15-859 Algorithms for Big Data — Fall 2017
Problem Set 2 Solutions

Problem 1: Composability of Sketching Matrices

(1) Let $y = b - Ux^*$, and $z = HDy$. In class we showed that for any fixed vector y, we have that $\|HDy\|_\infty = O(\sqrt{\log(nd)}\|y\|_2/\sqrt{n})$ with probability at least $1-1/200$, and so since $\|z\|_2 = \|y\|_2$, this implies $\|z\|_\infty = O(\sqrt{\log(nd)}\|z\|_2/\sqrt{n})$. We condition on this event in what follows, and let D be any fixed diagonal matrix for which this event holds. Let $V = HDU$. Since $U^Ty = 0$, we have $V^Tz = 0$. Also since U has orthonormal columns, and since H and D are orthonormal, $\|V^T\|_F^2 = \|U^T\|_F^2$, and so it suffices to show that with probability $1 - 1/100$ over the choice of P, $\|V^TP^TPz\|_2^2 = O(\epsilon/d)\|V^T\|_F^2\|z\|_2^2$.

Let $s = d\epsilon^{-1}\text{poly}(\log(nd))$ be the number of rows of P, and let R be the multi-set of s sampled indices in $[n]$, chosen by P. The i-th coordinate of V^TP^TPz is equal to $(n/s)\sum_{j\in R} v^i_j \cdot z_j$, where v^i is the i-th row of V^T. Since $\langle v^i, z\rangle = 0$, we have that $\mathbf{E}[(V^TP^TPz)_i] = 0$. Thus, $\mathbf{E}[(V^TP^TPz)_i^2] = \mathbf{Var}[(V^TP^TPz)_i]$ and since coordinates $j \neq j' \in R$ are independent, $\mathbf{Var}[(V^TP^TPz)_i] = (n^2/s^2)\cdot s \cdot \mathbf{Var}[v^i_j \cdot z_j]$, where j is a uniformly random index. Note also that $\mathbf{E}[v^i_j \cdot z_j] = 0$ and so $\mathbf{Var}[v^i_j \cdot z_j] = \mathbf{E}[(v^i_j \cdot z_j)^2]$.

We thus have,

$$\begin{aligned}
\mathbf{E}[(v^i_j \cdot z_j)^2] &= \sum_{k=1}^n (1/n)\cdot (v^i_k)^2 z_k^2 \\
&\leq (1/n)\|v^i\|_2^2\|z\|_\infty^2 \\
&\leq (1/n)\|v^i\|_2^2 O((\log(nd))\|z\|_2^2)/n \\
&= (O(\log(nd))/n^2)\|v^i\|_2^2\|z\|_2^2,
\end{aligned}$$

and so $\mathbf{E}[(V^TP^TPz)_i^2] \leq (O(\log(nd))/s)\|v^i\|_2^2\|z\|_2^2$. Consequently, $\mathbf{E}[\|V^TP^TPz\|_2^2] = O((\log(nd))/s)\|V^T\|_F^2\|z\|_2^2$. It follows for appropriate $s = O(d\epsilon^{-1}\log(nd))$, by a Markov bound we have that with probability at least $1-1/200$, $\|V^TP^TPz\|_2^2 = O(\epsilon/d)\|V^T\|_F^2\|z\|_2^2$.

(2) We showed in class that for a random CountSketch matrix T with $O(d^2)$ rows, it satisfies the first property above with probability 99/100. Along the way to showing this, we also showed that T satisfies the approximate matrix product property in class, where recall that if T has $O(d/\epsilon)$ rows, then the approximate matrix product property we showed is that with probability 99/100, $\|U^TS^TS(b-Ux^*)\|_2^2 = O(\epsilon/d)\|U^T\|_F^2\|Ux^*-b\|_2^2$.

(3) By the above, we just need to show that $S\cdot T$ satisfies properties (1) and (2). For property (1), notice that if $\|TAx\|_2 = (1\pm 1/10)\|Ax\|_2$ for all x and $\|STAx\|_2 = (1\pm 1/10)\|TAx\|_2$ for all x, then $\|STAx\|_2 = (1\pm 1/10)^2\|Ax\|_2 = (1\pm 1/2)\|Ax\|_2$ for all x, as desired.

