Stat260/CS294: Randomized Algorithms for Matrices and Data

Lecture 17 - 10/30/2013

Lecture 17: Toward Randomized Low-rank Approximation in Practice

Lecturer: Michael Mahoney *Scribe: Michael Mahoney*

Warning: these notes are still very rough. They provide more details on what we discussed in class, but there may still be some errors, incomplete/imprecise statements, etc. in them.

17 Toward Randomized Low-rank Approximation in Practice

Today, we will continue with the discussion of improved low-rank matrix approximation algorithms by describing a slightly different but much more powerful structural result that will allow us to reparameterize the low-rank approximation problem to obtain improved results both in theory and in practice. Here is reading for today.

- Lemma 2 (of arXiv-v2, or Lemma 4.2 of SODA) of: Boutsidis, Mahoney, and Drineas "An Improved Approximation Algorithm for the Column Subset Selection Problem"
- Theorem 9.1 of: Halko, Martinsson, and Tropp, "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions"

In particular, today we will cover the following topics.

- A discussion of theory-practice gap issues in low-rank matrix approximation algorithms.
- A finer structural result that we will use in the next few classes to bridge that gap.

17.1 Some Challenges for Low-rank Matrix Approximation in Practice

As with the LS problem and algorithms, here we also want to understand how these theoretical ideas for randomized low-rank matrix approximation can be used in practice. As we will see, just as with the LS problem and algorithms, the basic ideas do go through to practical situations, but some of the theory must be modified in certain ways. Among the issues that will come up for the randomized low-rank approximation situation are the following.

- It might be too expensive to sample $O\left(\frac{k\log(k)}{\epsilon^2}\right)$ rows/columns, and it might be difficult to do so if the constant in the big-O is left unspecified. Instead, we might want to choose exactly k, or we might want to choose $k+p$, where p is a small integer such as 5 or 10.
- In many applications, and in particular in those that are particularly interested moderate- to high-precision low-rank matrix approximation, e.g., numerical analysis and scientific computing applications, there are other goals of interest. For example, given a good approximation to

an orthogonal basis Q approximating A, one might want to find other types of matrix decompositions (e.g., various QR decompositions, thin SVDs, interpolative decompositions, etc.).

- One might want to parameterize problems/algorithms in terms of fixed rank version (where the input is a rank parameter, which is the approach we have taken), or one might want to parameterize problems/algorithms in terms of a fixed precision version (roughly, fix a prespecified precision level, e.g., near machine precision, and look for an approximation that provides that numerical error).
- If the spectrum decays somewhat slowly but not very slowly, then one might be interested in doing some sort of power iteration, which will help the spectrum to decay more quickly, and it might be of interest to incorporate this process directly into the algorithm.
- Rather than asking for a priori worst-case error bounds, one might be interested in doing a posteriori error estimation and deciding whether to continue with the algorithm based on the output of that estimation procedure.

We will briefly describe all of these issues—many of the issues are similar to those that arose when we discussed how RandNLA algorithms for the LS problem work in practice, but here we are considering the low-rank matrix approximation problem—but before we do that, let's give a more refined structural result. This result gives improved results in general; and, in particular, it makes it easier to perform these extensions.

17.2 A More Refined Structural Result for Low-rank Approximation

Recall that when we discussed the LS problem, we described a deterministic structural result, and then we showed how random sampling and random projections interface to that result. Moreover, how the randomization interfaced to that structure differed for algorithms that obtained the best results in worst-case theory versus those that obtained the best results in practice. For the versions of the low-rank approximation problem that we discussed in the last class, i.e., the $1 \pm \epsilon$ relative-error sampling and projection algorithms, we just related them to the LS problem. Thus, we really didn't take into account the low-rank structure, e.g., how the top and bottom subspaces of the input matrix interacted, in a particularly refined way. The reason was two-fold: (1) we were only interested in how the sample reproduced the top part of the spectrum and the top subspace of the matrix; and (2) we were willing to oversample to a level sufficient to obtain worst-case bounds. If we are interested in obtaining more refined results, as is common in practice, then we need a more refined structural result that takes into account how the top and bottom part of the spectrum of a matrix interact.

