CS 15-859: Algorithms for Big Data Fall 2017

Lecture 11-Part 2 – Thursday 11/16/2017

Prof. David Woodruff *Scribe: Dimitris Konomis*

1 Introduction

This lecture presents, at a survey level the following topics:

- Projection onto Complicated Objects and Gaussian Mean Width
- M-estimator Loss Functions for Regression
- Compressed Sensing

2 Projection onto Complicated Objects and Gaussian Mean Width

We have seen that least squares regression finds the closest point y in a subspace K to a given point b, with the subspace K being simply the column space of A.

$$y' = argmin_{y \in K} \|Sy - Sb\|_2 \tag{1}$$

$$\|y' - b\|_2 \leq (1 \pm \epsilon) \min_{y \in K} \|y - b\| \tag{2}$$

$$\|S(y - y')\|_2 = (1 \pm \epsilon)\|y - y'\|, \forall y, y' \in K \tag{3}$$

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

2.1 Example: Preserving Distances in a Set

What is the answer to the previous question, when K is:

- A set of n arbitrary points in $\mathbb{R}^d$?
- A set of n arbitrary points on a line in $\mathbb{R}^d$?

[**Hint**: Johnson-Lindensrauss Lemma]

2.2 Spherical Mean Width

Let K be a bounded subset in $\mathbb{R}^n$. The **width** in direction u for a unit vector u as:

Definition.

$$u = sup_{p,q \in K} \langle u, p - q \rangle \tag{4}$$

Then the **spherical mean width** of K is just the expectation over (infinitely many) directions u of the above width direction:

$$s(K) = E_u[sup_{p,q \in K} \langle u, p - q \rangle] \tag{5}$$

2.3 Gaussian Mean Width

We now formally define the **gaussian mean width**, $g(K)$, which be thought of as describing the ℓ_2 complexity of the subspace K.

Formally it is defined as:

Definition.

$$g(K) = E_g[sup_{p,q \in K} \langle g, p - q \rangle], \tag{6}$$

where $g \in \mathbb{R}^n$ is a i.i.d Gaussian vector $g \sim \mathcal{N}(0, I_n)$.

Is is straightforward to compute that the ratio of the gaussian mean width over the spherical mean width is:

$$\frac{g(K)}{s(K)} = \Theta(\sqrt{n}) \tag{7}$$

It is also straightforward to compute the guassian mean width for the following cases:

- $K = S^{n-1}$. In this case, $g(K) = \Theta(\sqrt{n})$.
- $K = \{u_1, \dots u_d\}$, where $u_i \in \mathbb{R}^n$, with $\|u_i\|_2 = 1, \forall i \in [d]$. In this case, $g(K) = \Theta(\sqrt{d})$.
- $K = \{u_1, \dots u_t\}$, where $u_i \in \mathbb{R}^n$, with $\|u_i\|_2 = 1, \forall i \in [t]$. In this case, $g(K) = \Theta(\sqrt{\log t})$.

We only present the analysis of the last case, for which it is technically harder to compute $g(K)$.

Indeed, let $u_1, u_2, \dots, u_t \in \mathbb{R}^n$, with $\|u_i\|_2 = 1, \forall i \in [t]$ be t arbitrary unit vectors in $\mathbb{R}^n$. Let also $g \in \mathbb{R}^n$ be an i.i.d. gaussian random vector. Also, let $Z_j, j \in [t]$ be the random variable:

$$Z_j = \langle u_j, g \rangle, \tag{8}$$

which is, by the 2-stability property of the normal distribution, a $\mathcal{N}(0, 1)$ random variable. We need to bound the quantity $E_g[\max_j Z_j]$.

