Randomized Kaczmarz with Averaging

Jacob Moorman, Thomas Tu, **Denali Molitor**, Deanna Needell

March 8, 2019

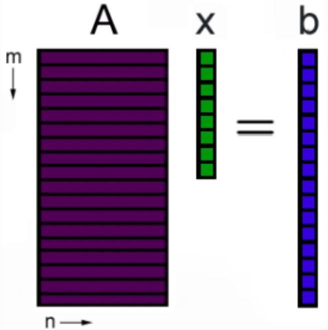
University of California, Los Angeles
Department of Mathematics

Problem of interest - Solving large linear systems

Goal: Solve large linear systems of the form

$$\mathbf{A}x = b,$$

where $\mathbf{A} \in \mathbb{R}^{m \times n}$ with $m \gg n$, $x \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$.

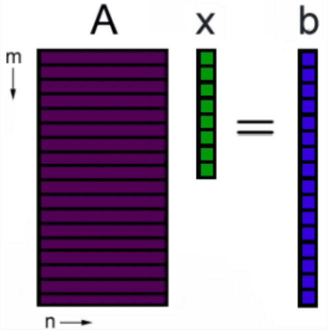


Problem of interest - Solving large linear systems

Goal: Solve large linear systems of the form

$$\mathbf{A}x = b,$$

where $\mathbf{A} \in \mathbb{R}^{m \times n}$ with $m \gg n$, $x \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$.



For large linear systems, it can be more efficient to use iterative methods than solving directly.

Outline

- Kaczmarz and Randomized Kaczmarz (RK)
- Relation to stochastic gradient descent
- Convergence of RK
- Parallel RK with averaging
- Experimental performance
- Hyperparameter optimization

Kaczmarz's Method

Kaczmarz:

Iteratively project onto the solution space with respect to the i^{th} row.

$$x^{k+1} = x^k - \frac{\mathbf{A}_i x^k - b_i}{\|\mathbf{A}_i\|^2} \mathbf{A}_i^\top,$$

where $i = k \mod m + 1$.

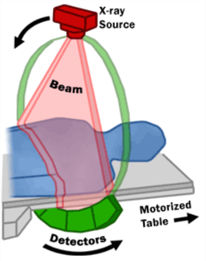
$\mathbf{A}_i$ denotes the i^{th} row of $\mathbf{A}$ (i.e. cycle through the rows of $\mathbf{A}$).

Proposed in 1937 by Stefan Kaczmarz.

Use in Medical Imaging

Rediscovered in 1970 as the Algebraic Reconstruction Technique (ART) by Gordon, Bender and Herman for Computed Tomography (CT).

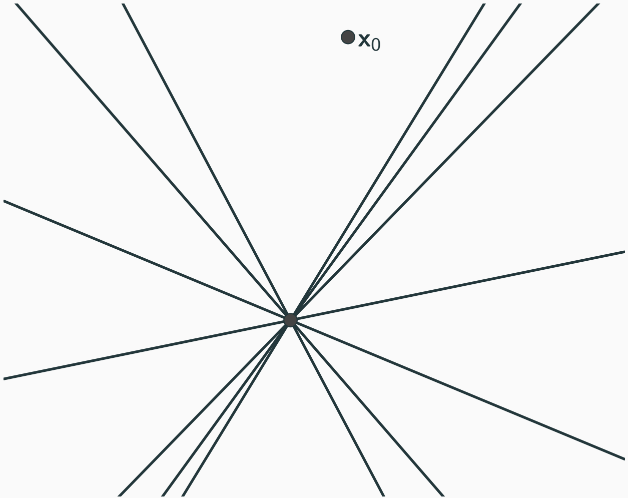
Implemented in a medical scanner in 1972.



images: www.fda.gov/radiation-emittingproducts/radiationemittingproductsandprocedures/medicalimaging/medicalx-rays/ucm115317.htm

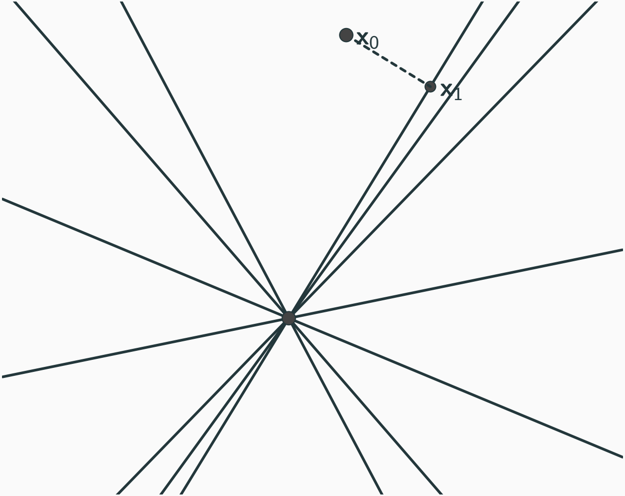
Kaczmarz Method

Lines represent solution spaces $H_i = \{x : \mathbf{A}_i x = b_i\}$.



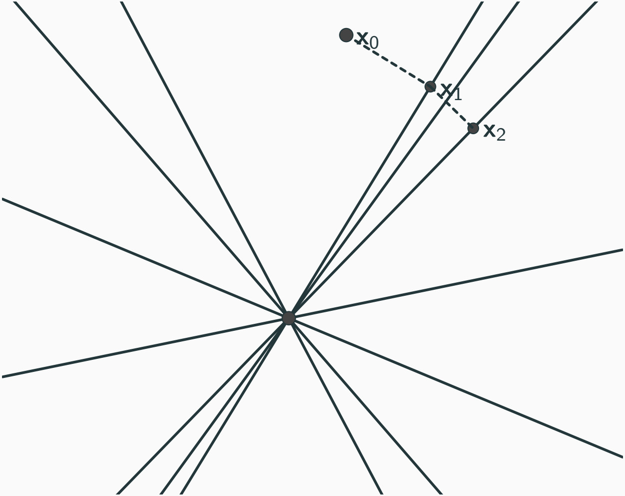
Kaczmarz Method

Lines represent solution spaces $H_i = \{x : \mathbf{A}_i x = b_i\}$.



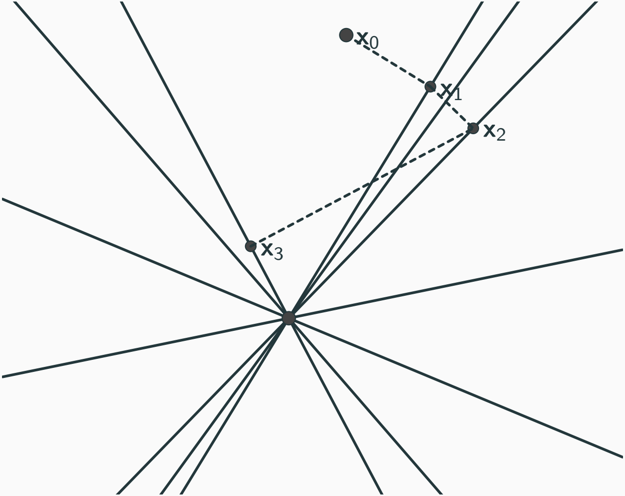
Kaczmarz Method

Lines represent solution spaces $H_i = \{x : \mathbf{A}_i x = b_i\}$.



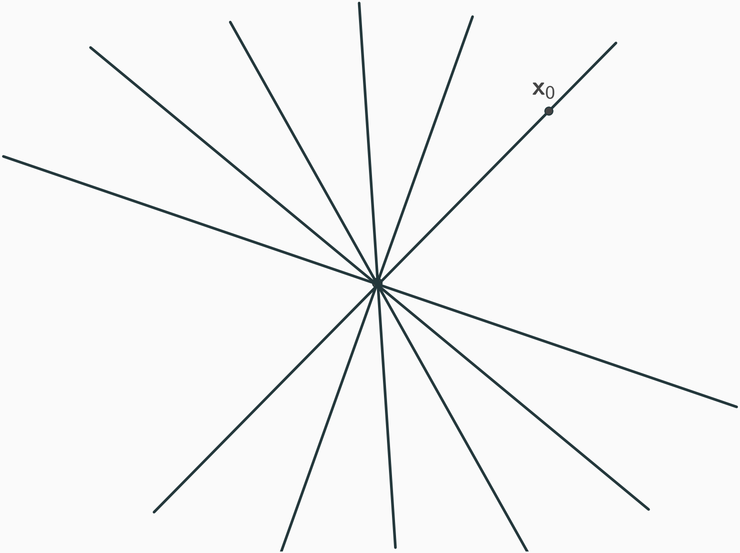
Kaczmarz Method

Lines represent solution spaces $H_i = \{x : \mathbf{A}_i x = b_i\}$.



