Unleashing the Power of Randomization in Auditing Differentially Private ML

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

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 \( R \):

\[
\mathbb{P}(\mathcal{A}(D_1) \in R) \leq e^\varepsilon \mathbb{P}(\mathcal{A}(D_0) \in R) + \delta 
\]

\( (1) \)

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 \( R = \{ \theta : \text{Loss(canary}; \theta) \leq \varepsilon \} \)

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

**Step 3: Bernoulli confidence intervals**

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

\[
\text{TPR} \approx \frac{1}{n} \sum_{i=1}^{n} \text{Loss}(D_i) \leq \varepsilon \Rightarrow \mathbb{V}_{\text{variance}}(\text{TPR}) \approx \frac{\text{variance}}{n} \leq \frac{\varepsilon}{\sqrt{n}} \]

\( (2) \)

**Experiments**

**Auditing a Gaussian mechanism**

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

**Result**:
- 4-16 \( \times \) 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**:

Variance reduction of LiDP (width of the confidence interval)
Net benefit of LiDP (balancing bias and variance)
Bias of LiDP (no higher-order estimators)

**Experiments**: FashionMNIST + MLP model

Gain in sample complexity from LiDP auditing

Training ~200 models instead of 1000

<table>
<thead>
<tr>
<th>( \varepsilon )</th>
<th>Data poison (JUO20)</th>
<th>Gradient poison (SKG+23)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>