DA 5001/6400 (July-Nov 2024): HW1

IIT Madras

Due Date: September 1, 2024 at 11:59 PM

Instructions

- The maximum score on the homework is 80 marks, including the bonus (extra credit) exercises.
- List the names of students you collaborated with for the homework. Also, cite any books, notes, or web resources you used for any problem. Failure to do so will be considered a violation of the honour code.
- If you used an LLM for help in one of the exercises, you must specify the name of the LLM and the exact prompt used.
- Corrections to the homework are shown in blue. Please note them.
- **Last updated**: Friday, August 30th at 8:30 am.

0 Honour Code (0 marks but mandatory)

Please read the full honour code on the course webpage and write "I ACCEPT THE HONOUR CODE" in your submission. **Your submission will not be graded if you do not accept the honour code.**

1 Utility of Randomized Response for Counting (10 marks)

In this exercise, we will calculate the utility of randomized response in computing the mean of true/false survey questions from a population in a differentially private manner.

Consider a binary input space $\mathcal{X} = \{0, 1\}$ representing the true response of participants to a survey question. We assume that the true underlying response x_i of each participant i is fixed and non-random. Suppose our n survey participants' responses are collected into a dataset $D = (x_1, \ldots, x_n) \in \mathcal{X}^n$.

Recall that randomized response at a privacy level $\varepsilon > 0$ privatizes each individual response, i.e., it returns $\mathcal{A}(D) = (y_1, \ldots, y_n)$ with

$$\mathbb{P}(y_i = x_i) = \frac{e^\varepsilon}{1 + e^\varepsilon} = 1 - \mathbb{P}(y_i = 1 - x_i)\,.$$

Our goal is to find the mean $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ of the dataset D. This problem has historically been referred to as "counting" because finding $\bar{x}$ is equivalent to counting the number of ones in the dataset D, which equals $n\bar{x}$.

Now, answer the following questions.

1. Find reals a, b such that $\mathbb{E}[y_i] = ax_i + b$ for all i (note that a, b must be independent of i but can depend on other problem parameters such as ε, n). Thus, establish that z_i, defined as

$$z_i = \frac{1}{a}(y_i - b)$$

satisfies $\mathbb{E}[z_i] = x_i$ for all i.
Context: We wish to estimate x_i from the privatized version y_i. An estimator $\hat{x}_i$ of x_i is said to be *biased* if $\mathbb{E}[\hat{x}_i] \neq x_i$ and *unbiased* otherwise. Unbiasedness is a highly desirable property meaning that we do not have any *systematic error* in estimation. Your calculations above show that y_i is a biased estimator of the truth x_i, but z_i is unbiased. Thus, we will use z_i going forward.

2. Consider the estimator $\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$ for the true mean $\bar{x}$. Show that $\mathbb{E}[\bar{z}] = \bar{x}$, i.e. $\bar{z}$ is an unbiased estimator of $\bar{x}$.

3. **Utility**: Find a number t_ε such that $|\bar{z} - \bar{x}| \leq t_\varepsilon$ with probability at least $1 - \nu$, for some $\nu > 0$ fixed.[1] Note that t should scale as $\text{poly}\log(1/\nu)$. **Hint**.[2]

4. Prove that

$$\lim_{\varepsilon \to 0^+} \frac{t_\varepsilon}{(\varepsilon\sqrt{n})^{-1}} = C\sqrt{\log(2/\nu)}\,,$$

where C is an absolute constant. In other words, $t_\varepsilon \approx \frac{1}{\varepsilon\sqrt{n}}$ as $\varepsilon \to 0$ (ignoring absolute constants and polylog factors).

5. Show that $\text{PrivLoss}(P, Q)$ is sub-Gaussian with $P = \text{Bernoulli}(p)$ and $Q = \text{Bernoulli}(1-p)$ (this shows up in randomized response with $p = e^\varepsilon/(1+e^\varepsilon)$). Find its variance proxy, i.e., the smallest number c^2 such that $\mathbb{E}_{Z\sim\text{PrivLoss}(P,Q)}[e^{\lambda(Z-\mathbb{E}[Z])}] \leq \exp(\lambda^2 c^2/2)$ for all $\lambda > 0$. How does this compare to its variance? **Correction**: The privacy loss has to be mean-centered.