To do that, observe that there are actually two related ways that we can break up the generalized LS problem. Given a matrix $A \in \mathbb{R}^{m \times n}$, where $\text{rank}(A) = k$ and a matrix $B \in \mathbb{R}^{m \times p}$, consider the generalized LS problem:

$$\text{argmin}_{X \in \mathbb{R}^{n \times p}} \left\| Z^T A X - Z^T B \right\|_\xi = \left(Z^T A \right)^\dagger Z^T B,$$

where $Z^T U$ is full rank (i.e., the rank $= k$). Then, we can split up the expression $\|AX_{opt} - B\|_\xi$ in one of two ways.

- $$\begin{aligned}\left\|A\left(Z^TA\right)^\dagger Z^TB-B\right\|_\xi &\leq \left\|U^\perp U^{\perp T}B\right\|_\xi \\ &+ \left\|U^TZZ^TU^\perp U^{\perp T}B\right\|_\xi \\ &+ \max_i\left|\sigma_i\left(Z^TU\right)-\sigma_i^{-1}\left(Z^TU\right)\right|\left\|Z^TU^\perp U^{\perp T}B\right\|_\xi\end{aligned}$$

- $$\left\|A\left(Z^TA\right)^\dagger Z^TB-B\right\|_\xi \leq \left\|U^\perp U^{\perp T}B\right\|_\xi + \left\|\left(U^TZ\right)^\dagger Z^TU^\perp U^{\perp T}B\right\|_\xi$$

Note that these two correspond to a generalization of the two related ways that we proved the tall LS result. Here, though, the two different ways to split up this expression will lead to two different structural results. One is the immediate generalization of the LS result that can be used to get $(1+\epsilon)$ relative-error bounds on the top part of the spectrum that we saw in the last class. The other can be used to do that, but it is more general; in particular, it can be used to get a more refined structural result that leads to better algorithms for the CSSP as well as for random projection algorithms with very aggressive downsampling.

The main issue is that the generalized LS algorithm we had assumes that the matrix is exactly rank k, which essentially means that it is exactly rectangular and just artificially fat. Then, we applied it to arbitrary matrices by carefully wedging projection matrices at various places, but the consequence of this is that we only got control on the top part of the spectrum. Now, let's do better by getting a structural result that says how the sampling operator interacts with both the top and bottom part of the spectrum. This structural result will hold for any sketching/sampling/projection matrix, and the randomness will enter only through it, so in that sense it will decouple the linear algebraic structure from the randomness.

Here is the basic setup. Let $A \in \mathbb{R}^{m\times n}$, and let its SVD, $A = U\Sigma V^T$, be represented as

$$A = U\begin{pmatrix}\Sigma_1 & 0\\ 0 & \Sigma_2\end{pmatrix}\begin{pmatrix}V_1^T\\ V_2^T\end{pmatrix}$$

where Σ_k is the $k\times k$ diagonal matrix consisting of the top k singular values, Σ_2 is the $(\min\{m,n\}-k)\times(\min\{m,n\}-k)$ diagonal matrix consisting of the bottom $\min\{m,n\}-k$ singular values, V_1^T and V_2^T are the matrices of the associated singular vectors, etc. (Note that we are using subscripts differently/inconsistently with respect to how we used them before, as well as how we will use them later; here, "1" and "2" refer to the top and bottom part of the spectrum, respectively.)

