COMPSCI 690RA: Problem Set 3

Due: 4/15 by 8pm in Gradescope.

Instructions:

- You are allowed to, and highly encouraged to, work on this problem set in a group of up to three members.
- Each group should **submit a single solution set**: one member should upload a pdf to Gradescope, marking the other members as part of their group in Gradescope.
- You may talk to members of other groups at a high level about the problems but **not work through the solutions in detail together**.
- You must show your work/derive any answers as part of the solutions to receive full credit.

1. Tighter Bounds for Trace Estimation (4 points)

Consider any matrix $A \in \mathbb{R}^{n \times n}$. Use the Hanson-Wright inequality to show that if $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \{-1, 1\}^n$ are chosen to have independent and uniformly distributed ± 1 entries, then for $m = O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$, $\bar{\mathbf{T}} = \frac{1}{m}\sum_{i=1}^m \mathbf{x}_i^T A \mathbf{x}_i$ satisfies,

$$\Pr\left[|\bar{\mathbf{T}} - \mathrm{tr}(A)| > \epsilon \|A\|_F\right] \leq \delta.$$

How does this compare to the bound proven in class using Chebyshev's inequality?

Let $B \in \mathbb{R}^{mn \times mn}$ be the block matrix with m on-diagonal blocks equal to $\frac{1}{m} \cdot A$. Let $\mathbf{x} \in \mathbb{R}^{nm}$ be the concatenation of $\mathbf{x}_1, \ldots, \mathbf{x}_m$. Then we can check that

$$\mathbf{x}^T B \mathbf{x} = \frac{1}{m}\sum_{i=1}^m \mathbf{x}_i^T A \mathbf{x}_i = \bar{\mathbf{T}}.$$

Now, $\mathrm{tr}(B) = m \cdot \mathrm{tr}(\frac{1}{m}) = \mathrm{tr}(A)$. Also $\|B\|_F^2 = m \cdot \|\frac{1}{m}A\|_F^2 = \frac{1}{m}\|A\|_F^2$. Further, $\|B\|_2 = \|\frac{1}{m}A\|_2 = \frac{1}{m}\|A\|_2 \leq \frac{1}{m}\|A\|_F$. So applying Hanson-Wright:

$$\begin{aligned}
\Pr\left[|\bar{\mathbf{T}} - \mathrm{tr}(A)| > \epsilon\|A\|_F\right] &= \Pr\left[|\mathbf{x}^T B \mathbf{x} - \mathrm{tr}(B)| > \epsilon \|A\|_F\right] \\
&\leq 2\exp\left(-c \cdot \min\left\{\frac{\epsilon^2\|A\|_F^2}{\|B\|_F^2}, \frac{\epsilon\|A\|_F}{\|B\|_2}\right\}\right) \\
&\leq 2\exp\left(-c \cdot \min\left\{\frac{\epsilon^2 m\|A\|_F^2}{\|A\|_F^2}, \frac{\epsilon m\|A\|_F}{\|A\|_F}\right\}\right) \\
&\leq 2\exp(-c \cdot \epsilon^2 m).
\end{aligned}$$

Thus, if we set $m = O(\log(1/\delta)/\epsilon^2)$ this probability is upper bounded by δ. This bound is similar to the one given by Chebyshev's inequality, but the dependence on $1/\delta$ is much better – logarithmic rather than linear. The proof via Hanson-Wright is also arguably much simpler.

2. Matrix Concentration from Scratch (8 points)

Consider a random symmetric matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$ where $\mathbf{M}_{ij} = \mathbf{M}_{ji}$ is set independently to 1 with probability 1/2 and -1 with probability 1/2. Let $\|\mathbf{M}\|_2 = \max_{x:\|x\|=1} \|\mathbf{M}x\|_2$ be the spectral norm of $\mathbf{M}$. Recall that $\|\mathbf{M}\|_2$ is equal to the largest singular value of $\mathbf{M}$, which equals the largest magnitude of one of its eigenvalues.

