CS 15-851: Algorithms for Big Data Spring 2024

Lecture 10 Part 2 — 03/28/2024

Prof. David Woodruff *Scribe: Cheng Ma*

1 Projection on Complicated Objects and Gaussian Mean Width

In previous lectures, we've seen that least squares regression finds the closest point y in a subspace K to a point b. Let us now consider a more general problem, where K is a possibly infinite set of points, and we have a point b and we want to compute the following (all norms here are Euclidean norms)

$$\min_{y \in K} ||y - b||_2$$

Similar to before, we can employ sketching. Let S be a sketching matrix, and we now want to output

$$y' = \arg\min_{y \in K} ||Sy - Sb||_2$$

whereby the arg minimizer y' in the sketch space also satisfies $||y' - b||_2 \leq (1+\varepsilon) \min_{y_k} |y - b|$. More generally, we want to preserve distances of all vectors in a set K, that is

$$\forall y, y' \in K |S(y - y')| = (1 \pm \varepsilon)|y - y'|$$

What properties of K determine the dimension and sparsity of S?

Let's consider the following examples to gain a sense of this question. What dimension of S is needed if K is:

- n arbitrary points in $\mathbb{R}^d$ If S is a matrix of i.i.d. Gaussians from the first lecture, we know that we need $O(\frac{\log n}{\varepsilon^2})$ rows in order to preserve distances as required by the Johnson-Lindenstrauss Lemma.
- n arbitrary points on a line in $\mathbb{R}^d$ Is S is a matrix of i.i.d. Gaussians, we only need $O(\frac{d}{\varepsilon^2})$ rows to preserve norms of all vectors in a d-dim subspace. Here $d = 1$ so we only need $O(1/\varepsilon^2)$ rows.

1.1 Spherical Mean Width

The previous examples motivate the Spherical Mean Width. Let K be a bounded subset in $\mathbb{R}^n$.

Definition (Width in direction u of a unit vector u)**.** The width in direction u for a unit vector u is defined as the following

$$\text{width in direction } u = \sup_{p,q \in K} < u, p - q >$$

You're taking two parallel hyperplanes whose normal vector is u and you're bringing them together until they're tangent with your set K.

Definition (Spherical mean width)**.** It's the average over all unit vectors u on the sphere is defined as

$$\mathop{\mathbb{E}}_{u}\left[\sup_{p,q\in K} < u, p-q >\right]$$

Intuitively, a bounded line has a large width only in one direction but on average it is very small. However, for a ball, on average the width is the same.

Definition (Gaussian Mean Width)**.** Let $g \sim N(0, I_n)$ be an i.i.d. Gaussian vector. Then the Gaussian mean width is defined as

$$g(L) = \mathop{\mathbb{E}}_{g}\left[\sup_{p,q\in K} < g, p-q >\right] = \Theta(\sqrt{n}) \cdot \text{spherical mean width}$$

Consider the following examples

- $K = S^{n-1}$, the unit sphere, has $g(K) = \Theta(\sqrt{n})$ because for any direction we consider, the width is 2.
- $K =$ set of unit vectors in a $d-$dimensional subspace of $\mathbb{R}^n$ has $g(K) = \Theta(\sqrt{d})$ because we can express the subspace as Ux where $U \in \mathbb{R}^{n\times d}$ has orthonormal columns. Hence when trying to calculate the Gaussian mean width, we're considering $\mathbb{E}_g\left[\sup_{\text{unit } x,y} < g, Ux - Uy >\right]$. But since U has orthonormal columns, we can recall from the first lecture that $gU = h$, a $d-$dimensional vector of i.i.d. Gaussians, we can express $< g, Ux - Uy >=< g, U(x-y) >\equiv< h, x-y >$, which gives us $\mathbb{E}_h\left[\sup_{x,y} < h, x-y >\right] = \Theta(\sqrt{d})$ since we've just reduced it to the previous problem and it's like taking the gaussian mean width over a $d-$dimensional unit ball.
- $K = t$ arbitrary unit vectors in $\mathbb{R}^n$ has $g(K) = \Theta(\log^{0.5} n)$ as explained below.

1.2 Gaussian Mean Width of t arbitrary unit vectors

Let $u^1, u^2, ..., u^t$ be t arbitrary unit vectors in $\mathbb{R}^n$. Let $g \in \mathbb{R}^n$ have i.i.d. $N(0,1)$ entries. Define random variables $Z_j =< u^j, g >$ which are $N(0,1)$ random variables. **We want to bound the quantity** $\mathbb{E}_g[\max_j Z_j]$

Fact: for an $N(0,1)$ random variable W and some $\lambda > 0$, $\mathbb{E}[e^{\lambda W}] = e^{\frac{\lambda^2}{2}}$

Firstly, we have that for any $\lambda > 0$,

$$\mathbb{E}[e^{\lambda \max_j Z_j}] \leq \sum_j \mathbb{E}[e^{\lambda Z_j}] \leq te^{\frac{\lambda^2}{2}}$$

The first inequality comes from the fact that quantity on the left only considers the max of Z_j, whereas the right quantity considers the sum of all Z_js. The second inequality comes from the above fact.

Secondly, for all $\lambda > 0$,

$$\mathop{\mathbb{E}}_{g}\left[\max_j Z_j\right] \leq (\frac{1}{\lambda}\log \mathbb{E}[e^{\lambda \max_j Z_j}]) \leq (\frac{\log t}{\lambda} + \frac{\lambda}{2}) \leq 2\sqrt{\log t}$$

The first inequality comes from Jensen's inequality and the concavity of log. The second inequality uses what we calculated from the first point and taking log of it. Finally, by setting $\lambda = \sqrt{\log t}$ we get the final inequality.

