CS 15-859: Algorithms for Big Data	Fall 2017
Lecture 1 — September 7, 2017	
Prof. David Woodruff	*Scribe: Rajesh Jayaram*

Part 2

Recall we are trying to solve the following problem:

Definition. Given a $n \times d$ matrix A and $b \in \mathbb{R}^n$, the **least squares linear regression problem** is to compute

$$\arg\min_{x \in \mathbb{R}^n} \|Ax - b\|_2^2$$

We are interested in obtaining an approximation solution. Namely, given any $\epsilon > 0$, we would like to find $x' \in \mathbb{R}^n$ such that $\|Ax' - b\|_2^2 = (1 \pm \epsilon) \min_{x \in \mathbb{R}^n} \|Ax - b\|_2^2$. Our approach has been to choose a $k \times n$ random matrix S of i.i.d. normal variables distributed $\mathcal{N}(0, 1/k)$, where $k = \frac{d}{\epsilon^2}$. We want to show that with high probability, for all $x \in \mathbb{R}^n$ we have $\|S(Ax - b)\|_2^2 = (1 \pm \epsilon)\|Ax - b\|_2^2$, meaning that distances are approximately preserved under multiplication by S.

Picking up where we left off: let g be any vector of normal random variables distributed $\mathcal{N}(0, 1/k)$.

Claim 1. If $u, v \in \mathbb{R}^n$ are orthogonal vectors (i.e. $\langle u, v, \rangle = 0$), then the random variables $\langle g, u \rangle$ and $\langle g, v \rangle$ are independant.

Proof. Since u, v are orthogonal, we can fix any rotation matrix R such that $Ru = \alpha e_1$ and $Rv = \beta e_2$, where $\alpha, \beta \in \mathbb{R}$ and e_1, e_2 are standard orthonormal basis vectors. Since we know that normal variables are rotationally invariant and rotations preserve inner products, then $\langle g, v \rangle = \langle Rg, Rv, \rangle = \langle h, \alpha e_1 \rangle = \alpha h_1$, where h is also $\mathcal{N}(0, 1/k)$ and h_i is the i-th coordinate. Similarly $\langle g, v \rangle = \langle Rg, Rv \rangle = \beta h_2$. Thus $\langle g, u \rangle$ and $\langle g, v \rangle$ are both independent normally distributed random variables $\alpha h_1, \beta h_2$ as desired. ■

We use this prior claim to show:

Proposition 1. *The matrix SA is a $k \times d$ matrix of i.i.d. $\mathcal{N}(0, 1/k)$ random variables.*

Proof. First observe that we can assume the columns of A are orthonormal. The justification is as follows. We want to prove our result for all x, and in the singular value decomposition of A we have $A = U\Sigma V^T$ where U has orthonormal columns. Thus if we prove that $\|SUx - Sb\|_2^2 = (1 \pm \epsilon)\|Ux - b\|_2^2$ for any x, then setting $x = \Sigma V^T y$ for any y proves $\|S(Ay - b)\|_2^2 = (1 \pm \epsilon)\|Ay - b\|_2^2$ as desired. Then similarly by scaling we can assume the columns of A are all unit vectors.

Now the rows of SA are each of the form $\langle g, A_1 \rangle, \langle g, A_2 \rangle, \ldots, \langle g, A_d \rangle$, which are independent by the prior claim. We have $\langle g, A_i \rangle = \sum_{j=1}^n g_j A_{i,j}$. Now each g_j is $\mathcal{N}(0, 1/k)$, so $\sum_{j=1}^n g_j A_{i,j}$ is normal $\mathcal{N}(0, \frac{1}{k} \sum_{j=1}^n A_{i,j}^2)$. But each A_i is a unit vector, thus $\langle g, A_i \rangle$ is normal $\mathcal{N}(0, 1/k)$. So each entry in SA is normal $\mathcal{N}(0, 1/k)$ as desired. ■

Subspace Embeddings

We now attempt to show that applying S to the subspace spanned by a matrix A results in low distortion of norms. In other words, the column space of A is approximately preserved under multiplication by S

Definition. A matrix S is a *Subspace Embedding* if with high probability, $\forall x \in \mathbb{R}^n$ we have $\|SAx\|_2^2 = (1 \pm \epsilon)\|Ax\|_2^2$.

