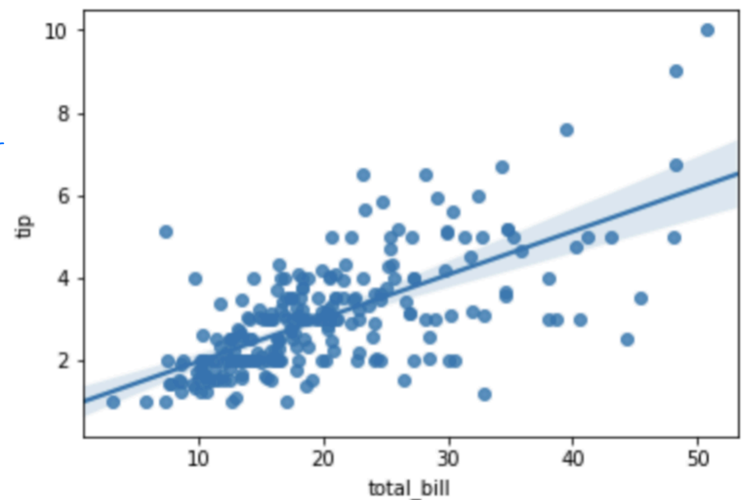
Lecture 16

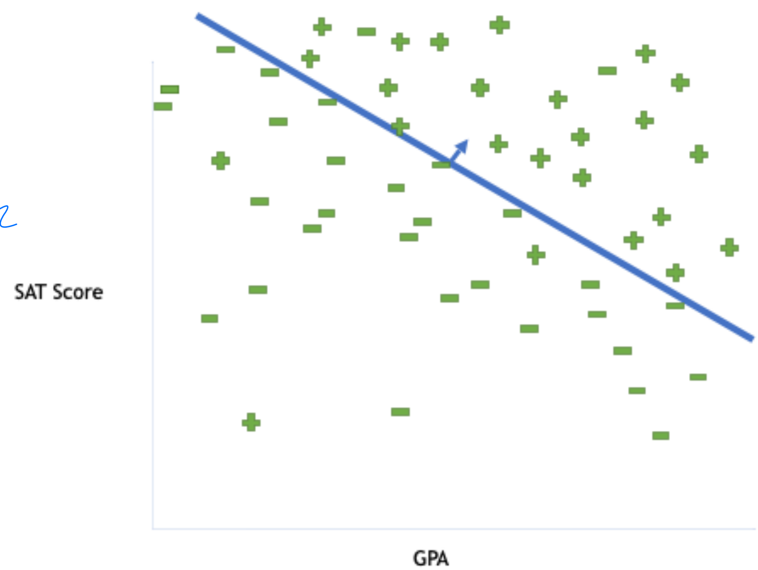
- Private Machine Learning
  - DP Gradient Descent
  - Privacy Analysis.

Optimization in ML

Linear Regression



Linear Classification



(Private) Optimization.

Given a data set $X = (x_1, \ldots, x_n)$

loss function: ℓ

feasible set of parameters (weights): $C \subseteq \mathbb{R}^d$

Empirical Risk Minimization (ERM):

$$\min_{w \in C} \underbrace{L(w; x)}_{\text{Empirical Risk.}} = \frac{1}{n} \sum_{i=1}^{n} \ell(w; x_i) \quad + \quad \text{Optional Regularization}$$

e.g., $\lambda \|w\|_1$

Often: $C = \mathbb{R}^d$

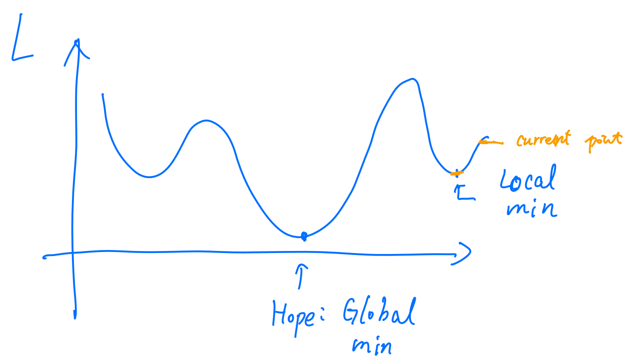
$$\ell : C \times X \longmapsto \mathbb{R}$$

