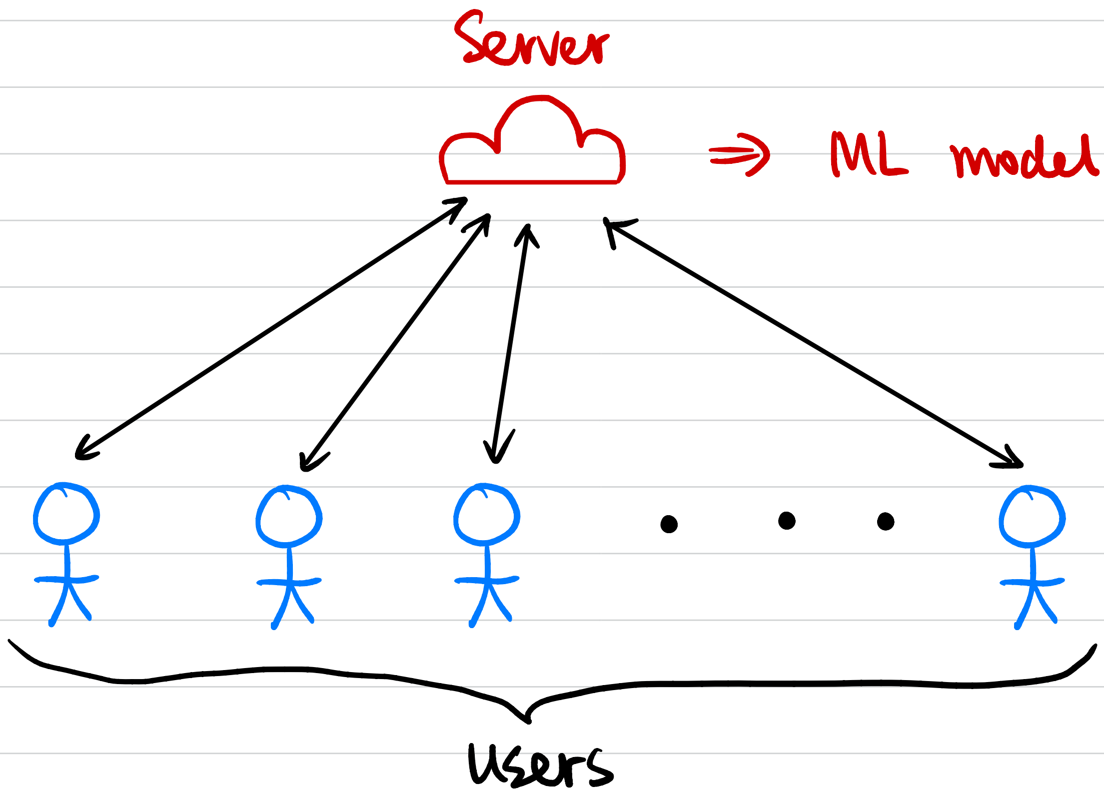
CSCI 1515 Applied Cryptography

This Lecture:

- Federated Learning
- Differential Privacy
- Elliptic Curve Cryptography

Federated Learning (FL)



