Schöpfer, Wenger, Magnor 2014]: The sequence x^k, when initialized with $x^0 = 0$, converges to the solution of

$$\min \|x\|_1 + \tfrac{1}{2\lambda}\|x\|_2^2 \ \text{ such that } \ Ax = b$$

  if every i appears infinitely often

Sparse Kaczmarz and linearized Bregman

$$z^{k+1} = z^k - \frac{a_{r(k)}^T x_k - b_{r(k)}}{\|a_{r(k)}\|_2^2} a_{r(k)}$$
$$x^{k+1} = S_\lambda(z^{k+1})$$

- Two interesting things:
  1. Very similar to Kaczmarz. Other “minimum-J-solutions” possible?
  2. Very similar to linearized Bregman iteration.

$$z^{k+1} = z^k - t_k A^T(Ax^k - b), \quad t_k \leq \frac{1}{\|A\|^2}$$

Sparse Kaczmarz and linearized Bregman

$$z^{k+1} = z^k - \frac{a_{r(k)}^T x_k - b_{r(k)}}{\|a_{r(k)}\|_2^2} a_{r(k)}$$
$$x^{k+1} = S_\lambda(z^{k+1})$$

- Two interesting things:
  1. Very similar to Kaczmarz. Other "minimum-J-solutions" possible?
  2. Very similar to linearized Bregman iteration.

$$z^{k+1} = z^k - t_k A^T(Ax^k - b), \quad t_k \leq \frac{1}{\|A\|^2}$$

- Approach taken here: "Split feasibility problems" will answer the first and explain the second point.

- The Kaczmarz method
- Randomization
- Sparsity
- **Split feasibility problems**
- Convergence rates

Convex split feasibility problems

- Split feasibility problem (SFP): Find x, such that

$$x \in \bigcap_{i=1}^{N_C} C_i, \qquad A_i x \in Q_i,\ \ i = 1, \dots, N_Q$$

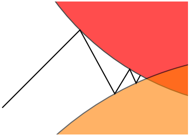
C_i, Q_i convex sets, A_i linear

Convex split feasibility problems

- Split feasibility problem (SFP): Find x, such that

$$x \in \bigcap_{i=1}^{N_C} C_i, \qquad A_i x \in Q_i,\ i = 1, \ldots, N_Q$$

  C_i, Q_i convex sets, A_i linear
- For a mere “feasibility problem”: Do alternating projections



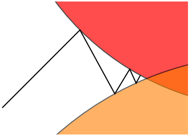
$x^{k+1} = P_{C_i}(x^k)$
$i = (k \mod N_C) + 1$ “control sequence”

Convex split feasibility problems

- Split feasibility problem (SFP): Find x, such that

$$x \in \bigcap_{i=1}^{N_C} C_i, \qquad A_i x \in Q_i,\ i = 1, \dots, N_Q$$

  C_i, Q_i convex sets, A_i linear
- For a mere "feasibility problem": Do alternating projections



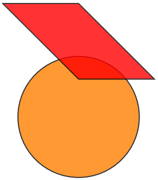
  $x^{k+1} = P_{C_i}(x^k)$
  $i = (k \mod N_C) + 1$ "control sequence"
- [1933 von Neumann (two subspaces), 1962 Halperin (several subspaces), Dijkstra, Censor, Combettes, Bauschke, Borwein, Deutsch, Lewis, Luke...]

Tackling split feasibility problems

- Projecting onto $\{x \mid Ax \in Q\}$: Project onto separating hyperplane

$$H^k = \{x \mid \langle Ax^k - P_Q(Ax^k), Ax - P_Q(Ax^k)\rangle \leq 0\}$$

(separates x^k from $\{x \mid Ax \in Q\}$)



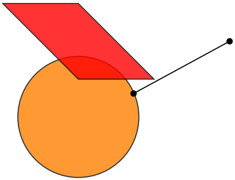
Tackling split feasibility problems

- Projecting onto $\{x \mid Ax \in Q\}$: Project onto separating hyperplane

$$H^k = \{x \mid \langle Ax^k - P_Q(Ax^k), Ax - P_Q(Ax^k)\rangle \leq 0\}$$

(separates x^k from $\{x \mid Ax \in Q\}$)

