Liza Rebrova

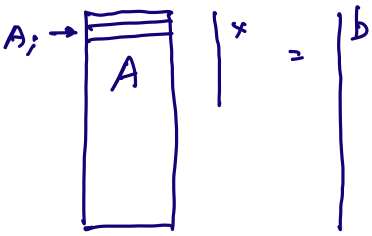
Randomized iterative methods and Non-asymptotic random matrix theory

RMT → random large high-dimensional objects are more predictable

non-asymptotic → for all large n (not $n \to \infty$)

randomized algorithms for large systems

e.g., linear systems



$Ax = b$

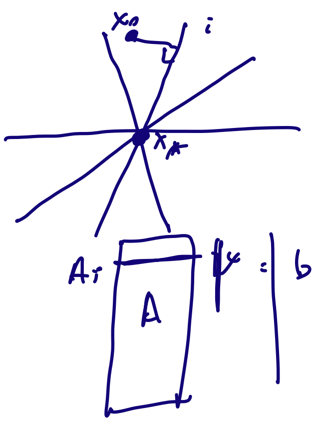
$A \in \mathbb{R}^{m \times n}$

$m \gg n$

Direct methods: $x = A^{\dagger} b$

Iterative methods: e.g. gradient descents, projection-based methods

Randomized Kaczmarz method:



1) initialize x_0

2) $x_{k+1} = x_k - \dfrac{(\langle A_i, x\rangle - b_i)\, A_i}{\|A_i\|_2^2}$

$i = i_0$ with prob $\dfrac{\|A_{i_0}\|^2}{\|A\|_F^2}$

3) until convergence

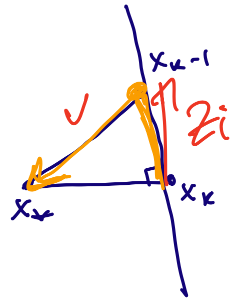
Convergence thm (Strohmer, Vershynin, '09)

$$\boxed{\mathbb{E}\,\|x_k - x_*\|_2^2 \le \Big(\underbrace{1 - \frac{\sigma_{\min}^2(A)}{\|A\|_F^2}}_{= \rho(A)}\Big)^k \|x_0 - x_*\|^2}$$

linear order of convergence under random selection rule

Proof idea:

$$\|x_k - x_*\|_2^2 = \|x_{k-1} - x_*\|_2^2 - \boxed{\|x_{k-1} - x_k\|_2^2}$$



$\|x_{k-1} - x_k\|_2^2 = \langle v, z_i \rangle, \quad v = x_{k-1} - x_*$

where $z_1, \dots z_n \dots$ is iid sample from

the discrete distribution

$$Z = \frac{A_i}{\|A_i\|_2} \text{ with prob } \frac{\|A_i\|^2}{\|A\|_F^2}$$

normal to ith equation

→ Let's assume all $\|A_i\| = 1$

$$\mathbb{E}\langle v, Z\rangle^2 = \sum_{i=1}^{m} \frac{1}{m} |\langle v, A_i \rangle|^2 \geq \frac{\boxed{\sigma_{\min}^2(A)}\|v\|_2^2}{\boxed{m}}$$

$$\boxed{\frac{\sigma_{\min}^2(A)}{m}} = \frac{\sigma_{\min}^2(A)}{\|A\|_F^2}$$

∎

Note: • for random iid models

$m \geq \delta n \qquad \mathbb{P}\{\sigma_{\min}(A) \leq c\sqrt{m}\} \leq e^{-cm}$

(no moments needed, Tikhomirov '14)

• coherent rows result in slow convergence

What if the system is not consistent?

$Ax = b + \varepsilon$

Convergence theorem (Needell, '10)

$$\mathbb{E}\|x_k - x_*\|^2 \leq (1 - \rho(A))\|x_{k-1} - x_*\|^2 + \frac{\|\varepsilon\|_\infty^2}{\rho(A)}$$

This works well for small noise

What if we have sparse adversarial corruptions?

Note:

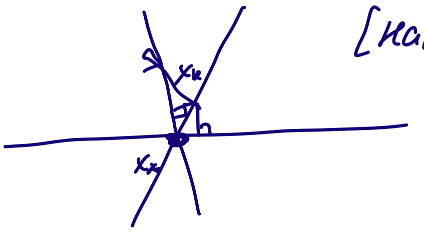
Robustness to noise → additive, multiplicative, … ?

