Journal of Machine Learning Research 26 (2025) 1-54 Submitted 9/24; Revised 10/25; Published 10/25

Fundamental Limits of Membership Inference Attacks on Machine Learning Models

Eric Aubinais ERIC.AUBINAIS@UNIVERSITE-PARIS-SACLAY.FR
Université Paris-Saclay, CNRS,
Laboratoire de mathématiques d'Orsay, 91405, Orsay, France

Elisabeth Gassiat ELISABETH.GASSIAT@UNIVERSITE-PARIS-SACLAY.FR
Université Paris-Saclay, CNRS,
Laboratoire de mathématiques d'Orsay, 91405, Orsay, France

Pablo Piantanida PABLO.PIANTANIDA@MILA.QUEBEC
ILLS - International Laboratory on Learning Systems,
MILA - Quebec AI Institute, Montreal (QC), Canada,
CNRS, CentraleSupélec — Université Paris-Saclay

Editor: Benjamin Guedj

Abstract

Membership inference attacks (MIA) can reveal whether a particular data point was part of the training dataset, potentially exposing sensitive information about individuals. This article provides theoretical guarantees by exploring the fundamental statistical limitations associated with MIAs on machine learning models at large. More precisely, we first derive the statistical quantity that governs the effectiveness and success of such attacks. We then theoretically prove that in a non-linear regression setting with overfitting learning procedures, attacks may have a high probability of success. Finally, we investigate several situations for which we provide bounds on this quantity of interest. Interestingly, our findings indicate that discretizing the data might enhance the learning procedure's security. Specifically, it is demonstrated to be limited by a constant, which quantifies the diversity of the underlying data distribution. We illustrate those results through simple simulations.

Keywords: Membership Inference Attacks, Statistical Limitations, Privacy, Theoretical Performance Bounds, Overfitting, Trustworthy Machine Learning.

1. Introduction

In today's data-driven era, machine learning models are designed to reach higher performance, and the size of new models will then inherently increase, therefore the information stored (or memorized) in the parameters (Hartley and Tsaftaris, 2022; Del Grosso et al., 2023). The protection of sensitive information is of paramount importance. Membership Inference Attacks (MIAs) have emerged as a concerning threat, capable of unveiling whether a specific data point was part of the training dataset of a machine learning model (Shokri et al., 2017; Song et al., 2017a; Nasr et al., 2019; Zhu et al., 2019). Such attacks can potentially compromise individual privacy and security by exposing sensitive information (Carlini et al., 2023b). Furthermore, the recent publication by Tabassi et al. (2019) from the National

Institute of Standards and Technology (NIST) explicitly highlights that a membership inference attack (MIA) which successfully identifies an individual as being part of the dataset used to train a target model constitutes a breach of confidentiality. This raises a crucial question: How should we evaluate and certify privacy in machine learning models?

To date, the most comprehensive defense mechanism against privacy attacks is Differential Privacy (DP), a framework initially introduced by Dwork et al. (2006). DP has shown remarkable adaptability in safeguarding the privacy of machine learning models during training, as demonstrated by the works of Jayaraman and Evans (2019); Hannun et al. (2021). However, it is worth noting that achieving a high level of privacy through differentially private training often comes at a significant cost to the accuracy of the model, especially when aiming for a low privacy parameter (Sablayrolles et al., 2019). Conversely, when evaluating the practical effectiveness of DP in terms of its ability to protect against privacy attacks empirically, the outlook is considerably more positive. DP has demonstrated its efficacy across a diverse spectrum of attacks, encompassing MIAs, attribute inference, and data reconstruction (see Guo et al. (2023) and references therein). DP has been extensively used to understand the performances of MIAs against learning systems Thudi et al. (2022) or how a mechanism could be introduced to defend oneself against MIAs He et al. (2022); Izzo et al. (2022).

Empirical evidence suggests that small models compared to the size of training set are often sufficient to thwart the majority of existent threats and empirically summarized in Baluta et al. (2022). Similarly, when the architecture of a machine learning model is overcomplex with respect to the size of the training set, model overfitting increases the effectiveness of MIAs, as has been identified by Shokri et al. (2017); Yeom et al. (2018); He et al. (2022); Del Grosso et al. (2023). However, despite these empirical findings, there remains a significant gap in our theoretical understanding of this phenomenon. Most of the existing literature on MIAs focuses on developing and analyzing specific, albeit efficient, attack strategies. While these contributions have been crucial in highlighting the privacy risks associated with MIAs, they fall short when it comes to auditing the privacy risks of ML models comprehensively. The failure of specific attacks does not guarantee that others will not succeed, meaning that individual attack strategies alone cannot provide a full certification of privacy in ML systems.

In this article, we explore the theoretical statistical principles underlying the privacy limitations of learning procedures in machine learning systems on a broad scale. Our investigation commences by establishing the **fundamental statistical quantity that governs the effectiveness and success of MIA attacks.** In the learning model we are examining, our primary focus is on learning procedures that are characterized as functions of the empirical distribution of their training data. Specifically, we concentrate on datasets of independent and identically distributed (*i.i.d.*) samples. To evaluate the feasibility and limitations of MIAs on a learning procedure, we will assess their **accuracy** by measuring the weighted probability of successfully determining membership. Notably, we assess the security of a learning procedure based on the highest level of accuracy achieved among all theoretically feasible MIAs. To this end, we delve into the intricacies of MIA and derive insights into the key factors that influence its outcomes. Subsequently, we explore various scenarios: overfitting learning procedures, empirical mean-based learning procedures and discrete data, among others, presenting bounds on this central statistical quantity

1.1 Contributions

In our research, we make theoretical contributions to the understanding of MIAs on machine learning models. Our key contributions can be summarized as follows:

- **Identification of the Central** Statistical Quantity: We introduce the critical statistical quantity denoted as $\Delta_{\nu,\lambda,n}(P,\mathcal{A})$, where ν is the theoretical fraction of *non-member* samples, λ is the importance accorded to the error of type I, n represents the size of the training dataset, P is the data distribution, and $\mathcal{A}$ is the underlying learning procedure. This quantity plays a pivotal role in assessing the feasibility and limitations of MIAs. We show that the quantity $\Delta_{\nu,\lambda,n}(P,\mathcal{A})$ is an $f-$divergence, which provides an intuitive measure of how distinct parameters of a model can be with respect to a sample in the training set, and as a result, it indicates the extent to which we can potentially recover sample membership through MIAs. Consequently, we demonstrate that when $\Delta_{\nu,\lambda,n}(P,\mathcal{A})$ is small, the accuracy of the best MIA is notably constrained. Conversely, when $\Delta_{\nu,\lambda,n}(P,\mathcal{A})$ approaches 1, the best MIA is successful with high probability. This highlights the importance of $\Delta_{\nu,\lambda,n}(P,\mathcal{A})$ in characterizing information disclosure in relation to the training set.

- **Lower Bounds for Overfitting Learning Procedures:** For learning procedures that overfit with high probability, we exhibit a lower bound on $\Delta_{\nu,\lambda,n}(P,\mathcal{A})$ (see Theorem 11). In a (non-)linear regression setting, we further theoretically demonstrate that learning procedures for which small training loss is reached, loss-based MIAs can achieve almost perfect inference, as illustrated in Section 6 by numerical experiments. Up to our knowledge, this is the first theoretical proof that overfitting indeed opens the way to successful MIAs.

- **Precise Upper Bounds for Empirical Mean-Based Learning Procedures:** For learning procedures that compute functions of empirical means, we establish upper bounds on $\Delta_{\nu,\lambda,n}(P,\mathcal{A})$. We prove that $\Delta_{\nu,\lambda,n}(P,\mathcal{A}))$ is bounded from above by a constant, determined by $(P,\mathcal{A})$, multiplied by $n^{-1/2}$. In practical terms, this means that having $\Omega(\varepsilon^{-2})$ samples in the dataset is sufficient to ensure that $\Delta_{\nu,\lambda,n}(P,\mathcal{A})$ remains below ε for any $\varepsilon \in (0,1)$.

- **Maximization of $\Delta_{\nu,\lambda,n}(P,\mathcal{A})$:** In scenarios involving discrete data (e.g., tabular data sets), we provide a precise formula for maximizing $\Delta_{\nu,\lambda,n}(P,\mathcal{A})$ across all learning procedures $\mathcal{A}$. Additionally, under specific assumptions, we determine that this maximization is proportional to $n^{-1/2}$ and to a quantity $C_K(P)$ which measures the diversity of the underlying data distribution. Interestingly, this result highlights the inherent properties of certain datasets that make them more vulnerable to MIAs, regardless of the underlying learning procedure. We illustrate this behavior with numerical experiments in Section 6.

The objective of the paper is therefore to highlight the central quantity of interest $\Delta_{\nu,\lambda,n}(P,\mathcal{A})$ governing the success of MIAs and propose an analysis in different scenarios.

1.2 Related Works

Privacy Attacks. The majority of cutting-edge attacks follow a consistent approach within a framework known as Black-Box. In this framework, where access to the data distribution is available, attacks assess the performance of a model by comparing it to a group of “shadow models”. These shadow models are trained with the same architecture but on an artificially and independently generated dataset from the same data distribution. Notably, loss evaluated on training samples are expected to be much lower than when evaluated on “test points”. Therefore, a significant disparity between these losses indicates that the sample in question was encountered during the training, effectively identifying it as a member. This is intuitively related to some sort of “stability” of the algorithm on training samples (Bousquet and Elisseeff, 2002). Interestingly, we explicitly identify the exact quantity controlling the accuracy of effective MIAs which may be interpreted as a measure of stability of the underlying algorithm. In fact, as highlighted by Rezaei and Liu (2021), it is important to note that MIAs are not universally effective and their success depends on various factors. These factors include the characteristics of the data distribution, the architecture of the model, particularly its size, the size of the training dataset, and others, as discussed recently by Shokri et al. (2017); Carlini et al. (2022a). Subsequently, there has been a growing body of research delving into Membership Inference Attacks (MIAs) on a wide array of machine learning models, encompassing regression models (Gupta et al., 2021), generation models (Hayes et al., 2018), and embedding models (Song and Raghunathan, 2020). A comprehensive overview of the existing body of work on various MIAs has been systematically compiled in a thorough survey conducted by Hu et al. (2022). While studies of MIAs through DP already reveal precise bounds, it is worth noting that these induce a significant loss of performance on the learning task. It is worth to emphasize that our work does not focus on differential privacy. We further discuss the relation between our quantity $\Delta_{\nu,\lambda,n}(P,\mathcal{A})$ and DP mechanisms in Section H. Interestingly, the findings of the Section 5 reveal a threshold on the minimum number of training samples to overcome the need of introducing DP mechanisms.

Overfitting Effects. The pioneering work by Shokri et al. (2017) has effectively elucidated the relationship between overfitting and the privacy risks inherent in many widely-used machine learning algorithms. These empirical studies clearly point out that overfitting can often provide attackers with the means to carry out membership inference attacks. This connection is extensively elaborated upon by Salem et al. (2018); Yeom et al. (2018), and later by He et al. (2022), among other researchers. Overfitting tends to occur when the underlying model has a complex architecture or when there is limited training data available, as explained in Baluta et al. (2022). Recent works (Yeom et al., 2018; Del Grosso et al., 2023) investigated the theoretical aspects of the overfitting effect on the performances of MIAs, showing that the MIA performances can be lower bounded by a function of the *generalization gap* under some assumptions on the loss function. In our paper, we explicitly emphasize these insights by quantifying the dependence of $\Delta_{\nu,\lambda,n}(P,\mathcal{A})$ either on the dataset size and underlying structural parameters, or explicitly on the overfitting probability of the learning model.

Memorization Effects. Machine learning models trained on private datasets may inadvertently reveal sensitive data due to the nature of the training process. This potential disclosure of sensitive information occurs as a result of various factors inherent to the training procedure, which include the extraction of patterns, associations, and subtle correlations from the data (Song et al., 2017a; Zhang et al., 2021). While the primary objective is to generalize from data and make predictions, there is a risk that these models may also pick up on, and inadvertently expose, confidential or private information contained within the training data. This phenomenon is particularly concerning as it can lead to privacy breaches, compromising the confidentiality and security of personal or sensitive data (Hartley and Tsaftaris, 2022; Carlini et al., 2022b, 2019; Leino and Fredrikson, 2020; Thomas et al., 2020). Recent empirical studies have shed light on the fact that, in these scenarios, it is relatively rare for the average data point to be revealed by learning models (Tirumala et al., 2022; Murakonda and Shokri, 2007; Song et al., 2017b). What these studies have consistently shown is that it is the outlier samples that are more likely to undergo memorization by the model (Feldman, 2020), leading to potential data leakage. This pattern can be attributed to the nature of learning algorithms, which strive to generalize from the data and make predictions based on common patterns and trends. Average or typical data points tend to conform to these patterns and are thus less likely to stand out. On the other hand, outlier samples, by their very definition, deviate significantly from the norm and may capture the attention of the model. So when an outlier sample is memorized, it means the model has learned it exceptionally well, potentially retaining the unique characteristics of that data point. As a consequence, when exposed to similar data points during inference, the model may inadvertently leak information it learned from the outliers, compromising the privacy and security of the underlying data. An increasing body of research is dedicated to the understanding of memorization effects in language models (Carlini et al., 2023a). In the context of our research, it is important to highlight that our primary focus is on understanding the accuracy of MIAs but not its relationship with memorization. Indeed, this connection remains an area of ongoing exploration and inquiry in our work.

2. Background and Problem Setup

In this paper, we focus on MIAs, the ability of recovering membership to a training dataset $\mathbf{z} := (z_1, \cdots, z_n) \in \mathcal{Z}^n$ of a test point $\tilde{z} \in \mathcal{Z}$ from a predictor $\hat{\mu} = \mu_{\hat{\theta}_n}$ in a model $\mathcal{F} := \{\mu_\theta : \theta \in \Theta\}$, where Θ is the space of parameters. The predictor is identified to its parameters $\hat{\theta}_n \in \Theta$ learned from $\mathbf{z}$ through a **learning procedure** $\mathcal{A} : \bigcup_{k>0} \mathcal{Z}^k \to \mathcal{P}' \subseteq \mathcal{P}(\Theta)$, that is $\hat{\theta}_n$ follows the distribution $\mathcal{A}(\mathbf{z})$ conditionally to $\mathbf{z}$, which we assume we have access to. Here, $\mathcal{P}(\Theta)$ is the set of all distributions on Θ, and $\mathcal{P}'$ is the range of $\mathcal{A}$.

This means that there exists a function g and a random variable ξ independent of $\mathbf{z}$ such that $\hat{\theta}_n = g(\mathbf{z}, \xi)$. When $\mathcal{A}$ takes values in the set of Dirac distributions, that is $\hat{\theta}_n$ is a deterministic function of the data, we shall identify the parameters directly to the output of the learning procedure $\hat{\theta}_n := \mathcal{A}(z_1, \cdots, z_n)$.

Throughout the paper, we will further assume that $\mathcal{A}$ can be expressed as a function of the empirical distribution of the training dataset. Letting $\mathcal{M}$ be the set of all discrete distributions on $\mathcal{Z}$, and $\hat{P}_n$ be the empirical distribution of the training dataset, it means that there exists a (randomized) function $G : \mathcal{M} \to \mathcal{P}'$ such that we have $\mathcal{A}(z_1, \cdots, z_n) = G(\hat{P}_n)$ (almost surely).

Interestingly, if a learning procedure minimizes an empirical cost, then it satisfies this as-

sumption. In particular, maximum likelihood based learning procedures or Bayesian methods from Sablayrolles et al. (2019) are special cases. Any instance of a learning procedure in what follows will satisfy these assumptions. We further discuss this assumption in Appendix A.
Within this framework, we consider MIAs as functions of the parameters and the test point whose outputs are 0 or 1.

Definition 1 (Membership Inference Attack - MIA) *Any measurable map $\phi : \Theta \times \mathcal{Z} \to \{0,1\}$ is called a* ***Membership Inference Attack****.*

Hereinafter, we assume that MIAs might access more information, including randomization, the learning procedure $\mathcal{A}$ and/or the distribution of the data P. This framework is usually referred to as white-box (Hu et al., 2022).
We encode membership to the training data set as 1. We assume that $z_1, \ldots, z_n$ are independent and identically distributed ($i.i.d.$) random variables with distribution P. Following Del Grosso et al. (2023) and Sablayrolles et al. (2019) frameworks, among others, we suppose that the test point $\tilde{z}$ is to be drawn from P independently from the samples $z_1, \ldots, z_n$ with probability $\nu \in (0,1)$. Otherwise, conditionally to $\mathbf{z}$, we set $\tilde{z}$ to any z_j each with uniform probability $1/n$.
Letting U be a random variable with distribution $\hat{P}_n := \frac{1}{n}\sum_{j=1}^{n} \delta_{z_j}$ conditionally to $\mathbf{z}$, z_0 to be drawn independently from P and T be a random variable having Bernoulli distribution with parameter ν and independent of any other random variables, we can state

$$\tilde{z} := Tz_0 + (1-T)U.$$

The probability ν represent the theoretical **fraction of non-member** samples when evaluating MIAs, which is usually unknown in practice. In particular, it reflects how the data are tested.
A sensible choice to evaluate the performance of an MIA would be to consider its probability of successfully guessing the membership, i.e.

$$\begin{aligned}\mathbb{P}(\phi(\hat{\theta}_n, \tilde{z}) = 1 - T) &= \mathbb{P}(\phi(\hat{\theta}_n, z_0) = 0, T = 1) + \mathbb{P}(\phi(\hat{\theta}_n, U) = 1, T = 0) \\ &= \mathbb{E}\left[\mathtt{TNR} + \mathtt{TPR}\right],\end{aligned}$$

where TPR (resp. TNR) is the True Positive Rate (resp. True Negative Rate) of the MIA ϕ. However, in an MIA setting, the TPR is arguably more important than the TNR. Therefore, we define the **importance of the TPR** as a real number $\lambda > 0$, and we measure the performance of an MIA by its weighted probability of successfully guessing the membership of the test point.

Definition 2 (Accuracy of an MIA) *The* ***accuracy of an MIA*** *ϕ is defined as*

$$Acc_{\nu,\lambda,n}(\phi; P, \mathcal{A}) := \mathbb{P}\left(\phi(\hat{\theta}_n, z_0) = 0, T = 1\right) + \lambda \mathbb{P}\left(\phi(\hat{\theta}_n, U) = 1, T = 0\right), \tag{1}$$

where the probability is taken over all randomness.

We have the following remarks:

- The accuracy of an MIA scales from 0 to $\nu + \lambda(1-\nu)$. Constant MIAs $\phi_0 \equiv 0$ and $\phi_1 \equiv 1$ have respectively an accuracy equal to ν and $\lambda(1-\nu)$, which means that we can always build an MIA with accuracy of at least $\max(\nu, \lambda(1-\nu))$ and any MIA performing worse than this quantity is irrelevant to use. Particularly, this means that we have
$$\max(\nu, \lambda(1-\nu)) \leq \sup_{\phi} \mathrm{Acc}_{\nu,\lambda,n}(\phi; P, \mathcal{A}) \leq \nu + \lambda(1-\nu).$$
Moreover, the probability in the definition of the accuracy is taken over the randomness of the learning procedure and the data. This means that $\mathrm{Acc}_{\nu,\lambda,n}(\phi; P, \mathcal{A})$ measures the accuracy of an MIA over the learning procedure $\mathcal{A}$ and the task P rather than over the trained model $\mu_{\hat{\theta}_n}$.

- Isolated attacks are insufficient to fully certify the privacy level of ML models. Instead, controlling the optimal achievable accuracy provides a way to audit the model's privacy. Specifically, when $\sup_\phi \mathrm{Acc}_{\nu,\lambda,n}(\phi; P, \mathcal{A})$ is low, it indicates that the model is secure. Conversely, if this value is high, it indicates a potential vulnerability to successful attacks, even if specific MIAs have not been particularly successful.

- When considering the balanced case $\lambda = 1$, one shall observe that the accuracy is simply the probability of successfully guessing the membership, i.e. $\mathbb{P}\left(\phi(\hat{\theta}_n, \tilde{z}) = 1 - T\right)$.

We now define the **Membership Inference Security** of a learning procedure as a quantity summarizing the amount of security of the system against MIAs.

Definition 3 (Membership Inference Security - MIS) *The Membership Inference Security of a learning procedure $\mathcal{A}$ is*

$$Sec_{\nu,\lambda,n}(P, \mathcal{A}) \coloneqq \frac{1}{\min(\nu, \lambda(1-\nu))} \left(\nu + \lambda(1-\nu) - \sup_{\phi} Acc_{\nu,\lambda,n}(\phi; P, \mathcal{A}) \right), \tag{2}$$

where the supremum is taken over all MIAs.

From the first point preceding Definition 3, we see that the MIS has been defined to ensure that it scales from 0 (the best MIA approaches perfect guess of membership) to 1 (MIAs can not do better than ϕ_0 and ϕ_1).

Remark 4 (Model-specific attack - limitations of this approach) *In the framework we introduced, the MIS represents the security of the learning procedure $\mathcal{A}$. One could understandably want to consider the security of a trained model, by changing the definition and considering the quantity $\sup_\phi d\left(\phi; z_1, \cdots, z_n\right)$ conditionally to the training dataset $\{z_1, \cdots, z_n\}$, for some metric d, where d scales from 0 to 1. Natural choices of d include $d(\phi; z_1, \cdots, z_n) = \mathbb{P}(\phi(\hat{\theta}_n, \tilde{z}) = 1 - T \mid z_1, \cdots, z_n)$ or $d(\phi; z_1, \cdots, z_n) = TPR(\phi; z_1, \cdots, z_n)$. Unfortunately, this problem is degenerate as for those two choices of metric d, it can be easily shown that,*

$$\sup_{\phi} d\left(\phi; z_1, \cdots, z_n\right) = 1,$$

for any learning procedure and any dataset. In other words, the underlying problem of studying the best performing MIA conditionally to the data has no relevant insight.
In this article, to avoid this degeneracy, we undertake the path of shifting the focus onto the learning procedure $\mathcal{A}$ and the task P.

