CS-448 Sublinear Algorithms for Big Data Analysis March 26, 2025

Lecture 6

Lecturer: Kshiteej Sheth

1 Least squares regression

The *exact* least squares regression is the following problem: given $A \in \mathbb{R}^{n\times d}$ and $b \in \mathbb{R}^n$, find

$$x^* = \text{argmin}_{x\in\mathbb{R}^d}||Ax - b||_2.$$

The least squares problem often comes up in the following setting. We are observing samples $a_i \in \mathbb{R}^d, i = 1, \dots, n$ (rows of A) together with a value of some unknown function f on the samples, perhaps corrupted by noise. The value of the function on the i-th sample is denoted by b_i. Then if the function f is linear in the attributes of the sample, i.e. coordinates of a_i, the least squares problem is asking to recover the coefficients x that allow one to predict b_i from a_i. In fact, in fairly general settings (e.g. when the vector b equals the value of the unknown linear function plus i.i.d. noise), a least squares fit is the best (unbiased) estimate of the linear function that one can obtain from the samples – see the Gauss-Markov theorem.

How do we solve least squares in general? The solution is $(A^TA)^+A^Tb$, where $(A^TA)^+$ is the Moore-Penrose pseudoinverse of A^TA, and can be computed via an SVD computation, taking $O(nd^2)$ time.

The *approximate* least squares problem is the following. We are given $A \in \mathbb{R}^{n\times d}, b \in \mathbb{R}^n$, $\epsilon \in (0, 1)$. Let $x^* := \text{argmin}_{x\in\mathbb{R}^d}||Ax - b||_2$. We would like to find $x' \in \mathbb{R}^d$ such that

$$||Ax' - b||_2 \le (1 + \epsilon)||Ax^* - b||_2. \tag{1}$$

We will solve least squares approximately using *subspace embeddings*:

Definition 1 *A random matrix $\Pi \in \mathbb{R}^{m\times n}$ is a (d, ϵ, δ)-subspace embedding if for every d-dimensional subspace $P \subseteq \mathbb{R}^n$ one has*

$$\textbf{Prob}[|||\Pi x||_2 - ||x||_2| \le \epsilon||x||_2 \textit{ for all } x \in P] \ge 1 - \delta.$$

The runtime of our final solution will be $O(\text{nnz}(A) + \text{poly}(d, 1/\epsilon))$, where $\text{nnz}(A)$ denotes the number of nonzeros in A. Note that the leading order term is normally $\text{nnz}(A)$ (if n is much larger than d), and hence this is much faster than the nd^2 time for SVD, especially if the matrix A is sparse.

Given a $(d + 1, \epsilon, \delta)$-subspace embedding Π, our algorithm will be simple: solve

$$x' := \text{argmin}_{x\in\mathbb{R}^d}||\Pi Ax - \Pi b||_2. \tag{2}$$

Why does this work? Consider the (at most) $(d + 1)$-dimensional subspace of $\mathbb{R}^n$ given by the span of $[A; b] \in \mathbb{R}^{n\times(d+1)}$. Since Π is a $(d + 1, \epsilon, \delta)$-subspace embedding, we have that with probability at least $1 - \delta$ over Π for very $x \in \mathbb{R}^d$

$$\left|||[A; b] \cdot [x^T; -1]^T||_2 - ||\Pi[A; b] \cdot [x^T; -1]^T||_2\right| \le \epsilon \left|||[A; b] \cdot [x^T; -1]^T||_2\right|. \tag{3}$$

Here $[x^T; -1]^T$ stands for a column vector obtained from $x \in \mathbb{R}^d$ by appending 1 as the last coordinate.