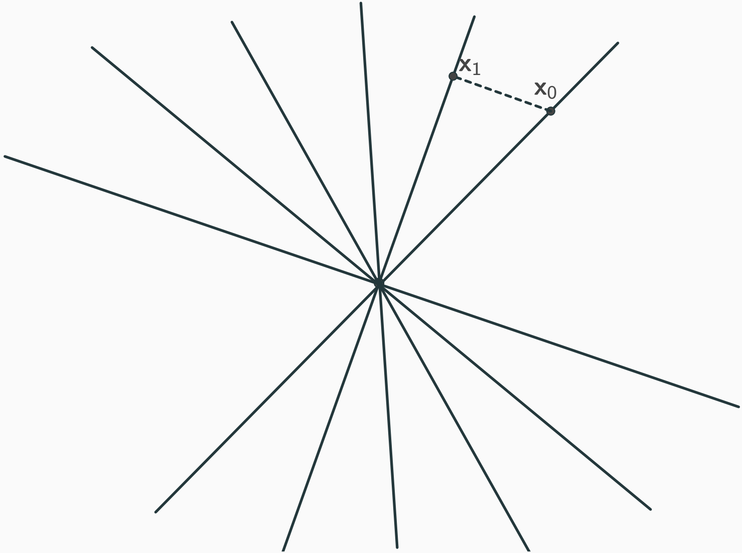
An Unlucky Ordering

Lines represent solution spaces $H_i = \{x : \mathbf{A}_i x = b_i\}$.



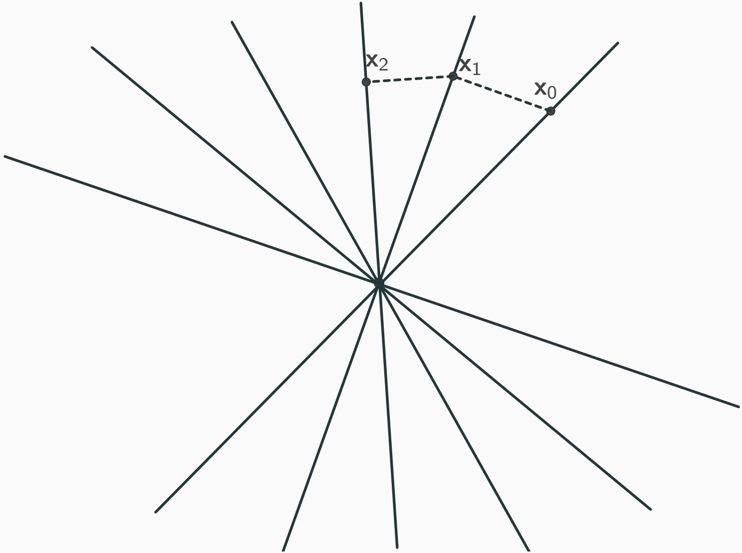
An Unlucky Ordering

Lines represent solution spaces $H_i = \{x : \mathbf{A}_i x = b_i\}$.



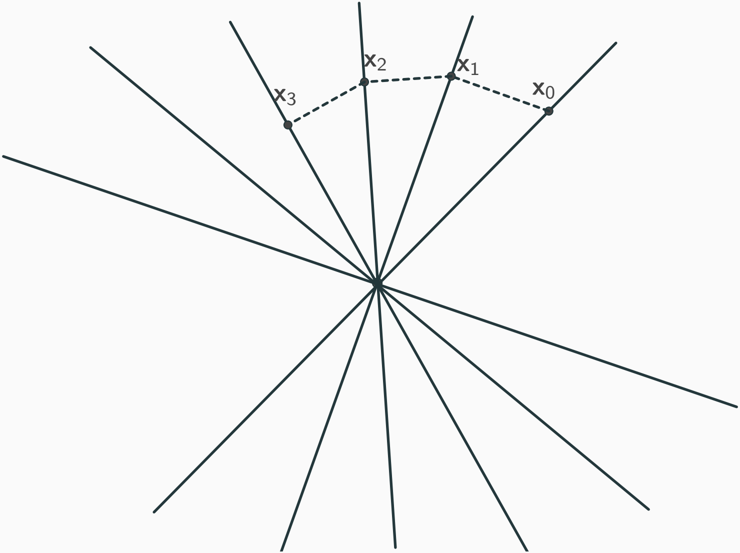
An Unlucky Ordering

Lines represent solution spaces $H_i = \{x : \mathbf{A}_i x = b_i\}$.



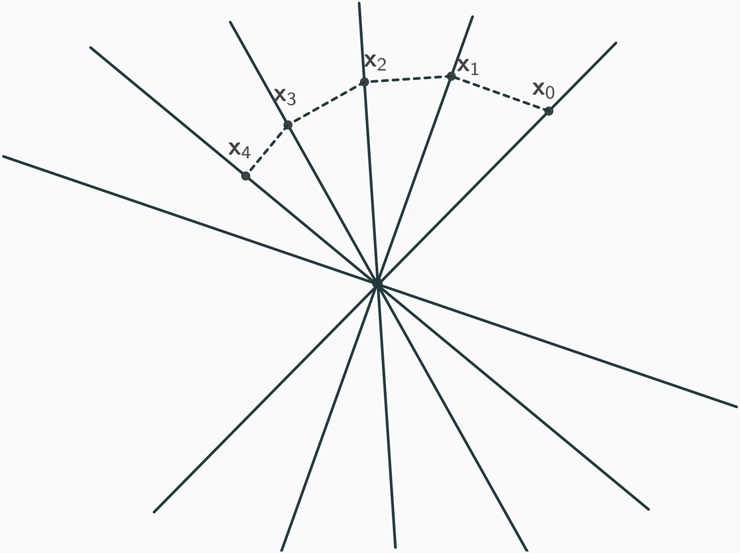
An Unlucky Ordering

Lines represent solution spaces $H_i = \{x : \mathbf{A}_i x = b_i\}$.



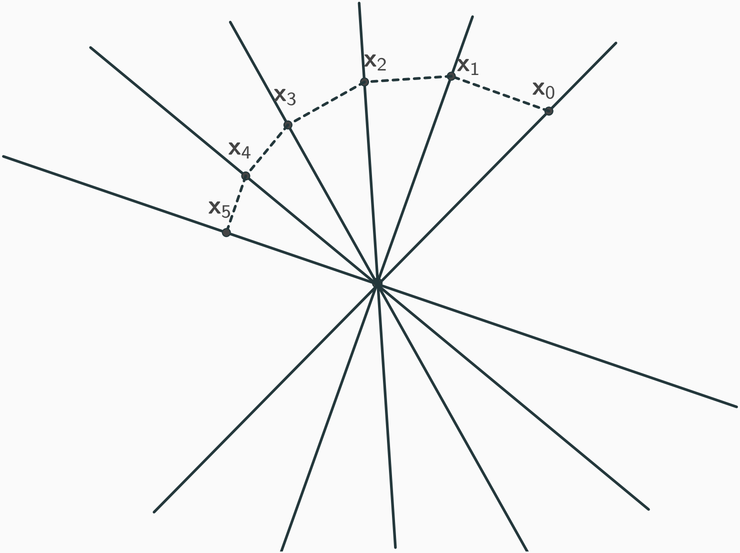
An Unlucky Ordering

Lines represent solution spaces $H_i = \{x : \mathbf{A}_i x = b_i\}$.



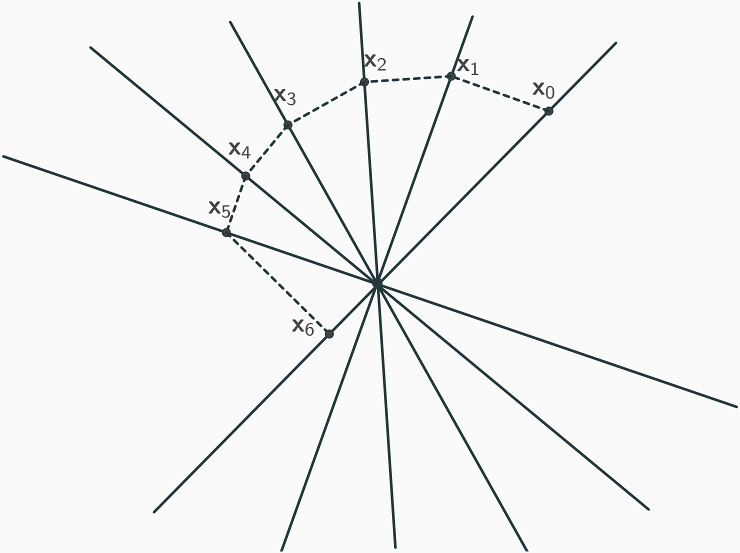
An Unlucky Ordering

Lines represent solution spaces $H_i = \{x : \mathbf{A}_i x = b_i\}$.



An Unlucky Ordering

Lines represent solution spaces $H_i = \{x : \mathbf{A}_i x = b_i\}$.



Randomized Kaczmarz

Randomized Kaczmarz (RK):
Iteratively project onto the solution space with respect to a single row.

$$x^{k+1} = x^k - \frac{\mathbf{A}_{i_k} x^k - b_{i_k}}{\left\|\mathbf{A}_{i_k}\right\|^2} \mathbf{A}_{i_k}^\top,$$

where $i_k \sim \mathcal{D}$.

$\mathbf{A}_i$ denotes the i^{th} row of $\mathbf{A}$.

- Randomization allows us to avoid being stuck with particularly bad orderings
- Don't get to take advantage of good orderings either though
- Often combined with heuristic choices, eg. sampling without replacement

Relation to stochastic gradient descent

Stochastic gradient descent (SGD)

Loss functions often take the form

$$F(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

Stochastic gradient descent (SGD)

Loss functions often take the form

$$F(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x)$$

Gradient descent:

$$x^{k+1} = x^k - \eta \nabla F(x^k),$$

where η is a step size.