First note that, since we are looking for a multiplicative error, by scaling we can assume that x is a unit vector ($x \in S^{d-1}$), and again as in the last proposition, we assume A has orthonormal columns. The following is standard;

Fact 1. If A has orthonormal columns then $\|Ax\|_2^2 = \|x\|_2^2$.

Now fix any $x \in \mathbb{R}^d$. Then $\|SAx\|_2^2 = \sum_{i=1}^k \langle g_i, x\rangle^2$ where g_i is the i-th row of SA (which is normal $\mathcal{N}(0, 1/k)$ as just proved). Now $\|x\|_2 = 1$, so just as before we have that each $\langle g_i, x\rangle^2$ is distributed $\mathcal{N}(0, 1/k)^2$, and $\mathbb{E}[\langle g_i, x\rangle^2] = \frac{1}{k}$, thus $\mathbb{E}[\|SAx\|_2^2] = 1$. Since $\|Ax\|_2^2 = 1$, we want $\|SAx\|_2^2$ to be tightly concentration around its expectation, so that w.h.p. $\|SAx\|_2^2 = (1 \pm \epsilon)$. To show this concentration, we invoke a classic result:

Theorem 1 (Johnson-Lindenstrauss)**.** *Suppose $h_1, \dots, h_k$ are i.i.d. $\mathcal{N}(0,1)$, and let $G = \sum_i h_i^2$. Then for $x > 0$:*

$$\mathbf{Pr}[G > k + 2\sqrt{kx} + 2x] < e^{-x}$$

$$\mathbf{Pr}[G < k - 2\sqrt{kx}] < e^{-x}$$

Note that $\mathbb{E}[G] = k$. Additionally observe that if we want a constant factor approximation of G, then setting $x = \Theta(k)$ will give the result with probability $e^{-\Theta(k)}$. By the union bound, setting $x = \frac{\epsilon^2 k}{16}$, then G will be $(1 \pm \epsilon)k$ with probability at least $1 - 2e^{-\epsilon^2 k/16}$. Setting $k = \Theta(\epsilon^2 \log(\delta^{-1}))$ gives the result with probability at least $(1 - \delta)$.

Now how can we apply this to our problem? Since $\|SAx\|_2^2 = \sum_{i=1}^k \langle g_i, x\rangle^2$ is a sum of squared normal random variables, applying the above theorem gives us

$$\mathbf{Pr}[\|SAx\|_2^2 = (1 \pm \epsilon)] \geq 1 - 2^{\Theta(d)}$$

Which is close to the result we want. Unfortunately, this holds for only one value of x, and we need it to hold for all values of x. Since there are infinitely many, we cannot union bound over all the vectors x that we need, thus we must construct a γ-*net* and union bound over it.

γ-nets

Definition (γ-net)**.** Let $\mathcal{M}$ be any metric space, and S a subset. Then a γ-net N of S is a subset of S such that $\forall x \in S, \exists y \in N$ such that $d(x, y) \leq \gamma$.

Since we need only consider unit vectors, we now construct a γ-net for the d-dimension unit sphere S^{d-1}. To do so, we utilize the following greedy approach:

Algorithm 1: Greedy Algorithm for γ-net

Input: $\gamma > 0$
Result: A γ-net N of S^{d-1}
$N \leftarrow \emptyset$
while $\exists x \in S^{d-1}$ *that is not* γ *close to any* $y \in N$ **do**
$\quad N \leftarrow N \cup \{x\}$
end
return N

Clearly the resulting set N is a γ-net, otherwise the algorithm would not have halted. We now show that N is not too large.

Claim 2. The γ-net N produced by the above greedy algorithm satisfies $|N| \leq \frac{(1+\gamma/2)^d}{(\gamma/2)^d}$.

