Fundamental Limits of Membership Inference Attacks on Machine Learning Models

Eric Aubinais, Elisabeth Gassiat, Pablo Piantanida

FUNDAMENTAL LIMITS OF MEMBERSHIP INFERENCE ATTACKS ON MACHINE LEARNING MODELS

Eric Aubinais & Elisabeth Gassiat
Université Paris-Saclay, CNRS
Laboratoire de mathématiques d'Orsay
Orsay, France
{eric.aubinais,elisabeth.gassiat}@universite-paris-saclay.fr

Pablo Piantanida
ILLS - International Laboratory on Learning Systems
MILA - Quebec AI Institute
CNRS, CentraleSupélec
Montreal, QC, Canada
pablo.piantanida@cnrs.fr

February 1, 2024

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. More precisely, we first derive the statistical quantity that governs the effectiveness and success of such attacks. We then deduce that in a very general regression setting with overfitting algorithms, attacks may have a high probability of success. Finally, we investigate several situations for which we provide bounds on this quantity of interest. Our results enable us to deduce the accuracy of potential attacks based on the number of samples and other structural parameters of learning models. In certain instances, these parameters can be directly estimated from the dataset.

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., 2023a]. Furthermore, a recent publication [Tabassi et al., 2019] from the National Institute of Standards and Technology (NIST) explicitly notes that an MIA that successfully identifies an individual as part of the dataset used for training the target model constitutes a breach of confidentiality.

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 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]. However, despite these empirical findings, there remains a significant gap in our theoretical understanding of this phenomenon. This article delves into the core statistical limitations surrounding MIAs on machine learning models at large.

Our investigation commences by establishing the fundamental statistical quantity that governs the effectiveness and success of these attacks. In the learning model we consider, our attention is directed towards **symmetric** algorithms that adhere to the **redundancy invariance property**, meaning that training on a dataset consisting of multiple repetitions of a dataset is equivalent to training on the same smaller dataset once. Specifically, we concentrate on datasets of independent and identically distributed (*i.i.d.*) samples. To assess the effectiveness of MIAs, we will gauge their **accuracy** by examining their success probability in determining membership. Notably, we assess the security of a model based on the highest level of accuracy achieved among MIAs. An MIA that attains the maximum accuracy will be referred to as an *oracle*.

We delve into the intricacies of MIA and derive insights into the key factors that influence its outcomes. Subsequently, we explore various scenarios: overfitting algorithms, empirical mean-based algorithms, discrete data and when the parameter space is quantized, among others, presenting bounds on this pivotal statistical quantity. These bounds provide crucial insights into the accuracy of potential attacks.

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 Crucial Statistical Quantity:** We introduce the critical statistical quantity denoted as $\Delta_n(P, \mathcal{A})$, where n represents the size of the training dataset, P is the data distribution, and $\mathcal{A}$ is the underlying algorithm. This quantity plays a pivotal role in assessing the accuracy of effective MIAs. The quantity $\Delta_n(P, \mathcal{A})$ 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 $\Delta_n(P, \mathcal{A})$ is small, the accuracy of the best MIA is notably constrained. Conversely, when $\Delta_n(P, \mathcal{A})$ approaches 1, the best MIA is successful with high probability. This highlights the importance of $\Delta_n(P, \mathcal{A})$ in characterizing information disclosure in relation to the training set.
- **Lower Bounds for Overfitting Algorithms** : For algorithms that overfit with high probability, we exhibit a lower bound on $\Delta_n(P, \mathcal{A})$, see Theorem 4.2. In a general regression setting, we further prove that whenever the algorithm overfits over its training set with probability at least $1 - \alpha_n \in (0, 1)$, the quantity $\Delta_n(P, \mathcal{A})$ is bounded from below by $1 - \alpha_n$, see Corollary 4.2.1. 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 Algorithms:** For algorithms that compute functions of empirical means, we establish precise upper bounds on $\Delta_n(P, \mathcal{A})$. We prove that $\Delta_n(P, \mathcal{A})$ 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 $\Delta_n(P, \mathcal{A})$ remains below ε for any $\varepsilon \in (0, 1)$.
- **Maximization of $\Delta_n(P, \mathcal{A})$:** In scenarios involving discrete data with an infinite parameter space, we provide a precise formula for maximizing $\Delta_n(P, \mathcal{A})$ across all algorithms $\mathcal{A}$. Additionally, when dealing with data that has a finite set of possible values, we determine that this maximization is proportional to a constant times $n^{-1/2}$. Furthermore, we reveal that there are distinct behaviors concerning the dependence on n when dealing with discrete data, which can include infinitely many values or when the parameter space is finite (e.g., machine learning models with quantized weights).

These contributions advance the theoretical understanding of MIAs on machine learning models, shedding light on the crucial role played by statistical quantities and their bounds in assessing the security and privacy of these models.

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. 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 $\Delta_n(P, \mathcal{A})$ 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., 2023b]. 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 $\mathbf{z} \coloneqq (\mathrm{z}_1, \cdots, \mathrm{z}_n) \in \mathcal{Z}^n$ of a test point $\tilde{\mathrm{z}} \in \mathcal{Z}$ from a predictor $\hat{\mu} = \mu_{\hat{\theta}_n}$ in a model $\mathcal{F} \coloneqq \{\mu_\theta : \theta \in \Theta\}$, where Θ is the space of parameters. The predictor is identified to its parameters $\hat{\theta}_n \in \Theta$ learned from $\mathbf{z}$ through an **algorithm** $\mathcal{A} : \bigcup_{k>0} \mathcal{Z}^k \to \mathcal{P}' \subseteq \mathcal{P}(\Theta)$, 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 algorithm $\hat{\theta}_n \coloneqq \mathcal{A}(\mathrm{z}_1, \cdots, \mathrm{z}_n)$.

We now consider MIAs as functions of the parameters and the test point whose outputs are 0 or 1.

Definition 2.1 (Membership Inference Attack - MIA)**.** *Any measurable map $\phi : \Theta \times \mathcal{Z} \to \{0, 1\}$ is called a **Membership Inference Attack**.*

We measure the accuracy of an MIA ϕ by its probability of successfully guessing the membership of the test point. For that purpose, we encode membership to the training data set as 1. We assume that $\mathrm{z}_1, \ldots, \mathrm{z}_n$ are independent and identically distributed (*i.i.d.*) random variables with distribution P. Following Del Grosso et al. [2023] or Sablayrolles et al. [2019] framework, we suppose that the test point $\tilde{\mathrm{z}}$ is to be drawn from P independently from the samples $\mathrm{z}_1, \ldots, \mathrm{z}_n$ with probability $\nu \in (0, 1)$. Otherwise, conditionally to $\mathbf{z}$, we set $\tilde{\mathrm{z}}$ to any z_j each with uniform probability $1/n$.

Letting U be a random variable with distribution $\hat{P}_n \coloneqq \frac{1}{n}\sum_{j=1}^n \delta_{\mathrm{z}_j}$ 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

$$\tilde{\mathrm{z}} \coloneqq T\mathrm{z}_0 + (1 - T)U.$$

Definition 2.2 (Accuracy of an MIA)**.** *The **accuracy of an MIA** ϕ is defined as*

$$\mathit{Acc}_n(\phi; P, \mathcal{A}) \coloneqq P\left(\phi(\hat{\theta}_n, \tilde{z}) = 1 - T\right), \tag{1}$$

where the probability is taken over all randomness.

The accuracy of an MIA scales from 0 to 1. Constant MIAs $\phi_0 \equiv 0$ and $\phi_1 \equiv 1$ have respectively an accuracy equal to ν and $1 - \nu$, which means that we always can build an MIA with accuracy of at least $\max(\nu, 1 - \nu)$ and any MIA performing worse than this quantity is irrelevant to use. We now define the **Membership Inference Security** of an algorithm as a quantity summarizing the amount of security of the system against MIAs.

