Towards User-level Differential Privacy at Scale

Krishna Pillutla
Google Research -> IIT Madras
Long live the revolution. Our next meeting will be at the docks at midnight on June 28.

Aha, found them!

When you train predictive models on input from your users, it can leak information in unexpected ways.
Models leak information about their training data.

Carlini et al. (USENIX Security 2021)
Models leak information about their training data reliably.

Carlini et al. (ICLR 2023)

Carlini et al. (USENIX Security 2021)
Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, Tom Goldstein

University of Maryland, College Park
{gowthami, vsingla, jgeiping, tomg}@cs.umd.edu
New York University
goldblum@nyu.edu

Generation

LAION-A Match
Differential privacy (DP)

Dwork, McSherry, Nissim, Smith. *Calibrating noise to sensitivity in private data analysis*. TCC 2006
Differential privacy (DP)

A randomized algorithm is $\epsilon$-differentially private if the addition of one unit of data does not alter its output distribution by more than $\epsilon$.
Example-level Differential privacy (DP)

A randomized algorithm is $\varepsilon$-differentially private if the addition of one example does not alter its output distribution by more than $\varepsilon$.
Differential privacy eliminates memorization

Which data do we use to train/finetune/align these models?

Test on **shifted distribution** (out-of-domain / OOD)

Test on **training distribution** (in-domain / ID)
Which data do we use to train/finetune/align these models?

Best training data = in-domain data
For many applications, in-domain data = **user data**
For many applications, in-domain data = user data

Each user can contribute multiple examples
A randomized algorithm is \( \varepsilon \)-differentially private if the addition of one \textit{example} does not alter its output distribution by more than \( \varepsilon \)
Example-level Differential privacy (DP)

A randomized algorithm is \( \epsilon \)-differentially private if the addition of one user’s data does not alter its output distribution by more than \( \epsilon \).
Why do we need user-level DP?
Why do we need user-level DP?

*Standard LLM finetuning pipelines are susceptible to user inference attacks!*

Nikhil Kandpal, KP, Alina Oprea, Peter Kairouz, Chris Choquette-Choo, Zheng Xu. Submitted (2024)
User Inference Attack

Attacker Has:

and

fresh i.i.d. samples from a user distribution

Model finetuned on user data

Attacker Wants to Infer:

Did samples come from one of

?
User inference is effective when the number of users is small and data per user is large.

More fine-tuning samples per user

More users
Short common phrases can exacerbate user inference
Example-level DP offers limited mitigation

AUROC:
- non-private: 88%
- $\varepsilon = 32$: 70%

Utility:
- DP model reaches what the private model achieves in 1/3 epoch

Example-level DP does not help here
A randomized algorithm is \( \varepsilon \)-differentially private if the addition of one user’s data does not alter its output distribution by more than \( \varepsilon \).
How do we realize user-level DP?
Outline: how do we realize user-level DP?

Learning algorithms:

(Anti-) correlated noise provably beats independent noise

For linear regression, dimension $d$ improves to problem-dependent effective dimension $d_{\text{eff}}$

<table>
<thead>
<tr>
<th>Independent noise</th>
<th>$\Theta(d)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlated noise</td>
<td>$\tilde{O}(d_{\text{eff}})$</td>
</tr>
<tr>
<td>Lower bound</td>
<td>$\Omega(d_{\text{eff}})$</td>
</tr>
</tbody>
</table>
Outline: how do we realize user-level DP?

Learning algorithms:

(Anti-) correlated noise provably beats independent noise

For linear regression, dimension $d$ improves to problem-dependent effective dimension $d_{eff}$

<table>
<thead>
<tr>
<th>Independent noise</th>
<th>$\Theta(d)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlated noise</td>
<td>$\tilde{O}(d_{eff})$</td>
</tr>
<tr>
<td>Lower bound</td>
<td>$\Omega(d_{eff})$</td>
</tr>
</tbody>
</table>

Auditing:

Randomness makes the audit more computationally efficient

Empirical lower bound on $\varepsilon$
Part 1: How do we learn with user-level DP?

(Anti-)correlated noise provably beats independent noise

ICLR 2024

Chris Choquette-Choo*
Dj Dvijotham*
Krishna Pillutla*
Arun Ganesh
Thomas Steinke
Abhradeep Thakurta

*Equal contribution, $\alpha\beta$-order
**DP-SGD**: How do we train models with example-level DP?

