# Random Walks on the Sphere and Linear Systems of Equations

(or: Stochastic Gradient Descent for Least Squares)

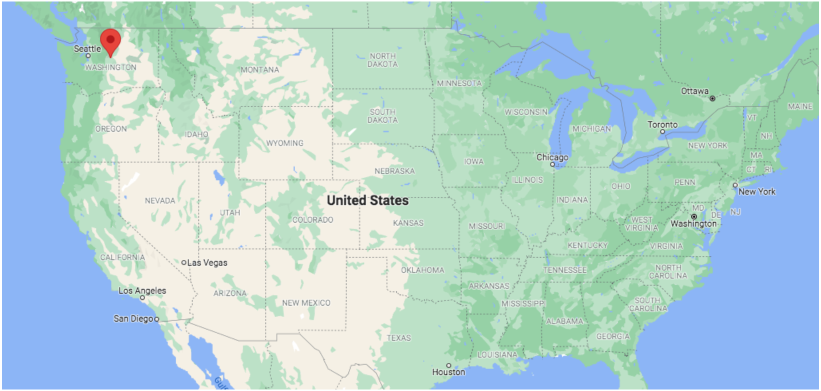
Stefan Steinerberger

Online ICCHA2021

UNIVERSITY of WASHINGTON

# It would be fun to be in Munich!

# It would be fun to be in Munich!




**A Bavarian Town**: Leavenworth, Washington.

# Leavenworth, WA

# Leavenworth, WA

# Leavenworth, WA

# Leavenworth, WA

# Outline

# Outline

The goal of this talk is to tell you about a nice way of (approximately) solving linear systems of equations.

# Outline

The goal of this talk is to tell you about a nice way of (approximately) solving linear systems of equations. It can be interpreted as Stochastic Gradient Descent applied to a classical Least Squares problem

# Outline

The goal of this talk is to tell you about a nice way of (approximately) solving linear systems of equations. It can be interpreted as Stochastic Gradient Descent applied to a classical Least Squares problem – and it can be analyzed rigorously!

# Outline

The goal of this talk is to tell you about a nice way of (approximately) solving linear systems of equations. It can be interpreted as Stochastic Gradient Descent applied to a classical Least Squares problem – and it can be analyzed rigorously!

I found it to be **mathematically rich** and naturally leading to **many(!)** open problems! (Some are mentioned in this talk.)

Throughout this talk, we will try to solve $Ax = b$ where $A \in \mathbb{R}^{n \times n}$, where $A$ is invertible.

Throughout this talk, we will try to solve $Ax = b$ where $A \in \mathbb{R}^{n\times n}$, where $A$ is invertible. We use $a_i \in \mathbb{R}^n$ to denote the $i-$th row, Throughout this talk, we will try to solve $Ax = b$ where $A \in \mathbb{R}^{n \times n}$, where $A$ is invertible. We use $a_i \in \mathbb{R}^n$ to denote the $i-$th row, so we can also write

$$\begin{pmatrix} a_1 \\ a_2 \\ \dots \\ a_n \end{pmatrix} x = b$$

Throughout this talk, we will try to solve $Ax = b$ where $A \in \mathbb{R}^{n \times n}$, where $A$ is invertible. We use $a_i \in \mathbb{R}^n$ to denote the $i-$th row, so we can also write

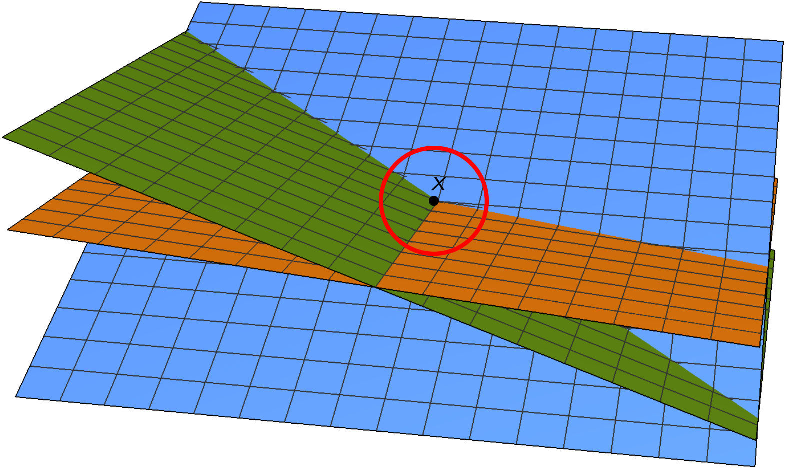
$$\begin{pmatrix} a_1 \\ a_2 \\ \dots \\ a_n \end{pmatrix} x = b$$

or

$$\forall\ 1 \leq i \leq n : \qquad \langle a_i, x \rangle = b_i.$$

# Linear Systems $\equiv$ Intersection of Hyperplanes

$$\forall\ 1 \leq i \leq n: \qquad \langle a_i, x \rangle = b_i$$



# The Kaczmarz Method

Stefan Kaczmarz
(1895 - 1939/1940)

Polish Mathematician

# The Kaczmarz Method

Stefan Kaczmarz
(1895 - 1939/1940)

Polish Mathematician
PhD in 1924 for Functional Equations

# The Kaczmarz Method


Stefan Kaczmarz
(1895 - 1939/1940)

Polish Mathematician
PhD in 1924 for Functional Equations
1930s: visit Hardy and Paley in Cambridge

# The Kaczmarz Method

Stefan Kaczmarz
(1895 - 1939/1940)

Polish Mathematician
PhD in 1924 for Functional Equations
1930s: visit Hardy and Paley in Cambridge
1937: *Approximate Solutions of Linear Equations* (3 pages)

# The Kaczmarz Method

Stefan Kaczmarz
(1895 - 1939/1940)

Polish Mathematician
PhD in 1924 for Functional Equations
1930s: visit Hardy and Paley in Cambridge
1937: *Approximate Solutions of Linear Equations* (3 pages)

*His colleagues described him as "tall and skinny", "calm and quiet", and a "modest man with rather moderate scientific ambitions".* (MacTutor Math Biographies)

# The Kaczmarz Method


Stefan Kaczmarz
(1895 - 1939/1940)

Polish Mathematician
PhD in 1924 for Functional Equations
1930s: visit Hardy and Paley in Cambridge
1937: *Approximate Solutions of Linear Equations* (3 pages)

*His colleagues described him as "tall and skinny", "calm and quiet", and a "modest man with rather moderate scientific ambitions".* (MacTutor Math Biographies)
Circumstances of death in WW2 unclear.

