Stat260/CS294: Randomized Algorithms for Matrices and Data

Lecture 13 - 10/16/2013

Lecture 13: Randomized Least-squares Approximation in Practice, Cont.

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.

13 Randomized Least-squares Approximation in Practice, Cont.

We continue with the disucssion from last time. There is no new reading, just the same as last class.

Today, we will focus on three things.

- We will discuss issues with good implementations that arise when downsampling more aggressively than worst-case theory needs.
- We will talk more about generalized eigenvalues and their relationship with perturbations of good preconditioners.
- We will talk about how those ideas are used in RandNLA solvers.

(This will wrap up our discussion of least squares. Next time, we will move onto RandNLA algorithms for low-rank matrix approximation.)

13.1 Issues with good implementations

Recall that last time we were discussing Blendenpik, which provides an implementation of a RandNLA algorithm to solve the very overdetermined LS problem. There is a bit of a theory-practice gap, and looking at how Blendenpik addresses some of those issues is illustrative more generally for other problems such as low-rank matrix approximation.

The basic idea is that rather than doing a QR decomposition of A, do a QR decomposition on ΠA, where Π is a FJLT (or some other, e.g., data-aware subspace embedding), and then use the $\tilde{R}^{-1}$ from QR on the subproblem as a preconditioner for an iterative algorithm on the original problem. We saw that if $\Pi A = \tilde{Q}\tilde{R}$ then $\kappa\left(A\tilde{R}^{-1}\right) = \kappa(SU)$. If we sample "enough," i.e., $\Omega(d\log(d)/\epsilon)$, then this condition number is $\leq 1+\epsilon$ and the very good subspace embedding provides a very good preconditioner.

While this sampling complexity is basically necessary if we want to use ΠA and Πb to solve the subproblem and obtain $1\pm\epsilon$ approximation guarantees, this might be overkill, if these methods are coupled with an iterative algorithm. This is fortunate, since in practice one typically down-samples more aggressively s.t. $\kappa\left(A\tilde{R}^{-1}\right)$ is still somewhat large (and/or that the sample looses rank, which

essentially means that it is infinite). In this situation, the worst-case TCS theory fails, but we we might still be able to use these RandNLA methods to construct a good preconditioner for a traditional NLA iterative algorithm. Even if we are not in such an extreme situation, constructing preconditioners, whether with deterministic or randomized methods, is expensive, and there is a tradeoff between very good preconditioners that need very few iterations and moderately good preconditioners that are less expensive but need more iterations. Having control over this tradeoff is very important in practice.

Said another way, here is the main issue.

- What if there are just a very few rows with very bad leverage scores (and we sample uniformly), or (if we sample non-uniformly) we get estimates of the leverage scores (via the fast leverage score algorithm, except that we down-sample more aggressively in the Π_1 projection inside that algorithm) that are good for most of the leverage scores but severely underestimate the leverage of a small number of rows? In this case, the following is true.
  - We don't have a go0d subspace embedding, and so solving the subproblem does not lead to a good solution, in worst case analysis.
  - Since we don't have a good subspace embedding, when using the sketch as a preconditioner, the condition number of the preconditioner is large (or infinite, if rank is lost), and so a naive bound that uses the condition to bound the number of iterations leads to poor results.
  - We can often still use this as a preconditioner for iterative methods such as LSQR and get good convergence in a small number of iterations in practice. Thus, it is still a reasonably good preconditioner, and there is theory to explain this—basically, the reason is since a small perturbation of it is a good preconditioner, and we will discuss that now.
- The basic idea here is that if there are a few rows that are missed (either with uniform sampling since they have high leverage, or with nonuniform sampling if we down-sample too aggressively), then the number of large singular values is bounded by the condition number of the large rows. Having just a few large rows means that $\kappa\left(A\tilde{R}^{-1}\right)$ is large, but is doesn't much affect the convergence properties of LSQR.
- Somewhat more precisely, the "R" factor from a QR decomposition of a perturbation $\tilde{A}$ of A is effective as a preconditioner for A; if is A is poorly conditioned, then it is ok if $\tilde{A}$ is well-conditioned. Ditto for A^TA and $\tilde{A}^T\tilde{A}$. Results of this form hold in general; and, for RandNLA, A is the matrix from the sample, and $\tilde{A}$ is some other matrix that isn't explicitly constructed.

