

Theory of Differential Privacy

Lecture 14:

DP-SGD and amplification by subsampling

EN.601.438/638 Fall 2026

Lydia Zakynthinou

Johns Hopkins University

Oct 19, 2026

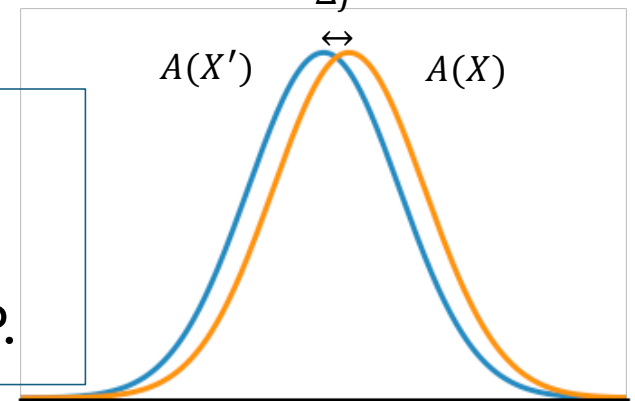
Gaussian Mechanism

- Let $f(X): \mathcal{X}^n \rightarrow \mathbb{R}^d$
- Let $\Delta f = \max_{\text{neighbors } X, X'} \{\|f(X) - f(X')\|_2\}$ be the ℓ_2 sensitivity of f

The Gaussian Mechanism returns

$$A(X) = f(X) + \mathcal{N}(0, \sigma^2 \mathbb{I}_d)$$

where $\sigma^2 = \frac{(\Delta f)^2 2 \log\left(\frac{1.25}{\delta}\right)}{\epsilon^2}$. It satisfies (ϵ, δ) -DP.
 $= c_{\epsilon, \delta}^2$



DP-(S)GD

Empirical loss minimization

Fix $x = (x_1, \dots, x_n)$.

$$L_x(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(\theta; x_i) \quad \theta^* = \operatorname{argmin}_{\theta \in \Theta} L_x(\theta).$$

How close is the empirical loss
of a $\hat{\theta}$ to $L_x(\theta^*)$?

- x fixed
- $\Theta \subseteq \mathbb{R}^d$ public, closed, convex

Projected Gradient Descent.

Assume $\text{diam}(\Theta) \leq R$

$\ell(\cdot; z)$ is convex and differentiable

$$\|\nabla \ell(\theta; z)\|_2 \leq G.$$

Define $g_t = \nabla L_x(\theta_t)$.

Initially, θ_0 .

In every round $t=1, \dots, T$: $\theta_{t+1} = \text{Proj}_{\Theta}(\theta_t - \eta g_t)$.

Output $\hat{\theta} = \frac{1}{T} \sum_{t=1}^T \theta_t$.

Assume $\text{diam}(\Theta) \leq R$

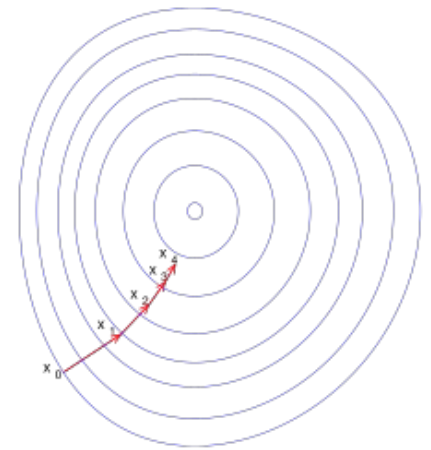
$\ell(\cdot; z)$ is convex and differentiable

$$\|\nabla \ell(\theta; z)\|_2 \leq G$$

Projected

Gradient Descent

- Parameter space $\Theta \in \mathbb{R}^d$,
- Loss function $\ell: \Theta \times \mathcal{X} \rightarrow \mathbb{R}$,
- Datapoints $x_i \in \mathcal{X} \subseteq \mathbb{R}^d$ for $i \in [n]$ (each user owns one datapoint)



Initialize θ_0 , $\|\theta^* - \theta_0\|_2 \leq R$ since $\theta^*, \theta_0 \in \Theta$

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

1. Let each gradient be $g_{t,i} = \nabla \ell(\theta_{t-1}, x_i) \Rightarrow \|g_{t,i}\|_2 \leq G$
2. Compute the average gradient: $\bar{g}_t = \frac{1}{n} \sum_{i \in [n]} g_{t,i}$
3. Take a step in the negative direction: $\theta_t = \underset{\Theta}{\text{Proj}} (\theta_{t-1} - \eta \bar{g}_t)$

