Learning, compression, and leakage: Minimising classification error via meta-universal compression principles

Fernando E. Rosas
Data Science Institute
Imperial College London
London, UK
f.rosas@imperial.ac.uk

Pedro A.M. Mediano
Department of Psychology
University of Cambridge
Cambridge, UK
pam83@cam.ac.uk

Michael Gastpar
School of Computer and Communication Sciences
École Polytechnique Fédérale de Lausanne
Lausanne, Switzerland
michael.gastpar@epfl.ch

Abstract—Learning and compression are driven by the common aim of identifying and exploiting statistical regularities in data, which opens the door for fertile collaboration between these areas. A promising group of compression techniques for learning scenarios is *normalised maximum likelihood* (NML) coding, which provides strong guarantees for compression of small datasets — in contrast with more popular estimators whose guarantees hold only in the asymptotic limit. Here we consider a NML-based decision strategy for supervised classification problems, and show that it attains heuristic PAC learning when applied to a wide variety of models. Furthermore, we show that the misclassification rate of our method is upper bounded by the *maximal leakage*, a recently proposed metric to quantify the potential of data leakage in privacy-sensitive scenarios.

Index Terms—Supervised learning; Universal Compression; Maximal Leakage; Normalised Maximum Likelihood

I. INTRODUCTION

Since compression and learning are both based on exploiting statistical regularities of the data, it is often possible to leverage compression techniques to enable novel learning methods. Examples of successful translations abound in the literature, including the use of universal compression methods such as Context Tree Weighting [1] for predicting time series via variable-order Markov chains [2].

Among the literature on universal compression, the work of Jorma Rissanen and the Minimum Description Length (MDL) community is particularly well-suited for statistical learning. There are two particularly attractive aspects of the MDL philosophy from a learning perspective (c.f. [3], [4]): a focus on the data itself and not on assumptions about related probabilistic models, and an emphasis on estimators that have useful properties for finite sample sizes. These ideas lead to the use of *normalised maximum likelihood* (NML) codes, previously introduced by Shtar'kov [5], to develop universal compression methods [6]. NML distributions provide minimax

optimal compression features for finite sample sizes — in contrast to e.g. distributions obtained via maximum likelihood estimation that only have guarantees in the asymptotic regime.

Despite of their attractive properties, NML distributions have not been much explored in the statistical learning literature. An important exception is the work reported in Refs. [7]–[9], which leverages conditional NML (cNML) distributions — originally introduced by Roos & Rissanen [10] — to address a supervised learning setting. The favourable properties of cNML-based learning strategies have been demonstrated for the cases of linear regression [8] and deep neural networks [9]. Unfortunately, the available theoretical guarantees for the learnability of cNML models are still limited.

Another important contribution of Rissanen was the development of the notion of *stochastic complexity*, a metric of model complexity that refines well-known model selection procedures such as the Akaike and Bayesian Information Criteria [6]. Stochastic complexity has a remarkable similarity to *maximal leakage*, a measure introduced in Ref. [11] to quantify leakage risk in privacy-sensitive scenarios. This formal similarity is particularly intriguing given the connection that exist between data privacy and learning: as privacy-preserving algorithms only process general properties of datasets without focusing on particular data samples (see [12]), they are less likely to fall prey to overfitting. This idea was first developed in the context of differential privacy [13], and recent reports have shown that maximal leakage can be used to bound the generalisation error in supervised learning scenarios [14], [15].

The goal of this paper is to establish a rigorous link between supervised learning, NML methods, and maximal leakage. For this, we employ a NML-based decision strategy based on meta-universal compression principles [4, Ch. 11.2], where the model is dynamically adapted according to the training data. We provide an upper bound, based on maximal leakage, to the performance gap between our NML strategy and the (optimal) MAP criterion (Theorem 1). Furthermore, using this bound we show that our NML strategy possesses strong learning guarantees that hold in various contexts (Theorem 2 and Proposition 2). Importantly, while most of the MDL

literature is based on logarithmic losses (including [7]–[9]), our approach quantifies performance in terms of classification accuracy, which is a more natural metric for supervised learning scenarios.

The rest of the paper is structured as follows. Section II introduces our supervised learning scenario and discusses fundamental notions of universal compression and information leakage. Section III presents our main technical results, and Section IV summarises our conclusions. The Appendices provide the proofs of our results, illustrate the findings in a simple scenario, and discuss some implementation issues.

II. PRELIMINARIES

A. Scenario

Let us consider a classification task where one needs to decide which class $Y \in \mathcal{Y} = \{c_1, \ldots, c_K\}$ a given observation $X \in \mathcal{X}$ belongs to. A *hypothesis* is a (possibly stochastic) mapping $h : \mathcal{X} \to \mathcal{Y}$, whose performance is measured using the *0–1 loss function* given by

$$\mathsf{Loss}\,(y, \tilde{y}) = \begin{cases} 1 & \text{if} \quad y \neq \tilde{y}, \\ 0 & \text{otherwise.} \end{cases}$$

The misclassification probability of x under h is calculated as

$$\begin{aligned} \mathsf{E}(h; x) &\coloneqq \mathbb{E}\{\mathsf{Loss}\,(Y, h(X)) \,|\, X = x\} \\ &= \mathbb{P}\,\{Y \neq h(X) | X = x\} \\ &= 1 - f(h(x)|x), \end{aligned} \quad (1)$$

where $f(y|x)$ is the conditional probability of $\{Y = y\}$ given $\{X = x\}$. The misclassification rate of h is defined as $\mathsf{E}(h) \coloneqq \mathbb{E}\{\mathsf{E}(h; X)\} = \mathbb{E}\{\mathsf{Loss}\,(Y, h(X))\}$. The well-known *maximum-a-posteriori* (MAP) rule, defined as[1]

$$h_{\text{MAP}}(x) \coloneqq \arg\max_{y \in \mathcal{Y}} f(y|x), \quad (2)$$

can be shown to attain a minimal misclassification rate given by $\mathsf{E}(h_{\text{MAP}}; x) = 1 - \max_{y \in \mathcal{Y}} f(y|x)$ [16]. Unfortunately, to build h_{MAP} one needs precise knowledge of $f(y|x)$, which is rarely available in most scenarios of practical interest.

Consider now n available samples for training denoted by $z_1 = (x_1, y_1), \ldots, z_n = (x_n, y_n)$, and denote the whole dataset by $z^n = (z_1, \ldots, z_n)$. Hypotheses that are built on training data correspond to functions $h : \mathcal{X} \times \mathcal{Z}^n \to \mathcal{Y}$, where $\mathcal{Z} \coloneqq \mathcal{X} \times \mathcal{Y}$. Then, a hypothesis $h(x, z^n)$ can be equivalently expressed as

$$h_q(x, z^n) = \arg\max_{y \in \mathcal{Y}} q(y|x, z^n), \quad (3)$$

where $q(y|x, z^n)$ is a (possibly not unique) suitable conditional probability distribution. The misclassification rate of h_q is

$$\begin{aligned} \mathsf{E}(h_q; x, z^n) &\coloneqq \mathbb{P}\,\{Y \neq h_q(X, Z^n) | X = x, Z^n = z^n\} \quad (4) \\ &= 1 - f(h_q(x, z^n)|x). \quad (5) \end{aligned}$$

Due to the optimality of h_{MAP}, $\mathsf{E}(h_{\text{MAP}}; x) \leq \mathsf{E}(h_q; x, z^n)$ holds for any hypotheses given by $q(y|x, z^n)$.

