Private Gradient Descent

- Recap: Private ERM Problem
  Exponential Mechanism
- (Projected) Gradient Descent
  Privacy, Convergence

HW2 Due on Weds

Formulation.

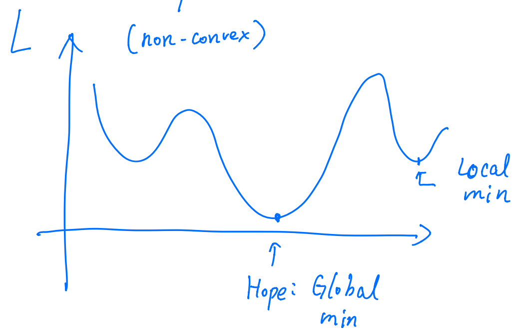
Input : $x_1, \dots, x_n$, $\ell : C \times X \longmapsto \mathbb{R}$

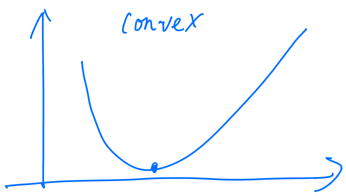
(feasible set of models/parameters)

$\ell(w, x)$ measures "loss"

$L : C \longmapsto \mathbb{R}$

$L(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(w, x_i)$





$\Pi_C(w) = \underset{w' \in C}{\arg\min} \|w - w'\|_2$

"Projection"

Goal: output $\hat{w}$ s.t.

$L(\hat{w}) \approx \min_{w \in C} L(w)$

Projected Gradient Descent (PGD)

learning rate

$PGD(L, C, \eta)$:

→ Init: $w_0 \in C$ arbitrary

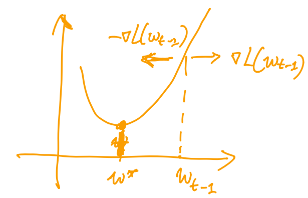
For $t = 1, \dots, T$:

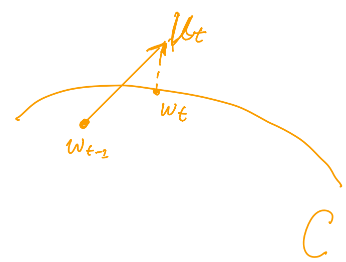
$g_t = \nabla L(w_{t-1})$ ← Gradient

$u_t \leftarrow w_{t-1} - \eta \cdot g_t$

$w_t \leftarrow \Pi_C (u_t)$.

→ Output $\hat{w} = \frac{1}{T} \sum_{t=1}^{T} w_t$





Robustness to noise in gradient estimation. (g_t)

→ For privacy: $\hat{g}_t = g_t + \underbrace{N(0, \sigma^2 I_d)}$ → d-dim i.i.d. Gaussian

$\mathbb{E}[\hat{g}_t] = g_t = \nabla L(w_{t-1})$

→ For efficiency: $\hat{g}_t = \nabla L(w_{t-1}, X_{I_t})$, $I_t \leftarrow unif\{1, .., n\}$

$\mathbb{E}[\hat{g}_t] = g_t = \nabla L(w_{t-1})$

"Unbiased"

Noisy/Private PGD $\left(\text{outputs } \hat{w} = \frac{1}{T}\sum_{t=1}^{T} w_t\right)$.

- Proof idea for privacy: think of releasing $w_1, \ldots, w_T$

  Suffices to release $w_0 \xleftrightarrow[\tilde{g}_1]{\text{diff}} w_1 \xleftrightarrow[\tilde{g}_2]{} w_2 \cdots$

  Suffices to release $\underbrace{\tilde{g}_1, \tilde{g}_2, \ldots, \tilde{g}_T}_{\text{Adaptive Composition of Gaussian Mechanism.}}$

Lemma. If ℓ is G-Lipschitz for every x. $(\|\nabla \ell(w;x)\| \le G)$ on C,
$\left(|\ell(w,x) - \ell(w',x)| \le G\|w-w'\|\right)$.

then Noisy PGD is (ε,δ)-DP when $\left[\sigma \ge \frac{2G}{n} \cdot \frac{\sqrt{2T\ln(1/\delta)}}{\varepsilon}.\right]$

Proof Ideas:

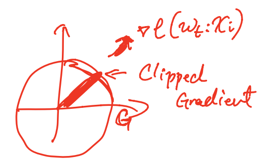
① Overall ℓ_2-sensitivity of all T gradients: $\frac{2G}{n} \cdot \sqrt{T}$

$$\nabla L(w_t) = \frac{1}{n}\sum_{i=1}^{n} \nabla \ell(w_t; x_i)$$

② $\underbrace{\text{Apply Advanced Composition}}_{\text{loose bound}}$ to T adaptive Gaussian Mech

Better techniques: Renyi DP accountant (Tensorflow)
Gaussian DP " — "
(zero) - Concentrated DP " — "

NB: Enforce low sensitivity by "Gradient Clipping" on $\nabla \ell(w_t; x_i)$



Private SGD.

learning rate

$\text{Private SGD}(L, C, \eta)$:

→ Init: $w_0 \in C$ arbitrary

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

Also . Subsampling a minibatch

$$I_t \leftarrow \text{unif}(\{1, \dots, n\}) ; \; g_t = \nabla \ell(w_{t-1}; x_{I_t})$$

$$\tilde{g}_t = g_t + N(0, \sigma^2 I_d)$$

$$u_t \leftarrow w_{t-1} - \eta \cdot \tilde{g}_t$$

$$w_t \leftarrow \Pi_C (u_t).$$

→ Output $\hat{w} = \frac{1}{T} \sum_{t=1}^{T} w_t$

SG Langevin Dynamics.

SGD+ Gaussian Noise.

How to analyze Privacy?

