Correlated Noise Provably Beats Independent Noise for DP learning

ICLR 2024

Krishna Pillutla (Google Research → IIT Madras)

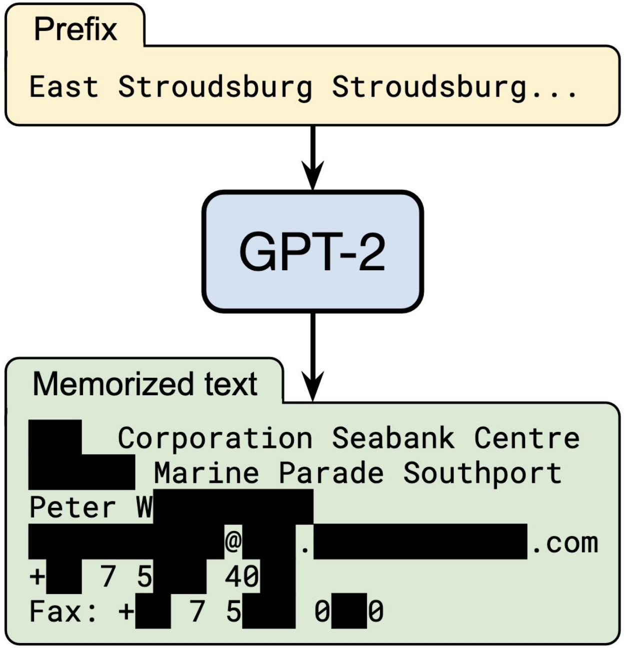
Presented at ***Laboratoire Jean Kuntzmann, UGA***

Joint work with Chris Choquette-Choo, Dj Dvijotham, Arun Ganesh, Thomas Steinke, Abhradeep Thakurta

Google Research

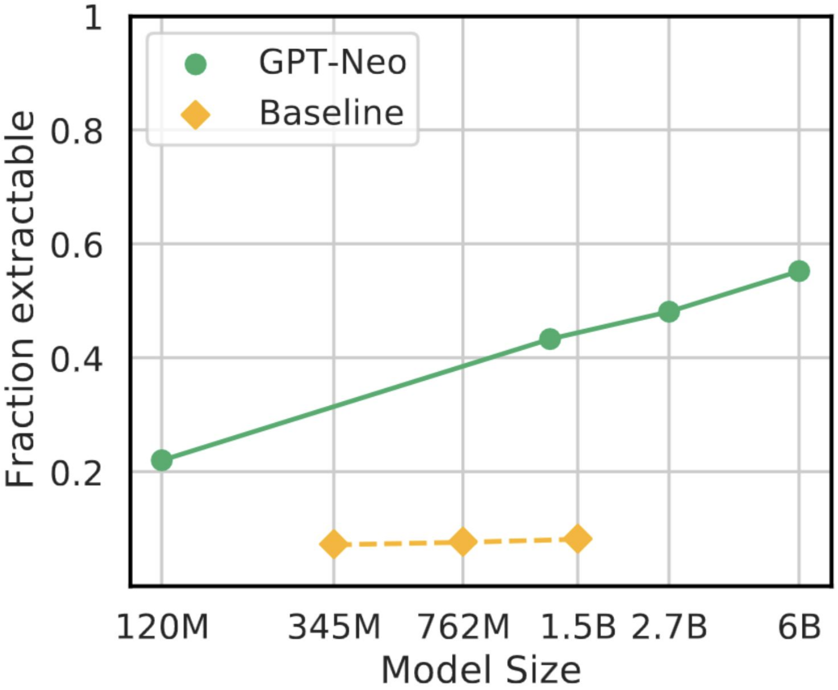
WHEN YOU TRAIN PREDICTIVE MODELS ON INPUT FROM YOUR USERS, IT CAN LEAK INFORMATION IN UNEXPECTED WAYS.

Models leak information about their training data

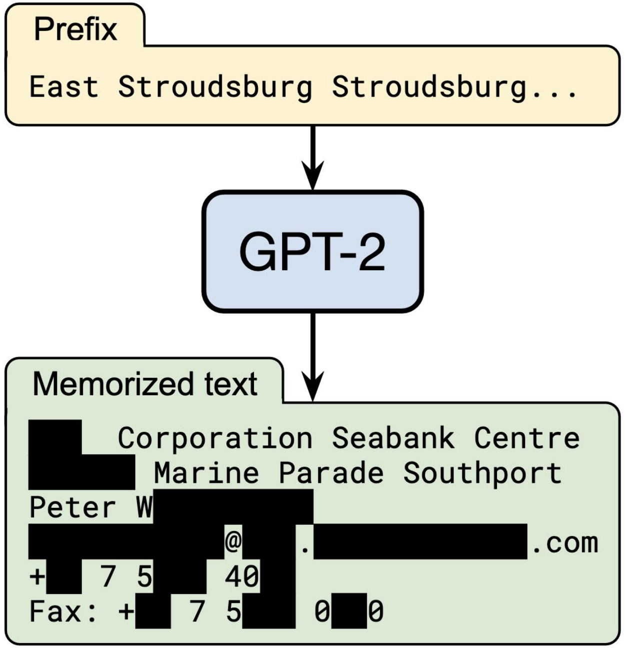


Carlini et al. (USENIX Security 2021)

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



Carlini et al. (ICLR 2023)



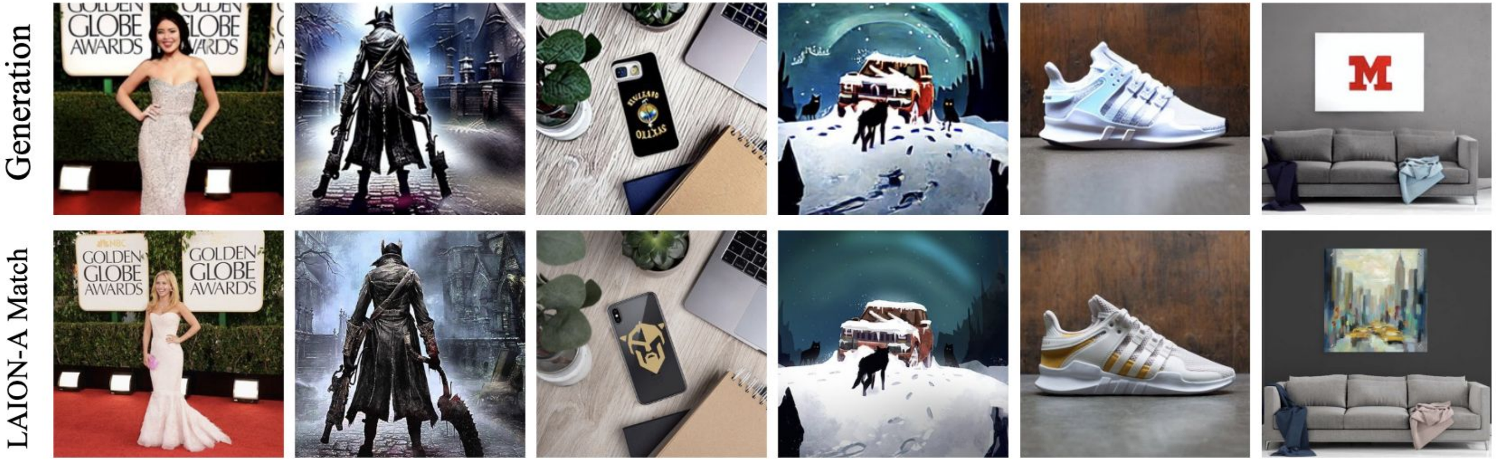
Carlini et al. (USENIX Security 2021)

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

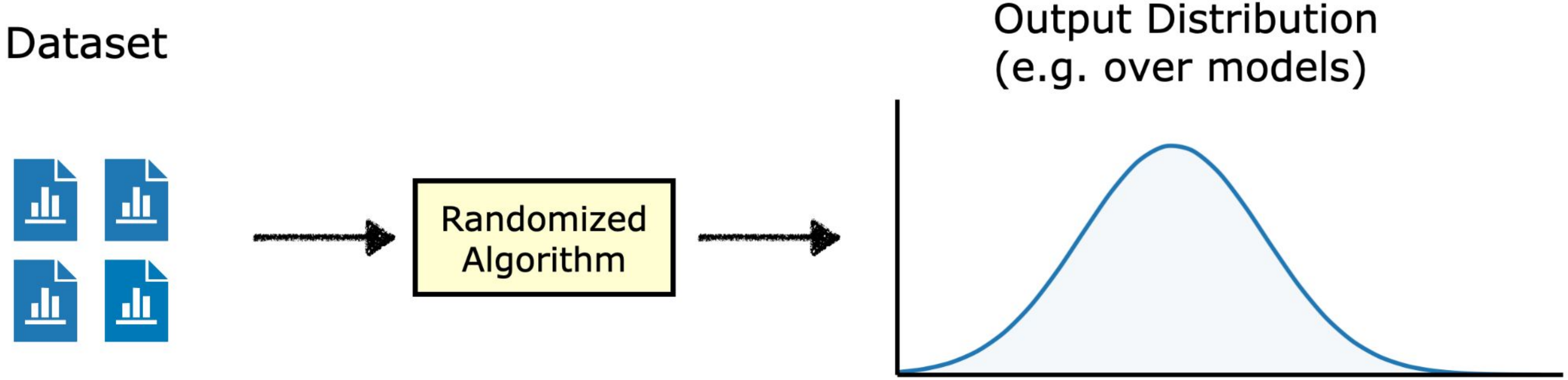
Gowthami Somepalli 🐢, Vasu Singla 🐢, Micah Goldblum 🗽, Jonas Geiping 🐢, Tom Goldstein 🐢