[1] In case there is more than one value of y that maximises (2), $h_{\text{MAP}}(x)$ assigns one of them randomly.

B. Universal compression

While elementary compression algorithms consider data coming from a single information source (i.e. i.i.d. data generated from symbols in the alphabet $\mathcal{Y}$ according to a given probability distribution $p(y)$), universal compression approaches aim to be suitable to compress data with respect to a statistical model class $\mathcal{M}$ — understood as a collection of probability distributions. The goal is to build distributions q that attain low values of

$$\text{REG}_{\max}(\mathcal{M}, q) \coloneqq \sup_{p \in \mathcal{M}} \max_{y \in \mathcal{Y}} \ln \frac{p(y)}{q(y)} = \sup_{p \in \mathcal{M}} R(p, q), \quad (6)$$

which stands for the "maximal regret" while using q to code data related to any model p in $\mathcal{M}$ [4].

A remarkable result from the MDL literature is that the minimiser of $\text{REG}_{\max}$ can often be written in closed form, and is given by an NML distribution of the form

$$q_{\text{NML}, \mathcal{M}}(y) = \frac{\sup_{p \in \mathcal{M}} p(y)}{Z_{\mathcal{M}}}, \quad (7)$$

where $Z_{\mathcal{M}} = \sum_{y \in \mathcal{Y}} \sup_{p \in \mathcal{M}} p(y)$ is a normalisation constant. The minimal regret is given by

$$\min_q \text{REG}_{\max}(\mathcal{M}, q) = \text{REG}_{\max}(\mathcal{M}, q_{\text{NML}, \mathcal{M}}) = \ln Z_{\mathcal{M}}, \quad (8)$$

being known as the *stochastic complexity* of $\mathcal{M}$ [6].

Note that the NML might not be well-defined if $Z_{\mathcal{M}}$ diverges. One solution to those cases is to employ submodels to reduce the minimal regret, since $\mathcal{M}' \subset \mathcal{M}$ implies $Z_{\mathcal{M}'} \leq Z_{\mathcal{M}}$. This approach is known as *meta-universal* coding, which includes a range of techniques developed in the literature [4, Section 11.2].

C. Quantifying information leakage

Consider a variable ϕ that parameterises the distributions $p_\phi(Y)$ that belong to $\mathcal{M}$. We are interested in quantifying how much information about ϕ can be extracted from observations of Y. Note that this highly non-trivial issue is not properly addressed by naive applications of Shannon's mutual information or differential privacy criteria [17], [18].

We follow Ref. [11] and consider a random variable U that is conditionally independent of Y given ϕ, and imagine guessing U from Y via $\hat{U}$, so that $U - \phi - Y - \hat{U}$ forms a Markov chain. Then, the *maximal leakage* between ϕ and Y,

$$\mathcal{L}(\phi \to Y) \coloneqq \sup_{U - \phi - Y - \hat{U}} \log \frac{\mathbb{P}\{U = \hat{U}\}}{\max_{u \in \mathcal{U}} \mathbb{P}\{U = u\}}, \quad (9)$$

characterizes the least protected secret U (that is, the worst case over U) of ϕ with respect to Y. A closed-form formula for $\mathcal{L}(\phi \to Y)$ is given by [18, Corollary 4]

$$\mathcal{L}(\phi \to Y) = \log \sum_{y \in \mathcal{Y}} \sup_{\theta \in \text{supp}(\phi)} f(y|\theta), \quad (10)$$

with $\text{supp}(\phi) \coloneqq \{\theta \in \Theta : \mathbb{P}\{\phi = \theta\} > 0\}$. This form is equivalent to the Sibson's mutual information of order infinity [19], and has a number of useful properties and an operational interpretation that are discussed in Ref. [18].

III. OPTIMIZING THE HYPOTHESIS BASED ON META-UNIVERSAL CODING PRINCIPLES

A. Learning based on universal source coding

We first focus on a parametric model $\mathcal{P}$, which is a set of conditional distributions $p_{\boldsymbol{\theta}}(y|x)$ indexed by $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d) \in \mathbb{R}^d$. Following meta-universal coding principles (c.f. Section II-B), we consider sub-models of the form

$$\mathcal{A}(z^n) = \{p_{\boldsymbol{\theta}}(\cdot|\cdot) \in \mathcal{P} : \boldsymbol{\theta} \in \Theta(z^n)\} \subset \mathcal{M}, \quad (11)$$

where $\Theta(z^n) \subset \mathbb{R}^d$ is a restriction in the space of parameters that depends on the traning set z^n. For the sub-model $\mathcal{A}(z^n)$, we define the following NML distribution:

$$q_{\mathrm{NML},\mathcal{A}}(y|x, z^n) := \frac{\sup_{\boldsymbol{\theta} \in \Theta(z^n)} p_{\boldsymbol{\theta}}(y|x)}{Z\big(x; \Theta(z^n)\big)}, \quad (12)$$

with $Z\big(x; \Theta(z^n)\big) = \sum_{y \in \mathcal{Y}} \sup_{\boldsymbol{\theta} \in \Theta(z^n)} p_{\boldsymbol{\theta}}(y|x)$. Please note that this type of NML construction has been considered before in Ref. [7, Sec. 5]. Importantly, $Z\big(x; \Theta(z^n)\big) < \infty$ due to the finiteness of $\mathcal{Y}$, and hence $q_{\mathrm{NML},\mathcal{A}}$ is well-defined for all $\mathcal{A}(z^n)$. The minimal regret attained by this NML distribution is $\ln Z\big(x; \Theta(z^n)\big)$, which corresponds to the stochastic complexity of model $\mathcal{A}(z^n)$.

When designing an NML distribution, choosing an adequate sub-model $\mathcal{A}(z^n)$ is critical — or, equivalently, to set adequate parameter restrictions $\Theta(z^n)$. To gain insight about the effect of $\Theta(z^n)$ on the corresponding NML distribution, let us study the stochastic complexity of the sub-model as a form of information leakage (c.f. Section II-C). For this, we consider a random variable ϕ that takes values in a subset of the parameter space $\Theta(z^n) \subset \mathbb{R}^d$, and assume it satisfies the Markov chain $\phi - Z^n - X$. Following Eq. (10), the maximal leakage from ϕ to Y for given $X = x$ and $Z^n = z^n$ is

$$\mathcal{L}(\phi \to Y|x; z^n) := \ln \left\{ \sum_{y \in \mathcal{Y}} \sup_{\boldsymbol{\theta} \in \mathrm{supp}(\phi|z^n)} p_{\boldsymbol{\theta}}(y|x) \right\}, \quad (13)$$

with $\mathrm{supp}(\phi|z^n) = \{\boldsymbol{\theta} \in \Theta(z^n) : \mathbb{P}\{\phi = \boldsymbol{\theta}|Z^n = z^n\} > 0\}$. This quantity has two useful properties:

1. *It corresponds to a stochastic complexity:* if ϕ is such that $\mathrm{supp}(\phi|z^n) = \Theta(z^n)$, then $\mathcal{L}(\phi \to Y|x; z^n) = \log Z\big(x; \Theta(z^n)\big)$.
2. *It is monotonous with $\mathrm{supp}(\phi|z^n)$, and does not depend on other details of its distribution:* if ϕ_1 and ϕ_2 are variables such that $\mathrm{supp}(\phi_1|z^n) \subseteq \mathrm{supp}(\phi_2|z^n)$, then $\mathcal{L}(\phi_1 \to Y|x; z^n) \le \mathcal{L}(\phi_2 \to Y|x; z^n)$.