1. (2 points) Give upper and lower bounds on $\|\mathbf{M}\|_2$ that hold deterministically – i.e., for any random choice of the entries of $\mathbf{M}$. **Hint:** You'll probably want to use $\|\mathbf{M}\|_F$, and its relation to the singular values to derive your bounds.

   We have $\|\mathbf{M}\|_2 = \sigma_1(\mathbf{M}) \leq \|\mathbf{M}\|_F = \left(\sum_{i=1}^n \sigma_i(\mathbf{M})^2\right)^{1/2}$. Additionally, we always have $\|\mathbf{M}\|_F = \sqrt{n^2} = n$. Thus, we always have $\|\mathbf{M}\|_2 \leq n$.

   Similarly, we have $\|\mathbf{M}\|_2^2 = \sigma_1(\mathbf{M})^2 \geq \frac{\|\mathbf{M}\|_F^2}{n} = \frac{\sum_{i=1}^n \sigma_i(\mathbf{M})^2}{n}$. I.e., the largest squared singular value is larger than the average squared singular value. Thus, $\|\mathbf{M}\|_2^2 \geq n^2/n = n$. So $\|\mathbf{M}\|_2 \geq \sqrt{n}$ always.

2. (2 points) Observe that you can also write $\|\mathbf{M}\|_2 = \max_{x:\|x\|=1} |x^T\mathbf{M}x|$. Show that for any $x \in \mathbb{R}^n$ with $\|x\|_2 = 1$, with probability $\geq 1-\delta$, $|x^T\mathbf{M}x| = c\sqrt{\log(1/\delta)}$ for some constant c.

   Hint: Use Hoeffding's inequality, which is a useful variant on the Bernstein inequality. For independent random variables $\mathbf{X}_1, \ldots, \mathbf{X}_n$, and scalars $a_1, \ldots, a_n, b_1, \ldots, b_n$ with $\mathbf{X}_i \in [a_i, b_i]$, $\Pr\left[\left|\sum_{i=1}^n \mathbf{X}_i - \mathbb{E}[\sum_{i=1}^n \mathbf{X}_i]\right| \geq t\right] \leq 2\exp\left(\frac{-2t^2}{\sum_{i=1}^n (b_i-a_i)^2}\right)$.

   For some fixed $x \in \mathbb{R}^n$ with $\|x\|_2^2 = 1$, we have $x^T\mathbf{M}x = \sum_{i=1}^n x(i)^2 \cdot \mathbf{M}_{ii} + \sum_{i\neq j} 2x(i)x(j)\mathbf{M}_{ij}$. Note that since the $\mathbf{M}_{ij}$ terms are all independent, the term in these sum are all idependent. We have $\mathbb{E}[\mathbf{M}_{ij}] = 0$ for all i,j so $\mathbb{E}[x^T\mathbf{M}x] = 0$. Additionally, each term in the first sum is bounded in the range $[-x(i)^2, x(i)^2]$ and in the second in the range $[-2x(i)x(j), 2x(i)x(j)]$. We can bound the sum of squared widths of these ranges as:

   $$\begin{aligned}\sum_{i=1}^n x(i)^4 + 4\sum_{i\neq j}^n x(i)^2x(j)^2 &\leq 2\sum_{i=1}^n x(i)^4 + 4\sum_{i\neq j}^n x(i)^2x(j)^2\\ &= 2\left(\sum_{i=1}^n x(i)^2\right)^2 \leq 2,\end{aligned}$$

   where the final bound follows from the fact that $\|x\|_2^2 = \sum_{i=1}^n x(i)^2 = 1$. Applying Hoeffding's inequality we then have:

   $$\Pr\left[\left|x^T\mathbf{M}x\right| \geq t\right] \leq 2\exp\left(\frac{-2t^2}{2}\right) \leq 2\exp(-t^2).$$

   Thus, if we set $t = \sqrt{\ln(2/\delta)} = c\sqrt{\ln(1/\delta)}$ for some constant c, this probability is bounded by $2\cdot\exp(-\ln(2/\delta)) = 2\cdot\delta/2 = \delta$, as required.

