Unleashing the Power of Randomization in Auditing Differentially Private ML

Krishna Pillutla, Galen Andrew, Peter Kairouz, H. Brendan McMahan, Alina Oprea, Sewoong Oh

Google Research · Northeastern University · University of Washington · Neural Information Processing Systems

Auditing DP: Standard Approach

Auditing: Empirically test whether the claimed DP guarantee is correct or tight

JE19, JUO20, NST+21, …

Step 1: DP definition

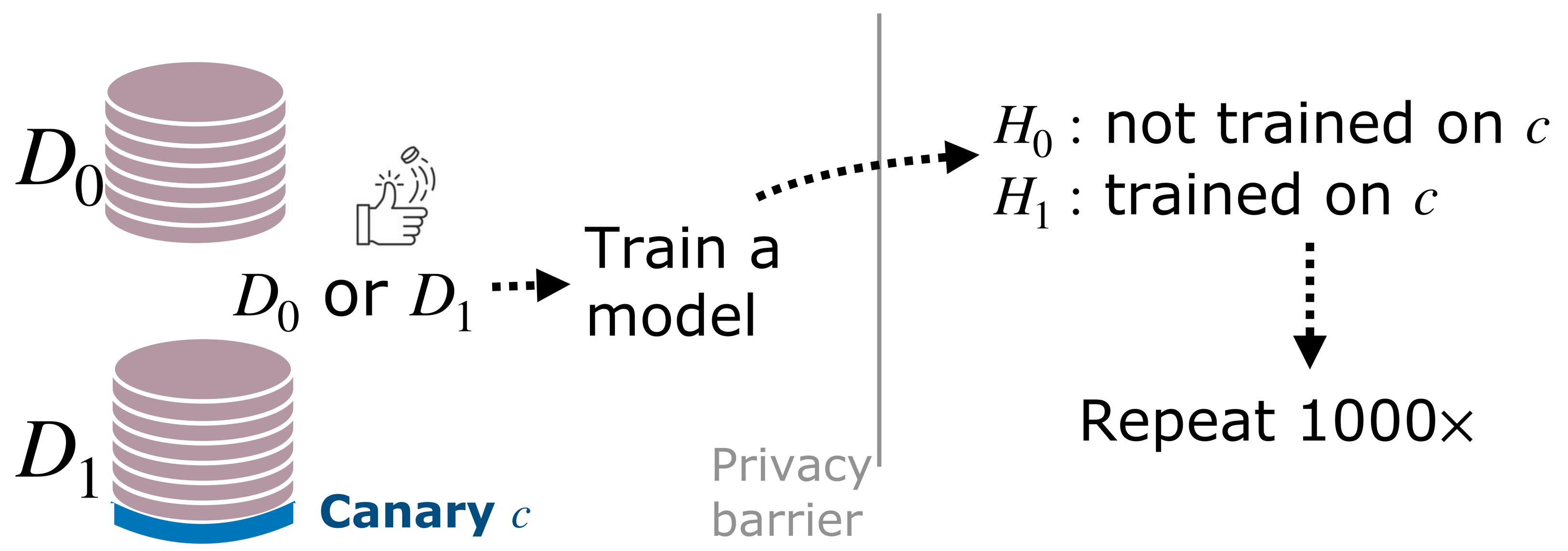
For all neighboring datasets D_0, D_1 and outcomes R:

$$\mathbb{P}(\mathscr{A}(D_1) \in R) \le e^{\varepsilon}\, \mathbb{P}(\mathscr{A}(D_0) \in R) + \delta \qquad \textbf{(1)}$$

True positive rate *False positive rate*

Step 2: Binary hypothesis tests

Take D_0 = dataset, $D_1 = D_0 \cup \{\text{canary}\}$ and the test statistic as $R = \{\theta : \text{Loss}(\text{canary}\,;\theta) \le \tau\}$



Step 3: Bernoulli confidence intervals

Run n trials (each trial = one model training run)

$$\text{TPR} \approx \frac{1}{n}\sum_{i=1}^{n} \text{Loss}_i(D_1) \le \tau) \pm \sqrt{\frac{\text{variance}}{n}} \qquad \textbf{(2)}$$

