Randomized Kaczmarz Algorithms: Exact MSE Analysis and Optimal Sampling Probabilities

Yue M. Lu
Signals, Information, and Networks Group (SING)
Harvard University

Joint work with Ameya Agaskar and Chuang Wang

April 17, 2015

Support: The Department of Air Force under Contract #FA8721T05TCT0002 and NSF CCF-1319140
\[ y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} -a_1^T \\ -a_2^T \\ \vdots \\ -a_m^T \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} z_1 \\ \vdots \\ z_m \end{bmatrix} \]

observation

system matrix

noise
New considerations:

- Massive sizes
- Streaming data
- Distributed storage
- Parallel computing platform
• Noiseless case: $y = Ax$ encodes a system of $m$ equations.
• Noiseless case: \( \mathbf{y} = A \mathbf{x} \) encodes a system of \( m \) equations.

\[
\begin{align*}
\mathbf{a}_1 \mathbf{x} &= \mathbf{y}_1 \\
\mathbf{a}_2 \mathbf{x} &= \mathbf{y}_2 \\
\mathbf{a}_3 \mathbf{x} &= \mathbf{y}_3
\end{align*}
\]
• Noiseless case: \( y = Ax \) encodes a system of \( m \) equations.

• Iteratively projects onto hyperplanes.
• Noiseless case: \( y = Ax \) encodes a system of \( m \) equations.

• Iteratively projects onto hyperplanes.

\[
\begin{align*}
\mathbf{a}_1 \mathbf{x} &= y_1 \\
\mathbf{a}_2 \mathbf{x} &= y_2 \\
\mathbf{a}_3 \mathbf{x} &= y_3
\end{align*}
\]
• Noiseless case: \( y = Ax \) encodes a system of \( m \) equations.

• Iteratively projects onto hyperplanes.
• Noiseless case: $y = Ax$ encodes a system of $m$ equations.

• Iteratively projects onto hyperplanes.
• Noiseless case: $y = Ax$ encodes a system of $m$ equations.

• Iteratively projects onto hyperplanes.
• Noiseless case: $y = Ax$ encodes a system of $m$ equations.

• Iteratively projects onto hyperplanes.
• Noiseless case: \( y = Ax \) encodes a system of \( m \) equations.

• Iteratively projects onto hyperplanes.
• Noiseless case: \( y = Ax \) encodes a system of \( m \) equations.

• Iteratively projects onto hyperplanes.
Kaczmarz Algorithm

- Iterative algorithm introduced by S. Kaczmarz (1937)
- Also known as algebraic reconstruction technique (ART)
- Special case of projection onto convex sets (POCS)

**Pseudocode**

Initialize arbitrary \( \mathbf{x}^{(0)} \)

For \( k = 1 \) to \( N_{\text{iter}} \):

\[
    r \leftarrow (k \mod m) + 1 \quad \text{select the next row}
\]

\[
    \mathbf{x}^{(k)} \leftarrow \mathbf{x}^{(k-1)} + \frac{y_r - \mathbf{a}_r^T \mathbf{x}^{(k-1)}}{||\mathbf{a}_r||^2} \mathbf{a}_r \quad \text{projection}
\]

\[
    \widehat{\mathbf{x}} \leftarrow \mathbf{x}^{(N_{\text{iter}})}
\]

- Can be extended to find least squares estimate from noisy measurements (Zouzias & Freris 2013)
Order matters!
Order matters!
Order matters!
Order matters!
Order matters!
Order matters!
Order matters!
Order matters!
Order matters!
Order matters!
Order matters!
IF YOU USE THE WRONG ROW ORDER

YOU'RE GONNA HAVE A BAD TIME
Solution: Randomized Kaczmarz

- Strohmer and Vershynin (2009) proposed randomizing the order:

  Choose row $a_i$ with probability proportional to $||a_i||^2$.

