CSC 2515 Lecture 11: Differential Privacy

Roger Grosse

University of Toronto

Overview

- So far, this class has been about getting algorithms to perform well according to some metric (e.g. prediction error).
  - Up until about 5 years ago, this is what almost the entire field was about.
- Now that AI is in widespread use by companies and governments, and used to make decisions about people, we have to ask: are we optimizing the right thing?
- The final two lectures are about AI ethics.
  - Focus is on *technical*, rather than *social/legal/political*, aspects. (I'm not qualified to talk about the latter.)

Overview

- **This lecture:** differential privacy
  - Companies, governments, hospitals, etc. are collecting lots of sensitive data about individuals.
  - Anonymizing data is surprisingly hard.
  - Differential privacy gives a way to analyze data that provably doesn't leak (much) information about individuals.
- **Next lecture:** algorithmic fairness
  - How can we be sure that the predictions/decisions treat different groups fairly? What does this even mean?
- Privacy and fairness are among the most common topics the Vector Institute is asked for advice about by local companies and hospitals.
- **Disclaimer:** I'm still learning this too.

Overview

- Many AI ethics topics we're leaving out
  - Explainability (people should be able to understand why a decision was made about them)
  - Accountability (ability for a third-party to verify that an AI system is following the regulations)
  - Bad side effects of optimizing for click-through?
  - How should self-driving cars trade off the safety of passengers, pedestrians, etc.? (Trolley problems)
  - Unemployment due to automation
  - Face recognition and other surveillance-enabling technologies
  - Autonomous weapons
  - Risk of international AI arms races
  - Long-term risks of superintelligent AI
- I'm focusing on privacy and fairness because these topics have well-established *technical* principles and techniques that address part of the problem.

Overview

An excellent popular book:

Why Is Anonymization Hard?

Why Is Anonymization Hard?

Some examples of anonymization failures (taken from *The Ethical Algorithm*)

- In the 1990s, a government agency released a database of medical visits, stripped of identifying information (names, addresses, social security numbers)
  - But it did contain zip code, birth date, and gender.
  - Researchers estimated that 87 percent of Americans are uniquely identifiable from this triplet.
- Netflix Challenge (2006), a Kaggle-style competition to improve their movie recommendations, with a $1 million prize
  - They released a dataset consisting of 100 million movie ratings (by "anonymized" numeric user ID), with dates
  - Researchers found they could identify 99% of users who rated 6 or more movies by cross-referencing with IMDB, where people posted reviews publicly with their real names

Why Is Anonymization Hard?

Not sufficient to prevent unique identification of individuals.

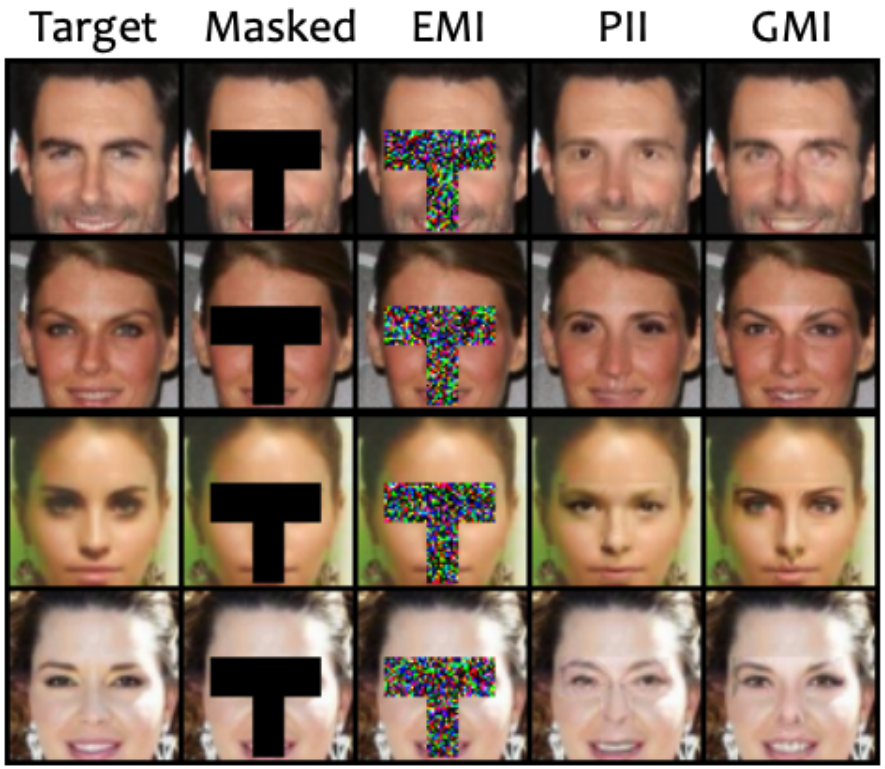
Name	Age	Gender	Zip Code	Smoker	Diagnosis
*	60–70	Male	191**	Y	Heart disease
*	60–70	Female	191**	N	Arthritis
*	60–70	Male	191**	Y	Lung cancer
*	60–70	Female	191**	N	Crohn's disease
*	60–70	Male	191**	Y	Lung cancer
*	***50–60***	***Female***	191**	N	HIV
*	50–60	Male	191**	Y	Lyme disease
*	50–60	Male	191**	Y	Seasonal allergies
*	***50–60***	***Female***	191**	N	Ulcerative colitis

Kearns & Roth, *The Ethical Algorithm*

From this (fictional) hospital database, if we know Rebecca is 55 years old and in this database, then we know she has 1 of 2 diseases.

Why Is Anonymization Hard?

- Even if you don't release the raw data, the weights of a trained network might reveal sensitive information.
- Model inversion attacks recover information about the training data from the trained model.
- Here's an example of reconstructing individuals from a face recognition dataset, given a classifier trained on this dataset and a generative model trained on an unrelated dataset of publicly available images.



- **Col 1:** training image. **Col 2:** prompt. **Col 4:** best guess from only public data. **Col 5:** reconstruction using classification network.
- **Source:** Zhang et al., "The secret revealer: Generative model-inversion attacks against deep neural networks." `https://arxiv.org/abs/1911.07135`

Why Is Anonymization Hard?