True rate *Empirical rate*

Overall, (1) + (2) $\Longrightarrow$

$$\varepsilon \ge \log\left(\frac{\text{TPR} - \delta}{\text{FPR}}\right) \ge \log\left(\frac{\widehat{\text{TPR}}_n - \frac{1}{\sqrt{n}} - \delta}{\widehat{\text{FPR}}_n + \frac{1}{\sqrt{n}}}\right)$$

Problem: the $1/\sqrt{n}$ term requires n large
How do we solve this? Add multiple canaries
Key: Avoid group privacy with *randomization*

Auditing Lifted DP

Step 1: Lifted DP (LiDP) definition

Def: $\mathscr{A}$ is (ε, δ)-LiDP if for all random $(\boldsymbol{D_0}, \boldsymbol{D_1}, \boldsymbol{R}) \sim \mathscr{P}$ independent of $\mathscr{A}$ s.t. $\boldsymbol{D}_0, \boldsymbol{D}_1$ are neighboring, we have

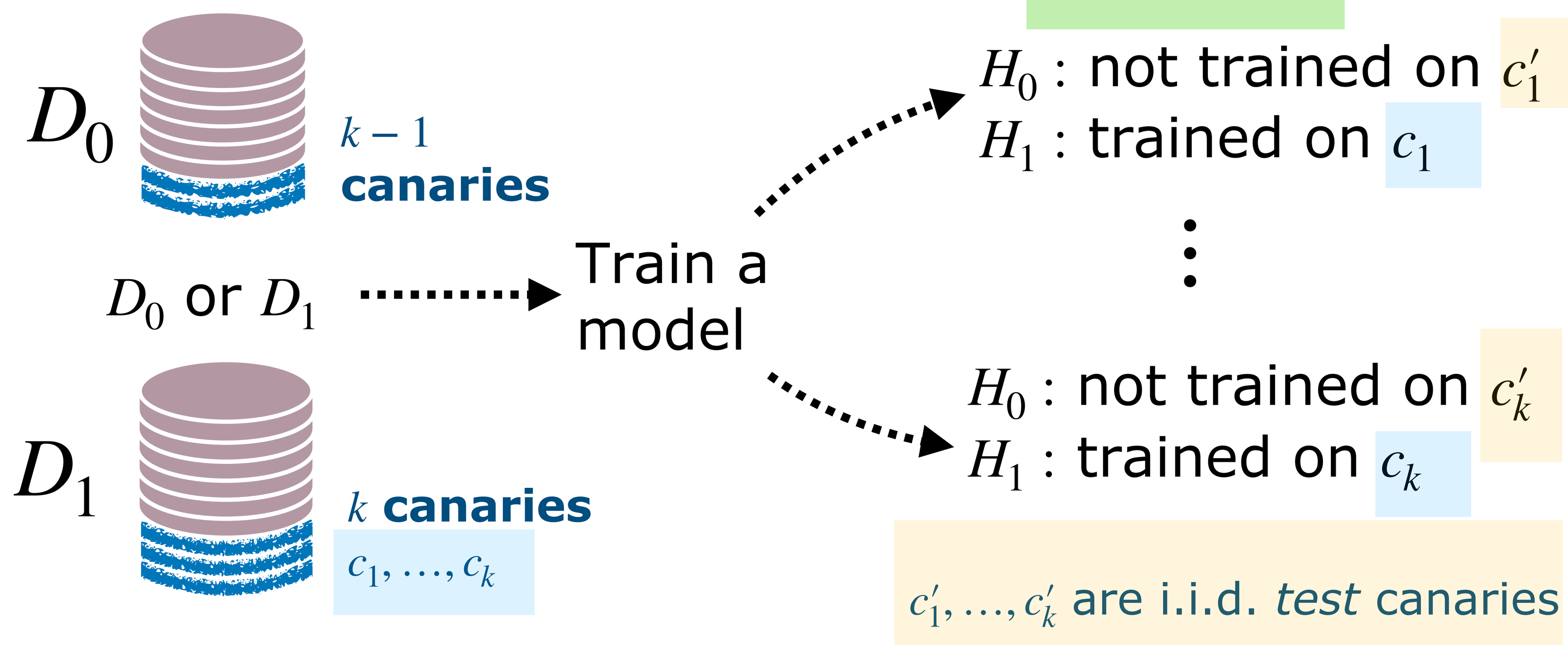
$$\mathbb{P}(\mathscr{A}(\boldsymbol{D}_1) \in \boldsymbol{R}) \le e^{\varepsilon}\, \mathbb{P}(\mathscr{A}(\boldsymbol{D}_0) \in \boldsymbol{R}) + \delta \qquad \textbf{(3)}$$