Definition 2.3 (Membership Inference Security - MIS)**.** *Let $\nu_* \coloneqq \min(\nu, 1 - \nu)$. The Membership Inference Security of an algorithm $\mathcal{A}$ is*

$$\mathit{Sec}_n(P, \mathcal{A}) \coloneqq \nu_*^{-1}\left(1 - \sup_\phi \mathit{Acc}_n(\phi; P, \mathcal{A})\right), \tag{2}$$

where the sup is taken over all MIAs.

The Membership Inference Security scales from 0 (the best MIA approaches perfect guess of membership) to 1 (MIAs can not do better than ϕ_0 and ϕ_1).

Throughout this paper, we focus on algorithms that are **symmetric** and **redundancy invariant**. An algorithm is symmetric if it is invariant under permutation of its inputs. On the other hand, an algorithm is redundancy invariant if for any input dataset, the output of the algorithm on the dataset would be the same as if the dataset was repeated.

Definition 2.4 (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 2.5 (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*

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

The redundancy invariance property states that no information can be gathered from giving the same dataset multiple times.

3 Performance Assessment of Membership Inference Attacks

In this section, we prove that the **Crucial Statistical Quantity** for the assessment of the accuracy of membership inference attacks is $\Delta_n(P, \mathcal{A})$, defined as

$$\Delta_n(P, \mathcal{A}) \coloneqq \left\| \mathcal{L}\big((\hat{\theta}_n, \mathrm{z}_1)\big) - \mathcal{L}\big((\hat{\theta}_n, \mathrm{z}_0)\big) \right\|_{\mathrm{TV}}, \tag{3}$$

which depends on P, n and $\mathcal{A}$. Here, for any random variable x, $\mathcal{L}(\mathrm{x})$ denotes its probability distribution, and for any distributions Q_1 and Q_2, $\|Q_1 - Q_2\|_{\mathrm{TV}}$ denotes the total variation distance between Q_1 and Q_2. One can interpret $\Delta_n(P, \mathcal{A})$ as quantifying some stability of the algorithm. We first prove that symmetric and redundancy invariant algorithms can be characterized as functions of the empirical distribution $\hat{P}_n$ of the training dataset. All proofs of this section are deferred to Appendix A.
Let $\mathcal{M}$ be the set of all discrete distributions on $\mathcal{Z}$.

Proposition 3.1. *Let $f : \bigcup_{k>0} \mathcal{Z}^k \to \mathcal{Z}'$ be a measurable map onto any space $\mathcal{Z}'$. Then the followings 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 $f(z_1, \cdots, z_k) = G\left(\frac{1}{k}\sum_{j=1}^k \delta_{z_j}\right)$.

In particular, we may apply Proposition 3.1 to any algorithms $\mathcal{A}$ with $\mathcal{Z}' = \mathcal{P}'$.
Interestingly, if an algorithm minimizes an empirical cost, then it is of the latter kind. In particular, maximum likelihood based algorithms or Bayesian methods from Sablayrolles et al. [2019] are special cases.
Proposition 3.1 (ii) gives us the combinatorial form of the problem, allowing us to study thoroughly discrete cases, which Thudi et al. [2022] and Izzo et al. [2022] are special cases[1].

Theorem 3.2 (Key bound on accuracy)**.** *Suppose P is any distribution and $\mathcal{A}$ is any symmetric redundancy invariant algorithm. Then the accuracy of any MIA ϕ satisfies:*

$$\nu_* - \nu_* \Delta_n(P, \mathcal{A}) \leq Acc_n(\phi; P, \mathcal{A}) \leq 1 - \nu_* + \nu_* \Delta_n(P, \mathcal{A}).$$

In particular,

$$Sec_n(P, \mathcal{A}) \geq 1 - \Delta_n(P, \mathcal{A}).$$

Theorem 3.2 shows that an upper bound on $\Delta_n(P, \mathcal{A})$ translates into a lower bound for the MIS of any algorithm. When $\nu = 1/2$, $\Delta_n(P, \mathcal{A})$ is the quantity that controls the best possible accuracy of MIAs as the result below proves.

Theorem 3.3. *Suppose P is any distribution, $\mathcal{A}$ is any symmetric redundancy invariant algorithm and $\nu = 1/2$. Then*

$$Sec_n(P, \mathcal{A}) = 1 - \Delta_n(P, \mathcal{A}).$$

We see that $\Delta_n(P, \mathcal{A})$ appears to be a key mathematical quantity for assessing the accuracy of MIAs. We thus study in Section 4 a control on $\Delta_n(P, \mathcal{A})$ when the algorithm overfits, and in Section 5 situations in which we are able to give precise controls on $\Delta_n(P, \mathcal{A})$.

Notice that there is no assumption on the data distribution P. For instance, we can take into account outliers by making P a mixture.

4 Overfitting Causes Lack of Security

In this section, we assume that $\mathcal{Z} \coloneqq \mathcal{X} \times \mathcal{Y}$ and that the algorithm $\mathcal{A}$ produces overfitting parameters $\hat{\theta}_n$. We then note $\mathrm{z} \coloneqq (\mathrm{x}, \mathrm{y})$. We consider learning systems minimizing $\theta \mapsto \sum_{i=1}^n l_\theta(x_i, y_i)$ where $l_\theta : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$ is a loss function.

Definition 4.1 (Overfitting)**.** *We say that an algorithm $\mathcal{A}$ is overfitting over a training set $(\mathrm{z}_1, \cdots, \mathrm{z}_n)$ with probability $(1 - \alpha_n) \in (0, 1)$ when*

$$P(l_{\hat{\theta}_n}(\mathrm{x}_1, \mathrm{y}_1) = 0) \geq 1 - \alpha_n. \tag{4}$$

[1] Both articles treat almost the same context through the DP prism. They both suppose a finite set $\mathcal{D}$ corresponding to the support of the distribution, from which they draw the training set.

When $\alpha_n = 0$, and when the algorithm $\mathcal{A}$ is symmetric, Equation 4 is equivalent to having $l_{\hat{\theta}_n}(\mathrm{x}_i, \mathrm{y}_i) = 0$ for all $i = 1, \ldots, n$. Also, note that α_n depends implicitly on the data distribution P and the algorithm $\mathcal{A}$.

We then get the following result :

Theorem 4.2 (Overfitting induces lack of security)**.** *Assume Equation 4 holds. Let $S_\theta := \{(x, y) \in \mathcal{X} \times \mathcal{Y} : l_\theta(x, y) = 0\}$ be the zero set of l_θ for all $\theta \in \Theta$. Then we have*

$$\Delta_n(P, \mathcal{A}) \geq 1 - \alpha_n - \int_\Theta P((\mathrm{x}, \mathrm{y}) \in S_\theta) d\mu_{\hat{\theta}_n}, \tag{5}$$

where $\mu_{\hat{\theta}_n}$ is the distribution of $\hat{\theta}_n$.

This theorem establishes an upper bound on the security under overfitting. The proof is given below.

Proof of Theorem 4.2. Let $S \coloneqq \{(\theta, x, y) : l_\theta(x, y) = 0\}$. From the definition of $\Delta_n(P, \mathcal{A})$, we have that

$$\begin{aligned}
\Delta_n(P, \mathcal{A}) &\geq P((\hat{\theta}_n, \mathrm{x}_1, \mathrm{y}_1) \in S) - P((\hat{\theta}_n, \mathrm{x}, \mathrm{y}) \in S) \\
&= P((\mathrm{x}_1, \mathrm{y}_1) \in S_{\hat{\theta}_n}) - P((\mathrm{x}, \mathrm{y}) \in S_{\hat{\theta}_n}) \\
&= 1 - \alpha - P((\mathrm{x}, \mathrm{y}) \in S_{\hat{\theta}_n}) \\
&= 1 - \alpha - \int_{\theta \in \Theta} P((\mathrm{x}, \mathrm{y}) \in S_\theta) d\mu_{\hat{\theta}_n}.
\end{aligned}$$

□

We now focus on the case where $\mathcal{Y} \coloneqq \mathbb{R}$ and there exists a family of functions $\Psi_\theta : \mathcal{X} \to \mathbb{R}$ such that for all $\theta \in \Theta, x \in \mathcal{X}$ and $y \in \mathbb{R}$, we have

$$l_\theta(x, y) = 0 \iff y = \Psi_\theta(x). \tag{6}$$

Equation 6 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 $\Psi_\theta(\mathrm{x})$ and y. For instance, this is the case when we consider a regression setting minimizing least squares error.

