Stochastic Gradient Descent Variants for Corrupted Systems of Linear Equations

Jamie Haddock

CISS,

March 27, 2020

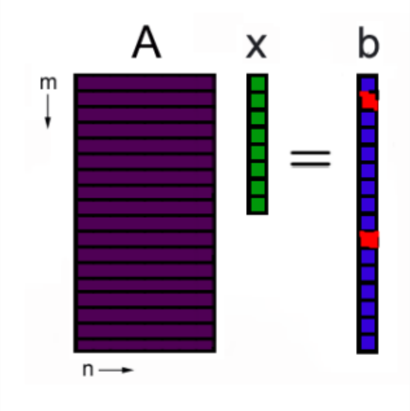
Computational and Applied Mathematics
UCLA

Problem

Solve an overdetermined system of equations

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

where some entries of b have been arbitrarily corrupted.



Set-up

- A is an $m \times n$ matrix with $m > n$ and normalized rows.

Set-up

- A is an $m \times n$ matrix with $m > n$ and normalized rows.
- We call $\mathbf{x}^*$ in $\mathbb{R}^n$ the pseudosolution.

Set-up

- A is an $m \times n$ matrix with $m > n$ and normalized rows.
- We call $\mathbf{x}^*$ in $\mathbb{R}^n$ the pseudosolution.
- $\mathbf{b}_C \in \mathbb{R}^n$ has at most βm nonzero entries (β is the fraction of corrupted entries).

Set-up

- A is an $m \times n$ matrix with $m > n$ and normalized rows.
- We call $\mathbf{x}^*$ in $\mathbb{R}^n$ the pseudosolution.
- $\mathbf{b}_C \in \mathbb{R}^n$ has at most βm nonzero entries (β is the fraction of corrupted entries).
- Given knowledge of A and the corrupted measurements $\mathbf{b} := A\mathbf{x}^* + \mathbf{b}_C$, we would like an algorithm to recover $\mathbf{x}^*$.

Set-up

- A is an $m \times n$ matrix with $m > n$ and normalized rows.
- We call $\mathbf{x}^*$ in $\mathbb{R}^n$ the pseudosolution.
- $\mathbf{b}_C \in \mathbb{R}^n$ has at most βm nonzero entries (β is the fraction of corrupted entries).
- Given knowledge of A and the corrupted measurements $\mathbf{b} := A\mathbf{x}^* + \mathbf{b}_C$, we would like an algorithm to recover $\mathbf{x}^*$.
- Moreover we would like to recover $\mathbf{x}^*$ via row-action methods (e.g. Randomized Kazcmarz, or SGD) which use rows of A, $\mathbf{a}_i^\top$.

Set-up

- A is an $m \times n$ matrix with $m > n$ and normalized rows.
- We call $\mathbf{x}^*$ in $\mathbb{R}^n$ the pseudosolution.
- $\mathbf{b}_C \in \mathbb{R}^n$ has at most βm nonzero entries (β is the fraction of corrupted entries).
- Given knowledge of A and the corrupted measurements $\mathbf{b} := A\mathbf{x}^* + \mathbf{b}_C$, we would like an algorithm to recover $\mathbf{x}^*$.
- Moreover we would like to recover $\mathbf{x}^*$ via row-action methods (e.g. Randomized Kazcmarz, or SGD) which use rows of A, $\mathbf{a}_i^\top$.
- For which matrices A can we obtain such a guarantee?

First Approach: Random Kaczmarz (RK)

RK

1. *Start with initial guess* $\mathbf{x}_0$
2. $\mathbf{x}_{k+1} = \mathbf{x}_k + (b_{i_k} - \mathbf{a}_{i_k}^T \mathbf{x}_k)\mathbf{a}_{i_k}$ *where* $i_k \in [m]$ *is chosen randomly*
3. *Repeat (2)*

First Approach: Random Kaczmarz (RK)

RK

1. *Start with initial guess* $\mathbf{x}_0$
2. $\mathbf{x}_{k+1} = \mathbf{x}_k + (b_{i_k} - \mathbf{a}_{i_k}^T \mathbf{x}_k)\mathbf{a}_{i_k}$ *where* $i_k \in [m]$ *is chosen randomly*
3. *Repeat (2)*

- Geometrically, each index i corresponds to a hyperplane in $\mathbb{R}^n$. RK projects orthogonally onto a randomly chosen hyperplane.