Intuitively, $\mathcal{L}(\phi \to Y|x; z^n)$ quantifies the information about ϕ that can still be leaked from Y after x and z^n have already been given.[2] Put simply, the leakage measures how much better the training would be with $n + 1$ samples, by considering all potential additional training samples of the form $z_{n+1} = (x, c_k)$ with $k = 1, \ldots, K$. Therefore, a high

---

[2]Note that $\mathcal{L}(\phi \to Y|x; z^n)$ is not a conditional leakage, but the leakage for given values of $X = x$ and $Z^n = z^n$. Conditional leakage has been defined in Ref. [18].

---

value of $\mathcal{L}(\phi \to Y|x; z^n)$ implies that the training enabled by z^n has not saturated yet and still has room for improvement.

We make this intuition precise with the analysis carried out below. Let us denote by $q_{\mathrm{NML},\phi}(y|x, z^n)$ the NML distribution for the model $\mathcal{P}$ with parameters restricted to $\mathrm{supp}(\phi|z^n)$, and consider the hypothesis given by

$$h_{\mathrm{NML},\phi}(x, z^n) = \arg\max_{y \in \mathcal{Y}} q_{\mathrm{NML},\phi}(y|x, z^n) \quad (14)$$

$$= \arg\max_{y \in \mathcal{Y}} \sup_{\boldsymbol{\theta} \in \mathrm{supp}(\phi|z^n)} p_{\boldsymbol{\theta}}(y|x). \quad (15)$$

Our first result identifies upper bounds to the performance of this hypothesis.

Theorem 1. *Consider a d-dimensional parametric model $\mathcal{P}$, and a conditional probability $f(y|x)$. Then, for any random variable $\phi \in \mathbb{R}^d$ that depends on a dataset $z^n \in \mathcal{Z}^n$, the following bound holds:*

$$\mathrm{E}(h_{\mathrm{NML},\phi}; x, z^n) - \mathrm{E}(h_{\mathrm{MAP}}; x)$$
$$\le \exp\big\{\Delta\big(f, \mathrm{supp}(\phi|z^n)|x\big) + \mathcal{L}(\phi \to Y|x; z^n)\big\} - 1 \,,$$

where $\Delta(f, \Theta|x) := \inf_{\boldsymbol{\theta} \in \Theta} \max_{y \in \mathcal{Y}} \ln \frac{f(y|x)}{p_{\boldsymbol{\theta}}(y|x)}$.

Proof. The proof proceeds in three steps. First, one proves that for any distribution $q(y|x, z^n)$ the following bound holds:

$$\mathrm{E}(h_q; x, z^n) - \mathrm{E}(h_{\mathrm{MAP}}; x) \le e^{R(f, q|x, z^n)} - 1 \,, \quad (16)$$

where $R(f, q|x, z^n) := \max_{y \in \mathcal{Y}} \ln \frac{f(y|x)}{q(y|x, z^n)}$ is the redundancy between f and q given x and the training sample z^n (c.f. Section II-B). Then, one proves a triangle inequality $R(f, q) \le \Delta(f, \Theta) + \mathrm{REG}_{\max}(\mathcal{A}, q)$ for any sub-model $\mathcal{A}$ with parameters in $\Theta \subseteq \mathbb{R}^d$. Finally, the two previous steps are combined using $q = q_{\mathrm{NML},\phi}$ and $\Theta = \mathrm{supp}(\phi|z^n)$ to show the desired result. The details of the proof can be found in Appendix A. $\square$

The above result reflects the trade-offs involved in the design of $h_{\mathrm{NML},\phi}$: on the one hand, having a variable $\phi|z^n$ with a large support provides a big model which reduces Δ, at the risk of introducing a substantial regret as measured by the leakage $\mathcal{L}$; on the other hand, having a reduced support of $\phi|z^n$ guarantees a small leakage, at the price of increasing Δ. This result shows, in turn, that the maximal leakage provides a natural measure of overfitting. In effect, if the model with variables in $\mathrm{supp}(\phi|z^n)$ is too large, then for each class c_k there exists a parameter $\boldsymbol{\theta}_k \in \mathrm{supp}(\phi|z^n)$ such that $p_{\boldsymbol{\theta}}(c_k|x) \approx 1$, and hence $\mathcal{L} \approx \log |\mathcal{Y}|$. This is an indication of overfitting, as — rewording Ref. [20, Ch. 6] — a hypothesis that can accommodate every possible outcome explains none of them. On the other extreme, if $\arg\max_{\boldsymbol{\theta} \in \mathrm{supp}(\phi|z^n)} p_{\boldsymbol{\theta}}(y_k|x)$ is approximately constant for all classes, then $\mathcal{L} \approx 0$, which implies that the hypothesis is trustable.

We conclude this section by presenting a method to bound $\mathcal{L}(\phi \to Y|x; z^n)$ when the Fisher information matrix of the family $\mathcal{P}$ is well-defined. The Fisher information matrix of the distribution $p_{\boldsymbol{\theta}}(y|x)$ can be defined to be the $d \times d$ matrix

$I(\boldsymbol{\theta}|x)$ whose component in the i-th row and j-th column is calculated as

$$\left[I(\boldsymbol{\theta}|x) \right]_{i,j} = \mathbb{E}\left\{ \frac{\partial}{\partial \theta_i} \ln p_{\boldsymbol{\theta}}(Y|x) \cdot \frac{\partial}{\partial \theta_j} \ln p_{\boldsymbol{\theta}}(Y|x) \right\} . \quad (17)$$

The maximal eigenvalue of $I(\boldsymbol{\theta}|x)$ is denoted as $\sigma_{\max}(\boldsymbol{\theta}|x)$.

Lemma 1. *If* $\mathrm{supp}(\phi|z^n)$ *is a convex set and the Fisher information matrix is well-defined, then*

$$\mathcal{L}(\phi \to Y|x; z^n) \le \ln\left\{ 1 + \sum_{k=2}^{K} ||\boldsymbol{\theta}_k - \boldsymbol{\theta}_1|| \sqrt{\sigma_{\max}(\tilde{\boldsymbol{\theta}}_k|x)} \right\},$$

with $\boldsymbol{\theta}_i = \arg\max_{\boldsymbol{\theta} \in \mathrm{supp}(\phi|z^n)} p_{\boldsymbol{\theta}}(y_i|x)$ *for* $i = 1, \ldots, K$ *with* $\mathcal{Y} = \{y_1, \ldots, y_K\}$, *and* $\tilde{\boldsymbol{\theta}}_j = \tau_j \boldsymbol{\theta}_1 + (1 - \tau_j)\boldsymbol{\theta}_j$ *with* $\tau_j \in [0, 1]$ *for* $j = 2, \ldots, K$.