- Guarantees exponential convergence:

\[
\frac{\mathbb{E}||x^{(N)} - x||^2}{||x^{(0)} - x||^2} \leq (1 - \kappa(A)^{-2})^N \\
\frac{||A||_F ||A^{-1}||_2}{||A||_F ||A^{-1}||_2}
\]

- Works for arbitrary probabilities by preconditioning, so we assume row $i$ chosen with probability $p_i$. 
• Strohmer and Vershynin (2009) proposed randomizing the order:

Choose row $a_i$ with probability proportional to $||a_i||^2$.

• Guarantees exponential convergence:

$$
\frac{\mathbb{E}||x^{(N)} - x||^2}{||x^{(0)} - x||^2} \leq (1 - \kappa(A)^{-2})^N
$$

• Works for arbitrary probabilities by preconditioning, so we assume row $i$ chosen with probability $p_i$.

**Related work**

• Error bounds for *inconsistent systems* (Needell 2012)
• *Almost-sure* convergence (Chen & Powell 2012)
• Extension to find *least-square solution* in noise (Zouzias & Freris 2013)
• *Block* Kaczmarz (Needell & Tropp 2014)
1. Exact MSE formula and decay rate

2. Optimization of row selection probabilities

3. “Quenched error exponent”
$$\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} + \frac{y_r - \mathbf{a}_r^T \mathbf{x}^{(k-1)}}{||\mathbf{a}_r||^2} \mathbf{a}_r$$
\[ x^{(k)} - x = x^{(k-1)} - x + \frac{y_r - \mathbf{a}_r^T x^{(k-1)}}{||\mathbf{a}_r||^2} \mathbf{a}_r \]
\[(x^{(k)} - x) = (x^{(k-1)} - x) - \frac{a^T_r (x^{(k-1)} - x)}{||a_r||^2} a_r\]
\[(x^{(k)} - x) = (x^{(k-1)} - x) - \frac{a_r^T (x^{(k-1)} - x)}{||a_r||^2} a_r\]

\[z^{(k)} = Q_k z^{(k-1)} \quad \text{where} \quad Q_k = \left(I - \frac{a_r a_r^T}{||a_r||^2}\right)\]
\[(x^{(k)} - x) = (x^{(k-1)} - x) - \frac{a_r^T (x^{(k-1)} - x)}{||a_r||^2} a_r\]

\[z^{(k)} = Q_k z^{(k-1)}\]

where

\[Q_k = \left( I - \frac{a_r a_r^T}{||a_r||^2} \right)\]

\[z^{(k)} = Q_k Q_{k-1} \cdots Q_1 z^{(0)}\]

product of random matrices
Proposition (A.-Wang-Lu 2014)

\[ \text{MSE}_N = \text{vec} \mathbf{I}^T (\mathbb{E} \mathbf{Q} \otimes \mathbf{Q})^N \text{vec}(\mathbf{z}^{(0)} \mathbf{z}^{(0)\,T}) \]

where \( \mathbb{E} \mathbf{Q} \otimes \mathbf{Q} = \sum_i p_i \left( \mathbf{I} - \frac{\mathbf{a}_i \mathbf{a}_i^T}{\| \mathbf{a}_i \|^2} \right) \otimes^2 \)

- \( \text{vec} \) — vectorization operator; stack columns of matrix into vector.
- \( \otimes \) — matrix Kronecker product.
\[ \mathbb{E} \| z^{(N)} \|^2 = \mathbb{E} \| Q_N Q_{N-1} \cdots Q_1 z^{(0)} \|^2 \]

MSE

\[ = \mathbb{E} z^{(0)T} Q_1 Q_2 \cdots Q_N Q_N \cdots Q_2 Q_1 z^{(0)} \]

\[ = \mathbb{E} \text{trace}(Q_1 Q_2 \cdots Q_N Q_N \cdots Q_2 Q_1 z^{(0)} z^{(0)T}) \]

\[ = \mathbb{E} \text{vec}(Q_1 Q_2 \cdots Q_N Q_N \cdots Q_2 Q_1)^T \text{vec}(z^{(0)} z^{(0)T}) \]
Proof sketch

