Instructions

- **Due:** Tuesday, March 3 in class.
- Turn in a handwritten or printed PDF in class. Ensure it contains your name and email.
- You may use any resources in completing this homework. These include other students, AI tools, textbooks, and course notes.
- While completing the assignment, follow these instructions for each subproblem:
  - Make sure you understand the question.
  - Try to solve it alone.
  - If you use outside help, make sure you understand and can explain the solution.
- **Collaborator Acknowledgment:** After each problem, discuss the help you received. The length of your discussion should correspond to the level of assistance. For example, if a friend gave you a small hint, you might only need a sentence or two. At the other extreme, heavy reliance on AI tools requires a detailed reflection on what information you were missing and how you verified the answer.

Problems

Problem 1. SmallDB: Synthetic Data for Linear Queries In this problem, we will see a technique for answering a huge number of linear queries with low error. Recall we are given a set of queries $f_1, f_2, \ldots, f_k : \mathcal{X}^n \to [0,1]$ where each $f_j = \frac{1}{n}\sum_{i=1}^n \varphi_j(x_i)$ and each $\varphi_j : \mathcal{X} \to \{0,1\}$ is a single count (or predicate). Assume that $\mathcal{X}$ is finite. We will measure absolute error: writing $F : \mathcal{X}^n \to [0,1]^k$ and our answers as $\tilde{a} = (\tilde{a}_1, \tilde{a}_2, \ldots, \tilde{a}_k)$, this is

$$\mathrm{err}(F, x) := \|F(x) - \tilde{a}\|_\infty = \max_j |f_j(x) - \tilde{a}_j|.$$

We will fix some $\alpha, \beta > 0$ and ask how large n needs to be ensure the error is at most α with probability at least $1 - \beta$.

Here's the algorithm: let $\mathcal{Z} = \mathcal{X}^m$ be the space of datasets with $m = \frac{2\log(2k)}{\alpha^2}$. Sample some $Z \in \mathcal{Z}$ from the exponential mechanism with utility $u(z; x) = -\|F(z) - F(x)\|_\infty$. This utility function has sensitivity $\Delta = \frac{1}{n}$. Thus, we sample

$$Z \sim p_x(z) \propto \exp\left\{\frac{\varepsilon}{2\Delta} \cdot u(z; x)\right\} = \exp\left\{\frac{\varepsilon n}{2} \cdot u(z; x)\right\}$$

Then return $\tilde{a} = F(Z)$.

(a) Establish that there exists some z^* with utility at least $-\frac{\alpha}{2}$. The simplest way is via the *probabilistic method*: define a distribution over $Z \in \mathcal{Z}$ by sampling rows of x with replacement and show that it has nonzero probability of generating a high-utility dataset.

You can reuse the Hoeffding bound from Homework 1: if $A_1, \ldots, A_m$ are i.i.d. random variables in $\{0,1\}$ with mean μ, then $\Pr[|\mu - \frac{1}{m}\sum_i A_i| \geq \alpha/2] \leq 2\exp(-n\alpha^2/2)$.[1]

[1] Remark: if you've taken a class on statistical learning theory, you might recognize that this is a statement about *uniform convergence*; one can replace $\log k$ with other measures of complexity such as the Vapnik–Chervonenkis dimension.

(b) Apply the standard utility guarantees for the exponential mechanism: when using the exponential mechanism over a finite space $\mathcal{Y}$, we have

$$\Pr\left[u(\tilde{Y};x) < u(y^*;x) - \frac{2\Delta\log(|\mathcal{Y}|/\beta)}{\varepsilon}\right] \leq \beta.$$

(c) How large does n need to be to ensure we achieve error at most α with probability at least $1-\beta$?

Problem 2. Exponential Mechanism for Optimization In this problem, we will consider the accuracy guarantees that arise from using the exponential mechanism to perform empirical risk minimization (i.e., loss minimization over a dataset). We will try to find a parameter $\theta \in \Theta \subseteq \mathbb{R}^d$ that approximately minimizes

$$L(\theta, x) = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta; x_i). \tag{1}$$

We make a few assumptions on the problem. First, the parameter space is $\Theta = \{\theta : \|\theta\|_2 \leq R\}$, i.e., the ℓ_2 ball of radius R. Next, assume the loss function is *M-Lipschitz* for all x:

$$\forall x, \theta, \theta' : \quad |\ell(\theta; x) - \ell(\theta'; x)| \leq M \cdot \|\theta - \theta'\|_2.$$

Finally, assume $\ell(0, x) = 0$ for all x. Our algorithm draws $\tilde{\theta} \in \Theta$ from the exponential mechanism with utility $u(\theta; x) = -L(\theta, x)$:

$$p_x(\theta) \propto \exp\left(-\frac{\varepsilon}{2\Delta} \cdot L(\theta; x)\right)$$

where Δ is the sensitivity (set below). We'll use $\hat{\theta} = \hat{\theta}(x)$ to denote the minimizer of Eq. (1) and $\tilde{\theta}$ to denote our private answer.