Proof. See Appendix B. □

B. Learning guarantees for well-specified models

We now consider the case where there exists a set of parameters $\boldsymbol{\theta}_0 \in \Theta \subset \mathbb{R}^d$ such that $f(y|x) = p_{\boldsymbol{\theta}_0}(y|x)$. Let us focus on the case where there is a consistent estimator $\hat{\theta} : \mathcal{Z}^n \to \Theta$ such that $\hat{\theta}(Z^n) \xrightarrow{p} \boldsymbol{\theta}_0$. Our next result is that, under these conditions, there exists a sequence of random variables ϕ_n such that the hypothesis h_{NML,ϕ_n} attains a form of agnostic probably approximately correct (PAC) learning [20], [21].

Theorem 2. *Consider* $f(y|x) = p_{\boldsymbol{\theta}_0}(y|x) \in \mathcal{P}$ *for some unknown parameter* $\boldsymbol{\theta}_0 \in \Theta \subset \mathbb{R}^d$, *and assume that there exists a consistent estimator* $\hat{\theta}(Z^n)$ *of* $\boldsymbol{\theta}_0$. *Also, assume that the Fisher matrix of* $\mathcal{P}$ *is well-defined over all* Θ, *and that* $\boldsymbol{\theta}_0$ *is an interior point. Then, for given* $x \in \mathcal{X}$ *and* $\epsilon, \delta > 0$, *there exists a random mapping* $\phi|\hat{\theta}$ *and* $n_0 \in \mathbb{N}$ *such that*

$$\mathrm{E}(h_{\mathrm{NML},\phi}; x, z^n) \le \mathrm{E}(h_{\mathrm{MAP}}; x) + \epsilon \quad (18)$$

for all $n \ge n_0$, *where the inequality holds for all* $z^n \in B \subset \mathcal{Z}^n$ *with* $\mathbb{P}\{Z^n \in B\} \ge 1 - \delta$.

Proof. One builds ϕ as a noisy version of a consistent estimator $\hat{\theta}(z^n)$, with the noise regulated by a parameter ρ. By carefully choosing ρ, one can use Theorem 1 and bound Δ using the properties of the consistent estimator, and control the leakage $\mathcal{L}$ using Lemma 1. The full proof is presented in Appendix C. □

Corollary 1 (Heuristic PAC learning). *If the assumptions required by Theorem 2 hold, then for given* $\delta, \epsilon > 0$ *there exists a random mapping* $\phi|\hat{\theta}$ *and an* n_0 *such that*

$$\mathbb{E}\{\mathrm{E}(h_{\mathrm{NML},\phi}; X, z^n)\} \le \mathbb{E}\{\mathrm{E}(h_{\mathrm{MAP}}; X)\} + \epsilon \quad (19)$$

for all $n \ge n_0$, *where the inequality holds for all* $z^n \in B \subset \mathcal{Z}^n$ *with* $\mathbb{P}\{Z^n \in B\} \ge 1 - \delta$.

Proof. See Appendix D. □

The conditions of Theorem 2 are satisfied if $\mathcal{P}$ is an exponential family (i.e. $p_{\boldsymbol{\theta}}(y|x)$ is an exponential family distribution for each $x \in \mathcal{X}$). Also, if $|\mathcal{X}| < \infty$ then any conditional

distribution $f(y|x)$ is just a collection of $2^{|\mathcal{X}|}$ multinomial distributions, and hence can be expressed using $|\mathcal{Y}| \cdot 2^{|\mathcal{X}|}$ parameters. In both cases, the corresponding parameters can be estimated via a maximum likelihood estimator, which is known to be consistent in these cases.[3] Please note that it is not straightforward to use our proof techniques to guarantee heuristic PAC learning to classification based directly on $\hat{\theta}$ (see Appendix E).

It would be useful to find explicit expressions for the dependency of δ, ϵ and n_0. For the particular case of models with a maximum likelihood estimator (MLE), one can prove additional properties of the $h_{\mathrm{NML},\phi}$ hypothesis. We leverage the fact that MLEs follow a central limit theorem:

$$\sqrt{n}\left(\hat{\theta}(z^n) - \boldsymbol{\theta}_0 \right) \xrightarrow{d} N\left(0, I^{-1}(\boldsymbol{\theta}_0) \right), \quad (20)$$

with $I(\boldsymbol{\theta}) = \mathbb{E}\{I(\boldsymbol{\theta}|X)\}$ being the unconditional Fisher matrix (with the average taken over both Y and X).

Proposition 1. *Consider a* d-*dimensional parametric model* $\mathcal{P}$ *with well-defined MLE* $\hat{\theta}(z^n)$ *and a positive-definite Fisher matrix* $I(\boldsymbol{\theta})$. *Then, for given* $\delta > 0$, $x \in \mathcal{X}$ *and* $z^n \in \mathcal{Z}^n$, *the following holds:*

$$\mathrm{E}(h_{\mathrm{NML},\boldsymbol{\psi}}; x, z^n) - \mathrm{E}(h_{\mathrm{MAP}}; x) \le e^{\mathcal{L}(\boldsymbol{\psi} \to Y|x; z^n)} - 1$$
$$\le \frac{1}{\sqrt{n}} K_{\delta,x} , \quad (21)$$

where $\boldsymbol{\psi} := \hat{\theta}(z^n) + W_\rho \in \mathbb{R}^d$ *with* W_ρ *uniformly distributed over a ball of radius* $\rho = \mathcal{O}(n^{-1/2})$ *and* $K_{\delta,x}$ *is a constant that does not depend on* n.

Proof. See Appendix F. □

Above, the first inequality provides a practical way to estimate the performance gap between $h_{\mathrm{NML},\boldsymbol{\psi}}$ and h_{MAP}. In effect, given that the radius ρ of the noise term of $\boldsymbol{\psi}$ has an explicit value, one can estimate the leakage $\mathcal{L}$. Additionally, the second inequality states that the performance gap reduces at least as $1/\sqrt{n}$ with the number of training samples.

C. Learning non-identifiable systems

In the previous section, we studied the PAC learning properties of NML estimators in the somewhat restrictive scenario in which the target function $f(y|x)$ belongs to the parametric family of models under consideration. This final subsection provides a generalisation of the main results presented above to more widely applicable settings.

We now consider a family of parametric models $\mathcal{P}$ that is capable of universal approximation, in the sense of Hornik [23]: in particular, for a given $f(y|x)$ with reasonable properties and $\epsilon > 0$, we assume that there exists a subset of parameter space $\Theta_{f,\epsilon} \subset \mathbb{R}^d$ such that $R(f, p_{\boldsymbol{\theta}}) < \epsilon$ for all $\boldsymbol{\theta} \in \Theta_{f,\epsilon}$.[4] Additionally, we consider that the system may

---

[3]For more information about existence of consistent estimators, see [22].

