

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/ULA, other versions eg Projected DP-GD, DP-SGD, state some theorems for basic optimization problems, etc

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

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) $\hat{g}_t \leftarrow \frac{1}{n} \sum_{i=1}^n \text{CLIP}_C(\nabla_w \ell(w_{t-1}; x_i))$

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

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

6) Return (w_1, \dots, w_T)

$$\text{CLIP}_\beta(x) \text{ is } \min \left\{ \frac{x}{\|x\|_2}, \beta \cdot x \right\}$$

A First Privacy Analysis

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

Claim If $\lambda \geq \frac{T \sqrt{\log 1/\delta}}{\epsilon n} \cdot C$, then Alg 1

is (ϵ, δ) -DP.

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

Constants

[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 $(\frac{\epsilon}{T}, \frac{\delta}{T})$ -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 $\epsilon' = T\epsilon$, $\delta' = T\delta$. This holds even if A_t is picked adaptively.

↑ Better/cleaner statement for adaptivity?

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

In Alg 1, we compute

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

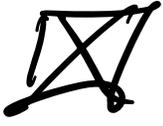
Global sensitivity? Fix two datasets X, X' .

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

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

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

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

So for $(\frac{\epsilon}{T}, \frac{\delta}{T})$ -DP, set $\lambda \geq \frac{2T \sqrt{\log^2 \frac{1}{\delta}}}{\epsilon} \cdot \Delta f$
 $= \frac{4TC \sqrt{\log^2 \frac{1}{\delta}}}{\epsilon n}$ 

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 \Rightarrow lots of
noise

Clipping too small \Rightarrow destroy signal

heavy price for increased T