Indeed, for any $\lambda > 0$:

$$\begin{aligned} E[e^{\lambda \max_j Z_j}] &\leq \sum_{j=1}^{t} E[e^{\lambda Z_j}] \\ &= te^{\lambda^2/2}, \end{aligned} \tag{9}$$

where the inequality follows from (?) and the equality uses the fact for a $\mathcal{N}(0,1)$ normal variable w:

$$E[e^{\lambda w}] = e^{\lambda^2/2} \tag{10}$$

Therefore, for any $\lambda > 0$:

$$\begin{aligned} E_g[\max_j Z_j] &\leq \frac{1}{\lambda} \log E[e^{\max_j Z_j}] \\ &\leq \frac{\log t}{\lambda} + \frac{\lambda}{2} \\ &= 2\sqrt{\log t}, \end{aligned} \tag{11}$$

2.4 Sketching Bounds

Theorem 1 (Gordon, 1988 [3]). *Let K be a subset of S^{n-1}. If G is a random Gaussian matrix with $s = \frac{g(K)^2}{\epsilon^2}$ rows, then for all $y, y' \in K$:*

$$\|G(y - y')\|_2^2 = (1 \pm \epsilon)\|y - y'\|_2^2 \tag{12}$$

Theorem 2 (Bourgen, Dirksen, Nelson, 2015 [1]). *S can have $m = \frac{g(K)^2 poly(\log n)}{\epsilon^2}$ rows and $s = \frac{poly(\log n)}{\epsilon^2}$ non-zeros per column if m and s satisfy a condition related to higher moments of $sup_{p,q}\langle g, p - q\rangle$.*

[1] contains similar results for finite and infinite union of subspaces.

3 M-Estimator Loss Functions for Regression

We have seen ways to use linear sketching and obtain fast, approximate randomized algorithms for ℓ_1 and ℓ_2 regression respectively. ℓ_1 regression can be solved efficiently using Linear Programming and is less sensitive to outliers than ℓ_2^2 regression, whereas ℓ_2^2 regression enjoys smoothness properties and has a closed form solution.

In practice, statisticians and data scientists use other fitness measures which try to combine the benefits of both ℓ_1 and ℓ_2^2 regression.

A very common loss used in practice is the so-called **Huber loss**:

$$M_H(x) = \begin{cases} \frac{x^2}{2c}, & |x| \leq c \\ |x| - \frac{c}{2}, & |x| > c \end{cases} \tag{13}$$

Other examples of loss functions include the **L1-L2 loss**:

$$M_{L1-L2}(x) = 2(\sqrt{1 + \frac{x^2}{2}} - 1), \tag{14}$$

the **Fair Estimator loss**:

$$M_F(x) = c^2(\frac{|x|}{c} - \log(1 + \frac{|x|}{c})) \tag{15}$$

and the **Tukey Estimator loss**:

$$M_T(x) = \begin{cases} \frac{c^2}{6}(1 - [1 - (\frac{x}{c})^2]^3), & |x| \leq c \\ \frac{c^2}{6}, & |x| > c \end{cases} \tag{16}$$

3.1 Nice M-Estimators

Definition. An M-estimator is called **nice** if it has at least linear growth and at most quadratic growth. Formally, there exists a constant $C_M > 0.$ such that for all a, a' with $|a| \geq |a'| > 0$, the following holds true:

$$C_M|\frac{a}{a'}| \leq \frac{M(a)}{M(a')} \leq |\frac{a}{a'}|^2 \tag{17}$$

Furthermore, an M-estimator is defined to have the value 0 at 0 ($M(0) = 0$).

Definition. An M-estimator is called **sketchable** if there is a distribution on matrices $S \in \mathbb{R}^{k \times n}$, where k is a slow-growing function of n and for which, with good probability, the following holds true:

$$\|Sx\|_M = \Theta(\|x\|_M) \tag{18}$$

It is relatively straightforward to prove that any **convex** M satisfies the lower bound of equation (17). Indeed:

$$\begin{aligned} M(a') &= M(\frac{a'}{a} \cdot a + (1 - \frac{a'}{a}) \cdot 0) \\ &\leq \frac{a'}{a}M(a) + (1 - \frac{a'}{a})M(0) \\ &= \frac{a'}{a}M(a) \end{aligned} \tag{19}$$

It can also be proved that any sketchable M satisfies the quadratic upper bound of equation (17).

3.2 Nice M-Estimator Theorem

The following theorem, due to Woodruff and Clarkson, guarantees the existence of a fast algorithm for approximate M-regression.

