CSCE 658: Randomized Algorithms

Lecture 22

Samson Zhou

Relevant Supplementary Material

- Chapter 3-4 of "The Algorithmic Foundations of Differential Privacy", by Cynthia Dwork and Aaron Roth (https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf)

Previously: Differential Privacy

- [DMNS06] Given $\varepsilon > 0$ and $\delta \in (0,1)$, a randomized algorithm $A: U^* \to Y$ is (ε, δ)-differentially private if, for every neighboring pair D and D' of datasets, and for all $E \subseteq Y$,

$$\Pr[A(D) \in E] \le e^{\varepsilon} \cdot \Pr[A(D') \in E] + \delta$$

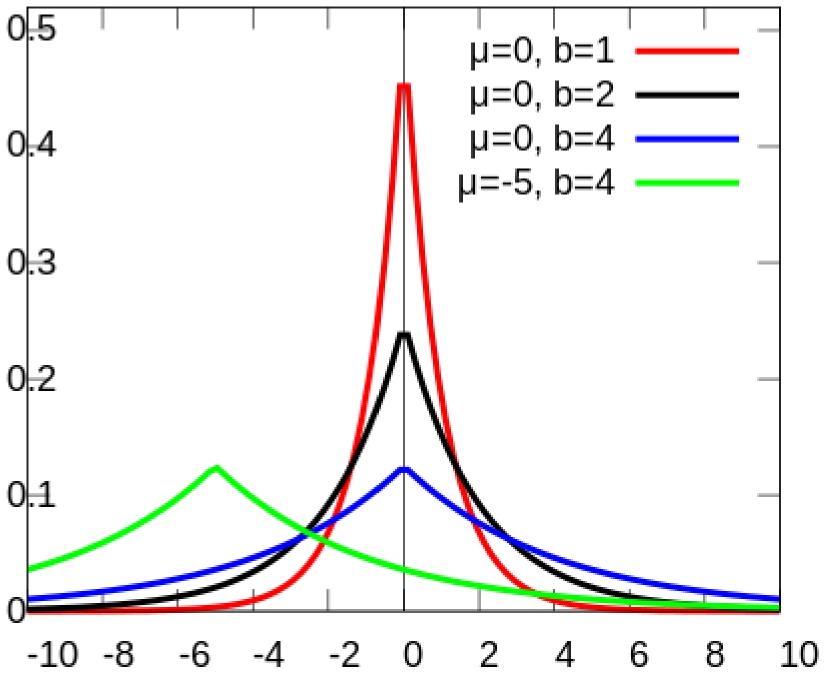
$A($    $) \approx A($    $)$

Previously: Laplace Mechanism

- Output: Algorithm computes $f(x)$ and releases $f(x) + Z$, where $Z \sim \mathrm{Lap}\left(\frac{\sigma_f}{\varepsilon}\right)$

- Laplacian distribution: Probability density function for $\mathrm{Lap}(b)$ is

$$p(x) = \frac{1}{2b} \exp\left(-\frac{|x|}{b}\right) = \frac{1}{2b} e^{\left(-\frac{|x|}{b}\right)}$$



Previously: Exponential Mechanism

- Choose a score function $S: (X^n, Y) \to \mathbb{R}$ and global sensitivity σ

- Sample $y \in Y$ with probability proportional to $\exp\left(\frac{\varepsilon}{2\sigma} S(x, y)\right)$

Mechanisms: Exponential vs. Laplace

- Laplace mechanism: Output $y = z + x$ with probability proportional to $\frac{\varepsilon}{2\sigma_f} e^{\left(-\frac{\varepsilon|x|}{\sigma_f}\right)}$
- Exponential mechanism: Output $y = z + x$ with probability proportional to $e^{\left(-\frac{\varepsilon|x|}{\sigma_f}\right)}$
- Recovers the Laplace mechanism!

Exponential Mechanism Drawbacks

- Sampling process may be inefficient

- Error can be large

Counting Queries

- How many people in this class have pets?
- How many people in this class were at Kyle Field last weekend?

- Suppose we have a set Q of q queries on a database D
- How do we privately answer the queries?

