15-859 Algorithms for Big Data — Fall 2022

Solutions for Problem Set 1

Problem 1: Subspace Embeddings for Other Norms

Part 1

Note that given a vector $x \in \mathbb{R}^d$, the d^p entries of $x^{\otimes p}$ are indexed by $(i_1, \ldots, i_p)$ with $1 \le i_j \le d$ for $1 \le j \le p$ and

$$(x^{\otimes p})_{(i_1,\ldots,i_p)} := x_{i_1} x_{i_2} \ldots x_{i_p}.$$

For two arbitrary vectors $x, y \in \mathbb{R}^d$, by definition of the inner product of two vectors, we have

$$\begin{aligned}
\langle x^{\otimes p}, y^{\otimes p} \rangle &= \sum_{1 \le i_1,\ldots,i_p \le d} (x^{\otimes p})_{(i_1,\ldots,i_p)} (y^{\otimes})_{(i_1,\ldots,i_p)} \\
&= \sum_{1 \le i_1,\ldots,i_p \le d} (x_{i_1} \cdots x_{i_p})(y_{i_1} \cdots y_{i_p}) \\
&= \sum_{1 \le i_1,\ldots,i_p \le d} (x_{i_1} y_{i_1}) \cdots (x_{i_p} y_{i_p}) \\
&= \left(\sum_{1 \le i_1 \le d} x_{i_1} y_{i_1} \right) \cdots \left(\sum_{1 \le i_p \le d} x_{i_p} y_{i_p} \right) \\
&= \left(\sum_{1 \le i \le d} x_i y_i \right)^p = \langle x, y \rangle^p.
\end{aligned}$$

Part 2

Give an $n \times d$ matrix A, let $A_i \in \mathbb{R}^d$ denote the i-th row of matrix A. Thus,

$$\|Ax\|_p^p = \sum_i |\langle A_i, x \rangle|^p = \sum_i (|\langle A_i, x \rangle|^{p/2})^2.$$

From above, we have that $\langle A_i, x \rangle^{p/2} = \langle A_i^{\otimes p/2}, x^{\otimes p/2} \rangle$. Now define an $n \times d^{p/2}$ matrix $A^{(p/2)}$ such that i-th row of $A^{(p/2)}$ is $A_i^{\otimes p/2}$. Now

$$\begin{aligned}
\|A^{(p/2)} x^{\otimes p/2}\|_2^2 &= \sum_i \langle (A^{(p/2)})_i, x^{\otimes p/2} \rangle^2 \\
&= \sum_i \langle A_i^{\otimes p/2}, x^{\otimes p/2} \rangle^2 \\
&= \sum_i (\langle A_i, x \rangle^{p/2})^2 = \sum_i \langle A_i, x \rangle^p \\
&= \|Ax\|_p^p.
\end{aligned}$$

Let S be a Gaussian matrix with $O(d^{p/2}/\epsilon^2)$ rows. As the matrix $A^{(p/2)}$ has $d^{p/2}$ columns we have $\text{rank}(A^{(p/2)}) \leq d^{p/2}$ and therefore with probability $\geq 9/10$, for all vectors x,

$$\|SA^{(p/2)}x^{\otimes p/2}\|^2 = (1 \pm \epsilon)\|A^{(p/2)}x^{\otimes p/2}\|_2^2 = (1 \pm \epsilon)\|Ax\|_p^p.$$

Further, for constant p, $A^{(p/2)}$ can be computed in time $n \cdot \text{poly}(d)$ and then the matrix product $SA^{(p/2)}$ can be computed in $n \cdot \text{poly}(d)/\epsilon^2$.

Part 3

As seen in class, if a Gaussian matrix S has $O(\text{rank}(A)/\epsilon^2)$ rows, then it is a subspace embedding for A with probability $\geq 9/10$. Using this property, we showed in the previous part that if S has $O(d^{p/2}/\epsilon^2)$ rows, then S is a subspace embedding for the matrix $A^{(p/2)}$. In this problem, we want to argue that if A is a Vandermonde matrix, then $\text{rank}(A^{(p/2)}) \leq dp/2$ which can be much smaller than the crude upper bound of $d^{p/2}$ given by the number of columns of the matrix $A^{(p/2)}$.

