Alpha-NML Universal Predictors

Marco Bondaschi, *Graduate Student Member, IEEE,* and Michael Gastpar, *Fellow, IEEE*

Abstract

Inspired by the connection between classical regret measures employed in universal prediction and Rényi divergence, we introduce a new class of universal predictors that depend on a real parameter $\alpha \geq 1$. This class interpolates two well-known predictors, the mixture estimators, that include the Laplace and the Krichevsky-Trofimov predictors, and the Normalized Maximum Likelihood (NML) estimator. We point out some advantages of this new class of predictors and study its benefits from two complementary viewpoints: (1) we prove its optimality when the maximal Rényi divergence is considered as a regret measure, which can be interpreted operationally as a middle ground between the standard average and worst-case regret measures; (2) we discuss how it can be employed when NML is not a viable option, as an alternative to other predictors such as Luckiness NML. Finally, we apply the α-NML predictor to the class of discrete memoryless sources (DMS), where we derive simple formulas to compute the predictor and analyze its asymptotic performance in terms of worst-case regret.

Index Terms

Universal prediction, universal compression, Normalized Maximum Likelihood, Sibson's mutual information, Rényi capacity, redundancy-capacity theorem.

I. Introduction

Prediction refers to the general problem of estimating the next symbols of a sequence given its past, and evaluating the confidence of such an estimate. This problem appears in a large number of research areas, such as information theory, statistical decision theory, finance, and machine learning. Some knowledge about the probability distribution that models the sequence one wishes to predict is clearly helpful. Unfortunately, in many practical applications such knowledge is missing. If this is the case, then one may wish to say something about the future of the sequence when the true model of the source that is producing the symbols is *any* of the models belonging to a certain class. This problem usually goes under the name of *universal prediction* [1]. It has applications in a wide range of areas, such as compression [2], [3], gambling [4] and machine learning [5], [6].

The formal statement of universal prediction that we consider in this work is the following. For any $n \geq 1$, we assume that a sequence $x^{n-1} = (x_1, x_2, \dots, x_{n-1})$ of $n-1$ symbols from a given (possibly infinite) alphabet $\mathcal{X}$ has been generated by some unknown (random or deterministic) source. Suppose that we design a predictor

The authors are with the School of Computer and Communication Sciences, École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland (e-mail: {marco.bondaschi, michael.gastpar}@epfl.ch).

This work was presented in part at the 2022 IEEE International Symposium on Information Theory, Espoo, Finland, Jun. 2022.

that, given the past symbols of the sequence, returns some numerical prediction about the next symbol x_n. This prediction may be an estimation $\hat{x}_n$ of next symbol itself, or it may also be something more informative, such as an estimation of the probability distribution of the next symbol. The latter case carries the additional information of the *confidence* associated to the estimation, in terms of how probable our best guess on the next symbol is.

In order to evaluate the quality of the prediction, one uses a so-called *loss function* ℓ that maps the pair formed by the prediction and the actual symbol x_n, to a real number. In this paper, we follow the classical perspective where the predictor assigns probabilities to the possible values of the next outcome x_n [1], [7]. In such a case, one usually chooses as a loss function some value that is inversely proportional to the estimated probability of x_n. The reason for such a choice is that, if the source generates frequently symbols to which our predictor assigned a low probability, then the measured loss is high, signaling that our predictor is bad; on the contrary, if the source generates symbols to which the predictor assigned high probability, then the loss is small.

A very popular choice, mainly due to its connection with universal compression, is the *logarithmic loss*. If $\hat{p}(\cdot|x_1^{n-1})$ is the probability distribution on the next symbol estimated by the predictor, then the associated logarithmic loss is defined as

$$\ell(\hat{p}, x_n) \triangleq \log \frac{1}{\hat{p}(x_n|x^{n-1})}. \tag{1}$$

If the quality of the predictor is measured on more than one symbol, for example the whole sequence x^n, then one can take as a performance measure the *cumulative loss* L, which is the sum of the losses of the n symbols. In the case of the logarithmic loss, one has

$$\begin{aligned} L(\hat{p}, x^n) &\triangleq \sum_{i=1}^{n} \ell(\hat{p}, x_i) && (2)\\ &= \sum_{i=1}^{n} \log \frac{1}{\hat{p}(x_i|x^{i-1})} && (3)\\ &= \log \frac{1}{\hat{p}(x^n)} && (4) \end{aligned}$$

where $\hat{p}(x^n) \triangleq \prod_{i=1}^{n} \hat{p}(x_i|x^{i-1})$ can be defined as the joint estimated probability of the entire sequence x^n. In the remainder of the paper, we take our loss function to be the logarithmic loss.

Let us now consider a given class of distributions $\mathcal{P} = \{p_\theta : \theta \in \Theta\}$ indexed by a parameter set Θ, and let us assume that the actual source belongs to this class, or, less strictly speaking, that this class is the one we want to compare our predictor to. Usually, Θ is a subset of $\mathbb{R}^d$ for some $d \geq 1$, and $\theta \in \Theta$ is the parameter vector of some parametric family, e.g., discrete memoryless sources, Markov sources of order k, auto-regressive sources, a certain exponential family, etc. When building a predictor for sequences of symbols, one needs a metric or a criterion that measures the quality of the predictor by taking into consideration the different possible sequences x^n, as well as the possible sources of the class $\mathcal{P}$. To construct such a measure, one usually starts from the difference between the logarithmic loss of the predictor $\hat{p}$ and that of a distribution p_θ in $\mathcal{P}$, that is,

$$\begin{aligned} R(\hat{p}, p_\theta, x^n) &\triangleq \log \frac{1}{\hat{p}(x^n)} - \log \frac{1}{p_\theta(x^n)} && (5)\\ &= \log \frac{p_\theta(x^n)}{\hat{p}(x^n)}, && (6) \end{aligned}$$

which is usually called *regret*. Two regret measures that are generally employed to assess the quality of a predictor are the *average regret*

$$\begin{aligned} R_{\mathrm{av}}(\hat{p}) &\triangleq \sup_{\theta\in\Theta} \mathbb{E}_\theta\left[R(\hat{p}, p_\theta, X^n)\right] && (7)\\ &= \sup_{\theta\in\Theta} \int_{\mathcal{X}^n} p_\theta(x^n) \log\frac{p_\theta(x^n)}{\hat{p}(x^n)}\, dx^n && (8)\\ &= \sup_{\theta\in\Theta} D(p_\theta \| \hat{p}), && (9) \end{aligned}$$

and the *worst-case regret*

$$\begin{aligned} R_{\max}(\hat{p}) &\triangleq \sup_{\theta\in\Theta} \sup_{x^n\in\mathcal{X}^n} R(\hat{p}, p_\theta, x^n) && (10)\\ &= \sup_{\theta\in\Theta} \sup_{x^n\in\mathcal{X}^n} \log\frac{p_\theta(x^n)}{\hat{p}(x^n)} && (11)\\ &= \sup_{\theta\in\Theta} D_\infty(p_\theta\|\hat{p}) && (12) \end{aligned}$$

where $D_\infty(p_\theta\|\hat{p})$ is the Rényi divergence of order infinity. The maximization over all parameters in Θ that appears in the considered definitions of regret comes from the fact that, in the universal prediction setting, one generally considers the case where no prior knowledge on the parameters is available, that is, no source in $\mathcal{P}$ is considered a better candidate to be the true one in advance.

It is well known [8] that the predictor that minimizes the worst-case regret is the Normalized Maximum Likelihood (NML) estimator, whenever it exists. Its formula is

$$\hat{p}_{\mathrm{NML}}(x^n) = \frac{\sup_{\theta\in\Theta} p_\theta(x^n)}{\int_{\mathcal{X}^n} \sup_{\theta\in\Theta} p_\theta(x^n)\, dx^n}. \tag{13}$$

Even if it has a nice closed-form expression, in general the NML has several disadvantages, including the fact that it may not exist since the integral in the denominator in (13) may not converge, and the necessity of computing a maximization over the parameter space Θ.

These limitations led researchers to look for good alternatives to the NML predictor. For the class of discrete memoryless sources over a finite alphabet $\mathcal{X} = \{1, 2, \dots, m\}$, such an alternative is the Krichevsky-Trofimov estimator [9], which assigns as a probability for the next symbol $k \in \{1, 2, \dots, m\}$ a value proportional to

$$\hat{p}_{\mathrm{KT}}(k|x^{n-1}) \propto n_k + \frac{1}{2}, \tag{14}$$

where n_k is the number of k's in the past sequence x^{n-1}.

As opposed to NML, the KT predictor is not affected by the disadvantages listed above. Furthermore, it turns out that it achieves, for the class of discrete memoryless sources, the same asymptotic regret, up to a constant term, as the NML when $n \to \infty$ [4]. However, no similar results are proved for other classes of distributions, and also, the NML estimator performs better in general when n is finite. For these reasons, the search for alternative predictors that have fewer drawbacks (e.g., broader applicability) with respect to the NML estimator is still important.

The contribution of this paper is the introduction of a class of predictors inspired by Sibson's α-mutual information and the connection between regret measures used in universal prediction and Rényi divergence, that we term α-

NML predictors. This class is parametrized by $\alpha \geq 1$ and its definition depends on the choice of a prior probability distribution w over the parameter space Θ:

$$\hat{p}_\alpha(x^n) \triangleq \frac{\left\{\int_\Theta w(\theta)\, p_\theta^\alpha(x^n)\, d\theta\right\}^{1/\alpha}}{\int_{\mathcal{X}^n} \left\{\int_\Theta w(\theta)\, p_\theta^\alpha(\bar{x}^n)\, d\theta\right\}^{1/\alpha}\, d\bar{x}^n}\,. \tag{15}$$

As an example, for DMS this class interpolates between the KT estimator and the NML. For $\alpha = 1$, our predictor gives the same probability estimation (14) as the KT predictor. For $\alpha = 2$, it assigns a probability that is proportional to

$$\hat{p}_{\alpha=2}(k|x^{n-1}) \propto \sqrt{\left(n_k + \frac{1}{4}\right)\left(n_k + \frac{3}{4}\right)}\,. \tag{16}$$

Finally, when $\alpha \to \infty$, we retrieve the classical NML formula. The general conditional formula as a function of α is given in Equation (100).

In the paper, we study the α-NML predictors from two complementary perspectives. The first one is to investigate its performance when Rényi divergence is used as a regret measure. We call such regret measures α-regret; they can be interpreted operationally as a middle ground between the classical average regret and the worst-case regret, which are widely employed in universal prediction. In fact, both the average regret and the worst-case regret can be written as a maximization of a Rényi divergence – see (9) and (12). If the maximization of a Rényi divergence of any order α between these two extreme cases is taken as a regret measure, then it turns out that α-NML is the optimal predictor, provided that the proper prior distribution on the parameter space Θ is chosen.

