Near-Optimal Private Learning with Correlated Noise Mechanisms

Krishna Pillutla

Jan. 12 2025 @ CMI (BIRS Workshop)

INDIAN INSTITUTE OF TECHNOLOGY MADRAS • सिद्धिर्भवति कर्मजा •

CeRAI Centre for Responsible AI

WSAI IITM

Joint work with

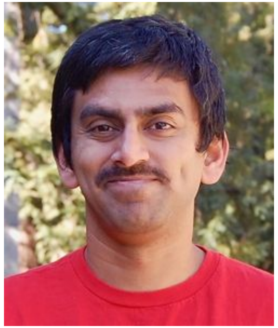

Chris Choquette-Choo


Dj Dvijotham


Brendan McMahan


Arun Ganesh

Thomas Steinke


Abhradeep Thakurta

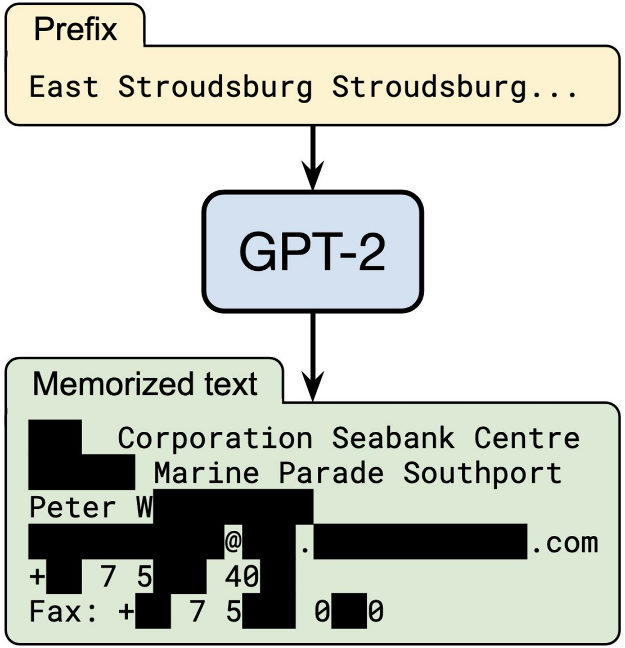
*****Equal contribution, $\alpha\beta$-order

Choquette-Choo*, Dvijotham*, ***P.****, Ganesh, Steinke, Thakurta.
Correlated Noise Provably Beats Independent Noise for Differentially Private Learning.
ICLR (2024).

Dvijotham, McMahan, ***P.***, Steinke, Thakurta.
Efficient and Near-Optimal Noise Generation for Streaming Differential Privacy.
FOCS (2024).

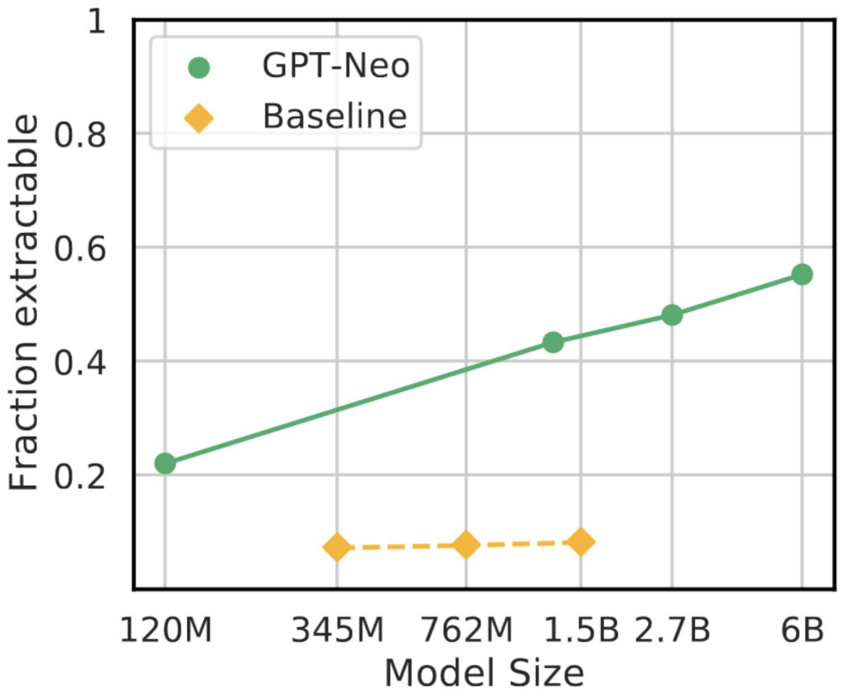
WHEN YOU TRAIN PREDICTIVE MODELS ON INPUT FROM YOUR USERS, IT CAN LEAK INFORMATION IN UNEXPECTED WAYS.

Models leak information about their training data

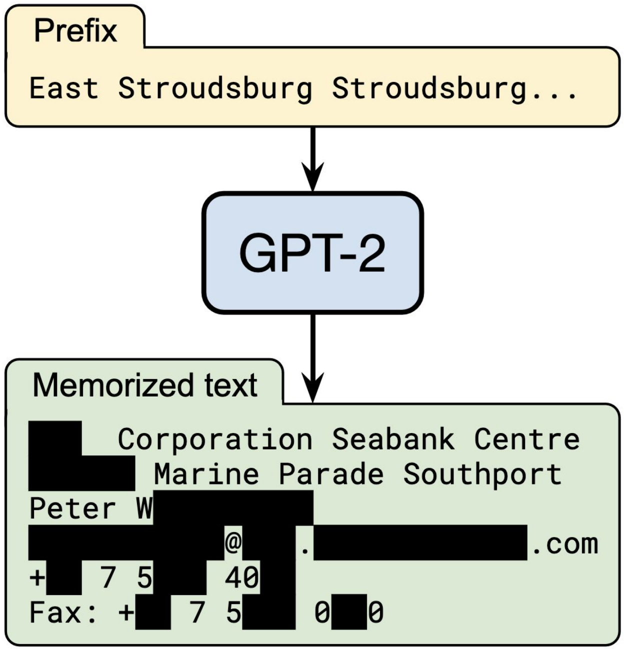


Carlini et al. (USENIX Security 2021)

Models leak information about their training data ***reliably***



Carlini et al. (ICLR 2023)



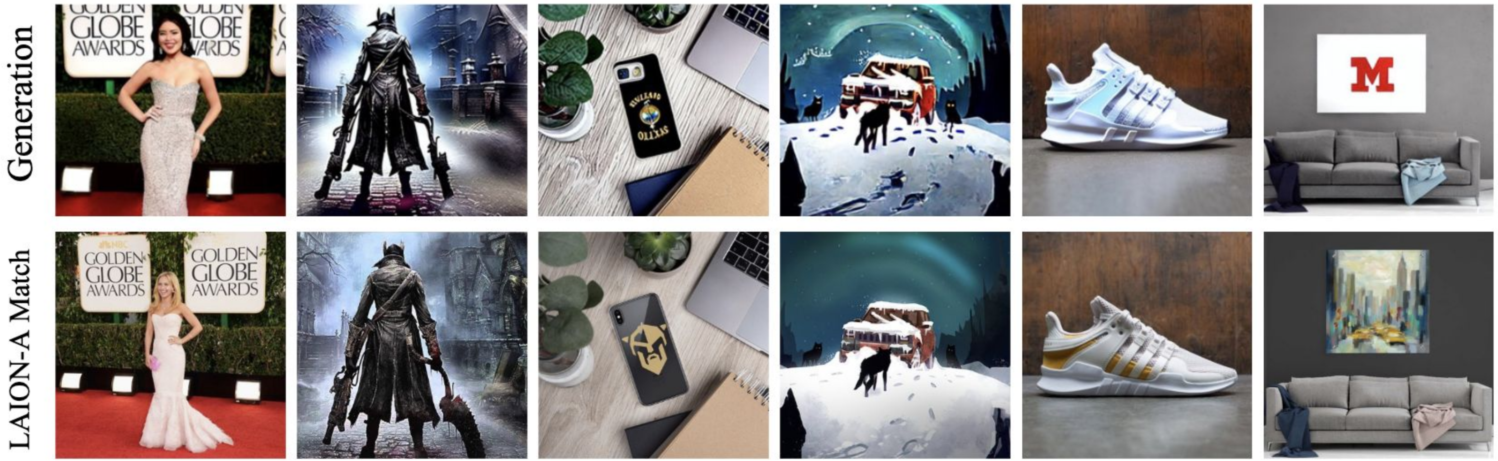
Carlini et al. (USENIX Security 2021)

Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models

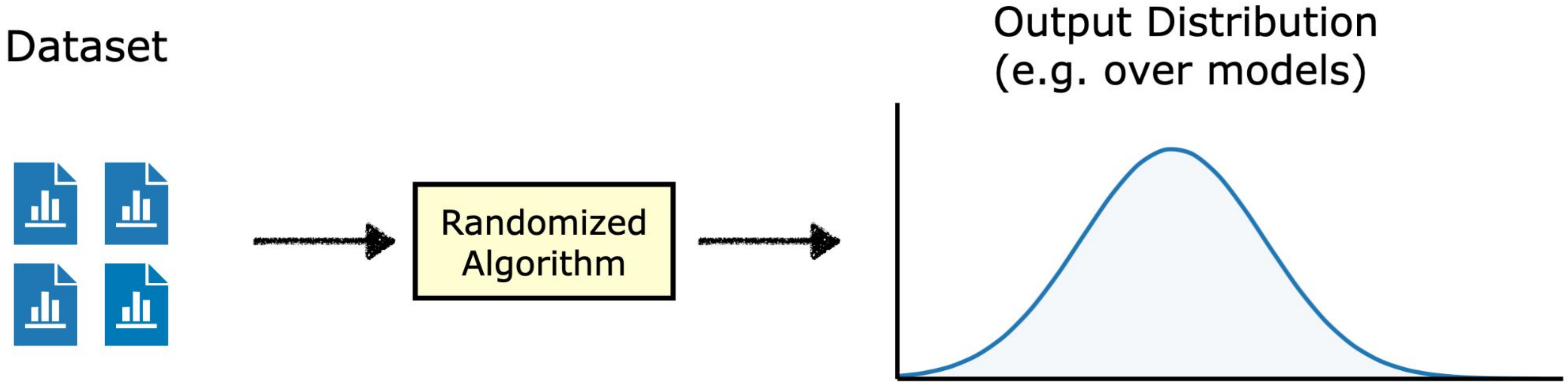
Gowthami Somepalli 🐢, Vasu Singla 🐢, Micah Goldblum 🗽, Jonas Geiping 🐢, Tom Goldstein 🐢

🐢 University of Maryland, College Park
{gowthami, vsingla, jgeiping, tomg}@cs.umd.edu

🗽 New York University
goldblum@nyu.edu

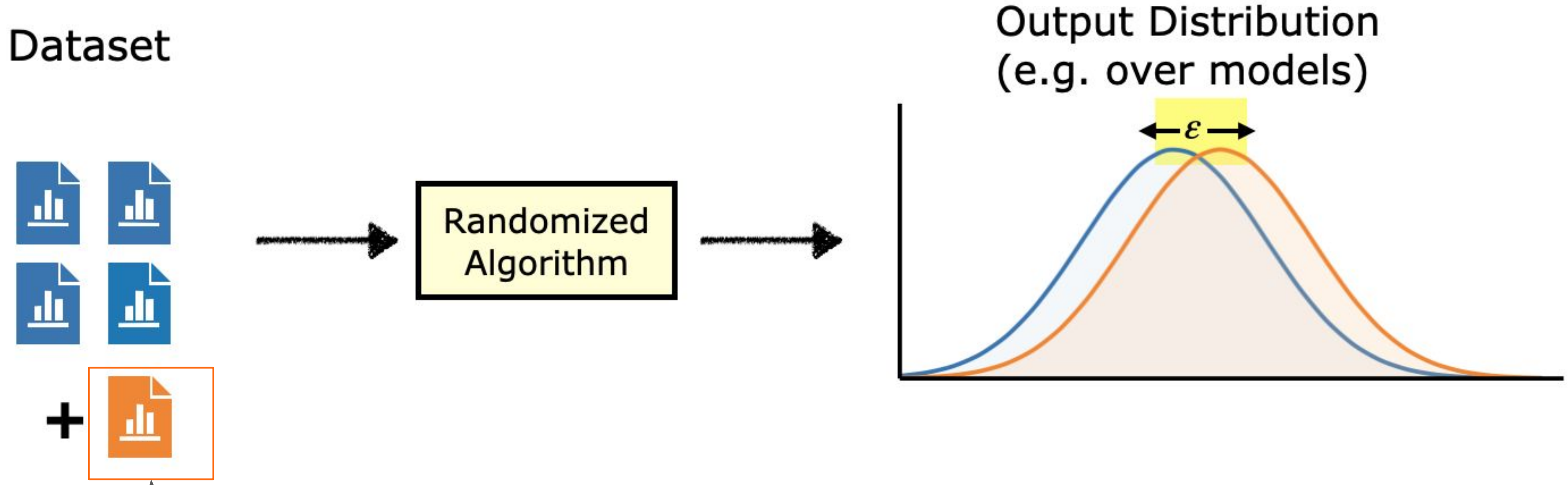


Differential privacy (DP)



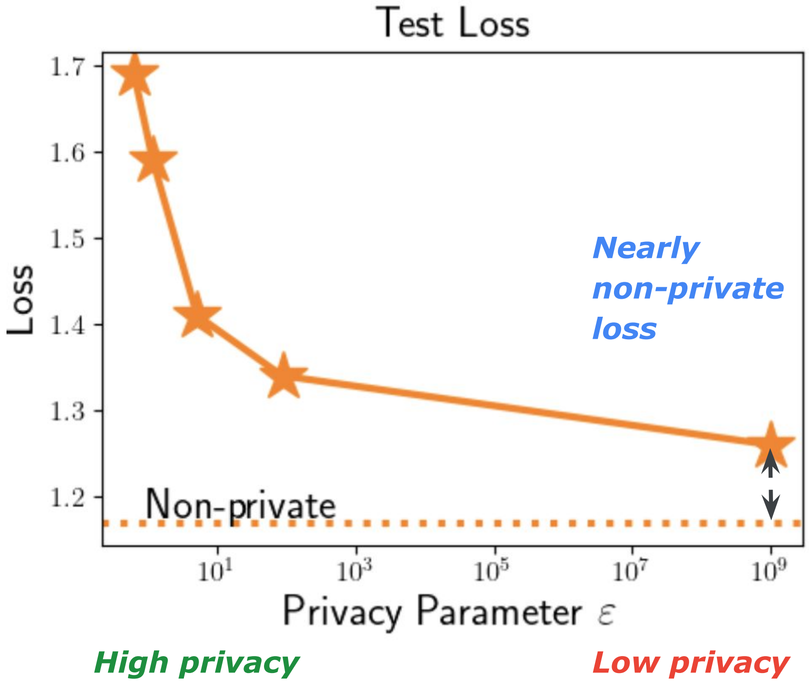
Dwork, McSherry, Nissim, Smith. **Calibrating noise to sensitivity in private data analysis**. TCC 2006

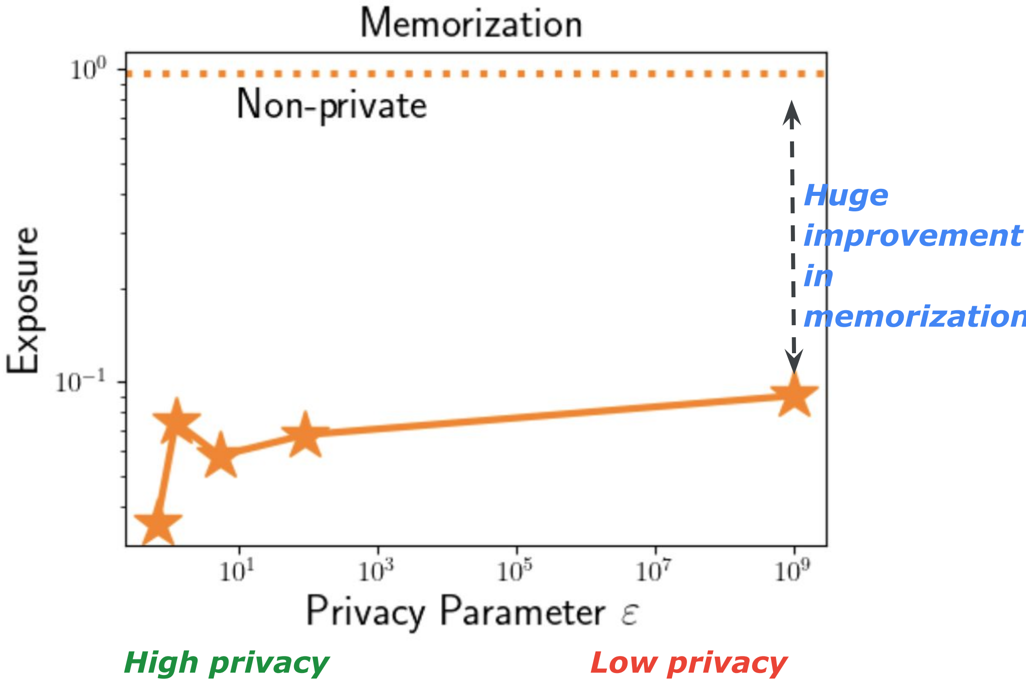
Differential privacy (DP)



A randomized algorithm is ε-**differentially private** if the addition of **one unit of data** does not alter its output distribution by more than ε

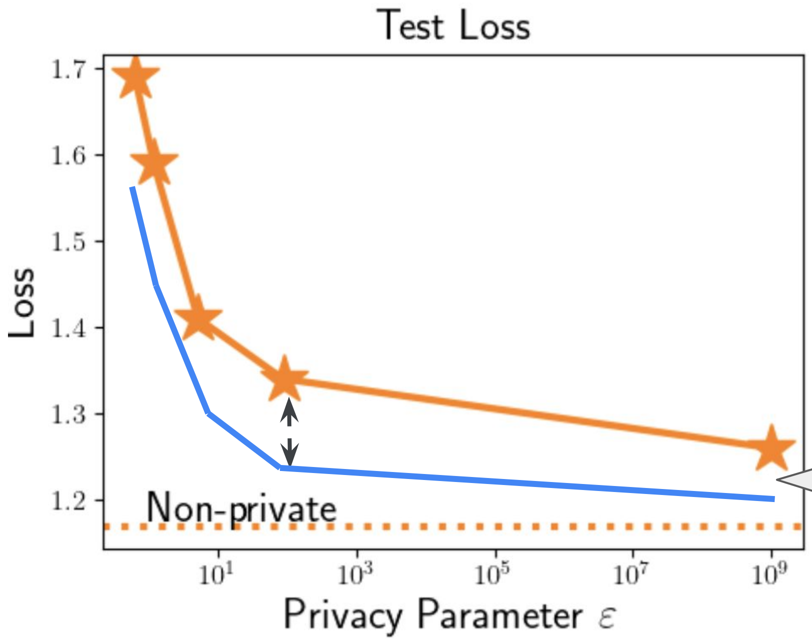
Differential privacy eliminates memorization





Carlini, Liu, Erlingsson, Kos, Song. **The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks.** USENIX Security 2019.

