0368-3248-01-Algorithms in Data Mining **Fall 2012**

Lecture 7: Matrix Sampling

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 Please do not cite this note as a reliable source.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

---

1 Ashwelde-Winter inequality

In their work [1] Ashwelde and Winter give an incredibly useful bound for the sums of independent random symmetric matrices. Here we recap the lemma statement. A short proof due to Roman Vershynin [2] is given as a reference.

Lemma 1.1. *Let X_i be independent random $d \times d$ symmetric matrices with mean zero s.t. $\|X_i\| \le 1$. Let $S_n = \sum_{i=1}^n X_i$, let $\sigma_i^2 = \|\mathrm{Var}[X_i]\|$ and $\sigma^2 = \sum_{i=1}^n \sigma_i^2$, then:*

$$\Pr[\|S_n - \mathbb{E}[S_n]\| \ge t] \le d \cdot \max\{e^{-\frac{t^2}{4\sigma^2}}, e^{\frac{-t}{2}}\}$$

2 Rank-k approximation

Here we try to approximate AA^T by sampling columns of the matrix A. From this point onwards we assume, w.l.o.g. that $\|A\|_{fro} = 1$.

Define n unit norm matrices $C_i = A_{(i)}A_{(i)}^T/\|A_{(i)}\|^2$ where A_i is the i'th column of A. Also define the random matrix valued variable Z which takes values C_i w.p. $p_i = \|A_{(i)}\|^2$. Note that p is a distribution since $\sum_{i=1}^n p_i = \sum_{i=1}^n \|A_{(i)}\|^2 = \|A\|_{fro}^2 = 1$. Let us compute the expectation of Z:

$$\mathbb{E}[Z] = \sum_{i=1}^n p_i C_i = \sum_{i=1}^n \|A_{(i)}\|^2 (A_{(i)}A_{(i)}^T/\|A_{(i)}\|^2) = \sum_{i=1}^n A_{(i)}A_{(i)}^T = AA^T$$

We will therefore try to approximate AA^T by averaging r independent copies of such variables $\frac{1}{r}\sum_{i=1}^r Z_i$.

$$\begin{aligned} \Pr[\|\frac{1}{r}\sum_{i=1}^r Z_i - AA^T\| &> \varepsilon\|AA^T\|] = \Pr[\|\sum_{i=1}^r (Z_i - AA^T)\| > r\varepsilon\|AA^T\|] && (1)\\ &= \Pr[\|\sum_{i=1}^r X_i\| > r\varepsilon\|AA^T\|/2] && (2) \end{aligned}$$

where we define $X_i = (Z_i - AA^T)/2$. To apply the matrix chernoff bound above we need to make sure that the variables X_i meet the conditions. First, they are clearly independent since Z_i are. Also, they have mean zero since $\mathbb{E}[Z_i] = AA^T$. Finally, $\|X_i\| = \|(Z_i - AA^T)/2\| \le \|Z_i\|/2 + \|AA^T\|/2 \le 1$. Thus, to apply the bound above we only need to compute $\sigma^2 = \sum_{i=1}^r \|\mathbb{E}[X_i^2]\|$.

$$\begin{aligned} \sigma_i^2 &\le \|\mathbb{E}[X_i^2]\| \le \|\mathbb{E}[(Z_i - AA^T)^2]\|/2 && (3)\\ &= \|\mathbb{E}[Z_i^2 - ZAA^T - AA^TZ + (AA^T)^2]\|/2 && (4)\\ &= \|AA^T - (AA^T)^2\|/2 \le \|AA^T\|/2 && (5) \end{aligned}$$

This gives that $\sigma^2 \leq r\|AA^T\|/2$.

$$\Pr[\|\sum_{i=1}^{r} X_i\| > r\varepsilon\|AA^T\|/2] \leq m \cdot e^{-\frac{r\varepsilon^2\|AA^T\|}{8}}$$

This gives us an ε approximation in the spectral norm with probability at least $1-\delta$ if $r \geq \frac{8}{\|AA^T\|\varepsilon^2}\log(m/\delta)$. Another trivial observation is that $1 = \|A\|_{fro} = tr(AA^T) \leq m\|AA^T\|$ which gives that $\frac{1}{\|AA^T\|} \leq m$. To recap, for any matrix, sampling $r = \frac{8m}{\varepsilon^2}\log(m/\delta)$ columns is sufficient in order to approximate AA^T in the 2-norm up to multiplicative factor $\varepsilon\|AA^T\|$.

3 Rank-k Approximation

What does this tell us about the SVD. Note that the matrix resulting from the sampling above can be thought of the matrix $\hat{A}\hat{A}^T$ where $\hat{A} \in \mathbb{R}^{m\times r}$ contains rescaled sampled columns of A. More accurately, $\hat{A}_{(i)} = \frac{1}{\sqrt{r}\|A_{(j)}\|}A_j$ if in step i we picked column j from A.

We want to say that $\hat{A}$ somehow represents A well. One way to say this is that the left singular vectors of $\hat{A}$ and A are "similar" (the right singular vectors are not in the same dimension) To make this more accurate we recap the property of the best rank-k approximation of A

$$\|A - P_k A\| = \sigma_{k+1}$$

Where the projection matrix $P_k = U_k U_k^T$ contains the top k left singular vectors of A. Now consider projecting A on the top left singular vectors of $\hat{A}$ instead, how much do we "loose" by that?

A lemma 4 from [3] makes this exact.

Lemma 3.1. *Let $\hat{P}_k$ be the projection on the top k left singular vectors of $\hat{A}$, then*

$$\|A - \hat{P}_k A\|^2 \leq \sigma_{k+1}^2 + 2\|\hat{A}\hat{A}^T - AA^T\|$$

Proof. To see this lets compute the supremum over values $\|x(A - \hat{P}_k A)\|$, clearly x is such that $x\hat{P}_k = 0$.

$$\begin{aligned}
\|A - \hat{P}_k A\|^2 &= \langle AA^T x, x\rangle && (6)\\
&= \langle (AA^T - \hat{A}\hat{A}^T)x, x\rangle + \langle \hat{A}\hat{A}^T x, x\rangle && (7)\\
&\leq ||AA^T - \hat{A}\hat{A}^T|| + \hat{\sigma}_{k+1}^2 && (8)
\end{aligned}$$

Where $\hat{\sigma}_{k+1}$ is the $k+1$'th singular value of $\hat{A}$. Since, $\hat{\sigma}_{k+1}^2 \leq \sigma_{k+1}^2 + ||AA^T - \hat{A}\hat{A}^T||$ we get the lemma. □

Finally, the SVD of $\hat{A}$ is a good approximation to the SVD of A in the sense that

$$\|A - \hat{P}_k A\| \leq \sigma_{k+1} + 2\varepsilon\|A\|_2$$

References

[1] Rudolf Ahlswede and Andreas Winter. Strong converse for identification via quantum channels. *IEEE Transactions on Information Theory*, 48(3):569–579, 2002.

[2] Roman Vershynin. A note on sums of independent random matrices after ahlswede-winter. *Lecture Notes.*

[3] Petros Drineas and Ravi Kannan. Pass efficient algorithms for approximating large matrices, 2003.