The second perspective is to look at the α-NML predictors as an alternative to NML, when the latter cannot be used because, e.g., it does not exist. In fact, α-NML converges to NML when $\alpha \to \infty$. However, if a finite α is chosen, it can be shown that α-NML exists for some classes of distributions for which the NML does not. As a further example, we analyze α-NML's performance in terms of worst-case regret, which is the measure under which the NML is optimal, and we investigate how much we pay in terms of regret with respect to NML, and how much we gain with respect to the KT predictor, as a function of the parameter α, when the class of discrete memoryless sources is considered. For the binary alphabet case, the performance improvement of the new predictor is illustrated numerically in Figure 1.

A. Related work

The worst-case regret and the Normalized Maximum Likelihood predictor were first studied in [8]. The Krichevsky-Trofimov predictor was introduced in [9] for binary sources, and it was generalized to general finite alphabets in [4], where its asymptotic worst-case regret is also analyzed. A summary of the properties of Sibson's α-mutual information can be found in [10]. The problem of maximizing Sibson's mutual information is studied in [11], where a result similar to Theorem 1 of this paper is derived by different means. In [12], a different regret measure, also based on the Rényi divergence, is introduced. In the same paper, the authors show that, in the case of discrete memoryless sources with finite alphabet, their regret measure is equivalent to the α-regret defined in (25). For this particular case, the authors also derive the asymptotical value of the regret as the sequence length goes to infinity. This result was used in the proof of Theorem 5 here. In [13], the authors study the minimax regret with the additional constraint that the point-wise regret be an integer, which naturally arises from a universal compression perspective, where codeword lengths are considered. Related but different modified regret measures have been studied in several works, including [14].

B. Overview

The remainder of the paper is organized as follows. In Section II we introduce α-regret, which is defined in terms of Rényi divergence and described from an operational point of view. In Section III, we introduce the class of α-NML predictors and we prove their optimality under α-regret. In Section IV, we discuss the advantages of using α-NML as an alternative to NML, when the latter cannot be used, and we compare it to other alternatives to NML such as Luckiness NML. In Section V, we apply α-NML to the parametric family of discrete memoryless sources (DMS), deriving some simple closed-form formulae to compute the probabilities estimated by the predictor, and studying its performance in terms of worst-case regret, discussing how much we pay by using α-NML instead of NML in this setting.

C. Notation

We use upper case letters X to denote random variables and lower case letters x to denote their realizations. A sequence of length n is denoted by a superscript X^n. Probability distributions are denoted by $p_\theta(x)$, where $\theta \in \Theta$ is a parameter. The Rényi divergence of order α between two probability distributions p_θ and $\hat{p}$ on some alphabet $\mathcal{X}^n$ is denoted as

$$D_\alpha(p_\theta \| \hat{p}) \triangleq \frac{1}{\alpha-1} \log \int_{\mathcal{X}^n} p_\theta(x^n) \left(\frac{p_\theta(x^n)}{\hat{p}(x^n)} \right)^{\alpha-1} dx^n. \tag{17}$$

Sibson's α-mutual information [15] (see also [10]) between two random variables $(X, Y) \sim p_X \times p_{Y|X}$ on $\mathcal{X} \times \mathcal{Y}$ is denoted as

$$I_\alpha(X, Y) \triangleq \frac{\alpha}{\alpha-1} \log \int_{\mathcal{Y}} \left\{ \int_{\mathcal{X}} p_X(x)\, p_{Y|X}^\alpha(y|x)\, dx \right\}^{1/\alpha} dy. \tag{18}$$

In the limit $\alpha \to \infty$, the two previous definitions become

$$D_\infty(p_\theta \| \hat{p}) = \sup_{x^n} \log \frac{p_\theta(x^n)}{\hat{p}(x^n)} \tag{19}$$

and

$$I_\infty(X, Y) = \log \int_{\mathcal{Y}} \sup_{x \in \operatorname{supp}(X)} p_{Y|X}(y|x)\, dy \tag{20}$$

where $\operatorname{supp}(X)$ is the support of the random variable X.

II. α-REGRET

The average regret (8) and the worst-case regret (11) are the two most widely employed regret measures in universal prediction. One can interpret them operationally as follows. Suppose that a source S_θ generates sequences of n symbols from an alphabet $\mathcal{X}$ according to a distribution p_θ on $\mathcal{X}^n$, for some parameter $\theta \in \Theta$. For a predictor $\hat{p}$, the regret, i.e., the difference in logarithmic loss, for a given sequence $x^n \in \mathcal{X}^n$ is $R_\theta(\hat{p}, x^n) =$

$\log \frac{1}{\hat{p}(x^n)} - \log \frac{1}{p_\theta(x^n)} = \log \frac{p_\theta(x^n)}{\hat{p}(x^n)}$. One then has the choice of how much weight to give to each sequence x^n with respect to the others. The two approaches that we discussed before are the following.

- Giving weight to each sequence x^n according to its probability. With this choice we obtain the average regret (8):
$$R_{\mathrm{av},\theta}(\hat{p}) \triangleq \mathbb{E}_\theta \left[\log \frac{p_\theta(X^n)}{\hat{p}(X^n)} \right] = \int_{\mathcal{X}^n} p_\theta(x^n) \log \frac{p_\theta(x^n)}{\hat{p}(x^n)} \, dx^n = D(p_\theta \| \hat{p}). \tag{21}$$
- Considering only the sequence x^n with the highest regret. In this case we get the worst-case regret (11):
$$R_{\max,\theta}(\hat{p}) \triangleq \sup_{x^n} \log \frac{p_\theta(x^n)}{\hat{p}(x^n)} = D_\infty(p_\theta, \hat{p}). \tag{22}$$

Essentially, in the first case we are measuring the goodness of the predictor $\hat{p}$ in terms of its average regret when the sequences are generated by S_θ according to p_θ, while in the second case we only consider the regret for the worst possible sequence. The two measures (21) and (22) can be recognized as two extreme cases. The former weighs the sequences only according to their probability, without taking into account the amount of regret each sequence carries. Instead, the latter puts all the weight on the worst sequence, without considering how probable it is for this sequence to actually occur. A natural interpolation between these two cases is given by the following α-regret, which is equal to the Rényi divergence of order α.

Definition 1: The α-regret of a predictor $\hat{p}$ with respect to a distribution p_θ is equal to
$$R_{\alpha,\theta}(\hat{p}) \triangleq D_\alpha(p_\theta \| \hat{p}) = \frac{1}{\alpha - 1} \log \int_{\mathcal{X}^n} p_\theta(x^n) \left(\frac{p_\theta(x^n)}{\hat{p}(x^n)} \right)^{\alpha - 1} dx^n. \tag{23}$$

Note that $R_{\alpha=1,\theta}(\hat{p}) = R_{\mathrm{av},\theta}(\hat{p})$ and $R_{\alpha=\infty,\theta}(\hat{p}) = R_{\max,\theta}(\hat{p})$. This regret measure takes into account all the sequences according to their probability, and at the same time gives some additional penalty to sequences with large regret. In fact, setting $\lambda = \alpha - 1$, one can rewrite this measure as
$$R_{1+\lambda,\theta}(\hat{p}) \triangleq D_{1+\lambda}(p_\theta \| \hat{p}) = \frac{1}{\lambda} \log \int_{\mathcal{X}^n} p_\theta(x^n) \exp \left(\lambda \log \frac{p_\theta(x^n)}{\hat{p}(x^n)} \right) dx^n \tag{24}$$
which is an exponential average of the regrets: the larger the value assigned to the parameter λ, the more importance is given to the regret of each sequence x^n in determining the weighting. Finally, maximizing the regret measures (21) and (22) over all possible sources gives back the definitions of $R_{\mathrm{av}}(\hat{p})$ and $R_{\max}(\hat{p})$ in Equations (8) and (11). Doing the same for α-regret gives
$$R_\alpha(\hat{p}) \triangleq \sup_{\theta \in \Theta} D_\alpha(p_\theta \| \hat{p}) = \sup_{\theta \in \Theta} \frac{1}{\alpha - 1} \log \int_{\mathcal{X}^n} p_\theta(x^n) \left(\frac{p_\theta(x^n)}{\hat{p}(x^n)} \right)^{\alpha - 1} dx^n \tag{25}$$
where the parameter α in (25) and λ in (24) are related by the equation $\alpha = 1 + \lambda$.

It is worth noting that the exponential dependency of α-regret is also related to the tilted losses considered in machine learning, see e.g. [14]. Moreover, it has a similar flavor to the codeword length measure that Campbell studied in [16] for compression. In fact, it is well known that there is a strong connection between prediction and compression when the logarithmic loss is used. A classical variable-length coding problem is to find a uniquely decodable code for the symbols in $\mathcal{X}^n$ that minimizes the expected length
$$L_{\mathrm{av},\theta} \triangleq \mathbb{E}_\theta[\ell(X^n)] = \int_{\mathcal{X}^n} p_\theta(x^n) \, \ell(x^n) \, dx^n \tag{26}$$

where $\ell(x^n)$ is the length of the codeword associated to the sequence x^n, and X^n is distributed according to a given p_θ. It is well known that the code that minimizes this quantity is any code that associates to the sequence x^n a codeword with length $\ell(x^n) = \log \frac{1}{p_\theta(x^n)}$. If the code is instead constructed using a different distribution $\hat{p}(x^n)$, assigning to x^n a codeword of length $\hat{\ell}(x^n) = \log \frac{1}{\hat{p}}(x^n)$, the penalty payed in number of bits would be exactly the regret $R_\theta(\hat{p}, x^n) = \log \frac{p_\theta(x^n)}{\hat{p}(x^n)}$ seen before. In [16], Campbell was looking for the optimal code that minimizes an alternative measure to (26) in which an exponential dependency on the codeword lengths $\ell(x^n)$ is introduced, instead of the usual expected codeword length. This is different than the setting we consider here, from which α-regret emerges. In fact, from a compression perspective, in this work we are still considering the optimal code with respect to the classical expected codeword length, and the exponential dependency is on the difference in bits (i.e., the regret) that the designed code uses with respect to the optimal one (in the usual expected codeword length sense).

III. α-NML PREDICTORS

A. *Definition*

With the following definition, we introduce a novel class of universal predictors, that we call α-NML predictors.

