ORF523

Dimension reduction for optimization problems

Sketching for dimension reduction

- **Iterative sketches:** light, small(er), exhaustive. Matters the distribution of all possible sketches
- **“Preserving” sketches:** one imprint of data preserving its crucial properties
- Intermediate regime exists, e.g. sketching SVD. One sketch but we might be willing to lose partial information.

Linear systems

Sketch-and-project

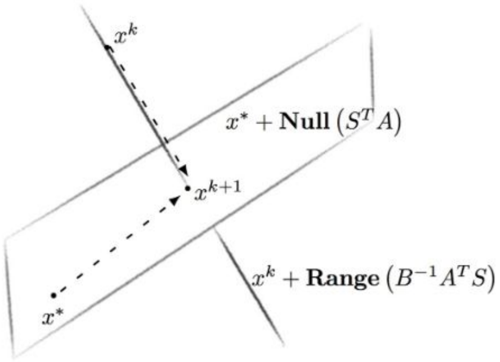
$S^T A x = S^T b$

$\boxed{\text{Instead of } A\mathbf{x} = \mathbf{b}, \text{ solve } S^T A\mathbf{x} = S^T\mathbf{b}}$

$m \boxed{S^T} \; n = \boxed{S}$

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

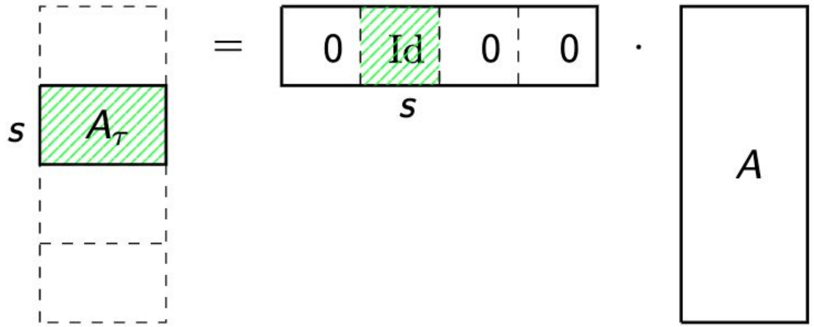
$$\mathbf{x}_k = \mathbf{x}_{k-1} + (S^T A)^\dagger (S^T\mathbf{b} - S^T A\mathbf{x}_k)$$



Discrete random sketches and Kaczmarz methods

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

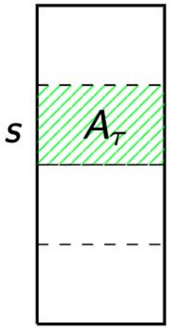
$$A_\tau = \left[\ 0 \,\middle|\, \mathrm{Id} \,\middle|\, 0\ \right] \cdot A = S^T A; \quad \mathbf{b}_\tau = S^T b$$



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

Block Kaczmarz Method

Assume that for all the rows $\|\mathbf{A_i}\| = 1$.



Starting at $\mathbf{x}_0 \in \mathbb{R}^n$:

1. Choose A_τ a block row subset at random, $\tau = \tau(k) \subset [m]$, $|\tau| = s$
2. Define $\mathbf{x}_k := \mathbf{x}_{k-1} + (A_\tau)^\dagger(\mathbf{b}_\tau - A_\tau \mathbf{x}_k)$
3. Continue until convergence (or for a certain number of steps).

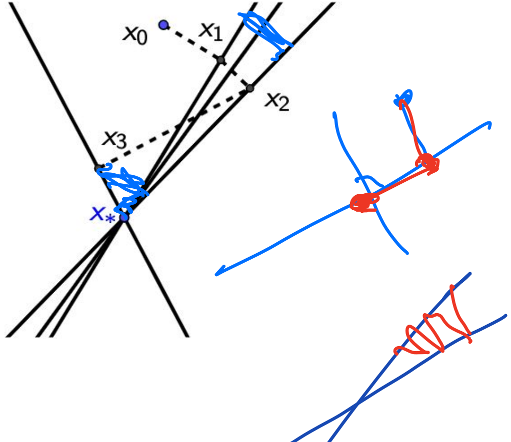
- Recall: sketch-and-project update rule $\mathbf{x}_k = \mathbf{x}_{k-1} + (S^T A)^\dagger (S^T \mathbf{b} - S^T A \mathbf{x}_k)$.
- Informally, this works if all the block subsets we can choose are well-conditioned. The existence of these *good pavings* is tightly related to Kadison-Zinger conjecture.
- For incoherent random models (say, subgaussian) any paving is good with high probability

Randomized Kaczmarz (RK) method

Assume that for all the rows $\|\mathbf{A_i}\| = 1$.