Stochastic gradient descent (SGD)

Loss functions often take the form

$$F(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

Gradient descent:

$$x^{k+1} = x^k - \eta \nabla F(x^k),$$

where η is a step size.

Stochastic gradient descent:

$$x^{k+1} = x^k - \eta \nabla f_i(x^k).$$

Stochastic gradient descent (SGD)

Loss functions often take the form

$$F(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

Gradient descent:

$$x^{k+1} = x^k - \eta \nabla F(x^k),$$

where η is a step size.
Stochastic gradient descent:

$$x^{k+1} = x^k - \eta \nabla f_i(x^k).$$

$\mathbb{E}\left[\nabla f_i(x^k)\right] = \nabla F(x^k).$

Randomized Kaczmarz and SGD

Needell, Srebro, and Ward 2015

Randomized Kaczmarz can be viewed as reweighted SGD with importance sampling applied to

$$F(x) = \frac{1}{2}||\mathbf{A}x - b||_2^2 = \sum_i \frac{1}{2}(\mathbf{A}_i x - b_i)^2.$$

Randomized Kaczmarz and SGD

Needell, Srebro, and Ward 2015

Randomized Kaczmarz can be viewed as **reweighted** SGD with **importance sampling** applied to

$$F(x) = \frac{1}{2}||\mathbf{A}x - b||_2^2 = \sum_i \frac{1}{2}(\mathbf{A}_i x - b_i)^2.$$

Importance sampling: Pick i with probability p_i.

Reweighted SGD: Use the weighted update

$$x^{k+1} = x^k - \frac{\eta}{np_i}\nabla f_i(x^k).$$

Randomized Kaczmarz and SGD

Note:

$$\nabla f_i(x) = \mathbf{A}_i^\top (\mathbf{A}_i x - b_i).$$

Randomized Kaczmarz and SGD

Note:

$$\nabla f_i(x) = \mathbf{A}_i^\top(\mathbf{A}_i x - b_i).$$

Choosing

$$p_i = \frac{||\mathbf{A}_i||_2^2}{||\mathbf{A}||_F^2},$$

reweighted SGD becomes

$$\begin{aligned} x^{k+1} &= x^k - \frac{\eta||\mathbf{A}||_F^2}{n||\mathbf{A}_i||_2^2}\nabla f_i(x^k) \\ &= x^k - \eta\frac{||\mathbf{A}||_F^2}{n}\frac{\mathbf{A}_i^\top(\mathbf{A}_i x - b_i)}{||\mathbf{A}_i||_2^2}. \end{aligned}$$

Randomized Kaczmarz and SGD

Note:

$$\nabla f_i(x) = \mathbf{A}_i^\top (\mathbf{A}_i x - b_i).$$

Choosing

$$p_i = \frac{||\mathbf{A}_i||_2^2}{||\mathbf{A}||_F^2},$$

reweighted SGD becomes

$$\begin{aligned} x^{k+1} &= x^k - \frac{\eta ||\mathbf{A}||_F^2}{n||\mathbf{A}_i||_2^2} \nabla f_i(x^k) \\ &= x^k - \eta \frac{||\mathbf{A}||_F^2}{n} \frac{\mathbf{A}_i^\top (\mathbf{A}_i x - b_i)}{||\mathbf{A}_i||_2^2}. \end{aligned}$$

Convergence of RK

Convergence of RK [Strohmer - Vershynin 2009]

$$x^{k+1} = x^k - \frac{\mathbf{A}_{i_k}^\top \left(\mathbf{A}_{i_k} x^k - b_{i_k}\right)}{\left\|\mathbf{A}_{i_k}\right\|^2}$$

Convergence of RK [Strohmer - Vershynin 2009]

$$x^{k+1} - x^\star = x^k - x^\star - \frac{\mathbf{A}_{i_k}^\top \left(\mathbf{A}_{i_k} x^k - b_{i_k}\right)}{\left\|\mathbf{A}_{i_k}\right\|^2}$$

Convergence of RK [Strohmer - Vershynin 2009]

$$
\begin{aligned}
x^{k+1} - x^\star &= x^k - x^\star - \frac{\mathbf{A}_{i_k}^\top \left(\mathbf{A}_{i_k} x^k - b_{i_k}\right)}{\left\|\mathbf{A}_{i_k}\right\|^2} \\
&= x^k - x^\star - \frac{\mathbf{A}_{i_k}^\top \mathbf{A}_{i_k} \left(x^k - x^\star\right)}{\left\|\mathbf{A}_{i_k}\right\|^2}
\end{aligned}
$$

Convergence of RK [Strohmer - Vershynin 2009]

$$
\begin{aligned}
x^{k+1} - x^\star &= x^k - x^\star - \frac{\mathbf{A}_{i_k}^\top \left(\mathbf{A}_{i_k} x^k - b_{i_k}\right)}{\left\|\mathbf{A}_{i_k}\right\|^2} \\
&= x^k - x^\star - \frac{\mathbf{A}_{i_k}^\top \mathbf{A}_{i_k} \left(x^k - x^\star\right)}{\left\|\mathbf{A}_{i_k}\right\|^2} \\
&= \left(\mathbf{I} - \frac{\mathbf{A}_{i_k}^\top \mathbf{A}_{i_k}}{\left\|\mathbf{A}_{i_k}\right\|^2}\right) \left(x^k - x^\star\right).
\end{aligned}
$$

Convergence of RK [Strohmer - Vershynin 2009]

$$
\begin{aligned}
x^{k+1} - x^\star &= x^k - x^\star - \frac{\mathbf{A}_{i_k}^\top \left(\mathbf{A}_{i_k} x^k - b_{i_k}\right)}{\left\|\mathbf{A}_{i_k}\right\|^2} \\
&= x^k - x^\star - \frac{\mathbf{A}_{i_k}^\top \mathbf{A}_{i_k} \left(x^k - x^\star\right)}{\left\|\mathbf{A}_{i_k}\right\|^2} \\
&= \left(\mathbf{I} - \frac{\mathbf{A}_{i_k}^\top \mathbf{A}_{i_k}}{\left\|\mathbf{A}_{i_k}\right\|^2}\right)\left(x^k - x^\star\right).
\end{aligned}
$$

Taking the norm,

$$
\|x^{k+1} - x^\star\|_2^2 = \left\| \left(\mathbf{I} - \frac{\mathbf{A}_{i_k}^\top \mathbf{A}_{i_k}}{\left\|\mathbf{A}_{i_k}\right\|^2}\right)\left(x^k - x^\star\right) \right\|_2^2
$$

Convergence of RK [Strohmer - Vershynin 2009]

$$\begin{aligned}
x^{k+1} - x^\star &= x^k - x^\star - \frac{\mathbf{A}_{i_k}^\top \left(\mathbf{A}_{i_k} x^k - b_{i_k}\right)}{\left\|\mathbf{A}_{i_k}\right\|^2} \\
&= x^k - x^\star - \frac{\mathbf{A}_{i_k}^\top \mathbf{A}_{i_k} \left(x^k - x^\star\right)}{\left\|\mathbf{A}_{i_k}\right\|^2} \\
&= \left(\mathbf{I} - \frac{\mathbf{A}_{i_k}^\top \mathbf{A}_{i_k}}{\left\|\mathbf{A}_{i_k}\right\|^2}\right) \left(x^k - x^\star\right).
\end{aligned}$$

Taking the norm,

$$\begin{aligned}
\left\|x^{k+1} - x^\star\right\|_2^2 &= \left\| \left(\mathbf{I} - \frac{\mathbf{A}_{i_k}^\top \mathbf{A}_{i_k}}{\left\|\mathbf{A}_{i_k}\right\|^2}\right) \left(x^k - x^\star\right) \right\|_2^2 \\
&= \left\|x^k - x^\star\right\|_2^2 - \left\| \frac{\mathbf{A}_{i_k}^\top \mathbf{A}_{i_k}}{\left\|\mathbf{A}_{i_k}\right\|^2} \left(x^k - x^\star\right) \right\|_2^2,
\end{aligned}$$

where the last step is by orthogonality.

Convergence of RK [Strohmer - Vershynin 2009]

If we select row i with probability $p_i = \frac{\|\mathbf{A}_i\|_2^2}{\|\mathbf{A}\|_F^2}$, taking the expectation conditioned on x^k we have

$$\mathbb{E}\left[\|x^{k+1} - x^\star\|_2^2 \mid x^k\right] = \mathbb{E}\left[\|x^k - x^\star\|_2^2 - \left\|\frac{\mathbf{A}_{i_k}^\top \mathbf{A}_{i_k}}{\|\mathbf{A}_{i_k}\|^2}\left(x^k - x^\star\right)\right\|_2^2 \middle| x^k\right]$$

.

