Lecture 17

- Private Machine Learning
  - DP SGD
  - Privacy Analysis
  - DPSGD in code.

Announcement : Release HW3 this week

includes : ① Written Component

② Programming component.

Recitation (in-person)

feasible set of parameters/weights w

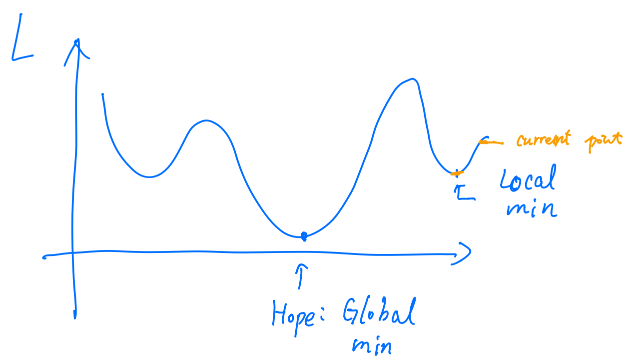
$$\ell : C \times X \longmapsto \mathbb{R}$$

C: feasible set of parameters/weights w; X: data domain

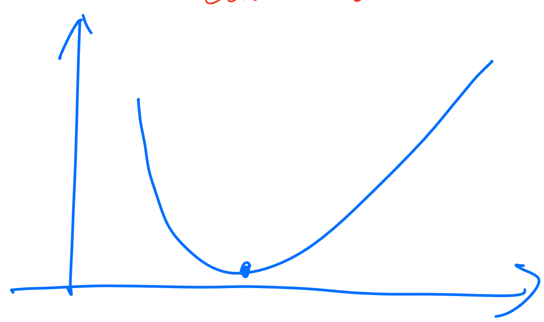
$\ell(w, x_i) \leftarrow$ loss of $w \in C$ on example x_i.

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

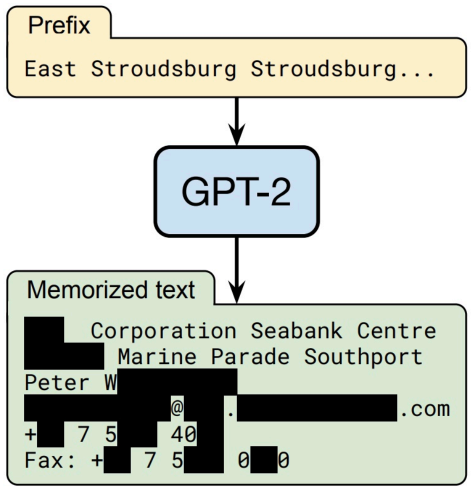
Non-convex loss



Convex loss



"Memorization" Attack.



Extracting Training Data from Large Language Models

Nicholas Carlini[1] Florian Tramèr[2] Eric Wallace[3] Matthew Jagielski[4]
Ariel Herbert-Voss[5,6] Katherine Lee[1] Adam Roberts[1] Tom Brown[5]
Dawn Song[3] Úlfar Erlingsson[7] Alina Oprea[4] Colin Raffel[1]
[1]*Google* [2]*Stanford* [3]*UC Berkeley* [4]*Northeastern University* [5]*OpenAI* [6]*Harvard* [7]*Apple*

Private SGD. (DP-SGD)

$$\text{Private SGD}\Big(\underset{\text{Loss}}{L(\cdot) = \frac{1}{n}\sum_{i=1}^{n} \ell(\cdot\,; x_i)}, \underset{\text{feasible set}}{C}, \underset{\text{learning rate}}{\eta}, \underset{\text{Noise rate}}{\sigma} \Big):$$

Init: $w_0 \in C$

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

Random subsample $B_t \subseteq \{1, \ldots, n\}$
"mini-batch"