Let's denote the learning rate as \( \eta \). The update rule for \( \theta_{t+1} \) is given by:

\[
\theta_{t+1} = \theta_t - \eta \left( g_t + z_t \right)
\]

- Stochastic gradient clipped to \( \|g\| \leq G \) per-example
- Independent Gaussian noise

References:
- Song et al. (2013)
- Bassily et al. (FOCS 2014)
- Abadi et al. (CCS 2016)
**DP-FedAvg**: How do we train models with user-level DP?

\[ \theta_{t+1} = \theta_t - \eta \left( g_t + z_t \right) \]

- Stochastic gradient clipped to \( \|g\| \leq G \) per-user
- \textit{Independent} Gaussian noise

**DP-SGD**: DP Training with *Independent* Noise

For $q$-zCDP, take noise variance $= \frac{G^2}{2\rho}$

($G = \text{gradient clip norm}$)

$$\theta_{t+1} = \theta_t - \eta \left( g_t + z_t \right)$$
**DP-FTRL**: DP Training with **Correlated** Noise

\[ \theta_{t+1} = \theta_t - \eta \left( g_t + \sum_{\tau=0}^{t} \beta_{t,\tau} z_{t-\tau} \right) \]

**Correlated** Gaussian noise \((z_t \text{ i.i.d. Gaussian})\)

Kairouz, McMahan, Song, Thakkar, Thakurta, Xu. **Practical and Private (Deep) Learning without Sampling or Shuffling**. ICML 2021.
**DP-FTRL**: DP Training with *Correlated* Noise

For $\varphi$-zCDP, take noise variance as:

$$\text{For } \varphi\text{-zCDP, take noise variance } = \frac{G^2}{2\rho} \max_{t} \| [B^{-1}]; t \|_2^2$$

$$B = \begin{pmatrix}
\beta_{0,0} & 0 & 0 & \ldots \\
\beta_{1,0} & \beta_{1,1} & 0 & \ldots \\
\beta_{2,0} & \beta_{2,1} & \beta_{2,2} & \ldots \\
\vdots & & & \\
\end{pmatrix}$$

**Correlated** Gaussian noise ($z_t$ i.i.d. Gaussian)

$$\theta_{t+1} = \theta_t - \eta \left( g_t + \sum_{\tau=0}^{t} \beta_{t,\tau} z_{t-\tau} \right)$$


“the first production neural network trained directly on user data announced with a formal DP guarantee.”

- Google AI Blog post, Feb 2022
Do we use *independent* or *correlated* noise?

**DP-SGD**  
**DP-FTRL**
**Prior work:** (Empirically) correlated noise outperforms independent noise

**Experiment:**
User-level DP with StackOverflow

**Graph:**
- DP-FTRL (+ amplification) uniformly beats DP-SGD

---

Choquette-Choo, Ganesh, McKenna, McMahan, Rush, Thakurta, Xu.
(Amplified) Banded Matrix Factorization: A unified approach to private training. NeurIPS 2023
Our goal: a *provable* gap between DP-SGD & DP-FTRL
## DP-FTRL vs. DP-SGD: Previous Theory

For convex & $G$-Lipschitz losses

<table>
<thead>
<tr>
<th>Method</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>DP-SGD</td>
<td>$\frac{Gd^{1/4}}{\sqrt{\rho T}}$</td>
</tr>
<tr>
<td>DP-FTRL</td>
<td>$\frac{Gd^{1/4}}{\sqrt{\rho^2 T}}$</td>
</tr>
</tbody>
</table>

$\rho$: privacy level (zCDP)  
$d$: dimension  
$T$: #iterations

Kairouz, McMahan, Song, Thakkar, Thakurta, Xu.  
Practical and Private (Deep) Learning without Sampling or Shuffling. ICML 2021.
Streaming setting: Suppose we draw a fresh data point $x_t \sim P$ in each iteration $t$ (i.e. only 1 epoch)
Toeplitz noise correlations: $\beta_{t,\tau} = \beta_\tau$

\[
\theta_{t+1} = \theta_t - \eta \left( g_t + \sum_{\tau=0}^{t} \beta_\tau z_{t-\tau} \right)
\]

\[
B = \begin{pmatrix}
\beta_{0,0} \\
\beta_{0,1} & \beta_{1,0} \\
\beta_{0,2} & \beta_{1,1} & \beta_{2,0} \\
\vdots & \vdots & \vdots & \ddots
\end{pmatrix}
\rightarrow
B = \begin{pmatrix}
\beta_0 \\
\beta_1 & \beta_0 \\
\beta_2 & \beta_1 & \beta_0 \\
\vdots & \vdots & \vdots & \ddots
\end{pmatrix}
\]

