CPSC 536N: Randomized Algorithms 2011-12 Term 2

Lecture 15

Prof. Nick Harvey *University of British Columbia*

1 Low-rank approximation of matrices

Let A be an arbitrary $n \times m$ matrix. We assume $n \leq m$. We consider the problem of approximating A by a low-rank matrix. For example, we could seek to find a rank s matrix B minimizing $\|A - B\|$.

It is known that a truncated singular value decomposition gives an optimal solution to this problem. Formally, let $A = U \Sigma V^{\mathsf{T}}$ be the singular value decomposition of A. Let $\sigma_1(A) \geq \cdots \geq \sigma_n(A)$ be the singular values (i.e., diagonal entries of Σ.) Let $u_1, \ldots, u_n$ be the left singular vectors (i.e., columns of U). Let $v_1, \ldots, v_n$ be the right singular vectors (i.e., columns of V).

Fact 1 $\sum_{i=1}^{s} \sigma_i u_i v_i^{\mathsf{T}}$ *is a solution to* $\min_{B \,:\, \mathrm{rank}(B) \leq s} \|A - B\|$*, and the minimum value equals* σ_{s+1}*.*

Another way of stating this same fact is as follows.

Fact 2 *Let* $V_s = [v_1, \ldots, v_s]$ *be the* $n \times s$ *matrix consisting of the top* s *left singular vectors. Let* $P_s = V_s V_s^{\mathsf{T}}$ *be the orthogonal projection onto the span of* $\{v_1, \ldots, v_s\}$*. Then* $B = AP_s$ *is a solution to* $\min_{B \,:\, \mathrm{rank}(B) \leq s} \|A - B\|$*. Furthermore,* $\|A - AP_s\| = \sigma_{s+1}$*.*

The SVD can be computed in $O(mn^2)$ time. (Strictly speaking, this is not correct — the singular values can be irrational, so realistically we can only compute an ϵ-approximate SVD.) With the recent trend towards analyzing "big data", a running time of $O(mn^2)$ might be too slow.

In the past 15 years there has been a lot of work on sophisticated algorithms to quickly compute low-rank approximations. For example, one could measure the approximate error in different norms, sample more or fewer vectors, improve the running time, reduce the number of passes over the data, improve numerical stability, etc. Much more information can be found in the survey of Mahoney, the review article of Halko-Martinsson-Tropp, the PhD thesis of Boutsidis, etc.

2 Rudelson & Vershynin's Algorithm

Let A be an $n \times m$ matrix. The **Frobenius norm** $\|A\|_F$ is defined by

$$\|A\|_F^2 \;:=\; \mathrm{tr}(AA^{\mathsf{T}}) \;=\; \sum_{i,j} A_{i,j}^2 \;=\; \sum_i \sigma_i^2.$$

The **stable rank** (or **numerical rank**) of A is

$$\frac{\|A\|_F^2}{\|A\|^2} \;=\; \frac{\sum_i \sigma_i^2}{\max_i \sigma_i^2}.$$

Clearly the stable rank cannot exceed the usual rank, which is the number of strictly positive singular values. The stable rank is a useful surrogate for the rank because it is largely unaffected by tiny singular values.

Let r denote the stable rank of A and let $a_1, \ldots, a_n$ be the rows of A. We consider the following algorithm for computing a low-rank approximation $\tilde{A}$ to A.

- Initially $\tilde{A}$ is the empty matrix.
- Fix any $k \geq 32r\log(n)/\epsilon^4$. (Here we are assuming that the algorithm knows r, or at least reasonable bounds on r.)
- For $i = 1, \ldots, k$
  - Pick a row a_i with probability proportional to $\|a_i\|^2/\|A\|_F^2$.
  - Add the row $\frac{\|A\|_F}{\sqrt{k}\,\|a_i\|}a_i$ to $\tilde{A}$.
- Compute the SVD of $\tilde{A}$.

The runtime of this algorithm is dominated two main tasks. (1) The computation of the sampling probabilities. This can be done in time linear in the number of non-zero entries of A. (2) Computing the SVD of $\tilde{A}$. Since $\tilde{A}$ has size $k \times m$, this takes $O(mk^2) = O(m \cdot \mathrm{poly}(r, \log n, 1/\epsilon))$ time.