$$g_t = \frac{1}{|B_t|} \sum_{i \in B_t} \nabla_w \ell(w_{t-1}; x_i)$$

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

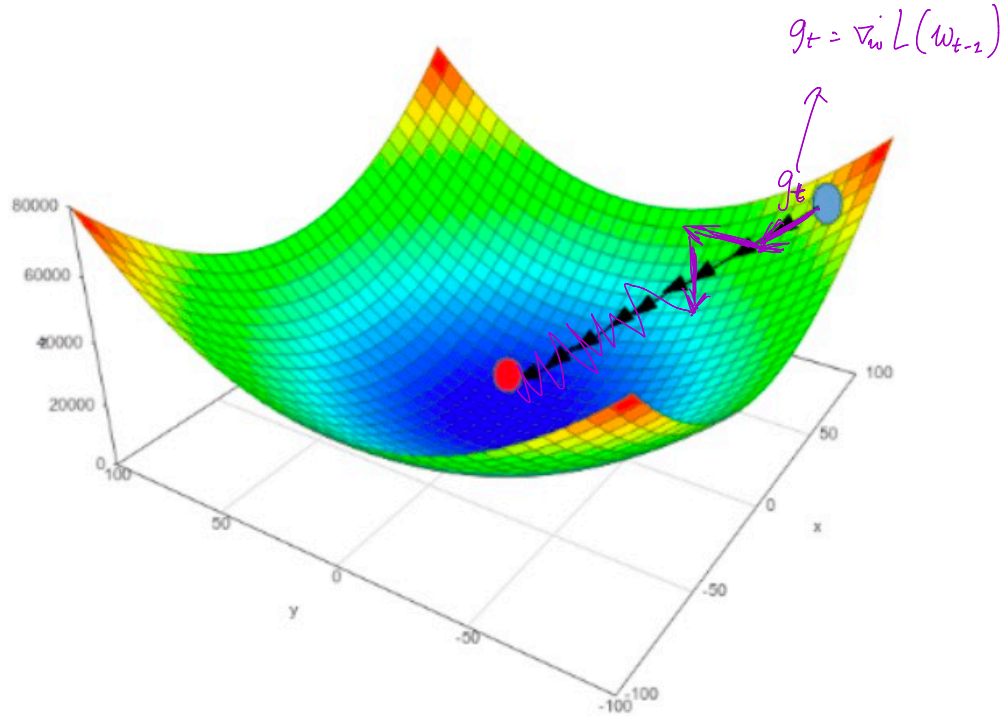
$$u_t = w_{t-1} - \eta \cdot \tilde{g}_t$$

$$w_t = \arg\min_{w \in C} \|w - u_t\|_2$$

Output: $\frac{1}{T} \sum_{t=1}^{T} w_t$

or

w_T



$g_t = \nabla_w L(w_{t-1})$
g_t
80000
60000
40000
20000
0
100
50
0
-50
-100
y
x

Privacy Proof

Proof idea: • Think of releasing $w_1, w_2, \ldots, w_T$.

• Suffices to release the update between iterates

$$w_0 \; [\longleftarrow \text{update} \longrightarrow] \; w_1 \; [\longleftarrow \text{update} \longrightarrow] \; w_2 \; \ldots\ldots \; w_T$$

• Suffices to release the sequence of gradient estimates

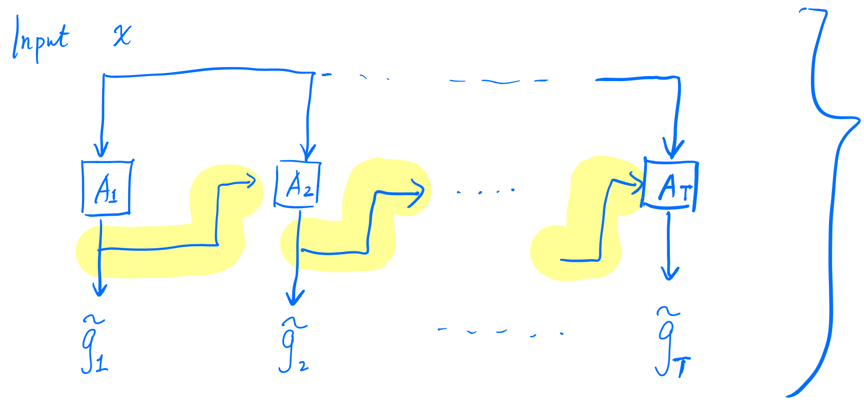
$$\underbrace{\tilde{g}_1, \tilde{g}_2, \ldots, \tilde{g}_T}$$

The output is a post-processing

Show releasing $(\tilde{g}_1, \tilde{g}_2, \ldots, \tilde{g}_T)$ satisfies DP.

1. Each step (releasing $\tilde{g}_t$) satisfies (ε, δ)-DP
2. Adaptive composition across T steps

Adaptive Composition.



Composition A is $(\tilde{\varepsilon}, \tilde{\delta})$-DP

Suppose each step $A_1, \dots, A_T$ is (ε, δ)-DP.

What are the values of $\tilde{\varepsilon}$ and $\tilde{\delta}$?