Matrix identities

$$\operatorname{vec}(ABC) = (C^T \otimes A) \operatorname{vec}(B)$$

$$\operatorname{trace}(AB) = \operatorname{vec}(A)^T \operatorname{vec}(B)$$

$$\operatorname{trace} AB = \operatorname{trace} BA$$

$$\mathbb{E} ||z||^2 = \mathbb{E} \operatorname{vec}(Q_1 Q_2 \cdots Q_N Q_N \cdots Q_2 Q_1)^T \operatorname{vec}(z^{(0)} z^{(0)^T})$$
\[ \text{MSE}_N \]

\[ = \mathbb{E}\{\text{vec}(Q_1 Q_2 \cdots Q_N Q_N \cdots Q_2 Q_1)\}^T \text{vec}(z(0) z(0)^T) \]

\[ = \mathbb{E}\{(Q_1 \otimes Q_1) \text{vec}(Q_2 \cdots Q_N Q_N \cdots Q_2)\}^T \text{vec}(z(0) z(0)^T) \]
\[ \text{MSE}_N = \mathbb{E}\{\text{vec}(Q_1 Q_2 \cdots Q_N Q_N \cdots Q_2 Q_1)\}^T \text{vec}(z^{(0)} z^{(0)T}) \]
\[ = \{(\mathbb{E}Q \otimes Q)^N \text{vec}(I)\}^T \text{vec}(z^{(0)} z^{(0)T}) \]
\[ \text{MSE}_N = \mathbb{E}\{\text{vec}(Q_1 Q_2 \cdots Q_N Q_N \cdots Q_2 Q_1)^T \text{vec}(z^{(0)} z^{(0)T}) \} \\
= \{(\mathbb{E}Q \otimes Q)^N \text{vec}(I)\}^T \text{vec}(z^{(0)} z^{(0)T}) \\
= \text{vec} I^T (\mathbb{E}Q \otimes Q)^N \text{vec}(z^{(0)} z^{(0)T}) \]
MSE decays exponentially:  \( \mathbb{E}||z^{(N)}||^2 = \exp(-\gamma_a N + o(N)) \)

**Error exponent**

\[
\gamma_a \overset{\text{def}}{=} \lim_{N \to \infty} - \frac{1}{N} \log \mathbb{E}||z^{(N)}||^2
\]

We can compute the error exponent:

\[
\gamma_a = - \log \lambda_{\text{max}} \left( \sum_i p_i \left( \mathbf{I} - \frac{a_i a_i^T}{||a_i||^2} \right)^2 \right)
\]

- Can be computed in \( O(mn^2) \) time
- Must be positive; exponential convergence confirmed
1. Exact MSE formula and decay rate

2. Optimization of row selection probabilities

3. “Quenched error exponent”
Convex optimization problem: minimize error exponent.

\[(p_1, \ldots, p_m) = \arg\min_p \lambda_{\text{max}} \left( \sum_i p_i \left( I - \frac{a_i a_i^T}{\|a_i\|^2} \right)^\otimes 2 \right)\]

semi-definite programming
n = 3 lets us easily visualize the optimal probabilities

**Intuition:** explorers of sparsely-populated regions chosen with higher probability
1. Exact MSE formula and decay rate

2. Optimization of row selection probabilities

3. “Quenched error exponent”
Simulations

150 x 20 matrix w/ Gaussian entries.
2.3. Error Exponents: Annealed vs. Quenched

It is not hard to see that the average error is not necessarily a good representative of the typical behavior of the algorithm. The empirical MSE is overlaid on the histogram, as we plotted on a logarithmic scale because of the wide range of realizations. It appears that our prediction matches the empirical value quite well.

Of the 100000 trials, and the prediction based on the annealed error exponent (cyan dashed line). We have also plotted the quenched average error trajectory, i.e. the annealed) average error trajectory (blue solid line) of all 30 representative of the typical behavior of the algorithm.

