Sketching for Motzkin's Iterative Method for Linear Systems

Liza Rebrova

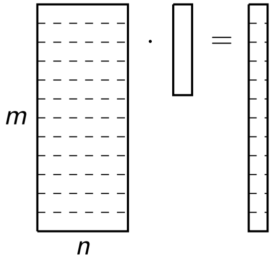
Department of Mathematics
UCLA

Asilomar Conference on Signals, Systems, and Computers, November 2019

Joint work with Deanna Needell

Model: overdetermined linear system

$$A \cdot x_* = b$$



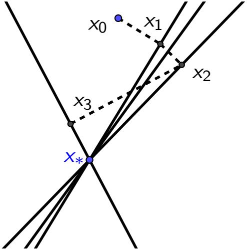
A is a tall $m \times n$ matrix $(m \gg n)$ assumed to have full column rank. Notations: A_i - rows of A,

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

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:*

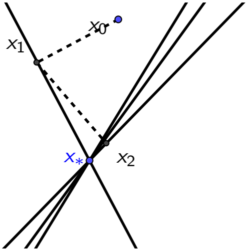
$$\mathbb{E}||x_k - x_*||_2^2 \leq \left(1 - \frac{1}{\tilde{\kappa}(A)}\right)^k ||x_0 - x_*||_2^2.$$

where $\tilde{\kappa}(A) = \frac{||A||_F^2}{\sigma_{min}^2(A)}$ *is a condition number of A.*

Relaxation (Motzkin's) method

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

1. Choose
   $i := \text{argmax}_{j \in [m]} (A_j x_k - b_j)^2$
2. Define $x_k := x_{k-1} + \frac{b_i - A_i^T x_{k-1}}{||A_i||^2} A_i$
3. Repeat until $\|Ax_k - b\|_2 < \varepsilon$



Theorem (Haddock - Needell 2018)

The randomized Kaczmarz converges to x_* linearly in expectation:

$$||x_k - x_*||_2^2 \leq \prod_{i=0}^{k-1} \left(1 - \frac{\sigma_{min}^2(A)}{4\gamma_i(A)}\right) ||x_0 - x_*||_2^2,$$

where $\gamma_i(A) = \|Ax_i - Ax_*\|_2^2 / \|Ax_i - Ax_*\|_\infty^2$ is a dynamic range of the i-th residual.

Randomized Kaczmarz method:

- could get stuck in “similar” equations
- iterations are fast
- provable (convergence in expectation)

Motzkin method:

- iterations make good progress
- slower iterations (search of the best equation is slow...)
- dynamic range is theoretically estimated only in some special cases

E.g., for A with independent standard normal entries

$$\gamma_k \sim \|A\|_F^2 n / \log(m-k),$$

which shows accelerated convergence when $\log(m-k) > n$.

Sketch-and-project framework

Gower - Richtárik (2015):

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

$S = m \times s$ sketch matrix, assume $s \ll m$

idea: to solve an easier $s \times n$ system instead of the original

Iteration:

- Pick random S from some distribution
- $x_k := x_{k-1} + (S^T A)^\dagger (S^T b - S^T A x_k)$

Note! Taking $S = e_k$ (randomly at each iteration) makes $S^T A = A_k$ (k-th row) and recovers randomized Kaczmarz method.

Sketching for Motzkin

Idea: instead of

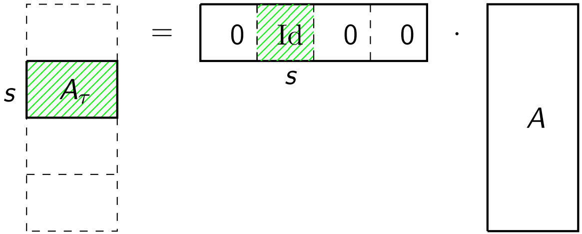
$$i := arg \max_{j \in [m]} (A_j x_k - b_j)^2$$

search over some smaller subset of indices.

Sketching for Motzkin

For example, for some subset $J \subset [m]$,

$$i := arg\max_{j \in J}(A_j x_k - b_j)^2 = arg\max_{j \in J}((A_J)_j x_k - (b_J)_j)^2$$



Block sketching is sketching A with

$S =$ randomly placed identity completed by zeroes.

Also known as SKM method.

(SKM) Sampling Kaczmarz Motzkin Method

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

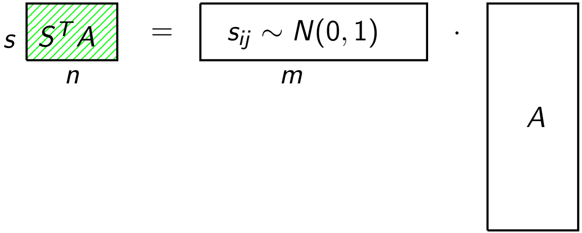
1. Choose $\tau_k \subset [m]$ to be a sample of size β constraints chosen uniformly at random among the rows of A.
2. From the β rows, choose $i := arg \max_j (A_j x_k - b_j)^2$
3. Define $x_k := x_{k-1} + \frac{b_i - A_i x_{k-1}}{||A_i||^2} A_i$
4. Repeat until convergence

DeLoera, Haddock, Needell (2019)
"A Sampling Kaczmarz-Motzkin Algorithm for Linear Feasibility"
SIAM Journal on Scientific Computing, vol. 39, 5, 66–87, 2017.