-
  - $x^{k+1} = P_{C_i}(x^k)$ for a constraint $u \in C_i$
  - $x^{k+1} = P_{H^k}(x^k)$ for a constraint $A_i x \in Q_i$



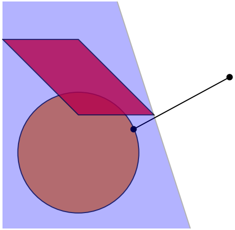
Tackling split feasibility problems

- Projecting onto $\{x \mid Ax \in Q\}$: Project onto separating hyperplane

$$H^k = \{x \mid \langle Ax^k - P_Q(Ax^k), Ax - P_Q(Ax^k)\rangle \leq 0\}$$

(separates x^k from $\{x \mid Ax \in Q\}$)

-
  - $x^{k+1} = P_{C_i}(x^k)$
    for a constraint $u \in C_i$
  - $x^{k+1} = P_{H^k}(x^k)$
    for a constraint $A_i x \in Q_i$



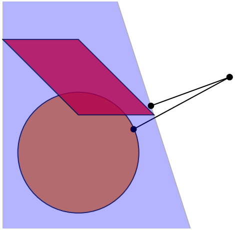
Tackling split feasibility problems

- Projecting onto $\{x \mid Ax \in Q\}$: Project onto separating hyperplane

$$H^k = \{x \mid \langle Ax^k - P_Q(Ax^k), Ax - P_Q(Ax^k)\rangle \leq 0\}$$

  (separates x^k from $\{x \mid Ax \in Q\}$)

-
  - $x^{k+1} = P_{C_i}(x^k)$
    for a constraint $u \in C_i$
  - $x^{k+1} = P_{H^k}(x^k)$
    for a constraint $A_i x \in Q_i$



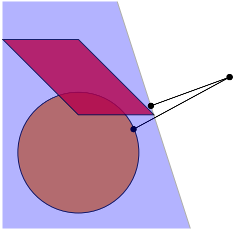
Tackling split feasibility problems

- Projecting onto $\{x \mid Ax \in Q\}$: Project onto separating hyperplane

$$H^k = \{x \mid \langle Ax^k - P_Q(Ax^k), Ax - P_Q(Ax^k)\rangle \leq 0\}$$

(separates x^k from $\{x \mid Ax \in Q\}$)

-
  - $x^{k+1} = P_{C_i}(x^k)$ for a constraint $u \in C_i$
  - $x^{k+1} = P_{H^k}(x^k)$ for a constraint $A_i x \in Q_i$



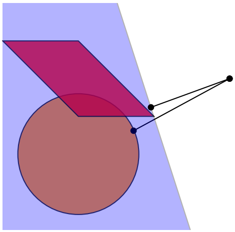
Tackling split feasibility problems

- Projecting onto $\{x \mid Ax \in Q\}$: Project onto separating hyperplane

$$H^k = \{x \mid \langle Ax^k - P_Q(Ax^k), Ax - P_Q(Ax^k)\rangle \leq 0\}$$

(separates x^k from $\{x \mid Ax \in Q\}$)

-
  - $x^{k+1} = P_{C_i}(x^k)$
    for a constraint $u \in C_i$
  - $x^{k+1} = P_{H^k}(x^k)$
    for a constraint $A_i x \in Q_i$
- Converges to feasible point.



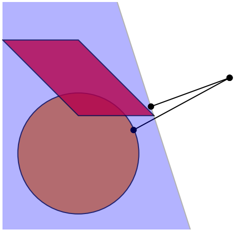
Tackling split feasibility problems

- Projecting onto $\{x \mid Ax \in Q\}$: Project onto separating hyperplane

$$H^k = \{x \mid \langle Ax^k - P_Q(Ax^k), Ax - P_Q(Ax^k)\rangle \le 0\}$$

  (separates x^k from $\{x \mid Ax \in Q\}$)

