On the Generalization Error of Differentially Private Algorithms via Typicality

Yanxiao Liu[†], Chun Hei Michael Shiu[⋆], Lele Wang[⋆] and Deniz Gündüz[†]

[†]Imperial College London
[⋆]The University of British Columbia

Overview

- *Generalization error* measures the gap between the training loss and the population loss of a learned algorithm.
- Information-theoretic generalization bounds: an algorithm that leaks little information generalizes well.
- Differentially private algorithms admit sharper generalization bounds.
- The information leakage can be measured by mutual information and maximal leakage.
- Existing typicality-based arguments (Rodríguez-Gálvez et al., 2021) provide easily computable bounds of mutual information.
- Firstly, we further refine these bounds by exploiting differential privacy.
- Secondly, we provide explicit bounds of maximal leakage.

Generalization Error

- $S = (Z_1, \ldots, Z_n)$ is the training dataset where $Z_i \sim P_Z$ are i.i.d..
- A stochastic learning algorithm $P_{W|S}$ can be considered as a *channel* $P_{W|S}$, mapping $S = (Z_1, \ldots, Z_n)$ to a hypothesis W.
- Given a dataset $S = s$, the algorithm outputs W following $P_{W|S=s}$.
- $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}$ is the loss function and usually assumed to be σ-sub-Gaussian, i.e., for all $\lambda \in \mathbb{R}$ and all w,

$$\mathbf{E}_{Z\sim P_Z}\left[\exp\left(\lambda(\mathbf{E}[\ell(w,Z)] - \ell(w,Z))\right)\right] \leq \exp\left(\lambda^2\sigma^2/2\right).$$

- The generalization error is

$$\mathrm{gen}(S,W) := \mathbb{E}_{P_Z}[\ell(W,Z)] - \frac{1}{n}\sum_{i=1}^{n}\ell(W,Z_i).$$

Information-Theoretic Generalization Error Bounds

- By Xu and Raginsky (2017); Russo and Zou (2016), $\mathrm{gen}(S, W)$ can be bounded in terms of mutual information $I(S, W)$:

$$|\mathbf{E}\left[\mathrm{gen}(W, S)\right]| \leq \sqrt{\frac{2\sigma^2}{n} I(S; W)}. \tag{1}$$

- Intuition: an algorithm that leaks little information about the dataset generalizes well (Bassily et al., 2018).

Information-Theoretic Generalization Error Bounds

- Besides expectation bound, one can also consider *high-probability* bounds: for $\eta \in (0,1)$ (Esposito et al., 2021),

$$\mathbf{P}\left(\left|\mathrm{gen}(W,S)\right| \geq \eta\right) \leq 2\exp\left(\mathcal{L}(S \to W) - \frac{n\eta^2}{2\sigma^2}\right). \tag{2}$$

- $\mathcal{L}(S \to W)$ is the maximal leakage (Issa et al., 2016): given a joint distribution P_{SW} on finite alphabets $\mathcal{S}, \mathcal{W}$,

$$\mathcal{L}(S \to W) = \log \sum_{w \in \mathcal{W}} \max_{s \in \mathcal{S}: P_S(s) > 0} P_{W|S}(w|s).$$

- (2) provides an exponential decay in the probability of having a large generalization error, and $\mathcal{L}(S \to W)$ depends on P_S only through its support.

Differential Privacy (DP)

- We consider differentially private algorithms (Dwork et al., 2006).
- Let $S, S' \in \mathcal{Z}^n$ be neighboring datasets, i.e., they differ in one data point. A randomized algorithm $\mathcal{A}$ is said to satisfy (ε, δ)-DP if for every measurable set $\mathcal{V} \subseteq \mathcal{W}$,

$$\mathbf{P}\big(\mathcal{A}(S) \in \mathcal{V}\big) \leq e^{\varepsilon}\mathbf{P}\big(\mathcal{A}(S') \in \mathcal{V}\big) + \delta.$$

- A *stable* algorithm should produce similar outputs on similar input datasets, and stable algorithms generalize well (Bassily et al., 2016).
- In this work we consider the case where $\delta = 0$, i.e., the pure DP.