Theorem: $\mathscr{A}$ is (ε,δ)-DP $\Longleftrightarrow$ $\mathscr{A}$ is (ε,δ)-LiDP

Consequence: We can have *random* canaries!

Step 2: Randomized hypothesis tests

Test for k vs. $k-1$ canaries that are drawn i.i.d. from P



Consequence: Get k statistics from each trial

Step 3: Adaptive higher-order confidence

Challenge: the statistics are *correlated* (not i.i.d.)

We derive novel CIs using *empirical* correlations!

$$\left|\text{TPR} - \widehat{\text{TPR}}_{n,k}\right| \le \sqrt{\frac{1}{n}\left(\text{corr.} + \frac{1}{k} + \sqrt{\frac{\text{4th moment}}{n}}\right)} \qquad \textbf{(4)}$$

If $\text{corr.} = O(1/k)$, **improvement**: (3) + (4) $\Longrightarrow$

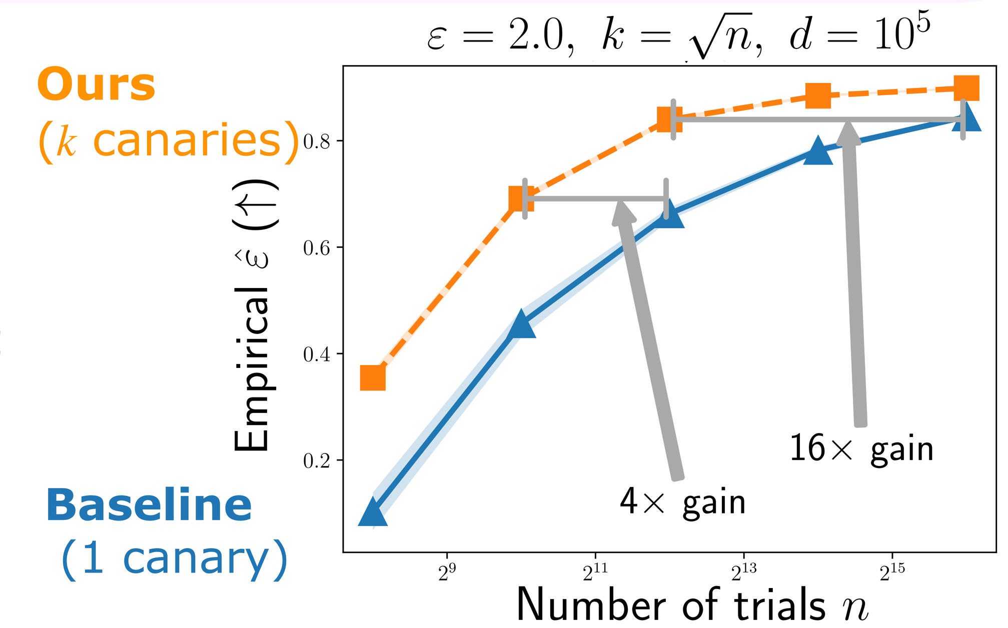
$$\varepsilon \ge \log\left(\frac{\text{TPR} - \delta}{\text{FPR}}\right) \ge \log\left(\frac{\widehat{\text{TPR}}_{n,k} - \frac{1}{\sqrt{nk}} - \frac{M_4}{n^{3/4}} - \delta}{\widehat{\text{FPR}}_{n,k} + \frac{1}{\sqrt{nk}} + \frac{M_4}{n^{3/4}}}\right)$$

