Theory of Differential Privacy

Lecture 14:

DP-FTRL

EN.601.438/638 Fall 2026

Lydia Zakynthinou

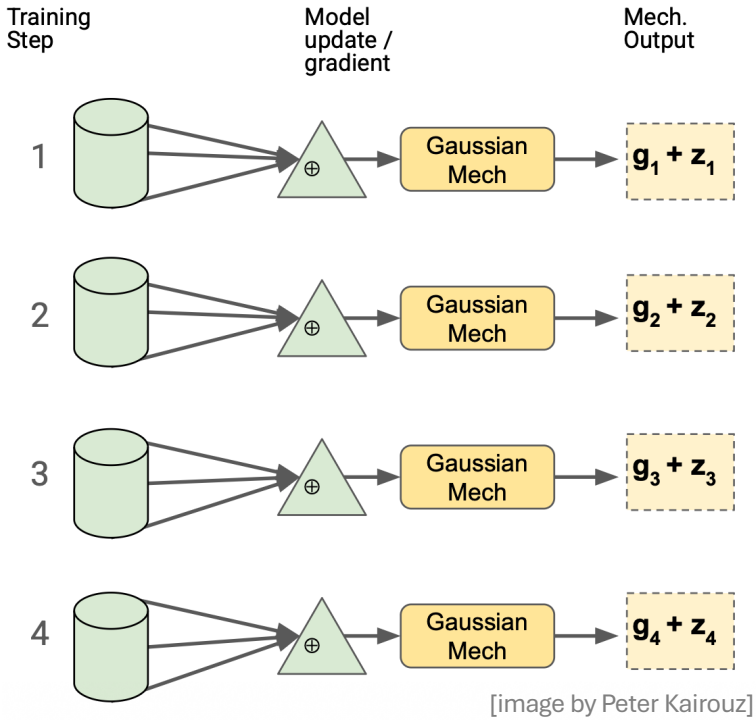
Johns Hopkins University

Oct 21, 2026

Through the lens of continual observation

$\Theta = \mathbb{R}^d$

Through the lens of continual observation



[image by Peter Kairouz]

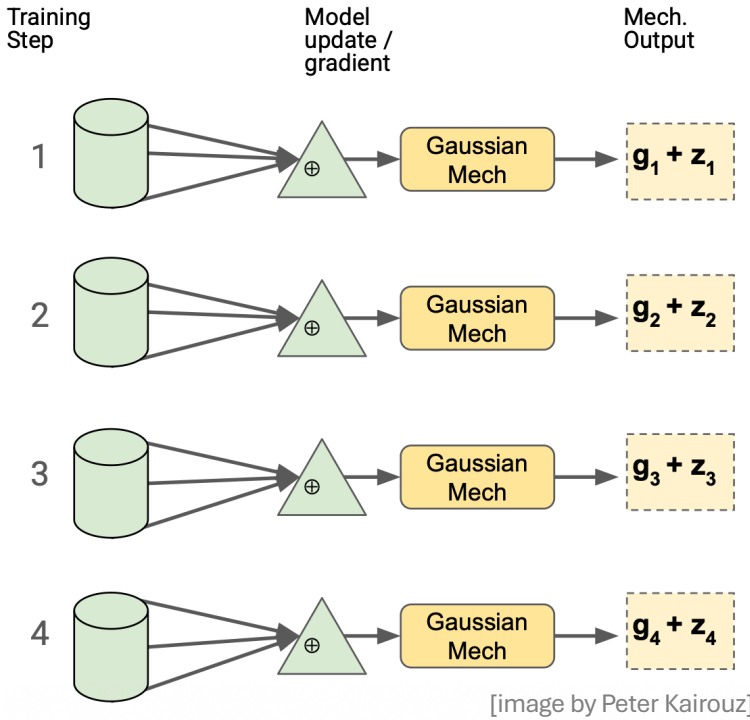
$$\theta_t = \theta_0 - \eta \underbrace{\sum_{\tau \in [t]} g_\tau}_{s_t}$$

Need a good approximation of prefix sums s_t, not each gradient g_t.