[4]To see why a universal approximator satisfies $R(f, p_{\boldsymbol{\theta}}) < \epsilon$, consider Theorem 1 in Ref. [24], stating that for any given target function $g(x)$, a parametrised approximator $G_{\boldsymbol{\theta}}(x)$, and an $\epsilon > 0$ there exists $\boldsymbol{\theta}$ such that $|g(x) - G_{\boldsymbol{\theta}}(x)| \le \epsilon$ for all x. Then, consider $g = \ln f$ and $G_{\boldsymbol{\theta}} = \ln p_{\boldsymbol{\theta}}$ to obtain the desired bound on $R(f, p_{\boldsymbol{\theta}})$.

be non-identifiable [25], in the sense that there are multiple $\boldsymbol{\theta}$ that minimise $\mathrm{E}(h_{p_{\boldsymbol{\theta}}})$, and in general the set $\Theta_{f,\epsilon} \subset \mathbb{R}^d$ might be non-convex. Moreover, we assume that there exists a (non-ergodic) estimator that converges to $\Theta_{f,\epsilon}$ in probability for any $\epsilon > 0$; i.e. a function $\hat{\theta} : \mathcal{Z}^n \to \mathbb{R}^d$ such that for all $\delta, \rho > 0$ there exists an $n_0(\delta, \rho) \in \mathbb{N}$ such that for all $n > n_0$ there is a set $B \subset \mathbb{R}^d$ of measure $\mathbb{P}\{Z^n \in B\} > 1 - \delta$ such that $\{\boldsymbol{\theta} \in \mathbb{R}^d : ||\boldsymbol{\theta} - \hat{\theta}(z^n)|| < \rho\} \cap \Theta_{f,\epsilon} \neq \varnothing$ for all $z^n \in B$. The next result shows that the desirable properties of our NML strategy still hold in this more general context.

Proposition 2. *Consider a conditional probability $f(y|x)$, and a universal approximator model $\mathcal{P}$ with well-defined Fisher matrix and a non-ergodic estimator $\hat{\theta}$ that converges in probability to $\Theta_{f,\epsilon}$ for any $\epsilon > 0$. Then, given $x \in \mathcal{X}$ and $z^n \in \mathcal{Z}^n$, for each $\epsilon, \delta > 0$, there exists $n_0 \in \mathbb{N}$ and a random mapping $\phi|\hat{\theta}$ such that for all $n > n_0$*

$$\mathrm{E}(h_{\mathrm{NML},\phi}; x, z^n) \leq \mathrm{E}(h_{\mathrm{MAP}}; x) + \epsilon \ . \tag{22}$$

Proof. See Appendix G. □

This result generalises the main result in Theorem 2 to the more practical setting of large non-identifiable models, like multi-layer neural networks, showing that NML can provide PAC guarantees even in the case of very general models.

IV. CONCLUSION

This paper provides a first step in the exploration of the potential of meta-universal coding and maximal leakage techniques for supervised learning theory. We have proposed an approach to build hypotheses based on Normalised Maximum Likelihood (NML) that can be applied to any standard learning algorithm. Crucially, we showed that models evaluated with this NML strategy attain heuristic PAC learning in a wide variety of contexts, and for specific cases we further showed that the performance gap between the NML approach and the optimal strategy decreases at least with the square-root of the number of samples.

In addition, we have provided an upper bound on the performance of our proposed NML strategy, and showed that this upper bound is directly determined by maximal leakage: a quantity used in the data privacy literature that we linked to the model's capacity to overfit. One interesting aspect of maximal leakage as a measure of overfitting is that it depends on the specific input to be classified, and hence could potentially be used to assess open problems in adversarial learning settings.

We hope this contribution may motivate further research efforts within the fascinating interface between learning, universal compression, and data privacy.

ACKNOWLEDGMENT

Proof. Consider the model class $\mathcal{M} = \{p_{\boldsymbol{\theta}} \in \mathcal{P} : \boldsymbol{\theta} \in \mathrm{supp}(\phi|z^n)\}$. By using Lemmas 2 and 3 (shown below) one can show that

$$\mathrm{E}(h_{\mathrm{NML},\phi}; x, z^n) - \mathrm{E}(h_{\mathrm{MAP}}; x) \leq \exp\big\{\Delta\big(f, \mathrm{supp}(\phi|z^n)\big) + \mathrm{REG}_{\max}\big(\mathrm{supp}(\phi|z^n), q_{\mathrm{NML},\phi}|x, z^n\big)\big\} - 1.$$

The Theorem is then proven by noting that

$$\mathrm{REG}_{\max}\big(\mathrm{supp}(\phi|z^n), q_{\mathrm{NML},\phi}|x, z^n\big) = \ln\big\{Z\big(x; \mathrm{supp}(\phi|z^n)\big)\big\}$$
$$= \mathcal{L}(\phi \to Y|x; z^n),$$

with $Z\big(x; \mathrm{supp}(\phi|z^n)\big)$ as defined in Eq. (12). □

Lemma 2. *For h_{MAP} and h_q as defined in Eqs. (2) and (3), the following holds:*

$$\mathrm{E}(h_q; x, z^n) - \mathrm{E}(h_{\mathrm{MAP}}; x) \leq e^{R(f,q|x,z^n)} - 1,$$

with $R(f, q|x, z^n) := \max_{y \in \mathcal{Y}} \ln \frac{f(y|x)}{q(y|x,z^n)}$.

Proof. Let us use $\delta := R(f, q|x, z^n)$ as a shorthand notation throughout the proof. Then, $\ln f(y|x) \leq \delta + \ln q(y|x)$ for all $y \in \mathcal{Y}$. Then, one can show that

$$\mathbb{P}\{Y = h_{\mathrm{MAP}}(X)|X = x\} = f(h_{\mathrm{MAP}}(x)|x)$$
$$\leq e^\delta q(h_{\mathrm{MAP}}(x)|x, z^n)$$
$$\leq e^\delta q(h_q(x, z^n)|x, z^n),$$

where the last equality holds because $h_q(x, z^n) = \arg\max_{y \in \mathcal{Y}} q(y|x, z^n)$. Now, note that for all $y_0 \in \mathcal{Y}$ one has that

$$q(y_0|x, z^n) = 1 - \sum_{y \neq y_0} q(y|x, z^n)$$
$$\leq 1 - e^{-\delta} \sum_{y \neq y_0} f(y|x)$$
$$= 1 - e^{-\delta}\Big(1 - f(y_0|x)\Big).$$

Then, this gives

$$\mathbb{P}\big\{Y = h_{\mathrm{MAP}}(X)|X = x\big\} \leq e^\delta\Big[1 - e^{-\delta} + e^{-\delta}f\big(h_q(x, z^n)|x\big)\Big]$$
$$= e^\delta - 1 + \mathbb{P}\{Y = h_q(X, Z^n)|X = x, Z^n = z^n\},$$

from where the desired result follows. □

Note that $R(f, q|x, z^n) \geq 0$ and hence $e^R - 1$ is non-negative, which is consistent with the optimality of the MAP hypothesis.

Lemma 3. *For any model class $\mathcal{M}$, the following bound holds:*

$$R(q, f|x, z^n) \leq \Delta(f, \mathcal{M}|x) + \mathrm{REG}_{\max}(q, \mathcal{M}|x, z^n),$$

with $R(q, f|x, z^n)$ as defined in Lemma 2.