3. Performance Assessment of Membership Inference Attacks

In this section, we highlight the **Central** Statistical Quantity for the assessment of the accuracy of membership inference attacks, and show some basic properties on it. For any $\alpha > 0$, define the function $\tilde{D}_\alpha$ as

$$\begin{aligned} \tilde{D}_\alpha(P, Q) &:= \frac{1}{\alpha} \sup_B \ [\alpha P(B) - Q(B)] \\ &= \frac{1}{\alpha} \sup_B \ [Q(B) - \alpha P(B)] + \left(1 - \frac{1}{\alpha}\right), \end{aligned} \tag{3}$$

and the function D_α as

$$D_\alpha(P, Q) := \max(1, \alpha) \left[\tilde{D}_\alpha(P, Q) - \left(1 - \frac{1}{\alpha}\right)_+ \right], \tag{4}$$

for any distributions P and Q, where the supremum is taken over all measurable sets. Defining $\gamma := \frac{\nu}{\lambda(1-\nu)}$, we will then show that the central statistical quantity $\Delta_{\nu,\lambda,n}(P, \mathcal{A})$ is defined as

$$\Delta_{\nu,\lambda,n}(P, \mathcal{A}) = D_\gamma \left(\mathbb{P}_{(\hat{\theta}_n, z_0)}, \mathbb{P}_{(\hat{\theta}_n, z_1)} \right), \tag{5}$$

which depends on ν, λ, P, n and $\mathcal{A}$. Here, for any random variable x, $\mathbb{P}_x$ denotes its distribution, and for any real number $a \in \mathbb{R}$, $a_+ = \max(0, a)$. The quantity D_α has the remarkable property of being an $f-$divergence (Rényi, 1961; Csiszár et al., 2004), which we formalize in the following proposition.

Proposition 5 *The map $(P, Q) \mapsto D_\alpha(P, Q)$ is an $f-$divergence between P and Q with as generator the function $f_\alpha(x) = \frac{1}{2} \max(1, \alpha) [|x - 1/\alpha| - |1 - 1/\alpha|]$. Additionally, it holds that*

$$0 \leq D_\alpha(P, Q) \leq 1. \tag{6}$$

If x_1 and x_2 are random variables with joint distribution $\mathbb{P}_{(x_1, x_2)}$, then for any function $f \in \{D_\alpha, \tilde{D}_\alpha\}$ it holds that

$$f(\mathbb{P}_{x_1} \otimes \mathbb{P}_{x_2}, \mathbb{P}_{(x_1, x_2)}) = \mathbb{E}_{x_2} \left[f \left(\mathbb{P}_{x_1}, \mathbb{P}_{x_1 | x_2} \right) \right], \tag{7}$$

where $\mathbb{P}_{x_1|x_2}$ is the distribution of x_1 conditionally to x_2.

We have the following comments:

- The quantity $\Delta_{\nu,\lambda,n}(P, \mathcal{A})$ can be interpreted as quantifying some stability of the learning procedure. Here, z_1 represents an arbitrary random sample from the training set, while z_0 denotes a random sample that has not been seen during training. Thus, our quantity $\Delta_{\nu,\lambda,n}(P, \mathcal{A})$ captures the sensitivity of the learning procedure to individual samples by quantifying the distance between the joint distribution $\mathbb{P}_{(\hat{\theta}_n, z_1)}$ and the product distribution $\mathbb{P}_{(\hat{\theta}_n, z_0)} = \mathbb{P}_{\hat{\theta}_n} \otimes \mathbb{P}_{z_0}$. When individual samples have little influence on the output model, replacing a single sample in the training dataset causes only a minor shift in the model's parameters. In such cases, the output parameters exhibit minimal dependence on any individual sample, leading to a low value of $\Delta_{\nu,\lambda,n}(P, \mathcal{A})$. Conversely, if it is highly sensitive to each sample, even small changes in the dataset result in significant alterations to the parameters of the model.

- It is worth noting that in the ML literature, standard measures typically rely on the joint distribution between the ML model and the entire dataset (such as privacy measures, generalization bounds, etc.). In contrast, our novel approach diverges from this by focusing solely on the joint distribution between the ML model and a single training sample. Also, our metric does not measure the privacy of a trained model, it rather measures the MIA-wise privacy of a learning procedure.

- The choice of z_1 is arbitrary and is only for simplicity purpose.

We now state the main theorem displaying the relation between $\mathrm{Sec}_{\nu,\lambda,n}(P, \mathcal{A})$ and $\Delta_{\nu,\lambda,n}(P, \mathcal{A})$.

Theorem 6 (Key bound on accuracy) *Suppose P is any distribution and $\mathcal{A}$ is any learning procedure. Then the accuracy of any MIA ϕ satisfies:*

$$Acc_{\nu,\lambda,n}(\phi; P, \mathcal{A}) - \max(\nu, \lambda(1-\nu)) \leq \min(1, 1/\gamma)\nu\Delta_{\nu,\lambda,n}(P, \mathcal{A}).$$

In particular, we have

$$Sec_n(P, \mathcal{A}) = 1 - \Delta_{\nu,\lambda,n}(P, \mathcal{A}).$$

Recall that $\max(\nu, \lambda(1-\nu))$ is the maximum accuracy between the constant MIAs ϕ_0 and ϕ_1. The first point of Theorem 6 states that if $\Delta_{\nu,\lambda,n}(P, \mathcal{A})$ is low, then no MIA can perform substantially better than the constant MIAs. The second point of Theorem 6 shows that $\Delta_{\nu,\lambda,n}(P, \mathcal{A})$ is the quantity that controls the best possible accuracy of MIAs.
We see that $\Delta_{\nu,\lambda,n}(P, \mathcal{A})$ appears to be the key mathematical quantity for assessing the accuracy of MIAs. Furthermore, it is worth to emphasize that there is no assumption on the data distribution P. For instance, we can take into account outliers by making P a mixture.

Remark 7 (Relation with other divergences) *Proposition 5 shows that D_α is a divergence. In particular this means that it satisfies a Data Processing Inequality and is invariant by translation and rescaling. When $\alpha = 1$, D_α coincides with the total variation distance $\|\cdot\|_{TV}$ and the inequality $D_\alpha(P, Q) \leq \max(1, \alpha)\|P - Q\|_{TV}$ holds for any $\alpha > 0$. In any case, the upper bound over D_α is reached when the supports of the distributions are disjoint. The lower bound is reached when $P = Q$.*
Using Pinsker's inequality, it is easy to show that we have the relation:

$$Sec_{\nu,\lambda,n}(P, \mathcal{A}) \geq 1 - \max(1, \gamma)\sqrt{I(\hat{\theta}_n; z_1)/2}, \tag{8}$$

where $I(\hat{\theta}_n; z_1)$ is the mutual information between the parameters $\hat{\theta}_n$ and one random sample z_1. This mutual information can be interpreted as a measure of how much the parameters $\hat{\theta}_n$ memorize the information about z_1. Nevertheless, Pinsker's inequality typically yields a loose bound, indicating that relying on mutual information might be overly conservative in many cases.

Remark 8 (Differential Privacy) *Interestingly, if an $(\varepsilon, \delta)-$differentially private (Dwork et al., 2014) mechanism $\mathbb{M}$ is used to secure the learning procedure $\mathcal{A}$ by composition $\mathbb{M} \circ \mathcal{A}$, then a bound heuristically similar to the bound obtained for the KL divergence by Dwork et al. (2010) can be stated. See also Duchi and Ruan (2024) for other metrics. Specifically, the following relation (proved in equation 53 in Section H.2) holds*

$$Sec_{\nu,\lambda,n}(P, \mathbb{M} \circ \mathcal{A}) \geq 1 - \max(1, \gamma) \left[(e^{\varepsilon} - 1/\gamma)_+ - (1 - 1/\gamma)_+ + \delta \right]. \tag{9}$$

Though more refined bounds can be obtained, this relation shows that in some scenarios, DP mechanisms might provide conservative but loose lower bounds on the security level. However, differential privacy is a tool to induce privacy into a model whereas the MIS a tool to measure the MIA-wise privacy of a learning procedure. Specifically, the two frameworks are not equivalent and do not convey the same message. We further discuss it in Section H.

In Section 4, we analyze how $\Delta_{\nu,\lambda,n}(P, \mathcal{A})$ is controlled when the learning procedure exhibits overfitting. In Section 5, we address scenarios where we can provide precise control on $\Delta_{\nu,\lambda,n}(P, \mathcal{A})$. Numerical experiments are presented in Section 6, and the proof of Theorem 6 is detailed in Appendix E.

4. Overfitting Causes Lack of Security

In this section, we assume that $\mathcal{Z} \coloneqq \mathcal{X} \times \mathcal{Y}$ and that the learning procedure $\mathcal{A}$ produces overfitting parameters $\hat{\theta}_n$. We then note $z \coloneqq (x, y)$. We consider learning systems minimizing $L_n : \theta \mapsto \frac{1}{n} \sum_{j=1}^{n} l_\theta(x_j, y_j)$ for some training dataset $(z_1, \cdots, z_n)$ where $l_\theta : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$ is a loss function. We defer all proofs of this section to Appendix F.

Definition 9 ($(\varepsilon, 1-\alpha)$-Overfitting) *We say that the learning procedure $\mathcal{A}$ is $(\varepsilon, 1-\alpha)$-overfitting for some $\varepsilon \in \mathbb{R}^+$ and $\alpha \in (0, 1)$ when*

$$\mathbb{P}\left(l_{\hat{\theta}_n}(x_1, y_1) \leq \varepsilon \right) \geq 1 - \alpha, \tag{10}$$

where the probability is taken over the data and the randomness of $\mathcal{A}$.

When $\alpha = 0$, equation 10 is equivalent to having $l_{\hat{\theta}_n}(x_j, y_j) \leq \varepsilon$ almost surely for all $j = 1, \ldots, n$. Furthermore, in many learning procedures, we give an additional stopping criteria taking the form $L_n \leq \eta$ for some $\eta \in \mathbb{R}^+$. Letting $\mathcal{A}_\eta$ such a learning procedure, we give a sufficient condition for equation 10 to hold:

Proposition 10 *For some fixed $\varepsilon \in \mathbb{R}^+$ and $\alpha \in (0, 1)$, let $\eta \coloneqq \varepsilon\alpha$ and suppose that $\mathcal{A}_\eta$ stops as soon as $L_n(\hat{\theta}_n) \leq \eta$. Then $\mathcal{A}_\eta$ is $(\varepsilon, 1-\alpha)$-overfitting.*

We will need an additional hypothesis for the following theorem.
Hypothesis (H1) : $\mathcal{Y} \coloneqq \mathbb{R}^s$ for some $s \geq 1$, and for all $\theta \in \Theta, x \in \mathcal{X}$ and $y \in \mathcal{Y}$, we have

$$l_\theta(x, y) = \omega(y, \Psi_\theta(x)), \tag{11}$$

for some family of functions $\Psi_\theta : \mathcal{X} \to \mathbb{R}$ and some continuous function $\omega : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$.

Theorem 11 (Overfitting induces lack of security) *Assume $\mathcal{A}$ is $(\varepsilon, 1-\alpha)$-overfitting for some fixed (ε, α). Then we have*

$$Sec_{\nu,\lambda,n}(P, \mathcal{A}) \leq \max(1, 1/\gamma) \left(\alpha + \gamma \mathbb{P}\left(l_{\hat{\theta}_n}(x, y) \leq \varepsilon\right)\right). \tag{12}$$

*Assume furthermore that **H1** holds. Assume that for all $\eta > 0$, $\mathcal{A}_\eta$ stops as soon as $L_n \leq \eta$ and that a version of the conditional distribution of y given x is absolutely continuous with respect to the Lebesgue measure, then*

$$\lim_{\eta \to 0^+} Sec_{\nu,\lambda,n}(P, \mathcal{A}_\eta) = 0. \tag{13}$$

The second point of Theorem 11 states that for regressors with reasonably low training loss on the dataset, a loss-based MIA $\phi_\varepsilon : (\theta, z) \mapsto 1_{l_\theta(z) \leq \varepsilon}$ would reach high success probability. This theoretically confirms the already well-known insight that overfitting implies poor security.
Hypothesis **H1** occurs when $\Psi_\theta(x)$ models the conditional expectation of y given x, in a setting where the loss function is defined as a distance between $\Psi_\theta(x)$ and y.

Remark 12 *Interestingly, if we only assume Definition 9 to hold without Proposition 10 to hold, then a much weaker version of the second point of Theorem 11 still holds. Indeed, for a fixed $\alpha \in (0, 1)$, given a sequence of learning procedures $(\mathcal{A}^\varepsilon)_{\varepsilon \in \mathbb{R}^+}$ that are $(\varepsilon, 1-\alpha)$-overfitting for all $\varepsilon > 0$, we have that $\lim_{\varepsilon \to 0} Sec_{\nu,\lambda,n}(P, \mathcal{A}^\varepsilon) \leq \max(1, 1/\gamma)\alpha$.*

Example 1 (non-linear regression Neural Network) *We consider here a (non-linear) regression setting, that is for all $j = 1, \ldots, n$, we have $y_j \coloneqq \Psi^*(x_j) + \zeta_j$, where ζ_j is some independent random noise and the function $\Psi^* : \mathcal{X} \to \mathbb{R}$ is arbitrary, fixed and unknown. We aim at estimating Ψ^* by some Neural Network $\Psi_\theta \in \mathcal{F}$, where $\mathcal{F}$ is some fixed model. For instance $\mathcal{F}$ can be the set of all 2-layers ReLU neural networks with fixed hidden layer width. The learning procedure $\mathcal{A}$ then learns by minimizing the MSE loss $L_n \coloneqq \frac{1}{n}\sum_{j=1}^n (y_j - \Psi_\theta(x_j))^2$. In this case, equation 11 holds. Under the further assumption that there is an arbitrarily close approximation Ψ_θ of Ψ^* in $\mathcal{F}$, one can construct the sequence of learning procedures $(\mathcal{A}_\eta)_{\eta \in \mathbb{R}^+}$ such that the hypotheses of the second point of Theorem 11 for equation 13 to hold. Refer to Section 6 for a numerical illustration.*

Example 2 (Linear regression) *We assume here a linear regression setting, that is $\mathcal{X} \coloneqq \mathbb{R}^d$ for some $d \in \mathbb{N}$, and $y_j \coloneqq \beta^T x_j + \zeta_j$, where ζ_j is some independent random noise and $\beta \in \mathbb{R}^d$ is fixed and unknown. Further assuming that ζ_j is absolutely continuous with respect to the Lebesgue measure, and that $d > n$, both equation 10 (with $\varepsilon, \alpha = 0$) and equation 11 hold. Then, the assumptions of the second point of Theorem 11 are satisfied, leading to $Sec_{\nu,\lambda,n}(P, \mathcal{A}) = 0$.*

5. Security is Data Size Dependent

In this section, we study the converse, where we aim at understanding when to expect $\mathrm{Sec}_{\nu,\lambda,n}(P,\mathcal{A})$ to be close to 1. All the proofs of the section can be found in Appendix G.

5.1 Empirical Mean based Learning Procedures

We first study the case of learning procedures for which the parameters $\hat{\theta}_n$ can be expressed in the form of functions of empirical means (e.g., linear regression with mean-squared error, method of moments...). Specifically, for any (fixed) measurable maps $L : \mathcal{Z} \to \mathbb{R}^d$ and $F : \mathbb{R}^d \to \mathbb{R}^q$ for some $d, q \in \mathbb{N}$, we consider that

$$\hat{\theta}_n \coloneqq F\left(\frac{1}{n}\sum_{j=1}^{n} L(z_j)\right). \tag{14}$$

Equation 14 states that the parameters are the result of the learning procedure $\mathcal{A} : (z_1, \cdots, z_n) \mapsto \delta_{F\left(\frac{1}{n}\sum_{j=1}^{n} L(z_j)\right)}$, where δ_θ stands for the Dirac mass at θ. We then have the following result.

Theorem 13 *Suppose that the distribution of $L(z_1)$ has a non zero absolutely continuous part with respect to the Lebesgue measure, and a third finite moment. Then*

$$Sec_{\nu,\lambda,n}(P,\mathcal{A}) \geq 1 - \max(1,\gamma)\left(c_{L,P} + \frac{\sqrt{d}}{2\sqrt{n}}\right)n^{-1/2}, \tag{15}$$

for some constant $c_{L,P}$ depending only on L and P.

Remark 13 : Theorem 13 implies that a sufficient condition to ensure $\mathrm{Sec}_{\nu,\lambda,n}(P,\mathcal{A})$ to be made larger than $1-\max(1,\gamma)\varepsilon$, is to have $n \geq \Omega(\varepsilon^{-2})$. The hidden constant only depends on the distribution data P and the parameters dimension d. See Appendix G for a proof.
We now provide examples for which Theorem 13 allows us to give an lower bound on $\mathrm{Sec}_{\nu,\lambda,n}(P,\mathcal{A})$.

Example 3 (solving equations) *We seek to estimate an (unknown) parameter of interest $\theta_0 \in \Theta \subseteq \mathbb{R}^d$. We suppose that we are given two functions $h : \Theta \to \mathbb{R}^l$ and $\psi : \mathcal{Z} \to \mathbb{R}^l$ for some $l \in \mathbb{N}$, and that θ_0 is solution to the equation*

$$h(\theta_0) = \mathbb{E}[\psi(z)], \tag{16}$$

where z is a random variable of distribution P. Having access to data samples $z_1, \ldots, z_n$ drawn independently from the distribution P, we estimate $\mathbb{E}[\psi(z)]$ by $\frac{1}{n}\sum_{j=1}^{n}\psi(z_j)$. Assuming that h is invertible, one can set $\hat{\theta}_n = h^{-1}\left(\frac{1}{n}\sum_{j=1}^{n}\psi(z_j)\right)$, provided that $\frac{1}{n}\sum_{j=1}^{n}\psi(z_j) \in \mathcal{I}m(h)$. In particular, when $\mathcal{Z} = \mathbb{R}$, this method generalizes the method of moments by setting $\psi(z) = (z, z^2, \cdots, z^l)$. We then may apply Theorem 13 to any estimators obtained by solving equations.

Example 4 (Linear Regression) *We consider here the same framework as in Example 2, where $d < n$ (hence Definition 9 can not be fulfilled with $\alpha = 0$). Let $\mathbb{X}$ be the $d \times n$ matrix whose i^{th} row is x_i, and $\mathbb{Y}$ be the column vector $(y_1, \cdots, y_n)^T$. We then recall that the estimator $\hat{\beta}_n$ of β is given by*

$$\hat{\beta}_n \coloneqq (\mathbb{X}\mathbb{X}^T)^{-1}\mathbb{X}\mathbb{Y}^T.$$

Based on equation 14, if we set $F(K, b) \coloneqq K^{-1}b^T$ and $L((x, y)) \coloneqq ((x^i x^j)_{i,j=1}^d, (x^i y)_{i=1}^d)$, where x^i is the i^{th} coordinate of x, then the estimator can be expressed as follows $\hat{\beta}_n = F\left(\frac{1}{n}\sum_{j=1}^n L((x_j, y_j))\right)$.

Interestingly, we see from Examples 2 and 4 that the security of least squares linear regression estimator is constant 0 up to $n = d$ (where d is both the dimension of the data and the dimension of the parameters), and then is increasing up to 1 provided that $n \to \infty$.

5.2 Discrete Data Distribution

We now consider the common distribution of the points in the data set to be $P \coloneqq \sum_{k=1}^K p_k \delta_{u_k}$ for some fixed $K \in \mathbb{N} \cup \{\infty\}$, some fixed probability vector $(p_1, \cdots, p_K)$ and some fixed points $u_1, \ldots, u_K$ in $\mathcal{Z}$. Without loss of generality, we may assume that $p_k > 0$ for all $k \in \{1, \cdots, K\}$. In the discrete distribution setting, we show that the convergence of the MIS toward 1 can occur at different rates (see Appendix C.1), depending on both the algorithm and the underlying data distribution. Therefore, in what follows, we are interested in studying the worst security among all learning procedures, namely $\min_{\mathcal{A}} \mathrm{Sec}_{\nu,\lambda,n}(P, \mathcal{A})$.

Theorem 14 *For $k = 1, \ldots, K$, let B_k be a random variable having Binomial distribution with parameters (n, p_k). Then,*

$$\min_{\mathcal{A}} Sec_{\nu,\lambda,n}(P, \mathcal{A}) = 1 - \max(1, \gamma)\frac{1}{2\gamma}\sum_{k=1}^K \mathbb{E}\left[\left|\frac{B_k}{n} - \gamma p_k\right|\right] + \max(1, \gamma)\frac{1}{2}\left|1 - \frac{1}{\gamma}\right|, \tag{17}$$

where the minimum is taken over all learning procedures and is reached on learning procedures of the form $\mathcal{A}(z_1, \cdots, z_n) = \delta_{F(\frac{1}{n}\sum_{j=1}^n \delta_{z_j})}$ for some injective maps F.

Theorem 14 provides an exact formula to accurately bound $\mathrm{Sec}_{\nu,\lambda,n}(P, \mathcal{A})$ for any learning procedure $\mathcal{A}$, including those that exhibit the most significant leakage. It is shown in Theorem 14 that the minimum is reached for deterministic learning procedures $\mathcal{A}$. We show below that the r.h.s. of equation 17 is tightly related to the quantity

$$C_K(P) \coloneqq \sum_{k=1}^K \sqrt{p_k(1 - p_k)}. \tag{18}$$

This quantity can thus be exploited to compare leakage between different datasets.
It is worth noting that $C_K(P)$ is a diversity measure, giving a control on the homogeneity of the data distribution. We show in Appendix C that it is comparable both to the Gini-Simpson and the Shannon Entropy.
Unlike λ begin a meta-parameter, the theoretical value of ν is never known in practice. In the following corollary, we exhibit the security of the worst learning procedure privacy-wise in the worst theoretical setup possible.

Corollary 15 *Assume that $C_K(P) < \infty$, $n \geq 5$ and $n > 1/p_k$ for all $k = 1, \ldots, K$. Then there exists universal constants $c \geq 0.29$ and $c' \leq 0.44$ such that*

$$\min_{\nu,\mathcal{A}} Sec_{\nu,\lambda,n}(P, \mathcal{A}) = 1 - \varepsilon_n,$$

where ε_n satisfies

$$cC_K(P)n^{-1/2} \leq \varepsilon_n \leq c'C_K(P)n^{-1/2}.$$

If $C_K(P) < \infty$ but the condition on n does not hold, we still have $\min_{\nu,\mathcal{A}} Sec_{\nu,\lambda,n}(P, \mathcal{A}) \geq 1 - \frac{C_K(P)}{2}n^{-1/2}$.