Starting at $\mathbf{x}_0 \in \mathbb{R}^n$:

1. Project current iterate to $\mathbf{A_i}$:
   $\mathbf{x}_{k+1} = \mathbf{x}_k - (\langle \mathbf{A_i}, \mathbf{x}_k \rangle - \mathbf{b}_i)\mathbf{A_i}$,
   where $i \sim \mathit{Unif}\{1, \ldots, m\}$;
2. Continue until convergence (or for a certain number of steps). ?



- Geometrically, each index i corresponds to a hyperplane in $\mathbb{R}^n$. RK projects orthogonally onto a randomly chosen hyperplane.
- If the rows are not normalize, the next i is chosen with the probability proportional to the L_2-norm of the i-th row.
- Convergence rate depends on $\sigma^2_{min}(\mathbf{A}) := \lambda_{min}(\mathbf{A}^T\mathbf{A})$.

Convergence rates

Theorem (Strohmer - Vershynin 2009)

For a system $A\mathbf{x}_* = b$, RK converges to x_* linearly in expectation:

$$\mathbb{E}||\mathbf{x}_k - \mathbf{x}_*||_2^2 \leq \left(1 - \frac{\sigma_{min}^2(A)}{\|A\|_F^2}\right)^k ||\mathbf{x}_0 - \mathbf{x}_*||_2^2.$$

Theorem (Needell - Tropp 2012)

The block Kaczmarz converges to $\mathbf{x}_*$ in expectation with accelerated rate

$$\mathbb{E}||\mathbf{x}_k - \mathbf{x}_*||_2^2 \leq \left(1 - c\frac{\sigma_{min}^2(A)}{||A||^2 \log m}\right)^k ||\mathbf{x}_0 - \mathbf{x}_*||_2^2,$$

if all blocks are well-conditioned: for some $\delta \in (0,1)$,

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

One sketch can be used for approximation

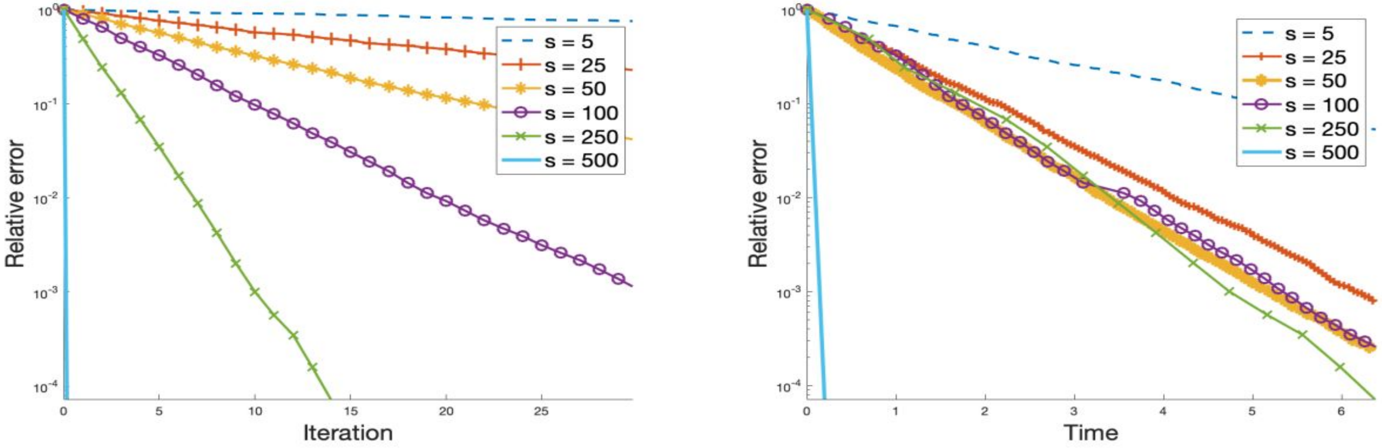


FIGURE 1. Gaussian model: iteration (left) and time (right) vs error for the varying block size s.