Definition 2: Let $\mathcal{P} = \{p_\theta : \theta \in \Theta\}$ be a parametric class of distributions on an alphabet $\mathcal{X}^n$. Let w be any probability distribution on Θ. For any $\alpha \geq 1$, the α-NML predictor is defined as

$$\hat{p}_\alpha(x^n) \triangleq \frac{\left\{\int_\Theta w(\theta)\, p_\theta^\alpha(x^n)\, d\theta\right\}^{1/\alpha}}{\int_{\mathcal{X}^n} \left\{\int_\Theta w(\theta)\, p_\theta^\alpha(\bar{x}^n)\, d\theta\right\}^{1/\alpha} d\bar{x}^n}. \tag{27}$$

Note that the definition of α-NML also depends on the class of distributions $\mathcal{P}$ and on the prior distribution w on the parameter space Θ. We omit this dependence to ease the notation, since it will be made clear from the context. The α-NML class is a continuous interpolation between the NML predictor and another very popular class of predictors. In fact, taking $\alpha = 1$ gives

$$\hat{p}_1(x^n) = \int_\Theta w(\theta)\, p_\theta(x^n)\, d\theta \tag{28}$$

which is the well-known class of *mixture estimators*. Taking instead the limit $\alpha \to \infty$ gives

$$\hat{p}_\infty(x^n) = \frac{\sup_{\theta \in \mathrm{supp}(w)} p_\theta(x^n)}{\int_{\mathcal{X}^n} \sup_{\theta \in \mathrm{supp}(w)} p_\theta(x^n)\, dx^n}. \tag{29}$$

When the support of the prior distribution w is the entire parameter space Θ, the α-NML for $\alpha = \infty$ is precisely the NML predictor defined in (13). It is important to highlight the role of Sibson's α-mutual information in how α-NML is defined. Note that the denominator of the NML predictor in (13) is $\exp I_\infty(\phi, X^n)$. The α-NML definition ideally follows by replacing Sibson's mutual information of order infinity with any other order $\alpha \geq 1$ in the predictor's denominator. In fact, the denominator in (27) is precisely $\exp \frac{\alpha-1}{\alpha} I_\alpha(\phi, X^n)$, where Sibson's mutual information I_α is defined in (18).

Definition 2 begs two fundamental questions. Namely, *(i),* the conditions on $\mathcal{P}, w(\theta)$ and α under which the α-NML predictor exists, and *(ii)*, the criteria for choosing $w(\theta)$ and α. We defer the discussion on this topic to Section IV, where we compare existence conditions for the α-NML to those for the NML, in order to investigate the use of α-NML as an alternative to NML, when the latter does not converge.

B. *Optimality with respect to α-regret*

It is a well-known fact (see, e.g., [17, Thm. 37]) that the NML defined in (13), whenever it exists, achieves the minimum worst-case regret (11), which is equal to

$$R_{\max}(\hat{p}_{\text{NML}}) = \min_{\hat{p}} R_{\max}(\hat{p}) = I_\infty(\phi, X^n) = \log \int_{\mathcal{X}} \sup_{\theta \in \Theta} p_\theta(x^n)\, dx^n. \tag{30}$$

Furthermore, it is known and it has been proved several times in different contexts (see [1] and references therein) that, under certain conditions, a mixture predictor of the form (28) is optimal under average regret as defined in (8), for a proper choice of the prior distribution $w(\theta)$. Since α-regret is an interpolation between average and worst-case regret, and, furthermore, the α-NML predictor is an interpolation between the mixture predictor and the NML, it is natural to ask whether the α-NML is actually optimal under α-regret. The following theorem shows that, under suitable conditions, this is indeed the case, when the proper prior distribution $w(\theta)$ is chosen. A similar result with a different proof is shown in [11], where the problem of the maximization of Sibson's α-mutual information is investigated.

Theorem 1: Let $\mathcal{X}$ be a (possibly infinite) alphabet set, and let $\mathcal{P} = \{p_\theta : \theta \in \Theta\}$ be a parametric class of distributions on $\mathcal{X}^n$. Let $(\phi, X^n) \sim w \times p_\phi(X^n)$ for some prior distribution w on Θ, and suppose that there exists a probability distribution w^* on Θ such that, for $(\phi^*, X^n) \sim w^* \times p_{\phi^*}(X^n)$

$$I_\alpha(\phi^*, X^n) = \sup_{w : \phi \sim w} I_\alpha(\phi, X^n). \tag{31}$$

Then, the α-NML defined in (27) with prior w^*, i.e.,

$$\hat{p}_\alpha(x^n) = \frac{\left\{\int_\Theta w^*(\theta) p_\theta^\alpha(x^n)\, d\theta\right\}^{1/\alpha}}{\sum_{x^n} \left\{\int_\Theta w^*(\theta) p_\theta^\alpha(x^n)\, d\theta\right\}^{1/\alpha}}, \tag{32}$$

minimizes $R_\alpha(\hat{p})$ over all probability distributions on $\mathcal{X}^n$. Furthermore, the minimal α-regret is equal to

$$\min_{\hat{p}} R_\alpha(\hat{p}) = I_\alpha(\phi^*, X^n). \tag{33}$$

Proof: The case for $\alpha = 1$ is well known and was first proved by Gallager in [18]. We prove the theorem for $\alpha > 1$, following an idea similar to Gallager's. Let $C_\alpha \triangleq \sup_w I_\alpha(\phi, X^n)$. We want to prove that

$$D_\alpha(p_\theta \| \hat{p}_\alpha) \leq C_\alpha \tag{34}$$

for every $\theta \in \Theta$. By contradiction, suppose that there exists $\bar{\theta} \in \Theta$ such that,

$$D_\alpha(p_{\bar{\theta}} \| \hat{p}_\alpha) > C_\alpha. \tag{35}$$

For any $0 \leq t \leq 1$, define the probability distribution

$$w^*_{\bar{\theta},t} = (1-t) w^* + t \delta_{\bar{\theta}} \tag{36}$$

where $\delta_{\bar{\theta}}$ is the singular distribution centered on $\bar{\theta}$. Then, we have

$$f(t) \triangleq (\alpha - 1) I_\alpha(\phi, X^n)\big|_{\phi \sim w^*_{\bar{\theta}}} = \alpha \log \int_{\mathcal{X}^n} \left\{ t p_{\bar{\theta}}^\alpha(x^n) + (1-t) \int_\Theta w^*(\theta) p_\theta^\alpha(x^n)\, d\theta \right\}^{1/\alpha} dx^n. \tag{37}$$

By the assumption that w^* is the maximizer of $I_\alpha(\phi, X^n)$, $f(t)$ is maximized at $t = 0$. Taking the derivative of $f(t)$ with respect to t gives

$$f'(t) = \frac{\int_{\mathcal{X}^n} \left(p_\theta^\alpha(x^n) - \int_\Theta w^*(\theta) p_\theta^\alpha(x^n) d\theta\right) \left\{\int_\Theta w^*(\theta) p_\theta^\alpha(x^n) d\theta\right\}^{\frac{1-\alpha}{\alpha}} dx^n}{\int_{\mathcal{X}^n} \left\{t p_\theta^\alpha(x^n) + (1-t) \int_\Theta w^*(\theta) p_\theta^\alpha(x^n) d\theta\right\}^{1/\alpha} dx^n} \tag{38}$$

Evaluating this derivative in $t = 0$ gives

$$\begin{aligned} f'(0) &= \frac{\int_{\mathcal{X}^n} p_\theta^\alpha(x^n) \left\{\int_\Theta w^*(\theta) p_\theta^\alpha(x^n) d\theta\right\}^{\frac{1-\alpha}{\alpha}} dx^n}{\int_{\mathcal{X}^n} \left\{\int_\Theta w^*(\theta) p_\theta^\alpha(x^n) d\theta\right\}^{1/\alpha} dx^n} - 1 & (39) \\ &= \exp\left\{(\alpha - 1)(D_\alpha(p_\theta \| \hat{p}_\alpha) - C_\alpha)\right\} - 1 > 0. & (40) \end{aligned}$$

This contradicts the fact that $f(t)$ is maximized at $t = 0$, so we proved that $D_\alpha(p_\theta \| \hat{p}_\alpha) \leq C_\alpha$ for every θ. Hence,

$$R_\alpha(\hat{p}_\alpha) = \max_\theta D_\alpha(p_\theta \| \hat{p}_\alpha) \leq C_\alpha. \tag{41}$$

However, it is known [12] that $\min_{\hat{p}} R_\alpha(\hat{p}) = C_\alpha$, which proves that $\hat{p}_\alpha$ with prior w^* is indeed a minimizer of $R_\alpha(\hat{p})$. Substituting (32) into the definition of α-regret (23) gives precisely $I_\alpha(\phi^*, X^n)$, proving (33). ∎

IV. α-NML AS AN ALTERNATIVE TO NML

An important feature of the class of α-NML predictors is that these predictors are able to solve some of the problems that afflict the classical NML. First of all, α-NML predictors do not require any maximization over the parameter space Θ. The maximization is in fact replaced by a weighted average of the distributions p_θ to the power of α. Furthermore, by choosing carefully the prior w and the parameter α, one is able to control the convergence of the integral at the denominator of (27). In this sense, the role of the prior w is similar to that of the luckiness function that appears in the definition of Luckiness NML [19], [20], an alternative predictor that was introduced in the literature to overcome the convergence problem of the NML estimator. It is defined as

$$\hat{p}_{\text{LNML}}(x^n) = \frac{\sup_{\theta \in \Theta} \pi(\theta) p_\theta(x^n)}{\sum_{\bar{x}^n \in \mathcal{X}^n} \sup_{\theta \in \Theta} \pi(\theta) p_\theta(\bar{x}^n)}. \tag{42}$$

where π is the luckiness function, a probability distribution on Θ which models how confident one is that a given $\theta \in \Theta$ is the true parameter of the source. Luckiness NML is the predictor that minimizes a regret measure related to the worst-case regret (11), the *worst-case luckiness regret*, which is defined as

$$R_{\max}(\pi, \hat{p}) = \max_{\theta \in \Theta} \max_{x^n} \frac{\pi(\theta) p_\theta(x^n)}{\hat{p}(x^n)}. \tag{43}$$

However, α-NML has three additional advantages with respect to Luckiness NML, as an alternative predictor to NML:

- it has a simple interpretation in terms of Rényi divergence, and it is provably optimal under α-regret measures;
- provided that $\text{supp}(w) = \Theta$, the α-NML converges to NML (if it exists) as $\alpha \to \infty$, while this is not the case for Luckiness-NML.
- it includes Luckiness NML as a special case, if the prior w is chosen properly as a function of α (see Subsection C later on).