Convergence of RK [Strohmer - Vershynin 2009]

If we select row i with probability $p_i = \frac{\|\mathbf{A}_i\|_2^2}{\|\mathbf{A}\|_F^2}$, taking the expectation conditioned on x^k we have

$$\begin{aligned}\mathbb{E}\left[\|x^{k+1} - x^\star\|_2^2 \mid x^k\right] &= \mathbb{E}\left[\|x^k - x^\star\|_2^2 - \left\|\frac{\mathbf{A}_{i_k}^\top \mathbf{A}_{i_k}}{\|\mathbf{A}_{i_k}\|^2}\left(x^k - x^\star\right)\right\|_2^2 \middle| x^k\right] \\ &= \|x^k - x^\star\|_2^2 - \sum_i \frac{\|\mathbf{A}_i\|_2^2}{\|\mathbf{A}\|_F^2}\left\|\frac{\mathbf{A}_i^\top \mathbf{A}_i}{\|\mathbf{A}_i\|_2^2}\left(x^k - x^\star\right)\right\|_2^2.\end{aligned}$$

Convergence of RK [Strohmer - Vershynin 2009]

If we select row i with probability $p_i = \frac{\|\mathbf{A}_i\|_2^2}{\|\mathbf{A}\|_F^2}$, taking the expectation conditioned on x^k we have

$$\mathbb{E}\left[\|x^{k+1} - x^\star\|_2^2 \mid x^k\right] = \mathbb{E}\left[\|x^k - x^\star\|_2^2 - \left\|\frac{\mathbf{A}_{i_k}^\top \mathbf{A}_{i_k}}{\|\mathbf{A}_{i_k}\|^2}\left(x^k - x^\star\right)\right\|_2^2 \Big| x^k\right]$$

$$= \|x^k - x^\star\|_2^2 - \sum_i \frac{\|\mathbf{A}_i\|_2^2}{\|\mathbf{A}\|_F^2}\left\|\frac{\mathbf{A}_i^\top \mathbf{A}_i}{\|\mathbf{A}_i\|_2^2}\left(x^k - x^\star\right)\right\|_2^2.$$

Note that

$$\left\|\frac{\mathbf{A}_i^\top \mathbf{A}_i}{\|\mathbf{A}_i\|_2^2}\left(x^k - x^\star\right)\right\|_2^2 = \frac{\|\mathbf{A}_i\|_2^2\left(\mathbf{A}_i(x^k - x^\star)\right)^2}{\|\mathbf{A}_i\|_2^4} = \frac{\left(\mathbf{A}_i(x^k - x^\star)\right)^2}{\|\mathbf{A}_i\|_2^2}.$$

Convergence of RK [Strohmer - Vershynin 2009]

If we select row i with probability $p_i = \frac{\|\mathbf{A}_i\|_2^2}{\|\mathbf{A}\|_F^2}$, taking the expectation conditioned on x^k we have

$$\begin{aligned}
\mathbb{E}\left[\|x^{k+1} - x^\star\|_2^2 \mid x^k\right] &= \mathbb{E}\left[\|x^k - x^\star\|_2^2 - \left\|\frac{\mathbf{A}_{i_k}^\top \mathbf{A}_{i_k}}{\|\mathbf{A}_{i_k}\|^2}\left(x^k - x^\star\right)\right\|_2^2 \middle| x^k\right] \\
&= \|x^k - x^\star\|_2^2 - \sum_i \frac{\|\mathbf{A}_i\|_2^2}{\|\mathbf{A}\|_F^2}\left\|\frac{\mathbf{A}_i^\top \mathbf{A}_i}{\|\mathbf{A}_i\|_2^2}\left(x^k - x^\star\right)\right\|_2^2 \\
&= \|x^k - x^\star\|_2^2 - \sum_i \frac{\left(\mathbf{A}_i(x^k - x^\star)\right)^2}{\|\mathbf{A}\|_F^2}
\end{aligned}$$

.

Convergence of RK [Strohmer - Vershynin 2009]

If we select row i with probability $p_i = \frac{\|\mathbf{A}_i\|_2^2}{\|\mathbf{A}\|_F^2}$, taking the expectation conditioned on x^k we have

$$\begin{aligned}
\mathbb{E}\left[\|x^{k+1} - x^\star\|_2^2 \mid x^k\right] &= \mathbb{E}\left[\|x^k - x^\star\|_2^2 - \left\|\frac{\mathbf{A}_{i_k}^\top \mathbf{A}_{i_k}}{\|\mathbf{A}_{i_k}\|^2}\left(x^k - x^\star\right)\right\|_2^2 \middle| x^k\right] \\
&= \|x^k - x^\star\|_2^2 - \sum_i \frac{\|\mathbf{A}_i\|_2^2}{\|\mathbf{A}\|_F^2}\left\|\frac{\mathbf{A}_i^\top \mathbf{A}_i}{\|\mathbf{A}_i\|_2^2}\left(x^k - x^\star\right)\right\|_2^2 \\
&= \|x^k - x^\star\|_2^2 - \sum_i \frac{\left(\mathbf{A}_i(x^k - x^\star)\right)^2}{\|\mathbf{A}\|_F^2} \\
&= \|x^k - x^\star\|_2^2 - \frac{\|\mathbf{A}(x^k - x^\star)\|_2^2}{\|\mathbf{A}\|_F^2}
\end{aligned}$$

.

Convergence of RK [Strohmer - Vershynin 2009]

If we select row i with probability $p_i = \frac{\|\mathbf{A}_i\|_2^2}{\|\mathbf{A}\|_F^2}$, taking the expectation conditioned on x^k we have

$$
\begin{aligned}
\mathbb{E}\left[\|x^{k+1} - x^\star\|_2^2 \mid x^k\right] &= \mathbb{E}\left[\|x^k - x^\star\|_2^2 - \left\|\frac{\mathbf{A}_{i_k}^\top \mathbf{A}_{i_k}}{\|\mathbf{A}_{i_k}\|^2}\left(x^k - x^\star\right)\right\|_2^2 \middle| x^k\right] \\
&= \|x^k - x^\star\|_2^2 - \sum_i \frac{\|\mathbf{A}_i\|_2^2}{\|\mathbf{A}\|_F^2}\left\|\frac{\mathbf{A}_i^\top \mathbf{A}_i}{\|\mathbf{A}_i\|_2^2}\left(x^k - x^\star\right)\right\|_2^2 \\
&= \|x^k - x^\star\|_2^2 - \sum_i \frac{\left(\mathbf{A}_i(x^k - x^\star)\right)^2}{\|\mathbf{A}\|_F^2} \\
&= \|x^k - x^\star\|_2^2 - \frac{\|\mathbf{A}(x^k - x^\star)\|_2^2}{\|\mathbf{A}\|_F^2} \\
&\leq \|x^k - x^\star\|_2^2 - \frac{\sigma_{\min}^2(\mathbf{A})}{\|\mathbf{A}\|_F^2}\|x^k - x^\star\|_2^2.
\end{aligned}
$$

Convergence Rate

Theorem (Strohmer - Vershynin 2009)
Let x be the solution to the consistent system of linear equations $\mathbf{A}x = b$. Then the Randomized Kaczmarz method converges to x exponentially in expectation. At each iteration,

$$\mathbb{E}\left[\left\|x^{k+1} - x^\star\right\|^2\right] \leq \left(1 - \frac{\sigma_{\min}^2(\mathbf{A})}{\|\mathbf{A}\|_F^2}\right) \mathbb{E}\left[\left\|x^k - x^\star\right\|^2\right].$$

- $\sigma_{\min}(\mathbf{A})$ *is the smallest singular value of* $\mathbf{A}$
- $\|\mathbf{A}\|_F^2 = \sum_{i,j} \mathbf{A}_{ij}^2$

Convergence Rate

Theorem (Strohmer - Vershynin 2009)
Let x be the solution to the consistent system of linear equations $\mathbf{A}x = b$. Then the Randomized Kaczmarz method converges to x exponentially in expectation. At each iteration,

$$\mathbb{E}\left[\left\|x^{k+1} - x^\star\right\|^2\right] \leq \left(1 - \frac{\sigma^2_{\min}(\mathbf{A})}{\|\mathbf{A}\|_F^2}\right) \mathbb{E}\left[\left\|x^k - x^\star\right\|^2\right].$$

Iterating the result above

$$\mathbb{E}\left[||x^k - x^\star||_2^2\right] \leq \left(1 - \frac{\sigma^2_{\min}(\mathbf{A})}{||\mathbf{A}||_F^2}\right)^k ||x^0 - x^\star||_2^2.$$

- $\sigma_{\min}(\mathbf{A})$ *is the smallest singular value of* $\mathbf{A}$
- $\|\mathbf{A}\|_F^2 = \sum_{i,j} \mathbf{A}_{ij}^2$

Inconsistent Systems

We assume $\mathbf{A}x = b$ is *overdetermined.*

Inconsistent Systems

We assume $\mathbf{A}x = b$ is *overdetermined.*

If no solution x exists, we seek the least-squares solution

$$x^\star = \operatorname*{argmin}_{x \in \mathbb{R}^n} \frac{1}{2} \|b - \mathbf{A}x\|_2^2.$$

Inconsistent Systems

We assume $\mathbf{A}x = b$ is *overdetermined.*

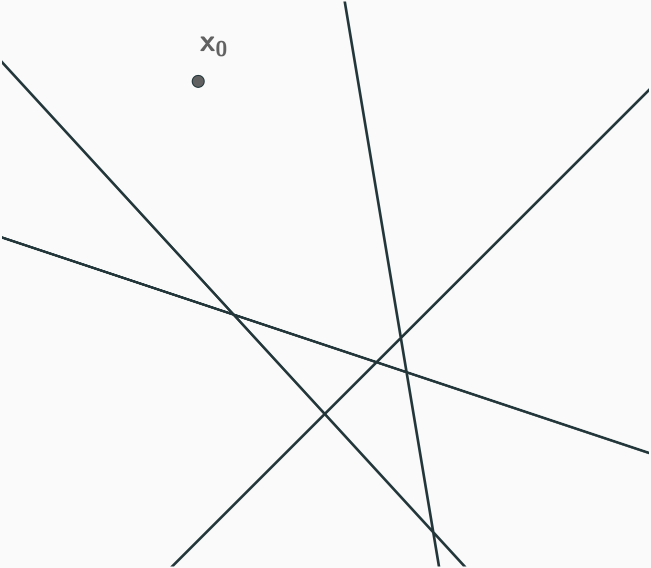
If no solution x exists, we seek the least-squares solution

$$x^\star = \operatorname*{argmin}_{x \in \mathbb{R}^n} \frac{1}{2} \| b - \mathbf{A}x \|_2^2.$$

