15-850: Advanced Algorithms **CMU, Spring 2024**

HW #3 (out Monday March 4, 2023) **Due:** Monday Mar 25, 2023

Same as HW2: Collaboration in a group of 2-3 is encouraged. Please solve *two* of the four problems.

1. **Nearly Orthonormal Vectors.** Call a set of unit vectors "near-orthonormal" if the inner product of any two of them is close to zero. In this problem we will show that while there are at most d orthonormal vectors in $\mathbb{R}^d$, there can be exponentially more near-orthonormal vectors. For vectors $x, y \in \mathbb{R}^d$, we use $\langle x, y\rangle = \sum_{i=1}^{d} x_i y_i$ to denote the inner product.

   (a) Let $x = (x_1, x_2, \ldots, x_d)$ and $y = (y_1, y_2, \ldots, y_d)$ be two independently and uniformly chosen vectors in $\{-1, 1\}^d$. (I.e., each bit x_i and y_i in each vector is independently and uniformly chosen from $\{-1, 1\}$.) Show that

   $$\mathbf{Pr}[|\langle x, y\rangle| \geq \varepsilon d] \leq 2\exp\left(-\varepsilon^2 d/6\right)$$

   (b) Given any constant $\varepsilon > 0$, a set S of unit vectors is called *ε-orthonormal* if for all $\vec{x}, \vec{y} \in S$,

   $$|\langle \vec{x}, \vec{y}\rangle| \leq \varepsilon.$$

   Show that there exist constants $c, d_0 > 0$ (possibly depending on ε) such that for any $d \geq d_0$, if you sample $N := \exp(c\varepsilon^2 d)$ random vectors independently and uniformly from the set $\{-\frac{1}{\sqrt{d}}, +\frac{1}{\sqrt{d}}\}^d$, this sampled set is ε-orthonormal with probability at least 1/2.

2. **An Approximate Counter, and the Median-of-Means Estimator.** Here is a way of maintaining an approximate counter. (Call this the *basic* counter.)

   Start with $X \leftarrow 0$. When an element arrives, increment X by 1 with probability 2^{-X}. When queried, return $N := 2^X - 1$.

   (a) Suppose the actual count is n, show that $\mathbf{E}[N] = n$, and $\mathbf{Var}(N) = \frac{n(n-1)}{2}$.

   Since its variance is large, average k independent basic counters $N_1, N_2, \ldots, N_k$, and output the sample average $\widehat{N} := \frac{1}{k}\sum_i N_i$. Call this the *k-mean counter.*

   (b) (Do not submit) Show that $\mathbf{Pr}[\widehat{N} \notin (1 \pm \varepsilon)n] \leq \frac{1}{2\varepsilon^2 k}$.

   Hence using $k = \frac{1}{2\varepsilon^2\delta}$ counters can make the failure probability at most δ. (I.e., your error is less than εn with "confidence" $1 - \delta$.) Here's a way to use only $K = O(\frac{1}{\varepsilon^2}\log\frac{1}{\delta})$ counters to get the same answer (and the approach is useful in many different contexts beyond this one). We call this counter the *median-of-means counter.*

   (c) Suppose Y is a real-valued random variable and let $I \subseteq \mathbb{R}$ denote an interval. Suppose $\mathbf{Pr}[Y \notin I] \leq 1/4$.

   Now, take a collection of ℓ-many independent copies of Y and let M denote the median of $Y_1, \ldots, Y_\ell$. Show that by taking $\ell = \Theta(\log(1/\delta))$, we get $\mathbf{Pr}[M \notin I] \leq \delta$. *Hint: what must happen if the median is too high? What is the chance of that?*

   (d) Using (c), conclude that by taking Y to be the k_0-mean counter from part (b) with $k_0 = \Theta(1/\epsilon^2)$, we have $\mathbf{Pr}[M \notin (1 \pm \epsilon)n] \leq \delta$.

3. **(The Fast and The Calm.)** We want to give a fast implementation of the Johnson-Lindenstrauss transform from $\mathbb{R}^D \to \mathbb{R}^k$. Assume for simplicity that D is a power of 2. All norms in this section are ℓ_2-norms, unless otherwise specified.

   (a) (Do not submit.) If a randomized algorithm produces A such that $\mathbf{Pr}[\|Ax\|^2 \in (1 \pm \varepsilon)\|x\|] \geq 1 - 1/n$ for any fixed unit vector x, then A has $1 - 1/n$ non-zero columns in expectation. (Hint: what happens for sparse vectors x?)

   (b) (Do not submit.) Define the Walsh-Hadamard matrices H_t as follows: $H_0 = (1)$, and $H_t = \left(\begin{smallmatrix} H_{t/2} & H_{t/2} \\ -H_{t/2} & H_{t/2} \end{smallmatrix}\right)$. Show that the rows and columns of H_D are orthogonal, and have ℓ_2-length $\sqrt{D}$.