2 Utility of the Laplace Mechanism for Counting (10 marks)

Consider the counting problem of Problem 1. Suppose we use the Laplace mechanism instead of randomized response. At a given privacy level of $\varepsilon > 0$, the Laplace mechanism outputs

$$\hat{x} = \bar{x} + W\,, \quad \text{where} \quad W \sim \text{Laplace}(0, \Delta/\varepsilon)\,,$$

where Δ is the sensitivity of the mean operation. We showed in class that this mechanism is ε-DP. Now answer the following questions.

1. Let us define two datasets $D = (x_i)_{i=1}^n$ and $D' = (x'_i)_{i=1}^n$ as neighbours if $x_i = x'_i$ for all $i \neq j$. That is, dataset D' neighbours D if we replace some $x_j \in D$ with another element x'_j, and all other elements remain equal.[3] Find the sensitivity $\Delta = \max_{D\simeq D'} |A(D) - A(D')|$ of the mean $A(D) = \frac{1}{n}\sum_{i=1}^n x_i$ under this notion of neighbourhood.

[1] This is equivalent to showing that $\mathbb{P}(|\bar{z} - \bar{x}| > t_\varepsilon) \leq \nu$.
[2] **Hint**: Use Hoeffding's inequality.
[3] This is known as the "replace-one" notion of neighbourhood, in contrast to the usual "add/remove" notion of neighbourhood.

Utility of the Laplace mechanism Our estimate $\hat{x}$ is clearly unbiased. The error in our estimate can be calculated by bounding $\mathbb{P}(|\hat{x} - x| > t) = \mathbb{P}(|\xi| > t)$. In other words, we need a tail bound on the Laplace distribution. Let us use the Chernoff bounding strategy to derive such a tail bound.

2. Find the moment generating function $\phi_W(\lambda) = \mathbb{E}[\exp(\lambda W)]$ for $W \sim \text{Laplace}(0, b)$ (here, $b > 0$ is the scale factor). Note that its PDF is $f_W(x) = \frac{1}{2b} \exp\left(-\frac{|x|}{b}\right)$.

3. Establish using the Chernoff bound the following:

$$\mathbb{P}_{W \sim \text{Laplace}(0,b)}(|W| > t) \leq 4 \exp\left(-\frac{t}{\sqrt{2}b}\right) .$$

   Context: Note that this tail bound has a scaling of $\exp(-t)$; this is a slower decrease than the $\exp(-t^2)$ scaling we expect from sub-Gaussian random variables. The Laplace distribution is not sub-Gaussian, but it falls under the family of sub-exponential distributions.

4. Conclude that the utility bound of the Laplace mechanism is

$$|\hat{x} - x| \leq O\left(\frac{1}{n\varepsilon} \log(1/\nu)\right)$$

   with probability at least $1 - \nu$. Here, $O(\cdot)$ denotes the big-O notation.

5. The calculations above show that the utility of the Laplace mechanism is better than that of randomized response. Intuitively, why do you think that is the case? **Hint**.[4]
   Context: Later in the course, we will study the federated learning setting where each x_i can reside with a different person/smartphone/physical location. The differences between the implementations of the two algorithms will be meaningful in that context, depending on the presence of a *trusted aggregator*.

3 Properties of the Privacy Loss Distribution (10 marks)

Consider two distributions P and Q. Let $Z \sim \text{PrivLoss}(P, Q)$ denote the privacy loss random variable. Unless specified otherwise, you may assume that P, Q are continuous distributions on $\mathbb{R}^d$ — in this case, Z is distributed as $\log(f_P(Y)/f_Q(Y))$ for $Y \sim P$, where f_P and f_Q are respectively the PDFs of P and Q.

1. Derive the privacy loss distribution for the Laplace mechanism (i.e., compute privacy loss distribution between $P = \text{Laplace}(0, \Delta/\varepsilon)$ and $Q = \text{Laplace}(\Delta, \Delta/\varepsilon)$). Plot the PDF of the privacy loss distribution.

2. Prove that $\mathbb{E}[Z] = \text{KL}(P\|Q)$ is the KL divergence between P and Q.

3. Prove that $\text{KL}(P\|Q) \geq 0$ for any distributions P, Q; thus, the privacy loss distribution has non-negative mean. You can restrict your proof to the case where P, Q are discrete distributions for this exercise (the proof for the general case is similar). **Hint**. [5]

[4] **Hint**: Think about how many bits are privatized for randomized response. How about the Laplace mechanism (assuming finite precision floating point arithmetic)?