Proof. First, note that

$$R(q, f|x, z^n) = \max_{y \in \mathcal{Y}}\left\{\ln \frac{f(y|x)}{p(y|x)} + \ln \frac{p(y|x)}{q(y|x, z^n)}\right\},$$

which holds for all $p \in \mathcal{M}$. This implies that

$$R(q, f|x, z^n) = \inf_{p \in \mathcal{M}} \max_{y \in \mathcal{Y}} \left\{ \ln \frac{f(y|x)}{p(y|x)} + \ln \frac{p(y|x)}{q(y|x, z^n)} \right\}$$

$$\leq \inf_{p \in \mathcal{M}} \max_{y \in \mathcal{Y}} \ln \frac{f(y|x)}{p(y|x)} + \sup_{p \in \mathcal{M}} \max_{y \in \mathcal{Y}} \ln \frac{p(y|x)}{q(y|x, z^n)},$$

proving the desired result. Note that, above, the last inequality is a consequence of the fact that

$$\inf_x \{ f(x) + g(x) \} \leq \inf_x \{ f(x) + \sup_x g(x) \}$$
$$= \inf_x f(x) + \sup_x g(x).$$

$\square$

APPENDIX B
PROOF OF LEMMA 1

Proof. For the second part, let us enumerate the possible classes as $\mathcal{Y} = \{y_1, \ldots, y_K\}$. Now, for given training data $z^n \in \mathcal{Z}^n$, we introduce the shorthand notation $\boldsymbol{\theta}_k := \arg\max_{\boldsymbol{\theta} \in \text{supp}(\phi|z^n)} p_{\boldsymbol{\theta}}(y_k|x)$ for $k = 1, \ldots, K$. Then,

$$\exp \left\{ \mathcal{L}(\phi \to Y|x, z^n) \right\} = \sum_{k=1}^{K} p_{\boldsymbol{\theta}_k}(y_k|x)$$

$$= 1 + \sum_{k=2}^{K} \left[p_{\boldsymbol{\theta}_k}(y_k|x) - p_{\boldsymbol{\theta}_1}(y_k|x) \right]$$

$$\leq 1 + \sum_{k=2}^{K} 2 d_{\text{TV}} \left(p_{\boldsymbol{\theta}_k}(y|x), p_{\boldsymbol{\theta}_1}(y|x) \right)$$

$$\leq 1 + \sum_{k=2}^{K} \sqrt{2 D \left(p_{\boldsymbol{\theta}_k}(y|x) || p_{\boldsymbol{\theta}_1}(y|x) \right)}.$$

Above, $d_{\text{TV}} \left(p(y|x), q(y|x) \right) := 1/2 \sum_{y \in \mathcal{Y}} |p(y|x) - q(y|x)|$ is the total variation distance, and the last inequality is a direct application of the well-known Pinsker inequality. To finish the proof, note that

$$\partial_i D \left(p_{\boldsymbol{\theta}}(Y|x) || p_{\boldsymbol{\theta}_0}(Y|x) \right) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = 0,$$

$$\partial^2_{i,j} D \left(p_{\boldsymbol{\theta}}(Y|x) || p_{\boldsymbol{\theta}_0}(Y|x) \right) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = I_{i,j}(\boldsymbol{\theta}_0|x).$$

Therefore, a first order Taylor expansion of the Kullback-Leibler divergence on $\boldsymbol{\theta}$ centered in $\boldsymbol{\theta}_0$ that expresses the reminder according to the Lagrange form [26] gives

$$D \left(p_{\boldsymbol{\theta}_k}(Y|x) || p_{\boldsymbol{\theta}_1}(Y|x) \right) = \frac{1}{2} (\boldsymbol{\theta}_k - \boldsymbol{\theta}_1)^T I(\tilde{\boldsymbol{\theta}}|x)(\boldsymbol{\theta} - \boldsymbol{\theta}_0),$$

where $\tilde{\boldsymbol{\theta}}_k = \tau_k \boldsymbol{\theta}_1 + (1 - \tau_k)\boldsymbol{\theta}_k$ for some $\tau_k \in (0, 1)$. Note that $\tilde{\boldsymbol{\theta}}_k \in \text{supp}(\phi|x, z^n)$ due to the convexity of the latter. The proof concludes by noting that

$$(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T I(\tilde{\boldsymbol{\theta}}; x)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \leq ||\boldsymbol{\theta} - \boldsymbol{\theta}_0||^2 \sigma_{\max}(\tilde{\boldsymbol{\theta}}|x),$$

due to the properties of the maximal eigenvalue $\sigma_{\max}(\tilde{\boldsymbol{\theta}}|x)$. $\square$

APPENDIX C
COMPLETE PROOF OF THEOREM 2

Proof. Let us consider a given $x \in \mathcal{X}$. As $\hat{\theta}$ is a consistent estimator of $\boldsymbol{\theta}_0$, then for given $\delta, \rho > 0$ there exists $n_{\boldsymbol{\theta}}(\delta, \rho) \in \mathbb{N}$ such that for all $n \geq n_{\boldsymbol{\theta}}(\delta, \rho)$ the following holds:

$$\mathbb{P} \left\{ ||\hat{\theta}(Z^n) - \boldsymbol{\theta}_0|| \geq \rho \right\} < \delta.$$

This implies that $B := \{ z^n \in \mathcal{Z}^n : ||\hat{\theta}(z^n) - \boldsymbol{\theta}_0|| < \rho \}$ satisfies $\mathbb{P} \{ Z^n \in B \} \geq 1 - \delta$. Also, by defining $\phi = \hat{\theta}(Z^n) + W_\rho$ with W_ρ distributing uniformly over $B(\rho) = \{ \boldsymbol{\theta} \in \mathbb{R}^d : ||\boldsymbol{\theta}|| < \rho \}$, then $\boldsymbol{\theta}_0 \in \text{supp}(\phi|z^n)$ for all $z^n \in B$. This implies, in turn, that $\Delta(f, \text{supp}(\phi|z^n)|x) = 0$. Therefore, using Theorem 1 one finds that for all $z^n \in B$ the following inequality holds:

$$\mathbb{E}(h_{\text{NML},\phi}; x, z^n) - \mathbb{E}(h_{\text{MAP}}; x) \leq \exp \left\{ \mathcal{L}(\phi \to Y|x; z^n) \right\} - 1. \quad (23)$$

To build a bound on $\mathcal{L}(\phi \to Y|x; z^n)$, let us define

$$\sigma_{\max}^{(\rho)}(\boldsymbol{\theta}_0|x) := \sup_{||\boldsymbol{\theta} - \boldsymbol{\theta}_0|| < \rho} \sigma_{\max}(\boldsymbol{\theta}|x). \quad (24)$$

By using the fact that $||\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}|| < 2\rho$ for any $\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \text{supp}(\phi|z^n)$, a direct application of Lemma 1 shows that

$$\exp \left\{ \mathcal{L}(\phi \to Y|x; z^n) \right\} \leq 1 + 2\rho K \sqrt{\sigma_{\max}^{(\rho)}(\hat{\theta}(z^n)|x)}. \quad (25)$$