First Approach: Random Kaczmarz (RK)

RK

1. *Start with initial guess* $\mathbf{x}_0$
2. $\mathbf{x}_{k+1} = \mathbf{x}_k + (b_{i_k} - \mathbf{a}_{i_k}^T \mathbf{x}_k)\mathbf{a}_{i_k}$ *where* $i_k \in [m]$ *is chosen randomly*
3. *Repeat (2)*

- Geometrically, each index i corresponds to a hyperplane in $\mathbb{R}^n$. RK projects orthogonally onto a randomly chosen hyperplane.
- RK has good convergence properties for well-conditioned, consistent systems

First Approach: Random Kaczmarz (RK)

RK

1. *Start with initial guess* $\mathbf{x}_0$
2. $\mathbf{x}_{k+1} = \mathbf{x}_k + (b_{i_k} - \mathbf{a}_{i_k}^T \mathbf{x}_k)\mathbf{a}_{i_k}$ *where* $i_k \in [m]$ *is chosen randomly*
3. *Repeat (2)*

- Geometrically, each index i corresponds to a hyperplane in $\mathbb{R}^n$. RK projects orthogonally onto a randomly chosen hyperplane.
- RK has good convergence properties for well-conditioned, consistent systems
- ... but handles corruptions very poorly

First Approach: Random Kaczmarz (RK)

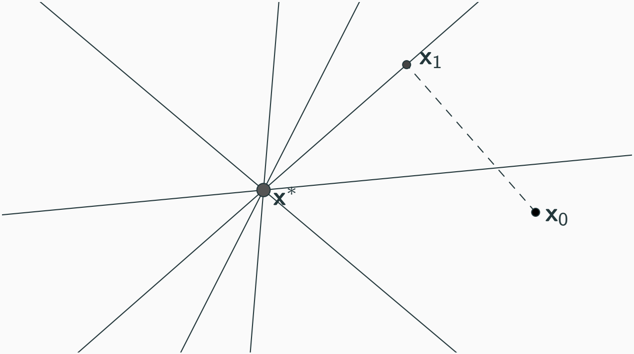
RK

1. *Start with initial guess* $\mathbf{x}_0$
2. $\mathbf{x}_{k+1} = \mathbf{x}_k + (b_{i_k} - \mathbf{a}_{i_k}^T \mathbf{x}_k)\mathbf{a}_{i_k}$ *where* $i_k \in [m]$ *is chosen randomly*
3. *Repeat (2)*

First Approach: Random Kaczmarz (RK)

RK

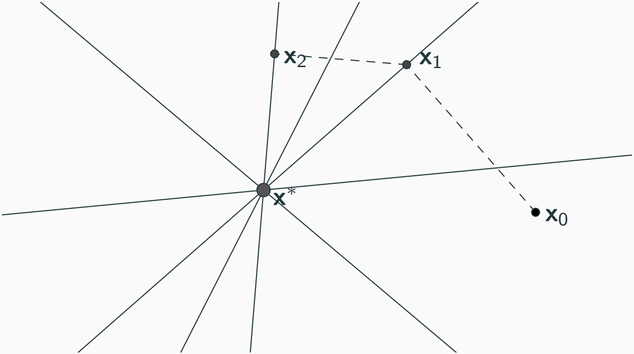
1. *Start with initial guess* $\mathbf{x}_0$
2. $\mathbf{x}_{k+1} = \mathbf{x}_k + (b_{i_k} - \mathbf{a}_{i_k}^T \mathbf{x}_k)\mathbf{a}_{i_k}$ *where* $i_k \in [m]$ *is chosen randomly*
3. *Repeat (2)*



First Approach: Random Kaczmarz (RK)