For property (2), we have that since S satisfies the generalization of property (2) given in the problem statement, that

$$\|U^T T^T S^T ST(b - Ux^*) - U^T T^T T(b - Ux^*)\|_2 \leq \frac{\sqrt{\epsilon}}{\sqrt{d}} \|U^T T^T\|_F \|T(b - Ux^*)\|_2.$$

Consequently, by the triangle inequality,

$$\|U^T T^T S^T ST(b - Ux^*)\|_2 \leq \|U^T T^T T(b - Ux^*)\|_2 + \frac{\sqrt{\epsilon}}{\sqrt{d}} \|U^T T^T\|_F \|T(b - Ux^*)\|_2.$$

Since T is a subspace embedding for U, we have $\|U^T T^T\|_F = \|TU\|_F \leq (1+1/2)\|U\|_F$. Also, since $\|Ty\|_2 \leq (1+1/2)\|y\|_2$ for a fixed vector y with probability 99/100, we have $\|T(b - Ux^*)\|_2 \leq (1+1/2)\|b - Ux^*\|_2$. Finally, since T satisfies property (2), $\|U^T T^T T(b - Ux^*)\|_2 \leq \frac{\sqrt{\epsilon}}{\sqrt{d}} \|U^T\|_F \|b - Ux^*\|_2$. Putting these statements together and plugging into (1), we have that with probability at least 24/25, $\|U^T T^T S^T ST(b - Ux^*)\|_2^2 = O(\frac{\sqrt{\epsilon}}{\sqrt{d}} \cdot \|U^T\|_F^2 \|b - Ux^*\|_2^2$, as desired.

Problem 2: Linear Dependence on ϵ for Low Rank Approximation

(1) We still need property (1), that S is a $(1 \pm 1/2)$-subspace embedding for the column span of A. The second property slightly changes in that now we need that if U is an orthonormal basis for the column span of A, then $\|U^T S^T S(B - UX^*)\|_F^2 = O(\epsilon/d)\|U^T\|_F^2 \|UX^* - B\|_F^2$, where X^* is the minimizer to $\min_X \|UX - B\|_F^2$. The rest of the proof, as in the solutions for problem set 1, goes through.

(2) This is a similar argument to the one given in class. We consider the hypothetical regression problem $\min_X \|A_k X - A\|_F^2$. By the previous part, letting U be an $n \times k$ orthonormal basis for the column span of A_k, we have that if ST is a subspace embedding for the column span of A_k, and if it satisfies $\|U^T S^T S(A - A_k)\|_F^2 = O(\epsilon/d)\|U^T\|_F^2 \|A - A_k\|_F^2$, then the minimizer X' to $\min_X \|STA_k X - STA\|_F^2$ satisfies $\|A_k X' - A\|_F^2 \leq (1+\epsilon)\|A - A_k\|_F^2$. Note that here, in the notation of the previous part, $\|UX^* - B\|_F^2 = \|A_k - A\|_F^2$, since UX^* is of rank k, and the optimal rank-k approximation to A is A_k. Importantly though, the minimizer X' can be written as $(STA_k)^- STA$, and so is a $(1+\epsilon)$-approximate rank-k approximation to A in the row span of $S \cdot T \cdot A$.

(3) By the previous part, we know that $\min_{\text{rank}-k X} \|XSTA - A\|_F^2 \leq (1+\epsilon)\|A - A_k\|_F^2$. In particular, there exists a rank-k matrix X' for which $\|X'STA - A\|_F^2 \leq (1+\epsilon)\|A - A_k\|_F^2$. Suppose we write $X' = YC$, where Y is $n \times k$ and C is $k \times s$. Now consider the problem $\min_Y \|YCSTA - A\|_F^2$. Suppose we apply a sketch $T'S'$ to the right of this problem, obtaining the problem $\min_Y \|YCSTAT'S' - AT'S'\|_F^2$. Then since $CSTA$ has rank k, we again have that $S'T'$ has the two properties (1) and (2) of the first part of this problem, and therefore if Y' is the minimizer to $\min_Y \|YCSTAT'S' - AT'S'\|_F^2$, then

