CS2080: Applied Privacy for Data Science Beyond Noise Addition & Synthetic Data

School of Engineering & Applied Sciences
Harvard University

March 5, 2025

BEYOND NOISE ADDITION

Beyond Noise Addition

Fact: Laplace (or geometric) mechanism has the optimal worst-case, additive accuracy for answering a single real-valued query with DP.

Q: Why might we want other mechanisms?

Approaches we'll see:

1. Decompose a computation into several queries that can be answered with Laplace (cf. HW4 #2)

2. The exponential mechanism: applies even to discrete outcomes (ex: median)

3. Approaches based on local sensitivity and restricted sensitivity ~~(ex: graph statistics → beyond tabular data)~~

DP Medians

Recall: median over $\mathcal{X} = [0,1]$ has global sensitivity $\geq 1/2$

- Laplace Mechanism is useless.
- Same applies for discrete data, e.g. education variable $\in \{bb, 1, 2, \dots, 16\}$ in PUMS dataset.

Q: what mechanism from prior weeks could use to estimate the median of a discrete variable?

A:

Exponential Mechanism

Given candidate space $\mathcal{C}$ and score function $s: \mathcal{X} \times \mathcal{C} \to \mathbb{R}$, the exponential mechanism is:

$$M(x): \text{output } c \in \mathcal{C} \text{ with probability} \propto \exp\left(\varepsilon \cdot \frac{s(x,c)}{2 \cdot \Delta s}\right),$$

$$\text{where } \Delta s \overset{\text{def}}{=} \max_{x \sim x', c} |s(x,c) - s(x',c)|.$$

$$= \max_{x \sim x'} \|s(x,\cdot) - s(x',\cdot)\|_\infty$$

Thm: above mechanism is ε-DP.

Q: what is a good score function for the median?
A:
Why?

Utility of the Exp Mech Median

Next time: notebook for experiments comparing using exponential mechanism vs. histograms for the median.

Theoretical explanation: whp

- Exp mech outputs c with $s(x,c) \geq -O(\log|\mathcal{C}|)/\varepsilon$.

- Histogram outputs y with $s(x,c) \geq -O(|\mathcal{C}|^{1/2})/\varepsilon$.

Q: How to implement for continuous $\mathcal{C} = [0,1]$?

Local Sensitivity

- Even when $\Delta q = \mathrm{GS}_q$ is large, the local sensitivity may often be small on many natural datasets x:
$$\mathrm{LS}_q(x) \stackrel{\text{def}}{=} \max_{x':x'\sim x} |q(x') - q(x)|.$$

- Adding noise proportional to local sensitivity is not DP (why?)

- But there are several DP methods that approximate this idea (smooth sensitivity, propose-test-release, privately bounding local sensitivity, restricted sensitivity) when local sensitivity is small on all "nearby" or "similar" datasets.

- Examples: DP graph analysis (Spr 2022 2/22 slides), HW5 1c
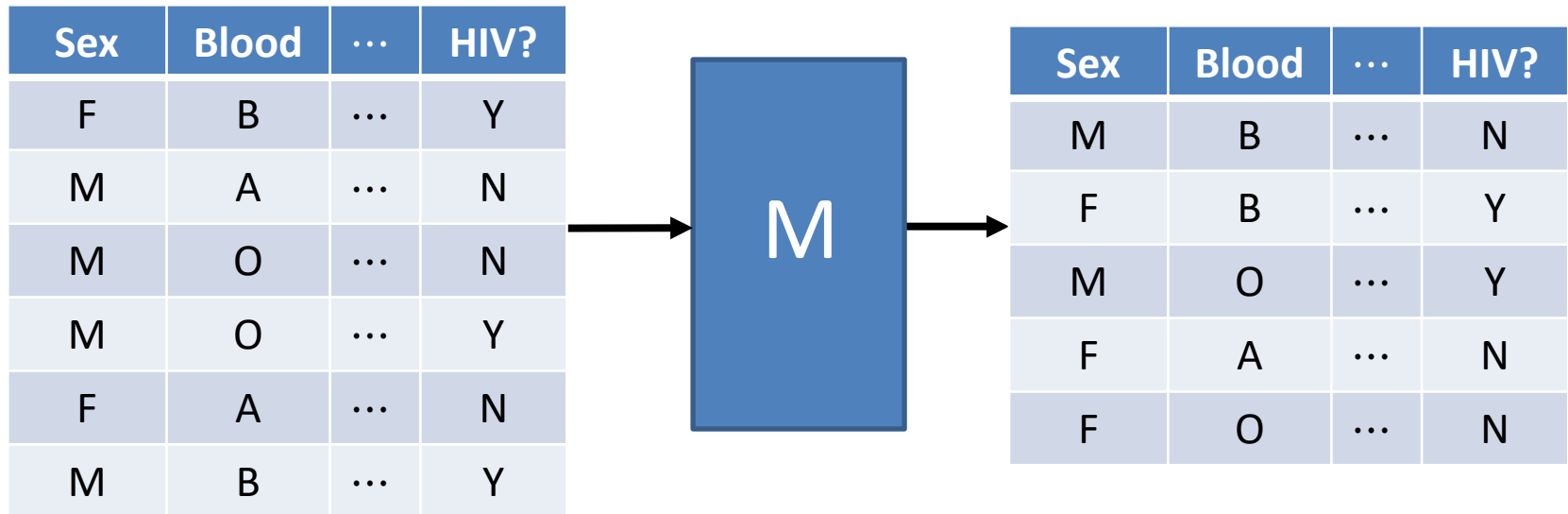
ONE-SHOT RELEASES: SYNTHETIC DATA