Output $\bar{\theta} = \frac{1}{T} \sum \theta_t$

Under assumptions, $L(\bar{\theta}) \leq L(\theta^*) + \frac{1}{\eta T} + \eta = L(\theta^*) + \frac{1}{\sqrt{T}}$

$\underset{\Theta}{\text{Proj}}$

$$\text{Empirical loss: } L(\bar{\theta}) \leq L(\theta^*) + \frac{R^2}{2\eta T} + \frac{\eta}{2} G^2 \underset{\text{opt. } \eta}{\approx} L(\theta^*) + \frac{RG}{\sqrt{T}}$$

Differentially Private Gradient Descent

- Parameter space $\Theta \in \mathbb{R}^d$,
- Loss function $\ell: \Theta \times \mathcal{X} \rightarrow \mathbb{R}$,
- Datapoints $x_i \in \mathcal{X} \subseteq \mathbb{R}^d$ for $i \in [n]$ (each user owns one datapoint)
- Operator $\Pi_C(y) = \min\{1, C/\|y\|_2\} \cdot y$

Initialize θ_0

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

1. Compute the average **clipped** gradient: $\bar{g}_t = \frac{1}{n} \sum_{i \in [n]} \Pi_C(g_{t,i})$
2. Add **noise**: $\tilde{g}_t = \bar{g}_t + z_t, z_t \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_d)$
3. Take a step in the negative direction: $\theta_t = \theta_{t-1} - \eta \tilde{g}_t$

Utility: $\mathbb{E}[\tilde{g}_t | \mathcal{H}_{t-1}] = \bar{g}_t = \nabla L_X(\theta_{t-1})$ noise

Assume $G \leq C$.

Simons Institute, Spring 2026

Federated and Collaborative Learning Bootcamp

8

$$\mathbb{E}[\|\tilde{g}_t\|_2^2 | \mathcal{H}_{t-1}] \leq V^2 \text{ (for some } V = \text{Var}(\mathcal{N}(\bar{g}_t, \sigma^2 \mathbb{I}_d)) \text{)}$$

$$\text{Similarly, empirical loss} \approx \frac{RV}{\sqrt{T}}$$

DP-GD



- Step 1+2: (ϵ, δ) -DP since $\Delta \bar{g}_t = \frac{2C}{n}$ (Gaussian Mechanism).
- Step 3: (ϵ, δ) -DP (Post-processing).
- Overall, DP-GD is $(\sqrt{T}\epsilon, T\delta)$ -DP (Advanced Composition).

Initialize θ_0

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

1. Compute the average clipped gradient: $\bar{g}_t = \frac{1}{n} \sum_{i \in [n]} \Pi_C(g_{t,i})$
2. Add noise: $\tilde{g}_t = \bar{g}_t + z_t, z_t \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_d)$ for $\sigma = c_{\epsilon, \delta} \cdot \frac{2C}{n}$
3. Take a step in the negative direction: $\theta_t = \underset{\text{Proj}_\Theta}{\hat{}}(\theta_{t-1} - \eta \tilde{g}_t)$

DP-GD



- DP-GD is (ϵ, δ) -DP (Advanced Composition).

- ★★☆ • Under assumptions, $L(\bar{\theta}) \leq L(\theta^*) + \frac{R^2}{\eta T} + \eta G^2 + \eta \frac{T d C^2 \text{opt.} \eta^\top}{\epsilon^2 n^2} = L(\theta^*) + \frac{\sqrt{d}}{\epsilon n}$ for $G = R = C = 1$



- Running time $Tn(\text{time}(\nabla) + d)$: Slow...

Initialize θ_0

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

1. Compute the average clipped gradient: $\bar{g}_t = \frac{1}{n} \sum_{i \in [n]} \Pi_C(g_{t,i})$
2. Add noise: $\tilde{g}_t = \bar{g}_t + z_t, z_t \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_d)$ for $\sigma \approx \frac{\sqrt{T \log(T/\delta)} R}{\epsilon n}$
3. Take a step in the negative direction: $\theta_t = \theta_{t-1} - \eta \tilde{g}_t$

