Trustworthy Machine Learning

Differential Privacy 2

Sangdon Park

POSTECH

Contents from

A preliminary version of this paper appears in the proceedings of the *23rd ACM Conference on Computer and Communications Security (CCS 2016)*. This is a full version.

Deep Learning with Differential Privacy

October 25, 2016

Martín Abadi* Andy Chu* Ian Goodfellow†
H. Brendan McMahan* Ilya Mironov* Kunal Talwar*
Li Zhang*

- (I guess) The first DP paper for deep learning
- This is a complicated application of the basic DP, so we will briefly see high-level ideas.

Difference?

- DP with convex loss
  - Add noise on the final model
  - Add noise before learning
  - Strategies in convex loss treat learning process as a block box
- DP with non-convex loss
  - Consider learning process as a white box for the careful(?) characterization of parameter updates.

Definition: Differential Privacy (Again)

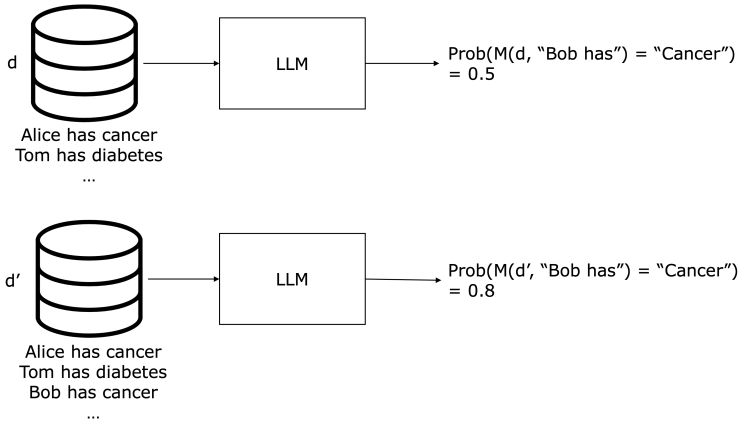
Definition

A randomized mechanism $\mathcal{M} : \mathcal{D} \to \mathcal{R}$ with domain $\mathcal{D}$ and range $\mathcal{R}$ satisfies (ε, δ)-differential privacy if for any two "adjacent" inputs $d, d' \in \mathcal{D}$ and for any subset of outputs $S \subseteq \mathcal{R}$ it holds that

$$\mathbb{P}\left\{\mathcal{M}(d) \in S\right\} \leq e^{\varepsilon}\mathbb{P}\left\{\mathcal{M}(d') \in S\right\} + \delta.$$

- Notations are slightly adjusted for learning.
- "adjacent" inputs: two inputs differ in a single labeled example.

A Toy Example



- Here, the mechanism $\mathcal{M}$ includes training an LLM over a dataset and querying a question.
- At least we know that d' has Bob's information (and he likely has cancer due to the high confidence).

Differentially Private SGD (DP-SGD)

Algorithm 1 Differentially private SGD (Outline)

Input: Examples $\{x_1, \ldots, x_N\}$, loss function $\mathcal{L}(\theta) = \frac{1}{N}\sum_i \mathcal{L}(\theta, x_i)$. Parameters: learning rate η_t, noise scale σ, group size L, gradient norm bound C.

Initialize θ_0 randomly

for $t \in [T]$ **do**

Take a random sample L_t with sampling probability L/N

Compute gradient

For each $i \in L_t$, compute $\mathbf{g}_t(x_i) \leftarrow \nabla_{\theta_t}\mathcal{L}(\theta_t, x_i)$

Clip gradient

$\bar{\mathbf{g}}_t(x_i) \leftarrow \mathbf{g}_t(x_i) / \max\left(1, \frac{\|\mathbf{g}_t(x_i)\|_2}{C}\right)$

Add noise

$\tilde{\mathbf{g}}_t \leftarrow \frac{1}{L}\left(\sum_i \bar{\mathbf{g}}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})\right)$

Descent

$\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{\mathbf{g}}_t$

Output θ_T and compute the overall privacy cost (ε, δ) using a privacy accounting method.