# The Kaczmarz method

The method is remarkably simple: we want

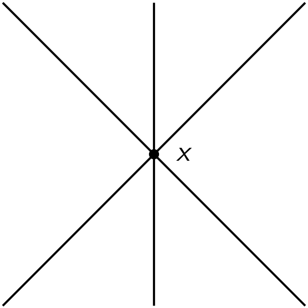
$$\forall\ 1 \leq i \leq n: \qquad \langle a_i, x \rangle = b_i.$$

# The Kaczmarz method

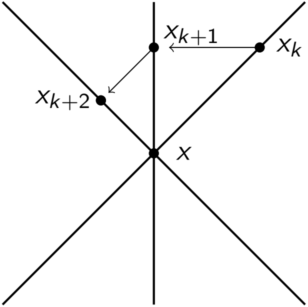
The method is remarkably simple: we want

$$\forall\ 1 \leq i \leq n: \qquad \langle a_i, x \rangle = b_i.$$

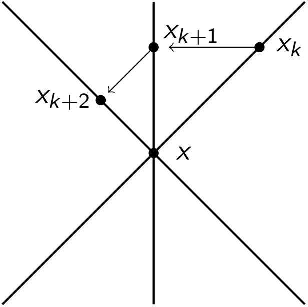
Geometrically, we want to find the intersection of hyperplanes.



# The Kaczmarz method



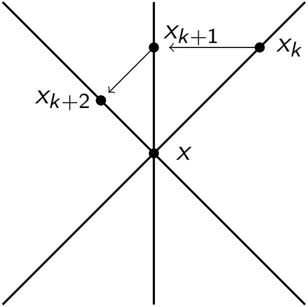
# The Kaczmarz method




Project iteratively on the hyperplanes given by

$$\langle a_i, x \rangle = b_i.$$

# The Kaczmarz method




Project iteratively on the hyperplanes given by

$$\langle a_i, x \rangle = b_i.$$

Pythageorean Theorem implies that the distance to the solution always decreases (unless you are already on that hyperplane).

# The Kaczmarz method

If we project $x_k$ onto the hyperplane given by the $i-$th equation $\langle a_i, x\rangle = b_i$ to obtain $x_{k+1}$, then

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

# The Kaczmarz method

If we project $x_k$ onto the hyperplane given by the $i-$th equation $\langle a_i, x\rangle = b_i$ to obtain $x_{k+1}$, then

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

- This is *cheap*: it's an inner product! We do not even have to load the full matrix into memory.

# The Kaczmarz method

If we project $x_k$ onto the hyperplane given by the $i-$th equation $\langle a_i, x\rangle = b_i$ to obtain $x_{k+1}$, then

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

- This is *cheap*: it's an inner product! We do not even have to load the full matrix into memory.
- This is thus useful for large matrices.

**Standard Kaczmarz**. We cycle through the indices $i$ and set

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

**Standard Kaczmarz**. We cycle through the indices $i$ and set

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

wird wiederum auf $L_1 = 0$ geworfen und gibt den Punkt $x_1^{\dots}, \dots x_n^{\dots}$, usw. Die Konvergenz des Verfahrens ist geometrisch ohne weiteres einleuchtend.

(The convergence of this method is geometrically obvious) – but the convergence *speed* is not.

**Standard Kaczmarz**. We cycle through the indices $i$ and set

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

wird wiederum auf $L_1 = 0$ geworfen und gibt den Punkt $x_1^{\dots}, \dots x_n^{\dots}$,
usw. Die Konvergenz des Verfahrens ist geometrisch ohne weiteres einleuchtend.

(The convergence of this method is geometrically obvious) – but the convergence *speed* is not.

**Random Kaczmarz.** We pick a *random* equation $i$ and set

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

**Standard Kaczmarz.** We cycle through the indices $i$ and set

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

wird wiederum auf $L_1 = 0$ geworfen und gibt den Punkt $x_1^{\dots}, \dots x_n^{\dots}$,
usw. Die Konvergenz des Verfahrens ist geometrisch ohne weiteres einleuchtend.

(The convergence of this method is geometrically obvious) – but the convergence *speed* is not.

**Random Kaczmarz.** We pick a *random* equation $i$ and set

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

- somehow behaves a little better

**Standard Kaczmarz.** We cycle through the indices $i$ and set

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

wird wiederum auf $L_1 = 0$ geworfen und gibt den Punkt $x_1, \ldots x_n$,
usw. Die Konvergenz des Verfahrens ist geometrisch ohne weiteres einleuchtend.

(The convergence of this method is geometrically obvious) – but the convergence *speed* is not.

**Random Kaczmarz.** We pick a *random* equation $i$ and set

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

- somehow behaves a little better
- used since the 1980s

**Standard Kaczmarz.** We cycle through the indices $i$ and set

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

wird wiederum auf $L_1 = 0$ geworfen und gibt den Punkt $x_1^{\dots}, \dots x_n^{\dots}$,
usw. Die Konvergenz des Verfahrens ist geometrisch ohne weiteres einleuchtend.

(The convergence of this method is geometrically obvious) – but the convergence *speed* is not.

**Random Kaczmarz.** We pick a *random* equation $i$ and set

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

- somehow behaves a little better
- used since the 1980s
- stochastic gradient descent for $\|Ax - b\|^2 \to \min$

**Theorem (Strohmer & Vershynin, 2007)**

*Pick the $i-$th equation with likelihood proportional to $\|a_i\|^2$, then*

### Theorem (Strohmer & Vershynin, 2007)

*Pick the $i-$th equation with likelihood proportional to $\|a_i\|^2$, then*

$$\mathbb{E}\,\|x_k - x\|_2^2 \le \left(1 - \frac{\sigma_n(A)^2}{\|A\|_F^2}\right)^k \|x_0 - x\|_2^2.$$

## Theorem (Strohmer & Vershynin, 2007)

*Pick the $i-$th equation with likelihood proportional to $\|a_i\|^2$, then*

$$\mathbb{E}\,\|x_k - x\|_2^2 \leq \left(1 - \frac{\sigma_n(A)^2}{\|A\|_F^2}\right)^k \|x_0 - x\|_2^2.$$

- $\|A\|_F$ is the Frobenius norm $\|A\|_F^2 = \sum_{i,j=1}^n a_{ij}^2$.
- $\sigma_n(A)$ is the smallest singular value of $A$.

# Sketch of the Proof

Strohmer & Vershynin's argument is short and elegant (certainly one of the reasons it has inspired **a lot** of subsequent work).