DP-SGD



- Running time $T(\text{time}(\nabla) + d)$: Better

Initialize θ_0

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

1. Choose a random user $i_t \sim \text{Uniform}([n])$
2. Compute the clipped gradient: $g_t = \Pi_C(g_{t,i_t})$
3. Add noise: $\tilde{g}_t = g_t + z_t, z_t \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_d)$ for $\sigma \approx \frac{\sqrt{T \log(T/\delta)} \mathcal{L}}{\varepsilon n}$
4. Take a step in the negative direction: $\theta_t = \underset{\text{Proj}_\Theta}{\left(\theta_{t-1} - \eta \tilde{g}_t \right)}$

DP-SGD $\Delta \bar{g}_t = \frac{2C}{n}$



- $\Delta g_t = 2C \Rightarrow$ privacy guarantee per round is n times worse!

Initialize θ_0

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

1. Choose a random user $i_t \sim \text{Uniform}([n])$
2. Compute the clipped gradient: $g_t = \Pi_C(g_{t,i_t})$
3. Add noise: $\tilde{g}_t = g_t + z_t, z_t \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_d)$ for $\sigma \approx \frac{\sqrt{T \log(T/\delta)} 2C}{\epsilon n}$
4. Take a step in the negative direction: $\theta_t = \underset{\text{Proj}}{(\theta_{t-1} - \eta \tilde{g}_t)}$

Not really. Only if i_t per round was also released.

sampling hides participation $\rightarrow$ amplifies privacy.

“Secrecy of the sample”: Smith [2009], Kasiviswanathan, Lee, Nissim, Raskhodnikova, Smith [2011]
Balle, Barthe, and Gaborardi [BBG18], Steinke [2022] survey

DP-SGD



Amplification by Subsampling:

S : chooses user randomly, each with probability p .

If A is (ϵ, δ) -DP, then $A \circ S$ is $(\ln(1 + (e^\epsilon - 1)p), p\delta)$ -DP.

Initialize θ_0

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

1. Choose a random user $i_t \sim \text{Uniform}([n])$
2. Compute the clipped gradient: $\bar{g}_t = \Pi_C(g_{t,i_t})$
3. Add noise: $\tilde{g}_t = g_t + z_t, z_t \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_d)$ for $\sigma \approx \frac{\sqrt{T \log(T/\delta) 2C}}{\epsilon n}$
4. Take a step in the negative direction: $\theta_t = \underset{\text{Proj}_\Theta}{\left(\theta_{t-1} - \eta \tilde{g}_t \right)}$

DP-SGD



Amplification by Subsampling:

S : chooses user randomly, each with probability p .

If A is (ϵ, δ) -DP, then $A \circ S$ is roughly $(p\epsilon, p\delta)$ -DP, if $\epsilon \leq 1$.

Initialize θ_0

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

1. Choose a random user $i_t \sim \text{Uniform}([n])$
2. Compute the clipped gradient: $\bar{g}_t = \Pi_C(g_{t,i_t})$
3. Add noise: $\tilde{g}_t = g_t + z_t, z_t \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_d)$ for $\sigma \approx \frac{\sqrt{T \log(T/\delta)} \mathcal{C}}{\epsilon n}$
4. Take a step in the negative direction: $\theta_t = (\theta_{t-1} - \eta \tilde{g}_t)$

2030

DP-SGD



Amplification by Subsampling:

S : chooses user randomly, each with probability $p = 1/n$.

If A is (ϵ, δ) -DP, then $A \circ S$ is roughly $(\epsilon/n, \delta/n)$ -DP, if $\epsilon \leq 1$.

Initialize θ_0

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

1. Choose a random user $i_t \sim \text{Uniform}([n])$
2. Compute the clipped gradient: $\bar{g}_t = \Pi_C(g_{t,i_t})$
3. Add **noise**: $\tilde{g}_t = g_t + z_t, z_t \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_d)$ for $\sigma \approx \frac{\sqrt{T \log(T/\delta)} \mathcal{L}}{\epsilon n}$
4. Take a step in the negative direction: $\theta_t = (\theta_{t-1} - \eta \tilde{g}_t)$

Proj_C

in round t : $\tilde{g}_t = g_t + \mathcal{N}(0, \sigma^2 \mathbb{I}_d)$ where $\sigma^2 = \frac{C^2 \log(1/\delta)}{\epsilon^2}$
 then round t is $(\epsilon \cdot \frac{1}{n}, \delta \cdot \frac{1}{n})$ -DP (by amplification).