For simplicity, we will assume that $\mathbf{A}$ *is full rank.*

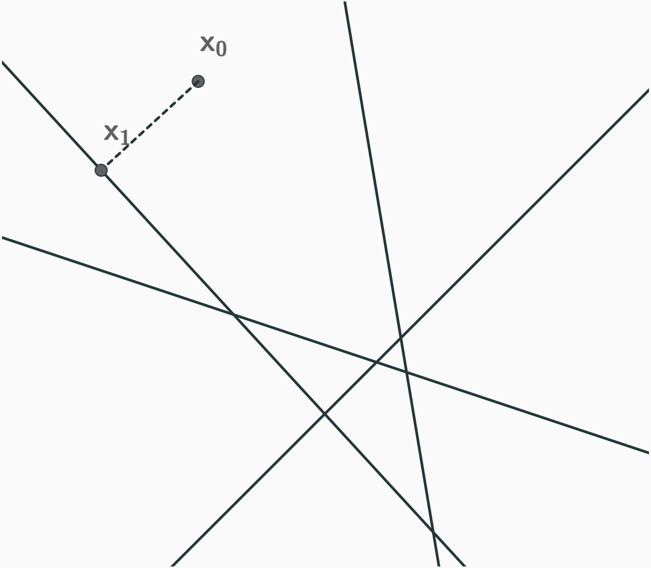
Inconsistent Systems

Solution spaces $H_i = \{x : \mathbf{A}_i x = b_i\}$ no longer all intersect.



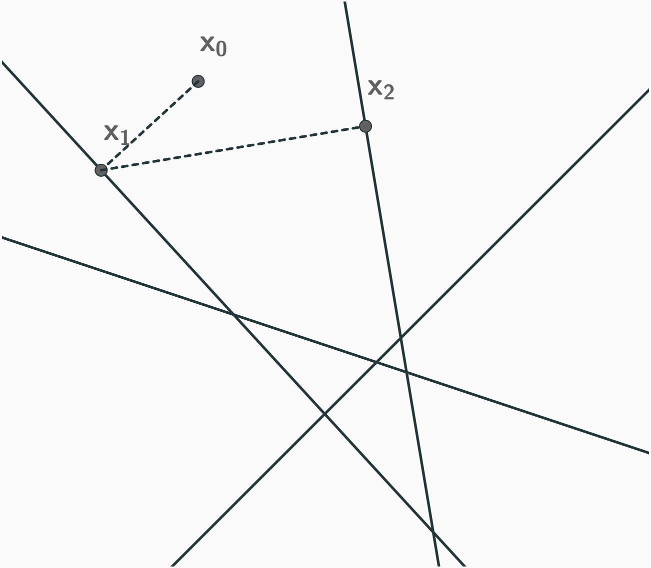
Inconsistent Systems

Solution spaces $H_i = \{x : \mathbf{A}_i x = b_i\}$ no longer all intersect.



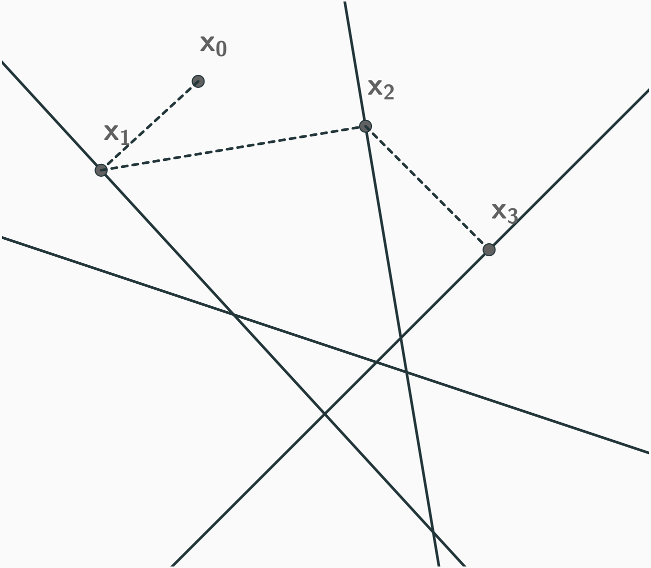
Inconsistent Systems

Solution spaces $H_i = \{x : \mathbf{A}_i x = b_i\}$ no longer all intersect.



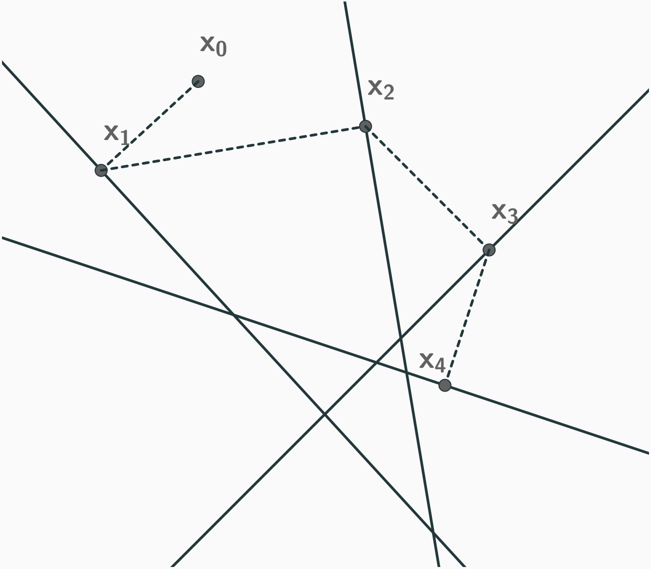
Inconsistent Systems

Solution spaces $H_i = \{x : \mathbf{A}_i x = b_i\}$ no longer all intersect.



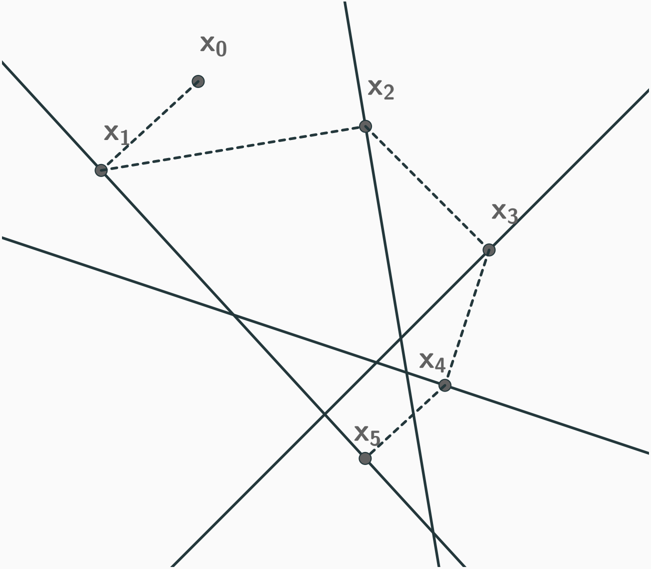
Inconsistent Systems

Solution spaces $H_i = \{x : \mathbf{A}_i x = b_i\}$ no longer all intersect.



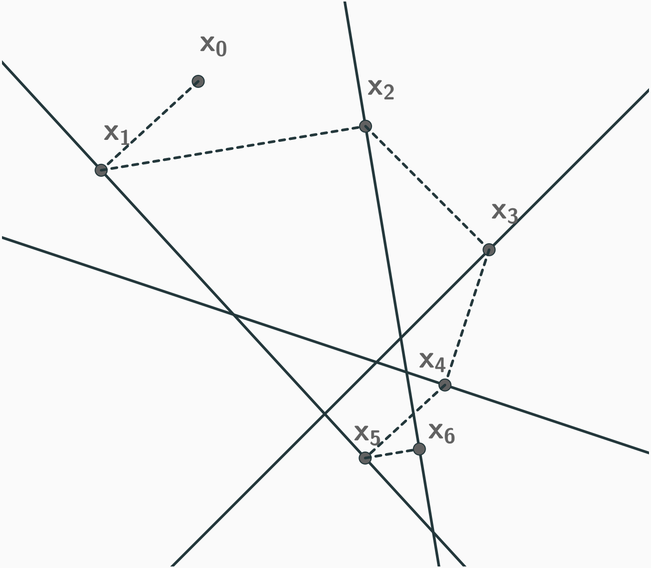
Inconsistent Systems

Solution spaces $H_i = \{x : \mathbf{A}_i x = b_i\}$ no longer all intersect.