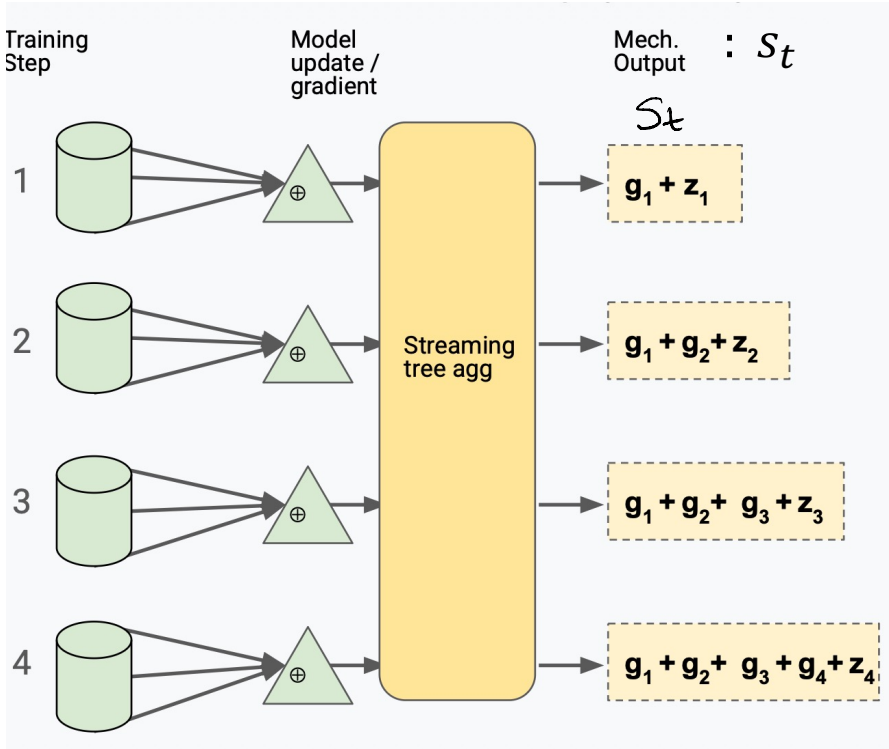
$$\theta_t = \Pi_\Theta(\theta_{t-1} - \eta g_t) = \theta_{t-1} - \eta g_t =$$

$$\frac{1}{n}\sum_{i=1}^{n} \nabla \ell(\theta_{t-1}, x_i)$$

Through the lens of continual observation



[image by Peter Kairouz]



Dwork, Naor, Pitassi, Rothblum [2010], Chan, Shi, Song [2011]

Binary Tree Mechanism

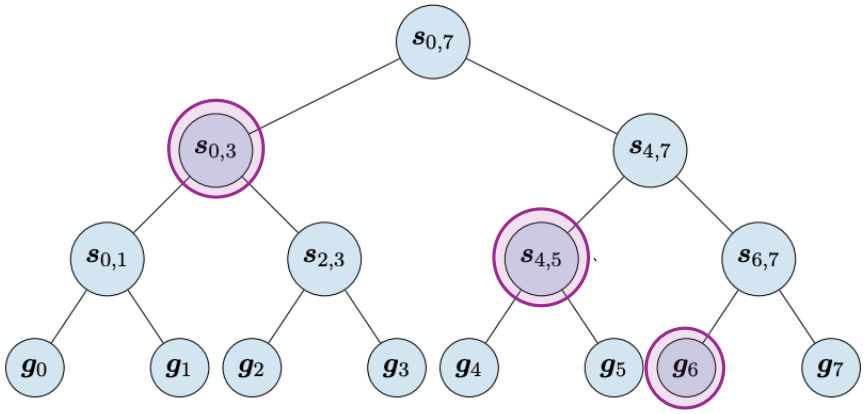
Input: stream $g_1, \dots, g_T, \forall t: \|g_t\| \le C = 1$
Output: prefix sums $\tilde{s}_1, \dots, \tilde{s}_T$

Every g_t affects $\log t$ nodes

$\Rightarrow \sigma \approx \frac{\log T \sqrt{\log(\frac{\log T}{\delta})}}{\varepsilon}$ ensures (ε, δ)-DP.