Typicality Bounds for DP Algorithms

- For information measures like $I(S;W)$ and $\mathcal{L}(S \to W)$, they are not easy to compute exactly.
- For generalization error, we intend to have easily computable bounds.
- A natural question is: can we upper bound

$$I(S;W) \quad \text{and} \quad \mathcal{L}(S \to W)$$

  by explicit functions of $(n, |\mathcal{Z}|)$ (or $(n, \varepsilon, |\mathcal{Z}|)$)?
- For $I(S;W)$, Rodríguez-Gálvez et al. (2021) used method of types and covering to derive explicit bounds.
- We will first explain their techniques, and then show our contributions.
- Similar to Rodríguez-Gálvez et al. (2021), we consider:
  - discrete algorithms;
  - permutation-invariant algorithms.
- Under permutation invariance, the algorithm depends on the dataset only through the multiset of samples, equivalently through its *type*.

Why Types Are Useful

- The *type* of a dataset S is its empirical distribution:

$$T_S(a) = \frac{\mathrm{N}(a \mid S)}{n}, \qquad a \in \mathcal{Z}.$$

- The number of possible datasets is exponential in n, but the number of possible types is only polynomial:

$$|\mathcal{T}_{\mathcal{Z},n}| \leq (n+1)^{|\mathcal{Z}|-1}.$$

- Instead of controlling all datasets, one controls only their types.
- We use the multiset distance

$$\mathrm{d}(s, s') := \frac{1}{2} \sum_{a \in \mathcal{Z}} \left| \mathrm{N}(a \mid s) - \mathrm{N}(a \mid s') \right| = \frac{n}{2} \| T_s - T_{s'} \|_1.$$

- If the algorithm is ε-DP, then nearby datasets should produce similar output laws.

Mutual Information Bounds

- We have the following relation (Polyanskiy and Wu, 2025):

$$I(S;W) \leq \mathbb{E}_{s\sim P_S}\left[D_{\mathrm{KL}}(P_{W|S=s}\|Q_W)\right],$$

  where the equality is achieved if and only if $Q_W = P_W$.
- The problem becomes: choose a suitable reference distribution Q_W and bound

$$D_{\mathrm{KL}}(P_{W|S=s}\|Q_W).$$

- The key idea of Rodríguez-Gálvez et al. (2021) is to choose Q_W as a mixture of output laws corresponding to a finite representative set of datasets.

Covering Construction

- Work with the count vector $\mathrm{N}_s = (\mathrm{N}(a \mid s) : a \in \mathcal{Z})$.
- Partition the cube into a $t \times \cdots \times t$ grid and choose one representative dataset from each nonempty cell.
- This produces a representative set (*covering*)

$$\mathcal{S}_0 = \{s_1, \ldots, s_M\}, \qquad M \leq t^{|\mathcal{Z}|-1}.$$

- Every dataset s is close to some representative s_i, with

$$\mathrm{d}(s, s_i) \lesssim (|\mathcal{Z}| - 1)\frac{n}{t}.$$

Background

- If we take the mixture over *all* types:

$$Q_W = \sum_{s' \in \mathcal{S}} \omega_{s'} P_{W|T_{s'}},$$

(Rodríguez-Gálvez et al., 2021, Lemma 2) gives

$$D_{\mathrm{KL}}(P_{W|S=s} \| Q_W) \leq \min_{s' \in \mathcal{S}} \left\{ D_{\mathrm{KL}}(P_{W|S=s} \| P_{W|T_{s'}}) - \log \omega_{s'} \right\}.$$

- If the weights are uniform, $\omega_s = |\mathcal{S}|^{-1}$ for all $s \in \mathcal{S}$, we have