Differentially Private SGD (DP-SGD)

Algorithm 1 Differentially private SGD (Outline)

Input: Examples $\{x_1, \ldots, x_N\}$, loss function $\mathcal{L}(\theta) = \frac{1}{N}\sum_i \mathcal{L}(\theta, x_i)$. Parameters: learning rate η_t, noise scale σ, group size L, gradient norm bound C.

Initialize θ_0 randomly

for $t \in [T]$ **do**

Take a random sample L_t with sampling probability L/N

Compute gradient

For each $i \in L_t$, compute $\mathbf{g}_t(x_i) \leftarrow \nabla_{\theta_t}\mathcal{L}(\theta_t, x_i)$

Clip gradient

$\bar{\mathbf{g}}_t(x_i) \leftarrow \mathbf{g}_t(x_i) / \max\left(1, \frac{\|\mathbf{g}_t(x_i)\|_2}{C}\right)$

Add noise

$\tilde{\mathbf{g}}_t \leftarrow \frac{1}{L}\left(\sum_i \bar{\mathbf{g}}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})\right)$

Descent

$\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{\mathbf{g}}_t$

Output θ_T and compute the overall privacy cost (ε, δ) using a privacy accounting method.

- $\mathcal{M}_t(d) \coloneqq \sum_{i \in L_t} \tilde{\mathbf{g}}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 I)$: the Gaussian mechanism (when $d \coloneqq L_t$)

Differentially Private SGD (DP-SGD)

Algorithm 1 Differentially private SGD (Outline)

Input: Examples $\{x_1, \ldots, x_N\}$, loss function $\mathcal{L}(\theta) = \frac{1}{N}\sum_i \mathcal{L}(\theta, x_i)$. Parameters: learning rate η_t, noise scale σ, group size L, gradient norm bound C.

Initialize θ_0 randomly

for $t \in [T]$ **do**

Take a random sample L_t with sampling probability L/N

Compute gradient

For each $i \in L_t$, compute $\mathbf{g}_t(x_i) \leftarrow \nabla_{\theta_t}\mathcal{L}(\theta_t, x_i)$

Clip gradient

$\bar{\mathbf{g}}_t(x_i) \leftarrow \mathbf{g}_t(x_i) / \max\left(1, \frac{\|\mathbf{g}_t(x_i)\|_2}{C}\right)$

Add noise

$\tilde{\mathbf{g}}_t \leftarrow \frac{1}{L}\left(\sum_i \bar{\mathbf{g}}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})\right)$

Descent

$\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{\mathbf{g}}_t$

Output θ_T and compute the overall privacy cost (ε, δ) using a privacy accounting method.

- $\mathcal{M}_t(d) \coloneqq \sum_{i \in L_t} \tilde{\mathbf{g}}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 I)$: the Gaussian mechanism (when $d \coloneqq L_t$)
- Why clipping?

Differentially Private SGD (DP-SGD)

Algorithm 1 Differentially private SGD (Outline)

Input: Examples $\{x_1, \ldots, x_N\}$, loss function $\mathcal{L}(\theta) = \frac{1}{N}\sum_i \mathcal{L}(\theta, x_i)$. Parameters: learning rate η_t, noise scale σ, group size L, gradient norm bound C.

Initialize θ_0 randomly

for $t \in [T]$ **do**

Take a random sample L_t with sampling probability L/N

Compute gradient

For each $i \in L_t$, compute $\mathbf{g}_t(x_i) \leftarrow \nabla_{\theta_t}\mathcal{L}(\theta_t, x_i)$

Clip gradient

$\bar{\mathbf{g}}_t(x_i) \leftarrow \mathbf{g}_t(x_i) / \max\left(1, \frac{\|\mathbf{g}_t(x_i)\|_2}{C}\right)$

Add noise

$\tilde{\mathbf{g}}_t \leftarrow \frac{1}{L}\left(\sum_i \bar{\mathbf{g}}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})\right)$

Descent

$\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{\mathbf{g}}_t$

Output θ_T and compute the overall privacy cost (ε, δ) using a privacy accounting method.