Computationally: store $O(T)$ coefficients instead of $O(T^2)$
**Asymptotics:** Iterates converge to a stationary distribution as $t \to \infty$
Asymptotics: Iterates converge to a stationary distribution as $t \to \infty$

Asymptotic error

$$F_\infty(\beta) = \lim_{t \to \infty} E\left[ F(\theta_t) - F(\theta_\star) \right]$$

Asymptotics at a fixed learning rate $\eta > 0$
Noisy-SGD/Noisy-FTRL: DP-SGD/DP-FTRL without clipping

\[ \| \text{clip}(g, G) \| \]

\[ G \]

\[ 0 \]

\[ \| g \| \]

Let us study the noise dynamics of the algorithms (do not satisfy DP guarantees)
Mean estimation in 1 dimension

\[
\min_{\theta} \left[ F(\theta) = \mathbb{E}_{x \sim P} (\theta - x)^2 \right]
\]

Data distribution
s.t. \(|x| \leq 1

Solve with stochastic optimization problem
with DP-SGD/DP-FTRL
Mean estimation in 1 dimension

**Informal Theorem:** The asymptotic error of a $\varphi$-zCDP sequence is

<table>
<thead>
<tr>
<th>Independent noise (DP-SGD)</th>
<th>$F_\infty(\beta^{\text{sgd}}) = \rho^{-1}\eta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlated noise (DP-FTRL)</td>
<td>$\inf_{\beta} F_\infty(\beta) = F_\infty(\beta^*) = \rho^{-1}\eta^2 \log^2 \frac{1}{\eta}$</td>
</tr>
</tbody>
</table>

$\eta$: learning rate (constant and non-zero)

$\varphi$: privacy level
DP-FTRL is always better than DP-SGD. DP-FTRL is significantly better at $\eta \to 0$ or $\eta \to 1$. The graph shows the suboptimality ratio for mean estimation, with $y \approx 0.54$. The y-axis represents the ratio of DP-FTRL to DP-SGD, and the x-axis represents the learning rate $\eta$. As $\eta$ approaches 0 or 1, the ratio increases significantly.
Proposition: The correlations \( \beta_0^* = 1, \quad \beta_t^* = -t^{-3/2}(1 - \eta)^t \)
attain the optimal error

\[
\inf_{\beta} F_\infty(\beta) = F_\infty(\beta^*) = \rho^{-1} \eta^2 \log^2 \frac{1}{\eta}
\]
Proposition: The correlations

\[ \beta_0^* = 1, \quad \beta_t^* = -t^{-3/2}(1 - \eta)^t \]

attain the optimal error

\[ \inf_{\beta} F_\infty(\beta) = F_\infty(\beta^*) = \rho^{-1} \eta^2 \log^2 \frac{1}{\eta} \]

\( \nu\text{-DP-FTRL} \)

For general problems, use

\[ \beta_0 = 1, \quad \beta_t = -t^{-3/2}(1 - \nu)^t \]

and tune the parameter \( \nu \)
Linear regression

$$\min_\theta \left[ F(\theta) = \mathbb{E}(y - \langle \theta, x \rangle)^2 \right]$$

where \( x \sim \mathcal{N}(0, H) \)

\( H \) is also the Hessian of the objective
Linear regression

\[
\min_\theta \left[ F(\theta) = \mathbb{E} (y - \langle \theta, x \rangle)^2 \right]
\]

where \( x \sim \mathcal{N}(0, H) \)

\[
y | x \sim \mathcal{N}(x^\top \theta_*, \sigma^2)
\]
**Informal Theorem**: The asymptotic error is