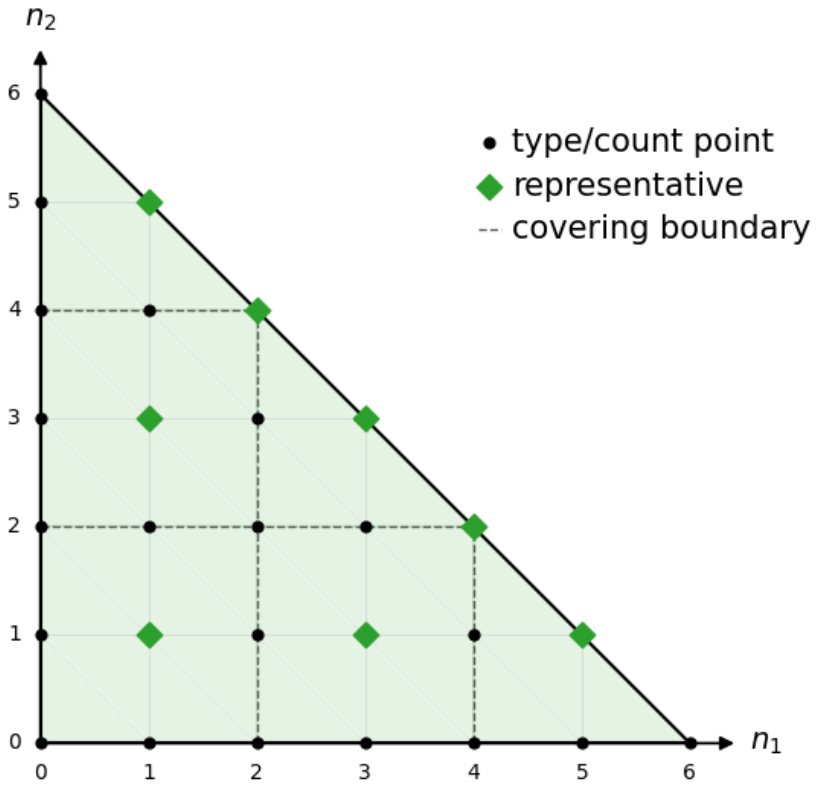
$$D_{\mathrm{KL}}(P_{W|S=s} \| Q_W) \leq (|\mathcal{Z}| - 1) \log(N + 1).$$

- For private (stable) algorithms, the idea is to choose one representative dataset per covering cell,

$$\mathcal{S}_0 = \{s_1, \ldots, s_M\}, \qquad Q_W^{(t)} := \frac{1}{M} \sum_{i=1}^{M} P_{W|S=s_i}.$$

Covering Argument

- This part is from (Rodríguez-Gálvez et al., 2021).
- All types T_s lie inside the hypercube $[0, N]^{|\mathcal{Z}|-1}$, since the last dim is determined by the preceding $|\mathcal{Z}| - 1$.
- Split $[0, N]$ into t parts, get a hypercube cover with side length $\ell := N/t$.
- Choose the *center atom* of each small hypercube, its distance to any other atom in the hypercube $\leq (\lfloor \ell \rfloor + 1)/2$.
- Hence the distance between any dataset s and its center s_i is $\leq (N/t)(|\mathcal{Z}| - 1)$.