Now, we'll go into more detail on this. In particular, we will outline some of the ideas from the ANT paper, but we will skip some of the details that are less relevant.

13.2 Using perturbed QR factorizations to solve linear LS problems

Before we go into some of the details, let's describe the notion of generalized eigenvalues and generalized condition numbers that we mentioned a few classes ago.

Recall that if we use CG with a preconditioner M, then we solve $Ax = b$ by solving $M^{-1}Ax = M^{-1}b$. If $M = A$, the we are basically solving the linear system with a direct method, while of $M = I$, then

we are basically solving it with unpreconditioned iterative method. So, the goal is to find an M such that M is easy to compute and $\kappa\left(M^{-1}A\right)$ is small. Recall also the definitions of generalized eigenvalues and generalized condition numbers.

Definition 1 *Let $S,T \in \mathbb{R}^{n\times n}$, then $\lambda = \lambda(S,T) \in \mathbb{R}^n$ is a* finite generalized eigenvalue *of the matrix pencil (S,T) if there exists a vector $v \neq 0$ such that $\begin{cases} Sv = \lambda Tv \\ Tv \neq 0 \end{cases}$. In addition, ∞ is an* infinite generalized eigenvalue *of (S,T) if there exists a $v \neq 0$ such that $\begin{cases} Tv = 0 \\ Sv \neq 0 \end{cases}$. Note that ∞ is a eigenvalue of (S,T) iff 0 is an eigenvalue of (T,S).*

Definition 2 *The finite and infinite eigenvalues of a pencil are* determined eigenvalues*, i.e., the eigenvector uniquely determines the eigenvalue. If $Sv = Tv = 0$ for some $v \neq 0$, then v is an* indeterminate eigenvector*, since $Sv = \lambda Tv$, for all $\lambda \in \mathbb{R}$. We can denote the set of determined eigenvalues of (S,T) by $\Lambda(S,T)$.*

Definition 3 *Let $S,T \in \mathbb{R}^{n\times n}$ have the same null space. The generalized condition number is*

$$\kappa(S,T) = \frac{\lambda_{max}(S,T)}{\lambda_{min}(S,T)},$$

where them max *and* min *are over the determined eigenvalues of (S,T).*

Here is a fact. The behavior of preconditioned iterative methods is determined by the clustering of the generalized eigenvectors, and the number of iterations is bounded by a quantity that is proportional to the generalized condition number.

- CG converges in $O\left(\sqrt{\kappa(A,M)}\right)$ iterations.
- LSQR on A preconditioned by R converges in $O\left(\sqrt{\kappa(A^TA, R^TR)}\right)$ iterations.

Note that these bounds provide sufficient conditions; but since these bounds come from a more refined analysis that depends on the entire spectrum, they are not necessary. We now turn to a theory that provides a more refined analysis.

In particular, here we describe a theory (from the ANT paper) that is more general than RandNLA preconditioning but which can be applied directly to RandNLA preconditioning.

Let A be a matrix and let $\hat{A} = \begin{pmatrix} A \\ B \end{pmatrix}$. Then,

$$\begin{aligned}
\left(\hat{A}^T\hat{A}\right)^{-1}A^TA &= \left(A^TA + B^TB\right)^{-1}A^TA \\
&= \left(A^TA + B^TB\right)^{-1}\left(A^TA + B^TB - B^TB\right) \\
&= I - \underbrace{\left(A^TA + B^TB\right)^{-1}B^TB}_{=\Omega}.
\end{aligned}$$

Here is a fact:

$$\operatorname{rank}(\Omega) \leq \operatorname{rank}(B).$$

So, in particular, if the matrix B is low-rank, then the matrix $\left(\hat{A}^T\hat{A}\right)^{-1} A^T A$ is a low-rank perturbation of the identity I.

The important consequence of this is that a symmetric rank-k perturbation of the identity I has $\leq k$ non-unit eigenvalues. In exact arithmetic, this is sufficient to guarantee the convergence in k iterations of several Krylov methods. So, in particular, for the LS problem, the Cholesky factor of $\hat{A}^T\hat{A}$, which is the R matrix from the QR decomposition of $\hat{A}$, is a good LS preconditioner for A.

The same analysis extends to other types of perturbations, e.g., to the case when the perturbation is such that rows of A are dropped. The Avron, Ng, and Toledo paper generalized this to other matrix perturbations.

- To when $\hat{A}$ is singular.
- To when rows are removed instead of added.
- To when columns are exchanged.
- To preconditioners for $\hat{A}$ other than the R factor.

They also bound the size of the non-unit eigenvalues, which is important when A is rank deficient.

Observe that the generalized spectrum of $\left(A^T A, A^T A\right)$ is very simple: the pencil has $\operatorname{rank}(A)$ eigenvalues that are 1 and the rest are indeterminate. In light of this, let's describe the spectra of the following perturbed pencils.

- $\left(A^T A, A^T A + B^T B - C^T C\right)$
- $\left(A^T A, \hat{A}^T\hat{A}\right)$, when $A = \begin{pmatrix} D & E \end{pmatrix}$ and $\hat{A} = \begin{pmatrix} D & F \end{pmatrix}$.

The perturbations of $A^T A$ shift some of the eigenvalues of $\left(A^T A, A^T A\right)$. Let's call the eigenvalues that move away from 1 *runaway* eigenvalues. We will analyze the runaway eigenvalues, which govern the convergence of LSQR when a factorization or approximation of a perturbed matrix is used as a preconditioner.

To start simple, let's give a result that bounds the number of runaway eigenvalues (and other aspects of the spectrum) when we add/subtract a symmetric product from a matrix.

Theorem 1 *Let $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{k\times n}$, $C \in \mathbb{R}^{r\times n}$, for some $1 \leq k+r \leq n$, and define $\chi = \begin{pmatrix} B \\ C \end{pmatrix}$.*

Then,

- *In the pencil $\left(A^T A, A^T A + B^T B - C^T C\right)$, at most $\operatorname{rank}(\chi) \leq k + r$ generalized determined eigenvalues may be different than 1.*

- *If* 1 *is not a generalized eigenvalue in the pencil* $\left(B^TB, C^TC\right)$*, and if* $A^TA + B^TB - C^TC$ *is full rank, then: (1) the pencil* $\left(A^TA, A^TA + B^TB + C^TC\right)$ *does not have indeterminate eigenvectors, and (2) the multiplicity of eigenvalue* 1 *is exactly dim-null*$(\chi) \geq n - k - r$*, and (3) the multiplicity of the zero eigenvalue is exactly dim-null*(A)*.*
- *The sum pencil* $\left(A^TA, A^TA + B^TB\right)$ *cannot have an infinite eigenvalue, and all of its eigenvalues are in the interval* $[0, 1]$*.*

Similar results can be obtained for modifying A in other ways, e.g., a set of columns of A. Here is one such result. To state it, denote the columns of A that are not modified by D, and denote the modified columns before and after by E and F, respectively.

Theorem 2 *Let* $D \in \mathbb{R}^{m\times n}$*,* $E \in \mathbb{R}^{m\times k}$*, and* $F \in \mathbb{R}^{m\times k}$*, for some* $1 \leq k < n$*. In addition, let*

$$\begin{aligned} A &= \begin{pmatrix} D & E \end{pmatrix} \in \mathbb{R}^{m\times(n+k)} \\ \hat{A} &= \begin{pmatrix} D & F \end{pmatrix} \in \mathbb{R}^{m\times(n+k)}. \end{aligned}$$

Then in the pencil $\left(A^TA, \hat{A}^T\hat{A}\right)$*, at least* $n - k$ *of the generalized finite eigenvalues are equal to* 1*.*

Similarly,

- If a preconditioner M is effective for a matrix A^TA, then it is also effective for the perturbed matrices $A^TA + B^TB - C^TC$ ad also $\hat{A}^T\hat{A}$.
- If the rank of the matrices B, C, E, and F is low, then most of the generalized eigenvalues of the perturbed preconditioned system will be bounded by the extreme generalized eigenvalues of the unperturbed preconditioned system.
- I.e., the number of runaway eigenvalues is small, but the non-runaway eigenvalues are not necessarily at 1, and they can move in an interval whose size determines the condition number of the original preconditioned system.