Corollary 15 implies that no matter the theoretical value of ν, a sufficient condition to ensure security larger than $1 - \varepsilon$ is to have at least $n \geq (C_K(P)/2\varepsilon)^2$.
Importantly, when designing a learning procedure, one can never have access to the true theoretical value ν. Consequently, Corollary 15 provides a practical way of measuring the security, without requiring unavailable knowledge.
In any case, this result indicates that discretizing the data allows the control on the security of any learning procedure, no matter the specified value of λ. The condition $n \geq 1/p_k$ shall be understood as the expected number of samples to observe the atom u_k within a dataset. Indeed, if Ξ_k is the random variable representing the required number of samples to observe the atom u_k, then Ξ_k follows a Geometric law with parameter p_k, hence its expectation is $\mathbb{E}[\Xi_k] = 1/p_k$. We further discuss it in Section C. Section 6 illustrates the impact of $C_K(P)$ through numerical experiments. We further discuss this Corollary in Section C.

6. Numerical Experiments

In this section, we propose two numerical experiments to illustrate our results in Sections 4 and 5. All simulations have been conducted with the Pytorch library. We refer to Appendix D for more details on the experiments.

6.1 Overfitting

We run three experiments to illustrate to illustrate the results of Section 4.
Regression task. We run a non-linear regression experiment to illustrate the results of Section 4 and specifically Example 1. We then consider the setting of Example 1 with $\Psi^*(x) = \sin(\pi\beta^T x)$ for some fixed β. During the training of the neural network $\Psi_{\hat{\theta}_n}$, at each iteration we evaluate the fraction of training (validation) data that achieves a loss below ε. Validation data correspond to a set of data independent from the training dataset. Figure 1 illustrates Theorem 11 by showing that for very small values of threshold ($\varepsilon = 10^{-6}$), we still reach 100% training accuracy after 2500 iterations whereas the validation accuracy (for $\varepsilon = 10^{-6}$) stabilizes at near 0%.

To illustrate Example 2, we train multiple linear regressors until complete overfitting (training loss near 0) and evaluate it at the end of its training. We perform this evaluation for various input dimensions from 10 to 2000. For each dimension, we evaluate the linear

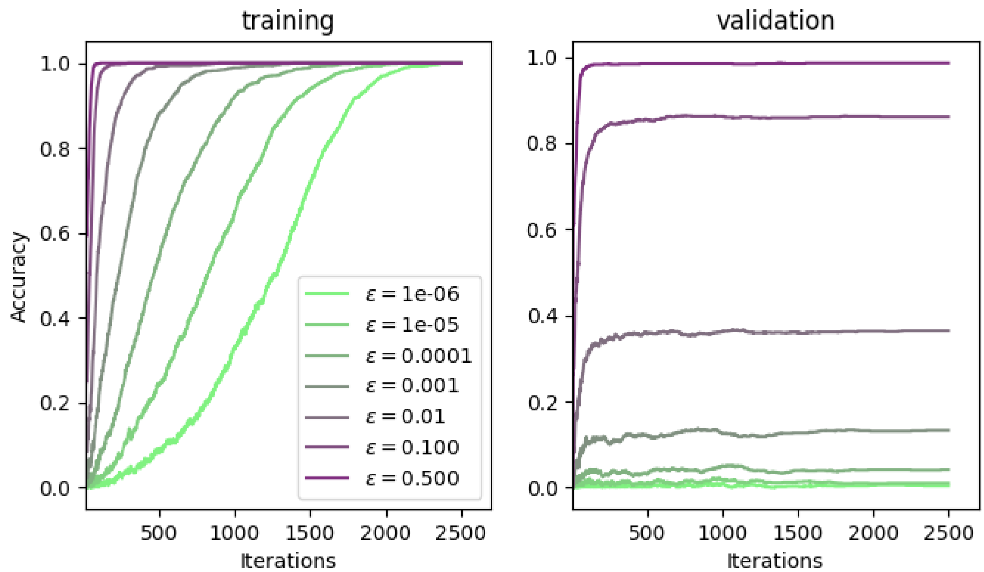
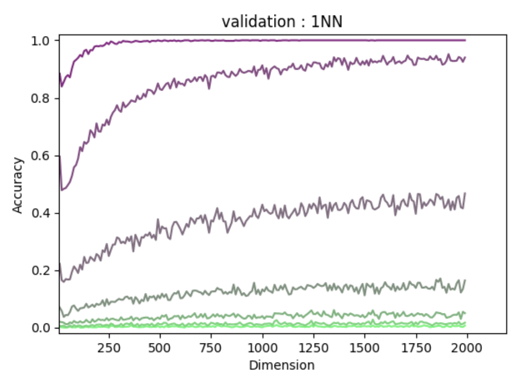


Figure 1: Shows the fraction of the Training/Validation dataset whose loss is under given thresholds during the training process. The left figure shows the **training accuracy**, and the right figure shows the **validation accuracy**.

(a) Linear regression.



(b) Nearest Neighbors (k=1)

Figure 2: Shows the fraction of the Validation dataset whose loss is under given thresholds at the end of the training process for different dimensions. The left figure shows the **validation accuracy** for the linear regression model , and the right figure shows the **validation accuracy** for the nearest neighbors model.

regressor and report the fraction of the validation dataset whose loss in under given thresholds. Figure 2a illustrates Theorem 11 by showing that for very small values of threshold ($\varepsilon = 10^{-6}$), the validation accuracy remains near 0% for any dimension.
Therefore, simple loss-based MIA with threshold ε would suffice to accurately predict membership most of the time both for the non-linear regression and the linear regression tasks.

The number of iterations being generally unknown to the MIA, the calibration of ε is a hard task to perform. Even though it seems that the loss-based attack with threshold $\varepsilon = 10^{-6}$ is a good candidate to achieve near perfect guess, it is worth noting that it would occur only if at least 2000 iterations (for the non-linear regression task) have been done during the training procedure.

Nearest Neighbors. We additionally illustrate overfitting through a $k-$Nearest Neighbors model, with $k = 1$. Similarly to the linear regression task, we perform the evaluation for various input dimensions from 10 to 2000. By construction of the Nearest Neighbors algorithm, the error main by the predictor on the training dataset is always 0. Figure 2b shows a similar behavior as for the regression tasks, meaning that simple loss-based MIAs would also suffice to to accurately predict membership.

6.2 Impact of $C_K(P)$ on accuracy

Corollary 15 indicates that discretizing the data distribution improves the security of the model. To illustrate the impact of a discretization through the constant $C_K(P)$, we trained several 3-layers neural networks to classify samples from the MNIST dataset (Deng, 2012). Before training, we fixed three discretizations[1] (clusterings). For each dataset (with varying size n), we trained a neural network on it, and three other neural networks on discretized versions of the original dataset (one for each clustering). We then numerically computed the quantity $C_K(P)$ for each discretization. Table 1 shows the accuracy of all neural networks,

n	raw dataset	$C_K(P) = 4.3$	$C_K(P) = 6.74$	$C_K(P) = 9.20$
1000	0.989 ± 0.0011	0.963 ± 0.0223 ($\Delta_{\nu,\lambda,n} \leq 0.07$)	0.968 ± 0.0137 ($\Delta_{\nu,\lambda,n} \leq 0.11$)	0.986 ± 0.0039 ($\Delta_{\nu,\lambda,n} \leq 0.15$)
5000	0.993 ± 0.0012	0.967 ± 0.0184 ($\Delta_{\nu,\lambda,n} \leq 0.03$)	0.971 ± 0.0282 ($\Delta_{\nu,\lambda,n} \leq 0.05$)	0.984 ± 0.0055 ($\Delta_{\nu,\lambda,n} \leq 0.07$)
10000	0.994 ± 0.0006	0.971 ± 0.0141 ($\Delta_{\nu,\lambda,n} \leq 0.02$)	0.977 ± 0.0082 ($\Delta_{\nu,\lambda,n} \leq 0.03$)	0.984 ± 0.0055 ($\Delta_{\nu,\lambda,n} \leq 0.05$)

Table 1: Shows the accuracy of classifiers on MNIST dataset. The column **n** displays the dataset size. The column **raw dataset** displays the accuracy of the neural network on the original dataset. Each column of the column **discretized datasets** displays the accuracy of a neural network on the discretized dataset associated to the constant $C_K(P)$.

and displays the impact of the discretization on the accuracy, depending on n and the value of $C_K(P)$. For a dataset of size $n = 1000$, our neural network reaches an accuracy of 0.989 when trained on the original dataset. When discretizing, Table 1 displays a loss of almost 2.5% of accuracy for the discretization having $C_K(P) = 4.30$, and a loss of 2% for the other discretizations. As discussed in Section C, increasing the number of clusters will increase the value of $C_K(P)$. Table 1 displays the intuition that smaller discretization (smaller $C_K(P)$) will lower simultaneously the accuracy and the quantity $\Delta_{\nu,\lambda,n}(P, \mathcal{A})$, which motivates the need to optimize the discretization to find a trade-off between security and accuracy.

7. Summary and Discussion

The findings presented in this article open gates to the theoretical understanding of MIAs, and partially confirm some of empirically observed facts. Specifically, we confirmed that

1. Many clustering algorithms exist, but we did not aim at optimizing the choice of the discretizations.

overfitting indeed induces the possibility of highly successful attacks. We further revealed a sufficient condition on the size of the training dataset to ensure control on the security of the learning procedure, when dealing with discrete data distributions or functionals of empirical means. We established that the rates of convergence consistently follow an order of $n^{-1/2}$. The constants established in the rates of convergence scale with the number of discrete data points and the dimension of the parameters in the case of functionals of empirical means. Interestingly, for discrete data, the diversity measure $C_K(P)$ highlights the use of data quantization to ensure privacy by design.

The security measure $\mathrm{Sec}_{\nu,\lambda,n}(P, \mathcal{A})$ can be estimated by approximating the optimal accuracy-maximizing attack with a binary classifier (e.g., a feed-forward neural network) and evaluating its classification accuracy. The optimal attack is theoretically known and is given by $\phi^*(\theta, z) = 1_{(\theta,z)\in B^*}$, where B^* denotes the set reaching the maximum in the definition. However, determining this set explicitly depends on the densities of the joint and product distributions between $\hat{\theta}_n$ and z_1. Direct evaluation of these densities is computationally demanding, highlighting the necessity of practical estimation approaches. We therefore believe that a thorough investigation into the estimation of $\mathrm{Sec}_{\nu,\lambda,n}(P, \mathcal{A})$ constitutes a substantial research effort that requires a dedicated full-length study.

In Section 5, our work is currently limited to the discrete data (e.g., tabular datasets) and empirical mean based learning procedures. We intend to extend further our research to the complete study of maximum likelihood estimation, empirical loss minimization and Stochastic Gradient Descent. Moreover, we studied the worst possible learning procedure by taking the supremum over all learning procedures. An interesting question would be to analyze the security of specific sets of learning procedures. We leave this to further works. Furthermore, we aim to explore the optimization of the trade-off discussed at the conclusion of Section 6.2.

Our findings about overfitting learning procedures do not cover classification learning procedures. We anticipate continuing our research in this direction to gain a comprehensive understanding of the impact of overfitting. Currently, we are able to extend these results to a minor scenario (see Appendix B Proposition 19). We anticipate that these assumptions may be relaxed in future investigations.

We have also assumed the data to be *i.i.d.*. Often, concentration inequalities and asymptotic theorems over *i.i.d.* samples have their counterparts in the dependent setting. Assuming some dependence in the data is therefore an interesting way to extend our results.

Acknowledgments

Elisabeth Gassiat is supported by Institut Universitaire de France, and by the Agence Nationale de la Recherche under projects ANR-21-CE23-0035-02 and ANR-23- CE40-0018-02.

We sincerely thank the reviewers for their valuable and constructive feedback.

Appendix A. More comments on Section 2

We further discuss here the assumption that the learning procedure shall be expressed as a function of the empirical distribution of the training dataset.
Usual definition of a learning procedure $\mathcal{A}$ asks for its domain to be $\bigcup_{n\geq 1}\mathcal{Z}^n$ which is similar to identifying it with a sequence of learning procedures $(\mathcal{A}_n : \mathcal{Z}^n \to \mathcal{P}')_{n\geq 1}$ where for each $n \geq 1$ we have that the restriction of $\mathcal{A}$ to $\mathcal{Z}^n$ coincides with $\mathcal{A}_n$. However, this definition allows not specifying its behavior through all values of n, and specifically having drastically different behaviors for different values of n. To rigorously study the characteristics of a learning procedure, it is natural to ask that its behavior is similar for all values of n, meaning that its behavior can be defined independently of n.
Our assumption solves this issue, as the function $G : \mathcal{M} \to \mathcal{P}'$ from Section 2 is defined for all discrete distribution on $\mathcal{Z}$. Furthermore, it is worth noting that this assumption holds for all learning procedures aiming at minimizing the empirical cost on its training dataset. For most learning procedures, it will still hold even when some weights are applied to the samples. Indeed, changing the distribution P by P' for some other distribution P', the training dataset size n by some other integer $n' \in \mathbb{N}$ and adequately adapt the learning procedure $\mathcal{A}$ into another learning procedure $\mathcal{A}'$ to take into account the changes, makes the study still valid as long as we consider $\Delta_{\nu,\lambda,n'}(P', \mathcal{A}')$ instead of $\Delta_{\nu,\lambda,n}(P, \mathcal{A})$. Therefore, it is sufficient to conduct the study under this hypothesis.
In particular, this assumption treats all points similarly, and is invariant with respect to the redundancy of the whole dataset. It is equivalent to saying that the learning procedure is **symmetric** and **redundancy invariant**, whose definitions are given below.

Definition 16 (Symmetric Map) *Given two sets $\mathcal{Z}_1$ and $\mathcal{Z}_2$ and an integer k, a map $f : \mathcal{Z}_1^k \to \mathcal{Z}_2$ is said to be **symmetric** if for any $(a_1, \cdots, a_k) \in \mathcal{Z}_1^k$ and any permutation σ on $\{1, \cdots, k\}$, we have*

$$f(a_1, \cdots, a_k) = f\left(a_{\sigma(1)}, \cdots, a_{\sigma(k)}\right).$$

Definition 17 (Redundancy Invariant Map) *Given two sets $\mathcal{Z}_1$ and $\mathcal{Z}_2$, a map $f : \bigcup_{k>0}\mathcal{Z}_1^k \to \mathcal{Z}_2$ is said to be **redundancy invariant** if for any integer m and any $\boldsymbol{a} = (a_1, \cdots, a_m) \in \mathcal{Z}_1^m$, we have*

$$f(\boldsymbol{a}) = f(\boldsymbol{a}, \cdots, \boldsymbol{a}).$$

We summarize the last claim in the following proposition.

Proposition 18 *Let $f : \bigcup_{k>0}\mathcal{Z}^k \to \mathcal{Z}'$ be a measurable map onto any space $\mathcal{Z}'$. Then the following statements are equivalent*

(i) f is redundancy invariant and for any $k \in \mathbb{N}$, the restriction of f to $\mathcal{Z}^k$ is symmetric.

(ii) There exists a function $G : \mathcal{M} \to \mathcal{Z}'$ such that for any $k \in \mathbb{N}$, for any $(z_1, \cdots, z_k) \in \mathcal{Z}^k$ we have $f(z_1, \cdots, z_k) = G\left(\frac{1}{k}\sum_{j=1}^k \delta_{z_j}\right)$.

Proof [Proof of Proposition 18] We only prove that (i) implies (ii). The fact that (ii) implies (i) is straightforward.
Let $f : \bigcup_{k>0} \mathcal{Z}^k \to \mathcal{Z}'$ be a measurable map satisfying condition (i). Let $\mathcal{M}^{\text{emp}}$ be the set of all possible empirical distributions, that is the subset of $\mathcal{M}$ containing all $\frac{1}{k}\sum_{j=1}^{k} \delta_{z_j}$ for all integer k and all $(z_1, \cdots, z_k) \in \mathcal{Z}^k$. We shall define G on $\mathcal{M}^{\text{emp}}$ such that (ii) holds true.
For any $Q \in \mathcal{M}^{\text{emp}}$, let $\{z_1, \cdots, z_m\}$ be its support and $q_1, \ldots, q_m \in (0, 1)$ be such that $Q = \sum_{j=1}^{m} q_j \delta_{z_j}$. Since Q is an empirical distribution, there exists positive integers $k_1, \ldots, k_m$ (for each j, k_j is the number of occurences of z_j in the sample from which Q is the empirical distribution) such that $q_j = \frac{k_j}{K}$, with $K = \sum_{j=1}^{m} k_j$.
Let $r = gcd(k_1, \ldots, k_m)$ be the greatest common divisor of the k_j's and define $k'_j = k_j/r$ for $j = 1, \ldots, m$. Then with $K' \coloneqq \sum_{j=1}^{m} k'_j$, we have $q_j = \frac{k'_j}{K'}$.
Now, for any other sequence of positive integers $\ell_1, \ldots, \ell_m$ such that $q_j = \frac{\ell_j}{L}$, with $L = \sum_{j=1}^{m} \ell_j$, we get for all j, $\ell_j = sk'_j$ with $s = gcd(\ell_1, \ldots, \ell_m)$. Thus we may define $G(Q) = f(\boldsymbol{z})$ where $\boldsymbol{z}$ is the dataset consisting of all z_j's with k'_j repetitions.
We now prove that such a G satisfies (ii). Indeed, for any integer k and any $Z \coloneqq (z'_1, \cdots, z'_k) \in \mathcal{Z}^k$, define $V \coloneqq ((\ell_1, z_1), \cdots, (\ell_m, z_m))$ where $(z_1, \cdots, z_m)$ are the distinct elements of Z and $(\ell_1, \cdots, \ell_m)$ are their occurrences. Define r as their greatest common divisor, and $(k_1, \ldots, k_m) = (\ell_1, \cdots, \ell_m)/r$. By using the fact that f is symmetric and redundancy invariant, we get that $f(Z) = f(Z_0) = G(Q)$ where Z_0 is the dataset consisting of all z_j's with k_j repetitions and $Q = \sum_{j=1}^{m} \frac{k_j}{K} \delta_{z_j} = \frac{1}{n}\sum_{j=1}^{n} \delta_{z'_j}$. Thus (ii) holds. ■

Appendix B. More comments on Overfitting

In this section, we give an extension of Theorem 11 to the setting of classification.

The second point of Theorem 11 requires the absolute continuity of the distribution of the label with respect to the Lebesgue measure, which makes it not straightforward to extend it to classifiers.
We discuss here one very specific framework in which we have been able to extend our results to the classification setting. The framework and the assumptions are all inspired from Vardi et al. (2022).

We assume that the data space is restrained to the binary classification setting with data in the sphere of radius $\sqrt{s}$, i.e. $\mathcal{Z} \coloneqq \left(\sqrt{s}\mathbb{S}^{s-1}\right) \times \{-1, 1\}$ where $\mathbb{S}^{s-1}$ is the unit sphere in $\mathbb{R}^s$. We assume our data $(z_1, \cdots, z_n) \coloneqq ((x_1, y_1), \cdots, (x_n, y_n))$ to be independently drawn on $\mathcal{Z}$ from a distribution P. We assume that the conditional law of x_1 given y_1 is absolutely continuous with respect to the Lebesgue measure on $\sqrt{s}\mathbb{S}^{s-1}$. We denote by $\mathcal{H}$ the latter hypothesis. Let $\Psi_\theta(x) = \sum_{j=1}^{l} v_j \sigma(w_j^T x + b_j)$ be a 2−ReLU network with parameters θ, i.e. $\theta = (v_j, w_j, b_j)_{j=1}^{l}$ with $l \in \mathbb{N}$ the width of the network and $\sigma(u) = \max(u, 0)$. We aim at learning a classifier $\Psi_{\hat{\theta}_n}$ on the data by minimizing

$$\mathcal{L} : \theta \mapsto \sum_{j=1}^{n} l(y_j \Psi_\theta(x_j)), \tag{19}$$

where $l : \mathbb{R} \to \mathbb{R}^+$ is either the exponential loss or the logistic loss. To reach the objective, we apply Gradient Flow on the objective equation 19, producing a trajectory $\theta_n(t)$ at time t. From Vardi et al. (2022) Theorem 3.1, there exists a 2−ReLU network classifying perfectly the training dataset, as long as $\max_{i \neq j} \left\{ |x_i^T x_j| \right\} < d$, which holds almost surely by $\mathcal{H}$. Let the initial point $\theta_n(0)$ be the parameters of this network.
Then by Vardi et al. (2022) Theorem 2.1, paraphrasing Lyu and Li (2019); Ji and Telgarsky (2020), $\frac{\theta_n(t)}{\|\theta_n(t)\|}$ converges as t tends to infinity to some vector $\bar{\theta}_n$ which is colinear to some KKT point of the following problem

$$\min_\theta \frac{1}{2} \|\theta\|^2 \text{ s.t. } \forall i = 1, \ldots, n; y_i \Psi_\theta(x_i) \geq 1. \tag{20}$$

Conditional to the event $E :=$ "$\max_{i \neq j} \left\{ |x_i^T x_j| \right\} \leq \frac{s+1}{3n} - 1$", by Vardi et al. (2022) Lemma C.1 we get that for all $j = 1, \ldots, n$, we have

$$y_j \Psi_{\bar{\theta}_n}(x_j) = \lambda(z_1, \cdots, z_n), \tag{21}$$

for some $\lambda(z_1, \cdots, z_n) > 0$.
We consider our learning procedure $\mathcal{A}$ to output

$$\mathcal{A}(z_1, \cdots, z_n) = \hat{\theta}_n := \frac{\bar{\theta}_n}{\sqrt{\lambda(z_1, \cdots, z_n)}},$$

which gives the same classifier as with $\bar{\theta}_n$.
We then get the following result.