$N = 6$, $t = 3$ and $|\mathcal{Z}| = 3$
Black dots: feasible datasets
Green squares: mixture representatives

Covering Argument

- We then use the previous lemma to bound $D_{\mathrm{KL}}(P_{W|S=s} \| Q_W)$.
- By group privacy and the covering uses at most $M \leq t^{|\mathcal{Z}|-1}$ representatives, the previous lemma yields:

  if $1/n \leq \varepsilon \leq 1$,

$$D_{\mathrm{KL}}(P_{W|S=s} \| Q_W) \leq (|\mathcal{Z}| - 1) \log(e\varepsilon n).$$

  If $\varepsilon n < 1$, instead set $t = 1$, giving

$$D_{\mathrm{KL}}(P_{W|S=s} \| Q_W) \leq (|\mathcal{Z}| - 1)\varepsilon n.$$

Our Improvement on the Bounds

- The step we want to improve is the mixture step

$$D_{\mathrm{KL}}(P\|Q) \le D_{\mathrm{KL}}(P\|Q_i) - \log \omega_i.$$

- This only uses the weight of the *one selected component* Q_i.
- For a DP algorithm, nearby output laws have substantial overlap.
- Let

$$Q = \sum_{b=1}^{M} \omega_b Q_b, \qquad \omega_b > 0, \qquad \sum_{b=1}^{M} \omega_b = 1.$$

- Fix an index i, and assume $Q_b(E) \ge \alpha_{b,i} Q_i(E),\ \forall E$ for some $\alpha_{b,i} \in [0,1]$.
- Then for any $P \ll Q_i$,

$$D_{\mathrm{KL}}(P\|Q) \le D_{\mathrm{KL}}(P\|Q_i) - \log\left(\sum_{b=1}^{M} \omega_b \alpha_{b,i}\right),$$

which is at least as good as (Rodríguez-Gálvez et al., 2021).

Why DP Fits the Overlap Lemma

- If $Q_b = P_{W|S=s_b}$ and $Q_i = P_{W|S=s_i}$ are ε-DP, then for all measurable E,

$$P_{W|S=s_b}(E) \geq e^{-\varepsilon \mathrm{d}(s_b, s_i)} P_{W|S=s_i}(E).$$

- Here the overlap lemma applies with

$$\alpha_{b,i} = e^{-\varepsilon \mathrm{d}(s_b, s_i)}.$$

- Using the same covering argument with our new mixture lemma, we obtain:

$$D_{\mathrm{KL}}(P_{W|S=s} \| Q_W) \leq \min_{t \in \{1,\ldots,n\}} \Big\{ (|\mathcal{Z}| - 1)\frac{\varepsilon n}{t} + (|\mathcal{Z}| - 1) \log t - \mathbb{1}_{\{t \geq 2\}} \log\left(1 + e^{-\varepsilon n}\right) \Big\}.$$

where $-\mathbb{1}_{\{t \geq 2\}} \log(1 + e^{-\varepsilon n})$ is from the overlap gain.

Further Improvements

- We also combine the overlap idea with the two refinements on the covering, both are used in (Rodríguez-Gálvez et al., 2021).
- **Refinement 1:** a better counting on the hypercubes.
- This replaces $(|\mathcal{Z}| - 1) \log t$ by the sharper term

$$(|\mathcal{Z}| - 1) \log \left(t + \frac{|\mathcal{Z}| - 2}{2} \right) - \log((|\mathcal{Z}| - 1)!).$$

- **Refinement 2:** restrict to the strongly typical set.
- This yields

$$I(S; W) \lesssim \min_t \left\{ \frac{\varepsilon |\mathcal{Z}| \sqrt{n \log n}}{t} + \text{counting term} - \text{overlap gain} \right\}.$$

- Our work in this part keeps their covering and typicality machinery, but strengthens the mixture step and analyses.

Comparison of Bounds

The solid lines are our bounds, and the same color refers to the same covering scheme.