- - $x^{k+1} = P_{C_i}(x^k)$
    for a constraint $u \in C_i$
  - $x^{k+1} = P_{H^k}(x^k)$
    for a constraint $A_i x \in Q_i$
- Converges to feasible point.
- E.g.: $Q = \{b\}$: $x^{k+1} = x^k + t_k A^T(Ax^k - b)$
  $\rightsquigarrow$ minimum norm solution of $Ax = b$

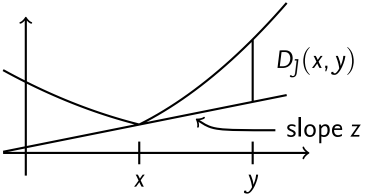


Towards sparse solutions with Bregman projections

- $P_C(x) = \operatorname{argmin}_{y \in C} \|x - y\|^2 \rightsquigarrow$ orthogonal projection
- $J : X \to \mathbf{R}$ convex, $z \in \partial J(x)$

$$D^z(x, y) = J(y) - J(x) - \langle z, y - x \rangle$$

Bregman distance



- Bregman distances $\rightsquigarrow$ *Bregman projection*:
$\Pi_C^z(x) = \operatorname{argmin}_{y \in C} D_J^z(x, y)$

Bregman projections

- Assume $J : \mathbf{R}^n \to \mathbf{R}$ continuous, α-strongly convex ($\implies \nabla J^*$ is α^{-1}-Lipschitz)
- Bregman projections onto hyperplanes $H = \{a^T x = \beta\}$ are simple: if $z \in \partial J(x)$

$$\Pi_H^z(x) = \nabla J^*(z - \bar{t}a), \quad \bar{t} = \operatorname*{argmin}_t J^*(z - ta) + t\beta$$

  Moreover: $z - \bar{t}a \in \partial J(\Pi_H^z(x))$ new subgradient in $\Pi_H^z(x)$.
- **RBPSFP**: Random Bregman projections for SFP $x \in \cap C_i$, $A_i x \in Q_i$:
  - Initialize $z_0 \in \partial J(x_0)$
  - $x^{k+1} = \Pi_{C_i}^{z^k}(x^k)$ or $x^{k+1} = \Pi_{H^i}^{z^k}(x^k)$, update $z^k \in \partial J(x^k)$
- random: every index appears infinitely often

Convergence

- **Theorem:** [Schöpfer, L., Wenger 2014] RBPSFP converges to a feasible point $\bar{x} \in C := \bigcap C_i \cap \{x \mid A_i x \in Q_i\}$.

Convergence

- **Theorem:** [Schöpfer, L., Wenger 2014] RBPSFP converges to a feasible point $\bar{x} \in C := \bigcap C_i \cap \{x \mid A_i x \in Q_i\}$.
- Application to
$$\min J(x) \text{ s.t. } Ax = b$$
Multiple possibilities, e.g.
  1. only one “difficult constraint”: $Ax \in Q = \{b\}$
  2. many simple constraints $C_i = \{a_i^T x = b_i\}$

Convergence

- **Theorem:** [Schöpfer, L., Wenger 2014] RBPSFP converges to a feasible point $\bar{x} \in C := \bigcap C_i \cap \{x \mid A_i x \in Q_i\}$.
- Application to

$$\min J(x) \text{ s.t. } Ax = b$$

  Multiple possibilities, e.g.

  1. only one "difficult constraint": $Ax \in Q = \{b\}$
  2. many simple constraints $C_i = \{a_i^T x = b_i\}$

- In both cases: Convergence to minimum-J solution

Sparse solutions

- $J(x) = \lambda \|x\|_1$ does not work - not strongly convex

Sparse solutions

- $J(x) = \lambda\|x\|_1$ does not work - not strongly convex
- $J(x) = \lambda\|x\|_1 + \frac{1}{2}\|x\|^2$: strongly convex with constant 1

Sparse solutions

- $J(x) = \lambda\|x\|_1$ does not work - not strongly convex
- $J(x) = \lambda\|x\|_1 + \frac{1}{2}\|x\|^2$: strongly convex with constant 1
- Bregman projection onto hyperplanes $H = \{a^T x = \beta\}$: if $z \in \partial J(x)$