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

🗽 New York University
goldblum@nyu.edu

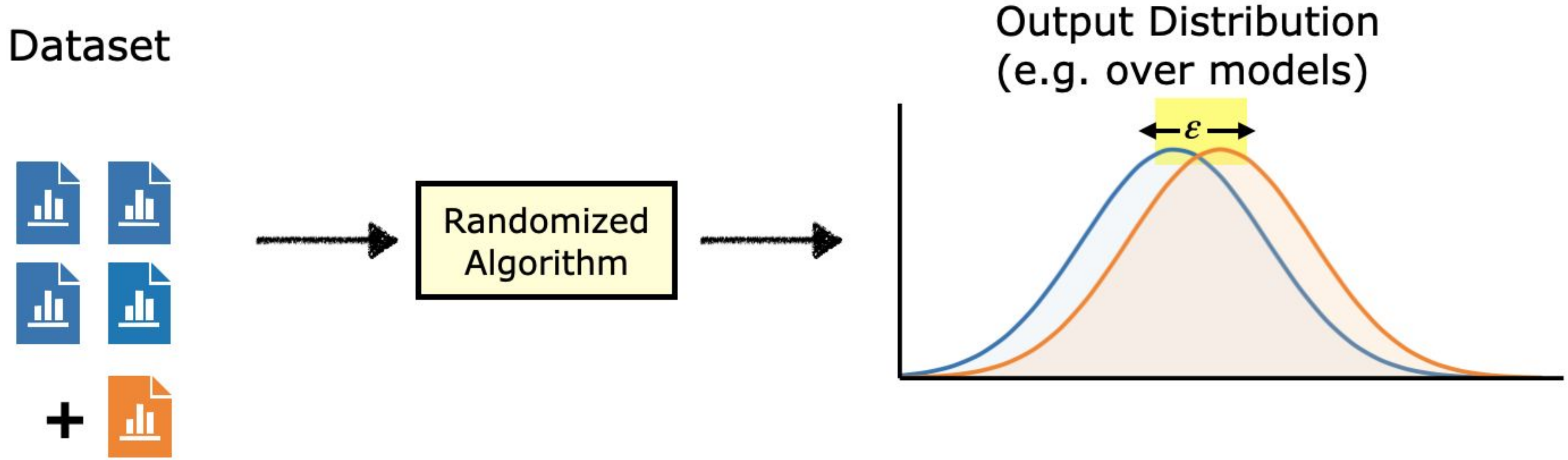


Differential privacy (DP)



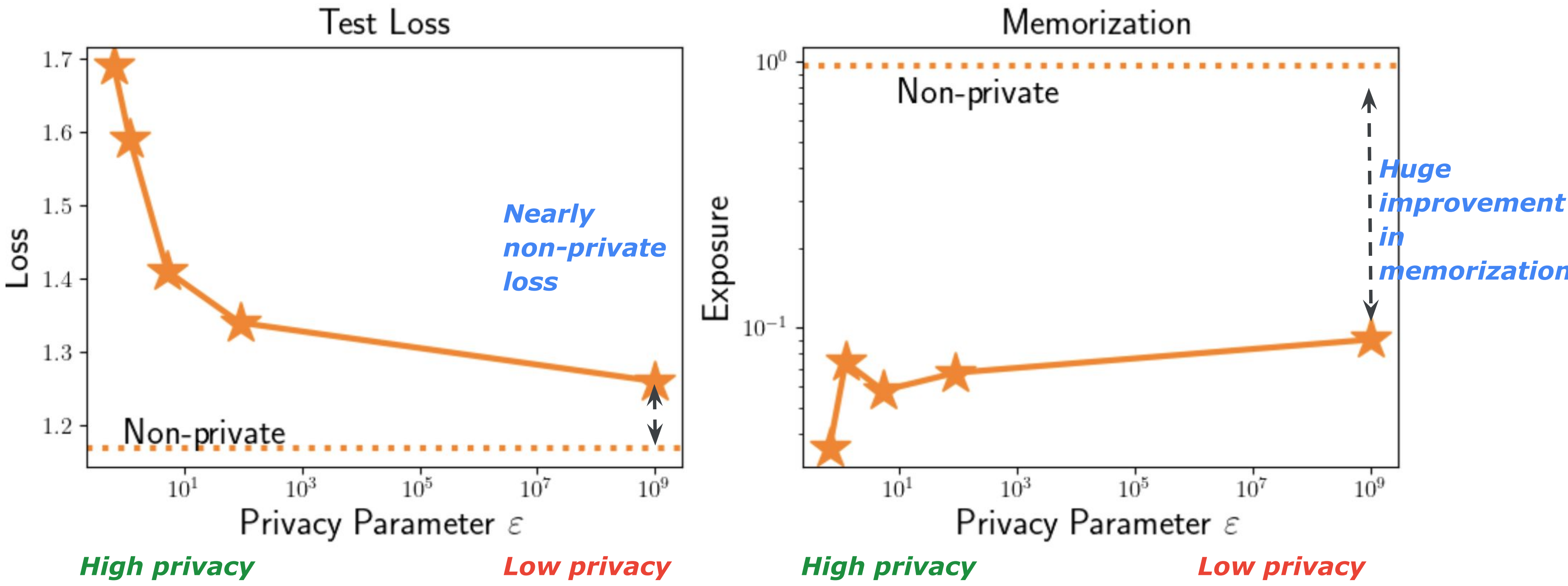
Dwork, McSherry, Nissim, Smith. **Calibrating noise to sensitivity in private data analysis**. TCC 2006

Differential privacy (DP)



A randomized algorithm is ε-**differentially private** if the addition of **one user's 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.

How do we train models with DP?

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

Model parameters

Loss function

Data

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

Gradient clipped to $\|g\| \leq G$

Independent Gaussian noise

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

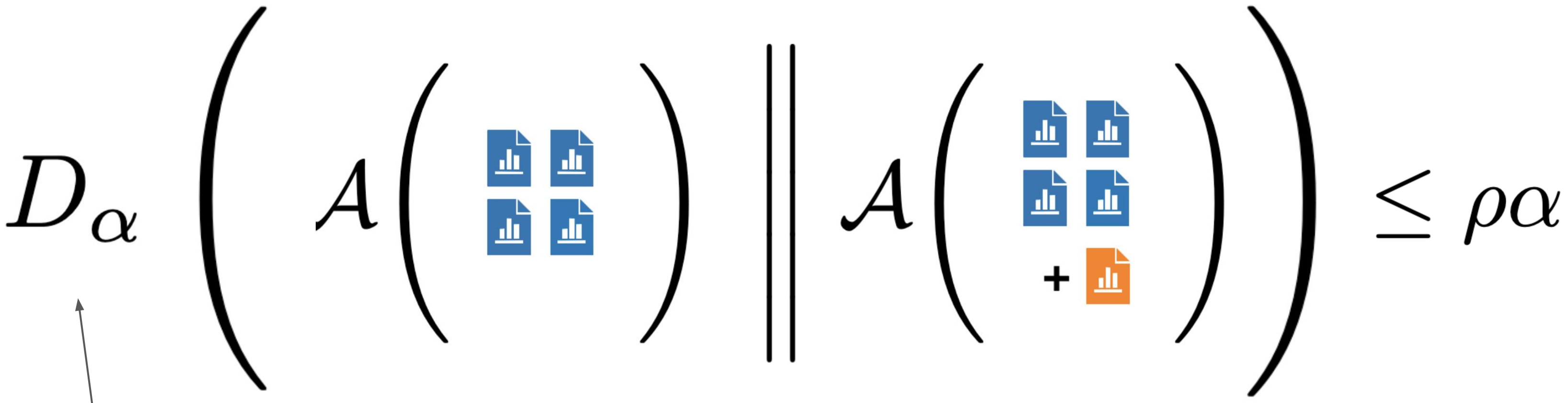
Learning rate

Google Research

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

Recall: ϱ-Zero-Concentrated DP (ϱ-zCDP)

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



Rényi α-divergence

Google Research

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

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

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

(G = gradient clip norm)

Independent Gaussian noise

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

Google Research

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

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

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

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

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.