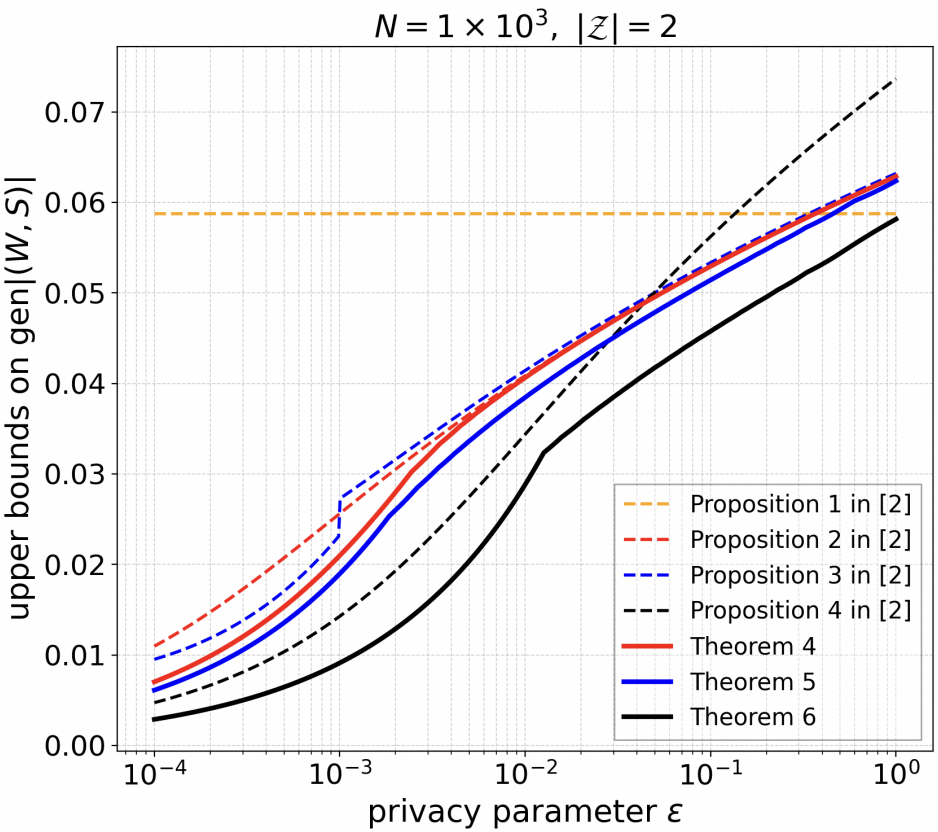


Figure: Dataset of size $n = 1 \times 10^3$ with $|\mathcal{Z}| = 2$.