Every s_t decomposes into $\log t$ nodes

$\Rightarrow \|\tilde{s}_t - s_t\| = \|z_t\| \approx \sigma\sqrt{d \log T} = \tilde{O}_{\varepsilon,\delta}(\log^{1.5} T)$



Each node has noise $v_{a,b} = \mathcal{N}(0, \sigma^2 \mathbb{I}_d)$

$$\begin{aligned}\tilde{s}_{0,6} &= \tilde{s}_{0,3} + \tilde{s}_{4,5} + \tilde{g}_6 \\ &= s_{0,3} + s_{4,5} + g_6 + v_{0,3} + v_{4,5} + v_6 \\ &= \textstyle\sum_{t=0}^{6} g_t + z_t\end{aligned}$$

Sum of $\log t$ independent d-dim Gaussians

Jain, Kothari, Thakurta [2012], Thakurta, Smith [2013],
Kairouz, McMahan, Song, Thakkar, Thakurta, Xu [2021]

DP-Follow The Regularized Leader

$n\,(d + \text{time}(\nabla))$

- Initialize θ_0
- For $t = 1, \ldots, T$: $= n$

1. Compute $\tilde{s}_t$ via Binary Tree Mechanism
2. Update step: $\theta_t = \text{argmin}_{\theta \in \Theta} \left(\langle \tilde{s}_t, \theta \rangle + \frac{1}{2\eta} \|\theta - \theta_0\|^2 \right)$ $= \theta_0 - \eta \tilde{s}_t$

Excess risk of $\bar{\theta} = \frac{1}{T}\sum \theta_t$ is $\approx \eta \text{MaxErr}(\tilde{s}) + \frac{\|\theta^* - \theta_0\|}{\eta n} + \eta \approx \tilde{O}\left(\sqrt{\frac{\sqrt{d}}{\varepsilon n}}\right)$

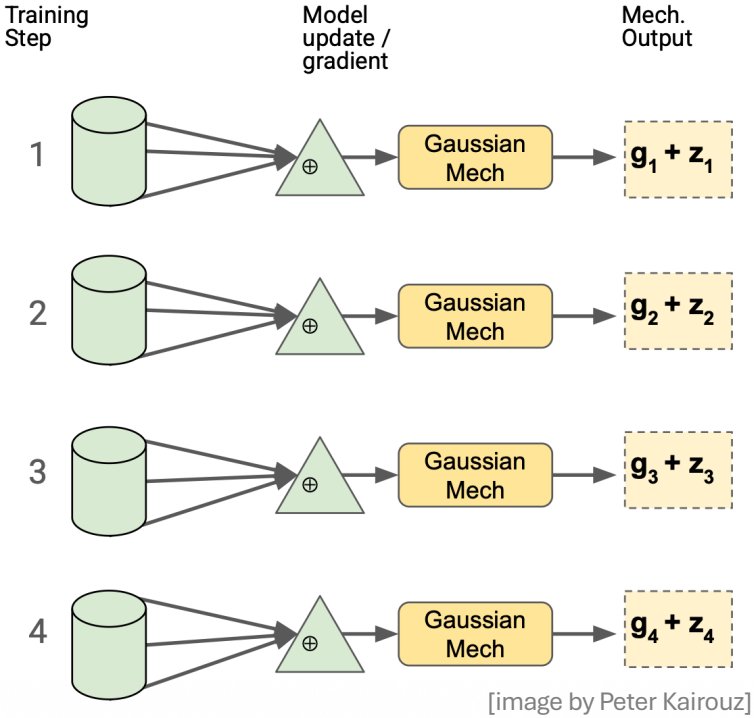
* Binary Tree for continual observation does not take into account that g_t depends on g_{t-1} etc). OK in this case because it's an adaptive process using Gaussian mech.