- $\mathcal{M}_t(d) \coloneqq \sum_{i \in L_t} \tilde{\mathbf{g}}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 I)$: the Gaussian mechanism (when $d \coloneqq L_t$)
- Why clipping?
- How to determine the noise level σ to satisfy (ε, δ)-DP?

Main Ingredient: Norm Clipping

Norm Clipping

$$\tilde{\mathbf{g}}_t(x_i) \leftarrow \frac{\mathbf{g}_t(x_i)}{\max\left(1, \frac{\|\mathbf{g}_t(x_i)\|_2}{C}\right)}$$

- Maintain the norm of gradients to be at most C, *i.e.*,

$$\frac{\mathbf{g}}{\max\left(1, \frac{\|\mathbf{g}\|_2}{C}\right)} = \begin{cases} \mathbf{g} & \text{if } \|\mathbf{g}\|_2 \leq C \\ \frac{C}{\|\mathbf{g}\|_2}\mathbf{g} & \text{if } \|\mathbf{g}\|_2 > C \end{cases}$$

Main Ingredient: Norm Clipping

Norm Clipping

$$\tilde{\mathbf{g}}_t(x_i) \leftarrow \frac{\mathbf{g}_t(x_i)}{\max\left(1, \frac{\|\mathbf{g}_t(x_i)\|_2}{C}\right)}$$

- Maintain the norm of gradients to be at most C, *i.e.,*

$$\frac{\mathbf{g}}{\max\left(1, \frac{\|\mathbf{g}\|_2}{C}\right)} = \begin{cases} \mathbf{g} & \text{if } \|\mathbf{g}\|_2 \leq C \\ \frac{C}{\|\mathbf{g}\|_2}\mathbf{g} & \text{if } \|\mathbf{g}\|_2 > C \end{cases}$$

- Limit "privacy loss" at each learning iteration for a tighter the DP guarantee
  - If the norm of gradients is "large", we need to add "large" noise to cover them (otherwise, privacy leaking)
  - Without clipping, we need to add noise proportional to the largest norm of gradients.
  - With clipping, (as we control the maximum of the norm) we can choose a smaller noise level.
  - Price to pay: clipping may hurt accuracy

Main Ingredient: Norm Clipping

Norm Clipping

$$\tilde{\mathbf{g}}_t(x_i) \leftarrow \frac{\mathbf{g}_t(x_i)}{\max\left(1, \frac{\|\mathbf{g}_t(x_i)\|_2}{C}\right)}$$

- Maintain the norm of gradients to be at most C, *i.e.,*

$$\frac{\mathbf{g}}{\max\left(1, \frac{\|\mathbf{g}\|_2}{C}\right)} = \begin{cases} \mathbf{g} & \text{if } \|\mathbf{g}\|_2 \leq C \\ \frac{C}{\|\mathbf{g}\|_2}\mathbf{g} & \text{if } \|\mathbf{g}\|_2 > C \end{cases}$$

- Limit "privacy loss" at each learning iteration for a tighter the DP guarantee
  - If the norm of gradients is "large", we need to add "large" noise to cover them (otherwise, privacy leaking)
  - Without clipping, we need to add noise proportional to the largest norm of gradients.
  - With clipping, (as we control the maximum of the norm) we can choose a smaller noise level.
  - Price to pay: clipping may hurt accuracy
- Clipping before averaging
  - may provide a tighter DP guarantee (why?)

Privacy Analysis: Is DP-SGD DP?

To this end, bound the *moments* of *privacy loss* in two steps!

