Private Gradient Descent

Recap:

- (Projected) Gradient Descent
  Privacy, Convergence
- Practical Aspects of Private Deep Learning (w/ slides)

$\boxed{\text{HW2 Due}}$

Projected Gradient Descent (PGD)

Constraint set

learning rate

$$\mathrm{PGD}(L, C, \eta):$$

→ Init: $w_0 \in C$ arbitrary

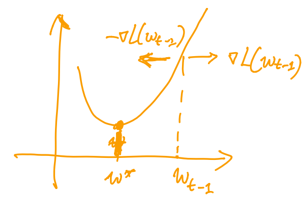
For $t = 1, \dots, T$:

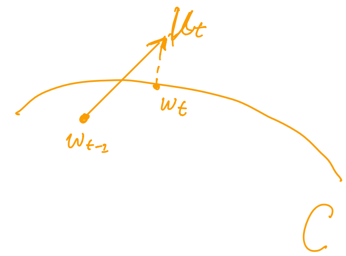
$$g_t = \nabla L(w_{t-1}) \leftarrow \text{Gradient}$$

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

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

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





Private SGD.

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

→ Init: $w_0 \in C$ arbitrary

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

Also · Subsampling a minibatch → $I_t \leftarrow \text{unif}(\{1, \ldots, n\})$; $g_t = \nabla \ell(w_{t-1}; x_{I_t})$

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

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

$w_t \leftarrow \Pi_C(u_t)$.

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

Privacy Amplification

- Keep I_t secret
- Use their randomness.

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

Computation at each step

take one data point

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

- A' is (ε', δ')-DP where

$$\varepsilon' = \ln\left(1 + \frac{e^{\varepsilon} - 1}{n}\right) \approx \boxed{\frac{\varepsilon}{n}} \quad \text{for } \varepsilon \leq 1$$

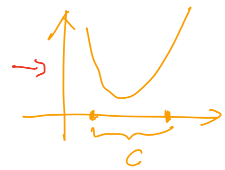
$$\delta' = \frac{\delta}{n}$$

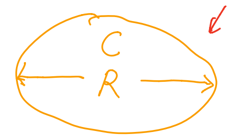
Can generalize to sub-sample of size $\boxed{m \leq n}$.

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

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

Convergence / Optimality





Theorem. Let $L : C \longrightarrow \mathbb{R}$ be convex and $\underline{G\text{-Lipschitz}}$

$C \subseteq \mathbb{R}^d$ be a closed and convex set with $\underline{\text{diameter } R}$

(Part a)

$\downarrow$

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

- For regular PGD, set $\eta = \frac{R}{G\sqrt{T}}$, then $L(\hat{w}) - L(w^*) \leq \boxed{\frac{RG}{\sqrt{T}}}$ $\downarrow 0$ $T \nearrow \infty$
- For noisy PGD, set η, T, δ^2 so that, $\mathbb{E}\left[L(\hat{w}) - L(w^*)\right] \leq O\left(\frac{RG\sqrt{d \ln(1/\delta)}}{n\varepsilon}\right)$

For theory: $T \approx n^2$

Practice: Trial-&-error. (:so is learning rate)

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

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

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).

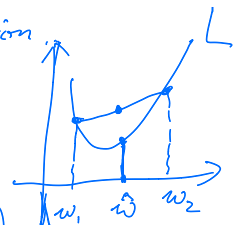
$$\underbrace{L(w_t) - L(w^*)}_{\text{Excess Risk}} \le \frac{\eta \cdot \|g_t\|^2}{2} + \frac{1}{2\eta}\Big(\underbrace{\|w_t - w^*\|^2 - \|w_{t+1} - w^*\|^2}_{\text{Reduction on Squared distances.}}\Big)$$

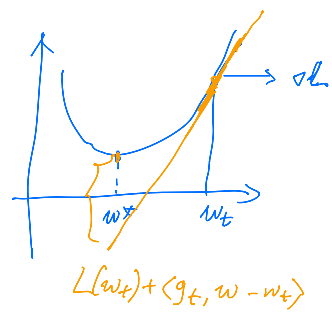
Proof for $\hat{w} = \frac{1}{T}\sum_{t=1}^{T} w_t$

By Jensen Inequality. for Convex function.

$$L(\hat{w}) \le \underbrace{\frac{1}{T}\sum_t L(w_t)}$$

Compare w/ $\frac{1}{T}\left(T \cdot L(w^*)\right)$





$$L(\hat{w}) - L(w^*) \le \frac{1}{T}\left(\sum_t \left(L(w_t) - L(w^*)\right)\right) \quad \leftarrow \text{use "Progress Claim"}$$

$$\le \frac{\eta}{2} \cdot \max_t \|g_t\|^2 + \frac{1}{2\eta T}\left(\|w_1 - w^*\|_2^2 - \|w_{T+1} - w^*\|^2\right)$$

$$\le \frac{\eta}{2} \cdot G^2 + \frac{1}{2\eta T}\left(\|w_1 - w^*\|_2^2\right)$$

$$\le \underbrace{\frac{\eta}{2} G^2 + \frac{R^2}{2\eta T}}_{\text{Equalize}} \underset{\text{Set } \eta = \frac{R}{G} \cdot \frac{1}{\sqrt{T}}}{=} \frac{GR}{\sqrt{T}}$$

Noisy / Private PGD.

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

"New" Progress Claim.

$$\mathbb{E}[L(w_t) - L(w^*)] \le \frac{\eta}{2}\mathbb{E}[\|\tilde{g}_t\|^2] + \frac{1}{2\eta}\mathbb{E}[\|w_t - w^*\|^2 - \|w_{t+1} - w^*\|^2]$$

Proof. $\mathbb{E}[L(w_t) - L(w^*)] \le \mathbb{E}_t[\langle \eta g_t, w_t - w^* \rangle]$

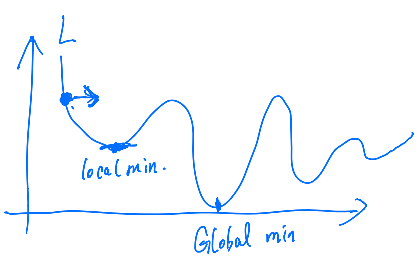
$$= \mathbb{E}_{w_t}[\langle \eta\, \mathbb{E}[\tilde{g}_t | w_t], w_t - w^* \rangle]$$

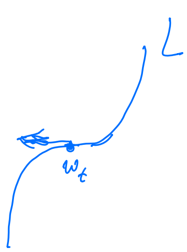
$$= \mathbb{E}[\langle \eta\, \tilde{g}_t, w_t - w^* \rangle]$$

$$\langle a, b \rangle = \|a\|_2^2 + \|b\|_2^2 - \|a-b\|_2^2$$

$$\mathbb{E}[\|\tilde{g}_t\|_2^2] \le \|g_t\|_2^2 + \boxed{d\sigma^2}$$

What about Nonconvex Case?





Smoothness.
(Lipschitz Gradient)

$$\|\nabla L(w) - \nabla L(w')\|_2 \le \beta \|w - w'\|_2$$

$$L(w') \le L(w) + \nabla L(w)^T (w' - w) + \frac{\beta}{2}\|w - w'\|^2.$$

Can Show: $w_1, \dots, w_T$

$$\frac{1}{T}\sum_t \|\nabla L(w_t)\|_2^2 \longrightarrow O\left(\frac{1}{\sqrt{T}}\right). \quad \text{(non-DP)}$$

$$\longrightarrow \frac{\sqrt{d}}{n\varepsilon}\sqrt{\ln\left(\frac{1}{\delta}\right)} \quad \text{(DP)}$$