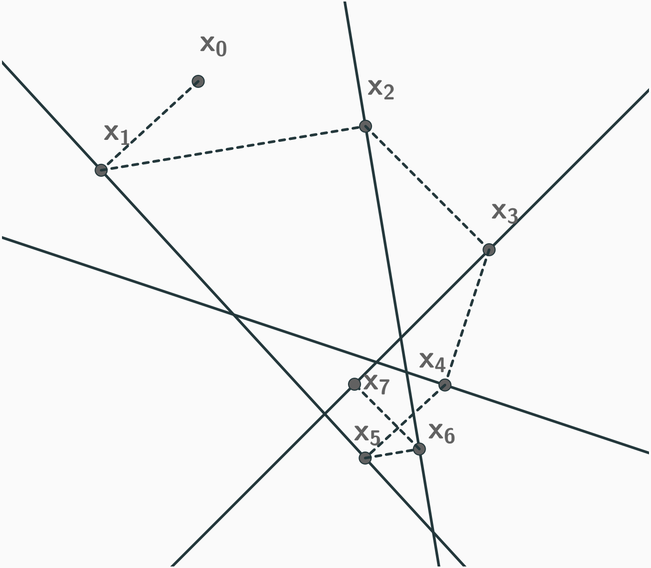
Inconsistent Systems

Solution spaces $H_i = \{x : \mathbf{A}_i x = b_i\}$ no longer all intersect.



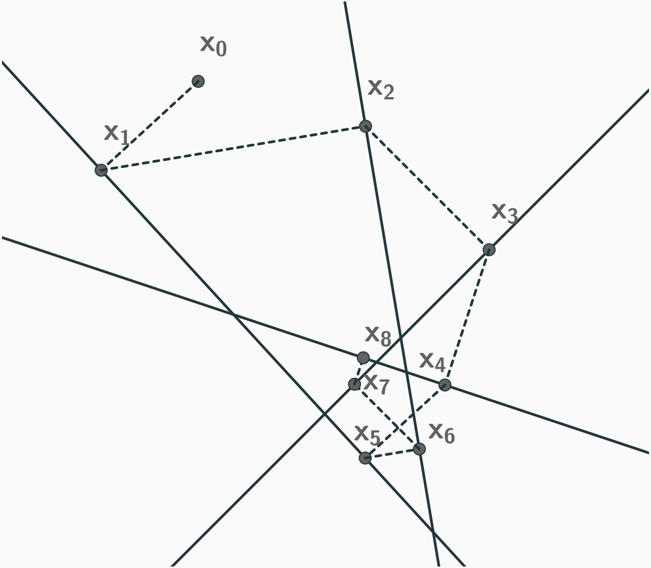
Inconsistent Systems

Solution spaces $H_i = \{x : \mathbf{A}_i x = b_i\}$ no longer all intersect.



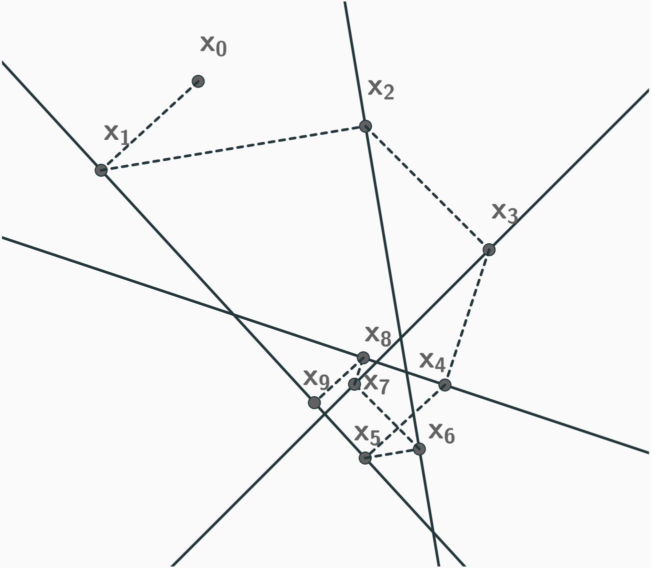
Inconsistent Systems

Solution spaces $H_i = \{x : \mathbf{A}_i x = b_i\}$ no longer all intersect.



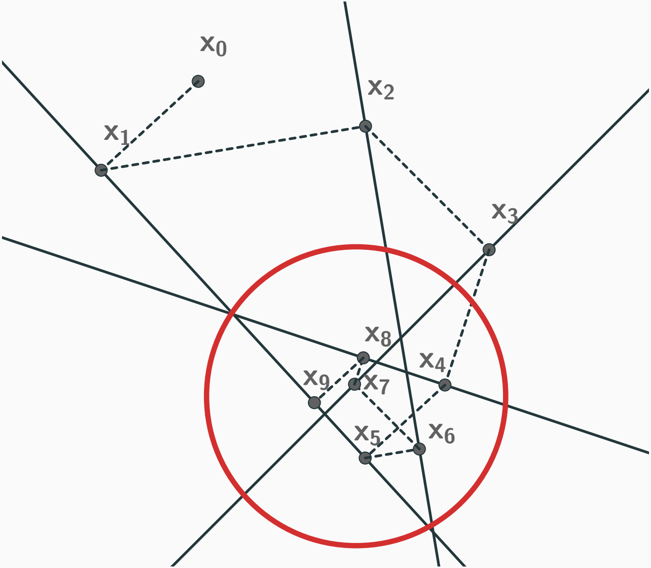
Inconsistent Systems

Solution spaces $H_i = \{x : \mathbf{A}_i x = b_i\}$ no longer all intersect.



Inconsistent Systems

Solution spaces $H_i = \{x : \mathbf{A}_i x = b_i\}$ no longer all intersect.



Convergence Rate for Inconsistent Systems

Theorem (Needell 2010, Zouzias-Freris 2013)

At each iteration,

$$\mathbb{E}\left[\left\|x^{k+1} - x^\star\right\|^2\right] \leq \left(1 - \frac{\sigma_{\min}^2(\mathbf{A})}{\|\mathbf{A}\|_F^2}\right) \mathbb{E}\left[\left\|x^k - x^\star\right\|^2\right] + \frac{\left\|r^\star\right\|^2}{\|\mathbf{A}\|_F^2}.$$

- $\sigma_{\min}(\mathbf{A})$ *is the smallest singular value of* $\mathbf{A}$
- $\|\mathbf{A}\|_F^2 = \sum_{i,j} \mathbf{A}_{ij}^2$
- $r^\star = \mathbf{A}x^\star - b$ *is the least-squares residual*
- $\dfrac{\left\|r^\star\right\|^2}{\sigma_{\min}^2(\mathbf{A})}$ *is referred to as the convergence horizon.*

Convergence Rate for Inconsistent Systems

Theorem (Needell 2010, Zouzias-Freris 2013)

At each iteration,

$$\mathbb{E}\left[\left\|x^{k+1}-x^\star\right\|^2\right] \le \left(1-\frac{\sigma_{\min}^2(\mathbf{A})}{\|\mathbf{A}\|_F^2}\right)\mathbb{E}\left[\left\|x^k-x^\star\right\|^2\right]+\frac{\left\|r^\star\right\|^2}{\|\mathbf{A}\|_F^2}.$$

Iterating the result above,

$$\mathbb{E}\left[\left\|x^k-x^\star\right\|^2\right] \le \left(1-\frac{\sigma_{\min}^2(\mathbf{A})}{\|\mathbf{A}\|_F^2}\right)^k\left\|x^0-x^\star\right\|^2+\frac{\left\|r^\star\right\|^2}{\sigma_{\min}^2(\mathbf{A})},$$

- $\sigma_{\min}(\mathbf{A})$ *is the smallest singular value of* $\mathbf{A}$
- $\|\mathbf{A}\|_F^2 = \sum_{i,j}\mathbf{A}_{ij}^2$
- $r^\star = \mathbf{A}x^\star - b$ *is the least-squares residual*
- $\dfrac{\left\|r^\star\right\|^2}{\sigma_{\min}^2(\mathbf{A})}$ *is referred to as the convergence horizon.*

RK with Averaging

Parallelizing RK

RK:

$$x^{k+1} = x^k - \frac{\mathbf{A}_{i_k} x^k - b_{i_k}}{\left\|\mathbf{A}_{i_k}\right\|^2} \mathbf{A}_{i_k}^\top$$

Parallelizing RK

Relaxed RK:

$$x^{k+1} = x^k - \lambda_{k,i_k} \frac{\mathbf{A}_{i_k} x^k - b_{i_k}}{\left\|\mathbf{A}_{i_k}\right\|^2} \mathbf{A}_{i_k}^\top$$

Parallelizing RK

Relaxed RK:

$$x^{k+1} = x^k - \lambda_{k,i_k} \frac{\mathbf{A}_{i_k} x^k - b_{i_k}}{\|\mathbf{A}_{i_k}\|^2} \mathbf{A}_{i_k}^\top$$

We consider a simple parallel extension in which we use a weighted average of independent updates.

RK with averaging:

$$x^{k+1} = x^k - \sum_{i \in \tau_k} \frac{w_i}{|\tau_k|} \frac{\mathbf{A}_i x^k - b_i}{\|\mathbf{A}_i\|^2} \mathbf{A}_i^\top$$

Weights w_i

Number of threads $|\tau_k|$

Some definitions

Normalization matrix

$$\mathbf{D} := \mathbf{Diag}(\|\mathbf{A}_1\|, \|\mathbf{A}_2\|, \ldots, \|\mathbf{A}_m\|),$$

so that the matrix $\mathbf{D}^{-1}\mathbf{A}$ has rows with unit norm.

