Solving Linear Systems of Equations via Random Kaczmarz/Stochastic Gradient Descent

Stefan Steinerberger

August 2020
Outline

1. The Kaczmarz method
2. Random Kaczmarz (or: Stochastic Gradient Descent)
3. What happens to the singular vectors?
4. Getting stuck between a rock and a hard place
5. Changing the likelihoods
6. The energy cascade
7. What’s next?
Throughout this talk, we will try to solve $Ax = b$ where $A \in \mathbb{R}^{n \times n}$.

Let us assume that $A$ is invertible and use $a_i \in \mathbb{R}^n$ to denote the $i$–th row, so we can also write

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

or

$$
\forall 1 \leq i \leq n : \quad \langle a_i, x \rangle = b_i.
$$
Throughout this talk, we will try to solve $Ax = b$ where $A \in \mathbb{R}^{n \times n}$.

Let us assume that $A$ is invertible and use $a_i \in \mathbb{R}^n$ to denote the $i$–th row, so we can also write

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

or

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

Most of the results are more general and apply to overdetermined systems that have a solution. I don’t quite know what happens when there are no solutions: do things converge properly to a least squares solution? (*Question:* how many arguments survive?)
The Kaczmarz method

Stefan Kaczmarz
(1895 - 1939/1940)

Polish Mathematician

The circumstances of his death in WW2 (either 1939 or 1940) are unclear.

His colleagues described him as “tall and skinny”, “calm and quiet”, and a “modest man with rather moderate scientific ambitions”. (bit strange, taken verbatim from MacTutor Math Biographies)
The Kaczmarz method

Polish Mathematician
PhD in 1924 for Functional Equations

Stefan Kaczmarz
(1895 - 1939/1940)
The Kaczmarz method

Stefan Kaczmarz
(1895 - 1939/1940)

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

His colleagues described him as "tall and skinny", "calm and quiet", and a "modest man with rather moderate scientific ambitions".

The circumstances of his death in WW2 (either 1939 or 1940) are unclear.
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

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". (bit strange, taken verbatim from MacTutor Math Biographies)

Stefan Kaczmarz
(1895 - 1939/1940)
The Kaczmarz method

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". (bit strange, taken verbatim from MacTutor Math Biographies)
The circumstances of his death in WW2 (either 1939 or 1940) are unclear.

Stefan Kaczmarz
(1895 - 1939/1940)
The Kaczmarz method

The method is remarkably simple: we want

$$\forall 1 \leq i \leq n : \quad \langle a_i, x \rangle = b_i.$$
The Kaczmarz method

The method is remarkably simple: we want

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

Geometrically, we want to find the intersection of hyperplanes.
The Kaczmarz method

Project iteratively on the hyperplanes given by

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

Pythagorean Theorem implies that the distance to the solution always decreases (unless you are already on that hyperplane).
The Kaczmarz method

This is very simple in terms of equations. If we project onto the hyperplane given by the $i$–th equation, we have

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

This is very simple in terms of equations. If we project onto the hyperplane given by the $i$–th equation, we have

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

▶ 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

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

(The convergence of this method is geometrically obvious) – but the convergence rate is not.
**Standard Kaczmarz.** We cycle through the indices $i$ and set

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

(The convergence of this method is geometrically obvious) – but the convergence rate 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 in Tomography
- stochastic gradient descent for $\|Ax - b\|^2 \rightarrow \min$
Theorem (Strohmer & Vershynin, 2007)

If we pick the \(i\)-th equation with likelihood proportional to \(\|a_i\|^2\), then

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

\(\|A\|_F\) is the Frobenius norm, i.e.

\[
\|A\|_F^2 = \sum_{i,j=1}^n a_{ij}^2.
\]

\(\|A^{-1}\|_2\) is the inverse of the smallest singular value, i.e.

\[
\|A^{-1}\|_2 = \inf_{x \neq 0} \frac{\|A^{-1}x\|}{\|x\|} = \inf_{x \neq 0} \frac{\|x\|}{\|Ax\|}.
\]
Sketch of the Proof

Strohmer & Vershynin’s argument is very short and elegant (one of the reasons it has inspired a lot of subsequent work).
Sketch of the Proof

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

\[ X \]

\[ X_k \]
Sketch of the Proof

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

\[ \|x_{k+1} - x\|_2^2 \leq \left(1 - \left| \left\langle \frac{x_k - x}{\|x_k - x\|}, Z \right\rangle \right|^2 \right) \|x_k - x\|_2^2. \]
Sketch of the Proof