Application: Google mobile keyboard prediction

Machine Learning Background

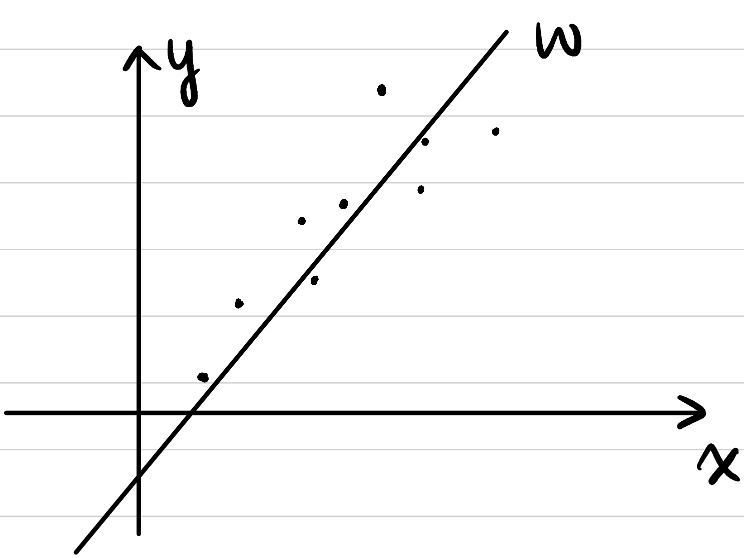
Linear Regression

Data Points $(\vec{x}, y)$

ML Model: coefficient vector $\vec{w}$

$$g(\vec{x}) = \langle \vec{x}, \vec{w} \rangle$$

Goal: Find $\vec{w}$ that minimizes $L(\vec{w})$.



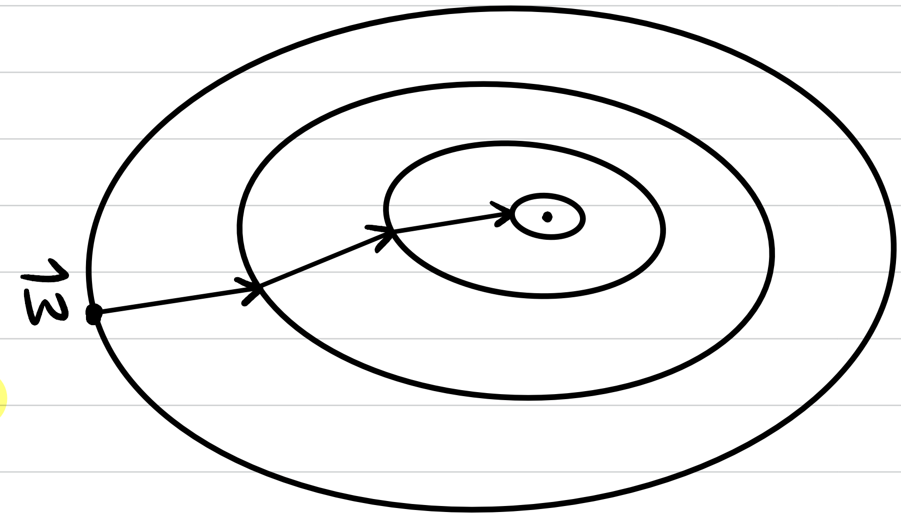
FL for Linear Regression

Stochastic Gradient Descent (SGD)

- $\vec{w}$ initialized with arbitrary value
- Given a data point $(\vec{x}_i, y_i)$:

$$\vec{w} \leftarrow \vec{w} - \eta \cdot \nabla L_i(\vec{w})$$

$$\vec{w} \leftarrow \vec{w} - \eta \cdot (\langle \vec{x}_i, \vec{w} \rangle - y_i) \cdot \vec{x}_i$$