Linear Queries

- Let D be a database for d features on n users
- Let $f \in \mathbb{R}^d$ be the corresponding frequency vector, so that $|f_i| \leq n$ for all $i \in [d]$

- A linear query is $\ell(f) = \frac{1}{n}\sum_{i=1}^{d} a_i f_i = \frac{1}{n}(a_1 f_1 + \ldots + a_d f_d)$
- Suppose we have a set Q of q queries on a database D
- How do we privately answer the queries?

Counting Queries

- How many people in this class have pets?
- How many people in this class besides the instructor have pets?

- Intuition: we do not need to use additional privacy budget to answer the second query

Linear Queries

- Let D be a database for d features on n users
- A linear query is $\ell(f) = \frac{1}{n}\sum_{i=1}^{d} a_i f_i = \frac{1}{n}(a_1 f_1 + \ \ldots + a_d f_d)$
- Suppose we have a set Q of q queries on a database D
- How might we answer the queries *non-privately*, say with target additive error α?

Linear Queries

- A linear query is $\ell(f) = \frac{1}{n}\sum_{i=1}^{d} a_i f_i = \frac{1}{n}(a_1 f_1 + \ \ldots + a_d f_d)$
- Let's normalize, so that $a_i, f_i \in [0,1]$ for all $i \in [d]$

- Suppose we have a set Q of q queries on a database D
- How might we answer the queries *non-privately*, say with target additive error α?

Linear Queries

- Suppose we sample $O\left(\frac{1}{\alpha^2}\right)$ items of the database D into a database $\widetilde{D}$ and answer $\ell(\tilde{f}) = \frac{1}{|D'|}\sum_{i=1}^{d} a_i \widetilde{f}_i$, where $\tilde{f}$ is the frequency vector for $\widetilde{D}$

Linear Queries

- Suppose we sample $O\left(\frac{1}{\alpha^2}\right)$ items of the database D into a database $\widetilde{D}$ and answer $\ell(\tilde{f}) = \frac{1}{|D'|}\sum_{i=1}^{d} a_i \widetilde{f_i}$, where $\tilde{f}$ is the frequency vector for $\widetilde{D}$

- We have $E\left[\ell(\tilde{f})\right] = \ell(f)$

Additive Chernoff Bound

- Additive Chernoff bound: Let $X_1, \dots, X_n \in [0,1]$ be independent random variables and let $X = X_1 + \cdots + X_n$ have expected value μ. Then for any $t \geq 0$:

$$\Pr[|X - \mu| \geq t] \leq 2e^{-2nt^2}$$

Linear Queries

- Suppose we sample $O\left(\frac{1}{\alpha^2}\right)$ items of the database D into a database $\widetilde{D}$ and answer $\ell(\tilde{f}) = \frac{1}{|D'|}\sum_{i=1}^{d} a_i \widetilde{f_i}$, where $\tilde{f}$ is the frequency vector for $\widetilde{D}$

- We have $E\left[\ell(\tilde{f})\right] = \ell(f)$ and thus by additive Chernoff bound, $\left|\ell(\tilde{f}) - \ell(f)\right| < \alpha$ with probability 0.99

Linear Queries

- This gives correctness for a single query
- How to handle a set Q of q queries?
- Sample $O\left(\frac{\log q}{\alpha^2}\right)$ items of the database and do median-of-means

SmallDB Algorithm

- Let V be the set of vectors v with $\|v\|_1 = O\left(\frac{\log q}{\alpha^2}\right)$
- Define $S(f, v) = -\max_{\ell \in Q} |\ell(f) - \ell(v)|$
- Sample and output $v \in V$ with the exponential mechanism with score function $S(f, v)$

SmallDB Summary

- The SmallDB mechanism is $(\varepsilon, 0)$-differentially private
- For the vector v output by the SmallDB algorithm (if f is sufficiently "large"), we have

$$\max_{\ell \in Q} |\ell(f) - \ell(v)| \leq \alpha$$

- Example of synthetic dataset

Approximate DP

- [DMNS06] Given $\varepsilon > 0$ and $\delta \in (0,1)$, a randomized algorithm $A: U^* \to Y$ is (ε, δ)-differentially private if, for every neighboring pair D and D' of datasets, and for all $E \subseteq Y$,
$$\Pr[A(D) \in E] \leq e^{\varepsilon} \cdot \Pr[A(D') \in E] + \delta$$