Least squares

Convergence rates for least squares

Theorem (Needell 2010)

For a least squares problem such that $e = \min \|A\mathbf{x} - b\|_2$, RK converges to a "horizon" around x_*:

$$\mathbb{E}||\mathbf{x}_k - \mathbf{x}_*||_2^2 \leq \left(1 - \frac{\sigma_{min}^2(A)}{\|A\|_F^2}\right)^k ||\mathbf{x}_0 - \mathbf{x}_*||_2^2 + \frac{n\|e\|_\infty^2}{\sigma_{\min}^2(A)}.$$

Theorem (Needell - Tropp 2012)

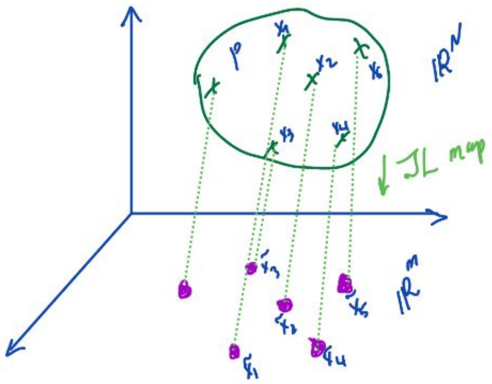
The block Kaczmarz converges to $\mathbf{x}_*$ in expectation with accelerated rate

$$\mathbb{E}||\mathbf{x}_k - \mathbf{x}_*||_2^2 \leq \left(1 - c\frac{\sigma_{min}^2(A)}{||A||^2 \log m}\right)^k ||\mathbf{x}_0 - \mathbf{x}_*||_2^2 + \frac{c\|e\|_2^2}{\sigma_{\min}^2(A)}.,$$

if all blocks are well-conditioned: for some $\delta \in (0,1)$,

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

Distance-preserving sketches



Johnson-Lindenstrauss lemma: There exists a linear function from $\mathbb{R}^N$ to $\mathbb{R}^m$ ϵ-preserves distances between p points for $m \geq c_\eta \epsilon^{-2} \ln p$.

- This function can be realized as an i.i.d. subgaussian random matrix
- Other matrix models work; these models are data-oblivious
- Works for all p-element sets

Johnnson-Lindenstrauss transform

Essentially, we need a matrix $S \in \mathbb{R}^{m \times N}$ such that

$$\left| \|Sx\|_2^2 - \|x\|_2^2 \right| \leq \epsilon \|x\|_2^2 \text{ for any } x \in Set - Set$$

Theorem

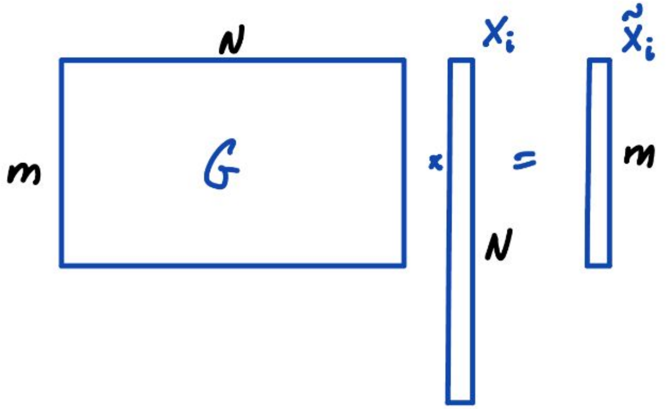
(Larsen, Nelson, 2016) For any $p, N \geq 2$, there exists a set of p vectors in $\mathbb{R}^N$ so that any linear map $\mathbb{R}^N \to \mathbb{R}^m$, ϵ-preserving distances between them, must have $m \gtrsim \epsilon^{-2} \ln p$.

