COMS E6998-9: Algorithms for Massive Data (Spring'19) Feb 14, 2019

Lecture 8: Fast dimension reduction, ℓ_1 regression

Instructor: *Alex Andoni* Scribes: *Darshan Thaker, Yuchen Mo*

1 Review

Recall from last class the Fast Johnson-Lindenstrauss Lemma, which tells us that there exists a linear map $\phi : \mathbb{R}^n \to \mathbb{R}^k$, $\phi(x) = Sx$ where $S = P \cdot H \cdot D$. Here, P is a sparse projection matrix, H is a Hadamard matrix, and D is a diagonal matrix with entries as random ± 1's. The total time complexity for computing S is as follows: $O(n)$ for (Dx), $O(n \log n)$ for $(H \cdot (Dx))$, $O(k)$ for $(P \cdot (HDx))$. In total, this means the time for dimensionality reduction is $O(n \log n + k)$. Recall the following lemma.

Lemma 1. *For vector* $y \in \mathbb{R}^n$, *if* $k = \Omega\left(\frac{\log \frac{1}{\delta}}{\epsilon^2} \cdot \frac{n \cdot ||y||_\infty^2}{||y||_2^2}\right)$, *then with probability* $\geq 1-\delta$, $||Py||_2 \in (1 \pm \epsilon)||y||_2$

2 Fast JL (cont.)

We want to show that if $y = H \cdot D \cdot x$, then $||y||_\infty$ cannot be too large. More formally:

Claim 2. *Let* $y = H \cdot D \cdot x$, $\Pr\left[||y||_\infty \leq c \cdot \sqrt{\frac{\log(n/\delta)}{n}} \cdot ||x||_2\right] \geq 1 - \delta$

We just need to prove that $\forall y_i, |y_i| \leq c \cdot \sqrt{\frac{\log(n/\delta)}{n}} \cdot ||x||_2$ with high probability.

Proof. Here A_j^i means ith row and jth column. We have that $r_j \in \left\{\frac{1}{\sqrt{n}}, -\frac{1}{\sqrt{n}}\right\}$

$$\begin{aligned} y_i = \langle H^i, D_i x \rangle &= \sum_j r_j x_j \\ &= \frac{1}{\sqrt{n}} \cdot (\pm x_1 \pm x_2 \pm \cdots \pm x_n) \end{aligned}$$

We will use the following theorem now.

Theorem 3. *Let* $a_1, \cdots, a_n \in \mathbb{R}$, $r_1, \cdots, r_n \in \{\pm 1\}$, $S_n = |\sum_j r_j a_j|$, *then* $\Pr\left[S_n \geq t \cdot \sqrt{\sum_j a_j^2}\right] \leq e^{-t^2/2}$. *(Proof: [1] equation 1.2)*

Thus, we have

$$\Pr\left[|y_i| \geq \sqrt{\frac{2\log(n/\delta)}{n}} \cdot ||x||_2\right] \leq \frac{\delta}{n}$$

By taking union bound:

$$\Pr\left[\forall i \in [n], |y_i| \leq \sqrt{\frac{2\log(n/\delta)}{n}} \cdot ||x||_2\right] \geq 1 - \delta$$

□

Subspace embedding

Recall the setup from last lecture: we are given $\{y \in \mathbb{R}^n | y = Ux\}$, $x \in \mathbb{R}^{n \times d}$, and we have that $d < n$. If $S \in \mathbb{R}^{k \times d}$, and for all $x \in \mathbb{R}^d$ and $y = Ux$, $||Sy||_2 \in (1 \pm \epsilon)||y||_2$, then S is a subspace embedding of U. From last time, we know the following claim:

Claim 4. *If S is defined as in the JL Lemma, i.e., $S_{i,j} \sim N\left(0, \frac{1}{k}\right)$, $k = \Omega\left(d/\epsilon^2\right)$, then S is a subspace embedding of U.*

The following claim also holds, but we do not prove it.