Corollary 4.2.1. *Assume Equations 4 and 6 hold. Assume moreover that a version of the conditional distribution of* y *given* x *has no atom. Then,*

$$\Delta_n(P, \mathcal{A}) \geq 1 - \alpha_n. \tag{7}$$

Corollary 4.2.1 states that for regressors overfitting with high probability, there possibly exists an MIA with high probability success. We now give examples for which Corollary 4.2.1 allows us to give a lower bound on $\Delta_n(P, \mathcal{A})$.

Example 4.3 (Linear regression)**.** *Let* x *be a random variable of distribution P_{x} taking values in $\mathbb{R}^d$, and* w *be a random variable independent of* x *whose distribution is absolutely continuous with respect to the Lebesgue measure. Let* $\mathrm{y} \coloneqq \beta^T \mathrm{x} + \mathrm{w}$ *for some fixed $\beta \in \mathbb{R}^d$. We seek to estimate β by minimizing least squares error, for which Equation 6 holds with $\Psi_\theta(x) = \theta^T x$. Assume we have access to data samples $((\mathrm{x}_1, \mathrm{y}_1), \cdots, (\mathrm{x}_n, \mathrm{y}_n))$ drawn independently from the joint distribution of* (x, y)*, with $d > n$. Then, Equation 4 holds with $\alpha_n = 0$, see Frei et al., with any regularization. Then, the assumptions of Corollary 4.2.1 are satisfied, leading to $\Delta_n(P, \mathcal{A}) = 1$.*

Example 4.4 (Neural Networks)**.** *It is worth mentioning that Ψ_θ may be a neural network with parameters $\theta \in \mathbb{R}^d$.*

Proof of Corollary 4.2.1. It suffices to prove from Theorem 4.2 that for any $\theta \in \Theta$, we have $P((\mathrm{x}, \mathrm{y}) \in S_\theta) = 0$. Let $\theta \in \Theta$, then we have

$$\begin{aligned}
P((\mathrm{x}, \mathrm{y}) \in S_\theta) &= \mathbb{E}\left[P((\mathrm{x}, \mathrm{y}) \in S_\theta \mid \mathrm{x})\right] \\
&= \mathbb{E}\left[P(y = \Psi_\theta(\mathrm{x}) \mid \mathrm{x})\right] \\
&= 0,
\end{aligned}$$

where the second equality comes from Equation 6, and the last equality comes from the hypothesis that the conditional distribution of y given x has no atom. □

5 Security is Data Size Dependent

In this section, we study the converse, where we aim at understanding when to expect $\Delta_n(P, \mathcal{A})$ to be close to 0. All the proofs of the section can be found in Appendix B.

5.1 Empirical Mean based algorithms

We first study the case of algorithms 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 the algorithm

$$\mathcal{A} : (z_1, \cdots, z_n) \mapsto \delta_{F\left(\frac{1}{n}\sum_{j=1}^n L(z_j)\right)}, \tag{8}$$

where δ_θ stands for the Dirac mass at θ.
Without loss of generality, we may assume that

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

Let $m_j \coloneqq \mathbb{E}\left[\left\|C^{-1/2}\Big\{L(\mathrm{z}_1) - \mathbb{E}\left[L(\mathrm{z}_1)\right]\Big\}\right\|_2^j\right]$ for any positive integer j, that is the expectation of the j-th power of the norm of the centered and reduced version of $L(\mathrm{z}_1)$, and C be the covariance matrix of $L(\mathrm{z}_1)$.

Theorem 5.1. *Suppose that the distribution of $L(\mathrm{z}_1)$ has a non zero absolutely continuous part with respect to the Lebesgue measure, and suppose $m_3 < \infty$. Then*

$$\Delta_n(P, \mathcal{A}) \leq \left(C(d)(1 + m_3) + \frac{m_1}{2}\right) n^{-1/2} + \frac{\sqrt{d}}{2n}, \tag{9}$$

for some constant $C(d)$ depending only on the dimension d of $L(\mathrm{z}_1)$.

Remark 5.1 : Theorem 5.1 implies that for distributions P satisfying the hypotheses, for any positive ϵ, for any algorithm $\mathcal{A}$ that can be expressed as a function of empirical means, $\mathrm{Sec}_n(P, \mathcal{A})$ can be made larger than $1 - \epsilon$ as soon as $\Delta_n(P, \mathcal{A}) \leq \epsilon$, which holds as soon as $n \geq \Omega(\varepsilon^{-2})$ is sufficient to ensure a security of at least $1 - \varepsilon$, where the hidden constant depend on the data distribution P and the parameters dimension d. See Appendix B for a proof.

We now provide examples for which Theorem 5.1 allows us to give an upper bound on $\Delta_n(P, \mathcal{A})$.

Example 5.2 (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*

$$h(\theta_0) = \mathbb{E}[\psi(\mathrm{z})]. \tag{10}$$

where z is a random variable of distribution P. Having access to data samples $\mathrm{z}_1, \ldots, \mathrm{z}_n$ drawn independently from the distribution P, we estimate $\mathbb{E}[\psi(\mathrm{z})]$ by $\frac{1}{n}\sum_{j=1}^n \psi(\mathrm{z}_j)$. The estimate $\hat{\theta}_n$ of θ_0 is then set to be the solution–provided that it exists–to the equation:

$$h(\hat{\theta}_n) = \frac{1}{n}\sum_{j=1}^n \psi(\mathrm{z}_j).$$

If the solution exists and h is invertible, one can set $\hat{\theta}_n = h^{-1}\left(\frac{1}{n}\sum_{j=1}^n \psi(\mathrm{z}_j)\right)$.

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 5.1 to any estimators obtained by solving equations.

Example 5.3 (Linear Regression)**.** *We consider here the same framework as in Example 4.3, where $d < n$ (hence the overfitting assumption can not be fulfilled with $\alpha_n = 0$). Let $\mathbb{X}$ be the $d \times n$ matrix whose i^{th} row is x_i, and $\mathbb{Y}$ be the column vector $(\mathrm{y}_1, \cdots, \mathrm{y}_n)^T$. 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 (8), 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 we can express the estimator as*

$$\hat{\beta}_n = F\left(\frac{1}{n}\sum_{j=1}^{n} L((\mathrm{x}_j, \mathrm{y}_j))\right).$$

We then may apply Theorem 5.1 to the least squares estimator for Linear Regression.

Interestingly, we see from Examples 4.3 and 5.3 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 $P \coloneqq \sum_{j=1}^{K} p_j \delta_{u_j}$ 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 $p_j > 0$ for all $j \in \{1, \cdots, K\}$.

Theorem 5.4. *For* $j = 1, \ldots, K$*, let* B_j *be a random variable having Binomial distribution with parameters* (n, p_j)*. Then,*

$$\max_{\mathcal{A}} \Delta_n(P, \mathcal{A}) = \frac{1}{2}\sum_{j=1}^{K} \mathbb{E}\left[\left|\frac{B_j}{n} - p_j\right|\right], \tag{11}$$

where the max *is taken over all symmetric and redundancy invariant algorithms and is reached on algorithms 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 5.4 allows to get upper bounds on $\Delta_n(P, \mathcal{A})$ for any algorithm $\mathcal{A}$, and consequently lower bounds on $\mathrm{Sec}_n(P, \mathcal{A})$ for any algorithm $\mathcal{A}$.

We now propose an analysis of the r.h.s. of (11). Define $C(P) \coloneqq \sum_{j=1}^{K} \sqrt{p_j(1-p_j)}$. We first give a general upper bound which is meaningful as soon as $C(P) < \infty$.