$\|Y'CSTA - A\|_F^2 \le (1+\epsilon)\min_Y \|YCSTA - A\|_F^2 \le (1+O(\epsilon))\|A - A_k\|_F^2$. Importantly though $Y' = AT'S'(CSTAT'S')^-$, and is therefore a rank-k matrix in the column span of $AT'S'$. Thus, $AT'S'(CSTAT'S)^-CSTA$ is a rank-k matrix in the column span of $AT'S'$ and in the row span of STA providing a $(1+O(\epsilon))$-approximate rank-k approximation. It follows that if X' is the solution to $\min_{\text{rank}-k\ X} \|AT'S'XSTA - A\|_F^2$, then $\|AT'S'X'STA - A\|_F^2 \le (1+O(\epsilon))\|A - A_k\|_F^2$.

(4) Given the previous part, we just need to solve the optimization problem

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

We can apply the technique of affine embeddings that we saw in class. Namely, suppose we choose two CountSketch matrices R_1 and R_2, where R_1 has $\text{poly}(k/\epsilon)$ rows and R_2 has $\text{poly}(k/\epsilon)$ columns. Then with arbitrarily large constant probability, for all X, $\|R_1AT'S'XSTA - R_1A\|_F^2 = (1 \pm \epsilon)\|AT'S'XSTA - A\|_F^2$. Also, for all X, $\|R_1AT'S'XSTAR_2 - R_1AR_2\|_F^2 = (1 \pm \epsilon)\|R_1AT'S'XSTA - R_1A\|_F^2 = (1 \pm O(\epsilon))\|AT'S'XSTA - A\|_F^2$. We can compute R_1A in $O(\text{nnz}(A))$ time. Noting that $\text{nnz}(R_1A) \le \text{nnz}(A)$, we can compute R_1AR_2 in $O(\text{nnz}(A))$ time as well. We can also compute TAR_2 in $O(\text{nnz}(A))$ time and R_1AT' in $O(\text{nnz}(A))$ time. The remaining products $R_1AT'S'$ and $STAR_1$ can each be computed in $\text{poly}(k/\epsilon)$ time. At this point the optimal rank-k X' is given by the formula in the hint: let $(R_1AT'S') = U\Sigma V^T$ be its SVD, and let $STAR_2 = AZB^T$ be its SVD. Then

$$X' = (R_2AT'S')^-(UU^T(R_1AR_2)BB^T)_k(STAR_2)^-.$$

Note that all operations are on low dimensional matrices and can be performed in $\text{poly}(k/\epsilon)$ time, giving a total of $\text{nnz}(A) + \text{poly}(k/\epsilon)$ time for this part of the problem.

We can compute AT' in $\text{nnz}(A)$ time. This matrix is $n \times O(k^2 + k/\epsilon)$. We can then compute $AT'S'$ in $\tilde{O}(n(k^2 + k/\epsilon))$ time. This matrix is $n \times \tilde{O}(k/\epsilon)$. Similarly, we can compute STA in $\text{nnz}(A) + \tilde{O}(d(k^2 + k/\epsilon))$ time. This matrix is $\tilde{O}(k/\epsilon) \times d$. We can compute the SVD of X', denoted by $U\Sigma V^T$, in $\text{poly}(k/\epsilon)$ time, where $U\Sigma$ is $\tilde{O}(k/\epsilon) \times k$, and V^T is $k \times \tilde{O}(k/\epsilon)$. We can compute $L = AT'S'(U\Sigma)$ in $\tilde{O}(nk^2/\epsilon)$ time and similarly compute $R = V^TSTA$ in $\tilde{O}(dk^2/\epsilon)$ time. In total, we can output L and R in $\text{nnz}(A) + \tilde{O}((n+d)k^2/\epsilon) + \text{poly}(k/\epsilon)$ time.