A. Existence of α-NML and choice of w and α

The main advantage of α-NML is that, while it is an approximation of NML that gets more and more accurate as $\alpha \to \infty$, α-NML may exist for finite α even when the NML does not converge. In fact, the existence of α-NML is determined by the convergence of the integral at the denominator of (27), i.e.,

$$\int_{\mathcal{X}^n} \left\{ \int_{\Theta} w(\theta)\, p_\theta^\alpha(x^n)\, d\theta \right\}^{1/\alpha} dx^n \,. \tag{44}$$

General conditions for its convergence are hard to find. However, the following sufficient conditions can be derived.

Theorem 2: For a given $\alpha \geq 1$, the α-NML predictor exists if any of the following conditions are satisfied.

1) If the NML exists, so does the α-NML, for any prior distribution w on Θ.
2) If $\mathcal{P}$ is an exponential family of distributions, and if $\mathrm{supp}(w)$ is an INECCSI[1] subset of Θ, then the α-NML always exists.
3) If Θ is countable, then the α-NML exists if

$$\sum_{\theta \in \Theta} w(\theta)^{\frac{1}{\alpha}} < \infty. \tag{45}$$

4) If either Θ or $\mathcal{X}$ are finite, then the α-NML always exists.

Proof: The proof follows from the observation that the logarithm of the denominator of α-NML, i.e., Equation (44), is Sibson's mutual information of order α (minus the multiplicative constant). The three propositions of the theorem can be proved using known properties of $I_\alpha(\phi, X^n)$.

1) It is known that $I_\alpha(\phi, X^n)$ is a non-decreasing function of α [10]. Therefore,

$$I_\alpha(\phi, X^n) \leq I_\infty(\phi, X^n) = \log \int_{\mathcal{X}} \sup_{\theta \in \mathrm{supp}(w)} p_\theta(x^n)\, dx^n \leq \log \int_{\mathcal{X}} \sup_{\theta \in \Theta} p_\theta(x^n)\, dx^n. \tag{46}$$

Since the last quantity is precisely the logarithm of the denominator of the NML, then, if the NML exists, i.e., if its denominator is finite, so does α-NML.

2) If $\mathrm{supp}(w) \subset \Theta$ is INECCSI, then

$$I_\alpha(\phi, X^n) \leq I_\infty(\phi, X^n) = \log \int_{\mathcal{X}} \sup_{\theta \in \mathrm{supp}(w)} p_\theta(x^n)\, dx^n \tag{47}$$

and the last quantity is known to be finite from [19, Theorem 7.1].

3) Due to the data processing inequality for Sibson's mutual information [10], we have that $I_\alpha(\phi, X^n) \leq I_\alpha(\phi, \phi)$. Furthermore, for countable Θ, [10] also shows that

$$I_\alpha(\phi, \phi) = H_{\frac{1}{\alpha}}(\phi) = \frac{\alpha}{\alpha - 1} \log \sum_\theta w(\theta)^{\frac{1}{\alpha}}. \tag{48}$$

Hence, if $\sum_\theta w(\theta)^{\frac{1}{\alpha}}$ is finite, then $I_\alpha(\phi, X^n)$ is finite and the α-NML exists.

4) If Θ is finite, then the existence of α-NML is guaranteed by point 3), since the sum in (45) contains finitely many bounded terms. If $\mathcal{X}$ is finite, then we have

$$\sum_{x^n} \left\{ \int_{\Theta} w(\theta)\, p_\theta^\alpha(x^n)\, d\theta \right\}^{1/\alpha} \leq \sum_{x^n} \left\{ \int_{\Theta} w(\theta)\, d\theta \right\}^{1/\alpha} = |\mathcal{X}|^n \tag{49}$$

[1] Following [19], an INECCSI subset of Θ is a set $\Theta_0 \subset \Theta$ such that the interior of Θ_0 is not empty, and its closure is a compact subset of the interior of Θ.

where the inequality is due to the fact that for finite $\mathcal{X}$ we have $p_\theta(x^n) \leq 1$ for every θ and x^n, and the final equality is due to the fact that w is a probability distribution on Θ.

■

Theorem 2 helps us to answer the two questions that arose in Section III. In fact, the theorem highlights the fundamental role of the prior distribution w on the existence of α-NML. Even when the NML does not exist, one can carefully choose w so that the α-NML converges. In particular, condition 3) shows an interesting interplay between the prior w and the parameter α, when the parameter space Θ is countable. Note that the sum in (45) is a non-decreasing function of α: when $\alpha = 1$, the sum always converges since w is a probability distribution, while it diverges in the limit $\alpha \to \infty$. Since the α-NML gets closer to the NML as α grows, Equation (45) suggests that, for a given w, one should choose the largest possible α such that $\sum_\theta w(\theta)^{1/\alpha}$ converges.

In general, however, good choices of w and α will depend on the detailed structure of parameter space Θ and the associated probability distributions. For example, for the special case $\alpha = 1$, we can connect to the classic literature on mixture estimators. Namely, when the parametric family under consideration is the class of discrete memoryless sources, one retrieves well-known estimators depending on the chosen prior w: when w is the uniform distribution, one obtains the Laplace estimator [21], [22], while the Krichevsky-Trofimov estimator is obtained when w is a Dirichlet distribution $D(\frac{1}{2}, \ldots, \frac{1}{2})$. The NML predictor is instead retrieved in the limit $\alpha \to \infty$, provided that for every $x^n \in \mathcal{X}^n$, $\sup_\theta p_\theta(x^n)$ is achieved for a θ such that $w(\theta) > 0$. This condition is achieved in particular for a prior w such that $w(\theta) > 0$ for every $\theta \in \Theta$.

It is important to point out that the NML does not exist for most parametric families with an infinite parameter space Θ [19]. By constrast, the α-NML may very well exist. In the sequel, we illustrate this fact by the aid of a number of examples. Consider for example the case where the distribution $p_\theta(x^n)$ is unbounded for some $x^n \in \mathcal{X}^n$ and $\theta \in \Theta$. It is clear that the NML never exists in such a case, since $\sup_\theta p_\theta(x^n) = +\infty$ for some x^n. However, the α-NML may exist nonetheless, for proper choices of the prior w, as in the following example.

Example 1: Consider the location family with parameter space $\Theta = \mathbb{R}$ and associated distributions

$$p_\theta(x) = \begin{cases} -\log(x-\theta), & \theta < x \leq 1+\theta \\ 0, & \text{otherwise.} \end{cases} \tag{50}$$

While it is easy to check that this is a valid probability density function for every $\theta \in \mathbb{R}$, the key observation in this example is that $\sup_\theta p_\theta(x) = +\infty$ for every $x \in \mathbb{R}$. Therefore, the NML does not exist. However, consider the α-NML predictor with prior $w(\theta) = \mathbb{1}_{\{0 \leq \theta \leq 1\}}$ and, e.g., $\alpha = 2$. Then, the α-NML esists. In fact, after some algebra, one can check that the integral in (44) equals

$$\int_{\mathcal{X}^n} \left\{ \int_\Theta w(\theta)\, p_\theta^\alpha(\bar{x}^n)\, d\theta \right\}^{1/\alpha} d\bar{x}^n = \tag{51}$$

$$= \int_0^1 \left(\sqrt{x^2 \ln^2 x - 2x \ln x + 2x} + \sqrt{2 - x^2 \ln^2 x + 2x \ln x - 2x} \right) dx \tag{52}$$

$$\approx 1.68. \tag{53}$$

Note that Luckiness NML as defined in [19], which is another alternative predictor to NML, does not exist either in this case, since for any $x \in (0,1)$, $\sup_\theta w(\theta)p_\theta(x) = +\infty$.

The following is an example of a class of bounded distributions, with countable alphabet $\mathcal{X}$, for which the NML does not exist, but the α-NML does for every $\alpha \geq 1$.

Example 2: Consider the family of geometric distributions

$$p_\theta(x) = (1-\theta)^x \theta \tag{54}$$

for $x \in \mathcal{X} = \mathbb{N}$. A simple calculation shows that

$$\sum_x \sup_\theta p_\theta(x) = \sum_x \frac{1}{x}\left(1 - \frac{1}{x+1}\right)^{x+1} = +\infty. \tag{55}$$

However, consider the α-NML with the prior w as the uniform distribution on $[0,1]$ and α as any positive integer. Then,

$$\int_0^1 w(\theta)p_\theta^\alpha(x)\,dx = \frac{1}{\alpha(x+1)+1} \cdot \frac{1}{\binom{\alpha(x+1)}{\alpha}} \tag{56}$$

due to properties of the Beta function. Now, since $\binom{\alpha(x+1)}{\alpha} \sim \frac{(\alpha(x+1))^\alpha}{\alpha!}$ for any fixed $\alpha \geq 1$ and large x, one has

$$\sum_x \left\{\int_0^1 w(\theta)p_\theta^\alpha(x)\,dx\right\}^{1/\alpha} = \sum_x \left\{\frac{1}{\alpha(x+1)+1}\frac{1}{\binom{\alpha(x+1)}{\alpha}}\right\}^{1/\alpha}. \tag{57}$$

This series converges if and only if the series

$$\sum_x \frac{1}{x^{1+\frac{1}{\alpha}}} \tag{58}$$

does, which is the case for every $1 \leq \alpha < \infty$. Note that with this choice of prior, Luckiness NML does not exist, either, since it is equal to the NML.

Finally, the following is an example of a parametric class with uncountable $\mathcal{X}$ and Θ for which the NML does not exist, but the α-NML does, for every $\alpha \geq 1$ and with a prior distribution w such that $\text{supp}(w) = \Theta$.

Example 3: Consider the normal location family with variance 1, where $\Theta = \mathbb{R}$ and $p_\theta(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2}}$, and consider the prior distribution $w(\theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{\theta^2}{2}}$. Then, we have

$$\begin{aligned}
\int_\Theta w(\theta)p_\theta^\alpha(x)\,d\theta &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-\frac{\theta^2}{2} - \frac{\alpha}{2}(x-\theta)^2} && (59)\\
&= \frac{1}{2\pi} e^{-\frac{\alpha}{2(1+\alpha)}x^2} \int_{-\infty}^{\infty} e^{-\frac{1+\alpha}{2}(\theta - \frac{\alpha x}{1+\alpha})^2} && (60)\\
&= \frac{1}{\sqrt{2\pi(1+\alpha)}} e^{-\frac{\alpha}{2(1+\alpha)}x^2}. && (61)
\end{aligned}$$

Hence, we have

$$\begin{aligned}
\int_{\mathcal{X}} \left\{\int_\Theta w(\theta)\,p_\theta^\alpha(x)\,d\theta\right\}^{1/\alpha} dx &= \frac{1}{(2\pi(1+\alpha))^{\frac{1}{2\alpha}}} \int_{-\infty}^{\infty} e^{-\frac{x^2}{2(1+\alpha)}}\,dx && (62)\\
&= (2\pi(1+\alpha))^{\frac{1}{2}(1-\frac{1}{\alpha})} && (63)
\end{aligned}$$

which is finite for every $1 \leq \alpha < \infty$. As expected, its value goes to infinity as $\alpha \to \infty$. In fact, for this parametric class, the NML does not exist, since $\sup_\theta p_\theta(x) = \frac{1}{\sqrt{2\pi}}$ is constant, and therefore its integral over $\mathcal{X} = \mathbb{R}$ is infinite.

