CS 15-851: Algorithms for Big Data Spring 2025

Lecture Lecture 2 Part 1 — Jan 23

Prof. David Woodruff *Scribe: Chris Crawford*

1 Sketching for fast linear regression – continued

We continue from where we left off in the previous lecture, discussing algorithms for least-squares linear regression, which is formalized as finding $\mathbf{x}^* = \text{argmin}_{\mathbf{x}} \|A\mathbf{x} - \mathbf{b}\|_2^2$. While it is possible to find an optimal solution by using the normal equations, actually computing such a solution takes $O(nd^2)$ time, which is unnaceptably long for the size of data we're considering in this class.

Our approach instead is to *approximate* the optimal solution within a factor of $(1 \pm \epsilon)$ with high probability. Specifically, we choosing a sketching matrix $S \in \mathbb{R}^{k \times n}$ of i.i.d. $N\left(0, \frac{1}{k}\right)$ random variables, and approximating the regression solution with the normal equations on the "sketched" matrices, i.e. $\mathbf{x}' := (SA)^-(S\mathbf{b})$.

1.1 Completing the net argument

Last time, we showed that for any fixed unit vector $\mathbf{x}$, $\mathbf{Pr}[\|SA\mathbf{x}\|_2^2 \in (1 \pm \epsilon)] \geq 1 - 2^{-\Theta(d)}$. We have also shown that for a $\frac{1}{2}$-net $M = \{A\mathbf{x} | \mathbf{x} \in N\}$, the dot product between two scaled vectors in M is preserved within a factor of $1 \pm \epsilon$ after sketching. In other words,

$$\forall \mathbf{y}, \mathbf{y}' \in M, \alpha, \beta \in \mathbb{R}, \langle \alpha\mathbf{y}, \beta\mathbf{y}' \rangle = \alpha\beta\langle S\mathbf{y}, S\mathbf{y}' \rangle + O(\alpha\beta\epsilon)$$

We want to generalize these guarantees to all vectors $\mathbf{y} = A\mathbf{x}$.

Let $\mathbf{y} = A\mathbf{x}$ for an arbitrary unit vector $\mathbf{x} \in S^{d-1}$, and pick $\mathbf{y}_1 \in M$ such that $\|\mathbf{y} - \mathbf{y}_1\|_2 \leq \gamma$. Now let $\alpha = \frac{1}{\|\mathbf{y} - \mathbf{y}_1\|_2}$ and observe that $\alpha \geq \frac{1}{\gamma}$. If we choose $\mathbf{y}_2' \in M$ such that $|\alpha(\mathbf{y} - \mathbf{y}_1) - \mathbf{y}_2'| \leq \gamma$, then $\left\|\mathbf{y} - \mathbf{y}_1 - \frac{\mathbf{y}_2'}{\alpha}\right\|_2 \leq \frac{\gamma}{\alpha} \leq \gamma^2$. Now we can let $\mathbf{y}_2 = \frac{\mathbf{y}_2'}{\alpha}$ and repeat the process again for $\mathbf{y}_3, \mathbf{y}_4, ...$ etc. Observe that $\|\mathbf{y} - \sum_{i=1}^n \mathbf{y}_i\|_2 \leq \gamma^n$ for all integers n, and that this sum $\sum_i \mathbf{y}_i$ is a linear combination of net vectors.