Theorem 3 (Woodruff, Clarkson, 2013)**.** *There exists an algorithm, that uses sketching, runs in $O(nnz(A) + poly(d \log n))$ time and outputs $x' \in \mathbb{R}^d$, such that for any constant $C > 1$, with probability at least 99%, the following holds true:*

$$\|Ax' - b\|_M \leq C \min_x \|Ax - b\|_M \tag{20}$$

It is important to point out the following remarks:

Remark 1. For convex nice M-estimators, we can solve the M regression problem using convex programming, in polynomial time ($poly(nd)$), that is slow in practice.

Remark 2. The sketch of the approximate algorithm is **universal**, in the sense that the same M-sketch works for all nice M-estimators.

3.3 M-Sketch

The universal M-sketch is the following block matrix:

$$T = \begin{bmatrix} S_0 \cdot D_0 \\ S_1 \cdot D_1 \\ S_2 \cdot D_2 \\ \vdots \\ S_{\log n} \cdot D_{\log n} \end{bmatrix} \tag{21}$$

The matrices $S_i, i \in [\log n]$ are independent CountSketch matrices with $poly(d)$ rows and n columns, where as the matrices $D_i, i \in [\log n]$ are diagonal and perform uniform sampling of the n rows, scaling by a factor of $\frac{1}{(\log n)^i}$.

The crucial property that we want T to satisfy is, for any $y = Ax - b$:

$$\|T(Ax - b)\|_{w,M} \approx \|Ax - b\|_M \tag{22}$$

It can be seen that both large coordinates and small coordinates can be efficiently sampled using T As an example, consider the vector $y = (n, 1, 1, \ldots 1)$. For further details, check [2].

4 Compressed Sensing

In the compressed sensing problem, we are trying to estimate a vector $x \in \mathbb{R}^n$ by having access (due to for example physical constraints) to random linear measurements of x instead.

In our context, we choose a random sketching matrix $S \in \mathbb{R}^{r \times n}$ and observe $S \cdot x$. We want to output a vector $x' \in \mathbb{R}^n$ such that:

$$\|x - x'\|_p = D \cdot \min_{k-sparse z} \|x - z\|_q, \tag{23}$$

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

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

- **Randomized** (“for-each") schemes
- **Deterministic** (“for-all") schemes

Let x_k denote the best k-sparse approximation to x, i.e. the vector containing the largest k coordinates in magnitude.

The famous CountSketch matrix provides with a randomized scheme, achieving the ℓ_2/ℓ_2 guarantee with high probability:

$$\|x - x'\|_2 = O(1) \cdot \|x - x_k\|_2 \tag{24}$$

4.1 CountSketch for Compressed Sensing

Indeed, multiplying by a CountSketch matrix S with $O(k \log n)$ rows can be thought of as $O(\log n)$ repetitions of hashing into $O(k)$ buckets. S is a random linear map. As we have already seen in the class, S estimates every coordinate, $x_i, i \in [n]$ of a $x \in \mathbb{R}^n$ vector up to an additive error of $\frac{\|x - x_K\|_2}{\sqrt{k}}$.

If we output the $2k$-sparse vector $x' \in \mathbb{R}^n$ that consists of the top $2k$ (with respect to magnitude) estimates given by CountSketch, we can prove that equation (24) is satisfied with very high probability. Before giving the proof of this claim, we define two useful notions.

Definition. A coordinate i is **heavy** if:

$$|x_i| \geq \frac{\|x - x_k\|_2}{\sqrt{k}} \tag{25}$$

It is easy to see that there can be at most $2k$ heavy coordinates.

Definition. A coordinate i is **super-heavy** if:

$$|x_i| \geq 3 \cdot \frac{\|x - x_K\|_2}{\sqrt{k}} \tag{26}$$

One can now observe that the set T of super-heavy coordinates is in the support of the vector x'. Therefore:

$$\begin{aligned}\|x-x'\|_2 &\le \|(x-x')_T\|_2 + \|(x-x')_{[n]\setminus T}\|_2 \\ &\le \sqrt{2}k \cdot \frac{\|x-x_K\|_2}{\sqrt{k}} + \|(x-x')_{[n]\setminus T}\|_2 \\ &\le \sqrt{2}\|x-x_k\|_2 + \|(x-x_k)_{[n]\setminus T}\|_2 + \|(x_k-x')_{[n]\setminus T}\|_2 \\ &= O(\|x-x_k\|_2), \end{aligned} \tag{27}$$

as desired.