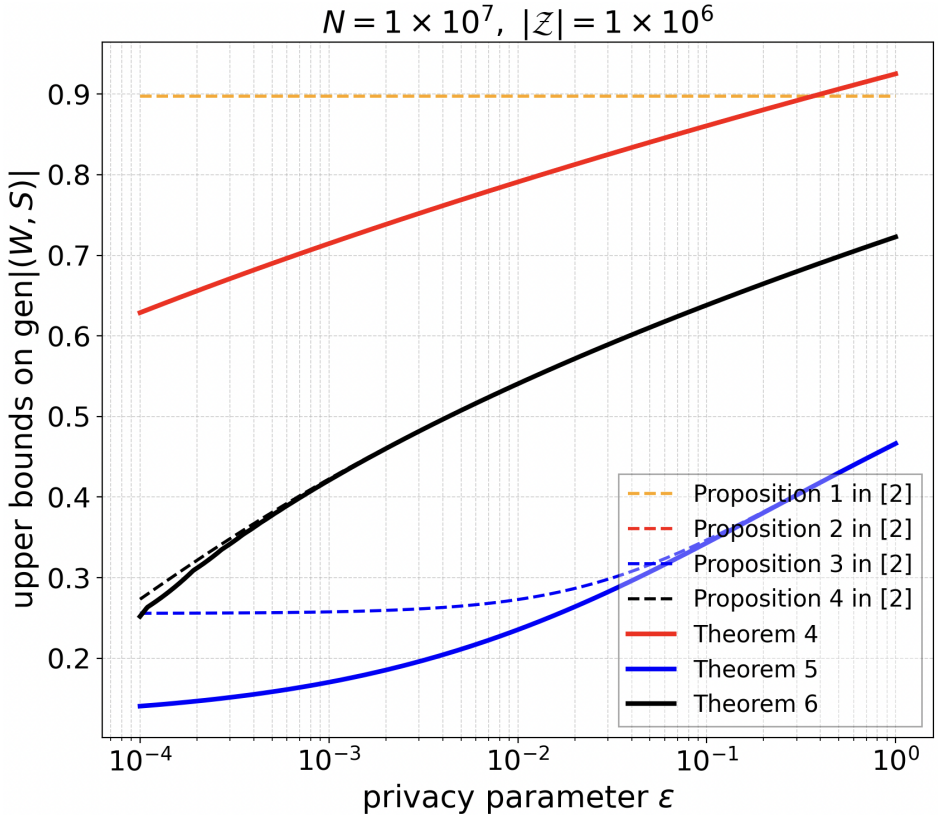


Figure: Dataset of size $n = 1 \times 10^7$ with $|\mathcal{Z}| = 1 \times 10^6$.

Maximal Leakage

- Rodríguez-Gálvez et al. (2021) focused only on mutual information.
- We derive explicit bounds for maximal leakage, which imply high-probability generalization guarantees via (Esposito et al., 2021)

$$\mathbb{P}(|\mathrm{gen}(S,W)| \geq \eta) \leq 2\exp\left(\mathcal{L}(S \to W) - \frac{n\eta^2}{2\sigma^2}\right).$$

- Since maximal leakage contains a pointwise maximum over datasets,

$$\mathcal{L}(S \to W) = \log \sum_{w \in \mathcal{W}} \max_{s: P_S(s)>0} P_{W|S=s}(w),$$

  the approach for mutual information cannot be used directly.
- We instead use the covering to reduce the maximum over all datasets to a maximum over representative datasets.

From Mixtures to Envelopes

- For mutual information, the representative set is used to construct a mixture

$$Q_W^{(t)} = \frac{1}{M} \sum_{i=1}^{M} P_{W|S=s_i}.$$

- For maximal leakage, we choose a representative set to control the envelope

$$\sum_{w \in \mathcal{W}} \max_{s: P_S(s)>0} P_{W|S=s}(w).$$

- If every dataset s is close to a representative s_i with $\mathrm{d}(s, s_i) \leq R$, we have

$$P_{W|S=s}(w) \leq e^{\varepsilon R} P_{W|S=s_i}(w).$$

- Hence

$$\max_s P_{W|S=s}(w) \leq e^{\varepsilon R} \max_{i \in [M]} P_{W|S=s_i}(w).$$

A Simple Tool for Maximal Leakage

- Let $Q_1, \ldots, Q_M$ be probability mass functions on $\mathcal{W}$.
- If, for every b, i,

$$Q_b(w) \geq \alpha_{b,i} Q_i(w), \qquad \forall w \in \mathcal{W},$$

then

$$\sum_{w \in \mathcal{W}} \max_{1 \leq i \leq M} Q_i(w) \leq \frac{M}{C}, \qquad C := \min_i \sum_{b=1}^{M} \alpha_{b,i}.$$

- For an ε-DP algorithm, taking $Q_i = P_{W|S=s_i}$ gives

$$\alpha_{b,i} = e^{-\varepsilon \mathrm{d}(s_b, s_i)}.$$

- Let $\mathcal{S}_0 = \{s_1, \ldots, s_M\}$ be a representative set with $\min_i \mathrm{d}(s, s_i) \leq R$ and define $C_{\mathcal{S}_0} := \min_i \sum_{b=1}^{M} e^{-\varepsilon \mathrm{d}(s_b, s_i)}$, we have

$$\mathcal{L}(S \to W) \leq \varepsilon R + \log M - \log C_{\mathcal{S}_0}.$$

- If we use the same covering method as we used, we have:

$$\mathcal{L}(S \to W) \leq \min_{1 \leq t \leq n} \left\{ \varepsilon \min \left\{ n, \frac{kn}{t} \right\} + k \log t - \log \left(1 + (t^k - 1) e^{-n\varepsilon} \right) \right\}.$$