S is a (ε, δ, s)-JL transform if for any s-element subset of $\mathbb{R}^n$

$$(1 - \varepsilon)\|x\|^2 \leq \|Sx\|_2^2 \leq (1 + \varepsilon)\|x\|_2^2$$

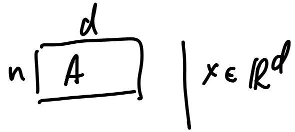
for any $x \in \mathcal{S}$ with probability at least $1 - \delta$.

Fast JL embeddings



- In the interesting regime for dimension reduction, N is large and $G \cdot X$ is heavy.
- Sparse and Fourier-based realizations of G, frequently with logarithmic losses in optimality ($\ln N$)
- There exists a (ε, δ, s) JL-transform with $m = O(\varepsilon^{-2} \log(s\delta^{-1})$ and $O(\varepsilon^{-1} \log(s\delta^{-1})$ non-zero entries per column.

Sketching least squares with JL transform



Theorem (Sarlos, 2006) Thm 12

Let $A \in \mathbb{R}^{n \times d}$ *and* $\hat{x} = \arg\min \|Ax - b\|_2$, *and* $x' = \arg\min \|SAx - Sb\|_2$, *where* $S \in \mathbb{R}^{m \times n}$ *is a* (ε, δ, S)-JL map such that $\underline{m \geq C\varepsilon^{-2} d \log d}$. *Then, with probability at least* $1/3$,

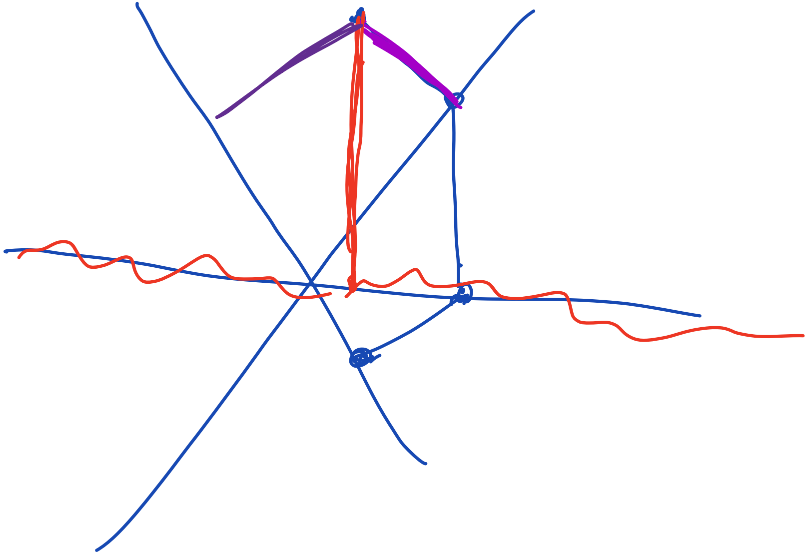
$$\|\hat{x} - x'\|_2 \leq \frac{\varepsilon}{\sigma_{\min}(A)} \min \|Ax - b\|_2.$$

Note: this defines S, also depending on how optimized a particular model of JL transform is.

For more details, there is a link to this paper on class website.

Side remark: geometry matters in iterative solvers too…

[Eldar, Needell]

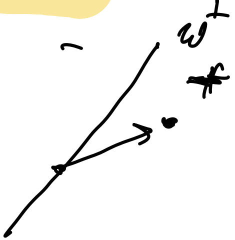


Application: Sketching SVMs

$z_i \in \{\pm 1\}$ (a_i, z_i) - classification data, $a_i \in \mathbb{R}^n$

$$w^* = \underset{w \in \mathbb{R}^n}{\arg\min} \left\{ \frac{1}{C} \sum_i [1 - z_i \langle w, a_i \rangle]_+^2 + \frac{1}{2} \|w\|_2^2 \right\}$$