RK

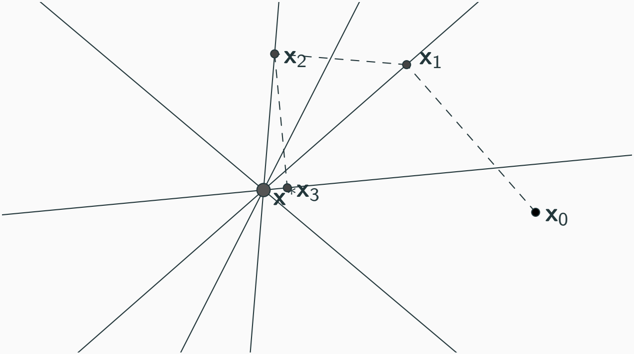
1. *Start with initial guess* $\mathbf{x}_0$
2. $\mathbf{x}_{k+1} = \mathbf{x}_k + (b_{i_k} - \mathbf{a}_{i_k}^T \mathbf{x}_k)\mathbf{a}_{i_k}$ *where* $i_k \in [m]$ *is chosen randomly*
3. *Repeat (2)*



First Approach: Random Kaczmarz (RK)

RK

1. *Start with initial guess* $\mathbf{x}_0$
2. $\mathbf{x}_{k+1} = \mathbf{x}_k + (b_{i_k} - \mathbf{a}_{i_k}^T \mathbf{x}_k)\mathbf{a}_{i_k}$ *where* $i_k \in [m]$ *is chosen randomly*
3. *Repeat (2)*



First Approach: Randomized Kaczmarz (RK)

RK

1. *Start with initial guess* $\mathbf{x}_0$
2. $\mathbf{x}_{k+1} = \mathbf{x}_k + (b_{i_k} - \mathbf{a}_{i_k}^T \mathbf{x}_k)\mathbf{a}_{i_k}$ *where* $i_k \in [m]$ *is chosen randomly*
3. *Repeat (2)*

First Approach: Randomized Kaczmarz (RK)

RK

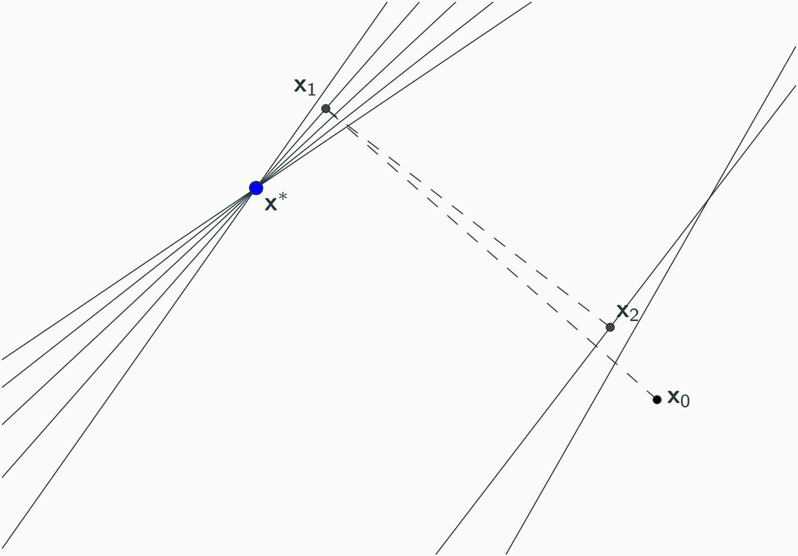
1. *Start with initial guess* $\mathbf{x}_0$
2. $\mathbf{x}_{k+1} = \mathbf{x}_k + (b_{i_k} - \mathbf{a}_{i_k}^T \mathbf{x}_k)\mathbf{a}_{i_k}$ *where* $i_k \in [m]$ *is chosen randomly*
3. *Repeat (2)*

Theorem (Strohmer-Vershynin, 2008)

