COMS W6998-5: Algorithms through Geometric Lens (Fall'17) Sep 27, 2017

Lecture 7 - Fast Dimension Reduction, Fast Linear Algebra

Instructors: *Alex Andoni, Ilya Razenshteyn* Scribes: *Parita Pooj*

1 Introduction

Let's continue from last time where we started with the Fast JL Transform result by Ailon and Chazelle. Last lecture, we discussed subsampling as an approach to solve the problem and saw the places where it would break.

2 Fast Dimension Reduction

Restating the theorem we want to arrive at -

Theorem 1. *Fast Johnson-Lindenstrauss Transform (Ailon & Chazelle '04) [1]*
For every ϵ, δ ¿ 0, there exists a distribution over matrices $A \in \mathbb{R}^{s \times d}$ such that $\forall$ x we have

$$\mathbb{P}[\|Ax\|_2^2 \in s(1 \pm \epsilon)\|x\|_2^2] \geq 1 - \delta$$

where $x \to Ax$ can be done in $O(d \log d)$ time, where $s = O\Big(\frac{\log()1/\delta)}{\epsilon^2} \log(d/\delta)\Big)$.
The drawback of the time saved in computation is reflected in the slight increase in the bound for dimension s required for the above to hold true.

In the last, we started with the naive approach to this problem by trying to apply a random orthogonal matrix for A, and we realized how that doesn't help improve the computational time. Picking up from there, we move on to the construction showed by Ailon and Chazelle that was proven to help improve the computational time

2.1 Construction using Hadamard Matrices

$$x \to HDx$$

This is two step construction where:

1. $x \to Dx$ where D is a diagonal matrix with every diagonal entry as ± 1 at random

$$D = \begin{bmatrix} \pm 1 & 0 & 0 & \dots & 0 \\ 0 & \pm 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \pm 1 \end{bmatrix}$$

2. $Dx \to HDx$ where
   H is a Walsh-Hadamard matrix which can be written as:
   for $d = 2^t$

$$H_t = \frac{1}{\sqrt{2}} \begin{bmatrix} H_{t-1} & H_{t-1} \\ H_{t-1} & H_{t-1} \end{bmatrix}$$

2.2 Proof for Running Time

We can prove that the computational time required for the above construction is $O(d \log d)$ based on the following two claims:

Claim 2. $\|Dx\|_2^2 = \|x\|_2^2 \ \forall x$
and $x \to Dx$ *can be done in time* $O(d)$

This claim is fairly straightforward and doesn't need a proof for it.

Claim 3. $\|Hx\| = \|x\|_2^2 \ \forall x$
and $x \to Hx$ *can be done in time* $O(d \log d)$

Proof. The former part of the claim is easy to follow considering that all elements of the Hadamard matrix are $\pm\frac{1}{\sqrt{d}}$ which would imply that

$$\|Hx\|_2^2 = \sum_{i=1}^{d} \frac{1}{d} x_i^2 = \|x\|_2^2$$

We can show that the computation for Hx can be done in $O(d \log d)$ time by observing that the Hadamard matrix H_t can be constructed by repeating H_{t-1}. Thus, a hadamard matrix can be broken into four blocks where each block is essentially the same (The fourth block (2, 2) is simply $-H_{t-1}$ and hence, is essentially the same)

To convert $x \to Hx$ where $x \in \mathbb{R}^{2t}$, we break x in two parts:
$x = [x_1 x_2]$ where $x_i \in \mathbb{R}^{2t-1}$
Compute $y_1 = H_{t-1}x_1$ and $y_2 = H_{t-1}x_2$
Thus,

$$H_t x = \left(\frac{y_1 + y_2}{\sqrt{2},} \frac{y_1 - y_2}{\sqrt{2}} \right)$$

Thus, the computation takes $O(d \log d)$ □

2.3 Proof for concentration

To prove Fast Johnson-Lindenstrauss Theorem based on the above construction, it would be enough to prove the following lemma:

Lemma 4. $\forall x : \|x\|_2 = 1$

$$\mathbb{P}_D \left[\|HDx\|_\infty \leq O\left(\sqrt{\frac{\log(d/\delta)}{d}} \right) \right] \geq 1 - \delta$$