(a) Prove that the sensitivity of this utility function is $\Delta = \frac{2MR}{n}$.
(Hint: for any θ and x we have $|\ell(\theta, x)| = |\ell(\theta, x) - \ell(0, x)| \leq M \cdot \|\theta - 0\|_2 \leq MR$.)

(b) Let $S_t \subseteq \Theta$ denote the set of θ with satisfy $L(\theta; x) - L(\hat{\theta}; x) \leq t$. Prove that, for any $t > 0$, S_t contains all θ which satisfy $\|\theta - \hat{\theta}\|_2 \leq \frac{t}{M}$.

(c) The volume of the ℓ_2 ball with radius r in d dimensions is $V_d \cdot r^d$, where V_d is a number that depends only on d. Use this to lower-bound the volume of S_t.[2]

(d) Use the standard utility analysis to upper-bound the probability we return a point with $L(\tilde{\theta}; x) - L(\hat{\theta}; x) > 2t$. Here's the start:

$$\begin{aligned}\Pr[L(\tilde{\theta}; x) - L(\hat{\theta}; x) > 2t] &\leq \frac{\Pr[L(\tilde{\theta}; x) - L(\hat{\theta}; x) > 2t]}{\Pr[L(\tilde{\theta}; x) - L(\hat{\theta}; x) \leq t]} \\ &\leq \frac{\mathrm{Vol}(\Theta)}{\mathrm{Vol}(S_t)} \cdot \frac{\exp(???)}{\exp(???)}.\end{aligned}$$

How large should t be to ensure this is at most β?[3]

[2] There is a technical complication that occurs when $\hat{\theta}$ is on or near the boundary of Θ. It doesn't materially affect the final answer. If you ignore this complication you will still get full credit. If you correctly deal with it I will write a star (worth zero points) on your assignment.

[3] When solving for t, you will end up with a t on one side of the expression and a $\log(1/t)$ on the other. You can simply drop the latter term and receive full credit or, for the rare double star (worth twice as much as a single star), resolve the issue.

Problem 3. Inverse Sensitivity Mechanism, Discrete Case In class we saw the *inverse sensitivity mechanism*, which is the usual exponential mechanism applied to the following utility measure: for function $f : \mathcal{X}^n \to \mathcal{Y}$ and for $x \in \mathcal{X}^n$ and $y \in \mathcal{Y}$,

$$u(x; y) = -\text{len}_f(x; y) := -\min_x \left\{ d(x, x') \mid f(x') = y \right\}.$$

Here $d(\cdot, \cdot)$ denotes Hamming distance. Also recall the definition of local sensitivity at distance k:

$$\text{LS}_f^k(x) = \max_{\substack{x' \\ d(x,x') \le k}} |f(x) - f(x')|.$$

In this problem, we will apply it to integer-valued functions $f : \mathcal{X}^n \to \mathcal{Y}$, with $\mathcal{Y} = \{1, 2, \ldots, d\}$. Let random variable $\tilde{Y}$ denote the output of this algorithm. We will prove that, for any $t > 0$,

$$\mathbb{E}\left[|\tilde{Y} - f(x)| \right] \le \text{LS}_f^{k'} + t,$$

where $k' = \frac{2}{\varepsilon} \log(2d^2/\varepsilon t)$.

(a) Let $\mathcal{Y}_x^k = \{y \in \mathcal{Y} : \text{len}_f(x, y) = k\}$. Prove that $\Pr[\tilde{Y} \in \mathcal{Y}_k] \le d \cdot e^{-k\varepsilon/2}$.
(Hint: relate the probability on these outputs to that on $f(x)$ itself.)

(b) Show that, for $y \in \mathcal{Y}_x^k$, $|y - f(x)| \le \text{LS}_f^k(x)$.

(c) Write out the expectation and prove that, for any $1 \le T \le n$,

$$\mathbb{E}\left[|\tilde{Y} - f(x)| \right] \le \text{LS}_f^T(x) + \frac{2d^2}{\varepsilon} e^{-\varepsilon T/2}.$$

(d) Finish the proof.