Some definitions

Normalization matrix

$$\mathbf{D} := \mathbf{Diag}(\|\mathbf{A}_1\|, \|\mathbf{A}_2\|, \ldots, \|\mathbf{A}_m\|),$$

so that the matrix $\mathbf{D}^{-1}\mathbf{A}$ has rows with unit norm.

Probability matrix

$$\mathbf{P} := \mathbf{Diag}(p_1, p_2, \ldots, p_m)$$

where $p_j = \mathbb{P}(i = j)$ with $i \sim \mathcal{D}$.

Some definitions

Normalization matrix

$$\mathbf{D} := \mathbf{Diag}(\|\mathbf{A}_1\|, \|\mathbf{A}_2\|, \ldots, \|\mathbf{A}_m\|),$$

so that the matrix $\mathbf{D}^{-1}\mathbf{A}$ has rows with unit norm.

Probability matrix

$$\mathbf{P} := \mathbf{Diag}(p_1, p_2, \ldots, p_m)$$

where $p_j = \mathbb{P}(i = j)$ with $i \sim \mathcal{D}$.

Weight matrix

$$\mathbf{W} := \mathbf{Diag}(w_1, w_2, \ldots, w_m).$$

Coupling between weights and probabilities

Recall the update

$$x^{k+1} = x^k - \sum_{i \in \tau_k} \frac{w_i}{|\tau_k|} \frac{\mathbf{A}_i x^k - b_i}{\|\mathbf{A}_i\|^2} \mathbf{A}_i^\top$$

$$= x^k - \mathbf{A}^\top \sum_{i \in \tau_k} \frac{w_i}{|\tau_k|} \frac{\mathbf{I}_i^\top \mathbf{I}_i}{\|\mathbf{A}_i\|^2} (\mathbf{A}x^k - b)$$

Coupling between weights and probabilities

Recall the update

$$x^{k+1} = x^k - \sum_{i \in \tau_k} \frac{w_i}{|\tau_k|} \frac{\mathbf{A}_i x^k - b_i}{\|\mathbf{A}_i\|^2} \mathbf{A}_i^\top$$
$$= x^k - \mathbf{A}^\top \sum_{i \in \tau_k} \frac{w_i}{|\tau_k|} \frac{\mathbf{I}_i^\top \mathbf{I}_i}{\|\mathbf{A}_i\|^2} (\mathbf{A}x^k - b)$$

As the number of threads $|\tau_k| \to \infty$,

$$x^{k+1} = x^k - \mathbf{A}^\top \mathbb{E}\left[w_i \frac{\mathbf{I}_i^\top \mathbf{I}_i}{\|\mathbf{A}_i\|^2} \right] (\mathbf{A}x^k - b).$$

Coupling between weights and probabilities

Recall the update

$$x^{k+1} = x^k - \sum_{i \in \tau_k} \frac{w_i}{|\tau_k|} \frac{\mathbf{A}_i x^k - b_i}{\|\mathbf{A}_i\|^2} \mathbf{A}_i^\top$$
$$= x^k - \mathbf{A}^\top \sum_{i \in \tau_k} \frac{w_i}{|\tau_k|} \frac{\mathbf{I}_i^\top \mathbf{I}_i}{\|\mathbf{A}_i\|^2} (\mathbf{A}x^k - b)$$

As the number of threads $|\tau_k| \to \infty$,

$$x^{k+1} = x^k - \mathbf{A}^\top \mathbb{E}\left[w_i \frac{\mathbf{I}_i^\top \mathbf{I}_i}{\|\mathbf{A}_i\|^2} \right] (\mathbf{A}x^k - b).$$

Note:

- This is a deterministic update.
- $\mathbb{E}\left[w_i \frac{\mathbf{I}_i^\top \mathbf{I}_i}{\|\mathbf{A}_i\|^2} \right] = \mathbf{PWD}^{-2}$.

Coupling between weights and probabilities

Since we want the method to converge to the least-squares solution, we should require that $x^\star$ be a fixed point of

$$x^{k+1} = x^k - \mathbf{A}^\top \mathbf{P} \mathbf{W} \mathbf{D}^{-2} (\mathbf{A} x^k - b).$$

Coupling between weights and probabilities

Since we want the method to converge to the least-squares solution, we should require that $x^\star$ be a fixed point of

$$x^{k+1} = x^k - \mathbf{A}^\top \mathbf{PWD}^{-2}(\mathbf{A}x^k - b).$$

Any fixed point x must solve

$$\mathbf{A}^\top \mathbf{PWD}^{-2}\mathbf{A}x = \mathbf{A}^\top \mathbf{PWD}^{-2}b.$$

Coupling between weights and probabilities

Since we want the method to converge to the least-squares solution, we should require that $x^\star$ be a fixed point of

$$x^{k+1} = x^k - \mathbf{A}^\top \mathbf{PWD}^{-2}(\mathbf{A}x^k - b).$$

Any fixed point x must solve

$$\mathbf{A}^\top \mathbf{PWD}^{-2}\mathbf{A}x = \mathbf{A}^\top \mathbf{PWD}^{-2}b.$$

These are the normal equations of the weighted least-squares problem

$$\text{minimize } \frac{1}{2}\|b - \mathbf{A}x\|^2_{\mathbf{PWD}^{-2}}, \quad \text{where} \quad \|\cdot\|^2_{\mathbf{M}} = \langle \cdot, \mathbf{M}\cdot\rangle.$$

Coupling between weights and probabilities

Since we want the method to converge to the least-squares solution, we should require that $x^\star$ be a fixed point of

$$x^{k+1} = x^k - \mathbf{A}^\top \mathbf{PWD}^{-2}(\mathbf{A}x^k - b).$$

Any fixed point x must solve

$$\mathbf{A}^\top \mathbf{PWD}^{-2}\mathbf{A}x = \mathbf{A}^\top \mathbf{PWD}^{-2} b.$$

These are the normal equations of the weighted least-squares problem

$$\text{minimize } \frac{1}{2}\|b - \mathbf{A}x\|^2_{\mathbf{PWD}^{-2}}, \quad \text{where} \quad \|\cdot\|^2_{\mathbf{M}} = \langle \cdot, \mathbf{M}\cdot \rangle.$$

For **inconsistent systems**, we require the following coupling between the probability matrix $\mathbf{P}$ and the weight matrix $\mathbf{W}$:

$$\mathbf{PWD}^{-2} = \alpha \mathbf{I}.$$

General convergence for RK with averaging

Theorem

Suppose $\mathbf{PWD}^{-2} = \frac{\alpha}{\|\mathbf{A}\|_F^2}\mathbf{I}$ *for relaxation parameter* $\alpha > 0$*. Then the error at each iteration of RK with averaging satisfies*

$$\mathbb{E}\left[\left\|e^{k+1}\right\|^2\right] \\ \leq \sigma_{\max}\left(\left(\mathbf{I} - \alpha\frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)^2 - \frac{\alpha^2}{|\tau_k|}\left(\frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)^2\right)\left\|e^k\right\|^2 + \frac{\alpha}{|\tau_k|}\frac{\left\|r^k\right\|_{\mathbf{W}}^2}{\|\mathbf{A}\|_F^2},$$

where

- $e^k = x^k - x^\star$
- $\|\cdot\|_{\mathbf{W}}^2 = \langle\cdot, \mathbf{W}\cdot\rangle$
- $\|\mathbf{A}\|_F^2 = \sum_{i,j}\mathbf{A}_{ij}^2$
- $r^k = b - \mathbf{A}x^k$ *is the* k^{th} *residual.*

Uniform Weights

When the weights are uniform, i.e. $\mathbf{W} = \alpha \mathbf{I}$,

$$\begin{aligned}
\|r^k\|_{\mathbf{W}}^2 &= \|b - \mathbf{A}x^k\|_{\mathbf{W}}^2 \\
&= \alpha^2 \|b + \mathbf{A}(-x^\star + x^\star - x^k)\|_2^2 \\
&= \alpha^2 \|r^\star + \mathbf{A}e^k\|_2^2.
\end{aligned}$$

Uniform Weights

When the weights are uniform, i.e. $\mathbf{W} = \alpha \mathbf{I}$,

$$\begin{aligned}\|r^k\|_{\mathbf{W}}^2 &= \|b - \mathbf{A}x^k\|_{\mathbf{W}}^2 \\ &= \alpha^2 \|b + \mathbf{A}(-x^\star + x^\star - x^k)\|_2^2 \\ &= \alpha^2 \|r^\star + \mathbf{A}e^k\|_2^2.\end{aligned}$$

Since $\mathbf{A}^\top r^\star = 0$,

$$\|r^k\|_{\mathbf{W}}^2 = \alpha^2 \left(\|r^\star\|_2^2 + \|\mathbf{A}e^k\|_2^2 \right)$$

Convergence for RK with averaging using uniform weights

Theorem

Suppose $p_i = \frac{\|\mathbf{A}_i\|^2}{\|\mathbf{A}\|_F^2}$ *and* $\mathbf{W} = \alpha \mathbf{I}$. *Then the expected error at each iteration of RK with averaging satisfies*

$$\mathbb{E}\left[\left\|e^{k+1}\right\|^2\right] \leq \sigma_{\max}\left[\left(\mathbf{I} - \alpha \frac{\mathbf{A}^\top \mathbf{A}}{\|\mathbf{A}\|_F^2}\right)^2 + \frac{\alpha^2}{|\tau_k|}\left(\mathbf{I} - \frac{\mathbf{A}^\top \mathbf{A}}{\|\mathbf{A}\|_F^2}\right)\frac{\mathbf{A}^\top \mathbf{A}}{\|\mathbf{A}\|_F^2}\right]\left\|e^k\right\|^2 + \frac{\alpha^2 \left\|r^\star\right\|^2}{|\tau_k| \|\mathbf{A}\|_F^2}.$$

Takeaways:

- One can solve for the optimal α for the convergence bound based on $\sigma_{\max}(\mathbf{A})$ and $\sigma_{\min}(\mathbf{A})$.
- Increasing α amplifies effect of noise.
- Increasing $|\tau_k|$ improves the convergence rate and decreases the convergence horizon.