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

For ϱ-zCDP, take noise variance = $\frac{G^2}{2\rho} \max_t \left\| [B^{-1}]_{:,t} \right\|_2^2$

sensitivity

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

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

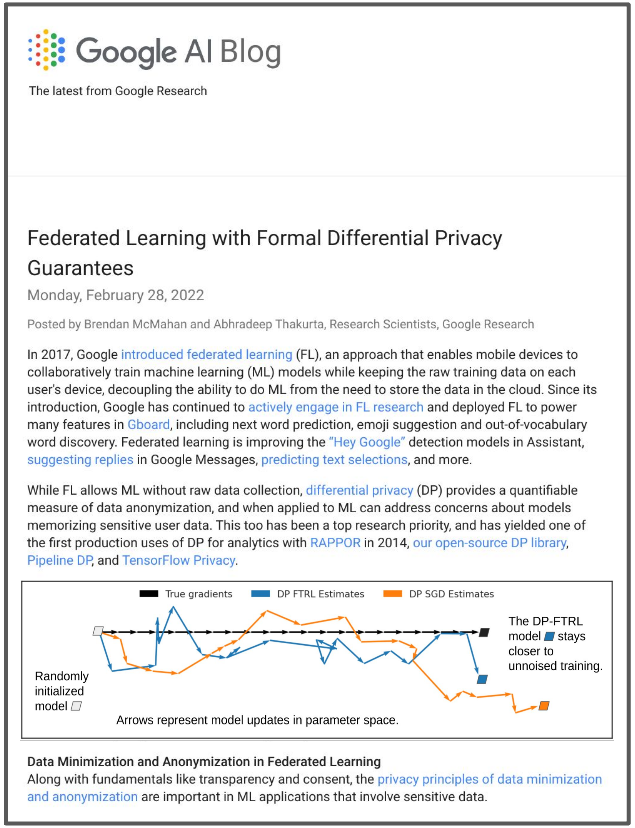
$$\theta_{t+1} = \theta_t - \eta \left(g_t + \sum_{\tau=0}^{t} \beta_{t,\tau} z_{t-\tau} \right)$$

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.

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.

Do we use independent or correlated noise?

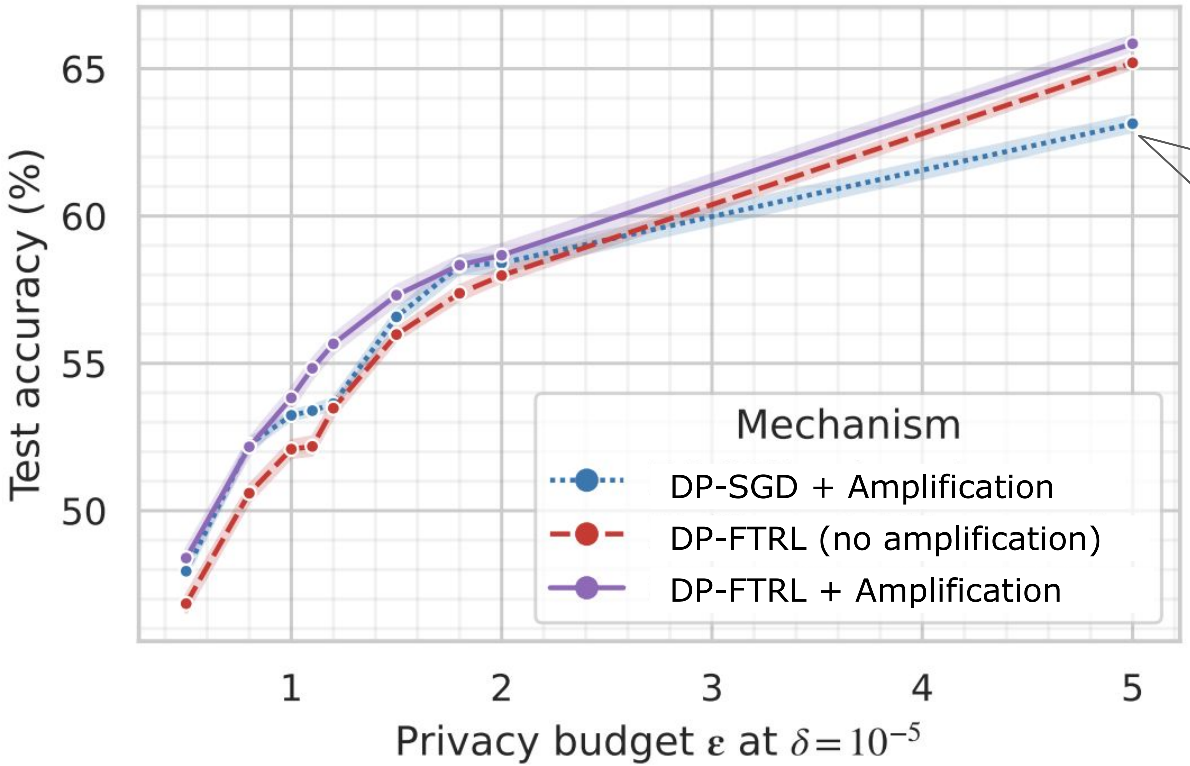
DP-SGD

DP-FTRL

Prior work: [Choquette-Choo et al. (NeurIPS '23)]

- (Empirically) correlated noise outperforms independent noise

Experiment:
DP learning with CIFAR-10



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

Google Research

Our contributions

Theory

- correlated noise is **provably** better

Google Research

Our contributions

What we show: For linear regression (without clipping) and learning rate η<1, the expected final error as $T\to\infty$ scales as

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

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

η: learning rate
ϱ: privacy level

Our contributions

Informal Theorem: For linear regression (without clipping) and learning rate η<1, the expected final error as $T \to \infty$ is

Independent noise (DP-SGD without clipping)	$\Theta(d \quad \rho^{-1} \eta)$
Correlated noise (DP-FTRL without clipping)	$\tilde{O}(d_{\text{eff}} \, \rho^{-1} \, \eta^2)$
Lower bound for any algorithm	$\Omega(d_{\text{eff}} \, \rho^{-1} \, \eta^2)$

Matches lower bound (upto polylog factors)

η: learning rate
ϱ: privacy level

Google Research

Prior work: [Choquette-Choo et al. (NeurIPS ’23)]

- Solve a semi-definite program (SDP) to find these correlations
- Cubic complexity $O(T^3)$ in the number of iterations T

$$\min_{X \succeq 0} \left\{ \mathbf{Tr}(AX^{-1}A^\top) \;:\; \mathrm{diag}(X) = 1 \right\}$$

$$A = \begin{pmatrix} 1 & & & \\ 1 & 1 & & \\ \vdots & & & \\ 1 & 1 & \cdots & 1 \end{pmatrix}_{T \times T}$$

Google Research

Our contributions

Empirical:

- computationally much more efficient:
  cubic $O(T^3) \rightarrow$ linear $O(T)$

Our contributions

Empirical:

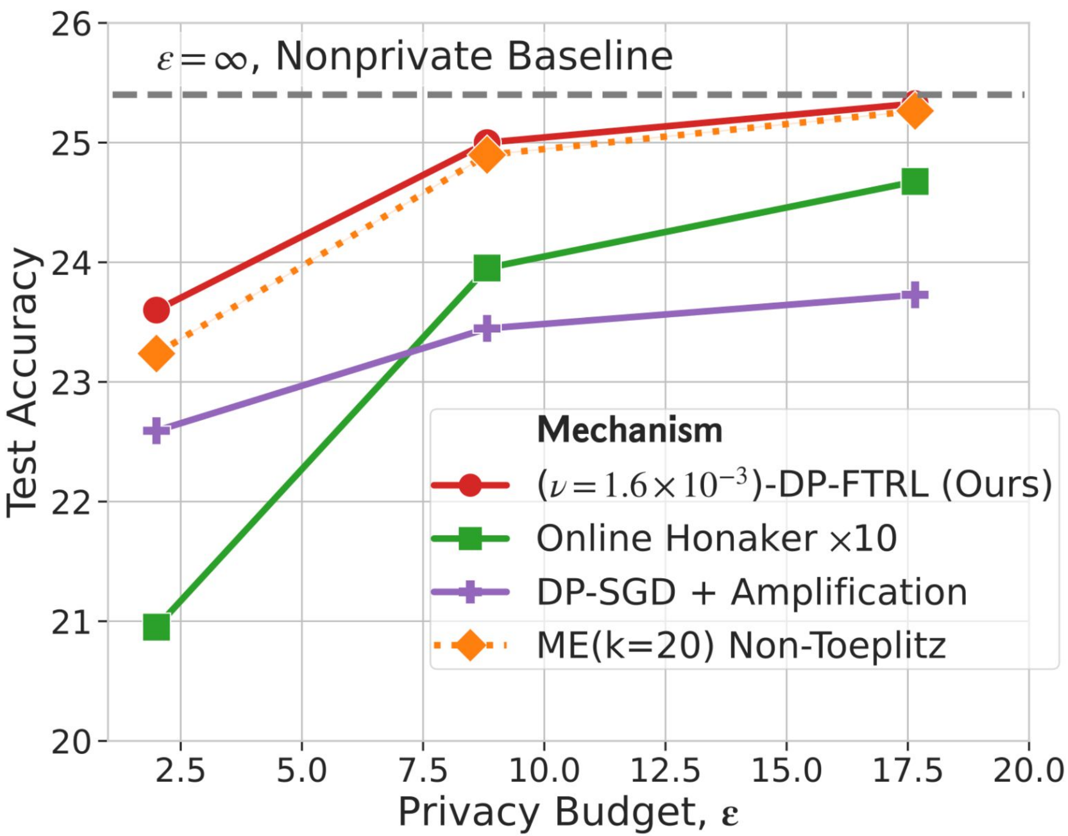
- computationally much more efficient:
  cubic $O(T^3) \rightarrow$ linear $O(T)$

Set $\beta_0 = 1, \quad \beta_\tau = -\tau^{-3/2}(1-\nu)^\tau$

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

The hyper-parameter ν is tuned

Empirical results for private deep learning



Ours matches SoTA!