Proposition 19 *Assume that $l \geq n$ and let $C := \max_{i \neq j} \left\{ |x_i^T x_j| \right\}$. Then, there exists an initialization $\theta_n(0)$ of the gradient flow for which it holds that*

$$\tilde{D}_\gamma \left(\mathbb{P}_{(\hat{\theta}_n, z_0)}, \mathbb{P}_{(\hat{\theta}_n, z_1)} \right) \geq \frac{1}{\gamma} \mathbb{P} \left(C \leq \frac{s+1}{3n} - 1 \right) + \left(1 - \frac{1}{\gamma} \right).$$

Moreover, if the marginal distribution of x is the uniform distribution on $\sqrt{s}\mathbb{S}^{s-1}$, then

$$\mathbb{P} \left(C \leq \frac{s+1}{3n} - 1 \right) \geq 1 - s^{3 - ln(s)/4},$$

as soon as $n \leq \frac{1}{3} \frac{s+1}{\sqrt{s ln(s)} + 1}$.

Proposition 19 with Theorem 6 show that the MIS is then upper bounded as

$$\text{Sec}_{\nu,\lambda,n}(P, \mathcal{A}) \leq \max(1, 1/\gamma) s^{3 - ln(s)/4}.$$

Proof [Proof of Proposition 19] By definition of Ψ_θ for any $\theta \in \Theta$, it holds that these networks are $2-$homogeneous, so that conditional to the event E, equation 21 leads to

$$y_j \Psi_{\hat{\theta}_n}(x_j) = 1, \tag{22}$$

for any $j = 1, \dots, n$.
Let $S := \{(\theta, x, y) \in \Theta \times \left(\sqrt{s}\mathbb{S}^{s-1}\right) \times \{-1, 1\} : y\Psi_\theta(x) = 1\}$. Then, by definition of $\Delta_n(P, \mathcal{A})$, we have

$$\begin{aligned}
\tilde{D}_\gamma\left(\mathbb{P}_{(\hat{\theta}_n, z_0)}, \mathbb{P}_{(\hat{\theta}_n, z_1)}\right) &\geq \frac{1}{\gamma}\mathbb{P}((\hat{\theta}_n, x_1, y_1) \in S) - \mathbb{P}((\hat{\theta}_n, x, y) \in S) + \left(1 - \frac{1}{\gamma}\right) \\
&= \frac{1}{\gamma}\mathbb{P}((\hat{\theta}_n, x_1, y_1) \in S \mid E)\mathbb{P}(E) + \frac{1}{\gamma}\mathbb{P}((\hat{\theta}_n, x_1, y_1) \in S \mid E^c)\mathbb{P}(E^c) + \left(1 - \frac{1}{\gamma}\right) \\
&- \mathbb{E}\left[\mathbb{P}(\Psi_{\hat{\theta}_n}(x) = y \mid \hat{\theta}_n, y)\right] \\
&\geq \frac{1}{\gamma}\mathbb{P}((\hat{\theta}_n, x_1, y_1) \in S \mid E)\mathbb{P}(E) + \left(1 - \frac{1}{\gamma}\right) - \mathbb{E}\left[\mathbb{P}(\Psi_{\hat{\theta}_n}(x) = y \mid \hat{\theta}_n, y)\right],
\end{aligned}$$

where we have lower bounded the second term by 0.
By equation 22, we have $\mathbb{P}((\hat{\theta}_n, x_1, y_1) \in S \mid E) = 1$. Now, by independence between (x, y) and $\hat{\theta}_n$, it is sufficient to show that for any $\theta \in \Theta$, we have $P(\Psi_\theta(x) = y \mid y) = 0$ almost surely. Without loss of generality, we may assume that $v_j \neq 0$ for any $j = 1, \dots, l$. We set $B_J(x, y) := \left\{\forall j \in J, w_j^T x + b_j > 0\right\} \cap \left\{\forall j \in J^c, w_j^T x + b_j \leq 0\right\} \cap \left\{\sum_{j \in J} v_j \left(w_j^T x + b_j\right) = y\right\}$ for any $J \subseteq [1, \cdots, l]$. We then get

$$\begin{aligned}
\mathbb{P}\left(\Psi_{\hat{\theta}_n}(x) = y \mid y\right) &= \sum_{J \subseteq [1, \cdots, l]} P\left(B_J(x, y) \mid y\right) \\
&\leq \sum_{J \subseteq [1, \cdots, l]} P\left(\sum_{j \in J} v_j \left(w_j^T x + b_j\right) = y \mid y\right).
\end{aligned}$$

Note that the space $H_{y,J} := \left\{x \in \mathbb{R}^s : \sum_{j \in J} v_j \left(w_j^T x + b_j\right) = y\right\}$ is an hyperplan of $\mathbb{R}^s$ for any $y \in \{-1, 1\}$ and any $J \subseteq [1, \cdots, l]$. Then the quantity $P(x \in H_{y,J} \mid y)$ equals 0 by $\mathcal{H}$. Hence,

$$\tilde{D}_\gamma(\mathbb{P}_{(\hat{\theta}_n, z_0)}, \mathbb{P}_{(\hat{\theta}_n, z_1)}) \geq \frac{1}{\gamma}\mathbb{P}(E) + \left(1 - \frac{1}{\gamma}\right).$$

Under the further hypothesis that x is uniformly distributed on the sphere, and that $n \leq \frac{1}{3}\frac{s+1}{\sqrt{s ln(s)+1}}$, it holds that $\frac{s+1}{3n} - 1 \geq \frac{\sqrt{s}}{ln(s)}$. Then Vardi et al. (2022) Lemma 3.1 concludes. ■

Appendix C. More comments on Section 5

We give here some more details about the behavior of $\Delta_{\nu,\lambda,n}(P,\mathcal{A})$ when the set of parameters Θ has finite cardinal. We also further discuss the quantity $C_K(P)$.

C.1 Different rates for $\Delta_{\nu,\lambda,n}(P,\mathcal{A})$

Corollary 15 gives a rate of $n^{-1/2}$ for $\max_{\nu,\mathcal{A}} \Delta_{\nu,\lambda,n}(P,\mathcal{A})$ when $C_K(P) < \infty$ and n is sufficiently large. In the case when $C_K(P)$ is infinite, it is interesting to note that we still have convergence to 1 of $\min_{\nu,\mathcal{A}} \mathrm{Sec}_{\nu,\lambda,n}(P,\mathcal{A})$ but at an arbitrarily slow rate. We formalize this result in the following lemma :

Lemma 20 *If $C_K(P) = \infty$, $\min_{\nu,\mathcal{A}} Sec_{\nu,\lambda,n}(P,\mathcal{A})$ still tends to 1 as n tends to infinity, but the (depending on P) rate can be arbitrarily slow.*

In this case, in order to find the minimum amount of data to get a control on $Sec_n(P,\mathcal{A})$ requires the estimation of the r.h.s. of equation 17 which is not obvious, as the condition $C_K(P) = \infty$ requires $K = \infty$.

Proof Remarking that $\tilde{D}_\gamma\left(\mathbb{P}_{(\hat{\theta}_n,z_0)}, \mathbb{P}_{(\hat{\theta}_n,z_1)}\right) \leq \|\mathbb{P}_{(\hat{\theta}_n,z_0)} - \mathbb{P}_{(\hat{\theta}_n,z_1)}\|_{TV} + (1-1/\gamma)$, it is a direct corollary of Lemmas 7 and 8 of Berend and Kontorovich (2013). ■

Corollary 15 also requires n to be at least $1/p_j$ for any $j = 1,\ldots,K$. Let $C_q(P) = \sum_{k=1}^q \sqrt{p_k(1-p_k)}$, where $p_1 \geq \ldots \geq p_K$. Without this assumption, Corollary 15 becomes

Lemma 21 *Assume $C_K(P) < \infty$ and $n \geq 5$. Assume without loss of generality that $p_1 \geq \cdots \geq p_K$, and let $k_n := max\{k : p_k \geq 1/n\}$. Then we have*

$$\min_{\nu,\mathcal{A}} Sec_{\nu,\lambda,n}(P,\mathcal{A}) = 1 - \varepsilon_n,$$

where ε_n satisfies

$$cn^{-1/2}C_{k_n}(P) + o(e^{-n}) \leq \varepsilon_n - \sum_{k>k_n} p_k(1-p_k)^n \leq c'n^{-1/2}C_{k_n}(P),$$

where c and c' are the universal constants from Corollary 15.

Lemma 21 shows that when the number of data n is less than the expected number of samples required to observe u_k, the influence of u_k in the computation is tightly related to its probability of appearing. Specifically, its probability of appearing is the probability of a Geometric random variable with parameter p_k of being at most n, i.e. exactly $p_k(1-p_k)^n$. When $n \geq 1/p_k$, the influence of u_k in $\min_{\mathcal{A}} \mathrm{Sec}_{\nu,\lambda,n}(P,\mathcal{A})$ changes, and is related to the diversity of the distribution $C_K(P)$.

Proof [Proof of Lemma 21] From the proof of Corollary 15, we know that the minimum is reached in $\gamma = 1$, i.e. $\nu = 1/(1+\lambda)$. In particular, by Lemma 29, we have

$$\frac{1}{2}\sum_{k=1}^{K}\mathbb{E}\left[\left|\frac{B_k}{n}-p_k\right|\right]=\frac{1}{n}\sum_{k=1}^{K}\psi(m_k,p_k),$$

where $\psi(m_k,p_k)=\binom{n}{m_k+1}(m_k+1)p_k^{m_k+1}(1-p_k)^{n-m_k}$ with $m_k=\lfloor np_k\rfloor$. For all $k>k_n$, it holds that $m_k=0$ and therefore $\psi(m_k)=p_k(1-p_k)^n$. The result follows from the proof of Corollary 15. ■

We conclude this section by providing an example in which $\Delta_{\nu,\lambda,n}(P,\mathcal{A})$ has a much faster rate than $n^{-1/2}$.

Lemma 22 *Let P be the Bernoulli distribution with parameter $p\in(0,1)$ and let $\hat{\theta}_n:=\sup_j z_j$. Then*

$$\Delta_{\nu,\lambda,n}(P,\mathcal{A})=(1/2)\max(1,1/\gamma)o((1-p)^n).$$

Proof [Proof of Lemma 22] From the integral form of $\tilde{D}_\alpha$ (see equation 26), one has for any distributions P and Q,

$$\tilde{D}_\alpha(P,Q)=\frac{1}{2\alpha}\int|\alpha p-q|\,d\zeta+\frac{1}{2}\left(1-\frac{1}{\alpha}\right).$$

Additionally, from equation 7, one has

$$\tilde{D}_\gamma\left(\mathbb{P}_{(\hat{\theta}_n,z_0)},\mathbb{P}_{(\hat{\theta}_n,z_1)}\right)=\mathbb{E}_{z_1}\left[\frac{1}{2\gamma}\int\left|\gamma p_{\hat{\theta}_n}-p_{\hat{\theta}_n|z_1}\right|d\zeta+\frac{1}{2}\left(1-\frac{1}{\gamma}\right)\right],$$

where $p_{\hat{\theta}_n}$ (resp. $p_{\hat{\theta}_n|z_1}$) is the density of $\mathbb{P}_{\hat{\theta}_n}$ (resp. $\mathbb{P}_{\hat{\theta}_n|z_1}$ conditional to z_1) with respect to ζ. As the distributions are discrete, for any $b\in\{0,1\}$, one has,

$$\begin{aligned}\int\left|\gamma p_{\hat{\theta}_n}-p_{\hat{\theta}_n|z_1}\right|d\zeta&=\left|\gamma\mathbb{P}\left(\hat{\theta}_n=1\right)-\mathbb{P}\left(\hat{\theta}_n=1\mid z_1=b\right)\right|\\&\quad+\left|\gamma\mathbb{P}\left(\hat{\theta}_n=0\right)-\mathbb{P}\left(\hat{\theta}_n=0\mid z_1=b\right)\right|\\&=\left|\gamma(1-(1-p)^n)-\begin{cases}1&\text{, if } b=1\\1-(1-p)^{n-1}&\text{, if } b=0\end{cases}\right|\\&\quad+\left|\gamma(1-p)^n-\begin{cases}0&\text{, if } b=1\\(1-p)^{n-1}&\text{, if } b=0\end{cases}\right|\\&=\begin{cases}\gamma(1-p)^n+|\gamma(1-p)^n-(\gamma-1)|&\text{, if } b=1\\(1-p)^n|\gamma-1/(1-p)|+\left|\gamma(1-p)^n-(\gamma-1)-(1-p)^{n-1}\right|&\text{, if } b=0\end{cases}.\end{aligned}$$

Taking the expectation over z_1 gives

$$\begin{aligned}
\mathbb{E}_{z_1}\left[\int \left|\gamma p_{\hat{\theta}_n} - p_{\hat{\theta}_n|z_1}\right| d\zeta\right] &= p\left[\gamma(1-p)^n + |\gamma(1-p)^n - (\gamma-1)|\right] \\
&+ (1-p)\left[(1-p)^n|\gamma - 1/(1-p)| + \left|\gamma(1-p)^n - (\gamma-1) - (1-p)^{n-1}\right|\right] \\
&= (1-p)^n\left[\gamma p + (1-p)|\gamma - 1/(1-p)|\right] \\
&+ p|\gamma-1|\left|1 - (1-p)^n\frac{\gamma}{\gamma-1}\right| \\
&+ (1-p)|\gamma-1|\left|1 + \frac{(1-p)^{n-1}}{\gamma-1} - (1-p)^n\frac{\gamma}{\gamma-1}\right| \\
&= |\gamma-1| + o((1-p)^n).
\end{aligned}$$

Combining everything gives

$$\begin{aligned}
\Delta_{\nu,\lambda,n}(P,\mathcal{A}) &= \max(1,\gamma)\left[\tilde{D}_\gamma\left(\mathbb{P}_{(\hat{\theta}_n,z_0)}, \mathbb{P}_{(\hat{\theta}_n,z_1)}\right) - \left(1-\frac{1}{\gamma}\right)_+\right] \\
&= \max(1,\gamma)\left[\frac{1}{2}\left|1-\frac{1}{\gamma}\right| + \frac{1}{2\gamma}o((1-p)^n) + \frac{1}{2}\left(1-\frac{1}{\gamma}\right) - \left(1-\frac{1}{\gamma}\right)_+\right] \\
&= (1/2)\max(1,1/\gamma)o((1-p)^n),
\end{aligned}$$

where the last equality comes from the fact that for any $a \in \mathbb{R}$,

$$(1/2)(1-a) - (1-a)_+ = -(1/2)|1-a|.$$

■

C.2 Relation of $C_K(P)$ with the Gini-Simpson Entropy and the Shannon's Entropy

We recall that for a discrete random variable X with distribution $P = \sum_{k=1}^K p_k \delta_{u_k}$, the Gini-Simpson Entropy is given by (see Bhargava and Doyle (1974); Rao (1982))

$$\text{G-S}(X) := 1 - \sum_{k=1}^K p_k^2,$$

and the Shannon Entropy is given by

$$H(X) := -\sum_{k=1}^K p_k \log(p_k).$$

From the inequality $\frac{1}{2}\sqrt{p(1-p)} \geq p(1-p)$ for all $p \in [0,1]$, we have

$$\frac{C_K(P)}{2} = \frac{1}{2}\sum_{k=1}^K \sqrt{p_k(1-p_k)} \geq \sum_{k=1}^K p_k(1-p_k) = \text{G-S}(X). \tag{23}$$

From the concavity of the square root, we also have

$$\frac{C_K(P)}{K} = \frac{1}{K}\sum_{k=1}^{K}\sqrt{p_k(1-p_k)} \leq \sqrt{\frac{1}{K}\sum_{k=1}^{K}p_k(1-p_k)} = \sqrt{\frac{1}{K}\text{G-S}(X)}. \tag{24}$$

The Gini-Simpson index can be interpreted as the expected distance between two randomly selected individuals when the distance is defined as zero if they belong to the same category and one otherwise (Rao et al., 1981), that is $\mathbb{P}(X \neq Y)$ for X and Y i.i.d.. The inequality mentioned above suggests that as the Gini-Simpson index increases (e.g., higher diversity of the data), security decreases and thus, the MIAs are expected to be more successful. Another, such commonly used diversity measure is Shannon entropy. Interestingly, $C_K(P)$ can be also upped and lower bounded by the Shannon entropy as follows:

$$H(X) \leq C_K(P) \leq \sqrt{K}\sqrt{H(X)}. \tag{25}$$

These bounds easily follow by noticing that

$$C_K(P) \leq \sqrt{K\left[1-\exp\left(-H(X)\right)\right]} \leq \sqrt{KH(X)},$$

by upper bounding:

$$-\log\left(\sum_{k=1}^{K}p_k^2\right) \leq H(X).$$

Similarly,

$$-p_k\log p_k \leq \sqrt{p_k(1-p_k)}, \quad \text{for all } 0 \leq p_k \leq 1,$$

and thus, $C_K(P) \geq H(X)$.
The Gini-Simpson Entropy and the Shannon Entropy are maximized by the uniform distribution. This is also the case for $C_K(P)$, as proved below.

Lemma 23 *Let $P := \sum_{k=1}^{K} p_k\delta_{u_k}$ be a discrete distribution with finite K. Let $\mathcal{M}_K$ be the set of all such distributions. We then have the following properties on $C_K(P)$. For fixed $K \geq 2$, we have*

- $\max\limits_{P\in\mathcal{M}_K} C_K(P) = \sqrt{K-1}$.
- $\underset{P\in\mathcal{M}_K}{argmax}\ C_K(P) = Unif(u_1, \cdots, u_K)$.

Interestingly, one can observe that for a fixed and finite number of atoms K, the sub-levels of $C_K(P)$ are tightly controlled by the value of $\max\limits_k p_k$. More precisely, at fixed $\max\limits_k p_k := \delta$ for some fixed $\delta \in [1/K; 1)$, the width of the interval $(inf\{C_K(P)\}, max\{C_K(P)\}]$ is entirely determined by δ and K. Specifically, the further δ is from $1/2$, the thinner the interval gets. We first summarize this comment below, and give the proof of Lemma 23 after.

Lemma 24 *Let $\delta \in [1/K, 1)$. Then the following statements hold :*

$$\max \left\{ C_K(P) : P \in \mathcal{M}_K, \max_k p_k = \delta \right\} = \sqrt{\delta(1-\delta)} + \sqrt{(1-\delta)(K-2+\delta)},$$
$$\inf \left\{ C_K(P) : P \in \mathcal{M}_K, \max_k p_k = \delta \right\} = \sqrt{\delta(1-\delta)}\lfloor \delta^{-1} \rfloor + \sqrt{\delta \lfloor \delta^{-1} \rfloor (1 - \delta \lfloor \delta^{-1} \rfloor)}.$$

Proof [Proof of Lemma 23] Denote by $M_K(\delta) \coloneqq \sqrt{\delta(1-\delta)} + \sqrt{(1-\delta)(K-2+\delta)}$. By Lemma 24, we have that $\max_{P \in \mathcal{M}_K} C_K(P) = \max_{\delta \in [1/K, 1)} M_K(\delta) = \sqrt{K-1}$ reached at $\big(\underbrace{\frac{1}{K}, \cdots, \frac{1}{K}}_{K}\big) \in \mathcal{M}_K(1/K)$. ∎

Proof [Proof of Lemma 24] Let $\mathcal{M}_K(\delta) \coloneqq \{P \in \mathcal{M}_K : \max_k p_k = \delta\}$. Without loss of generality, we always assume that $p_K = \max_k p_k$. Let $f : p \mapsto \sqrt{p(1-p)}$.
First notice that f is a concave function, so that for any $p_1, \cdots, p_m \in [0,1]$ we have $\frac{1}{m}\sum_{k=1}^m f(p_k) \leq f\left(\frac{1}{m}\sum_{k=1}^m p_k\right)$. In particular, for any $P \in \mathcal{M}_K(\delta)$, we have $\sum_{k=1}^{K-1} p_k = 1 - \delta$ giving

$$C_K(P) \leq C_K\left(\underbrace{\frac{1-\delta}{K-1}, \cdots, \frac{1-\delta}{K-1}}_{K-1}, \delta\right).$$

Evaluating the r.h.s. of the last inequality gives us the first result.

If $\delta \geq 1/2$, using again the concavity of f gives us that for any $P \in \mathcal{M}_K(\delta)$, we have

$$C_K(P) \geq C_K(\underbrace{0, \cdots, 0}_{K-2}, 1-\delta, \delta),$$

where the last quantity evaluates to $2\sqrt{\delta(1-\delta)}$.
If $\delta < 1/2$, using the concavity of f gives us that for any $P \in \mathcal{M}_K(\delta)$, we have

$$C_K(P) \geq C_K(0, \cdots, 0, 1 - \delta\lfloor \delta^{-1} \rfloor, \underbrace{\delta, \cdots, \delta}_{\lfloor \delta^{-1} \rfloor}),$$

evaluating at $\sqrt{\delta(1-\delta)}\lfloor \delta^{-1} \rfloor + \sqrt{\delta \lfloor \delta^{-1} \rfloor (1 - \delta \lfloor \delta^{-1} \rfloor)}$ for any numbers of zeros. Combining the two results give the last equality of the lemma. ∎

Appendix D. Setting for Section 6

We provide here additional values used for the numerical experiments presented in Section 6. All experiments have been conducted with PyTorch library.

D.1 Details for Section 6.1

Non-linear regression. We consider synthetically generated data $x_1, \cdots, x_n \overset{i.i.d.}{\sim} \mathrm{Unif}(\mathbb{S}_{d-1})$ with $d = 10$ and $n = 1000$. For any $j = 1, \ldots, n$, we set $y_j = \Psi^*(x_j) + \zeta_j$, where $\zeta_j \sim \mathcal{N}(0, 0.01)$ is some independent noise and Ψ^* is defined for all $x \in \mathbb{R}^d$ by $\Psi^*(x) = \sin(\pi\beta^T x)$ for some fixed $\beta \in \mathbb{S}_{d-1}$. For the sake of simplicity, we took $\beta = (1, 0, \cdots, 0)$, as the distribution of the data is rotation-invariant. We aim at estimating Ψ^* by a 2 layers ReLU neural network, whose hidden layer has width 4096. We train the neural network Ψ_θ by minimizing the MSE loss with the *Adam* optimizer and learning rate 0.1 for 2500 iterations. To evaluate the accuracy, we generated 10000 test samples $x_1^{test}, \cdots, x_{10000}^{test} \overset{i.i.d.}{\sim} \mathrm{Unif}(\mathbb{S}_{d-1})$ independently from the training dataset.