3. (4 points) Prove that with probability $1 - \frac{1}{n^{c_1}}$, $\|\mathbf{M}\|_2 \leq c_2\sqrt{n\log n}$ for some fixed constants c_1, c_2. **Hint:** Use an ϵ-net and part (1).

Let $\mathcal{N}$ be a $\frac{1}{n}$-net for the unit ℓ_2 fall in $\mathbb{R}^n$. We can construct $\mathcal{N}$ with $|\mathcal{N}| \leq (4n)^n$. Then applying part (2) with $\delta = \frac{1}{n^{c_1}\cdot|\mathcal{N}|}$, via a union bound, with probability at least $1-1/n^{c_1}$, for all $x' \in \mathcal{N}$,

$$x'^T\mathbf{M}x \leq c\sqrt{\log(1/\delta)} = c\sqrt{\log(4n^n \cdot n^{c_1})} = c_2\sqrt{n\log n},$$

for some constant c_2. Now, for any $x \in \mathbb{R}^n$, there is some $x' \in \mathcal{N}$ with $\|x - x'\|_2 \le \frac{1}{n}$. Let $e \stackrel{\text{def}}{=} x - x'$ so $\|e\| \le 1/n$. We have:

$$\begin{aligned} x^T \mathbf{M} x &= (x' + e)^T M (x' + e) = x'^T M x + 2e^T M x' + e^T M e \\ &\le c_2 \sqrt{n \log n} + 2e^T M x' + e^T M e. \end{aligned}$$

Now, $e^T M x' \le \|e\|_2 \cdot \|M\|_2 \cdot \|x'\|_2 \le \frac{1}{n} \cdot n$, since x' is unit norm and via our bound in part (1). Similarly, $e^T M e \le \|e\|_2^2 \cdot \|M\| \le \frac{1}{n}$. Plugging into our bound above we have:

$$\begin{aligned} x^T \mathbf{M} x &\le c_2 \sqrt{n \log n} + 2 + \frac{1}{n} \\ &\le (c_2 + 3)\sqrt{n \log n}, \end{aligned}$$

assuming that $n \log n > 1$. This completes the proof.

3. Randomized Preconditioning (12 points)

One way that subspace embeddings are often used in practice are within *preconditioned iterative methods* for linear regression. Here we will see how to analyze one such method. Given $A \in \mathbb{R}^{n \times d}$ and $b \in \mathbb{R}^n$, the goal is to find an approximate minimizer $x \in \mathbb{R}^d$ of the least squares loss function $\|Ax - b\|_2^2$.

1. (2 points) Assume that $\mathbf{S} \in \mathbb{R}^{m \times n}$ is an 1/4-subspace embedding for $A \in \mathbb{R}^{n \times d}$. I.e., for all $x \in \mathbb{R}^d$, $\frac{3}{4}\|Ax\|_2 \le \|\mathbf{S}Ax\|_2 \le \frac{5}{4}\|Ax\|_2$. Prove that all eigenvalues of $(A^T \mathbf{S}^T \mathbf{S} A)^{-1} A^T A$ lie in the range $[1/2, 2]$.

   Hint: You may assume that $A^T A$ has full rank. You may also want to use that for any two matrices $M, N \in \mathbb{R}^{d \times d}$, the non-zero eigenvalues of MN are equal to those of NM.

   Observe that since $A^T A$ is full rank, all its eigenvalues are non-zero and thus $x^T A^T A x > 0$ for all x. Thus, the eigenvalues of $A^T \mathbf{S}^T \mathbf{S} A$ must also be non-zero, since otherwise there would be some x with $\|\mathbf{S}A\|_2^2 = x^T A^T \mathbf{S}^T \mathbf{S} A x = 0$, violating the subspace embedding guarantee.

