Maximum Likelihood vs. Sequential Normalized Maximum Likelihood in On-line Density Estimation

Wojciech Kotłowski
Centrum Wiskunde & Informatica
kotlowsk@cwi.nl

Peter Grünwald
Centrum Wiskunde & Informatica
pdg@cwi.nl

Abstract

The paper considers sequential prediction of individual sequences with log loss (online density estimation) using an exponential family of distributions. We first analyze the regret of the maximum likelihood ("follow the leader") strategy. We find that this strategy is (1) suboptimal and (2) requires an additional assumption about boundedness of the data sequence. We then show that both problems can be be addressed by adding the currently predicted outcome to the calculation of the maximum likelihood, followed by normalization of the distribution. The strategy obtained in this way is known in the literature as the *sequential normalized maximum likelihood* or *last-step minimax* strategy. We show for the first time that for general exponential families, the regret is bounded by the familiar $(k/2)\log n$ and thus optimal up to $O(1)$. We also show the relationship to the Bayes strategy with Jeffreys' prior.

1 Introduction

The game of sequential prediction of individual sequences with log loss (online density estimation) is defined in the following way. Let $x_1, x_2, \ldots \in \mathcal{X}^*$, be a sequence of outcomes revealed one at a time. After observing $x^n = x_1, x_2, \ldots, x_n$, a forecaster assigns a probability distribution on $\mathcal{X}$, denoted $P(\cdot \mid x^n)$. Then, after x_{n+1} is revealed, the forecaster incurs the *log loss* $-\log P(x_{n+1} \mid x^n)$. The performance of the strategy is measured relative to the best in a reference set of strategies, which we call the *model* $\mathcal{P}$. The difference between the accumulated loss of the prediction strategy and the best strategy in the model is called the *regret.* The goal is to minimize the regret in the worst case over all possible data sequences.

We assume the model $\mathcal{P} = \{P_\theta \mid \theta \in \Theta\}$ is an exponential family of distributions, examples of which include normal, Bernoulli, multinomial, Gamma, Poisson, Pareto, geometric distributions and many others. If there is a known time horizon n of the game (maximal number of outcomes), the minimax strategy for the game is the *normalized maximum likelihood* (NML) strategy (Shtarkov, 1987, Rissanen, 1996). If the parameter space of a k-dimensional exponential family is constrained to a compact subset Θ_0, NML achieves regret $\frac{k}{2}\log n + O(1)$ for all data sequences. NML, however, requires knowledge of the time horizon and is impractical to calculate in many situations. A particularly simple and popular prediction strategy is the *maximum likelihood* (ML) strategy, (also known as "follow the leader"), which predicts the next outcome x_n by using the distribution $P_{\hat{\mu}_{n-1}}$, with $\hat{\mu}_{n-1}$ being the ML estimator based on the $n-1$ past outcomes. The ML strategy, contrary to NML, belongs to the family of *plug-in* strategies which in each iteration predict with one of the strategies from the model.

Despite the popularity of the "follow the leader" approach, guarantees on the regret of the ML strategy were only obtained for some special exponential families, such as normal, Bernoulli and Gamma distributions (Freund, 1996, Azoury and Warmuth, 2001). In this paper, we prove general bounds which hold for any exponential family. We show (Theorem 1, Section 3) that if the parameter space is constrained to a compact subset Θ_0, and if the outcomes are bounded within a ball of radius B, the regret can be upper bounded by $C\log n + O(1)$, where C is a constant depending on B and Θ_0. We also prove (Theorem 2) that the bound is essentially tight. In other words, (1) the ML strategy requires boundedness of the data sequence, and (2) the rate of the regret growth is still logarithmic, but the constant in front of $\log n$ can be very large, especially when B is large. Moreover, Theorem 2 implies that those two drawbacks are shared among all plug-in strategies.

We also show, however, that both problems can be be addressed by adding the currently predicted outcome to the calculation of the maximum likelihood, followed by normalization of the distribution. Typically, the new strategy predicts with a distribution $P(x_n|x^{n-1})$ proportional to $P_{\hat{\mu}_n}(x_n)$. Usually, this distribution will not be equal to any of the distributions P_μ within the model, so the new strategy is not plug-in. The strategy obtained in this way is known as *sequential normalized maximum likelihood* (Rissanen and Roos, 2007, Roos and Rissanen, 2008) (SNML). It was discovered with a different motivation in mind: Rissanen and Roos noticed that its predictions coincide with those of the NML distribution under the assumption that the current iteration is the last iteration. Therefore, it can be viewed as an approximation to NML for which the time horizon of the game does not need to be known. A similar idea, though restricted to strategies within the model (plug-in strategies), was introduced by Takimoto and Warmuth (2000) under the name *last-step minimax*.