Finally, for given $\delta, \epsilon > 0$ one calculates

$$\rho_\epsilon(x, z^n) = \min \{ \epsilon, \epsilon/\mathsf{C}_\epsilon(x; z^n) \}$$

with $\mathsf{C}_\epsilon(x; z^n) := 2K \sqrt{\sigma_{\max}^{(\epsilon)}(\hat{\theta}(z^n)|x)}$, which is well defined for small ϵsien la as $\boldsymbol{\theta}_0$ is an interior point. Then, noting that $\rho \leq \epsilon$ implies that $\sigma_{\max}^{(\rho)}(\hat{\theta}(z^n|x)) \leq \sigma_{\max}^{(\epsilon)}(\hat{\theta}(z^n|x))$, one can find that for all $n \geq n_{\boldsymbol{\theta}}(\rho_\epsilon(x; z^n), \delta)$ it is guaranteed that

$$\mathbb{E}(h_{\text{NML},\phi}; x, z^n) - \mathbb{E}(h_{\text{MAP}}; x) \leq \epsilon,$$

where the inequality holds for all $z^n \in B$. $\square$

APPENDIX D
PROOF OF COROLLARY 1

Proof. Let us denote as $\mathsf{T}(\boldsymbol{\theta}|x) := \sum_{i=1}^{d} \left[I(\boldsymbol{\theta}|x) \right]_{i,i}$ the trace of $I(\boldsymbol{\theta}|x)$, and $\mathsf{T}(\boldsymbol{\theta}) = \mathbb{E}\{\mathsf{T}(\boldsymbol{\theta}|X)\}$. Moreover, let us define

$$\mathsf{T}^{(\rho)}(\boldsymbol{\theta}_0|x) := \sup_{||\boldsymbol{\theta} - \boldsymbol{\theta}_0|| < \rho} \mathsf{T}(\boldsymbol{\theta}|x). \quad (26)$$

Then, by considering Eqs. (23) and (25) and noting that $\sigma_{\max}(\boldsymbol{\theta}|x) \leq \mathsf{T}(\boldsymbol{\theta}|x)$, one can show that

$$\mathbb{E} \{ \mathbb{E}(h_{\text{NML},\phi}; X, z^n) \} - \mathbb{E} \{ \mathbb{E}(h_{\text{MAP}}; X) \}$$

$$\leq 2\rho K \mathbb{E} \left\{ \sqrt{\sigma_{\max}^{(\rho)}(\hat{\theta}(z^n); X)} \right\}$$

$$\leq 2\rho K \mathbb{E} \left\{ \sqrt{\mathsf{T}^{(\rho)}(\hat{\theta}(z^n); X)} \right\}$$

$$\leq 2\rho K \sqrt{\mathsf{T}^{(\rho)}(\hat{\theta}(z^n))}.$$

The last step uses the well-known Jensen inequality. Finally, the corollary is proven by selecting $\rho_\epsilon(z^n) = \min\{\epsilon, \epsilon/\mathsf{D}_\epsilon(z^n)\}$ with $\mathsf{D}_\epsilon(z^n) := 2K \sqrt{\mathsf{T}^{(\epsilon)}(\hat{\theta}(z^n))}$. $\square$

Consider the plug-in hypothesis, which corresponds to our NML strategy with $\phi = \hat{\theta}$ and hence $h_{\text{NML},\phi} = h_{p_{\hat{\theta}}}$. Here we show that our proof of heuristic PAC learning cannot be applied — at least directly — to this case.

Let us consider how Theorem 1 could be used in this scenario. As in this case ϕ defines a particularly narrow model, i.e. $\text{supp}(\phi|z^n) = \{\hat{\theta}\}$, then is direct to verify that $\mathcal{L}(\phi \to Y|x, z^n) = 0$ and $\Delta(p_{\theta_0}, \text{supp}(\phi|z^n)|x) = \max_{y \in \mathcal{Y}} \ln f(y|x)/p_{\hat{\theta}}(y|x)$. While the consistency of $\hat{\theta}$ guarantees the convergence to zero of $\Delta(p_{\theta_0}, \text{supp}(\phi|z^n)|x)$ for each $x \in \mathcal{X}$, guaranteeing stronger types of convergence (which would be needed to prove heuristic PAC learning) is not straightforward. In particular, notice that to guarantee the convergence of $\sup_{x \in \mathcal{X}} \Delta(f, \text{supp}(\phi|z^n)|x)$ to zero as n grows, as one would need a function $C(\theta)$ such that for large n the following holds for all $x \in \mathcal{X}, y \in \mathcal{Y}$:

$$\ln p_{\theta_0}(y|x) - \ln p_{\hat{\theta}}(y|x) \le C(\theta_0) \cdot ||\theta_0 - \hat{\theta}||. \quad (27)$$

However, if the cardinality of $\mathcal{X}$ is infinite, it is possible to build examples where no such $C(\theta)$ exists, even if $p_{\hat{\theta}}(y|x) \to p_{\theta_0}(y|x)$ for each $x \in \mathcal{X}, y \in \mathcal{Y}$. This is a consequence of the fact that the derivative of the logarithm is unbounded within the interval $(0,1)$.

Proof. Under appropriate assumptions, the MLE $\hat{\theta}(Z^n)$ satisfies the Berry-Esseen bound [27]

$$\left| \mathbb{P}\left\{ \left\| \sqrt{n} I^{1/2}(\theta_0)\left(\hat{\theta}(Z^n) - \theta_0\right) \right\|^2 \le G_d^{-1}(t) \right\} - t \right| \le \frac{c}{\sqrt{n}},$$

where $G_d(\cdot)$ is the CDF of the chi-squared distribution with d degrees of freedom, and c is an absolute constant. Therefore, using the fact that

$$\sqrt{\sigma_{\min}(\theta_0)} \cdot ||\hat{\theta}(Z^n) - \theta_0|| \le ||I^{1/2}(\theta_0)\left(\hat{\theta}(Z^n) - \theta_0\right)||,$$

one can show that

$$\mathbb{P}\left\{ n\sigma_{\min}(\theta_0) \left\| \hat{\theta}(Z^n) - \theta_0 \right\|^2 \le G_d^{-1}(t) \right\} \ge t - \frac{c}{\sqrt{n}}.$$

Then, by taking $t = 1 - \delta + cn^{-1/2}$, and assuming that n is large enough so that $t \in [0,1]$, then one can find that

$$\mathbb{P}\left\{ \left\| \hat{\theta}(Z^n) - \theta_0 \right\| \le \sqrt{\frac{G_d^{-1}(1 - \delta + cn^{-1/2})}{n\sigma_{\min}(\theta_0)}} \right\} \ge 1 - \delta.$$

Note that $\sigma_{\min}(\theta_0) > 0$ because $I(\theta_0)$ is assumed to be positive definite.

Therefore, by considering $\psi := \hat{\theta}(z^n) + W$ with W uniformly distributed over a ball of radius

$$\rho_n := \sqrt{\frac{G_d^{-1}(1 - \delta + cn^{-1/2})}{n\sigma_{\min}(\theta_0)}}, \quad (28)$$

then $\theta_0 \in \text{supp}(\psi|z^n)$ for all $z^n \in B \subset \mathcal{Z}^n$ with $\mathbb{P}\{Z^n \in B\} = 1 - \delta$. Then, $\Delta(\text{supp}(\psi|z^n), f) = 0$ for all $z^n \in B$, and hence the first inequality in (21) can be proven using Theorem 1.