B. Worst-case regret of α-NML

As we mentioned earlier, the NML is the optimal predictor under worst-case regret (12), i.e., α-regret for $\alpha = \infty$. It is therefore interesting to study the worst-case regret of α-NML, in order to assess how much one loses under this metric as a function of $\alpha \geq 1$. The following formula highlights the depencence of the worst-case regret for the α-NML predictor on Sibson's α-mutual information.

Lemma 1: The worst-case regret of the α-NML predictor with prior w can be written as

$$R_{\max}(\hat{p}_\alpha) = \frac{\alpha-1}{\alpha} I_\alpha(\phi, X^n) + W_\alpha(\mathcal{P}) \tag{64}$$

where $I_\alpha(\phi, X^n)$ is the α-mutual information for $(\phi, X^n) \sim w(\phi)\, p_\phi(X^n)$, and

$$W_\alpha(\mathcal{P}) \triangleq \sup_{x^n \in \mathcal{X}^n} \log \frac{\sup_{\theta \in \Theta} p_\theta(x^n)}{\left\{\int_\Theta w(\theta)\, p_\theta^\alpha(x^n)\, d\theta\right\}^{1/\alpha}}. \tag{65}$$

Proof: Starting from (11) and substituting the definition of α-NML given by Equation (27), we have

$$\begin{aligned}
R_{\max}(\hat{p}_\alpha) &= \sup_{\theta \in \Theta} \sup_{x^n \in \mathcal{X}^n} \log \frac{p_\theta(x^n)}{\hat{p}_\alpha(x^n)} && (66)\\
&= \sup_{\theta \in \Theta} \sup_{x^n \in \mathcal{X}^n} \log \frac{p_\theta(x^n)}{\frac{\left\{\int_\Theta w(\theta)\, p_\theta^\alpha(x^n)\, d\theta\right\}^{1/\alpha}}{\int_{\mathcal{X}^n} \left\{\int_\Theta w(\theta)\, p_\theta^\alpha(\bar{x}^n)\, d\theta\right\}^{1/\alpha} d\bar{x}^n}} && (67)\\
&= \log \int_{\mathcal{X}^n} \left\{\int_\Theta w(\theta)\, p_\theta^\alpha(x^n)\, d\theta\right\}^{1/\alpha} dx^n + \max_{x^n} \log \frac{\sup_\theta p_\theta(x^n)}{\left\{\int_\Theta w(\theta)\, p_\theta^\alpha(x^n)\, d\theta\right\}^{1/\alpha}} && (68)\\
&= \frac{\alpha-1}{\alpha} I_\alpha(\phi, X^n) + W_\alpha(\mathcal{P}) && (69)
\end{aligned}$$

where in the last step we used the definitions of $I_\alpha(\phi, X^n)$ and $W_\alpha(\mathcal{P})$ in Equations (18) and (65) respectively. ■

Since in the limit $\alpha \to \infty$ the α-NML predictor becomes equal to the NML, it follows that $R_{\max}(\hat{p}_\alpha)$ tends to the optimal worst-case regret $I_\infty(\phi, X^n)$ when α goes to infinity. However, in general, it is not clear from (64) what is the behavior of $R_{\max}(\hat{p}_\alpha)$ as a function of α, i.e., it is not clear if the regret is monotonically decreasing with α or not, and this might depend on the actual class of distributions that is considered. In fact, the first term in (64) is increasing with α, due to known properties of Sibson's α-mutual information [10]. However, the overall behavior of the regret is certainly not increasing with α, since it reaches its minimum when $\alpha \to \infty$. This proves the critical role of the second term $W_\alpha(\mathcal{P})$ in the overall behavior of the worst-case regret. While this term is in general of difficult analysis, in Section V we show that it can be written in a simple form for the class of discrete memoryless sources, for which a complete asymptotic analysis of the worst-case regret can be carried out.

C. Connection between α-NML and Luckiness NML

The previous sections already made clear that the choice of the prior distribution w in (27) in defining the α-NML distribution is of fundamental importance. For example, the choice of a Dirichlet prior in (87) is what makes α-NML almost optimal for the case of discrete memoryless sources, and the choice of the correct prior is necessary for the optimality of α-NML under the α-regret (25). When we discussed in Section III that the α-NML interpolates the mixture predictors (28) and the NML (13), we assumed w to be fixed and independent of α. It turns out that if one chooses w carefully as a function of α, the α-NML can also approximate other predictors related to the NML, in particular the Luckiness NML predictor defined in (42). In fact, let π be the luckiness function used in the definition of Luckiness NML, and consider as a regret measure the expectation over the parameters in Θ according to π, and then the expectation over sequences distributed according to p_θ. The result is what we may call *average luckiness regret*, which is formally defined as

$$R_{\mathrm{av}}(\pi, \hat{p}) = \mathbb{E}_{\theta \sim \pi} \left[\mathbb{E}_{X^n \sim p_\theta} \left[\log \frac{p_\theta(X^n)}{\hat{p}(X^n)} \right] \right] . \tag{70}$$

It is easy to prove that the predictor minimizing this regret is a mixture predictor whose weighting function is π, i.e.,

$$\hat{p}(x^n) = \int_\Theta \pi(\theta) p_\theta(x^n) d\theta \, . \tag{71}$$

A possible interpolation between the Luckiness NML defined in (42) and the mixture predictor in (71) is again given by α-NML of Equation (27), if one chooses the proper prior distribution w. In fact, for any given $\alpha \geq 1$, one can take the tilted prior distribution

$$w(\theta) = \frac{\pi(\theta)^\alpha}{\int_\Theta \pi(\theta)^\alpha d\theta} \tag{72}$$

provided that the integral in the denominator converges. With such a choice of prior, the α-NML becomes

$$\hat{p}_\alpha(x^n) = \frac{\left\{ \int_\Theta w(\theta) \, p_\theta^\alpha(x^n) \, d\theta \right\}^{1/\alpha}}{\sum_{x^n} \left\{ \int_\Theta w(\theta) \, p_\theta^\alpha(x^n) \, d\theta \right\}^{1/\alpha}} \tag{73}$$

$$= \frac{\left\{ \int_\Theta (\pi(\theta) \, p_\theta(x^n))^\alpha \, d\theta \right\}^{1/\alpha}}{\sum_{x^n} \left\{ \int_\Theta (\pi(\theta) \, p_\theta(x^n))^\alpha \, d\theta \right\}^{1/\alpha}} . \tag{74}$$

For convenience, we can call this predictor *Luckiness α-NML*. However, it is important to note that this predictor is not something different from the already defined α-NML. In fact, it is simply a particular instance of that same predictor, where one chooses a particular prior distribution w – in this case, it is the one in Equation (72). In this sense, the α-NML is able to link the standard NML and the Luckiness NML under the same, more general, object. By taking $\alpha = 1$, one retrieves the mixture predictor in (71), while in the limit $\alpha \to \infty$, one gets the luckiness NML that is defined in (42).

Similarly to the case of the α-regret of Equation (25), there also exists an interpolation between the worst-case luckiness regret in (43) and the average luckiness regret in (70) for which the Luckiness α-NML is the optimal predictor. In fact, consider the luckiness regret defined for any $\alpha \geq 1$ by

$$R_\alpha(\pi, \hat{p}) = \frac{1}{\alpha - 1} \log \mathbb{E}_{\theta \sim \pi_\alpha} \left[\mathbb{E}_{X^n \sim p_\theta} \left[\left(\frac{p_\theta(X^n)}{\hat{p}(X^n)} \right)^{\alpha - 1} \right] \right] \tag{75}$$

where

$$\pi_\alpha(\theta) = \frac{\pi(\theta)^\alpha}{\int_\Theta \pi(\bar{\theta})^\alpha d\bar{\theta}} . \tag{76}$$

Notice that the exponent of α inside the expectation is the same as in (25), as well as the normalization factor $\frac{1}{\alpha-1} \log$ in front. However, while the interpolation of α-regret discussed in Section II acts only on how the sequences x^n are considered, in the luckiness case, instead, the interpolation also occurs on how the parameters in Θ are weighted. In fact, in the worst-case luckiness regret (43) there is a maximization over θ, while in the average luckiness regret

(70) there is an expectation according to π. The way this interpolation is handled in (75) is through an expectation over a tilted version of the luckiness function π, which equals π when $\alpha = 1$, and it assigns probability one to the maximal θ when $\alpha \to \infty$. The optimal predictor for the luckiness α-regret is the Luckiness α-NML.

Theorem 3: For any given $\alpha \geq 1$ and any given luckiness function π over Θ, the Luckiness α-NML defined in (74) with prior distribution w taken as in (72), minimizes $R_\alpha(\pi, \hat{p})$ over all probability distributions on $\mathcal{X}^n$, under the assumption that the prior distribution converges, i.e., if

$$\int_\Theta \pi(\theta)^\alpha d\theta < \infty. \tag{77}$$

Proof: See Appendix A. ■

Alternatively, one could define a different average luckiness regret measure such as

$$\tilde{R}_{\text{av}}(\pi, \hat{p}) = \max_{\theta \in \Theta} \mathbb{E}_{X^n \sim p_\theta} \left[\log \frac{\pi(\theta) p_\theta(X^n)}{\hat{p}(X^n)} \right]. \tag{78}$$

Then, a simple interpolation between this regret and (43) is

$$R_\alpha(\pi, \hat{p}) = \sup_{\theta \in \Theta} \frac{1}{\alpha - 1} \log \mathbb{E}_{X^n \sim p_\theta} \left[\left(\frac{\pi(\theta) p_\theta(X^n)}{\hat{p}(X^n)} \right)^{\alpha - 1} \right] \tag{79}$$

which is strongly related to the α-regret in Equation (25). In fact, one can prove a result similar to Theorem 3 for this regret measure. In this context, it can be used to show that there exists a distribution $w^*(\theta)$ on Θ such that, the α-NML defined in (27), with prior equal to

$$w(\theta) = \frac{w^*(\theta)\pi^\alpha(\theta)}{\int_\Theta w^*(\bar{\theta})\pi^\alpha(\bar{\theta}) d\bar{\theta}} \tag{80}$$

is the predictor that minimizes (79).