- A neural net language model trained on Linux source code learned to output the exact text of the GPL license.

```
/*
 *  Copyright (c) 2006-2010, Intel Mobile Communications.  All rights reserved.
 *
 *   This program is free software; you can redistribute it and/or modify it
 * under the terms of the GNU General Public License version 2 as published by
 * the Free Software Foundation.
 *
 *        This program is distributed in the hope that it will be useful,
 * but WITHOUT ANY WARRANTY; without even the implied warranty of
 *   MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.  See the
```

http://karpathy.github.io/2015/05/21/rnn-effectiveness/

- Gmail uses language models for email autocompletion. Imagine if the autocomplete feature spits out the entire text of one of your past emails.

Why Is Anonymization Hard?

- It's hard to guess what capabilities attackers will have, especially decades into the future.
  - **Analogy with crypto:** Cryptosystems today are designed based on what quantum computers might be able to do in 30 years.
  - To defend against unknown capabilities, we need mathematical guarantees.
- **Want to guarantee:** no individual is directly harmed (e.g. through release of sensitive information) by being part of the database, even if the attacker has tons of data and computation.

An Intuition Pump: Randomized Response

Randomized Response

- **Intuition:** Randomized response is a survey technique that ensures some level of privacy.
- **Example:** Have you ever dodged your taxes?
  - Flip a coin.
  - If the coin lands Heads, then answer truthfully.
  - If it lands Tails, then flip it again.
    - If it lands Heads, then answer Yes.
    - If it lands Tails, then answer No.
- **Probability of responses:**

	Yes	No
Dodge	3/4	1/4
No Dodge	1/4	3/4

Randomized Response

- Tammy the Tax Investigator assigns a prior probability of 0.02 to Bob having dodged his taxes. Then she notices he answered Yes to the survey. What is her posterior probability?

$$\begin{aligned}\Pr(\mathrm{Dodge} \mid \mathrm{Yes}) &= \frac{\Pr(\mathrm{Dodge})\Pr(\mathrm{Yes} \mid \mathrm{Dodge})}{\Pr(\mathrm{Dodge})\Pr(\mathrm{Yes} \mid \mathrm{Dodge}) + \Pr(\mathrm{NoDodge})\Pr(\mathrm{Yes} \mid \mathrm{NoDodge})} \\ &= \frac{0.02 \cdot \frac{3}{4}}{0.02 \cdot \frac{3}{4} + 0.98 \cdot \frac{1}{4}} \\ &\approx 0.058\end{aligned}$$

- So Tammy's beliefs haven't shifted too much.
- More generally, randomness turns out to be a really useful technique for preventing information leakage.

Randomized Response

- How accurately can we estimate μ, the population mean?
- Let $X_{\mathrm{T}}^{(i)}$ denote individual i's response if they respond truthfully, and $X_{\mathrm{R}}^{(i)}$ individual i's response under the RR mechanism.
- Maximum likelihood estimate, if everyone responds truthfully:

$$\hat{\mu}_{\mathrm{T}} = \frac{1}{N} \sum_{i=1}^{N} X_{\mathrm{T}}^{(i)}$$

- Variance of the ML estimate:

$$\begin{aligned} \mathrm{Var}(\hat{\mu}_{\mathrm{T}}) &= \frac{1}{N} \mathrm{Var}(X_{\mathrm{T}}^{(i)}) \\ &= \frac{1}{N} \mu(1-\mu). \end{aligned}$$

Randomized Response

- How to estimate μ from the randomized responses $\{X_{\mathrm{R}}^{(i)}\}$?

$$\mathbb{E}[X_{\mathrm{R}}^{(i)}] = \frac{1}{4}(1-\mu) + \frac{3}{4}\mu$$
$$\Rightarrow \hat{\mu}_{\mathrm{R}} = \frac{2}{N}\sum_i X_{\mathrm{R}}^{(i)} - \frac{1}{2}$$

- Variance of the estimator:

$$\begin{aligned}\mathrm{Var}(\hat{\mu}_{\mathrm{R}}) &= \frac{4}{N}\mathrm{Var}(X_{\mathrm{R}}^{(i)}) \\ &\geq \frac{4}{N}\mathrm{Var}(X_{\mathrm{T}}^{(i)}) \\ &= 4\,\mathrm{Var}(\hat{\mu}_{\mathrm{T}})\end{aligned}$$

- The variance decays as $1/N$, which is good.
- But it is at least 4x larger because of the randomization. Can we do better?

Differential Privacy

Differential Privacy

Basic setup:

- There is a database $\mathcal{D}$ which potentially contains sensitive information about individuals.
- The database curator has access to the full database. We assume the curator is trusted.
- The data analyst wants to analyze the data. She asks a series of queries to the curator, and the curator provides a response to each query.
- The way in which the curator responds to queries is called the mechanism. We'd like a mechnism that gives helpful responses but avoids leaking sensitive information about individuals.

Differential Privacy

- Two databases $\mathcal{D}_1$ and $\mathcal{D}_2$ are neighbouring if they agree except for a single entry.
- **Idea:** if the mechanism behaves nearly identically for $\mathcal{D}_1$ and $\mathcal{D}_2$, then an attacker can't tell whether $\mathcal{D}_1$ or $\mathcal{D}_2$ was used (and hence can't learn much about the individual).
- **Definition:**
  - A mechanism $\mathcal{M}$ is ε-differentially private if for any two neighbouring databases $\mathcal{D}_1$ and $\mathcal{D}_2$, and any set $\mathcal{R}$ of possible responses