(GSM) Gaussian sketched Motzkin

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

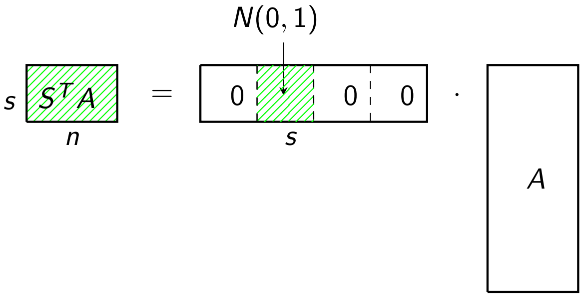
1. Sketch the system: $A_S := S^T A$ and $b_S := S^T b$, where S is an $m \times s$ standard normal matrix.
2. Choose $i := \text{argmax}_{j \in [s]}((A_S)_j x_k - (b_S)_j)^2$
3. Define $x_k := x_{k-1} + \frac{(b_S)_i - (A_S)_i x_{k-1}}{||(A_S)_i||^2}(A_S)_i$
4. Repeat until convergence



(sGSM) Sparse Gaussian Sketched Motzkin

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

1. Sketch the system: $A_S := S^T A$ and $b_S := S^T b$, where S has an $s \times s$. gaussian block.
2. Choose $i := \text{argmax}_{j \in [s]}((A_S)_j x_k - (b_S)_j)^2$
3. Define $x_k := x_{k-1} + \frac{(b_S)_i - (A_S)_i x_{k-1}}{||(A_S)_i||^2}(A_S)_i$
4. Repeat until convergence



Theoretical bounds

Theorem (Rebrova Needell 2019)

The GSM converges to x_* *linearly in expectation:*

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

The sGSM converges to x_* *linearly in expectation:*

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

Here, $c > 0$ *is an absolute constant,* $\tilde{\kappa}(A) = \frac{\|A\|_F^2}{\sigma_{min}^2(A)}$ *is the standard rate, and* A' *is the worst conditioned* $s \times n$ *row submatrix of* A*, namely,*

$$\tilde{\kappa}(A') = \operatorname{argmin}_{A_\tau} \tilde{\kappa}(A_\tau).$$

Well-conditioned sub-blocks

Every standardized matrix admits a good row paving:

Theorem

Let A be a $m \times n$ matrix. For any $\delta \in (0,1)$, there exists a partition on at most $\|A\|^2 \log m\delta^{-2}$. blocks, such that for every block A_τ

$$1 - \delta \leq \sigma_{min}(A_\tau) \leq \sigma_{max}(A_\tau) \leq 1 + \delta.$$

Moreover,

- Good paving could be constructed in poly-time
- For incoherent matrices a random row partition is likely a good paving

(Tropp, Popa, Bourgain, Tzafriri, Vershynin)

Bits of proof

- An analogue of the dynamic range

$$\mathbb{E}\frac{\|b_S - A_S x_k\|_\infty^2}{\|(A_S)_i\|_2^2} \geq \frac{(\mathbb{E}\|b_S - A_S x_k\|_\infty)^2}{\mathbb{E}(\|(A_S)_i\|_2^2)}$$

  by Jensen's inequality: the function $(x, y) \mapsto x^2/y$ is convex on the positive orthant

- Numerator:

$$\mathbb{E}_S\|S^T A(x_* - x_k)\|_\infty = \mathbb{E}\max_{i\in[s]}\langle S_i, v\rangle \geq c\|v\|_2\sqrt{\log s}$$

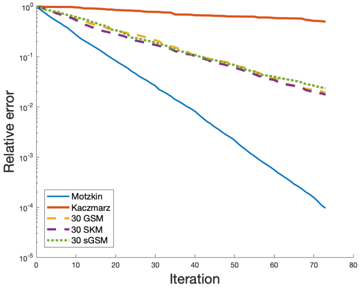
  estimate for the max of independent $N(0,1)$ random variables

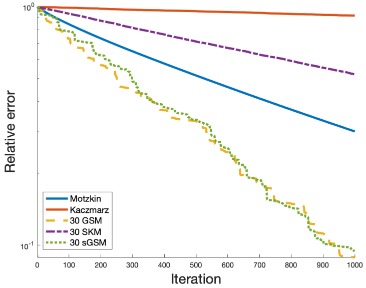
- Denominator:

$$\mathbb{E}\|(S^T A)_i\|_2^2 = \|A\|_F^2 \text{ (direct computation)}$$

Experiments on artificial datasets

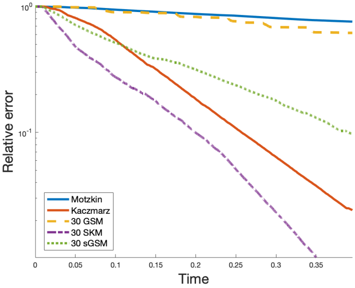
$A = 5000 \times 100$ i.i.d. matrix:
$N(0, 1)$ model and $Unif\,[0.8, 1]$ model