# Sketch of the Proof

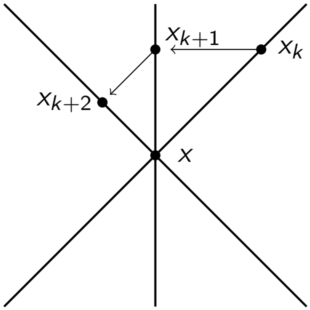
Strohmer & Vershynin's argument is short and elegant (certainly one of the reasons it has inspired **a lot** of subsequent work).

$$
\begin{aligned}
\mathbb{E}\left|\left\langle \frac{x_k - x}{\|x_k - x\|}, Z \right\rangle\right|^2 &= \sum_{i=1}^{m} \frac{\|a_j\|_2^2}{\|A\|_F^2} \left\langle \frac{x_k - x}{\|x_k - x\|}, \frac{a_j}{\|a_j\|_2} \right\rangle^2 \\
&= \frac{1}{\|A\|_F^2} \sum_{i=1}^{m} \left\langle \frac{x_k - x}{\|x_k - x\|}, a_j \right\rangle^2 \\
&= \frac{1}{\|A\|_F^2} \left\| A \frac{x_k - x}{\|x_k - x\|} \right\|^2 \\
&\geq \frac{1}{\|A\|_F^2} \frac{1}{\|A^{-1}\|_2^2}
\end{aligned}
$$

# 3. A Refined Analysis

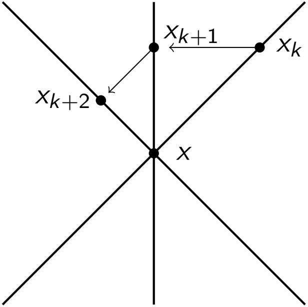
# 3. A Refined Analysis

Here's what I really wanted to know: what does $x_k - x$ do? Looking at the picture, it should be sort of jumping around.

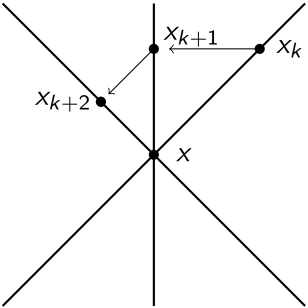


# 3. A Refined Analysis

Here's what I really wanted to know: what does $x_k - x$ do? Looking at the picture, it should be sort of jumping around.




But in numerical experiments, I didn't see that.




Empirically, the (random) sequence of vectors

$$\frac{x_k - x}{\|x_k - x\|}$$

tends to mainly a linear combination of singular vectors with small singular values.

Theorem (Small Singular Values Dominate, SIMAX 2021)

*Let $v_\ell$ be a (right) singular vector of $A$ associated to the singular value $\sigma_\ell$.*

Theorem (Small Singular Values Dominate, SIMAX 2021)

*Let $v_\ell$ be a (right) singular vector of $A$ associated to the singular value $\sigma_\ell$. Then*

$$\mathbb{E}\left\langle x_k - x, v_\ell \right\rangle = \left(1 - \frac{\sigma_\ell^2}{\|A\|_F^2}\right)^k \left\langle x_0 - x, v_\ell \right\rangle .$$

**Theorem (Small Singular Values Dominate, SIMAX 2021)**

*Let $v_\ell$ be a (right) singular vector of $A$ associated to the singular value $\sigma_\ell$. Then*

$$\mathbb{E}\left\langle x_k - x, v_\ell \right\rangle = \left(1 - \frac{\sigma_\ell^2}{\|A\|_F^2}\right)^k \left\langle x_0 - x, v_\ell \right\rangle .$$

- Different rate of contraction in different subspaces.

**Theorem (Small Singular Values Dominate, SIMAX 2021)**

*Let $v_\ell$ be a (right) singular vector of $A$ associated to the singular value $\sigma_\ell$. Then*

$$\mathbb{E}\left\langle x_k - x, v_\ell \right\rangle = \left(1 - \frac{\sigma_\ell^2}{\|A\|_F^2}\right)^k \left\langle x_0 - x, v_\ell \right\rangle .$$

- Different rate of contraction in different subspaces.
- The slowest rate of decay is given by the smallest singular value $\sigma_n$.

**Theorem (Small Singular Values Dominate, SIMAX 2021)**

*Let $v_\ell$ be a (right) singular vector of $A$ associated to the singular value $\sigma_\ell$. Then*

$$\mathbb{E}\left\langle x_k - x, v_\ell \right\rangle = \left(1 - \frac{\sigma_\ell^2}{\|A\|_F^2}\right)^k \left\langle x_0 - x, v_\ell \right\rangle .$$

- Different rate of contraction in different subspaces.
- The slowest rate of decay is given by the smallest singular value $\sigma_n$. This recovers Strohmer-Vershynin.

**Theorem (Small Singular Values Dominate, SIMAX 2021)**
*Let $v_\ell$ be a (right) singular vector of $A$ associated to the singular value $\sigma_\ell$. Then*

$$\mathbb{E}\left\langle x_k - x, v_\ell \right\rangle = \left(1 - \frac{\sigma_\ell^2}{\|A\|_F^2}\right)^k \left\langle x_0 - x, v_\ell \right\rangle .$$

- Different rate of contraction in different subspaces.
- The slowest rate of decay is given by the smallest singular value $\sigma_n$. This recovers Strohmer-Vershynin.
- **Open Problem:** Only Expectation, what can one say about the variance...?

**Theorem (Small Singular Values Dominate, SIMAX 2021)**

*Let $v_\ell$ be a (right) singular vector of $A$ associated to the singular value $\sigma_\ell$. Then*

$$\mathbb{E}\left\langle x_k - x, v_\ell \right\rangle = \left(1 - \frac{\sigma_\ell^2}{\|A\|_F^2}\right)^k \left\langle x_0 - x, v_\ell \right\rangle .$$

- Different rate of contraction in different subspaces.
- The slowest rate of decay is given by the smallest singular value $\sigma_n$. This recovers Strohmer-Vershynin.
- **Open Problem:** Only Expectation, what can one say about the variance...? Or some other form of deviation from mean?

This suggests that the method can be used to find the smallest singular vector of a matrix:

This suggests that the method can be used to find the smallest singular vector of a matrix: solve the problem $Ax = 0$.

This suggests that the method can be used to find the smallest singular vector of a matrix: solve the problem $Ax = 0$. Then $x_k - x = x_k$ converges to a linear combination of singular vectors corresponding to small singular values.

This suggests that the method can be used to find the smallest singular vector of a matrix: solve the problem $Ax = 0$. Then $x_k - x = x_k$ converges to a linear combination of singular vectors corresponding to small singular values.