One-Shot Releases

Sex	Blood	⋯	HIV?
F	B	⋯	Y
M	A	⋯	N
M	O	⋯	N
M	O	⋯	Y
F	A	⋯	N
M	B	⋯	Y

M

DP mechanism

data analysts
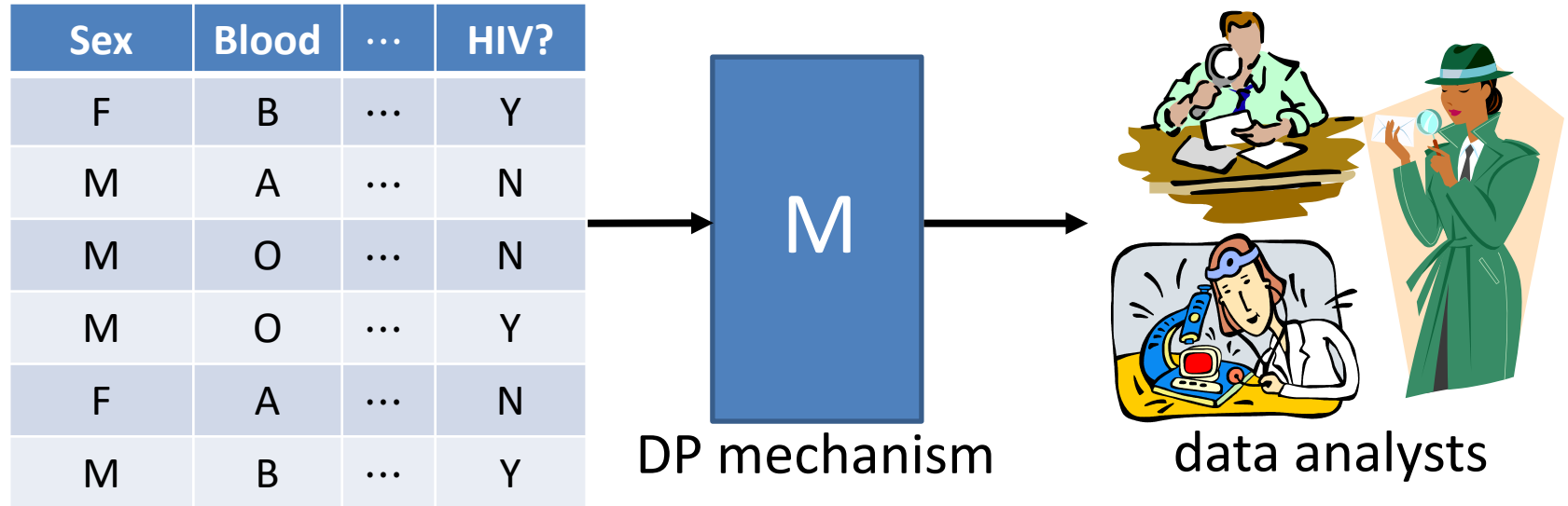
Goal: release as much useful info as possible given privacy budget
- Ideally support unforeseen analyses
- Summary statistics
- ML model
- Synthetic data

Differentially Private Synthetic Data

Sex	Blood	⋯	HIV?
F	B	⋯	Y
M	A	⋯	N
M	O	⋯	N
M	O	⋯	Y
F	A	⋯	N
M	B	⋯	Y

M

Sex	Blood	⋯	HIV?
M	B	⋯	N
F	B	⋯	Y
M	O	⋯	Y
F	A	⋯	N
F	O	⋯	N

$M: \mathcal{X}^n \rightarrow \mathcal{X}^m$ such that:

– $M(x)$ has the same schema as a real dataset.