1. Bounding the moment for each learning iteration
2. Bounding the moments for all learning iterations

Then, what is

- privacy loss? An surrogate for measuring DP.
- the moments of the privacy loss?

Measuring DP: Privacy Loss

Privacy Loss

$$\ell(o; \mathcal{M}, \mathbf{aux}, d, d') \coloneqq \log \frac{\mathbb{P}\left\{\mathcal{M}(\mathbf{aux}, d) = o\right\}}{\mathbb{P}\left\{\mathcal{M}(\mathbf{aux}, d') = o\right\}}$$

- $d, d' \in \mathcal{D}$: neighboring datasets
- $\mathcal{M}$: a mechanism
- $\mathbf{aux}$: an auxiliary input, *e.g.,* previous gradients
- $o \in \mathcal{R}$: an outcome
- How to capture the properties of the privacy loss?
  - Consider o as a random variable, *i.e.,* $o \sim \mathcal{M}(\mathbf{aux}, d)$.
  - Analyze the privacy loss via moments.

Measuring DP: Moments of Privacy Loss

Moment

$$\alpha_{\mathcal{M}}(\lambda) = \max_{\mathsf{aux},d,d'} \alpha_{\mathcal{M}}(\lambda; \mathsf{aux}, d, d') \quad \text{where}$$

$$\alpha_{\mathcal{M}}(\lambda; \mathsf{aux}, d, d') := \ln \mathbb{E}_{o\sim\mathcal{M}(\mathsf{aux},d)} e^{\lambda \ell(o;\mathcal{M},\mathsf{aux},d,d')}$$

Measuring DP: Moments of Privacy Loss

Moment

$$\alpha_{\mathcal{M}}(\lambda) = \max_{\mathsf{aux},d,d'} \alpha_{\mathcal{M}}(\lambda; \mathsf{aux}, d, d') \quad \text{where}$$

$$\alpha_{\mathcal{M}}(\lambda; \mathsf{aux}, d, d') := \ln \mathbb{E}_{o\sim\mathcal{M}(\mathsf{aux},d)} e^{\lambda \ell(o;\mathcal{M},\mathsf{aux},d,d')}$$

- The *moment-generating function* (or moments) of a real-valued random variable X, denoted by $M_X(\lambda)$, captures the useful properties of the corresponding distribution.

$$\begin{aligned} M_X(\lambda) &:= \mathbb{E}\{e^{\lambda X}\} \\ &= \mathbb{E}\left\{1 + \lambda X + \frac{\lambda^2 X^2}{2!} + \frac{\lambda^3 X^3}{3!} + \cdots\right\} \\ &= 1 + \lambda \mathbb{E}\{X\} + \frac{\lambda^2 \mathbb{E}\{X^2\}}{2!} + \frac{\lambda^3 \mathbb{E}\{X^3\}}{3!} + \cdots \end{aligned}$$

  - To obtain mean, differentiating $M_X(\lambda)$ once with respect to λ and setting $\lambda = 0$.

From the Moments to the DP Guarantee

Theorem

For any $\varepsilon > 0$, the mechanism $\mathcal{M}$ is (ε, δ)-DP where

$$\delta = \min_{\lambda} e^{\alpha_{\mathcal{M}}(\lambda) - \lambda\varepsilon}.$$

- Connect (ε, δ)-DP to $\alpha_{\mathcal{M}}(\lambda)$
- Given δ, if we know the moments $\alpha_{\mathcal{M}}(\lambda)$, the privacy parameter ε is determined.
- How to compute or bound $\alpha_{\mathcal{M}}(\lambda)$?