The matrix $A^{(p/2)}$ has $d^{p/2}$ columns indexed by $(i_1, \ldots, i_{p/2})$ for $1 \leq i_j \leq d$ and $(i_1, \ldots, i_{p/2})$-th column of $A^{(p/2)}$ is given by

$$(y_1^{i_1-1+i_2-1+\cdots+i_{p/2}-1}, \ldots, y_n^{i_1-1+i_2-1+\cdots+i_{p/2}-1}) = (y_1^{(\sum_j i_j)-p/2}, \ldots, y_n^{(\sum_j i_j)-p/2}).$$

Hence the $(i_1, \ldots, i_{p/2})$-th column of $A^{(p/2)}$ is purely a function of $\sum_{j \leq p/2} i_j$. As $\sum_{j \leq p/2} i_j \leq dp/2$, there are at most $dp/2$ distinct columns in the matrix $A^{(p/2)}$ and therefore $\text{rank}(A^{(p/2)}) \leq dp/2$.

Thus, if S is a Gaussian matrix with $O(dp/\epsilon^2)$ rows, then with probability $\geq 9/10$, for all x,

$$\|SA^{(p/2)}x^{\otimes p/2}\|_2^2 = (1 \pm \epsilon)\|A^{(p/2)}x^{\otimes p/2}\|_2^2 = (1 \pm \epsilon)\|Ax\|_p^p.$$

For constant p, the matrix $SA^{(p/2)}$ can again be computed in time $n \cdot \text{poly}(d)/\epsilon^2$.

Additional Remarks

Given a time budget of $n \cdot \text{poly}(d)$, we can even compute subspace embeddings with $\epsilon = 0$ for the above problems for constant p. For example, in part 3, we showed that the matrix $A^{(p/2)}$ has a rank at most $dp/2$ which implies that the matrix $(A^{(p/2)})^T A^{(p/2)}$ has a rank at most $dp/2$ as well. Using the eigen value decomposition of the matrix $(A^{(p/2)})^T A^{(p/2)}$, we can obtain another matrix B with $\text{rank}(A^{(p/2)}) \leq dp/2$ rows such that $B^T B = (A^{(p/2)})^T A^{(p/2)}$ and therefore,

$$\begin{aligned}
\|Bx^{\otimes p/2}\|_2^2 &= (x^{\otimes p/2})^T B^T B x^{\otimes p/2} \\
&= (x^{\otimes p/2})^T (A^{(p/2)})^T A^{(p/2)} x^{\otimes p/2} \\
&= \|A^{(p/2)}x^{\otimes p/2}\|_2^2 \\
&= \|Ax\|_p^p.
\end{aligned}$$

For a constant p, the $n \times d^{p/2}$ matrix $A^{(p/2)}$ can be computed in $n \cdot \text{poly}(d)$ and then $(A^{(p/2)})^T A^{(p/2)}$ can be computed in $n \cdot \text{poly}(d)$. The eigen decomposition of the $d^{(p/2)} \times d^{(p/2)}$ matrix $(A^{(p/2)})^T A^{(p/2)}$ can then be computed in $\text{poly}(d)$ time to obtain the matrix $dp/2 \times d^{p/2}$ matrix B.

Problem 2: Randomized Rounding for Sparsification

Following the hint, we bound $\|A - \hat{A}\|_2$ and $\|\hat{A} - \tilde{A}\|_2$ separately and then use triangle inequality for operator norm to obtain the upper bound. As $\hat{A}$ is formed by zeroing the entries with absolute values of at most $\epsilon/2n$ in A, we have that the entries of $A - \hat{A}$ have an absolute value of at most $\epsilon/2n$ and therefore

$$\|A - \hat{A}\|_2^2 \leq \|A - \hat{A}\|_F^2 \leq (\epsilon/2n)^2 n^2 \leq \epsilon^2/4.$$

Thus, $\|A - \hat{A}\|_2 \leq \epsilon/2$. We now bound $\|\hat{A} - \tilde{A}\|_2$. For $t = 1, \ldots, s$, we define the random matrix X_t as

$$X_t := \frac{\hat{A}_{i_t,j_t}}{p_{i_t,j_t}} e_{i_t} e_{j_t}^T$$

where (i_t, j_t) is the entry sampled in the t-th iteration of the algorithm. We have $\tilde{A} = (1/s)\sum_{t=1}^{s} X_t$. Clearly, for all $t = 1, \ldots, s$

$$\begin{aligned}
\mathbf{E}[X_t] &= \sum_{i,j} \frac{\hat{A}_{i,j}}{p_{i,j}} e_i e_j^T \mathbf{Pr}[i_t = i, j_t = j] = \sum_{i,j} \frac{\hat{A}_{i,j}}{p_{i,j}} e_i e_j^T p_{i_t,j_t} \\
&= \sum_{i,j} \hat{A}_{i,j} e_i e_j^T \\
&= \hat{A}.
\end{aligned}$$

We use the following theorem to bound $\|\hat{A} - \tilde{A}\|_2$.