- (Basic) Composition : $\tilde{\varepsilon} = T\varepsilon$, $\tilde{\delta} = T\delta$.
- Advanced Composition : $\tilde{\varepsilon} = \varepsilon \cdot \sqrt{2T \ln\left(\frac{1}{\delta'}\right)} + T \cdot \varepsilon \frac{e^{\varepsilon} - 1}{e^{\varepsilon} + 1}$

$$\tilde{\delta} = T\delta + \delta'$$

If $\varepsilon < \frac{1}{\sqrt{T}}$

$(\varepsilon\sqrt{T})^2$ is "smaller" than $\varepsilon\sqrt{T}$

Then $\tilde{\varepsilon}$ is in the order of $\varepsilon \cdot \sqrt{T \ln\left(\frac{1}{\delta'}\right)}$

$\ll \varepsilon \cdot T$

for large T.

Numeric Example.

$\varepsilon = \frac{1}{1000}$, δ

$T = 500$,

Basic Composition : $\tilde{\varepsilon} = 0.5$, $\tilde{\delta} = T \cdot \delta$

Advanced Composition : $\tilde{\varepsilon} \leq 0.1$, $\tilde{\delta} = 10^{-6} + T\delta$

Privacy Proof

Proof idea: • Think of releasing $w_1, w_2, \ldots, w_T$.

• Suffices to release the update between iterates

$$w_0 \, [\longleftarrow \text{update} \longrightarrow] \, w_1 \, [\longleftarrow \text{update} \longrightarrow] \, w_2 \ldots\ldots \, w_T$$

• Suffices to release the sequence of gradient estimates

$$\underbrace{\tilde{g}_1, \tilde{g}_2, \ldots, \tilde{g}_T}$$

The output is a post-processing

Show releasing $(\tilde{g}_1, \tilde{g}_2, \ldots, \tilde{g}_T)$ satisfies DP.

This step. $\longrightarrow$ ① Each step (releasing $\tilde{g}_t$) satisfies (ε, δ)-DP

② Adaptive composition across T steps ✓

Multivariate Gaussian Mechanism

e.g. average gradient

$$f : \mathcal{X}^n \mapsto \mathbb{R}^d$$

$$\Delta_2(f) = \max_{\substack{x, x' \\ \text{neighbors}}} \| f(x) - f(x') \|_2$$

L_2 sensitivity.

$$A(x) = f(x) + N\Big(0, \underbrace{\frac{2\Delta_2^2 \log(2/\delta)}{\varepsilon^2}}_{\sigma^2} I_{d\times d}\Big)$$

per-coordinate variance.

$$f(x) \quad \begin{bmatrix} f(x)_1 \\ f(x)_2 \\ \vdots \\ f(x)_d \end{bmatrix} \begin{array}{l} + N(0, \sigma^2) \\ + N(0, \sigma^2) \\ \vdots \\ + N(0, \sigma^2) \end{array}$$

"fresh" Independent noise

Theorem: $\forall\ \varepsilon \le 1,\ \delta > 0$

$A(\cdot)$ satisfies (ε, δ)-DP.

What is Δ_2?

$$f : \mathcal{X}^n \mapsto \mathbb{R}^d$$

$$\Delta_2(f) = \max_{\substack{x, x' \\ \text{neighbors}}} \| f(x) - f(x') \|_2$$

Average gradient

$$f \longleftarrow \frac{1}{|B_t|} \sum_{i \in B_t} \nabla_w \ell(w; x_i)$$

Suppose $|B_t| = 1$.

$$f \leftarrow \nabla_w \ell(w; x_i)$$

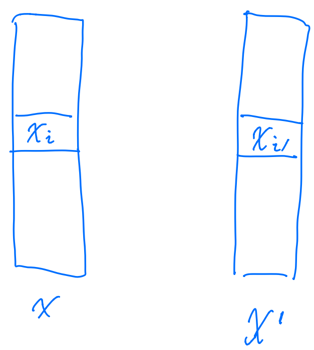
$$\Delta_2 = \max_{x_i, x_i'} \| \nabla_w \ell(w; x_i) - \nabla_w \ell(w; x_i') \|_2$$

In theory, we make assumption on ℓ so that $\Delta_2 \leq G$

Lipschitzness

$\simeq$ Changing w a bit does not change $\ell(w; x_i)$ too much for $x_i \in \mathcal{X}$.

Privacy Amplification by Sub-sampling.



→ first sample minibatch B_t of size 1.

→ Compute gradient using B_t.

