15-859 Algorithms for Big Data — Fall 2017

Problem Set 2

Due: 23:59, Saturday, October 14

Please see the following link for collaboration and other homework policies:
`http://www.cs.cmu.edu/afs/cs/user/dwoodruf/www/teaching/15859-fall17/grading.pdf`

Problem 1: Composability of Sketching Matrices (10 points) We saw in problem 2.3 on the first problem set that a random Gaussian matrix S with $O(d/\epsilon)$ rows can be used for solving the regression problem $\min_x \|Ax-b\|_2$ up to a $(1+\epsilon)$-factor with constant probability, by outputting $x = (SA)^{-}Sb$. The only properties needed of S were that with constant probability it (1) provides a $(1 \pm 1/2)$-factor subspace embedding for the column span of A, and (2) if U is an orthonormal basis for the column span of A, then $\|U^T S^T S(b - Ux^*)\|_2^2 = O(\epsilon/d)\|U^T\|_F^2 \|Ux^* - b\|_2^2$, where x^* is the minimizer to $\min_x \|Ux - b\|_2^2$. See the solutions given for problem set 1 if you are unclear on this. It follows that *any* sketching matrix S which satisfies properties (1) and (2) can be used for solving the regression problem. In this problem we will obtain analogous results for other sketching matrices.

(1) (5 points) We would like to obtain the analogous result for $S = P \cdot H \cdot D$ being the Subsampled Randomized Hadamard Transform (the Fast Johnson Lindenstrauss Transform discussed in class). We saw in class that if S has $d \cdot \text{poly}(\log(nd))$ rows then with probability at least 99/100, simultaneously for all x, $\|SAx\|_2 = (1 \pm 1/2)\|Ax\|_2$, and thus S satisfies property (1) above. Show that if S has $d\epsilon^{-1} \cdot \text{poly}(\log(nd))$ rows, then it also satisfies property (2) above with probability at least 99/100.

To solve this problem, it may be useful to express $\|U^T S^T S(b - Ux^*)\|_2^2$ in terms of $V = HDU$ and $z = HD(b - Ux^*)$, compute the expectation of this quantity, and apply Markov's bound. You may want to first condition on a random D (in the PHD expression for S) so that z has a certain nice property that we discussed in class when applying HD to vectors. Then you can think of P as uniformly sampling entries and trying to approximate the inner products between rows of V^T and z.

(2) (1 point) State why, if we instead choose a random CountSketch matrix T with $O(d^2 + d/\epsilon)$ rows, then it satisfies properties (1) and (2), each with probability at least 99/100.

(3) (4 points) Show that if we choose a random CountSketch matrix $T \in \mathbb{R}^{t \times n}$ with $t = O(d^2 + d/\epsilon)$ rows, and an independent Subsampled Randomized Hadamard Transform $S \in \mathbb{R}^{s \times t}$ with $s = d\epsilon^{-1} \cdot \text{poly}(\log(d/\epsilon))$ rows, then if $x' = (STA)^{-}STb$, it holds that $\|Ax' - b\|_2 \leq (1+\epsilon)\min_x \|Ax - b\|_2$ with probability at least 24/25. What is the overall time complexity for computing x'?

For this problem, you can assume a generalization of the first part of the problem, namely, that S satisfies that if S has $d\epsilon^{-1} \cdot \text{poly}(\log(nd))$ rows, then it also satsisfies the following with probability at least 99/100: for any fixed matrix A and vector b,

$$\|A^T S^T Sb - A^T b\|_2^2 \leq \frac{\epsilon}{d}\|A\|_F^2\|b\|_2^2.$$

You may also find it helpful to use the fact that we showed in class that for any fixed vector y, $\|Ty\|_2 = (1 \pm 1/2)\|y\|_2$ with probability at least 99/100.

Problem 2: Linear Dependence on ϵ for Low Rank Approximation (25 points) In class we saw how, given an $n \times d$ matrix A, we can output a rank-k matrix B for which $\|A-B\|_F^2 \leq (1+\epsilon)\|A-A_k\|_F^2$ with probability at least 9/10, in $\text{nnz}(A)+(n+d)\cdot\text{poly}(k/\epsilon)$ time. Here A_k is the best rank-k approximation to A. Our proof was based on affine embeddings, which generalize subspace embeddings. In this problem we will give a somewhat different proof and improve the time complexity to $\text{nnz}(A) + \tilde{O}((n+d)k^2/\epsilon) + \text{poly}(k/\epsilon)$, where for a function f, $\tilde{O}(f)$ denotes $f \cdot \text{poly}(\log(f))$.