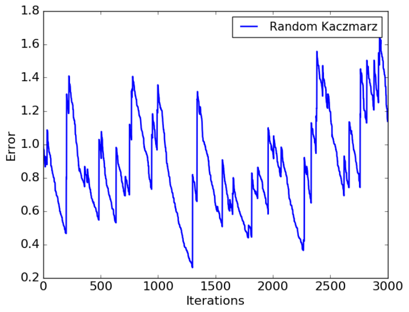
If $A\mathbf{x} = \mathbf{b}$ *is consistent and RK is used with* $\mathbb{P}[i_k = j] = \|\mathbf{a}_j\|^2/\|A\|_F^2$ *then iterates converge linearly in expectation with*

$$\mathbb{E}\|\mathbf{x}_k - \mathbf{x}^*\|^2 \leq \left(1 - \frac{1}{\|A\|_F^2\|A^{-1}\|^2}\right)^k \|\mathbf{x}_0 - \mathbf{x}^*\|^2.$$

RK with corruptions



RK with corruptions



- 50000×100 Gaussian system with 1000 corruptions.

Median Thresholding

- Idea: If a sampled hyperplane looks corrupted, don't project!

Median Thresholding

- Idea: If a sampled hyperplane looks corrupted, don't project!
- Consider the set of distances $\{d_1, \ldots d_m\}$ from $\mathbf{x}_k$ to the hyperplanes. If d_i is unusually large among these distances, then don't project onto that hyperplane.

Median Thresholding

- Idea: If a sampled hyperplane looks corrupted, don't project!
- Consider the set of distances $\{d_1, \dots d_m\}$ from $\mathbf{x}_k$ to the hyperplanes. If d_i is unusually large among these distances, then don't project onto that hyperplane.
- To quantify: Don't project if d_i is larger than the median of $\{d_1, \dots, d_m\}$.

Median Thresholding

- Idea: If a sampled hyperplane looks corrupted, don't project!
- Consider the set of distances $\{d_1, \dots d_m\}$ from $\mathbf{x}_k$ to the hyperplanes. If d_i is unusually large among these distances, then don't project onto that hyperplane.
- To quantify: Don't project if d_i is larger than the median of $\{d_1, \dots, d_m\}$.
- (Nothing too special about the median – other quantiles are possible.)

Median Thresholding

- Idea: If a sampled hyperplane looks corrupted, don't project!
- Consider the set of distances $\{d_1, \ldots d_m\}$ from $\mathbf{x}_k$ to the hyperplanes. If d_i is unusually large among these distances, then don't project onto that hyperplane.
- To quantify: Don't project if d_i is larger than the median of $\{d_1, \ldots, d_m\}$.
- (Nothing too special about the median – other quantiles are possible.)
- For efficiency, it is useful to subsample a collection of rows when computing the median.

Median RK

Method 1 Median RK

procedure $\textsc{MedRK}(A, \mathbf{b}, N, T)$
  $\mathbf{x}_0 = \mathbf{0}$
  for $j = 1, \ldots, N$ **do**
    sample $i_1, \ldots i_T \sim$ Uniform$(1, \ldots, m)$
    sample $k \sim$ Uniform$(1, \ldots, m)$
    if $|\mathbf{a}_k^\top \mathbf{x}_{j-1} - b_k| \leq \text{median}\{|\mathbf{a}_i^\top \mathbf{x}_{j-1} - b_i| : i \in i_1, ..., i_T\}$ **then**
      $\mathbf{x}_j = \mathbf{x}_{j-1} - (\mathbf{a}_k^\top \mathbf{x}_{j-1} - b_k)\mathbf{a}_k$
    else
      $\mathbf{x}_j = \mathbf{x}_{j-1}$

Convergence Result

Theorem

Let A be a random $m \times n$ matrix with rows sampled uniformly over S^{n-1}. With probability $1 - e^{-c_1 n}$ the median RK algorithm with $T = m$ satisfies

$$\mathbb{E}(\|\mathbf{x}_k - \mathbf{x}^*\|^2) \leq \left(1 - \frac{c}{n}\right)^k \|\mathbf{x}_0 - \mathbf{x}^*\|^2,$$

provided that the fraction of corrupted entries β is smaller than some positive constant, and that n and m/n are larger than fixed constants. The corrupted entries and values may be chosen adversarially.

