15-859 Algorithms for Big Data — Fall 2017
Problem Set 1 Solutions

Problem 1: High Probability Matrix Product and Embeddings

(1) Let $[\ell]$ denote the set $\{1, 2, 3, \ldots, \ell\}$. For each $i \in [\ell]$, we compute

$$s_i = \text{median}_{j\in[\ell]} \|A(S^i)(S^i)^T B - A(S^j)(S^j)^T B\|_F.$$

We output the index i^* whose value s_{i^*} is the smallest. We need to show

$$\Pr[\|A(S^{i^*})(S^{i^*})^T B - AB\|_F > \epsilon\|A\|_F\|B\|_F] \leq \delta.$$

By Chernoff bounds, for an appropriate $\ell = \Theta(\log(1/\delta))$ and $r = \Theta(1/\epsilon^2)$, with probability at least $1-\delta$, there is a subset $T \subseteq [\ell]$ of size at least $\frac{3\ell}{5}$ for which for all $i \in T$, $\|A(S^i)(S^i)^T B - AB\|_F \leq (\epsilon/3)\|A\|_F\|B\|_F$. We call this event $\mathcal{E}$, and condition on it occurring. For any $i, j \in T$, by the triangle inequality,

$$\begin{aligned}\|A(S^i)(S^i)^T B - A(S^j)(S^j)^T B\|_F &\leq \|A(S^i)(S^i)^T B - AB\|_F + \|AB - A(S^j)(S^j)^T B\|_F \\ &\leq (2\epsilon/3)\|A\|_F\|B\|_F.\end{aligned}$$

Since $|T| > \ell/2$, and we take the median value when forming s_i and s_j, we have $s_i, s_j \leq (2\epsilon/3)\|A\|_F\|B\|_F$ and so $s_{i^*} \leq (2\epsilon/3)\|A\|_F\|B\|_F$. Since we take a median value to form s_{i^*} and $|T| > \ell/2$, there exists a $j \in T$ for which

$$\|A(S^{i^*})(S^{i^*})^T B - A(S^j)(S^j)^T B\|_F \leq s_{i^*} \leq (2\epsilon/3)\|A\|_F\|B\|_F.$$

Hence, for this $j \in T$, by the triangle inequality,

$$\begin{aligned}\|A(S^{i^*})(S^{i^*})^T B - AB\|_F &\leq \|A(S^{i^*})(S^{i^*})^T - A(S^j)(S^j)^T B\|_F + \|A(S^j)(S^j)^T B - AB\|_F \\ &\leq \frac{2\epsilon}{3}\|A\|_F\|B\|_F + \frac{\epsilon}{3}\|A\|_F\|B\|_F \\ &\leq \epsilon\|A\|_F\|B\|_F.\end{aligned}$$

The only event we conditioned on was $\mathcal{E}$, so this holds with probability at least $1-\delta$.

(2) Given an $i \in [\ell]$ for which $\text{rank}(S^i A) = d$, we first show how to test for another $j \in [\ell]$ if $\|S^i Ax\|_2 = (1 \pm \varepsilon)\|S^j Ax\|_2$ for all x. $S^i A = U^i \Sigma^i (V^i)^T$ in its singular value decomposition (SVD), the condition that $\|S^i Ax\|_2 = (1 \pm \varepsilon)^2\|S^j Ax\|_2$ for all x is equivalent to the condition that $\|\Sigma^i(V^i)^T x\|_2 = (1 \pm \varepsilon)^2\|\Sigma^j(V^j)^T x\|_2$ for all x. Since $S^i A$ has rank d, $\Sigma^i(V^i)^T$ is an invertible $d \times d$ matrix, and so we may make the change of variables $y = \Sigma^i(V^i)^T x$, and so this condition is equivalent to $\|y\|_2 = (1 \pm \varepsilon)^2\|\Sigma^j(V^j)^T V^i(\Sigma^i)^{-1} y\|_2$ for all y. The latter condition is equivalent to all singular values of $\Sigma^j(V^j)^T V^i(\Sigma^i)^{-1}$ being in the range $[(1-\varepsilon)^2, (1+\varepsilon)^2]$. Thus, by this chain

of equivalences, we have that $\|S^iAx\|_2 = (1 \pm \varepsilon)^2 \|S^jAx\|_2$ if and only if all singular values of $\Sigma^j (V^j)^T V^i (\Sigma^i)^{-1}$ are in the range $[(1-\varepsilon)^2, (1+\varepsilon)^2]$.

