CS 596: Theoretical Machine Learning

Lecturer: Pranjal Awasthi
Scribe: Dong Yang

Lecture #
November 17th, 2016

1 Previous Lecture

The singular value decomposition (SVD) of $M_{n\times n}$ can be expressed as $M = U\Sigma V^T$, where $U \in \mathbb{R}^{n\times r}$ and $V \in \mathbb{R}^{n\times r}$ are orthonormal matrices corresponding to the left and right singular vectors. r is the rank of the matrix M. The singular values are collected in the diagonal matrix $\Sigma = \text{diag}\,(\sigma_1, \sigma_2, ..., \sigma_r)$, where σ_i is the ith largest singular value of M, i.e., $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r$. Here $r << n$, so M is a low-rank matrix.

Question. *Give the sampled matrix $P_\Omega(M)$ under uniform sampling, how to infer M?*

$$P_\Omega(M) = \begin{cases} M_{i,j} & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}$$

In the previous lecture we saw the following guarantee achieved by r-SVD.

Theorem 1. *Let $\tilde{M} = P_\Omega(M)$, and $M' = \frac{1}{p}\tilde{M} = \sum_{i=1}^{n} \sigma_i u_i v_i^T$, then the output matrix is $M'_r = \sum_{i=1}^{r} \sigma_i u_i v_i^T$. If $p > \frac{c\log n}{n}$, then with the probability larger than $1 - \frac{1}{n^3}$,*

$$\left\|M'_r - M\right\|_F \leq O(\sqrt{\frac{r}{np}}).$$

(truncated SVD, works for any matrix)

2 Today's Lecture

In this lecture, we are going to apply the *alternate minimization* algorithm to improve the above bound. The output matrix M'_r can be factorized as following.

$$M'_r = X_0 \cdot Y_0^T,$$

where matrix X_0 is n by r, Y_0 is n by n. We can try to improve Mr' by the following procedure

for $t = 1, 2, ...$ **do**
  $Y_t \leftarrow \underset{Y}{\text{argmin}} \left\|P_\Omega\left(M - X_{t-1}Y^T\right)\right\|_F^2$
  $X_t \leftarrow \underset{X}{\text{argmin}} \left\|P_\Omega\left(M - XY_{t-1}^T\right)\right\|_F^2$
end for

Remark. *The alternate minimization would fail for some extreme cases of the matrix. For example,*

$$\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}.$$

This matrix cannot be fully completed from a sampling of its entries unless we see all the entries. For most sampling sets, we could only see zeros and it is almost impossible to infer that it is a non-zero matrix.

We will show the alternate minimization works if M is μ-incoherent (which successfully avoids the previous bad case).

2.1 Special Case: rank-1 matrix

In this section, we discuss about the special case of $M = \sigma_1 v^* (v^*)^T$, where $\|v^*\| = 1$ and M is μ-incoherent. In other words, M is a rank-1 matrix.

$$M = \begin{bmatrix} v_1^* \\ v_2^* \\ \vdots \\ v_n^* \end{bmatrix} [\sigma_1] \begin{bmatrix} v_1^* & v_2^* & \cdots & v_n^* \end{bmatrix}$$

Based on the definition of μ-incoherency, we have that $|v_i^*| \leq \sqrt{\frac{\mu}{n}}$. We will first show that because of incoherence X_0, is weakly correlated with v^*.

Lemma 1. *If $p \geq c\frac{\log n}{n}$, then with probability $\geq 1 - \frac{1}{n^3}$, we have that x_0, the output of 1-SVD, satisfies $\|x_0 - v^*\| \leq \frac{\mu}{\log n}$.*

Proof. Assume M is μ-incoherent, define $R = \frac{\tilde{M}}{p} - M$. Based on the definition of $\tilde{M}$, we have that $\mathrm{E}[R] = \frac{pM}{p} - M = 0$. Then with high probability, the spectral norm $\|R\| \leq \sigma\sqrt{n}$, where $\sigma = \max_{i,j} \mathrm{Var}(R_{i,j})$. Furthermore, because of incoherence of v^* we have that

$$\begin{aligned} \mathrm{E}[R_{i,j}^2] &= p\left(\frac{M_{i,j}}{p} - M_{i,j}\right)^2 + (1-p)(-M_{i,j})^2 \\ &\leq \frac{1}{p} \cdot \frac{\mu}{n} + (1-p)\frac{\mu^2}{n^2} \end{aligned}$$

Since $|v_i^*| \leq \sqrt{\frac{\mu}{n}}$, $M_{i,j} = \sigma_1 v_i v_j \leq \sigma\frac{\mu}{n}$. So we get that with probability larger than $1 - \frac{1}{n^3}$, $\|R\| \leq \sqrt{\frac{\sigma_1^2}{p} \cdot \frac{\mu^2}{n^2}} \cdot \sqrt{n} = \frac{\sigma_1 \mu}{\sqrt{np}}$. Next we use the Davis-Kahan theorem to deduce that

$$\|v^* - v_0\| \leq \frac{\left\|M - \frac{\tilde{M}}{p}\right\|}{\sigma_1} = \frac{\mu}{\sqrt{np}} \leq \frac{\mu}{\sqrt{\log n}}$$

□

In the rank 1 symmetric case, alternate minimization take a much simpler form.