Theorem 3 *Fix any $s \in \{1, \ldots, n\}$. Let P_s be the orthogonal projection onto top s right singular vectors of $\tilde{A}$. With probability at least $1 - 2/n$,*

$$\|A - AP_s\| \;\leq\; \sigma_{s+1}(A) + \epsilon\|A\|.$$

In other words, the best rank s projection obtained from $\tilde{A}$ does nearly as well as the best rank s projection obtained from A. (Compare against Fact 2.) Since our algorithm explicitly computes the SVD of $\tilde{A}$, so it can easily compute the matrix P_s. We can then use P_s to efficiently compute an approximate SVD of A as well; see the survey of Halko, Martinsson and Tropp.

Corollary 4 *Set $s = r/\epsilon^2$. Let P_s be the orthogonal projection onto top s right singular vectors of $\tilde{A}$. Then, with probability at least $1 - 2/n$,*

$$\|A - AP_s\| \;\leq\; 2\epsilon\|A\|. \tag{1}$$

Let us contrast the error guarantee of (1) with the guarantee we achieved in the previous lecture. Last time we sampled a matrix by sampled a matrix L_w and showed it approximates L_G in the sense that

$$\left|\, x^\mathsf{T} L_G x - x^\mathsf{T} L_w x \,\right| \;\leq\; \epsilon x^\mathsf{T} L_G x \qquad \forall x \in \mathbb{R}^n.$$

We say that our result from last time achieves "multiplicative error guarantee". In contrast, Corollary 4 only guarantees that

$$\left|\, \|Ax\| - \|AP_s x\| \,\right| \;\leq\; 2\epsilon\|A\| \qquad \forall x \in \mathbb{R}^n \text{ with } \|x\| = 1,$$

even though $\|Ax\|$ may be significantly smaller than $\epsilon\|A\|$. We say that today's theorem only achieves "additive error guarantee".

To prove today's theorem, we will use a version of the Ahlswede-Winter inequality that provides an additive error guarantee.

Theorem 5 *Let Y be a random, symmetric, positive semi-definite $n \times n$ matrix such that $\|\mathrm{E}[Y]\| \leq 1$. Suppose $\|Y\| \leq R$ for some fixed scalar $R \geq 1$. Let $Y_1, \ldots, Y_k$ be independent copies of Y (i.e., independently sampled matrices with the same distribution as Y). For any $\epsilon \in (0,1)$, we have*

$$\Pr\left[\left\| \frac{1}{k} \sum_{i=1}^{k} Y_i - \mathrm{E}[Y_i] \right\| > \epsilon \right] \;\leq\; 2n \cdot \exp(-k\epsilon^2/4R).$$

The proof is almost identical to the proof of Theorem 1 in Lecture 13. The only difference is that the final sentence of that proof should be deleted.

Our Theorem 3 is actually weaker than Rudelson & Vershynin's result. They show that one can take k to be roughly $r \log r/\epsilon^4$, which is quite remarkable because it is "dimension free": the number of samples does not depend on the dimension n. Unfortunately our proof, which uses little more than the Ahlswede-Winter inequality, does not give that stronger bound because the failure probability in the Ahlswede-Winter inequality depends on the dimension. Rudelson & Vershynin prove an (additive error) variant of Ahlswede-Winter which avoids avoids this dependence on the dimension. Oliveira 2010 and Hsu-Kakade-Zhang 2011 give further progress in this direction.

2.1 Proofs

The proof of Theorem 3 follows quite straightforwardly from Theorem 5 and the following lemma, which we prove later.

Lemma 6 *Let B be an arbitrary $n \times m$ matrix. Let P_s be the orthogonal projection onto top s left singular vectors of B. Then*

$$\|A - AP_s\|^2 \;\leq\; \sigma_{s+1}(A)^2 + 2\|A^\mathsf{T} A - B^\mathsf{T} B\|.$$

Proof:(of Theorem 3.) Note that everything is invariant under scaling A. So we can assume $\|A\| = 1$. We defined $a_1, \ldots, a_n$ to be the rows of A, but let us now transpose them to become column vectors, so

$$A^\mathsf{T} A \;=\; \sum_{i=1}^{n} a_i a_i^\mathsf{T}.$$

Let $Y_1, \ldots, Y_k$ be independent, identically distributed random matrices with the following distribution:

$$\Pr\left[\, Y_j \;=\; \|A\|_F^2 \cdot \frac{a_i a_i^\mathsf{T}}{a_i^\mathsf{T} a_i} \,\right] \;=\; \frac{a_i^\mathsf{T} a_i}{\|A\|_F^2}.$$

This is indeed a probability distribution because

$$\|A\|_F^2 \;=\; \mathrm{tr}(A^\mathsf{T} A) \;=\; \mathrm{tr}(\sum_i a_i a_i^\mathsf{T}) \;=\; \sum_i a_i^\mathsf{T} a_i.$$

Note that the change to $\tilde{A}^\mathsf{T} \tilde{A}$ during the jth iteration of the algorithm is Y_j/k.