Corollary 5.4.1. *In all cases (*K *finite or infinite), if* $C(P) < \infty$*, then*

$$\max_{\mathcal{A}} \Delta_n(P, \mathcal{A}) \leq \frac{C(P)}{2} n^{-1/2}.$$

If $K = \infty$ *and* $C(P) = \infty$*,* $\max_{\mathcal{A}} \Delta_n(P, \mathcal{A})$ *still tends to* 0 *as* n *tends to infinity, but the (depending on* P*) rate can be arbitrarily slow.*

Corollary 5.4.1 implies that for distributions P such that $C(P) < \infty$, for any positive ϵ, for any algorithm $\mathcal{A}$, $\mathrm{Sec}_n(P, \mathcal{A})$ can be made larger than $1 - \epsilon$ as soon as the data set contains more than $(C(P)/2\epsilon)^2$ points. Notice that $C(P)$ can be estimated using the dataset also. However, for distributions with $C(P) = \infty$, finding the amount of data needed to get the same control on $\mathrm{Sec}_n(P, \mathcal{A})$ requires the estimation of the r.h.s. of (11) which is not obvious.

Notice that when K is finite, $C(P)$ is also finite. We now prove that in this case, the way $\Delta_n(P, \mathcal{A})$ depends in the number of data is indeed $n^{-1/2}$.

Corollary 5.4.2. *Suppose* $K < \infty$*. For all* $n \geq 2$*,*

$$\max_{\mathcal{A}} \Delta_n(P, \mathcal{A}) \geq \frac{\exp(-13/6)}{\sqrt{2\pi}}(C(P) - 1/\sqrt{2n})n^{-1/2}.$$

When the algorithm deterministically maps $\mathbf{z}$ to some element of Θ, that is $\mathcal{A}$ can be written as $\mathcal{A}(z_1, \cdots, z_n) = \delta_{F(\frac{1}{n}\sum_{j=1}^{n}\delta_{z_j})}$ for some map F, and the support Θ of the image distribution has finite cardinal $L \in \mathbb{N}$, it is not always possible to construct functions F that are injective. In this case, one may rewrite Theorem 5.4 as follows.

Lemma 5.5. *Let* $\mathbf{r} \coloneqq (N_1, \cdots, N_K)$ *be a random vector having multinomial distribution with parameters* $(n; p_1, \cdots, p_K)$*. There exists a partition* $(D_l)_{l=1\cdots L}$ *of the support of* $\mathbf{r}$ *such that*

$$\Delta_n(P, \mathcal{A}) = \frac{1}{2}\sum_{j=1}^{K}\sum_{l=1}^{L}\left|\mathbb{E}\left[\left\{p_j - \frac{N_j}{n}\right\} 1\{\mathbf{r} \in D_l\}\right]\right|.$$

Lemma 5.5 is a tool to understand the behaviour of $\Delta_n(P, \mathcal{A})$ depending on the structure of the algorithm $\mathcal{A}$. Although there is a strong similarity between Lemma 5.5 and Theorem 5.4, the value of $\Delta_n(P, \mathcal{A})$ in Lemma 5.5 is smaller than the right hand side of Equation 11. This could informally mean that discretizing/quantizing an algorithm improves its security.

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

Lemma 5.6. *Let P be the Bernoulli distribution with parameter $p \in (0, 1)$ and let $\hat{\theta}_n \coloneqq \sup_j \mathrm{z}_j$. Then,*

$$\Delta_n(P, \mathcal{A}) = 2p(1-p)^n.$$

From Lemma 5.6 and the second part of Corollary 5.4.1, one see that in the case of algorithms with finite support output, $\Delta_n(P, \mathcal{A})$ may have many different behaviours.

6 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 overfitting indeed induces the possibility of highly successful attacks. We further revealed a sufficient condition on the number of data samples to ensure control of the security of the learning system, when dealing with discrete data distributions or functionals of empirical means. For the latter scenario, we established that the rates of convergence consistently follow an order of $n^{-1/2}$, at which information acquired by attackers becomes irrelevant. 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.

Limitations and perspectives for future work. In Section 5, our work is currently limited to the discrete case and empirical mean based algorithms. We intend to extend further our research, and specifically to flow of empirical means, resulting to the complete study of maximum likelihood estimation, empirical loss minimization and Stochastic Gradient Descent.

In the presence of overfitting, we have established a lower bound of $1 - \alpha_n$. A precise control of α_n would require a thorough understanding of overfitting.

Moreover, our findings do not straightforwardly generalize to classification algorithms. We anticipate continuing our research in this direction to gain a comprehensive understanding of the impact of overfitting. We have specific plans to broaden the scope of our results to include classification algorithms. Currently, we can establish a similar result for $2-$ReLU binary classification networks, albeit under very stringent assumptions. Specifically, we study the case when the data are concentrated on the sphere of radius $\sqrt{s}$ in $\mathbb{R}^s$ and in a high-dimensional setting $s \geq n$. When the algorithm outputs a classifier whose parameters are in the direction of the gradient flow minimizing the exponential loss or the logistic loss, we prove that $\Delta_n(P, \mathcal{A})$ is lower bounded by the probability of the data to be not far from orthogonality, see Appendix C Proposition C.1. We anticipate that these assumptions may be relaxed in future investigations.

References

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

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

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, 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.

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.

Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances*

in Neural Information Processing Systems, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/60a6c4002cc7b29142def88715312

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, 2023a.

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

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.

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.

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.

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.

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.

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

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.

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

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.

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.

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.

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.

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.

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.

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

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.

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.

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.

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.

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, 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.

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.

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.

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.

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

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.

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.

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*, 2023b. URL https://openreview.net/forum?id=TatRHT_1cK.

Spencer Frei, Vidya Muthukumar, Fanny Yang, and Daniel Hsu. Reconsidering Overfitting in the Age of Overparametrized Models.

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.

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.

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.

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.

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

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.

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.

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

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

A Proofs of Section 3

Proof of Proposition 3.1. 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. ☐

Proof of Theorem 3.2 and Theorem 3.3. From the law of total probability, we have

$$\begin{aligned}
\text{Acc}_n(\phi; P, \mathcal{A}) &= P(\phi(\hat{\theta}_n, \tilde{z}) = 1 - T) \\
&= P(T=1)P\left(\phi(\hat{\theta}_n, \tilde{z}) = 1 - T | T = 1\right) + P(T=0)P\left(\phi(\hat{\theta}_n, \tilde{z}) = 1 - T | T = 0\right) \\
&= \nu P\left(\phi(\hat{\theta}_n, \mathrm{z}_0) = 0\right) + (1-\nu)P\left(\phi(\hat{\theta}_n, \mathrm{z}_1) = 1\right),
\end{aligned}$$

where the third equality comes from the definition of $\tilde{\mathrm{z}}$ and T. We now define $B \coloneqq \{(\theta, z) \in \Theta \times \mathcal{Z} : \phi(\theta, z) = 1\}$ and rewrite $\text{Acc}_n(\phi; P, \mathcal{A})$ as

$$\text{Acc}_n(\phi; P, \mathcal{A}) = \nu\left(1 - P\left((\hat{\theta}_n, \mathrm{z}_0) \in B\right)\right) + (1-\nu)P\left((\hat{\theta}_n, \mathrm{z}_1) \in B\right). \tag{12}$$

Taking the maximum over all MIAs ϕ then reduces to taking the maximum of the r.h.s. of Equation (12) over all measurable sets B. Setting $\gamma \coloneqq \frac{\nu}{1-\nu}$, we then get

$$\max_{\phi} \text{Acc}_n(\phi; P, \mathcal{A}) = (1-\nu)\max_{B}\Big[P\left((\hat{\theta}_n, \mathrm{z}_0) \in B\right) - \gamma P\left((\hat{\theta}_n, \mathrm{z}_1) \in B\right)\Big] + \nu, \tag{13}$$

where the maximum is taken over all measurable sets B. Let now ζ be a dominating measure of the distributions of $(\hat{\theta}_n, \mathrm{z}_0)$ and $(\hat{\theta}_n, \mathrm{z}_1)$ (for instance their average). We denote by p (resp. q) the density of the distribution of $(\hat{\theta}_n, \mathrm{z}_0)$ (resp. $(\hat{\theta}_n, \mathrm{z}_1)$) with respect to ζ. Then, the involved maximum in the r.h.s. of Equation (13) is reached on the set