Proof. We are essentially trying to prove that the $L - \infty$ norm is concentrated around its mean over the random distribution for D. We rely on one of the basic properties of Hadamard matrix here which we have already seen above - all entries of the Hadamard matrix are $\pm\frac{1}{\sqrt{d}}$

We prove this by first proving that the first entry of HDx is concentrated around its mean, and then extend it for the norm by using the union bound property:

Definition 5. *Union Bound Property: for any finite or countable set of events $A_1, A_2, A_3, \ldots$, the probability that at least one of the events happens is no greater than the sum of the probabilities of the individual events.*

$$\mathbb{P}\Big(\bigcup_i A_i\Big) \leq \sum_i \mathbb{P}(A_i).$$

The first entry in the matrix (HDx) can be written as:

$$(HDx)_1 = \frac{1}{\sqrt{d}}(\pm x_1 \pm x_2 \pm \ldots \pm x_d)$$

$$\mathbb{E}(HDx)_1 = 0$$

We need to now show that the value of $(HDx)_1$ is concentrated around its mean. We can use the lemma:

Lemma 6 (Eq 1.2 from [3])**.**

$$\mathbb{P}[|x_1 \pm x_2 \pm \ldots \pm x_d| \geq t\|x\|_2] \leq e^{-\Omega(t^2)}$$

$$\mathbb{P}\left[|(HDx)_1| \leq C\sqrt{\frac{\log(d/\delta)}{d}}\right] \leq \frac{\delta}{10d}$$

Extending the above using union bound property, it follows that:

$$\mathbb{P}_D\left[\|HDx\|_\infty \leq O\left(\sqrt{\frac{\log(d/\delta)}{d}}\right)\right] \geq 1 - \delta$$

□

2.4 Final Construction for Fast JLT

The final construction by Ailon & Chazelle for the Fast Johnson-Lindenstrauss Theorem involves one additional step. We subsample the result from HDx using the subsampling matrix Π. Thus, we can reformulate it as:

$$\forall x \quad \mathbb{P}\Big[\|Ax\|_2^2 \in (1 \pm \epsilon)\|x\|_2^2\Big] \geq 1 - \delta$$

Here, $x \to Ax$, where $A = \Pi HD$ can be achieved with $s = O\Big(\log(d/\delta).\frac{\log(1/\delta)}{\epsilon^2}\Big)$ in $O(d \log d)$ time

We arrive at this using Lemma 6, and by using the result of Lemma 12 from subsampling discussed during last class which proved the following using Bernstien's inequality theorem:

$$\mathbb{P}\left[\|Ax\|_2^2 \in (1 \pm \epsilon)\frac{s}{d}\|x\|_2^2\right] \geq 1 - \delta \text{ for dimension } s = \Omega\left(\frac{log(1/\delta)}{\epsilon^2} d \frac{\|x\|_\infty^2}{\|x\|_2^2}\right)$$

Thus, the addition of projection matrix, with the analysis from the previous lecture extends Lemma 6 to prove the Fast Johnson-Lindenstrauss Theorem, with the assumption that $\frac{\|x\|_\infty^2}{\|x\|_2^2}$ is small, which is observed due to the transformation $x \to HDx$.

As we can see, we get the same guarantee as the original Johnson-Lindenstrauss Theorem with worse bound on s - reduced dimension or number of rows.

2.5 Analysis

Now, if we analyze the above, we can see that HD rotation "spreads out" the vector. For example, if we consider a vector $x = e_1 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$, which is a sparse matrix (kind of equivalent to a delta function), we get: $\forall i \quad (HDx)_i = \pm\frac{1}{\sqrt{d}}$ which is spread out. Here the diagonal matrix D is important because if Hadamard was applied directly, a uniformly spread out vector would become sparse. The randomness in D fixes it for the few dense x where it could have failed.

3 Fast Linear Algebra

3.1 Fast Algorithms for Linear Regression

Linear Regression is a method to model the relationship between variables using a linear model. Linear Least Squares is the most common approach to the problem. It is also commonly used in other areas of Machine Learning as well - for example, to find the Maximum Likelihood Estimate (for Gaussian models).