We will apply Theorem 5 to $Y_1, \ldots, Y_k$. We have

$$\mathrm{E}[Y_j] \;=\; \sum_{i=1}^{n} \frac{a_i^\mathsf{T} a_i}{\|A\|_F^2} \cdot \left(\|A\|_F^2 \cdot \frac{a_i a_i^\mathsf{T}}{a_i^\mathsf{T} a_i} \right) \;=\; \sum_{i=1}^{n} a_i a_i^\mathsf{T} \;=\; A^\mathsf{T} A,$$

so $\|\mathrm{E}[Y_j]\| \leq 1$. We may take $R := \|Y_j\| = \|A\|_F^2 = r$, the stable rank of A (since we assume $\|A\| = 1$). Since $\tilde{A}^\mathsf{T}\tilde{A} = \sum_{j=1}^k Y_j/k$ and $\mathrm{E}[Y_j] = A^\mathsf{T}A$, we get

$$\Pr\left[\ \left\|\tilde{A}^\mathsf{T}\tilde{A} - A^\mathsf{T}A\right\| > \epsilon^2/2\ \right] \ = \ \Pr\left[\ \left\|\frac{1}{k}\sum_{j=1}^k Y_j - \mathrm{E}[Y_j]\right\| > \epsilon^2/2\ \right] \ \leq \ 2n \cdot \exp(-k\epsilon^4/16r) \ = \ 2/n,$$

by Theorem 5.

Now we apply Lemma 6 to A and $\tilde{A}$. So, with probability at least $1 - 2/n$,

$$\|A - AP_s\|^2 \ \leq \ \sigma_{s+1}(A)^2 + \epsilon^2 \ \leq \ (\sigma_{s+1}(A) + \epsilon)^2.$$

Taking square roots completes the proof. □

Proof:(of Corollary 4). By the ordering of the singular values,

$$s\sigma_s(A)^2 \ \leq \ \sum_{i=1}^s \sigma_s(A)^2 \ \leq \ \|A\|_F^2,$$

implying $\sigma_s(A)^2 \leq \|A\|_F^2/s$. In particular, if $s = \|A\|_F^2/\epsilon^2\|A\|^2 = r/\epsilon^2$ then $\sigma_s(A) \ \leq \ \epsilon\|A\|$. □

Proof:(of Lemma 6). Let Q_s be the orthogonal projection onto the kernel of P_s (i.e., the span of the bottom $n - s$ left singular vectors of B). Then

$$\|A - AP_s\| \ = \ \|A(I - P_s)\| \ = \ \|AQ_s\| \ = \ \sup_{x\,:\,\|x\|=1} \|AQ_s x\| \ = \ \sup_{x \in \mathrm{span}(Q_s)\,:\,\|x\|=1} \|Ax\|.$$

So,

$$\begin{aligned}
\|A - AP_s\|^2 \ &= \ \sup_{x \in \ker(P_s)\,:\,\|x\|=1} \|Ax\|^2 \\
&= \ \sup_{x \in \ker(P_s)\,:\,\|x\|=1} \langle A^\mathsf{T}Ax, x\rangle \\
&\leq \ \sup_{x \in \ker(P_s)\,:\,\|x\|=1} \langle (A^\mathsf{T}A - B^\mathsf{T}B)x, x\rangle \ + \ \sup_{x \in \ker(P_s)\,:\,\|x\|=1} \langle B^\mathsf{T}Bx, x\rangle \\
&\leq \ \|A^\mathsf{T}A - B^\mathsf{T}B\| \ + \ \sigma_{s+1}(B^\mathsf{T}B) \\
&\leq \ 2\|A^\mathsf{T}A - B^\mathsf{T}B\| \ + \ \sigma_{s+1}(A^\mathsf{T}A),
\end{aligned}$$

where the last step uses the following fact. □

Fact 7 *Let X and Y be symmetric matrices of the same size. Let $\lambda_i(X)$ and $\lambda_i(Y)$ respectively denote the ith largest eigenvalues of X and Y. Then*

$$\max_i |\lambda_i(X) - \lambda_i(Y)| \ \leq \ \|X - Y\|.$$

Proof: See Horn & Johnson, "Matrix Analysis", page 370. □