Proof. Let $B(x, r)$ be the open ball of radius r centered at x. Since every time we added an x to N in the algorithm, x was not contained in any ball of radius γ centered at any other point in N. Therefore the set of balls $\mathcal{B} = \{B(x, \gamma/2) \mid x \in N\}$ is pairwise disjoint (if it were not, one element of N would be contained in a ball of radius γ around another). Furthermore, the set $\mathcal{B}$ is contained within the ball $B(0, 1+\gamma/2)$, which has volume $C(1+\gamma/2)^d$, where C is some constant depending on d. Similarly, we have $\mathrm{Vol}(\mathcal{B}) = |N|C(\gamma/2)^d$, and since $\mathrm{Vol}(\mathcal{B}) \leq \mathrm{Vol}(B(0, 1+\gamma/2))$ by containment, it follows that $|N| \leq \frac{(1+\gamma/2)^d}{(\gamma/2)^d}$. ■

Now let $M = \{Ax \mid x \in N\}$ where N is the net generated by the greedy algorithm. Then M is the image of the γ-net N under multiplication by A. Clearly $|M| \leq |N|$, and in fact this holds at equality by the orthonormality assumption on A. We would now like to show that M is a γ-net for the subspace spanned by A.

Claim 3. For all $x \in S^{d-1}$, there exists a $y \in M$ so that $\|Ax - y\|_2 \leq \gamma$.

Proof. Fix such an x, and let x' be s.t. $\|x - x'\|_2 \leq \gamma$. Then $\|Ax - Ax'\| = \|x - x'\| \leq \gamma$ by orthonormality of A. Thus $y = Ax' \in M$ suffices. ■

Now let us recall where we are. We have proven for a fixed x that $\mathbf{Pr}[\|ASx\|_2^2 = (1 \pm \epsilon)] \geq 1 - 2^{-\Theta(d)}$, and accordingly for any fixed pair $x, x' \in S^{d-1}$ the values $\|SAx\|_2^2, \|SAx'\|_2^2$, and $\|SA(x - x')\|_2^2$ are preserved up to a $(1 \pm \epsilon)$ factor with probability at least $1 - 2^{-\Theta(d)}$. Now write:

$$\|SA(x - x')\|_2^2 = \|SAx\|_2^2 + \|SAx'\|_2^2 - 2\langle SAx, SAx' \rangle$$
$$\|A(x - x')\|_2^2 = \|Ax\|_2^2 + \|Ax'\|_2^2 - 2\langle Ax, Ax' \rangle$$

Because $A(x - x')$ has bounded norm, it follows that $\|SA(x - x')\|_2^2 = (1 \pm \epsilon)\|A(x - x')\|_2^2 = \|A(x - x')\|_2^2 \pm O(\epsilon)$, and the same result applies to each of $\|SAx\|_2^2, \|SAx'\|_2^2$. Thus each *norm* term in the above two equations is preserved up to an additive $O(\epsilon)$ term. It follows that

$$\mathbf{Pr}[\langle Ax, Ax' \rangle = (1 \pm \epsilon)\langle SAx, SAx' \rangle \pm O(\epsilon)] \geq 1 - 2^{-\Theta(d)}$$

Therefore, with the above probability, S preserves inner products up to an additive $O(\epsilon)$ factor (for any fixed x, x'). Now fix a 1/2-net N of S^{d-1}, and let $M = \{Ax \mid x \in N\}$ be its image under A (which is again a 1/2 net of $A(S^{d-1})$ as proven earlier). We know $|M| \leq 5^d$ by our earlier upper bound. Now by the union bound, we have

$$\mathbf{Pr}[\forall y, y' \in M, \langle y, y' \rangle = \langle Sy, Sy' \rangle \pm O(\epsilon)] \geq 1 - 2^{-\Theta(d)} \tag{1}$$