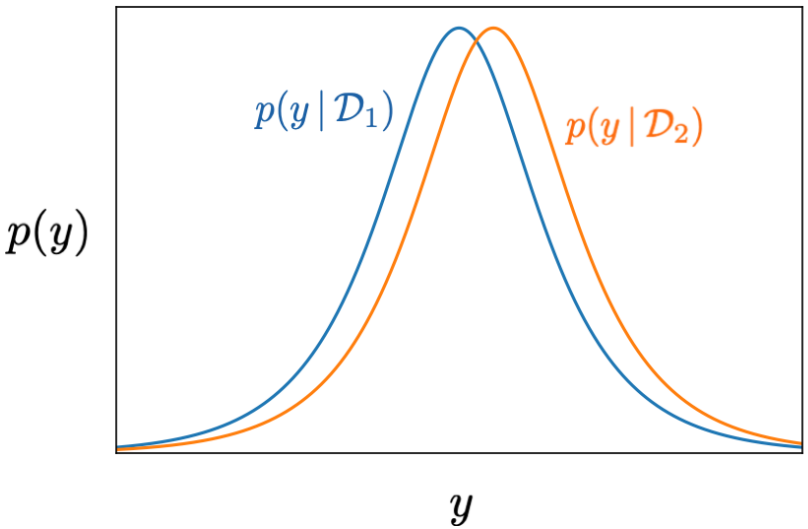
$$\Pr(\mathcal{M}(\mathcal{D}_1) \in \mathcal{R}) \leq \exp(\varepsilon) \Pr(\mathcal{M}(\mathcal{D}_2) \in \mathcal{R}).$$

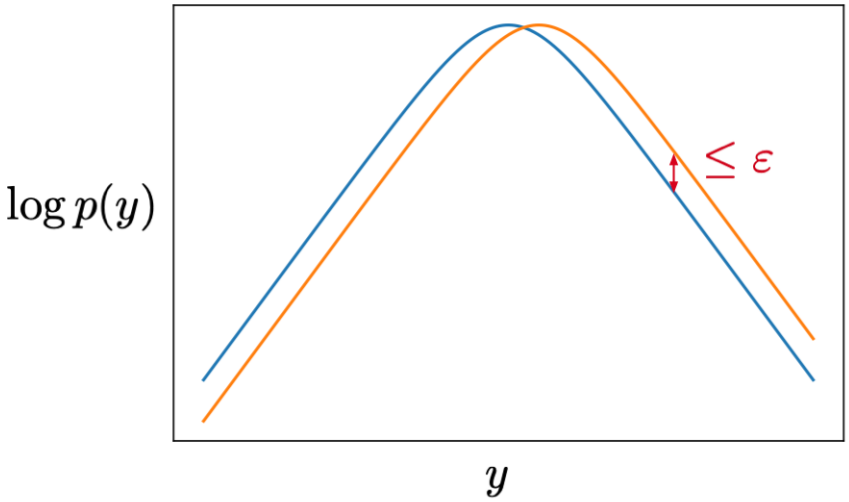
- **Note:** for small ε, $\exp(\varepsilon) \approx 1 + \varepsilon$.
- **A consequence:** for any possible response y,

$$\exp(-\varepsilon) \leq \frac{\Pr(\mathcal{M}(\mathcal{D}_1) = y)}{\Pr(\mathcal{M}(\mathcal{D}_2) = y)} \leq \exp(\varepsilon)$$

Differential Privacy

Visually:





Notice that the tail behavior is important.

Differential Privacy

- Anna is an attacker who wants to figure out if Patrick (x) is in the cancer database $\mathcal{D}$. Her prior probability for him being in the database is 0.4. $\mathcal{D}$ is ε-differentially private. She makes a query and gets back $y = \mathcal{M}(\mathcal{D})$.
- She's narrowed it down to two possible databases $\mathcal{D}_1$ and $\mathcal{D}_2$, which are identical except that $x \in \mathcal{D}_1$ and $x \notin \mathcal{D}_2$.
- After observing y, she computes her posterior probability using Bayes' Rule:

$$\begin{aligned}
\Pr(x \in \mathcal{D} \mid y) &= \frac{\Pr(x \in \mathcal{D}) \Pr(y \mid x \in \mathcal{D})}{\Pr(x \in \mathcal{D}) \Pr(y \mid x \in \mathcal{D}) + \Pr(x \notin \mathcal{D}) \Pr(y \mid x \notin \mathcal{D})} \\
&\geq \frac{\Pr(x \in \mathcal{D}) \Pr(y \mid x \in \mathcal{D})}{\Pr(x \in \mathcal{D}) \Pr(y \mid x \in \mathcal{D}) + \exp(\varepsilon) \Pr(x \notin \mathcal{D}) \Pr(y \mid x \in \mathcal{D})} \\
&= \frac{\Pr(x \in \mathcal{D})}{\Pr(x \in \mathcal{D}) + \exp(\varepsilon) \Pr(x \notin \mathcal{D})} \\
&\geq 0.4 \exp(-\varepsilon)
\end{aligned}$$

- Similarly, $\Pr(x \in \mathcal{D} \mid y) \leq 0.4 \exp(\varepsilon)$. So Anna hasn't learned much about Patrick.

Differential Privacy

- In what sense does this definition guarantee privacy?
- Suppose a data analyst takes the result $y = \mathcal{M}(\mathcal{D})$ and further processes it with some algorithm f (without peeking at the data itself). Is it still private?
- Let $\mathcal{R}$ be a set of possible outputs, and $\mathcal{R}'$ be the pre-image under f, i.e. $\mathcal{R}' = \{y : f(y) \in \mathcal{R}\}$.

$$\begin{aligned}\Pr(f(\mathcal{M}(\mathcal{D}_1)) \in \mathcal{R}) &= \Pr(\mathcal{M}(\mathcal{D}_1) \in \mathcal{R}') \\ &\leq \exp(\varepsilon) \Pr(\mathcal{M}(\mathcal{D}_2) \in \mathcal{R}') \\ &= \exp(\varepsilon) \Pr(f(\mathcal{M}(\mathcal{D}_2)) \in \mathcal{R})\end{aligned}$$

- Hence, the composition $f \circ \mathcal{M}$ is also ε-differentially private. No matter how clever the analyst is, or the resources she throws at it, she can't learn more than ε about an individual entry!

