0368-3248-01-Algorithms in Data Mining **Fall 2013**

Lecture 9: Matrix approximation continued

Lecturer: Edo Liberty

Warning: *This note may contain typos and other inaccuracies which are usually discussed during class. Please do not cite this note as a reliable source. If you find mistakes, please inform me.*

Recap

Last class was dedicated to approximating matrices by sampling individual entries from them. We used the following useful concentration result for sums of random matrices.

Lemma 0.1 (Matrix Bernstein Inequality [1]). *Let $X_1, \ldots, X_s$ be independent $m \times n$ matrix valued random variables such that*

$$\forall_{k\in[s]} \quad \mathbb{E}[X_k] = 0 \quad \textit{and} \quad \|X_k\| \leq R$$

Set $\sigma^2 = \max\{\|\sum_k \mathbb{E}[X_k X_k^T]\|, \|\sum_k \mathbb{E}[X_k^T X_k]\|\}$ *then*

$$\Pr[\|\sum X_k\| > t] \leq (m+n)e^{-\frac{t^2}{\sigma^2 + Rt/3}}$$

We obtained an approximate sampled matrix B which was sparser than A. For example, for a matrix A containing values in $\{1, 0, -1\}$ it sufficed that B contains only s entries and

$$s \in O\left(\frac{nr\log(n/\delta)}{\varepsilon^2}\right).$$

Moreover, we got that $\|A - B\| \leq \varepsilon\|A\|$. We also claimed that we can compute the PCA projection of B, P_k^B instead of that of A, P_k and still have:

$$\sigma_{k+1} = \|A - P_k A\| \leq \|A - P_k^B A\| \leq \sigma_{k+1} + \varepsilon\|A\|$$

In this class we will reduce the amount of space needed to be independent of n. We will do this by approximating AA^T directly. Here we follow the ideas in [2] and give a much simpler proof which, unfortunately, obtains slightly worse bounds.

PCA with an approximate covariance matrix

Here we see that approximating $BB^T \sim AA^T$ is sufficient in order to compute an approximate PCA projection (Lemma 3.8 in [2]). Let P_k^B denote the projection on the top k left singular values of B.

$$\begin{aligned}
\|A - P_k^B A\|^2 &\leq \sup_{x,\|x\|=1} \|xA - xP_k^B A\|^2 && (1)\\
&= \sup_{x\in null(P_k^B),\|x\|=1} \|xA\|^2 && (2)\\
&= \sup_{x\in null(P_k^B),\|x\|=1} \|xAA^T\| && (3)\\
&\leq \sup_{x\in null(P_k^B),\|x\|=1} \|x(AA^T - BB^T)\| + \|xBB^T\|^2 && (4)\\
&\leq \|AA^T - BB^T\| + \sigma_{k+1}\mathbb{E}(BB^T) && (5)\\
&\leq 2\|AA^T - BB^T\| + \sigma_{k+1}(AA^T) && (6)
\end{aligned}$$

The last transition is due to the fact that $\sigma_{k+1}(BB^2) \leq \sigma_{k+1}(AA^T) + \|AA^T - BB^T\|$. By taking the square root and recalling that $\sigma_{k+1}(AA^T) = \sigma^2_{k+1}(A) \doteq \sigma^2_{k+1}$ we get:

$$\|A - P^B_k A\| \leq \sqrt{\sigma^2_{k+1} + 2\|AA^T - BB^T\|} \leq \sigma_{k+1} + \sqrt{2\|AA^T - BB^T\|}$$

Therefore, we have that

$$\sqrt{2\|AA^T - BB^T\|} \leq \varepsilon\|A\| \Longrightarrow \|A - P^B_k A\| \leq \sigma_{k+1} + \varepsilon\|A\|$$

Column subset selection by sampling

From this point on, our goal is to find a matrix B such that $\|AA^T - BB^T\|$ is small and B is as sparse or small as possible. Note that

$$AA^T = \sum_{j=1}^{n} A_j A_j^2$$

where we denote by A_j the j'th column of the matrix A. Let

$$B \leftarrow A_j/\sqrt{p_j} \text{ with probability } p_j$$

Computing the expectation of BB^T we have

$$\mathbb{E}[BB^T] = \sum_{j=1}^{n} p_j (A_j/\sqrt{p_j})(A_j/\sqrt{p_j})^T = \sum_{j=1}^{n} A_j A_j^T = AA^T$$

Clearly, we cannot hope to approximate the matrix A with only one column. We therefore define B to be s such sampled columns from A side by side.

$$B = \frac{1}{\sqrt{s}}[B_1|\ldots|B_s]$$

Just to clarify, B is an $m \times s$ matrix containing s columns from A (rescaled) and $B_k \leftarrow A_j/\sqrt{p_j}$ with probability p_j. Computing the expectation of BB^T we get that

$$\mathbb{E}[BB^T] = \sum_{k=1}^{s} \mathbb{E}[\frac{1}{s} B_k B_k^T] = \mathbb{E}[B_k B_k^T] = AA^T$$