→ random
- Gaussian
  - heavy-tailed…
- Byzantine, …

→ arbitrary under some conditions
- bounded
- adversarial and sparse

e.g. Dalayan, Thompson '19 vector estimation in linear models

---



[Haddock, Needell '18]

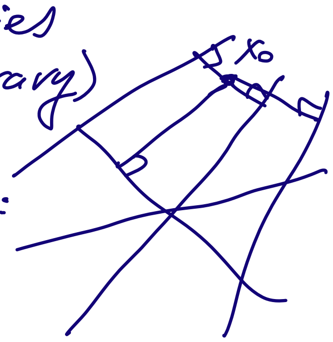
Key: Residual information (distances to hyperplanes)

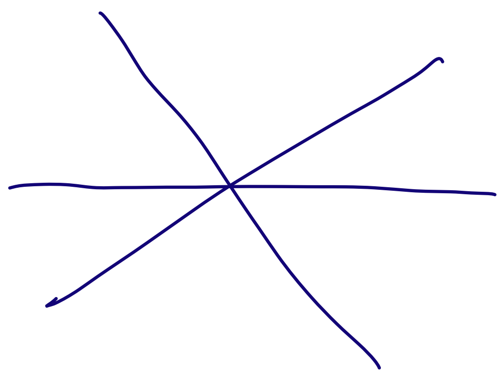
Issues:
1. With more corruptions we do not progress towards solution
2. Some corrupted equations have small residuals

Quantile RK

$$A x = b + \varepsilon$$

ε ← at most βm non-zero entries (arbitrary)





Algorithm (Quantile RK):

1) At each step x_k:
- Compute a sample of residuals for $i_1, \dots i_T$

$\{ |\langle A_{i_\ell}, x_k\rangle - b_{i_\ell}| , \ \ell = 1, \dots T \}$

- Compute q-Quantile of this set $Q_k(T)$
- For the next sampled index i
  - if $|\langle A_i, x_k\rangle - b_i| \leq Q_k(T)$ proceed with the projection
  - otherwise, stay

Convergence thm (Haddock, Needell, Rebrova, Swartworth '20)

If • A is $m \times n$ matrix, $m \geq Cn$

- x_*: $Ax_* = \tilde{b}$ and we have access to b
  $b = \tilde{b} + \varepsilon$, $\text{supp}(\varepsilon) \leq \beta n$
- A has isotropic, independent, subgaussian, unit norm rows
- A_{ij} have centered and bounded density functions

Then
$$\mathbb{E}\, \|x_k - x_*\|^2 \leq \left(1 - C_q \frac{\sigma_{\min}^2(A)}{\|A\|_F^2}\right)^k \|x_0 - x_*\|^2$$

if β is smaller than some positive constant and $q \leq \frac{1}{2} - \beta$ w/prob $1 - ce^{-cqm}$.

Plan:

1) Proof idea and RMT
2) De-randomization (Steinberger '21)
3) Quantile SGD for linear systems

Proof idea

a) Quantiles of the residual concentrate at $\mathcal{O}\left(\frac{\|x_k - x_*\|}{\beta\sqrt{n}}\right)$ with high probability (uniformly over

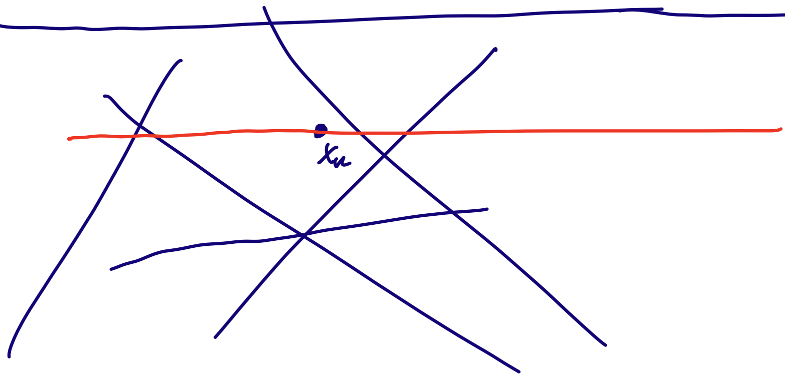
βm all $x = x_k - x_*$)