Goal: Better privacy-utility trade-offs



High privacy ***Low privacy***

Goal: Achieve *smaller loss* at each privacy level

How do we train models with DP?

Loss function

Model parameters

$$\min_{\theta} \left[F(\theta) = \mathbb{E}_{x \sim P} \left[f(\theta; x) \right] \right]$$

Data

DP-SGD: How do we train models with DP?

Independent
Gaussian noise

Stochastic gradient
clipped to $\|g\|_2 \leq 1$
per-example

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

Learning
rate

Song et al. (2013), Bassily et al. (FOCS 2014), Abadi et al. (CCS 2016)

DP-FTRL: DP Training with ***Correlated*** Noise

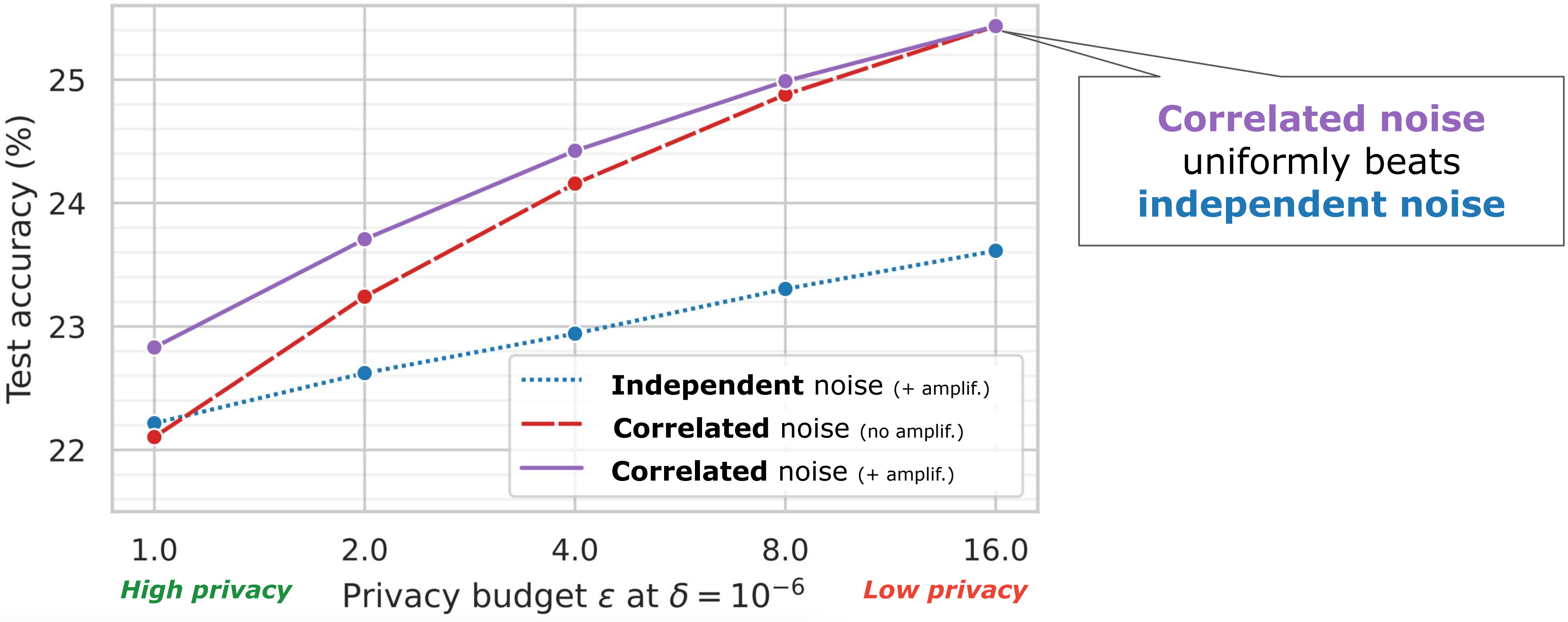
(Anti-)correlated
Gaussian noise
(z_t i.i.d. Gaussian)

$$\theta_{t+1} = \theta_t - \eta\Big(g_t + \boxed{z_t - \sum_{\tau=1}^{t} \beta_\tau z_{t-\tau}}\Big)$$

Kairouz, McMahan, Song, Thakkar, Thakurta, Xu. **Practical and Private (Deep) Learning without Sampling or Shuffling**. ICML 2021.
Denisov, McMahan, Rush, Smith, Thakurta. **Improved Differential Privacy for SGD via Optimal Private Linear Operators on Adaptive Streams**. NeurIPS 2022.

Prior work: (Empirically) correlated noise outperforms independent noise

Experiment: DP language modeling
Dataset: StackOverflow

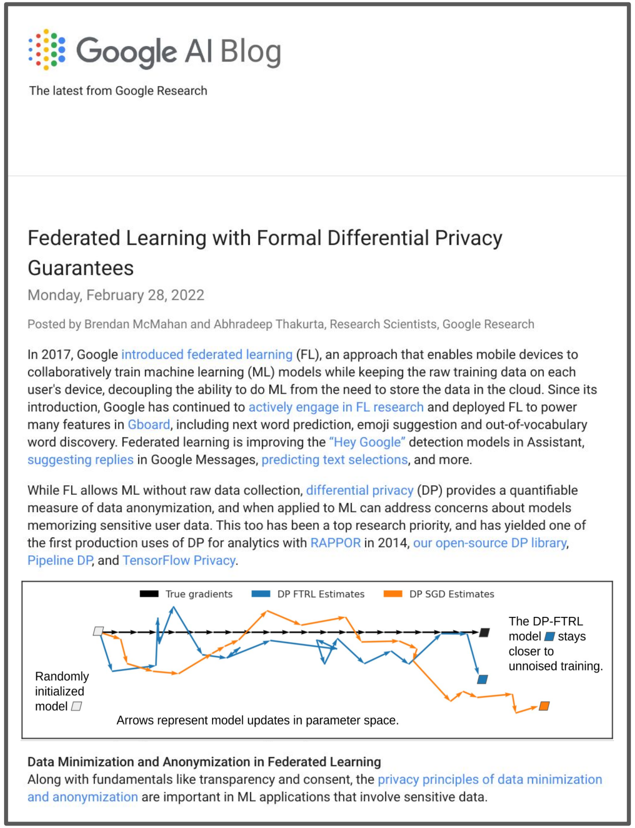


Choquette-Choo, Ganesh, McKenna, McMahan, Rush, Thakurta, Xu.
(Amplified) Banded Matrix Factorization: A unified approach to private training. NeurIPS 2023

Production Training

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

- [Google AI Blog post](), Feb 2022



Google AI Blog

The latest from Google Research

Federated Learning with Formal Differential Privacy Guarantees

Monday, February 28, 2022

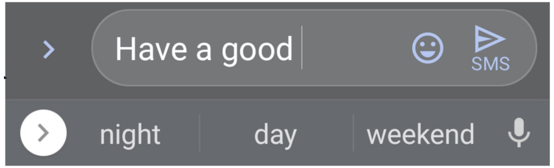
Posted by Brendan McMahan and Abhradeep Thakurta, Research Scientists, Google Research

In 2017, Google introduced federated learning (FL), an approach that enables mobile devices to collaboratively train machine learning (ML) models while keeping the raw training data on each user's device, decoupling the ability to do ML from the need to store the data in the cloud. Since its introduction, Google has continued to actively engage in FL research and deployed FL to power many features in Gboard, including next word prediction, emoji suggestion and out-of-vocabulary word discovery. Federated learning is improving the "Hey Google" detection models in Assistant, suggesting replies in Google Messages, predicting text selections, and more.

While FL allows ML without raw data collection, differential privacy (DP) provides a quantifiable measure of data anonymization, and when applied to ML can address concerns about models memorizing sensitive user data. This too has been a top research priority, and has yielded one of the first production uses of DP for analytics with RAPPOR in 2014, our open-source DP library, Pipeline DP, and TensorFlow Privacy.

Data Minimization and Anonymization in Federated Learning

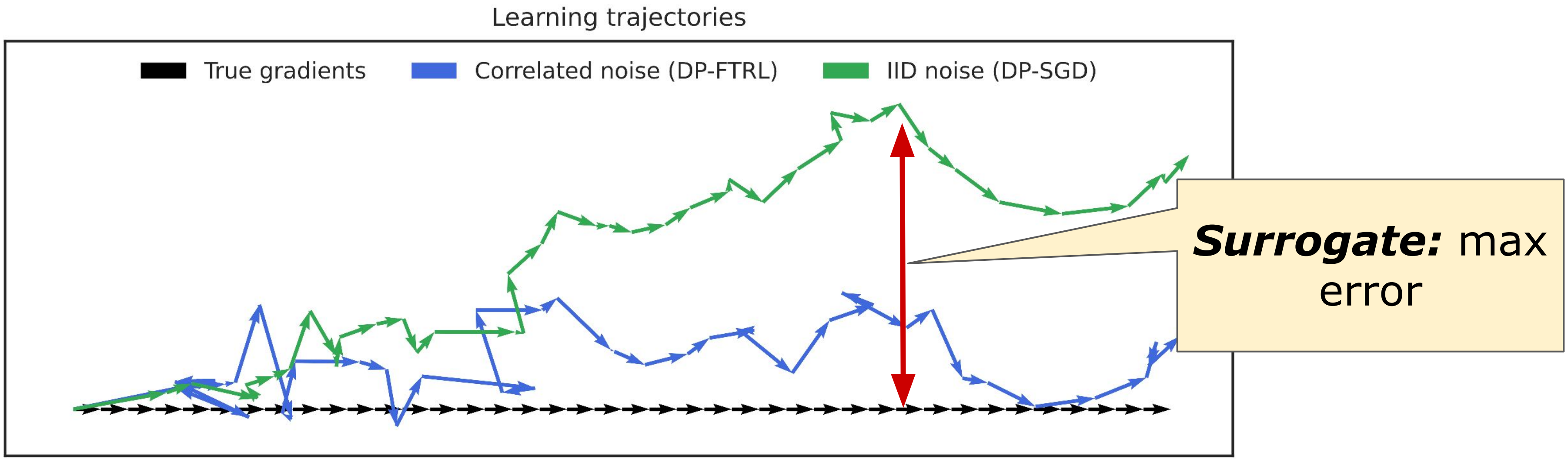
Along with fundamentals like transparency and consent, the privacy principles of data minimization and anonymization are important in ML applications that involve sensitive data.



How do we find the noise coefficients?

How do we find the noise coefficients?

Current Approach:
Find the noise coefficients β_t to ***minimize the cumulative noise*** added to the learning trajectory (such that a given DP constraint is satisfied)



How do we find the noise coefficients?

$$\theta_{t+1} = \theta_t - \eta\Big(g_t + \underbrace{\boxed{z_t - \sum_{\tau=1}^{t} \beta_\tau z_{t-\tau}}}_{=:w_t}\Big)$$

Find the noise coefficients β_t to ***minimize the max error*** (i.e. cumulative noise added to the learning trajectory):

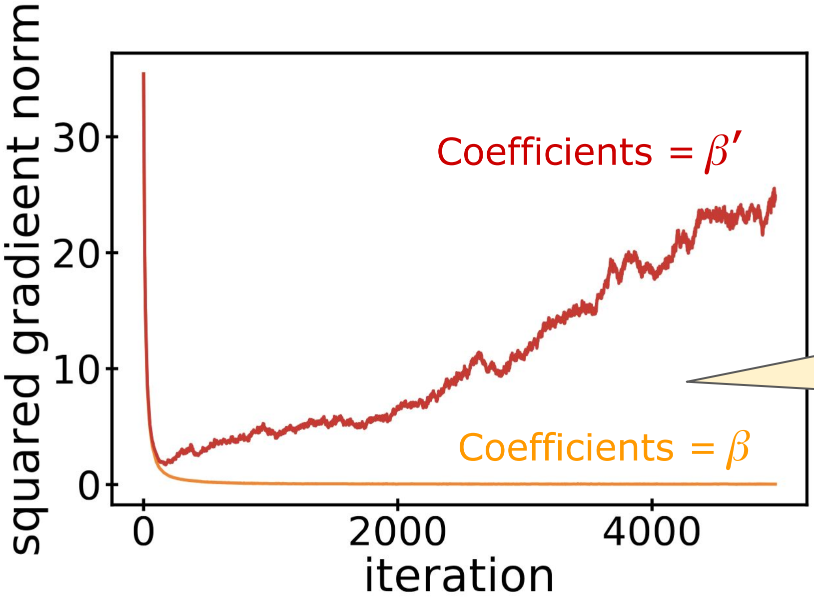
Surrogate Objective

$$\mathcal{E}(\beta)^2 = \max_{t \le n} \mathbb{E}_{z_\tau \sim \mathcal{N}(0,\sigma^2 I)} \left\| \sum_{\tau=0}^{t} w_\tau \right\|_2^2$$

where the variance σ^2 is chosen so that θ_t's satisfy a given DP constraint

Part 1: Is correlated noise provably better for learning problems?