Google Research

Outline

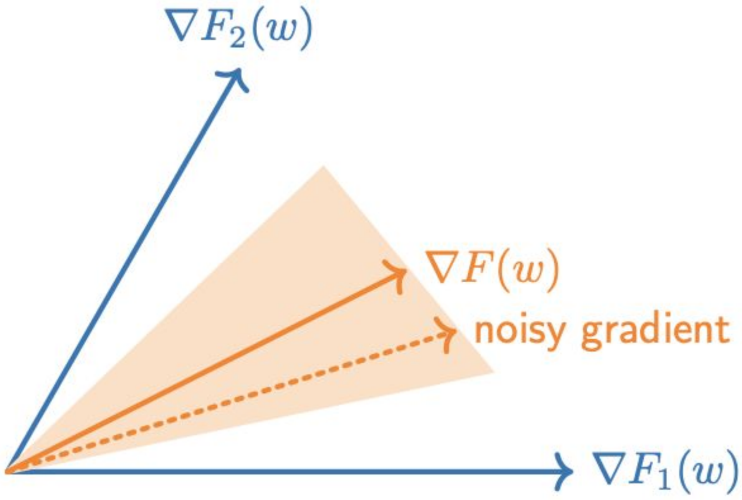
- Background
- Theoretical Results
- Empirical Results

Google Research

Outline

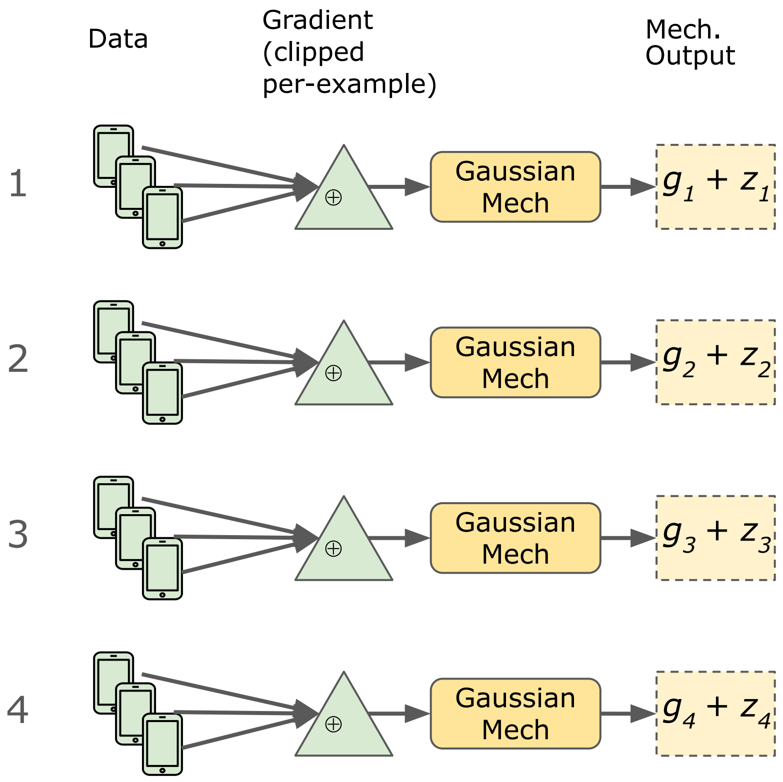
- **Background**
- Theoretical Results
- Empirical Results

DP-SGD's primitive: private mean estimation of minibatch (clipped) gradients in each iteration



Google Research

DP-SGD adds independent noise in each iteration



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

Why DP-FTRL?

DP-SGD requires privacy amplification by random sampling for good practical performance

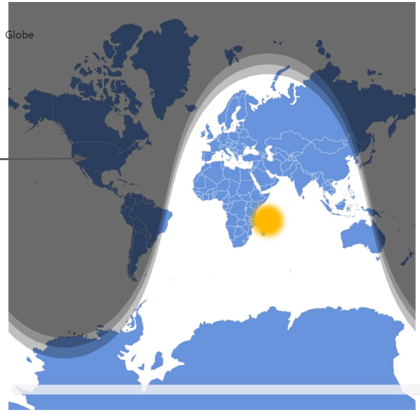
Google Research

Why DP-FTRL?

DP-SGD requires privacy amplification by random sampling for good practical performance

(Provable) Random sampling not possible in applications such as federated learning

Charging/WiFi required for federated learning (usually at *night*)



Google Research

DP-FTRL: privatize prefix sums of gradients

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

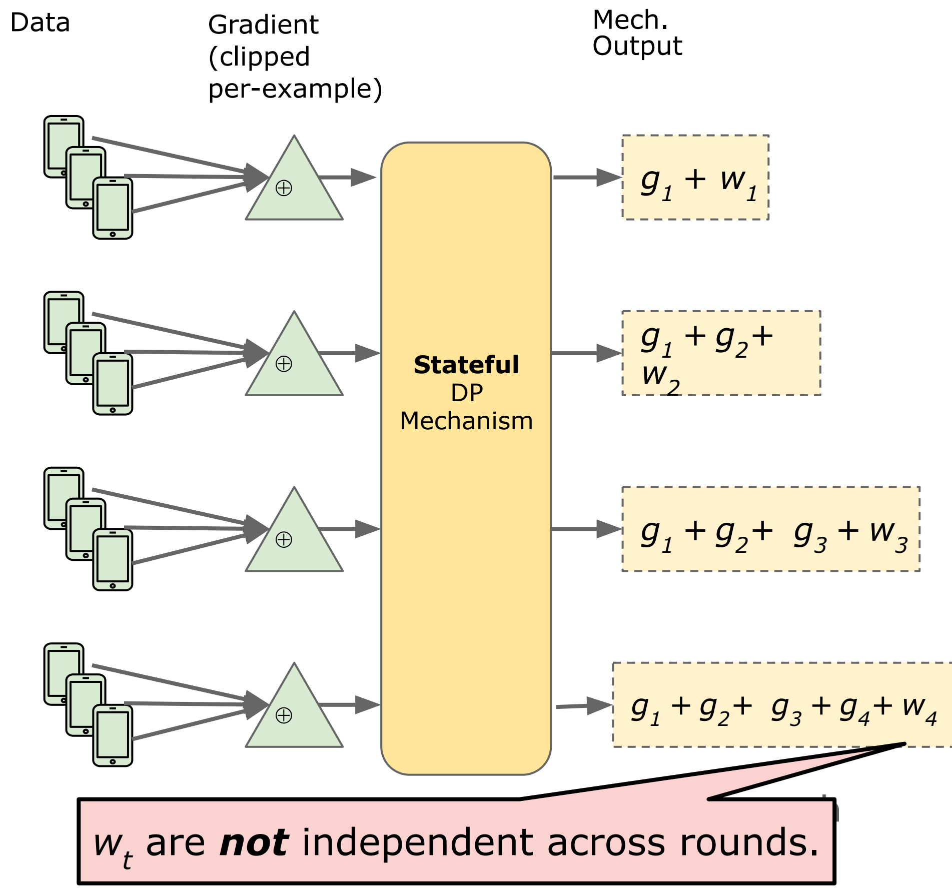
SGD update (without noise)

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

DP-FTRL: privatize prefix sums of gradients

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

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

Empirically, DP-FTRL (without amplification)
is competitive with DP-SGD + amplification

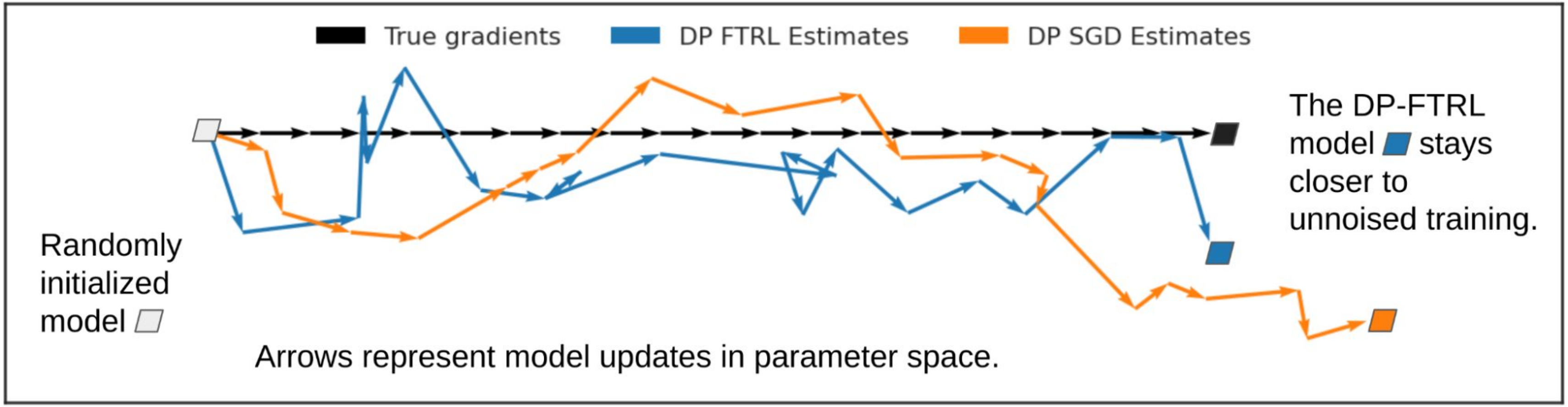


Figure: [Google AI Blog post]