Theorem 1 (Matrix-Bernstein) *Consider a finite sequence $\{S_k\}$ of independent, random matrices with common dimension $d_1 \times d_2$. Assume that for $\mathbf{E}[S_k] = 0$ and $\|S_k\|_2 \leq L$ for each index k. Let $Z = \sum_k S_k$ and let*

$$\nu(Z) = \max(\|\mathbf{E}[ZZ^T]\|_2, \|\mathbf{E}[Z^T Z]\|_2).$$

For all $t \geq 0$,

$$\mathbf{Pr}[\|Z\|_2 \geq t] \leq (d_1 + d_2) \exp\left(-\frac{t^2/2}{\nu(Z) + Lt/3}\right).$$

Note that the random variables $X_1, \ldots, X_t$ are identically distributed but are not centered. So we define $Y_t = X_t - \mathbf{E}[X_t] = X_t - \hat{A}$. Note $\mathbf{E}[Y_t] = \mathbf{E}[X_t] - \hat{A} = 0$. We now bound $\|Y_t\|_2$ as follows:

$$\begin{aligned}
\|Y_t\|_2 &\leq \|X_t - \hat{A}\|_2 \\
&\leq \|X_t\|_2 + \|\hat{A}\|_2 \\
&\leq \left\| \frac{\hat{A}_{i_t,j_t}}{p_{i_t,j_t}} e_{i_t} e_{j_t}^T \right\|_2 + \|\hat{A}\|_2 = \frac{\|\hat{A}\|_F^2}{|\hat{A}_{i_t,j_t}|} + \|\hat{A}\|_2.
\end{aligned}$$

Using the fact that $|\hat{A}_{i,j}| \geq \epsilon/2n$ for all nonzero entries of $\hat{A}$ and as the randomly sampled (i_t, j_t) corresponds to a nonzero entry of $\hat{A}_{i,j}$ with probability 1, we have that

$$\|Y_t\|_2 \leq (2n/\epsilon)\|\hat{A}\|_F^2 + \|\hat{A}\|_2 \leq (2n/\epsilon)\|\hat{A}\|_F^2 + \|\hat{A}\|_F.$$

As $\epsilon \leq \|\hat{A}\|_F$ for a non-vacuous guarantee, we have $\|\hat{A}\|_F^2/\epsilon \geq \|\hat{A}\|_F$ and therefore that $\|Y_t\|_2 \leq (3n/\epsilon)\|\hat{A}\|_F^2$. Now define

$$Z = \sum_t (Y_t) = \sum_t (X_t - \hat{A}).$$

We have that $\nu(Z) = \max(\|\mathbf{E}[ZZ^T]\|_2, \|\mathbf{E}[Z^T Z]\|_2)$. We first bound $\|\mathbf{E}[ZZ^T]\|_2$.

$$\mathbf{E}[ZZ^T] = \mathbf{E}[(\sum_t Y_t)(\sum_t Y_t^T)] = \mathbf{E}[\sum_{t,t'} Y_t Y_{t'}^T].$$

If $t \neq t'$, then by independence of Y_t and $Y_{t'}$, we have $\mathbf{E}[Y_t Y_{t'}] = \mathbf{E}[Y_t]\mathbf{E}[Y_{t'}] = 0$. Hence,

$$\mathbf{E}[ZZ^T] = \sum_t \mathbf{E}[Y_t Y_t^T] = s\mathbf{E}[Y_1 Y_1^T].$$

Now, $\mathbf{E}[Y_1 Y_1^T] = \mathbf{E}[(X_1 - \hat{A})(X_1 - \hat{A})^T] = \mathbf{E}[X_1 X_1^T] - \hat{A}\hat{A}^T$ by linearity of expectation. We further obtain that $\|\mathbf{E}[Y_1 Y_1^T]\|_2 \leq \|\mathbf{E}[X_1 X_1^T]\|_2 + \|\hat{A}\|_2^2$ by triangle inequality. We now have,