The surrogate objective is not related to the learning objective



$$\mathcal{E}(\beta) = \mathcal{E}(\beta')$$

Same ***max error*** but different ***learning performance***

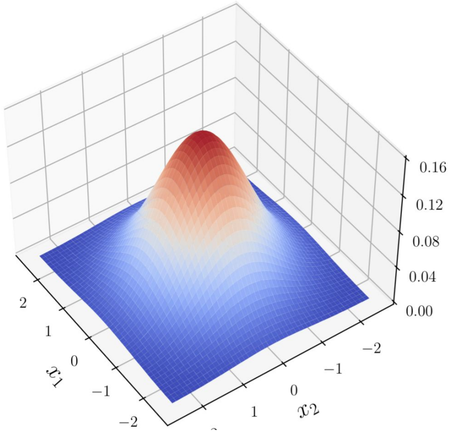
Koloskova, McKenna, Charles, Rush, McMahan. **Gradient Descent with Linearly Correlated Noise: Theory and Applications to Differential Privacy**. NeurIPS 2023

Part 1: Correlated noise *is* provably better for learning problems

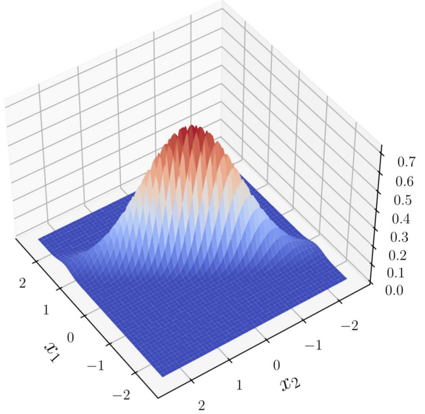
(Anti-) correlated noise provably beats independent noise

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

Independent noise	$\Theta(d)$
Correlated noise	$\tilde{O}(d_{\mathrm{eff}})$
Lower bound	$\Omega(d_{\mathrm{eff}})$



High effective dimension



Low effective dimension

Part 2: Noise generation time complexity

$$\theta_{t+1} = \theta_t - \eta\Big(g_t + z_t - \boxed{\sum_{\tau=1}^{t} \beta_\tau z_{t-\tau}}\Big)$$

Quadratic time complexity:
Noise generation requires *O(t)* time in iteration *t*

Part 2: Near-optimal noise generation time complexity

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

Our approach: Approximate the noise coefficients as

$$\beta_t \approx \beta'_t = \sum_{i=1}^{d} \alpha_i \lambda_i^{t-1}$$

Per-iteration time:
O(*d* x dimension)

Part 2: Near-optimal noise generation time complexity

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

Our approach: Approximate the noise coefficients as

$$\beta_t \approx \beta'_t = \sum_{i=1}^{d} \alpha_i \lambda_i^{t-1}$$

Per-iteration time:
O(*d* x dimension)

Error: If *d* = *O*(*log*^{2}(*n*/*c*)), then the error is $\mathcal{E}(\beta') \leq \mathcal{E}(\beta) + c$

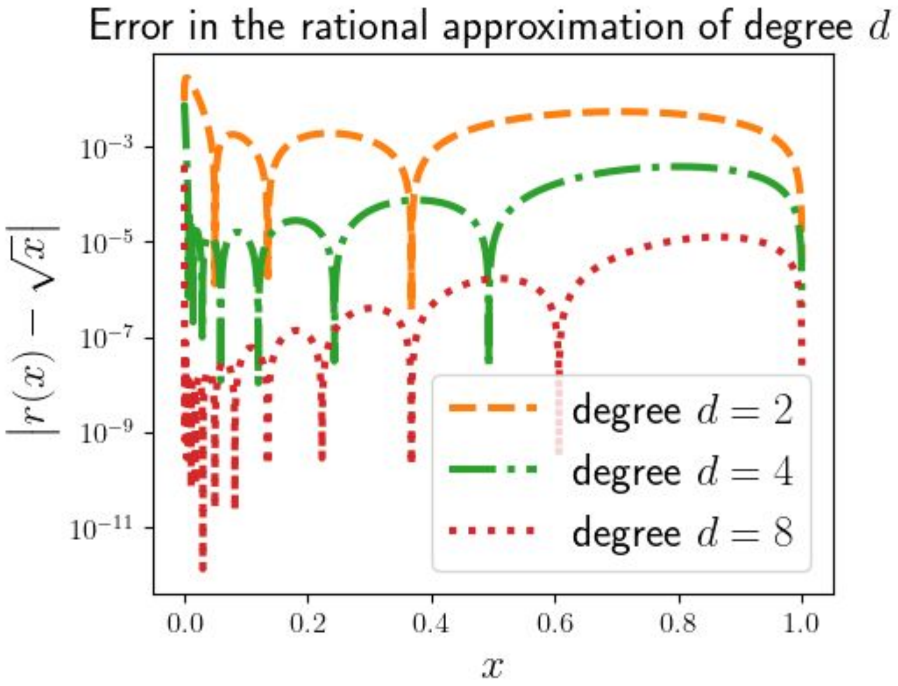
(*n* = Number of steps)

Part 2: Near-optimal noise generation time complexity

Key insight: Approximation theory

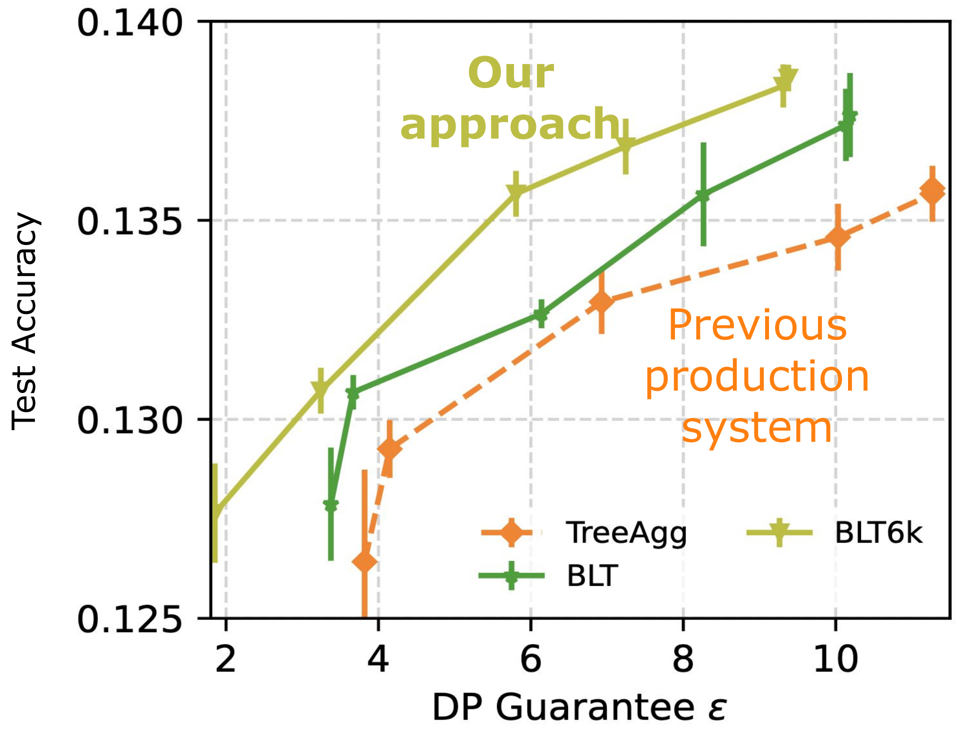
There exists a
rational function *r(x)* of degree-*d*
that satisfies the approximation:

$$\sup_{x\in[0,1]} |r(x) - \sqrt{x}| \leq 3 \cdot \exp(-\sqrt{d}).$$



Newman. **Rational approximation to |*x*|**. Michigan Math. J. (1964)

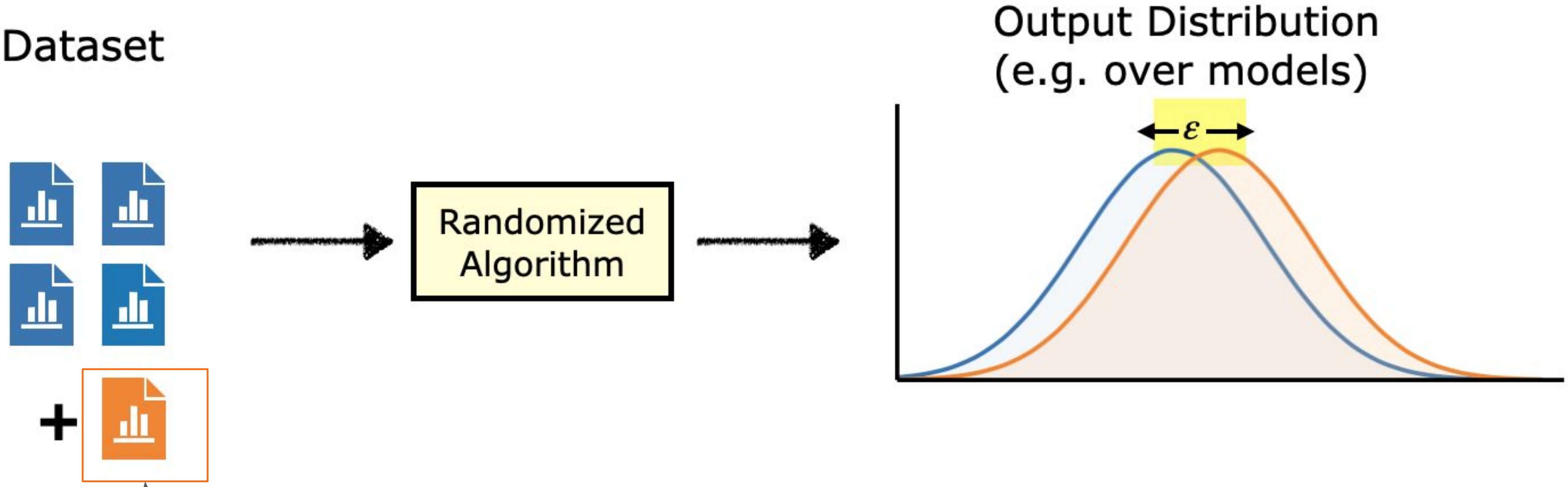
Practical Impact:
Google's production language model (Portuguese)



Plot: McMahan, Xu, Zhang (2024)

Background

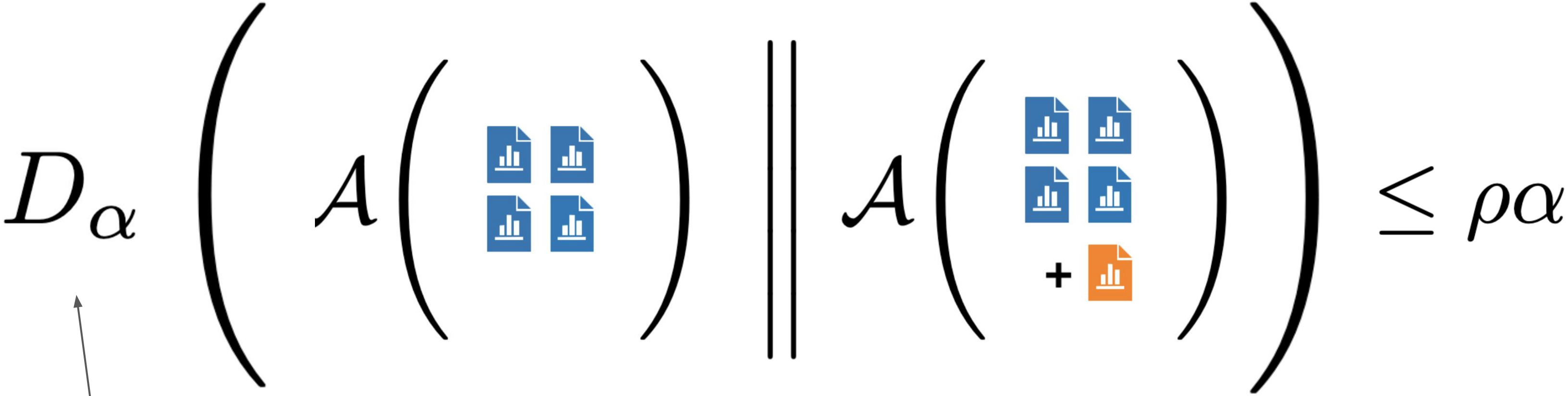
Differential privacy (DP)



A randomized algorithm is ε-**differentially private** if the addition of **one unit of data** does not alter its output distribution by more than ε

ϱ-Zero-Concentrated DP (ϱ-zCDP)

For all $0 < \alpha < \infty$, we have

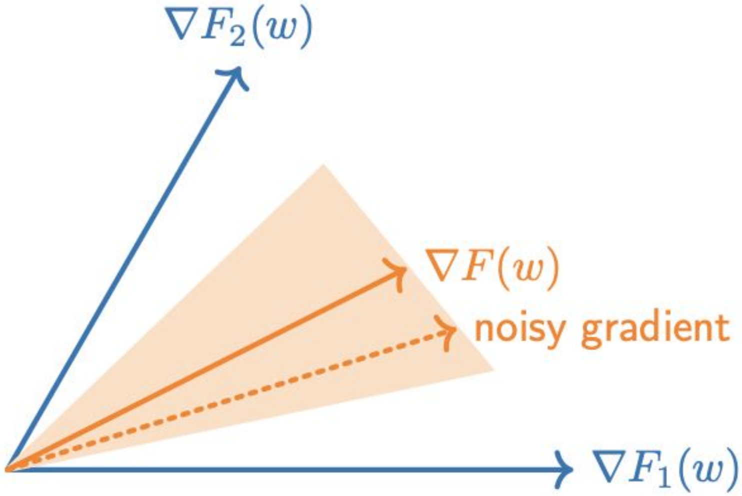


Rényi α-divergence

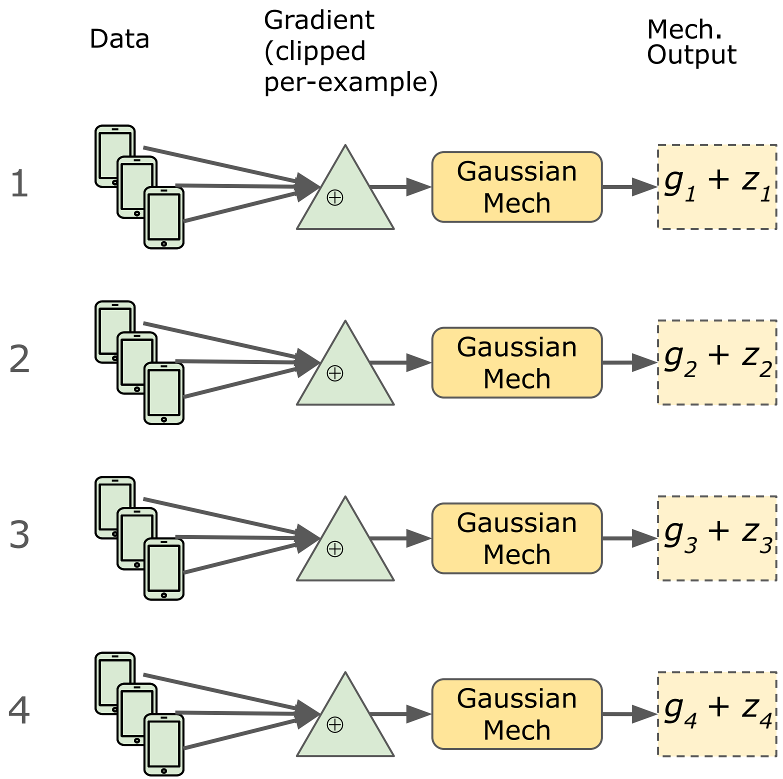
Bun & Steinke. **Concentrated Differential Privacy: Simplifications, Extensions, and Lower Bounds**. TCC 2016

DP-SGD / Independent noise

Primitive: private mean estimation of minibatch (clipped) gradients in each iteration



DP-SGD adds independent noise in each iteration



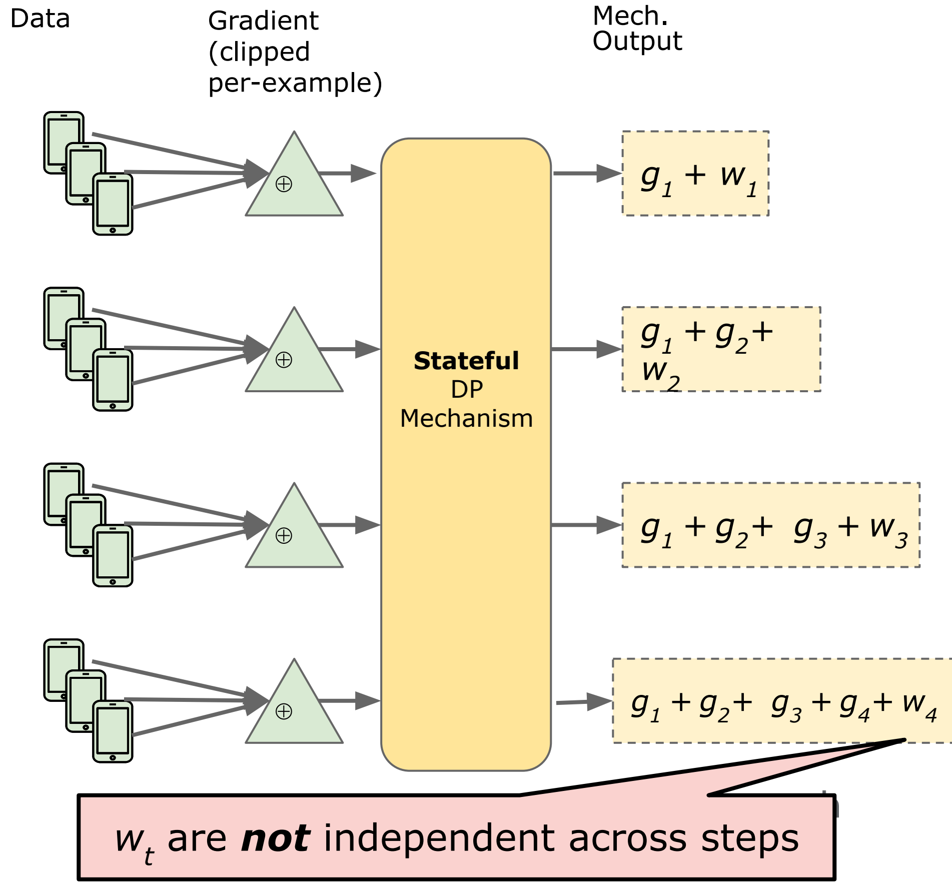
For ϱ-zCDP, take
$\mathrm{Var}(z_t) = 1/(2\varrho)$

Abadi et. al., **Deep Learning with Differential Privacy**, CCS 2016.

DP-FTRL: privatize prefix sums of gradients

$$\theta_t - \theta_0 = -\sum_{\tau=0}^{t-1} g_\tau$$

SGD update (without noise)



Kairouz, McMahan, Song, Thakkar, Thakurta, Xu. **Practical and Private (Deep) Learning without Sampling or Shuffling**. ICML 2021.