Since x' is the optimum in (2), we have

$$||\Pi Ax' - \Pi b||_2 \leq ||\Pi Ax^* - \Pi b||_2. \tag{4}$$

By (3) applied with $x = x^*$ we have

$$||Ax^* - b||_2 \geq ||\Pi Ax^* - \Pi b||_2 - \epsilon||Ax^* - b||_2$$

so

$$||\Pi Ax^* - \Pi b||_2 \leq (1+\epsilon)||Ax^* - b||_2.$$

Similarly by (3) applied with $x = x'$ we have

$$||Ax' - b||_2 \leq ||\Pi Ax' - \Pi b||_2 + \epsilon||Ax' - b||_2$$

so

$$||Ax' - b||_2 \leq \frac{1}{1-\epsilon}||\Pi Ax' - \Pi b||_2.$$

Putting these two bounds together with (4), we get

$$||Ax' - b||_2 \leq \frac{1}{1-\epsilon}||\Pi Ax' - \Pi b||_2 \leq \frac{1}{1-\epsilon}||\Pi Ax^* - \Pi b||_2 \leq \frac{1+\epsilon}{1-\epsilon}||Ax^* - b||_2,$$

and hence x' from (2) satisfies (1), as required.

2 CountSketch is a subspace embedding

Recall that the CountSketch matrix is defined as follows. Fix a number B of buckets, a hash function $h : [n] \to [B]$ and a sign function $\sigma : [n] \to \{-1, +1\}$. For $r \in [B]$ and $a \in [n]$ let

$$S_{ra} = \begin{cases} \sigma(a) & \text{if } h(a) = r \\ 0 & \text{o.w.} \end{cases}$$

In other words, the CountSketch matrix multiplies an input vector by a diagonal sign matrix, and then hashes elements of the resulting vector into B buckets. We will need h to be pairwise independent, and σ to be four-wise independent.

We will show that for every subspace $U \in \mathbb{R}^{n \times d}$, if B is sufficiently large as a function of $d, 1/\epsilon$ and δ, then

$$\mathbf{Prob}[|||\Pi x||_2 - ||x||_2| \leq \epsilon||x||_2 \text{ for every } x \text{ in the span of the columns of } U] \geq 1 - \delta. \tag{5}$$

Our plan is as follows. We first show that in order to achieve the subspace embedding property in (5), it suffices to show that the matrix $U^T\Pi^T\Pi U$ is spectrally close to the identity matrix. We then note that proving that $U^T\Pi^T\Pi U$ is close to the identity in the Frobenius norm is even stronger, and prove the required upper bound on the Frobenius norm of $U^T\Pi^T\Pi U - I$ in the next section.

The final result (see section 2.3) will be that CountSketch with B buckets is a $(d, \epsilon, 2d^2/(\epsilon^2 B))$-subspace embedding. Setting $B = Cd^2/\epsilon^2$ for a large enough absolute constant C gives a subspace embedding with large constant probability.

2.1 Reducing to an upper bound on $||U^T S^T SU - I||_F$

We start by noting that (5) is equivalent to

$$(1-\epsilon)||x||_2 \le ||\Pi x||_2 \le (1+\epsilon)||x||_2 \ \text{ for every } x \text{ in the span of the columns of } U.$$

Rescaling ϵ appropriately we get that it suffices to ensure that for every x in the span of the columns of U

$$(1-\epsilon)||x||_2^2 \le ||\Pi x||_2^2 \le (1+\epsilon)||x||_2^2.$$

Writing $x = Uy$ for $y \in \mathbb{R}^d$, we get that the latter condition is equivalent

$$(1-\epsilon)y^T y \le yU^T \Pi^T \Pi U y \le (1+\epsilon)y^T y \quad \text{for all } y \in \mathbb{R}^d,$$

which is equivalent to

$$||U^T \Pi^T \Pi U - I_d||_2 \le \epsilon.$$

Since Frobenius norm upper bounds spectral norm, it suffices to show that if B is sufficiently large, then with high probability

$$||U^T \Pi^T \Pi U - I_d||_F \le \epsilon.$$

2.2 Upper bounding $||U^T S^T SU - I||_F$

In what follows we will show that for every orthonormal $U \in \mathbb{R}^{n\times d}$, if $\Pi = S$, with S a CountSketch matrix with B buckets, one has with probability at least $1 - 2d^2/(\epsilon^2 B)$ over S

$$||U^T S^T SU - I||_F \le \epsilon \tag{6}$$

We start by noting that for every $i, j \in [1:d]$ the matrix $M := U^T S^T SU$ satisfies