Denisov, McMahan, Rush, Smith, Thakurta [2022], Choquette-Choo, McMahan, Rush, Thakurta [2023],
Choquette-Choo, Dvijotham, Pillutla, Ganesh, Steinke, and Thakurta [2024].

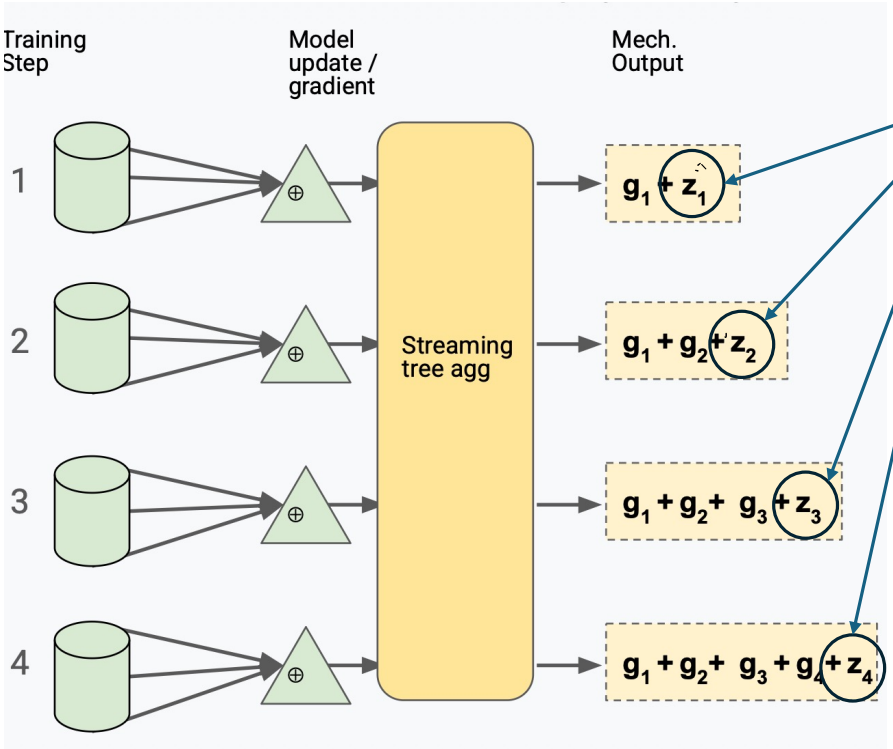
$$\text{Regret}_T = \sum_{t=1}^{T} \ell(\theta_t, x_t) - \sum_{t=1}^{T} \ell(\theta_t, w^*) \le \eta \max_t \text{Err}(\tilde{s}) + \frac{\|\theta^* - \theta_0\|}{\eta n}$$

$$\bar{\theta} = \frac{1}{T}\sum_{t=1}^{T} \theta_t \Rightarrow L(\bar{\theta}) \le \frac{1}{T}\text{Regret}_T$$

Through the lens of continual observation



[image by Peter Kairouz]