DP-FTRL: privatize prefix sums of gradients

For ϱ-zCDP, take

$$\mathrm{Var}(z_t) = \frac{1}{2\rho}\left(\max_{t \le n} \|(B^{-1})[:,t]\|_2^2\right)$$

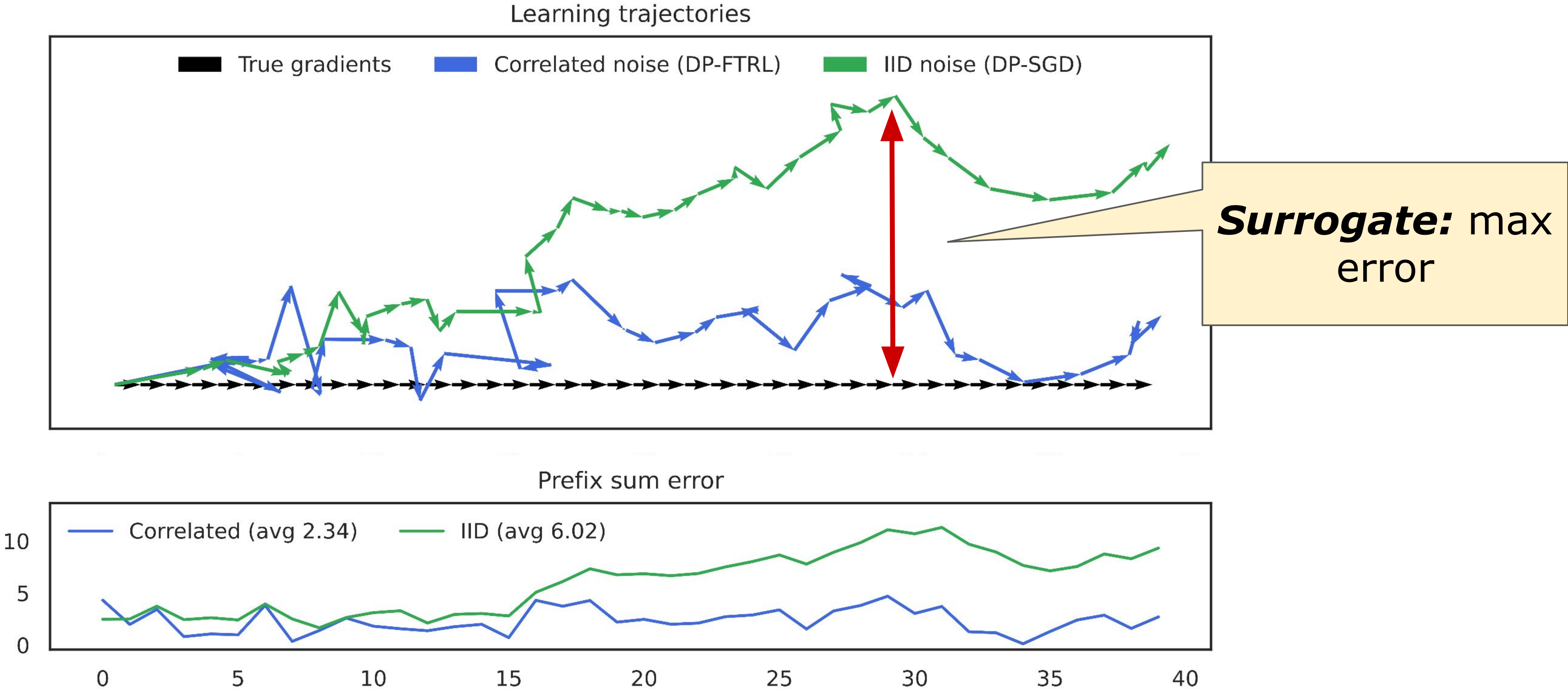
sensitivity

$$B = \begin{pmatrix} 1 & & & & \\ -\beta_1 & 1 & & & \\ -\beta_2 & -\beta_1 & 1 & & \\ \vdots & \vdots & \vdots & \ddots & \\ -\beta_{n-1} & -\beta_{n-2} & \cdots & \cdots & 1 \end{pmatrix}$$

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

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

Kairouz, McMahan, Song, Thakkar, Thakurta, Xu. **Practical and Private (Deep) Learning without Sampling or Shuffling**. ICML 2021.
Denisov, McMahan, Rush, Smith, Thakurta. **Improved Differential Privacy for SGD via Optimal Private Linear Operators on Adaptive Streams**. NeurIPS 2022.


Learning trajectories
True gradients
Correlated noise (DP-FTRL)
IID noise (DP-SGD)
Surrogate: max error
Prefix sum error
Correlated (avg 2.34)
IID (avg 6.02)
10
5
0
0
5
10
15
20
25
30
35
40

Toeplitz mechanism: optimal max error

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

Theorem

[Fichtenberger, Henzinger, Upadhyay (ICML ‘23); Dvijotham, McMahan, ***P***., Steinke, Thakurta (FOCS ‘24)]

For any number n of steps, the optimal max error is obtained by coefficients $\beta_t^* = t^{-3/2}$ and satisfies the bounds

$$\mathcal{E}(\beta^*) = \frac{\log n}{\pi} + \text{constant}$$

Toeplitz mechanism: optimal max error

$$\theta_{t+1} = \theta_t - \eta\Big(g_t + z_t - \sum_{\tau=1}^{t} \boxed{\beta_\tau} z_{t-\tau}\Big)$$

Theorem

[Fichtenberger, Henzinger, Upadhyay (ICML '23); Dvijotham, McMahan, ***P***., Steinke, Thakurta (FOCS '24)]

For any number *n* of steps, the optimal max error is obtained by coefficients $\beta_t^* = t^{-3/2}$ and satisfies the bounds

$$\mathcal{E}(\beta^*) = \frac{\log n}{\pi} + \text{constant}$$

Exponential improvement over independent noise

$$\mathcal{E}(\beta^{\mathrm{SGD}}) = \Theta(\sqrt{n})$$

Part 1: Learning guarantees

(Anti-)correlated noise **provably** beats independent noise

ICLR 2024

DP-FTRL vs. DP-SGD: Previous Theory

For convex & *G*-Lipschitz losses

Independent Noise	$\frac{Gd^{1/4}}{\sqrt{\rho T}}$
Correlated Noise	$\frac{Gd^{1/4}}{\sqrt{\rho^2 T}}$

ϱ: privacy level (zCDP)
d: dimension
T: #iterations

Kairouz, McMahan, Song, Thakkar, Thakurta, Xu. **Practical and Private (Deep) Learning without Sampling or Shuffling**. ICML 2021.

Setting and Simplifications

Loss function

Model parameters

$$\min_{\theta} \left[F(\theta) = \mathbb{E}_{x \sim P} \left[f(\theta; x) \right] \right]$$

Data

Streaming setting: Suppose we draw a fresh data point $x_t \sim P$ in each iteration t (i.e. only 1 epoch)

Asymptotics: Iterates converge to a stationary distribution as $t \to \infty$

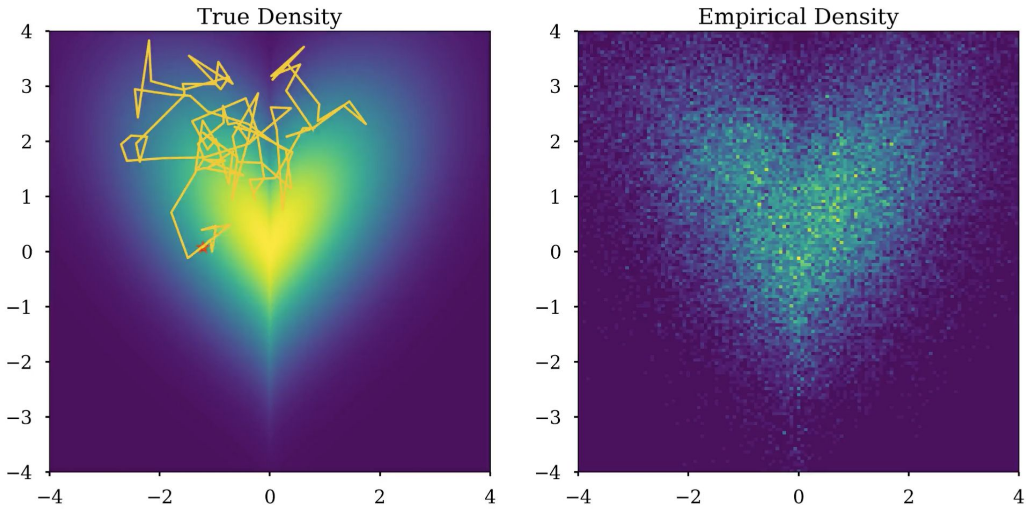


Image credit:
[Abdul Fatir Ansari]

Asymptotics: Iterates converge to a stationary distribution as $t \to \infty$

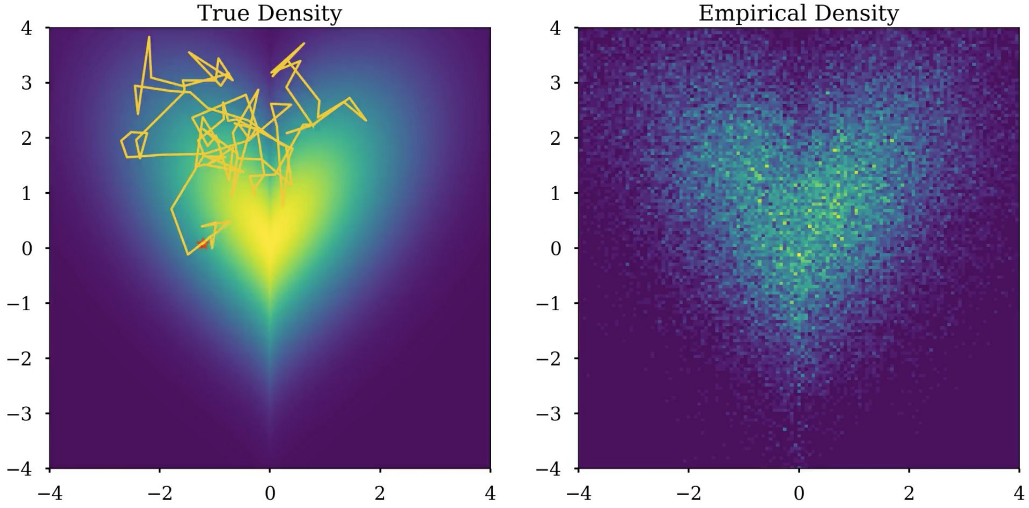


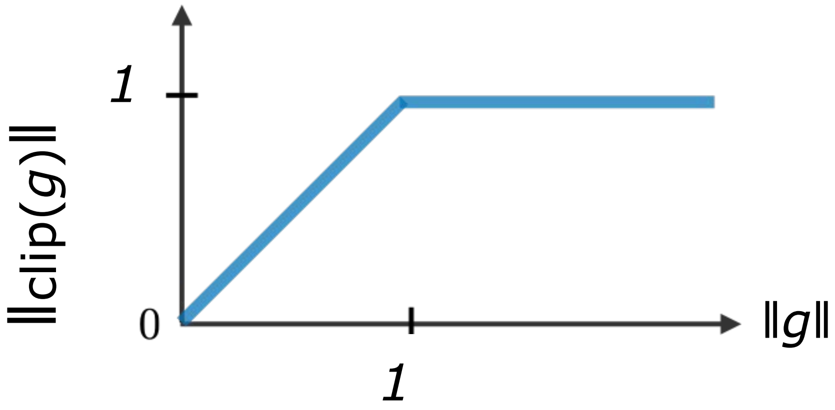
Image credit: Abdul Fatir Ansari

Asymptotic error

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

Asymptotics at a fixed learning rate $\eta > 0$

Noisy-SGD/Noisy-FTRL: DP-SGD/DP-FTRL without clipping



Lets us study the noise dynamics of the algorithms
(do not satisfy DP guarantees)

Mean estimation in 1 dimension

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

Data distribution
s.t. $|\boldsymbol{x}| \leq 1$

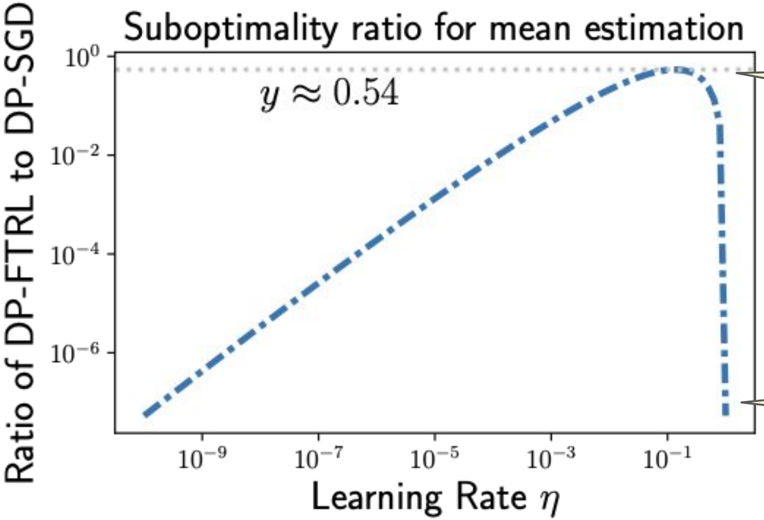
Solve with stochastic optimization problem
with DP-SGD/DP-FTRL

Mean estimation in 1 dimension

Informal Theorem: The asymptotic error of a ϱ-zCDP sequence is

Independent noise (DP-SGD)	$F_\infty(\beta^{\mathrm{sgd}}) = \rho^{-1}\eta$
Correlated noise (DP-FTRL)	$\inf_\beta F_\infty(\beta) = F_\infty(\beta^\star) = \rho^{-1}\eta^2 \log^2 \frac{1}{\eta}$

η: constant learning rate in (0, 1)
ϱ: privacy level



DP-FTRL is always better than DP-SGD

DP-FTRL is significantly better at $\eta \to 0$ or $\eta \to 1$

Closed form correlations for mean estimation

Proposition: The correlations $\beta_t = t^{-3/2}(1-\eta)^t$ attain the optimal error

$$\inf_\beta F_\infty(\beta) = F_\infty(\beta^\star) = \rho^{-1}\eta^2 \log^2 \frac{1}{\eta}$$

Closed form correlations for mean estimation

Proposition: The correlations $\beta_t = t^{-3/2}(1 - \eta)^t$ attain the optimal error

$$\inf_{\beta} F_\infty(\beta) = F_\infty(\beta^\star) = \rho^{-1}\eta^2 \log^2 \frac{1}{\eta}$$

ν-DP-FTRL

For general problems, use $\beta_t = t^{-3/2}(1 - \nu)^t$ and tune the parameter ν

Linear regression

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

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

H is also the Hessian of the objective

Linear regression

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

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

$$y | x \sim \mathcal{N}(x^\top \theta_\star, \sigma^2)$$

Well-specified
linear model

Informal Theorem: The asymptotic error is

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

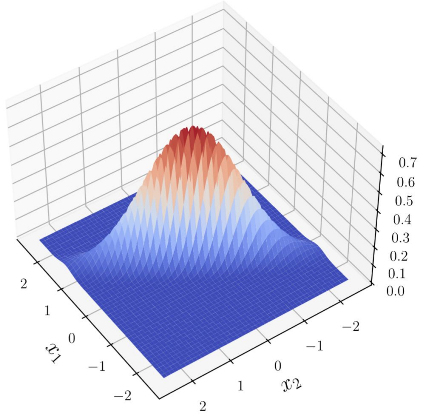
Improve dimension d to problem-dependent
effective dimension d_{eff}

Effective dimension

$$d_{\text{eff}} = \mathbf{Tr}(H)/\|H\|_2 \leq d$$

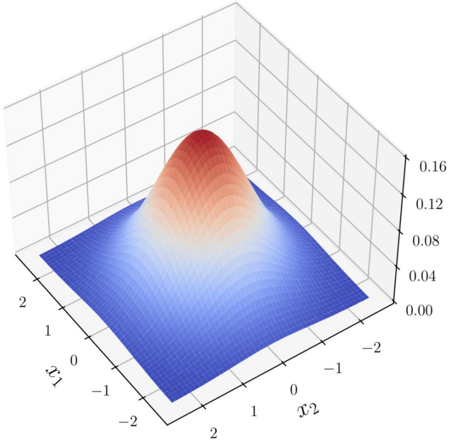
Low effective dimension

$$\lambda_1 = 1, \lambda_2 = \cdots = \lambda_d = 1/d$$



High effective dimension

$$\lambda_1 = \lambda_2 = \cdots = \lambda_d = 1$$



Closely connected to **numerical/stable rank**

SAMPLING FROM LARGE MATRICES: AN APPROACH THROUGH GEOMETRIC FUNCTIONAL ANALYSIS

MARK RUDELSON AND ROMAN VERSHYNIN

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:

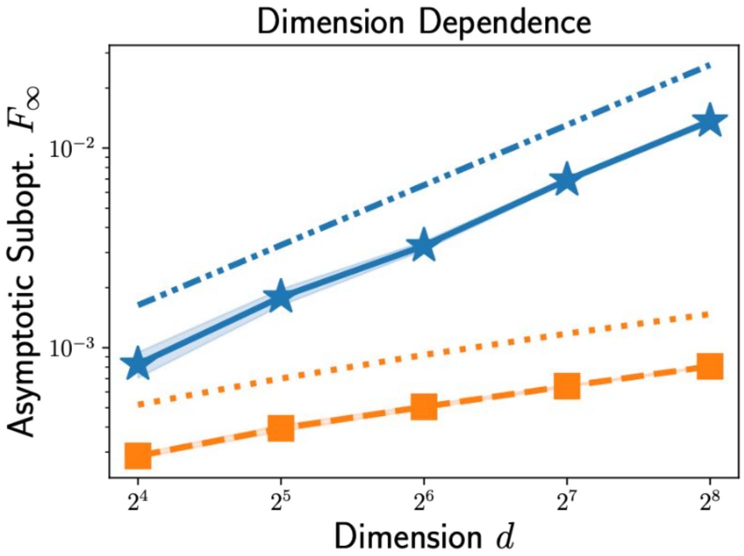
- Numerical linear algebra (e.g. randomized matrix multiplications) [Tropp (2014), Cohen-Nelson-Woodruff (2015)]
- Matrix concentration [Hsu-Kakade-Zhang (2012), Minsker (2017)]
- …

Informal Theorem: The asymptotic error is

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

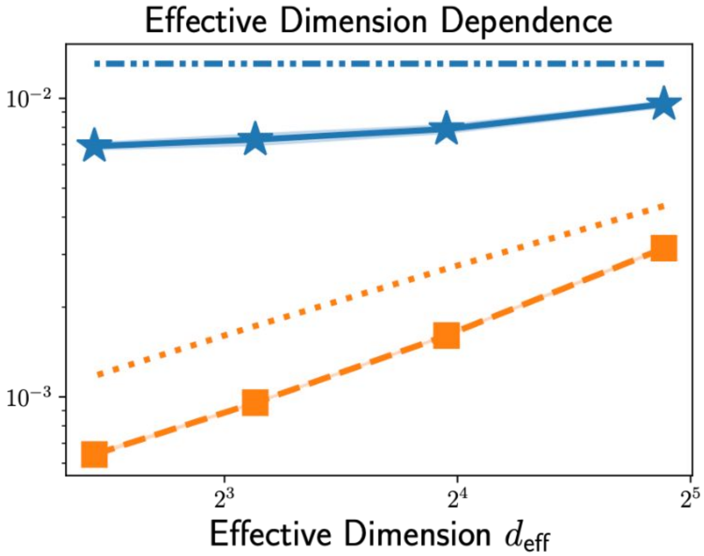
Improve dimension d to problem-dependent ***effective dimension*** d_{eff}

Linear regression: theory predicts simulations



Noisy-SGD scales with d

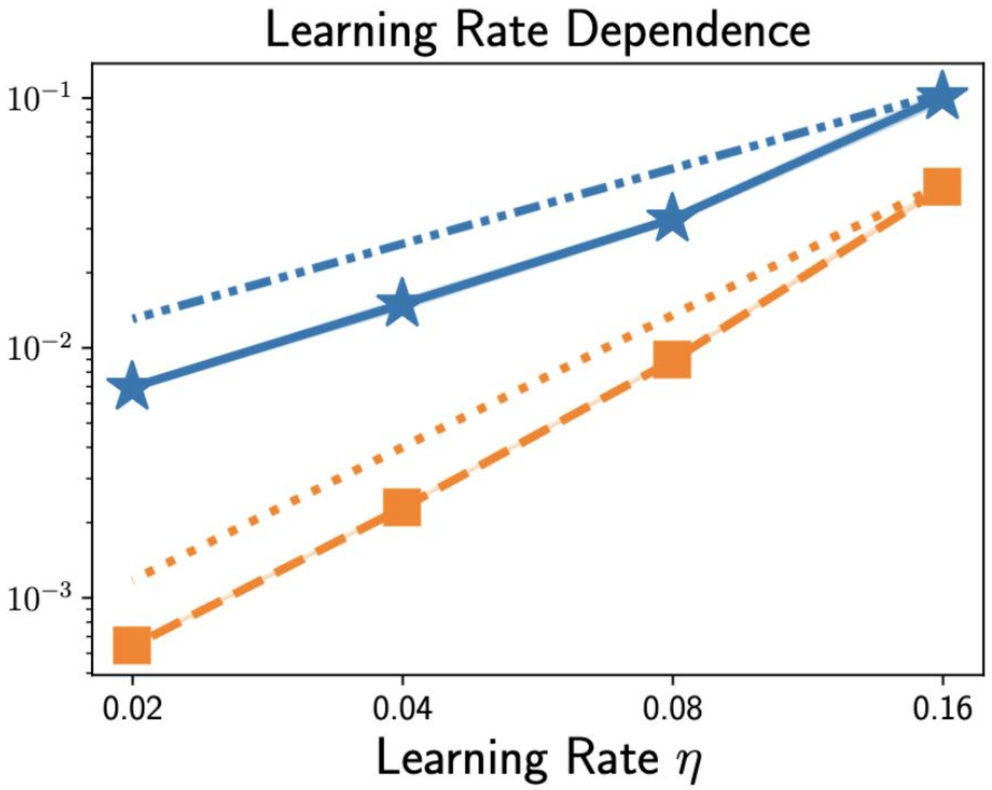
Noisy-FTRL scales with d_{eff}



Informal Theorem: The asymptotic error for $0 < \eta < 1$ is

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

Improved dependence on the learning rate η



Noisy-SGD scales as η

ν-**Noisy-FTRL**
scales as η^2

Noisy-FTRL $\gg$ **Noisy-SGD** at small η

Anticorrelated Noise Injection for Improved Generalization

Antonio Orvieto [*1] **Hans Kersting** [*2] **Frank Proske** [3] **Francis Bach** [2] **Aurelien Lucchi** [4]

Anti-PGD [Orvieto et al. (ICML '22)] corresponds to β_1=1

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

Subtract out the previous noise

Anticorrelated Noise Injection for Improved Generalization

Antonio Orvieto [*1] Hans Kersting [*2] Frank Proske [3] Francis Bach [2] Aurelien Lucchi [4]

Anti-PGD [Orvieto et al. (ICML ‘22)] corresponds to β_1=1

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

Asymptotic error = ∞ (as sensitivity scales of $O(t)$ for t iterations)

Anti-PGD can be adapted for DP by damping: take $\beta_1 = \nu$ ($0 < \nu < 1$)

$$\theta_{t+1} = \theta_t - \eta \left(g_t + z_t - \nu z_{t-1} \right)$$

Asymptotic error = $\sqrt{d d_{\text{eff}}}\ \rho^{-1}\ \eta^{3/2}$

Geometric mean of Noisy-SGD and lower bound

Finite-time rates with DP: Linear Regression

Independent noise (DP-SGD)	$\frac{1}{\rho T} + \frac{1}{T}$
Correlated noise (ν-DP-FTRL)	$\frac{1}{\rho T^2} + \frac{1}{T}$

Privacy error

T: number of iterations
ϱ: privacy level
η: learning rate is optimized

Proof sketch for Mean Estimation

Usual stochastic gradient proof patterns do not work:

No Markovian/martingale structure in the noise

Our approach: Analysis the ***Fourier*** domain

Letting $\boldsymbol{\delta}_t = \theta_t - \theta_*$, the DP-FTRL update can be written as

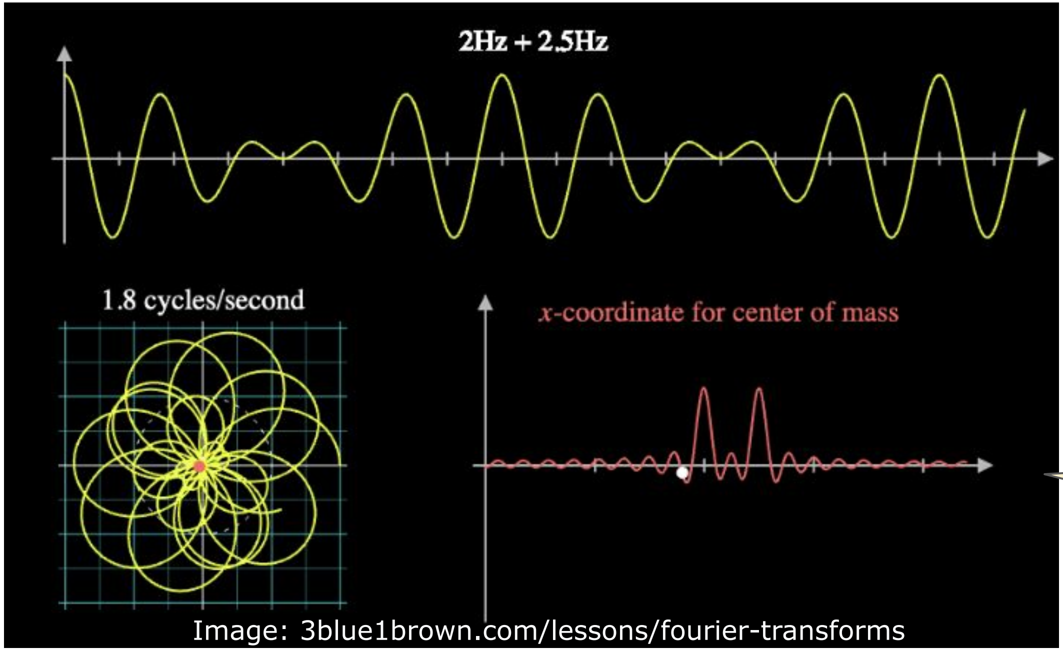
Linear Time-Invariant (LTI) system

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

Convolution of the noise

Fourier analysis can give the stationary variance of $\boldsymbol{\delta}_t$ in terms of the **discrete-time Fourier transform** $B(\omega) = \sum_{t=0}^{\infty} \beta_t e^{i\omega t}$ of the convolution weights β

Frequency



Time-domain description

Frequency-domain description

Letting $\boldsymbol{\delta}_t = \theta_t - \theta_*$, the DP-FTRL update can be written as

Linear Time-Invariant (LTI) system

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

Convolution of the noise

The stationary variance of $\boldsymbol{\delta}_t$ can be given as

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

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

sensitivity

For ϱ-zCDP, take

$$\mathbb{E}[z_t^2] = \frac{1}{2\rho} \max_t \left\|[B^{-1}]_{:,t}\right\|_2^2 = \frac{1}{2\rho}\int_{-\pi}^{\pi}\frac{\mathrm{d}\omega}{2\pi|B(\omega)|^2}$$

$$B = \begin{pmatrix} 1 & & & & \\ -\beta_1 & 1 & & & \\ -\beta_2 & -\beta_1 & 1 & & \\ \vdots & \vdots & \vdots & \ddots & \\ -\beta_{n-1} & -\beta_{n-2} & \cdots & \cdots & 1 \end{pmatrix}$$

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

Requires $|B(\omega)|$ **small**

sensitivity

For ϱ-zCDP, take
$$\mathbb{E}[z_t^2] = \frac{1}{2\rho} \max_t \left\|[B^{-1}]_{:,t}\right\|_2^2 = \frac{1}{2\rho}\int_{-\pi}^{\pi} \frac{\mathrm{d}\omega}{2\pi|B(\omega)|^2}$$

$$B = \begin{pmatrix} 1 & & & & \\ -\beta_1 & 1 & & & \\ -\beta_2 & -\beta_1 & 1 & & \\ \vdots & \vdots & \vdots & \ddots & \\ -\beta_{n-1} & -\beta_{n-2} & \cdots & \cdots & 1 \end{pmatrix}$$

Requires $|B(\omega)|$ **large**

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

Requires $|B(\omega)|$ **small**

sensitivity

For ϱ-zCDP, take

$$\mathbb{E}[z_t^2] = \frac{1}{2\rho} \max_t \left\| [B^{-1}]_{:,t} \right\|_2^2 = \frac{1}{2\rho}\int_{-\pi}^{\pi} \frac{\mathrm{d}\omega}{2\pi|B(\omega)|^2}$$

$$B = \begin{pmatrix} 1 & & & & \\ -\beta_1 & 1 & & & \\ -\beta_2 & -\beta_1 & 1 & & \\ \vdots & \vdots & \vdots & \ddots & \\ -\beta_{n-1} & -\beta_{n-2} & \cdots & \cdots & 1 \end{pmatrix}$$

Requires $|B(\omega)|$ **large**

Optimizing for $|B(\omega)|$ gives the theorem

For linear regression:

$$\boldsymbol{\theta}'_{t+1} = \big(\boldsymbol{I} - \eta(\boldsymbol{x}_t \otimes \boldsymbol{x}_t)\big)\boldsymbol{\theta}'_t + \eta\,\xi_t \boldsymbol{x}_t - \eta \sum_{\tau=0}^{\infty} \beta_\tau \boldsymbol{w}_{t-\tau}\,. \tag{25}$$

Multiplicative
noise

$$\boldsymbol{\theta}'_{t+1} = \big(\boldsymbol{I} - \eta(\boldsymbol{x}_t \otimes \boldsymbol{x}_t)\big)\boldsymbol{\theta}'_t + \eta\, \xi_t \boldsymbol{x}_t - \eta \sum_{\tau=0}^{\infty} \beta_\tau \boldsymbol{w}_{t-\tau}\,. \tag{25}$$

Decomposition:

$$\begin{aligned}
\boldsymbol{\theta}^{(0)}_{t+1} &= (\boldsymbol{I} - \eta \boldsymbol{H})\boldsymbol{\theta}^{(0)}_t + \eta \xi_t \boldsymbol{x}_t - \eta \sum_{\tau=0}^{\infty} \beta_\tau \boldsymbol{w}_{t-k}\,, \\
\boldsymbol{\theta}^{(r)}_{t+1} &= (\boldsymbol{I} - \eta \boldsymbol{H})\boldsymbol{\theta}^{(r)}_t + \eta(\boldsymbol{H} - \boldsymbol{x}_t \otimes \boldsymbol{x}_t)\boldsymbol{\theta}^{(r-1)}_t \;\text{ for } r > 0\,, \\
\boldsymbol{\delta}^{(r)}_{t+1} &= (\boldsymbol{I} - \eta \boldsymbol{x}_t \otimes \boldsymbol{x}_t)\boldsymbol{\delta}^{(r)}_t + \eta(\boldsymbol{H} - \boldsymbol{x}_t \otimes \boldsymbol{x}_t)\boldsymbol{\theta}^{(r)}_t\,.
\end{aligned}$$

$$\boldsymbol{\theta}'_t = \textstyle\sum_{r=0}^{m} \boldsymbol{\theta}^{(r)}_t + \boldsymbol{\delta}^{(m)}_t.$$

Aguech, Moulines, Priouret. **On a Perturbation Approach for the Analysis of Stochastic Tracking Algorithms**. SIAM J. Control. Optim., 2000
Bach and Moulines. **Non-Strongly-Convex Smooth Stochastic Approximation with Convergence Rate *O(1/n)***. NeurIPS 2013.

$$\boldsymbol{\theta}'_{t+1} = \big(\boldsymbol{I} - \eta(\boldsymbol{x}_t \otimes \boldsymbol{x}_t)\big)\boldsymbol{\theta}'_t + \eta\, \xi_t \boldsymbol{x}_t - \eta \sum_{\tau=0}^{\infty} \beta_\tau \boldsymbol{w}_{t-\tau}\,. \tag{25}$$

Decomposition:

$$\boldsymbol{\theta}_{t+1}^{(0)} = (\boldsymbol{I} - \eta \boldsymbol{H})\boldsymbol{\theta}_t^{(0)} + \eta \xi_t \boldsymbol{x}_t - \eta \sum_{\tau=0}^{\infty} \beta_\tau \boldsymbol{w}_{t-k}\,,$$

$$\boldsymbol{\theta}_{t+1}^{(r)} = (\boldsymbol{I} - \eta \boldsymbol{H})\boldsymbol{\theta}_t^{(r)} + \eta(\boldsymbol{H} - \boldsymbol{x}_t \otimes \boldsymbol{x}_t)\boldsymbol{\theta}_t^{(r-1)} \;\; \text{for } r > 0\,,$$

$$\boldsymbol{\delta}_{t+1}^{(r)} = (\boldsymbol{I} - \eta \boldsymbol{x}_t \otimes \boldsymbol{x}_t)\boldsymbol{\delta}_t^{(r)} + \eta(\boldsymbol{H} - \boldsymbol{x}_t \otimes \boldsymbol{x}_t)\boldsymbol{\theta}_t^{(r)}\,.$$

$$\boldsymbol{\theta}'_t = \textstyle\sum_{r=0}^{m} \boldsymbol{\theta}_t^{(r)} + \boldsymbol{\delta}_t^{(m)}.$$

Aguech, Moulines, Priouret. **On a Perturbation Approach for the Analysis of Stochastic Tracking Algorithms**. SIAM J. Control. Optim., 2000
Bach and Moulines. **Non-Strongly-Convex Smooth Stochastic Approximation with Convergence Rate *O(1/n)***. NeurIPS 2013.

Key idea: $\mathbb{E}\left[\boldsymbol{\delta}_0^{(m)} \otimes \boldsymbol{\delta}_0^{(m)}\right] \to \mathbf{0}$ as $m \to \infty$.