Our algorithm simply outputs any $i \in [\ell]$ for which there are at least $\frac{3\ell}{5}$ indices $j \in [\ell]$ for which $\|S^iAx\|_2 = (1 \pm \varepsilon)^2 \|S^jAx\|_2$ for all x, using the procedure above. If there is no such $i \in [\ell]$, we output FAIL. Let $\mathcal{E}$ be the event that there is a set $T \subseteq [\ell]$ of size at least $\frac{3\ell}{5}$ for which for all $i \in T$, $\|S^iAx\|_2 = (1 \pm \epsilon)\|Ax\|_2$ simultaneously for all $x \in \mathbb{R}^d$. By Chernoff bounds, $\Pr[\mathcal{E}] \geq 1 - \delta$, and we condition on $\mathcal{E}$ occurring. Note that conditioned on $\mathcal{E}$, we will not output FAIL, since any $i \in T$ satisfies $\text{rank}(S^iA) = d$ and that there are at least $\frac{3\ell}{5}$ indices $j \in [\ell]$ for which $\|S^iAx\|_2 = (1 \pm \varepsilon)^2 \|S^jAx\|_2$ for all x, so the procedure in the previous paragraph finds all such j. On the other hand, for any $i \in [\ell]$ for which there are at least $\frac{3\ell}{5}$ indices $j \in [\ell]$ for which $\|S^iAx\|_2 = (1 \pm \varepsilon)^2 \|S^jAx\|_2$ for all x, by the pigeonhole principle there is a $j \in T$ for which $\|S^iAx\|_2 = (1 \pm \varepsilon)^2 \|S^jAx\|_2$ for all x, and since $\|S^jAx\|_2 = (1 \pm \varepsilon)\|Ax\|_2$ for all x, we have $\|S^iAx\|_2 = (1 \pm \varepsilon)^3 \|Ax\|_2$ for all x, and so $\|S^iAx\|_2 = (1 \pm \Theta(\varepsilon))\|Ax\|_2$ for all x, as needed. Since the only event we conditioned on was $\mathcal{E}$, which occurs with probability at least $1 - \delta$, our output is successful with probability at least $1 - \delta$.

Problem 2: Linear Dependence on ϵ in Regression

(1) Since U is an orthonormal basis for the column span of A, we can write $y' = Ux$ for some $x \in \mathbb{R}^r$. Consequently, $\|SUx' - Sb\|_2 \leq \|SAy' - Sb\|_2$. We can also write $x' = Ay$ for some $y \in \mathbb{R}^d$ since U and A have the same column span, so $\|SAy' - Sb\|_2 \leq \|SUx' - Sb\|_2$, and so $\|SU'x - Sb\|_2 = \|SAy' - Sb\|_2$. A similar argument shows that $\min_x \|Ux - b\|_2 = \min_y \|Ay - b\|_2$. It now follows that if $\|Ux' - b\|_2 \leq (1+\epsilon)\min_x \|Ux - b\|_2$, then

$$\|Ay' - b\|_2 = \|Ux' - b\|_2 \leq (1+\epsilon)\min_x \|Ux - b\|_2 = (1+\epsilon)\min_y \|Ay - b\|_2.$$

(2) By the Pythagorean theorem, $\|Ux' - b\|_2^2 = \|UU^Tb - b\|_2^2 + \|Ux' - UU^Tb\|_2^2$, that is, the squared distance from b to a vector Ux' in the column span of U is the sum of the squared distance of b to its projection onto the column span of U and the squared distance of its projection to Ux'. We also know that $x^* = U^Tb$ by the normal equations for regression. Plugging this expression in for x^* completes the proof.