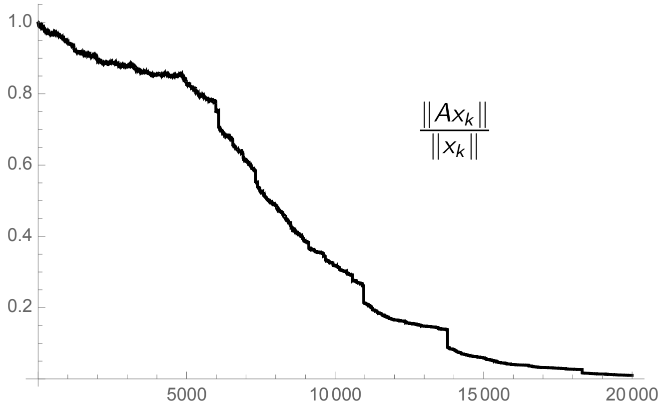
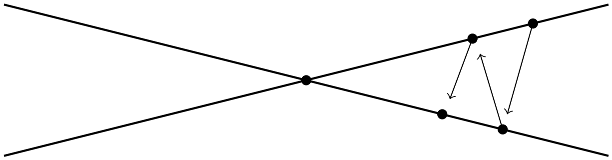
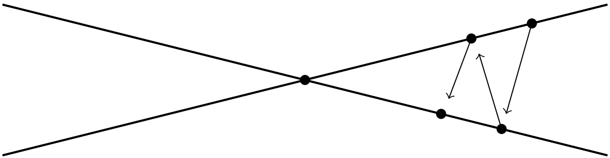



Figure: A sample evolution of $\|Ax_k\|/\|x_k\|$.

# 4. Stuck between a rock and a hard place

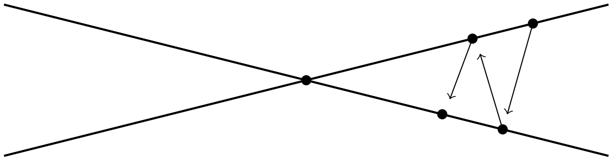


# 4. Stuck between a rock and a hard place



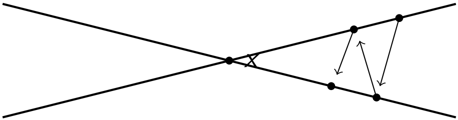
You get trapped in the narrow regions and it's hard to escape.

# 4. Stuck between a rock and a hard place



You get trapped in the narrow regions and it's hard to escape. This seems strange because, after all, it is a random process and you might end up on any hyperplane at any point in time.

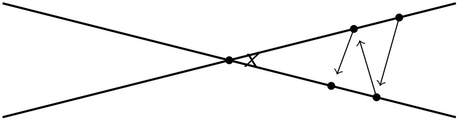
# Stuck between a rock and a hard place




Theorem (Slowing down in Bad Regions, SIMAX 2021)

*If* $x_k \neq x$ *and* $\mathbb{P}(x_{k+1} = x) = 0$, *then*

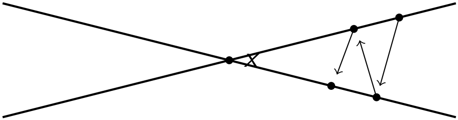
$$\mathbb{E}\left\langle \frac{x_k - x}{\|x_k - x\|}, \frac{x_{k+1} - x}{\|x_{k+1} - x\|} \right\rangle^2 =$$




Theorem (Slowing down in Bad Regions, SIMAX 2021)

*If* $x_k \neq x$ *and* $\mathbb{P}(x_{k+1} = x) = 0$, *then*

$$\mathbb{E}\ \left\langle \frac{x_k - x}{\|x_k - x\|}, \frac{x_{k+1} - x}{\|x_{k+1} - x\|} \right\rangle^2 = 1 - \frac{1}{\|A\|_F^2} \left\| A \frac{x_k - x}{\|x_k - x\|} \right\|^2 .$$

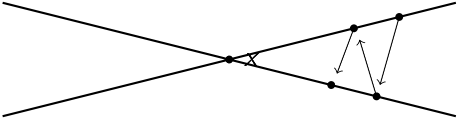



Theorem (Slowing down in Bad Regions, SIMAX 2021)

*If* $x_k \neq x$ *and* $\mathbb{P}(x_{k+1} = x) = 0$, *then*

$$\mathbb{E} \left\langle \frac{x_k - x}{\|x_k - x\|}, \frac{x_{k+1} - x}{\|x_{k+1} - x\|} \right\rangle^2 = 1 - \frac{1}{\|A\|_F^2} \left\| A \frac{x_k - x}{\|x_k - x\|} \right\|^2.$$

Once $x_k - x$ is mainly a linear combination of small singular vectors, this quantity changes very little! We stay trapped!

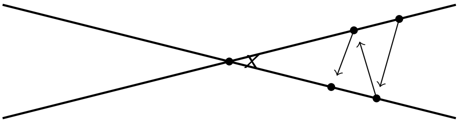



Theorem (Slowing down in Bad Regions, SIMAX 2021)

*If* $x_k \neq x$ *and* $\mathbb{P}(x_{k+1} = x) = 0$, *then*

$$\mathbb{E} \left\langle \frac{x_k - x}{\|x_k - x\|}, \frac{x_{k+1} - x}{\|x_{k+1} - x\|} \right\rangle^2 = 1 - \frac{1}{\|A\|_F^2} \left\| A \frac{x_k - x}{\|x_k - x\|} \right\|^2.$$

Once $x_k - x$ is mainly a linear combination of small singular vectors, this quantity changes very little! We stay trapped!
**Open Problem:** What about variance?




Theorem (Slowing down in Bad Regions, SIMAX 2021)

*If* $x_k \neq x$ *and* $\mathbb{P}(x_{k+1} = x) = 0$, *then*

$$\mathbb{E} \left\langle \frac{x_k - x}{\|x_k - x\|}, \frac{x_{k+1} - x}{\|x_{k+1} - x\|} \right\rangle^2 = 1 - \frac{1}{\|A\|_F^2} \left\| A \frac{x_k - x}{\|x_k - x\|} \right\|^2.$$

Once $x_k - x$ is mainly a linear combination of small singular vectors, this quantity changes very little! We stay trapped!
**Open Problem:** What about variance?
**Open Problem 2:** How do we escape?

$$\mathbb{E} \left\langle \frac{x_k - x}{\|x_k - x\|}, \frac{x_{k+1} - x}{\|x_{k+1} - x\|} \right\rangle^2 = 1 - \frac{1}{\|A\|_F^2} \left\| A \frac{x_k - x}{\|x_k - x\|} \right\|^2 .$$

**Proof.** Well, it's an identity, how hard can it be?

$$\mathbb{E}\left\langle \frac{x_k - x}{\|x_k - x\|}, \frac{x_{k+1} - x}{\|x_{k+1} - x\|} \right\rangle^2 = 1 - \frac{1}{\|A\|_F^2} \left\| A \frac{x_k - x}{\|x_k - x\|} \right\|^2.$$