Thus, $\|\boldsymbol{\theta}'_t\| \le \sum_{r=0}^{\infty} \left\|\boldsymbol{\theta}_t^{(r)}\right\|$

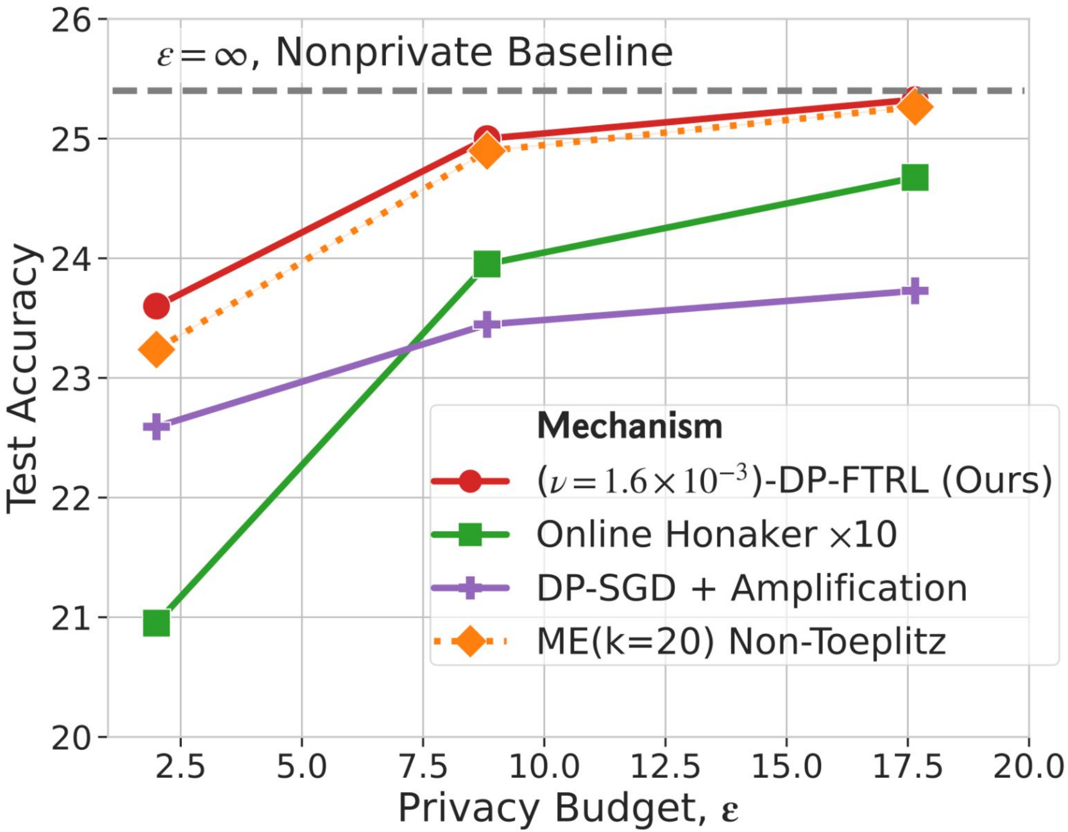
Experiments

***ν*-DP-FTRL**

For general problems, use $\beta_t = t^{-3/2}(1 - \nu)^t$

and tune the parameter ν

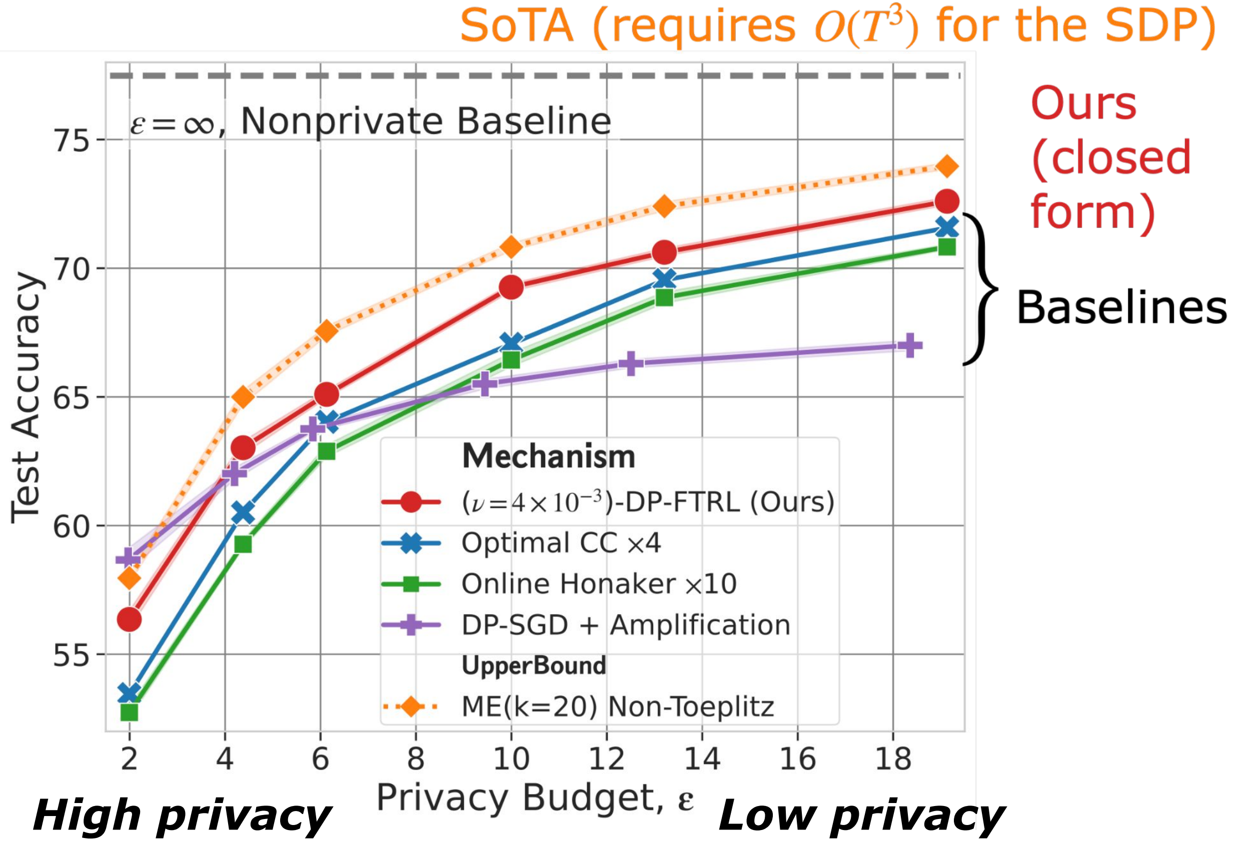
Language modeling with Stack Overflow | User-level DP



High privacy ***Low privacy***

Ours matches SoTA!

Image classification with CIFAR-10 | Example-level DP



Part 2: Efficient noise generation

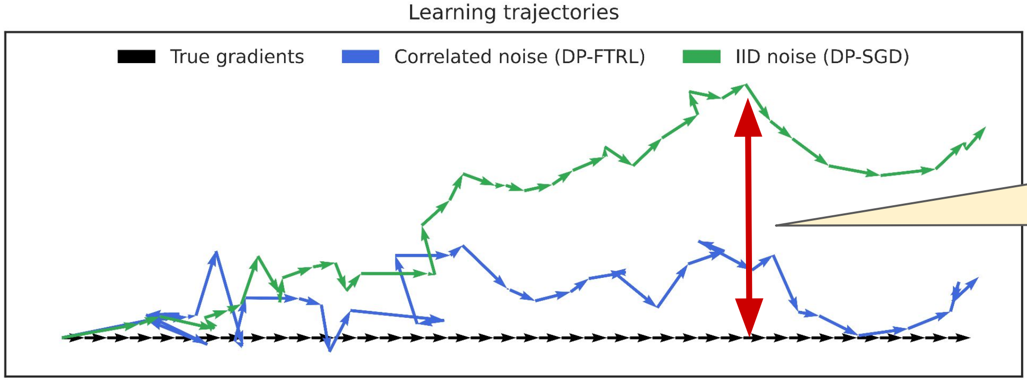
With near-optimal privacy-utility trade-offs

FOCS 2024

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

$\beta_t = t^{-3/2}$

Quadratic time complexity:
Noise generation requires $O(t)$ time in iteration t



Surrogate: max error

	Max Error	***Noise generation time*** (in iteration t)
Independent noise	$\Theta(\sqrt{n})$	$O(\mathrm{dim})$
Optimal correlated noise	$\frac{\log n}{\pi} + c$	$O(t \cdot \mathrm{dim})$

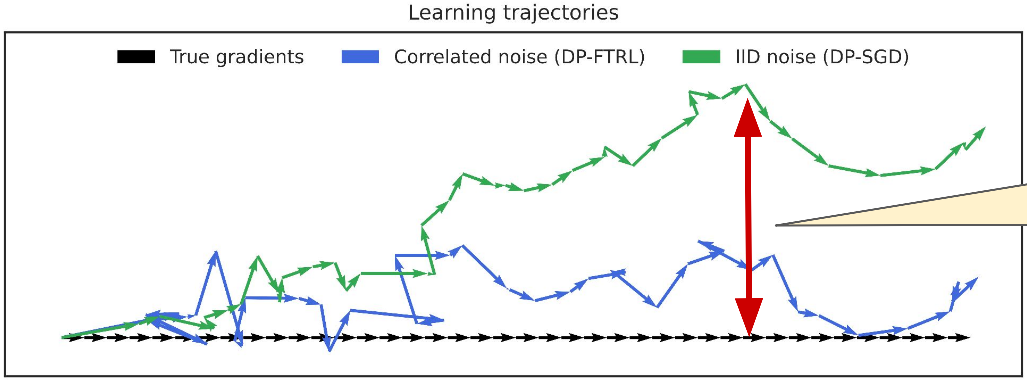
A first attempt: the banded mechanism

Set $\beta_t = 0$ for $t > b$

Then, we only have to sum *b* terms in $\sum_{\tau=1}^{b} \beta_\tau z_{t-\tau}$

Linear complexity:
Noise generation requires $O(b)$ time in each iteration

Choquette-Choo, Ganesh, McKenna, McMahan, Rush, Thakurta, Xu.
(Amplified) Banded Matrix Factorization: A unified approach to private training. NeurIPS 2023



Surrogate: max error

	Max Error	***Noise generation time*** (in iteration *t*)
Independent noise	$\Theta(\sqrt{n})$	$O(\text{dim})$
Optimal correlated noise	$\frac{\log n}{\pi} + c$	$O(t \cdot \text{dim})$
b-***Banded***	$O\left((\sqrt{n/b} - 1) \log b\right)$	$O(b \cdot \text{dim})$

Kalinin and Lampert. **Banded Square Root Matrix Factorization for Differentially Private Model Training**. NeurIPS 2024

Our approach: Intuition

Consider an exponentially decaying sequence $\beta_t = \alpha\, \lambda^{t-1}$.

Then, we can compute the correlated noise $w_t = \sum_{\tau=1}^{t} \beta_\tau z_{t-\tau}$

using the recurrence $w_{t+1} = \alpha\, z_t + \lambda\, w_{t-1}$

Linear complexity:
Noise generation requires O(dim) time in each iteration

Our approach: Intuition

Consider an exponentially decaying sequence $\beta_t = \alpha\, \lambda^{t-1}$.

Then, we can compute the correlated noise $w_t = \sum_{\tau=1}^{t} \beta_\tau z_{t-\tau}$

using the recurrence $w_{t+1} = \alpha\, z_t + \lambda\; w_{t-1}$

Linear complexity:
Noise generation requires O(dim) time in each iteration

Our approach: Intuition

Consider sums of exponentials: $\beta_t = \alpha_1 \lambda_1^{t-1} + \alpha_2 \lambda_2^{t-1}$

Then, we can compute the correlated noise $w_t = \sum_{\tau=1}^{t} \beta_\tau z_{t-\tau}$

using the recurrences

$$s_{t+1}^{(1)} = z_t + \lambda_1 z_{t-1} + \cdots = s_t^{(1)} + \lambda_1 z_t$$

Our approach: Intuition

Consider sums of exponentials: $\beta_t = \alpha_1 \lambda_1^{t-1} + \alpha_2 \lambda_2^{t-1}$

Then, we can compute the correlated noise $w_t = \sum_{\tau=1}^{t} \beta_\tau z_{t-\tau}$

using the recurrences

$$s_{t+1}^{(1)} = z_t + \lambda_1 z_{t-1} + \cdots = s_t^{(1)} + \lambda_1 z_t$$
$$s_{t+1}^{(2)} = z_t + \lambda_2 z_{t-1} + \cdots = s_t^{(2)} + \lambda_2 z_t$$

Our approach: Intuition

Consider sums of exponentials: $\beta_t = \alpha_1 \lambda_1^{t-1} + \alpha_2 \lambda_2^{t-1}$

Then, we can compute the correlated noise $w_t = \sum_{\tau=1}^{t} \beta_\tau z_{t-\tau}$

using the recurrences

$$s_{t+1}^{(1)} = z_t + \lambda_1 z_{t-1} + \cdots = s_t^{(1)} + \lambda_1 z_t$$
$$s_{t+1}^{(2)} = z_t + \lambda_2 z_{t-1} + \cdots = s_t^{(2)} + \lambda_2 z_t$$
$$w_{t+1} = \alpha_1 \, s_{t+1}^{(1)} + \alpha_2 \, s_{t+1}^{(2)}$$

Our approach: Intuition

Consider sums of exponentials: $\beta_t = \alpha_1 \lambda_1^{t-1} + \alpha_2 \lambda_2^{t-1}$

Then, we can compute the correlated noise $w_t = \sum_{\tau=1}^{t} \beta_\tau z_{t-\tau}$

using the recurrences

$$s_{t+1}^{(1)} = z_t + \lambda_1 z_{t-1} + \cdots = s_t^{(1)} + \lambda_1 z_t$$
$$s_{t+1}^{(2)} = z_t + \lambda_2 z_{t-1} + \cdots = s_t^{(2)} + \lambda_2 z_t$$
$$w_{t+1} = \alpha_1 \, s_{t+1}^{(1)} + \alpha_2 \, s_{t+1}^{(2)}$$

Linear time + **space**

Our approach: **B**uffered **L**inear **T**oeplitz (BLT) Mechanism

Approximate the optimal noise coefficients with *d* exponentials as

$$\beta_t \approx \beta'_t = \sum_{i=1}^{d} \alpha_i \lambda_i^{t-1}$$

Time & space complexity:
O(*d* x dimension)

Our approach: **B**uffered **L**inear **T**oeplitz (BLT) Mechanism

	Max Error	***Noise generation time*** (in iteration *t*)
Independent noise	$\Theta(\sqrt{n})$	$O(\mathrm{dim})$
Optimal correlated noise	$\frac{\log n}{\pi} + c$	$O(t \cdot \mathrm{dim})$
b-***Banded***	$O\Big((\sqrt{n/b} - 1)\log b\Big)$	$O(b \cdot \mathrm{dim})$
BLT of degree *d*	***??????***	$O(d \cdot \mathrm{dim})$

Our approach: **B**uffered **L**inear **T**oeplitz (BLT) Mechanism

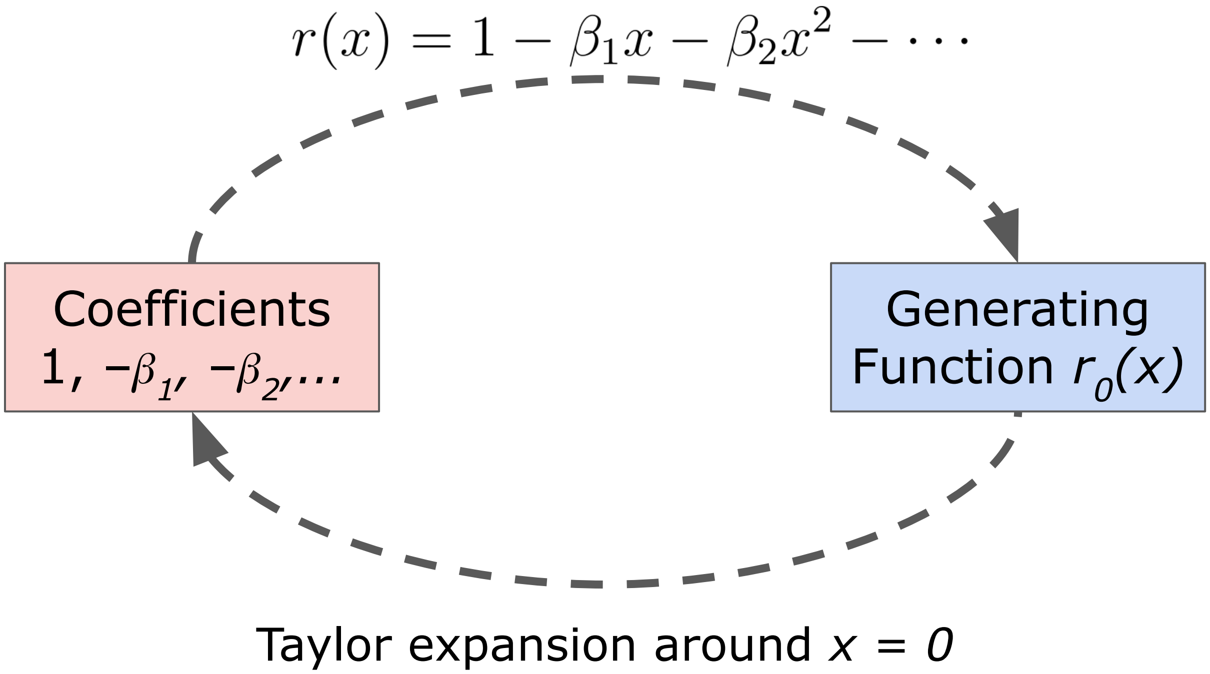
	Max Error	***Noise generation time*** (in iteration *t*)
Independent noise	$\Theta(\sqrt{n})$	$O(\mathrm{dim})$
Optimal correlated noise	$\frac{\log n}{\pi} + c$	$O(t \cdot \mathrm{dim})$
b-***Banded***	$O\Big((\sqrt{n/b} - 1)\log b\Big)$	$O(b \cdot \mathrm{dim})$
BLT of degree *d*	***??????***	$O(d \cdot \mathrm{dim})$

Approximation Theory!!