Google Research

DP-FTRL in Equations

DP-FTRL: Incorporating Correlated Noise

$$-\begin{pmatrix} \theta_1 - \theta_0 \\ \theta_2 - \theta_1 \\ \vdots \\ \theta_t - \theta_{t-1} \end{pmatrix} = \begin{pmatrix} g_0 \\ g_1 \\ \vdots \\ g_{t-1} \end{pmatrix}$$

SGD update (without noise)

Google Research

DP-FTRL: Incorporating Correlated Noise

$$-\begin{pmatrix} \theta_1 - \theta_0 \\ \theta_2 - \theta_1 \\ \vdots \\ \theta_t - \theta_{t-1} \end{pmatrix} = \begin{pmatrix} g_0 \\ g_1 \\ \vdots \\ g_{t-1} \end{pmatrix} + \begin{pmatrix} z_0 \\ z_1 \\ \vdots \\ z_{t-1} \end{pmatrix}$$

DP-SGD update (with independent noise)

DP-FTRL: Incorporating Correlated Noise

Noise correlation matrix

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

$$-\begin{pmatrix} \theta_1 - \theta_0 \\ \theta_2 - \theta_1 \\ \vdots \\ \theta_t - \theta_{t-1} \end{pmatrix} = \begin{pmatrix} g_0 \\ g_1 \\ \vdots \\ g_{t-1} \end{pmatrix} + B \begin{pmatrix} z_0 \\ z_1 \\ \vdots \\ z_{t-1} \end{pmatrix}$$

DP-FTRL update (with correlated noise)

DP-FTRL: Incorporating Correlated Noise

Noise correlation matrix

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

$$-\begin{pmatrix} \theta_1 - \theta_0 \\ \theta_2 - \theta_1 \\ \vdots \\ \theta_t - \theta_{t-1} \end{pmatrix} = BB^{-1} \begin{pmatrix} g_0 \\ g_1 \\ \vdots \\ g_{t-1} \end{pmatrix} + B \begin{pmatrix} z_0 \\ z_1 \\ \vdots \\ z_{t-1} \end{pmatrix}$$

DP-FTRL update (with correlated noise)

DP-FTRL: Incorporating Correlated Noise

$$-\begin{pmatrix} \theta_1 - \theta_0 \\ \theta_2 - \theta_1 \\ \vdots \\ \theta_t - \theta_{t-1} \end{pmatrix} = B \left(B^{-1} \begin{pmatrix} g_0 \\ g_1 \\ \vdots \\ g_{t-1} \end{pmatrix} + \begin{pmatrix} z_0 \\ z_1 \\ \vdots \\ z_{t-1} \end{pmatrix} \right)$$

Privatize $B^{-1}G$ with the Gaussian mechanism

DP-FTRL: Incorporating Correlated Noise

$$-\begin{pmatrix} \theta_1 - \theta_0 \\ \theta_2 - \theta_1 \\ \vdots \\ \theta_t - \theta_{t-1} \end{pmatrix} = B \left(B^{-1} \begin{pmatrix} g_0 \\ g_1 \\ \vdots \\ g_{t-1} \end{pmatrix} + \begin{pmatrix} z_0 \\ z_1 \\ \vdots \\ z_{t-1} \end{pmatrix} \right)$$

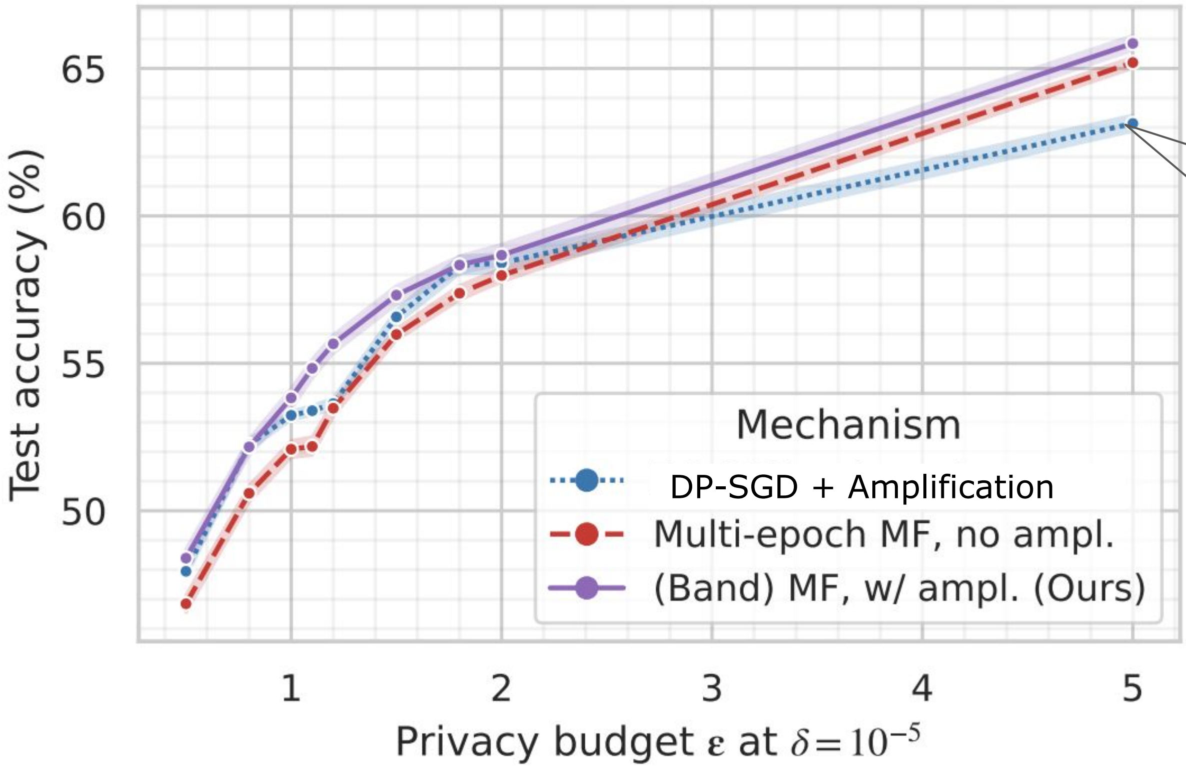
Privatize $B^{-1}G$ with the Gaussian mechanism

For ϱ-zCDP, take noise variance = $\frac{G^2}{2\rho} \max_t \left\| [B^{-1}]_{:,t} \right\|_2^2$

sensitivity

Google Research

DP-FTRL vs. DP-SGD: Empirical



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

Google Research

DP-FTRL vs. DP-SGD: Theory

For convex & *G*-Lipschitz losses

DP-SGD	$\frac{Gd^{1/4}}{\sqrt{\rho T}}$
DP-FTRL	$\frac{Gd^{1/4}}{\sqrt{\rho^2 T}}$

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.

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

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

Google Research

Towards a provable gap between DP-SGD & DP-FTRL

Model parameters

Loss function

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

Data

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

Toeplitz noise correlations: $\beta_{t,\tau} = \beta_\tau$

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

$$B = \begin{pmatrix} \beta_{0,0} & & & \\ \beta_{0,1} & \beta_{1,0} & & \\ \beta_{0,2} & \beta_{1,1} & \beta_{2,0} & \cdots \\ \vdots & & & \end{pmatrix} \longrightarrow B = \begin{pmatrix} \beta_0 & & & \\ \beta_1 & \beta_0 & & \\ \beta_2 & \beta_1 & \beta_0 & \cdots \\ \vdots & & & \end{pmatrix}$$

Computationally: store $O(T)$ coefficients instead of $O(T^2)$

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

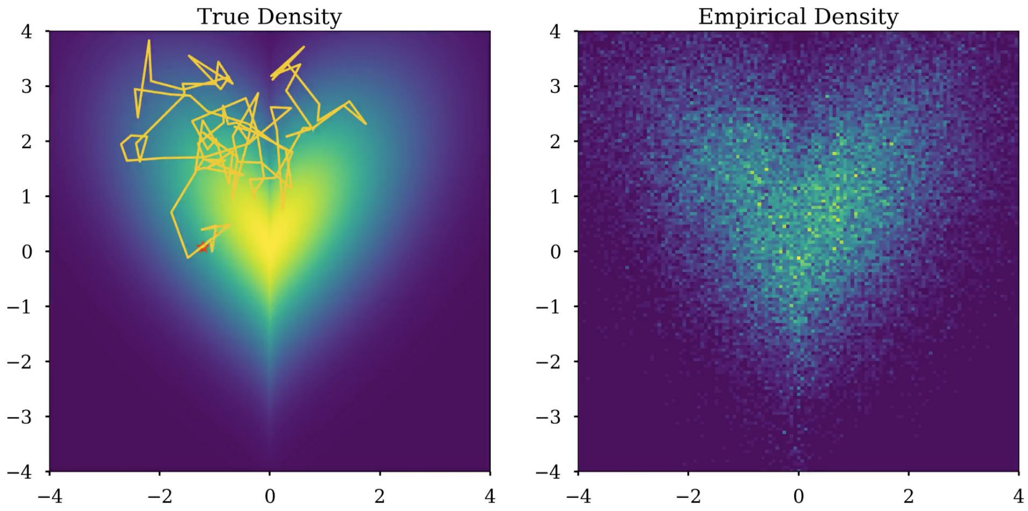


Image credit:
[Abdul Fatir Ansari]