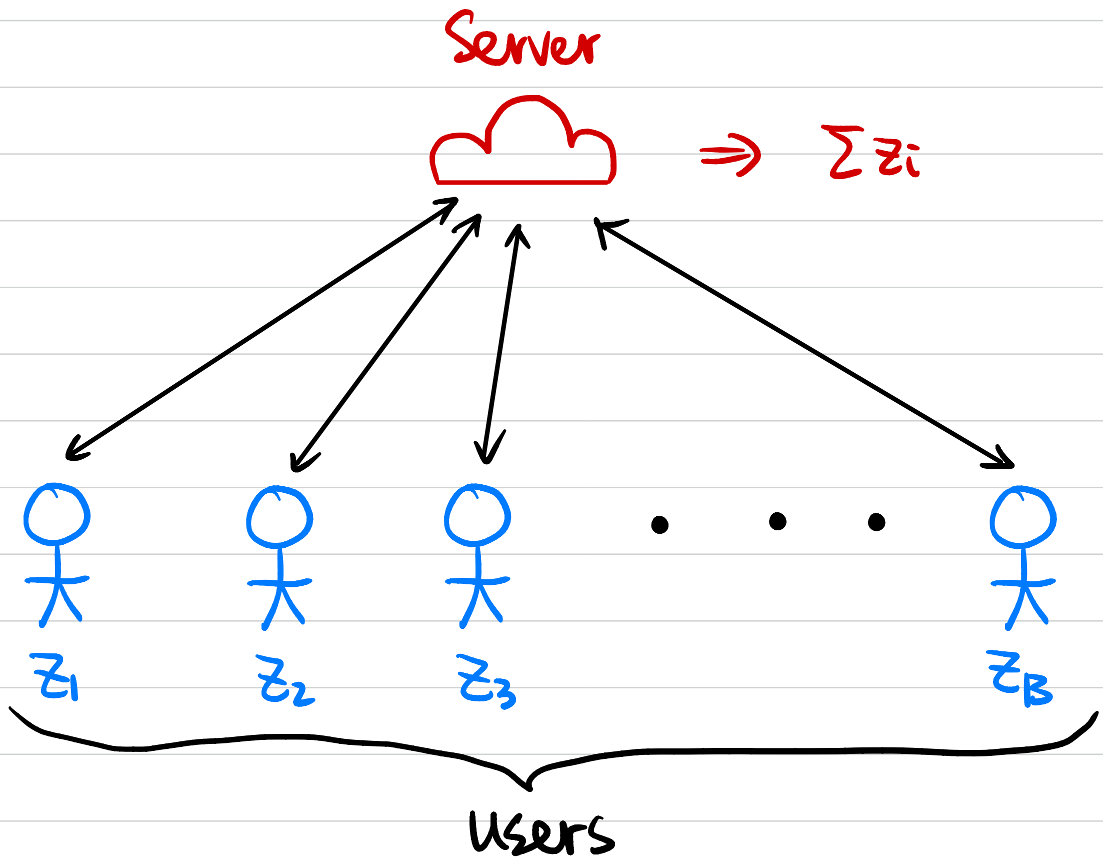
Batch SGD:

$$\vec{w} \leftarrow \vec{w} - \frac{\eta}{B} \cdot \sum_{i \in [B]} \nabla L_i(\vec{w})$$

$$\vec{w} \leftarrow \vec{w} - \frac{\eta}{B} \cdot X_B^T \cdot (X_B \cdot \vec{w} - Y_B)$$

$$\begin{bmatrix} | & & \\ \vec{x}_1^T & & \\ | & & \end{bmatrix} \cdot \left(\begin{bmatrix} - & \vec{x}_1 & - \\ & & \\ & & \end{bmatrix} \begin{bmatrix} | \\ \vec{w} \\ | \end{bmatrix} - \begin{bmatrix} y_1 \\ \\ \\ \end{bmatrix} \right)$$

Secure Aggregation



Potential Issues ?

FL for Logistic Regression

SGD:

$$\vec{w} \leftarrow \vec{w} - \eta \cdot \nabla L_i(\vec{w})$$

$$\vec{w} \leftarrow \vec{w} - \eta \cdot \left(f\left(\langle \vec{x}_i, \vec{w} \rangle \right) - y_i \right) \cdot \vec{x}_i$$

Batch SGD:

$$\vec{w} \leftarrow \vec{w} - \frac{\eta}{B} \cdot \sum_{i \in [B]} \nabla L_i(\vec{w})$$

$$\vec{w} \leftarrow \vec{w} - \frac{\eta}{B} \cdot X_B^T \cdot \left(f\left(X_B \cdot \vec{w} \right) - Y_B \right)$$

$$\begin{bmatrix} | & & \\ \vec{x}_1^T & & \\ | & & \end{bmatrix} \cdot \left(f\left(\begin{bmatrix} - & \vec{x}_1 & - \\ & & \\ & & \end{bmatrix} \begin{bmatrix} | \\ \vec{w} \\ | \end{bmatrix} \right) - \begin{bmatrix} y_1 \\ \\ \\ \end{bmatrix} \right)$$

Differential Privacy

Name	Age	Gender	Race	Weight	ZIP	Disease
Alice						
Bob						
Charlie						
David						
Emily						
Fiona						

Want to make the (sensitive) data public / available to others (e.g. for medical study).