Convergence Result

Theorem

Let A be a random $m \times n$ matrix with rows sampled uniformly over S^{n-1}. With probability $1 - e^{-c_1 n}$ the median RK algorithm with $T = m$ satisfies

$$\mathbb{E}(\|\mathbf{x}_k - \mathbf{x}^*\|^2) \leq \left(1 - \frac{c}{n}\right)^k \|\mathbf{x}_0 - \mathbf{x}^*\|^2 ,$$

provided that the fraction of corrupted entries β is smaller than some positive constant, and that n and m/n are larger than fixed constants. The corrupted entries and values may be chosen adversarially.

- "When A has incoherent rows, the convergence bound for RK holds up to constants."

Convergence Result

Theorem

Let A be a random $m \times n$ matrix with rows sampled uniformly over S^{n-1}. With probability $1 - e^{-c_1 n}$ the median RK algorithm with $T = m$ satisfies

$$\mathbb{E}(\|\mathbf{x}_k - \mathbf{x}^*\|^2) \leq \left(1 - \frac{c}{n}\right)^k \|\mathbf{x}_0 - \mathbf{x}^*\|^2,$$

provided that the fraction of corrupted entries β is smaller than some positive constant, and that n and m/n are larger than fixed constants. The corrupted entries and values may be chosen adversarially.

- "When A has incoherent rows, the convergence bound for RK holds up to constants."
- Result essentially holds with subsampling as well.

Convergence Result

Theorem

Let A be a random $m \times n$ matrix with rows sampled uniformly over S^{n-1}. With probability $1 - e^{-c_1 n}$ the median RK algorithm with $T = m$ satisfies

$$\mathbb{E}(\|\mathbf{x}_k - \mathbf{x}^*\|^2) \leq \left(1 - \frac{c}{n}\right)^k \|\mathbf{x}_0 - \mathbf{x}^*\|^2,$$

provided that the fraction of corrupted entries β is smaller than some positive constant, and that n and m/n are larger than fixed constants. The corrupted entries and values may be chosen adversarially.

- "When A has incoherent rows, the convergence bound for RK holds up to constants."
- Result essentially holds with subsampling as well.
- Can be generalized to other notions of incoherent rows.

Convergence Result

Theorem

Let A be a random $m \times n$ matrix with rows sampled uniformly over S^{n-1}. With probability $1 - e^{-c_1 n}$ the median RK algorithm with $T = m$ satisfies

$$\mathbb{E}(\|\mathbf{x}_k - \mathbf{x}^*\|^2) \leq \left(1 - \frac{c}{n}\right)^k \|\mathbf{x}_0 - \mathbf{x}^*\|^2,$$

provided that the fraction of corrupted entries β is smaller than some positive constant, and that n and m/n are larger than fixed constants. The corrupted entries and values may be chosen adversarially.

- "When A has incoherent rows, the convergence bound for RK holds up to constants."
- Result essentially holds with subsampling as well.
- Can be generalized to other notions of incoherent rows.

Proof Idea

Proof Idea

1. Show that $\text{median}\{\left|\mathbf{a}_i^\top \mathbf{x} - b_i\right| : i \in [m]\}$ is well concentrated around $\frac{1}{\sqrt{n}} \|\mathbf{x} - \mathbf{x}^*\|_2$ *for all* $\mathbf{x} \in \mathbb{R}^n$.

Proof Idea

1. Show that $\mathrm{median}\{\left|\mathbf{a}_i^\top \mathbf{x} - b_i\right| : i \in [m]\}$ is well concentrated around $\frac{1}{\sqrt{n}} \|\mathbf{x} - \mathbf{x}^*\|_2$ *for all* $\mathbf{x} \in \mathbb{R}^n$.
2. Condition on choosing a good row that the median algorithm projects onto. Show that this projection is fairly helpful in expectation.