DP-SGD



- DP-SGD is (ϵ, δ) -DP (Amplification + Adv Composition),
- its utility is roughly the same (esp. re: error due to privacy),
- *and* it runs faster than DP-GD!

Initialize θ_0

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

1. Choose a random user $i_t \sim \text{Uniform}([n])$
2. Compute the clipped gradient: $\bar{g}_t = \Pi_C(g_{t,i_t})$
3. Add **noise**: $\tilde{g}_t = g_t + z_t, z_t \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_d)$ for $\sigma \approx \frac{\sqrt{T \log(T/\delta) 2C}}{\epsilon n}$
4. Take a step in the negative direction: $\theta_t = \theta_{t-1} - \eta \tilde{g}_t$

noise

Privacy Accountants

Advanced Composition is tight, for black-box mechanisms. But DP-SGD is not one.

Kairouz, Oh, Viswanathan [2015], Murtagh, Vadhan [2016]

- Privacy loss RV for neighboring datasets X, X' and outcome y :

$$L_M(y) = \log \frac{\Pr[M(X) = y]}{\Pr[M(X') = y]}$$

- Log-moment (log MGF) at order $\lambda > 0$: $\alpha_M(\lambda) = \log \mathbb{E}_{y \sim M(X)}[e^{\lambda L_M(y)}]$
- Composition is additive: for $M = M_1 \circ \dots \circ M_T$, $\alpha_M(\lambda) \leq \sum_t \alpha_{M_t}(\lambda)$
- Convert to (ϵ, δ) via a Chernoff/Markov tail bound: $\delta(\epsilon) = \min_{\lambda > 0} e^{\alpha_M(\lambda) - \lambda \epsilon}$

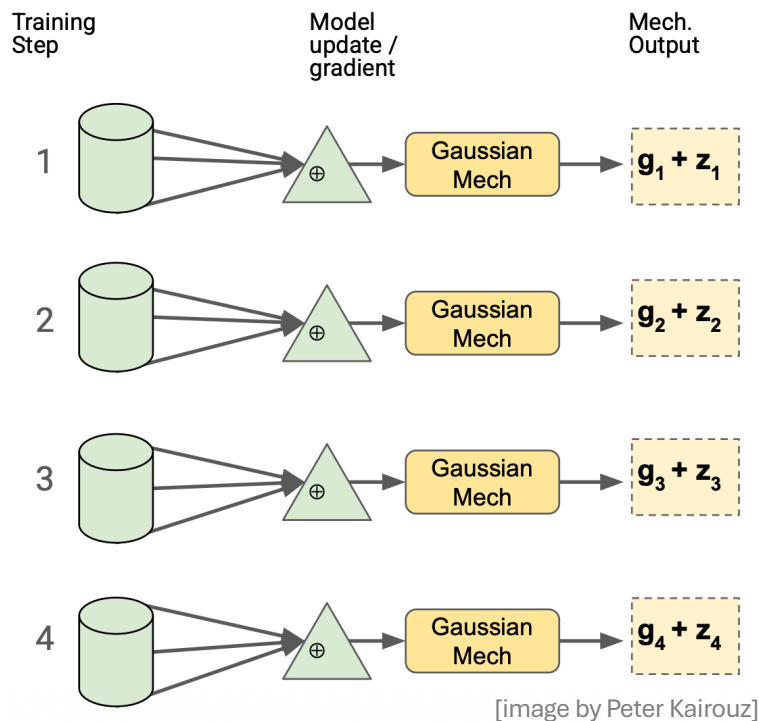


- DP-SGD with $\sigma \approx \frac{C\sqrt{T\log(1/\delta)}}{n\epsilon}$ is (ϵ, δ) -DP

Privacy Accountants

- Rényi [Mironov 2017] moments accountant: closed-form bounds
Mironov, Talwar, Zhang [2019] Wang, Balle, Kasiviswanathan [2018]
- PLD / FFT accountants: approximate the *full privacy loss distribution* and compose via (FFT) convolutions. Koskela, Jälkö, Honkela [2019]
- f-DP / Gaussian DP accountants: hypothesis-testing tradeoff function, interpretable privacy “curve.” Dong, Roth, Su [2019]

Challenges with DP-SGD



- Sampling assumptions of theoretical analyses and practical methods rarely match.

Chua, Ghazi, Kamath, Kumar, Manurangsi, Sinha, Zhang [2024]

- The regime for which privacy amplification works requires a *tiny* privacy parameter.
[see differentialprivacy.org post by Thomas Steinke]