Special Topics in Theoretical Computer Science Spring 2019

Lecture Notes — February 28, 2019

Prof. Huy Nguyen *Scribe: Peter Bernstein*

1 Sampling approach to compute $A^T B$

Last time we wanted to compute $A^T B$ where $A \in \mathbb{R}^{n\times d}, B \in \mathbb{R}^{n\times p}$. The idea we used was instead of computing this product exactly, we would write

$$A^T B = \sum_{i=1}^{n} a_i b_i^T$$

where a_i is the ith row of A and b_i is the ith row of B. We then said that we could sample the ith term with probability p_i. If the ith term is in fact picked, we add the term $\frac{1}{p_i} a_i b_i^T$ to a sum. That's how we got the estimate

$$C = \sum_{i=1}^{n} \frac{x_i}{p_i} a_i b_i^T$$

where x_i is an indicator of whether i is picked or not.

Last time we also showed that

$$\mathbb{E}[C] = A^T B \text{ and } \mathbb{E}[\|C - A^T B\|_F^2] = \sum_i \left(\frac{1}{p_i} - 1\right) \|a_i\|^2 \|b_i\|$$

and that the optimal choice of p is $p \sim \|a\|\|b\|$, i.e. p should be proportional to the product of the norms of a and b. This implies that

$$\mathbb{E}[\|C - A^T B\|_F^2] \leq \left(\sum_i \|a_i\|\|b_i\|\right)^2.$$

To improve our estimate from just using C, we can pick m samples and compute

$$\hat{C} = \frac{1}{m}\left(C_1 + C_2 + ... + C_m\right)$$

and note that $\left(\sum_i \|a_i\|\|b_i\|\right)^2 \leq \left(\sum_i \|a_i\|^2\right)\left(\sum_i \|b_i\|^2\right) = \|A\|_F^2 \|B\|_F^2$.

If we want

$$\|\hat{C} - A^T B\|_F^2] \leq \epsilon \|A\|_F^2 \|B\|_F^2 \text{ with probability } \frac{9}{10} \tag{1}$$

then we need $m = \Theta\left(\frac{1}{\epsilon^2}\right)$ samples.

Ideally we want a probability of $1-\delta$ instead of just $\frac{9}{10}$. In that case we could try our usual method of repeating the experiment $O(\log(1/\delta))$ times to get $\hat{C}_1, ..., \hat{C}_t$. Then if the $\hat{C}_i$ were numbers, we could take the median. But what should the analogous operation be on matrices?

The idea is to find a point close to a lot of other points and we will have with high probability that it is no more than 4ϵ from the optimum. By the Chernoff inequality, at least $\frac{9}{10}$ of the $\hat{C}_i$s satisfy (1). Now suppose $\hat{C_i}^*$ is within distance 2ϵ from $\frac{9}{10}$ of the others. This would imply that $\hat{C_i}^*$ is within 2ϵ of some "good" point, which implies $\hat{C_i}^*$ is within 4ϵ from $A^T B$.

The time for this procedure is $O(nd + np) + O(\frac{1}{\epsilon^2}dp + \log^2 \frac{1}{\delta}$. It is an open question whether we can speed up the second term in this expression. The term $\log^2 \frac{1}{\delta}$ seems high when typically we see a dependency on accuracy is $\log \frac{1}{\delta}$. The open question at a high level is this: can you find a "median" in $O(\log \frac{1}{\delta})$.

2 Sketching Approach

Definition 1 (JL moment property)**.** *Let D be a distribution over matrices $\Pi \in \mathbb{R}^{m \times n}$. D satisfies the $(\epsilon, \delta, p)-$JL moment property if for any x of unit norm ($\|x\|_2 = 1$),*

$$\mathop{\mathbb{E}}_{\Pi \sim D} \left[|\|\Pi x\|^2 - 1|^p \right] \leq \epsilon^p \delta.$$

Note that many things satisfy this property. For example, a matrix with iid Gaussian entries satisfies the $(\epsilon, \delta, \log \frac{1}{\delta})-$JL moment property with $n = \Theta\left(\frac{1}{\epsilon^2} \log \frac{1}{\delta}\right)$.

How should we use this to estimate $A^T B$? The idea is to operate with matrices $(A^T \Pi^T)(\Pi B)$. But first, let's introduce a lemma:

Lemma 2. *If D satisfies the $(\epsilon, \delta, p) - JL$ moment property then for any vectors x, y with unit lengths,*

$$\mathop{\mathbb{E}}_{\Pi \sim D} [|\langle \Pi x, \Pi y \rangle - \langle x, y \rangle|^p] \leq (3\epsilon)^p \delta.$$

Proof. We first calculate $\langle x, y \rangle$, $\langle \Pi x, \Pi y \rangle$, and then bound $|\langle \Pi x, \Pi y \rangle - \langle x, y \rangle|$ using the triangle inequality:

$$\begin{aligned} \langle x, y \rangle &= \frac{1}{4} \left(\|x + y\|^2 - \|x - y\|^2 \right) \\ \langle \Pi x, \Pi y \rangle &= \frac{1}{4} \left(\|\Pi(x + y)\|^2 - \|\Pi(x - y)\|^2 \right) \\ |\langle \Pi x, \Pi y \rangle - \langle x, y \rangle| &\leq \frac{1}{4} \left(\|\Pi(x + y)\|^2 - \|x + y\|^2 - \|\Pi(x - y)\|^2 + \|x - y\|^2 \right) \end{aligned}$$

So

$$\left(\mathop{\mathbb{E}}_{\Pi} |\langle \Pi x, \Pi y \rangle - \langle x, y \rangle|^p \right)^{\frac{1}{p}} \leq \left(\mathop{\mathbb{E}}_{\Pi} \left[\left| \left\| \Pi \left(\frac{x + y}{2} \right) \right\|^2 - \left\| \frac{x + y^2}{2} \right\| \right|^p \right] \right)^{\frac{1}{p}} + \left(\mathop{\mathbb{E}}_{\Pi} \left[\left| \left\| \Pi \left(\frac{x - y}{2} \right) \right\|^2 - \left\| \frac{x - y}{2} \right\|^2 \right|^p \right] \right)^{\frac{1}{p}}$$

which by the JL-moment property is bounded:

$$\left(\mathop{\mathbb{E}}_{\Pi} |\langle \Pi x, \Pi y \rangle - \langle x, y \rangle|^p \right)^{\frac{1}{p}} \leq \epsilon \delta^{\frac{1}{p}} + \epsilon \delta^{\frac{1}{p}} = 2\epsilon \delta^{\frac{1}{p}}$$

□

Theorem 3. *Suppose D satisfies the (ϵ, δ, p)-JL moment property. For any $A \in \mathbb{R}^{n\times a}, B \in \mathbb{R}^{n\times b}$,*

$$\mathop{\mathbb{P}}_{\Pi}\left(\left|\|A^T B - (\Pi A)^T(\Pi B)\|_F\right| > 3\epsilon\|A\|_F\|B\|_F\right) \leq \delta$$

and gives time $ab\frac{1}{\epsilon^2}\log\frac{1}{\delta}$ and to compute ΠA is $na\frac{1}{\epsilon^2}\log\frac{1}{\delta}$. Note the single log factor here.

We ran out of time just after starting this proof. We will continue with this next class.