$$\begin{aligned}
\mathbf{E}[X_1 X_1^T] &= \sum_{i,j} \frac{\hat{A}_{i,j}^2}{p_{i,j}^2} e_i e_j^T e_j e_i^T \mathbf{Pr}[(i_t = i, j_t = j)] \\
&= \sum_{i,j} \frac{\hat{A}_{i,j}^2}{p_{i,j}} e_i e_i^T = \sum_{i,j} \|\hat{A}\|_F^2 e_i e_i^T \\
&= n\|\hat{A}\|_F^2 I.
\end{aligned}$$

Thus, $\|\mathbf{E}[Y_1 Y_1^T]\|_2 \leq \|\mathbf{E}[X_1 X_1^T]\|_2 \leq n\|\hat{A}\|_F^2 + \|\hat{A}\|_2^2 \leq 2n\|\hat{A}\|_F^2$ and $\|\mathbf{E}[ZZ^T]\|_2 = s\|\mathbf{E}[Y_1 Y_1^T]\|_2 \leq 2sn\|\hat{A}\|_F^2$. We similarly obtain that $\|\mathbf{E}[Z^T Z]\|_2 \leq 2sn\|\hat{A}\|_F^2$. Using the above theorem,

$$\mathbf{Pr}[\|\sum_t Y_t\|_2 \geq \Delta] \leq 2n \exp\left(-\frac{\Delta^2/2}{2sn\|\hat{A}\|_F^2 + (n/\epsilon)\|\hat{A}\|_F^2 \Delta}\right)$$

For $\Delta = \epsilon s/2$,

$$\mathbf{Pr}[\|\sum_t Y_t\|_2 \geq \epsilon s/2] \leq 2n \exp\left(-\frac{\epsilon^2 s^2/8}{3sn\|\hat{A}\|_F^2}\right) = 2n \exp(-\epsilon^2 s/24n\|\hat{A}\|_F^2).$$

For $s \geq 24n\|\hat{A}\|_F^2 \ln(2n^2)$, we have that with probability $\geq 1 - 1/n$,

$$\|\tilde{A} - \hat{A}\|_2 = \|(1/s)\sum_t Y_t\|_2 \leq \epsilon/2.$$

Finally, by triangle inequality, $\|A - \tilde{A}\|_2 \leq \|A - \hat{A}\|_2 + \|\hat{A} - \tilde{A}\|_2 \leq \epsilon$ with probability $\geq 1 - 1/n$.

Additional Remarks

Frobenius Norm Sparsification vs Spectral Norm Sparsification In this problem, we bound the error of randomized rounding for spectral norm sparsification. Given a matrix A, if we

want a matrix $\tilde{A}$ with s nonzero entries that minimizes $\|A - \tilde{A}\|_F$, then we can see that the best $\tilde{A}$ is obtained by keeping the s entries of highest absolute values in A but this algorithm is not correct for obtaining the matrix $\tilde{A}$ with s nonzero entries that minimizes $\|A - \tilde{A}\|_2$. Think of some examples using small matrices.

Matrix approximation in Frobenius norm and Spectral norm can be very different as showed by the following example. Let $M(x)$ denote the following matrix:

$$M(x) = \begin{bmatrix} x & 1 \\ 1 & -1 \end{bmatrix}$$

We have $\operatorname{argmin}_x \|M(x)\|_F = 0$ but $\|M(1)\|_2 < \|M(0)\|_2$. So, $\operatorname{argmin}_x \|M(x)\|_2 \neq 0$.

Removing Small Elements before sampling Notice that removing the elements with absolute value at most $\epsilon/2n$ before the sampling process helped us in obtaining an upper bound on $\|Y_t\|_2$ in the above proof. It is possible to perform sparsification without the removal step too but we have to use a slightly different sampling distribution. See Section 6.2 from "An Introduction to Matrix Concentration Inequalities" by Joel A. Tropp.