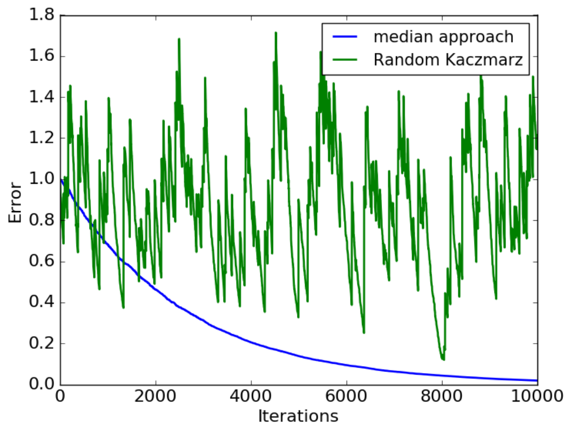
Proof Idea

1. Show that $\text{median}\{\left|\mathbf{a}_i^\top \mathbf{x} - b_i\right| : i \in [m]\}$ is well concentrated around $\frac{1}{\sqrt{n}} \left\|\mathbf{x} - \mathbf{x}^*\right\|_2$ *for all* $\mathbf{x} \in \mathbb{R}^n$.
2. Condition on choosing a good row that the median algorithm projects onto. Show that this projection is fairly helpful in expectation.
3. Condition on choosing a corrupted row that the median algorithm projects onto. Show that this projection doesn't hurt too much.

Proof Idea

1. Show that median$\{\left|\mathbf{a}_i^\top \mathbf{x} - b_i\right| : i \in [m]\}$ is well concentrated around $\frac{1}{\sqrt{n}} \|\mathbf{x} - \mathbf{x}^*\|_2$ *for all* $\mathbf{x} \in \mathbb{R}^n$.
2. Condition on choosing a good row that the median algorithm projects onto. Show that this projection is fairly helpful in expectation.
3. Condition on choosing a corrupted row that the median algorithm projects onto. Show that this projection doesn't hurt too much.

A Typical Run



- 50000×100 Gaussian system with 1000 corruptions.

Another Approach: l^1 SGD

- Under reasonable conditions, recovering $\mathbf{x}^*$ is equivalent to solving

$$\operatorname{argmin}_{\mathbf{x}} \|A\mathbf{x} - \mathbf{b}\|_0 .$$

Another Approach: l^1 SGD

- Under reasonable conditions, recovering $\mathbf{x}^*$ is equivalent to solving

$$\operatorname{argmin}_{\mathbf{x}} \left\| A\mathbf{x} - \mathbf{b} \right\|_0 .$$

- NP-hard in general, so solve the convex relaxation instead

$$\operatorname{argmin}_{\mathbf{x}} \left\| A\mathbf{x} - \mathbf{b} \right\|_1 .$$

Another Approach: l^1 SGD

- Under reasonable conditions, recovering $\mathbf{x}^*$ is equivalent to solving

$$\operatorname{argmin}_{\mathbf{x}} \left\| A\mathbf{x} - \mathbf{b} \right\|_0 .$$

- NP-hard in general, so solve the convex relaxation instead

$$\operatorname{argmin}_{\mathbf{x}} \left\| A\mathbf{x} - \mathbf{b} \right\|_1 .$$

- In many situations, the solutions to these problems coincide exactly (Candes, Tao '05; Candes, Rudelson, Tao, Vershynin '05).

Another Approach: l^1 SGD

- Under reasonable conditions, recovering $\mathbf{x}^*$ is equivalent to solving

$$\operatorname{argmin}_{\mathbf{x}} \|A\mathbf{x} - \mathbf{b}\|_0 .$$

- NP-hard in general, so solve the convex relaxation instead

$$\operatorname{argmin}_{\mathbf{x}} \|A\mathbf{x} - \mathbf{b}\|_1 .$$

- In many situations, the solutions to these problems coincide exactly (Candes, Tao '05; Candes, Rudelson, Tao, Vershynin '05).
- We would like to use SGD with respect to this objective,

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta_k \mathrm{sign}(\mathbf{a}_i^\top \mathbf{x}_k - b_i)\mathbf{a}_i,$$

where $i \in [m]$ is sampled uniformly.