Claim 5. *If S is defined as in fast JL, i.e., $S = PHD$, $k = \Omega\left(\frac{d+\log n}{\epsilon^2}\log(d)\right)$, then S is a subspace embedding of U.*

3 Least absolute deviation regression

We first introduce the least absolute deviation regression problem, also known as the ℓ_1 regression problem.

Definition 6. *The least absolute deviation regression is as follows: Given $A \in \mathbb{R}^{n \times d}$, $b \in \mathbb{R}^d$, we wish to find $\min_{x \in \mathbb{R}^d} ||Ax - b||_1$. Here $||y||_1 = \sum_i |y_i|$ for a vector y (i.e. the ℓ_1 norm).*

This can be reformulated into a linear program, where our objective is to minimize $\sum_i^n t_i$, subject to the following linear constraints: $\forall i \in [n], -t_i \leq \langle A^i, x\rangle \leq t_i$ where $t_i \geq 0,\ x_i \geq 0$.

Observation 7. *There are $n+d$ variables and $2n$ constraints. Thus, using linear programming techniques to solve it directly takes $\mathrm{Poly}(n \cdot d)$ time, which is too slow for large n and d.*

Suppose there is a linear mapping $\phi : \mathbb{R}^n \to \mathbb{R}^k$, i.e., $\phi(y) = S \cdot y,\ S \in \mathbb{R}^{k \times n}$, such that $\forall y \in \mathrm{span}(U)$, $||\phi(y)||_1 \in (1 \pm \epsilon) \cdot ||y||_1$, where $U \in \mathbb{R}^{n \times (d+1)}$ is a basis of the subspace spanned by $(A_1, \ldots, A_d, b)$.

Then to solve the original ℓ_1 regression,

1. Solve $\min_x ||\phi\left(Ax - b\right)||_1$. Let x' be the solution to this reduced ℓ_1 case.

2. Show x' is also a solution to the original problem.

Proof. We'll show that x' is an $1 + 2\epsilon$ approximation to the original problem. Let x^* be the optimal solution of the original $\min_x ||Ax - b||_1$ problem.

$$\begin{aligned} ||Ax' - b||_1 &\le \frac{1}{1-\epsilon} \cdot ||SAx' - Sb||_1 && (1) \\ &\le \frac{1}{1-\epsilon} \cdot ||SAx^* - Sb||_1 && (2) \\ &\le \frac{1+\epsilon}{1-\epsilon} \cdot ||Ax^* - b||_1 && (3) \\ &\le (1+2\epsilon)||Ax^* - b||_1 && (4) \end{aligned}$$

Details:

1: $\forall y \in \text{span}(U),\ ||Sy||_1 \ge (1-\epsilon) \cdot ||y||_1$

2: x' minimizes $||SAx - Sb||_1$

3: let $y = Ax^* - b$, $||Sy||_1 \le (1+\epsilon)||y||_1$

4: $\varepsilon \in (0, 1/2)$ □

Thus the total running time will be reduced to $\mathrm{O}(DR) + \mathrm{Poly}(kd)$, where DR is the time to perform dimensionality reduction. We now see how efficient performing this dimensionality reduction actually is.

3.1 Sampling-based Method

Our goal then becomes the following: given $U \in \mathbb{R}^{n \times d}$, a basis of the subspace spanned by the columns of A and b, find the L1 subspace embedding. Note that here, d includes the columns of A as well as b, but is simply renamed to d for convenience. Recall that a L1 subspace embedding is a matrix S such that for all $x \in \mathbb{R}^d$, $\|SUx\|_1 \in (1 \pm \epsilon)\|Ux\|_1$. Let $y = Ux$.