From the Moments to the DP Guarantee: A Proof Sketch

- Recall the privacy loss ℓ

$$\ell(o; \mathcal{M}, \mathbf{aux}, d, d') \coloneqq \ln \frac{\mathbb{P}\left\{\mathcal{M}(\mathbf{aux}, d) = o\right\}}{\mathbb{P}\left\{\mathcal{M}(\mathbf{aux}, d') = o\right\}}$$

- Let an (bad) event $B \coloneqq \ell(o; \cdot) \geq \varepsilon$
- For any S, we have

$$\begin{aligned}
\mathbb{P}\left\{\mathcal{M}(d) \in S\right\} &= \mathbb{P}\left\{\mathcal{M}(d) \in S \cap B^c\right\} + \mathbb{P}\left\{\mathcal{M}(d) \in S \cap B\right\} \\
&\leq e^{\varepsilon}\mathbb{P}\left\{\mathcal{M}(d') \in S \cap B^c\right\} + \mathbb{P}\left\{\mathcal{M}(d) \in S \cap B\right\} \\
&\leq e^{\varepsilon}\mathbb{P}\left\{\mathcal{M}(d') \in S\right\} + \mathbb{P}\left\{\mathcal{M}(d) \in B\right\} \\
&\leq e^{\varepsilon}\mathbb{P}\left\{\mathcal{M}(d') \in S\right\} + e^{\alpha_{\mathcal{M}}(\lambda) - \lambda\varepsilon},
\end{aligned}$$

Here, the last inequality holds since

$$\mathbb{P}_{o\sim\mathcal{M}(d)}\left\{\ell(o;\cdot) \geq \varepsilon\right\} = \mathbb{P}_{o\sim\mathcal{M}(d)}\left\{e^{\lambda\ell(o;\cdot)} \geq e^{\lambda\varepsilon}\right\} \leq \frac{\mathbb{E}_{o\sim\mathcal{M}(d)}\{e^{\lambda\ell(o;\cdot)}\}}{e^{\lambda\varepsilon}} \leq e^{\alpha_{\mathcal{M}}(\lambda) - \lambda\varepsilon},$$

where the first inequality holds due to the Markov's inequality and the last inequality holds due to the definition of $\alpha_{\mathcal{M}}$.

Back to Mechanisms in DP-SGD

One-step Mechanism

$$\mathcal{M}_t(d) \coloneqq \sum_{i \in L_t} \tilde{\mathbf{g}}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 I)$$

- This is the Gaussian mechanism along with sampling from d to get L_t.
- It is DP (see Lemma 3 in this paper).
- However, this is a mechanism for at a given time step.

Multi-step Mechanism

$$\mathcal{M}(d) \propto \sum_{t=1}^{T} (-\eta_t) \mathcal{M}_t(d)$$

- Recall the DP-SGD update rule, *i.e.*, $\theta_T \leftarrow \theta_0 + \sum_{t=1}^{T} (-\eta_t) \mathcal{M}_t(d)$
- This is the composition of the Gaussian mechanisms.
- Is it DP?

Composibility Theorem

Theorem

Suppose that a mechanism $\mathcal{M}$ consists of a sequence of adaptive mechanisms, i.e., $\mathcal{M} := (\mathcal{M}_1, \ldots, \mathcal{M}_T)$, where $\mathcal{M}_t : \mathcal{R}_1 \times \cdots \times \mathcal{R}_{t-1} \times \mathcal{D} \to \mathcal{R}_t$. Then, for any $\lambda > 0$

$$\alpha_{\mathcal{M}}(\lambda) \leq \sum_{t=1}^{T} \alpha_{\mathcal{M}_t}(\lambda)$$

- "Adaptive" mechanism: a mechanism that depends on all previous mechanisms

$$\begin{aligned} \mathbf{aux}_2 &= \mathcal{M}_1(\mathbf{aux}_1, d) \\ \mathbf{aux}_3 &= \mathcal{M}_2(\mathbf{aux}_2, d) = \mathcal{M}_2(\mathcal{M}_1(\mathbf{aux}_1, d), d) \\ &\ldots \end{aligned}$$

- $\mathcal{M}$: *e.g.*, T-step gradient aggregation
- $\mathcal{M}_t$: *e.g.*, one-step gradient aggregation
- This theorem shares similar philosophy as a union bound.