– $M(x)$ reflects many statistical properties of x.

– M is differentially private.

Synthetic Data via DP Histograms

1. Use singleton bins $B_y = \{y\}$ for each $y \in \mathcal{X}$.

2. Construct a DP histogram $\left(a_1, \ldots, a_{|\mathcal{X}|}\right) \leftarrow M_{\text{hist}}(x)$, where $a_y \approx \#\{i : x_i = y\}$.

3. Output synthetic dataset $\hat{x}$ with a_y copies of each element y.

Difficulties?

Stability-Based Histogram

1. Let $B_1, \dots, B_k \subseteq \mathcal{X}$ be disjoint bins.

2. Define $q_j : \mathcal{X}^n \to \{0,1\}$ by $q_j(x) = \#\{i : x_i \in B_j\}$.

3. For each j s.t. $q_j(x) > 0$:

   a) Let $a_j = q_j(x) + Z_j$ for $Z_j \sim \text{Geo}(2/\varepsilon)$.

   b) If $a_j > \left\lceil \frac{2}{\varepsilon} \cdot \ln \frac{1}{\delta} \right\rceil$, output (j, a_j).

4. Treat all other bins as having a zero count.

Intuition for (ε, δ)-DP:

Stability-Based Histogram

1. Let $B_1, \ldots, B_k \subseteq \mathcal{X}$ be disjoint bins.

2. Define $q_j : \mathcal{X}^n \to \{0,1\}$ by $q_j(x) = \#\{i : x_i \in B_j\}$.

3. For each j s.t. $q_j(x) > 0$:

   a) Let $a_j = q_j(x) + Z_j$ for $Z_j \sim \mathrm{Geo}(2/\varepsilon)$.

   b) If $a_j > \left\lceil \frac{2}{\varepsilon} \cdot \ln \frac{1}{\delta} \right\rceil$, output (j, a_j).

4. Treat all other bins as having a zero count.

Benefits:

- Computation and output size linear in n rather than $|\mathcal{X}|$.
- Max error $O((1/\varepsilon) \cdot \ln(1/\delta))$ whp, independent of $|\mathcal{X}|$.
- But still can have poor utility when $|\mathcal{X}|$ large. (Why?)
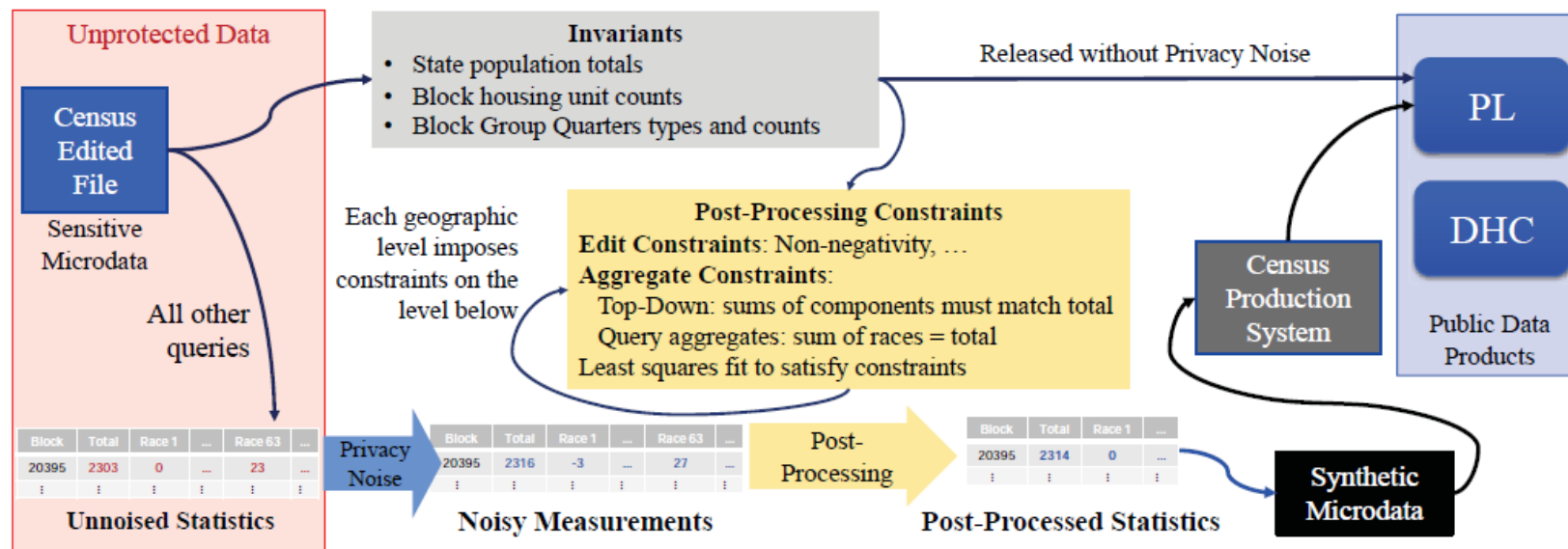
CASE STUDY: 2020 U.S. CENSUS