Problem 3: Spectral Norm Low Rank Approximation

(1) Consider $\|xA - x[AP_B]_k|_2^2$ for an arbitrary unit vector x. We thus have,

$$\begin{aligned}
\|xA - x[AP_B]_k\|_2^2 &= \|xA - xAP_B\|_2^2 + \|xAP_B - x[AP_B]_k\|_2^2 \\
&\leq \|xA - x\tilde{A}\|_2^2 + \|xAP_B - x[AP_B]_k\|_2^2 \\
&\leq \|A - \tilde{A}\|_2^2 + \|AP_B - [AP_B]_k\|_2^2 \\
&\leq \|A - \tilde{A}\|_2^2 + \|AP_B - \tilde{A}\|_2^2 \\
&= \|A - \tilde{A}\|_2^2 + \|(A - \tilde{A})P_B\|_2^2 \\
&\leq \|A - \tilde{A}\|_2^2 + \|A - \tilde{A}\|_2^2 \\
&= 2\|A - \tilde{A}\|_2^2,
\end{aligned}$$

where the first equality is the Pythagorean theorem, the first inequality holds since xAP_B is the closest point to xA in the row span of B while $x\tilde{A}$ is just an arbitrary point in the row span of B, the second inequality just uses the definition of the operator norm being the supremum over all unit vectors x, the third inequality uses that $[AP_B]_k$ is the best rank-k approximation to AP_B given by the SVD whereas $\tilde{A}$ is just an arbitrary rank-k matrix, the second equality holds since $\tilde{A}P_B = \tilde{A}$ since the rows of $\tilde{A}$ are already in the row span of B, and the last inequality holds since projections cannot increase spectral norm.

(2) We note that $B = A^r$, and so $\|B - B_k\|_2 = \sigma_{k+1}(A)^r$. Given that we have $\|A - PA\|_2 \leq \|B - PB\|_2^{1/r}$, we have

$$\begin{aligned}
\|A - PA\|_2 &\leq \|B - PB\|_2^{1/r} \\
&\leq \|B - B_k\|_2^{1/r}(\mathrm{poly}(n))^{1/(2r)} \\
&= \|B - B_k\|_2^{1/r} 2^{\Theta(\log n)/r} \\
&= (1 + \epsilon)\|A - A_k\|_2,
\end{aligned}$$

where we have used that $2^{\Theta(\log n)/r} \leq (1 + \epsilon)$ for an appropriate $r = O((\log n)/\epsilon)$.

(3) We are given that $\|B - PB\|_2^2 \leq \|B - B_k\|_2^2 + \|(B - B_k)G(V_k^T G)^-\|_2^2$. Suppose we write $B - B_k = U_{n-k}\Sigma_{n-k}V_{n-k}^T$ in its SVD. Then we have using sub-multiplicativity of the operator norm,

$$\begin{aligned}
\|(B - B_k)G(V_k^T G)^-\|_2^2 &= \|U_{n-k}\Sigma_{n-k}V_{n-k}^T G(V_k^T G)^-\|_2^2 \\
&\leq \|U_{n-k}\Sigma_{n-k}\|_2^2 \|V_{n-k}^T G\|_2^2 \|(V_k^T G)^-\|_2^2 \\
&= \|B - B_k\|_2^2 \|V_{n-k}^T G\|_2^2 \|(V_k^T G)^-\|_2^2,
\end{aligned}$$

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

(4) We showed in class that the Gaussian distribution is rotationally invariant, and therefore $V_{n-k}^T G$ is an $n - k \times k$ matrix of i.i.d. normal random variables, and so using the given fact, we have $\|V_{n-k}^T G\|_2^2 = O(n)$ with probability at least $99/100$. Similarly, $V_k^T G$

is a $k \times k$ matrix of i.i.d. normal random variables, and so using the given fact, we have $\sigma_k(V_k^T G)^2 \geq (C')^2/k$ with probability at least $99/100$, for a constant $C' > 0$. Consequently, $\|(V_k^T G)^-\|_2^2 = \frac{1}{\sigma_k(V_k^T G)^2} = O(k)$ with this probability. Hence, with probability at least $9/10$, we have $\|B - PB\|_2^2 \leq \|B - B_k\|_2^2(1 + O(nk))$. Using the second part of this problem, we conclude that with this probability $\|A - PA\|_2 \leq (1+\epsilon)\|A - A_k\|_2$.