When does the guarantee $\|A - \tilde{A}\|_2 \leq \epsilon$ make sense? We need $\epsilon < \|A\|_2$ for the guarantee to be non-trivial as $\tilde{A} = 0$ would otherwise satisfy $\|A - \tilde{A}\|_2 = \|A\|_2 \leq \epsilon$. So, the above algorithm has to sample $\Omega(n(\|A\|_F^2/\|A\|_2^2)\ln n)$ entries to give a non trivial guarantee. The quantity $\|A\|_F^2/\|A\|_2^2$ is called the "stable rank" of A and is a continuous relaxation of rank that is resilient to small perturbations. Thus, matrices with *low* stable rank can be approximated well in spectral norm by sparse matrices.

Problem 3: Sparse Deterministic Matrix Product

As in the hint, we first prove a similar result for vectors. Let $a, b \in \mathbb{R}^n$ be arbitrary vectors. Define $\bar{a}$ as follows : If $|a_i| > (\epsilon/2)\|a\|_1$, then set $\bar{a}_i = a_i$ and otherwise set $\bar{a}_i = 0$. Define the vector $\bar{b}$ similarly. Now,

$$\langle a, b\rangle = \langle \bar{a}, b\rangle + \langle a - \bar{a}, b\rangle = \langle \bar{a}, \bar{b}\rangle + \langle \bar{a}, b - \bar{b}\rangle + \langle a - \bar{a}, b\rangle.$$

Thus,

$$\begin{aligned} |\langle a, b\rangle - \langle \bar{a}, \bar{b}\rangle| &\leq |\langle \bar{a}, b - \bar{b}\rangle| + |\langle a - \bar{a}, b\rangle| \\ &\leq \|\bar{a}\|_1 \|b - \bar{b}\|_\infty + \|b\|_1 \|a - \bar{a}\|_\infty. \end{aligned}$$

Here $\|x\|_\infty := \max_i |x_i|$, the largest coordinate in absolute value in the vector x. The above inequality can be seen as follows: $|\langle x, y\rangle| = |\sum_i x_i y_i| \leq \sum_i |x_i y_i| \leq \sum_i |x_i| \|y\|_\infty \leq \|x\|_1 \|y\|_\infty$. As we zeroed only those coordinates in a that have magnitude at most $(\epsilon/2)\|a\|_1$, we have that $\|a - \bar{a}\|_\infty \leq (\epsilon/2)\|a\|_1$. Similarly, $\|b - \bar{b}\|_1 \leq (\epsilon/2)\|b\|_1$. Also note that $\|\bar{a}\|_1 \leq \|a\|_1$. Thus,

$$|\langle a, b\rangle - \langle \bar{a}, \bar{b}\rangle| \leq \epsilon \|a\|_1 \|b\|_1.$$

Now,

$$\begin{aligned} \|A \cdot B - \bar{A} \cdot \bar{B}\|_1 &= \sum_{i,j} |(AB)_{i,j} - (\bar{A}\bar{B})_{i,j}| \\ &= \sum_{i,j} |\langle A_{i*}, B_{*j}\rangle - \langle \bar{A}_{i*}, \bar{B}_{*j}\rangle| \\ &\leq \sum_{i,j} \epsilon \|A_{i*}\|_1 \|B_{*j}\|_1. \end{aligned}$$

In the last inequality, we used the bound for vectors we proved above. Now, $\sum_{i,j} \|A_{i*}\|_1 \|B_{*j}\|_1 = (\sum_i \|A_{i*}\|_1)(\sum_j \|B_{*j}\|_1) = \|A\|_1 \|B\|_1$. Hence,

$$\|A \cdot B - \bar{A} \cdot \bar{B}\|_1 \leq \epsilon \|A\|_1 \|B\|_1.$$

Problem 4: Computing the Best Cost Regression

We estimate the optimal cost of each of the regression problems and then obtain an approximation to the best cost regression. Consider the i-th regression problem:

$$\mathrm{OPT}_i = \min_x \|Ax - B_i\|_2.$$

Let S be a CountSketch matrix with $O(d^2)$ rows. As seen in class, with probability $\geq 9/10$, the matrix S is a $1/5$ subspace embedding for the column space of $[A, B_i]$ i.e., with probability $\geq 9/10$, for all vectors x,

$$\|S(Ax - B_i)\|_2 = (1 \pm 1/5)\|Ax - B_i\|_2.$$

Conditioned on the above event, if $\tilde{x} := (SA)^-(SB_i)$ is the optimal solution for the *sketched* regression problem $\min_x \|(SA)x - SB_i\|_2$ and $x^* := A^-B_i$ for the original regression problem, then

$$\|S(A\tilde{x} - B_i)\|_2 \leq \|S(Ax^* - B_i)\|_2 \leq (5/4)\|Ax^* - B_i\|_2 = (5/4)\mathrm{OPT}_i,$$

and

$$\|S(A\tilde{x} - B_i)\|_2 \geq (4/5)\|A\tilde{x} - B_i\|_2 \geq (4/5)\|Ax^* - B_i\|_2 \geq (4/5)\mathrm{OPT}_i.$$