4.2 No Deterministic Scheme for ℓ_2/ℓ_2 distortion

Let us consider $k=1$ and suppose, for the sake of contradiction, that S is a deterministic sketching matrix with $r=o(n)$ rows. It suffices to show that there exists a vector $x \in ker(S)$ which, for any constant $C>0$ satisfies:

$$\|x\|_\infty \ge C\|x-x_1\|_2 \tag{28}$$

Without loss of generality, assume that S has orthonormal rows. Since $\sum_i \|Se_i\|_2^2 = r$, therefore there exists a coordinate j with $\|Se_j\|_2^2 \le \frac{r}{n}$. Let x be:

$$x = e_j - S^T S e_j, \tag{29}$$

from which it is clear that $x \in ker(S)$. Then:

$$\begin{aligned}\|x\|_\infty^2 &\ge |x_j|^2 \\ &= (e^e_{jj} - e_j^T S^T S e_j)^2 \\ &\ge (1-\frac{r}{n})^2, \end{aligned} \tag{30}$$

while at the same time:

$$\begin{aligned}\|x-x_1\|_2 &\le \|x-e_j\|_2 \\ &= \|S^T S e_j\|_2 \\ &= \|Se_j\|_2 \\ &\le \sqrt{\frac{r}{n}} \\ &= o(1) \end{aligned} \tag{31}$$

4.3 Deterministic Schemes for ℓ_2/ℓ_1 distortion

Definition. Matrix S has the (ϵ, k)**-restricted isometry property (RIP)**, if for all k-sparse vectors $x \in \mathbb{R}^n$, the following holds:

$$(1-\epsilon)\|x\|_2^2 \leq \|Sx\|_2^2 \leq (1+\epsilon)\|x\|_2^2 \tag{32}$$

It can be shown, that if S has the (ϵ, k)-RIP property, then one can efficiently output an $x' \in \mathbb{R}^n$ for which:

$$\|x - x'\|_2 = O(\frac{1}{\sqrt{k}})\|x - x_k\|_1, \tag{33}$$

by solving the following Linear Program:

$$\mathcal{RIP} - \mathcal{LP} : \begin{cases} \min\limits_{z \in \mathbb{R}^n} & \|z\|_1 \\ \text{s.t} & Sz = Sx \end{cases} \tag{34}$$

The proof that $x' \in \mathbb{R}^n$ satisfies equation (33) uses the (ϵ, k)-RIP property and elementary norm manipulations.

There are deterministic, but not explicit matrices S with $O(k \log(\frac{n}{k}))$ rows that have the (ϵ, k)-RIP property for constant ϵ.

A major **open question** in the area remains if there exists an explicit matrix with (ϵ, k)-RIP property that has only $o(k^2)$ rows. Bourgain et al [4] can get $k^{2-\gamma}$ rows for a constant $\gamma > 0$ and $k \approx \sqrt{n}$.

References

[1] Jean Bourgain, Sjoerd Dirksen, Jenali Nelson. Toward a unified theory of sparse dimensionality reduction in Euclidean space. arxiv preprint arXiv:1311.2542, 2014.

[2] Kenneth L. Clarkson, David P. Woodruff. Input Sparsity for Robust Subspace Approximation. arxiv preprint arxiv:1510.06073

[3] Gordon, Y.: On Milmanâs inequality and random subspaces which escape through a mesh in R^n. Geometric aspects of functional analysis (1986/87), Lecture Notes in Math., vol. 1317, pp. 84â106. Springer, Berlin (1988). DOI 10.1007/BFb0081737. URL http://dx.doi.org/10.1007/BFb0081737

[4] J.Bourgain, S.Dilworth, K.Ford, S.Konyagin and D.Kutzarova, Explicit construction of RIP matrices and related problems, Duke Mathematical Journal, Vol.159 (2011), no.1, pages 145-185.