Parallel Sketch and Project Method [Richtárik and Takáč 2017]

For uniform weights and in the consistent case, this method was analyzed by Richtárik and Takáč under a more general framework.

Parallel Sketch and Project Method [Richtárik and Takáč 2017]

For uniform weights and in the consistent case, this method was analyzed by Richtárik and Takáč under a more general framework.

Sketch and project methods: Randomized iterative solvers for linear systems.

Iteratively project on to the solution space of

$$\mathbf{S}_i^\top \mathbf{A} x = \mathbf{S}_i^\top b,$$

where $\mathbf{S}_i \in \mathbb{R}^{m \times \tau}$ and $i \sim \mathcal{D}$.

Choosing $\mathbf{S}_i = e_i$, the i^{th} coordinate vector, recovers RK.

Experiments

Effect of number of threads $|\tau_k|$

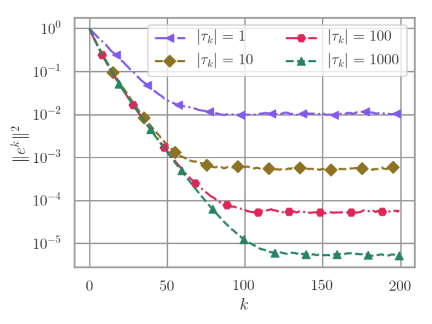


Figure 1: Uniform weights $w_i = 1$ and probabilities proportional to squared row norms $p_i = \frac{\|\mathbf{A}_i\|^2}{\|\mathbf{A}\|_F^2}$.

$$\mathbf{PWD}^{-2} \propto \mathbf{I}$$

Effect of number of threads $|\tau_k|$

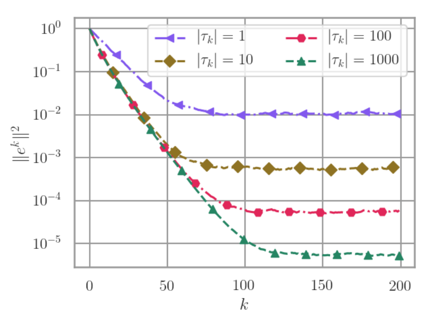


Figure 1: Uniform weights $w_i = 1$ and probabilities proportional to squared row norms $p_i = \frac{\|\mathbf{A}_i\|^2}{\|\mathbf{A}\|_F^2}$.

$$\mathbf{PWD}^{-2} \propto \mathbf{I}$$

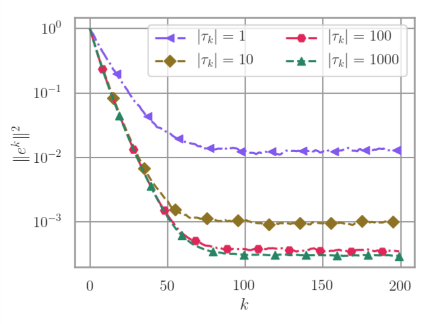


Figure 2: Uniform weights $w_i = 1$ and uniform probabilities $p_i = \frac{1}{m}$.

$$\mathbf{PWD}^{-2} \not\propto \mathbf{I}$$

Effect of number of threads $|\tau_k|$

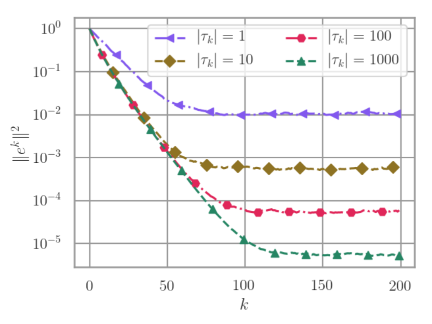


Figure 1: Uniform weights $w_i = 1$ and probabilities proportional to squared row norms $p_i = \frac{\|\mathbf{A}_i\|^2}{\|\mathbf{A}\|_F^2}$.

$$\mathbf{PWD}^{-2} \propto \mathbf{I}$$

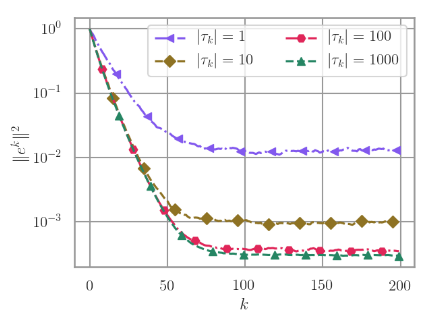


Figure 2: Uniform weights $w_i = 1$ and uniform probabilities $p_i = \frac{1}{m}$.

$$\mathbf{PWD}^{-2} \not\propto \mathbf{I}$$

$$\text{minimize } \frac{1}{2}\|b - \mathbf{A}x\|^2_{\mathbf{PWD}^{-2}}.$$

Effect of relaxation parameter α

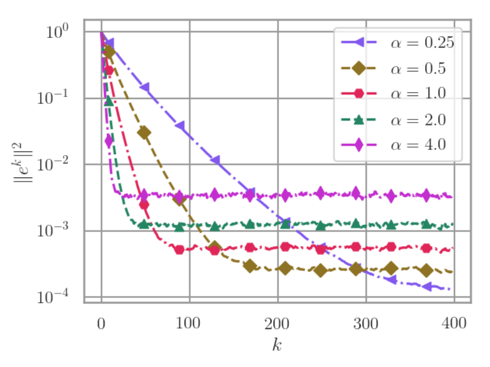


Figure 3: Uniform weights $w_i = \alpha$, probabilities proportional to squared row norms $p_i = \frac{\|\mathbf{A}_i\|^2}{\|\mathbf{A}\|_F^2}$, and number of threads $|\tau_k| = 10$.

Optimal choice for α

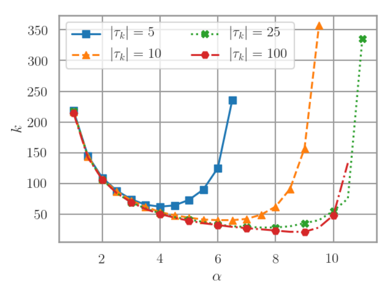


Figure 4: Uniform weights $w_i = \alpha$ and probabilities proportional to squared row norms $p_i = \frac{\|\mathbf{A}_i\|^2}{\|\mathbf{A}\|_F^2}$.

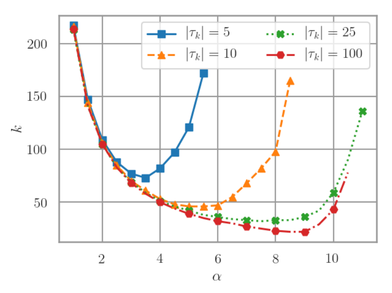


Figure 5: Weights proportional to squared row norms $w_i = \alpha m \frac{\|\mathbf{A}_i\|^2}{\|\mathbf{A}\|_F^2}$ and uniform probabilities $p_i = \frac{1}{m}$.

Summary

- Analyze an RK method with averaging that takes advantage of parallel computation
- Find a natural coupling between the probability matrix $\mathbf{P}$ and weight matrix $\mathbf{W}$
- Prove the expected convergence rate per iteration in the general case and a more interpretable rate for uniform weights
- Prove and demonstrate improved convergence with increasing $|\tau_k|$

Thanks!

References

- T. Strohmer and R. Vershynin. A randomized Kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications, 15(2):262278, 2009.*
- *D. Needell. Randomized Kaczmarz solver for noisy linear systems. BIT Numerical Mathematics, 50(2):395 403, 2010.*
- *A. Zouzias and N. M. Freris. Randomized extended Kaczmarz for solving least-squares. SIAM Journal on Matrix Analysis and Applications, 34(2):773 793, 2013.*
- *D. Needell, R. Ward, and N. Srebro. Stochastic gradient descent, weighted sampling, and the randomized Kaczmarz algorithm. Advances in Neural Information Processing Systems. 2014.*
- *D. Needell and R. Ward. "Batched stochastic gradient descent with weighted sampling." International Conference Approximation Theory. Springer, Cham, 2016.*
- *P. Richtárik and M. Takáč. Stochastic reformulations of linear systems: Algorithms and convergence theory. arXiv e-prints, page arXiv:1706.01108, June 2017.*
- *K. Yuan, Q. Ling, and W. Yin. "On the convergence of decentralized gradient descent." SIAM Journal on Optimization 26.3 (2016): 1835-1854.*