Prefix sums as linear queries

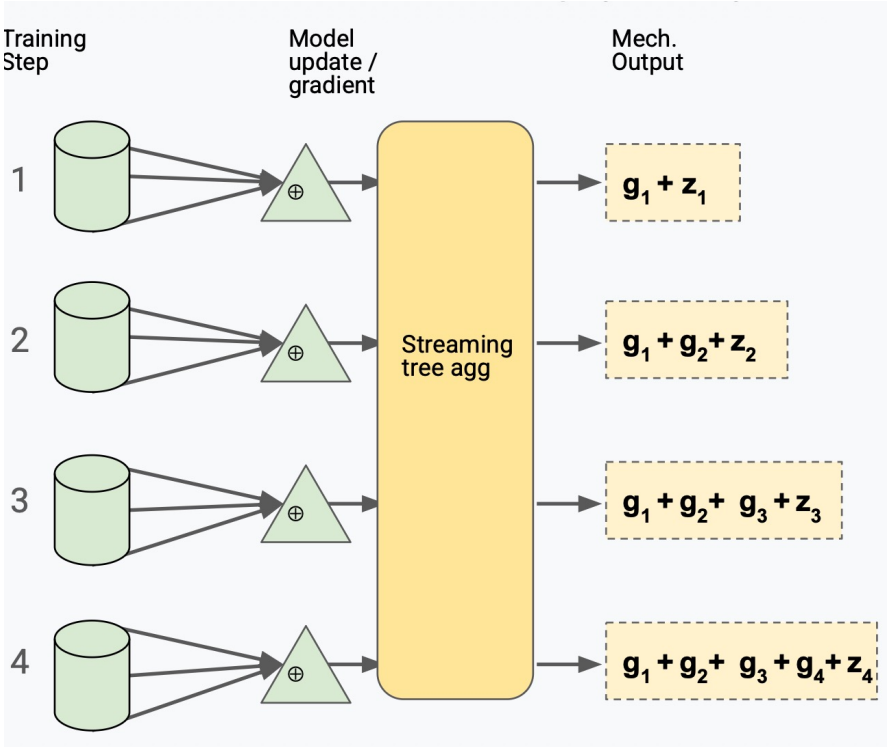
Prefix sums are linear queries

- Let $G = (g_1, \dots, g_n)^T \in \mathbb{R}^{n \times d}$ be the stack of gradients
- Let $A^{pre} \in \mathbb{R}^{n \times n}$ be the (lower-triangular) prefix sum matrix

$$A^{pre}[t, \tau] = \begin{cases} 1, & t \leq \tau \\ 0, & t > \tau \end{cases}$$

- Then $S = \mathrm{A}^{pre} G \in \mathbb{R}^{n \times d}$ is the stack of prefix sums s_t!

Release $S = \mathrm{A}^{pre} G$ privately



[image by Peter Kairouz]

$$A^{pre} \cdot G = \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_T \end{bmatrix}$$

Li, Miklau, Hay, McGregor, Rastogi [2015] “matrix mechanism”
Nikolov, Talwar, Zhang [2016] “factorization mechanism”

Correlated Noise Mechanisms

Compute $S = AG \in \mathbb{R}^{n \times d}$

inputs $G = (g_1, \dots, g_n)^T \in \mathbb{R}^{n \times d}$

workload matrix $A \in \mathbb{R}^{n \times n}$

Strategy:

1. Choose factorization $A = BC$
2. Add correlated noise $G + C^{-1}Z,\ Z \sim \mathcal{N}_{n \times d}(0, \sigma^2)$ (sensitivity depends on C)
3. Return $\mathrm{A}(G + C^{-1}Z) = S + BZ$ (utility depends on B)

Correlated Noise Mechanisms

Strategy:

1. Choose factorization $A = BC$
2. Add correlated noise $G + C^{-1}Z,\ Z \sim \mathcal{N}_{n\times d}(0, \sigma^2)$ (sensitivity depends on C)
3. Return $\mathrm{A}(G + C^{-1}Z) = S + BZ$ (utility depends on B)

- The mechanism $\mathcal{M}(G) = A(G + C^{-1}Z) = B(CG + Z)$ is a post-processing.
- Sensitivity $\max_{G,G'} \|C(G - G')\|_F \leq 2\|C\|_{col}$
- Setting $\sigma = c_{\varepsilon,\delta}\|C\|_{col}$ guarantees (ε, δ)-DP.