[5] **Hint**: There are multiple proofs for this. The easiest one is an invocation of Jensen's inequality for the convexity of the function $t \mapsto -\log(t)$. Alternatively, you may prove this using the log-sum inequality, which you proved in HW0.

4. Prove that $\mathbb{E}[\exp(-Z)] = 1$.

5. Suppose that $P = \mathcal{A}(D)$ and $Q = \mathcal{A}(D')$ are the output distributions of a randomized algorithm $\mathcal{A}(\cdot)$ that satisfies ε-DP. Prove that $\mathbb{P}(Z = \varepsilon) \leq \frac{e^\varepsilon}{1+e^\varepsilon}$. In other words, randomized response has the worst privacy loss. You may assume that Z is a discrete distribution supported on $2 \leq k < \infty$ points. **Hint**.[6] **Background**.[7]
Context: The basic composition result we proved will be tight if $\mathbb{P}(Z = \varepsilon) = 1$ was possible for some P, Q. But the result above shows that this is never possible. Indeed, it is this necessary slack which makes basic composition loose due to rare events.

4 Privacy Loss to DP Conversions (10 marks)

Consider two distributions P, Q over some set $\mathcal{Y}$ such that $\mathrm{PrivLoss}(P, Q)$ is well-defined. Define the hockey-stick divergence H_α as

$$H_\alpha(P\|Q) = \sup_{S \subset \mathcal{Y}} \left\{ P(S) - \alpha\, Q(S) \right\}.$$

Note that $\alpha = 1$ is related to the total variation distance, which we explore further in Problem 5.

1. (ε, δ)-DP in terms of the hockey-stick divergence: Prove that a randomized algorithm $\mathcal{A}(\cdot)$ is (ε, δ)-DP iff $H_{e^\varepsilon}\big(\mathcal{A}(D)\|\mathcal{A}(D')\big) \leq \delta$ for all neighbouring datasets $D \simeq D'$.

2. Prove that

$$H_{e^\varepsilon}(P\|Q) = \mathbb{P}_{Z \sim \mathrm{PrivLoss}(P,Q)}(Z > \varepsilon) - e^\varepsilon\ \mathbb{P}_{Z' \sim \mathrm{PrivLoss}(Q,P)}(Z' < -\varepsilon)\,.$$

   Hint.[8]

3. Compute the hockey-stick divergence $H_{e^\varepsilon}(\mathcal{N}(0,1), \mathcal{N}(\Delta, 1))$ using the expression above. You can express your answer in terms of the standard normal tail probability $\Phi(t) = \mathbb{P}\big(\mathcal{N}(0,1) > t\big)$.

4. Consider a deterministic function $A : \mathcal{X}^* \to \mathbb{R}$ with sensitivity $\Delta = \max_{D \simeq D'} |A(D) - A(D')|$. Then, the randomized algorithm $\mathcal{A}(D) = \mathcal{N}\big(A(D), \sigma^2\big)$ satisfies (ε, δ)-DP with

$$\delta = \delta(\varepsilon) = \Phi\left(\frac{\varepsilon - \rho}{\sqrt{2\rho}}\right) - e^\varepsilon\ \Phi\left(\frac{\varepsilon + \rho}{\sqrt{2\rho}}\right),$$

   where $\rho = \frac{\Delta^2}{2\sigma^2}$ and the $\Phi(t) = \mathbb{P}\big(\mathcal{N}(0,1) > t\big)$ is the standard normal tail probability.

5. Set $\Delta = 1$ and $\sigma = 1$. Plot $\delta(\varepsilon)$ from the expression above vs. ε (in log-log scale) as you vary ε between 0.1 and 10.

[6] **Hint**: Connect the privacy loss to the TPR and FPR (or equivalently, the Type-I and Type-II errors) of a hypothesis test. If $\mathbb{P}(Z = \varepsilon)$ is larger than the claimed bound, does it contradict the fundamental limits on the TPR/FPR that are possible under DP?