Experiments

Auditing a Gaussian mechanism

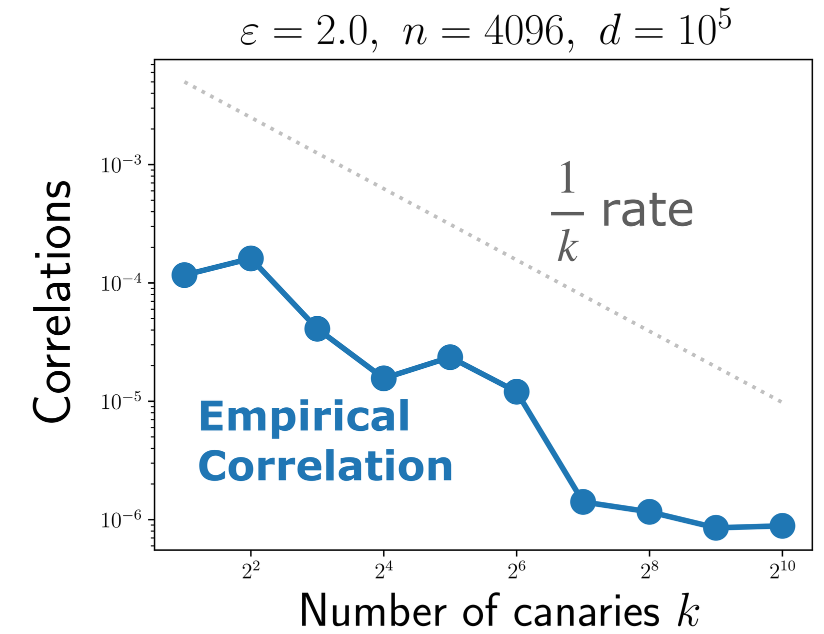
Setup:
- Sum query
- Canaries: uniform over unit sphere
- Test: inner product

