CS 15-859: Algorithms for Big Data Fall 2017

Lecture lecture 4 — 09/08/17

Prof. David Woodruff *Scribe: Hui Han Chin*

High Precision Regession

The *Sketch-and-Solve* approach to solving the regression problem of $\min_x |Ax - b|$ has runtime of the form $nnz(A) + (n+d)poly(d/\epsilon)$. However, there is a dependence on $poly(1/\epsilon)$ error term and this is not desirable as it would take very long to get low error solutions. The goal is to remove this dependence and ideally, get machine precision.

We would show a technique of finding x' for which $|Ax' - b|_2 \in (1+\epsilon)\min_x |Ax - b|_2$ with high probability but in time $nnz(A) + (n+d)poly(d)\log(1/\epsilon)$. The main idea is to sketch and solve to find for a initial crude solution before using gradient descent to improve it. The key insight is that *sketching can improve the condition number* of A.

Definition. Condition number, κ, of a matrix A

$$\kappa(A) = \frac{\sup_{|x|_2=1} |Ax|_2}{\inf_{|x|_2=1} |Ax|_2} = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}$$

The condition number estimate the inaccuracy of the approximate solution. A large condition number of A means that a small change in the target vector b will result a large change in the solution vector x. Thus is it desired to have small condition number, also known as a *well conditioned problem*. Many algorithms' time complexity, such as Precondition Conjugate Gradient, depend on the condition number.

Small QR Decomposition

1. Assume A is invertible. For general A, find the linearly independent columns and proceed.
2. Suppose S is a ϵ_0 subspace embedding for A, compute SA. We use ϵ_0 to denote the initial solution's error factor as oppose to ϵ, which is the final solution error factor.
3. Find R via QR decomposition such that $SA = QR^{-1} \Rightarrow SAR = Q$. QR decomposition can be done in $poly(d)$ time with no dependency on ϵ_0. Also we needed A to be invertible to ensure that R is invertible.
4. Observe that $\kappa(SAR) = \kappa(Q) = 1$ since Q is orthonormal.

Claim 1. $\kappa(AR) = \frac{1+\epsilon_0}{1+\epsilon_0}$
Proof: For all unit x, $|SARx|_2 = |Qx|_2 = 1$ as Q preserves norms. Since S is a subspace embedding, $|SARx|_2 \in (1 \pm \epsilon_0)|ARx|_2$

$$(1-\epsilon_0)|ARx|_2 \leq 1 \leq (1+\epsilon_0)|ARx|_2 \Rightarrow \kappa(AR) = \frac{1+\epsilon_0}{1-\epsilon_0}$$

Gradient Descent to Improve to a Constant Factor Solution

Gradient Descent update rule For a fixed ϵ_0, e.g. $1/2$, solving for $x_0 = \min_x |SARx - Sb|_2$ takes $nnz(A) + poly(d)$. We would find R as well in the same time. Note that x_0 is a constant factor approximation of x^*. Now apply Gradient Descent to improve x_0.

Definition. Update Rule

$$x_{m+1} \leftarrow x_m + R^T A^T (b - ARx_m)$$

Want to show at each update step, x_m improves by a constant factor.
Let the SVD be of AR be $AR = U\Sigma V^T$. Show that the distance to x^* decreases.

$$\begin{aligned}
AR(x_{m+1} - x^*) &= AR(x_m + R^T A^T (b - ARx_m) - x^*) \\
&= AR(x_m + R^T A^T ARx^* - R^T A^T ARx_m) - x^*) \quad \text{By normal equations} \\
&= (AR - ARR^T A^T AR)(x_m - x^*) \\
&= (U\Sigma V^T - (U\Sigma V^T)(V\Sigma U^T)(U\Sigma V^T)(x_m - x^*) \\
&= U(\Sigma - \Sigma^3)V^T(x_m - x^*) \\
|AR(x_{m+1} - x^*)|_2 &= |U(\Sigma - \Sigma^3)V^T(x_m - x^*)|_2 \\
&= O(\epsilon_0)|U\Sigma V^T(x_m - x^*)|_2 \quad \text{as } \kappa(AR) = \frac{1+\epsilon_0}{1-\epsilon_0} \to \sigma(AR) \leq O(1+\epsilon) \\
&= O(\epsilon_0)|AR(x_m - x^*)|_2
\end{aligned}$$

Thus at each round, the solution improves by a factor of ϵ_0. Combining with

$$|ARx_m - b|_2^2 = |AR(x_m - x^*)|_2^2 + |ARx^* - b|_2^2 = O((\epsilon_0)^m)|AR(x_0 - x^*)|_2 + |ARx^* - b|_2^2$$

since the initial solution is constant ϵ_0 approximation and each round the solution improves by factor of ϵ_0, thus the overall dependency on ϵ is $\log(\epsilon^-1)$.

Note this results holds with high probability. If the initial sketching SA failed to produce a good conditioner R, as $\kappa(AR)$ would be big, then this procedure would take much longer but it will arrive at the eventual right estimate x'.

Leverage Score Sampling

The subspace embedding methods so far are oblivious. For example if we use a `CountSketch` matrix S to embed A, if A has sparse rows, then SA has sparse rows too, which is not efficient. There is an alternative subspace method based on sampling the "important" rows of the matrix. This is called *Leverage Score Sampling*. Let $A \in R^{n\times d} = U\Sigma V^T$ be a rank-d matrix.