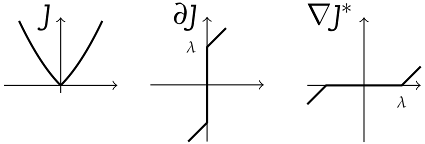
$$\Pi_H^z(x) = \nabla J^*(z - \bar{t}a), \quad \bar{t} = \operatorname*{argmin}_t J^*(z - ta) + t\beta$$

Sparse solutions

- $J(x) = \lambda\|x\|_1$ does not work - not strongly convex
- $J(x) = \lambda\|x\|_1 + \frac{1}{2}\|x\|^2$: strongly convex with constant 1
- Bregman projection onto hyperplanes $H = \{a^T x = \beta\}$: if $z \in \partial J(x)$

$$\Pi_H^z(x) = \nabla J^*(z - \bar{t}a), \quad \bar{t} = \operatorname*{argmin}_t J^*(z - ta) + t\beta$$

- $\nabla J^*(x) = (\partial J)^{-1}(x) = S_\lambda(x)$:



Basic algorithm and special cases:

- Variant 1: One difficult constraint $Ax = b$
- Variant 2: Many simple constraints $a_r^T x = b_r$
- In general: Block-processing $A_r x = b_r$

Basic algorithm and special cases:

- Variant 1: One difficult constraint $Ax = b$
- Variant 2: Many simple constraints $a_r^T x = b_r$
- In general: Block-processing $A_r x = b_r$

Iteration:

- Calculate

$$
\begin{aligned}
z^{k+1} &= z^k - t_k A_r^T w^k \\
x^{k+1} &= \nabla J^*(z^{k+1})
\end{aligned}
$$

with appropriate stepsize t_k (depending on w^k and β_k)

Basic algorithm and special cases:

- Variant 1: One difficult constraint $Ax = b$
- Variant 2: Many simple constraints $a_r^T x = b_r$
- In general: Block-processing $A_r x = b_r$

Iteration:

- Calculate

$$z^{k+1} = z^k - t_k A_r^T w^k$$
$$x^{k+1} = \nabla J^*(z^{k+1})$$

with appropriate stepsize t_k (depending on w^k and β_k)

- $J(x) = \|x\|_2^2/2$, variant 1.: Landweber iteration
- $J(x) = \|x\|_2^2/2$, variant 2.: Kaczmarz method
- $J(x) = \lambda\|x\|_1 + \|x\|_2^2/2$, variant 1.: Linearized Bregman
- $J(x) = \lambda\|x\|_1 + \|x\|_2^2/2$, variant 2.: Sparse Kaczmarz

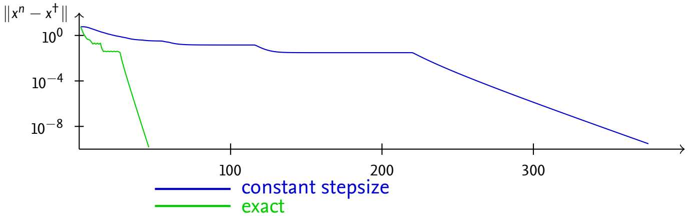
Inexact stepsizes are allowed

- Instead of projecting exactly, it suffices to move close enough
- Linearized Bregman:

$$t_k = \frac{\|Ax^k - b\|^2}{\|A^T(Ax^k - b)\|^2}, \quad \text{or} \quad t_k \leq \frac{1}{\|A\|^2}$$

- However: To compute exact stepsize, solve one-dimensional piecewise quadratic optimization problem (for $J(x) = \lambda\|x\|_1 + \|x\|_2^2/2$ can be done in $\mathcal{O}(n \log n)$, usually faster).