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

\[
\|x_{k+1} - x\|_2^2 \leq \left(1 - \left| \left\langle \frac{x_k - x}{\|x_k - x\|}, Z \right\rangle \right|^2 \right) \|x_k - x\|_2^2.
\]

\[
\mathbb{E} \left| \left\langle \frac{x_k - x}{\|x_k - x\|}, Z \right\rangle \right|^2 = \sum_{i=1}^{m} \frac{\|a_j\|_2^2}{\|A\|_F^2} \left( \left\langle \frac{x_k - x}{\|x_k - x\|}, \frac{a_j}{\|a_j\|_2} \right\rangle \right)^2
\]

\[
= \frac{1}{\|A\|_F^2} \sum_{i=1}^{m} \left\langle \frac{x_k - x}{\|x_k - x\|}, a_j \right\rangle^2
\]

\[
= \frac{1}{\|A\|_F^2} \left( A \frac{x_k - x}{\|x_k - x\|} \right)^2.
\]
3. What happens to the singular vectors?
3. What happens to the singular vectors?

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.
Numerically, 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. Of course it can jump around but it doesn’t seem to do it very much.
Theorem (Small Singular Values Dominate, (S. 2020))

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

$$\mathbb{E} \langle x_k - x, v_\ell \rangle = \left(1 - \frac{\sigma_\ell^2}{\|A\|_F^2}\right)^k \langle x_0 - x, v_\ell \rangle.$$
Theorem (Small Singular Values Dominate, (S. 2020))

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

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

The slowest rate of decay is given by the smallest singular value $\sigma_n$. Since

$$\sigma_n = \frac{1}{\|A^{-1}\|_2}$$

the rate is

$$1 - \frac{1}{\|A\|_F^2 \|A^{-1}\|_2^2}$$

which is the (optimal) rate of Strohmer & Vershynin.
Theorem (Small Singular Values Dominate, (S. 2020))

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

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

- The slowest rate of decay is given by the smallest singular value $\sigma_n$. Since

$$
\sigma_n = \frac{1}{\|A^{-1}\|_2} \quad \text{the rate is} \quad 1 - \frac{1}{\|A\|_F^2 \|A^{-1}\|_2^2}
$$

which is the (optimal) rate of Strohmer & Vershynin.

- **Open Problem**: Only Expectation, what about variance...?
Theorem (Small Singular Values Dominate, (S. 2020))

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

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

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

![Graph showing $\frac{\|Ax_k\|}{\|x_k\|}$ over iterations](graph.png)
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. This seems strange because, after all, it is a random process and you might end up on any hyperplane.
4. Stuck between a rock and a hard place

Theorem (Slowing down in Bad Areas, (S. 2020))

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

\[
\mathbb{E} \left< \frac{x_k - x}{\|x_k - x\|}, \frac{x_{k+1} - x}{\|x_{k+1} - x\|} \right>^2 = 1 - \frac{1}{\|A\|_F^2} \left\| A \frac{x_k - x}{\|x_k - x\|} \right\|^2.
\]
Theorem (Slowing down in Bad Areas, (S. 2020))

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

\[
\mathbb{E} \left( \frac{x_k - x}{\| x_k - x \|}, \frac{x_{k+1} - x}{\| x_{k+1} - x \|} \right)^2 = 1 - \frac{1}{\| A \|_F^2} \left\| A \frac{x_k - x}{\| x_k - x \|} \right\|^2.
\]

- The left-hand side corresponds to how much you change your angle, how much you change the direction from which you are approaching.
- The right-hand side checks how your current angle is related to singular vector.

Once you approach from small singular vectors, you slow down and are unlikely to change directions. That’s bad! How to fix it?
\[ \mathbb{E} \left( \frac{x_k - x}{\|x_k - x\|}, \frac{x_{k+1} - x}{\|x_{k+1} - x\|} \right)^2 = \frac{1}{\|A\|_F^2} \|Ax_k - x\|^2. \]

**Proof.** Littlewood's Lemma: Identities are always trivial.
5. Changing the likelihoods

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

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

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), ...
5. Changing the likelihoods

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

\[ \forall 1 \leq i \leq n : \quad \langle a_i, x \rangle = b_i \]

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), ...

*Wait a second... those inner products are expensive.*
5. Changing the likelihoods

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

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

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), ...