★★☆

Utility "max loss":

$$\max_{t\in[n]} \mathbb{E}\|\tilde{s}_t - s_t\|^2 = \max_{t\in[n]} \mathbb{E}\big((BZ)[t,:]\big)^2 = d\sigma^2 \max_{t\in[n]} \|B[t,:]\|^2 \propto \|C\|_{col}^2 \cdot \|B\|_{row}^2$$

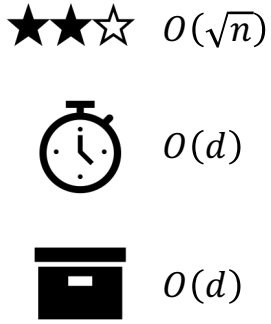
$\max \|\tilde{s}_t - s_t\| \leq \|C\|_{col} \|B\|_{row}$

Baseline: DP-SGD

Initialize θ_0

For $t = 1, \dots, n$:

1. Get clipped gradient (in unit ball): g_t
2. Sample iid noise $z_t = \mathcal{N}(0, \sigma^2 \mathbb{I}_d)$
3. Compute correlated noise $\tilde{z}_t = \sum_{\tau \in [t]} (C^{-1})[t, \tau] z_\tau$
4. $\tilde{g}_t = g_t + \tilde{z}_t$
5. Take a step in the negative direction: $\theta_t = \theta_{t-1} - \eta \tilde{g}_t$



- $C = \mathbb{I}, B = A$
- $\sigma = c_{\varepsilon,\delta} \|C\|_{col} \approx \sigma = c_{\varepsilon,\delta}$

- Prefix sums

$$\tilde{s}_t = \sum_{\tau \in [t]} A[t, \tau] \tilde{g}_\tau = \sum_{\tau \in [t]} g_\tau + \sum_{\tau \in [t]} z_\tau$$

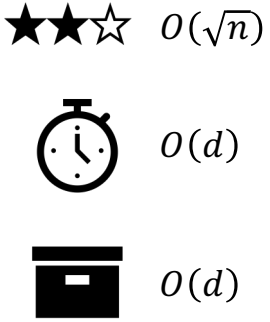
More Baselines

Input perturbation (DP-SGD)

- $C = \mathbb{I}, B = A$
- $\sigma = c_{\varepsilon,\delta}\|C\|_{col} \approx \sigma = c_{\varepsilon,\delta}$

- Prefix sums

$$\tilde{s}_t = \sum_{\tau \in [t]} g_\tau + \sum_{\tau \in [t]} z_\tau$$

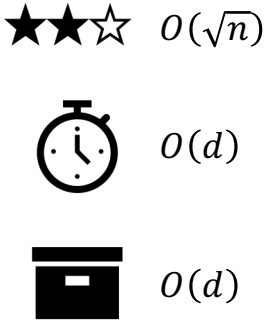


Output perturbation

- $C = A, B = \mathbb{I}$
- $\sigma \approx c_{\varepsilon,\delta}\|A\|_{col} \approx c_{\varepsilon,\delta}\sqrt{n}$

- Prefix sums

$$\tilde{s}_t = \sum_{\tau \in [t]} g_\tau + z_t$$



More Baselines

Output perturbation

- $C = A, B = \mathbb{I}$
- $\sigma \approx c_{\varepsilon,\delta}\|A\|_{col} \approx c_{\varepsilon,\delta}\sqrt{n}$

- Prefix sums

$$\tilde{s}_t = \sum_{\tau \in [t]} A[t,\tau] g_\tau + z_t$$

★★☆ $O(\sqrt{n})$

$O(d)$

$O(d)$

Binary Tree Mechanism

- C, B $\log n$-sparse

- Prefix sums

$$\tilde{s}_t = \sum_{\tau \in [t]} g_\tau + B[t,\tau] z_t$$

★★★ $\log_2 n + 2$

$O(d \log n)$

$O(d \log n)$

Optimal Correlated Noise Mechanisms

Given a factorization $A = BC$, $\mathcal{M}(G) = B(CG + Z)$.

