EE 381V: Large Scale Learning **Spring 2013**

Lecture 19 — March 26

Lecturer: Caramanis & Sanghavi *Scribe: Hongbo Si*

19.1 Reviews of Last Lecture

From the first lecture of this course, we focused on theory and algorithms with respect to large scale learning problems. In this process, basic issues in the field of linear algebra emerged, for instance matrix multiplication and factorization. Recall that in earlier lectures, we have discussed *Direct Methods*, which were exact but with high cost. To this end, *Iterative Methods* were introduced, where the solution could be exact after large number of iterations. These type of algorithms usually cost much less in each iteration and exploited special structure, as well as terminated early with a prior bound. Lanczos and Power methods are typical examples with all the properties.

From last lecture (Lecture 17) on, we began to discuss *Randomized Methods*. In particular, we concerned a factorization approximation problem. Given a matrix $A \in \mathbb{R}^{m\times n}$, assume its SVD is given by $A = U\Sigma V^*$, and its best rank-k approximation (with respect to $\|\cdot\|_2$ or $\|\cdot\|_F$) is $U_k U_k^* A \triangleq A_k$, where U_k is the first k columns of U. We want to approximately compute A_k by some other matrix X, such that the error is bounded by

$$\|A - X\|_2^2 \leq \|A - A_k\|_2^2 + \epsilon\|A\|_F^2,$$

or

$$\|A - X\|_F^2 \leq \|A - A_k\|_F^2 + \epsilon'\|A\|_F^2,$$

with running time linear in n and m. Note that the first term at the right hand side is the best error we can make, which cannot be evaded by any algorithm, and the second term is additional error, which is expected to be small in our algorithm.

Several most common randomized approximation schemes have been introduced in last lecture:

1. Dimension Reduction.
2. Column Selection.
3. Sparsification.
4. Approximation by sub-matrices.

Last time, we have seen the theoretical analysis of *Dimension Reduction* method, which is basically induced by Johnson-Lindenstrauss lemma. Today, we move on to discuss *Column Selection* and *Sparsification* methods.

19.2 Column Selection Method

In *Column Selection Method*, a set of columns are selected randomly with designed probability. More precisely, consider an aiming matrix $A \in \mathbb{R}^{m\times n}$. Design a set of i.i.d. random vectors $\{Y_i\}$, where $1 \le i \le l$ and $l \ge k$, such that

$$Y_i = \frac{1}{\sqrt{lp_j}} A_j \quad \text{with probability } p_j = \frac{\|A_j\|_2^2}{\|A\|_F^2},\ 1 \le j \le n, \tag{19.1}$$

where A_j is the jth column of A. Perform SVD to the constructed $Y \in \mathbb{R}^{m\times l}$, where the ith column of Y is Y_i, i.e. $Y = Q\Sigma_Y V_Y^*$. Output $X = Q_k Q_k^* A$, where Q_k is consisted of top k columns of Q.

Note that the choice of p_j as (19.1) is crucial. The key idea of designing it is revealed as follows. First, p_j behaves as a normalization parameter such that every column of Y has the same l_2-norm. In particular, we have for every $1 \le i \le l$,

$$\|Y_i\|_2^2 = \frac{1}{lp_j}\|A_j\|_2^2 = \frac{1}{l}\|A\|_F^2. \tag{19.2}$$

Secondly, which is more important, the choice of p_j guarantees that the expectation of YY^* equals AA^*, which is the desired expectation required by the algorithm. Mathematically, we have

$$\begin{aligned}
\mathbb{E}[YY^*] &= \mathbb{E}[\sum_{i=1}^{l} Y_i Y_i^*] \\
&= l\mathbb{E}[Y_1 Y_1^*] \\
&= l\sum_{j=1}^{n} p_j \left(\frac{1}{lp_j} A_j A_j^*\right) \\
&= \sum_{j=1}^{n} A_j A_j^* \\
&= AA^*.
\end{aligned} \tag{19.3}$$

To this end, YY^* is a summation of i.i.d. matrices, with desired expectation, which inspires us to use concentration bound to control the fluctuation. More precisely, we aim to show $\|AA^* - YY^*\|_2$ is small with high probability, which is feasible by concentration result. The reason that we need this result is revealed by the following lemma.

Lemma 19.1.

$$\|A - Q_k Q_k^* A\|_2^2 \le \|A - A_k\|_2^2 + 2\|AA^* - YY^*\|_2. \tag{19.4}$$

Note that the output of our algorithm is $X = Q_k Q_k^* A$, so a good bound on $\|AA^* - YY^*\|_2$ guarantees the performance of the algorithm. Before proving this lemma, we introduce a general fact about singular value first. For this, we give the definition of Lipschitz continuous.