Thus, with probability $\geq 9/10$, the optimal cost of the sketched problem approximates the optimal regression cost of the i-th regression problem. We need a smaller failure probability than $1 - 9/10 = 1/10$ so as to be able to approximate the optimal costs of all m regression problems simultaneously with high probability. To obtain an approximation that is correct with high probability, we consider $O(\log m)$ independent instances of the CountSketch matrices and take the median of the estimates of optimal cost. Let $S^{(1)}, \ldots, S^{(r)}$ be $r = O(\log m)$ independent instances of the CountSketch matrix. From above, for each $j = 1, \ldots, r$, with probability $\geq 9/10$,

$$\|S^{(j)}A(S^{(j)}A)^-S^{(j)}B_i - S^{(j)}B_i\|_2 \in [(4/5)\mathrm{OPT}_i, (5/4)\mathrm{OPT}_i].$$

Thus, by a Chernoff bound, with probability $\geq 1 - 1/\mathrm{poly}(m)$,

$$\mathrm{median}(\|S^{(j)}A(S^{(j)}A)^-S^{(j)}B_i - S^{(j)}B_i\|_2)_{j\in[r]} \in [(4/5)\mathrm{OPT}_i, (5/4)\mathrm{OPT}_i].$$

By a union bound, with probability $\geq 99/100$, simultaneously for all regressions problems,

$$\mathrm{median}(\|S^{(j)}A(S^{(j)}A)^-S^{(j)}B_i - S^{(j)}B_i\|_2)_{j\in[r]} \in [(4/5)\mathrm{OPT}_i, (5/4)\mathrm{OPT}_i].$$

Notice how we can use the same sketching matrices $S^{(j)}$ for all the regression problem. So the best cost regression problem essentially reduces to estimating the norms of columns of the matrices

$$(I - S^{(j)}A(S^{(j)}A)^-)S^{(j)}B$$

for $j = 1, \ldots, r$. For each j, $S^{(j)}A$ and $S^{(j)}B$ can be computed in time $\mathrm{nnz}(A)$ and $\mathrm{nnz}(B)$ respectively. The pseudoinverse $(S^{(j)}A)^-$ can be computed in time $O(d^4)$. But it takes $O(\mathrm{nnz}(S^{(j)}B)d^2)$ time to compute the entire matrix—the time budget we do not have. Note that we only need to estimate the column norms of the matrix and do not need to compute the columns of the matrix

exactly. Thus, we can use an appropriately scaled Gaussian matrix $\mathbf{G}$ with only a few rows to estimate the column norms.

In class, we saw that a Gaussian matrix with $O(\log N/\varepsilon^2)$ rows preserves the norms of N vectors simultaneously with probability $\geq 99/100$. Here, we want to estimate the norms of $m \cdot r$ vectors up to constant factors. Hence a Gaussian matrix with $O(\log m + \log r) = O(\log m)$ rows suffices to estimate norms of all the columns of $(I - S^{(j)}A(S^{(j)}A)^{-})S^{(j)}B$ by computing the column norms of the matrix

$$\mathbf{G}(I - S^{(j)}A(S^{(j)}A)^{-})S^{(j)}B.$$

For each j, computing the above matrix takes $O(\mathrm{nnz}(S^{(j)}B)\log m + d^3 \log m)$. As for a CountSketch matrix $S^{(j)}$, we have $\mathrm{nnz}(S^{(j)}M) \leq \mathrm{nnz}(M)$ for any matrix M. Overall in time

$$O(\mathrm{nnz}(A)\log m + \mathrm{nnz}(B)(\log m)^2 + d^4 \log m + d^3(\log m)^2),$$

with probability $\geq 9/10$, we can approximate the regression cost of all the m regression problems and obtain a 2 approximation for the best cost regression.