BU CS599 S1
Foundations of Private Data Analysis
Spring 2025

Lecture 14: DP-FTRL

Adam Smith

BU

Today

- What’s wrong with DP-SGD and amplification
- Gradient descent as online sums
- Tree mechanism recall
- DP-FTRL mechanism
- Error analysis

What's wrong with DP-SGD?

- DP-SGD requires either
  - Evaluating the entire gradient at each step (time $\Omega(nd)$), or
  - Amplification by subsampling
    - Requires having all the data in one place, randomly permuting it
    - Data needs to fit in memory ☹

- Today: algorithm with weaker guarantees but similar empirical performance
  - Basic idea: gradient descent is a continually updated sum

$$w_{t+1} = w_t + \eta g_t \qquad \text{where } g_t \approx \nabla L(w_t)$$

$$= w_1 + \eta \sum_{i=1}^{t} g_t$$

  - Idea: At every time step, approximate $s_t \coloneqq \sum_{i=1}^{t} g_t$.

Tree Mechanism (Vector Version)

- Receives stream of update vectors
$$g_1, g_2, \dots, g_T \in \mathbb{R}^d \text{ with } \|g_i\| \le 1$$
(we use $T = n$)
- Outputs $\widetilde{s_t} = s_t + Z_t$ where Z_t is tree mechanism noise and $s_t = \sum_{i=1}^{t} g_i$.
  - ➢ Each node $[i, j]$ in binary tree releases its sum plus noise $V_u \sim N(0, \sigma^2 \mathbb{I}_d)$
  - ➢ $Z_t = V_{u_1} + \cdots + V_{u_k}$ where $k \approx \log_2 T$
- **Accuracy:** For all $\beta > 0$, with probability $\ge 1 - \beta$, the Z_t's satisfy
$$\frac{1}{T}\sum_{t=1}^{T} \|Z_t\| \le \sigma\sqrt{d \log T \ln 1/\beta}$$
- **Privacy:** If
→ g_t depends only on x_t and previous outputs $\tilde{s}_1, \dots, \tilde{s}_{t-1}$,
→ and $\sigma = 2\sqrt{\log T \log 1/\delta}\,/\varepsilon$,
then Tree Mechanism satisfies (ε, δ)-DP (and $\left(\sigma\sqrt{\log T}\right)$-GDP).

DP-FTRL for continuous optimization

(with batch size 1 and the tree mech)

$\mathcal{A}(x_1, \dots, x_n)$:

- $w_1 = \vec{0}$
- For $t = 1$ to $T = n$:
  - $g_t = clip\big(\nabla \ell(w_t; x_t)\big)$
  - $\tilde{s}_t =$ (TreeMechanism with update g_t)
  - $w_{t+1} = \arg\min_{w \in C} \left(\langle \tilde{s}_t, w \rangle + \frac{\lambda}{2} \|w\|^2 \right)$
    (This is the same as $w_1 - \frac{1}{\lambda} \tilde{s}_t$ when $C = \mathbb{R}^d$)
- Return $\overline{w} = \frac{1}{T} \sum_{i=1}^{T} w_t$

Privacy: TreeMechanism. (Each data point only affects 1 tree entry. Subtle point: inputs are adaptive. Ok because of Gaussian mechanism.)

Analyzing convergence via regret

Theorem: Suppose ℓ is 1-Lipschitz and convex, and C is convex.

For all $x_1, \ldots, x_n$, and all $w^* \in C$, w.p. $\geq 1 - \beta$:

$$\boldsymbol{R} = \frac{1}{n}\sum_t \ell(w_t; x_t) - \frac{1}{n}\sum_t \ell(w^*, x_t)$$

$$= \frac{(Tree\ mech.\ error)}{\lambda} + \frac{1}{\lambda} + \frac{\lambda}{2n}\|w^*\|$$

where $(Tree\ mech\ error) = O\left(\frac{1}{\varepsilon}\sqrt{d \log^2 n \log\frac{1}{\delta} \log\frac{1}{\beta}}\right)$.