Attempt 1: "Anonymize" the data.

Delete personally identifiable information (PII): name, DOB, ...

Attempt 2: Only answer aggregate statistics queries.

Privacy Guarantee?

Access to the output shouldn't enable one to learn anything about an individual compared to one without access.

Is this possible?

Privacy vs. Correctness / Utility

Privacy Guarantee?

Access to the output shouldn't enable one to learn ~~anything~~ much more about an individual compared to one ~~without access~~.

with access to the output computed on a database without the individual.

Differential Privacy

x_1
x_2
$\vdots$
x_n

randomized

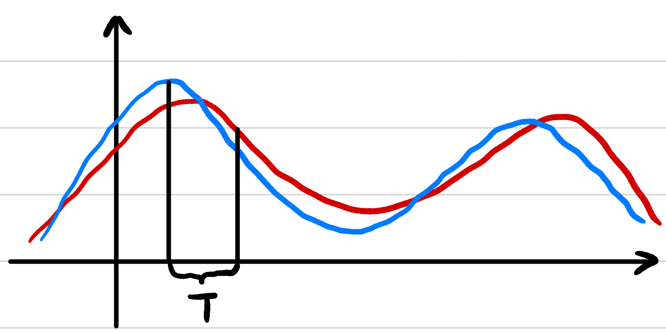
$$D \in X^n \longrightarrow \boxed{M} \longrightarrow M(D)$$

Def ε-Differential Privacy for a randomized mechanism:

$\forall$ neighboring datasets D_1 & D_2 (differing in one row),

$\forall\, T \subseteq \text{range}(M)$,

$$\Pr[M(D_1) \in T] \le e^{\varepsilon} \cdot \Pr[M(D_2) \in T]$$



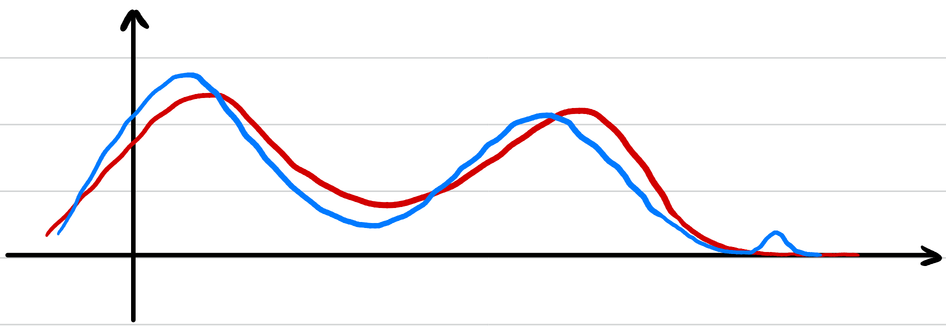
Differential Privacy

Def (ε, δ) - Differential Privacy for a randomized mechanism:

$\forall$ neighboring datasets D_1 & D_2 (differing in one row),

$\forall\, T \subseteq \text{range}(M)$,

$$\Pr[M(D_1) \in T] \leq e^{\varepsilon} \cdot \Pr[M(D_2) \in T] + \delta$$



Is a bigger ε better for privacy, or worse? Worse

Is a bigger δ better for privacy, or worse? Worse

Randomized Response

Counting query: What percentage $\overset{\alpha}{\swarrow}$ of individuals satisfy predicate P?

> For each row x_i:
>
> ① Sample $b \xleftarrow{\$} \{0,1\}$
>
> ② If $b=0$, then $y_i := P(x_i)$
>
> Otherwise, $y_i \xleftarrow{\$} \{0,1\}$
>
> $M(D) := (y_1, y_2, \cdots, y_n)$ ← fraction of 1s = β

Thm Randomized Response is $\ln 3$-DP.