<table>
<thead>
<tr>
<th>Noise Type</th>
<th>Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Independent noise (Noisy-SGD)</td>
<td>$d \rho^{-1} \eta$</td>
</tr>
<tr>
<td>Correlated noise ($\nu$-Noisy-FTRL)</td>
<td>$d_{\text{eff}} \rho^{-1} \eta^2 \log^2 \left( \frac{1}{\eta \mu} \right)$</td>
</tr>
<tr>
<td>Lower bound for any algorithm</td>
<td>$d_{\text{eff}} \rho^{-1} \eta^2$</td>
</tr>
</tbody>
</table>

**Improve dimension $d$ to problem-dependent effective dimension $d_{\text{eff}}$.**
Effective dimension

\[ d_{\text{eff}} = \frac{\text{Tr}(H)}{\|H\|_2} \leq d \]

**Low effective dimension**
\[ \lambda_1 = 1, \lambda_2 = \cdots = \lambda_d = 1/d \]

**High effective dimension**
\[ \lambda_1 = \lambda_2 = \cdots = \lambda_d = 1 \]

Closely connected to **numerical/stable rank**
Remark 1.3 (Numerical rank). The numerical rank $r = r(A) = \|A\|_F^2 / \|A\|_2^2$ in Theorem 1.1 is a relaxation of the exact notion of rank. Indeed, one always has $r(A) \leq \text{rank}(A)$. But as opposed to the exact rank, the numerical rank is stable under small perturbations of the matrix $A$. In particular, the numerical rank of $A$ tends to be low when $A$ is close to a low rank matrix, or when $A$ is sufficiently sparse.

$$d_{\text{eff}} = \text{srank}(H^{1/2})$$

[Rudelson & Vershynin (J. ACM 2007)]
The stable rank appears in:

- Numerical linear algebra (e.g. randomized matrix multiplications) [Tropp (2014), Cohen-Nelson-Woodruff (2015)]
- Matrix concentration [Hsu-Kakade-Zhang (2012), Minsker (2017)]
- …
**Informal Theorem:** The asymptotic error is

| Independent noise (Noisy-SGD) | $= d \rho^{-1} \eta$ |
| Correlated noise ($\nu$-Noisy-FTRL) | $\leq d_{\text{eff}} \rho^{-1} \eta^2 \log^2 \left( \frac{1}{\eta \mu} \right)$ |
| Lower bound for any algorithm | $\geq d_{\text{eff}} \rho^{-1} \eta^2$ |

**Improve dimension $d$ to problem-dependent effective dimension $d_{\text{eff}}$.**
Linear regression: theory predicts simulations

**Noisy-SGD** scales with $d$

**Noisy-FTRL** scales with $d_{\text{eff}}$
**Informal Theorem**: The asymptotic error for $0 < \eta < 1$ is

<|table|>
| **Independent noise** (Noisy-SGD) | $= d \rho^{-1} \eta$ |
| **Correlated noise** ($\nu$-Noisy-FTRL) | $\leq d_{\text{eff}} \rho^{-1} \eta^2 \log^2 \left( \frac{1}{\eta \mu} \right)$ |
| **Lower bound** for any algorithm | $\geq d_{\text{eff}} \rho^{-1} \eta^2$ |

**Improved dependence on the learning rate $\eta$**
Noisy-SGD scales as $\eta$

$\nu$-Noisy-FTRL scales as $\eta^2$

Noisy-FTRL $\gg$ Noisy-SGD at small $\eta$
**Finite-time rates with DP: Linear Regression**

<table>
<thead>
<tr>
<th>Noise Type</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Independent noise</td>
<td>$\frac{1}{\rho T} + \frac{1}{T}$</td>
</tr>
<tr>
<td>Correlated noise</td>
<td>$\frac{1}{\rho T^2} + \frac{1}{T}$</td>
</tr>
</tbody>
</table>

Privacy error

$T$: number of iterations

$\varphi$: privacy level

$\eta$: learning rate is optimized
Proof sketch for Mean Estimation

Updates are not Markovian (key for all stochastic gradient proofs)

**Our approach**: Analysis the Fourier domain
Letting $\delta_t = \theta_t - \theta_*$, the DP-FTRL update can be written as

$$\delta_{t+1} = (1 - \eta)\delta_t - \eta \sum_{\tau=0}^{t} \beta_{\tau} z_{t-\tau}$$