In this case, assume that we have the sketching matrix $S \in \mathbb{R}^{\ell\times k}$, which could be a sampling or projection or some other matrix, and where $\ell \geq k$. For example, $\ell = k$ or $\ell = k+p$ for $p = 5$ or $p = 10$, or $\ell = O\left(k\log(k)/\epsilon^2\right)$ are three regimes of particular interest to us. Then, we can define

$$\begin{cases}\Omega_1 = V_1^TS\\ \Omega_2 = V_2^TS\end{cases}$$

to be the perturbed version of the singular subspaces. To obtain good low-rank matrix approximation, we will want to control the singular subspaces of Ω_1 and Ω_2. In the absence of sketching, they are orthogonal, i.e., $V_2^TV_1 = 0$, and thus we will want to show that the sketched versions of

the subspaces are approximately orthogonal. This is different than before, where we just needed to show that

$$\left\|\Omega_1\Omega_1^T - I\right\|_2 = \left\|V_1^T S S^T V_1 - I\right\|_2 < 1/2,$$

i.e., that the top part of the subspace is well-behaved. That is, here we want to control both the top and bottom part of the spectrum as well as how they interact with each other via the sketching matrix S.

To do this, let $C = AS$, in which case we can write

$$C = U \begin{pmatrix} \Sigma_1 V_1^T S \\ \Sigma_2 V_2^T S \end{pmatrix} = U \begin{pmatrix} \Sigma_1 \Omega_1^T \\ \Sigma_2 \Omega_2^T \end{pmatrix},$$

where $\Sigma_1 V_1^T S$ is $k \times \ell$ and $\Sigma_2 V_2^T S$ is $(n-k) \times \ell$. (Note that C does not need to be actual columns, unless S is a sampling matrix, but instead it is any sketch of the columns.)

If Q is an orthonormal basis for the range of C (in this discussion, we are *not* filtering through the best rank k approximation to C, which corresponds to the "easier" situation before), then $QQ^T = P_C$, and we want to bound

$$\left\|A - QQ^T A\right\|_\xi = \left\|(I - P_C)\, A\right\|_\xi .$$

One can then prove the following, which is our main structural result for low-rank matrix approximation via randomized algorithms.

Theorem 1 *Given the above setup, then assuming that $\Omega_1 = V_1^T S$ has full rank, then*

$$\left\|(I - P_C)\, A\right\|_\xi \leq \left\|A - A_k\right\|_\xi + \left\|\Sigma_2 \Omega_2 \Omega_1^\dagger\right\|_\xi ,$$

where $\Omega_1 = V_1^T S$ and $\Omega_2 = V_2^T S$.

Remark. This structural result was first established and proven by Boutsidis et al. in the context of the Column Subset Selection Problems, and it was reproved with more complicated methods by Halko et al. in the context of parameterizing random projection algorithms for high-quality implementations. Gittens, Gu, and several others have used it since then in one form or another. That and other prior work which used this structural result only established if for the spectral and Frobenius norms, but it actually holds for any unitarily-invariant norm. This result is due to Drineas and Mahoney, but we haven't published it yet, so I'll include it here.

Remark. The $\Omega_2\Omega_1^\dagger$ term describes the interaction between the top and bottom part of the spectrum. The "unsketched" version of this is $V_2^T V_1^{T\dagger} = V_2^T V_1 = 0$, in which case there is no interaction between these orthogonal subspaces.

Remark. The assumption that Ω_1 is full rank is very nontrivial. Indeed, the entire point of using leverage-based sampling or random projections for the overdetermined LS problem is to ensure that. Here, it holds for worst-case input if we use leverage-based sampling or if we use random projections, with parameters set appropriately. Of course, if one can do an after-the-fact check to confirm that it is true (which is what one often does in practice), then one can use this theorem.