[7] **Background**: Although out of the scope of this exercise, the limit of $k \to \infty$ can be used to reason about the continuous case. This is a typical argument in measure-theoretic probability — refer to Sec. 1.6.3 of Durrett's book if interested.

[8] **Hint**: We derived in class the set $S_\star$ that maximizes the right side of the hockey-stick divergence and the value of $P(S_\star)$. You need to calculate $Q(S_\star)$ similarly and plug it into $H_\alpha(P\|Q) = P(S_\star) - \alpha Q(S_\star)$.

5 Alternate Definitions of DP That Do Not Work (10 marks)

In this exercise, we will investigate the pitfalls of common statistical divergences like the KL divergence (natural choice for ML folks) or the total variation distance (natural for cryptographers) as notions of privacy.

Total Variation (TV) Distance Recall that the total variation distance between two probability distributions P, Q over a space $\mathcal{Y}$ is defined as

$$\mathrm{TV}(P, Q) = \sup_{S \subset \mathcal{Y}} \{P(S) - Q(S)\} \ .$$

It can be shown that $\mathrm{TV}(P, Q) = \frac{1}{2} \int_{\mathcal{Y}} |f_P(y) - f_Q(y)| \, dy$ for continuous P, Q with densities f_P, f_Q, while for discrete distributions, an analogous expression holds where the integral is replaced by the sum and the densities by the probability mass functions.

The TV distance quantifies an additive error in the probabilities of any outcome, while ε-DP quantifies the multiplicative error. Suppose we define an alternative to DP based on the TV distance.

Definition 1. *We say that a randomized algorithm $\mathcal{A}(\cdot)$ is δ-TV private if $\mathrm{TV}\big(\mathcal{A}(D), \mathcal{A}(D')\big) \leq \delta$ for any neighbouring datasets $D \simeq D'$.*

Let us consider what goes wrong with such a definition.

1. If a randomized algorithm $\mathcal{A}(\cdot)$ is ε-DP, then show that it is δ-TV private with $\delta = 1 - e^{-\varepsilon}$.

2. Consider the "replace-one" notion of neighbourhood we considered in Problem 2. That is, $D \simeq D'$ are considered neighbours if we can obtain D' by substituting one of the elements of D so that $|D \cap D'| + 1 = |D| = |D'|$.
   Consider a randomized algorithm $\mathcal{A}_0 : \mathcal{X}^n \to \mathcal{X}$ such that $\mathcal{A}_0(x_1, \ldots, x_n)$ releases one of its inputs uniformly at random.[9] Prove that $\mathcal{A}_0$ is δ-TV private with $\delta = 1/n$.

Any algorithm that releases one datapoint is not what we would intuitively think of as "private". Thus, our notion of δ-TV privacy does not make sense of δ around $1/n$ or larger. Let us consider what happens when $\delta < 1/2n$ is small. For this, we will need some additional properties of TV.

3. Prove that TV is a *distance metric* (in the mathematical sense of the term). That is, prove that it satisfies:
   (a) Symmetry: $\mathrm{TV}(P, Q) = \mathrm{TV}(Q, P)$.
   (b) Non-negativity: $\mathrm{TV}(P, Q) \geq 0$ with $\mathrm{TV}(P, Q) = 0$ if and only if $P = Q$.
   (c) Triangle Inequality: $\mathrm{TV}(P, Q) \leq \mathrm{TV}(P, S) + \mathrm{TV}(S, Q)$ for any probability distributions P, Q, S.

4. Suppose that $\mathcal{A}(\cdot)$ satisfies δ-TV privacy. Then, show that for any *arbitrary pairs of datasets* $D, D' \in \mathcal{X}^n$ (that are not required to be neighbouring), we have $\mathrm{TV}\big(\mathcal{A}(D), \mathcal{A}(D')\big) \leq n\delta$. **Hint**.[10]

[9] This is also known colloquially as the "name and shame" mechanism.

[10] **Hint**. Construct a series of datasets $D = D_0 \simeq D_1 \simeq \cdots \simeq D_k = D'$ for some $k \leq n$ by swapping out one element at a time. Then, apply the triangle inequality.

Here is why δ-TV privacy with small δ fails. If $\delta < 1/2n$, then we have that $\mathrm{TV}\big(\mathcal{A}(D), \mathcal{A}(D')\big) \leq 1/2$. That is, with probability at least $1/2$, the output $\mathcal{A}(D)$ is independent of D (since the output should match that of any arbitrary dataset D'). In other words, such a randomized algorithm is useless. Thus, the TV distance is not a good metric for defining differential privacy (i.e. between neighbouring datasets).