Result: 4-16 × gain in sample complexity

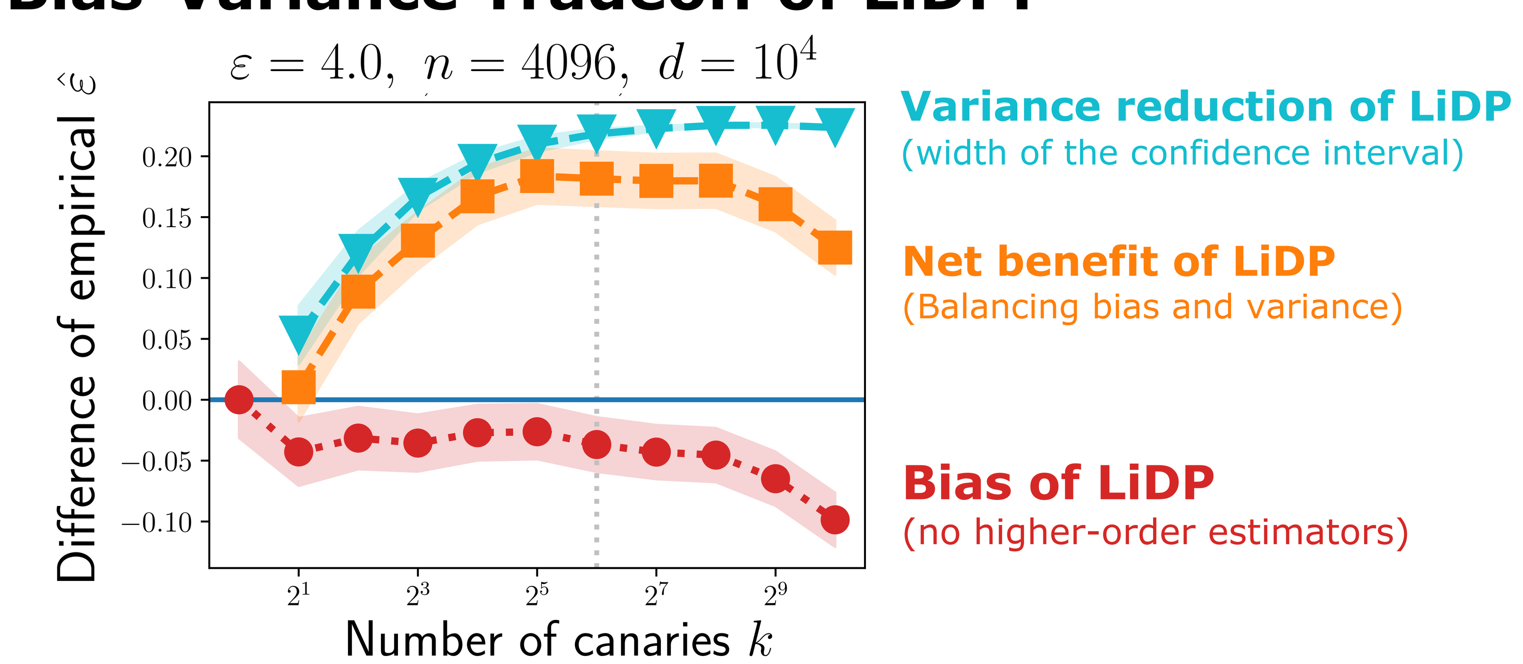


Analysis:
Empirical canary correlations are small, so LiDP auditing gives large wins.

Practical Guidance:
Multiple canaries should be "orthogonal"

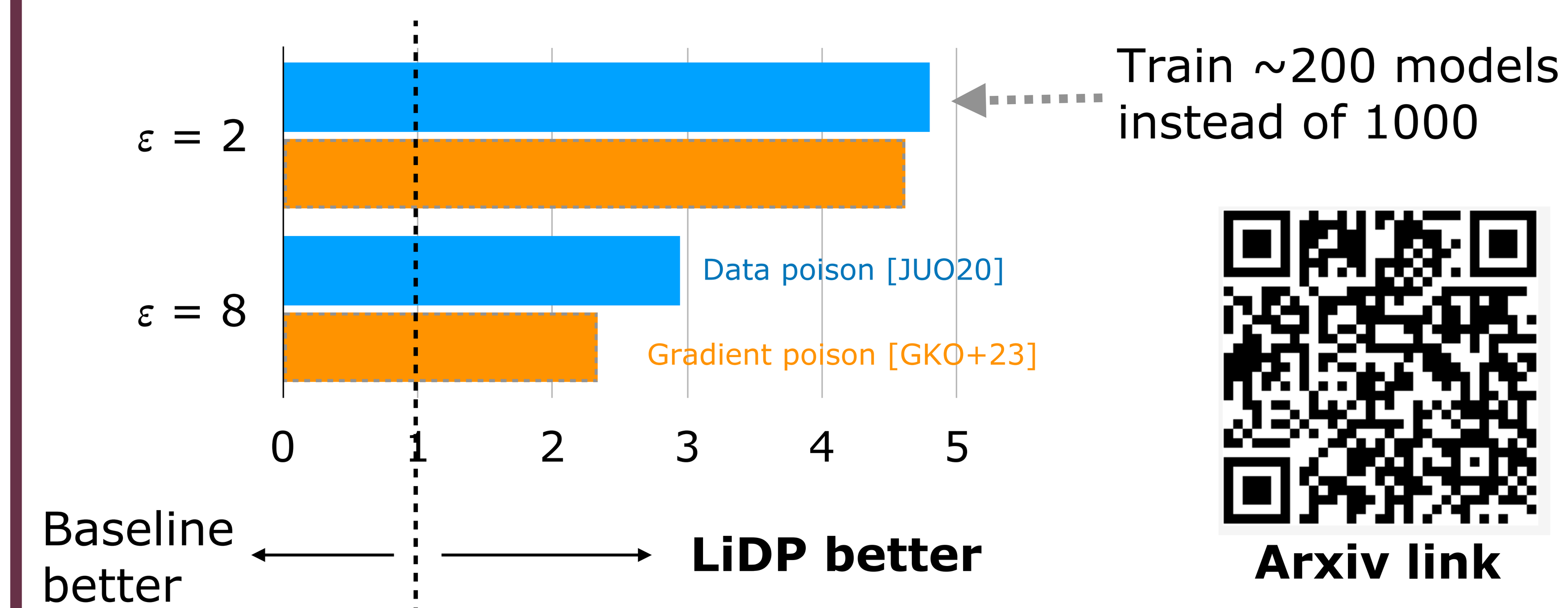


Bias-Variance Tradeoff of LiDP:



Experiments: FashionMNIST + MLP model

Gain in sample complexity from LiDP auditing



Arxiv link