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

Lecture 8: Matrix sampling and rank-k approximation

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 Please do not cite this note as a reliable source.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.*

In this lesson we will try to approximate a matrix A by a sparser matrix B. Clearly , B will be easier to store and more computationally efficient when applied as an operator. Before we begin thought, let us see how such an approximation allows us to compute an approximate PCA for A.

Remember that the best approximation of A of rank k is $A_k = P_k A$ where P_k is the projection on the top k singular values of A.

$$\min_{rank(A')=k} \|A - A'\| = \|A - A_k\| = \|A - P_k A\| = \sigma_{k+1}$$

where σ_{k+1} is the $(k+1)$'th largest singular value of A. Now, assume we are given a matrix B such that $\|A - B\| \leq \varepsilon \|A\|$ for some small enough ε. If we compute the PCA of B and look at the projection on its top k singular values P_k^B, what can we say about $\|A - P_k^B A\|$? To compute $\|A - P_k^B A\|$ we multiply it by a test vector from the left x.

$$\begin{aligned}
\|A - P_k^B A\| &\leq \sup_x \|xA - xP_k^B A\| && (1)\\
&= \sup_{x \in null(P_k^B)} \|xA - xP_k^B A\| && (2)\\
&= \sup_{x \in null(P_k^B)} \|xA\| && (3)\\
&\leq \sup_{x \in null(P_k^B)} \|xA - xB\| + \|xB\| && (4)\\
&\leq \|A - B\| + \sigma_{k+1}^B && (5)\\
&\leq \sigma_{k+1} + 2\varepsilon \|A\|_2 && (6)
\end{aligned}$$

The last line stems from the fact that $\sigma_{k+1}^B \leq \sigma_{k+1} + \|A - B\|$. What we conclude is that the PCA computed for B is a good approximation to the actual PCA of A in the sense that

$$\|A - P_k^B A\| = \sigma_{k+1} + 2\varepsilon \|A\|_2$$

Element-wise sampling

In this section we will argue that it is sufficient to sample single entries from a matrix to approximate it. There is a very simple proof of for this fact [1] but we will follow a slightly more involved one which gives better results [2].

In this section we will denote by $a_{i,j}$ the value of entry (i,j) in the matrix A. We will denote by $A_{i,j}$ the $m \times n$ matrix whose entries are all zeros except entry (i,j) which is set to $a_{i,j}$. In other words, $A = \sum_{i,j} A_{i,j}$. Given A our goal is to produce another matrix B such that $\|A - B\|$ is small and that B is much sparser than A.

Let us define the B to take the value $A_{i,j}/p_{i,j}$ with probability $p_{i,j}$.

$$\mathbb{E}[B] = \sum_{i,j} p_{i,j}(A_{i,j}/p_{i,j}) = \sum_{i,j} A_{i,j} = A$$

Of course, we cannot hope to approximate A with a matrix with only 1 non zero. We therefore average s such variables

$$B = \frac{1}{s}\sum_{k=1}^{s} B_k$$

We still have that $\mathbb{E}[B] = A$ but now we can use matrix-chernoff bounds to argue about $\|A - B\|$.

Lemma 0.1 (Matrix Bernstein Inequality [3]). *Let $X_1, \ldots, X_s$ be independent $m \times n$ matrix valued random variables such that*

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

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

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

To use the lemma we convert $\|A - B\|$ to a sum of mean zero matrices

$$\|A - B\| = \|\frac{1}{s}\sum_{k=1}^{s} B_k - A\| = \|\sum_{k=1}^{s}(B_k - A)/s\|$$