Google Research

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

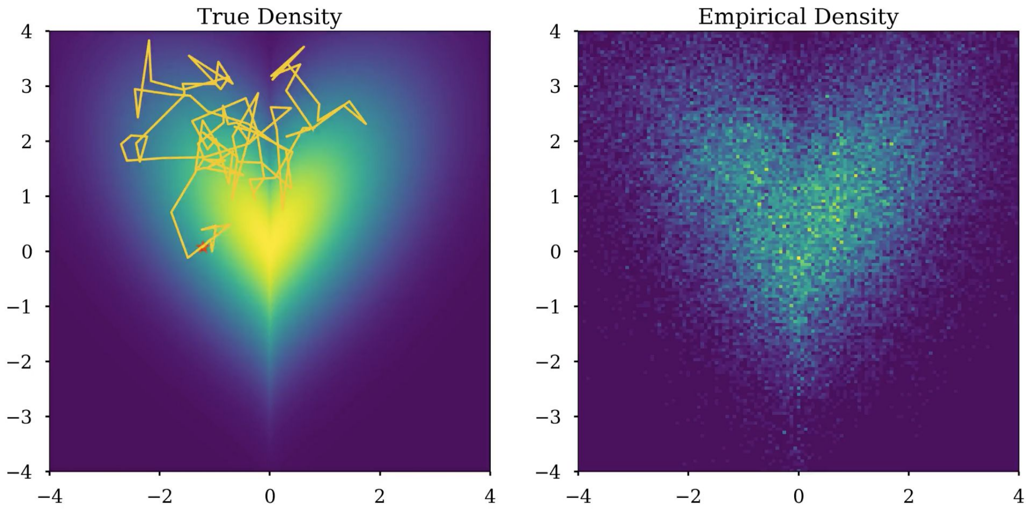
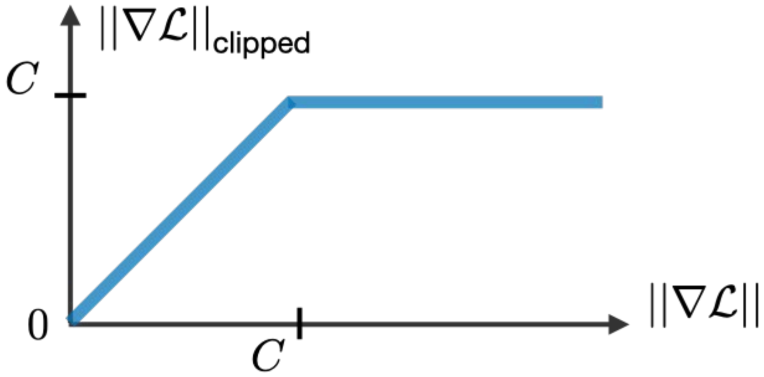


Image credit:
Abdul Fatir Ansari

Asymptotic error

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

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



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

Outline

- Background
- **Theoretical Results**
- Empirical Results

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

Google Research

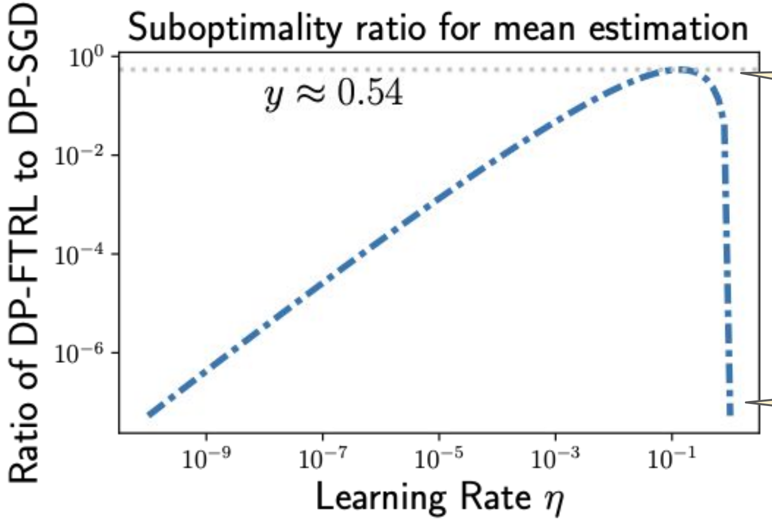
Mean estimation in 1 dimension

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

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

η: learning rate
ϱ: privacy level

Google Research



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 $\beta_0^\star = 1, \quad \beta_t^\star = -t^{-3/2}(1-\eta)^t$ 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 $\beta_0^\star = 1, \quad \beta_t^\star = -t^{-3/2}(1-\eta)^t$
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 $\beta_0 = 1, \quad \beta_t = -t^{-3/2}(1-\nu)^t$

and tune the parameter ν

Google Research

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

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

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

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

Well-specified
linear model

Google Research

Informal Theorem: The asymptotic error for linear regression with $\lambda_{\max}(H) = 1$ and $0 < \eta < 1$

Independent noise (Noisy-SGD)	$=$	$d \; \rho^{-1} \eta$
Correlated noise (ν-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_{eff}

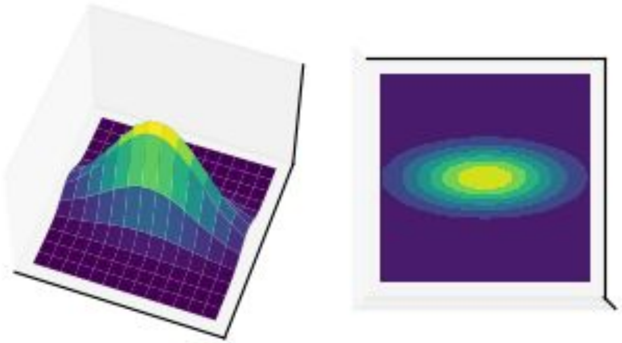
Google Research

Effective dimension

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

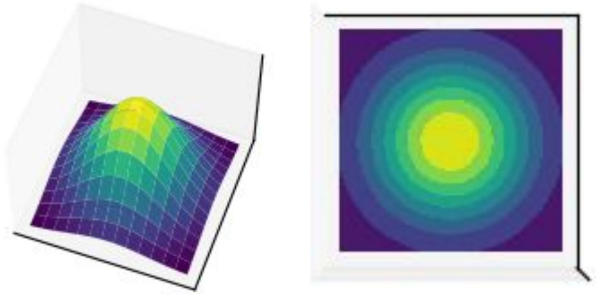
Low effective dimension

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



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.

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

Google Research

[Rudelson & Vershynin (J. ACM 2007)]

The stable rank appears in:

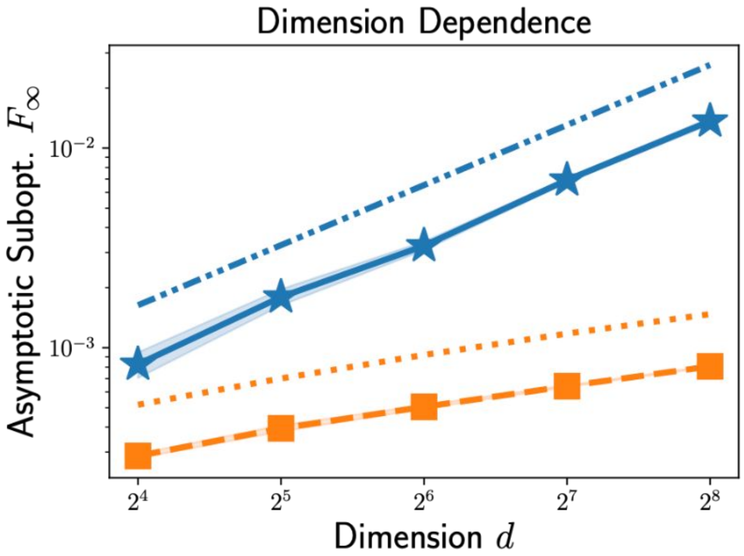
- Numerical linear algebra (e.g. randomized matrix multiplications) [Tropp (2014), Cohen-Nelson-Woodruff (2015)]
- Matrix concentration [Hsu-Kakade-Zhang (2012), Minsker (2017)]
- …

Informal Theorem: The asymptotic error for linear regression with $\lambda_{max}(H) = 1$ and $0 < \eta < 1$

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

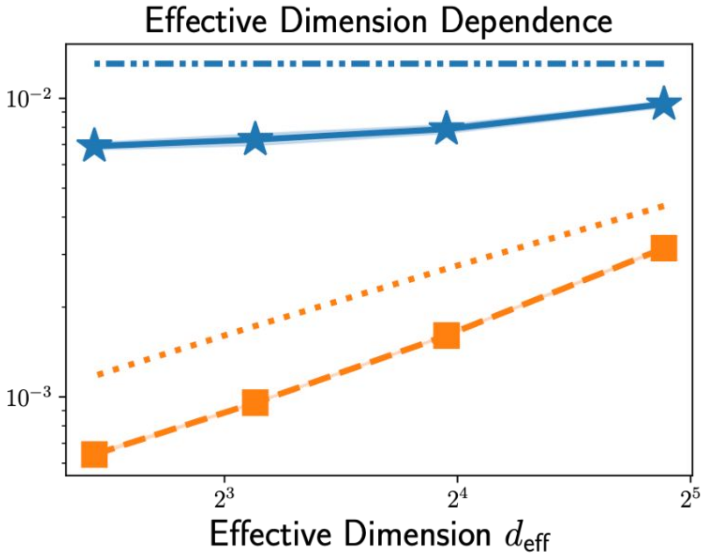
Improve dimension d to problem-dependent ***effective dimension*** d_{eff}