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

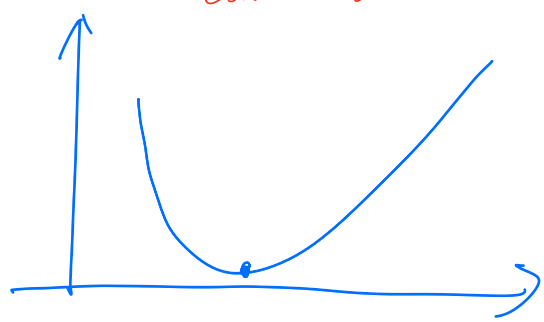
$$L : C \longmapsto \mathbb{R}$$

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

Non-convex loss



Convex loss



Projected Gradient Descent (PGD)

$$\mathrm{PGD}\big(\overset{\text{Loss}}{L},\ \overset{\text{Feasible Set}}{C},\ \overset{\text{Learning Rate}}{\eta},\ \overset{\text{\# rounds}}{T}\big):$$

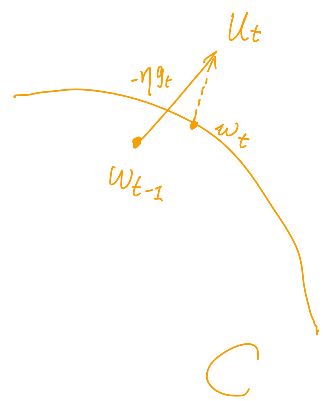
Init: $w_0 \in C$ (any point)

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

gradient: $g_t = \nabla L(w_{t-1})$

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

projection: $w_t \leftarrow \underset{w \in C}{\mathrm{argmin}} \|w - u_t\|_2$

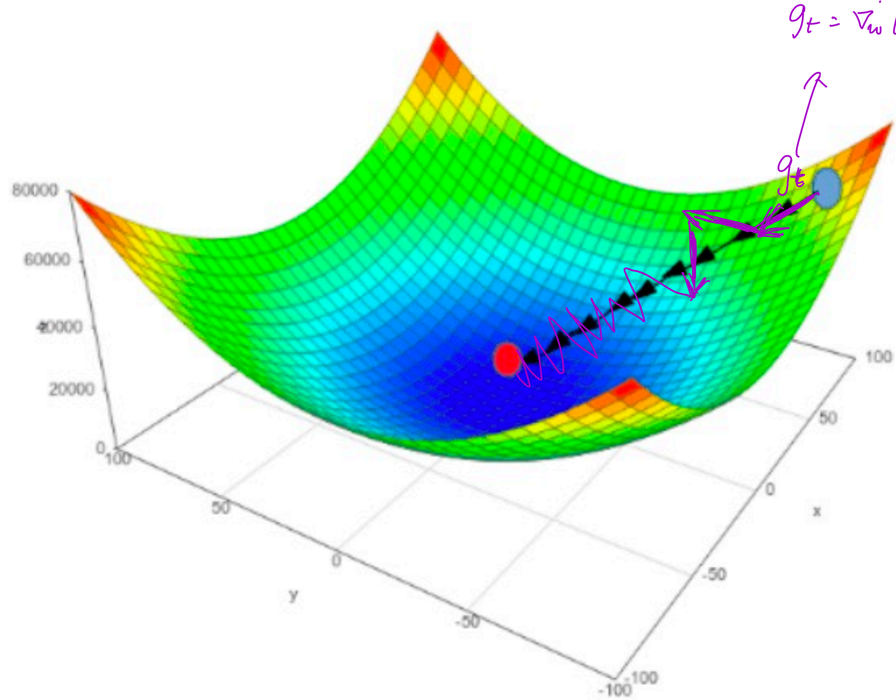


Output: w_T ← Last iterate

or

$\frac{1}{T}\sum_{t=1}^{T} w_t$ ← Average iterate.