Given two norm spaces $(\mathfrak{X}, \|\cdot\|_{\mathfrak{X}})$ and $(\mathfrak{Y}, \|\cdot\|_{\mathfrak{Y}})$, a mapping $f : \mathfrak{X} \to \mathfrak{Y}$ is called Lipschitz continuous if there exists a real constant L such that, for all x_1 and x_2 in $\mathfrak{X}$,

$$\|f(x_1) - f(x_2)\|_{\mathfrak{Y}} \leq L\|x_1 - x_2\|_{\mathfrak{X}}.$$

Any such L is referred to as a Lipschitz constant for f.

Lemma 19.2. *Singular value (as a function of matrix) is Lipschitz continuous with Lipschitz constant 1 with respect to operator norm, i.e. for any matrices $Z_1, Z_2 \in \mathbb{R}^{m\times n}$,*

$$|\sigma_k(Z_1) - \sigma_k(Z_2)| \leq \|Z_1 - Z_2\|_2, \tag{19.5}$$

where $\sigma_k(\cdot)$ denotes the kth singular value, and $k \leq \min(m, n)$.

Proof: (to Lemma 19.2) The proof bases on Courant-Fischer variational formulas for singular values (similar to Problem 6 in Homework 2), which says

$$\sigma_k(Z) = \max_{V\in\mathcal{V}_k} \min_{\substack{x\in V \\ \|x\|_2=1}} \|Zx\|_2, \tag{19.6}$$

where $\mathcal{V}_k$ is the set of all subspaces in $\mathbb{R}^n$ of dimension k. Assume $\sigma_k(Z_1) \geq \sigma_k(Z_2)$ without loss of generality. Define

$$V^* = \arg\max_{V\in\mathcal{V}_k} \min_{\substack{x\in V \\ \|x\|_2=1}} \|Z_1x\|_2,$$

and

$$x^* = \arg \min_{\substack{x\in V^* \\ \|x\|_2=1}} \|Z_2x\|_2,$$

then we have

$$\begin{aligned}
|\sigma_k(Z_1) - \sigma_k(Z_2)| &= \max_{V\in\mathcal{V}_k} \min_{\substack{x\in V \\ \|x\|_2=1}} \|Z_1x\|_2 - \max_{V\in\mathcal{V}_k} \min_{\substack{x\in V \\ \|x\|_2=1}} \|Z_2x\|_2 \\
&\leq \min_{\substack{x\in V^* \\ \|x\|_2=1}} \|Z_1x\|_2 - \min_{\substack{x\in V^* \\ \|x\|_2=1}} \|Z_2x\|_2 \\
&\leq \|Z_1x^*\|_2 - \|Z_2x^*\|_2 \\
&\leq \|(Z_1 - Z_2)x^*\|_2 \\
&\leq \|Z_1 - Z_2\|_2.
\end{aligned}$$

□

With this result, we are ready to prove Lemma 19.1.

Proof: (to Lemma 19.1) First, we want to show

$$\|A - Q_k Q_k^* A\|_2 = \sup_{\substack{Q_k^* z = 0 \\ \|z\|_2 = 1}} \|A^* z\|_2. \tag{19.7}$$

Assume $x = x_Q + x_{Q^\perp}$, where x_Q is in the range of Q_k, i.e. $x_Q = Q_k y$ for some y, and $x_{Q^\perp}$ is in the perpendicular subspace, i.e. $Q_k^* x_{Q^\perp} = 0$. Hence,

$$(I - Q_k Q_k^*) x_Q = Q_k y - Q_k y = 0. \tag{19.8}$$

Using (19.8), and the fact that for any matrix Z, $\|Z\|_2 = \|Z^*\|_2$, we have

$$\begin{aligned}
\|(I - Q_k Q_k^*) A\|_2 &= \|A^*(I - Q_k Q_k^*)\|_2 \\
&= \sup_{\|x\|_2 = 1} \|A^*(I - Q_k Q_k^*) x\|_2 \\
&= \sup_{\|x\|_2 = 1} \|A^*(I - Q_k Q_k^*) x_{Q^\perp}\|_2 \\
&= \sup_{\substack{Q_k^* z = 0 \\ \|z\|_2 = 1}} \|A^* z\|_2,
\end{aligned}$$

where the last step holds by letting $z = x_{Q^\perp}$.