We are now ready to use the matrix Bernstein inequality above:

$$\|BB^T - AA^T\| = \|\sum_{k=1}^{s} \frac{1}{s}(B_k B_k^T - AA^T)\| = \|\sum_{k=1}^{s} X_k\|$$

To make things simpler we pick $p_j = \|A_j\|^2/\|A\|_F^2$. In words, the columns are picked with probability proportional to their squared 2 norm.

$$R = \max\|X_k\| \leq \max_{j=1}^{n} \|\frac{1}{s} A_j A_j^T/p_j\| + \|\frac{1}{s} AA^t\| = \frac{1}{s}\|A\|_F^2 + \frac{1}{s}\|A\|_2^2$$

$$\sigma^2 = \|\sum_k \mathbb{E}[X_k X_k^T]\| = \frac{1}{s}\|\mathbb{E}[B_k B_k^T B_k B_k^T - AA^T AA^T]\| \quad (7)$$

$$\leq \frac{1}{s}\|\sum_{j=1}^{n} p_j A_j A_j^T A_j A_j^T/p_j^2\| + \frac{1}{s}\|A\|_2^4 = \frac{1}{s}\|A\|_F^2\|A\|_2^2 + \frac{1}{s}\|A\|_2^4 \quad (8)$$

Plugging both expressions into the matrix Chernoff bound above we get that

$$\Pr[\|BB^T - AA^T\| > \varepsilon^2\|A\|_2^2/2] \leq 2me^{-\frac{s\varepsilon^4\|A\|_2^4/4}{\|A\|_F^2\|A\|_2^2+\|A\|_2^4+\varepsilon^2\|A\|_2^2\|A\|_F^2/3+\varepsilon^2\|A\|_2^4/3}}$$

By using that $\|A\|_F \geq \|A\|_2$ and that $\varepsilon \leq 1$ and denoting the numeric rank of A by $r = \|A\|_F^2/\|A\|_2^2$ we can simplify this to be

$$\Pr[\|BB^T - AA^T\| > \varepsilon^2\|A\|_2^2/2] \leq 2me^{-\frac{s\varepsilon^4}{16r}}$$

To conclude, it is enough to sample

$$s \geq \frac{16r}{\varepsilon^4}\log(2m/\delta)$$

columns from A (with probability proportional to their squared 2 norm) to form a matrix B such that $\|BB^T - AA^T\| \leq \varepsilon^2\|A\|_2^2/2$ with probability at least $1-\delta$. According to above this also gives us that $\|A - P_k^B A\| \leq \sigma_{k+1} + \varepsilon\|A\|$ which completes our claim. Note that the number of non zeros in B is bounded by $s \cdot m$ which is independent of n, the number of columns. This is potentially significantly better than the results obtained in last week's class.

Remark 0.1. *More elaborate column selection algorithms exist which provide better approximation but these will not be discussed here. See [3] for the latest result that I am aware of.*

Deterministic Lossy SVD

I this section we will see that the above approximation can be improved using a simple trick [4]. The algorithm keeps an $m \times s$ sketch matrix B which is updated every time a new column from the input matrix A is added. It maintains the invariant that the last column in the sketch is always zero. When a new input row is added it is places in the last (all zeros) column of the sketch. Then, using its SVD the sketch is rotated from the right so that its columns are orthogonal. Finally, the sketch column norms are 'shrunk' so that the last column is again all zeros. In the algorithm we denote by $[U, \Sigma, V] \leftarrow SVD(B)$ the Singular Value Decomposition of B. We use the convention that $U\Sigma V^T = B$, $U^TU = V^TV = I$, and $\Sigma = \text{diag}([\sigma_1, \ldots, \sigma_s])$, $\sigma_1 \geq \ldots \geq \sigma_s$. The notation I stands for the $s \times s$ identity matrix while B_s denotes the s'th column of B (similarly A_j).

Input: s, $A \in \mathbb{R}^{m\times n}$
$B^0 \leftarrow$ all zeros matrix $\in \mathbb{R}^{m\times s}$
for $i \in [n]$ **do**
  $C^i = B^{i-1}$
  $C^i_s = A_i$ — Put A_i in the last (all zeros) column of C^i
  $[U^i, \Sigma^i, V^i] \leftarrow SVD(C^i)$
  $D^i \leftarrow U^i\Sigma^i$
  $\delta_i \leftarrow \Sigma^i_{s,s}$
  $W^i \leftarrow \sqrt{I - \Sigma^{-2}\delta_i^2}$
  $B^i \leftarrow D^iW^i$
end for
Return: $B \leftarrow B^n$