V. α-NML FOR DMS

We now focus on the important class of discrete memoryless sources taking values in a finite but arbitrary alphabet[2]. This class has been the focus of a large part of the literature on universal prediction and compression. The main reasons for this are that this class is the simplest non-trivial example for which one can get a sense of how a predictor behaves, and at the same time prove rigorously some results in terms of performance of a predictor compared to the optimal.

A. The Krichevsky-Trofimov estimator

The most important result on universal prediction for this class of distributions is possibly the Krichevsky-Trofimov estimator. Let the source alphabet be $\mathcal{X} = \{1, 2, \ldots, m\}$. Let also

$$\Theta = \left\{ \boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m) : \sum_{i=1}^m \theta_i = 1 \text{ and } \theta_i \geq 0 \text{ for every } i \right\} \tag{81}$$

be the parameter set. For each parameter $\boldsymbol{\theta}$ and sequence $x^n = (x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$, the source indexed by $\boldsymbol{\theta}$ generates the sequence x^n with probability

$$p_{\boldsymbol{\theta}}(x^n) = \prod_{i=1}^m \theta_i^{n_i}, \tag{82}$$

[2] In part of the literature this class also goes under the name of *constant experts* — see, e.g., [7].

where

$$n_i = |\{1 \leq j \leq n : x_j = i\}| \,. \tag{83}$$

For the class of discrete memoryless sources described above, the Krichevsky-Trofimov predictor is a simple mixture estimator,

$$\hat{p}_{\mathrm{KT}}(x^n) \triangleq \int_{\Theta} w_{\mathrm{D}}(\boldsymbol{\theta})\, p_{\boldsymbol{\theta}}(x^n)\, d\boldsymbol{\theta} \tag{84}$$

where the prior distribution on the parameter space is $w_{\mathrm{D}} \sim D(\frac{1}{2}, \ldots, \frac{1}{2})$, i.e., the Dirichlet distribution with parameters equal to $\frac{1}{2}$,

$$w_{\mathrm{D}}(\boldsymbol{\theta}) = \frac{\Gamma\left(\frac{m}{2}\right)}{\pi^{m/2}} \prod_{i=1}^{m} \frac{1}{\sqrt{\theta_i}} \,. \tag{85}$$

This estimator has arguably three major advantages.

1) Its probability estimates $\hat{p}_{\mathrm{KT}}(x^n)$ can be computed easily in closed form. In fact, substituting the definitions of w_{D} and of $p_{\boldsymbol{\theta}}$ into (84) and using properties of the Gamma function $\Gamma(t)$, one is able to derive the simple formula

$$\hat{p}_{\mathrm{KT}}(x^n) = \frac{\Gamma\left(\frac{m}{2}\right)}{\pi^{m/2}} \frac{\prod_{i=1}^{m} \Gamma\left(n_i + \frac{1}{2}\right)}{\Gamma\left(n + \frac{m}{2}\right)} \,. \tag{86}$$

2) It is asymptotically optimal in n up to a constant term, in terms of both worst-case regret $R_{\max}$ and average regret R_{av}.
3) Simple formulae exist for the computation of the conditional probability of a new symbol given the previous ones.

B. The α-NML for DMS

We now compare the Krichevsky-Trofimov estimator with the α-NML for the class of discrete memoryless sources. We will use as a prior distribution w for the α-NML the same Dirichlet $D(\frac{1}{2}, \ldots, \frac{1}{2})$ of the KT estimator. It is important to notice that other prior distributions could also be considered. In this work, we focus on the Dirichlet prior distribution mainly for two reasons: (1) it makes easier to compare our predictor to the Krichevsky-Trofimov, since with such a choice, the α-NML is precisely the KT predictor when $\alpha = 1$; (2) it is easier to handle mathematically and to get closed-form formulas for the estimated probabilities. Furthermore, the Dirichlet distribution $D(\frac{1}{2}, \ldots, \frac{1}{2})$ is the so-called *Jeffreys' prior distribution* [23] for the class of discrete memoryless sources. It is known that a mixture predictor with prior distribution equal to Jeffrey's prior has an asymptotically optimal regret, for exponential families of distributions and for most sequences x^n in $\mathcal{X}^n$ (see, e.g., [19, Section 8.1] and references therein, for a more precise account of these results). Nevertheless, it is likely that other prior distributions would improve the performance of the predictor, at the cost of additional complexity of implementation.

In the case of the Dirichlet distribution w_D as in (85), the α-NML predictor takes the form

$$\hat{p}_{\alpha}(x^n) = \frac{1}{Z_n(\alpha)} \left\{ \int_{\Theta} \prod_{i=1}^{m} \theta_i^{\alpha n_i - \frac{1}{2}} \, d\boldsymbol{\theta} \right\}^{1/\alpha} \,, \tag{87}$$

where the normalization constant $Z_n(\alpha)$ is equal to

$$Z_n(\alpha) \triangleq \sum_{x^n} \left\{ \int_{\Theta} \prod_{i=1}^{m} \theta_i^{\alpha n_i - \frac{1}{2}} \, d\boldsymbol{\theta} \right\}^{1/\alpha} \,. \tag{88}$$

The integral on the right is known in the literature as the multivariate Beta function, and it has the closed-form expression

$$\int_{\Theta} \prod_{i=1}^{m} \theta_i^{\alpha n_i - \frac{1}{2}} \, d\boldsymbol{\theta} = \frac{\prod_{i=1}^{m} \Gamma\left(\alpha n_i + \frac{1}{2}\right)}{\Gamma\left(\alpha n + \frac{m}{2}\right)} \,, \tag{89}$$

so that the probability estimates given by the α-NML predictor can be written as

$$\hat{p}_\alpha(x^n) = \frac{1}{Z_n(\alpha)} \left\{ \frac{\prod_{i=1}^{m} \Gamma\left(\alpha n_i + \frac{1}{2}\right)}{\Gamma\left(\alpha n + \frac{m}{2}\right)} \right\}^{1/\alpha} \tag{90}$$

where

$$Z_n(\alpha) = \sum_{x^n} \left\{ \frac{\prod_{i=1}^{m} \Gamma\left(\alpha n_i + \frac{1}{2}\right)}{\Gamma\left(\alpha n + \frac{m}{2}\right)} \right\}^{1/\alpha} \,. \tag{91}$$

We now want to briefly discuss the computational complexity of α-NML. Notice that in principle the sum that appears in $Z_n(\alpha)$ contains an exponential number of terms in n, which may be of concern from a computational point of view. However, it can be seen that the actual terms in the sum only depend on the number of symbols $\boldsymbol{n} = (n_1, n_2, \ldots, n_m)$. Therefore, one can group equal terms together to get

$$Z_n(\alpha) = \sum_{\boldsymbol{n}} \binom{n}{n_1, \ldots, n_m} \left\{ \frac{\prod_{i=1}^{m} \Gamma\left(\alpha n_i + \frac{1}{2}\right)}{\Gamma\left(\alpha n + \frac{m}{2}\right)} \right\}^{1/\alpha} \,. \tag{92}$$

Written in this way, the sum contains only a polynomial number of terms, since the number of different vectors $\boldsymbol{n}$ is upper-bounded by $(n+1)^{m-1}$. In particular, when the alphabet is binary — i.e., when $m = 2$, — the number of terms is linear in n. Furthermore, the computation of the multinomial coefficients is also not a problem, since they can be computed recursively from the previous ones with a constant number of operations.

Finally, the Gamma terms in (90) and (91) can also be computed efficiently, when $\alpha \geq 1$ is restricted to be an integer. In such a case, one can use the recurrence formula for the Gamma function

$$\Gamma(z+1) = z\Gamma(z) \tag{93}$$

to compute each of the Gamma terms in the two formulae, e.g.,

$$\Gamma\left(\alpha n_i + \frac{1}{2}\right) = \left(\alpha n_i - \frac{1}{2}\right)\left(\alpha n_i - \frac{3}{2}\right) \cdots \frac{3}{2} \cdot \frac{1}{2} \cdot \sqrt{\pi} \,, \tag{94}$$

where we used the well-known fact that $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$. Similar computations can be used to calculate the denominator term $\Gamma\left(\alpha n + \frac{m}{2}\right)$. As one can see, the number of operations required for each term of the sum in (91) is linear in αn. Therefore, for any positive integer α, the number of operations required to compute $Z_n(\alpha)$ and $\hat{p}_\alpha(x^n)$ is polynomial in n and linear in α. As we will see later on, a small value of α is already enough to improve significantly the worst-case regret of the α-NML predictor, and to get close to the optimal regret achieved by the NML.

When α is a positive integer, one can also derive simple formulae for the conditional probability of the next symbol when a sequence of length $n-1$ is already given. Consider the setting where a fixed sequence $x^{n-1} \in \mathcal{X}^{n-1}$

has been revealed, and we want to estimate the conditional probability of symbol $k \in \mathcal{X}$ given x^{n-1}, where $\mathcal{X} = \{1, 2, \ldots, m\}$. As an intermediate step, let us compute the ratio $\hat{p}_\alpha(x^{n-1}, k)/\hat{p}_\alpha(x^{n-1})$.