Then, it is straightforward to see

$$\|A^* z\|_2^2 = \langle z, AA^* z \rangle = \langle z, (AA^* - YY^*) z \rangle + \langle z, YY^* z \rangle, \tag{19.9}$$

and from $Q_k z = 0$, where Q_k is the top k columns of Y's SVD, we have

$$\langle z, YY^* z \rangle \le \sigma_{k+1}(Y)^2 = \sigma_{k+1}(YY^*), \tag{19.10}$$

Using Lemma 19.2 by choosing $Z_1 = AA^*$ and $Z_2 = YY^*$, we have

$$\sigma_{k+1}(YY^*) \le \sigma_{k+1}(AA^*) + \|AA^* - YY^*\|_2. \tag{19.11}$$

Combining (19.9), (19.10) and (19.11), we have

$$\begin{aligned}
\|A^* z\|_2^2 &\le \|AA^* - YY^*\|_2 + \sigma_{k+1}(AA^*) + \|AA^* - YY^*\|_2 \\
&= 2\|AA^* - YY^*\|_2 + \sigma_{k+1}(AA^*).
\end{aligned} \tag{19.12}$$

Note that (19.12) holds for any z with $\|z\|_2 = 1$ and $Q_k^* z = 0$, thus combining with (19.7) and observing $\|A - A_k\|_2 = \sigma_{k+1}(A)$, we have

$$\begin{aligned}
\|A - Q_k Q_k^* A\|_2^2 &= \sup_{\substack{Q_k^* z = 0 \\ \|z\|_2 = 1}} \|A^* z\|_2^2 \\
&\le \sigma_{k+1}(AA^*) + 2\|AA^* - YY^*\|_2 \\
&= \|A - A_k\|_2^2 + 2\|AA^* - YY^*\|_2.
\end{aligned}$$

□

Lemma 19.1 tells that it is sufficient to bound $\|AA^*-YY^*\|_2$. From the former analysis on the choice of p_j, we have seen YY^* is a summation of i.i.d. matrices with desired expectation, and recall that *Matrix Bernstein Inequality* (in Lecture 7) is a powerful tool to give this type of bound.

Theorem 19.3. *(Matrix Bernstein Inequality) $Z_1, \dots, Z_l \in \mathbb{R}^{m\times n}$ are independent matrices, with $\mathbb{E}[Z_i] = 0$, and $\|Z_i\|_2 \le R$ for any $1 \le i \le l$, as well as $\|\sum_{i=1}^l \mathbb{E}[Z_i]^2\|_2 \le \sigma^2$. Define $Z = \sum_{i=1}^l Z_l$, then*

$$Pr\{\|Z\|_2 > t\} \le m \cdot \exp\left\{-\frac{t^2}{6(Rt+\sigma^2)}\right\}. \tag{19.13}$$

In order to use this theorem directly, we need to modify the random matrices such that the four ingredients in theorem are all satisfied. In this sense, assume the column we choose for ith step is j_i, where $1 \le i \le l$ and $1 \le j_i \le n$, then construct

$$Z_i = \frac{1}{lp_{j_i}} A_{j_i} A_{j_i}^* - \frac{1}{l} AA^*. \tag{19.14}$$

Note that the only randomness comes from the subindex j_i, and using $Y_i = \frac{1}{\sqrt{lp_{j_i}}} A_{j_i}$, it is simple to see

$$\sum_{i=1}^l Z_i = \sum_{i=1}^l \left(Y_i Y_i^* - \frac{1}{l} AA^*\right) = YY^* - AA^*. \tag{19.15}$$

Now we need to check the four conditions for Theorem 19.3 to hold.

- *Independence.* Since the randomness only comes from j_i, and each choice of column is independent, the constructed Z_is are also independent.
- *Zero mean.* By using (19.3), we have for any $1 \le i \le l$,

$$\mathbb{E}[Z_i] = \mathbb{E}[Y_i Y_i^*] - \frac{1}{l} AA^* = 0.$$

- *Absolute bound.* Using (19.2), we have for any $1 \le i \le l$,

$$\|Z_i\|_2 = \|Y_i Y_i^* - \frac{1}{l} AA^*\|_2 \le \|Y_i Y_i^*\|_2 + \frac{1}{l}\|AA^*\|_2 \le \frac{2}{l}\|A\|_F^2 \triangleq R. \tag{19.16}$$

- *Variance.* Note that

$$\begin{aligned}
\mathbb{E}[Z_i^2] &= \mathbb{E}\left[\left(\frac{1}{lp_{j_i}}A_{j_i}A_{j_i}^* - \frac{1}{l}AA^*\right)^2\right] \\
&= \mathbb{E}\left[\frac{1}{l^2p_{j_i}^2}\|A_{j_i}\|_2^2 A_{j_i}A_{j_i}^* + \frac{1}{l^2}(AA^*)^2 - \frac{2}{l^2p_{j_i}}A_{j_i}A_{j_i}^*AA^*\right] \\
&= \frac{1}{l^2}\left\{\|A\|_F^2\mathbb{E}\left[\frac{A_{j_i}A_{j_i}^*}{p_{j_i}}\right] + (AA^*)^2 - 2\mathbb{E}\left[\frac{A_{j_i}A_{j_i}^*}{p_{j_i}}\right]AA^*\right\} \\
&= \frac{1}{l^2}\left\{\|A\|_F^2\mathbb{E}\left[lY_iY_i^*\right] + (AA^*)^2 - 2\mathbb{E}\left[lY_iY_i^*\right]AA^*\right\} \\
&= \frac{1}{l^2}\left\{\|A\|_F^2AA^* + (AA^*)^2 - 2(AA^*)(AA^*)\right\} \\
&= \frac{1}{l^2}\left\{\|A\|_F^2AA^* - (AA^*)^2\right\}.
\end{aligned}$$