Linear regression and nearest neighbors. For both of the simulations, we have varied the dimension of input from $d = 10$ up to $d = 2000$ with uneven steps between the dimensions. For the nearest neighbors model, we have kept the same data as for the non-linear regression. For the linear regression, we have change Ψ^* simply by $\Psi^*(x) = \beta^T x$ for the same β. The nearest neighbor model has been implemented using the class **KNeighborsRegressor** of the *scikit-learn* library. The learning of the linear regressor has been made similarly to the training of the non-linear regressor.

D.2 Details for Section 6.2

We consider here the whole MNIST dataset, and use the given separation between training and test. We aimed at classifying the dataset by learning on a 3-layers ReLU neural network, with both internal layers having width 256. We trained the neural networks by minimizing the cross-entropy loss with the *Adam* optimizer and learning 0.01 for 500 iterations. To create the discretizations, we drew three times 1000 samples from the dataset (that have been used only for that purpose) and we constructed the clusterizing learning procedure based on the **MiniBatchKMeans** function from *scikit-learn* library. We used $K = 20, 50, 100$ clusters respectively. For each dataset size $n = 1000, 5000, 10000$, we performed the following steps :

- We draw a dataset D_n of size n not containing the samples used for the clusterings.
- We trained a first neural network on D_n (column **raw dataset**).
- For each clustering, we discretized D_n and then trained a neural network on the new dataset.

Appendix E. Proofs of Section 3

We give in this section the proofs relative to Section 3.

Proof [Proof of Proposition 5] Let p (resp. q) be the density of P (resp. Q) with respect to any dominating measure ζ. Let $B^* = \{\alpha p \geq q\}$ be the set reaching the supremum in the definition of $\tilde{D}_\gamma$, namely

$$\tilde{D}_\alpha(P, Q) = \frac{1}{\alpha}\int_{B^*} (\alpha p - q)\, d\zeta.$$

By taking the complementary in the supremum, we see that

$$\begin{aligned}\tilde{D}_\alpha(P, Q) &= \frac{1}{\alpha}\sup_B Q(B) - \alpha P(B) + \left(1 - \frac{1}{\alpha}\right) \\ &= \frac{1}{\alpha}\int_{(B^*)^c} (q - \alpha p) d\zeta + \left(1 - \frac{1}{\alpha}\right).\end{aligned}$$

Taking the average of both gives,

$$\tilde{D}_\alpha(P, Q) = \frac{1}{2\alpha}\int |\alpha p - q|\, d\zeta + \frac{1}{2}\left(1 - \frac{1}{\alpha}\right), \tag{26}$$

from which we deduce

$$\begin{aligned}D_\alpha(P, Q) &= \max(1, \alpha)\left[\frac{1}{2\alpha}\int |\alpha p - q|\, d\zeta + \frac{1}{2}\left(1 - \frac{1}{\alpha}\right) - \left(1 - \frac{1}{\alpha}\right)_+\right] &(27)\\ &= \int f_\alpha(p/q) q d\zeta, &(28)\end{aligned}$$

where the second equality comes from $(1/2)(1 - 1/\alpha) - (1 - 1/\alpha)_+ = -(1/2)|1 - 1/\alpha|$.
Equation 7 follows immediately from equations 26 and 27.
From equation 28, f_α being convex, continuous and satisfies $f_\alpha(1) = 0$, D_α is an $f-$divergence.
We now show the inequalities. We only prove the lower bound as the upper bound is trivial.
By definition of $\tilde{D}_\alpha$, for any measurable set B, we have,

$$\tilde{D}_\alpha(P, Q) \geq \frac{1}{\alpha}\left(\alpha P(B) - Q(B)\right).$$

Then taking B as the whole space, we get,

$$\tilde{D}_\alpha(P, Q) \geq \frac{1}{\alpha}(\alpha - 1) = 1 - \frac{1}{\alpha}.$$

Additionally, taking B as the null set, we have trivially $\tilde{D}_\alpha \geq 0$, which implies $\tilde{D}_\alpha(P, Q) \geq \left(1 - \frac{1}{\alpha}\right)_+$, which leads to the lower bound. ■

Proof [Proof of Theorem 6]
By the *i.i.d.* assumption of the data, and by the independence of T to the rest of the random variables, we have

$$\mathrm{Acc}_{\nu,\lambda,n}(\phi; P, \mathcal{A}) = \nu\mathbb{P}\left(\phi(\hat{\theta}_n, z_0) = 0\right) + \lambda(1 - \nu)\mathbb{P}\left(\phi(\hat{\theta}_n, z_1) = 1\right).$$

We now define $B \coloneqq \{(\theta, z) \in \Theta \times \mathcal{Z} : \phi(\theta, z) = 1\}$ and rewrite $\mathrm{Acc}_{\nu,\lambda,n}(\phi; P, \mathcal{A})$ as

$$\mathrm{Acc}_{\nu,\lambda,n}(\phi; P, \mathcal{A}) = \nu \left(1 - \mathbb{P}\left((\hat{\theta}_n, z_0) \in B\right)\right) + \lambda(1-\nu)\mathbb{P}\left((\hat{\theta}_n, z_1) \in B\right). \tag{29}$$

Taking the maximum over all MIAs ϕ then reduces to taking the maximum of the r.h.s. of equation 29 over all measurable sets B. We then get

$$\max_{\phi} \mathrm{Acc}_{\nu,\lambda,n}(\phi; P, \mathcal{A}) = \lambda(1-\nu)\max_{B}\Big[\mathbb{P}\left((\hat{\theta}_n, z_1) \in B\right) - \gamma\mathbb{P}\left((\hat{\theta}_n, z_0) \in B\right)\Big] + \nu, \tag{30}$$

where the maximum is taken over all measurable sets B. Let now ζ be a dominating measure of the distributions of $(\hat{\theta}_n, z_0)$ and $(\hat{\theta}_n, z_1)$ (for instance their average). We denote by p (resp. q) the density of the distribution $\mathbb{P}_{(\hat{\theta}_n, z_0)}$ of $(\hat{\theta}_n, z_0)$ (resp. $\mathbb{P}_{(\hat{\theta}_n, z_1)}$ for $(\hat{\theta}_n, z_1)$) with respect to ζ. Then, the involved maximum in the r.h.s. of equation 30 is reached on the set

$$B^* \coloneqq \{q/p \geq \gamma\}.$$

The maximum being taken over all measurable sets in equation 30, we may consider replacing B by its complementary B^c in the expression giving

$$\max_{\phi} \mathrm{Acc}_{\nu,\lambda,n}(\phi; P, \mathcal{A}) = \lambda(1-\nu)\max_{B}\left[\gamma\mathbb{P}\left((\hat{\theta}_n, z_0) \in B\right) - \mathbb{P}\left((\hat{\theta}_n, z_1) \in B\right)\right] + \lambda(1-\nu), \tag{31}$$

which comes from the definition of γ. In this case the maximum is reached on the set

$$B^{*c} \coloneqq \{q/p < \gamma\}.$$

In particular, by the definition of $\tilde{D}_\alpha$, we then get,

$$\max_{\phi} \mathrm{Acc}_{\nu,\lambda,n}(\phi; P, \mathcal{A}) = \lambda(1-\nu) + \nu\tilde{D}_\gamma\left(\mathbb{P}_{(\hat{\theta}_n, z_0)}, \mathbb{P}_{(\hat{\theta}_n, z_1)}\right).$$

Therefore, if $\lambda(1-\nu) \geq \nu$, it holds $\gamma \leq 1$, and in particular $\tilde{D}_\gamma\left(\mathbb{P}_{(\hat{\theta}_n, z_0)}, \mathbb{P}_{(\hat{\theta}_n, z_1)}\right) = \Delta_{\nu,\lambda,n}(P, \mathcal{A})$. Therefore, we have

$$\begin{aligned}\max_{\phi} \mathrm{Acc}_{\nu,\lambda,n}(\phi; P, \mathcal{A}) - \lambda(1-\nu) &= \nu\tilde{D}_\gamma\left(\mathbb{P}_{(\hat{\theta}_n, z_0)}, \mathbb{P}_{(\hat{\theta}_n, z_1)}\right) \\ &= \nu\Delta_{\nu,\lambda,n}(P, \mathcal{A}).\end{aligned}$$

On the converse, if $\lambda(1-\nu) \leq \nu$, it holds $\gamma \geq 1$, and we have

$$\begin{aligned}\max_{\phi} \mathrm{Acc}_{\nu,\lambda,n}(\phi; P, \mathcal{A}) - \nu &= \lambda(1-\nu) + \nu\left(\tilde{D}_\gamma\left(\mathbb{P}_{(\hat{\theta}_n, z_0)}, \mathbb{P}_{(\hat{\theta}_n, z_1)}\right) - 1\right) \\ &= \lambda(1-\nu) + \nu\left(\frac{1}{\gamma}\Delta_{\nu,\lambda,n}(P, \mathcal{A}) + \left(1 - \frac{1}{\gamma}\right) - 1\right) \\ &= \nu\frac{1}{\gamma}\Delta_{\nu,\lambda,n}(P, \mathcal{A}),\end{aligned}$$

hence the inequality.
Now, by the definition of the MIS, we have

$$\begin{aligned}
\mathrm{Sec}_{\nu,\lambda,n}(P,\mathcal{A}) &= \frac{1}{\min(\nu,\lambda(1-\nu))}\left(\nu+\lambda(1-\nu)-\left[\lambda(1-\nu)+\nu\tilde{D}_\gamma\left(\mathbb{P}_{(\hat{\theta}_n,z_0)},\mathbb{P}_{(\hat{\theta}_n,z_1)}\right)\right]\right)\\
&= \frac{1}{\min(1,1/\gamma)}\left(1-\tilde{D}_\gamma\left(\mathbb{P}_{(\hat{\theta}_n,z_0)},\mathbb{P}_{(\hat{\theta}_n,z_1)}\right)\right)\\
&= 1-\frac{1}{\min(1,1/\gamma)}\left(\min(1,1/\gamma)-1+\tilde{D}_\gamma\left(\mathbb{P}_{(\hat{\theta}_n,z_0)},\mathbb{P}_{(\hat{\theta}_n,z_1)}\right)\right)\\
&= 1-\frac{1}{\min(1,1/\gamma)}\left(\tilde{D}_\gamma\left(\mathbb{P}_{(\hat{\theta}_n,z_0)},\mathbb{P}_{(\hat{\theta}_n,z_1)}\right)-\left(1-\frac{1}{\gamma}\right)_+\right)\\
&= 1-\Delta_{\nu,\lambda,n}(P,\mathcal{A}),
\end{aligned}$$

hence the equality. ■

Appendix F. Proofs of Section 4

We give here the proofs for the Section 4.

Proof [Proof of Proposition 10] Let $l_j \coloneqq l_j(\hat{\theta}_n) = l_{\hat{\theta}_n}(x_j,y_j)$. The learning procedure $\mathcal{A}_{\varepsilon,\alpha}$ stops as soon as $\frac{1}{n}\sum_{j=1}^n l_{\hat{\theta}_n}(x_j,y_j) \le \varepsilon\alpha$.
Let $B_\varepsilon \coloneqq \left\{j : l_{\hat{\theta}_n}(x_j,y_j) \le \varepsilon\right\}$ be the set of samples with loss not larger than ε at the end of the training. We then have the following sequence of inequalities:

$$n\varepsilon\alpha \ge \sum_{j=1}^n l_{\hat{\theta}_n}(x_j,y_j) \ge \sum_{j\in B_\varepsilon} l_{\hat{\theta}_n}(x_j,y_j) + (n-\#B_\varepsilon)\varepsilon \ge (n-\#B_\varepsilon)\varepsilon.$$

From the two extremes, we get that $\sum_{j=1}^n 1\{l_j \le \varepsilon\} \coloneqq \#B_\varepsilon \ge n(1-\alpha)$. From the *i.i.d.* hypothesis on the data $z_1,\cdots,z_n$, taking the expectation gives the result. ■

Proof [Proof of Theorem 11] Let $S_\theta^\varepsilon \coloneqq \{(x,y)\in\mathcal{X}\times\mathcal{Y} : l_\theta(x,y)\le\varepsilon\}$ be the ε-sub-level set of l_θ for all $\theta\in\Theta$. We begin by proving the first point. Let $\mathcal{A}$ be an $(\varepsilon,1-\alpha)$-overfitting learning procedure, and let $S^\varepsilon \coloneqq \{(\theta,x,y) : l_\theta(x,y)\le\varepsilon\}$. From the definition of $\tilde{D}_\gamma$, and by taking the complementary in the supremum, we have that

$$\begin{aligned}
\tilde{D}_\gamma(\mathbb{P}_{(\hat{\theta}_n,z_0)},\mathbb{P}_{(\hat{\theta}_n,z_1)}) &\ge \frac{1}{\gamma}\mathbb{P}((\hat{\theta}_n,x_1,y_1)\in S^\varepsilon) - \mathbb{P}((\hat{\theta}_n,x,y)\in S^\varepsilon) + \left(1-\frac{1}{\gamma}\right)\\
&= \frac{1}{\gamma}\mathbb{P}((x_1,y_1)\in S^\varepsilon_{\hat{\theta}_n}) - \mathbb{P}((x,y)\in S^\varepsilon_{\hat{\theta}_n}) + \left(1-\frac{1}{\gamma}\right)\\
&\ge \frac{1}{\gamma}(1-\alpha) - \mathbb{P}((x,y)\in S^\varepsilon_{\hat{\theta}_n}) + \left(1-\frac{1}{\gamma}\right)\\
&= 1-\frac{1}{\gamma}\alpha - \int_{\theta\in\Theta} P((x,y)\in S^\varepsilon_\theta)d\mu_{\hat{\theta}_n},
\end{aligned}$$

which proves the first point by

$$\begin{aligned}
\mathrm{Sec}_{\nu,\lambda,n}(P,\mathcal{A}) &= 1-\Delta_{\nu,\lambda,n}(P,\mathcal{A}) \\
&= 1-\max(1,\gamma)\left(\tilde{D}_{\gamma}(\mathbb{P}_{(\hat{\theta}_n,z_0)},\mathbb{P}_{(\hat{\theta}_n,z_1)})-\left(1-\frac{1}{\gamma}\right)_{+}\right) \\
&= \begin{cases} 1-\gamma\left[1-\frac{1}{\gamma}\alpha-\mathbb{P}\left((x,y)\in S^{\varepsilon}_{\hat{\theta}_n}\right)-\left(1-\frac{1}{\gamma}\right)\right] & \text{, if } \gamma\geq 1 \\ 1-\left[1-\frac{1}{\gamma}\alpha-\mathbb{P}\left((x,y)\in S^{\varepsilon}_{\hat{\theta}_n}\right)\right] & \text{, if } \gamma\leq 1 \end{cases} \\
&= \begin{cases} \alpha+\gamma\mathbb{P}\left((x,y)\in S^{\varepsilon}_{\hat{\theta}_n}\right) & \text{, if } \gamma\geq 1 \\ \frac{1}{\gamma}\left(\alpha+\gamma\mathbb{P}\left((x,y)\in S^{\varepsilon}_{\hat{\theta}_n}\right)\right) & \text{, if } \gamma\leq 1 \end{cases} \\
&\leq \max(1,1/\gamma)\left(\alpha+\gamma\mathbb{P}\left((x,y)\in S^{\varepsilon}_{\hat{\theta}_n}\right)\right).
\end{aligned}$$

Now assume that we have a sequence of learning procedures $(\mathcal{A}_\eta)_{\eta\in\mathbb{R}^+}$ that stop as soon as $L_n \leq \eta$. Assume the additional hypotheses given in the second point of Theorem 11 hold. Let $\alpha \in (0,1)$ be a fixed scalar. By Proposition 10, $\mathcal{A}_\eta$ is $(\eta/\alpha, 1-\alpha)$-overfitting, so that by the first point proven above, we have

$$\mathrm{Sec}_{\nu,\lambda,n}(P,\mathcal{A}_\eta) \leq \max(1,1/\gamma)\left(\alpha+\gamma\mathbb{P}\left((x,y)\in S^{\eta/\alpha}_{\hat{\theta}_n}\right)\right).$$

For any $\theta\in\Theta$, we have

$$\begin{aligned}
P((x,y)\in S^{\eta/\alpha}_{\theta}) &= \mathbb{E}\left[P((x,y)\in S^{\eta/\alpha}_{\theta}\mid x)\right] \\
&= \mathbb{E}\left[P(\omega(y,\Psi_\theta(x))\leq \eta/\alpha\mid x)\right] \\
&\overset{\eta\to 0^+}{\to} 0,
\end{aligned}$$

where the limit comes from the continuity of ω and the absolute continuity of the distribution of y given x. In particular, for any $\alpha\in(0,1)$, we then have

$$\lim_{\eta\to 0}\mathrm{Sec}_{\nu,\lambda,n}(P,\mathcal{A}_\eta)\leq \max(1,1/\gamma)\alpha.$$

Taking the infimum over α gives the result.

■

Appendix G. Proofs of Section 5

We give here the proofs for the Section 5.

Proof [Proof of Theorem 13] First, note that we have for any distributions P and Q,

$$\begin{aligned}\tilde{D}_\alpha(P,Q) &= \frac{1}{\alpha}\sup_B \,[\alpha P(B) - \alpha Q(B) + (\alpha-1)Q(B)] \\ &\leq \begin{cases} \|P-Q\|_{TV} & \text{, if } \alpha < 1 \\ \|P-Q\|_{TV} + \left(1-\frac{1}{\alpha}\right) & \text{, if } \alpha \geq 1 \end{cases} \\ &= \|P-Q\|_{TV} + \left(1-\frac{1}{\alpha}\right)_+ .\end{aligned}$$

By Theorem 6, we have

$$\mathrm{Sec}_{\nu,\lambda,n}(P,\mathcal{A}) \geq 1 - \max(1,\gamma)\left\|\mathbb{P}_{(\hat{\theta}_n,z_0)} - \mathbb{P}_{(\hat{\theta}_n,z_1)}\right\|_{TV}.$$

We will therefore prove an upper bound on $\left\|\mathbb{P}_{(\hat{\theta}_n,z_0)} - \mathbb{P}_{(\hat{\theta}_n,z_1)}\right\|_{TV}$. For any positive integer j, we let $m_j := \mathbb{E}\left[\left\|C^{-1/2}\left\{L(z_1) - \mathbb{E}\left[L(z_1)\right]\right\}\right\|_2^j\right]$ be the expectation of the j-th power of the norm of the centered and reduced version of $L(z_1)$, where C is the covariance matrix of $L(z_1)$.
Setting $L_n := \frac{1}{n}\sum_{j=1}^n L(z_j)$, by the data processing inequality (Ziv and Zakai, 1973) applied to the total variation distance, for any measurable map $g : \mathbb{R}^d \times \mathcal{Z} \to \mathcal{Z}'$ taking values in any measurable space $\mathcal{Z}'$, we have

$$\|\mathcal{L}(g(L_n,z_1)) - \mathcal{L}(g(L_n,z_0))\|_{\mathrm{TV}} \leq \|\mathcal{L}((L_n,z_1)) - \mathcal{L}((L_n,z_0))\|_{\mathrm{TV}}.$$

The inequality holds in particular for g defined for all (l,z) in $\mathbb{R}^d \times \mathcal{Z}$ by $g(l,z) = (F(l),z)$, from which we get

$$\begin{aligned}\left\|\mathbb{P}_{(\hat{\theta}_n,z_1)} - \mathbb{P}_{(\hat{\theta}_n,z_0)}\right\|_{TV} &\leq \|\mathcal{L}((L_n,z_1)) - \mathcal{L}((L_n,z_0))\|_{\mathrm{TV}} \\ &= \mathbb{E}\left[\|\mathcal{L}(L_n \mid z_1) - \mathcal{L}(L_n)\|_{\mathrm{TV}}\right],\end{aligned}$$

in which the expectation is taken over z_1. Here, for any random variable x, $\mathcal{L}(x)$ denotes its distribution.
For $j = 1,\ldots,n$, denote by $\mathrm{v}_j := C^{-1/2}(L(z_j) - \mathbb{E}[L(z_j)])$ the centered and reduced version of $L(z_j)$. The total variation distance being invariant by translation and rescaling, we shall

write

$$
\begin{aligned}
\|\mathcal{L}(L_n \mid z_1) - \mathcal{L}(L_n)\|_{\mathrm{TV}} &= \left\|\mathcal{L}(L_n - \mathbb{E}[L(z_1)]) - \mathcal{L}(L_n - \mathbb{E}[L(z_1)] \mid z_1)\right\|_{\mathrm{TV}} \\
&= \left\|\mathcal{L}\left(\frac{1}{n}\sum_{j=1}^{n}(L(z_j) - \mathbb{E}[L(z_j)])\right) - \mathcal{L}\left(\frac{1}{n}\sum_{j=1}^{n}(L(z_j) - \mathbb{E}[L(z_j)]) \;\Big|\; z_1\right)\right\|_{\mathrm{TV}} \\
&= \left\|\mathcal{L}\left(\frac{C^{-1/2}}{\sqrt{n}}\sum_{j=1}^{n}(L(z_j) - \mathbb{E}[L(z_j)])\right) - \mathcal{L}\left(\frac{C^{-1/2}}{\sqrt{n}}\sum_{j=1}^{n}(L(z_j) - \mathbb{E}[L(z_j)]) \;\Big|\; z_1\right)\right\|_{\mathrm{TV}} \\
&= \left\|\mathcal{L}\left(\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\mathrm{v}_j\right) - \mathcal{L}\left(\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\mathrm{v}_j \;\Big|\; \mathrm{v}_1\right)\right\|_{\mathrm{TV}}.
\end{aligned}
$$

Denoting by $\mathcal{N}_d(\beta, \Sigma)$ the $d-$dimensional normal distribution with parameters (β, Σ), it holds almost surely that

$$
\begin{aligned}
\left\|\mathcal{L}\left(\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\mathrm{v}_j\right) - \mathcal{L}\left(\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\mathrm{v}_j \;\Big|\; \mathrm{v}_1\right)\right\|_{\mathrm{TV}} &\leq \left\|\mathcal{L}\left(\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\mathrm{v}_j\right) - \mathcal{N}_d(0, \boldsymbol{I}_d)\right\|_{\mathrm{TV}} \\
&+ \left\|\mathcal{L}\left(\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\mathrm{v}_j \;\Big|\; \mathrm{v}_1\right) - \mathcal{N}_d\left(\frac{1}{\sqrt{n}}\mathrm{v}_1, \frac{n-1}{n}I_d\right)\right\|_{\mathrm{TV}} \\
&+ \left\|\mathcal{N}_d\left(0, \boldsymbol{I}_d\right) - \mathcal{N}_d\left(\frac{1}{\sqrt{n}}\mathrm{v}_1, \frac{n-1}{n}\boldsymbol{I}_d\right)\right\|_{\mathrm{TV}} \\
&= \left\|\mathcal{L}\left(\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\mathrm{v}_j\right) - \mathcal{N}_d(0, \boldsymbol{I}_d)\right\|_{\mathrm{TV}} \\
&+ \left\|\mathcal{L}\left(\frac{1}{\sqrt{n-1}}\sum_{j=1}^{n-1}\mathrm{v}_j\right) - \mathcal{N}_d(0, \boldsymbol{I}_d)\right\|_{\mathrm{TV}} \\
&+ \left\|\mathcal{N}_d\left(0, \boldsymbol{I}_d\right) - \mathcal{N}_d\left(\frac{1}{\sqrt{n}}\mathrm{v}_1, \frac{n-1}{n}\boldsymbol{I}_d\right)\right\|_{\mathrm{TV}}.
\end{aligned}
$$

Applying Theorem 2.6 of Bally and Caramellino (2016) with variable v_j and parameter $r = 2$, one can upper bound the first two terms by some constant $C(d)(1+m_3)$ times $n^{-1/2}$. The constant $C(d)$ here depends only on the dimension of the parameters d. We may upper bound the last term by the following proposition.