$$\begin{aligned}
\frac{\hat{p}_\alpha(x^{n-1}, k)}{\hat{p}_\alpha(x^{n-1})} &= \frac{\frac{1}{Z_n(\alpha)} \left\{ \frac{\Gamma(\alpha(n_k+1)+\frac{1}{2}) \prod_{i \neq k} \Gamma(\alpha n_i + \frac{1}{2})}{\Gamma(\alpha n + \frac{m}{2})} \right\}^{\frac{1}{\alpha}}}{\frac{1}{Z_{n-1}(\alpha)} \left\{ \frac{\Gamma(\alpha n_k + \frac{1}{2}) \prod_{i \neq k} \Gamma(\alpha n_i + \frac{1}{2})}{\Gamma(\alpha(n-1) + \frac{m}{2})} \right\}^{\frac{1}{\alpha}}} && (95) \\
&= \frac{Z_{n-1}(\alpha)}{Z_n(\alpha)} \left\{ \frac{\Gamma(\alpha n_k + \alpha + \frac{1}{2}) \Gamma(\alpha n - \alpha + \frac{m}{2})}{\Gamma(\alpha n_k + \frac{1}{2}) \Gamma(\alpha n + \frac{m}{2})} \right\}^{\frac{1}{\alpha}} && (96) \\
&= \frac{Z_{n-1}(\alpha)}{Z_n(\alpha)} \left\{ \prod_{j=0}^{\alpha-1} \frac{\alpha n_k + \frac{1}{2} + j}{\alpha n - \alpha + \frac{m}{2} + j} \right\}^{\frac{1}{\alpha}}, && (97)
\end{aligned}$$

where in the last step we used (93) recursively. Finally, we can obtain the conditional probability of k given x^{n-1} as

$$\begin{aligned}
\hat{p}_\alpha(k | x^{n-1}) &\triangleq \frac{\hat{p}_\alpha(x^{n-1}, k)}{\sum_{i=1}^m \hat{p}_\alpha(x^{n-1}, i)} && (98) \\
&= \frac{\frac{\hat{p}_\alpha(x^{n-1}, k)}{\hat{p}_\alpha(x^{n-1})}}{\sum_{i=1}^m \frac{\hat{p}_\alpha(x^{n-1}, i)}{\hat{p}_\alpha(x^{n-1})}} && (99) \\
&= \frac{\prod_{j=0}^{\alpha-1} (\alpha n_k + \frac{1}{2} + j)^{1/\alpha}}{\sum_{i=1}^m \prod_{j=0}^{\alpha-1} (\alpha n_i + \frac{1}{2} + j)^{1/\alpha}} && (100)
\end{aligned}$$

for any $k \in \mathcal{X}$. As one can see from (100), the computational complexity of each of these probabilities is linear in α and m and does not depend on n. For $\alpha = 1$, one obtains the known formula for the conditional probabilities of the Krichevsky-Trofimov estimator

$$\hat{p}_{\mathrm{KT}}(k | x^{n-1}) = \frac{n_k + \frac{1}{2}}{n + \frac{m}{2} - 1}. \qquad (101)$$

while, e.g., for $\alpha = 2$, one gets the formula mentioned in the Introduction in Equation (16).

C. Worst-case regret of α-NML for DMS

We now want to discuss the performance of α-NML in terms of worst-case regret, with the primary objective of analyzing how much the regret of α-NML improves upon that of the Krichevsky-Trofimov estimator, and how it compares to the optimal NML. In order to do this, we start by finding the asymptotical value of the worst-case regret for α-NML, starting from formula (64). This formula has two major advantages in the discrete memoryless case. First, the asymptotics of the α-mutual information term, which would be in general hard to study, can actually be computed using known results in the literature, once one recognizes the optimality of the Dirichlet prior. Second, the maximization over sequences in $\mathcal{X}^n$ in the $W_\alpha(\mathcal{P})$ term, that would be complicated to evaluate in general, can be resolved explicitly for this particular class of distributions.

Theorem 4: For the class of discrete memoryless sources, the $W_\alpha(\mathcal{P})$ term defined in (65) is equal to

$$W_\alpha(\mathcal{P}) = \frac{1}{\alpha} \log \frac{\Gamma(\alpha n + \frac{m}{2})}{\Gamma(\alpha n + \frac{1}{2})} + \frac{1}{2\alpha} \log \pi - \frac{1}{\alpha} \Gamma\left(\frac{m}{2}\right). \qquad (102)$$

Proof: See Appendix B. ■

With the help of this result, we can prove the asymptotics of the worst-case regret for the α-NML estimator.

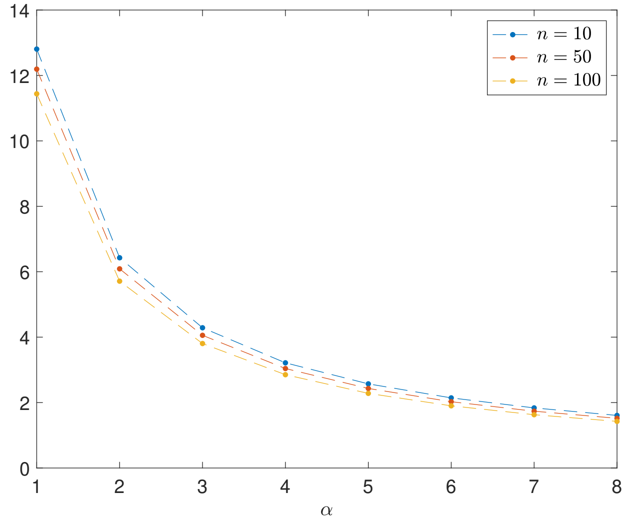

Fig. 1. Percentage of increase of $R_{\max}(\hat{p}_\alpha)$ with respect to the optimal value $R_{\max}(\hat{p}_{\text{NML}})$, as a function of α, for binary sequences of length $n = 10, 50, 100$ and integer values of α. The value at $\alpha = 1$ corresponds to the regret of the Krichevsky-Trofimov estimator.

Theorem 5: For the class of discrete memoryless sources, the worst-case regret of the α-NML predictor is equal to

$$R_{\max}(\hat{p}_\alpha) = \frac{m-1}{2}\log\frac{n}{2} + \frac{1}{2}\log\pi - \log\Gamma\left(\frac{m}{2}\right) + \frac{m-1}{2\alpha}\log 2 + o(1) \tag{103}$$

where $o(1) \to 0$ as $n \to \infty$.

Proof: See Appendix C. ∎

From (103) it can be seen that the asymptotic behavior of the worst-case regret of α-NML has the same dependence on n for every $\alpha \geq 1$, while the terms that do not depend on n strictly decrease as α increases. Therefore, the α-NML has an asymptotic advantage with respect to the Krichevsky-Trofimov estimator only in the constant term. However, for finite length, computer evaluation of the worst-case regret show that the advantage of α-NML over the KT estimator is larger. For example, Figure 1 shows some of these results for binary alphabet. Since asymptotically the difference of the regret of the α-NML (and in particular the Krichevsky-Trofimov estimator) and that of the NML is a constant, one expects the percentage of increase of the regret to tend to zero as n goes to infinity, for every α. However, as one can see from Figure 1, this decrease appears to be very slow, an additional indication that the (almost) optimality of the Krichevsky-Trofimov estimator in terms of worst-case regret is only asymptotical, while for finite-length sequences the difference is actually substantial. However, precise analysis of finite-length regret remains difficult.

VI. Conclusion

In this paper, we introduced a new class of general predictors dependent on a real parameter $\alpha \geq 1$, which is shown to interpolate to mixture predictors and the NML. The idea for this class of predictors comes from the connection between the worst-case regret achieved by the NML predictor, and Sibson's α-mutual information.

We proved the optimality of α-NML under α-regret, a general regret measure linked to Rényi divergence, that interpolates between the classical average and worst-case regrets. Also, we discussed examples that prove the broad applicability of α-NML also for families of distributions for which the NML does not exist, and we compared α-NML to other alternatives to NML such as Luckiness NML. Furthermore, we showed that for the popular family of discrete memoryless sources, one is able to derive some simple formulas to compute the probabilities estimated by the new class of predictors, when the parameter α is a positive integer. For this class of distributions, we also derive an asymptotic expression for the worst-case regret of α-NML, which interpolates between those of the Krichevsky-Trofimov estimator and the NML.

Appendix A
Proof of Theorem 3

Notice that one can rewrite (75) as

$$R_\alpha(\pi, \hat{p}) = D_\alpha(\pi_\alpha \times p_\theta \| \pi_\alpha \times \hat{p}). \tag{104}$$

Thanks to [10, Equation (32)], it follows that the minimum regret over all predictors $\hat{p}$ satisfies

$$\min_{\hat{p}} R_\alpha(\pi, \hat{p}) = \min_{\hat{p}} D_\alpha(\pi_\alpha \times p_\theta \| \pi_\alpha \times \hat{p}) = I_\alpha(\pi_\alpha, p_\theta). \tag{105}$$

Furthermore, one can check by substituting the definition of luckiness α-NML (74) with prior π_α into (75), that the regret of the luckiness α-NML is equal to

$$R_\alpha(\pi, \hat{p}_\alpha) = I_\alpha(\pi_\alpha, p_\theta). \tag{106}$$

From (105) and (106) it follows that the Luckiness α-NML minimizes the regret.

Appendix B
Proof of Theorem 5

For the discrete memoryless sources case, one can rewrite (65) as

$$\begin{aligned} W_\alpha(\mathcal{P}) &= \max_{\boldsymbol{n}} \log \frac{\max_{\boldsymbol{\theta}} \prod_{i=1}^m \theta_i^{n_i}}{\left\{ \frac{\Gamma(\frac{m}{2})}{\pi^{m/2}} \frac{\prod_{i=1}^m \Gamma(\alpha n_i + \frac{1}{2})}{\Gamma(\alpha n + \frac{m}{2})} \right\}^{1/\alpha}} && (107) \\ &= \max_{\boldsymbol{n}} \log \frac{\prod_{i=1}^m (\frac{n_i}{n})^{n_i}}{\left\{ \frac{\Gamma(\frac{m}{2})}{\pi^{m/2}} \frac{\prod_{i=1}^m \Gamma(\alpha n_i + \frac{1}{2})}{\Gamma(\alpha n + \frac{m}{2})} \right\}^{1/\alpha}} && (108) \\ &= \frac{1}{\alpha} \log \frac{\pi^{m/2} \Gamma(\alpha n + \frac{m}{2})}{\Gamma(\frac{m}{2})} - n \log n \\ &\qquad + \max_{\boldsymbol{n}} \sum_{i=1}^m \left(n_i \log n_i - \frac{1}{\alpha} \log \Gamma\left(\alpha n_i + \frac{1}{2}\right)\right) && (109) \end{aligned}$$

where the maximization is over vectors $\boldsymbol{n} = (n_1, n_2, \ldots, n_m)$ with integer entries such that $\sum_{i=1}^m n_i = n$ and $n_i \geq 0$ for every i. Notice that to prove the theorem, it suffices to show that the quantity

$$\sum_{i=1}^m \left(n_i \log n_i - \frac{1}{\alpha} \log \Gamma\left(\alpha n_i + \frac{1}{2}\right)\right) \tag{110}$$

is maximized for $n_m = n$ and $n_i = 0$ for every $i \neq m$, for every $n \geq 1$ and $m \geq 2$. We prove this by induction on m. For $m = 2$, let $t = \frac{n_1}{n}$, $0 \leq t \leq 1$. Then, we wish to prove that the function

$$f(t) = nt\log(nt) - \frac{1}{\alpha}\log\Gamma\left(\alpha nt + \frac{1}{2}\right) + n(1-t)\log(n(1-t)) - \frac{1}{\alpha}\log\Gamma\left(\alpha n(1-t) + \frac{1}{2}\right) \quad (111)$$

is maximized at $t = 1$ for $0 \leq t \leq 1$. Notice that $f(t)$ is symmetrical around $t = \frac{1}{2}$. Hence, it suffices to prove that $f(t)$ is convex for $0 \leq t \leq 1$, and to prove this it is enough to show that