For proving the second inequality, note that a direct application of Lemma 1 shows that

$$\exp\left\{ \mathcal{L}(\psi \to Y|x; z^n) \right\} \le 1 + \frac{2}{\sqrt{n}} \cdot \sqrt{\frac{\sigma_{\max}^{(\rho_n)}(\hat{\theta}(z^n)|x)}{\sigma_{\min}(\theta_0)}},$$

with $\sigma_{\max}^{(\rho)}$ defined as in Eq. (24). Furthermore, by noting that by construction of ρ_n it is guaranteed that $||\hat{\theta}(Z^n) - \theta_0|| \le \rho_n$ with probability $1 - \delta$, then $\sigma_{\max}^{(\rho_n)}(\hat{\theta}(z^n)|x) \le \sigma_{\max}^{(2\rho_n)}(\theta_0|x)$, which in turn implies that

$$\exp\left\{ \mathcal{L}(\psi \to Y|x; z^n) \right\} \le 1 + \frac{2}{\sqrt{n}} \cdot \sqrt{\frac{\sigma_{\max}^{(2\rho_n)}(\theta_0|x)}{\sigma_{\min}(\theta_0)}}.$$

Finally, the proof concludes by noting that ρ_n, and hence also $\sigma_{\max}^{(\rho_n)}$, decrease with n.

$\square$

Proof. Let us consider $\epsilon, \delta > 0$, and define $\phi := \hat{\theta}(Z^n) + W_\rho \in \mathbb{R}^d$ with W_ρ distributed uniformly over a d-dimensional ball of radius $\rho > 0$. By the properties of $\hat{\theta}$, there exists $n_0(\delta, \rho) \in \mathbb{N}$ such that for all $n > n_0$ then there exists $\theta_0 \in \Theta_f \cap \text{supp}(\phi|z^n)$, for all $z^n \in B$ with $\mathbb{P}\{Z^n \in B\} > 1 - \delta$. Then, it is direct to check that $\Delta(f, \text{supp}(\phi|z^n)) < \epsilon_0$ for all $z^n \in B$. Additionally, following the proof of Theorem 2 (in particular, the derivation that leads to Eq.(25)), one can check that the fact that $\text{supp}(\phi|z^n)$ has a bounded support implies that

$$\exp\left\{ \mathcal{L}(\phi \to Y|x; z^n) \right\} \le 1 + 2\rho K \sqrt{\sigma_{\max}^{(\rho)}(\hat{\theta}(z^n)|x)},$$

with $\sigma_{\max}^{(\rho)}(\theta|x)$ as defined in Eq. (24).

With these results at hand, let us now choose $\rho(x; z^n) = \min\{\epsilon_0, \epsilon_0/\text{C}_{\epsilon_0}(x; z^n)\}$ with $\text{C}_{\epsilon_0}(x; z^n) := 2K\sqrt{\sigma_{\max}^{(\epsilon_0)}(\hat{\theta}(z^n)|x)}$. Using these results and Theorem 1, and the fact that $e^x \approx 1 + x$ for $1 \gg |x|$, one finds that for all $n > n_0(\delta, \rho(x; z^n))$ then

$$\mathbb{E}(h_{\text{NML},\phi}; x, z^n) - \mathbb{E}(h_{\text{MAP}}; x) \le e^{\Delta(f, \text{supp}(\phi|z^n))} e^{\mathcal{L}(\phi \to Y|x; z^n)} - 1$$
$$\le 2\epsilon_0 + \epsilon_0^2.$$

Finally, the proof concludes by selecting ϵ_0 such that

$$2\epsilon_0 + \epsilon_0^2 < \epsilon. \quad (29)$$

$\square$

REFERENCES

[1] F. M. Willems, Y. M. Shtarkov, and T. J. Tjalkens, "The context-tree weighting method: Basic properties," *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 653–664, 1995.

[2] R. Begleiter, R. El-Yaniv, and G. Yona, "On prediction using variable order Markov models," *Journal of Artificial Intelligence Research*, vol. 22, pp. 385–421, 2004.

[3] J. Rissanen, *Stochastic Complexity in Statistical Inquiry.* World Scientific, 1989.

[4] P. D. Grünwald, *The Minimum Description Length Principle.* MIT press, 2007.

[5] Y. M. Shtar'kov, "Universal sequential coding of single messages," *Problemy Peredachi Informatsii*, vol. 23, no. 3, pp. 3–17, 1987.

[6] J. Rissanen, "Fisher information and stochastic complexity," *IEEE Transactions on Information Theory*, vol. 42, no. 1, pp. 40–47, Jan 1996.

[7] Y. Fogel and M. Feder, "Universal learning of individual data," in *2019 IEEE International Symposium on Information Theory (ISIT)*, July 2019, pp. 2289–2293.

[8] K. Bibas, Y. Fogel, and M. Feder, "A new look at an old problem: A universal learning approach to linear regression," in *2019 IEEE International Symposium on Information Theory (ISIT)*, July 2019, pp. 2304–2308.

[9] K. Bibas, Y. Fogel, and M. Feder, "Deep pNML: Predictive normalized maximum likelihood for deep neural networks," *arXiv:1904.12286*, 2019.

[10] T. Roos and J. Rissanen, "On sequentially normalized maximum likelihood models," *Compare*, vol. 27, no. 31, p. 256, 2008.

[11] I. Issa, S. Kamath, and A. B. Wagner, "An operational measure of information leakage," in *2016 Annual Conference on Information Science and Systems (CISS)*, March 2016, pp. 234–239.

[12] B. Rassouli, F. E. Rosas, and D. Gündüz, "Data disclosure under perfect sample privacy," *IEEE Transactions on Information Forensics and Security*, pp. 1–1, 2019.

[13] C. Dwork, A. Roth *et al.*, "The algorithmic foundations of differential privacy," *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[14] I. Issa, A. R. Esposito, and M. Gastpar, "Strengthened information-theoretic bounds on the generalization error," in *2019 IEEE International Symposium on Information Theory (ISIT)*, July 2019, pp. 582–586.

[15] A. R. Esposito, M. Gastpar, and I. Issa, "A new approach to adaptive data analysis and learning via maximal leakage," *arXiv:1903.01777*, 2019.

[16] H. V. Poor, *An Introduction to Signal Detection and Estimation.* Springer Science & Business Media, 2013.

[17] F. du Pin Calmon and N. Fawaz, "Privacy against statistical inference," in *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2012, pp. 1401–1408.

[18] I. Issa, A. B. Wagner, and S. Kamath, "An operational approach to information leakage," *arXiv:1807.07878*, 2018.

[19] R. Sibson, "Information radius," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, vol. 14, no. 2, pp. 149–160, Jun 1969.

[20] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms.* Cambridge University Press, 2014.

[21] D. Haussler, *Probably Approximately Correct Learning.* University of California, Santa Cruz, Computer Research Laboratory, 1990.

[22] K. Knight, *Mathematical Statistics.* Chapman and Hall/CRC, 1999.

[23] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural Networks*, vol. 4, no. 2, pp. 251–257, 1991.

[24] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of Control, Signals, and Systems*, vol. 2, pp. 303–314, 1989.

[25] H. White, "Learning in artificial neural networks: A statistical perspective," *Neural Computation*, vol. 4, pp. 425–464, 1989.

[26] S. Lang, *Calculus of Several Variables.* Springer Science & Business Media, 2012.

[27] J. Pfanzagl, "The accuracy of the normal approximation for estimates of vector parameters," *Z. Wahrscheinlichkeitstheorie verw. Geb*, vol. 25, pp. 171–198, 1973.