From sequences to functions

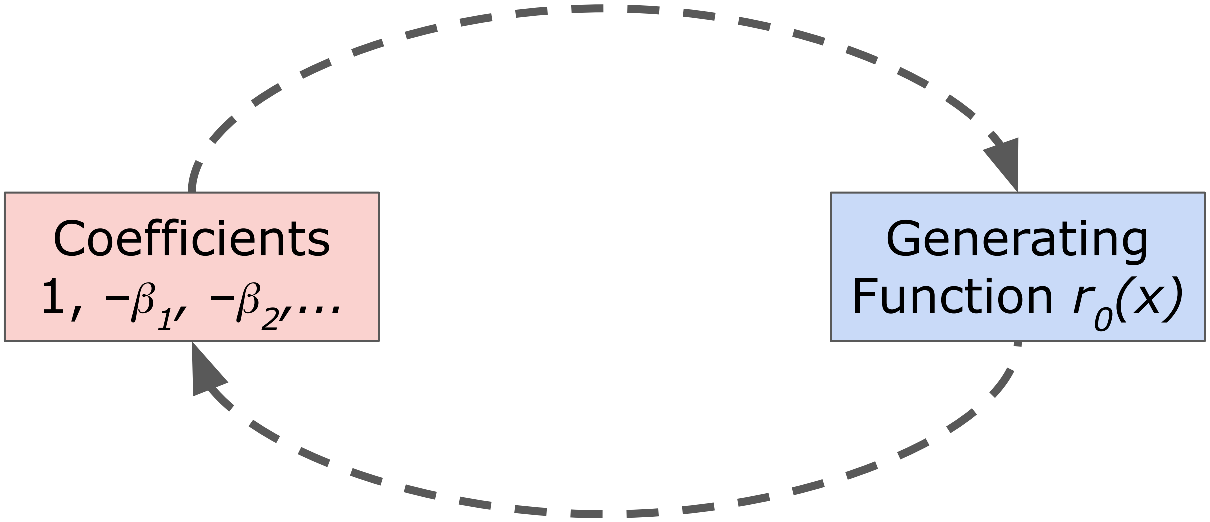
$$r(x) = 1 - \beta_1 x - \beta_2 x^2 - \cdots$$

From sequences to functions



From sequences to functions

$$r(x) = 1 - \beta_1 x - \beta_2 x^2 - \cdots$$



Taylor expansion around $x = 0$

Coefficients $\beta_t = \Theta(t^{-3/2})$ $\Leftrightarrow$ generating function $r_0(x) = (1 - x)^{1/2}$

From sequences to functions

$$r(x) = 1 - \beta_1 x - \beta_2 x^2 - \cdots$$

Coefficients
1, $-\beta_1$, $-\beta_2$,...

Generating Function $r_0(x)$

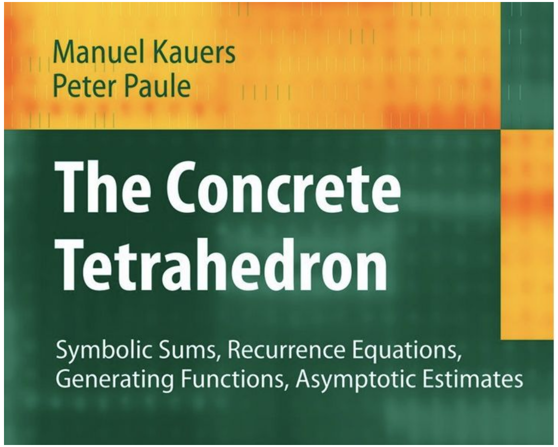
Taylor expansion around $x = 0$

BLT

$$\beta_t = \sum_{i=1}^{d} \alpha_i \lambda_i^{t-1}$$

????

BLT generating functions



Theorem (Informal):

The following properties are equivalent:

1. β's are a (complex) BLT sequence: $$\beta_t = \sum_{i=1}^{d} \alpha_i \lambda_i^{t-1}$$

1. Its generating function $r(x)$ is a **rational function** of degree d

1. β's satisfy a linear recurrence $\beta_t = \sum_{i=1}^{d} q_i \, \beta_{t-i}$

From functions to efficient noise generation

Generating Function
$r_0(x) = (1 - x)^{1/2}$

How good is it?

Rational approximation
$r(x) \approx (1 - x)^{1/2}$

Theorem [Dvijotham, McMahan, ***P***., Steinke, Thakurta 2024]

The max error of a sequence (β_t) with generating function $r(x)$ is

$$\mathcal{E}(\beta) \leq \frac{\log n}{\pi} + O(n \cdot \mathrm{err}(r))$$

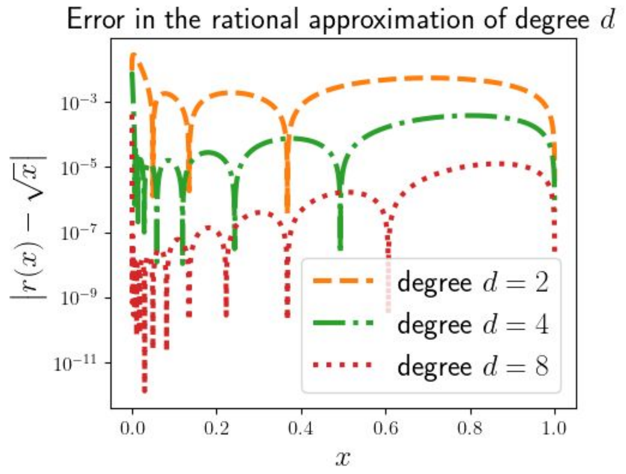
where err(r) quantifies the **approximation quality**

$$\mathrm{err}(r) = \max_{x \in \mathbb{C} : |x| = 1 - n^{-1}} \left| r(x) - \sqrt{1-x} \right|$$

There exists a degree-*d* rational function that satisfies the tight approximation bound:

$$\sup_{x \in [0,1]} |r(x) - \sqrt{x}| \leq 3 \cdot \exp(-\sqrt{d}).$$

Newman. **Rational approximation to |*x*|**. Michigan Math. J. (1964)



where err(*r*) quantifies the **approximation quality**

$$\mathrm{err}(r) = \max_{x \in \mathbb{C}\,:\,|x| = 1 - n^{-1}} \left| r(x) - \sqrt{1-x} \right|$$

Our approach: **B**uffered **L**inear **T**oeplitz (BLT) Mechanism

	Max Error	***Noise generation time*** (in iteration *t*)
Independent noise	$\Theta(\sqrt{n})$	$O(\mathrm{dim})$
Optimal correlated noise	$\frac{\log n}{\pi} + c$	$O(t \cdot \mathrm{dim})$
b-***Banded***	$O\Big((\sqrt{n/b} - 1)\log b\Big)$	$O(b \cdot \mathrm{dim})$
BLT of degree *d*	$\frac{\log n}{\pi} + O\big(n \cdot \exp(-\sqrt{d})\big)$	$O(d \cdot \mathrm{dim})$

Suffices to take $d = O(\log^2 n)$!

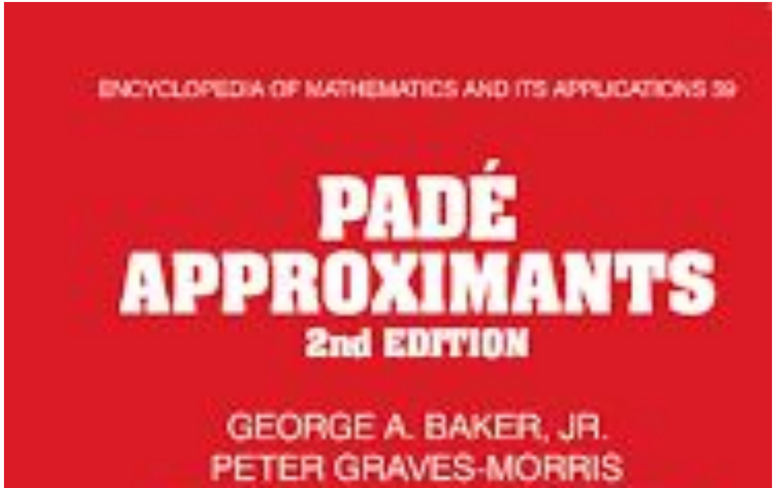
Our approach: **B**uffered **L**inear **T**oeplitz (BLT) Mechanism

	Error	***Noise generation time*** (*n*: # iterations)
Independent	$\Theta(\sqrt{n})$	$O(\mathrm{dim})$
Optimal correlated	$\frac{\log n}{\pi} + c$	$O(n \cdot \mathrm{dim})$
Ours	$\frac{\log n}{\pi} + c$	$O(\log^2(n) \cdot \mathrm{dim})$

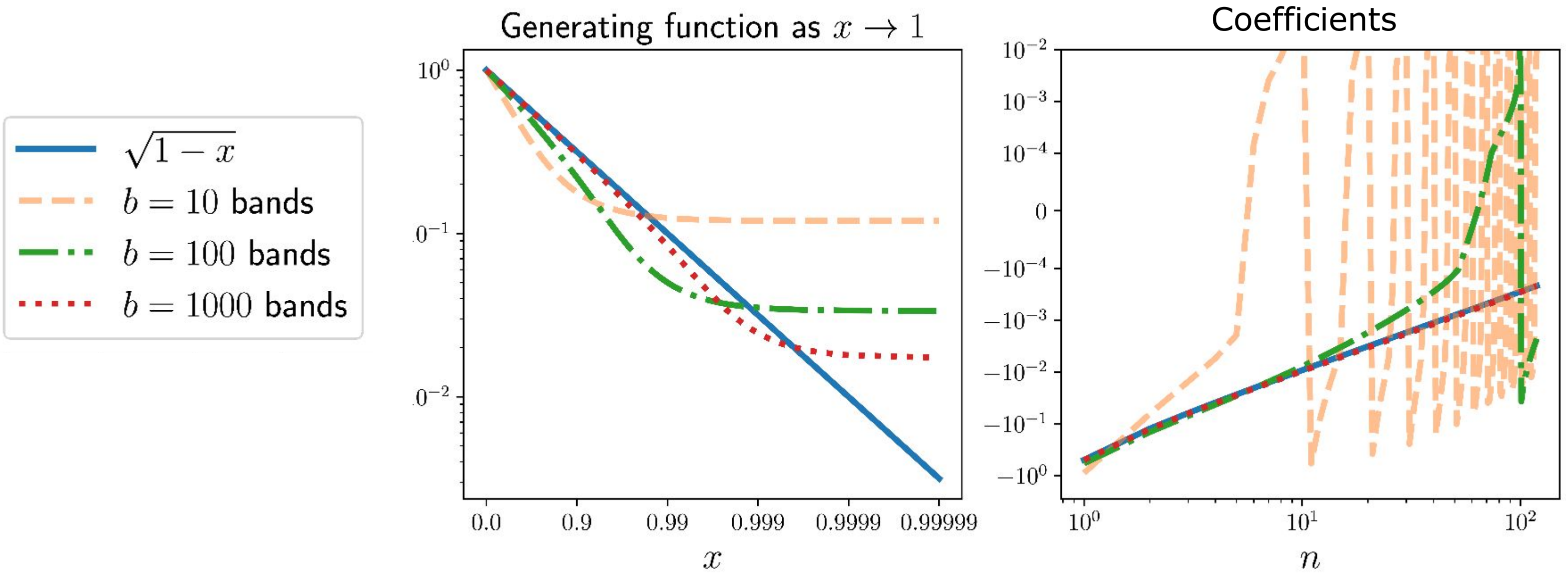
Key difference: approximation quality

Banded: Set $\beta_t = 0$ for $t > b$ $\Rightarrow$ polynomial approximation

BLT: $\beta_t = \sum_{i=1}^{d} \alpha_i \lambda_i^{t-1}$ $\Rightarrow$ rational approximation

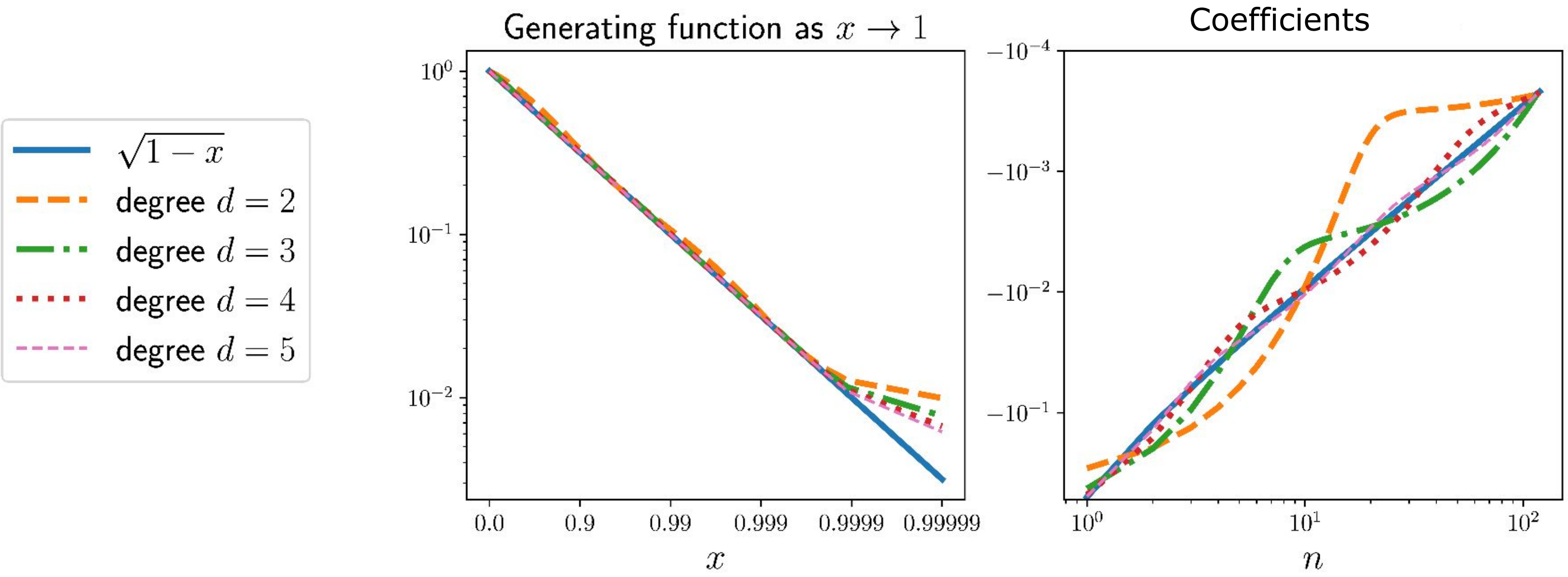


Approximation quality: banded mechanism



Note: here, we use a polynomial approximation to 1 / (1 — x)$^{1/2}$ rather than (1 — x)$^{1/2}$