$$g(t) = nt\log(nt) - \frac{1}{\alpha}\log\Gamma\left(\alpha nt + \frac{1}{2}\right) \quad (112)$$

is convex for $0 \leq t \leq 1$. Notice that

$$\begin{aligned} g'(t) &= n\log(nt) + n - n\,\psi\left(\alpha nt + \frac{1}{2}\right) && (113) \\ &= n - n\log\alpha + n\log(\alpha nt) - n\,\psi\left(\alpha nt + \frac{1}{2}\right) && (114) \\ &= n - n\log\alpha + n\,h(\alpha nt) && (115) \end{aligned}$$

where $\psi(x) = \frac{d}{dx}\log\Gamma(x)$ is the digamma function, and

$$h(x) = \log(x) - \psi\left(x + \frac{1}{2}\right). \quad (116)$$

By [24, Theorem 4.2], it follows that $h'(x) \geq 0$ for every $x \geq 0$. Therefore, one has

$$g''(t) = \alpha n^2 h'(\alpha nt) \geq 0 \quad (117)$$

for every $0 \leq t \leq 1$, i.e., $g(t)$ is convex. Hence, $f(t)$ is maximized at $t = 1$, and the case $m = 2$ is proved. Assume now that the case $m = k$ is true, i.e., that (110) is maximized for $n_k = n$ and $n_i = 0$ for $i \neq k$, for every $n \geq 1$. Consider the case $m = k + 1$. For every $\boldsymbol{n} = (n_1, n_2, \ldots, n_{k+1})$, one has

$$\begin{aligned} &\sum_{i=1}^{m}\left(n_i\log n_i - \frac{1}{\alpha}\log\Gamma\left(\alpha n_i + \frac{1}{2}\right)\right) && (118) \\ &= \sum_{i=1}^{k}\left(n_i\log n_i - \frac{1}{\alpha}\log\Gamma\left(\alpha n_i + \frac{1}{2}\right)\right) \\ &\qquad + n_{k+1}\log n_{k+1} - \frac{1}{\alpha}\log\Gamma\left(\alpha n_{k+1} + \frac{1}{2}\right) && (119) \\ &\leq -\sum_{i=1}^{k-1}\frac{1}{\alpha}\log\Gamma\left(\frac{1}{2}\right) + (n - n_{k+1})\log(n - n_{k+1}) \\ &\qquad - \frac{1}{\alpha}\log\Gamma\left(\alpha(n - n_{k+1}) + \frac{1}{2}\right) + n_{k+1}\log n_{k+1} \\ &\qquad - \frac{1}{\alpha}\log\Gamma\left(\alpha n_{k+1} + \frac{1}{2}\right) && (120) \\ &\leq -\sum_{i=1}^{k}\frac{1}{\alpha}\log\Gamma\left(\frac{1}{2}\right) + n\log n - \frac{1}{\alpha}\log\Gamma\left(\alpha n + \frac{1}{2}\right), && (121) \end{aligned}$$

where the first inequality follows from the case $m = k$, and the second inequality follows from the case $m = 2$. Thus, (121) shows that (118) is maximized for $n_{k+1} = n$, as desired. Hence, the case $m = k + 1$ is proved, and the theorem follows.

Appendix C
Proof of Theorem 6

We start from Equation (64). The asymptotics of the α-mutual information term indirectly follows from the proof of Theorem 2 in [12]. In fact, the theorem states that

$$\sup_{w \in \mathcal{P}(\Theta)} I_\alpha(\phi, X^n) = \frac{m-1}{2} \log \frac{n}{2} + \frac{1}{2} \log \pi - \log \Gamma\left(\frac{m}{2}\right) - \frac{m-1}{2(\alpha-1)} \log \alpha + o(1)\,, \tag{122}$$

from which it follows that

$$I_\alpha(\phi, X^n) \leq \frac{m-1}{2} \log \frac{n}{2} + \frac{1}{2} \log \pi - \log \Gamma\left(\frac{m}{2}\right) - \frac{m-1}{2(\alpha-1)} \log \alpha + o(1) \tag{123}$$

for $(\phi, X^n) \sim w(\phi)\, p_\theta(X^n)$ and w taken as the Dirichlet distribution $\mathrm{Dir}(1/2, \ldots, 1/2)$. However, again in [12], from Equation (80) onwards they also prove that

$$I_\alpha(\phi, X^n) \geq \frac{m-1}{2} \log \frac{n}{2} + \frac{1}{2} \log \pi - \log \Gamma\left(\frac{m}{2}\right) - \frac{m-1}{2(\alpha-1)} \log \alpha + o(1)\,. \tag{124}$$

Therefore, equations (123) and (124) show that

$$I_\alpha(\phi, X^n) = \frac{m-1}{2} \log \frac{n}{2} + \frac{1}{2} \log \pi - \log \Gamma\left(\frac{m}{2}\right) - \frac{m-1}{2(\alpha-1)} \log \alpha + o(1)\,. \tag{125}$$

We are now left with the $W_\alpha(\mathcal{P})$ term. Starting from (102), we want to find the asymptotics of the first logarithm, which is the only term dependent on n. From [25] we have that

$$\lim_{t \to \infty} t^{b-a} \frac{\Gamma(t+a)}{\Gamma(t+b)} = 1 \tag{126}$$

for all real numbers a and b. Therefore, we also have

$$\lim_{n \to \infty} \left[\log \frac{\Gamma(\alpha n + \frac{m}{2})}{\Gamma(\alpha n + \frac{1}{2})} - \frac{m-1}{2} \log(\alpha n) \right] \tag{127}$$

$$= \lim_{n \to \infty} \log \left[(\alpha n)^{\frac{1}{2} - \frac{m}{2}} \frac{\Gamma(\alpha n + \frac{m}{2})}{\Gamma(\alpha n + \frac{1}{2})} \right] = 0\,, \tag{128}$$

or equivalently,

$$\log \frac{\Gamma(\alpha n + \frac{m}{2})}{\Gamma(\alpha n + \frac{1}{2})} = \frac{m-1}{2} \log(\alpha n) + o(1)\,. \tag{129}$$

Plugging this into (102) gives

$$W_\alpha(\mathcal{P}) = \frac{m-1}{2\alpha} \log(\alpha n) + \frac{1}{2\alpha} \log \pi - \frac{1}{\alpha} \Gamma\left(\frac{m}{2}\right) + o(1)\,. \tag{130}$$

Finally, plugging this and (125) into (64) leads to (103).

REFERENCES

[1] N. Merhav and M. Feder, “Universal prediction,” *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2124–2147, 1998.

[2] J. Ziv and A. Lempel, “A universal algorithm for sequential data compression,” *IEEE Trans. Inf. Theory*, vol. 23, no. 3, pp. 337–343, 1977.

[3] F. Willems, Y. Shtarkov, and T. Tjalkens, “The context-tree weighting method: basic properties,” *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 653–664, 1995.

[4] Q. Xie and A. Barron, “Asymptotic minimax regret for data compression, gambling, and prediction,” *IEEE Trans. Inf. Theory*, vol. 46, no. 2, pp. 431–445, 2000.

[5] Y. Fogel and M. Feder, “Universal learning of individual data,” in *Proc. 2019 IEEE Int. Symp. Inf. Theory (ISIT)*, 2019, pp. 2289–2293.

[6] F. E. Rosas, P. A. M. Mediano, and M. Gastpar, “Learning, compression, and leakage: Minimising classification error via meta-universal compression principles,” in *Proc. 2020 IEEE Inf. Theory Workshop (ITW)*, 2021.

[7] N. Cesa-Bianchi and G. Lugosi, *Prediction, learning, and games*. Cambridge, United Kingdom: Cambridge University Press, 2006.

[8] Y. M. Shtarkov, “Universal sequential coding of single messages,” *Problems Inform. Transmission*, vol. 23, no. 3, pp. 175–186, 1987.

[9] R. Krichevsky and V. Trofimov, “The performance of universal encoding,” *IEEE Trans. Inf. Theory*, vol. 27, no. 2, pp. 199–207, 1981.

[10] S. Verdú, “α-mutual information,” in *Proc. 2015 IEEE Inf. Theory Appl. Workshop (ITA)*, 2015.

[11] C. Cai and S. Verdú, “Conditional Rényi divergence saddlepoint and the maximization of α-mutual information,” *Entropy*, vol. 21, no. 10, 2019.

[12] S. Yagli, Y. Altuğ, and S. Verdú, “Minimax Rényi redundancy,” *IEEE Trans. Inf. Theory*, vol. 64, no. 5, pp. 3715–3733, 2018.

[13] M. Drmota and W. Szpankowski, “Precise minimax redundancy and regret,” *IEEE Transactions on Information Theory*, vol. 50, no. 11, pp. 2686–2707, 2004.

[14] T. Li, A. Beirami, M. Sanjabi, and V. Smith, “On tilted losses in machine learning: Theory and applications,” *Journal of Machine Learning Research*, vol. 24, no. 142, pp. 1–79, 2023. [Online]. Available: http://jmlr.org/papers/v24/21-1095.html

[15] R. Sibson, “Information radius,” *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, no. 14, pp. 149–160, 1969.

[16] L. Campbell, “A coding theorem and Rényi’s entropy,” *Information and Control*, vol. 8, no. 4, pp. 423–429, 1965.

[17] T. van Erven and P. Harremoës, “Rényi divergence and Kullback-Leibler divergence,” *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 3797–3820, 2014.

[18] R. G. Gallager, “Source coding with side information and universal coding.” [Online]. Available: https://web.mit.edu/gallager/www/papers/paper5.pdf

[19] P. D. Grünwald, *The minimum description length principle*. Cambridge, MA: MIT Press, 2007.

[20] P. Grünwald and T. Roos, “Minimum description length revisited,” *International Journal of Mathematics for Industry*, vol. 11, no. 01, 2019.

[21] L. Davisson, “Universal noiseless coding,” *IEEE Trans. Inf. Theory*, vol. 19, no. 6, pp. 783–795, 1973.

[22] J. Rissanen, “Complexity of strings in the class of Markov sources,” *IEEE Trans. Inf. Theory*, vol. 32, no. 4, pp. 526–532, 1986.

[23] H. Jeffreys, “An invariant form for the prior probability in estimation problems,” *Proceedings of the Royal Society A*, vol. 186, pp. 453–461, 1946.

[24] H. Alzer and C. Berg, “Some classes of completely monotonic functions, II,” *The Ramanujan Journal*, vol. 11, no. 2, pp. 225–248, 2006.

[25] M. Abramowitz and I. A. Stegun, *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. Gaithersburg, MD: National Bureau of Standards, 1970.