**Proof.** Well, it's an identity, how hard can it be?

$$\begin{aligned}
\mathbb{E}\left\langle x_k, \frac{x_{k+1}}{\|x_{k+1}\|} \right\rangle^2 &= \sum_{i=1}^m \frac{\|a_i\|^2}{\|A\|_F^2} \left\langle x_k, \frac{x_k - \frac{\langle a_i, x_k \rangle}{\|a_i\|^2} a_i}{\|x_k - \frac{\langle a_i, x_k \rangle}{\|a_i\|^2} a_i\|} \right\rangle^2 = \sum_{i=1}^m \frac{\|a_i\|^2}{\|A\|_F^2} \frac{\left\langle x_k, x_k - \frac{\langle a_i, x_k \rangle}{\|a_i\|^2} a_i \right\rangle^2}{\|x_k - \frac{\langle a_i, x_k \rangle}{\|a_i\|^2} a_i\|^2} \\
&= \sum_{i=1}^m \frac{\|a_i\|^2}{\|A\|_F^2} \frac{\left\langle x_k, x_k - \frac{\langle a_i, x_k \rangle}{\|a_i\|^2} a_i \right\rangle^2}{\|x_k - \frac{\langle a_i, x_k \rangle}{\|a_i\|^2} a_i\|^2} = \sum_{i=1}^m \frac{\|a_i\|^2}{\|A\|_F^2} \frac{\|x_k - \frac{\langle a_i, x_k \rangle}{\|a_i\|^2} a_i\|^4}{\|x_k - \frac{\langle a_i, x_k \rangle}{\|a_i\|^2} a_i\|^2} \\
&= \sum_{i=1}^m \frac{\|a_i\|^2}{\|A\|_F^2} \left\| x_k - \frac{\langle a_i, x_k \rangle}{\|a_i\|} \frac{a_i}{\|a_i\|} \right\|^2 = \sum_{i=1}^m \frac{\|a_i\|^2}{\|A\|_F^2} \left( 1 - \frac{\langle a_i, x_k \rangle^2}{\|a_i\|^2} \right) \\
&= \frac{1}{\|A\|_F^2} \sum_{i=1}^m \left( \|a_i\|^2 - \langle a_i, x_k \rangle^2 \right) = 1 - \frac{1}{\|A\|_F^2} \sum_{i=1}^m \langle a_i, x_k \rangle^2 = 1 - \frac{\|Ax_k\|^2}{\|A\|_F^2}.
\end{aligned}$$

$$\mathbb{E}\left\langle \frac{x_k - x}{\|x_k - x\|}, \frac{x_{k+1} - x}{\|x_{k+1} - x\|} \right\rangle^2 = 1 - \frac{1}{\|A\|_F^2}\left\| A\frac{x_k - x}{\|x_k - x\|}\right\|^2.$$

**Proof.** Well, it's an identity, how hard can it be?

$$\begin{aligned}
\mathbb{E}\left\langle x_k, \frac{x_{k+1}}{\|x_{k+1}\|}\right\rangle^2 &= \sum_{i=1}^{m}\frac{\|a_i\|^2}{\|A\|_F^2}\left\langle x_k, \frac{x_k - \frac{\langle a_i, x_k\rangle}{\|a_i\|^2}a_i}{\|x_k - \frac{\langle a_i, x_k\rangle}{\|a_i\|^2}a_i\|}\right\rangle^2 = \sum_{i=1}^{m}\frac{\|a_i\|^2}{\|A\|_F^2}\frac{\left\langle x_k, x_k - \frac{\langle a_i, x_k\rangle}{\|a_i\|^2}a_i\right\rangle^2}{\|x_k - \frac{\langle a_i, x_k\rangle}{\|a_i\|^2}a_i\|^2}\\
&= \sum_{i=1}^{m}\frac{\|a_i\|^2}{\|A\|_F^2}\frac{\left\langle x_k, x_k - \frac{\langle a_i, x_k\rangle}{\|a_i\|^2}a_i\right\rangle^2}{\|x_k - \frac{\langle a_i, x_k\rangle}{\|a_i\|^2}a_i\|^2} = \sum_{i=1}^{m}\frac{\|a_i\|^2}{\|A\|_F^2}\frac{\|x_k - \frac{\langle a_i, x_k\rangle}{\|a_i\|^2}a_i\|^4}{\|x_k - \frac{\langle a_i, x_k\rangle}{\|a_i\|^2}a_i\|^2}\\
&= \sum_{i=1}^{m}\frac{\|a_i\|^2}{\|A\|_F^2}\left\|x_k - \frac{\langle a_i, x_k\rangle}{\|a_i\|}\frac{a_i}{\|a_i\|}\right\|^2 = \sum_{i=1}^{m}\frac{\|a_i\|^2}{\|A\|_F^2}\left(1 - \frac{\langle a_i, x_k\rangle^2}{\|a_i\|^2}\right)\\
&= \frac{1}{\|A\|_F^2}\sum_{i=1}^{m}\left(\|a_i\|^2 - \langle a_i, x_k\rangle^2\right) = 1 - \frac{1}{\|A\|_F^2}\sum_{i=1}^{m}\langle a_i, x_k\rangle^2 = 1 - \frac{\|Ax_k\|^2}{\|A\|_F^2}.
\end{aligned}$$

**Open Problem:** It would be nice to have more such identities.

# 5. Changing the likelihoods

New idea: maybe we shouldn't pick the likelihoods randomly.

# 5. Changing the likelihoods

New idea: maybe we shouldn't pick the likelihoods randomly. We want

$$\forall\, 1 \leq i \leq n: \qquad \langle a_i, x \rangle = b_i$$

so maybe we should pick equations where $|\langle a_i, x \rangle - b_i|$ is large?

# 5. Changing the likelihoods

New idea: maybe we shouldn't pick the likelihoods randomly. We want

$$\forall\ 1 \leq i \leq n : \qquad \langle a_i, x \rangle = b_i$$

so maybe we should pick equations where $|\langle a_i, x \rangle - b_i|$ is large?

This is known as the maximum residual method. It is known since (at least) the 1990s that this is faster (Feichtinger, Cenker, Mayer, Steier and Strohmer, 1992), (Griebel and Oswald, 2012), ...