Linear regression: theory predicts simulations



Noisy-SGD
scales with d

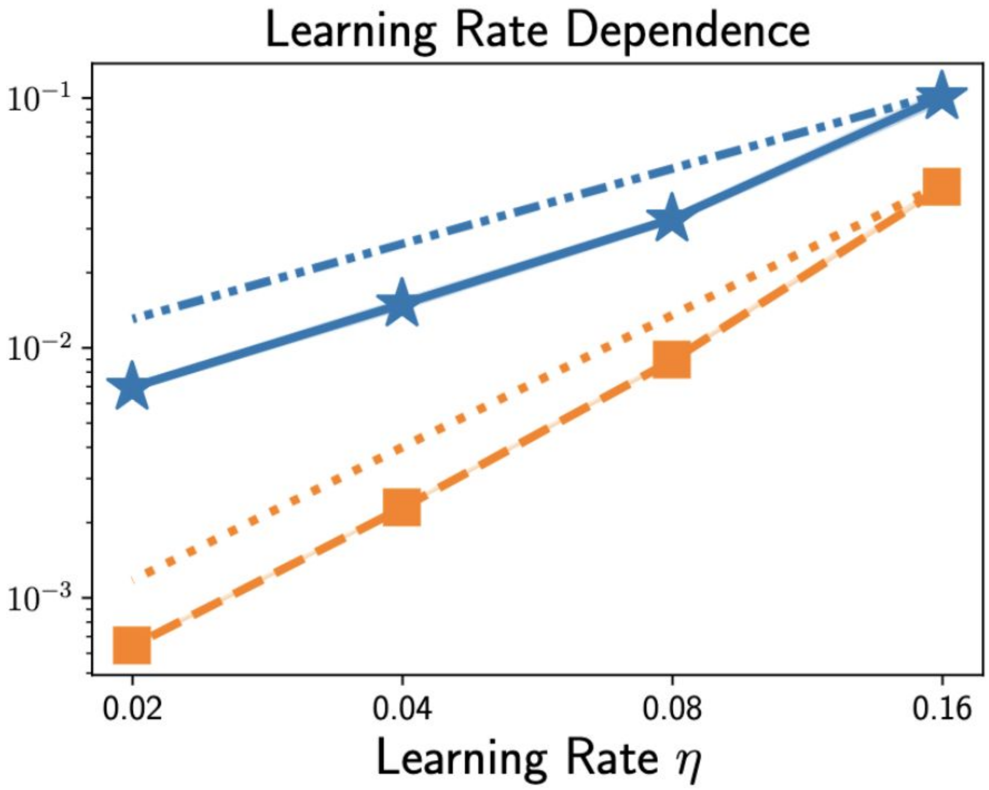
Noisy-FTRL
scales with d_{eff}



Informal Theorem: The asymptotic error for linear regression with $\lambda_{max}(H) = 1$ and $0 < \eta < 1$

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

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

Google Research

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_0=1, 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_0=1, 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)

Google Research

Anti-PGD can be adapted for DP by damping: take β_0=1, β_1=-ν (0 < ν < 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

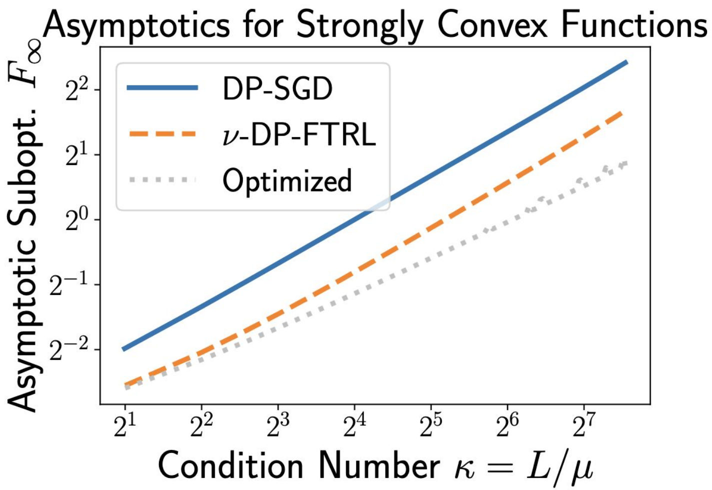
Rates with DP

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

Extensions

- Gap between DP-FTRL & DP-SGD for general strongly convex functions



Google Research

Proof sketch for Mean Estimation

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

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

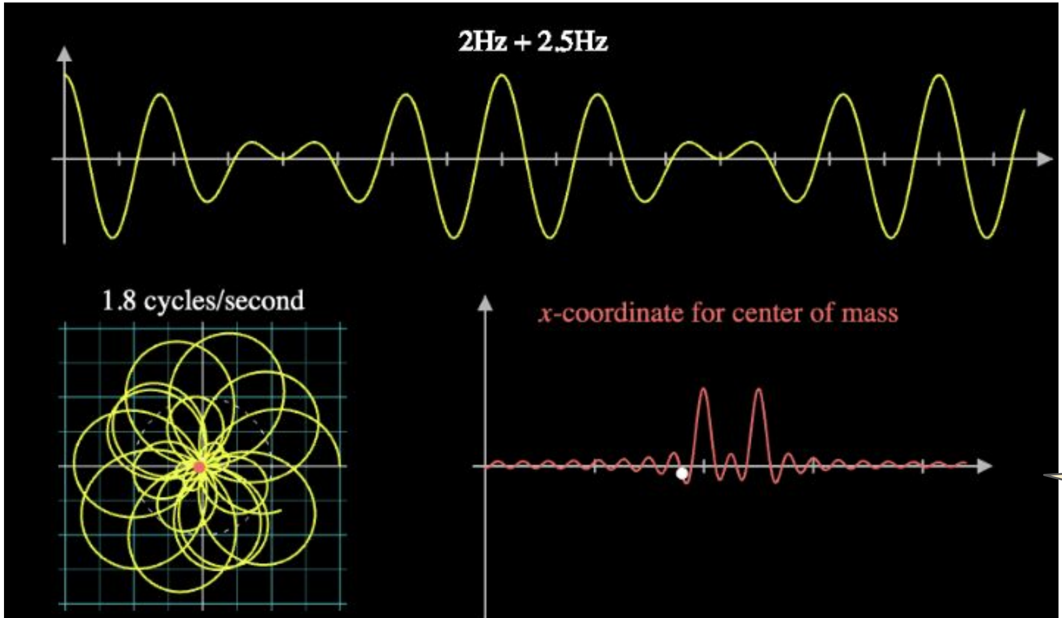


Image: 3blue1brown.com/lessons/fourier-transforms

Time-domain description

Frequency-domain description

Google Research

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

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

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

sensitivity

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

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

$$B = \begin{pmatrix} \beta_0 & & & \\ \beta_1 & \beta_0 & & \\ \beta_2 & \beta_1 & \beta_0 & \cdots \\ \vdots & & & \end{pmatrix}$$

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

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

sensitivity

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

$$B = \begin{pmatrix} \beta_0 & & & \\ \beta_1 & \beta_0 & & \\ \beta_2 & \beta_1 & \beta_0 & \cdots \\ \vdots & & & \end{pmatrix}$$

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

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

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

sensitivity

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

$$B = \begin{pmatrix} \beta_0 & & & \\ \beta_1 & \beta_0 & & \\ \beta_2 & \beta_1 & \beta_0 & \cdots \\ \vdots & & & \end{pmatrix}$$

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:

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

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

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

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

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:

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

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

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

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

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.