Stepsize comparison - linearized Bregman



$A \in \mathbf{R}^{1000\times 2000}$ Gaussian distriuted entries, $x^\dagger$ 20 non-zeros (also Gausian distributed)

- The Kaczmarz method
- Randomization
- Sparsity
- Split feasibility problems
- **Convergence rates**

Convergence rates for RBPSFP

Theorem (Schöpfer, L. 2018)

RBPSFB with $C = \bigcap C_i \cap \{x \mid A_i x \in Q_i\}$ *converges with a rate*

$$\mathbf{E}(\operatorname{dist}(x^k, C)) = \mathcal{O}(1/\sqrt{k})$$

if $\{C_k\}_k$ *and each* $\{Q_i, \operatorname{rg}(A_i)\}$ *is boundedly linearly regular and* J *is strongly convex.*
If, additionally, J *is piecewise linear quadratic, then method converges linearly, i.e.*

$$\mathbf{E}(\operatorname{dist}(x^k, C)) = \mathcal{O}(q^k).$$

Proof based on error bounds...
Corollary: The randomized sparse Kaczmarz (RaSK) method converges linearly.

Taylores results for randomized sparse Kaczmarz

Theorem (Schöpfer, L. 2018)

For RaSK with exact steps (ERaSK) for a consistent overdetermined system $Ax = b$ it holds that

$$\mathbf{E}(\|x^k - x^*\|_2) \le (1-\epsilon)^{k/2}\sqrt{2\lambda\|\hat{x}\|_1 + \|\hat{x}\|_2^2}$$

with

$$\epsilon = \frac{\tilde{\sigma}^2_{\min}(A)}{2\|A\|_F^2}\frac{|\hat{x}|_{\min}}{|\hat{x}|_{\min} + 2\lambda}$$

where $\tilde{\sigma}_{\min} = \min\{\sigma_{\min}(A_J) \mid A_J \neq 0 \textit{ submatrix}\}$,
$|\hat{x}|_{\min} = \min\{|\hat{x}_j| \mid \hat{x}_j \neq 0\}$.

Randomized sparse Kaczmarz for noisy data

Following [Needell 2010] and [Lai, Yin 2013]:

Theorem

For $Ax = b^\delta$ with $\|b^\delta - b\|_2 \leq \delta$ it holds for RaSK

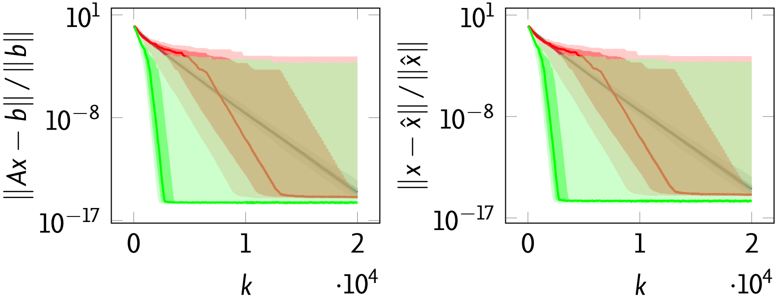
$$\mathbf{E}(\|x^k - x^*\|_2 \leq (1-\epsilon)^{k/2}\sqrt{2\lambda\|\hat{x}\|_1 + \|\hat{x}\|_2^2} + \sqrt{\frac{2|\hat{x}|_{\min} + 4\lambda}{|\hat{x}|_{\min}}}\frac{\delta}{\tilde{\sigma}_{\min}(A)}$$

and for ERaSK the upper bound is

$$(1-\epsilon)^{k/2}\sqrt{2\lambda\|\hat{x}\|_1 + \|\hat{x}\|_2^2} + \sqrt{\frac{2|\hat{x}|_{\min} + 4\lambda}{|\hat{x}|_{\min}}}\frac{\delta}{\tilde{\sigma}_{\min}(A)}\sqrt{1 + \frac{4\|A\|_{2,1}}{\delta}}$$

Sparsity also helps for overdetermined systems

- 200 columns, 1000 rows, **consistent** system $Ax = b$, unique solution $x^\dagger$, $\mathrm{nnz}(x^\dagger) = 25$

