Revisiting Differentially Private Hyper-parameter Tuning

Zihang Xiang*, Tianhao Wang+, Cheng-Long Wang*, Di Wang*
*: KAUST
+: University of Virginia

Background

Differential privacy (DP)

Two datasets $X, X' \subseteq \mathcal{X}$ are adjacent if they differ by only one data sample. A randomized algorithm $\mathcal{M}$ is $(\varepsilon, \delta) - DP$ if for all adjacent dataset $X, X' \subseteq \mathcal{X}$ and for all possible event S in the output space of $\mathcal{M}$, we have:

$$\Pr(\mathcal{M}(X) \in S) \leq e^{\varepsilon} \Pr(\mathcal{M}(X') \in S) + \delta$$

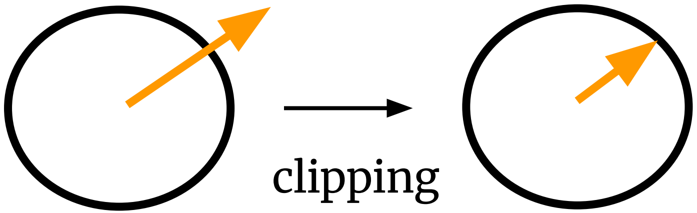
Algorithm that is DP provably defend a wide range of privacy attacks

Background

The famous DP-SGD algorithm guarantee DP, at its core:

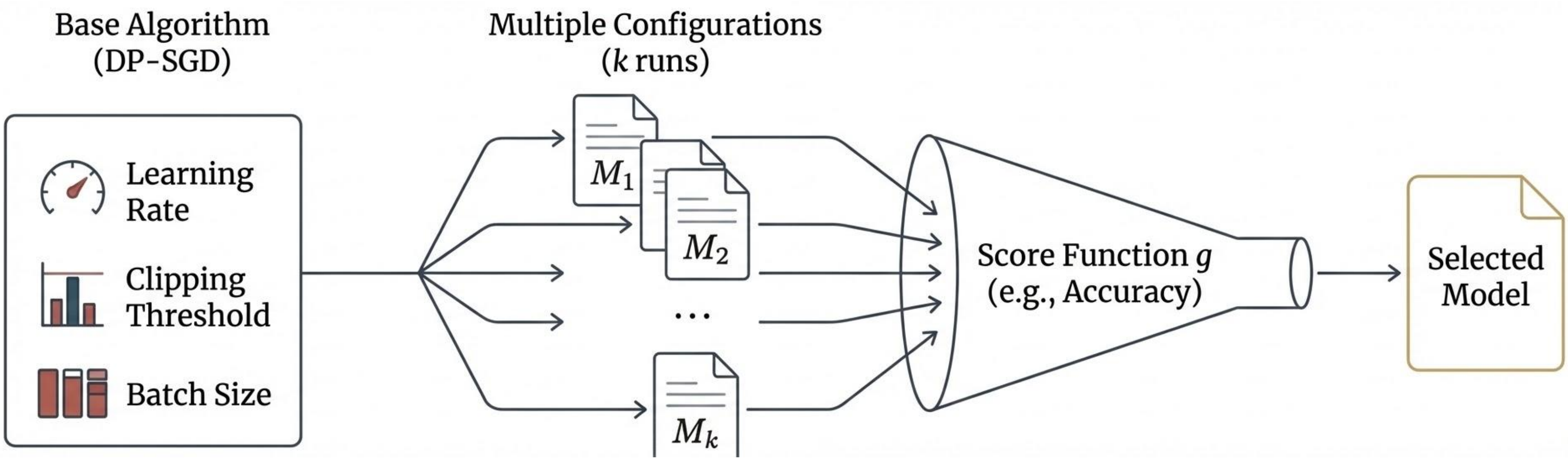
$$\theta_t = \theta_{t-1} - \eta \cdot g_{t-1}$$

$$\text{Noise} + \sum \text{Clipped per-example gradient}$$



[Abadi et al., CCS 2022]

Motivation



- DP-SGD often needs to tune parameters
  - Select one best model out of many
- Privacy analysis for a single run is well studies

Motivation

How to compute the privacy bound under hyperparameter tuning?