And we now condition on this event. By the linearity of the inner product, for any scalars α, β and $y, y' \in M$, we have $\langle \alpha y, \beta y' \rangle = \alpha\beta \langle Sy, Sy' \rangle \pm O(\epsilon\alpha\beta)$. Thus S preserves all inner products and scalings of vectors in our net M. Now let $y = Ax$ or any $x \in S^{d-1}$. Our goal will now to be to find a sequence of scaled vectors $y_1, y_2, \ldots$ from M whose sum converges to y. This will be done as follows:

Procedure for generating $y_1, y_2, \ldots$

1. First, pick $y_1 \in M$ such that $\|y - y_1\|_2 \leq \frac{1}{2}$, which we can do by the $\frac{1}{2}$-net property.
2. Let $\alpha > 0$ be such that $\|\alpha(y - y_1)\|_2 = 1$ (α is just $\frac{1}{\|y-y_1\|_2}$). Then $\alpha(y - y_1) \in S^{d-1}$, so
3. Let $y_2' \in M$ be such that $\|\alpha(y - y_1) - y_2'\|_2 \leq \frac{1}{2}$, which we can do again by the net property. Then because $\alpha = \frac{1}{\|y-y_1\|_2} \geq 2$, we have
$$\|y - y_1 - \frac{y_2'}{\alpha}\|_2 \leq \frac{1/2}{\alpha} \leq \frac{1}{2^2}$$
4. Set $y_2 = \frac{y_2'}{\alpha}$, and repeat to obtain $y_1, y_2, y_3, \ldots$

In general, the result of this is that $\|y - \sum_{i=1}^k y_i\| \leq \frac{1}{2^i}$, thus the sum $\sum_{i=1}^\infty y_i$ converges to y as desired. We now argue the following:

Proposition 2. *For any $x \in \mathbb{R}$, conditioned on equation (1), we have $\|SAx\|_2^2 = (1 \pm \epsilon)\|Ax\|_2$.*

Proof. Writing $y_i = (-y + y_1 + \cdots + y_i) + (y - y_1 - \cdots - y_{i-1})$, by the triangle inequality we obtain

$$\begin{aligned}
\|y_i\|_2 &\leq \| - y + y_1 + \cdots + y_i\|_2 + \|y - y_1 - \cdots - y_{i-1}\|_2 \\
&\leq \frac{1}{2^i} + \frac{1}{2^{i-1}} \\
&\leq \frac{1}{2^{i-2}}
\end{aligned}$$

Thus we have now that $y = \sum_{i=1}^\infty y_i$ and $\|y_i\|_2 \leq \frac{1}{2^{i-2}}$, so, expanding out, we write

$$\begin{aligned}
\|Sy\|_2^2 &= \|S \sum_{i=1}^\infty y_i\|_2^2 \\
&= \sum_{i=1}^\infty \|Sy_i\|_2^2 + 2\sum_{i=1}^\infty \sum_{j=1}^\infty \langle Sy_i, Sy_j \rangle \\
&= \sum_{i=1}^\infty \|y_i\|_2^2 + 2\sum_{i=1}^\infty \sum_{j=1}^\infty \langle y_i, y_j \rangle \pm O(\epsilon) \sum_{i=1}^\infty \sum_{j=1}^\infty \|y_i\|_2 \|y_j\|_2
\end{aligned}$$

But note that since $\|y_i\|_2 \leq \frac{1}{2^{i-2}}$, the sum $\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\|y_i\|_2\|y_j\|_2$ is doubly geometric, and therefore a constant So the above quantity is just

$$\sum_{i=1}^{\infty}\|y_i\|_2^2 + 2\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\langle y_i, y_j\rangle \pm O(\epsilon)$$

$$= \|y\|_2^2 \pm O(\epsilon)$$

$$= 1 \pm O(\epsilon)$$

and since this was for any $y = Ax$, where $x \in S^{d-1}$, by linearity we can scale and it follows that for all $x \in \mathbb{R}$ we have $\|SAx\|_2^2 = (1 \pm \epsilon)\|Ax\|_2$, which completes the proof. ■

Back to Regression

So we have shown that S is a subspace embedding. We now come back to our problem of regression, namely finding x such that $\|Ax - b\|_2 \leq (1+\epsilon)\min_{y\in\mathbb{R}^n}\|Ay - b\|_2$.