We can show by the triangle inequality that each $\|\mathbf{y}_i\|_2 \leq \gamma^i + \gamma^{i-1} \leq 2\gamma^{i-1}$. Now we can rewrite $\mathbf{y}$ as the sum $\sum_i \mathbf{y}_i$. Observe that since $|M|$ is finite, the sum must have finitely many terms. Now consider the value of $\|S\mathbf{y}\|_2^2$:

$$\begin{aligned} \|S\mathbf{y}\|_2^2 &= \left\| S \sum_i \mathbf{y}_i \right\|_2^2 \\ &= \sum_i \|S\mathbf{y}_i\|_2^2 + 2\sum_{i,j} \langle S\mathbf{y}_i, S\mathbf{y}_j \rangle \end{aligned}$$

Since the the norms of sketched net vectors are preserved within $1 \pm \epsilon$, and dot products of two sketched net vectors have the property $\langle S\mathbf{y}_i, S\mathbf{y}_j \rangle = \langle \mathbf{y}_i, \mathbf{y}_j \rangle \pm O(\epsilon)\|\mathbf{y}_i\|_2\|\mathbf{y}_j\|_2$, we get that

$$\begin{aligned}
&= \sum_i \|\mathbf{y}_i\|_2^2 + 2\sum_{i,j}\langle \mathbf{y}_i, \mathbf{y}_j \rangle \pm O(\epsilon)\sum_{i,j}\|\mathbf{y}_i\|_2\|\mathbf{y}_j\|_2 \\
&= \left\|\sum_i \mathbf{y}_i\right\|_2^2 \pm O(\epsilon) \\
&= \|\mathbf{y}\|_2^2 \pm O(\epsilon) \\
&= 1 \pm O(\epsilon) && \text{(since } \mathbf{x} \text{ is unit and } A \text{ is orthonormal)}
\end{aligned}$$

Since this holds for an arbitrary $\mathbf{y} = A\mathbf{x}$ for a unit vector $\mathbf{x}$, by linearity it follows that $\|SA\mathbf{x}\|_2 = (1 \pm \epsilon)\|A\mathbf{x}\|_2$. ■

1.2 Back to regression

We have now shown that S is a subspace embedding, meaning that for all $\mathbf{x}$, $\|SA\mathbf{x}\|_2 = (1 \pm \epsilon)\|A\mathbf{x}\|_2$. Recall the linear regression problem: $\mathbf{x}^* = \text{argmin}_{\mathbf{x}}\|A\mathbf{x} - \mathbf{b}\|_2$.

Now we can take $\mathbf{x}' = \text{argmin}_{\mathbf{x}}\|SA\mathbf{x} - S\mathbf{b}\|_2 = \text{argmin}_{\mathbf{x}}\|S(A\mathbf{x} - \mathbf{b})\|_2$. Observe that $A\mathbf{x} - \mathbf{b}$ lives in a $(d+1)$-dimensional subspace of $\mathbb{R}^n$, which we will call L. More precisely, $L = \text{colspan}(A) \cup \text{span}(\mathbf{b})$. By the result proven above, for all $\mathbf{y} \in L$ we have that $\|S\mathbf{y}\|_2 = (1 \pm \epsilon)\|\mathbf{y}\|_2$. It follows then that $(1-\epsilon)\|A\mathbf{x}' - \mathbf{b}\|_2 \leq \|S(A\mathbf{x}' - \mathbf{b})\|_2$, and we also have that $(1-\epsilon)\|A\mathbf{x}^* - \mathbf{b}\|_2 \leq (1-\epsilon)\|A\mathbf{x}' - \mathbf{b}\|_2$ from the definition of $\mathbf{x}^*$. From our definition of $\mathbf{x}'$, we can also get that $\|S(A\mathbf{x}' - \mathbf{b})\|_2 \leq \|S(A\mathbf{x}^* - \mathbf{b})\|_2$, and then apply the subspace embedding property again to get that $\|S(A\mathbf{x}^* - \mathbf{b})\|_2 \leq (1+\epsilon)\|A\mathbf{x}^* - \mathbf{b}\|_2$. Now we need only use transitivty to get the desired bounds:

$$(1-\epsilon)\|A\mathbf{x}^* - \mathbf{b}\|_2 \leq \|S(A\mathbf{x}' - \mathbf{b})\|_2 \leq (1-\epsilon)\|A\mathbf{x}^* - \mathbf{b}\|_2$$

■

For our choice of $k = O(d/\epsilon^2)$, we can compute the value of $\mathbf{x}'$ in poly $\left(\frac{d}{\epsilon}\right)$ time given the values of SA and $S\mathbf{b}$. However, computing SA for an arbitrary random matrix of i.i.d. $N\left(0, \frac{1}{k}\right)$ entries S still takes $O(nd^2)$ time, so computing the approximate solution from scratch in this way is no better than computing the exact solution!