Optimal Step Size

- The optimal step size η_k^* on iteration k will minimize

$$\mathbb{E}(\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2^2).$$

Optimal Step Size

- The optimal step size η_k^* on iteration k will minimize

$$\mathbb{E}(\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2^2).$$

- η_k^* is easy to compute analytically: $\eta_k^* = \mathbb{E}(\mathrm{sign}(\mathbf{a}_i^\top \mathbf{x}_k - b_i)(\mathbf{x}_k - \mathbf{x}^*)^\top \mathbf{a}_i)$.
- For this step size

$$\mathbb{E}(\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2^2) = \left(1 - \left(\frac{\eta_k^*}{\|\mathbf{x}_k - \mathbf{x}^*\|_2}\right)^2\right) \|\mathbf{x}_k - \mathbf{x}^*\|_2^2 .$$

Optimal Step Size

- The optimal step size η_k^* on iteration k will minimize

$$\mathbb{E}(\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2^2).$$

- η_k^* is easy to compute analytically: $\eta_k^* = \mathbb{E}(\text{sign}(\mathbf{a}_i^\top \mathbf{x}_k - b_i)(\mathbf{x}_k - \mathbf{x}^*)^\top \mathbf{a}_i)$.
- For this step size

$$\mathbb{E}(\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2^2) = \left(1 - \left(\frac{\eta_k^*}{\|\mathbf{x}_k - \mathbf{x}^*\|_2}\right)^2\right) \|\mathbf{x}_k - \mathbf{x}^*\|_2^2 .$$

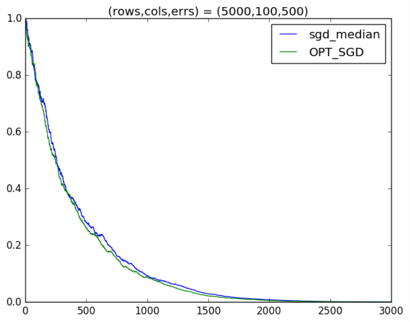
- Approximating η^* to within a small constant factor is sufficient to obtain a near-optimal guarantee.

Median SGD Algorithm

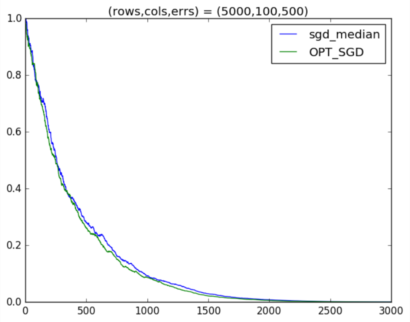
Method 2 Median SGD

1: **procedure** $\textsc{MedianSGD}(A, \mathbf{b}, \mathbf{x}_0, N)$
2: **for** j = 1, . . . , N **do**
3: sample $i_1, \ldots i_T \sim \text{Uniform}(1, \ldots, m)$
4: $\eta_j = \text{median}\{\left|\mathbf{a}_{i_l}^\top \mathbf{x}_{j-1} - b_{i_l}\right| : l \in [T]\}$
5: $\mathbf{x}_j = \mathbf{x}_{j-1} - \eta_j \text{sign}(\mathbf{a}_i^\top \mathbf{x} - b_i)\mathbf{a}_i$
return x_N

OPT vs. Median Step Sizes

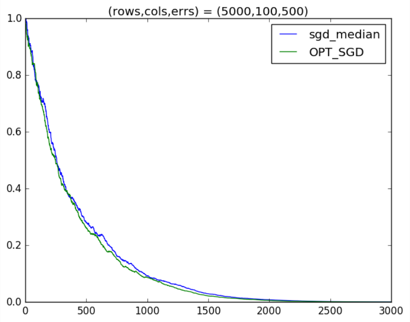


OPT vs. Median Step Sizes



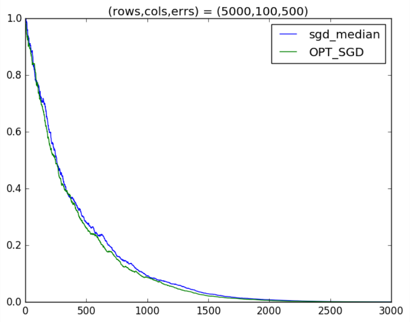
- Smaller 5000×100 system with 500 corruptions