Linear Time-Invariant (LTI) system

Convolution of the noise
Fourier analysis can give the stationary variance of $\delta_t$ in terms of the **discrete-time Fourier transform** of the convolution weights $\beta$

$$B(\omega) = \sum_{t=0}^{\infty} \beta_t e^{i\omega t}$$

Frequency-domain description

Time-domain description

Image: 3blue1brown.com/lessons/fourier-transforms
Letting $\delta_t = \theta_t - \theta^*$, the DP-FTRL update can be written as

$$
\delta_{t+1} = (1 - \eta) \delta_t - \eta \sum_{\tau=0}^{t} \beta_\tau z_{t-\tau}
$$

The stationary variance of $\delta_t$ can be given as

$$
\lim_{t \to \infty} \mathbb{E}[\delta_t^2] = \frac{\eta^2}{2\pi} \left( \int_{-\pi}^{\pi} \frac{|B(\omega)|^2}{\left| 1 - \eta - e^{i\omega} \right|^2} \, d\omega \right) \mathbb{E}[z_t^2]
$$
\[
\lim_{t \to \infty} \mathbb{E}[\delta_t^2] = \frac{\eta^2}{2\pi} \left( \int_{-\pi}^{\pi} \frac{|B(\omega)|^2}{|1 - \eta - e^{i\omega}|^2} \, d\omega \right) \mathbb{E}[z_t^2]
\]

For \( \varphi \)-zCDP, take
\[
\mathbb{E}[z_t^2] = \frac{1}{2\rho} \max_t \| [B^{-1}]_t \|^2_2
\]
\[
= \frac{1}{2\rho} \int_{-\pi}^{\pi} \frac{d\omega}{2\pi |B(\omega)|^2}
\]

sensitivity

\[
B = \begin{pmatrix}
\beta_0 \\
\beta_1 \\
\beta_0 \\
\beta_2 \\
\beta_1 \\
\beta_0 \\
\cdots
\end{pmatrix}
\]
For $\rho$-zCDP, take

\[
\lim_{t \to \infty} \mathbb{E}[\delta_t^2] = \frac{\eta^2}{2\pi} \left( \int_{-\pi}^{\pi} \frac{|B(\omega)|^2}{|1 - \eta - e^{i\omega}|^2} \, d\omega \right) \mathbb{E}[z_t^2]
\]

**sensitivity**

Requires $|B(\omega)|$ small

For $\rho$-zCDP, take

\[
\mathbb{E}[z_t^2] = \frac{1}{2\rho} \max_t \| [B^{-1}]_{:,t} \|_2^2
\]

\[
= \frac{1}{2\rho} \int_{-\pi}^{\pi} \frac{d\omega}{2\pi |B(\omega)|^2}
\]

Requires $|B(\omega)|$ large

\[
B = \begin{pmatrix}
\beta_0 \\
\beta_1 \\
\beta_0 \\
\beta_2 \\
\beta_1 \\
\beta_0 \\
\vdots
\end{pmatrix}
\]
For $\varrho$-zCDP, take

$$\lim_{t \to \infty} \mathbb{E}[\delta_t^2] = \frac{\eta^2}{2\pi} \left( \int_{-\pi}^{\pi} \frac{|B(\omega)|^2}{|1 - \eta - e^{i\omega}|^2} \, d\omega \right) \mathbb{E}[z_t^2]$$

sensitivity

Requires $|B(\omega)|$ small

For $\varrho$-zCDP, take

$$\mathbb{E}[z_t^2] = \frac{1}{2\rho} \max_t \| [B^{-1}]_{:,t} \|_2^2$$

$$= \frac{1}{2\rho} \int_{-\pi}^{\pi} \frac{d\omega}{2\pi|B(\omega)|^2}$$

Requires $|B(\omega)|$ large

Optimizing for $|B(\omega)|$ gives the theorem
Language modeling with Stack Overflow | User-level DP

$\epsilon = \infty$, Nonprivate Baseline

Test Accuracy vs Privacy Budget, $\epsilon$

- Ours matches SoTA!