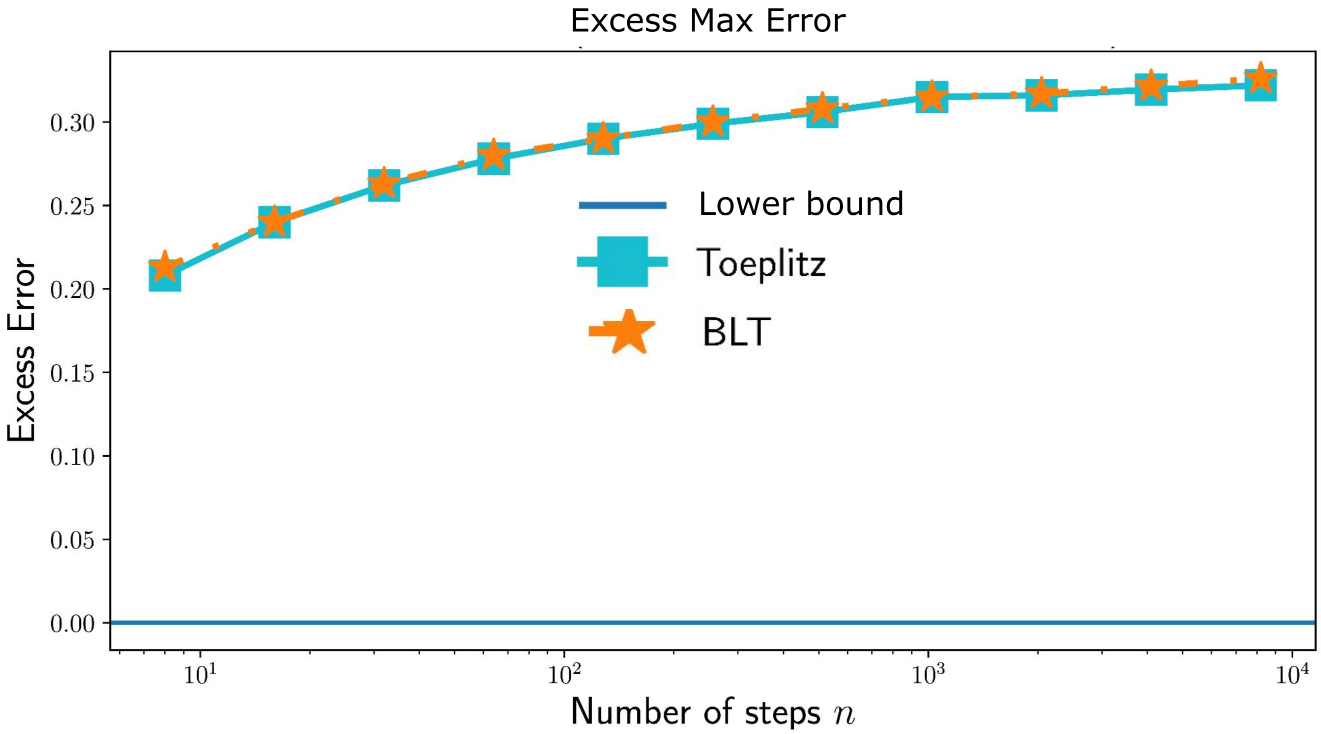
Approximation quality: BLT mechanism



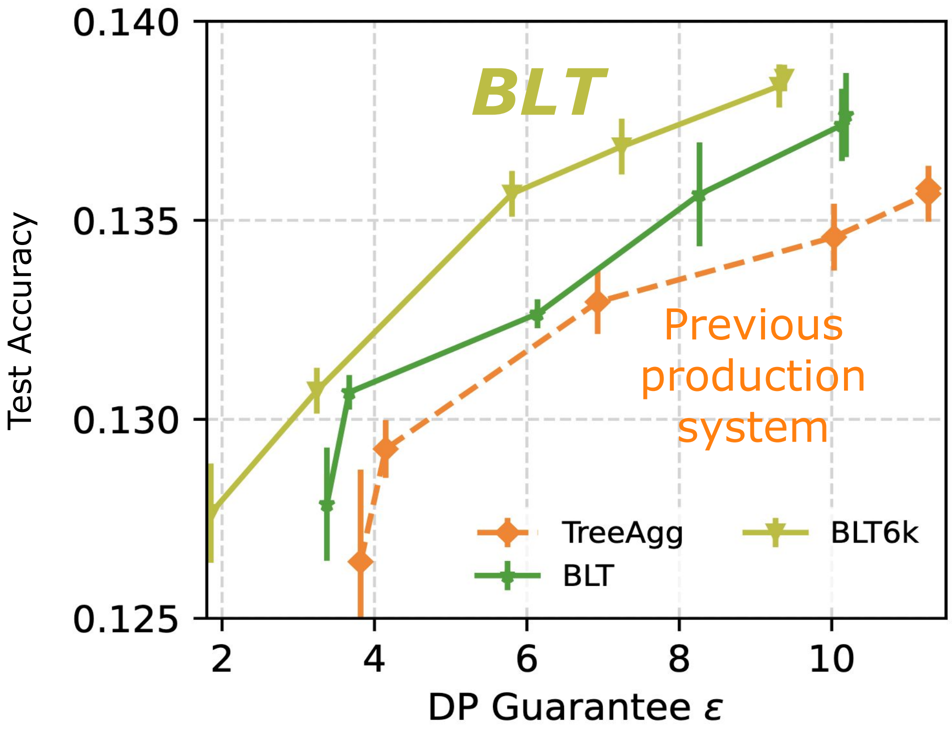
Note: BLT approximation of 1 / $(1 - x)^{1/2}$ $\Leftrightarrow$ BLT approximation of $(1 - x)^{1/2}$

[McMahan and ***P***. (2025)]

Empirical Results



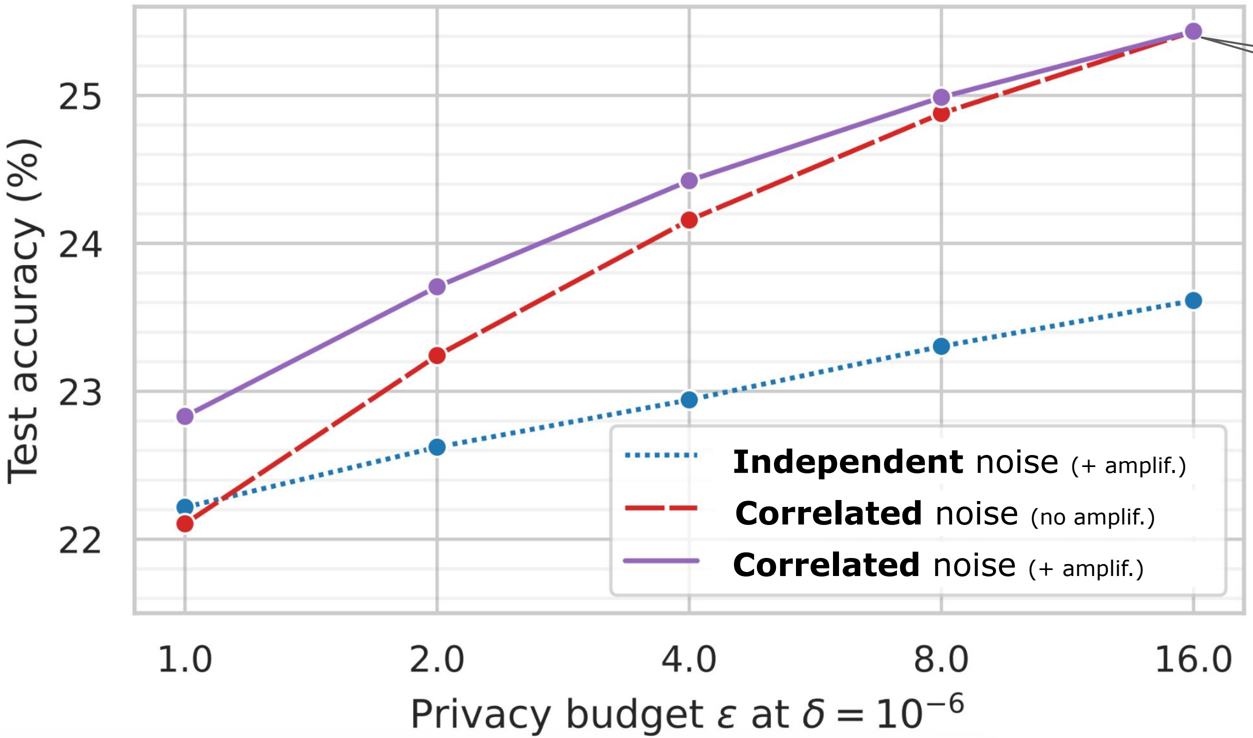
Practical Impact:
Google's production language model (Portuguese)



Plot: McMahan, Xu, Zhang (2024)

Conclusion and Open Problems

Goal: Better privacy-utility trade-offs



Correlated noise
uniformly beats
independent noise

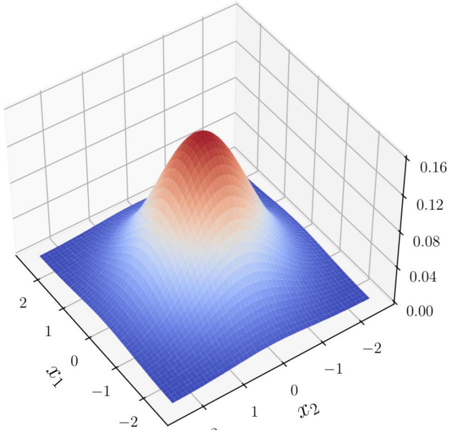
Choquette-Choo, Ganesh, McKenna, McMahan, Rush, Thakurta, Xu.
(Amplified) Banded Matrix Factorization: A unified approach to private training. NeurIPS 2023

Part 1: Correlated noise algorithms ***are*** provably better

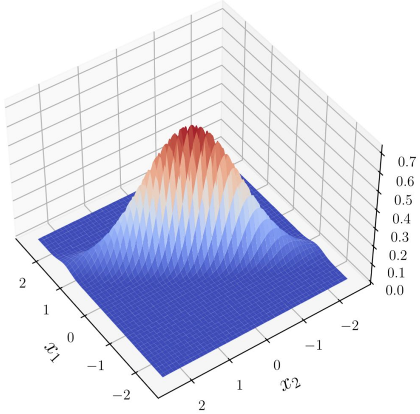
(Anti-) correlated noise provably beats independent noise

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

Independent noise	$\Theta(d)$
Correlated noise	$\tilde{O}(d_{\mathrm{eff}})$
Lower bound	$\Omega(d_{\mathrm{eff}})$



High effective dimension



Low effective dimension

Part 2: Near-optimal noise generation time complexity

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

Our approach: Approximate the noise coefficients as

$$\beta_t \approx \beta'_t = \sum_{i=1}^{d} \alpha_i \lambda_i^{t-1}$$

Time & space:
$O(d$ x dimension)

Error: If $d = O(\log^2(n/c))$, then the error is $\mathcal{E}(\beta') \leq \mathcal{E}(\beta) + c$

(n = Number of steps)

Coming soon: Survey/tutorial on correlated noise mechanisms!

Contents

Open Problem: Continuous time limits

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

Proceedings of Machine Learning Research vol 195:1–44, 2023 36th Annual Conference on Learning Theory

Universality of Langevin Diffusion for Private Optimization, with Applications to Sampling from Rashomon Sets

Arun Ganesh ARUNGANESH@GOOGLE.COM
Google Research

Abhradeep Thakurta ATHAKURTA@GOOGLE.COM
Google DeepMind

Jalaj Upadhyay JALAJ.UPADHYAY@RUTGERS.EDU
Rutgers University

Precise analysis (better rates)

Algorithm design

Open Problem: Multi-epoch Learning Guarantees

***Assumptions in this talk*:**

Streaming setting: Suppose we draw a fresh data point $x_t \sim P$ in each iteration t (i.e. only 1 epoch)

Precise analysis

Better algorithms

Open Problem: Adaptive Gradient Algorithms

SGD update (without noise)

$$\theta_t - \theta_0 = -\boxed{\sum_{\tau=0}^{t-1} g_\tau}$$

Adam update (without noise)

$$v_t = (1-\beta_1)v_{t-1} + \beta_1 g_t$$
$$s_t = (1-\beta_2)s_{t-1} + \beta_2 g_t^2$$
$$\theta_{t+1} = \theta_t - \eta \boxed{\frac{v_t}{\sqrt{s_t} + \delta}}$$

Non-linear functions of the injected noise

Advertisement: MS/PhD Openings in my group at IIT Madras

- Areas of interest in ML/AI:
  - Privacy-preserving AI
  - Making (generative) AI more robust
  - Applications in healthcare + public good

- Flavour:
  - Theoretical foundations +
  - State of the art empirical performance +
  - Real-world applications

Thank you!

Future Work

Theory

- Averaged iterate analysis + precise finite time bounds
- Analysis for non-Toeplitz systems

Ruppert. **Efficient Estimations from a Slowly Convergent Robbins-Monro Process**. 1998

Polyak and Juditsky. **Acceleration of Stochastic Approximation by Averaging**. SIAM J Control Optim. (1992)

Jain, Kakade, Kidambi, Netrapalli, Sidford. **Parallelizing Stochastic Gradient Descent for Least Squares Regression: Mini-batching, Averaging, and Model Misspecification**. JMLR (2018).

DP-FTRL: privatize prefix sums of gradients

SGD update (without noise) $$\theta_t - \theta_0 = -\sum_{\tau=0}^{t-1} g_\tau$$

$$\begin{pmatrix} g_0 \\ g_0 + g_1 \\ \vdots \\ g_0 + \cdots + g_{n-1} \end{pmatrix} = \boxed{\begin{pmatrix} 1 & & & \\ 1 & 1 & & \\ \vdots & \vdots & \ddots & \\ 1 & 1 & \cdots & 1 \end{pmatrix}} \begin{pmatrix} g_0 \\ g_1 \\ \vdots \\ g_{n-1} \end{pmatrix}$$

Call this matrix as *A*

Gradient Descent with Linearly Correlated Noise: Theory and Applications to Differential Privacy

NeurIPS 2023

Anastasia Koloskova* EPFL, Switzerland

Ryan McKenna Google Research

Zachary Charles Google Research

Keith Rush Google Research

Brendan McMahan Google Research

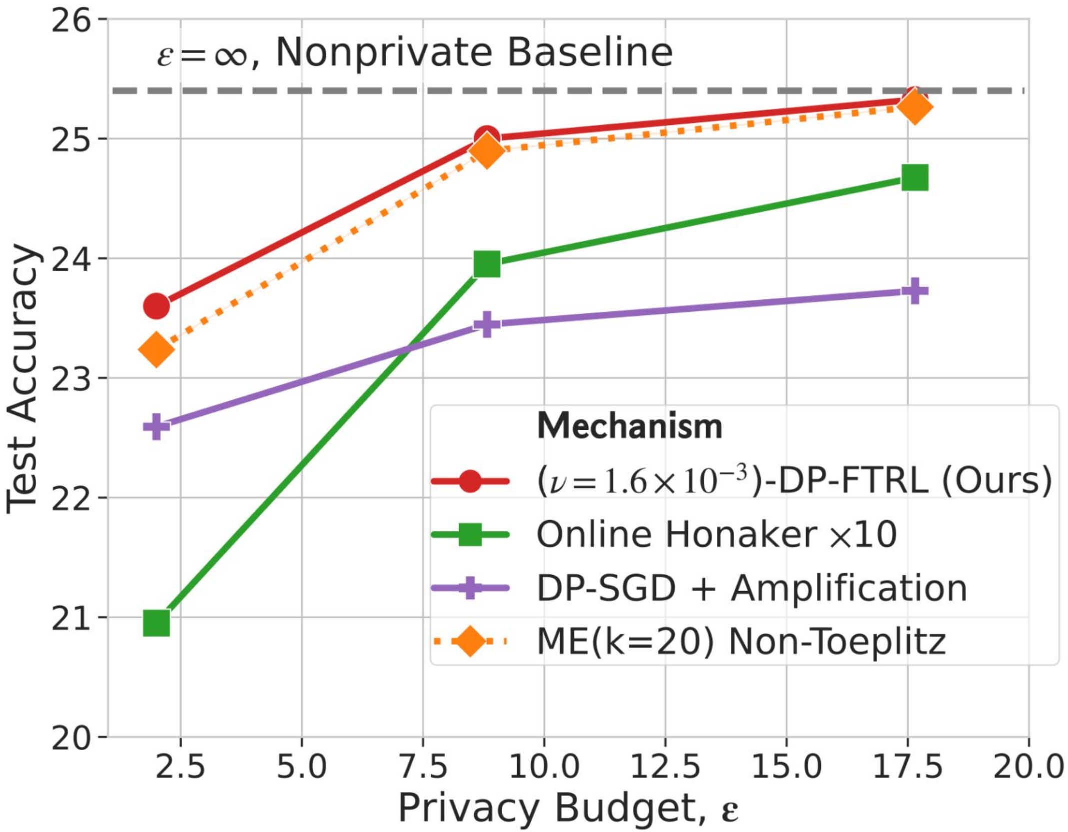
Theorem 4.7 (convex). *Under Assumptions 4.1, 4.2, and 4.3, if $\gamma \leq 1/4L$ and $\tau = \tilde{\Theta}(1/\gamma L)$, then* (7) *produces iterates with average error $(T+1)^{-1}\sum_{t=0}^{T}\mathbb{E}\left[f(\mathbf{x}_t) - f^\star\right]$ upper bounded by*

$$\tilde{\mathcal{O}}\left(\frac{\|\mathbf{x}_0 - \mathbf{x}^\star\|^2}{\gamma T} + \frac{\sigma^2}{TL\tau} \times \left[\frac{1}{\tau}\sum_{t=1}^{T}\left\|\mathbf{b}_t - \mathbf{b}_{\lfloor \frac{t}{\tau}\rfloor \tau}\right\|^2 + \sum_{\substack{1\leq t\leq T \\ t=0 \bmod \tau}}\|\mathbf{b}_t - \mathbf{b}_{t-\tau}\|^2 + \left\|\mathbf{b}_{\lfloor \frac{T}{\tau}\rfloor \tau}\right\|^2\right]\right).$$

Improved analysis of DP-FTRL

No provable gap between DP-SGD & DP-FTRL (same as previous)

Empirical results for private language modeling



Ours matches SoTA!