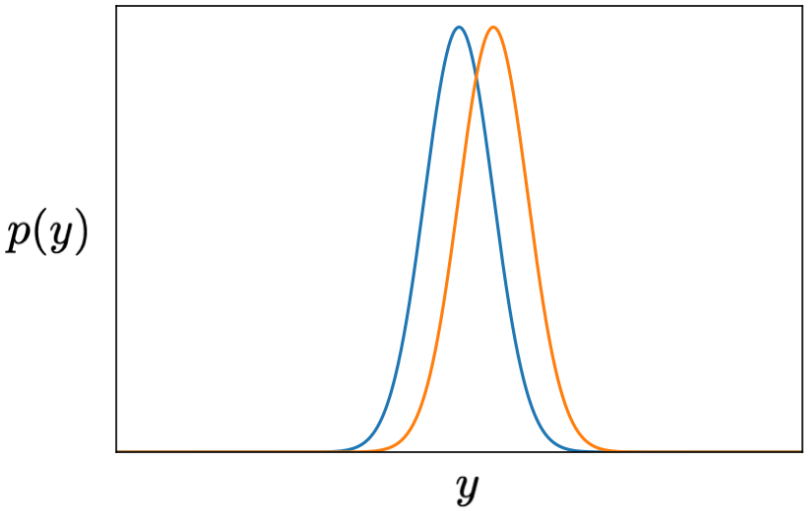
Laplace Mechanism

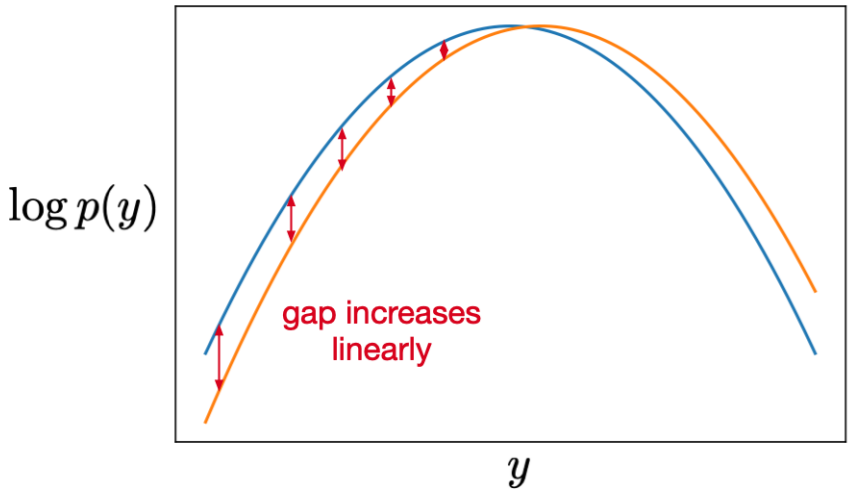
Laplace Mechanism

- A lot of queries we might want to ask can be seen as counting queries, i.e. counting the number of entries which have property $\mathcal{P}$.
  - E.g. naive Bayes, decision trees
- **Idea:** Maybe the mechanism can return noisy counts which are accurate enough for whatever analysis we're trying to do.

Laplace Mechanism

Attempt 1: Gaussian noise



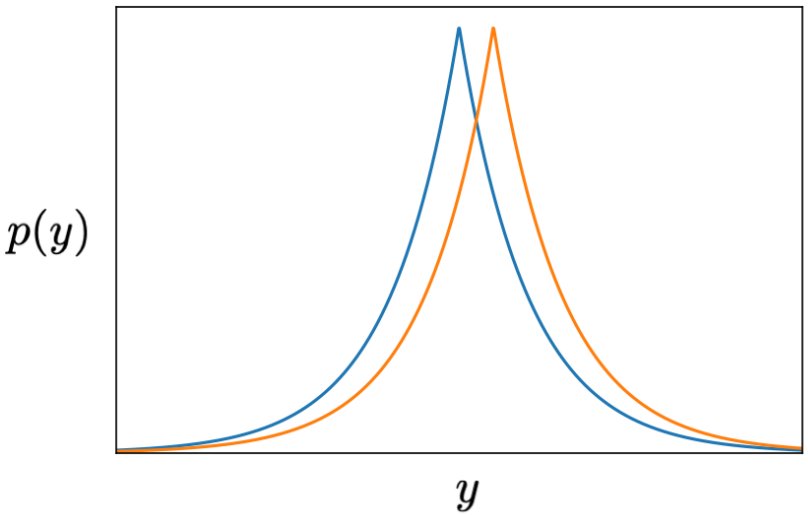


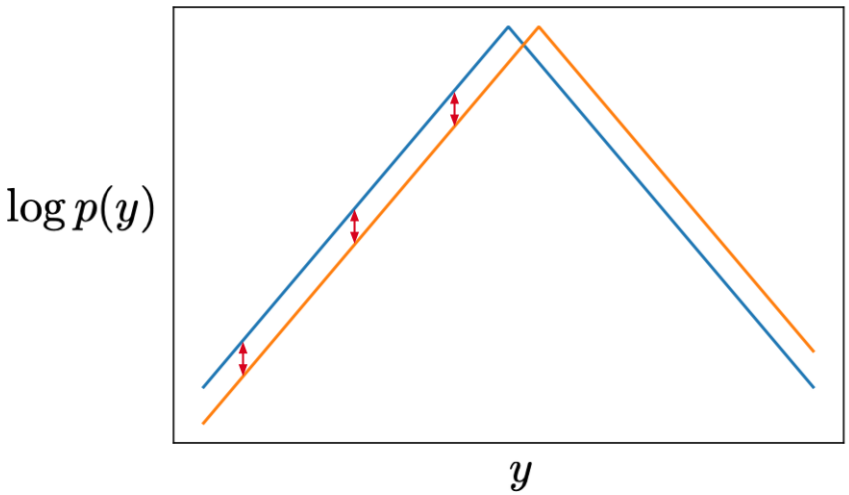
Gaussian noise violates our definition, but only because of the tails. It satisfies a different definition of differential privacy which allows violating the ε constraint with small probability, but that's beyond the scope of this lecture.