Note, however, that (ε, δ)-DP is a combination of δ-TV with ε-DP, so the TV distance is not entirely unusable for privacy. Comparing the definition of the hockey stick divergence with that of TV distance makes the connection apparent.

KL Divergence We will now see how the KL divergence is an unsuitable notion of differential privacy because it is not sensitive enough to low probability bad events.

5. Construct four discrete distributions P_1, Q_2, P_2, Q_2 with a support size of k of your choice such that their privacy loss random variables $Z_1 \sim \mathrm{PrivLoss}(P_1, Q_1)$ and $Z_2 \sim \mathrm{PrivLoss}(P_2, Q_2)$ satisfy the following:
   (a) $\mathbb{E}[Z_1] = \mathbb{E}[Z_2]$ or equivalently, we have $\mathrm{KL}(P_1, Q_1) = \mathrm{KL}(P_2, Q_2)$.
   (b) For some $\varepsilon > 0$, we have $\mathbb{P}(Z_1 \leq \varepsilon) = 1$ and $\mathbb{P}(Z_2 > \varepsilon) \geq 0.1$.

Can you think about why this is bad from a privacy perspective?

6 Properties of the Rényi Divergences and DP (10 marks)

Recall that the Rényi divergence of order $\alpha \in (0, \infty) \setminus \{1\}$ between two discrete distributions P and Q is defined as

$$R_\alpha(P\|Q) = \frac{1}{\alpha - 1} \log \left(\mathbb{E}_{X \sim P} \left(\frac{P(X)}{Q(X)} \right)^{\alpha - 1} \right) = \frac{1}{\alpha - 1} \log \left(\mathbb{E}_{X \sim Q} \left(\frac{P(X)}{Q(X)} \right)^{\alpha} \right) .$$

For continuous distributions P, Q with respective PDFs f_P, f_Q, the likelihood ratio $P(X)/Q(X)$ is replaced by the density ratio $f_P(X)/f_Q(X)$. The Rényi divergence for orders $\alpha = 0, 1, \infty$ is defined as the limit of the expression above.

We will establish a number of properties of the Rényi divergence. You may assume that $P = (P_1, \ldots, P_k)$ and $Q = (Q_1, \ldots, Q_k)$ are discrete distributions over k items unless specified otherwise.

1. Prove that $\lim_{\alpha \to 1} R_\alpha(P\|Q) = \mathrm{KL}(P\|Q)$ recovers the KL divergence. **Hint**.[11]
2. Prove that $R_\infty(P\|Q) := \lim_{\alpha \to \infty} R_\alpha(P\|Q) = \log \left(\max_{i=1,\ldots,k} \frac{p_i}{q_i} \right)$. Hence, prove that a randomized algorithm $\mathcal{A}(\cdot)$ is ε-DP iff $R_\infty\big(\mathcal{A}(D)\|\mathcal{A}(D')\big) \leq \varepsilon$ for all neighbouring datasets $D \simeq D'$.
3. Prove that $R_\alpha(P\|Q)$ is a non-decreasing function of $\alpha > 1$ for any pair of fixed distributions P, Q. **Hint**.[12]

[11] **Hint**: Use L'Hôpital's rule.

[12] **Hint**: There are many ways to prove this. One option is to use Jensen's inequality on the convex function $f(t) = t^{\frac{\alpha' - 1}{\alpha - 1}}$ for $1 < \alpha < \alpha' < \infty$ to show that

$$\exp\big((\alpha' - 1) R_\alpha(P\|Q)\big) \leq \exp\big((\alpha' - 1) R_{\alpha'}(P\|Q)\big) .$$

An alternate option is to define $h(\alpha) := R_\alpha(P\|Q)$ (with arbitrary P, Q fixed), note that it is continuous and differentiable and show that its derivative $h'(\alpha)$ is non-negative.

Context: As α gets larger, $R_\alpha(P\|Q)$ focuses more on tail events, i.e. worst-case sets S where $P(S)$ is large and $Q(S)$ is small. Indeed, $\alpha \to 1$ corresponds to the KL divergence, which is not sensitive to the tails (Problem 5), while $\alpha \to \infty$ recovers ε-DP which is overly sensitive to low-probability tail events. Rényi divergence is flexible enough to focus the analysis on intermediate values of α where we can capture the best of both worlds.