Maximal Leakage Bound

- If we improve the covering scheme, we have a strictly tighter bound:

$$\mathcal{L}(S \to W) \leq \min_{1 \leq t \leq n} \left\{ \varepsilon \min\left\{ n, \frac{kn}{t} \right\} + \log S_k(t) - \log\left(1 + (S_k(t) - 1)e^{-n\varepsilon}\right) \right\}.$$

  where $S_k(t) := \binom{t+k-1}{k}$ and $k = |\mathcal{Z}| - 1$.

- If we use strong typicality, we have

$$\mathcal{L}(S \to W) \lesssim \min_{t} \left\{ \varepsilon \min\left\{ n, \frac{m\sqrt{n \log n}}{t} \right\} + \log A_m(t) - \log\left(1 + (A_m(t) - 1)e^{-n\varepsilon}\right) \right\}$$

  where $A_m(t) := \min\{t^m, mt^{m-1}\}$ and $m := |\mathcal{Z}|$.

Comparison of Maximal Leakage Bounds

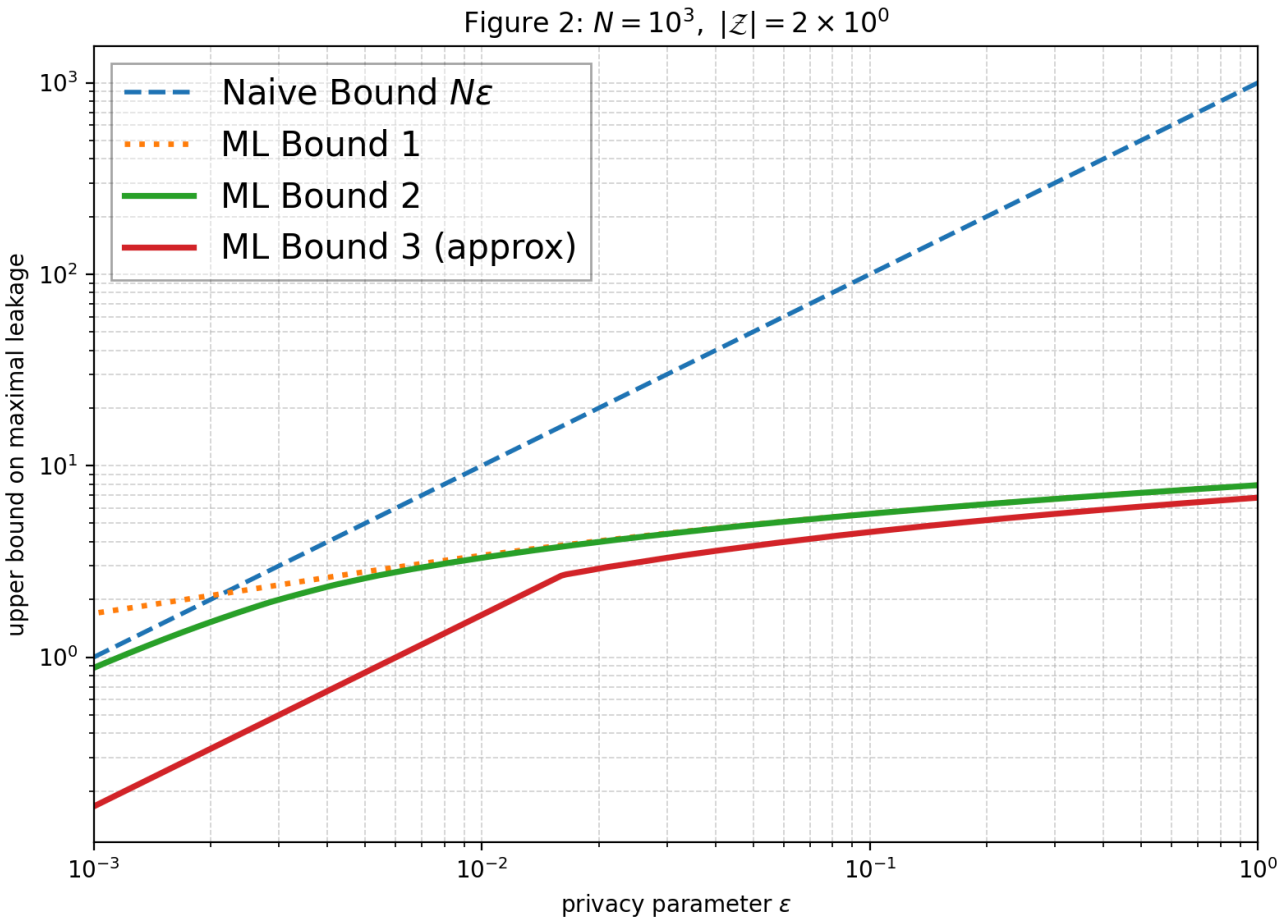


Figure: $n = 1 \times 10^3$, $|\mathcal{Z}| = 2$.

Summary

- Generalization error can be bounded in terms of mutual information and maximal leakage.
- We are interested in explicit, easily computable bounds of both.
- We improve the mutual information bounds in Rodríguez-Gálvez et al. (2021) by further exploiting the property of differential privacy.
- We provide new maximal leakage bounds.

Future Directions

- We only considered pure ε-DP algorithms, there exist other privacy measures, e.g., Gaussian DP (Dong et al., 2022) and Rényi DP (Mironov, 2017), to be considered.
- We only considered discrete algorithms, one can consider extending the analysis to continuous algorithms as well.
- We only considered mutual information and maximal leakage, there exist generalization bounds via other information measures.

Acknowledgement

Thank you!

This work was partially supported by UKRI under the projects AI-R (EP/X030806/1) and INFORMED-AI (EP/Y028732/1), and also supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant No. RGPIN-2019-05448.