Proposition 25 *Let n be an integer and $\beta \in \mathbb{R}^d$ be any $d-$dimensional vector. Then it holds that*

$$
\left\|\mathcal{N}_d(0, \boldsymbol{I}_d) - \mathcal{N}_d\left(\frac{1}{\sqrt{n}}\beta, \frac{n-1}{n}\boldsymbol{I}_d\right)\right\|_{TV} \leq \frac{\sqrt{d}}{2n} + \frac{1}{2\sqrt{n}}\|\beta\|_2.
$$

Applying Proposition 25 to the last quantity, it holds that

$$\left\|\mathcal{N}_d\left(0, \boldsymbol{I}_d\right) - \mathcal{N}_d\left(\frac{1}{\sqrt{n}}\mathrm{v}_1, \frac{n-1}{n}\boldsymbol{I}_d\right)\right\|_{\mathrm{TV}} \leq \frac{\sqrt{d}}{2n} + \frac{1}{2\sqrt{n}}\|\mathrm{v}_1\|_2,$$

and the result follows from taking the expectation, with $c_{L,P} = C(d)(1+m_3) + \frac{m_1}{2}$. ■

Proof [Proof of Proposition 25]
Applying Proposition 2.1 of Devroye et al. (2018), it holds almost surely that

$$\begin{aligned}
&\left\|\mathcal{N}_d(0, \boldsymbol{I}_d) - \mathcal{N}_d\left(\frac{1}{\sqrt{n}}\beta, \frac{n-1}{n}\boldsymbol{I}_d\right)\right\|_{\mathrm{TV}} \\
&\leq \frac{1}{2}\sqrt{tr\left(\boldsymbol{I}_d\frac{n-1}{n}\boldsymbol{I}_d - \boldsymbol{I}_d\right) + \frac{1}{n}\|\beta\|_2^2 - \ln\left(det\left(\frac{n-1}{n}\boldsymbol{I}_d\right)\right)} \\
&= \frac{1}{2}\sqrt{-\frac{d}{n} + \frac{1}{n}\|\beta\|_2^2 - d\ln\left(\frac{n-1}{n}\right)} \\
&\leq \frac{1}{2}\sqrt{-d\left(\frac{1}{n} + \ln\left(\frac{n-1}{n}\right)\right)} + \frac{1}{2}\sqrt{\frac{1}{n}\|\beta\|_2^2} \\
&\leq \frac{\sqrt{d}}{2n} + \frac{1}{2\sqrt{n}}\|\beta\|_2,
\end{aligned}$$

where $tr(\cdot)$ is the trace operator and $det(\cdot)$ is the matrix determinant operator. The third inequality is due to $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for positive scalars a and b. The first term in the last inequality comes from the fact that $x - 1 - \ln(x) \leq (x-1)^2$ if $x \geq 1/3$ which holds with $x = \frac{n-1}{n}$ for $n \geq 2$. ■

Proof [Proof of Remark 13] From equation 15, we have that

$$cn^{-1/2} + \frac{\sqrt{d}}{2}n^{-1} \leq \varepsilon,$$

is sufficient to ensure $\left\|\mathbb{P}_{(\hat{\theta}_n, z_1)} - \mathbb{P}_{(\hat{\theta}_n, z_0)}\right\|_{TV} \leq \varepsilon$, hence a security of at least $1 - \max(1, \gamma)\varepsilon$. Setting $x \coloneqq n^{-1/2}$, it is equivalent to

$$cx + \frac{\sqrt{d}}{2}x^2 - \varepsilon \leq 0.$$

From the the study of the above quadratic function, as $x \geq 0$ is assumed, we get that this is equivalent to

$$
\begin{aligned}
n^{-1/2} &\leq \frac{-c + \sqrt{c^2 + 2\varepsilon\sqrt{d}}}{\sqrt{d}} \\
\iff n &\geq \frac{d}{2c^2 + 2\varepsilon\sqrt{d} - 2c\sqrt{c^2 + 2\varepsilon\sqrt{d}}} \\
&= \frac{d}{2c^2} \frac{1}{1 + \frac{\varepsilon\sqrt{d}}{c^2} - \sqrt{1 + 2\frac{\varepsilon\sqrt{d}}{c^2}}}.
\end{aligned}
$$

From the mean-value form of Taylor theorem of order 2 at 0, there exists $0 \leq \bar{u} \leq u := \frac{\varepsilon\sqrt{d}}{c^2}$ such that

$$
\sqrt{1+2u} = 1 + u - \frac{1}{2}(1+2\bar{u})^{-3/2}.
$$

Therefore, the condition becomes

$$
\begin{aligned}
n &\geq \frac{d}{2c^2} \frac{2(1+2\bar{u})^{3/2}}{u^2} \\
&= \varepsilon^{-2} c^2 (1+2\bar{u})^{3/2}.
\end{aligned}
$$

As $\bar{u} \leq u \leq \frac{\sqrt{d}}{c^2}$, $n \geq \varepsilon^{-2}c^2(1 + \frac{\sqrt{d}}{c^2})^{3/2}$ ensures the above condition, hence the result. ■

We recall from Proposition 5 that the following equation holds

$$
\Delta_{\nu,\lambda,n}(P, \mathcal{A}) = \mathbb{E}_{z_1}\left[D_\gamma\left(\mathbb{P}_{\hat{\theta}_n}, \mathbb{P}_{\hat{\theta}_n|z_1}\right)\right].
$$

Proof [Proof of Theorem 14] The proof will be divided in two steps. First, we will prove the inequality

$$
\Delta_{\nu,\lambda,n}(P, \mathcal{A}) \leq \max(1, \gamma)\left[\frac{1}{2\gamma}\sum_{k=1}^{K}\mathbb{E}\left[\left|\frac{B_k}{n} - \gamma p_k\right|\right] - \frac{1}{2}\left|1 - \frac{1}{\gamma}\right|\right], \tag{32}
$$

for any distribution P and learning procedure $\mathcal{A}$. Second, we prove that this upper bound is reached for learning procedures that map any data set to a Dirac mass, summarized in the following lemma.

Lemma 26 *For $k = 1, \ldots, K$, let B_k be random variables having Binomial distribution with parameters (n, p_k). Suppose that $\mathcal{A}(z_1, \cdots, z_n) = \delta_{F\left(\frac{1}{n}\sum_{j=1}^n \delta_{z_j}\right)}$ for any $n \in \mathbb{N}$ and $z_1, \ldots, z_n \in \mathcal{Z}$, for some measurable map $F : \mathcal{M} \to \Theta$ with infinite range $|\Theta| = \infty$, i.e. $\hat{\theta}_n \stackrel{\mathcal{L}}{=} F\left(\frac{1}{n}\sum_{j=1}^n \delta_{z_j}\right)$. Then we have*

$$
\max_F \Delta_{\nu,\lambda,n}(P, \mathcal{A}) = \max(1, \gamma)\left[\frac{1}{2\gamma}\sum_{k=1}^{K}\mathbb{E}\left[\left|\frac{B_k}{n} - \gamma p_k\right|\right] - \frac{1}{2}\left|1 - \frac{1}{\gamma}\right|\right].
$$

Lemma 26 and equation 32 will imply Theorem 14 as follows,

$$\begin{aligned}\min_{\mathcal{A}} \mathrm{Sec}_{\nu,\lambda,n}(P,\mathcal{A}) &= 1 - \max_{\mathcal{A}} \Delta_{\nu,\lambda,n}(P,\mathcal{A}) \\ &= 1 - \max(1,\gamma)\left[\frac{1}{2\gamma}\sum_{k=1}^{K}\mathbb{E}\left[\left|\frac{B_k}{n} - \gamma p_k\right|\right] - \frac{1}{2}\left|1 - \frac{1}{\gamma}\right|\right].\end{aligned}$$

We first prove equation 32.
Since $\hat{\theta}_n$ has distribution $G(\hat{P}_n)$ conditionally on $\mathbf{z}$, where $\hat{P}_n := \frac{1}{n}\sum_{j=1}^{n}\delta_{z_j}$ is the empirical distribution of the data set, from Proposition 18, we have

$$\begin{aligned}\mathbb{P}(\hat{\theta}_n \in B) &= \mathbb{E}[\mathbb{P}(\hat{\theta}_n \in B|\mathbf{z})] \\ &= \mathbb{E}[G(\hat{P}_n)(B)]\end{aligned} \tag{33}$$

$$\mathbb{P}(\hat{\theta}_n \in B|z_1) = \mathbb{E}[G(\hat{P}_n)(B)|z_1], \tag{34}$$

for any measurable set B.
Recall that $u_1,\ldots,u_K$ are the (fixed) support points of P. For any $k \in \{1,\cdots,K\}$, let $\hat{P}_n^k := \frac{1}{n}\left(\delta_{u_k} + \sum_{j=2}^{n}\delta_{z_j}\right)$. Using equations 7, 33 and 34 we may rewrite $\tilde{D}_\gamma$ as

$$\tilde{D}_\gamma\left(\mathbb{P}_{(\hat{\theta}_n,z_0)},\mathbb{P}_{(\hat{\theta}_n,z_1)}\right) = \sum_{k=1}^{K} p_k \frac{1}{\gamma}\sup_B\left(\gamma\mathbb{E}[G(\hat{P}_n)(B)] - \mathbb{E}[G(\hat{P}_n^k)(B)]\right). \tag{35}$$

For any integer n, let $\mathcal{M}_n$ be the set of all possible empirical distributions for data sets with n points and let $\mathcal{G}_n = G(\mathcal{M}_n)$. Since P has at most countable support, then $\mathcal{G}_n$ is at most countable and equation 35 gives

$$\tilde{D}_\gamma\left(\mathbb{P}_{(\hat{\theta}_n,z_0)},\mathbb{P}_{(\hat{\theta}_n,z_1)}\right) = \sum_{k=1}^{K} p_k \frac{1}{\gamma}\sup_B \sum_{g\in\mathcal{G}_n} g(B)\left[\gamma\mathbb{P}(G(\hat{P}_n) = g) - \mathbb{P}(G(\hat{P}_n^k) = g)\right]. \tag{36}$$

For some fixed $g \in \mathcal{G}_n$, let us denote by $\mathcal{M}_n(g) = G^{-1}(\{g\}) \cap \mathcal{M}_n$ the set of possible empirical distributions Q in $\mathcal{M}_n$ such that $G(Q) = g$. Then we have for any $g \in \mathcal{G}_n$,

$$g(B)\left(\gamma\mathbb{P}(G(\hat{P}_n) = g) - \mathbb{P}(G(\hat{P}_n^k) = g)\right) = \sum_{Q\in\mathcal{M}_n(g)} G(Q)(B)\left(\gamma P^n(\hat{P}_n = Q) - P^n(\hat{P}_n^k = Q)\right),$$

so that summing over all g gives

$$\tilde{D}_\gamma\left(\mathbb{P}_{(\hat{\theta}_n,z_0)},\mathbb{P}_{(\hat{\theta}_n,z_1)}\right) = \sum_{k=1}^{K} p_k \frac{1}{\gamma}\sup_B \sum_{Q\in\mathcal{M}_n} G(Q)(B)\left[\gamma P^n(\hat{P}_n = Q) - P^n(\hat{P}_n^k = Q)\right], \tag{37}$$

since $(\mathcal{M}_n(g))_{g\in\mathcal{G}_n}$ is a partition of $\mathcal{M}_n$. As the distribution is discrete, any possible value Q of $\hat{P}_n$ is uniquely determined by a $K-$tuple $(k_1,\cdots,k_K)$ (if $K = \infty$ then by a sequence

$(k_1, k_2, \cdots))$ of non-negative integers such that $\sum_{j=1}^{K} k_j = n$ and $Q = \frac{1}{n}\sum_{j=1}^{K} k_j \delta_{u_j}$. The $K-$tuple (or sequence) corresponds to the distribution of the samples among the atoms, that is, if we define, for $j = 1, \ldots, K$, the random variable N_j as the number of samples in the dataset equal to u_j, then for such Q,

$$P^n(\hat{P}_n = Q) = \mathbb{P}(N_j = k_j;\ j = 1, \ldots, K).$$

Since the samples are i.i.d., we get for such Q

$$P^n(\hat{P}_n = Q) = \binom{n}{k_1, \cdots, k_K} \prod_{j=1}^{K} {p_j}^{k_j}, \tag{38}$$

where $\binom{n}{k_1, \cdots, k_m} = \frac{n!}{k_1! \cdots k_m!}$ is the multinomial coefficient. Notice that when $K = +\infty$, only a finite number m of integers k_j are non zero, so that equation 38 can be understood to hold also when $K = +\infty$ by keeping only the terms involving the positive integers k_j.

Let us now compute $P^{n-1}(\hat{P}_n^1 = Q)$. If $k_1 = 0$, then $P^{n-1}(\hat{P}_n^1 = Q) = 0$. Else,

$$\begin{aligned} P^{n-1}(\hat{P}_n^1 = Q) &= \mathbb{P}(N_1 = k_1 - 1,\ N_j = k_j;\ j = 2, \ldots, K) \\ &= \binom{n-1}{k_1 - 1, k_2, \cdots, k_K} \left(\prod_{j=2}^{K} p_j^{k_j}\right) p_1^{k_1 - 1} \\ &= \frac{k_1}{np_1} \binom{n}{k_1, \cdots, k_K} \prod_{j=1}^{K} p_j^{k_j}, \end{aligned}$$

which again is understood to hold also when $K = +\infty$.
Therefore in both cases, we get

$$P^{n-1}(\hat{P}_n^1 = Q) = \frac{k_1}{np_1} \binom{n}{k_1, \cdots, k_K} \prod_{j=1}^{K} p_j^{k_j}. \tag{39}$$

Now, using 38 and 39, denoting by g_N the image by G of the distribution determined by the $K-$tuple $N = (k_1, \cdots, k_K)$, we get

$$\begin{aligned} \sum_{Q \in \mathcal{M}_n} G(Q)(B) \left[\gamma P^n(\hat{P}_n = Q) - P^{n-1}(\hat{P}_n^1 = Q)\right] &= \sum_{k_1 + \cdots + k_K = n} g_N(B) \binom{n}{k_1, \cdots, k_K} \prod_{j=1}^{K} p_j^{k_j} \left(\gamma - \frac{k_1}{np_1}\right) \\ &= \mathbb{E}\left[\left(\gamma - \frac{N_1}{np_1}\right) g_N(B)\right], \end{aligned}$$

where $N = (N_1, \cdots, N_K)$ follows a multinomial distribution of parameters $(n; p_1, \cdots, p_K)$. The computation being similar for any $k = 1, \ldots, K$, we easily obtain,

$$\sum_{Q \in \mathcal{M}_n} G(Q)(B) \left[\gamma P^n(\hat{P}_n = Q) - P^{n-1}(\hat{P}_n^k = Q)\right] = \mathbb{E}\left[\left(\gamma - \frac{N_k}{np_k}\right) g_N(B)\right].$$

Now, plugging it into equation 37 gives,

$$\tilde{D}_\gamma\left(\mathbb{P}_{(\hat{\theta}_n,z_0)},\mathbb{P}_{(\hat{\theta}_n,z_1)}\right)=\sum_{k=1}^{K}p_k\frac{1}{\gamma}\sup_B\mathbb{E}\left[\left(\gamma-\frac{N_k}{np_k}\right)g_N(B)\right]. \tag{40}$$

We get from equation 40

$$\begin{aligned}\tilde{D}_\gamma\left(\mathbb{P}_{(\hat{\theta}_n,z_0)},\mathbb{P}_{(\hat{\theta}_n,z_1)}\right)&\le\sum_{k=1}^{K}p_k\frac{1}{\gamma}\mathbb{E}\left[\sup_B\left(\gamma-\frac{N_k}{np_k}\right)g_N(B)\right]\\&=\sum_{k=1}^{K}p_k\frac{1}{\gamma}\mathbb{E}\left[\left(\gamma-\frac{N_k}{np_k}\right)_+\right]\end{aligned} \tag{41}$$

$$\tilde{D}_\gamma\left(\mathbb{P}_{(\hat{\theta}_n,z_0)},\mathbb{P}_{(\hat{\theta}_n,z_1)}\right)\le\sum_{k=1}^{K}p_k\frac{1}{\gamma}\mathbb{E}\left[\left(\gamma-\frac{N_k}{np_k}\right)_-\right]+\left(1-\frac{1}{\gamma}\right), \tag{42}$$

where the equality in equation 41 comes from the fact that the supremum is reached on null sets when $\gamma-\frac{N_k}{np_k}$ is negative, and on sets of mass 1 when it is positive. Equation 42 is obtained by replacing B by its complementary B^c in the supremum and remarking that $\mathbb{E}[\gamma-\frac{N_k}{np_k}]=\gamma-1$. Taking the average of equations 41 and 42 gives

$$\tilde{D}_\gamma\left(\mathbb{P}_{(\hat{\theta}_n,z_0)},\mathbb{P}_{(\hat{\theta}_n,z_1)}\right)\le\frac{1}{2\gamma}\sum_{k=1}^{K}\mathbb{E}\left[\left|\frac{N_k}{n}-\gamma p_k\right|\right]+\frac{1}{2}\left(1-\frac{1}{\gamma}\right),$$

hence we get

$$\begin{aligned}\Delta_{\nu,\lambda,n}(P,\mathcal{A})&\le\max(1,\gamma)\left[\frac{1}{2\gamma}\sum_{k=1}^{K}\mathbb{E}\left[\left|\frac{N_k}{n}-\gamma p_k\right|\right]+\frac{1}{2}\left(1-\frac{1}{\gamma}\right)-\left(1-\frac{1}{\gamma}\right)_+\right]\\&=\max(1,\gamma)\left[\frac{1}{2\gamma}\sum_{k=1}^{K}\mathbb{E}\left[\left|\frac{N_k}{n}-\gamma p_k\right|\right]-\frac{1}{2}\left|1-\frac{1}{\gamma}\right|\right],\end{aligned}$$

which proves Equation 32. ■

Proof [Proof of Lemma 26] For some fixed $\theta\in\Theta$, we similarly denote by $\mathcal{M}_n(\theta)=F^{-1}(\{\theta\})\cap\mathcal{M}_n$ the set of possible empirical distributions Q in $\mathcal{M}_n$ such that $F(Q)=\theta$. Using equation 7, equation 26, and following similar steps as in equations 35, 36 and 37, by triangular inequality, we get that

$$
\begin{aligned}
\tilde{D}_\gamma\left(\mathbb{P}_{(\hat{\theta}_n,z_0)},\mathbb{P}_{(\hat{\theta}_n,z_1)}\right) &= \frac{1}{\gamma}\sum_{k=1}^{K}\frac{p_k}{2}\sum_{g\in\mathcal{G}_n}\left|\gamma\mathbb{P}(\delta_{F(\hat{P}_n)}=g)-\mathbb{P}(\delta_{F(\hat{P}^k_n)}=g)\right|+\frac{1}{2}\left(1-\frac{1}{\gamma}\right)\\
&= \frac{1}{\gamma}\sum_{k=1}^{K}\frac{p_k}{2}\sum_{\theta\in\Theta}\left|\gamma\mathbb{P}(F(\hat{P}_n)=\theta)-\mathbb{P}(F(\hat{P}^k_n)=\theta)\right|+\frac{1}{2}\left(1-\frac{1}{\gamma}\right)\\
&= \frac{1}{\gamma}\sum_{k=1}^{K}\frac{p_k}{2}\sum_{\theta\in\Theta}\left|\sum_{Q\in\mathcal{M}_n(\theta)}\left(\gamma P^n(\hat{P}_n=Q)-P^{n-1}(\hat{P}^k_n=Q)\right)\right|+\frac{1}{2}\left(1-\frac{1}{\gamma}\right) \qquad (43)\\
&\le \frac{1}{\gamma}\sum_{k=1}^{K}\frac{p_k}{2}\sum_{Q\in\mathcal{M}_n}\left|\gamma P^n(\hat{P}_n=Q)-P^{n-1}(\hat{P}^k_n=Q)\right|+\frac{1}{2}\left(1-\frac{1}{\gamma}\right),
\end{aligned}
$$

since $(\mathcal{M}_n(\theta))_{\theta\in\Theta}$ is a partition of $\mathcal{M}_n$. We now prove that when taking the maximum over all possible measurable maps F having range Θ, the inequality becomes an equality. Indeed, since Θ is infinite, it is possible to construct F such that F is an injection from $\bigcup_{n\in\mathbb{N}}\mathcal{M}_n$ to Θ, in which case for all $\theta\in\Theta$, $\mathcal{M}_n(\theta)$ is either the empty set or a singleton. Thus, equation 43 gives

$$
\max_F \tilde{D}_\gamma\left(\mathbb{P}_{(\hat{\theta}_n,z_0)},\mathbb{P}_{(\hat{\theta}_n,z_1)}\right) = \frac{1}{2\gamma}\sum_{k=1}^{K}p_k\sum_{Q\in\mathcal{M}_n}\left|\gamma P^n(\hat{P}_n=Q)-P^{n-1}(\hat{P}^k_n=Q)\right|+\frac{1}{2}\left(1-\frac{1}{\gamma}\right),
$$

and the lemma follows from equations 38 and 39 and the same steps as in the proof of Theorem 14. ■

To prove Corollary 15, we need the following intermediary results.