Since $\mathbb{E}[B_k] = A$ we can set $X_k = (B_k - A)/s$ and satisfy $\mathbb{E}[X_k] = 0$. Now, to compute R and σ^2 we need to set $p_{i,j}$. It makes sense to sample larger elements in the matrix with higher probability. But, there is a large number of ways to do that, see for example references inside [2]. For the sake of this class we'll pick a simple distribution which will make our derivation easier. We set $p_{i,j} = |a_{i,j}|/|A|_1$ where $|A|_1 = \sum_{i,j} |a_{i,j}|$. Clearly p is a valid distribution since $\sum_{i,j} p_{i,j} = \sum_{i,j} a_{i,j}/|A|_1 = 1$. Let's start with computing $R = \max_k \|X_k\|$

$$\max_k \|X_k\| = \max \|(A_{i,j}/p_{i,j} - A)/s\| \leq |A|_1/s + \|A\|_2/s$$

To compute σ^2 we start by computing $\|\sum_k \mathbb{E}[X_k X_k^T]\|$:

$$\|\sum_k \mathbb{E}[(B_k - A)(B_k - A)^T/s^2]\| = \|\mathbb{E}[(B_k - A)(B_k - A)^T/s]\| = \|\mathbb{E}[B_k B_k^T] - AA^T\|/s \leq \|\mathbb{E}[B_k B_k^T]\|/s + \|A\|_2^2/s.$$

To compute $\mathbb{E}[B_k B_k^T]$ we recall that B_k contains only one non zero. $B_k = A_{i,j}/p_{i,j}$ w.p. $p_{i,j} = a_{i,j}/|A|_1$. Therefore, $B_k B_k^T$ also contains only one entry but on the diagonal. Namely with probability $a_{i,j}/|A|_1$ we have that

$$B_k B_k^T = A_{i,j} A_{i,j}^T / p_{i,j}^2 = |A|_1^2 e_{i,i}$$

Where $e_{i,i}$ is a matrix holding the value 1 in position (i,i) and zero everywhere else. By remembering that the norm of a diagonal matrix is the maximal value on its diagonal we can compute the expectation

$$\|\mathbb{E}[B_k B_k^T]\| = \|\sum_{i,j}(|a_{i,j}|/|A|_1)|A|_1^2 e_{i,i}\| = |A|_1 \|\sum_{i,j} |a_{i,j}| e_{i,i}\| = |A|_1 \max_i \sum_j |a_{i,j}| = |A|_1 \rho$$

Here we define $\rho = \max_i \sum_j |a_{i,j}|$ to be the maximal ℓ_1 norm of a row in A. Putting this all together we get:

$$\Pr[\|A - B\| > t] \leq (m+n)e^{-\frac{st^2}{|A|_1\rho + \|A\|_2^2 + |A|_1 t/3 + \|A\|_2 t/3}}$$

Setting $t = \varepsilon\|A\|_2$, and demanding a failure probability of at most δ we get:

$$s \geq \frac{\log((m+n)/\delta)}{\varepsilon^2}\left(\frac{|A|_1\rho}{\|A\|_2^2} + 1 + \frac{\varepsilon|A|_1}{3\|A\|_2} + \frac{\varepsilon}{3}\right)$$

It is hard to immediately see what this means since we need to quantify $\|A\|_1$ and ρ and their relation to $\|A\|_2$. As an example, let us consider a matrix A which contains S non zero values and $a_{i,j} \in \{1, 0, -1\}$.

For such matrices $|A|_1 = S = \|A\|_f^2 = \sqrt{S}\|A\|_f$. Moreover, we recall that the numerical rank of A is defined as $\|A\|_f^2/\|A\|_2^2$. The above reduces to the following

$$s \geq \frac{\log((m+n)/\delta)}{\varepsilon^2}\left(r\rho + \varepsilon\sqrt{Sr} + O(1)\right)$$

Moreover, assuming $n \geq m$ we have that both $\rho < n$ and and $\sqrt{S} \leq n$. So, this could be reduced even further to

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

Note that this is does not depended on the number of rows m!

References

[1] Sanjeev Arora, Elad Hazan, and Satyen Kale. A fast random sampling algorithm for sparsifying matrices. In *Proceedings of the 9th international conference on Approximation Algorithms for Combinatorial Optimization Problems, and 10th international conference on Randomization and Computation*, APPROX'06/RANDOM'06, pages 272–279, Berlin, Heidelberg, 2006. Springer-Verlag.

[2] Petros Drineas and Anastasios Zouzias. A note on element-wise matrix sparsification via matrix-valued chernoff bounds. *CoRR*, abs/1006.0407, 2010.

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