# Correlated Noise Provably Beats Independent Noise for Differentially Private Learning

**Christopher Choquette-Choo***, **Krishnamurthy Dvijotham***, **Krishna Pillutla***,
Arun Ganesh, Thomas Steinke, Abhradeep Thakurta

Google DeepMind
Google Research

## Differentially Private Learning

### Summary

How to add **noise** to SGD for differentially private learning?

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

**Main result**: (anti-) correlated noise is **provably** better

**Standard Approach:** SGD with **independent** noise

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

Learning Rate — Clipped gradient s.t. $\|g_t\|_2 \le G$ — **Independent** Gaussian noise

For streaming data, take $z_t \sim \mathcal{N}(0, G^2/2\rho)$ for a desired level of privacy $\rho$ zero-concentrated DP

**Recent Work:** SGD with **correlated** noise

[Kairouz et al. (ICML '21), Denisov et al. (NeurIPS '22), …]

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

**Correlated** Gaussian noise

**Sensitivity**

$$\text{Variance}(z_t) = \frac{G^2}{2\rho} \max_t \left\| [B^{-1}]_{:,t} \right\|_2^2$$

required for the same privacy

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

**Empirically:** Correlated noise ≫ independent noise
[Choquette-Choo et al. (NeurIPS '23)]

**Compute:** Solve a semi-definite program for $B$

**This work**: *first* clear separation in theory!

Asymptotic suboptimality: $F_\infty(\beta) = \lim_{t\to\infty} \mathbb{E}\big[F(\theta_t) - F(\theta_\star)\big]$

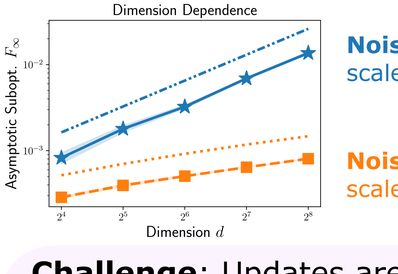
## Theoretical Results

**Mean Estimation in 1D** with $f(\theta; x) = (\theta - x)^2$

**Theorem**: $F_\infty(\beta^{\text{sgd}}) = \rho^{-1}\eta$ **(DP-SGD)**

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

Optimal correlations are $\beta_0^\star = 1$, $\beta_t^\star = -t^{-3/2}(1-\eta)^t$.




DP-FTRL is always better

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

**$\nu$-DP-FTRL**: use $\beta_t^\nu = -t^{-3/2}(1-\nu)^t$ with a tunable $\nu$

### Linear Regression

$f(\theta; x, y) = (y - \langle \theta, x \rangle)^2$ with $x \sim \mathcal{N}(0, H)$

**Theorem**: For algorithms without clipping, we have

$$F_\infty(\beta^{\text{sgd}}) = d\ \rho^{-1}\eta \qquad \textbf{(Noisy-SGD)}$$

$$F_\infty(\beta^\nu) \le d_{\text{eff}}\ \rho^{-1}\eta^2 \log^2 \frac{1}{\eta\mu} \qquad \textbf{(Noisy-FTRL)}$$

$$\inf_\beta F_\infty(\beta) \ge d_{\text{eff}}\ \rho^{-1}\eta^2 \qquad \textbf{(Lower bound)}$$

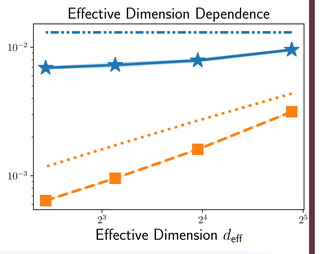
**Effective** dimension $d_{\text{eff}} = \text{Tr}(H)/\|H\|_2 \le d$

Near-optimal up to log factors




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

**Noisy-FTRL** scales with $d_{\text{eff}}$

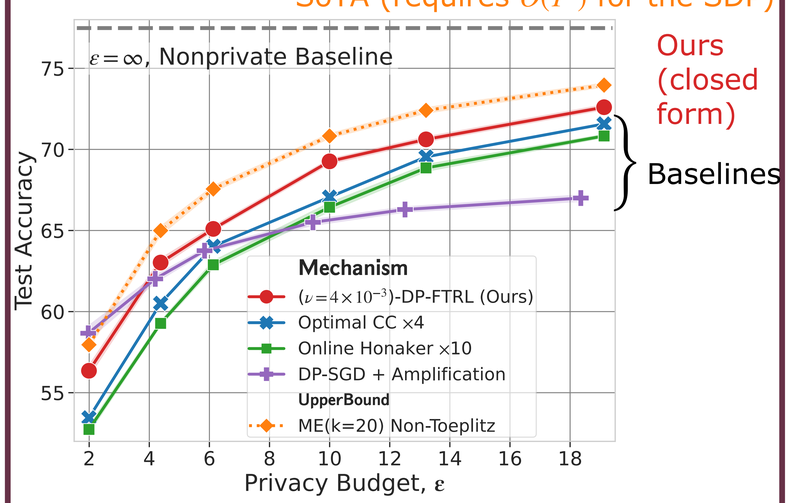



**Challenge**: Updates are no longer Markovian

**Solution**: Analysis in the Fourier domain

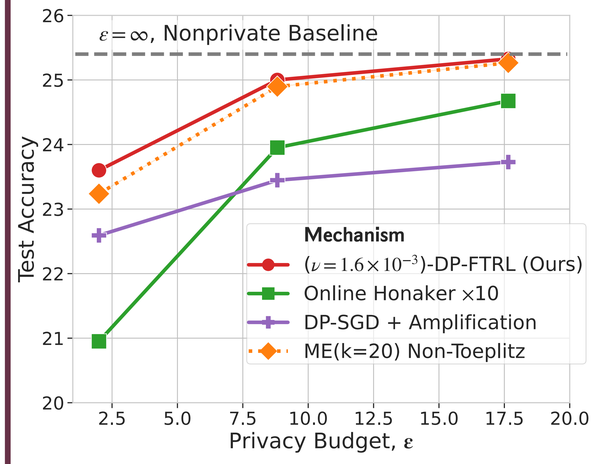
## Experiments: Private Deep Learning

**Image Classification** (CIFAR-10)




**Language Modeling** (StackOverflow)

Federated learning + user-level DP




Ours matches SoTA!

**Extensions** (see paper)

- Finite-time DP bounds for linear regression
- Bounds for general strongly convex functions

**Arxiv link**