Theorem 3 *Let* $A \in \mathbb{R}^{m\times n}$*,* $B \in \mathbb{R}^{k\times n}$*, and* $C \in \mathbb{R}^{r\times n}$*, for some* $1 \leq k + r < n$*. Let* $M \in \mathbb{R}^{n\times n}$ *be SPSD, and assume*

$$\begin{aligned} null(M) &\subseteq null(A^TA) \\ null(M) &\subseteq null(B^TB) \\ null(M) &\subseteq null(C^TC) \end{aligned}$$

Then, if we assume that

$$\alpha \leq \lambda_1\left(A^TA, M\right) \leq \lambda_{rank(M)}\left(A^TA, M\right) \leq \beta,$$

then it follows that

$$\begin{aligned} \alpha &\leq \lambda_{r+1}\left(A^TA + B^TB - C^TC, M\right) \\ &\leq \lambda_{rank(M)-k}\left(A^TA + B^TB - C^TC, M\right) \\ &\leq \beta \end{aligned}$$

Corollary 1 *Let $A \in \mathbb{R}^{n\times d}$, $B \in \mathbb{R}^{k\times k}$, for $1 \leq k \leq d$, be full rank matrices, and let $M \in \mathbb{R}^{n\times n}$ be SPSD. If the eigenvalues of $\left(A^TA, M\right)$ are in the interval (α, β), where α and β are numbers, then the $d-k$ smallest eigenvalues of $\left(A^TA + B^TB, M\right)$ are in the same interval (α, β).*

Similar results can be stated when columns are modified and for other types of perturbations.

13.3 Back to our preconditioned RandNLA LS solver

There are several applications of these ideas for preconditioned LS solvers: drop rows for sparsity; updating (adding rows) and down-dating (drop rows); adding rows to help solve rank deficient problems. We will describe how they are used in the Blendenpik solver; other RandNLA solvers like LSRN do similar things.

A key aspect of implementations is that they project onto (say) $2d$ rows, rather than (say) $10d\log(d)/\epsilon$ rows. In that case, we might loose rank or have other problems; but the theory we just outlined means that we can still obtain a good preconditioner. Here is an outline of how that happens.

What if we project onto (say) $2d$ rows, so that we don't uniforms the leverage scores, i.e., so that there are still a few bad coherence rows? Relatedly, what if the input matrix has only a very few high leverage rows and that we miss them in the random sample? Then, the sample can (sometimes) still lead to a good preconditioner. The basic reason for this is that a few rows with large norm may allow a few singular values of the preconditioned system $A\hat{R}^{-1}$ to be very large, but the number of large singular values is bounded by the number of large rows (and those can be dealt with with a few extra iterations of the iterative method). That is, a few large singular values can cause the condition number of AR^{-1} to be large, and they can even lead to subspace-nonpreservation, leading the worst-case bounds to fail, but they don't much affect the convergence of LSQR.

Here is a basic lemma.

Lemma 1 *Let $A \in \mathbb{R}^{n\times d}$, with $n \geq d$, and suppose that $A = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix}$, where A_2 has $\ell \leq \min\{n-d, d\}$. In addition, let $S \in \mathbb{R}^{k\times(n-\ell)}$ be a matrix such that SA_1 is full rank, and let its QR decomposition be $SA_1 = \hat{Q}\hat{R}$. Then, at least $n-\ell$ of the singular values of $A\hat{R}^{-1}$ are in the interval $\left(\sigma_{min}\left(A_1\hat{R}^{-1}\right), \sigma_{max}\left(A\hat{R}^{-1}\right)\right)$.*

Proof: The singular values $\sigma_i\left(A_1\hat{R}^{-1}\right)$ are the square roots of the generalized eigenvalues of $\left(A_1^TA_1, (SA_1)^TSA_1\right)$; and the singular values $\sigma_i\left(A\hat{R}^{-1}\right)$ are the square roots of the generalized eigenvalues of $\left(A_1^TA_1 + A_2^TA_2, (SA_1)^TSA_1\right)$. The matrix $A^TA = A_1^TA_1 + A_2^TA_2$ is a rank ℓ perturbation of $A_1^TA_1$. So, by the corollary above, we know that at least $d-\ell$ eigenvalues of $\left(A^TA, (SA_1)^TSA_1\right)$ are in the interval between the smallest and largest of the generalized eigenvalues of $\left(A_1^TA_1, (SA_1)^TSA_1\right)$.

$\diamond$

So, what this lemma says is that as long as the number of runaway eigenvalues is small, then we can still use it as a good preconditioner. If we precondition more aggressively, etc., such that we

loose rank, then similar ideas apply if we perturb the matrix to make it full rank. We won't go into the details now, except to say that there are many other variants of this possible, e.g., see the related solver LSRN.