Main DP Theorem for DP-SGD

Theorem

There exist constants c_1 and c_2 so that given the sampling probability $q = L/N$ and the number of steps T, for any $\varepsilon < c_1 q^2 T$, Algorithm 1 is (ε, δ)-differentially private for any $\delta > 0$ if we choose

$$\sigma \geq c_2 \frac{q\sqrt{T \log 1/\delta}}{\varepsilon}$$

- Provide intuition on tuning nobs.
- $\varepsilon \propto T$: privacy-accuracy trade-off
- With the known "strong composition" (*i.e.*, a baseline), we need

$$\sigma = \Omega\left(\frac{q\sqrt{T \log(1/\delta) \log(T/\delta)}}{\varepsilon}\right)$$

  - This is one without clipping.
  - This difference will be justified in experiments.

Practical Guideline to Compute ε

- The moments bound:

$$\alpha_{\mathcal{M}}(\lambda) \leq \sum_{i=1}^{T} \alpha_{\mathcal{M}_i}(\lambda)$$

- For the Gaussian mechanism with random sampling

$$\alpha_{\mathcal{M}_i}(\lambda) \leq (\text{computable upper bound})$$

  - See the paper for details.

- From the "Moment-DP" theorem, $\mathcal{M}$ is (ε, δ)-DP if

$$\min_{\lambda} e^{\alpha_{\mathcal{M}}(\lambda) - \lambda\varepsilon} \leq \min_{\lambda} e^{\sum_{i=1}^{T} \alpha_{\mathcal{M}_i}(\lambda) - \lambda\varepsilon} \leq \delta.$$

  - The above assumes that we can compute $\alpha_{\mathcal{M}_i}(\cdot)$ exactly.
  - If T, q, σ, and δ are given and conduct greedy search over ε (and solving $\min_\lambda$ via greedy search) to find the minimum ε.

(Proposed) Moments Accountant v.s. (Standard) Strong Composition

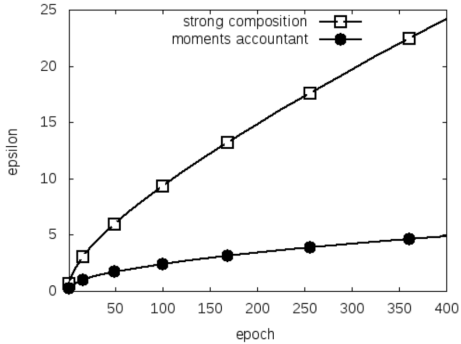


Figure 2: The ε value as a function of epoch E for $q = 0.01$, $\sigma = 4$, $\delta = 10^{-5}$, using the strong composition theorem and the moments accountant respectively.

(Proposed) Moments Accountant v.s. (Standard) Strong Composition

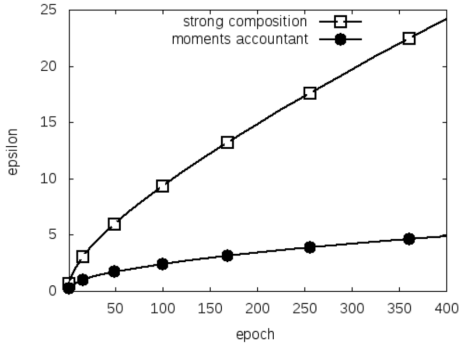


Figure 2: The ε value as a function of epoch E for $q = 0.01$, $\sigma = 4$, $\delta = 10^{-5}$, using the strong composition theorem and the moments accountant respectively.

- How about the comparison of model accuracy? Clipping may hurt accuracy.

Conclusion

- The proposed "Moments Accountant" has a stronger DP guarantee.
  - Why? partially due to practical treatments on clipping
- Nice connection between a moments bound and the DP guarantee.