4. Alternate form of concentrated DP: Prove that a randomized algorithm $\mathcal{A}(\cdot)$ is ρ-zCDP iff

$$R_\alpha\big(\mathcal{A}(D)\|\mathcal{A}(D')\big) \le \rho\alpha$$

for all $\alpha > 1$ and for all neighbouring datasets $D \simeq D'$.

5. Derive from first principles the following formula for d-dimensional Gaussians with a common covariance matrix $\Sigma \in \mathbb{S}^d_{++}$:

$$R_\alpha\big(\mathcal{N}(\mu,\Sigma)\|\mathcal{N}(\mu',\Sigma)\big) = \frac{\alpha}{2}\,(\mu-\mu')^\top\Sigma^{-1}(\mu-\mu')\,.$$

Context: This linear scaling of the α-Rényi divergence with α for Gaussians has some direct implications. First, this directly implies the zCDP property of the Gaussian mechanism. Second, the behaviour is qualitatively the same for all α, i.e., zCDP fully captures all properties of the Gaussian mechanism. As we see in class, subsampling plus composition leads to non-linear behavior with α, where it makes sense to focus on intermediate values of α.

7 From Rényi Divergences to Properties of DP (10 marks)

We will continue establishing properties of the Rényi divergence from the previous exercise. These properties directly imply desirable properties of DP variants such as Rényi DP and zCDP.

1. Prove that $R_\alpha(P_1 \times P_2\|Q_1 \times Q_2) = R_\alpha(P_1\|Q_1) + R_\alpha(P_2\|Q_2)$. Recall that $P_1 \times P_2$ denotes the product distribution of P_1 and P_2; a sample from $P_1 \times P_2$ can be obtained by sampling $X_1 \sim P_1$, $X_2 \sim P_2$ and returning (X_1, X_2).
Context: This leads to a composition result for Rényi DP and hence, zCDP as well. Note that this is identical to the zCDP composition result we proved in class.

2. Prove the triangle-like inequality: for all $1 < \alpha < \alpha' < \infty$ and distributions P, Q, S with $R_\infty(S\|Q) < \infty$, we have

$$R_\alpha(P\|Q) \le \frac{\alpha'}{\alpha'-1}R_\beta(P\|S) + R_{\alpha'}(S\|Q)\,,$$

where $\beta = \frac{\alpha(\alpha'-1)}{\alpha'-\alpha}$. You may use Hölder's inequality which states that

$$\mathbb{E}[XY] \le \big(\mathbb{E}[X^p]\big)^{1/p}\,\big(\mathbb{E}[Y^q]\big)^{1/q}$$

for all random variables X, Y (such that the expectations above are well-defined) and all $p, q > 1$ satisfying $1/p + 1/q = 1$. **Hint**.[13]

[13] **Hint**: Show using Hölder's inequality that for all $p, q > 1$ with $1/p + 1/q = 1$, we have

$$\exp\big((\alpha-1)R_\alpha(P\|Q)\big) \le \exp\left(\frac{\alpha p - 1}{p}R_{\alpha p}(P\|S)\right)\,\exp\big((\alpha-1)R_{(\alpha-1)q+1}(S\|Q)\big)\,.$$

Now, rearrange and substitute appropriate values for p, q. Note the correction in blue in the equation above.

3. Prove the triangle inequality for zCDP: if $R_\alpha(P\|S) \le \rho_1\alpha$ and $R_\alpha(S\|Q) \le \rho_2\alpha$ for all $\alpha > 1$, then we have

$$R_\alpha(P\|Q) \le \left(\sqrt{\rho_1} + \sqrt{\rho_2}\right)^2\alpha \quad \forall\alpha > 1.$$

Hint.[14]
Context: This triangle inequality can be used to prove a property known as group privacy. If $D \simeq D'$ and $D' \simeq D''$ are two pairs of neighbouring datasets, then D and D'' differ by two elements. But, the zCDP property can be extended to hold for all such pairs D, D'' using the triangle inequality above. In general, this can be extended to pairs of datasets that differ by k elements and is known as "group privacy". This is because the algorithm is now private w.r.t. changes in *groups* of k datapoints.