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

The maximum being taken over all measurable sets in Equation (13), we may consider replacing B by its complementary B^c in the expression giving

$$\max_{\phi} \text{Acc}_n(\phi; P, \mathcal{A}) = (1-\nu)\max_{B}\left[\gamma P\left((\hat{\theta}_n, \mathrm{z}_1) \in B\right) - P\left((\hat{\theta}_n, \mathrm{z}_0) \in B\right)\right] + (1-\nu), \tag{14}$$

where in this case the maximum is reached on the set

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

Taking the average on Equations (13) and (14), we get

$$\max_{\phi} \mathrm{Acc}_n(\phi; P, \mathcal{A}) = \frac{1}{2} + \frac{1}{2}\int \Big|(1-\nu)p - \nu q\Big| d\zeta. \tag{15}$$

By the triangular inequality, we may obtain the two following inequalities:

$$\max_{\phi} \mathrm{Acc}_n(\phi; P, \mathcal{A}) \leq \frac{1}{2} + \frac{|1-2\nu|}{2}\int q d\zeta + \frac{1-\nu}{2}\int \Big|p - q\Big| d\zeta,$$
$$\max_{\phi} \mathrm{Acc}_n(\phi; P, \mathcal{A}) \leq \frac{1}{2} + \frac{|1-2\nu|}{2}\int p d\zeta + \frac{\nu}{2}\int \Big|p - q\Big| d\zeta.$$

With $\int q d\zeta = \int p d\zeta = 1$, it holds that when $\nu \leq 1/2$, we have $1 - 2\nu \geq 0$ so that $1/2 + |1-2\nu|/2 = 1 - \nu$. Similarly, we get $1/2 + |1-2\nu|/2 = \nu$ when $\nu \geq 1/2$. Then, setting $\nu_* \coloneqq \min\{\nu, 1-\nu\}$ we have in both cases

$$1/2 + |1-2\nu|/2 = 1 - \nu_*.$$

Since $\Delta_n(P, \mathcal{A}) = (1/2)\int |p - q| d\zeta$ from the definition of the total variation distance, by taking the minimum over the two previous expressions, we have

$$\max_{\phi} \mathrm{Acc}_n(\phi; P, \mathcal{A}) \leq 1 - \nu_* + \nu_* \Delta_n(P, \mathcal{A}),$$

from which we deduce

$$\mathrm{Sec}_n(P, \mathcal{A}) \geq 1 - \Delta_n(P, \mathcal{A}).$$

Following the same steps for the minimum, we have

$$\min_{\phi} \mathrm{Acc}_n(\phi; P, \mathcal{A}) \geq \nu_* - \nu_* \Delta_n(P, \mathcal{A}),$$

hence Theorem 3.2. Theorem 3.3 comes from plugging $\nu = 1/2$ in Equation (15).

□

B Proofs of Section 5

Proof of Theorem 5.1. Setting $L_n \coloneqq \frac{1}{n}\sum_{j=1}^n L(\mathrm{z}_j)$, 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

$$\|\mathcal{L}(g(L_n, \mathrm{z}_1)) - \mathcal{L}(g(L_n, \mathrm{z}_0))\|_{\mathrm{TV}} \leq \|\mathcal{L}((L_n, \mathrm{z}_1)) - \mathcal{L}((L_n, \mathrm{z}_0))\|_{\mathrm{TV}}.$$

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

$$\Delta_n(P, \mathcal{A}) \leq \|\mathcal{L}((L_n, \mathrm{z}_1)) - \mathcal{L}((L_n, \mathrm{z}_0))\|_{\mathrm{TV}} = \mathbb{E}\left[\|\mathcal{L}(L_n \mid \mathrm{z}_1) - \mathcal{L}(L_n)\|_{\mathrm{TV}}\right],$$

in which the expectation is taken over z_1.
For $j = 1, \ldots, n$, denote by $\mathrm{v}_j \coloneqq C^{-1/2}(L(\mathrm{z}_j) - \mathbb{E}[L(\mathrm{z}_j)])$ the centered and reduced version of $L(\mathrm{z}_j)$. The total variation distance being invariant by translation and rescaling, we shall write

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

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

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

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 B.1. *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 B.1 to the last quantity, it holds that

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

and the result follows from taking the expectation. □

Proof of Proposition B.1. Applying Proposition 2.1 of Devroye et al. [2018], it holds almost surely that

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

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 of Remark 5.1. Setting $c \coloneqq C(d)(1+m_3)+\frac{m_1}{2}$, from Equation 9, we have that

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

is sufficient to ensure $\Delta_n(P,\mathcal{A})\leq\varepsilon$, hence a security of at least $1-\varepsilon$.

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 $0\leq\bar{u}\leq u\coloneqq\frac{\varepsilon\sqrt{d}}{c^2}$ 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.

□

In all proofs below, a symmetric and redundancy invariant algorithm $\mathcal{A}$ is identified with the function G given by Proposition 3.1. We shall also use the fact that

$$\Delta_n(P,\mathcal{A})=\mathbb{E}\left[\left\|\mathcal{L}(\hat{\theta}_n)-\mathcal{L}(\hat{\theta}_n|\mathrm{z}_1)\right\|_{\mathrm{TV}}\right], \tag{16}$$

where $\mathcal{L}(\hat{\theta}_n|\mathrm{z}_1)$ is the distribution of $\hat{\theta}_n$ conditional to z_1, and the expectation is taken on the random variable z_1.

Proof of Theorem 5.4. The proof will be divided in two steps. First, we will prove the inequality

$$\Delta_n(P,\mathcal{A})\leq\frac{1}{2}\sum_{j=1}^{K}\mathbb{E}\left[\left|\frac{B_j}{n}-p_j\right|\right], \tag{17}$$

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

Lemma B.2. *For $j=1,\dots,K$, let B_j be random variables having Binomial distribution with parameters (n,p_j). 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 $z_1,\dots,z_n\in\mathcal{Z}$, for some measurable map $F:\mathcal{M}\to\Theta$ with infinite range $|\Theta|=\infty$, i.e. $\hat{\theta}_n\overset{\mathcal{L}}{=}F\left(\frac{1}{n}\sum_{j=1}^{n}\delta_{\mathrm{z}_j}\right)$. Then we have*

$$\max_F\Delta_n(P,\mathcal{A})=\frac{1}{2}\sum_{j=1}^{K}\mathbb{E}\left[\left|\frac{B_j}{n}-p_j\right|\right].$$

Theorem 5.4 will simply follow from Lemma B.2 and Equation (17).

Let us first prove Equation (17).
Since $\hat{\theta}_n$ has distribution $G(\hat{P}_n)$ conditionally on $\mathbf{z}$, where $\hat{P}_n \coloneqq \frac{1}{n}\sum_{j=1}^n \delta_{\mathrm{z}_j}$ is the empirical distribution of the data set, from Proposition 3.1, we have

$$\begin{aligned} P(\hat{\theta}_n \in B) &= \mathbb{E}[P(\hat{\theta}_n \in B|\mathbf{z})] \\ &= \mathbb{E}[G(\hat{P}_n)(B)] \end{aligned} \tag{18}$$

$$P(\hat{\theta}_n \in B|\mathrm{z}_1) = \mathbb{E}[G(\hat{P}_n)(B)|\mathrm{z}_1], \tag{19}$$

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 $\hat{P}_n^k \coloneqq \frac{1}{n}\left(\delta_{u_k} + \sum_{j=2}^n \delta_{\mathrm{z}_j}\right)$.
Using (16), (18) and (19) we may rewrite $\Delta_n(P, \mathcal{A})$ as

$$\Delta_n(P, \mathcal{A}) = \sum_{k=1}^K p_k \sup_B \left(\mathbb{E}[G(\hat{P}_n)(B)] - \mathbb{E}[G(\hat{P}_n^k)(B)]\right). \tag{20}$$

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 (20) gives

$$\Delta_n(P, \mathcal{A}) = \sum_{k=1}^K p_k \sup_B \left(\sum_{g\in\mathcal{G}_n} g(B) P(G(\hat{P}_n) = g) - \sum_{g\in\mathcal{G}_n} g(B) P(G(\hat{P}_n^k) = g)\right). \tag{21}$$

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 $g \in \mathcal{G}_n$,

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

so that summing over all g gives

$$\Delta_n(P, \mathcal{A}) = \sum_{k=1}^K p_k \sup_B \left(\sum_{Q\in\mathcal{M}_n} G(Q)(B)\left[P(\hat{P}_n = Q) - P(\hat{P}_n^k = Q)\right]\right), \tag{22}$$

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,

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

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

$$P(\hat{P}_n = Q) = \binom{n}{k_1, \cdots, k_K} \prod_{j=1}^K {p_j}^{k_j}, \tag{23}$$

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 (23) can be understood to hold also when $K = +\infty$ by keeping only the terms involving the positive integers k_j.