$$\begin{aligned} M_{ij} &= \sum_{r=1}^{B} \sum_{a=1}^{n} \sum_{b=1}^{n} S_{r,a} U_{a,i} S_{r,b} U_{b,j} \\ &= \sum_{a=1}^{n} U_{a,i} U_{a,j} \left(\sum_{r=1}^{B} S_{r,a}^2 \right) + \sum_{r=1}^{B} \sum_{a=1}^{n} \sum_{b=1, b\neq a}^{n} S_{r,a} U_{a,i} S_{r,b} U_{b,j} \\ &= \delta_{i,j} + \sum_{r=1}^{B} \sum_{\substack{a,b=1,\\ a\neq b}}^{n} S_{r,a} U_{a,i} S_{r,b} U_{b,j}, \end{aligned}$$

where $\delta_{i,j}$ equals 1 if $i = j$ and equals 0 otherwise. We thus have, for every $i, j \in [1:d]$, that

$$(M - I)_{ij} = \sum_{r=1}^{B} \sum_{\substack{a,b=1,\\ a\neq b}}^{n} S_{r,a} U_{a,i} S_{r,b} U_{b,j},$$

We prove (6) by first upper bounding the expectation of $||M - I||_F^2$, and then applying Markov's inequality. We have

$$\begin{aligned}
\mathbf{E}[||M-I||_F^2] &= \sum_{i=1}^{d}\sum_{j=1}^{d}\mathbf{E}[((M-I)_{ij})^2] \\
&= \sum_{i=1}^{d}\sum_{j=1}^{d}\mathbf{E}\left[\left(\sum_{r=1}^{B}\sum_{\substack{a,b=1,\\ a\neq b}}^{n} S_{r,a}U_{a,i}S_{r,b}U_{b,j}\right)^2\right] \\
&= \sum_{i=1}^{d}\sum_{j=1}^{d}\mathbf{E}\left[\sum_{r=1}^{B}\sum_{r'=1}^{B}\sum_{\substack{a,b=1,\\ a\neq b}}^{n}\sum_{\substack{a',b'=1,\\ a'\neq b'}}^{n} S_{r,a}U_{a,i}S_{r,b}U_{b,j}\cdot S_{r',a'}U_{a',i}S_{r',b'}U_{b',j}\right] \\
&= \sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{r=1}^{B}\sum_{r'=1}^{B}\sum_{\substack{a,b=1,\\ a\neq b}}^{n}\sum_{\substack{a',b'=1,\\ a'\neq b'}}^{n}\mathbf{E}\left[S_{r,a}S_{r,b}S_{r',a'}S_{r',b'}\right]U_{a,i}U_{b,j}U_{a',i}U_{b',j}
\end{aligned}$$

The set $\{a, b, a', b'\}$ must contain every element an even number of times if $\mathbf{E}\left[S_{r,a}S_{r,b}S_{r',a'}S_{r',b'}\right] \neq 0$, since the random sign function raised to an odd power has zero expectation, and the signs are four-wise independent by assumption.

Thus, it suffices to consider two cases.

Case 1: $a = a'$**,** $b = b'$**.** Note that we must have $r = r'$, as otherwise $S_{r,a} \cdot S_{r',a} = 0$. We thus get

$$\begin{aligned}
&\sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{r=1}^{B}\sum_{\substack{a,b=1,\\ a\neq b}}^{n}\mathbf{E}\left[S_{r,a}^2S_{r,b}^2\right]U_{a,i}^2U_{b,j}^2 \\
&= \frac{1}{B^2}\sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{r=1}^{B}\sum_{\substack{a,b=1,\\ a\neq b}}^{n}U_{a,i}^2U_{b,j}^2 \qquad \text{(since } \mathbf{E}[S_{r,a}^2] = \mathbf{Prob}[h(a)=r] = 1/B \text{ and } h \text{ is pairwise independent)} \\
&\leq \frac{1}{B}\sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{a,b=1}^{n}U_{a,i}^2U_{b,j}^2 \\
&= \frac{1}{B}\sum_{i=1}^{d}\sum_{j=1}^{d}(\sum_{a=1}^{n}U_{a,i}^2)(\sum_{b=1}^{n}U_{b,j}^2) \\
&\leq \frac{d^2}{B}.
\end{aligned}$$

In going from line 1 to line 2 we used the fact that

$$\mathbf{E}[S_{r,a}^2S_{r,b}^2] = \mathbf{Prob}[h(a) = r \ \text{ and } \ h(b) = r] = 1/B^2$$

by pairwise independence of h. In going from line 2 to line 3 we used the fact that all terms in the summation are nonnegative. In going from line 4 to line 5 we used the fact that $\sum_{a=1}^{n} U_{a,i}^2 = \sum_{b=1}^{n} U_{b,j}^2 = 1$ by orthonormality of columns of U.