8 Concentration of Measure: Simulation (10 marks)

In this exercise, we will simulate the concentration of the empirical mean to the population mean.

Here are the steps to follow:

- For each trial $j = 1, \ldots, m$, draw n i.i.d. samples $X_1^{(j)}, \ldots, X_n^{(j)}$ for a given distribution $P = \text{Bernoulli}(0.7)$.
- Calculate the deviation $\Delta_n^{(j)} = \left|\frac{1}{n}\sum_{i=1}^n X_i^{(j)} - \mathbb{E}_{X\sim P}[X]\right|$ for each trial j.
- Denote $\hat{q}_n(\nu)$ be the empirical $1-\nu$ quantile of the $\Delta_n^{(j)}$. That is, $\hat{q}_n(\nu)$ is the number which is larger than $(1-\nu)$-fraction of the deviations and smaller than ν-fraction of the deviations.
- Compare this empirical observation to the following analytical bounds:
  (a) Chebyshev's Inequality: Compute the upper bound $t_n(\nu)$ implied by Chebyshev's inequality such that $\mathbb{P}\big(|(1/n)\sum_{i=1}^n X_i - \mathbb{E}[X]| > t_n(\nu)\big) \le \nu$ holds.
  (b) Hoeffding Inequality: Same as above but with Hoeffding's inequality.
  (c) Gaussian approx: If the Gaussian approximation from the central limit theorem is correct, then

$$\left|\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}[X]\right| \le C_\nu\sqrt{\frac{\text{Var}(P)}{n}}$$

  holds with probability at least $1-\nu$, where C_ν denote the $(\nu/2)$-quantile of the standard Gaussian. Call this right side as $g_n(\nu)$.
  (d) Plot the empirical quantile $\hat{q}_n(\nu)$ vs. $\nu \in [0.1, 0.05, 0.01, 0.005, 0.001]$ with $n = 10^3$ and $m = 10^4$. Compare this to the upper bounds $t_n(\nu)$ implied by (a) Chebyshev and (b) Hoeffding inequalities, and (c) the Gaussian approximation $g_n(\nu)$. Make this plot with the x-axis in log-scale but the y-axis in linear-scale.
  (e) Repeat the above plot but vary $n \in [10, 10^2, 10^3, 10^4]$, while keeping $\nu = 0.01$ fixed. Plot this in log-log scale and notice the straight lines. What is the slope of each of the lines?[15]

You may also wish to view the plots without the Chebyshev line to observe the finer details.

[14] **Hint**: Use the triangle-like inequality you just proved for Rényi divergences and find (using calculus) the value of α' that gives the smallest upper bound for each α.

[15] Note that the curve $y = cx^a$ appears as a straight line of slope a in log-log scale.

9 [Bonus] Chebyshev's Inequality is Tight in the Worst Case (10 marks)

We saw how the Chernoff bounds are significantly tighter than Chebyshev bounds. However, it turns out that Chebyshev bounds are the best one can do for random variables with finite variance, as we shall see now.

Consider the random variable X defined on $(2, \infty)$ with the PDF

$$f_X(x) = C \, \frac{2\log(x) + 3}{x^3 \log^4 x},$$

where $C < \infty$ is a normalizing constant so that $\int_2^\infty f_X(x)dx = 1$. We will show that Chebyshev's inequality is nearly tight for this random variable.

Prove the statements below from first principles (you may verify this by numerical integration). Please note that ChatGPT is known to get some of these integrals wrong at the time of this writing.

1. Show that $\int_2^\infty \frac{2\log(x)+3}{x^3 \log^4 x} < \infty$ so that the PDF $f_X(x)$ is well-defined.
2. Show that $\mathbb{E}[X] < \infty$, i.e. X has finite mean.
3. Show that $\mathbb{E}[X^2] < \infty$, i.e. X has finite variance.
4. Show that $\mathbb{E}[X^3] = \infty$, i.e. X does not admit a third moment
5. Show that $\mathbb{P}(X > t) \geq \frac{c}{t^2 \log^3(t)}$ for all $t > 4$ for some absolute constant $c > 0$. Thus, the $1/t^2$ bound given by Chebyshev's inequality is tight up to log factors in this case.