2 Alternative choices for S

If we want significant improvement in time complexity, then we need to construct the sketching matrix S in a way that the product SA can be computed efficiently. We can reduce this problem to efficiently multiplying S with an arbitrary vector, since matrix multiplication can be thought of as repeated matrix-vector multiplication for each column in the matrix.

2.1 Subsampled Randomized Hadamard Transform

Consider $S = PHD$, the product of three matrices, where...

- $D \in \mathbb{R}^{n \times n}$ is a diagonal matrix with diagonals $d_{ii} \in \{1, -1\}$ chosen with uniform probability,
- $H \in \mathbb{R}^{n \times n}$ is the Hadamard matrix, given by $h_{ij} = \frac{(-1)^{\langle i,j \rangle}}{\sqrt{n}}$ where $\langle i, j \rangle$ represents the dot product of i and j expressed as binary vectors
- $P \in \mathbb{R}^{k \times n}$ is a sampling matrix that chooses a random subset of rows and rescales them by a factor of $\sqrt{\frac{n}{k}}$.

Claim 1. The product $SA = PHDA$ can be computed in $O(nd \log n)$ time.

Let $\mathbf{z} \in \mathbb{R}^n$ be arbitrary. It is clearly efficient to compute the product $D\mathbf{z}$ because D is diagonal, so $(D\mathbf{z})_i = d_{ii} z_i$. The can be computed in $O(n)$ time.

Computing the product $P\mathbf{z}$ is also efficient: since each row of P contains exactly one nonzero entry, it suffices to add the scaled value of the selected row of $\mathbf{z}$ to the output. For instance, let's say the ith sample selects row j, so $p_{ij} = \sqrt{\frac{n}{k}}$, and $p_{i\ell} = 0$ for all $\ell \neq j$. Then $(P\mathbf{z})_i = \sqrt{\frac{n}{k}}(z_j)$, so this can be computed in $O(k)$ time, which is also within $O(n)$ since $k \ll n$.

Consider the product $H\mathbf{z}$. If we assume n to be a power of 2, observe that the Hadamard matrix can be broken down as follows:

$$H = \left[\begin{array}{c|c} H' & H' \\ \hline H' & -H' \end{array}\right]$$

where $H' \in \mathbb{R}^{\frac{n}{2} \times \frac{n}{2}}$ is the Hadamard matrix of size $\frac{n}{2}$, scaled to have entries of magnitude $\frac{1}{\sqrt{n}}$. This makes sense since for all $i, j < \frac{n}{2}$, the binary representations of $\frac{n}{2} + i$ and $\frac{n}{2} + j$ will differ by one bit from i and j respectively, and the dot product will be unchanged for all but the bottom right submatrix, which will yield $\frac{(-1)^{\langle i+n/2, j+n/2 \rangle}}{\sqrt{n}} = (-1)\frac{(-1)^{\langle i,j \rangle}}{\sqrt{n}}$. This gives us two subproblems: multiplying H' by the top half of $\mathbf{z}$, and multiplying H' by the bottom half of $\mathbf{z}$, after which we can combine and add the results to get $H\mathbf{z}$. This gives us a recurrence of $T(n) = 2T\left(\frac{n}{2}\right) + O(n)$, which can be simplified to $T(n) = O(n \log n)$.