Census DAS Process

Figure 5-1: Process used to produce privacy-protected data products.

Consistency & Optimization

- Structural Zeroes: Enforced by edit and imputation, DP can't reintroduce it

- Invariants: public statistics with exact values

- Optimizing accuracy: for a set $\mathcal{Q}$ of queries

Census Bureau's Use of DP

Excerpts from:

[Michael Hawes and Michael Ratcliffe. "Understanding the 2020 Census Disclosure Avoidance System: Differential Privacy 201 and the TopDown Algorithm," Census Webinar, May 13, 2021.](#)
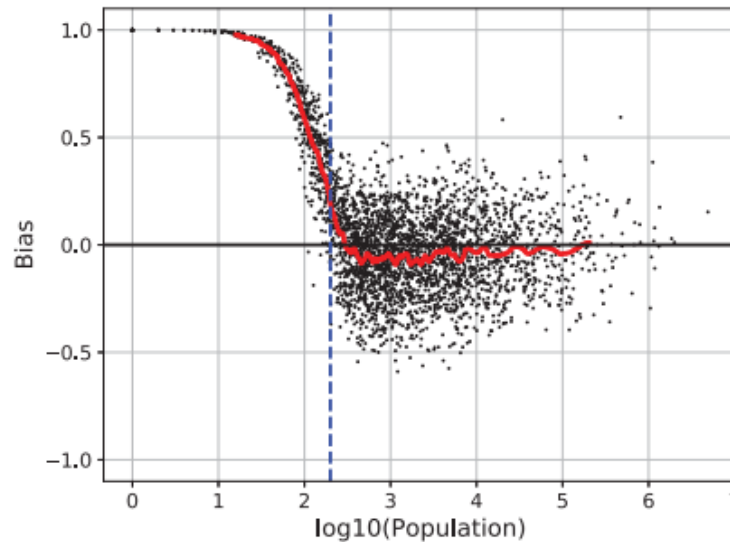
Bias

Figure 6-7: The calculated bias, $B = (\sigma_+^2 - \sigma_-^2)/(\sigma_+^2 + \sigma_-^2)$, calculated for every county, as a function of its Hispanic voting age population. The vertical dashed blue line indicates the value of σ for the noise distribution, which is $\sigma = 200$ for this graph. The solid red line is an average calculated over a window that of the nearest 100 points by population. Note that for all populations less than σ are biased towards positive values, resulting in larger post-processed values than the enumerated ones. Correspondingly, the non-negativity constraint combined with the aggregate constraint means that the populations larger than σ have a small negative bias.