Privacy Amplification

- Keep I_t secret
- Use their randomness.

In general: $A : \mathcal{X} \mapsto \mathcal{Y}$ is (ε, δ)-DP.

take one data point

- Consider: $A' : \mathcal{X}^n \mapsto \mathcal{Y}$
$$\begin{cases} I \leftarrow \text{unif}(\{1, \dots, n\}) \\ \text{Return } A(x_I) \end{cases}$$

- A' is (ε', δ')-DP where
$$\varepsilon' = \ln\left(1 + \frac{e^{\varepsilon} - 1}{n}\right) \approx \frac{\varepsilon}{n} \quad \text{for } \varepsilon \le 1$$
$$\delta' = \frac{\delta}{n}$$

Can generalize to sub-sample of size $\boxed{m \le n}$.

$$\varepsilon' \approx \frac{m}{n} \varepsilon$$
$$\delta' \approx \frac{m}{n} \delta.$$

- Adding Gaussian noise w/ $\sigma = \frac{2G}{\varepsilon}\sqrt{2\ln(1/\delta)}$ is (ε,δ)-DP.

$$\underline{\left[\nabla \ell(w_{t-1}; x) + N(0, \sigma^2 I)\right]}$$

- Subsampling + Gaussian is (ε', δ')-DP

$$\underline{\varepsilon' = \frac{\varepsilon}{n}, \quad \delta' = \frac{\delta}{n}.}$$

- Advanced Composition. over T iterations.

$$\left(\frac{\varepsilon\sqrt{T}}{n}, \frac{\delta \cdot T}{n}\right)\text{-DP}$$

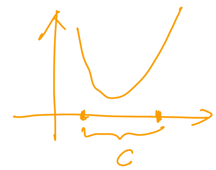
(ignoring constants).

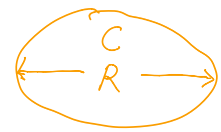
Set δ to be smaller than $\frac{1}{n}$.

Can $\delta = \frac{1}{n^5}$

Can Run $T = n^2$ iterations!

Convergence / Optimality.





Theorem. Let $L : C \longrightarrow \mathbb{R}$ be convex and G-Lipschitz, $C \subseteq \mathbb{R}^d$ be a closed and convex set with diameter R

(Part a)
↓

$$\boxed{w^* \in \underset{w \in C}{\arg\min}\ L(w)}$$

- For regular PGD, set $\eta = \frac{R}{G\sqrt{T}}$, then $L(\hat{w}) - L(w^*) \leq \frac{RG}{\sqrt{T}}$ $\downarrow 0$

- For noisy PGD, set η, T, σ^2 so that, $\mathbb{E}\left[L(\hat{w}) - L(w^*)\right] \leq O\left(\frac{RG\sqrt{d \ln(1/\delta)}}{n\varepsilon}\right)$

"Cost of privacy" Gap: $\frac{\sqrt{d}}{n\varepsilon}$ ← "tight" in the worst-case

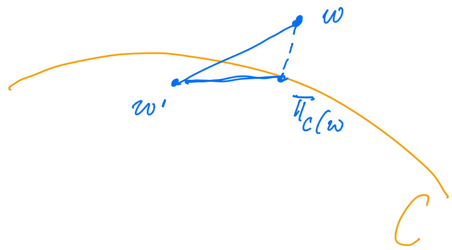
Gap for EM: $\frac{d}{n\varepsilon}$

Projection Lemma

Lemma If $C \subseteq \mathbb{R}^d$ is closed and convex.

Then for any $w \in \mathbb{R}^d$, $w' \in C$

$$\| \Pi_C(w) - w' \| \leq \| w - w' \|$$



Proof (for regular PGD).

$$w^* = \underset{w \in C}{\operatorname{argmin}} \; L(w)$$

2 Key Quantities

Excess Risk

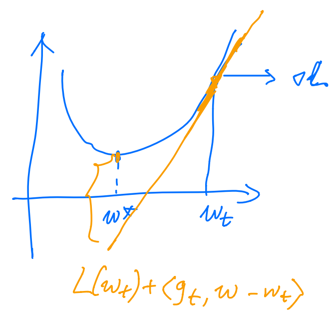
Distance to w^*

Claim. (Measure of Progress).

$$\underbrace{L(w_t) - L(w^*)}_{\text{Excess Risk}} \le \frac{\eta \cdot \|g_t\|^2}{2} + \frac{1}{2\eta}\left(\|w_t - w^*\|^2 - \|w_{t+1} - w^*\|^2\right)$$

Squared distances.

Proof. $L(w^*) \ge L(w_t) + \langle g_t, w^* - w_t \rangle$



$$L(w_t) - L(w^*) \le \underline{\frac{1}{\eta} \langle \eta g_t, w_t - w^* \rangle}$$

$$\boxed{\begin{array}{l} \forall a, b \in \mathbb{R}^d \\ \quad \langle a, b \rangle = \frac{1}{2}\left(\|a\|^2 + \|b\|^2 - \|a-b\|^2\right) \end{array}}$$

$$\begin{aligned} L(w_t) - L(w^*) &\le \frac{1}{2\eta}\left(\|\eta g_t\|_2^2 + \|w_t - w^*\|^2 - \|w_t - w^* - \eta g_t\|^2\right) \\ &= \frac{\eta \cdot \|g_t\|^2}{2} + \frac{1}{2\eta}\left(\|w_t - w^*\|^2 - \|u_t - w^*\|^2\right) \\ &\le \frac{\eta \cdot \|g_t\|^2}{2} + \frac{1}{2\eta}\left(\|w_t - w^*\|^2 - \|w_{t+1} - w^*\|^2\right) \quad \text{(Projection)} \end{aligned}$$

Noisy / Private PGD.