Lemma 27 *The function* $f^+ : x\mapsto \sum_{k=1}^{K}\mathbb{E}\left[\left|x\frac{B_k}{n}-p_k\right|\right]-(x-1)$ *is non-increasing on* $[1,\infty)$.

Lemma 28 *For any* $t,p\in\mathbb{R}^+$*, the function* $f^-_{t,p} : x\mapsto \frac{1}{x}|xt-p|-p\left(\frac{1}{x}-1\right)$ *is non-decreasing on* $(0,1]$.

Lemma 29 *Define for any* $m=0,\dots,n$ *and any* $p\in(0,1)$ *the function* $\psi(m,p)=\binom{n}{m+1}(m+1)p^{m+1}(1-p)^{n-m}$*. Then letting* $m_k=\lfloor\gamma n p_k\rfloor$*, it holds that*

$$
\frac{1}{2\gamma}\sum_{k=1}^{K}\mathbb{E}\left[\left|\frac{B_k}{n}-\gamma p_k\right|\right]=\frac{1}{\gamma n}\sum_{k=1}^{K}\psi(m_k,p_k)+\frac{1}{2}\left|1-\frac{1}{\gamma}\right|-\left|1-\frac{1}{\gamma}\right|o(e^{-n}). \qquad (44)
$$

Lemma 30 *Let* $m=\lfloor np\rfloor$ *for some probability* $p\in(0,1)$ *such that* $1\le m\le n-2$*. Then there exist universal constants* $c>0.29$ *and* $c'<0.44$ *such that,*

$$
c\sqrt{n}\sqrt{p(1-p)}\le\psi(m,p)\le c'\sqrt{n}\sqrt{p(1-p)}. \qquad (45)
$$

Additionally, to prove Lemma 29, we need a result from De Moivre (1730) which we state and prove at the end of the section for completion.

Lemma 31 (De Moivre (1730)) *Let $n \geq 1$ be an integer, and $p \in (0,1)$ be a real number. Let $m = 0, \cdots, n$ be any integer. Then it holds that*

$$\sum_{k=0}^{m} \binom{n}{k} p^k (1-p)^{n-k} (np-k) = (m+1) \binom{n}{m+1} p^{m+1} (1-p)^{n-m}, \tag{46}$$

where $\binom{n}{k}$ is the binomial coefficient. For generality, we set $\binom{n}{n+1} = 0$.

Proof [Proof of Corollary 15] By Theorem 14, we have

$$\max_{\mathcal{A}} \Delta_{\nu,\lambda,n}(P, \mathcal{A}) = \max(1,\gamma) \left[\frac{1}{2\gamma} \sum_{k=1}^{K} \mathbb{E}\left[\left| \frac{B_k}{n} - \gamma p_k \right| \right] - \frac{1}{2} \left| 1 - \frac{1}{\gamma} \right| \right].$$

Now, from Lemma 27, for any $\gamma \leq 1$, we have $1/\gamma \geq 1$, and it holds that

$$\begin{aligned} \max_{\mathcal{A}} \Delta_{\nu,\lambda,n}(P, \mathcal{A}) &= \frac{1}{2} \sum_{k=1}^{K} \mathbb{E}\left[\left| \frac{1}{\gamma} \frac{B_k}{n} - p_k \right| \right] - \frac{1}{2} \left(\frac{1}{\gamma} - 1 \right) \\ &= \frac{1}{2} f_+(1/\gamma) \\ &\leq \frac{1}{2} f_+(1) \\ &= \frac{1}{2} \sum_{k=1}^{K} \mathbb{E}\left[\left| \frac{B_k}{n} - p_k \right| \right]. \end{aligned}$$

From Lemma 28, for any $\gamma \geq 1$, we have $1/\gamma \leq 1$, and it holds that

$$\begin{aligned} \max_{\mathcal{A}} \Delta_{\nu,\lambda,n}(P, \mathcal{A}) &= \frac{1}{2} \sum_{k=1}^{K} \mathbb{E}\left[\left| \frac{B_k}{n} - \gamma p_k \right| \right] - \frac{1}{2} (\gamma - 1) \\ &= \frac{1}{2} \sum_{k=1}^{K} \mathbb{E}\left[\gamma \left| \frac{1}{\gamma} \frac{B_k}{n} - p_k \right| - p_k (\gamma - 1) \right] \\ &= \frac{1}{2} \sum_{k=1}^{K} \mathbb{E}\left[f^-_{B_k/n, p_k}\left(\frac{1}{\gamma} \right) \right] \\ &\leq \frac{1}{2} \sum_{k=1}^{K} \mathbb{E}\left[f^-_{B_k/n, p_k}(1) \right] \\ &= \frac{1}{2} \sum_{k=1}^{K} \mathbb{E}\left[\left| \frac{B_k}{n} - p_k \right| \right]. \end{aligned}$$

In particular, this means that the minimum security is reached for $\gamma = 1$, i.e. $\nu = 1/(1+\lambda)$. On the one hand, if $\gamma = 1$ and for any $k = 1, \ldots, K$, $n > 1/p_k$, then also for any $k = 1, \ldots, K$,

$n > 1/(1-p_k)$. This implies that for any $k = 1, \dots, K$, we have $1 \le m_k \le n-2$. Then Theorem 14, Lemma 29 and Lemma 30 yield the result.
On the other hand, if the condition on n does not hold, we still have for any $k = 1, \dots, K$ by Cauchy-Schwartz inequality,

$$\mathbb{E}\left[\left|\frac{B_k}{n} - p_j\right|\right] \le \sqrt{Var(B_k/n)} = n^{-1/2}\sqrt{p_k(1-p_k)},$$

which shows the lower bound on the MIS when the condition on n does not hold..

■

Proof [Proof of Lemma 27]
Let $q_j^k = \mathbb{P}(B_k = j)$. Then we have

$$f^+(x) = \sum_{k=1}^{K}\sum_{j=1}^{n} q_j^k \left|x\frac{j}{n} - p_k\right| - (x-1).$$

Now note that f^+ is continuous and almost everywhere differentiable. Letting $m_j = \lfloor np_j/x \rfloor$, on any point of differentiability, we have,

$$\begin{aligned}
(f^+)'(x) &= \sum_{k=1}^{K}\sum_{j=1}^{n} q_j^k \frac{\partial}{\partial x}\left|x\frac{j}{n} - p_k\right| - 1 \\
&= -1 + \sum_{k=1}^{K}\sum_{j=1}^{n} q_j^k \begin{cases} -j/n & \text{, if } j < np_k/x \\ j/n & \text{, if } j > np_k/x \end{cases} \\
&= -1 + \sum_{k=1}^{K}\sum_{j=1}^{n} q_j^k \left[-\frac{j}{n}1_{j \le m_k} + \frac{j}{n}1_{j > m_k+1}\right] \\
&= -1 + \sum_{k=1}^{K}\left[-2\mathbb{E}\left[\frac{B_k}{n}1_{B_k \le m_k}\right] + p_k\right] \\
&= -2\sum_{k=1}^{K}\mathbb{E}\left[\frac{B_k}{n}1_{B_k \le m_k}\right] \le 0,
\end{aligned}$$

which shows that f^+ is non-increasing by continuity of the function. ■

Proof [Proof of Lemma 28] The proof is similar to the proof of Lemma 27. In particular, on any point of differentiability, we have

$$
\begin{aligned}
(f^-_{t,p})'(x) &= \frac{-1}{x^2}|xt-p| + \frac{1}{x}\frac{\partial}{\partial x}|xt-p| + \frac{p}{x^2} \\
&= \frac{1}{x^2}\left[p - |xt-p|\right] + x\begin{cases} -t & \text{, if } x < p/t \\ t & \text{, if } x > p/t \end{cases} \\
&= \frac{1}{x^2}\begin{cases} p-(p-xt)-xt & \text{, if } x < p/t \\ p-(xt-p)+xt, \text{ if } x > p/t \end{cases} \\
&= \frac{2p}{x^2}1_{x>p/t} \geq 0,
\end{aligned}
$$

which shows that $f^-_{t,p}$ is non-decreasing, by continuity of the function. ■

Proof [Proof of Lemma 29] By Lemma 31, for any $m_k \leq n$, we have $\mathbb{E}[(np_k - B_k)1_{B_k \leq m_k}] = \psi(m_k, p_k)$. Additionally, using the generalization $\binom{n}{n+k} = 0$ for $k > 0$, the equality still holds using the fact that $\mathbb{E}[np_k - B_k] = 0$. Then we have

$$
\begin{aligned}
\mathbb{E}\left[|B_k - \gamma np_k|\right] &= \mathbb{E}\left[(\gamma np_k - B_k)\left(1_{B_k\leq m_k} - 1_{B_k > m_k}\right)\right] \\
&= \mathbb{E}\left[(\gamma np_k - \gamma B_k)\left(1_{B_k\leq m_k} - 1_{B_k > m_k}\right)\right] + (\gamma-1)\,\mathbb{E}\left[B_k\left(1_{B_k\leq m_k} - 1_{B_k > m_k}\right)\right] \\
&= 2\gamma\mathbb{E}\left[(np_k - B_k)\,1_{B_k\leq m_k}\right] + (\gamma-1)\,\mathbb{E}\left[2B_k 1_{B_k\leq m_k} - np_k\right] \\
&= 2\gamma\psi(m_k,p_k) + (\gamma-1)\,\mathbb{E}\left[2(B_k - np_k)1_{B_k\leq m_k} + 2np_k\mathbb{P}(B_k\leq m_k) - np_k\right] \\
&= 2\gamma\psi(m_k,p_k) + (\gamma-1)\left[-2\psi(m_k,p_k) + 2np_k\mathbb{P}(B_k\leq m_k) - np_k\right] \\
&= 2\psi(m_k,p_k) + (\gamma-1)\,np_k\left(2\mathbb{P}(B_k\leq m_k) - 1\right),
\end{aligned}
$$

which gives,

$$
\begin{aligned}
\frac{1}{2\gamma}\sum_{k=1}^K \mathbb{E}\left[\left|\frac{B_k}{n} - \gamma p_k\right|\right] &= \sum_{k=1}^K\left(\frac{1}{\gamma n}\psi(m_k,p_k) + \frac{1}{2}\left(1-\frac{1}{\gamma}\right)p_k\left(2\mathbb{P}(B_k\leq m_k) - 1\right)\right) \\
&= \frac{1}{\gamma n}\sum_{k=1}^K\psi(m_k,p_k) + \left(1-\frac{1}{\gamma}\right)\left[-1/2 + \sum_{k=1}^K p_k\mathbb{P}(B_k\leq m_k)\right].
\end{aligned}
\tag{47}
$$

By Hoeffding Inequality, we have that,

$$
\sum_{k=1}^K p_k\mathbb{P}(B_k\leq m_k)\begin{cases} \geq 1 - \sum_{k=1}^K p_k e^{-np_k^2(1-\gamma)^2} & \text{, if } \gamma > 1 \\ \leq \sum_{k=1}^K p_k e^{-np_k^2(1-\gamma)^2} & \text{, if } \gamma \leq 1 \end{cases},
$$

which gives

$$
\begin{aligned}
\frac{1}{2\gamma}\sum_{k=1}^K \mathbb{E}\left[\left|\frac{B_k}{n} - \gamma p_k\right|\right] &= \frac{1}{\gamma n}\sum_{k=1}^K\psi(m_k,p_k) + \left(1-\frac{1}{\gamma}\right)\begin{cases} 1/2 - o(e^{-n}) & \text{, if } \gamma > 1 \\ -1/2 + o(e^{-n}) & \text{, if } \gamma \leq 1 \end{cases} \\
&= \frac{1}{\gamma n}\sum_{k=1}^K\psi(m_k,p_k) + \frac{1}{2}\left|1-\frac{1}{\gamma}\right| - \left|1-\frac{1}{\gamma}\right| o(e^{-n}),
\end{aligned}
$$

hence the result. ■

Proof [Proof of Lemma 30] We shall bound $\psi(m,p)$ by Robbins (1955), which states that for any integer $k \geq 1$, it holds that

$$\sqrt{2\pi}k^{k+1/2}e^{-k}e^{1/12(k+1)} < k! < \sqrt{2\pi}k^{k+1/2}e^{-k}e^{1/12k}. \tag{48}$$

Set $a := \exp\left(\frac{1}{12(n+1)} - \frac{1}{12(m+1)} - \frac{1}{12(n-(m+1))}\right)$. Since $1 \leq m \leq n-2$, we have

$$a \geq \exp(-1/6). \tag{49}$$

One may apply Inequality 48 to get

$$\begin{aligned}
\binom{n}{m+1} &> \frac{\sqrt{2\pi}n^{n+1/2}}{\sqrt{2\pi}(m+1)^{m+1+1/2}\sqrt{2\pi}(n-(m+1))^{n-(m+1)+1/2}}\frac{e^{-n}}{e^{-(m+1)}e^{-(n-(m+1))}}a \\
&= \frac{\sqrt{n}}{\sqrt{2\pi}}n^n\left[m+1\right]^{-(m+1+1/2)}\left[n-(m+1)\right]^{-(n-(m+1)+1/2)}a \\
&:= ba.
\end{aligned}$$

Now,

$$\begin{aligned}
b(m+1)p^{m+1}(1-p)^{n-m} &= \frac{\sqrt{n}}{\sqrt{2\pi}}n^n\left[m+1\right]^{-(m+1+1/2-1)}\left[n-(m+1)\right]^{-(n-(m+1)+1/2)} \\
&\times p^{m+1}(1-p)^{n-m} \\
&= \frac{\sqrt{n}}{\sqrt{2\pi}}n^n(np)^{-(m+1/2)}\left[\frac{m+1}{np}\right]^{-(m+1/2)} \\
&\times (n(1-p))^{-(n-(m+1/2))}\left[\frac{n-(m+1)}{n(1-p)}\right]^{-(n-(m+1/2))} \\
&\times p^{m+1}(1-p)^{n-m} \\
&= \frac{\sqrt{n}}{\sqrt{2\pi}}\sqrt{p(1-p)}\left[\frac{m+1}{np}\right]^{-(m+1/2)} \\
&\times \left[\frac{n-(m+1)}{n(1-p)}\right]^{-(n-(m+1/2))} \\
&:= \frac{\sqrt{n}}{\sqrt{2\pi}}\sqrt{p(1-p)}d,
\end{aligned}$$

which finally implies

$$\frac{\sqrt{n}}{\sqrt{2\pi}}\sqrt{p(1-p)}da < \binom{n}{m+1}(m+1)p^{m+1}(1-p)^{n-m}.$$

Similarly, set $\tilde{a} := \exp\left(\frac{1}{12n} - \frac{1}{12(m+2)} - \frac{1}{12(n-m)}\right)$. We then get

$$\tilde{a} \leq e^{-1/36}. \tag{50}$$

From the same steps as for the lower bound, we have,

$$\binom{n}{m+1}(m+1)p^{m+1}(1-p)^{n-m} < \frac{\sqrt{n}}{\sqrt{2\pi}}\sqrt{p(1-p)}d\tilde{a}.$$

Define $\epsilon \in [0,1)$ such that $np = m + \epsilon$. Then

$$d = \exp\left\{\left(m+\frac{1}{2}\right)\ln\left(\frac{m+\epsilon}{m+1}\right) + \left(n-m-\frac{1}{2}\right)\ln\left(\frac{n-m-\epsilon}{n-m-1}\right)\right\}.$$

For any $m \in \{1, \ldots, n-2\}$ and $\epsilon \in [0,1)$, define

$$f(m,\epsilon) = \left(m+\frac{1}{2}\right)\ln\left(\frac{m+\epsilon}{m+1}\right) + \left(n-m-\frac{1}{2}\right)\ln\left(\frac{n-m-\epsilon}{n-m-1}\right).$$

By studying the function $\epsilon \mapsto f(m,\epsilon)$ we get that for all $\epsilon \in [0,1)$, $f(m,\epsilon) \geq \min\{f(m,0), 0\}$. By studying the function $m \mapsto f(m,0)$ we get that for all $m \in \{1; \ldots; n-2\}$, $f(m,0) \geq \min\{f(1,0), f(n-2,0)\}$. But

$$f(1,0) = -f(n-2,0) = -\frac{3}{2}\log(2) + \left(n-\frac{3}{2}\right)\log\left(1+\frac{1}{n-2}\right).$$

Now, Taylor expansion of $\log(1+u)$ allows to prove

$$-\frac{3}{2}\log(2) + 1 - \frac{1}{4(n-2)^2} \leq f(1,0) \leq -\frac{3}{2}\log(2) + 1 + \frac{1}{2(n-2)}.$$

When $n \geq 5$, it is easy to see that $-\frac{3}{2}\log(2) + 1 - \frac{1}{4(n-2)^2} > 0$, so that using equation 49, and setting

$$c = \frac{\exp\left(\frac{3}{2}\log(2) - 1 - 1/3\right)}{\sqrt{2\pi}},$$

we get,

$$\psi(m,p) \geq c\sqrt{n}\sqrt{p(1-p)}.$$

A rough approximation gives $c > 0.29$.
By similar arguments, we obtain

$$\log(d) \leq \frac{7}{2}\log(7) - 2\log(3) - \frac{13}{2}\log(2). \tag{51}$$

Using equation 50 and setting

$$c' = \frac{\exp(\frac{7}{2}\log(7) - 2\log(3) - \frac{13}{2}\log(2) - 1/36)}{\sqrt{2\pi}},$$

we get,

$$\psi(m,p) \leq c'\sqrt{n}\sqrt{p(1-p)}.$$

A rough approximation gives $c' < 0.44$. ■

Proof [Proof of Lemma 31] We will prove the result by recursion. First, for $m = 0$, note that we have,

$$\binom{n}{0} p^0 (1-p)^n (np - 0) = np(1-p)^n$$
$$(0+1)\binom{n}{1} p^{0+1}(1-p)^n = np(1-p)^n,$$

which proves the initial statement. Assume that the result holds for some $m < n$, then we have,

$$\begin{aligned}
\sum_{k=0}^{m+1} \binom{n}{k} p^k (1-p)^{n-k}(np-k) &= (m+1)\binom{n}{m+1} p^{m+1}(1-p)^{n-m} \\
&\quad + \binom{n}{m+1} p^{m+1}(1-p)^{n-(m+1)}(np-(m+1)) \\
&= \binom{n}{m+1} p^{m+1}(1-p)^{n-(m+1)} \left[(m+1)(1-p) + np - (m+1)\right] \\
&= p^{m+2}(1-p)^{n-(m+1)}(n-(m+1)) \frac{n!}{(m+1)!(n-(m+1))!} \\
&= (m+2)\binom{n}{m+2} p^{m+2}(1-p)^{n-(m+1)},
\end{aligned}$$

where the first equality comes from the recursion hypothesis.
By recursion, the lemma holds. ■

Appendix H. Differential Privacy and MIS

We further discuss here the relation between differential privacy guarantees and the MIS studied in the paper. Specifically, we discuss the fact that guarantees given by differential privacy are not equivalent to guarantees given by the MIS. We first briefly describe differential privacy, and then compare it to the MIS.

H.1 Differential Privacy

For completeness, we first give the definition of $(\varepsilon, \delta)-$differential privacy.

Definition 32 ($(\varepsilon, \delta)-$differential privacy (Dwork et al., 2014)) *Let $\mathcal{A}$ be a learning procedure. We say that $\mathcal{A}$ satisfied $(\varepsilon, \delta)-$differential privacy if for any $S \subseteq \Theta$, and any neighboring datasets D and D', it holds*

$$\mathbb{P}(\mathcal{A}(D) \in S \mid D) \leq e^{\varepsilon} \mathbb{P}(\mathcal{A}(D') \in S \mid D') + \delta. \tag{52}$$

In the previous definition, we say that D and D' are neighboring if they differ by at most one entry. For small values of (ε, δ), one shall interpret differential privacy as some stability measure of the learning procedure. Indeed, differential privacy states that the conditional distribution of $\mathcal{A}(D)$ given D is "almost" indistinguishable from the conditional distribution of $\mathcal{A}(D')$ given D'.

In particular, Differential Privacy is a framework that considers the worst case scenario. Equation 52 must be satisfied for any measurable set S and any neighboring datasets D and D', including possible pathological datasets. In other words, Differential Privacy requires not to differentiate importance between datasets.

A key strength of differential privacy is its stability under composition. More specifically, if a learning procedure $\mathcal{A}$ is $(\varepsilon, \delta)-$DP, then it is $(k\varepsilon, k\delta)-$DP under $k-$fold composition (Dwork et al., 2006; Dwork and Lei, 2009). Future work have provided more precise characterization of the composition theorems (Kairouz et al., 2015). One shall interpret these "composition theorems" as procedures to privatize a learning procedure. Given an $(\varepsilon, \delta)-$DP Mechanism $\mathbb{M}$ and a learning procedure $\mathcal{A}$, one can use $\mathbb{M} \circ \mathcal{A}$ to transform $\mathcal{A}$ into an $(\varepsilon, \delta)-$DP learning procedure.