Naive method: Linear addup

Bound degradation scales with k,
or $\sqrt{k}$ if use advance composition

Unacceptable if k is large

[Papernot and Steinke, ICLR 2022]

SOTA private selection alg.

Algorithm 2 Private Selection Protocol $\mathcal{H}$ [42], [30]

Input: Dataset X; algorithms Ω; distribution ξ; score function g

1: Draw a sample: $k \leftarrow \xi$
2: $Y \leftarrow$ **Null**, $S \leftarrow -\infty$
3: **for** $i = 1, 2, \cdots, k$ **do**
4: Uniformly randomly fetch one element $\mathcal{M}_i$ from Ω
5: $y_i \leftarrow \mathcal{M}_i(X)$ ▷ Run $\mathcal{M}_i$ on dataset X
6: **If** $g(y_i) > S$: $Y \leftarrow y_i$, $S \leftarrow g(y_i)$ ▷ Selecting the "best"
7: **end for**

Output: Y

1. Run DP-SGD a random number of times
2. Then release the best one run
3. The privacy bound is better than linear addup

Motivation

Now we are interested in the following questions

1. Because **<u>only one run</u>** is released,

   Bound for privacy selection = Bound of base DP-SGD?

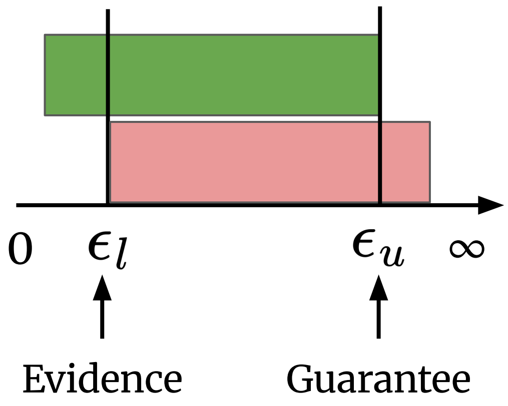
2. Although *[Papernot and Steinke, ICLR 2022]* have proposed private selection algorithm and improved bound, is it the best upper bound ?

Method

ϵ_u The privacy **upper** bound, given by theoretical analysis

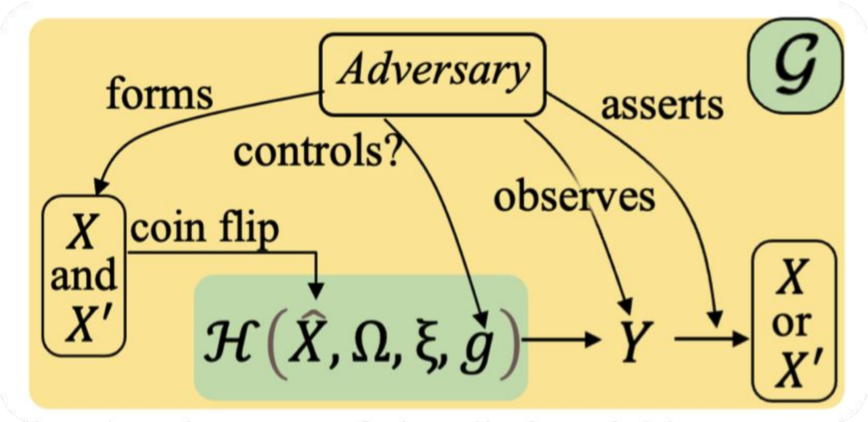
ϵ_l The privacy **lower** bound, given by privacy audit