References I

Bassily, R., Moran, S., Nachum, I., Shafer, J., and Yehudayoff, A. (2018). Learners that use little information. In *Algorithmic Learning Theory*, pages 25–55. PMLR.

Bassily, R., Nissim, K., Smith, A., Steinke, T., Stemmer, U., and Ullman, J. (2016). Algorithmic stability for adaptive data analysis. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 1046–1059.

Dong, J., Roth, A., and Su, W. J. (2022). Gaussian differential privacy. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):3–37.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Third Theory of Cryptography Conference, TCC 2006, New York, NY, March*, pages 265–284.

Esposito, A. R., Gastpar, M., and Issa, I. (2021). Generalization error bounds via Rényi-, f-divergences and maximal leakage. *IEEE Transactions on Information Theory*, 67(8):4986–5004.

Issa, I., Kamath, S., and Wagner, A. B. (2016). An operational measure of information leakage. In *Conference on Information Science and Systems (CISS)*, pages 234–239. IEEE.

Mironov, I. (2017). Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pages 263–275. IEEE.

Polyanskiy, Y. and Wu, Y. (2025). *Information theory: From coding to learning*. Cambridge University Press.

Rodríguez-Gálvez, B., Bassi, G., and Skoglund, M. (2021). Upper bounds on the generalization error of private algorithms for discrete data. *IEEE Transactions on Information Theory*, 67(11):7362–7379.

Russo, D. and Zou, J. (2016). Controlling bias in adaptive data analysis using information theory. In *Artificial Intelligence and Statistics*, pages 1232–1240. PMLR.

Xu, A. and Raginsky, M. (2017). Information-theoretic analysis of generalization capability of learning algorithms. *Advances in Neural Information Processing Systems*, 30.