Lecture 4: DP Gradient Descent

Today
- Adaptive composition
- Introduce "vanilla" DP-GD
- A first privacy analysis
- A first utility analysis

Might cover: connections to sampling/VLA, other versions eg Projected DP-GD, DPSGD, state some theorems for basic optimization problems, etc

Setup: want to minimize some loss fn $L(w, X) = \frac{1}{n}\sum_{i=1}^{n} \ell(w; x_i)$ over $w \in \mathbb{R}^d$.

Alg: DP-GD

Input: dataset $X \in \mathcal{X}^n$, clipping threshold $C > 0$, noise scale $\lambda > 0$, num steps T, step size $\eta > 0$, loss fn ℓ

~~Output: iterates $(w_0, w_1, \dots w_T)$~~

1) $w_0 \leftarrow 0$

2) for $t = 1, \dots, T$:

3) $\quad \hat{g}_t \leftarrow \frac{1}{n}\sum_{i=1}^{n} \text{CLIP}_\gamma\left(\nabla_w \ell(w_{t-1}; x_i)\right)$

4) $\tilde{g}_t \leftarrow \hat{g}_t + \mathcal{N}(0, \lambda^2 \mathbb{I}_d)$

5) $w_t \leftarrow w_{t-1} - \eta \tilde{g}_t$

6) Return $(w_1, \dots, w_T)$

$\text{Clip}_B(x)$ is $\min\left\{\frac{x}{\|x\|_2}, B\right\} \cdot x$

A First Privacy Analysis

For $\varepsilon, \delta \in (0,1)$,

Claim If $\lambda \geq \frac{T\sqrt{\log\frac{1}{\delta}}}{\varepsilon n} \cdot C$, (Constants) then Alg 1 is (ε, δ)-DP.

$\lambda \geq \frac{T\sqrt{2\log\frac{2T/\delta}{\varepsilon n}}}{\varepsilon n}$, then

[Before proving, discuss statement
[note proof could be difficult, lots of dependencies]

Proof

Use basic composition: treat each time step as running one alg

$$w_{t+1} \leftarrow A_{GD}(X, w_t)$$

Prove A_{GD} is $(\frac{\varepsilon}{T}, \frac{\delta}{T})$-DP, then overall is (ε, δ)-DP.

"ε's & δ's add up"

Claim [Basic Composition] If $A_1, A_2, \dots, A_T$ are all (ε, δ)-DP, then overall $(A_1(X), A_2(X), \dots, A_T(X))$ is (ε', δ')-DP for $\varepsilon' = T\varepsilon$, $\delta' = T\delta$. This holds even if A_t is picked adaptively.

↖ Better/cleaner statement for adaptivity?

Claim [GaussianMechanism is DP] If $f: X^n \to \mathbb{R}^d$ has ℓ_2-global sensitivity $\leq \Delta_f$, then for $\varepsilon, \delta \in (0,1)$ if $C \geq \frac{2\Delta_f \sqrt{\log 2/\delta}}{\varepsilon}$ then $f(x) + \mathcal{N}(0, C^2 \mathbb{I}_d)$ is (ε, δ)-DP.

In Alg 1, we compute

$$f(x,w) = \frac{1}{n} \sum_{i=1}^{n} \text{CLIP}_\gamma(\nabla_w \ell(w, x_i))$$

Global sensitivity? Fix two datasets X, X'.

$$\|f(x,w) - f(x',w)\|_2$$

$$= \left\| \frac{1}{n} \sum_{i=1}^{n} \text{CLIP}_c(\nabla \ell(w, x_i)) - \frac{1}{n} \sum_{i=1}^{n} \text{CLIP}_c(\nabla \ell(w, x_i')) \right\|_2$$

$$= \frac{1}{n} \left\| \text{CLIP}_c(\nabla \ell(w, x_1)) - \text{CLIP}_c(\nabla \ell(w, x_2)) \right\|_2$$

$$\leq \frac{2C}{n}.$$

So for $\left(\frac{\varepsilon}{T}, \frac{\delta}{T}\right)$-DP, set $\lambda \geq 2T\frac{\sqrt{\log 2T/\delta}}{\varepsilon} \cdot \Delta_f$

$= \frac{4TC\sqrt{\log 2T/\delta}}{\varepsilon n}$. ☒

Talk about what is informal here:
releasing gradients vs new params, postprocessing, etc.

A FIRST Utility Analysis

to be continued

- Lots of DP optimization theory, mirroring non-private literature (not complete!)
- Mostly gradient-based, not entirely
- Considerations for DP-GD:

clipping too big ⇒ lots of noise

clipping too small ⇒ destroy signal

heavy price for increased T