Proposed fix: choose the $i-$th equation with likelihood proportional to

$$\mathbb{P}(\text{we choose equation } i) = \frac{|\langle a_i, x_k \rangle - b|^p}{\|Ax_k - b\|_{\ell^p}^p}.$$

Proposed fix: choose the $i-$th equation with likelihood proportional to

$$\mathbb{P}(\text{we choose equation } i) = \frac{|\langle a_i, x_k \rangle - b|^p}{\|Ax_k - b\|_{\ell^p}^p}.$$

- for $p = 0$, every equation is picked with equal likelihood

Proposed fix: choose the $i-$th equation with likelihood proportional to

$$\mathbb{P}(\text{we choose equation } i) = \frac{|\langle a_i, x_k \rangle - b|^p}{\|Ax_k - b\|_{\ell^p}^p}.$$

- for $p = 0$, every equation is picked with equal likelihood
- for $p$ large, the large deviations are more likely to be picked

Proposed fix: choose the $i-$th equation with likelihood proportional to

$$\mathbb{P}(\text{we choose equation } i) = \frac{|\langle a_i, x_k \rangle - b|^p}{\|Ax_k - b\|_{\ell^p}^p}.$$

- for $p = 0$, every equation is picked with equal likelihood
- for $p$ large, the large deviations are more likely to be picked
- in practice, no difference between $p = 20$ and $p = 10^{100}$

Proposed fix: choose the $i-$th equation with likelihood proportional to

$$\mathbb{P}(\text{we choose equation } i) = \frac{|\langle a_i, x_k \rangle - b|^p}{\|Ax_k - b\|_{\ell^p}^p}.$$

- for $p = 0$, every equation is picked with equal likelihood
- for $p$ large, the large deviations are more likely to be picked
- in practice, no difference between $p = 20$ and $p = 10^{100}$
- the method 'converges' to maximum residual as $p \to \infty$.

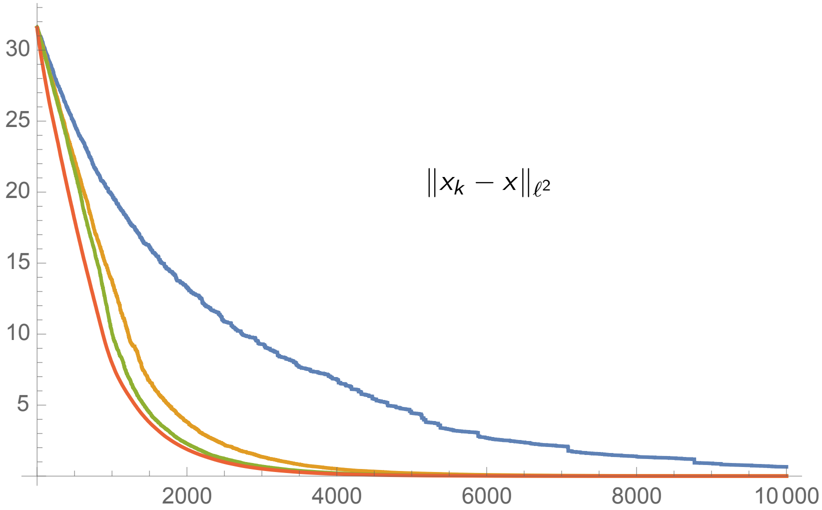



Figure: $\|x_k - x\|_{\ell^2}$ for the Randomized Kaczmarz method (blue), for $p = 1$ (orange), $p = 2$ (green) and $p = 20$ (red).

**Theorem (Weighting is better, Math. Comp, 2021)**

*Let $0 < p < \infty$, let $A$ be normalized to having the norm of each row be $\|a_i\| = 1$. Then*

$$\mathbb{E}\, \|x_k - x\|_2^2 \leq \left(1 - \inf_{x \neq 0} \frac{\|Ax\|_{\ell^{p+2}}^{p+2}}{\|Ax\|_{\ell^p}^{p} \|x\|_{\ell^2}^2}\right)^k \|x_0 - x\|_2^2.$$

**Theorem (Weighting is better, Math. Comp, 2021)**

*Let $0 < p < \infty$, let $A$ be normalized to having the norm of each row be $\|a_i\| = 1$. Then*

$$\mathbb{E}\,\|x_k - x\|_2^2 \leq \left(1 - \inf_{x \neq 0} \frac{\|Ax\|_{\ell^{p+2}}^{p+2}}{\|Ax\|_{\ell^p}^{p}\,\|x\|_{\ell^2}^2}\right)^k \|x_0 - x\|_2^2.$$

This is at least the rate of Randomized Kaczmarz ($p = 0$):

$$\inf_{x \neq 0} \frac{\|Ax\|_{\ell^{p+2}}^{p+2}}{\|Ax\|_{\ell^p}^{p}\,\|x\|_{\ell^2}^2} \geq \frac{\sigma_n^2}{\|A\|_F^2}.$$

### Theorem (Weighting is better, Math. Comp, 2021)

*Let $0 < p < \infty$, let $A$ be normalized to having the norm of each row be $\|a_i\| = 1$. Then*

$$\mathbb{E}\,\|x_k - x\|_2^2 \leq \left(1 - \inf_{x \neq 0} \frac{\|Ax\|_{\ell^{p+2}}^{p+2}}{\|Ax\|_{\ell^p}^p \|x\|_{\ell^2}^2}\right)^k \|x_0 - x\|_2^2.$$

This is at least the rate of Randomized Kaczmarz ($p = 0$):

$$\inf_{x \neq 0} \frac{\|Ax\|_{\ell^{p+2}}^{p+2}}{\|Ax\|_{\ell^p}^p \|x\|_{\ell^2}^2} \geq \frac{\sigma_n^2}{\|A\|_F^2}.$$

**Open Problem.** Is there any structure in $x_k - x$?

Theorem (Weighting is better, Math. Comp, 2021)

*Let $0 < p < \infty$, let $A$ be normalized to having the norm of each row be $\|a_i\| = 1$. Then*

$$\mathbb{E}\,\|x_k - x\|_2^2 \leq \left(1 - \inf_{x \neq 0} \frac{\|Ax\|_{\ell^{p+2}}^{p+2}}{\|Ax\|_{\ell^p}^p \|x\|_{\ell^2}^2}\right)^k \|x_0 - x\|_2^2.$$

This is at least the rate of Randomized Kaczmarz ($p = 0$):

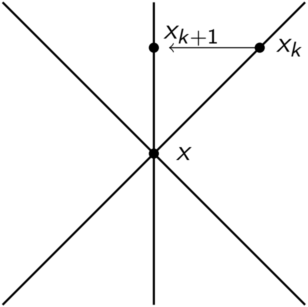
$$\inf_{x \neq 0} \frac{\|Ax\|_{\ell^{p+2}}^{p+2}}{\|Ax\|_{\ell^p}^p \|x\|_{\ell^2}^2} \geq \frac{\sigma_n^2}{\|A\|_F^2}.$$

**Open Problem.** Is there any structure in $x_k - x$?

**Open Problem 2.** The method is *a priori* specified: are there any smarter ways of adapting dynamically along the flow?