Laplace Mechanism

The Laplace distribution is just what we need.

$$p(y; \mu, b) = \frac{1}{2b} \exp\left(-\frac{|y - \mu|}{b}\right)$$





b is a parameter which determines the scale of the distribution.

Variance: $2b^2$

Laplace Mechanism

- Let f be a deterministic vector-valued function of a database. The L^1 sensitivity of f is defined as:

$$\Delta f = \max_{\substack{\mathcal{D}_1, \mathcal{D}_2 \\ \text{neighbours}}} \|f(\mathcal{D}_1) - f(\mathcal{D}_2)\|_1.$$

- Recall that $\|\mathbf{x}\|_1 = \sum_i |x_i|$.
- Suppose f returns the vector of counts of individuals who fall into k disjoint buckets. What is the L^1 sensitivity of f? (Ans: 1)

Laplace Mechanism

- Laplace mechanism: return a vector $\mathbf{y}$ whose entries are independently sampled from Laplace distributions

$$y_i \sim \text{Laplace}\left(f(\mathcal{D})_i,\ \frac{\Delta f}{\varepsilon}\right),$$

  where $f(\mathcal{D})_i$ denotes the ith entry of $f(\mathcal{D})$.
- The noise is calibrated to the privacy requirement: higher sensitivity queries and tighter privacy constraints imply more noise.

Laplace Mechanism

- **Claim:** the Laplace mechanism is differentially private.
- Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be two neighboring databases, and $y = \mathcal{M}(\mathcal{D})$.

$$
\begin{aligned}
\frac{p(\mathbf{y} \mid \mathcal{D}_1)}{p(\mathbf{y} \mid \mathcal{D}_2)} &= \frac{\prod_i \frac{\varepsilon}{2\Delta f} \exp(-\frac{\varepsilon |f(\mathcal{D}_1)_i - y_i|}{\Delta f})}{\prod_{i=1}^k \frac{\varepsilon}{2\Delta f} \exp(-\frac{\varepsilon |f(\mathcal{D}_2)_i - y_i|}{\Delta f})} \\
&= \prod_i \exp\left(\frac{\varepsilon(|f(\mathcal{D}_2)_i - y_i| - |f(\mathcal{D}_1)_i - y_i|)}{\Delta f}\right) \\
&\leq \prod_i \exp\left(\frac{\varepsilon(|f(\mathcal{D}_2)_i - f(\mathcal{D}_1)_i|)}{\Delta f}\right) && \text{(triangle ineq.)} \\
&= \exp\left(\frac{\varepsilon \sum_i |f(\mathcal{D}_2)_i - f(\mathcal{D}_1)_i|}{\Delta f}\right) \\
&= \exp\left(\frac{\varepsilon \|f(\mathcal{D}_2) - f(\mathcal{D}_1)\|_1}{\Delta f}\right) \\
&\leq \exp(\varepsilon) && \text{(defn. of } \Delta f)
\end{aligned}
$$

Laplace Mechanism

- **Example:** What fraction of Canadians have blue eyes?
- Mechanism returns the counts (ξ_1, ξ_2) of Canadians with and without blue eyes, plus Laplace noise. We'd like to satisfy a privacy constraint of $\varepsilon = 0.1$. How much Laplace noise should we add?
  - Ans: $\Delta f / \varepsilon = 1/0.1 = 10$.
- The noise scale is independent of the population size!
- I.e., you can answer the query to within about ± 10 people, out of the population of Canada. So you can obtain very accurate answers to queries over large populations.

Laplace Mechanism

Comparison to randomized response

- Recall the randomized response method:

	Yes	No
Dodge	3/4	1/4
No Dodge	1/4	3/4

- For what ε is this ε-differentially private? (Ans: $\log 3$)
- **Recall:** ML estimate from truthful responses has variance $\frac{1}{N}\mu(1-\mu)$ and estimate from randomized responses has variance at least 4x larger.
- **Laplace mechanism:** add Laplace noise η with scale $\Delta f/\varepsilon = 1/\log 3 \approx 0.91$

$$\hat{\mu}_{\mathrm{L}} = \frac{1}{N}\Big(\sum_{i=1}^{N} X_{\mathrm{T}}^{(i)} \; + \; \eta\Big)$$
$$= \hat{\mu}_{\mathrm{T}} + \frac{\eta}{N}$$

- The added noise has variance $\mathcal{O}(1/N^2)$, compared with the statistical error, which is $\mathcal{O}(1/N)$. So we lose almost no accuracy.

Laplace Mechanism

- **Example:** Naïve Bayes
- Suppose you have a target t which takes K_t possible values, and you have D different features x_j, each of which takes K_j possible values.
- Recall that to fit a naïve Bayes classifier, we need to calculate the counts of all the joint configurations (t, x_j) for each x_j.
- What is the scale of Laplace noise we should add to each count to make this differentially private with $\varepsilon = 0.1$?
  - The sensitivity is $\Delta f = D$, so we need $\Delta f / \varepsilon = 10D$.

Exponential Mechanism

Exponential Mechanism

- Suppose the goal of the analysis is to make a decision Y.
- We have a loss function $\mathcal{L}(Y, \mathcal{D})$ which determines how unhappy we are with any particular Y as a response for database $\mathcal{D}$.
- The exponential mechanism tries to pick a reasonably good decision subject to a privacy constraint. We do this by picking Y randomly as:

$$\Pr(Y = y) \propto \exp\left(-\frac{\varepsilon}{2\Delta\mathcal{L}}\mathcal{L}(y, \mathcal{D})\right)$$

- $\Delta\mathcal{L}$ is the sensitivity of $\mathcal{L}$, just like for the Laplace mechanism.
- The resulting probabilities are basically a softmax of $-\mathcal{L}$. Distributions of this form are also called Boltzmann distributions (from statistical mechanics).