Lemma 0.2. *Let B be the output of the above algorithm for a matrix A and an integer s then:*

$$\|AA^T - BB^T\| \leq \|A\|_F^2/s$$

Proof. We start by bounding $\|AA^T - BB^T\|$ from above:

$$\|AA^T - BB^T\| = \max_{\|x\|=1}(x^TAA^Tx - x^TBB^Tx) = \max_{\|x\|=1}(\|x^TA\|^2 - \|x^TB\|^2).$$

We open this with a simple telescopic sum $\|x^T B\|^2 = \sum_{j=1}^n \|x^T B^i\|^2 - \|x^T B^{i-1}\|^2$ and by replacing $\|x^T A\|^2 = \sum_{j=1}^n (x^T A_j)^2$.

$$\begin{aligned}
\|x^T A\|^2 - \|x^T B\|^2 &= \sum_{j=1}^n [(x^T A_j)^2 - (\|x^T B^i\|^2 - \|x^T B^{i-1}\|^2)] && (9)\\
&= \sum_{j=1}^n [((x^T A_j)^2 + \|x^T B^{i-1}\|^2) - \|x^T B^i\|^2] && (10)\\
&= \sum_{j=1}^n [\|x^T C^i\|^2 - \|x^T B^i\|^2] && (11)\\
&= \sum_{j=1}^n [\|x^T D^i\|^2 - \|x^T D^i W^i\|^2] && (12)\\
&= \sum_{j=1}^n x^T [D^i (D^i)^T - D^i (W^i)^2 (D^i)^T] x && (13)\\
&= \sum_{j=1}^n x^T [\delta_i^2 D^i (\Sigma^i)^{-2} (D^i)^T] x && (14)\\
&= \sum_{j=1}^n x^T [\delta_i^2 U^i (U^i)^T] x && (15)\\
&\leq \sum_{j=1}^n \delta_i^2 \|U^i (U^i)^T\| \leq \sum_{j=1}^n \delta_i^2 && (16)
\end{aligned}$$

For this to mean anything we have to bound the term $\sum_{j=1}^n \delta_i^2$. We do this computing the Frobenius norm of B.

$$\begin{aligned}
\|B^n\|_F^2 &= \sum_{i=1}^n \|B^i\|_F^2 - \|B^{i-1}\|_F^2 && (17)\\
&= \sum_{i=1}^n [\|B^i\|_F^2 - \|D^i\|_F^2] + [\|D^i\|_F^2 - \|C^i\|_F^2] + [\|C^i\|_F^2 - \|B^{i-1}\|_F^2] && (18)
\end{aligned}$$

Let us deal with each term separately:

$$\begin{aligned}
\|B^i\|_F^2 - \|D^i\|_F^2 &= \mathrm{tr}(B^i (B^i)^T - D^i (D^i)^T) = \mathrm{tr}(\delta_i^2 U^i (U^i)^2) = s\delta_i^2 && (19)\\
\|D^i\|_F^2 - \|C^i\|_F^2 &= 0 && (20)\\
\|C^i\|_F^2 - \|B^{i-1}\|_F^2 &= \|A_i\|^2 && (21)
\end{aligned}$$

Putting these together we get

$$\|B\|_F^2 = \|B^n\|_F^2 = \|A\|_F^2 - s \sum \delta_i^2$$

Since $\|B\|_F^2 \geq 0$ we conclude that $\sum \delta_i^2 \leq \|A\|_F^2 / s$ Combining with the above:

$$\|BB^T - AA^T\| \leq \|A\|_F^2 / s$$

□

Finally, we recall that to achieve the approximation guarantee $\|A - P_k^B A\| \leq \sigma_{k+1} + \varepsilon \|A\|$ it is sufficient to require $\|BB^T - AA^T\| \leq \varepsilon^2 \|A\|_2^2 / 2$ which is obtained when

$$s \geq \frac{2r}{\varepsilon^2}$$

This completes the claim.

Discussion

Note that while the deterministic lossy SVD procedure requires less space it also requires more operations per column insertion. Namely, it computes the SVD of the sketch in every iteration. This can be somewhat improved (see [4]) but it is still far from being as efficient as column sampling. Making this algorithm faster while maintaining its approximation guaranty is an open problem. Another interesting problem is to make this algorithm take advantage of the matrix sparsity.

References

[1] Emmanuel Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Commun. ACM*, 55(6):111–119, June 2012.

[2] Mark Rudelson and Roman Vershynin. Sampling from large matrices: An approach through geometric functional analysis. *J. ACM*, 54(4), July 2007.

[3] Christos Boutsidis, Petros Drineas, and Malik Magdon-Ismail. Near optimal column-based matrix reconstruction. In *Proceedings of the 2011 IEEE 52nd Annual Symposium on Foundations of Computer Science*, FOCS '11, pages 305–314, Washington, DC, USA, 2011. IEEE Computer Society.

[4] Edo Liberty. Simple and deterministic matrix sketching. *CoRR*, abs/1206.0594, 2012.