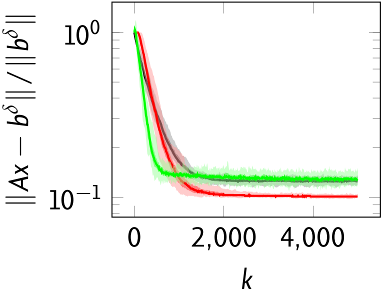


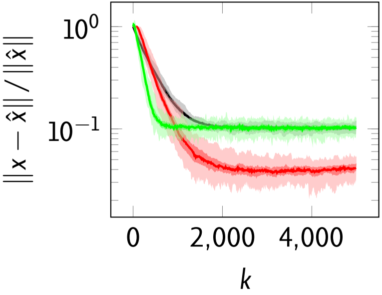
Black: Randomized Kaczmarz, Red: Randomized sparse Kaczmarz, Green: Exact-step randomized sparse Kaczmarz

Sparsity also helps for overdetermined systems

- 200 columns, 1000 rows, **inconsistent** system $Ax = b$,

10% relative error

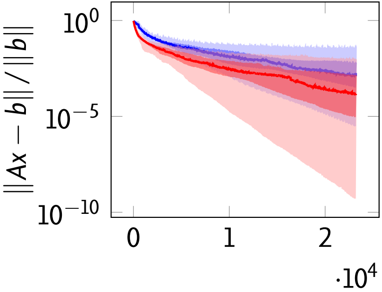


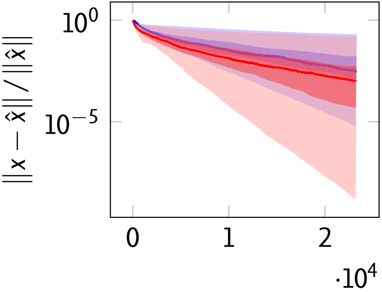


Black: Randomized Kaczmarz, Red: Randomized sparse Kaczmarz,
Green: Exact-step randomized sparse Kaczmarz

Randomization also helps

- Matrix from fan-beam CT, consistent system $Ax = b$, unique solution $x^\dagger$, 100 columns, 1164 rows, $\mathrm{nnz}(x^\dagger) = 20$

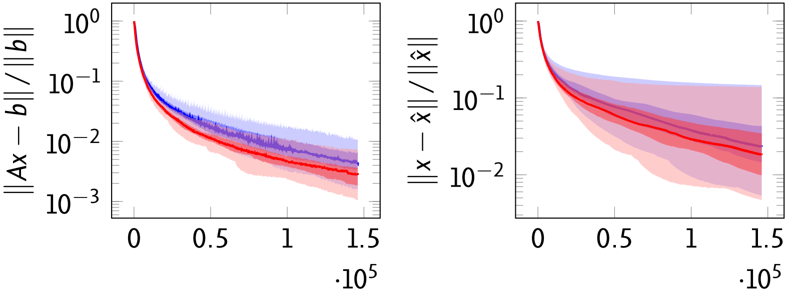




Blue: Sparse Kaczmarz, Red: Randomized sparse Kaczmarz

Randomization also helps

- Matrix from fan-beam CT, consistent system $Ax = b$, unique solution $x^\dagger$, 900 columns, 3660 rows, $\mathrm{nnz}(x^\dagger) = 180$



Blue: Sparse Kaczmarz, Red: Randomized sparse Kaczmarz

Conclusion

- Randomization gives uniform expected progress,
  hence convergence rates
- Randomization usually improves (random reshuffle also works)
- Extension to sparse solutions simple; exact stepsize matters, though
- Convergence of RaSK and ERaSK linear
- Exacts steps faster, lower accuracy for noisy data

References

- Dirk A. Lorenz, Frank Schöpfer, and Stephan Wenger, *The linearized Bregman method via split feasibility problems: Analysis and generalizations*, SIAM Journal on Imaging Sciences **2** (2014), no. 7, 1237–1262, [doi, arXiv].
- Dirk A Lorenz, Stephan Wenger, Frank Schöpfer, and Marcus Magnor, *A sparse Kaczmarz solver and a linearized Bregman method for online compressed sensing*, Image Processing (ICIP), 2014 IEEE International Conference on, IEEE, 2014, [doi, arXiv], pp. 1347–1351.
- Frank Schöpfer and Dirk A. Lorenz, *Linear convergence of the Randomized Sparse Kaczmarz method*, To appear in *Mathematical Programming*, 2018, [doi, arXiv].