z_i: classification label



↓ dual

$$x^* = \arg\min \|Bx\|_2^2 \; : \; x \geq 0 \quad \sum_{i=1}^{d} x_i = 1$$

where $B = \left[(AD)^T \cdot \frac{1}{C} I \right]$, $D = \mathrm{diag}(z_i)$

Simplex-constrained quadratic program

$$\hat{x} = \arg\min \|SBx\|_2^2 \quad x \geq 0, \; \sum x_i = 1$$

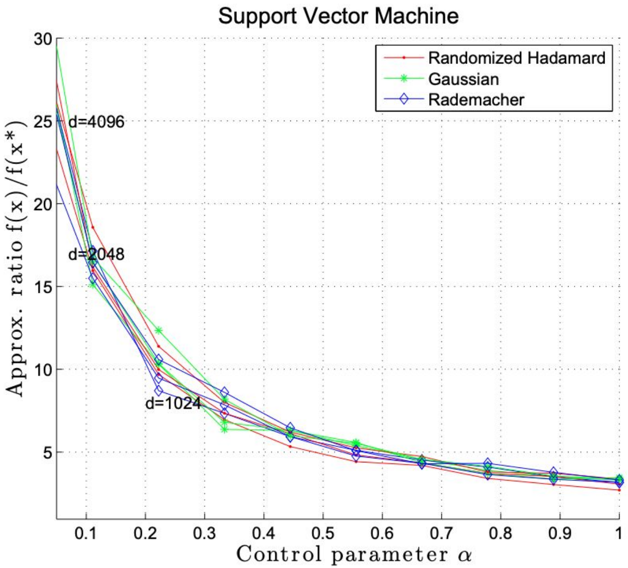
Theorem (Pilanci Wainwright, '14)

A sub-Gaussian sketch with

$$m \geq \frac{c_0}{\varepsilon^2} \|x^*\|_0 \log(N) \cdot \max \frac{\|a_j\|_2^2}{\sigma_k^-(A)}$$

produces ε-optimal solution with probability

$$1 - c_1 \exp(-c_2 m \varepsilon^2)$$



$$\sigma_k^-(A) = \min_{\substack{\|z\|_2 = 1 \\ \|z\|_1 \leq 2\sqrt{k}}} \|Az\|_2^2$$

restricted smallest singular value

Sketching SDPs [Bluhm, Franca, 18]

$$\max \quad \mathrm{Tr}(CX)$$

$$\text{s.t.} \quad \mathrm{Tr}(A_i X) \leq b_i$$

$$X \succeq 0$$

Standard form?

Sketching symmetric matrices

$\Phi(X) = SXS^T$, S is a JL transform

Equivalent JL definition: $\forall\, v, w \in \text{Set}$

$$|\langle Sv, Sw\rangle - \langle v, w\rangle| \leq \varepsilon \|v\|_2 \cdot \|w\|_2$$

Lemma 1 If $Q_1, \dots Q_k \in \mathrm{Sym}(n)$, $m \geq \sum_{i=1}^{k} \mathrm{rk}(Q_i)$, S is (ε, δ, m)-JL

Then $\mathbb{P}[\forall i,j \; \mathrm{Tr}(SQ_iS^TSQ_jS^T) - \mathrm{Tr}(Q_iQ_j)) <$

$< 3\varepsilon \|Q_i\|_1 \cdot \|Q_j\|_1] \leq 1-\delta$

Shatten-1 norm $(\sum |\lambda_i|)$

But! To have $\varepsilon \|Q_i\|_2 \|Q_j\|_2$ on the right one needs

$m = O(n)$!

Approximate SDP problem

Algorithm: S is a (ε, δ, m)-JLT,

$m \geq rk(X^*) + rk(C) + \sum_{i=1}^{..} rk(A_i)$

Solve a smaller relaxed problem...

$$\boxed{\begin{aligned} &\max \ Tr(SCS^T Y) \\ &\text{s.t.} \ Tr(SA_iS^T Y) \leq b_i + 3\varepsilon \|A_i\|_1 \|X^*\|_1 \\ &\qquad Y \succeq 0 \end{aligned}}$$

Why? $Tr(SCS^TSX^*S^T) \geq \underbrace{Tr(CX^*)}_{\alpha} - 3\varepsilon\|X^*\|_1\|C\|_1$

optimal

Upper Bound on sketched solution:

$\alpha_S + 3\varepsilon\|X^*\|_1\|C\|_1 \geq \alpha$

Lower bound depends on stability

$\frac{\alpha_S}{1+3\varepsilon k\eta} \leq \alpha \quad (\eta = tr X^*, \ k = \max_i \|A_i\|_1$

All SDP problems cannot be sketches this way

Thm: If Φ is a random linear map that estimates a value of any SDP within $1 \le \tau \le \frac{2}{\sqrt{3}}$ factor with high probability $\Rightarrow$

$$m = O(n^2)$$

Follows from hardness of estimating of the operator norm of a matrix

[Woodruff Sketching as a tool for numerical linear algebra '14]

Sketching Convex Programs:

- *Dimensionality reduction of SDPs through sketching* A. Bluhm, D. Stilck Franca (2018)
- *Scalable Semidefinite Programming* A. Yurtsever J. Tropp, O. Fercoq, Madeleine Udell, and Volkan Cevher (2021)
- *Randomized Sketches of Convex Programs with Sharp Guarantees* M. Pilanci, M. J. Wainwright (2014)
- *Randomized Projection Methods for Convex Feasibility* I.Necoara, P. Richtarik, A. Patascu (2018)

Convex feasibility and iterative sketching

Exactness $\quad \operatorname{dist}^2_{\mathcal{X}}(x) \le k \;\mathbb{E}\left[\operatorname{dist}^2_{\mathcal{X}_S}(x)\right] \quad \forall x \in \mathbb{R}^n \quad (*)$

Thm $\quad \mathcal{X}$ - convex set with non-empty interior

$\bar{x} \in \mathcal{X}: \quad B_\rho(\bar{x}) \in \mathcal{X}$

$\mathcal{X}_S$ is a family of stochastic approximations

(S is random
$\rho = \{p_S\}$ is distribution) $\Rightarrow (*)$ holds with

$$k = \frac{\max \|x - \bar{x}\|^2}{\rho^2 \min\limits_{S \in \Omega} p_S}$$

Example : $\mathcal{X} = \bigcap \mathcal{X}_S$

Convex and conic optimization outlook

- Convexity tends to make optimization problems easier
- But there are hard convex problems and easy non-convex
- A family of tractable convex problems

$$\mathrm{LP} \subset \mathrm{QP} \subset \mathrm{QCQP} \subset \mathrm{SOCP} \subset \mathrm{SDP} \;\Big\} \subset \mathrm{CP}$$

- For “easy” (P-) problem: polynomial algorithm might be still slow…
- For hard problems:
  - complexity theory can justify (NP-)hardness
  - special cases can be solved exactly
  - convex relaxations give bounds (SDP can be more efficient than LP!)
  - approximate solutions are possible, randomization can help build them
  - and more: e.g., sequential convex programs

Many applications considered…

- Machine learning (SVMs)
- Signal detection (probabilistic estimates)
- Control (finding stabilizing controllers)
- Combinatorial optimization (graph problems)
- Compressed sensing (low-rank fitting, matrix completion)
- Approximation theory (randomized rounding)
-

And more, including finance (Markowich portfolio optimization), information theory …

Looking backwards…

~~Tentative~~ list of topics

- Math review
- **Unconstrained nonlinear optimization**: first and second order optimality conditions
- **Convex analysis and convex optimization problems**
- **Duality and certificates of infeasibility**
- From linear programs to **positive semidefinite programs**
- **Relaxations to linear and** SDP **problems**
- **Complexity theory**
- Applications to combinatorial optimization, data science etc
- Approximate solutions and randomization
- Convex feasibility problems and randomization
- Randomized sketching and dimensionality reduction for convex optimization

Final exam

- Take home May 4 (9am) -- May 11 (noon)
- No collaborations
- You can only refer to the material proved in class/notes, everything else must be justified in your work
- Coding component

- Liza's office hour: 11am-12:30pm May 3

Thanks for your attention!