   Suppose for contradiction that $(A^T \mathbf{S}^T \mathbf{S} A)^{-1} A^T A$ had some (non-zero) eigenvalue $\lambda \notin [1/2, 2]$. Then by the hint, $(A^T \mathbf{S}^T \mathbf{S} A)^{-1/2} A^T A (A^T \mathbf{S}^T \mathbf{S} A)^{-1/2}$ also has some non-zero eigenvalue $\lambda \notin [1/2, 2]$. Here $(A^T \mathbf{S}^T \mathbf{S} A)^{-1/2}$ is the matrix with $(A^T \mathbf{S}^T \mathbf{S} A)^{-1/2} (A^T \mathbf{S}^T \mathbf{S} A)^{-1/2} = (A^T \mathbf{S}^T \mathbf{S} A)^{-1}$. It can be obtained e.g. by writing $(A^T \mathbf{S}^T \mathbf{S} A)^{-1} = V \Lambda V^T$ in its eigendecomposition and then letting $(A^T \mathbf{S}^T \mathbf{S} A)^{-1/2} = V \Lambda^{1/2} V^T$.

   This means that there is some eigenvector v with $\|v\|_2 = 1$ such that

   $$v^T (A^T \mathbf{S}^T \mathbf{S} A)^{-1/2} A^T A (A^T \mathbf{S}^T \mathbf{S} A)^{-1/2} v \notin [1/2, 2].$$

   Let $y = (A^T \mathbf{S}^T \mathbf{S} A)^{-1/2} v$.Then this means that

   $$\|Ay\|_2^2 = y^T A^T A y \notin [1/2, 2].$$

   But observe that

   $$\|\mathbf{S}Ay\|_2^2 = y^T (A^T \mathbf{S}^T \mathbf{S} A) y = z^T z = \|z\|_2^2 = 1.$$

   Thus, $\|Ay\|_2^2 \notin \left[\frac{1}{2} \cdot \|\mathbf{S}Ay\|_2^2, 2\|\mathbf{S}Ay\|_2^2\right]$. But this contradicts the subspace embedding guarantee which ensures that

   $$\|Ay\|_2^2 \le \left(\frac{4}{3}\right)^2 \|\mathbf{S}Ay\|_2^2 \le 2\|\mathbf{S}Ay\|_2^2$$

and

$$\|Ay\|_2^2 \geq \left(\frac{4}{5}\right)^2 \|\mathbf{S}Ay\|_2^2 \geq \frac{1}{2}\|\mathbf{S}Ay\|_2^2.$$

2. (2 points) Consider solving least squares regression iteratively, starting with some guess $x_0 \in \mathbb{R}^d$ and repeatedly applying the iteration $x_{i+1} = x_i - \eta A^T(Ax_i - b)$, where $\eta \in (0,1)$ is some step size. Let $x_* = \arg\min_{x\in\mathbb{R}^d} \|Ax-b\|_2^2$. Prove that this iteration is equivalent to:

$$x_{i+1} = (I - \eta A^T A)(x_i - x_*) + x_*.$$

Hint: Prove that $A^T A x_* = A^T b$.

We first prove the hint that $A^T A x_* = A^T b$. There are many ways to prove this. One way is to observe that since $\|Ax-b\|_2^2 = x^T A^T A x - 2x^T A^T b + b^T b$, the gradient of this function is $\nabla\|Ax-b\|_2^2 = 2A^T Ax - 2A^T b$. At an optimum, this gradient must be 0, so we must have $2A^T A x_* - 2A^T b = 0 \implies A^T A x_* = A^T b$.

Now, using that $A^T A x_* = A^T b$, we have:

$$\begin{aligned} x_{i+1} &= x_i - \eta A^T(Ax_i - b) \\ &= x_i - \eta A^T A x_i + \eta A^T A x_* \\ &= (I - \eta A^T A)(x_i - x_*) + x_*, \end{aligned}$$

which gives the claim.