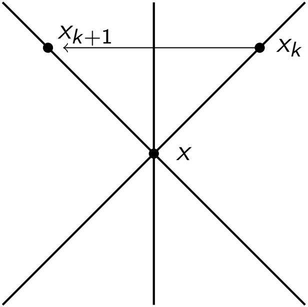
# 6. The Magic Bubble

# 6. The Magic Bubble



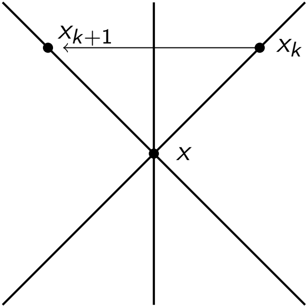

Right now we are projecting...

# 6. The Magic Bubble

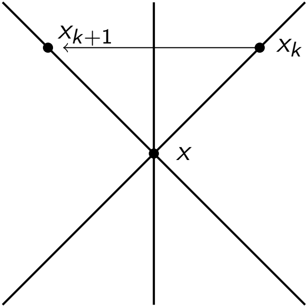



...but we could also be reflecting.

# 6. The Magic Bubble



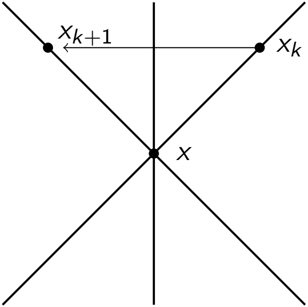

...but we could also be reflecting. Reflection doesn't get us any closer to the solution but it does something else.




We get that, again from Pythagoras,

$$\|x_k - x\| = \|x_{k+1} - x\|.$$

The distance to the true solution stays exactly preserved!




We get that, again from Pythagoras,

$$\|x_k - x\| = \|x_{k+1} - x\|.$$

The distance to the true solution stays exactly preserved! The formula stays simple

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

This gives us a new approach to the problem.

This gives us a new approach to the problem.

- Start with some arbitrary $x_0 \in \mathbb{R}^n$.

This gives us a new approach to the problem.

- Start with some arbitrary $x_0 \in \mathbb{R}^n$.
- Generate a sequence of vectors in $\mathbb{R}^n$ via

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

  You can pick the $i$ any way you like.

This gives us a new approach to the problem.

- Start with some arbitrary $x_0 \in \mathbb{R}^n$.
- Generate a sequence of vectors in $\mathbb{R}^n$ via

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

  You can pick the $i$ any way you like.
- Do this for a while until you are happy. You end up with a set $\{x_0, \ldots, x_n\}$ such that

$$\|x_k - x\| \qquad \text{is constant.}$$

This gives us a new approach to the problem.

- Start with some arbitrary $x_0 \in \mathbb{R}^n$.
- Generate a sequence of vectors in $\mathbb{R}^n$ via

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

  You can pick the $i$ any way you like.
- Do this for a while until you are happy. You end up with a set $\{x_0, \dots, x_n\}$ such that

$$\|x_k - x\| \qquad \text{is constant.}$$

  **They are all on a sphere around the true solution.**

This gives us a new approach to the problem.

- Start with some arbitrary $x_0 \in \mathbb{R}^n$.
- Generate a sequence of vectors in $\mathbb{R}^n$ via

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

  You can pick the $i$ any way you like.
- Do this for a while until you are happy. You end up with a set $\{x_0, \dots, x_n\}$ such that

$$\|x_k - x\| \qquad \text{is constant.}$$

  **They are all on a sphere around the true solution.**
- **Open Problem:** Reconstruct a good approximation of the center of a sphere from knowing many points on the sphere.

One could certainly do exact reconstruction.

One could certainly do exact reconstruction. Suppose we have $x_1, x_2, \ldots, x_{n+1}$ all on a sphere in $\mathbb{R}^n$.

One could certainly do exact reconstruction. Suppose we have $x_1, x_2, \ldots, x_{n+1}$ all on a sphere in $\mathbb{R}^n$.

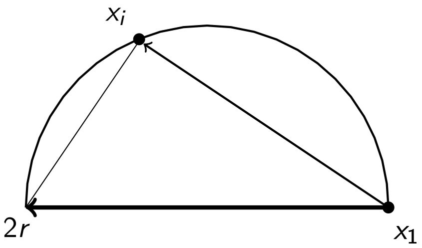



Figure: Thales' Theorem guarantees $\langle x_i - x_1, 2r \rangle = \|x_i - x_1\|^2$.

One could certainly do exact reconstruction. Suppose we have $x_1, x_2, \ldots, x_{n+1}$ all on a sphere in $\mathbb{R}^n$.

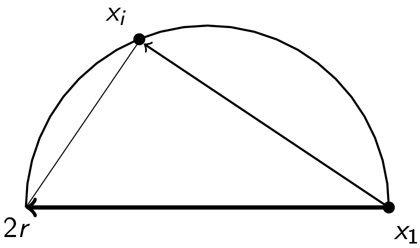



Figure: Thales' Theorem guarantees $\langle x_i - x_1, 2r \rangle = \|x_i - x_1\|^2$.

So we end up with another linear system for $r$.

One could certainly do exact reconstruction. Suppose we have $x_1, x_2, \ldots, x_{n+1}$ all on a sphere in $\mathbb{R}^n$.

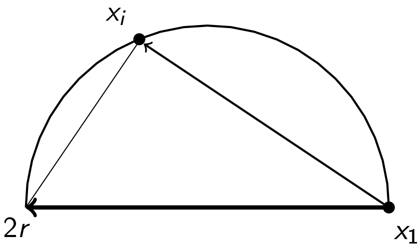



Figure: Thales' Theorem guarantees $\langle x_i - x_1, 2r \rangle = \|x_i - x_1\|^2$.

So we end up with another linear system for $r$.
**Open Problem:** Can this be used for 'upgrading' the quality of the system? It seems that yes, maybe.