Exponential Mechansim

- **Claim:** The exponential mechanism is ε-differentially private.
- For two neighboring databases $\mathcal{D}_1$ and $\mathcal{D}_2$, and any value y,

$$\frac{p(y \mid \mathcal{D}_1)}{p(y \mid \mathcal{D}_2)} = \frac{\frac{\exp\left(-\frac{\varepsilon}{2\Delta\mathcal{L}}\mathcal{L}(y,\mathcal{D}_1)\right)}{\sum_{y'}\exp\left(-\frac{\varepsilon}{2\Delta\mathcal{L}}\mathcal{L}(y',\mathcal{D}_1)\right)}}{\frac{\exp\left(-\frac{\varepsilon}{2\Delta\mathcal{L}}\mathcal{L}(y,\mathcal{D}_2)\right)}{\sum_{y'}\exp\left(-\frac{\varepsilon}{2\Delta\mathcal{L}}\mathcal{L}(y',\mathcal{D}_2)\right)}}$$

$$= \underbrace{\frac{\exp\left(-\frac{\varepsilon}{2\Delta\mathcal{L}}\mathcal{L}(y,\mathcal{D}_1)\right)}{\exp\left(-\frac{\varepsilon}{2\Delta\mathcal{L}}\mathcal{L}(y,\mathcal{D}_2)\right)}}_{\leq \exp(\varepsilon/2)} \cdot \underbrace{\frac{\sum_{y'}\exp\left(-\frac{\varepsilon}{2\Delta\mathcal{L}}\mathcal{L}(y',\mathcal{D}_2)\right)}{\sum_{y'}\exp\left(-\frac{\varepsilon}{2\Delta\mathcal{L}}\mathcal{L}(y',\mathcal{D}_1)\right)}}_{\leq \exp(\varepsilon/2)}$$

- Both inequalities are straightforward applications of the definition of $\Delta\mathcal{L}$.
- Hence, $\frac{p(y \mid \mathcal{D}_1)}{p(y \mid \mathcal{D}_2)} \leq \exp(\varepsilon)$, so we're done.

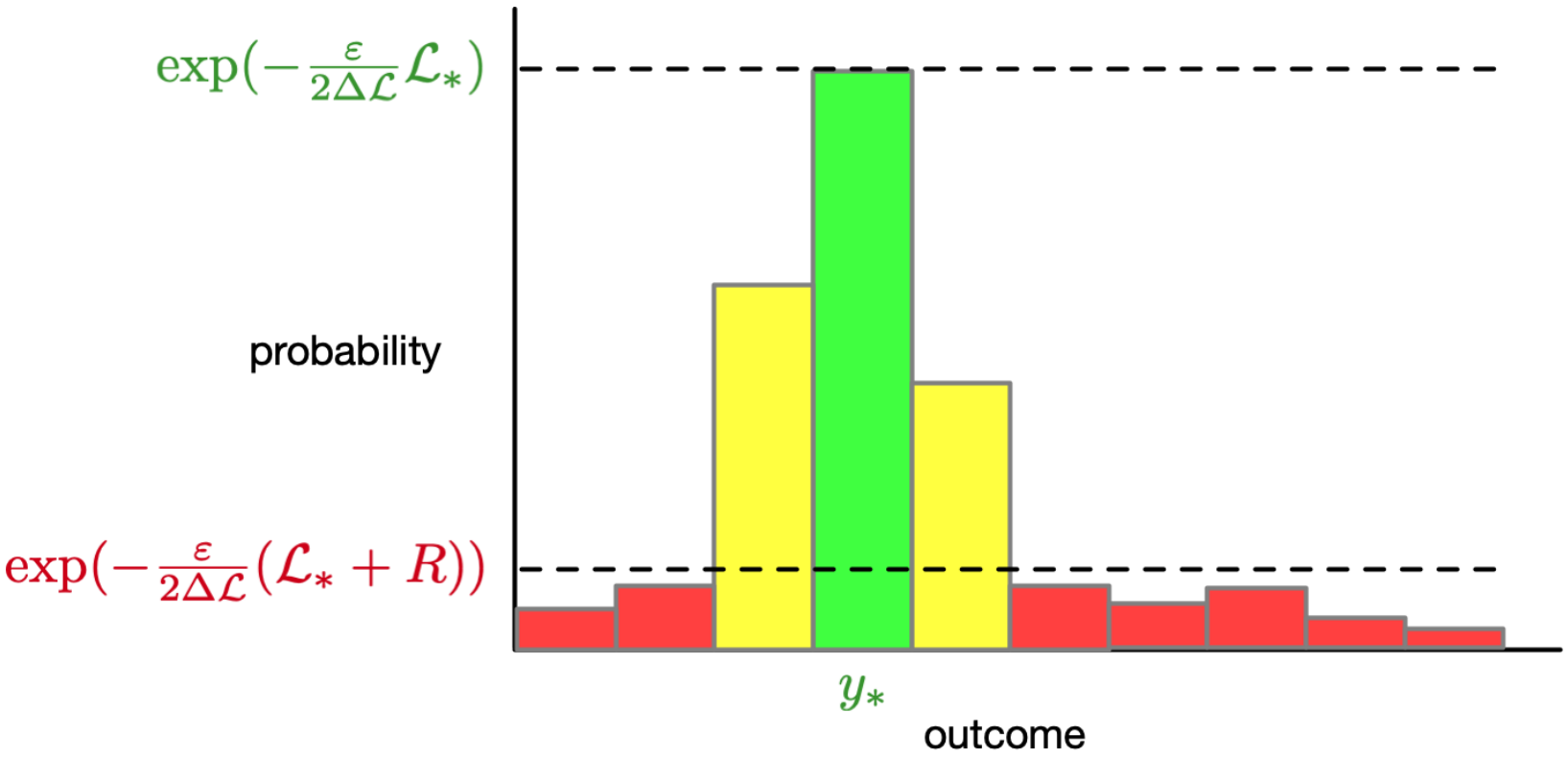
Exponential Mechanism

- **Claim:** For discrete Y, the exponential mechanism is unlikely to choose Y to be much worse than optimal.
- Let $y_* = \arg\min_y \mathcal{L}(y, \mathcal{D})$ and $\mathcal{L}_* = \mathcal{L}(y_*, \mathcal{D})$.
- Consider all the values y which are suboptimal by more than R, i.e. which have $\mathcal{L}(y, \mathcal{D}) \geq \mathcal{L}(y_*, \mathcal{D}) + R$.

