Iterative linear solvers and random matrices

New bounds for the block Gaussian sketch-and-project method

Liza Rebrova

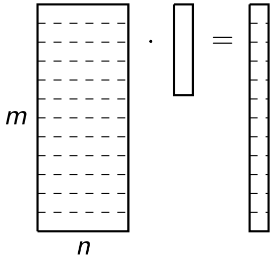
Department of Mathematics
UCLA

SOCAMS, Caltech, April 2019

Joint work with Deanna Needell

Model: overdetermined linear system

$$A \cdot x_* = b$$



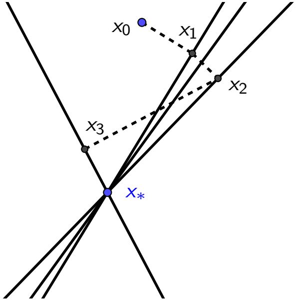
A is a tall $m \times n$ matrix ($m \gg n$) assumed to have full column rank. Notations: A_i - rows of A,

$$\sigma_{min}^2 = eig_{min}(A^T A) = 1/\|A^{-1}\|^2{}_{L_2 \to L_2}$$

Randomized Kaczmarz method

Starting at some $x_0 \in \mathbb{R}^n$:

1. Choose $i = i(k) \in [m]$ with probability $\|A_i\|_2^2 / \|A\|_F^2$
2. Define $x_k := x_{k-1} + \frac{b_i - A_i^T x_{k-1}}{||A_i||^2} A_i$
3. Repeat until $\|Ax_k - b\|_2 < \varepsilon$ (some threshold)



Convegence theorem (Strohmer - Vershynin 2009)

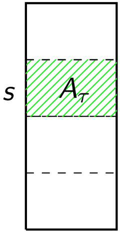
The randomized Kaczmarz converges to x_* *linearly in expectation:*

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

Block Kaczmarz Method

Starting at $x_0 \in \mathbb{R}^n$:

1. Choose A_τ a block row subset at random, $\tau = \tau(k) \subset [m]$, $|\tau| = s$
2. Define $x_k := x_{k-1} + (A_\tau)^\dagger(b_\tau - A_\tau x_k)$
3. Repeat until $\|Ax_k - b\|_2 < \varepsilon$



Convegence theorem (Needell - Tropp 2012)

The block Kaczmarz converges to x_* in expectation with accelerated rate

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

if all blocks are well-conditioned: for some $\delta \in (0,1)$,
number of blocks $\cdot\, max_\tau \|A_\tau\|_2^2 \lesssim \|A\|_2^2 \log(m)\frac{1}{\delta^2} \cdot (1+\delta)$.

Sketch-and-project methods

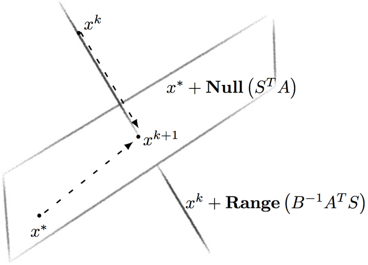
Gower - Richtárik (2015):

instead of $Ax = b$, solve $S^T Ax = S^T b$

$S = m \times s$ sketch matrix, if $s \ll m$ (sketched system is easier)

Iteration:

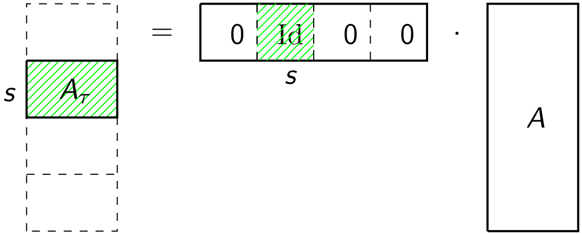
$$x_k := x_{k-1} + (S^T A)^\dagger (S^T b - S^T A x_k)$$
$$= (\mathrm{Id} - (S^T A)^\dagger S^T A) x_k + (S^T A)^\dagger S^T b.$$



Discrete random sketches and Kaczmarz methods

$$A_i = (0, \ldots, 0, 1, 0, \ldots, 0) \cdot A$$

$$A_\tau = \left[\ 0 \ \middle| \ \mathrm{Id} \ \middle| \ 0 \ \right] \cdot A = S^T A; \quad b_\tau = S^T b$$



Sketch-and-project methods with $S =$ (randomly placed identity completed by zeroes) are randomized Kaczmarz methods

Gaussian sketching

$$A_\xi := \xi^T \cdot A, \ \text{where } \xi \sim N(0, \mathrm{Id})$$

$$A_S := S^T \cdot A, \ \text{where } S \text{ is } m \times s \text{ gaussian random matrix}$$

$$\boxed{A_S} := \boxed{N(0,1) \text{ i.i.d.}} \cdot \boxed{A}$$

Gaussian sketch-and-project method takes gaussian random matrices S with i.i.d. entries as sketches.