Find $(B^*, C^*) = \underset{A=BC}{\operatorname{argmin}} \|C\|_{col}^2 \cdot \|B\|_{row}^2$

SDP+Cholesky decomposition: $O(n^3)$ to compute C^* (offline)

★★★ $\frac{1}{\pi}\ln n + \mathrm{c}$

$O(dn)$

$O(d + n^2)$

For prefix sums, imposing lower-triangular structure:

Non-constructive solution achieves $\frac{1}{\pi}\ln n + 1$ Mathias [1993]

Lower bound for Gaussian Z: $\frac{1}{\pi}\ln n$ Matoušek, Nikolov, Talwar [2020]

More efficient relaxations

Impose Toeplitz structure:

$$\boldsymbol{C} = \begin{pmatrix} c_0 & 0 & 0 & \cdots & 0 \\ c_1 & c_0 & 0 & \cdots & 0 \\ c_2 & c_1 & c_0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ c_{n-1} & c_{n-2} & c_{n-3} & \cdots & c_0 \end{pmatrix}$$

★★★ $\frac{1}{\pi}\ln n + \mathrm{c}$

$O(dn)$

$O(d+n)$

For prefix sums, closed-form $B_{Toep} = C_{Toep} = (A^{pre})^{1/2}$

⋮

Impose Banded Toeplitz structure

Choquette-Choo, Ganesh, McKenna, McMahan, Rush, Guha Thakurta, Xu [2023]

Dvijotham, McMahan, Pillutla, Steinke, Thakurta [2024]

More efficient relaxations

Impose BLT (Buffered Linear Toeplitz) structure:

$$\mathsf{BLT}(\boldsymbol{\alpha}, \boldsymbol{\lambda}) := \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ \sum_{i=1}^{d} \alpha_i & 1 & \cdots & \cdots & 0 \\ \sum_{i=1}^{d} \alpha_i \lambda_i & \sum_{i=1}^{d} \alpha_i & 1 & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{d} \alpha_i \lambda_i^{n-2} & \sum_{i=1}^{d} \alpha_i \lambda_i^{n-3} & \sum_{i=1}^{d} \alpha_i \lambda_i^{n-4} & \cdots & 1 \end{pmatrix}$$