(3) We have $x' = (SU)^- Sb$ and since S is an $O(1)$-approximate subspace embedding for the column span of U, which has linearly independent columns, we have that SU has linearly independent columns. So, $(SU)^- = ((SU)^T SU)^{-1}(SU)^T = (U^TS^TSU)^{-1}U^TS^T$ and $x' = (U^TS^TSU)^{-1}U^TS^TSb$. We also have $x^* = U^Tb$. So,

$$\begin{aligned} \|U(x' - x^*)\|_2^2 &= O(1)\|U(U^TS^TSU)^{-1}U^TS^TSb - UU^Tb\|_2^2 \\ &= O(1)\|(U^TS^TSU)^{-1}U^TS^TSb - U^Tb\|_2^2. \end{aligned}$$

Since S is a $(1 \pm 1/2)$-subspace embedding with probability at least $9/10$ by property (1), all singular values of $(U^T S^T S U)^{-1}$ are in the range $[2/3, 2]$, and thus

$$\begin{aligned}
\|(U^T S^T S U)^{-1} U^T S^T S b - U^T b\|_2^2 &= O(1)\|(U^T S^T S U)((U^T S^T S U)^{-1} U^T S^T S b - U^T b)\|_2^2 \\
&= O(1)\|U^T S^T S b - U^T S^T S U U^T b\|_2^2 \\
&= O(1)\|U^T S^T S(b - U x^*)\|_2^2.
\end{aligned}$$

We now use the approximate matrix product property, which says with probability at least $9/10$,

$$\|U^T S^T S(b - U x^*)\|_2^2 = O(\epsilon/d)\|U^T\|_F^2 \cdot \|U x^* - b\|_2^2 = O(\epsilon)\|U x^* - b\|_2^2,$$

which therefore holds with probability at least $1 - 1/10 - 1/10 = 4/5$.

Problem 3: CountSketch Preserves Frobenius Norm We give an elementary argument based on Chebyshev's inequality. Let A_i denote the i-th column of A, for $i \in [d]$. For each of the d rows i of S, let $h(i) \in [r]$ denote the location of the single non-zero entry of S in the i-th row, and let $\sigma_i \in \{-1, 1\}$ be this entry. Then

$$\|AS\|_F^2 = \sum_{j \in [r]} \|\sum_{i \in [d] \text{ such that } h(i)=j} \sigma_i A_i\|_2^2 = \sum_{j \in [r]} \sum_{i, i' \in [d] \text{ such that } h(i)=j} \sigma_i \sigma_{i'} \langle A_i, A_i \rangle.$$

For any fixed h, taking expectation over σ we have that $\mathbf{E}[\sigma_i \sigma_{i'}] = 0$ unless $i = i'$, in which case $\mathbf{E}[\sigma_i \sigma_{i'}] = 1$. It follows by linearity of expectation that

$$\mathbf{E}[\|AS\|_F^2] = \sum_{j \in [r]} \sum_{i \text{ such that } h(i)=j} \|A_i\|_2^2 = \|A\|_F^2.$$

We also have

$$\|AS\|_F^4 = \sum_{j_1, j_2 \in [r]} \sum_{i_1, i_2 \text{ such that } h(i_1)=h(i_2)=j_1} \sigma_{i_1} \sigma_{i_2} \langle A_{i_1}, A_{i_2} \rangle \sum_{i_3, i_4 \text{ such that } h(i_3)=h(i_4)=j_2} \sigma_{i_3} \sigma_{i_4} \langle A_{i_3, i_4} \rangle.$$

Let $\delta(h(i_1) = j_1)$ be 1 if $h(i_1) = j_1$, and be 0 otherwise. Then we can write $\mathbf{E}[\|AS\|_F^4]$ as

$$\sum_{j_1, j_2 \in [r], i_1, i_2, i_3, i_4 \in [d]} \mathbf{E}[\delta(h(i_1) = j_1)\delta(h(i_2) = j_1)\delta(h(i_3) = j_2)\delta(h(i_4) = j_2)\sigma_{i_1}\sigma_{i_2}\sigma_{i_3}\sigma_{i_4}]$$

$$\cdot \langle A_{i_1}, A_{i_2} \rangle \langle A_{i_3}, A_{i_4} \rangle$$

Taking expectation only with respect to σ, to have a non-zero expectation, we must be able to partition $\{i_1, i_2, i_3, i_4\}$ into equal pairs. This drives the analysis behind the following cases.