Results 1: convergence rate

Convegence theorem (R - Needell 2019)

The gaussian block Kaczmarz method converges to x_* with the rate

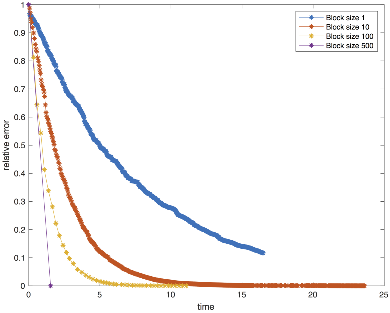
$$\mathbb{E}\|x_k - x_*\|_2^2 \le \left(1 - \frac{s\sigma_{min}^2(A)}{(9\sqrt{s}\|A\| + C\|A\|_F)^2}\right)^k \|x_0 - x_*\|_2^2,$$

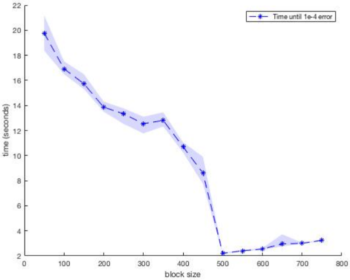
where $1 \le s \le m$ is the dimension of the gaussian sketch S.

- recovers "standard rate" $\sigma_{min}^2(A)/\|A\|_F^2$ for $s = 1$
- per iteration performance improves with increasing s
- actually, cputime performance also improves with increasing s

Better convergence for bigger sketch size

For $A = 50000 \times 500$ i.i.d. gaussian matrix:





Left: time(s) vs relative error for the varying sketch size $s = 1, 10, 100, 500$;
right: block size vs average time until relative error $1e - 4$

Proof ideas: random matrices

1. We need to estimate $\mathbb{E}\|(S^TA)^\dagger \cdot S^TAx\|_2^2$ from below - a product of two (dependent!) random matrices

2. S is $m \times s$ standard normal i.i.d. matrix.

$$\mathbb{E}\|S^TAx\|_2^2 = s\|Ax\|_2^2 \geq s\sigma_{min}^2(A)$$

But we need a high probability statement for any $s \geq 1$:

$$\mathbb{P}(\|S^Tv\|_2^2 > \|v\|^2 s/10) \geq 0.5$$

for any $v \in \mathbb{R}^m$ and $s \geq 1$ - Cramér's concentration theorem.

3.

$$\mathbb{E} \sup_{x \in S^{n-1}} \|S^T A x\|_2 \leq \sqrt{m}\|A\|_2$$

Can we get a better estimate? Yes!

$$\mathbb{E} \sup_{x \in S^{n-1}} \|S^T A x\|_2 = \mathbb{E} \sup_{w \in AS^{n-1}} \|S^T w\|_2 \leq \sqrt{s}\|A\| + C\|A\|_F.$$

To show 3.: apply matrix deviation inequality:

$$\mathbb{E} \sup_{w \in U} \|S^T w\|_2 \leq \sqrt{s} \sup_{w \in U} \|w\|_2 + C\gamma(U),$$

to the ellipse $U := AS^{n-1}$. Here, $\gamma(U)$ is gaussian complexity of the set U:

$$\gamma(U) := \mathbb{E} \sup_{w \in U} |\langle \xi, w \rangle|, \text{ where } \xi \sim N(0, I_n)$$

Results 2: sampling sketches from finite collection

We could select sketches from the pre-sampled collection of gaussian random matrices

Theorem (R - Needell)

Let $\mathcal{S} = \{S_1, \dots, S_N\}$ *be a set of* $m \times s$ *random matrices with i.i.d. standard normal entries,* $m^{2.5} \leq N \leq e^{m/3}$*. Then, with probability at least* $1 - 3/m$*, for any initial estimate* x_0*, finite block gaussian Kaczmarz method converges with the rate*

$$\mathbb{E}\|x_k - x_*\|_2^2 \leq \left(1 - \frac{s}{36m\kappa^2(A)}\right)^k \|x_0 - x_*\|_2^2.$$

In practice, the collection $\mathcal{S}$ can be much smaller, about $|\mathcal{S}| \sim m/s$

Results 3: Solving noisy systems

If the system is inconsistent, we can search for least-squares problem solution with gaussian block Kaczmarz method:

$$x_* = \text{argmin}_x \|Ax - b\|_2^2$$

and the error (noise) $e := Ax_* - b$.