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

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

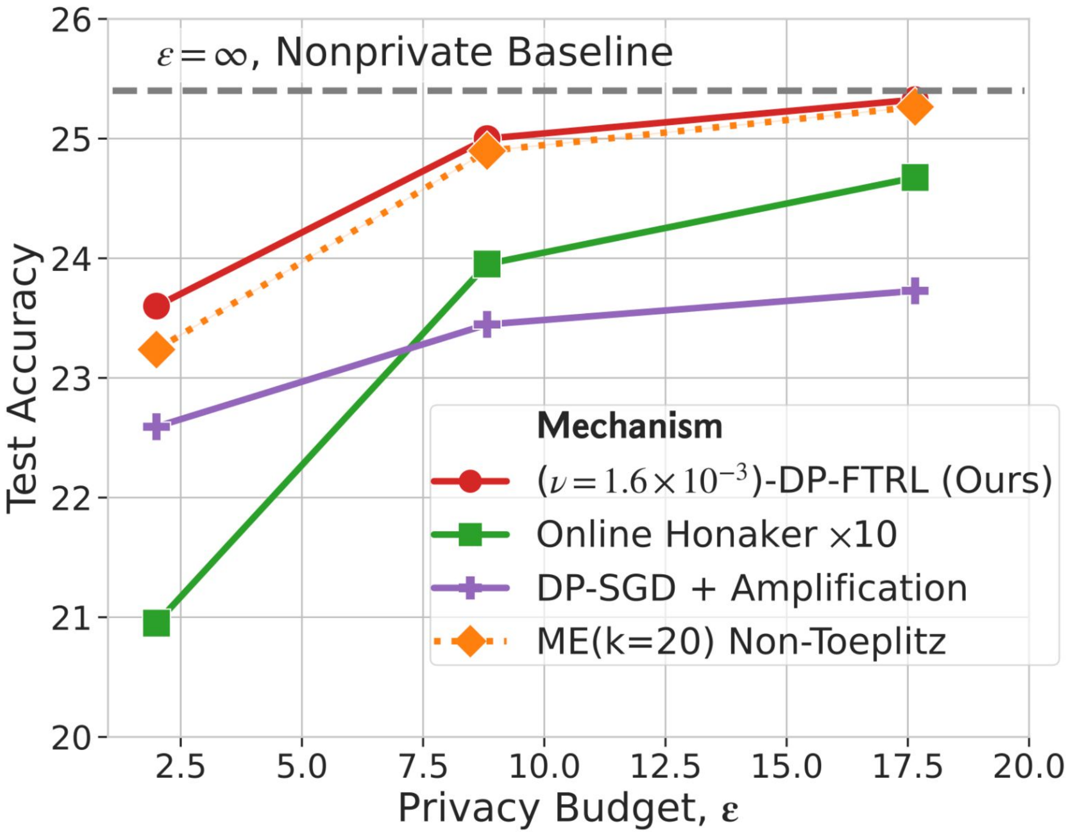
Google Research

Outline

- Background
- Theoretical Results
- **Empirical Results**

Empirical Results

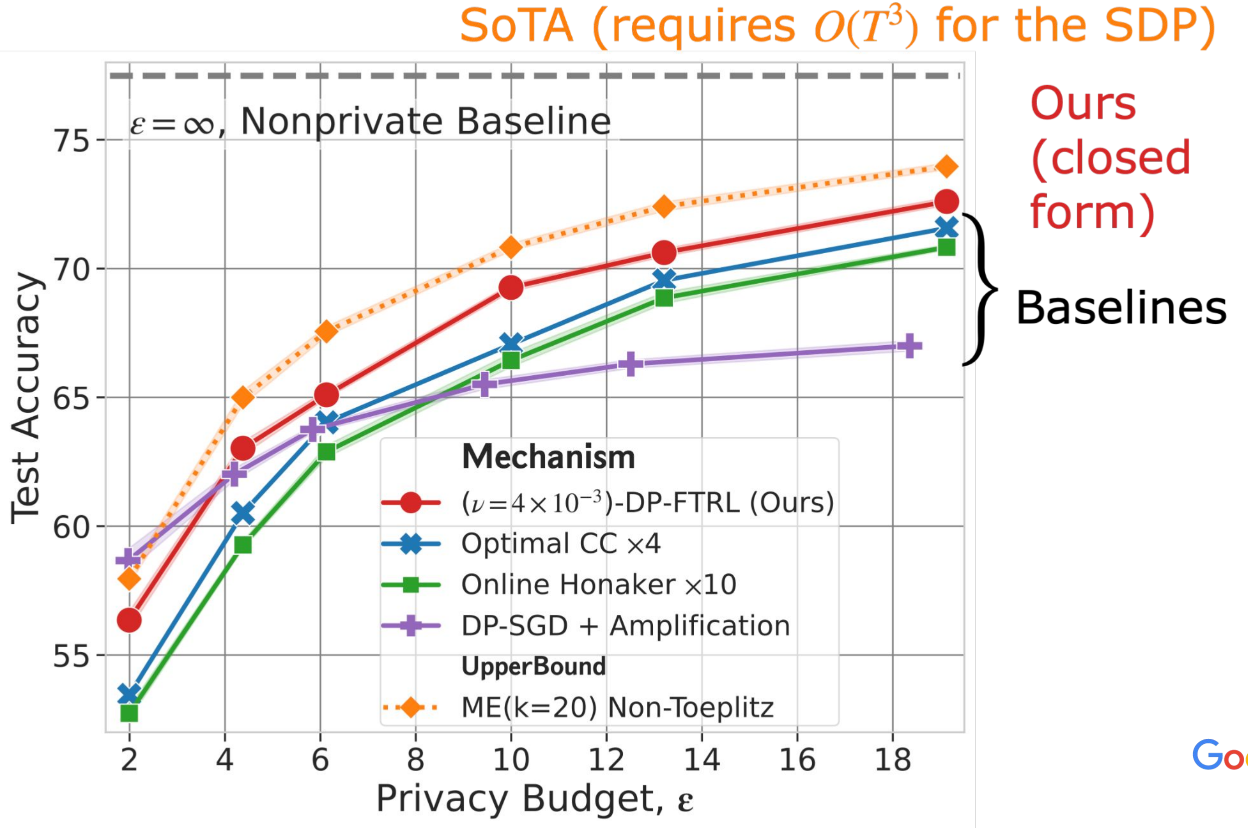
Language modeling with Stack Overflow



Ours matches SoTA!

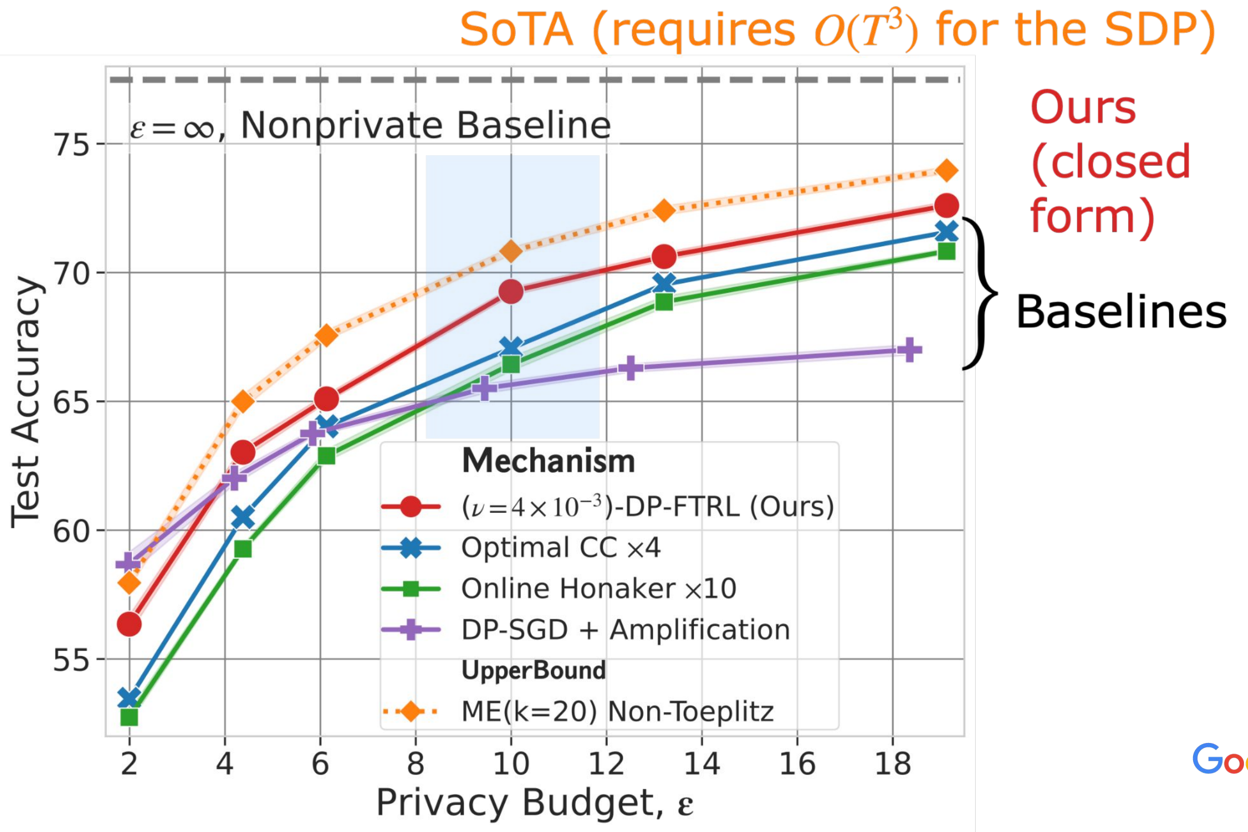
Google Research

Image classification with CIFAR-10



Google Research

Image classification with CIFAR-10



Google Research

Summary

Theory

- correlated noise is **provably** better
- Depends on effective dimension instead of dimension
- Matches lower bounds

Empirical:

- computationally much more efficient that SoTA (cubic → constant)
- nearly matches SoTA empirically

Google Research

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

Google Research

Future Work

Algorithms

- Natively support adaptive gradient methods

Future Work

Practical:

- *Efficient approximation*:
  - Currently, running time = $O(T^2)$ for T iterations
  - "Low rank" approx: $O(k)$ runtime, $O(kd)$ memory
  - Approximation theory of rational functions

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

Google Research

Thank you! Questions?

Arxiv link

https://arxiv.org/pdf/2310.06771.pdf

Google Research