Data Producer vs. Consumer

Data Producer's View
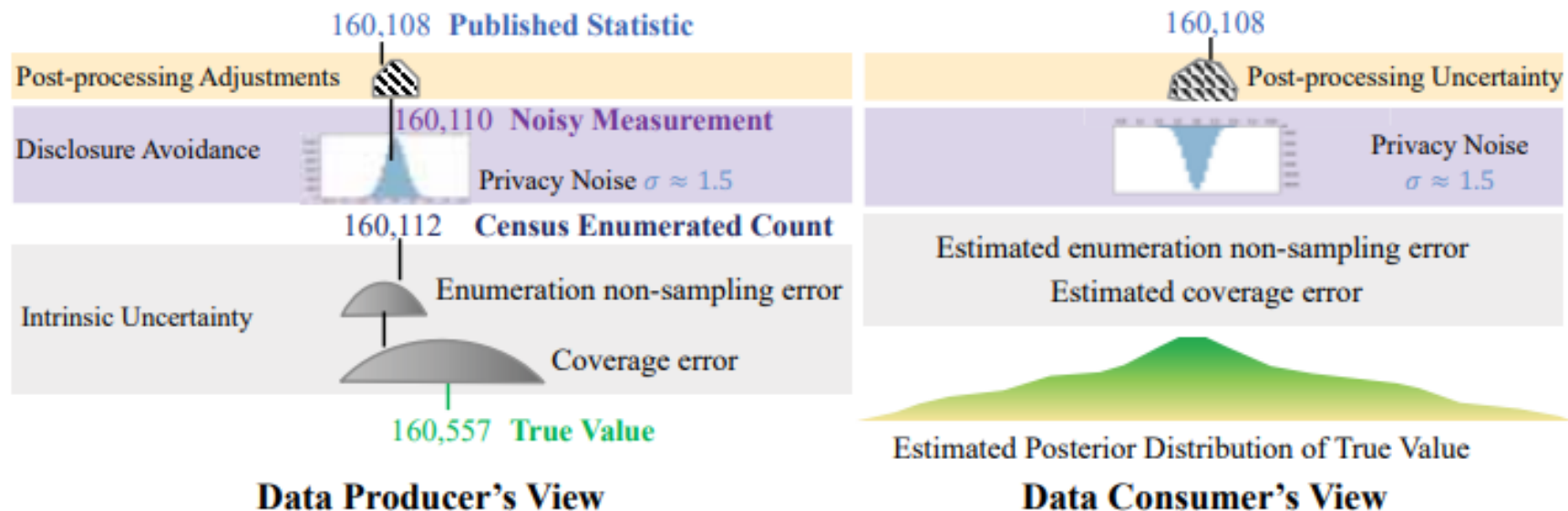
Data Consumer's View

Figure 6-1: Accuracy from the perspective of the data producer (Census Bureau) and data consumer. (The 160,112 census enumerated count is the population of Chattanooga City, Tennessee as enumerated by the 2010 census (Table 6-9), but is just used as an arbitrary example. All the other values and scales are for illustrative purposes only, and do not represent real data.)

Some challenges raised by JASON

- Detailed queries not used directly in published statistics but consuming privacy-loss budget