Definition. Leverage Score $l(i)$ of the ith row of A

$$l(i) = |U_{i,*}|_2^2$$

Observe that if A is orthonormal, then $\sum_i l(i) = d$ since it just reordering the sum of squares of the entries in A. Also note that $l(i)$ is independent of the orthonormal basis U of A. Suppose both U, U' are both othornormal bases of A.

Claim 2. $\forall i, |e_iU|_2^2 = |e_iU'|_2^2$
Proof: Since both U and U' have the same column space and are bases, then there is some change of basis matrix Z such that $U = U'Z$. Since U, U' are orthonormal, then Z is a rotation matrix meaning it has both orthonormal rows and columns. Then, $|e_iU|_2^2 = |e_iU'Z|_2^2 = |e_iU'|_2^2$

Intuitively, the leverage score correspond to the importance of a row to the matrix that we are solving. Sampling according to the leverage score will allow us to pick the important rows. However computing the actual leverage score would require an expansive SVD. The idea is to find another distribution $q(i)$ that approximates the leverage scores $l(i)$. Let $q(i) \geq \frac{\beta l(i)}{d}$, with β as some scaling parameter.

Definition. Approximate Leverage Score Sampling matrix $S = D\Omega^T$, where D is a $k \times k$ diagonal rescaling matrix and Ω is a $n \times k$ sampling matrix.For each column j of Ω, D, independently, and with replacement, pick a row index $i \in [n]$ with probability q_i. Set $\Omega_{i,j} = 1$ and $D_{j,j} = \frac{1}{\sqrt{q_i k}}$

Note that both Ω, D can be computed in $O(nd + n + k\log k)$ time.

Leverage Score gives Subspace Embedding

We want to show $S = D \cdot \Omega^T$ is a $(1 \pm \epsilon)$ embedding for $|A|$. Taking SVD, $A = U\Sigma V^T$, this is equivalent of showing, with high probability,

$$\forall y, |SUy|_2^2 = (1 \pm \epsilon)|Uy|_2^2 = (1 \pm \epsilon)|y|_2^2, \quad \text{or} \quad |U^TS^TSU - I|_2 \leq \epsilon$$

To analyze $U^TS^TSU - I$, we would use Matrix Chernoff Bound.

Definition. Matrix Chernoff Bound. Let $X_1, \cdots, X_k$ be independent copies of a symmetric random matrix $X \in R^{d\times d}$ with $\mathbb{E}[X] = 0, |X|_2 \leq \gamma, |\mathbb{E}\left[X^TX\right]|_2 \leq \sigma^2$. Let $W = \frac{1}{k}\sum_{j\in[k]} X_j$. For any $\epsilon > 0$

$$\mathbf{Pr}\Big[|W|_2 > \epsilon\Big] \leq 2d \cdot \exp(-\frac{k\epsilon^2}{\sigma^2 + \gamma\epsilon/3})$$

Note that since W is symmetric $|W|_2 = \sup_{|x|_2=1} x^TWx$

Do a change of variable to X such that $W = X^T X = U^T S^T S U - I$
Let $i(j)$ denote the index of the row of U sampled in the jth trial.
Let I_d be the $d \times d$ Identity matrix and 0_d be the $d \times d$ all-0 matrix.
Let $X_j = I_d - \frac{U_{i(j)}^T U_{i(j)}}{q_{i(j)}}$ and note that X_j are independent copies of the symmetric matrix random variable. We have the following properties.

$$\begin{aligned}
\mathbb{E}[X_j] &= I_d - \sum_i q_i \left(\frac{U_i^T U_i}{q_i}\right) && i\text{th row sampled with } q_i \\
&= I_d - \sum_i (U_i^T U_i) \\
&= I_d - I_d && U \text{ orthonormal} \\
&= 0_d \\
|X_j|_2 &\leq |I_d|_2 + \frac{|U_{i(j)}^T U_{i(j)}|_2}{q_{i(j)}} && \text{triangle ineq} \\
&\leq 1 + \max_i \frac{|U_i|_2^2}{q_i} \\
&\leq 1 + \frac{d}{\beta} && \text{by def of } q_i
\end{aligned}$$

For 2 matrices A, B, let the *Loewner Order* $A \leq B$ be such that $x^T A x \leq x^T B x, \forall x$.

$$\begin{aligned}
\mathbb{E}\left[X^T X\right] &= I_d - 2\mathbb{E}\left[\frac{U_{i(j)}^T U_{i(j)}}{q_{i(j)}}\right] + \mathbb{E}\left[\frac{(U_{i(j)}^T U_{i(j)})^2}{q_{i(j)}^2}\right] \\
&= 2(I_d - \mathbb{E}\left[\frac{U_{i(j)}^T U_{i(j)}}{q_{i(j)}}\right]) - I_d + \sum_i q_i \frac{(U_{i(j)}^T U_{i(j)})^2}{q_{i(j)}^2} \\
&\leq 2(0_d) - I_d + (\frac{d}{\beta}) \sum_i U_i^T U_i && \text{from } \mathbb{E}[X_j] \\
&\leq (\frac{d}{\beta} - 1) I_d && U \text{ orthonormal}
\end{aligned}$$

Hence we get

$$|\mathbb{E}\left[X^T X\right]|_2 \leq \frac{d}{\beta} - 1$$