H.2 Comparison to the MIS

We begin by discussing core conceptual differences between differential privacy and the MIS.
Differential privacy is a tool to induce privacy into a model. Given a learning procedure $\mathcal{A}$ and a known $(\varepsilon, \delta)-$DP mechanism $\mathbb{M}$, one can transform $\mathcal{A}$ into an $(\varepsilon, \delta)-$DP learning procedure by composition $\mathbb{M} \circ \mathcal{A}$. However, for an arbitrary learning procedure $\mathcal{A}$, it does not provide a way of measuring the privacy of $\mathcal{A}$.
The MIS is a tool to measure the MIA-wise privacy of a learning procedure. For any learning procedure $\mathcal{A}$ it measures its MIA-wise privacy. However, it does not provide a way to induce privacy into a model.
Consequently, differential privacy is a tool to induce privacy whereas the MIS is a metric to quantify the privacy. In particular, though the respective objectives might intersect, their utility are not equivalent. Differential privacy gives a broad framework to define and control privacy, which "*intuitively will guarantee that a randomized learning procedure behaves similarly on similar input databases*" (Dwork et al., 2014). However, this does not formally answer a specific question, and shall have as many interpretations as there are privacy questions. On the converse, the MIS presented in this article gives a privacy metric specific to the MIA framework, arguably narrower than the differential privacy guarantees. The MIS is a metric that naturally comes into sight from the study of MIA accuracy, and results from considering the best performing MIA. Although this metric could be understood as quantifying some stability of the learning process, interpreting the MIS for other questions can not be done trivially.
In other words, differential privacy and the MIS does not convey the same message, and are simply to different frameworks.

Additionally, there are differences between their very definitions:

- **Conditional vs Marginal.** For differential privacy to hold, equation 52 must be satisfied for any measurable set S and any neighboring datasets D and D'. Fundamentally, differential privacy does not differentiate between natural datasets and pathological datasets. On the contrary, the MIS integrates over all possible datasets, and therefore pathological datasets would influence negligibly the output metric.

- **Privacy of the Learning Procedure vs. Privacy of the Model.** Given a dataset $\mathcal{D}$, for differential privacy to hold, the relation

$$\sup_S \sup_{D'} \left(\mathbb{P}(\mathcal{A}(D) \in S \mid D) - e^{\varepsilon}\mathbb{P}(\mathcal{A}(D') \in S \mid D')\right) \leq \delta,$$

must hold for any neighboring datasets $\mathcal{D}'$. Particularly, this shall be interpreted as any trained model $\mu_{\mathcal{A}(\mathcal{D})}$ must be (almost) indistinguishable from a neighboring trained model $\mu_{\mathcal{A}(\mathcal{D}')}$. On the other hand, letting $D = \{z_1, \cdots, z_n\}$ and $D' = \{z_1', z_2, \cdots, z_n\}$, then up to affine transformation, the central statistical quantity can be written as (see equation 7),

$$\mathbb{E}_{z_1} \sup_S \mathbb{E}_{D'} \left[\mathbb{P}\left(\mathcal{A}(D) \in S \mid D\right) - \frac{1}{\gamma}\mathbb{P}\left(\mathcal{A}(D') \in S \mid D'\right)\right].$$

Requiring a similar upper bound $\Delta_{\nu,\lambda,n}(P, \mathcal{A}) \leq \delta$ means that the learning procedure $\mathcal{A}$ must be stable over the task P.

Although the paradigm of differential privacy differs from the one of the MIS, they still are comparable.

From differential privacy to the MIS. When the learning procedure is known to be $(\varepsilon, \delta)-$DP for some $\varepsilon \in \mathbb{R}^+$ and $\delta \in (0, 1)$, then setting $D = \{z_1, \cdots, z_n\}$ and $D' = \{z_1', z_2, \cdots, z_n\}$, one can easily derive the equation of Remark 8 by,

$$\begin{aligned} \tilde{D}_\gamma\left(\mathbb{P}_{(\hat{\theta}_n, z_0)}, \mathbb{P}_{(\hat{\theta}_n, z_1)}\right) &= E_{z_1}\left[\frac{1}{\gamma}\sup_S \gamma\mathbb{P}\left(\hat{\theta}_n \in S \mid z_1\right) - \mathbb{P}\left(\hat{\theta}_n \in S\right)\right] \qquad (53) \\ &= E_{z_1}\left[\frac{1}{\gamma}\sup_S \mathbb{E}_{D'}\left[\gamma\mathbb{P}\left(\mathcal{A}(D) \in S \mid D\right) - \mathbb{P}\left(\mathcal{A}(D') \in S \mid D'\right)\right]\right] \\ &\leq E_{z_1}\left[\frac{1}{\gamma}\sup_S \gamma\left(e^{\varepsilon}\mathbb{P}\left(\hat{\theta}_n \in S\right) + \delta\right) - \mathbb{P}\left(\hat{\theta}_n \in S\right)\right] \\ &= \left(e^{\varepsilon} - \frac{1}{\gamma}\right)_+ + \delta, \end{aligned}$$

where the first equality comes from equation 7. From that, we get

$$\mathrm{Sec}_{\nu,\lambda,n}(P, \mathcal{A}) \geq 1 - \max(1, \gamma)\left(\left(e^{\varepsilon} - \frac{1}{\gamma}\right)_+ + \delta - \left(1 - \frac{1}{\gamma}\right)_+\right).$$

Notably, this simple relation shows that $\Delta_{\nu,\lambda,n}(P, \mathcal{A})$ can be controlled under differential privacy. It is worth noting that the bound we derived is heuristically similar to bounds already obtained for the KL divergence (Dwork et al., 2010). This bound could be potentially further refined by using Pinsker inequality and results on the reduction in KL divergence caused by passing data threough private channels (see Theorem 1 of Duchi et al. (2018) or Theorem 3.1 of Amorino and Gloter (2023)). Additionally, in many contexts, contraction results have been proved in the differential privacy framework. A remarkable result discussed in Remark I of Ghazi and Issa (2024) states that letting $\mathbb{A}_{\varepsilon,\delta} \coloneqq \{\mathcal{A} : \mathcal{A} \text{ is } (\varepsilon,\delta)\text{-DP}\}$ the set of all $(\varepsilon,\delta)-$DP learning procedures, when $\gamma = 1$, it holds that $\sup_{\mathcal{A}\in\mathbb{A}_{\varepsilon,\delta}} \Delta_{\nu,\lambda,n}(P,\mathcal{A}) = \delta + (1-\delta)\mathtt{tanh}(\varepsilon/2)$ where $\mathtt{tanh}$ is the hyperbolic tangent. This result still holds for Local Differentially Private (LDP) learning procedures (Dwork et al., 2006), namely letting $\mathbb{A}_{\varepsilon} \coloneqq \{\mathcal{A} : \mathcal{A} \text{ is } \varepsilon\text{-LDP}\}$ be the set of all $\varepsilon-$ LDP learning procedures, then when $\gamma = 1$ it holds that $\sup_{\mathcal{A}\in\mathbb{A}_{\varepsilon}} \Delta_{\nu,\lambda,n}(P,\mathcal{A}) = \mathtt{tanh}(\varepsilon/2)$ (Kairouz et al., 2016).

When $\gamma \neq 1$, one has the upper bound $\Delta_{\nu,\lambda,n}(P,\mathcal{A}) \leq \max(1,\gamma)\|\mathbb{P}_{(\hat{\theta}_n,z_0)} - \mathbb{P}_{(\hat{\theta}_n,z_1)}\|_{TV}$. This means that if $\mathcal{A} \in \mathbb{A}_{\varepsilon}$, then we have $\Delta_{\nu,\lambda,n}(P,\mathcal{A}) \leq \max(1,\gamma)\mathtt{tanh}(\varepsilon/2)$ (and similarly for $\mathcal{A} \in \mathbb{A}_{\varepsilon,\delta}$).

From MIS to differential privacy. On the converse, we discuss here that differential privacy is not required to hold relevantly for a learning procedure to be secure against MIAs. When the MIS is known to have some value $\mathrm{Sec}_n(P,\mathcal{A}) = s$, then it does not necessarily imply that there exist convenient values (ε,δ) for which the learning procedure $\mathcal{A}$ is $(\varepsilon,\delta)-$DP. For instance, differential privacy does not extend nicely to deterministic learning procedures.

Lemma 33 *Assume $\mathcal{A}$ is deterministic, then for any finite $\varepsilon \geq 0$ and $\delta \in (0,1)$, $\mathcal{A}$ is not $(\varepsilon,\delta)-$DP.*

However, as Theorem 14 together with Corollary 15 show that for any learning procedure $\mathcal{A}$, we have $\mathrm{Sec}_{\nu,\lambda,n}(P,\mathcal{A}) \underset{n\to\infty}{\to} 1$.

Another result which does not require the learning procedure to be deterministic can be stated. Assume the hypotheses of Section 5.1. Let $L : \mathcal{Z} \to \mathbb{R}^d$ and $F : \mathbb{R}^d \to \mathbb{R}^q$. Hence, we consider that the learning procedure outputs

$$\hat{\theta}_n = F\left(\frac{1}{n}\sum_{j=1}^{n} L(z_j)\right).$$

Further assume that the support of z is the whole space. Additionally, we may assume that the image of L, and F is the whole space as well.

Lemma 34 *Assume that $F(x) = \tilde{F}(x) + \eta N$, for some deterministic function $\tilde{F}$ and N follows a Normal distribution. Then for any finite $\varepsilon \geq 0$ and any $\delta \in (0,1)$, $\mathcal{A}$ is not $(\varepsilon,\delta)-$DP.*

Theorem 13 and Lemma 34 show that differential privacy can not provide relevant information, whereas the MIS can.

Proof [Proof of Lemma 33] Let fix D and D'. As $\mathcal{A}$ is assumed to be deterministic, we have that $\hat{\theta} = \mathcal{A}(D)$ and $\hat{\theta}' = \mathcal{A}(D')$ are constants conditionally to D and D'. As the differential privacy property must hold for any measurable set S, it must hold for singletons $\{\theta\} \subseteq \Theta$. If $\hat{\theta} \neq \hat{\theta}'$, then there exists S such that $\mathbb{P}(\hat{\theta} \in S \mid D) = 1$ and $\mathbb{P}(\hat{\theta}' \in S \mid D') = 0$. Hence, $\mathcal{A}$ can not be $(\varepsilon, \delta)-$DP for any $\delta < 1$. ■

Proof [Proof of Lemma 34] Let fix D and D'. Let $m = \tilde{F}\left(\frac{1}{n}\sum_{z \in D} L(z)\right)$ and $m' = \tilde{F}\left(\frac{1}{n}\sum_{z \in D'} L(z)\right)$. Conditionally to D (resp. D'), $\mathcal{A}(D)$ (resp. $\mathcal{A}(D')$) follows a Normal distribution with mean m (resp. m') and with covariance matrix $\eta^2 I_q$. Let $S_t = \{v \in \mathbb{R}^q : v_1 \leq t\}$. Then for $\mathcal{A}$ to satisfy $(\varepsilon, \delta)-$DP, it must hold, for any $t \in \mathbb{R}$, that

$$\mathbb{P}\left(\mathcal{A}(D) \in S_t \mid D\right) \leq e^{\varepsilon}\mathbb{P}\left(\mathcal{A}(D') \in S_t \mid D'\right) + \delta. \tag{54}$$

It is easy to verify that $\mathbb{P}\left(\mathcal{A}(D) \in S_t \mid D\right) = \Phi\left((t - m_1)/\eta\right)$ and $\mathbb{P}\left(\mathcal{A}(D') \in S_t \mid D'\right) = \Phi\left((t - m_1')/\eta\right)$, where Φ is the cumulative distribution function of the Standard Gaussian distribution. Now, as equation 54 must be satisfied for any neighboring datasets D and D', it must be satisfied when swapping roles between D and D'. Therefore, without loss of generality, we assume that $m_1' \geq m_1$. From the fact that the support of z is the whole space and the image of L and F is the whole space as well, we can construct neighboring datasets D and D' for which $|m_1 - m_1'|$ is arbitrarily large. In particular, we can consider m_1 arbitrarily large negatively and m_1' arbitrarily large positively which gives for any $t \in \mathbb{R}$,

$$\sup_{m_1, m_1'} \left| \Phi\left(\frac{t - m_1}{\eta}\right) - \Phi\left(\frac{t - m_1'}{\eta}\right) \right| = 1.$$

In particular, this means that there can not exist a finite ε and $\delta < 1$ such that $\mathcal{A}$ is $(\varepsilon, \delta)-$DP.

■

References

Chiara Amorino and Arnaud Gloter. Minimax rate for multivariate data under componentwise local differential privacy constraints. *arXiv preprint arXiv:2305.10416*, 2023.

Vlad Bally and Lucia Caramellino. Asymptotic development for the CLT in total variation distance. *Bernoulli*, 22(4):2442–2485, 2016. ISSN 1350-7265,1573-9759. doi: 10.3150/15-BEJ734. URL `https://doi.org/10.3150/15-BEJ734`.

Teodora Baluta, Shiqi Shen, S Hitarth, Shruti Tople, and Prateek Saxena. Membership inference attacks and generalization: A causal perspective. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 249–262, 2022.

Daniel Berend and Aryeh Kontorovich. A sharp estimate of the binomial mean absolute deviation with applications. *Statist. Probab. Lett.*, 83(4):1254–1259, 2013. ISSN 0167-7152,1879-2103. doi: 10.1016/j.spl.2013.01.023. URL `https://doi.org/10.1016/j.spl.2013.01.023`.

T.N. Bhargava and P.H. Doyle. A geometric study of diversity. *Journal of Theoretical Biology*, 43(2):241–251, 1974. ISSN 0022-5193. doi: https://doi.org/10.1016/S0022-5193(74)80057-3. URL https://www.sciencedirect.com/science/article/pii/S0022519374800573.

Olivier Bousquet and André Elisseeff. Stability and Generalization. *Journal of Machine Learning Research*, 2(Mar):499–526, 2002. ISSN ISSN 1533-7928. URL http://www.jmlr.org/papers/v2/bousquet02a.html.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 267–284, 2019.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1897–1914. IEEE, 2022a.

Nicholas Carlini, Matthew Jagielski, Chiyuan Zhang, Nicolas Papernot, Andreas Terzis, and Florian Tramer. The Privacy Onion Effect: Memorization is Relative. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 13263–13276. Curran Associates, Inc., 2022b.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying Memorization Across Neural Language Models. In *The Eleventh International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=TatRHT_1cK.

Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5253–5270, 2023b.

Imre Csiszár, Paul C Shields, et al. Information theory and statistics: A tutorial. *Foundations and Trends® in Communications and Information Theory*, 1(4):417–528, 2004.

Abraham De Moivre. *Miscellanea Analytica de Seriebus et Quadraturis.* J. Thonson and J. Watts, London, 1730.

Ganesh Del Grosso, Georg Pichler, Catuscia Palamidessi, and Pablo Piantanida. Bounding information leakage in machine learning. *Neurocomputing*, 534:1–17, 2023.

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The total variation distance between high-dimensional Gaussians with the same mean. *arXiv preprint arXiv:1810.08693*, 2018.

John C Duchi and Feng Ruan. The right complexity measure in locally private estimation: It is not the fisher information. *The Annals of Statistics*, 52(1):1–51, 2024.

John C Duchi, Michael I Jordan, and Martin J Wainwright. Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, 113(521): 182–201, 2018.

Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 371–380, 2009.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pages 265–284. Springer, 2006.

Cynthia Dwork, Guy N Rothblum, and Salil Vadhan. Boosting and differential privacy. In *2010 IEEE 51st annual symposium on foundations of computer science*, pages 51–60. IEEE, 2010.

Cynthia Dwork, Aaron Roth, et al. *The algorithmic foundations of differential privacy*, volume 9. Now Publishers, Inc., 2014.

Vitaly Feldman. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 954–959, 2020.

Elena Ghazi and Ibrahim Issa. Total variation meets differential privacy. *IEEE Journal on Selected Areas in Information Theory*, 2024.

Chuan Guo, Alexandre Sablayrolles, and Maziar Sanjabi. Analyzing Privacy Leakage in Machine Learning via Multiple Hypothesis Testing: A Lesson From Fano. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 11998–12011. PMLR, 23–29 Jul 2023. URL `https://proceedings.mlr.press/v202/guo23e.html`.

Umang Gupta, Dmitris Stripelis, Pradeep Lam, Paul M. Thompson, J. Ambite, and Greg Ver Steeg. Membership Inference Attacks on Deep Regression Models for Neuroimaging. In *International Conference on Medical Imaging with Deep Learning*, 2021. URL `https://api.semanticscholar.org/CorpusID:233864706`.

Awni Y. Hannun, Chuan Guo, and Laurens van der Maaten. Measuring Data Leakage in Machine-Learning Models with Fisher Information. In *Conference on Uncertainty in Artificial Intelligence*, 2021. URL `https://api.semanticscholar.org/CorpusID:232013768`.

John Hartley and Sotirios A Tsaftaris. Measuring unintended memorisation of unique private features in neural networks. *arXiv preprint arXiv:2202.08099*, 2022.

Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. Membership Inference Attacks Against Generative Models. 2018. URL `https://api.semanticscholar.org/CorpusID:202588705`.

Xinlei He, Zheng Li, Weilin Xu, Cory Cornelius, and Yang Zhang. Membership-Doctor: Comprehensive Assessment of Membership Inference Against Machine Learning Models. *arXiv preprint arXiv:2208.10445*, 2022.

Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S Yu, and Xuyun Zhang. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, 54(11s):1–37, 2022.

Zachary Izzo, Jinsung Yoon, Sercan O Arik, and James Zou. Provable Membership Inference Privacy. *arXiv preprint arXiv:2211.06582*, 2022.

Bargav Jayaraman and David Evans. Evaluating Differentially Private Machine Learning in Practice. In *Proceedings of the 28th USENIX Conference on Security Symposium*, SEC'19, page 1895–1912, USA, 2019. USENIX Association. ISBN 9781939133069.

Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. *Advances in Neural Information Processing Systems*, 33:17176–17186, 2020.

Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. In *International conference on machine learning*, pages 1376–1385. PMLR, 2015.

Peter Kairouz, Sewoong Oh, and Pramod Viswanath. Extremal mechanisms for local differential privacy. *Journal of Machine Learning Research*, 17(17):1–51, 2016.

Klas Leino and Matt Fredrikson. Stolen memories: Leveraging model memorization for calibrated {White-Box} membership inference. In *29th USENIX security symposium (USENIX Security 20)*, pages 1605–1622, 2020.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. *arXiv preprint arXiv:1906.05890*, 2019.

SK Murakonda and R Shokri. ML Privacy Meter: Aiding Regulatory Compliance by Quantifying the Privacy Risks of Machine Learning., 2007.

Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 739–753, 2019. doi: 10.1109/SP.2019.00065.

C.R. Rao, University of Pittsburgh. Institute for Statistics, and Applications. *Gini-Simpson Index of Diversity: A Characterization, Generalization and Applications*. Technical report (University of Pittsburgh. Institute for Statistics and Applications). University of Pittsburgh, 1981. URL `https://books.google.ca/books?id=d7laNQAACAAJ`.

C.Radhakrishna Rao. Diversity and dissimilarity coefficients: A unified approach. *Theoretical Population Biology*, 21(1):24–43, 1982. ISSN 0040-5809. doi: https://doi.org/10.1016/0040-5809(82)90004-1. URL `https://www.sciencedirect.com/science/article/pii/0040580982900041`.

Alfréd Rényi. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics*, volume 4, pages 547–562. University of California Press, 1961.

Shahbaz Rezaei and Xin Liu. On the difficulty of membership inference attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7892–7900, 2021.

Herbert Robbins. A remark on Stirling's formula. *Amer. Math. Monthly*, 62:26–29, 1955. ISSN 0002-9890,1930-0972. doi: 10.2307/2308012. URL https://doi.org/10.2307/2308012.

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Hervé Jégou. White-box vs black-box: Bayes optimal strategies for membership inference. In *International Conference on Machine Learning*, pages 5558–5567. PMLR, 2019.

Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. *arXiv preprint arXiv:1806.01246*, 2018.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.

Congzheng Song and Ananth Raghunathan. Information Leakage in Embedding Models. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, CCS '20, page 377–390, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370899. doi: 10.1145/3372297.3417270. URL https://doi.org/10.1145/3372297.3417270.

Congzheng Song, Thomas Ristenpart, and Vitaly Shmatikov. Machine Learning Models That Remember Too Much. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS '17, page 587–601, New York, NY, USA, 2017a. Association for Computing Machinery. ISBN 9781450349468. doi: 10.1145/3133956.3134077. URL https://doi.org/10.1145/3133956.3134077.

Congzheng Song, Thomas Ristenpart, and Vitaly Shmatikov. Machine learning models that remember too much. In *Proceedings of the 2017 ACM SIGSAC Conference on computer and communications security*, pages 587–601, 2017b.

Elham Tabassi, Kevin Burns, Michael Hadjimichael, Andres Molina-Markham, and Julian Sexton. A taxonomy and terminology of adversarial machine learning, 10 2019.

Aleena Anna Thomas, David Ifeoluwa Adelani, Ali Davody, Aditya Mogadala, and Dietrich Klakow. Investigating the Impact of Pre-trained Word Embeddings on Memorization in Neural Networks. In *Workshop on Time-Delay Systems*, 2020. URL https://api.semanticscholar.org/CorpusID:220658693.

Anvith Thudi, Ilia Shumailov, Franziska Boenisch, and Nicolas Papernot. Bounding membership inference. *arXiv preprint arXiv:2202.12232*, 2022.

Kushal Tirumala, Aram Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. Memorization without overfitting: Analyzing the training dynamics of large language models. *Advances in Neural Information Processing Systems*, 35:38274–38290, 2022.

Gal Vardi, Gilad Yehudai, and Ohad Shamir. Gradient methods provably converge to non-robust networks. *Advances in Neural Information Processing Systems*, 35:20921–20932, 2022.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pages 268–282. IEEE, 2018.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding Deep Learning (Still) Requires Rethinking Generalization. *Commun. ACM*, 64 (3):107–115, feb 2021. ISSN 0001-0782. doi: 10.1145/3446776. URL `https://doi.org/10.1145/3446776`.

Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. *Advances in neural information processing systems*, 32, 2019.

J. Ziv and M. Zakai. On Functionals Satisfying a Data-Processing Theorem. *IEEE Trans. Inf. Theor.*, 19(3):275–283, may 1973. ISSN 0018-9448. doi: 10.1109/TIT.1973.1055015. URL `https://doi.org/10.1109/TIT.1973.1055015`.