1.3 Sketching Bounds

Theorem 1 (Gordon's Theorem)**.** *Let K be a subset of S^{n-1}. A random Gaussian sketching matrix S with $\frac{g(K)^2}{\varepsilon^2}$ rows satisfies $\forall y, y' \in K, |S(y-y')|^2 = (1 \pm \varepsilon)|y-y'|^2$*

Essentially, the sketching dimension is determined by the Gaussian mean width. Tying this back to previous lectures, for a $d-$dimensional subspace, since $g(K) = \Theta(\sqrt{(}d))$, when we plug this back in we get that S needs d/ε^2 rows. For n arbitrary points, $g(k) = \Theta(\sqrt{n})$, using Gordon's Theorem shows we need $\frac{\log n}{\varepsilon^2}$ rows which gives us the JL-Lemma. These are special cases of Gordon's theorem.

What about sparse sketching matrices?

Theorem 2 (Bourgain, Dirksen, Nelson)**.** *S can have $m = g(K)^2 poly(\log n)/eps^2$ rows and $s = poly(\log n)/eps^2$ non-zeros per column in m and s satisfy a condition related to higher moments of $\sup_{p,q} < g, p-q >$*

2 Compressed Sensing

We are trying to estimate the vector $x \in \mathbb{R}^n$ by taking random "linear measurements". In the context of this class, we choose a random $r \times n$ sketching matrix S and observe Sx. We want to output a vector x' such that

$$||x - x'||_p = D \cdot \min_{k-\text{sparse} z} ||x - z||_q$$

where D is the distortion, also known as the ℓ_p/ℓ_q-guarantee.

Let x_k be the best $k-$sparse approximation to x. When x is a vector, it's just the vector containing the largest k coordinates in magnitude.

There are two schemes for estimating $x' \in \mathbb{R}^n$:

- Randomized "for-each" scheme
- Deterministic "for-all" scheme

CountSketch is a randomized scheme that achieves the ℓ_2/ℓ_2 guarantee with high probability:

$$||x - x'||_2 = O(1)||x - x_k||_2 (*)$$

2.1 CountSketch for Compressed Sensing

When S is a CountSketch matrix, it is a random linear map with $O(k \log n)$ rows. However, we can view it as $O(\log n)$ reptitions of hashing into $O(k)$ buckets. From previous lectures, we've seen that S estimates every coordinate x_i of x up to an additive error of $\frac{||x - x_k||_2}{\sqrt{k}}$.

If we output x' as a $2k-$sparse vector that consists of the top $2k$ estimates (with respect to magnitude) given by CountSketch, we can satisfy $(*)$ with high property.

Proof that the above scheme satisfies $(*)$:

Definition (Coordinate i is **heavy** if)**.**

$$|x_i| \geq \frac{|x - x_k|_2}{\sqrt{k}}$$

Note that **there can be at most 2k heavy coordinates**

Definition (Coordinate i is **super-heavy** if)**.**

$$|x_i| \geq 3\frac{||x - x_k||_2}{\sqrt{k}}$$

We claim that the set T of super-heavy coordinates is in the support of x'

$$\begin{aligned} |x - x'|_2 &\leq |(x - x')_T|_2 + |(x - x')_{[n]\setminus T}|_2 && \text{(By the triangle inequality)} \\ &\leq \sqrt{|T|} \cdot \frac{|x - x_k|_2}{\sqrt{k}} + |(x - x_k)_T|_2 + |(x_k - x')_{[n]\setminus T}|_2 \end{aligned}$$

(The first time comes from the previous line. The 2nd and 3rd term come from the triangle inequality.)

$$= O(|x - x_k|_2) \qquad \text{(As desired)}$$

Clearly the first two terms are $O(|x-x_k|_2)$. However, we can show that the last term $|(x_k-x')_{[n]\setminus T}|_2 \leq \sqrt{3k}\max_{i\in[n]\setminus T}(x_k - x')_i < O(\frac{|x - x_k}{\sqrt{k}})$. The $3k$ comes from the fact that x_k has k non-zero entries, x' has $2k$ non-zero entries, therefore their difference has at most $3k$ non-zero entries. The final inequality requires casing:

1. The difference is bounded by additive error if the index is in the support of both vectors
2. The difference is bounded by the heavy guarantee if the coordinate is in x_k.
3. If the coordinate is in x' then we need to show it is bounded by the quantity plus some additive error.

2.2 No Deterministic Algorithm Achieves ℓ_2/ℓ_2

Let's consider the case of $k = 1$. AFSOC that S is a deterministic sketching matrix with $r = o(n)$ rows. It sufficies to show that there exists a vector x in the kernel of S with $|x|_\infty \geq C|x - x_1|_2$ for any constant $C > 0$.

The idea here is that we only see that $Sx = 0$. Since we need to satisfy multiplicative error, we don't know if 0 was the input or some other vector. So we need to output $x' = 0$ and have that $|x - x'|_2 = |x|_2$ is our error. We are in trouble if $|x|_2 > O(1)|x - x_1|_2$ for $k = 1$.

WLOG, we can assume that S has orthonormal rows. Since $\sum_i ||Se_i||_2^2 = r$ we know there exists an i with $|Se_i|_2^2 \leq \frac{r}{n}$. Let $x = e_i - S^T Se_i$ which means that x is in the kernel of S since $Sx = Se_i - (SS^T)Se_i = 0$. Note here that $S^T S$ is the projection matrix.

But since

$$|x|_\infty^2 \geq |x_i|^2 = (e_i^t e_i - e_i^T S^T Se_i)^2 \geq (1 - \frac{r}{n})^2$$

While simultaneously we have that

$$|x - x_1|_2 \leq |x - e_i|_2 = |S^T Se_i|_2 = |Se_i|_2 \leq \sqrt{\frac{r}{n}} = o(1)$$