$$\begin{aligned} p(y \mid \mathcal{D}) &= k \exp\left(-\frac{\varepsilon}{2\Delta\mathcal{L}}\mathcal{L}(y, \mathcal{D})\right) \\ &\leq k \exp\left(-\frac{\varepsilon}{2\Delta\mathcal{L}}(\mathcal{L}(y_*, \mathcal{D}) + R)\right) \\ &= k \exp\left(-\frac{\varepsilon}{2\Delta\mathcal{L}}\mathcal{L}(y_*, \mathcal{D})\right) \exp\left(-\frac{\varepsilon R}{2\Delta\mathcal{L}}\right) \\ &= p(y_* \mid \mathcal{D}) \exp\left(-\frac{\varepsilon R}{2\Delta\mathcal{L}}\right) \end{aligned}$$

- k is the normalizing constant that makes the probabilities sum to 1.
- There are at most $|Y|$ such values, where $|Y|$ is the size of Y's domain. Hence, their total probability is $|Y| \exp\left(-\frac{\varepsilon R}{2\Delta\mathcal{L}}\right)$.
- Hence, the probability of suboptimality by R decays exponentially in R, and you're unlikely to be suboptimal by more than $\mathcal{O}((\Delta\mathcal{L}/\varepsilon)\log|Y|)$.

Exponential Mechanism



$$|Y| \exp(-\tfrac{\varepsilon}{2\Delta\mathcal{L}}(\mathcal{L}_* + R)) \leq \delta \exp(-\tfrac{\varepsilon}{2\Delta\mathcal{L}}\mathcal{L}_*) \quad \Leftrightarrow \quad R > \tfrac{2\Delta\mathcal{L}}{\varepsilon}(\log|Y| - \log\delta)$$

Exponential Mechanism

- **Example:** inferring the parameter of a Bernoulli distribution
- Suppose we have a dataset $\mathcal{D} = \{x_1, \ldots, x_N\}$ of coin flips, and we want to estimate the bias θ while protecting the privacy of each individual coin flip with $\varepsilon = 0.1$.
- Our loss is negative log-likelihood:

$$\mathcal{L}(\hat{\theta}, \mathcal{D}) = -\log \prod_{i=1}^{N} p(x_i; \hat{\theta})$$

- What is the sensitivity $\Delta\mathcal{L}$?
  - Ans: $\Delta\mathcal{L} = \infty$, because an observation $x_i = 1$ has probability 1 under $\hat{\theta} = 1$ and probability 0 under $\hat{\theta} = 0$.
  - Hence, we can't use the exponential mechanism without further assumptions.

Exponential Mechanism

- Now suppose we restrict $\hat{\theta}$ to be in the interval $(0.1, 0.9)$. Now what is the sensitivity?
  - Ans: $\Delta\mathcal{L} = -\log 0.1 \approx 2.3$.
- The exponential mechanism samples $\hat{\theta}$ as

$$\begin{aligned} p(\hat{\theta} \mid \mathcal{D}) &\propto \exp\left(-\frac{\varepsilon}{2\Delta\mathcal{L}}\mathcal{L}(\hat{\theta}, \mathcal{D})\right) \\ &= \exp\left(0.022 \log \prod_{i=1}^{N} p(x_i; \hat{\theta})\right) \\ &= \prod_{i=1}^{N} p(x_i; \hat{\theta})^{0.022} \\ &= \hat{\theta}^{0.022 N_H}(1-\hat{\theta})^{0.022 N_T} \end{aligned}$$

- **Note:** This is a beta distribution with parameters $a = 1 + 0.022\, N_H$ and $b = 1 + 0.022\, N_T$, truncated to $(0.1, 0.9)$.
  - Hence, $\hat{\theta}$ is a lot like a Bayesian posterior sample, except that each observation only counts for 0.022.

Exponential Mechanism

- Let's compare the Laplace and exponential mechanisms for estimating $\hat{\theta}$.
- **Laplace mechanism**: compute the counts N_H and N_T, then add Laplace noise with scale $\Delta\mathcal{L}/\varepsilon = 22$.
  - $\hat{\theta} = \frac{\hat{N}_H}{\hat{N}_H + \hat{N}_T}$
  - Can show $\mathrm{Var}(\hat{\theta} \mid \mathcal{D}) = \mathcal{O}(1/N^2)$
- **Exponential mechanism**:
  $\hat{\theta} \sim \mathrm{TruncatedBeta}(1 + 0.022\, N_H, 1 + 0.022\, N_T)$
  - Can show $\mathrm{Var}(\hat{\theta} \mid \mathcal{D}) = \mathcal{O}(1/N)$
- So the Laplace mechanism is much more accurate in this case. But the exponential mechanism is still useful in cases that aren't easily formulated as counts. We'll see an elegant example later in this lecture.

Composition Rules

Composition Rules

- So far, we've been looking at one query in isolation. What if we want to answer more than one question from the data we've collected?
- Can't just repeatedly use the same mechanism independently
  - Suppose the analyst asks the same counting query K times, and the curator always responds independently using the Laplace mechanism.
  - The analyst can get arbitrarily accurate counts by averaging the responses, rendering the privacy guarantee meaningless.
- Can we relate the privacy of multiple queries to the privacy of a single query? Such a result is known as a composition rule.

Composition Rules

- The easiest case is when the queries are non-adaptive, i.e. the analyst(s) make the queries without seeing the results of previous queries.
- **Claim:** Querying an ε-differentially private mechanism K times non-adaptively is $K\varepsilon$-differentially private.
- Letting y_1, y_2 be the responses, we have $y_1 \perp\!\!\!\perp y_2 \,|\, \mathcal{D}$. So,

$$\begin{aligned}\frac{p(y_1, y_2 \,|\, \mathcal{D}_1)}{p(y_1, y_2 \,|\, \mathcal{D}_2)} &= \frac{p(y_1 \,|\, \mathcal{D}_1)}{p(y_1 \,|\, \mathcal{D}_2)} \frac{p(y_2 \,|\, \mathcal{D}_1)}{p(y_2 \,|\, \mathcal{D}_2)} \\ &\leq \exp(\varepsilon) \cdot \exp(\varepsilon) \\ &= \exp(2\varepsilon)\end{aligned}$$

- **Corrollary:** if your privacy budget is ε, you should make sure the privacy parameters of the individual queries sum up to ε.