---

*Wait a second... those inner products are expensive. If you can precompute an $$N \times N$$ matrix, then it’s cheap. An iteration step gets basically twice as expensive.*
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
- 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 (S. 2020))

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

$$
\mathbb{E} \|x_k - x\|_2^2 \leq \left(1 - \inf_{x \neq 0} \frac{\|Ax\|_{\ell^{p+2}}}{\|Ax\|_{\ell^p} \|x\|_{\ell^2}}\right)^k \|x_0 - x\|_2^2.
$$
Theorem (Weighting is better (S. 2020))

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

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

This is at least the rate of Randomized Kaczmarz since

$$
\inf_{x \neq 0} \frac{\|Ax\|_{\ell^p}^{p+2}}{\|Ax\|_{\ell_p}^{p} \|x\|_2^{2}} \geq \frac{1}{\|A\|_F^2 \cdot \|A^{-1}\|_2^2}
$$

with equality if and only if the singular vector $v_n$ corresponding to the smallest singular value of $A$ has the property that $Av_n$ is a constant vector.
6. The Energy Cascade

Back to Randomized Kaczmarz (*Open Question*: how much of this is true for the weighted case?)
Figure: The size of $\|Ax_k - b\|_{\ell^2}$ for $k = 1, \ldots, 10000$. We observe rapid initial decay which then slows down.
The underlying story is somehow clear: the part of $x_k - x$ that belongs to larger singular vectors decays faster. But the Theorem above does not actually rigorously prove this: it’s a statement about expectation (no concentration!).
No concentration indeed!

Figure: The evolution of the normalized residual against the leading singular vector $v_1$: fluctuations around the mean.
Theorem (Energy Cascade (S, 2020))

Abbreviating

\[
\alpha = \max_{1 \leq i \leq n} \frac{\|Aa_i\|^2}{\|a_i\|^2},
\]

we have

\[
E \|Ax_{k+1} - b\|_2^2 \leq \left(1 + \frac{\alpha}{\|A\|_F^2}\right) \|Ax_k - b\|_2^2 - \frac{2}{\|A\|_F^2} \left\|A^T(Ax_k - b)\right\|_2^2.
\]

The constant \(\alpha\) is usually quite small, maybe \(\alpha \sim 2\).

The most important part is that there is an \(A^T\) term in the decay quantity: this makes contributions from large singular vectors even bigger and leads to even more decay.
Theorem (Energy Cascade (S, 2020))

Abbreviating

\[ \alpha = \max_{1 \leq i \leq n} \frac{\| Aa_i \|^2}{\| a_i \|^2}, \]

we have

\[ \mathbb{E} \| Ax_{k+1} - b \|_2^2 \leq \left( 1 + \frac{\alpha}{\| A \|_F^2} \right) \| Ax_k - b \|_2^2 - \frac{2}{\| A \|_F^2} \| A^T (Ax_k - b) \|_2^2. \]

- The constant \( \alpha \) is usually quite small, maybe \( \alpha \sim 2 \).
- The most important part is that there is an \( A^T \) term in the decay quantity: this makes contributions from large singular vectors even bigger and leads to even more decay.
\[ \|Ax_k - b\|_{\ell^2} \]

**Figure:** The Theorem in Action: numbers work out.
7. What’s next?
7. What’s next?

Summary

- Random Kaczmarz converges preferably along small singular vectors (we had several theorems about this).

This is good for $\|A(x_k - x)\|$ and bad for $\|x_k - x\|$ (the cascade phenomenon).
7. What’s next?

Summary

▶ Random Kaczmarz converges preferably along small singular vectors (we had several theorems about this).

This is good for \( \|A(x_k - x)\| \) and bad for \( \|x_k - x\| \) (the cascade phenomenon).

▶ We also proved some good results for the weighted variant showing that its better.
So now we have a lot better understanding: how can we improve things?
So now we have a lot better understanding: how can we improve things? The main problem is somehow: how do you avoid the bad subspaces? Maybe detect them and try to remove them?
So now we have a lot better understanding: how can we improve things? The main problem is somehow: how do you avoid the bad subspaces? Maybe detect them and try to remove them?

Strohmer & Vershynin remark in their paper that in many instances, it is better not to project onto the hyperplane but go a little bit further.

\[ x_{k+1} = x_k - 1.1 \frac{b_i - \langle a_i, x_k \rangle}{\|a_i\|^2} a_i. \]

Why?


Thank you!