How to estimate the query output?

$$\frac{1}{2}\cdot\frac{1}{2}+\frac{1}{2}\cdot\alpha=\beta \Rightarrow \alpha=\left(\beta-\frac{1}{4}\right)\cdot 2$$

How to make the mechanism more private?

$$b \begin{cases} \nearrow 0 & 10\% \\ \searrow 1 & 90\% \end{cases}$$

Laplace Mechanism

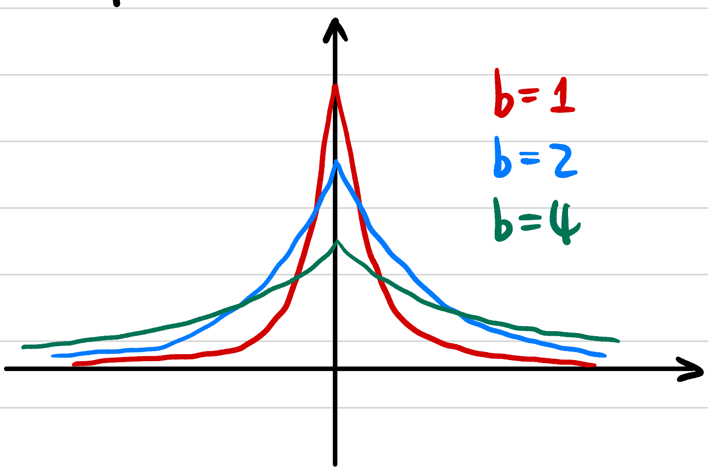
Def Sensitivity of a function $f: X^n \to \mathbb{R}$

$$\Delta f := \max_{D_1 \sim D_2} |f(D_1) - f(D_2)|$$

Laplace Mechanism: $M(D) = f(D) + \mathrm{Lap}(\Delta f / \varepsilon)$

Thm The Laplace Mechanism is ε-DP.

Laplace distribution:



probability distribution function

$$\mathrm{PDF}(x) = \frac{1}{2b} \cdot \exp\left(-\frac{|x|}{b}\right)$$

For $X \sim \mathrm{Lap}(b)$, $\Pr[|X| \geq bt] \leq \exp(-t)$

Is a bigger b better for privacy, or worse?

Better

Composition Theorems

Thm (post-processing) If $M: X^n \to Y$ is (ε, δ)-DP,

$f: Y \to Z$ is an arbitrary randomized function,

then $f \circ M: X^n \to Z$ is also (ε, δ)-DP.

Thm (group privacy) If $M: X^n \to Y$ is $(\varepsilon, 0)$-DP,

then M is $(k \cdot \varepsilon, 0)$-DP for groups of size k.

Thm (composition) If $M_i: X^n \to Y$ is $(\varepsilon_i, \delta_i)$-DP $\forall i \in [k]$,

then $M(D) := (M_1(D), \dots, M_k(D))$ is $\left(\sum_{i \in [k]} \varepsilon_i, \sum_{i \in [k]} \delta_i\right)$-DP.