Theorem (R - Needell)

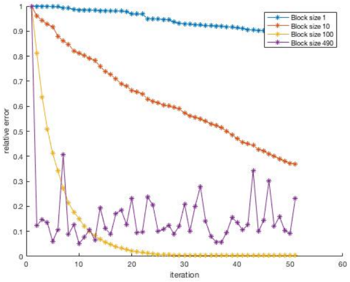
The gaussian block Kaczmarz method converges to x_* with the rate:

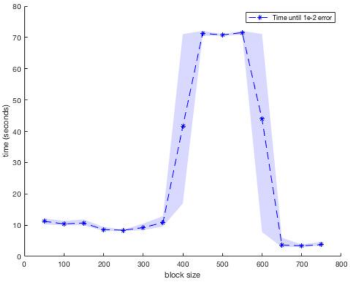
$$\mathbb{E}\|x_k - x_*\|_2^2 \leq r^k \|x_0 - x_*\|_2^2 + \frac{\|e\|_2^2}{s_{min}^4(A)} \cdot \left[\frac{(9\sqrt{s}\|A\| + C\|A\|_F)^2}{(\sqrt{n} - \sqrt{s})^2} \right]$$

Structurally differs from the noiseless case: diverges when $s \sim n$

Dependence on the block size in the noisy case

$$A = 50000 \times 500 \text{ i.i.d. gaussian matrix,}$$
$$e = \text{random gaussian noise, normalized: } \|e\|_2 = 0.05 * \|b\|_2$$

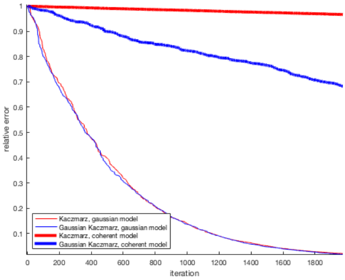


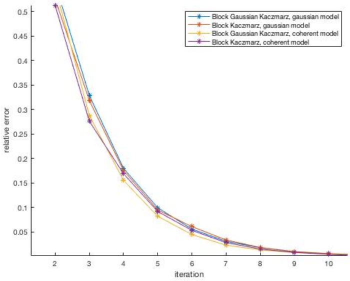


Left: iteration vs relative error for the sketch size $s = 1, 10, 100, 490$; right: block size vs average time until relative error $1e-2$; 70 sec is max allowed time

Is gaussian sketching practical?

$A = 50000 \times 500$ i.i.d. matrix:
$N(0,1)$ model (thin) and $Unif[0.8, 1]$ model (bold lines)



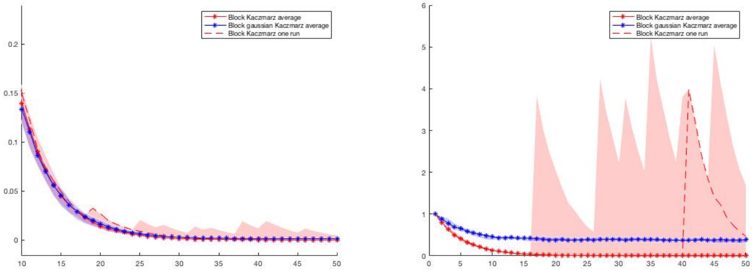


Left: $s = 1$, right: $s = 223$; blue = with gaussian sketching, red = without it

Gaussian sketching improves regular Kaczmarz for highly coherent systems when $s = 1$, but loses the advantage on bigger block sizes

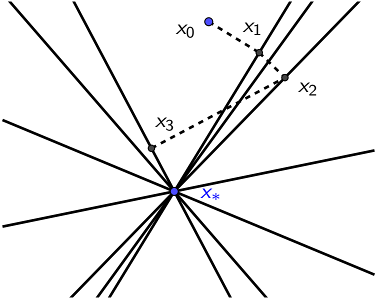
Gaussian sketching reduces variance

$$A = 50000 \times 500 \text{ i.i.d. matrix},$$
$$e = \text{spiky noise, 10 random spikes of size 50.}$$



Iteration vs relative error (median and range over 10 runs). Left: gaussian model, right: coherent model; blue = with gaussian sketching, red = without it.

Thanks for your attention!



Thanks for the pictures: Jamie Haddock, Gower&Richtarik "Randomized iterative methods ...", Matlab 2018b