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

an algo that takes input a data set of size 1.
(e.g., add gaussian noise to the gradient of one example)

Consider $A' : \mathcal{X}^n \mapsto \mathcal{Y}$

Random index

$I \leftarrow \text{unif}\{1, \dots, n\}$

Return $A(x_I)$

A' is (ε', δ')-DP where

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

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

Can generalize to $|B_t| > 1$.

$$\varepsilon' \approx \frac{|B_t|}{n} \varepsilon, \quad \delta' \approx \frac{|B_t|}{n} \cdot \delta$$

Wrapping up the privacy proof.

- For each step: sub-sampled Gaussian mechanism
- Apply Adaptive Composition.

DP-SGD (in Theory)

Init: $w_0 \in C$

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

Random subsample $B_t \subseteq \{1, \ldots, n\}$
"mini-batch"

$$g_t = \frac{1}{|B_t|} \sum_{i \in B_t} \nabla_w \ell(w_{t-1}; x_i)$$

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

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

$$w_t = \arg\min_{w \in C} \|w - u_t\|_2$$

Assume
ℓ is Lipschitz
or
gradient $\nabla_w \ell(w; x_i)$
for every $w \in C$ & $x_i \in X$

DP-SGD (in practice)

For $t = 1, \ldots T$

Sample mini-batch $B_t \subseteq \{1, \ldots, n\}$

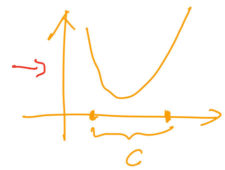
$$g_t = \frac{1}{|B_t|} \sum_{i \in B_t} \text{Clip}(\nabla_w \ell(w; x_i), G)$$

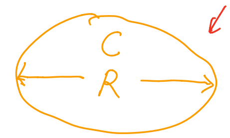
shrink gradient if too large.

$$\text{Clip}(g, G) = g \min\left(1, \frac{G}{\|g\|_2}\right).$$

$$\tilde{g}_t = g_t + \text{Gaussian Noise}.$$

Convergence / Optimality.





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

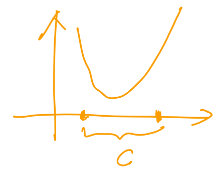
$C \subseteq \mathbb{R}^d$ be a closed and convex set with diameter R

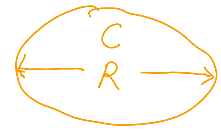
(Part a)

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

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

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)

$\downarrow$

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

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

"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).

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

2 Key Quantities: Excess Risk, Distance to w^*

Claim. (Measure of Progress).

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

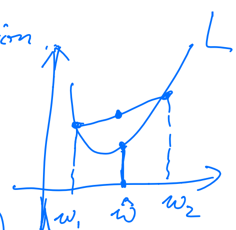
Reduction on Squared distances.

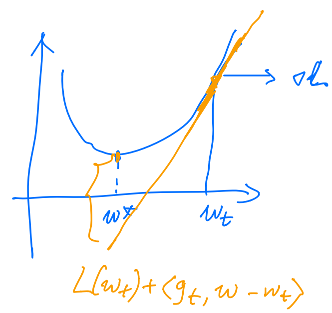
Proof for $\hat{w} = \frac{1}{T}\sum_{t=1}^{T} w_t$

By Jensen Inequality for convex function.

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

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





$$L(\hat{w}) - L(w^*) \le \frac{1}{T}\left(\sum_t \left(L(w_t) - L(w^*)\right)\right)$$ ← 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.

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

Proof.

$$E[L(w_t) - L(w^*)] \le E_t[\langle \eta g_t, w_t - w^* \rangle]$$

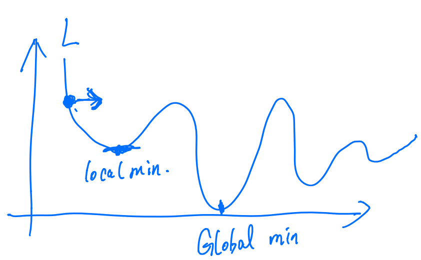
$$= E_{w_t}[\langle \eta E[\tilde{g}_t | w_t], w_t - w^* \rangle]$$

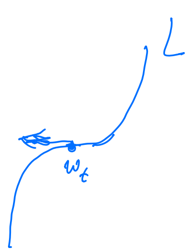
$$= E[\langle \eta \tilde{g}_t, w_t - w^* \rangle]$$

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

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

What about Noncovex 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: $w_1, \dots, w_T$

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

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