Elliptic Curve Cryptography

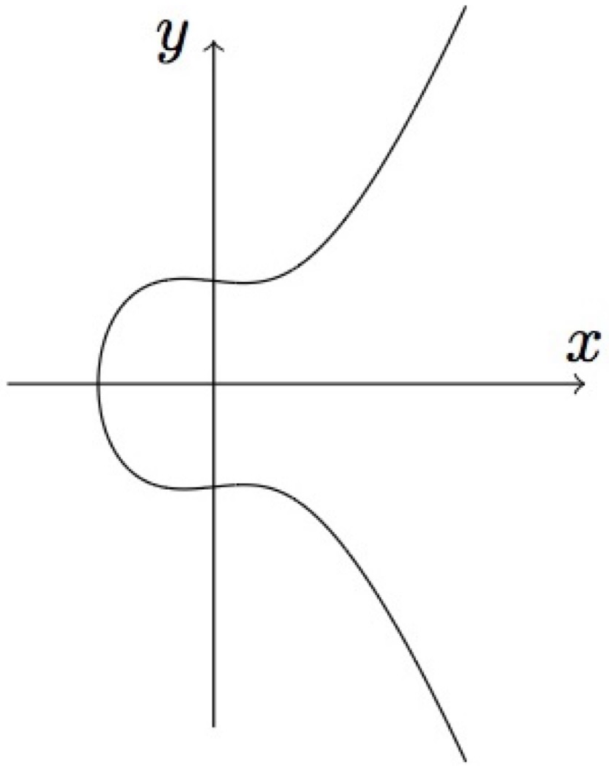
Cyclic group $\mathbb{G}$ of order q with generator g where DLOG / CDH / DDH holds.

How large is q? (128-bit security)

- Integer groups: $q \sim 2048$ bits
- Elliptic curve groups: $q \sim 256$ bits

Additional structure: bilinear pairings

Elliptic Curves



$$y^2 = x^3 + ax + b$$

$$(4a^3 + 27b^2 \neq 0)$$

Example: $y^2 = x^3 - x + 9$

points: $(0, \pm 3)$
$(1, \pm 3)$
$(-1, \pm 3)$

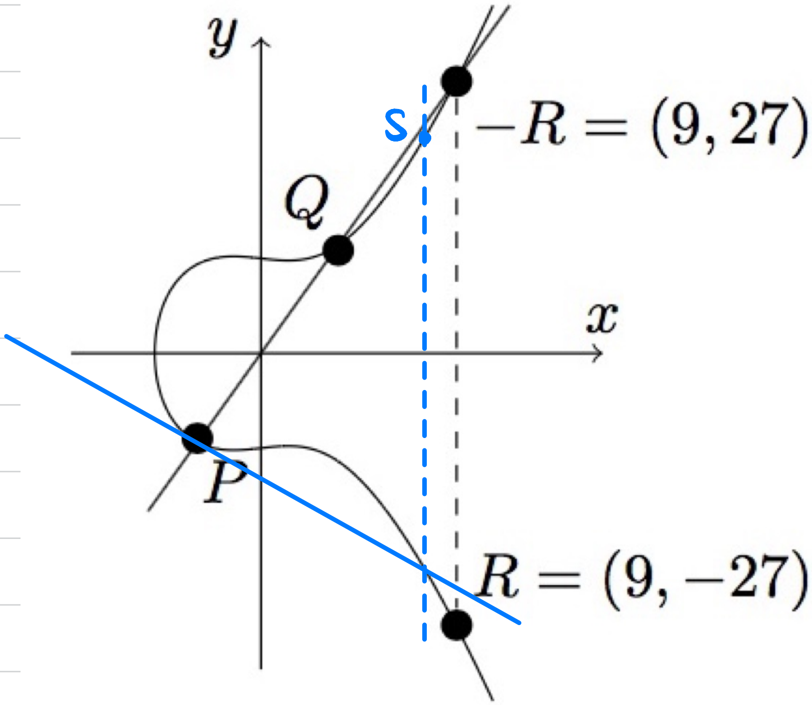
How to find rational points $(x, y) \in \mathbb{Q}^2$ on the curve?

$x = \frac{s}{t}$, $y = \frac{u}{v}$

$s, t, u, v \in \mathbb{Z}$

Elliptic Curves

How to find rational points $(x,y) \in \mathbb{Q}^2$ on the curve?



① Chord method $\quad R := P \boxplus Q$

$P = (-1, -3)$
$Q = (1, 3)$ $\Rightarrow$ $y = 3x$

$\Downarrow$

$(3x)^2 = x^3 - x + 9$

$x^3 - 9x^2 - x + 9 = 0$

Why is the third root rational?

Example: $y^2 = x^3 - x + 9$

② tangent method $\quad S := P \boxplus P$