$$\epsilon_l \leqslant \text{True privacy loss} \leqslant \epsilon_u$$



Method

We first estimate ϵ_u via ϵ_l by simulating the following game



Simulate it many times

Based on the simulations, we derive ϵ_l via existing tools

Experiment setups

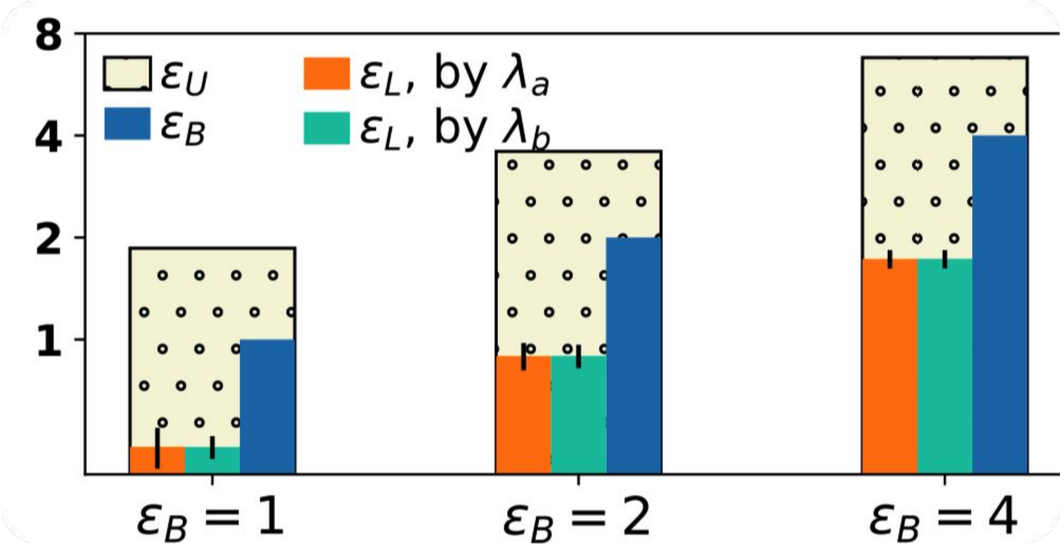
Setups	Adv. score	Adv. cap.
NTNV	No	Weak
NTCV	Yes	mid.
ETCV	Yes	strong

1. Normal training and normal validation (NTNV).
2. Normal training and controlled validation (NTCV)
3. Empty training and controlled validation (ETCV)

Method

Experiment result: **NTNV**

ϵ_B is the base algorithms upper bound

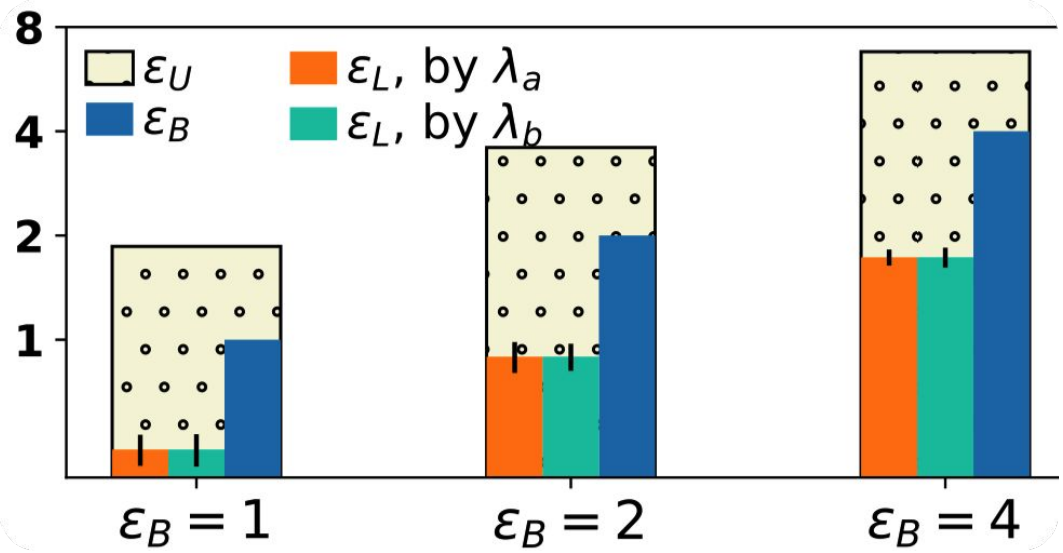


Finding:

1. We can't say much about ϵ_u
2. In practice, tuning DP-SGD leaks limited privacy

Experiment result: **NTCV**

ϵ_B is the base algorithms upper bound

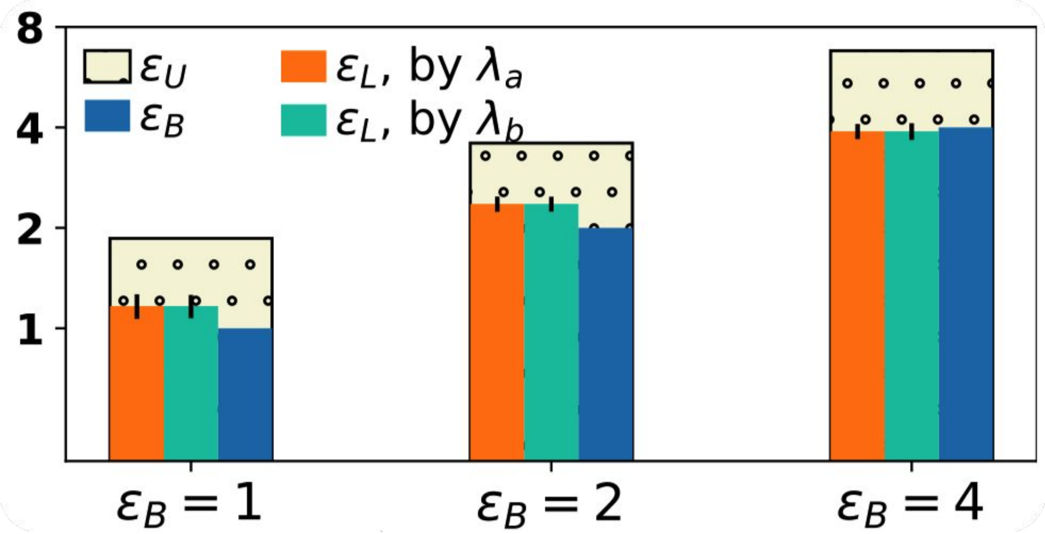


Finding:

1. Controlling the score function, i.e., selection criteria, does not help much.
2. In practice, adversarial selection will not expose more privacy risk

Experiment result: **ETCV**

ϵ_B is the base algorithms upper bound

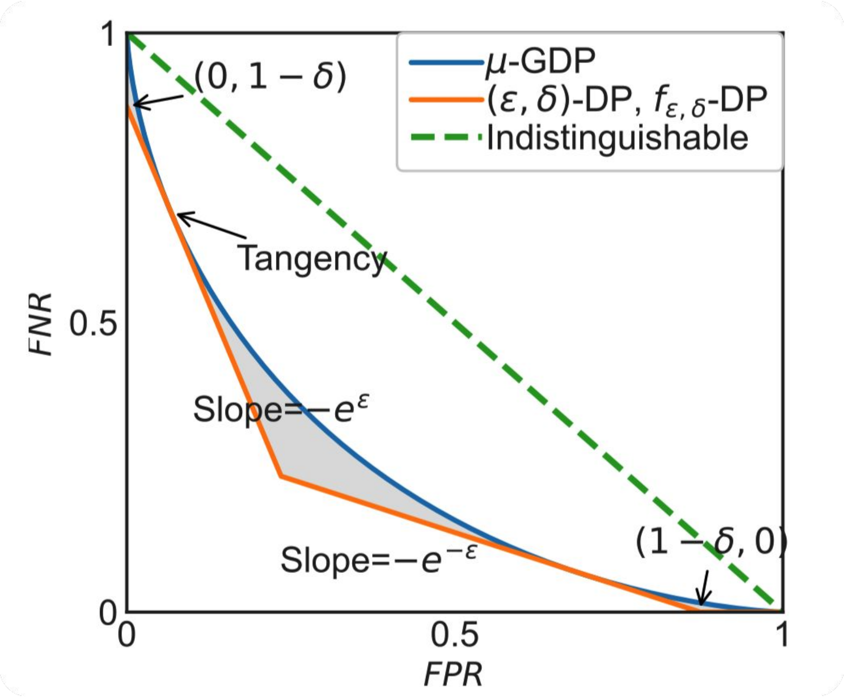


Finding:

1. $\epsilon_l > \epsilon_B$
2. The action of selection does incur more privacy leakage than the base DP-SGD
3. We then know that $\epsilon_u > \epsilon_B$

Method

Modeling the base algorithm the right way is the key



Key point, f-DP framework:

1. Previous $(\varepsilon, \delta) - DP$ claim for is loose for Gaussian mechanism
2. Gaussian mechanism has its tight trade-off function.