This is called an **external regret** bound.

Can choose λ to get $\boldsymbol{R} = \tilde{O}\left(\sqrt{\frac{\sqrt{d}}{\varepsilon n}}\right)$. Why is this useful?

Why is this useful?

- Suppose data drawn i.i.d. from P
- Define $L_P(w) := \mathbb{E}_{X\sim P}\ell(w;X)$
- **Lemma** ("online to batch conversion"):

$$\mathbb{E}_{\substack{X_1,\ldots,X_n\sim P\\ coins\ of\ A}}\left(L_P(\overline{w})\right) \le L_P(w^*) + (*).$$

- Proof: Exercise. (Hint: x_t is independent of w_t.)

Proving Main Theorem: Noise Terms

- Consider, for each t, the counterfactual nonprivate iterate

$$w'_t = \arg\min_{w \in C} \left(\sum_i^t g_i + \frac{\lambda}{2} \|w\|^2 \right), \text{roughly } \frac{s_t}{\lambda}.$$

- $\boldsymbol{R} = \frac{1}{n} \sum_t \ell(w_t; x_t) - \frac{1}{n} \sum_t \ell(w^*; x_t) \leq \frac{1}{n} \sum_t \langle g_t, w_t - w^* \rangle$
  - Use the linear tangent to $\ell(\cdot; x_t)$ at w_t
- $\frac{1}{n} \sum_t \langle g_t, w_t - w^* \rangle = \underbrace{\boxed{\frac{1}{n} \sum_t \langle g_t, w_t - w'_t \rangle}}_{\text{“noise”}} + \underbrace{\boxed{\frac{1}{n} \sum_t \langle g_t, w'_t - w^* \rangle}}_{\text{“regret”}}$

- Let's bound the noise term when $C = \mathbb{R}$, for simplicity:
  - $\frac{1}{n} \sum_t \langle g_t, w_t - w'_t \rangle = \frac{1}{n} \sum_t \left\langle g_t, \frac{s_t}{\lambda} - \frac{\tilde{s}_t}{\lambda} \right\rangle = -\frac{1}{n} \sum_t \frac{\langle g_t, Z_t \rangle}{\lambda}$

$$\leq \frac{n}{n} \cdot \frac{\|g_t\| \|Z_t\|}{\lambda} \leq \frac{(Tree\ Mech.\ error)}{\lambda}.$$

Proving Main Theorem: Regret

- We need to bound $\frac{1}{n}\sum_t \langle g_t, w_t' - w^* \rangle$
  - Let $r_t := \langle g_t, w_t' - w^* \rangle$, so we are bounding $\frac{1}{n}\sum r_t$
- Again, assume $C = \mathbb{R}$ for simplicity
- Consider potential function $\Phi_t = \frac{1}{2}\|w_t' - w^*\|^2$.
- **Claim:** $\Phi_t - \Phi_{t+1} \geq \frac{1}{\lambda} r_t - \frac{1}{2\lambda^2}\|g_t\|^2$
  - Proof by expanding Φ_t's as squares
  - Equivalent statement: $r_t \leq \lambda(\Phi_t - \Phi_{t+1}) + \frac{1}{2\lambda}\|g_t\|^2$.

- Thus $\frac{1}{n}\sum r_t \leq \frac{\lambda}{n}\left(\sum_t (\Phi_t - \Phi_{t+1}) + \frac{1}{2\lambda^2}\sum_t \|g_t\|^2\right)$

$$\leq \frac{\lambda \Phi_1}{n} + \frac{1}{2\lambda} = \frac{\lambda\|w^*\|}{n} + \frac{1}{2\lambda}.$$

QED ☺

Today

- What's wrong with DP-SGD and amplification
- Gradient descent as online sums
- Tree mechanism recall
- DP-FTRL mechanism
- Error analysis

Takeaway: DP-FTRL provides

- an alternative to DP-SGD with similar computational performance
- without requiring randomly ordered / sampled data for privacy.

Consequently, it is much easier to implement at scale.