3.2 Introducing Linear Least Squares Problem

Definition 7. *Linear Least Squares Problem:*

$$\min_x \|Ax - b\|_2^2$$

where $n >> d$, $A \in \mathbb{R}^{n \times d}$, $x \in \mathbb{R}^d$ *and* $b \in \mathbb{R}^n$

One possible solution to the above is using a Moore-Penrose pseudoinverse. It is derived by setting the gradient w.r.t. x to zero:

$$\nabla_x \|Ax - b\|_2^2 = 2A^T(Ax - b) = 0$$

$$x = (A^T A)^{-1} A^T b$$

Here, $A^+ = (A^T A)^{-1} A^T$ is referred to as the pseudoinverse.

The bottleneck in computational complexity for this solution comes from computing $A^T A$ which takes $\theta(nd^2)$ time. Naive running time can be considered as $O(nd^2 + poly(d))$. (Note: There have been better bounds for this but the above can be considered in most reasonable settings).

Can we do better? Yes, with randomization and approximation analysis. Intuitively, the best we can hope for would be $\approx O(nd)$, which would be our aim when coming up with an algorithm.

3.3 Strategy

We can convert the problem
$\min_x \|Ax - b\|_2^2 \quad \rightarrow \quad \min_x \|\tilde{A}x - \tilde{b}\|_2^2$
where $\tilde{A} \in \mathbb{R}^{s \times d}$, $\tilde{b} \in \mathbb{R}^s$, $s << n$

In other words, we reduce n-dimensional vectors to s-dimensions to ease the computational complexity.

To do this reduction, we can introduce a random matrix $S \in \mathbb{R}^{s \times n}, s << n$, such that approximate inequality is preserved between $\|SAx - Sb\|_2^2$ and $\|Ax - b\|_2^2$

We can achieve the required result if S is a random matrix such that following property is satisfied:

$$\mathbb{P}_S\Big[\forall x \ \|SAx - Sb\|_2^2 \in (1 \pm \epsilon)\|Ax - b\|_2^2\Big] \geq 0.9 \tag{1}$$

We cannot rely on minimizing $\|SAx - Sb\|_2^2$ with the below property:

$$\forall x \ \mathbb{P}_S\Big[\|SAx - Sb\|_2^2 \in (1 \pm \epsilon)\|Ax - b\|_2^2\Big] \geq 0.9$$

because there could be another x which does not follow the above property, thus requiring another minimum.

3.4 Oblivious Subspace Embeddings (OSE)

Definition 8. *[2] An oblivious subspace embedding is a random matrix S such that for any fixed d-dimensional subspace $U \subset \mathbb{R}^n$, $d << n$, $\epsilon > 0$*

$$\mathbb{P}\Big[\forall y \in U \ \|Sy\|_2^2 \in (1 \pm \epsilon)\|y\|_2^2\Big] \geq 0.9$$

The equation (1) follows from the property of an Oblivious Subspace Embeddings.

U can be thought of as a subspace spanning $\langle$columns of $A, b\rangle$, dim $U \leq d + 1$

3.5 OSE embedding for S

Thus, we can use an OSE as S. We have -

$$\tilde{A} = SA, \tilde{b} = Sb$$

- How many rows should be in S?
- How quickly can we compute $y \rightarrow Sy$?

To answer the first question: **How many rows should be in S?**

Claim 9. *S must have $\geq d$ rows.*

This is straightforward to see since the system will become underdetermined if not.

A simple idea would be take S as an i.i.d. scaled gaussian, as we did for JLT.

Theorem 10. *If S is $s \times n$ matrix independently and identically distributed over a scaled Gaussian $\mathcal{N}(0, 1)$, the property given by the equation (1) holds if $s \geq \frac{d}{\epsilon^2}$*

Here, we will prove for a weaker bound for s, $s = O\Big(\frac{d\log(1/\epsilon)}{\epsilon^2}\Big)$.