3. (2 points) Let $\lambda_{\max}(A^T A), \lambda_{\min}(A^T A)$ be the largest and small eigenvalues of $A^T A$ respectively, and let $\kappa = \frac{\lambda_{\max}(A^T A)}{\lambda_{\min}(A^T A)}$. Prove that if we set $\eta = \frac{1}{\lambda_{\max}(A^T A)}$, then the t^{th} iterate satisfies:

$$\|x_t - x_\star\|_2 \leq \left(1 - \frac{1}{\kappa}\right)^t \cdot \|x_0 - x_\star\|_2.$$

Hint: Bound the eigenvalues of $I - \eta A^T A$.

The eigenvalues of $\eta A^T A$ are equal to η times the eigenvalue of $A^T A$, which lie in the range $[\lambda_{\min}(A^T A), \lambda_{\max}(A^T A)]$. Thus, for $\eta = \frac{1}{\lambda_{\max}(A^T A)}$, the eigenvalues of $\eta A^T A$ lie in the range $[1/\kappa, 1]$. In turn, the eigenvalues of $I - \eta A^T A$ lie in the range $[0, 1 - 1/\kappa]$.

Thus, $\|(I - \eta A^T A)(x_i - x_*)\|_2 \leq (1 - 1/\kappa)\|x_i - x_*\|_2$. Using part (2) we can thus conclude:

$$\|x_t - x_*\|_2 = \|(I - \eta A^T A)(x_{t-1} - x_*) + x_* - x_*\|_2 \leq (1 - 1/\kappa)\|x_{t-1} - x_*\|_2.$$

Thus,

$$\|x_t - x_*\|_2 \leq (1 - 1/\kappa)^t \cdot \|x_0 - x^*\|_2.$$

4. (2 points) Use the above to show for any $\epsilon \geq 0$, after $t = O(\kappa \cdot \log(1/\epsilon))$ iterations, the t^{th} iterate satisfies $\|x_t - x_\star\|_2 \leq \epsilon\|x_\star\|_2$, assuming that we initialize with $x_0 = 0$.

Setting $x_0 = \vec{0}$, after t iterations we have

$$\begin{aligned} \|x_t - x_\star\|_2 &\leq \left(1 - \frac{1}{\kappa}\right)^t \cdot \|x_0 - x_\star\|_2 \\ &= \left(1 - \frac{1}{\kappa}\right)^t \cdot \|x_\star\|_2. \end{aligned}$$

Now, $(1-1/\kappa)^\kappa \leq e^{-1}$ so $(1-1/\kappa) \leq e^{-1/\kappa}$. Thus, $(1-1/\kappa)^t \leq e^{-t/\kappa}$. So if we set $t = \kappa \cdot \ln(1/\epsilon)$ we will have $\|x_t - x_\star\|_2 \leq \epsilon \|x_\star\|_2$.

5. (2 points) κ is known as the condition number of $A^T A$, and when it is large, the performance of this, and many other iterative methods for linear regression degrade. To avoid this we will instead consider a *preconditioned* update: let $\mathbf{S} \in \mathbb{R}^{m \times n}$ be random sketching matrix. And update: $x_{i+1} = x_i - \eta(A^T \mathbf{S}^T \mathbf{S} A)^{-1} A^T (Ax_i - b)$. Following the analysis above, and using part (1), show that if $\mathbf{S}$ is an 1/4-subpsace embedding for A, then this preconditioned method with an appropriately chosen η, has $\|x_t - x_\star\|_2 \leq \epsilon \|x_\star\|_2$ after $t = O(\log(1/\epsilon))$ iterations. That is, there is no dependence on κ.