$v_0 \leftarrow$ first singular vector of M'
$x_0 \leftarrow v_0$
for $t = 1, 2, \ldots$ **do**
  $y_t \leftarrow \underset{Y}{\mathrm{argmin}} \left\|P_\Omega\left(M - x_{t-1}y^T\right)\right\|_F^2$ {draw fresh samples every time}
  $x_t \leftarrow \frac{y_t}{\|y_t\|}$
end for

Let $V_\perp$ is the projection matrix perpendicular to v^*. The projection length of x_t is $\|V_\perp^* x_t\| = \sqrt{1-(v^* \cdot x_t)^2}$. Then we have the following guarantee

Theorem 2. *If $p \geq \frac{\mu^2}{n}\log^2(n)\log\left(\frac{1}{\epsilon}\right)$, then with the probability larger than $1-\frac{1}{n^3}$,*

$$\|V_\perp^* x_t\| \leq \epsilon, \text{ after } t = \log\left(\frac{1}{\epsilon}\right) \text{ steps}$$

Proof. It is not hard to see that x_0 can be made 2μ-incoherent by truncating large entries. A nice feature of the algorithm is that all the resulting iterates will also be $O(\mu)$-incoherent. This is not hard to prove. From now on we will assume that all the iterates x_t are $O(\mu)$-incoherent. Let's analyze a particular step of the algorithm. We have

$$y_t \leftarrow \underset{y}{\operatorname{argmin}} \left\|P_\Omega\left(M - x_{t-1}y^T\right)\right\|_F^2.$$

Define $L(y) = \underset{y}{\operatorname{argmin}} \sum_{(i,j)\in\Omega}\left(M_{i,j} - x_{t-1,i}y_j\right)^2$. In order to achieve the optimal solution, we have $\frac{\partial L(Y)}{\partial y_j} = 0$. Then we have the following. $\sum_{i,(i,j)\in\Omega}\left(M_{i,j} - x_{t-1,i}y_j\right)(-x_{i-1,t}) = 0$. This gives

$$\begin{aligned}
y_j &= \frac{\sum_{i,(i,j)\in\Omega} M_{i,j}x_{t-1,i}}{\sum_{i,(i,j)\in\Omega} x_{t-1,i}^2} \\
&= \sum_{i,(i,j)\in\Omega} M_{i,j}x_{t-1,i} + \frac{\sum_{i,(i,j)\in\Omega} M_{i,j}x_{t-1,i} - \sum_{i,(i,j)\in\Omega} x_{t-1,i}^2 \sum_{i,(i,j)\in\Omega} M_{i,j}x_{t-1,i}}{\sum_{i,(i,j)\in\Omega} x_{t-1,i}^2} \\
&= \left(Mx_{t-1}\right)_j + B_j^{-1}\left(\left(P_\Omega\left(\frac{M}{p}\right) - MB\right)x_{t-1}\right)_j
\end{aligned}$$

where $B_j^{-1} = \frac{p}{\sum_{i,(i,j)\in\Omega} x_{t-1,i}^2}$,

$$B = \begin{bmatrix} B_1 & & & \\ & B_2 & & \\ & & \ddots & \\ & & & B_n \end{bmatrix}.$$

Therefore, we have the equation $y_t = Mx_{t-1} + B_j^{-1}\left(P_\Omega\left(\frac{M}{p}\right) - MB\right)x_{t-1}$. Mx_{t-1} is similar to the term we encountered in the power method. Assume $x_t = \alpha v^* + V_\perp$, then

$$\begin{aligned}
Mx_t &= \sigma_1 v^*\left(v^*\right)^T\left(\alpha v^* + v_\perp\right) \\
&= \sigma_1 \alpha v^*.
\end{aligned}$$

Thus, MX_t only contributes on the direction of V^*.