$A($    $) \approx A($    $)$

Approximate DP

- δ denotes is an additive term between two probability distributions
- Interpretation: Outside of probability δ, the algorithm must satisfy Pure DP
- Rephrasing: Privacy may fail with probability δ

Approximate DP

- Can enable better utility

- Gaussian mechanism: Algorithm computes $f(x)$ and releases $f(x) + Z$, where $Z \sim \mathcal{N}\left(\frac{\sigma_f^2}{\varepsilon}\right)$, where σ_f is the L_2 sensitivity of f

$$\sigma_f = \max_{\text{neighbors } x,x'} \|f(x) - f(x')\|_2$$

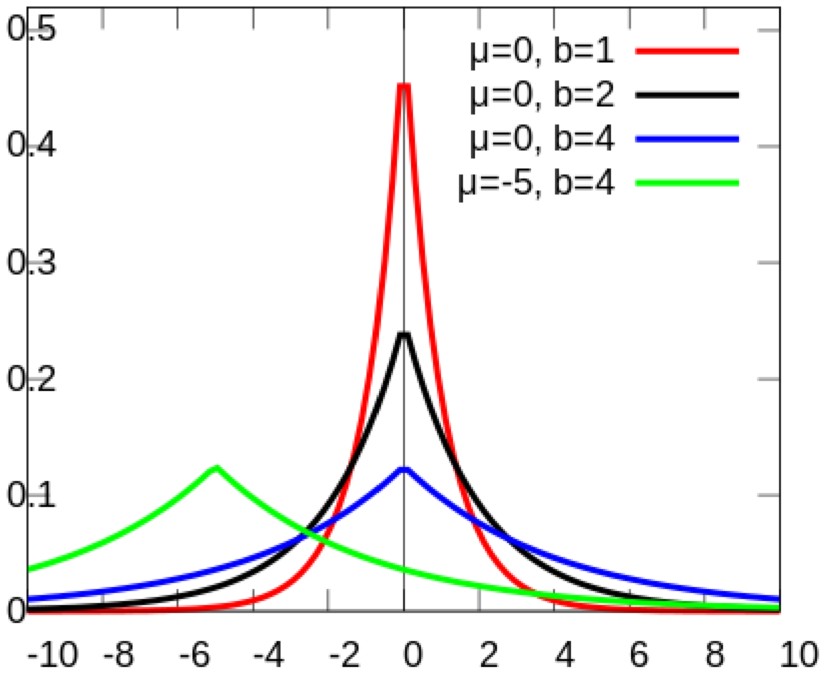
- Can result in smaller error, but only approximate DP

Laplace Mechanism

- Output: Algorithm computes $f(x)$ and releases $f(x) + Z$, where $Z \sim \mathrm{Lap}\left(\frac{\sigma_f}{\varepsilon}\right)$

- Laplacian distribution: Probability density function for $\mathrm{Lap}(b)$ is

$$p(x) = \frac{1}{2b}\exp\left(-\frac{|x|}{b}\right) = \frac{1}{2b}e^{\left(-\frac{|x|}{b}\right)}$$



Catastrophic Mechanism

- Output: Algorithm computes $f(x)$

- With probability δ, release $f(x)$
- With probability $1 - \delta$, release $f(x) + Z$, where $Z \sim \text{Lap}\left(\frac{\sigma_f}{\varepsilon}\right)$

Catastrophic Mechanism

- Catastrophic mechanism is (ε, δ)-DP
- Can release the entire dataset in the clear!
- Fortunately, most (ε, δ)-DP mechanisms do not fail catastrophically

Martingales

- Concentration inequalities when the random variables are not independent

- Azuma, Doob, etc.

Semester Review: Why Randomized Algorithms?

- Polynomial identity testing

- Karger's min-cut algorithm

Semester Review: Why Randomized Algorithms?