Method

We also need to find the worse-case score function

Theorem 2 (Necessary worst-case g, proof in Appendix E)**.** *Let distribution P be over some finite alphabets Γ, and define a distribution $F_{k,g}$ as follows.*

First, make $k > 0$ independent samples $\{x_1, x_2, \cdots, x_k\}$ from P; second, output x_i such that the score $g(x_i)$ computed by a score function $g : \Gamma \to \mathbb{R}$ is the maximum over these samples. Similarly, we define another distribution P' over the same alphabets Γ and derive a distribution $F'_{k,q}$ as the counterpart to $F_{k,g}$.

*For any score function $\hat{g}$, which is **not** a one-to-one mapping (hence a randomized tie-breaking is needed), there always exists a one-to-one mapping g^* satisfying*

$$\mathcal{D}_\alpha(F_{k,\hat{g}}||F'_{k,\hat{g}}) \leq \mathcal{D}_\alpha(F_{k,g^*}||F'_{k,g^*}). \tag{15}$$

Moreover, similar inequality also holds when k follows a general distribution ξ.

TL;DR:

1. One-to-one mapping score function leaks privacy the most
2. This is the necessary condition used to find the upper bound for privacy selection

Method

General form to compute the upper bound for any base algorithm

Theorem 3 (General form, proof in Appendix F). *Suppose the base algorithm is f-DP, then $\mathcal{H}$ is $(\varepsilon_{\mathcal{H}}, \delta_{\mathcal{H}})$-DP where*

$$\varepsilon_{\mathcal{H}} = \varepsilon + \max_{a \in [0,1]} \log \frac{\omega_\xi(1-a)}{\omega_\xi(b)}, \tag{19}$$

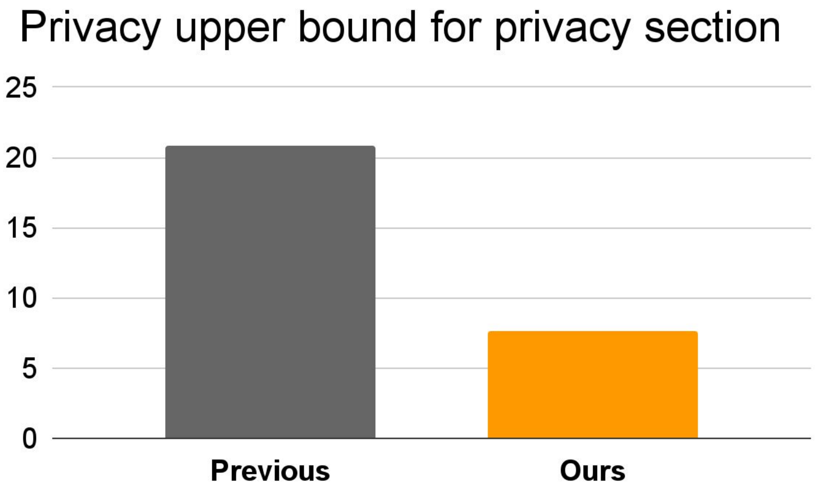
where $b = f(a)$ and ε is computed by Algorithm 3 whose two input arguments are the trade-off function f and $\delta = \delta_{\mathcal{H}}/\omega_\xi(1)$ (ω_ξ is defined in Equation (17)).

TL;DR:

1. The upper bound for private selection deteriorate by an a**dditive factor**
2. The factor is determined by the distribution specification in private selection algorithm

Result

Improved result for Gaussian mechanism



Base algorithm DP-SGD is (4.36, 1e-5)-DP

Improved empirical gain at the same privacy level for free

ε_B	$\varepsilon_{\mathcal{H}}^{O}$	Previous $\rightarrow$ Ours			
		MNIST	FMNIST	CIFAR10	SVHN
1	1.83	$0.921 \rightarrow 0.934$	$0.768 \rightarrow 0.793$	$0.412 \rightarrow 0.448$	$0.636 \rightarrow 0.661$
2	3.43	$0.942 \rightarrow 0.956$	$0.779 \rightarrow 0.802$	$0.467 \rightarrow 0.486$	$0.706 \rightarrow 0.745$
4	6.69	$0.951 \rightarrow 0.958$	$0.791 \rightarrow 0.817$	$0.504 \rightarrow 0.531$	$0.762 \rightarrow 0.786$

Conclusion

Audit shows that private selection indeed leak more privacy

Theoretical upper bound by modeling the base alg. via f-DP

Experiment shows that improve upper bound gives utility gain