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

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

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

$$
P(\hat{P}_n^1 = Q) = \frac{k_1}{n p_1} \binom{n}{k_1, \cdots, k_K} \prod_{j=1}^{K} p_j^{k_j}. \tag{24}
$$

Now, using (23) and (24), denoting by g_N the image by G of the distribution determined by the $K-$tuple $N = (k_1, \cdots, k_K)$, we get

$$
\begin{aligned}
\sum_{Q \in \mathcal{M}_n} G(Q)(B) \left(P(\hat{P}_n = Q) - P(\hat{P}_n^1 = Q) \right) &= \sum_{k_1 + \cdots + k_K = n} g_N(B) \binom{n}{k_1, \cdots, k_K} \prod_{j=1}^{K} p_j^{k_j} \left(1 - \frac{k_1}{n p_1} \right) \\
&= \mathbb{E}\left[\left(1 - \frac{N_1}{n p_1} \right) g_N(B) \right],
\end{aligned}
$$

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 $k = 1, \dots, K$, we easily obtain that for any $k = 1, \dots, K$

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

Now, plugging it into Equation (22) gives

$$
\Delta_n(P, \mathcal{A}) = \sum_{k=1}^{K} p_k \sup_B \mathbb{E}\left[\left(1 - \frac{N_k}{n p_k} \right) g_N(B) \right]. \tag{25}
$$

For any real number $x \in \mathbb{R}$, we denote by $(x)_+ = \max(x, 0)$ its positive part and $(x)_- = \max(0, -x)$ its negative part. We get from Equation (25)

$$
\begin{aligned}
\Delta_n(P, \mathcal{A}) &\leq \sum_{k=1}^{K} p_k \mathbb{E}\left[\sup_B \left(1 - \frac{N_k}{n p_k} \right) g_N(B) \right] \\
&= \sum_{k=1}^{K} p_k \mathbb{E}\left[\left(1 - \frac{N_k}{n p_k} \right)_+ \right] \qquad (26) \\
\Delta_n(P, \mathcal{A}) &\leq \sum_{k=1}^{K} p_k \mathbb{E}\left[\left(1 - \frac{N_k}{n p_k} \right)_- \right] \qquad (27)
\end{aligned}
$$

where the equality in Equation (26) comes from the fact that the supremum is reached on null sets when $1 - N_k/np_k$ is negative, and on sets of mass 1 when it is positive. Equation (27) is obtained by replacing B by its complementary B^c in the supremum and remarking that $\mathbb{E}[1 - N_k/np_k] = 0$. Combining Equations (26) and (27) gives

$$
\begin{aligned}
\Delta_n(P; \mathcal{A}) &\leq \sum_{k=1}^{K} p_k \min\left\{ \mathbb{E}\left[\left(1 - \frac{N_k}{n p_k} \right)_+ \right]; \mathbb{E}\left[\left(1 - \frac{N_k}{n p_k} \right)_- \right] \right\} \\
&\leq \frac{1}{2} \sum_{k=1}^{K} \mathbb{E}\left[\left| 1 - \frac{N_k}{n p_k} \right| \right],
\end{aligned}
$$

which proves Equation (17).

□

Proof of Lemma B.2. 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 (16), and following similar steps as in Equations (20), (21) and (22), by triangular inequality, we get that

$$\begin{aligned}
\Delta_n(P, \mathcal{A}) &= \sum_{k=1}^{K} \frac{p_k}{2} \sum_{g \in \mathcal{G}_n} \left| P(\delta_{F(\hat{P}_n)} = g) - P(\delta_{F(\hat{P}_n^k)} = g) \right| \\
&= \sum_{k=1}^{K} \frac{p_k}{2} \sum_{\theta \in \Theta} \left| P(F(\hat{P}_n) = \theta) - P(F(\hat{P}_n^k) = \theta) \right| \\
&= \sum_{k=1}^{K} \frac{p_k}{2} \sum_{\theta \in \Theta} \left| \sum_{Q \in \mathcal{M}_n(\theta)} \Big(P(\hat{P}_n = Q) - P(\hat{P}_n^k = Q) \Big) \right| \\
&\leq \sum_{k=1}^{K} \frac{p_k}{2} \sum_{Q \in \mathcal{M}_n} \left| P(\hat{P}_n = Q) - P(\hat{P}_n^k = Q) \right|,
\end{aligned} \tag{28}$$

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 emptyset or a singleton. Thus, Equation (28) gives

$$\max_F \Delta_n(P, \mathcal{A}) = \sum_{k=1}^{K} \frac{p_k}{2} \sum_{Q \in \mathcal{M}_n} \left| P(\hat{P}_n = Q) - P(\hat{P}_n^k = Q) \right|,$$

and the lemma follows from Equations (23) and (24) and the same steps as in the proof of Theorem 5.4.

□

Proof of Corollary 5.4.1. The upper bound comes from Cauchy-Schwartz's inequality and Theorem 5.4 since for any $j = 1, \ldots, K$,

$$\mathbb{E}\left[\left| \frac{B_j}{n} - p_j \right| \right] \leq \sqrt{Var(B_j/n)} = \sqrt{p_j(1-p_j)} n^{-1/2}.$$

When $C(P) = \infty$, this bound is trivial. Invoking Lemmas 7 and 8 of Berend and Kontorovich [2013], we get that $\max_{\mathcal{A}} \Delta_n(P, \mathcal{A})$ tends to 0 when n tends to infinity, with any possible rate depending on P. □

Proof of Corollary 5.4.2. Define $m_k \coloneqq \lfloor np_k \rfloor$, $k = 1, \ldots, K$. Using Theorem 5.4 , and De Moivre [1730], it holds that

$$\max_{\mathcal{A}} \Delta_n(P, \mathcal{A}) = \frac{1}{n} \sum_{k=1}^{K} \binom{n}{m_k + 1} (m_k + 1) p_k^{m_k+1} (1 - p_k)^{n - m_k}, \tag{29}$$

where $\binom{n}{m_k+1} = \frac{n!}{(m_k+1)!(n-(m_k+1))!}$ is a binomial coefficient. We shall approximate this binomial coefficient by Robbins [1955], which states that for any integer $k \geq 1$, it holds that

$$\sqrt{2\pi} k^{k+1/2} e^{-k} e^{1/12(k+1)} < k! < \sqrt{2\pi} k^{k+1/2} e^{-k} e^{1/12k}. \tag{30}$$

Note first that for any $k = 1, \cdots, K$, it holds that $m_k + 1 \geq 1$. For any $k = 1 \cdots K$, we have $n - (m_k + 1) \geq 1$ if and only if $n > 1/(1 - p_k)$.

When $n > 1/(1-p_k)$, we set $a_k := \exp\left(\frac{1}{12n+1} - \frac{1}{12(m_k+1)} - \frac{1}{12(n-(m_k+1))}\right)$. One may apply Inequality 30 to get

$$\begin{aligned}\binom{n}{m_k+1} &\overset{30}{>} \frac{\sqrt{2\pi}n^{n+1/2}}{\sqrt{2\pi}(m_k+1)^{m_k+1+1/2}\sqrt{2\pi}(n-(m_k+1))^{n-(m_k+1)+1/2}} \frac{e^{-n}}{e^{-(m_k+1)}e^{-(n-(m_k+1))}} a_k \\ &= \frac{\sqrt{n}}{\sqrt{2\pi}} n^n \left[m_k+1\right]^{-(m_k+1+1/2)} \left[n-(m_k+1)\right]^{-(n-(m_k+1)+1/2)} a_k \\ &:= c_k a_k.\end{aligned}$$

Now,

$$\begin{aligned}c_k(m_k+1)p_k^{m_k+1}(1-p_k)^{n-m_k} &= \frac{\sqrt{n}}{\sqrt{2\pi}} n^n \left[m_k+1\right]^{-(m_k+1+1/2-1)} \left[n-(m_k+1)\right]^{-(n-(m_k+1)+1/2)} \\ &\quad \times p_k^{m_k+1}(1-p_k)^{n-m_k} \\ &= \frac{\sqrt{n}}{\sqrt{2\pi}} n^n (np_k)^{-(m_k+1/2)} \left[\frac{m_k+1}{np_k}\right]^{-(m_k+1/2)} \\ &\quad \times (n(1-p_k))^{-(n-(m_k+1/2))} \left[\frac{n-(m_k+1)}{n(1-p_k)}\right]^{-(n-(m_k+1/2))} \\ &\quad \times p_k^{m_k+1}(1-p_k)^{n-m_k} \\ &= \frac{\sqrt{n}}{\sqrt{2\pi}} \sqrt{p_k(1-p_k)} \left[\frac{m_k+1}{np_k}\right]^{-(m_k+1/2)} \\ &\quad \times \left[\frac{n-(m_k+1)}{n(1-p_k)}\right]^{-(n-(m_k+1/2))} \\ &:= \frac{\sqrt{n}}{\sqrt{2\pi}} \sqrt{p_k(1-p_k)} d_k,\end{aligned}$$

which finally implies

$$\frac{\sqrt{n}}{\sqrt{2\pi}} \sqrt{p_k(1-p_k)} d_k a_k < \binom{n}{m_k+1} (m_k+1) p_k^{m_k+1}(1-p_k)^{n-m_k}.$$

Let us compute a lower bound on $d_k a_k$. We start with d_k.
Define $\epsilon_k \in [0,1)$ such that $np_k = m_k + \epsilon_k$. Then

$$\begin{aligned}d_k &= \exp\left\{\left(m_k+\frac{1}{2}\right)\ln\left(\frac{m_k+\epsilon_k}{m_k+1}\right) + \left(n-m_k-\frac{1}{2}\right)\ln\left(\frac{n-m_k-\epsilon_k}{n-m_k-1}\right)\right\} \\ &= \exp\left\{\left(m_k+\frac{1}{2}\right)\ln\left(1-\frac{1-\epsilon_k}{m_k+1}\right) + \left(n-m_k-\frac{1}{2}\right)\ln\left(1+\frac{1-\epsilon_k}{n-m_k-1}\right)\right\}.\end{aligned}$$

Using that for all $u \geq 0$, $\ln(1-u) \geq -u - u^2/2$, we get

$$d_k \geq \exp\left\{-(1-\epsilon_k) - \left(\frac{(1-\epsilon_k)^2}{m_k+1}\right)\right\} \geq \exp(-2).$$

Moreover, one obviously has $a_k \geq \exp(-1/6)$.

Now let us compute a lower bound for when $n \leq 1/(1-p_k) \iff p_k \geq 1 - 1/n$, which happens at most once since $n \geq 2$. In this case, one has $m_k = n-1$ and therefore

$$\binom{n}{m_k+1}(m_k+1)p_k^{m_k+1}(1-p_k)^{n-m_k} = \sqrt{n}\sqrt{p_k(1-p_k)}\sqrt{n}p_k^{n-1/2}(1-p_k)^{1/2}. \tag{31}$$

The function $\zeta : p \mapsto p^{n-1/2}(1-p)^{1/2}$ is increasing up to $1-1/2n$ and then is decreasing. Since $p_k \geq 1-1/n$, when $p_k \leq 1-1/2n$, one can lower bound the r.h.s. of 31 by the evaluation of ζ at $p = 1-1/n$

$$\begin{aligned}\sqrt{n}\sqrt{p_k(1-p_k)}\sqrt{n}p_k^{n-1/2}(1-p_k)^{1/2} &\geq \sqrt{n}\sqrt{p_k(1-p_k)}\sqrt{n}(1-1/n)^{n-1/2}n^{-1/2}\\ &= \sqrt{n}\sqrt{p_k(1-p_k)}(1-1/n)^{n-1/2}\\ &\geq \sqrt{n}\sqrt{p_k(1-p_k)}\frac{1}{2\sqrt{2}},\end{aligned}$$

where the last inequality comes from the fact that $n \mapsto (1-1/n)^{n-1/2}$ is increasing and $n \geq 2$.
In any case, if $p_k \leq 1 - 1/2n$ for all $k = 1, \cdots, K$, we get

$$\begin{aligned}\max_{\mathcal{A}} \Delta_n(P, \mathcal{A}) &\geq \min\left\{\frac{1}{2\sqrt{2}}; \frac{exp(-13/6)}{\sqrt{2\pi}}\right\} C(P)n^{-1/2}\\ &= \frac{exp(-13/6)}{\sqrt{2\pi}}C(P)n^{-1/2}.\end{aligned}$$

If for some j^* we have $p_{j^*} > 1 - 1/2n$, then we may lower bound the r.h.s. of 31 by 0. Without loss of generality, suppose that $j^* = K$. We then have

$$\begin{aligned}\max_{\mathcal{A}} \Delta_n(P, \mathcal{A}) &\geq \frac{exp(-13/6)}{\sqrt{2\pi}} \sum_{j=1}^{K-1} \sqrt{p_j(1-p_j)}n^{-1/2}\\ &\geq \frac{exp(-13/6)}{\sqrt{2\pi}} \left(C(P) - \frac{1}{\sqrt{2n}}\right) n^{-1/2},\end{aligned}$$

where the second inequality comes from $\sqrt{p_{j^*}(1-p_{j^*})} \leq \sqrt{2n-1}/(2n) \leq 1/\sqrt{2n}$.

□

Proof of Lemma 5.5. We recall that the range of the algorithm is finite. Without loss of generality, we identify the set of parameters to $\{1, \cdots, L\}$ for some $L \in \mathbb{N}$. Here Equation (21) writes

$$2\Delta_n(P, \mathcal{A}) = \sum_{k=1}^{K} p_k \sum_{l=1}^{L} \left| P(F(\hat{P}_n) = l) - P(F(\hat{P}_n^k) = l) \right|. \tag{32}$$

By the symmetry of $\mathcal{A}$, only the distribution of the data among the possible values is relevant, that is the random variables $N_1 \coloneqq \sum_{i=1}^{n} \mathbf{1}_{z_i=1}, \cdots, N_K \coloneqq \sum_{i=1}^{n} \mathbf{1}_{z_i=K}$. Note that for $j = 1, \cdots, K$, the random variable N_j is the number of occurrences of u_j in the dataset.
By the *i.i.d.* assumption of the data, the random vector $\mathbf{r} \coloneqq (N_1, \cdots, N_K)$ follows a multinomial distribution with parameters $(n; p_1, \cdots, p_K)$.
For any $k = 1, \cdots, K$, let $D_k \subseteq \{(n_1, \cdots, n_K) \in \{0, \cdots, n\}^K : \sum_{q=1}^{K} n_q = n\}$ such that $F(\hat{P}_n) = k$ if and only if $\mathbf{r} \in D_k$.
Since $\mathbf{r}$ follows a multinomial distribution, using Equations (32), (23) and (24), one gets

$$\begin{aligned}2\Delta_n(P, \mathcal{A}) &= \sum_{k=1}^{K} p_k \sum_{l=1}^{L} |P(\mathbf{r} \in D_l) - P(\mathbf{r} \in D_l \mid \mathbf{z}_1 = u_k)|\\ &= \sum_{k=1}^{K} p_k \sum_{l=1}^{L} \left| \sum_{(n_1, \cdots, n_K) \in D_l} \binom{n}{n_1 \cdots n_K} \left(\prod_{q=1}^{K} p_q^{n_q}\right) \left\{1 - \frac{n_k}{np_k}\right\} \right|\\ &= \sum_{k=1}^{K} \sum_{l=1}^{L} \left| \mathbb{E}\left[\left\{ p_k - \frac{N_k}{n} \right\} \mathbf{1}_{\mathbf{r} \in D_l} \right] \right|.\end{aligned}$$