Proof. A naive approach to proving would be to prove the theorem for one point and extend it using union bound. But since, we have infinite number of points, it won't be possible to do that. Instead, we try to discretize the space smartly. Let's assume that the space is a unit sphere in subspace U

$$\mathbb{S}^U = \{y \in U | \|y\|_2 = 1\}$$

Let $N_\epsilon \subset \mathbb{S}^U$ be the discretized "net" over the unit sphere, i.e., $|N_\epsilon| < \infty$

We can approach to the proof stepwise by proving the below in sequence:

1. $\forall y \in N_\epsilon$,
$$\mathbb{P}\Big[\|Sy\|_2^2 \in (1 \pm \epsilon)\|y\|_2^2\Big] \geq 1 - \frac{1}{10|N_\epsilon|}$$
2. With 1. proven, we can prove the below using union bound property
$$\mathbb{P}\Big[\forall y \in N_\epsilon : \|Sy\|_2^2 \in (1 \pm \epsilon)\|y\|_2^2\Big] \geq 0.9$$
3. After we prove 2. over the discretized space, we need to extend it so it holds true for all points in the subspace
$$\mathbb{P}\Big[\forall y \in U : \|Sy\|_2^2 \in (1 \pm \epsilon)\|y\|_2^2\Big] \geq 0.9$$

For the discretization to work, the number of rows $s = O\Big(\frac{\log|N_\epsilon|}{\epsilon^2}\Big)$
We define N_ϵ as an ϵ-net, formally defined as:

Definition 11. *N_ϵ is an ϵ-net of the unit sphere $\mathbb{S}^U$ if*

- $N_\epsilon \subset \mathbb{S}^U$
- $\forall y \in \mathbb{S}^U, \exists y' \in N_\epsilon$ *s.t.* $\|y' - y\|_2 \leq \epsilon$

The N_ϵ can be visualized as a tessellation in the subspace of the unit sphere $\mathbb{S}^U$ such that any point in the sphere can be ϵ-approximated by some point on N_ϵ.

Claim 12.
$$\exists\ \epsilon\text{-net } N_\epsilon \quad |N_\epsilon| = \left(\frac{c}{\epsilon}\right)^d$$

Claim 13. *For N_ϵ, proving 2. $\implies$ 3.*

Proof for Claim 12. We use the volume bound technique to prove that $\exists$ ϵ-net of $\mathbb{S}^{d-1} \subset \mathbb{R}^d$ of size $(\frac{c}{\epsilon})^d$. Consider the following approach to construct N_ϵ:

- Start with $Z = \{\}$ - empty set
- Add $z_1, z_2, \ldots \in S^{d-1}$ to Z greedily, until no point left y' left on the sphere that would break the requirement for an ϵ-net

After the construction, we have $Z = [z_1, z_2, \ldots, z_T]$.

The claim states that we can do the above construction with $T \leq \left(\frac{c}{\epsilon}\right)^d$. We can prove it by considering bowls (neighbourhoods) around each point, and consider their volumes. Let $v_i = B(z_i, \epsilon/2)$ be the volume of a bowl defined by the point z_i bounded by a neighbourhood of distance $(\epsilon/2)$. We know that -

- All v_i's must be disjoint
- $v_i \subset B(0, 1 + \epsilon/2)$

Bounding the total volume:

$$T.\left(\frac{\epsilon}{2}\right)^d \leq \left(1 + \frac{\epsilon}{2}\right)^d$$

$$T \leq \left(\frac{\left(1 + \frac{\epsilon}{2}\right)}{\epsilon/2}\right)^d$$

$$T \leq \left(\frac{2}{\epsilon} + 1\right)^d$$

$$T \leq \left(\frac{c}{\epsilon}\right)^d$$

Thus, we proved Claim 12. □

The rest of the proof for Theorem 10 will be continued in the next class. □

References

[1] Nir Ailon and Bernard Chazelle. Approximate nearest neighbors and the fast johnson-lindenstrauss transform. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, pages 557–563. ACM, 2006.

[2] Michael B Cohen. Nearly tight oblivious subspace embeddings by trace inequalities. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 278–287. Society for Industrial and Applied Mathematics, 2016.

[3] Iosif Pinelis. An asymptotically gaussian bound on the rademacher tails. *Electron. J. Probab.*, 17:22 pp., 2012.