(1) (2 points) In problem 1 above, we saw that for the sketching matrix ST, if $x' = (STA)^- STb$, then $\|Ax' - b\|_2 \leq (1+\epsilon)\min_x \|Ax - b\|_2$ with probability at least 24/25. Suppose we instead had the problem $\min_x \|AX - B\|_F$, where X is a $d \times z$ matrix of unknowns, A is a given $n \times d$ matrix, and B is a given $n \times z$ matrix. Explain why, if S and T are the same matrices and of the same dimensions as before, we have that $X' = (STA)^- STB$ satisfies $\|AX' - B\|_F^2 \leq (1+\epsilon)\min_X \|AX - B\|_F^2$ with probability at least 24/25. You don't need to give a formal proof of this since it is almost the same as before. Just state the differences you need to properties (1) and (2) above, if any, and explain briefly why S and T still satisfy these properties.

(2) (3 points) Suppose $T \in \mathbb{R}^{t \times n}$ is a random CountSketch matrix with $t = O(k^2 + k/\epsilon)$ rows, and $S \in \mathbb{R}^{s \times t}$ is an independent Subsampled Randomized Hadamard Transform with $s = k\epsilon^{-1}\text{poly}(\log(k/\epsilon))$ rows. Show that with probability at least 24/25, the row span of $S \cdot T \cdot A$ contains a $(1+\epsilon)$-approximate rank-k approximation to A, namely, that $\min_{\text{rank-}k\ X \in \mathbb{R}^{n \times s}} \|XSTA - A\|_F^2 \leq (1+\epsilon)\|A - A_k\|_F^2$.

To solve this problem, it may be helpful to consider a similar hypothetical question $\min_X \|A_k X - A\|_F^2$ that we considered in class.

(3) (10 points) Suppose $T' \in \mathbb{R}^{d \times t}$ is a random CountSketch matrix and $S' \in \mathbb{R}^{t \times s}$ is a Subsampled Randomized Hadamard Transform. Suppose further that S', T', S, and T are independent, where S and T are as above. Show that with probability at least 9/10, if X' is the solution to

$$\min_{\text{rank-}k\ X} \|AT'S'XSTA - A\|_F^2,$$

then $\|AT'S'X'STA - A\|_F^2 \leq (1+O(\epsilon))\|A - A_k\|_F^2$. Partial credit will be given if you instead show this for T' and S' with a larger $\text{poly}(k/\epsilon)$ number of columns.

To solve this problem, it may be helpful to consider the hypothetical regression problem used in solving the previous part twice, once on the left and once on the right.

(4) (10 points) Assuming part (3) and given only A, show how to find X' with probability at least $4/5$ in $\text{nnz}(A)+\text{poly}(k/\epsilon)$ time. Also, show how to write $AT'S'X'STA$ as $L \cdot R$, for some matrices $L \in \mathbb{R}^{n\times k}$ and $R \in \mathbb{R}^{k\times d}$, in $\text{nnz}(A) + \tilde{O}((n+d)k^2/\epsilon) + \text{poly}(k/\epsilon)$ time.

For this part of the problem, you will need to use the following: let $A \in \mathbb{R}^{m\times n}$, $C \in \mathbb{R}^{m\times c}$, and $D \in \mathbb{R}^{d\times n}$, and $k \leq \min(c,d)$ be an integer. Consider the following optimization problem:

$$X' = \text{argmin}_{X,\text{rank}(X)\leq k} \|A - CXD\|_F^2.$$

Then one possible solution is given by $X' = C^-[UU^TABB^T]_kD^-$, where $C = U\Sigma V^T$ and $D = EZB^T$ are the singular value decompositions of C and D, respectively.

To solve the problem, think about using affine embeddings to try to make the problem smaller.

Problem 3: Spectral Norm Low Rank Approximation (15 points) In this problem we will study *spectral low rank approximation*, that is, we want to find a rank-k matrix B for which $\|A - B\|_2 \leq (1+\epsilon)\|A - A_k\|_2$, where $A_k = \min_{\text{rank }-k\ B} \|A - B\|_2$, and $\|A - A_k\|_2 = \sigma_{k+1}(A)$, where $\sigma_{k+1}(A)$ is the $(k+1)$-st singular value of A.

For low rank approximation of a given $n \times d$ matrix A with respect to the Frobenius norm, a key idea of the algorithm we saw in class was to first compute SA, where S is a random $\text{poly}(k/\epsilon)\times n$ matrix from an appropriate family of matrices (Gaussian, Fast Johnson Lindenstrauss, CountSketch), and then find the best rank-k approximation to A inside of the row span of SA. By the Pythagorean theorem, the best rank-k approximation to A inside of the row span of SA is given by first projecting the rows of A onto the row span of SA, and then computing the SVD of the projected points. Formally, this best rank-k approximation is given by $[AP_{SA}]_k$, where for a matrix B, $[B]_k$ denotes its best rank-k approximation given by the SVD, and $P_{SA} = (SA)^-SA$ is the projection matrix onto the row span of SA.

(1) (6 points) Let $\tilde{A}$ be the best spectral rank-k approximation to a matrix A in the row span of another matrix B, that is, $\tilde{A} = YB$, where $Y = \text{argmin}_{\text{rank-k } Y'} \|Y'B - A\|_2$.