  Hence,

$$\begin{aligned}
\|\sum_{i=1}^{l}\mathbb{E}[Z_i^2]\|_2 &= \|\sum_{i=1}^{l}\frac{1}{l^2}\left\{\|A\|_F^2AA^* - (AA^*)^2\right\}\|_2 \\
&= \frac{1}{l}\|(\|A\|_F^2I - AA^*)AA^*\|_2 \\
&\le \frac{1}{l}\|(\|A\|_F^2I - AA^*)\|_2\|AA^*\|_2 \\
&\le \frac{1}{l}\|A\|_F^2\|A\|_2^2 \\
&\le \frac{1}{l}\|A\|_F^4 \triangleq \sigma^2, \qquad (19.17)
\end{aligned}$$

  where we have use a general result that $\|I - ZZ^*\|_2 \le 1$ for any matrix Z. This is true because if λ is an eigenvalue of ZZ^*, then $1-\lambda$ is an eigenvalue of $I - ZZ^*$. Since $\lambda \ge 0$, we have $1 - \lambda \le 1$, hence the largest eigenvalue is also no larger than 1.

Based on these conditions, it is straightforward to use Theorem 19.3. In particular, by choosing $t = \epsilon\|A\|_F^2$, we have

$$\begin{aligned}
\Pr\{\|AA^* - YY^*\|_2 \ge \epsilon\|A\|_F^2\} &\le m\cdot\exp\left\{-\frac{\epsilon^2\|A\|_F^4}{6(R\epsilon\|A\|_F^2 + \sigma^2)}\right\} \\
&= m\cdot\exp\left\{-\frac{\epsilon^2\|A\|_F^4}{6(\frac{2}{l}\|A\|_F^2\epsilon\|A\|_F^2 + \frac{1}{l}\|A\|_F^4)}\right\} \\
&= m\cdot\exp\left\{-\frac{l\epsilon^2}{6(2\epsilon+1)}\right\} \\
&\le m\cdot\exp\left\{-\frac{l\epsilon^2}{10}\right\} \\
&\le \delta,
\end{aligned}$$

for

$$\epsilon \geq \sqrt{\frac{10}{l}} \log\left(\frac{m}{\delta}\right). \tag{19.18}$$

To this end, we have bounded the approximation error by

$$\|A - Q_k Q_k^* A\|_2^2 \leq \|A - A_k\|_2^2 + 2\epsilon \|A\|_F^2,$$

where $Q_k Q_k^* A = X$ is the output of algorithm, and this holds with probability at least $1-\delta$, if choosing ϵ as (19.18). Thus, if the number of randomly sampled columns l is large enough, we can get a theoretical bound on approximation error with high probability.

19.3 Sparsification Method

Sparsification is another randomized scheme to approximate SVD. More precisely, consider a matrix $A \in \mathbb{R}^{m\times n}$. Another matrix Y, a parse version of A, is constructed, where each entry of A is randomly chosen, i.e.

$$Y_{ij} = \begin{cases} \frac{A_{ij}}{P_{ij}}, & \text{with probability } P_{ij}, \\ 0, & \text{with probability } 1 - P_{ij}, \end{cases} \tag{19.19}$$

where $P_{ij} = \frac{|E|}{mn}$. Perform SVD to the constructed $Y \in \mathbb{R}^{m\times n}$, i.e. $Y = Q\Sigma_Y V_Y^*$. Output $X = Q_k Q_k^* A$, where Q_k is consisted of top k columns of Q.

Note that because Y is sparse, faster iteration algorithms could be adopted for SVD. Moreover, $Y = \sum_{i,j} Y_{ij} e_i e_j^*$, which is also a form of summation of independent matrices, and $\mathbb{E}[Y] = \sum_{i,j} P_{ij} Y_{ij} e_i e_j^* = A$, which is the desired expectation. Hence, similar reason (Berstein inequality) gives the performance guarantee of this algorithm.

Theorem 19.4. *Assume $m \leq n$, $\max_{i,j} |A_{ij}| \leq A_{\max}$. If $|E| \geq Cn \log n$, then*

$$\|A - Q_k Q_k^* A\|_2 \leq \|A - A_k\|_2 + CA_{\max}\sqrt{mn}\sqrt{\frac{n}{|E|}}, \tag{19.20}$$

with probability at least $1 - \frac{1}{n^3}$.