OPT vs. Median Step Sizes



- Smaller 5000×100 system with 500 corruptions
- As long as the number of corruptions isn't too big, the median step size performs nearly optimally in practice

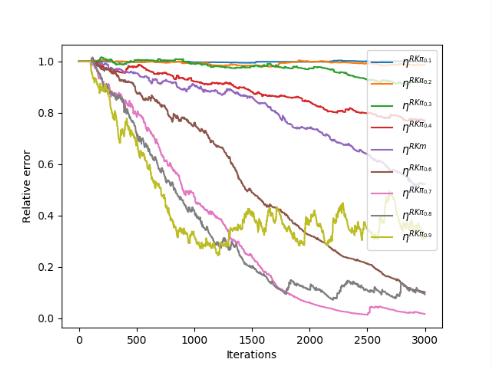
OPT vs. Median Step Sizes



- Smaller 5000×100 system with 500 corruptions
- As long as the number of corruptions isn't too big, the median step size performs nearly optimally in practice

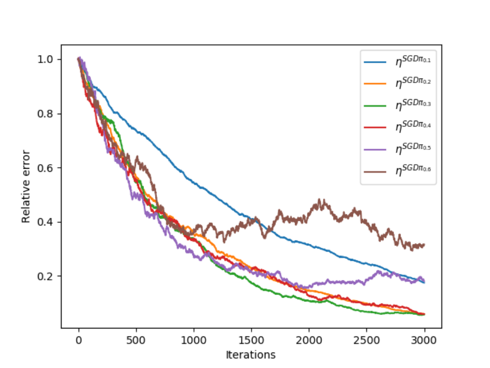
Experiments

Does the quantile for median RK matter?



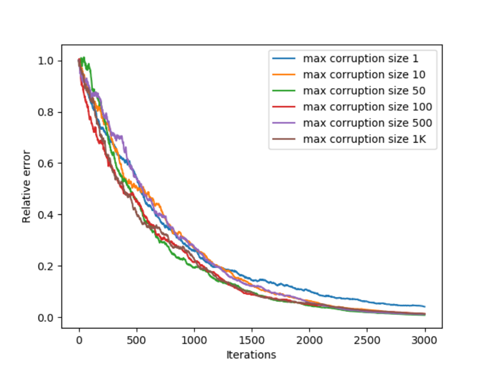
- 50000×100 Gaussian system, 30 percent corrupted entries

Does the quantile for median SGD matter?



- 50000×100 Gaussian system, 30 percent corrupted entries
- Note that choosing too small a step size hurts less than choosing too large a step size (can see from theory)

Does the size of corruptions matter?



- 50000×100 system, 30 percent corruptions, 30^{th} percentile SGD

Future Work/Open Questions

- How does the analysis of median RK extend to matrices with correlated rows?

Future Work/Open Questions

- How does the analysis of median RK extend to matrices with correlated rows?
- The analysis of median RK is qualitatively correct, but gives bad constants. Is there a better analysis that gives constants which match empirical results?

Future Work/Open Questions

- How does the analysis of median RK extend to matrices with correlated rows?
- The analysis of median RK is qualitatively correct, but gives bad constants. Is there a better analysis that gives constants which match empirical results?
- A greedy variant of median RK works quite well in practice. (If βm corruptions, then project onto hyperplane corresponding to $\beta m + 1$ largest residual.) Can we justify this approach theoretically?

Future Work/Open Questions

- How does the analysis of median RK extend to matrices with correlated rows?
- The analysis of median RK is qualitatively correct, but gives bad constants. Is there a better analysis that gives constants which match empirical results?
- A greedy variant of median RK works quite well in practice. (If βm corruptions, then project onto hyperplane corresponding to $\beta m + 1$ largest residual.) Can we justify this approach theoretically?