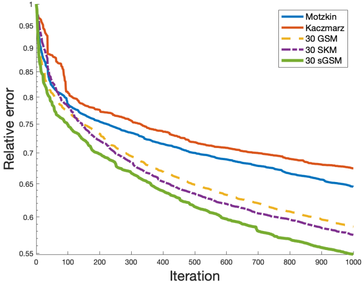
Experiments on artificial datasets-2

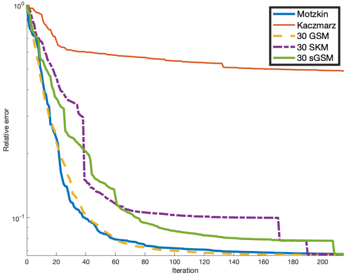
$A = 5000 \times 100$ i.i.d. matrix:
Unif $[0.8, 1]$ model



Experiments on real world datasets

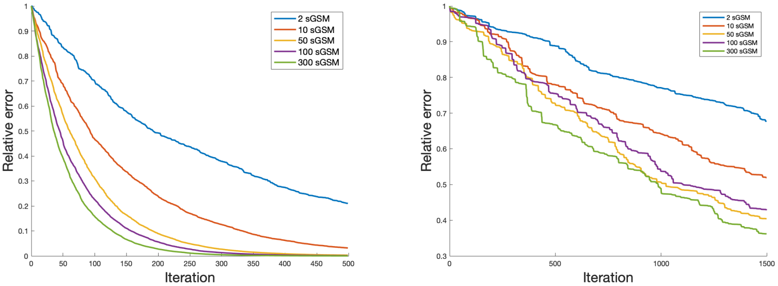
Left: GAS dataset (1000×128)
Right: COVTYPE dataset (5000×54)





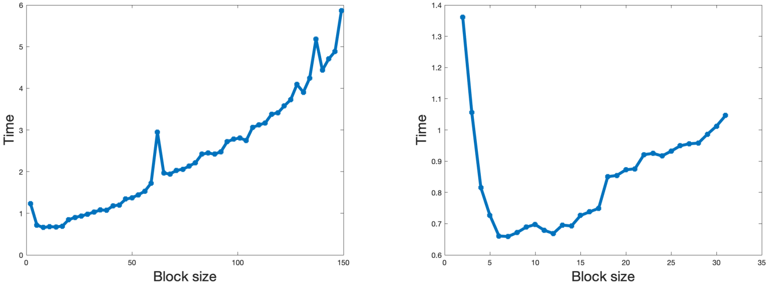
GSM method: dependence on sketch size

$$A = 5000 \times 500 \text{ i.i.d. matrix:}$$
Left: $N(0,1)$ model
Right: $Unif[0.8, 1]$ model



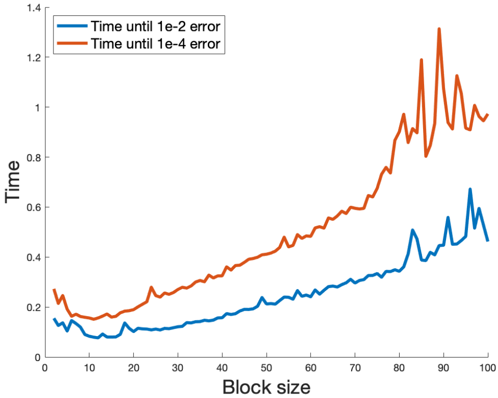
SparseGSM method: dependence on sketch size

$A = 5000 \times 500$ i.i.d. matrix:
$Unif[0.8, 1]$ model
Time until 1e-1 error, averaged over 10 iterations



SparseGSM method: dependence on sketch size

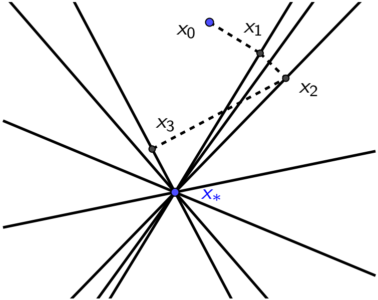
$A = 5000 \times 100$ i.i.d. matrix:
$Unif[0.8, 1]$ model
Time until 1e-2/1e-4 error, averaged over 20 iterations



Conclusions

- We consider 3 ways to sketch Motzkin's iterative method: SKM, GSM and sparseGSM
- We provide theoretical guarantees for the accelerated convergence of GSM (and sparseGSM for a well-conditioned matrix)
- We demonstrate experimentally some cases when sketched methods work better than both Kaczmarz and Motzkin (and when gaussian sketches outperform SKM)
- We investigate experimentally optimal block size for the sparseGSM method

Thanks for your attention!



Thanks for the pictures: Jamie Haddock, Deanna Needell, Matlab 2018b