In this paper, we develop bounds on the worst-case regret for SNML for general exponential families (such bounds had been unknown so far). As our main result, in Theorem 4, we prove that the regret of the SNML strategy is at most $\frac{k}{2}\log n + O(1)$, which matches, up to the $O(1)$ term, the minimax regret bound. This issue is important from a practical point of view, as SNML constitutes an interesting and effective algorithm for online density estimation and model selection. However, our results are also interesting from a conceptual point of view, as the answer to the following question: how much do we loose if we base our decision in a given moment by looking only one step ahead instead of looking at the whole possible future up to a given time horizon? Our results suggest that we do not loose anything substantial, at least asymptotically; for some models, it turns out that we don't even loose anything: in Section 5 we show that in some cases (but not always) the SNML strategy coincides with the *Bayes* strategy, when the prior distribution is chosen to be a *Jeffreys' prior*. Moreover, we prove that when the two strategies are equal, they are also equal to the NML strategy and thus minimax optimal.

Related Work Sequential prediction with log loss has been extensively studied in learning theory, in the framework of *prediction with expert advice* (Cesa-Bianchi and Lugosi, 2006). It also plays an important role in information theory: a key result based on the Kraft inequality (Cover and Thomas, 1991) states that, ignoring rounding issues, for every length function L of a uniquely decodable code, there is a probability distribution P such that $L(x) = -\log P(x)$ and vice versa. Thus, at least when $\mathcal{X}$ is countable, any prediction strategy can also be thought of as a *universal source coding algorithm*; the cumulative logarithmic loss corresponds exactly to the incurred codelength. As Rissanen's theory of Minimum Description Length (MDL) learning (Barron et al., 1998, Grünwald, 2007) is based on universal coding, a sequential prediction strategy with log loss defines an MDL model selection criterion. Similarly, in statistics, Dawid's theory of prequential model assessment (Dawid, 1984) is based on sequential prediction.

The ML strategy for exponential families was considered by Freund (1996) and Azoury and Warmuth (2001), with regret bounds proven for the particular cases of normal, Bernoulli and Gamma distributions. Grünwald and de Rooij (2005) showed the following: let the model $\mathcal{P}$ be an arbitrary 1-dimensional exponential family ($k = 1$). Suppose the outcomes are i.i.d. by some distribution P^*, possibly outside the model. Then the expected regret of the ML plug-in strategy is $(1/2c)\log n + O(1)$ where c is the variance of an outcome under the true distribution P^* divided by the variance under the element of the model P_θ that minimizes the Kullback-Leibler divergence $D(P^*\|P_\theta)$. In general, c can be much smaller than 1. Moreoever, it was shown by Grünwald and Kotłowski (2010) that *no* plug-in estimator can achieve $c = 1$. Our Theorems 1 and 2 are essentially extensions of this result to individual-sequence settings. Dasgupta and Hsu (2007) considered Gaussian density estimation with unknown mean and covariance matrix and obtained a much worse linear bound on the regret, excluding the possibility of a logarithmic bound. Their results do not contradict ours, since the set of reference strategies (distributions in the exponential family) was not constrained to be in a compact subset of the parameter space, which is necessary to obtain logarithmic bounds (interestingly, if one considers the regret conditioned on the first outcome then for some models it is possible after all to get logarithmic bound even with full parameter space, as we show in Section 5; we pose as an open problem whether this result extends to arbitrary exponential families). Raginsky et al. (2009) considered a plug-in strategy based on Bregman projections and proved regret bounds for general exponential families; their strategy, however, is different from those considered here. Hazan et al. (2007) proved logarithmic regret bounds on the follow the leader strategy in online convex optimization, however the assumptions of their theorem do not match online density estimation with exponential families. Kotłowski et al. (2010) considered "folowing the 'flattened' leader", an improvement over the ML strategy, "slightly" outside the model, achieving the optimal regret bound. However, this flattened-leader strategy still requires boundedness of the data sequence.

The idea of including the current observation to the calculation of maximum likelihood was considered by (Rissanen and Roos, 2007, Roos and Rissanen, 2008), though with a different motivation in mind. The regret bounds were not given apart from specific cases. A similar idea, though restricted to strategies within the model (plug-in strategies), was introduced by Takimoto and Warmuth (2000) under the name *last-step minimax* (the relation is made precise in Section 4).

The paper is organized as follows. We introduce the mathematical context for our results in Section 2. We then analyze the ML strategy in Section 3, proving the regret bounds which reveal suboptimal behavior in the worst case. Then, we introduce the SNML strategy in Section 4 and prove optimal regret bounds. We give some examples for particular exponential families in Section 5 and discuss the relationship between SNML and Bayes with Jeffreys' prior in Section 6. We end with a conclusion in Section 7.