Composition Rules

- **Example:** Recall that for naïve Bayes, we made a counting query that requests the joint counts of (t, x_j) for each feature x_j.
  - We concluded that $\Delta f = D$, so the Laplace mechanism adds Laplace noise with scale D/ε.
- We can alternatively formulate this as D different queries, chosen non-adaptively, each of which asks for the joint counts (t, x_j) for *one* feature x_j.
  - To satisfy a privacy budget of ε, each query should be $\frac{\varepsilon}{D}$-differentially private.
  - The sensitivity of each query is $\Delta f_j = 1$.
  - So we should add Laplace noise with scale $\Delta f_j/(\varepsilon/D) = D/\varepsilon$.
- Hence, the composition rule agrees with the basic Laplace mechanism for this example.

Small Database Mechanism

Small Database Mechanism (optional)

- You might notice a problem: if you have a privacy budget of ε and need to make lots of queries, then don't you need a ridiculously small privacy budget for each one?
- **Idea:** You can answer lots of queries as long as you remember to tell the same lies every time.
- E.g., if the analyst asks the same query K times, and the curator gives the same answer every time, then there's no additional privacy loss.
- But what to do about queries that are just slightly different?

Small Database Mechanism (optional)

- Assume we're given a set of scalar-valued counting queries (all at once) $\{f_k\}_{k=1}^K$, each of which estimates the expectation of some function $\phi_k(x)$ with values in $[0, 1]$.

$$f_k(\mathcal{D}) = \frac{1}{N} \sum_i \phi_k(x^{(i)}),$$

  where N is the number of entries. Note: each $\Delta f_k \approx 1/N$.
- Small database mechanism: construct a fake database $\hat{\mathcal{D}}$ in a differentially private way, and then use $\hat{\mathcal{D}}$ to answer all the queries.
- We'll select $\hat{\mathcal{D}}$ (from the set of all possible databases of a certain size $\hat{N}$) using the exponential mechanism.
- The loss is the maximum error for any query:

$$\mathcal{L}(\hat{\mathcal{D}}, \mathcal{D}) = \max_k |f_k(\hat{\mathcal{D}}) - f_k(\mathcal{D})|$$

- What is the sensitivity $\Delta\mathcal{L}$? (Ans: $1/N$)

Small Database Mechanism (optional)

- Suppose there are K queries and you want to answer them all to an error of at most α.
- Set the size of the small database to $\hat{N} = \log_2 K/\alpha^2$.
- The exponential mechanism automatically satisfies differential privacy. The curator could even release the small database!
- The hard part is showing that the results are accurate.

Small Database Mechanism (optional)

- **Fact:** there exists at least one database $\hat{\mathcal{D}}$ of size $\hat{N}$ such that

$$\mathcal{L}(\hat{\mathcal{D}}, \mathcal{D}) = \max_k |f_k(\hat{\mathcal{D}}) - f_k(\mathcal{D})| < \alpha.$$

  Hence, $\mathcal{L}_* \leq \alpha$.
  - Elegant combinatorial proof in Dwork & Roth (section 4.1)
- Now we apply our previous result showing the exponential mechanism produces a result with loss not much more than $\mathcal{L}_*$.

Small Database Mechanism (optional)

- Number of small databases: $|\mathcal{Y}| = |\mathcal{X}|^{\log_2 K/\alpha^2}$, where $\mathcal{X}$ is the domain of the entries. E.g., $|\mathcal{X}| = 2^D$ for D binary features.
- Showed earlier that with probability $1 - \delta$,

$$\mathcal{L}(y, \mathcal{D}) < \mathcal{L}_* + \frac{2\Delta\mathcal{L}}{\varepsilon}(\log|\mathcal{Y}| - \log\delta)$$

- Plugging in $\mathcal{L}_* < \alpha$, $\Delta\mathcal{L} = 1/N$, and $|\mathcal{Y}| = |\mathcal{X}|^{\log_2 K/\alpha^2}$, we have that with probability $1 - \delta$,

$$\mathcal{L}(\hat{\mathcal{D}}, \mathcal{D}) < \alpha + \frac{2}{\varepsilon N}\left(\frac{\log_2 K}{\alpha^2}\log|\mathcal{X}| - \log\delta\right)$$

- Notice that R is proportional to $\log K/N$. Hence, the number of queries we can answer accurately is *exponential* in N!

Odds and Ends

Federated Learning (optional)

- So far, we've assumed there's a curator who we trust with access to all the raw data.
- What if a company (say Google) wants to learn a classifier from the images stored on everyone's phones, but without having to send the images to Google?
- Federated learning: learning a model without any centralized entity having access to all the data
  - Google sends the phone the current weights of the network
  - The phone does a small number of steps of gradient descent, and communicates the local update back to Google
  - Google updates their network by adding the local update
- Does this satisfy differential privacy?
  - Not automatically, but the local updates could be randomized in a way that makes them differentially private.
- `https://ai.googleblog.com/2017/04/federated-learning-collaborative.html`

Recap

- A lot of ML models are trained on datasets containing sensitive information about individuals, and database reconstruction attacks can be surprisingly effective.
- Differential privacy gives a way of *provably* preventing (much) information about individuals from leaking.
- Building blocks of differential privacy
  - Laplace mechanism (add noise to counts)
  - Exponential mechanism (randomize a selection)
  - Composition rules (combine multiple private queries)
- Sometimes differentially private algorithms can accurately answer queries for large populations.
- The 2020 US Census will use differential privacy: `https://www.youtube.com/watch?v=yUyCYC6rb_4`