   Following (2) we have:

   $$x_{i+1} = (I - \eta(A^T \mathbf{S}^T \mathbf{S} A)^{-1} A^T A)(x_i - x_*) + x_*.$$

   Now, by part (1), the eigenvalues of $(A^T \mathbf{S}^T \mathbf{S} A)^{-1} A^T A$ lie in $[1/2, 2]$. Thus, if we set $\eta = 1/2$, the eigenvalues of $(I - \eta(A^T \mathbf{S}^T \mathbf{S} A)^{-1} A^T A)$ lie in $[0, 3/4]$. So $\|(I - \eta(A^T \mathbf{S}^T \mathbf{S} A)^{-1} A^T A)(x_i - x_*)\|_2 \leq 3/4 \cdot \|x_i - x_*\|_2$. Following the analysis of part (3) we thus have:

   $$\|x_t - x_*\|_2 \leq (3/4)^t \|x_0 - x_*\|_2.$$

   Setting $x_0 = 0$ and $t = O(\log(1/\epsilon))$, we have $\|x_t - x_*\|_2 \leq \epsilon \|x_*\|_2$ as required.

6. (2 points) How large must m be so that $\mathbf{S}$ satisfies the required subspace embedding property with probability at least 99/100? Assuming that $\mathbf{S}A \in \mathbb{R}^{m \times d}$ is already computed, how long does it take to compute $(A^T \mathbf{S}^T \mathbf{S} A)^{-1}$? And how long does each iteration of the preconditioned method take? How does this compare to the non-preconditioned method? How about to directly solving the system using an exact method? Assume that $n \gg d$ in your discussion.

By the analysis in class, we need $m = O\left(\frac{d + \log(1/\delta)}{\epsilon^2}\right) = O(d)$ when $\delta = 1/100$ and $\epsilon = 1/4$. Given this, with $\mathbf{S}A$ in hand, it requires $O(d^3)$ time to compute $A^T \mathbf{S}^T \mathbf{S} A$ and further $O(d^3)$ time to invert this matrix (e.g., via Gaussian elimination).

Once this inverse is computed, each iteration of the preconditioned method requires $O(nd)$ time to compute $Ax_i - b$, then $O(nd)$ time to multiply this vector by A^T to get $A^T(Ax_i - b)$, then $O(d^2)$ time to multiple by $(A^T \mathbf{S}^T \mathbf{S} A)^{-1}$ and $O(d)$ time to subtract from x_i. So the time per iteration is $O(nd + d^2) = O(nd)$, assuming that $n \geq d$. This is the same asymptotic runtime as required for the non-preconditioned algorithm per iteration.

Overall this gives runtime $O(d^3 + nd \log(1/\epsilon))$ for the preconditioned method vs. $O(nd\kappa \log(1/\epsilon))$ for the non-preconditioned method. Thus the preconditioned method can be much faster when $n \gg d$ and when κ is large. Solving the system directly would require $O(nd^2)$ time, which, except for very small ϵ will always be slower than the pre-conditioned method, typically much slower.

4. Compressed Sensing From Subspace Embedding (6 points)

Given a vector $x \in \mathbb{R}^n$ and a random matrix $\mathbf{S} \in \mathbb{R}^{m \times n}$, consider computing $\mathbf{y} = \mathbf{S}x$. If $m < n$, you can in general not determine $x \in \mathbb{R}^n$ from $\mathbf{y} \in \mathbb{R}^m$, since $\mathbf{S}$ is not an invertible map. Here, we will argue that you can recover x, assuming that it is k-sparse for small enough k. I.e., that it has at most k nonzero entries. This is known as *compressed sensing* or *sparse recovery*.

1. (2 points) Assume that $\mathbf{S}$ satisfies the distributional JL lemma/subspace embedding theorem proven in class. I.e., for any $A \in \mathbb{R}^{n \times d}$, if $m = O\left(\frac{d+\log(1/\delta)}{\epsilon^2}\right)$, then with probability at least $1-\delta$, $\mathbf{S}$ is an ϵ-subspace embedding for A. Prove that if $m = O\left(\frac{k\log(n/k)+\log(1/\delta)}{\epsilon^2}\right)$, with probability $\geq 1-\delta$, for all $z \in \mathbb{R}^n$ such that z is k-sparse, $(1-\epsilon)\|z\|_2 \leq \|\mathbf{S}z\|_2 \leq (1+\epsilon)\|z\|_2$.
Hint: Show that with high probability, $\mathbf{S}$ is an ϵ-subspace embedding simultaneously for $\binom{n}{k}$ different matrices.