We introduce a sampling-based method to find S. Specifically, let S be a diagonal matrix, where each entry on the diagonal is:

$$S_{ii} = \begin{cases} \frac{1}{p_i} & \text{with probability } p_i \\ 0 & else \end{cases} \tag{5}$$

We calculate the expected value of $\|Sy\|_1$, which is $\mathbb{E}[\|Sy\|_1] = \sum_{i=1}^{n} p_i \cdot \frac{1}{p_i} \cdot |y_i| + 0 = \|y\|_1$. This looks promising, however, there are still certain edge cases that can reduce the effectiveness of $\|Sy\|_1$ as an approximation to $\|y\|_1$. Specifically, consider the case when y is very sparse, say with only one non-zero entry. To accurately estimate the norm, we must find the non-zero entry and with high probability, the norm of Sy will be zero. However, intuitively, we should sample each coordinate proportional to the value of y_i. We don't exactly know each value of y_i, but the subspace constrains y_i in a certain way, so we can pick p_i in a more careful way.

Definition 8. *For all $x \in \mathbb{R}^d$, $\|x\|_2 \leq \|Ux\|_1 \leq \kappa \cdot \|x\|_2$, then the condition number of U is κ.*

Define $p_i = \min\left(1, c \cdot \left[\log\left(\frac{1}{\delta}\right)/\epsilon^2\right] \cdot \|U^i\|_2\right)$, where U^i represents the ith row of U. We want to show that with probability at least $1-\delta$, $\|Sy\|_1 \in (1 \pm \epsilon)\|y\|_1$. Before proceeding with this proof, we state a generalization of the Chernoff bound, known as *Bernstein's inequality.*

Theorem 9 (Bernstein's inequality)**.** *Suppose $X_1, \ldots, X_n$ are n independent random variables (not necessarily identically distributed). For all $i \in [n]$, we have that $|X_i| \leq M$, then for any $t > 0$,*

$$\Pr\left[\left|\sum_{i=1}^n X_i - \mathbb{E}\left[\sum_{i=1}^n X_i\right]\right| > t\right] \leq 2 \cdot \exp\left\{-\frac{0.5t^2}{\sum_{i=1}^n \mathrm{Var}[X_i] + 1/3 \cdot Mt}\right\}$$

This allows us to prove the following claim.

Claim 10. *If S is sampled as described in Equation 5 with $p_i = \min\left(1, c \cdot \left[\log\left(\frac{1}{\delta}\right)/\epsilon^2\right] \cdot \|U^i\|_2\right)$ for sufficiently large constant c, then with probability at least $1-\delta$, $\|Sy\|_1 \in (1 \pm \epsilon)\|y\|_1$*

Proof. Observe that by definition of L1 norm, we know that $\|Sy\|_1 = \sum_{i=1}^n |S_{ii}y_i|$. Define $X_i = |S_{ii}y_i|$, so $\|Sy\|_1 = \sum_{i=1}^n X_i$. Recall that $\mathbb{E}\left[\sum_{i=1}^n X_i\right] = \|y\|_1$. Without loss of generality, assume that $\|y\|_1 = 1$ - we can simply rescale y if not.

To apply Bernstein's inequality, we first try bound to the sum of variances of X_i as follows:

$$\begin{aligned}
\sum_{i=1}^n \mathrm{Var}[X_i] &\leq \sum_{i=1}^n \mathbb{E}[X_i^2] \\
&= \sum_{i=1}^n p_i \cdot \left(\frac{y_i}{p_i}\right)^2 \\
&= \sum_{i=1}^n \left|\frac{y_i}{p_i}\right| \cdot |y_i| \\
&\leq \max_i \left|\frac{y_i}{p_i}\right| \cdot \|y\|_1 \\
&= \max_i \left|\frac{y_i}{p_i}\right| \\
&\leq \frac{\epsilon^2}{c\log\left(\frac{1}{\delta}\right)} \cdot \frac{|\langle U^i, x\rangle|}{\|U^i\|_2} && \text{(Expand out } p_i\text{)} \\
&\leq \frac{\epsilon^2}{c\log\left(\frac{1}{\delta}\right)} \cdot \frac{\|U^i\|_2\|x\|_2}{\|U^i\|_2} && \text{(Cauchy-Schwarz)} \\
&\leq \frac{\epsilon^2}{c\log\left(\frac{1}{\delta}\right)} \cdot \|x\|_2
\end{aligned}$$