$g_t = \nabla_w L(w_{t-1})$



Robustness to noise in gradient estimation. (g_t)

Two sources of noise :

→ For efficiency :

Sample a minibatch $\underset{50}{B} \subseteq \{1, 2, \dots, \underset{50 \text{ million}}{n}\}$

gradient estimate $\tilde{g}_t = \frac{1}{|B|} \sum_{i \in B} \nabla_w L(w_{t-1}, x_i)$.

→ For privacy : Add Gaussian Noise

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

from Gaussian mech.

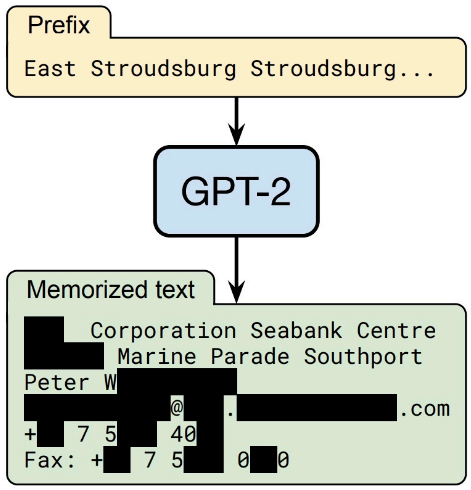
In both cases,

$\tilde{g}_t$ is an unbiased estimate of g_t

$$\mathbb{E}[\tilde{g}_t] = g_t$$

Stochastic Gradient Descent (SGD)

"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)

loss — feasible set — learning rate — Noise rate

$$\text{Private SGD}\Big(L(\cdot) = \frac{1}{n}\sum_{i=1}^{n} \ell(\cdot\,; x_i),\ C,\ \eta,\ \sigma \Big):$$

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.

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

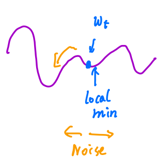
$$w_t = \arg\min_{w \in C} \| w - u_t \|_2$$

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

or

w_T

Stochastic Gradient Langevin Dynamics (SGLD)



Privacy Proof

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

• Suffices to release the update between iterates

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

• Suffices to release the sequence of gradient estimates

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

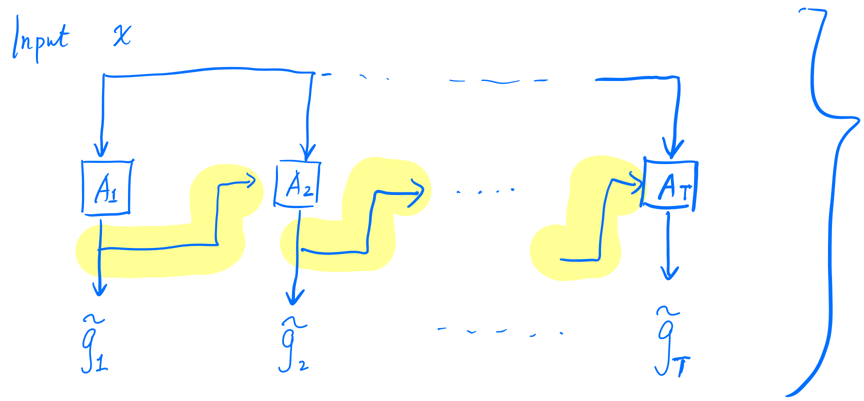
The output is a post-processing

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: $\tilde{\varepsilon} = \varepsilon \cdot \sqrt{2T \ln(\frac{1}{\delta'})} + T \cdot \varepsilon \frac{e^{\varepsilon} - 1}{e^{\varepsilon} + 1}$

$\tilde{\delta} = T\delta + \underbrace{\delta'}_{>0}$

for $\varepsilon \le 1$, $e^{\varepsilon} \approx 1 + \varepsilon \Rightarrow \frac{e^{\varepsilon}-1}{e^{\varepsilon}+1} \approx \frac{\varepsilon}{2}$

$\Rightarrow T \cdot \varepsilon \frac{e^{\varepsilon}-1}{e^{\varepsilon}+1} \approx \frac{T}{2} \cdot \varepsilon^2$

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

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

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

Numeric Example.

$\varepsilon = \frac{1}{1000}$, δ

$T = 500$,

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

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