- Release of noisy measurements

- Consistency between and within data products

- Threats, risks, and protections not sufficiently quantified

- Better communication with data users

- Re-interpretation of Title 13

STATE OF ART SYNTHETIC DATA GENERATION

Private Multiplicative Weights

[Blum-Ligett-Roth `08,…,Hardt-Rothblum `10]
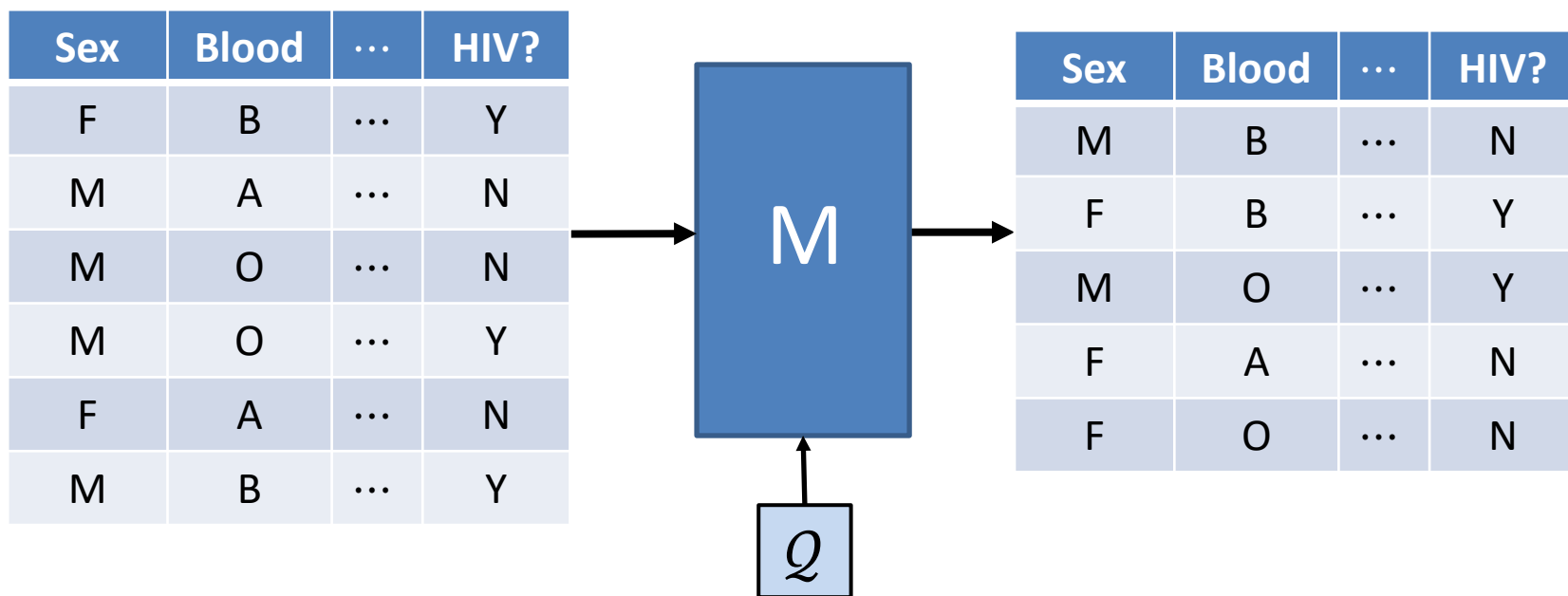
(ε, δ)-DP $M: \mathcal{X}^n \to \mathcal{X}^m$ such that $\forall q \in \mathcal{Q}, \ q: \mathcal{X} \to [0,1]$

$$\left| \frac{1}{n} \sum_{i=1}^{n} q(x_i) - \frac{1}{m} \sum_{i=1}^{m} q(M(x)_i) \right| \leq O \left(\frac{\sqrt{\log|\mathcal{X}| \cdot \log(1/\delta)} \cdot \log|\mathcal{Q}|}{\varepsilon n} \right)^{1/2}$$

Private Multiplicative Weights

[Hardt-Rothblum `10]

(ε, δ)-DP $M: \mathcal{X}^n \rightarrow \mathcal{X}^m$ such that $\forall q \in \mathcal{Q}, \; q: \mathcal{X} \rightarrow [0,1]$

$$\left| \frac{1}{n} \sum_{i=1}^n q(x_i) - \frac{1}{m} \sum_{i=1}^m q(M(x)_i) \right| \leq O \left(\frac{\sqrt{\log|\mathcal{X}| \cdot \log(1/\delta)} \cdot \log|\mathcal{Q}|}{\varepsilon n} \right)^{1/2}$$

Approach:

- DP online learning of a synthetic data distribution, playing against a "query" player trying to distinguish it from dataset

Problem: computation time $\text{poly}(n, |\mathcal{X}|, |\mathcal{Q}|)$.

- Exponential in dimensionality of data and query family.
- Inherent in the worst case (cf. "Complexity of DP").

Practical Approaches

- Use DP queries to learn a model of the data distribution, and use the model to generate synthetic data

- Models (from more structured to less):
  - Multivariate Gaussian ($\leftrightarrow$ means and (co)variances)
  - Graphical models/Markov Random Fields/Bayes Nets
  - Generative Adversarial Networks

Some Recent Developments

Table 2: Taxonomy of select-measure-generate mechanisms.

Name	Year	Workload Aware	Data Aware	Budget Aware	Efficiency Aware
Independent	-				✓
Gaussian	-	✓			
PrivBayes [54]	2014		✓	✓	✓
HDMM+PGM [40]	2019	✓			
PrivBayes+PGM [40]	2019		✓	✓	✓
MWEM+PGM [40]	2019	✓	✓		
PrivSyn [57]	2020		✓	✓	✓
MST [37]	2021		✓		✓
RAP [3]	2021	✓	✓		✓
GEM [33]	2021	✓	✓		✓
PrivMRF [7]	2021		✓	✓	✓
AIM [This Work]	2022	✓	✓	✓	✓

McKenna et al. "AIM: An Adaptive & Iterative Mechanism…" 2022

- Vietri et al. "Private Synthetic Data for Multitask Learning…" 2022.

- Liu et al. "Generating Private Synthetic Data with Genetic Algorithms" 2023.