This shows that multiplying $PHD\mathbf{z}$ can be done in $O(n \log n) + O(n) + O(k) = O(n \log n)$ time. Since A has d columns, it can be computed by multiplying out each of the d column vectors, giving us a final time complexity of $O(nd \log n)$. ■

It is worth noting that sampling from P alone is not generally effective: you could have $\mathbf{z}$ with only one nonzero entry, and have samples in P that only pick the zeros. However, HD acts as a rotation matrix[1], applying a rotation to the vector so that sampling from $HD\mathbf{z}$ is effective.

Theorem 1. *For all matrices $A \in \mathbb{R}^{n \times d}$ with orthonormal columns and all unit vectors $\mathbf{x}$, we have that $\|PHDA\mathbf{x}\|_2^2 \in (1 \pm \epsilon)$*

Since HD is a rotation matrix, we have that $\|HDA\mathbf{x}\|_2^2 = \|A\mathbf{x}\|_2^2 = 1$. We will let $\mathbf{y} = A\mathbf{x}$ for the remainder of the proof.

[1]It may be worth noting that the rows and columns of H are orthonormal. We can prove this by observing that WLOG for any $j \neq j'$, there exists some k where the kth bit differs between j and j'. The full proof was not covered in class, but the result is that the bit flips cause the signs in the dot product to cancel, which gives us $\langle H_{*j}, H_{*j'} \rangle = 0$.

Lemma 1 (Flattening Lemma)**.** *For any fixed* $\mathbf{y}$*,*

$$\mathbf{Pr}\left[\|HD\mathbf{y}\|_\infty \geq C\frac{\sqrt{\log\left(\frac{nd}{\delta}\right)}}{\sqrt{n}}\right] \leq \frac{\delta}{2d}$$

Recall that the infinity norm is defined as $\|\mathbf{z}\|_\infty = \max_i |z_i|$.

2.1.1 Proof of the Flattening Lemma

Let $C > 0$ be constant. We will show that for fixed $i \in [n]$, $\mathbf{Pr}\left[|(HD\mathbf{y})_i| \geq C\frac{\sqrt{\log\left(\frac{nd}{\delta}\right)}}{\sqrt{n}}\right] \leq \frac{\delta}{2nd}$. If we can prove this, we can get the desired result by applying a union bound over all i.

By the Asuma-Hoeffding bound, for independent zero-mean random variables Z_j, we have that $\mathbf{Pr}\left[\left|\sum_j Z_j\right| > t\right] \leq 2\exp\left[-\left(\frac{t^2}{2\sum_j \beta_j^2}\right)\right]$ where $|Z_j| \leq \beta_j$ with probability 1 for all j.

Write $(HD\mathbf{y})_i = \sum_j h_{ij}d_{jj}y_j$ and consider $Z_j = h_{ij}d_{jj}y_j$. Note that Z_j has a mean of zero, and that $|Z_j| \leq \frac{|y_j|}{\sqrt{n}}$ with probability 1. We can set $\beta_j = \frac{|y_j|}{\sqrt{n}}$, and observe that since $\mathbf{y}$ is unit, $\sum_j \beta_j^2 = \frac{1}{n}$. Setting $t = C\frac{\sqrt{\log\left(\frac{nd}{\delta}\right)}}{\sqrt{n}}$, we get

$$\mathbf{Pr}\left[\left|\sum_j Z_j\right| > C\frac{\sqrt{\log\left(\frac{nd}{\delta}\right)}}{\sqrt{n}}\right] \leq 2e^{-\frac{C^2\log(nd/\delta)}{2}} \leq \frac{\delta}{2nd}$$

By union bound over i, we get that $\mathbf{Pr}\left[\|HD\mathbf{y}\|_\infty > C\frac{\sqrt{\log\left(\frac{nd}{\delta}\right)}}{\sqrt{n}}\right] \leq 2e^{-\frac{C^2\log(nd/\delta)}{2}} \leq \frac{\delta}{2d}$, as desired. ■

The consequences of this lemma on proving the subspace embedding property for an SRHT sketch matrix are discussed in the next part of the lecture.