Case: $j_1 \neq j_2$. Then the set $\{i_1, i_2\}$ must be disjoint from $\{i_3, i_4\}$ since we cannot have $h(i) = j_1$ and $h(i) = j_2$ for some $j_1 \neq j_2$. It follows that $i_1 = i_2$ and $i_3 = i_4$ and $i_1 \neq i_3$ are the only terms which contribute to the expectation. It follows that the total contribution from terms for which $j_1 \neq j_2$ is

$$\sum_{j_1 \neq j_2 \in [r], i_1 \neq i_3 \in [d]} \frac{1}{r^2} \|A_{i_1}\|_2^2 \|A_{i_3}\|_2^2 \leq \|A\|_F^4 - \sum_i \|A_i\|_2^4.$$

Case: $j_1 = j_2$**, and** $i_1 = i_2 = i_3 = i_4$**.** The total contribution from these terms is

$$\sum_{j_1 \in [r], i_1 \in [d]} \frac{1}{r} \|A_{i_1}\|_2^4 = \sum_i \|A_i\|_2^4.$$

Case: $j_1 = j_2$**, and** $i_1 = i_2$**,** $i_3 = i_4$**,** $i_1 \neq i_3$**.** The total contribution from these terms is

$$\sum_{j_1 \in [r], i_1 \neq i_3 \in [d]} \frac{1}{r^2} \|A_{i_1}\|_2^2 \|A_{i_3}\|_2^2 = O(1/r) \|A\|_F^4.$$

Case: $j_1 = j_2$**, and** $i_1 = i_3$**,** $i_2 = i_4$**,** $i_1 \neq i_2$**.** The total contribution from these terms is

$$\sum_{j_1 \in [r], i_1 \neq i_2 \in [d]} \frac{1}{r^2} \langle A_{i_1}, A_{i_2} \rangle^2 = O(1/r) \|A\|_F^4.$$

Case: $j_1 = j_2$**, and** $i_1 = i_4$**,** $i_2 = i_3$**,** $i_1 \neq i_2$**.** This case is the same as the previous case, and contributes $O(1/r)\|A\|_F^4$.

In total, we have $\mathbf{E}[\|AS\|_F^4] = \|A\|_F^4 + O(1/r)\|A\|_F^4$. Hence, $\mathbf{Var}[\|AS\|_F^2] = \mathbf{E}[\|AS\|_F^4] - \mathbf{E}^2[\|AS\|_F^2] = O(1/r)\|A\|_F^4$. By Chebyshev's inequality,

$$\Pr[|\|AS\|_F^2 - \|A\|_F^2| \geq \epsilon \|A\|_F^2] = \frac{O(1/r)\|A\|_F^4}{\epsilon^2 \|A\|_F^4} \leq \frac{1}{10},$$

for suitably chosen $r = \Theta(1/\epsilon^2)$.

Problem 4: Sketching Structured Regression Problems

(1) Consider a family $\mathcal{F}_m$ of pairs (A, b) defined as follows. Let A^o be the $n \times d$ matrix with upper $d \times d$ matrix the $d \times d$ identity matrix, and $A^o_{i,j} = 1/d$ for all $i \in \{d+1, d+2, \ldots, d+m/d-1\}$ and all $j \in \{1, 2, \ldots, d\}$. For $i' \in \{d+1, \ldots, d+m/d-1\}$ and $j' \in \{1, 2, \ldots, d\}$, let $A^{i',j'} = A^o + (3n - 1/d)e_{i',j'}$, where $e_{i',j'}$ is the matrix with a single 1 in the (i', j')-th entry, and zeros in all remaining entries. Let $b_i = 1$ for $i \in \{1, 2, \ldots, d+m/d-1\}$, and $b_i = 0$ for $i \in \{d+m/d, \ldots, n\}$. Define $\mathcal{F}_m$ to be the union of (A^o, b) and $(A^{i',j'}, b)$ for $i' \in \{d+1, \ldots, d+m/d-1\}$ and $j' \in \{1, 2, \ldots, d\}$.