So, we banned all the large steps

Formally,

$\boxed{\text{Lemma}}$

$$\mathbb{P}\left(\begin{array}{l}\text{for every } x \in \mathbb{R}^n, \ |\langle A_i, x\rangle| \le \frac{C\|x\|}{\beta\sqrt{n}} \\ \text{for all but } \beta m \text{ indices } i \in [m]\end{array}\right) \geq 1 - e^{-cn}$$



b) As long as $\{\beta < \frac{1}{2},\ q < \frac{1}{2} - \beta\}$, probability to accept uncorrupted row p is $> \frac{1}{2}$

- Conditioned on projecting onto a bad equation, we do not lose too much

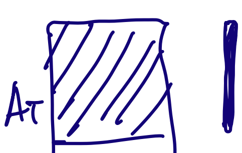
- Conditioned on projecting onto a good equation, we have the convergence rate of regular Kaczmarz algorithm <u>on the submatrix</u>

$\boxed{\text{Lemma}}$

$$\mathbb{P}\Big(\inf_{\substack{T \subset [m] \\ |T| \geq \alpha m}} \sigma_{\min}(A_T) \gtrsim \frac{\alpha^{3/2}}{24 D}\, \sigma_{\min}(A)\Big) \geq 1 - 3e^{-c\alpha m}$$

uniform smallest singular value

density constant!



Cf: restricted smallest singular value (Oymak Tropp '17)

Note: union over (known) bounds for fixed sub-mat. is too large

Proof sketch

$$\sigma_{min}(A_T) = \inf_{y\in S^{n-1}} \|A_T y\|_2 \geq \inf_{\underbrace{x\in \mathcal{N}}_{\varepsilon\text{-net}}} \|A_T x\| - \varepsilon \underbrace{\|A_T\|}_{\text{bounded}}$$

For fixed $x \in \mathcal{N}$

$$\mathcal{E}_i^x = \left\{ |\langle a_i, x\rangle|^2 < \frac{\alpha^2}{64 D^2}\cdot \frac{1}{n} \right\} \qquad \mathbb{P}(\mathcal{E}_i^x) \leq \frac{\alpha}{4}$$

Chernoff's bound: only few of $\mathcal{E}_i^x$ hold

So, $\|A_T x\|_2$ is large enough for any fixed x

Taking union bound over a net in S^{n-1} is much cheaper than over $\binom{m}{\alpha m}$ submatrices. ☺

∎

This result fails for discrete distributions!

E.g. consider symmetric Rademacher matrix

and $x = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0, \dots, 0)$

$\langle A_i, x\rangle = 0$ with prob $\frac{1}{2}$

For $\alpha < \frac{1}{2}$, x lies in a kernel of some $\alpha m \times n$ submatrix with high prob.

[Steinberger '21] De-randomization

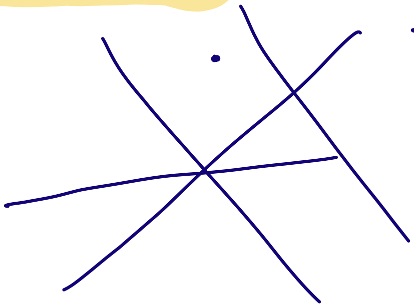
① $\|A\|$ and $\min_{|S| = \alpha m} \sigma_{min}(A_S)$ are the only 2 quantities that parametrize convergence

/ in sub-gaussian bounded density case we recover uncorrupted Kaczmarz rate /

Q's
- Upper bound for $\min_{|S| = \alpha m} \sigma_{min}(A_S)$?
- Heavy-tails?

- What additional restrictions on x can regularize $\min \sigma_{min}(A_S)$?
  $|S| = \alpha m$

Quantile SGD



Idea: control step size via the residual quantile $Q_k(T)$

Algorithm (Quantile SGD):

1) At each step x_k:
   - Compute a sample of residuals for $i_1, \dots i_T$
     $$\{ | \langle A_{i_\ell}, x_k \rangle - b_{i_\ell} | , \ \ell = 1, \dots T \}$$
   - Compute q-Quantile of this set $Q_k(T)$
   - Project cautiously:
     $$x_{k+1} = x_k - Q_k(T) \cdot \operatorname{sgn}(\langle A_i, x \rangle - b_i) A_i$$

<u>Convergence theorem</u> (Haddock - Needell - R - Swartworth '20)

Under the same conditions for A and $m > Cn \log n$,

$$\mathbb{E} \| x_k - x_* \|^2 \leq \left(1 - C_q \frac{\sigma_{min}^2(A)}{\|A\|_F^2} \right)^k \|x_0 - x_*\|_2^2$$