   (c) *Spreading the mass around.* Define the "flip" matrix F, which is diagonal with each diagonal entry being an independent Rademacher (± 1 with probability half each). For any unit vector $x \in \mathbb{R}^D$, define $y := \frac{1}{\sqrt{D}} HFx$. Show that $\|y\| = 1$. Moreover, use a Chernoff bound to show that there exists some constant $c > 0$ such that

$$\mathbf{Pr}\left[\exists i \in [D] \; s.t. \; y_i \geq c\sqrt{\frac{\log(nD)}{D}}\right] \leq 1/n^2.$$

   (d) *Flattening "spread" vectors.* Suppose $y \in \mathbb{R}^D$ has $\|y\|_2 = 1$ and $\|y\|_\infty = \max_i |y_i| \leq a$ for some $a > 0$. Define

$$q = \min(1, \Theta(a \log n)),$$

   and $k = \Theta(\log n/\varepsilon^2)$, as in the JL theorem. Construct matrix $M \in \mathbb{R}^{D\times k}$ with each entry being an independent $N(0, 1/q)$ with probability q, and zero otherwise. Define $Ay = \frac{1}{\sqrt{k}} My$. Show that $\|Ay\|^2 \in (1 \pm \varepsilon)$ with high probability.

   (e) (Do not submit.) Combine the above two parts to show that the linear transformation

$$\Phi(x) := \frac{1}{\sqrt{D}} AHFx$$

   is map $\mathbb{R}^D \to \mathbb{R}^k$ which preserves distances with high probability. Moreover, A has $O(k \log n)$ non-zero entries, with high probability.

   Finally, using the fact that multiplying by H can be done in $O(D \log D)$ time (you don't have to prove this, of course), and that multiplying a vector y by a sparse matrix can be done fast too, show that $\Phi(x)$ can be computed in $O(d \log d + k \log n)$ time.

4. **(Chernoff meets Matrices.)** In Lecture 13 we mentioned a very general theorem about matrix-valued Chernoff bounds for symmetric matrices. In this problem we'll take the first steps towards it. Assume eigenvalues are numbered so that $\lambda_1 \geq \ldots \geq \lambda_n$. Given a symmetric matrix X, define the matrix exponential e^X by its Taylor series expansion $e^X = I + X + \frac{1}{2!}X^2 + \frac{1}{3!}X^3 + \cdots$, which you may assume always converges. We'll prove:

   Theorem 1. *Let $X_1, X_2, \ldots, X_n$ be independent symmetric $d\times d$ matrices. Let $S_n = \sum_{i=1}^n X_i$. Then for any $t \geq 0$ and any $\ell \in \mathbb{R}$,*

$$\mathbf{Pr}\left[\lambda_1(S_n) \geq \ell\right] \leq d \cdot e^{-t\ell} \cdot \prod_{i=1}^n \lambda_1(\mathbf{E}[e^{tX_i}]). \tag{1}$$

$$\mathbf{Pr}\left[\lambda_d(S_n) \leq -\ell\right] \leq d \cdot e^{-t\ell} \cdot \prod_{i=1}^n \lambda_1(\mathbf{E}[e^{-tX_i}]). \tag{2}$$

Recall: the trace of A is $\text{tr}(A) := \sum_{i=1}^n a_{ii}$. You may use the following facts without proof.

(i) $\text{tr}(A) = \sum_{i=1}^n \lambda_i(A)$.

(ii) $\lambda_i(e^A) = e^{\lambda_i(A)}$.

(iii) The Golden-Thompson inequality: $\text{tr}(e^{A+B}) \leq \text{tr}(e^A \cdot e^B)$.

(iv) For positive semi-definite (psd) matrices A, B, $\text{tr}(AB) \leq \text{tr}(A) \cdot \lambda_1(B)$. (Recall that a symmetric matrix is psd iff all its eigenvalues are nonnegative.)

(v) Expectations and trace commute: i.e., $\mathbf{E}[\text{tr}(X)] = \text{tr}(\mathbf{E}[X])$.

Let us prove Theorem 1.

(a) Show that for any $t \geq 0$,

$$\mathbf{Pr}\left[\lambda_1(S_n) \geq \ell\right] \leq \mathbf{Pr}\left[\text{tr}(e^{tS_n}) \geq e^{t\ell}\right] \leq e^{-t\ell} \cdot \mathbf{E}[\text{tr}(e^{tS_n})].$$

(b) Show that

$$\mathop{\mathbf{E}}_{X_1,\ldots,X_n}\left[\text{tr}(e^{tS_n})\right] \leq \mathop{\mathbf{E}}_{X_1,\ldots,X_{n-1}}\left[\text{tr}(e^{tS_{n-1}})\right] \cdot \lambda_1\left(\mathbf{E}[e^{tX_n}]\right).$$

(Hint: why can you use (iv) above even if X_n is not psd?)

(c) Use (a)-(b) to prove (1).

(d) Use the same arguments on $(-S_n) = \sum_i (-X_i)$ to prove (2).

Note that Theorem 1 is the "Markov inequality" part of showing a Chernoff bound. The rest of the proof requires understanding $\mathbf{E}[e^{tX_n}]$, which requires linear algebra beyond the scope of this course. If you are curious, see the reference: Introduction to Random Matrix Theory.