Show that for any matrices A and B, we have $\|A - [AP_B]_k\|_2^2 \leq 2\|A - \tilde{A}\|_2^2$, where P_B is the projection onto the row span of B.

To solve this problem, try to think about $\|xA - x[AP_B]_k\|_2^2$ geometrially for an arbitrary unit vector x, and argue why you can express it as $\|xA - xAP_B\|_2^2 + \|xAP_B - x[AP_B]_k\|_2^2$. Then try to upper bound each of these terms by $\|A - \tilde{A}\|_2^2$.

Unfortunately, it turns out that the bound in part (1) can be tight, and this rules out an algorithm for spectral low rank approximation which, analogous to our algorithm for Frobenius low rank approximation, finds a matrix SA whose span contains a good rank-k approximation with respect to the spectral norm, and then approximately computes $[AP_{SA}]_k$. Further, it turns out that such matrices SA with S having $\text{poly}(k/\epsilon)$ rows and containing a $(1+\epsilon)$-approximate spectral rank-k low rank approximation, do not exist, if S is a linear sketch oblivious to A. We will not prove these statements here and you can think of them as side remarks.

Instead, what is often used for spectral low rank approximation is a variant of the power method which works as follows. Assume for simplicity that A is an $n \times n$ symmetric matrix. Let G be an $n \times k$ matrix of i.i.d. Gaussian random variables. Let $r = O((\log(n))/\epsilon)$. Compute BG, where $B = A^r$. Then BG is an $n \times k$ matrix. Let P be the projection onto the column span of BG. Output PA.

Using a convexity argument that you do not need to prove, one can show that $\|A - PA\|_2 \leq \|B - PB\|_2^{1/r}$, which you can assume in the following.

(2) (3 points) Show that if $\|B - PB\|_2^2 \leq \|B - B_k\|_2^2 \cdot \text{poly}(n)$, for a $\text{poly}(n)$ factor that does not depend on r, then for an appropriate $r = O(\epsilon^{-1} \log n)$ it holds that $\|A - PA\|_2 \leq (1+\epsilon)\|A - A_k\|_2$.

To solve this problem, try to think of how the SVD of B is related to that of A.

Given part (2), we can focus on matrix B, and our goal is just to show that $\|B - PB\|_2^2 \leq \|B - B_k\|_2^2 \cdot \text{poly}(n)$. Suppose $B = U\Sigma V^T$ is written in its singular value decomposition. Let V_k^T denote the top k rows of V^T (corresponding to the top k singular values) and V_{n-k}^T denote the bottom $n-k$ rows of V^T. We are going to use the following powerful statement, which says that for any $n \times k$ matrix S for which $\text{rank}(V_k^T S) = k$, if P is the projection onto the columns of BS, then

$$\|B - PB\|_2^2 \leq \|B - B_k\|_2^2 + \|(B - B_k)S(V_k^T S)^{-}\|_2^2. \tag{1}$$

We will not prove (1), but for intuition, suppose that $S = V_k$, from which it follows that $P = U_k U_k^T$, where U_k consists of the top k left singular vectors of B. Then the left hand side of (1) is $\|B - B_k\|_2^2$. Also, the term $\|(B - B_k)S(V_k^T S)^{-}\|_2^2$ vanishes since $(B - B_k)V_k = 0$. So the $\|(B - B_k)S(V_k^T S)^{-}\|_2^2$ term measures "how far" P is to $U_k U_k^T$.

Using (1) in our context, we have that $\|B - PB\|_2^2 \leq \|B - B_k\|_2^2 + \|(B - B_k)G(V_k^T G)^{-}\|_2^2$.

(3) (3 points) Show that $\|B - PB\|_2^2 \leq \|B - B_k\|_2^2(1 + \|V_{n-k}^T G\|_2^2 \|(V_k^T G)^{-}\|_2^2)$.

For this problem, the sub-multiplicativity of the operator norm, namely, that $\|A \cdot B \cdot C\|_2 \leq \|A\|_2 \cdot \|B\|_2 \cdot \|C\|_2$ may be helpful.

(4) (3 points) Argue with probability $9/10$ over G, we have $\|V_{n-k}^T G\|_2^2 = O(n)$ and $\|(V_k^T G)^{-}\|_2^2 = O(k)$. Conclude that if this happens then $\|A - PA\|_2 \leq (1+\epsilon)\|A - A_k\|_2$.

You may use several facts about Gaussian matrices which you do not need to prove: for an $r \times s$ matrix G of i.i.d. $N(0,1)$ random variables:

- $\|G\|_2 \leq C\sqrt{\max(r,s)}$ with probability at least $99/100$ for an appropriate constant $C > 0$, and
- if $r = s$, then $\sigma_r(G) \geq C'/\sqrt{r}$ with probability at least $99/100$, for an appropriate constant $C' > 0$.