Connection between Kaczmarz algorithm and stochastic gradient descent

[Needell, Srebro, Ward '13]:

RK can be viewed as SGD with L_2-loss and a certain step size

$$\left\{\begin{array}{l} \text{solving } x_* = \arg\min_x \|Ax - b\|_2^2 \\ x_{k+1} = x_k - (\langle A_i, x_k\rangle - b_i)\cdot A_i \end{array}\right\}$$

Here, consider L_1-loss:

$$\boxed{x_* = \arg\min_x \|Ax - b\|_1}$$

standard relaxation of the minimization of L_0-norm

[Candes, Tao]'05

[Candes, Rudelson, Tao, Vershynin]'05

SGD with L_1-loss:

$$x_{k+1} = x_k - \eta_k \cdot \operatorname{sgn}(\langle A_i, x_k\rangle - b_i)\, A_i$$

Quantile!

<u>Convergence proof idea</u>:

1) The optimal choice for η_k minimizes

$$\mathbb{E}(\|x_{k+1} - x_*\|^2).$$

Analytically,

- $\eta_k^* = \frac{1}{m}\sum_{j=1}^{m} \operatorname{sgn}(\langle A_i, x_k\rangle - b_i)\langle e_k, a_i\rangle$

$e_k = x_k - x_*$

not known

- $\mathbb{E}(\|e_{k+1}\|_2^2) = \left(1 - \left(\frac{\eta_k^*}{\|e_k\|}\right)^2\right)\|e_k\|^2$
- $\frac{\eta_k^*}{\|e_k\|} \geq \frac{1}{m} \left(\sim \frac{\sigma_{\min}^2(A)}{\|A\|_F^2}\right)$ for subgaussian matrices

So, the rest is concentration results:

a) step-sizes of similar size result

in the same order of convergence $\frac{1}{n}$

3) Quantile statistic is within factor 2 from η_k^* with high probability

Further directions:

- mixed corruption models
- amount of oversampling and dependence on it
- robustness beyond linear systems.

Special comment:

Blocking equations together results in significant speed-up:

$$\boxed{x_{k+1} = x_k + (A_\tau)^{+} \cdot (b_\tau - A_\tau x_k)}$$

randomly chosen sub-block of A

Convergence theorem (Needell, Tropp '12)

$$\mathbb{E}(\|x_k - x_*\|_2^2) \le \left(1 - c\,\frac{\sigma_{\min}^2(A)}{\|A\|^2 \log m}\right)^k \|x_0 - x_*\|^2$$

if all the blocks are well-conditioned

$$\frac{m}{|\tau|} \cdot \max_\tau \|A_\tau\|_2^2 \lesssim \|A\|^2 \log(m) \frac{1}{\delta^2}(1+\delta)$$

Existence of good partitions (pavings) is related to Kadison-Zinger conjecture

Sketching with continuously distributed matrix (e.g. Gaussian) leads to more explicit estimates for any block size.

$$x_k = x_{k-1} + (S^T A)^+ (S^T b - S^T A x_k)$$

<u>Convergence thm</u> (Rebrova, Needell, '19)

$$\mathbb{E}\|x_k - x_*\|_2^2 \le \left(1 - \frac{s\,\sigma_{\min}^2(A)}{(9\sqrt{s}\|A\| + C\|A\|_F)^2}\right)^k \|x_0 - x_*\|_2^2$$

for Gaussian sketch S of the size $s \times m$

<u>Key step:</u>

$$s_{\min} \mathbb{E}\, \|(S^T A)^+ \cdot S^T A\|_2^2$$

- decoupling of 2 dependent matrices
  + matrix deviation inequality

$$\mathbb{E} \sup_{x \in S^{n-1}} \|S^T A x\|_2 = \mathbb{E} \sup_{w \in AS^{n-1}} \|S^T w\|_2$$

$$\le \sqrt{s}\,\|A\| + C\,\|A\|_F$$

- alternative approach, via rank-one update formula

  [Derezinski, Liang, Liao, Mahoney '20]

$$\left(1 - \frac{C_1}{\sqrt{r}}\right) \bar{P}_\perp \preceq \mathbb{E}\left(I - (S^T A)^+ S^T A\right) \preceq \left(1 + \frac{C_2}{\sqrt{r}}\right) \bar{P}_\perp$$

where the surrogate matrix $\bar{P}_\perp$ bounds the spectrum of the inverse projection