Case 2: $a = b'$**,** $b = a'$**.** Note that we must have $r = r'$, as otherwise $S_{r,a} \cdot S_{r',a} = 0$. We thus get

$$\left| \sum_{i=1}^{d} \sum_{j=1}^{d} \sum_{r=1}^{B} \sum_{\substack{a,b=1, \\ a \neq b}}^{n} \mathbf{E}\left[S_{r,a}^2 S_{r,b}^2\right] U_{a,i} U_{a,j} U_{b,i} U_{b,j} \right|$$

$$= \left| \frac{1}{B} \sum_{i=1}^{d} \sum_{j=1}^{d} \sum_{\substack{a,b=1, \\ a \neq b}}^{n} U_{a,i} U_{a,j} U_{b,i} U_{b,j} \right| \qquad \text{(since } \mathbf{E}[S_{r,a}^2] = \mathbf{Prob}[h(a) = r] = 1/B \text{ and } h \text{ is pairwise independent)}$$

$$\leq \frac{1}{B} \sum_{i=1}^{d} \sum_{j=1}^{d} \sum_{\substack{a,b=1, \\ a \neq b}}^{n} |U_{a,i}| \cdot |U_{a,j}| \cdot |U_{b,i}| \cdot |U_{b,j}| \qquad \text{(by triangle inequality)}$$

$$\leq \frac{1}{B} \sum_{i=1}^{d} \sum_{j=1}^{d} \sum_{a,b=1}^{n} |U_{a,i}| \cdot |U_{a,j}| \cdot |U_{b,i}| \cdot |U_{b,j}| \qquad \text{(since all terms in the summation are nonnegative)}$$

$$= \frac{1}{B} \sum_{i=1}^{d} \sum_{j=1}^{d} (\sum_{a=1}^{n} |U_{a,i}||U_{a,j}|)(\sum_{b=1}^{n} |U_{b,i}||U_{b,j}|)$$

$$\leq \frac{1}{B} \sum_{i=1}^{d} \sum_{j=1}^{d} \sum_{a=1}^{n} U_{a,i}^2 \sum_{b=1}^{n} U_{b,j}^2 \qquad \text{(by Cauchy-Schwarz)}$$

$$\leq \frac{d^2}{B} \quad \text{(since } \sum_{a=1}^{n} U_{a,i}^2 = \sum_{b=1}^{n} U_{b,j}^2 = 1 \text{ by orthonormality of } U)$$

2.3 Putting it together

Putting the bounds from previous sections together, we get that

$$\mathbf{E}[||M - I||_F^2] \leq 2d^2/B,$$

and thus by Markov's inequality

$$\mathbf{Prob}[||M - I||_F^2 > \epsilon^2] \leq 2d^2/(\epsilon^2 B),$$

and since $||M - I||_2 \leq ||M - I||_F$,

$$\mathbf{Prob}[||M - I||_2 > \epsilon] \leq 2d^2/(\epsilon^2 B).$$

We thus get that CountSketch with B buckets is a $(d, \epsilon, 2d^2/(\epsilon^2 B))$-subspace embedding. Setting $B = Cd^2/\epsilon^2$ for a large enough absolute constant C gives a subspace embedding with large constant probability.

How efficiently can we solve least squares using this approach? We need to find

$$\text{argmin}_{x \in \mathbb{R}^d} ||SAx - Sb||_2.$$

The matrix SA is a $B \times d$ matrix, and hence this computation can be done in time poly(d) using SVD. How much time does it take to form the matrix SA and the vector Sb? Since every column of S has exactly one nonzero, the runtime of this is proportional to the number of nonzeros in the matrix A and the vector b. Thus, if the matrix is sparse, this is very efficient! The final runtime is $O(nnz(A) + \text{poly}(d, 1/\epsilon))$, which is a significant improvement over $O(nd^2)$ for direct SVD.