Mechanism:
- $(\nu = 1.6 \times 10^{-3})$-DP-FTRL (Ours)
- Online Honaker $\times 10$
- DP-SGD + Amplification
- ME($k=20$) Non-Toeplitz

Google Research
Image classification with CIFAR-10 | Example-level DP

SoTA (requires $O(T^3)$ for the SDP)

$\epsilon = \infty$, Nonprivate Baseline

Ours (closed form)

Baselines

Test Accuracy

Privacy Budget, $\epsilon$

Mechanism:
- $(\nu = 4 \times 10^{-3})$-DP-FTRL (Ours)
- Optimal CC $\times 4$
- Online Honaker $\times 10$
- DP-SGD + Amplification
- UpperBound
- ME($k=20$) Non-Toeplitz
Image classification with CIFAR-10 | Example-level DP

SoTA (requires $O(T^3)$ for the SDP)

Ours (closed form)

Baselines

Test Accuracy

Privacy Budget, $\epsilon$

Mechanism

- $(\nu = 4 \times 10^{-3})$-DP-FTRL (Ours)
- Optimal CC $\times 4$
- Online Honaker $\times 10$
- DP-SGD + Amplification
- UpperBound
- ME($k=20$) Non-Toeplitz

Google Research
Computational cost

- **SoTA**: cubic complexity to generate the $\beta$'s
- **Ours**: linear complexity (closed form)
  - nearly matches SoTA empirically
Summary

- Correlated noise is provably better
- Depends on effective dimension instead of dimension
- Matches lower bounds
Part 2: How audit user-level DP?

Unleashing the power of randomness in auditing DP

NeurIPS 2023

Krishna Pillutla  Galen Andrew  Peter Kairouz  Brendan McMahan  Alina Oprea  Sewoong Oh
Empirical privacy auditing

Provable analytic DP $\varepsilon$ (often loose)

Real privacy leakage

$\varepsilon$ empirical lower bound

Our focus
Why empirical privacy auditing?

To verify that we actually provide the guarantee we claim (no bugs in proofs/implementation)

Tramèr et al. *Debugging Differential Privacy: A Case Study for Privacy Auditing*. Preprint 2022
Gap between DP guarantees and empirical behavior: Memorization

Empirical Privacy Auditing requires many samples

- Trained w/ \((0.21, 10^{-5})\)-DP but empirically \(\varepsilon > 2.79\) with confidence \(1 - 10^{-8}\) ⇒ bug in implementation

- This required training \(n = 200,000\) models

Tramèr et al. *Debugging Differential Privacy: A Case Study for Privacy Auditing*. Preprint 2022
Our goal: make empirical privacy auditing more *sample-efficient*
Standard approaches for auditing privacy: **binary hypothesis testing**

\[ D_0 \]

\[ D_1 \]

Training data

Privacy barrier

\[ D_0 \text{ or } D_1 \]

Train model with a mechanism in question

---


Jagielski, Ullman, Oprea. **Auditing differentially private machine learning: How private is private SGD?** NeurIPS 2020
Standard approaches for auditing privacy: **binary hypothesis testing**

Training data

Train model with a mechanism in question

Privacy barrier

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

True Positive Rate

False Positive Rate

Repeat many times and measure privacy leakage

\( D_0 \) (Null Hypothesis) or \( D_1 \) (Alternative Hypothesis)?

---


Jagielski, Ullman, Oprea. **Auditing differentially private machine learning: How private is private SGD?** NeurIPS 2020
Bottleneck: Bernoulli confidence intervals

- Confidence intervals based on $n$ trials

$$TPR \approx \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(\text{Guess } i \text{ correct}) + \sqrt{\frac{\text{Variance}}{n}}$$

Actual TPR/FPR

Empirical TPR/FPR

Sample size $n$ needs to be large for good estimates

Actual TPR/FPR

$\varepsilon \geq \log \left( \frac{TPR - \delta}{FPR} \right)$

$\geq \log \left( \frac{T\hat{PR}_n - \frac{c}{\sqrt{n}} - \delta}{F\hat{PR}_n + \frac{c}{\sqrt{n}}} \right)$
Our approach: leverage randomness

- **Lifted DP**: Equivalent notion of DP with randomized datasets
- Multiple randomized hypothesis tests
- **Adaptive confidence intervals** capitalizing on low correlations
Multiple hypothesis tests for auditing Lifted DP

- **Leave-One Out** construction with i.i.d. random canaries

Random $D_1$ or $D_0$ Train Model

Is $c_1$ in $D_1$? Is $c'_1$ in $D_0$?