★★★ $\frac{1}{\pi} \ln n + \mathrm{c}$

$O(d \log^2 n)$

$O(d \log^2 n)$

Almost as time-/memory-efficient as Binary Tree, with optimal approximation

Correlated Noise Mechanisms for Differentially Private Learning

A Tutorial on Mathematical Foundations and Practical Aspects

Krishna Pillutla
IIT Madras

Jalaj Upadhyay
Rutgers University

Christopher A. Choquette-Choo
Google DeepMind

Krishnamurthy Dvijotham
ServiceNow Research

Arun Ganesh
Google Research

Monika Henzinger
IST, Austria

Jonathan Katz
Google

Ryan McKenna
Google Research

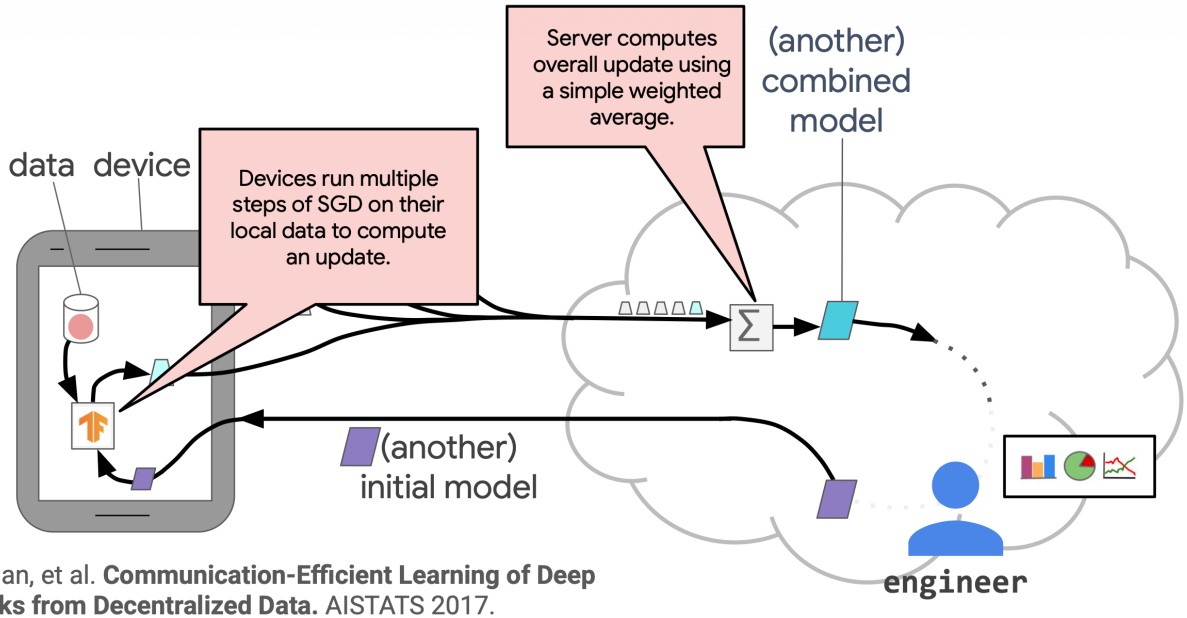
H. Brendan McMahan
Google Research

Keith Rush
Google DeepMind

Thomas Steinke
Google DeepMind

Abhradeep Thakurta
Google DeepMind

Privacy Unit

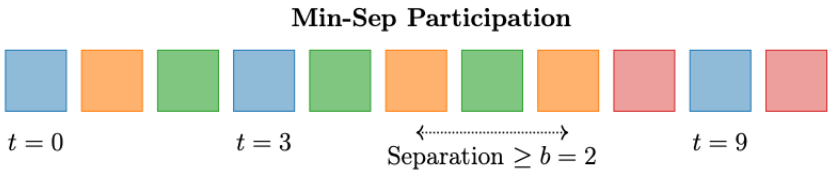


McMahan, et al. **Communication-Efficient Learning of Deep Networks from Decentralized Data.** AISTATS 2017.

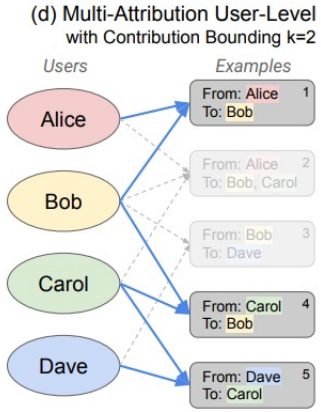
[image by Peter Kairouz]

Privacy Unit

- User-level in streaming



  - Each user participates at most k times ⇒ sensitivity $\times\sqrt{k}$
- Network data Fang, Dong, Yi [2022], Ganesh, McKenna, McMahan, Smith, Wu [2025]



Other aspects: convexity and higher-order

- Optimal rate for DP-SCO: Choquette-Choo, Ganesh, Thakurta [2024]
- Last-iterate advantage:
  - Ye, Shokri [2022], Altschuler, Talwar [2023]
  - Steinke, Nasr, Ganesh, Balle, Choquette-Choo, Jagielski, Hayes, Guha Thakurta, Smith, Terzis [2024]
- Second-order methods:
  - Ganesh, Haghifam, Steinke, Guha Thakurta [2023]

Other aspects: Public data

- Public pretraining and private fine-tuning
  - Ganesh, Haghifam , Nasr, Oh, Steinke, Thakkar, Thakurta, Wang [2023]
- Extracting useful info from public data
  - “Handcrafted” features for LR can be better than non-convex models Tramèr, Boneh [2021]
  - AdaGrad with publicly estimated subspaces Kairouz, Ribero, Rush , Thakurta [2021]
  - Narrowing the data universe (US Census)
- PATE
  - Papernot, Abadi, Erlingsson, Goodfellow, Talwar [2017]