□

Proof of Lemma 5.6. From Equation (16), one has

$$\Delta_n(P,\mathcal{A}) = \mathbb{E}\left(\left\|\mathcal{L}(F(\hat{P}_n)) - \mathcal{L}(F(\hat{P}_n)|\mathrm{z}_1)\right\|_{\mathrm{TV}}\right).$$

By definition of the total variation for discrete distributions, for $b \in \{0,1\}$, one has

$$\begin{aligned}
2\left\|\mathcal{L}(F(\hat{P}_n)) - \mathcal{L}(F(\hat{P}_n)|\mathrm{z}_1 = b)\right\|_{\mathrm{TV}} &= |P(F(\hat{P}_n)=1) - P(F(\hat{P}_n)=1|\mathrm{z}_1=b)| \\
&\quad + |P(F(\hat{P}_n)=0) - P(F(\hat{P}_n)=0|\mathrm{z}_1=b)| \\
&= \left|1-(1-p)^n - \left\{\begin{array}{ll} 1 & \text{si } b=1 \\ 1-(1-p)^{n-1} & \text{si } b=0 \end{array}\right.\right| \\
&\quad + \left|(1-p)^n - \left\{\begin{array}{ll} 0 & \text{if } b=1 \\ (1-p)^{n-1} & \text{if } b=0 \end{array}\right.\right| \\
&= 2(1-p)^{n-1}\left|(1-p) - \mathbf{1}_{b=0}\right| \\
&= 2(1-p)^{n-1}\left|p - \mathbf{1}_{b=1}\right|.
\end{aligned}$$

Taking expectation over z_1 gives

$$\begin{aligned}
\Delta_n(P,\mathcal{A}) &= (1-p)^{n-1}\mathbb{E}[|p - \mathbf{1}_{\mathrm{z}_1=1}|] \\
&= (1-p)^{n-1}[2p(1-p)] \\
&= 2p(1-p)^n,
\end{aligned}$$

which concludes the proof. □

C Extension to Classification

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 $(\mathrm{z}_1,\cdots,\mathrm{z}_n) \coloneqq ((\mathrm{x}_1,\mathrm{y}_1),\cdots,(\mathrm{x}_n,\mathrm{y}_n))$ 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 $\Psi_\theta(x) = \sum_{j=1} v_j \sigma(w_j^T x + b_j)$ be a $2-$ReLU network with parameters θ, i.e. $\theta = (v_j, w_j, b_j)_{j=1}$ with $\in \mathbb{N}$ 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

$$\mathcal{L} : \theta \mapsto \sum_{j=1}^{n} l(y_j \Psi_\theta(x_j)), \tag{33}$$

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 33, 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 $\max_{i\neq j}\left\{|\mathrm{x}_i^T \mathrm{x}_j|\right\} < d$, 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

$$\min_\theta \frac{1}{2}\|\theta\|^2 \ \text{ s.t. } \ \forall i = 1,\ldots,n; y_i \Psi_\theta(x_i) \geq 1. \tag{34}$$

Conditional to the event $E \coloneqq$ "$\max_{i\neq j}\left\{|\mathrm{x}_i^T \mathrm{x}_j|\right\} \leq \frac{s+1}{3n} - 1$", by Vardi et al. [2022] Lemma C.1 we get that for all $j = 1,\ldots,n$, we have

$$\mathrm{y}_j \Psi_{\bar{\theta}_n}(\mathrm{x}_j) = \lambda(\mathrm{z}_1, \cdots, \mathrm{z}_n), \tag{35}$$

for some $\lambda(\mathrm{z}_1, \cdots, \mathrm{z}_n) > 0$.

We consider our algorithm $\mathcal{A}$ to output

$$\mathcal{A}(\mathrm{z}_1, \cdots, \mathrm{z}_n) = \hat{\theta}_n \coloneqq \frac{\bar{\theta}_n}{\sqrt{\lambda(\mathrm{z}_1, \cdots, \mathrm{z}_n)}},$$

which gives the same classifier as with $\bar{\theta}_n$.

We then get the following result.

Proposition C.1. *Assume that* $\geq n$ *and let* $C \coloneqq \max_{i \neq j} \left\{ |\mathrm{x}_i^T \mathrm{x}_j| \right\}$. *Then, there exists an initialization* $\theta_n(0)$ *of the gradient flow for which it holds that*

$$\Delta_n(P, \mathcal{A}) \geq P\left(C \leq \frac{s+1}{3n} - 1 \right).$$

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

$$\Delta_n(P, \mathcal{A}) \geq 1 - s^{3 - ln(s)/4},$$

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

Proof of Proposition C.1. 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 35 leads to

$$\mathrm{y}_j \Psi_{\hat{\theta}_n}(\mathrm{x}_j) = 1, \tag{36}$$

for any $j = 1, \ldots, n$.

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

$$\begin{aligned}
\Delta_n(P, \mathcal{A}) &\geq P((\hat{\theta}_n, \mathrm{x}_1, \mathrm{y}_1) \in S) - P((\hat{\theta}_n, \mathrm{x}, \mathrm{y}) \in S) \\
&= P((\hat{\theta}_n, \mathrm{x}_1, \mathrm{y}_1) \in S \mid E) P(E) + P((\hat{\theta}_n, \mathrm{x}_1, \mathrm{y}_1) \in S \mid E^c) P(E^c) - \mathbb{E}\left[P(\Psi_{\hat{\theta}_n}(\mathrm{x}) = \mathrm{y} \mid \hat{\theta}_n, \mathrm{y}) \right] \\
&\geq P((\hat{\theta}_n, \mathrm{x}_1, \mathrm{y}_1) \in S \mid E) P(E) - \mathbb{E}\left[P(\Psi_{\hat{\theta}_n}(\mathrm{x}) = \mathrm{y} \mid \hat{\theta}_n, \mathrm{y}) \right],
\end{aligned}$$

where we have lower bounded the second term by 0.

By Equation 36, we have $P((\hat{\theta}_n, \mathrm{x}_1, \mathrm{y}_1) \in S \mid E) = 1$. Now, by independence between (x, y) and $\hat{\theta}_n$, it is sufficient to show that for any $\theta \in \Theta$, we have $P(\Psi_\theta(\mathrm{x}) = \mathrm{y} \mid \mathrm{y}) = 0$ almost surely. Without loss of generality, we may assume that $v_j \neq 0$ for any $j = 1, \ldots, .$ We then get

$$\begin{aligned}
P\left(\Psi_{\hat{\theta}_n}(\mathrm{x}) = \mathrm{y} \mid \mathrm{y} \right) &= \sum_{J \subseteq [1, \cdots,]} P\left(\left\{ \forall j \in J, w_j^T \mathrm{x} + b_j > 0 \right\} \cap \left\{ \forall j \in J^c, w_j^T \mathrm{x} + b_j \leq 0 \right\} \cap \left\{ \sum_{j \in J} v_j \left(w_j^T \mathrm{x} + b_j \right) = \mathrm{y} \right\} \mid \mathrm{y} \right) \\
&\leq \sum_{J \subseteq [1, \cdots,]} P\left(\sum_{j \in J} v_j \left(w_j^T \mathrm{x} + b_j \right) = \mathrm{y} \mid \mathrm{y} \right).
\end{aligned}$$

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

$$\Delta_n(P, \mathcal{A}) \geq P(E).$$

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.

□