Suppose we take the simple average
$$\overline{x} = \frac{1}{m}\sum_{k=1}^{m} x_k.$$

Suppose we take the simple average

$$\overline{x} = \frac{1}{m} \sum_{k=1}^{m} x_k.$$

Theorem (Applied Mathematics Quarterly, 2021)

If the $i-$th hyperplane is picked with likelihood proportional to $\|a_i\|^2$, the arising random sequence of points $(x_k)_{k=1}^{\infty}$ satisfies

$$\mathbb{E} \left\| x - \frac{1}{m} \sum_{k=1}^{m} x_k \right\| \leq \frac{1 + \|A\|_F \|A^{-1}\|}{\sqrt{m}} \cdot \|x - x_1\|.$$

Suppose we take the simple average

$$\overline{x} = \frac{1}{m}\sum_{k=1}^{m} x_k.$$

Theorem (Applied Mathematics Quarterly, 2021)

If the $i-$th hyperplane is picked with likelihood proportional to $\|a_i\|^2$, the arising random sequence of points $(x_k)_{k=1}^{\infty}$ satisfies

$$\mathbb{E}\left\| x - \frac{1}{m}\sum_{k=1}^{m} x_k \right\| \le \frac{1 + \|A\|_F \|A^{-1}\|}{\sqrt{m}} \cdot \|x - x_1\|.$$

So you need roughly $m \sim \|A\|_F^2 \|A^{-1}\|^2$ to decrease by a fixed factor. $\boxed{\textbf{Same as Kaczmarz.}}$

**Question.** How to reconstruct a good approximation of the center of a sphere from knowing many points on the sphere?

**Question.** How to reconstruct a good approximation of the center of a sphere from knowing many points on the sphere?

**Concrete Question.** You are given $100n$ points in $\mathbb{R}^n$ that lie on a sphere. How do you approximate the center?

**Question.** How to reconstruct a good approximation of the center of a sphere from knowing many points on the sphere?

**Concrete Question.** You are given $100n$ points in $\mathbb{R}^n$ that lie on a sphere. How do you approximate the center?

**Simple Averaging already leads to something as good as Random Kaczmarz!**

## Theorem (Applied Mathematics Quarterly, 2021)

If the $i-$th hyperplane is picked with likelihood proportional to $\|a_i\|^2$, the arising random sequence of points $(x_k)_{k=1}^{\infty}$ satisfies

$$\mathbb{E}\left\| x - \frac{1}{m}\sum_{k=1}^{m} x_k \right\| \leq \frac{1 + \|A\|_F \|A^{-1}\|}{\sqrt{m}} \cdot \|x - x_1\|.$$

## Theorem (Applied Mathematics Quarterly, 2021)

If the $i-$th hyperplane is picked with likelihood proportional to $\|a_i\|^2$, the arising random sequence of points $(x_k)_{k=1}^{\infty}$ satisfies

$$\mathbb{E}\left\| x - \frac{1}{m}\sum_{k=1}^{m} x_k \right\| \leq \frac{1 + \|A\|_F \|A^{-1}\|}{\sqrt{m}} \cdot \|x - x_1\|.$$

**Flavor of the Proof.**

- We can assume w.l.o.g. that $x = 0$ and that the sphere has radius 1. What can we say about

$$\mathbb{E}\left\| \frac{1}{m}\sum_{k=1}^{m} x_k \right\|?$$

## Theorem (Applied Mathematics Quarterly, 2021)

If the $i-$th hyperplane is picked with likelihood proportional to $\|a_i\|^2$, the arising random sequence of points $(x_k)_{k=1}^{\infty}$ satisfies

$$\mathbb{E}\left\| x - \frac{1}{m}\sum_{k=1}^{m} x_k \right\| \leq \frac{1 + \|A\|_F \|A^{-1}\|}{\sqrt{m}} \cdot \|x - x_1\|.$$

**Flavor of the Proof.**

- We can assume w.l.o.g. that $x = 0$ and that the sphere has radius 1. What can we say about
$$\mathbb{E}\left\| \frac{1}{m}\sum_{k=1}^{m} x_k \right\|?$$
- Let us use $R$ to denote the random reflection operator. Then
$$\frac{1}{m}\sum_{k=1}^{m} x_k = \frac{1}{m}\sum_{k=1}^{m} R^k x_0.$$

# The Flavor of the Proof

- $$\left\| \sum_{k=1}^{m} R^k x_0 \right\|^2 = \sum_{k,\ell=1}^{m} \left\langle R^k x_0, R^\ell x_0 \right\rangle$$

# The Flavor of the Proof

- $$\left\| \sum_{k=1}^{m} R^k x_0 \right\|^2 = \sum_{k,\ell=1}^{m} \left\langle R^k x_0, R^\ell x_0 \right\rangle$$
- So the relevant question is really, what can we say about
$$\mathbb{E} \left\langle R^k x_0, R^{\ell-k}(R^k x_0) \right\rangle .$$

# The Flavor of the Proof

- $$\left\| \sum_{k=1}^{m} R^k x_0 \right\|^2 = \sum_{k,\ell=1}^{m} \left\langle R^k x_0, R^\ell x_0 \right\rangle$$
- So the relevant question is really, what can we say about

$$\mathbb{E} \left\langle R^k x_0, R^{\ell-k}(R^k x_0) \right\rangle .$$

## A Decorrelation Lemma

We have, for any $x \in \mathbb{R}^n$, and any $k \in \mathbb{N}$,

$$\left| \mathbb{E} \left\langle x, R^k x \right\rangle \right| \leq \left( 1 - \frac{2\sigma_n^2}{\|A\|_F^2} \right)^k \|x\|^2.$$

*(Proof by Induction).*

# Summary

- The Kaczmarz method is a geometrically beautiful iterative method for solving linear system.

# Summary

- The Kaczmarz method is a geometrically beautiful iterative method for solving linear system.
- By replacing projection with reflection, we introduce a random reflection process on the sphere that is pretty interesting.

# Summary

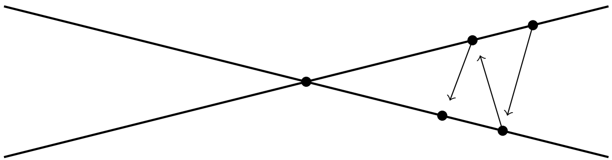
- The Kaczmarz method is a geometrically beautiful iterative method for solving linear system.
- By replacing projection with reflection, we introduce a random reflection process on the sphere that is pretty interesting.
- **Given points on a sphere, how do you estimate the location of the center of the sphere?**

# Summary

- The Kaczmarz method is a geometrically beautiful iterative method for solving linear system.
- By replacing projection with reflection, we introduce a random reflection process on the sphere that is pretty interesting.
- **Given points on a sphere, how do you estimate the location of the center of the sphere?**
- Taking the average leads to a method that is as good as Random Kaczmarz. *Anything better leads to a better method.*

# References


1. Randomized Kaczmarz converges along small singular vectors, SIMAX 2021
2. A Weighted Randomized Kaczmarz Method for Solving Linear Systems, Math. Comp. 2021
3. Surrounding the solution of a Linear System of Equations from all sides, Appl. Math. Quart 2021

THANK YOU!