Notice that setting $x = 1^d$ allows for $A^o x = b$, and so the regression cost is 0 in this case. Moreover, $x = 1^d$ is the unique solution giving cost 0, and so must be returned by any regression algorithm achieving relative error if the algorithm succeeds. On the other hand for $x = 1^d$, $\|A^{i',j'}x - b\|_2^2 \geq (3n-1)^2$ for any $i' \in \{d+1, \ldots, d+m/d-1\}$

and $j' \in \{1, 2, \ldots, d\}$, but setting $x = 0^d$ gives $\|A^{i',j'}x - b\|_2^2 = \|b\|_2^2 \leq n$, and so $x = 1^d$ does not provide a 2-approximate solution. It follows that the output of the regression problem can distinguish if the matrix A is A^o or if it is $A^{i',j'}$ for some i', j'.

We define two distributions μ and ν: μ just has support equal to (A^o, b), and so a sample from μ always equals (A^o, b). On the other hand, ν is the distribution obtained by choosing uniformly random and independent $i' \in \{d+1, \ldots, d+m/d-1\}$ and $j' \in \{1, 2, \ldots, d\}$ and outputting $(A^{i',j'}, b)$. By Yao's minimax principle, if there is a randomized algorithm which reads $o(m)$ entries in expectation to solve the approximate regression problem with probability 3/4, then there is a deterministic algorithm which reads $o(m)$ entries in expectation to solve the approximate regression problem given a random input from distribution $(\mu + \nu)/2$. By Markov's bound, this implies there exists a deterministic algorithm for solving the approximate regression problem with probability at least 2/3 from a random input from $(\mu + \nu)/2$, and which *always* reads $o(m)$ entries. By the previous paragraph, this deterministic algorithm succeeds, with probability at least 2/3, in deciding if the input comes from μ or from ν. We assume such an algorithm exists and derive a contradiction.

We can assume the deterministic algorithm only queries entries in rows numbered $d+1, \ldots, d+m/d-1$, since all other rows have the same entries for all matrices in all pairs in $\mathcal{F}_m$. Further, the algorithm can only distinguish the two distributions if it reads an entry of value $3n$, and when it does, it can correctly output that (A, b) was drawn from ν. Thus, we can identify the deterministic algorithm with a subset S of $o(m)$ entries in these rows. However, the probability that a matrix $A^{i',j'}$ from a pair in ν satisfies $(i', j') \in S$ is $|S|/m = o(1)$, and therefore with probability $1 - o(1)$ the algorithm only reads entries of value $1/d$. Thus, the correctness probability of the algorithm can be at most $(1 + o(1))/2 < 2/3$, a contradiction.

(2) Let S be an $r \times n$ CountSketch matrix, for $r = O(d/\epsilon^2)$. Let $h : [n] \to [r]$ and $\sigma : [n] \to \{-1, 1\}$ be the associated hash and sign functions. We know that if we compute $S \cdot A$ and $S \cdot b$, then if $x' = (SA)^- Sb$, we have $\|Ax' - b\|_2 \leq (1+\epsilon) \min_{x \in \mathbb{R}^d} \|Ax - b\|_2$. Also given SA, one can compute Sb in $O(n)$ time and then solve for x' in $\text{poly}(d/\epsilon)$ time. Thus, it suffices to show how to compute SA in $(n+d) \cdot \text{poly}(\log n)$ time. For each $i \in [r]$, let A^i be the matrix formed by A by removing all rows A_j for which $h(j) \neq i$. Let σ^i be the vector formed from $(\sigma_1, \ldots, \sigma_n) \in \{-1, 1\}^n$ by removing all entries for which $h(j) \neq i$. Then, the i-th row of (SA), denoted $(SA)_i$, satisfies $(SA)_i = \sigma^i A^i$. Observe that A^i, being a subset of rows of A, is itself a Vandermonde matrix. Therefore, by the hint, one can compute $\sigma^i A^i$ in $(r_i + d) \cdot \text{poly}(\log(r_i d))$ time, where r_i is the number of rows of A^i. It follows that SA can be computed in time

$$\sum_i (r_i + d) \cdot \text{poly}(\log(r_i d)) \leq (n + rd) \cdot \text{poly}(\log(nd)) \leq n \cdot \text{poly}(\log n) + \text{poly}(d(\log n)/\epsilon).$$