$B_j^{-1}\left(P_\Omega\left(\frac{M}{p}\right) - MB\right)x_{t-1}$ is the noise term, correlated with the subspace we are trying to achieve. The noise would hurt a bit, but the total amount would decrease. We sort the above equation and obtain

$$y_t = Mx_{t-1} + B_j^{-1}\left(P_\Omega\left(\frac{M}{p} - M\right) - M(I-B)\right)x_{t-1}.$$

We have already argued tha $\|P_\Omega\left(\frac{M}{p}-M\right)\| \leq \frac{\sigma_1\mu}{\sqrt{\log n}}$. Based on the definition $B_j = \frac{\sum_{i,(i,j)\in\Omega} x_{t-1,i}^2}{p}$, then B is very close to the identity matrix since $\|x_{t-1}\| = \sum_i x_{t-1,i}^2 = 1$.

$$B_j = \frac{1}{p}\sum_{i,(i,j)\in\Omega} x_{t-1,i}^2$$

Since x_{t-1} is also incoherent, then $|x_{t-1,i}| \leq \sqrt{\frac{\mu}{n}}$. Using Bernstein inequality, with high probability we have

$$|B_j - 1| \leq \frac{1}{20}.$$

We apply this fact to the previous formulation.

$$\begin{aligned} y_t \cdot V^* &= Mx_{t-1}\cdot V^* + B_j^{-1}\left(P_\Omega\left(\frac{M}{p}\right) - MB\right)x_{t-1}\cdot V^* \\ &\geq \sigma_1\left(v^*\cdot x_{t-1}\right) - \frac{\sigma_1}{20}\left(v^*\cdot x_{t-1}\right) \end{aligned}$$

Similarly, $y_t \cdot V_\perp^* \leq \frac{\sigma_1}{20}\|V_\perp^*\cdot x_{t-1}\|$. Because $X_t = \frac{Y_t}{\|Y_t\|}$, it is easy to prove that

$$\|V_\perp^*\cdot X_t\| \leq \frac{1}{4}\|V_\perp^*\cdot X_{t-1}\|$$

. Therefore, the convergence would be achieved after $\log(\frac{1}{\epsilon})$ iterations with very high probability. □

2.2 General Case

For general symmetric matrix M, we have the factorization

$$M = \sigma_1 V_1 V_1^T + \sigma_2 V_2 V_2^T + \cdots + \sigma_n V_n V_n^T.$$

Then the update of alternate minimization becomes

$Y_{t+1} \leftarrow MX_t + G_t$

$X_{t+1} \leftarrow \text{orthogonalize}\left(Y_{t+1}\right)$ {Gram Schmidt}

We can prove that $\|G_t\| \leq \frac{1}{32}\|V_\perp^* X_{t-1}\|$ and get the following guarantee

Theorem 3. *If $p \geq \frac{\mu^2}{n}\log^2(n)\log\left(\frac{1}{\epsilon}\right)\left(\frac{\sigma_1}{\sigma_k}\right)^2$, then with high probability after $t = \log\left(\frac{1}{\epsilon}\right)$ steps,*

$$\|V_\perp^* X_t\| \leq \epsilon$$

The dependence on $\left(\frac{\sigma_1}{\sigma_k}\right)^2$ can be removed via semi definite programming that minimizes the following program

$$\begin{aligned} \text{minimize} \quad & \left\|M'\right\|_* = \sum_i \sigma_i \\ \text{subject to} \quad & M'_{i,j} = M_{i,j} \text{ for any } (i,j)\in\Omega, \end{aligned}$$

Additional Readings

- The paper on viewing alternate minimization as noisy power method. `https://arxiv.org/abs/1212.0467`.
- Improvements to the above result. `https://arxiv.org/abs/1312.0925`.

3 Appendix

[Frobenius Norm]

$$\|A_{m\times n}\|_F \equiv \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{i,j}|^2}$$

[Nuclear Norm]

$$\|A_{m\times n}\|_* \equiv \sum_{i=1}^{\min m,n} \sigma_i = \text{trace}\left(A^T \cdot A\right)$$

[Spectral Norm]

$$\|A\| = \max_{\|x\|_2=1} \|Ax\|_2$$

[Incoherent]
A matrix $M \in \mathbb{R}^{m\times n}$ is incoherent with parameter μ if:

$$\left\|e_i^T U\right\|_2^2 \leq \frac{\mu r}{m}, \forall i \in [m]$$

$$\left\|V^T e_j\right\|_2^2 \leq \frac{\mu r}{n}, \forall j \in [n],$$

where $M = U\Sigma V^T$ is the SVD of M and $e_i^T U$,$V^T e_j$ denote the i-th row of U and the j-th row of V respectively.
[Bernstein Inequality]
Let X_1, X_2, ..., X_n be independent random variable. Suppose that $|X_i| < M$ for all i.

$$\sum_{i=1}^{n} \text{Var}\left(x_i\right) = \sigma^2$$

Then, for all t,

$$\Pr\left(\left(\sum_i X_i - \text{E}\left[\sum_i X_i\right]\right) > t\right) \leq e^{\frac{-t^2}{\frac{Mt}{3}+\sigma^2}}.$$