Theorem 2. *If S is a random $k \times n$ matrix of i.i.d. $\mathcal{N}(0, 1/k)$ normal variables, then with probability $1 - 2^{-\Theta(d)}$ we have $\min_{x\in\mathbb{R}^n}\|S(Ax - b)\|_2 \leq (1+\epsilon)\min_{x\in\mathbb{R}^n}\|(Ax - b)\|_2$.*

Proof. Since A was any matrix in the prior arguement, we now consider the subspaced spanned by both A and b, and let y be any vector in this subspace. By the subspace embedding property of S, we have $\|Sy\|_2 = (1 \pm \epsilon)\|y\|_2$, thus $\|S(Ax - b)\|_2 = (1 \pm \epsilon)\|Ax - b\|_2$ for all $x \in \mathbb{R}^n$. Thus $\min_{x\in\mathbb{R}^n}\|S(Ax-b)\|_2 \leq (1+\epsilon)\min_{x\in\mathbb{R}^n}\|(Ax-b)\|_2$, as desired. So by solving $\arg\min_{x\in\mathbb{R}^n}\|S(Ax - b)\|_2$, we obtain a $(1+\epsilon)$ approximate solution to the regression problem. ■

Choosing the right sketching matrix S

We have now shown that solving the problem $\arg\min_{x\in\mathbb{R}^n}\|S(Ax - b)\|_2$ gives us a adequate approximate solution to our regression problem. Unfortunately, computing the product SA can take $O(nd^2)$ time. Since we can solve the problem exactly in same time, we have seemingly gotten nowhere. However, if we cleverly choose S from a family of random matrices which still satisfies the subspace embedding properties we have just shown for $\mathcal{N}(0, 1/k)$ matrices here, then we may be able to do better. Namely, we will choose an S such that the computation SA can be done in $O(nd\log(n))$ time, which is an improvement for $d = \omega(\log(n))$. We first introduce a matrix with useful symmetry properties.

Definition. For $n = 2^k$, the $n \times n$ Hadamard matrix H is defined by:

$$H_{i,j} = \frac{1}{\sqrt{n}}(-1)^{\langle i,j\rangle}$$

where $\langle i, j\rangle$ is the dot product of k-bit binary represetions of i and j over the field $\mathbb{F}_2$.

Now let D be a diagonal $n \times n$ matrix of random ± 1 entries. We claim:

Claim 4. The family of matrices $S = PHD$, where P is a matrix which selects a random subset of rows of HD, satisfies the subspace embedding property.

To begin, we first prove the following fact:

Proposition 3. *The rows of the Hadamard matrix H are orthonormal.*

Proof. Let H_i be the i-th row of H. First note that for any $i \neq j$

$$\langle H_i, H_j \rangle = \sum_{\ell=1}^{n} H_{i,\ell} H_{j,\ell}$$

$$= \frac{1}{n} \sum_{\ell=1}^{n} (-1)^{\langle \ell, i+j \rangle}$$

Now since $i \neq j$, we can fix a coordinate $q \in [k]$ such that $i_q \neq j_q$. Thus $(i+j)_q = 1$. Now consider any $\ell \in [n]$, and let $\ell' \in [n]$ be the value ℓ but with the q-th bit flipped. Then $(-1)^{\langle \ell, i+j \rangle} + (-1)^{\langle \ell', i+j \rangle} = 0$, since the values of the dot product are all the same except for the q-th position, where they differ. Thus each value $\ell \in [n]$ cancels with the value $\ell' \in [n]$ for which the q-th bit is flipped. Thus $\sum_{\ell=1}^{n} (-1)^{\langle \ell, i+j \rangle} = 0$, so $\langle H_i, H_j \rangle = 0$, which proves that the columns of H are pairwise orthogonal. ■