Proof:[of theorem] First note that

$$\left\|A - P_C A\right\|_\xi = \left\|A - AS\,(AS)^\dagger\, A\right\|_\xi \tag{1}$$

and also that

$$(AS)^\dagger = \text{argmin}_{X \in \mathbb{R}^{k \times n}} \|A - ASX\|_\xi \tag{2}$$

and also that these two results hold for any unitarily invariant matrix norm. So, we can replace $(AS)^\dagger$ in (1) with any other $k \times n$ matrix and replace the equality (=) with an inequality ($\leq$). In particular, we will replace $(AS)^\dagger A$ with $(A_k S)^\dagger A_k$. Doing this, we get the following.

$$\begin{aligned}
\|A - P_C A\|_\xi &= \left\|A - AS\,(AS)^\dagger A\right\|_\xi \\
&\leq \left\|A - AS\,(A_k S)^\dagger A_k\right\|_\xi \\
&= \left\|A - A_k + A_k - (A - A_k + A_k)\,S\,(AS)^\dagger A\right\|_\xi \\
&\leq \underbrace{\left\|A_k - A_k S\,(A_k S)^\dagger A_k\right\|_\xi}_{\gamma_1} + \underbrace{\|A - A_k\|_\xi}_{\gamma_2} + \underbrace{\left\|(A - A_k)\,S\,(A_k S)^\dagger A_k\right\|_\xi}_{\gamma_3}.
\end{aligned}$$

Let's bound each of those three terms. Since γ_2 is simply $\|A - A_k\|_\xi$, we'll bound the other two terms. First, bound γ_1 as follows:

$$\begin{aligned}
\gamma_1 &= \left\|A_k - A_k S\,(A_k S)^\dagger A_k\right\|_\xi \\
&= \left\|A_k - A_k S\left(U_k \Sigma_k V_k^T S\right)^\dagger A_k\right\|_\xi \\
&= \left\|A_k - A_k S\left(V_k^T S\right)^\dagger (U_k \Sigma_k)^\dagger A_k\right\|_\xi \quad \text{(since both } V_k^T S \text{ and } U_k \Sigma_k \text{ are full rank)} \\
&= \left\|A_k - U_k \Sigma_k \underbrace{V_k^T S\left(V_k^T S\right)^\dagger}_{I_k} \underbrace{(U_k \Sigma_k)^\dagger U_k \Sigma_k}_{I_k} V_k^T\right\|_\xi \\
&= \left\|A_k - U_k \Sigma_k V_k^T\right\|_\xi \\
&= 0.
\end{aligned}$$

Next, bound γ_3 as follows:

$$\begin{aligned}
\gamma_3 &= \left\|(A - A_k)\,S\,(A_k S)^\dagger A_k\right\|_\xi \\
&= \left\|(A - A_k)\,S\left(U_k \Sigma_k V_k^T S\right)^\dagger A_k\right\|_\xi \\
&= \left\|(A - A_k)\,S\left(V_k^T S\right)^\dagger (U_k \Sigma_k)^\dagger A_k\right\|_\xi \quad \text{(since both matrices are full rank)} \\
&= \left\|U_{k,\perp} \Sigma_{k,\perp} V_{k,\perp}^T S\left(V_k^T S\right)^\dagger \underbrace{(U_k \Sigma_k)^\dagger U_k \Sigma_k}_{I_k} V_k^T\right\|_\xi \\
&= \left\|U_{k,\perp} \Sigma_{k,\perp} V_{k,\perp}^T S\left(V_k^T S\right)^\dagger V_k^T\right\|_\xi \\
&\leq \left\|\Sigma_{k,\perp} V_{k,\perp}^T S\left(V_k^T S\right)^\dagger\right\|_\xi \quad \text{(since the orthogonal matrices can be dropped)}
\end{aligned}$$

Here, we use $U_{k,\perp}$, $\Sigma_{k,\perp}$, $V_{k,\perp}$ to refer to the parts of U, Σ, and V that are orthogonal to the best rank k approximation to A.

The theorem then follows.

◇

Two final remarks. First, you can prove the generalization of this result to when there is a square on the norm using more sophisticated methods. Second, you can prove the generalization of this result to when $A_k = AV_kV_k^T$ is replaced with AYY^T is any approximation to V_k. This is of interest, since one can then choose Y to be any approximation to V_k, e.g., one constructed with a random projection algorithm.