The quintessential heavy-tailed distribution is the log-normal distribution. So let us assume that the error distribution is log-normal. Then the location of Strømmer et al. [5] is also shown (black dashed line).

Under these assumptions, simply. The quintessential heavy-tailed distribution is the log-normal distribution. So let us assume that the error distribution is log-normal. Then the location of Strømmer et al. [5] is also shown (black dashed line).

The log-normal distribution is quite simple. The quintessential heavy-tailed distribution is the log-normal distribution. So let us assume that the error distribution is log-normal. Then the location of Strømmer et al. [5] is also shown (black dashed line).
Simulations

150 x 20 matrix w/ Gaussian entries.
150 x 20 matrix w/ Gaussian entries.
IF YOU WANT TO MEASURE TYPICAL PERFORMANCE
DON'T USE the MSE!
Average performance:

**Annealed error exponent**

\[
\gamma_a \overset{\text{def}}{=} \lim_{N \to \infty} -\frac{1}{N} \log \mathbb{E}\|z^{(N)}\|^2
\]

Typical performance:

**Quenched error exponent**

\[
\gamma_q \overset{\text{def}}{=} \lim_{N \to \infty} -\frac{1}{N} \mathbb{E} \log \|z^{(N)}\|^2
\]

- Much more difficult to analyze.
- Known to physicists as the top *Lyapunov exponent*.
- They use heuristics to solve.
Physicists have their own intuition for this trick, but we can get the same result by assuming the error is log-normal:

Assume

$$\log \| z^{(N)} \|^2 \sim \mathcal{N}(N\mu, N\sigma^2).$$

Then

$$\gamma_q = -\mu$$

$$\mathbb{E}\| z^{(N)} \|^2 = \exp(N[\mu + \frac{1}{2}\sigma^2])$$

$$\mathbb{E}\| z^{(N)} \|^4 = \exp(N[2\mu + 2\sigma^2])$$

**Naive replica method**

$$\log Z = \lim_{n \to 0} \frac{Z^n - 1}{n}$$
Solve

\[ \mu = \frac{1}{N} \left[ 2 \log \mathbb{E} \| z^{(N)} \|^2 - \frac{1}{2} \log \mathbb{E} \| z^{(N)} \|^4 \right] \]

\[ \gamma_q \approx 2 \gamma_a - \frac{1}{2} \gamma_a^{(2)} \]

where \( \gamma_a^{(2)} = -\log \lambda_{\text{max}} \left( \sum_i p_i \left( I - \frac{a_i a_i^T}{\| a_i \|^2} \right)^\otimes 4 \right) \)

**Quenched error exponent**

\[ \gamma_q \overset{\text{def}}{=} \lim_{N \to \infty} -\frac{1}{N} \mathbb{E} \log \| z^{(N)} \|^2 \]
Simulation

A simulation was conducted with 500 random input vectors. These vectors were then iteratively reduced using a matrix, as shown in the figure. The error, represented by the logarithm of the squared errors, is plotted against the iteration number. The figure includes both the annealed and quenched error trajectories, as well as the prediction based on Proposition 1.

The empirical MSE is overlaid on the histogram (red solid line), and the prediction based on the annealed error exponent (cyan dashed line). We have overlaid these trajectories on a logarithmic scale because of the wide range of results.

The figure shows that the average error is not necessarily a great representation of the worst realization. It appears that our prediction matches the empirical value quite well. However, it is also clear that there is more to the story.

It is clear that our prediction matches the empirical value quite well, but there are occasional, rare, extreme failures that cause the average error to be much higher than the “typical” error. In reality, there are frequent, rare, extreme failures.
• **Exact MSE formula for randomized Kaczmarz algorithms (and its generalizations)**
  
  Lifting!

• **Annealed and quenched error exponents give decay rate**
  
  Average vs. typical performance

• **Finding optimal row selection probabilities**
  
  Convex optimization