- Relevant study material: Randomized Algorithms and Probabilistic Analysis, by Greg Valiant (http://web.stanford.edu/class/cs265/)

- Relevant study material: Chapter 1, Randomized Algorithms, by Rajeev Motwani and Prabhakar Raghavan

Semester Review: Probability Unit

- Basic probability (conditional probability, joint probability)
- Expectation, variance, moments
- Concentration inequalities (Markov, Chebyshev, exponential tail bounds, e.g., Chernoff, Bernstein)

Trivia Question #1 (Birthday Paradox)

- Suppose we have a fair n-sided die. How many times should we roll the die before the probability we see a repeated outcome among the rolls is at least $\frac{1}{2}$? Example: $1, 5, 2, 4, 5$

- $\Theta(1)$
- $\Theta(\log n)$
- $\Theta(\sqrt{n})$
- $\Theta(n)$

Trivia Question #3 (Max Load)

- Suppose we have a fair n-sided die that we roll n times. “On average”, what is the largest number of times any outcome is rolled? Example: $1, 5, 2, 4, 1, 3, 1$ for $n = 7$

- $\Theta(1)$
- $\tilde{\Theta}(\log n)$
- $\tilde{\Theta}(\sqrt{n})$
- $\tilde{\Theta}(n)$

Trivia Question #4 (Coupon Collector)

- Suppose we have a fair n-sided die. "On average", how many times should we roll the die before we see all possible outcomes among the rolls? Example: $1, 5, 2, 4, 1, 3, 1, 6$ for $n = 6$

- $\Theta(n)$
- $\Theta(n \log n)$
- $\Theta(n\sqrt{n})$
- $\Theta(n^2)$

Semester Review: Probability Unit

- Relevant study material: Randomized Algorithms and Probabilistic Analysis, by Greg Valiant (http://web.stanford.edu/class/cs265/)

- Relevant study material: Chapters 3-4, Randomized Algorithms, by Rajeev Motwani and Prabhakar Raghavan

Semester Review: Big Data Unit

- Dimensionality reduction (Johnson-Lindenstrauss, coresets)

- Streaming algorithms (insertion, insertion-deletion)
  - Reservoir Sampling
  - Heavy-Hitters (Misra-Gries, CountMin, CountSketch)
  - Norm estimation (AMS)
  - Information theory, lower bounds
  - Clustering

Semester Review: Big Data Unit

- Relevant study material: Data Stream Algorithms, by Amit Chakrabarti (https://www.cs.dartmouth.edu/~ac/Teach/CS35-Fall23/)

- Relevant study material: Chapters 3-4, Data Streams: Algorithms and Applications, by S. Muthukrishnan

Semester Review: Probabilistic Method

- Suppose we want to argue the existence of a certain desirable object
- Existential argument, non-constructive
- If there is an algorithm that can find it, it must exist!
- A random variable cannot always be less than its expected value
- A random variable cannot always be more than its expected value

Semester Review: Probabilistic Method

- Relevant study material: Chapter 5, Randomized Algorithms, by Rajeev Motwani and Prabhakar Raghavan

Semester Review: LLL

- Approach to argue the existence of something that satisfies a large number of constraints

- Probabilistic method (existence, not constructive)

Semester Review: LLL

- Relevant study material: Chapter 5, Randomized Algorithms, by Rajeev Motwani and Prabhakar Raghavan

Semester Review: Linear Programming

- Formulating LPs
- Duality
- Integer linear programs and rounding

Semester Review: Linear Programming

- Chapter 29 in “Introduction to Algorithms”, by Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein

- Chapters 5.1-5.5 in “The Design of Approximation Algorithms”, by David P. Williamson and David B. Shmoys

Semester Review: Online Learning

- Weighted majority
- Randomized weighted majority
- Multiplicative weights
- Hedge

Semester Review: Online Learning

- Lecture 13 of "Advanced Algorithms" Course Notes (http://www.cs.cmu.edu/afs/cs.cmu.edu/academic/class/15850-f20/www/notes/lec14.pdf), by Anupam Gupta

- Lecture 14 of "Advanced Algorithms" Course Notes (http://www.cs.cmu.edu/afs/cs.cmu.edu/academic/class/15850-f20/www/notes/lec15.pdf), by Anupam Gupta

Semester Review: Differential Privacy

- Randomized response
- Basic properties of DP (composition, post-processing)
- Laplace mechanism
- Exponential mechanism

Semester Review: Differential Privacy

- Chapters 3-4 of “The Algorithmic Foundations of Differential Privacy”, by Cynthia Dwork and Aaron Roth (https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf)