CS 388R: Randomized Algorithms, Fall 2023 November 27th, 2023

Lecture 25: Randomized Numerical Linear Algebra-2

Prof. Eric Price *Scribe: Liangchen Liu, Ziheng Chen*

NOTE: THESE NOTES HAVE NOT BEEN EDITED OR CHECKED FOR CORRECTNESS

1 Overview

In the last lecture, we have introduced the basic setup for least square problems in the language of linear algebra, where we discussed the computational efficiency for direct methods (Gaussian elimination) and iterative methods (gradient descent and conjugate gradient method). We also showed that an ϵ-subspace embedding leads to an $\mathcal{O}(\epsilon)$-sub-optimal solution via the sketch and solve method. In this lecture, we'd like to study the following questions:

1. how to find an oblivious subspace embedding S (or **OSE** for short)?
2. how to evaluate the matrix product SA quickly?
3. furthermore, how can an OSE be applied to derive a good pre-conditioner?

2 Recap on problem setup

For the least square problem

$$x^* := \arg\min \|Ax - b\|^2$$

with $A \in \mathbb{R}^{n\times d}, b \in \mathbb{R}^n$ and $n \gg d$, the optimal solution has the following closed form

$$x^* = A^\dagger b = \left(A^T A\right)^{-1} A^T b \tag{1}$$

if $\text{rank}A = d$. The computational cost for Eqn. 1 is $\mathcal{O}\left(nd^2 + d^3\right)$ without fast matrix multiplication and $\mathcal{O}\left(nd^{1.38} + d^{2.38}\right)$ with (fancy) fast matrix multiplication.

To improve the quadratic (or super-linear) dependency on d, we propose the following "sketch & solve" algorithm:

Definition 1. A random matrix S is considered a $\underline{d\text{-dimensional } (\epsilon, \delta)\text{-OSE}}$ (oblivious subspace embedding) if

$$\|Sx\| = (1 \pm \epsilon)\|x\|, \forall x \in U$$

holds for for any d-dimensional U with probability $1 - \delta$.

Algorithm 2 (Sketch & Solve)**.** *Given d and accuracy tolerance $\epsilon > 0$, choose a $(d+1)$-dimensional OSE (oblivious subspace embedding) $S \in \mathbb{R}^{m\times n}$ with $m = \widetilde{\mathcal{O}}\left(d/\epsilon^2\right)$ and solve*

$$\widehat{x} := \arg\min \|SAx - Sb\|_2.$$

Remark 3. Recall that a subspace embedding is capable of preserving the norm up to a relative error of ϵ, for any input x from a given subspace. Here, the subspace is considered to be spanned by the columns of $(A|b)$, having a dimension of $(d+1)$. However, we don't want the choice of S to rely on the coefficient matrix A, thus we introduce the notion of "oblivious" that requires (the random matrix) S to be a subspace embedding of any matrix A, with sufficiently high probability.

In the previous lecture, we have showed that a good OSE leads to a good approximated solution in the sense that

Proposition 4. *If S is an OSE of ϵ accuracy, the sketch & solve solution is good in the sense that*

$$\|A\widehat{x} - b\|_2 \leq (1 + \mathcal{O}(\epsilon)) \|Ax^* - b\|_2.$$

3 Distributional-JL

3.1 Yielding OSEs

OSEs are hard to construct since we need to preserve the norm of all vectors in the subspace at the same time. Instead, we start with a weaker family that preserves each vector at a time, called the distributional-JL (or **d-JL** for short) property. Then, we study the geometry of high dimensional balls to yield a union-bound that bridges d-JL and OSE.

Definition 5. A random matrix S is considered (ϵ, δ)-d-JL if

$$\|Sx\| = (1 \pm \epsilon) \|x\|$$

holds for any x with probability $1 - \delta$.

In the previous class, we have showed the existence of an ϵ_1-net of the unit ball in d dimensions.

Proposition 6. *Let $B := \left\{x \in \mathbb{R}^d : \|x\| \leq 1\right\}$ be the unit ball. Then, there exists an ϵ_1-net $N \subset B$ of size $|N| < \left(1 + \frac{2}{\epsilon_1}\right)^d$, namely*

$$\forall x \in B, \exists y \in N \text{ s.t. } \|x - y\| \leq \epsilon_1.$$

Remark. We fix $\epsilon_1 = 0.1$ in the following passage which is sufficient to derive desired matrix projection.

The following theorem connects d-JL with OSE. The intuition behind this theorem is that a d-dimensional subspace has $2^{\mathcal{O}(d)}$ "effectively different points".

Theorem 7. *If S is (ϵ, δ)-d-JL, then for any $x \in B$, $\|Sx\| = (1 \pm \mathcal{O}(\epsilon)) \|x\|$ holds with probability $1 - \delta \cdot |N|^2$. In other words, S is an $\left(\mathcal{O}(\epsilon), \delta \cdot |N|^2\right)$-OSE.*

Proof. Without loss of generality, we scale $x \in N$ properly so that it is on the unit sphere (i.e. $\|x\| = 1$). We iteratively define $x_0, x_1, \cdots \in \mathbb{R}^d$ and $r_1, r_2, \cdots \in \mathbb{R}^d$ by the following procedure:

1. Since N is an $\frac{1}{10}$-net of B, there exists $x_0 \in N$ s.t. $\|x - x_0\| \leq \frac{1}{10}$; let $r_1 := x - x_0$.

2. Similarly, since $\|10r_1\| = 10\|x - x_0\| \leq 1$, it can be approximated by $x_1 \in N$ s.t.

$$r_2 := r_1 - \frac{1}{10}x_1 \Rightarrow \|r_2\| \leq \frac{1}{100}.$$

3. We iteratively define x_k as the element in N that approximates $10^k r_k$ and $r_{k+1} := r_k - \frac{1}{10^k}x_k$.

We can telescope the definitions for r_k and arrive at

$$x = x_0 + r_1 = x_0 + \frac{1}{10}x_1 + r_2 = \cdots = \sum_{i=0}^{\infty} 10^{-i}x_i.$$

A naive attempt to bound the magnitude of x via triangular inequality, which leads to a sub-optimal result:

$$\|Sx\| \leq \sum_{i=0}^{\infty} 10^{-i}\|Sx_i\| \leq \frac{10}{9}(1+\epsilon),$$
$$\|Sx\| \geq \|Sx_0\| - \sum_{i=1}^{\infty} 10^{-i}\|Sx_i\| \geq \frac{8}{9} - \frac{10}{9}\epsilon.$$

Although one can require N to be an ϵ-net instead of a $\frac{1}{10}$-net, the extra cost is eventually added to $m = \mathcal{O}\left(\frac{d}{\epsilon^2}\log\frac{1}{\epsilon}\right)$ which is not optimal. To circumvent this issue, one should bound the norm via inner product. Let $y_i := \frac{x_i}{10^i}$, then

$$\|Sx\|_2^2 = \left\langle \sum_{i=0}^{\infty} Sy_i, \sum_{i=0}^{\infty} Sy_i \right\rangle = \sum_{i=0}^{\infty}\|Sy_i\|_2^2 + \sum_{i<j} 2\langle Sy_i, Sy_j\rangle.$$

Due to polarization identity

$$\langle z, w\rangle = \frac{1}{4}\left(\|z+w\|^2 - \|z-w\|^2\right),$$

once S preserves all y_i and $y_i \pm y_j$ with probability $1 - \delta \cdot |N|^2$, we have

$$\begin{aligned}
\|Sx\|_2^2 &= \sum_{i=0}^{\infty}\|y_i\|_2^2 \pm \epsilon\sum_{i=0}^{\infty}\|y_i\|_2^2 + \sum_{i<j} 2\langle y_i, y_j\rangle \pm \sum_{i<j} 2\epsilon\|y_i\|\|y_j\| \\
&= \left(\sum_{i=0}^{\infty}\|y_i\|_2^2 + \sum_{i<j} 2\langle y_i, y_j\rangle\right) \pm \epsilon\left(\sum_{i=0}^{\infty}\|y_i\|_2^2 + \sum_{i<j} 2\|y_i\|\|y_j\|\right) \\
&= \|x\|_2^2 \pm \epsilon\left(\sum_{i=0}^{\infty}\|y_i\|_2\right)^2 \\
&= 1 \pm 1.3\epsilon.
\end{aligned}$$

□

Example 8. In the last class, we mentioned that with $m = \mathcal{O}\left(\epsilon^{-2}\log\delta^{-1}\right)$ and $S \in \mathbb{R}^{m\times n}$ having i.i.d. entries from $\mathcal{N}\left(0, \frac{1}{m}\right)$, S is (ϵ, δ)-d-JL. Thus, increasing m to $\mathcal{O}\left(d\epsilon^{-2}\log\delta^{-1}\right)$ gives an (ϵ, δ)-OSE family.

3.2 Fast d-JL

We introduce a few state-of-art techniques.

Definition 9. A random matrix $B \in \mathbb{R}^{m\times n}$ is called an ϵ-RIP (restricted identity property) matrix with order k if

$$\|Bx\| = (1 \pm \epsilon)\|x\|$$

for any k-sparse x (i.e. x has less than k non-zero entries).

Theorem 10 ([KW11] Thm 3.1)**.** *If $B \in \mathbb{R}^{m\times n}$ is an ϵ-RIP matrix with order k and D is a diagonal random matrix with $D_{ii} \in \{-1, 1\}$ uniformly, then BD is $\left(\mathcal{O}(\epsilon), 2^{-\Omega(k)}\right)$-d-JL.*

Let us consider the Fourier transform matrix $F \in \mathbb{R}^{n\times n}$ with entries

$$F_{ij} := e^{2\pi\sqrt{-1}ij/n}.$$

We now sub-sample m rows from F to form matrix $\widetilde{F}$, then

Proposition 11 ([KW11] Sec 4.1)**.** *If $m = \mathcal{O}\left(d\epsilon^{-2}\log\delta^{-1}\cdot\log^4 d\right)$, then $\widetilde{F}$ is $\left(d\log\delta^{-1}, \epsilon\right)$-RIP. Thus, FD is $\left(\mathcal{O}(\epsilon), \delta 2^{-\Omega(d)}\right)$-d-JL and $(\mathcal{O}(\epsilon), \delta)$-OSE.*

4 OSE as pre-conditioner

The sketch & solve algorithm is able to produce an ϵ-approximation solution with a cost of $\widetilde{\mathcal{O}}\left(nd + d^3\epsilon^{-2}\right)$, while the conjugate gradient method requires $\mathcal{O}\left(nd\log\frac{n}{\epsilon}\kappa(A)\right)$. To combine the advantages of both methods, one can derive a pre-conditioner from the sketch & solve method that reduces the condition number $\kappa(A)$. Furthermore, since the OSE preserves the norm so well, the conjugate gradient method is no longer needed and a simple gradient descent suffices.

We first introduce the basics of pre-conditioning. For a fixed, invertible matrix $R \in \mathbb{R}^{d\times d}$, we solve

$$y^* = \arg\min\|ARy - b\|_2$$

that recovers x^* by Ry^*. If we manage to find a good R s.t. $\kappa(AR)$ is small (say 1.1), then the overall computational cost is lowered.

In the most ideal case, we perform an SVD decomposition to A:

$$A = U\Sigma V^T$$

where $U \in \mathbb{R}^{n\times n}, V \in \mathbb{R}^{d\times d}$ are orthogonal matrices and $\Sigma \in \mathbb{R}^{n\times d}$ has diagonal entries only. Then, picking $R := V\Sigma^{-1}$ leads to

$$AR = U \Rightarrow \kappa(AR) = 1.$$

However, an SVD factorization takes $\mathcal{O}\left(nd^2\right)$ time which defeats the purpose of adopting sketching in the first place.

The alternative approach is to pick an OSE S and perform SVD on SA. The overall computational time is the sum of computing SA and a smaller scale decomposition, namely

$$\text{total time} = \mathcal{O}(n\log n\, d) + \mathcal{O}\left(md^2\right) = \widetilde{\mathcal{O}}\left(nd + d^3\right).$$

Then, for any $y \in \mathbb{R}^d$ with unit norm, we have

$$(1-\epsilon)\|ARy\| \geq \|SARy\| \geq (1+\epsilon)\|ARy\|$$

where $\|SARy\| = 1$, so AR is an approximately orthogonal matrix, leading to

$$\kappa(AR) = 1 + \mathcal{O}(\epsilon) \leq 2.$$

Eventually, conjugate gradient takes $\mathcal{O}\left(nd \log \frac{n}{\epsilon_2}\right)$ time to achieve an ϵ_2-accurate solution.

In fact, AR is so close to being orthogonal that there is no need to adopt conjugate gradient method. Since

$$\|R^T A^T AR - I\| \leq \mathcal{O}(\epsilon),$$

intuitively we have

$$y \approx R^T A^T ARy \approx R^T A^T b.$$

One can show that the following iterative scheme

$$\begin{aligned} y^{(1)} &:= R^T A^T b, \\ y^{(2)} &:= y^{(1)} + R^T A^T \left(b - ARy^{(1)}\right), \\ &\cdots \end{aligned}$$

converges to y^* fast!

References

[KW11] Felix Krahmer, Rachel Ward. New and improved Johnson-Lindenstrauss embeddings via the Restricted Isometry Property. *SIAM Journal on Mathematical Analysis*, 43(3):1269–1281, 2011.