   Let $\mathcal{I}$ be the set of all matrices $I' \in \mathbb{R}^{n \times k}$ consisting of a subset of k columns of the identity matrix – i.e. k standard basis vectors. $|\mathcal{I}| = \binom{n}{k} \leq \left(\frac{en}{k}\right)^k$ by `https://www.johndcook.com/blog/2008/11/10/bounds-on-binomial-coefficients/`. Thus, by a union bound, for

$$m = O\left(\frac{k + \log\left(\frac{\left(\frac{en}{k}\right)^k}{\delta}\right)}{\epsilon^2}\right) = O\left(\frac{k\log(n/k) + \log(1/\delta)}{\epsilon^2}\right),$$

   with probability at least $1-\delta$, $\mathbf{S}$ is an ϵ-subspace embedding for all $I' \in \mathcal{I}$.

   Now, observe that for any k-sparse z, if we just let $z' \in \mathbb{R}^k$ be z restricted to its non-zero entries, $z = I'z'$ where I' contains the standard basis vectors corresponding to those non-zero entries. Similarly, $\mathbf{S}z = \mathbf{S}I'z'$. Thus, by the above subspace embedding property, with probability $\geq 1-\delta$, for all k-sparse $z \in \mathbb{R}^n$:

$$(1-\epsilon)\|I'z'\|_2 \leq \|\mathbf{S}I'z'\|_2 \leq (1+\epsilon)\|I'z'\|_2 \implies (1-\epsilon)\|z\|_2 \leq \|\mathbf{S}z\|_2 \leq (1+\epsilon)\|z\|_2,$$

   as required.

2. (2 points) Use the above result, applied with $k' = 2k$, to show that if $m = O(k\log(n/k) + \log(1/\delta))$, and $x \in \mathbb{R}^n$ is k-sparse, then with probability $\geq 1-\delta$, x can be recovered exactly from $\mathbf{y} = \mathbf{S}x$.
Hint: Consider solving the equation $\mathbf{y} = \mathbf{S}x$, under the restriction that x is k-sparse. Show that there is a unique solution.

   It suffices to show that for any $\mathbf{y} = \mathbf{S}x$, there is a unique solution for x under the assumption that x is k-sparse. Suppose for contradiction there are two k-sparse solutions x and x' with $\mathbf{y} = \mathbf{S}x = \mathbf{S}x'$. Then $\mathbf{S}(x-x') = 0$. Further, $\|x-x'\|_2 > 0$ since $x \neq x'$, and $x-x'$ is at most a $2k$-sparse vector.

   By part (1), for $m = O(k\log(n/k) + \log(1/\delta))$, with probability $\geq 1-\delta$, for all $2k$-sparse vectors $z \in \mathbb{R}^n$,

$$1/2\|z\|_2 \leq \|\mathbf{S}z\|_2 \leq 3/2\|z\|_2.$$

   Thus, since $\|x-x'\|_2 > 0$, $\|\mathbf{S}(x-x')\|_2 > 0$. But this is a contradiction since we have claimed that $\mathbf{S}(x-x') = 0$.

3. (2 points) Argue that the above result is nearly optimal in terms of how much x is compressed. In particular, prove that for any function $f : \mathbb{R}^n \to \{0,1\}^{o(k\log(n/k))}$, given $f(x)$ for some k-sparse $x \in \mathbb{R}^n$, one cannot recover x uniquely, even under the assumption that all entries of x are either 0 or 1.

   We argue this via the pigeonhole principal. $f(x)$ can take $2^{o(k\log(n/k))} = o\left(\left(\frac{n}{k}\right)^k\right)$ possible values. However, there are $\binom{n}{k} \geq \left(\frac{n}{k}\right)^k$ possible k-sparse vectors with entries equal to 0,1. Thus, there must be two such vectors x, x' with $f(x) = f(x')$. So, we cannot uniquely recover a k-sparse vector from the value of f applied to that vector.