$$\leq \frac{\epsilon^2}{c\log\left(\frac{1}{\delta}\right)} \cdot \|y\|_1 = \frac{\epsilon^2}{c\log\left(\frac{1}{\delta}\right)} \qquad (\|x\|_2 \leq \|Ux\|_1 \leq \kappa\|x\|_2)$$

Next, we find a bound on each $|X_i|$, which is less than $\left|\frac{y_i}{p_i}\right|$. Applying the same inequalities as above, we conclude that $|X_i| \leq \frac{\epsilon^2}{c\log\left(\frac{1}{\delta}\right)}$. Now, can apply Bernstein's inequality with $t = \epsilon$, which gives

$$\Pr\left[\left|\sum_{i=1}^{n} X_i - \mathbb{E}\left[\sum_{i=1}^{n} X_i\right]\right| > \epsilon\right] \leq 2 \cdot \exp\left\{-\frac{0.5\epsilon^2}{\frac{\epsilon^2}{c\log(1/\delta)} + \frac{\epsilon^3}{3c\log(1/\delta)}}\right\}$$

For constant c large enough, we can make this probability at most δ. Thus, with probability at least $1-\delta$, $\|Sy\|_2 \in (1 \pm \epsilon)\|y\|_1$. $\square$

To prove that we can obtain a subspace embedding across the entire subspace, we can discretize the L1 norm ball and show the following result: For any fixed vector $y = Ux$ such that $\|Sy\|_1 \in (1 \pm \epsilon)\|y\|_1$ with probability at least $1 - \left(\frac{10n}{\epsilon}\right)^{-d}$, then we can show that $\Pr_S[\forall y = Ux, \|Sy\|_1 \in (1 \pm \epsilon)\|y\|_1] \geq 0.9$. We do not prove this.

Finally, we claim that the expected number of non-zero elements in S is not too large. Let $\delta = \left(\frac{10n}{\epsilon}\right)^{-d}$, and let κ be the condition number of U.

$$\begin{aligned}
\mathbb{E}[\#\text{ of non-zero elements in } S] &= \sum_{i=1}^{n} p_i \\
&\leq \sum_{i=1}^{n} \frac{\log(1/\delta)}{\epsilon^2} \cdot \|U^i\|_2 \\
&\leq O\left(\frac{d}{\epsilon^2} \cdot \log\left(\frac{n}{\epsilon}\right)\right) \cdot \sum_{i=1}^{n}\sum_{i=1}^{n} \|U^i\|_2 \\
&\leq O\left(\frac{d}{\epsilon^2} \cdot \log\left(\frac{n}{\epsilon}\right)\right) \cdot \sum_{i,j} |U_{ij}| \\
&\leq O\left(\frac{d}{\epsilon^2} \cdot \log\left(\frac{n}{\epsilon}\right)\right) \cdot d \cdot \kappa \\
&\leq O\left(\frac{d^2}{\epsilon^2} \cdot \log\left(\frac{n}{\epsilon}\right)\right) \cdot \kappa
\end{aligned}$$

The second to last step above follows from the fact that $\|x\|_2 \leq \|Ux\|_1 \leq \kappa\|x\|_2$ for all x. Thus, we can choose x to be the vector of all-zeros and a single 1, and the statement follows.

We state the following fact without the proof. For any d dimensional subspace, there is always a basis U whose condition number κ is at most $\text{poly}(d)$.

References

[1] Pinelis, Iosif. "An asymptotically Gaussian bound on the Rademacher tails." Electronic Journal of Probability 17 (2012).