Is $c_2$ in $D_1$? Is $c'_2$ in $D_0$?

Is $c_3$ in $D_1$? Is $c'_3$ in $D_0$?

Average test statistics

$k$ Random canaries $c_1, c_2, c_3$

$k$-1 Random canaries
Multiple hypothesis tests for auditing Lifted DP

If the statistics are independent ⇒ better confidence intervals

Unfortunately, they are **dependent** (but highly uncorrelated)
Novel higher-order confidence interval

- 2nd-order confidence interval using empirical correlations between two tests

\[ |\text{TPR} - \widehat{\text{TPR}}_{n,k}| \lesssim \sqrt{\frac{1}{n} \left( \text{Correlation} + \frac{1}{k} + \sqrt{\frac{\text{4th moment}}{n}} \right)} \]

- Ideally, when correlation=\( O(1/k) \), the confidence interval improves as

\[ |\text{TPR} - \widehat{\text{TPR}}_{n,k}| \lesssim \sqrt{\frac{1}{nk} + \frac{1}{n^{3/4}}} \]
Takeaway: **Reduces variance** from randomness in trials

Standard approach: \[ \varepsilon \geq \log \left( \frac{\hat{TPR}_n - \frac{c}{\sqrt{n}} - \delta}{\hat{FPR}_n + \frac{c}{\sqrt{n}}} \right) \]

- \( c \) - Universal constant
- \( c' \) - Data-dependent constant

Lower variance =>
**Tighter confidence intervals**

Our approach: \[ \varepsilon \geq \log \left( \frac{\hat{TPR}_{n,k} - \frac{c}{\sqrt{nk}} - \frac{c'}{n^{3/4}} - \delta}{\hat{FPR}_{n,k} + \frac{c}{\sqrt{nk}} + \frac{c'}{n^{3/4}}} \right) \]
Proof of concept with Gaussian mechanisms

- Sum query with sensitivity 1
- Gaussian mechanism
- $k$ canaries uniformly random on the sphere
- **Test statistic** is inner product

Dwork, Smith, Steinke, Ullman, Vadhan. *Robust traceability from trace amounts*. FOCS 2015
Gain in sample complexity (FashionMNIST)

Suffices to train **200 models** instead of 1000 models

- $\varepsilon = 2$
- $\varepsilon = 4$
- $\varepsilon = 8$
- $\varepsilon = 16$

**Baseline is better**

**Equal**

**LiDP is better**

- Data poison
- Gradient poison
Privacy Auditing with One (1) Training Run

Thomas Steinke*
Google DeepMind
steinke@google.com

Milad Nasr*
Google DeepMind
srxzr@google.com

Matthew Jagielski*
Google DeepMind
jagielski@google.com
Bias-variance tradeoff in the number of canaries $k$

$\varepsilon = 4.0$, $n = 4096$, $d = 10^4$

\[
|TPR - \widehat{TPR}_{n,k}| \lesssim \sqrt{\frac{1}{nk}} + \frac{1}{n^{3/4}}
\]
Summary

- **Auditing Lifted DP** (equivalent to usual DP) using multiple **i.i.d. random canaries** to improve sample dependence of the confidence intervals

- Can integrate with existing recipes for designing canaries
Other highlights: large-scale group-stratified datasets

Dataset Grouper

Library for creating group-structured datasets.

- **Scalable:** can handle millions of clients ✅
- **Flexible:** any custom partition function on any TFDS/HuggingFace dataset ✅
- **Platform-agnostic:** works with TF, PyTorch, JAX, NumPy, ... ✅

NeurIPS D&B 2023
New federated LLM datasets: longer sequences

Largest previous datasets: Reddit, Stack Overflow

Typical sequence length of LLMs

Our datasets: FedC4, FedBookCO
New federated LLM datasets: more words & groups

Largest previous datasets

Our datasets

Total words

30x larger

Reddit Stack Overflow FedC4 FedBookCO

Largest previous datasets Our datasets

Total groups

10x larger

Reddit Stack Overflow FedC4 FedBookCO

Largest previous datasets Our datasets
Thank you!