2 Notation and Definitions

2.1 Exponential Family

Let $\mathcal{X}$ be a set of outcomes, taking values either in a finite or countable set, or in a subset of Euclidean space. Exponential family models (Barndorff-Nielsen, 1978) are families of distributions on $\mathcal{X}$ with densities $P_\theta(x) = e^{\theta^T \phi(x) - \psi(\theta)} h(x)$, defined relative to a random variable $\phi : \mathcal{X} \to \mathbb{R}^k$ (called *sufficient statistic*) and a function $h : \mathcal{X} \to [0, \infty)$. The function $\psi(\theta) = \log \int_{x \in \mathcal{X}} e^{\theta^T \phi(x)} h(x) \, \mathrm{d}x$ (the integral to be replaced by a sum for countable $\mathcal{X}$) is called a *partition function*, and $\Theta = \{\theta \in \mathbb{R}^k : \psi(\theta) < \infty\}$ is called the *natural parameter space.* We only consider *regular* exponential families, when Θ is an open and convex subset of $\mathbb{R}^k$, and the representation is *minimal*, i.e. the functions $\phi_i(x), i = 1, \ldots, k$, are linearly independent. Moreover, without loss of generality, we will make the simplifying assumption that $\phi(x) \equiv x$, i.e. the exponential family is in the canonical form. All results in this paper are valid for a more general ϕ. The function $\psi(\theta)$ is differentiable infinitely often, and strictly convex on Θ. A standard result for exponential families states (Barndorff-Nielsen, 1978) that the gradient $\mu = \nabla_\theta \psi(\theta)$ is the mean value vector of x, $\mu = \mathbb{E}_\theta[x]$, while the Hessian $\nabla^2_\theta \psi(\theta) = E_\theta[-\nabla^2_\theta \log P_\theta(x)] = I(\theta)$ coincides with the *Fisher information* matrix in the natural parameterization, is positive definite and equal to the covariance matrix $\mathrm{Cov}_\theta(x)$. Strict convexity of ψ implies that the function $\mu(\theta)$ is invertible, and thus suggests reparameterizing the distribution by μ. The function $\mu(\theta)$ maps parameters in the natural parameterization to the *mean value* parameterization $\Xi = \mu(\Theta)$. It is a diffeomorphism (Barndorff-Nielsen, 1978), and thus Ξ is also an open convex set of $\mathbb{R}^k$. The inverse $\theta(\mu)$ maps back to the natural parametrization. Moreover, $\nabla_\mu \theta(\mu) = E_\mu[-\nabla^2_\mu \log P_\mu(x)] = I(\mu)$ is the Fisher information in the mean-value parametrization, which is equal to to the inverse covariance matrix $\mathrm{Cov}^{-1}_\mu(x)$.

The KL-divergence between distributions P_θ and $P_{\theta'}$:

$$\mathbb{E}_\theta \left[\log \frac{P_\theta(x)}{P_{\theta'}(x)} \right] = \mathbb{E}_\mu \left[\log \frac{P_\mu(x)}{P_{\mu'}(x)} \right] = (\theta - \theta')^T \mu - \psi(\theta) + \psi(\theta'), \tag{1}$$

where $\mu = \mu(\theta)$ and $\mu' = \mu(\theta')$, is denoted by $D(\theta \| \theta')$ or by $D(\mu \| \mu')$, depending on the context.

Let $\Theta_0 \subseteq \Theta$ be any nonempty convex subset of Θ. Given the data sequence x^n, the *maximum likelihood* (ML) estimate $\hat{\theta}_n$ relative to Θ_0 is defined as:

$$\hat{\theta}_n = \arg \max_{\theta \in \Theta_0} P_\theta(x^n) = \arg \min_{\theta \in \Theta_0} - \log P_\theta(x^n), \tag{2}$$

or equivalently as:

$$\hat{\mu}_n := \mu(\hat{\theta}_n) = \arg \min_{\mu \in \Xi_0} - \log P_\mu(x^n),$$

where $\Xi_0 = \mu(\Theta_0)$ is also convex. By rewriting $-\log P_\theta(x^n) = -n(\theta^T \bar{x}_n - \psi(\theta)) - \log h(x^n)$, where $\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i$, we see that $\nabla_\theta - \log P_\theta(x^n) = -n(\bar{x}_n - \mu(\theta))$. This means than when $\bar{x}_n \in \Xi_0$, then $\hat{\mu}_n = \bar{x}_n$. More generally, we can exploit the fact that $-\log P_\theta(x^n)$ is a convex function of θ, and thus the necessary condition for a minimum of a convex function on a convex set states (Boyd and Vandenberghe, 2004) that $\nabla_{\hat{\theta}_n} - \log P_{\hat{\theta}_n}(x^n)(\theta - \hat{\theta}_n) \geq 0$ for all $\theta \in \Theta_0$, which implies:

$$(\hat{\mu}_n - \bar{x}_n)^T (\theta - \hat{\theta}_n) \geq 0 \tag{3}$$

for all $\theta \in \Theta_0$. Condition (3) has a nice interpretation in terms of Bregman projections. Assuming $\bar{x}_n \in \Xi$, we can rewrite (3) as:

$$D(\bar{x}_n \| \mu) \geq D(\hat{\mu}_n \| \mu) + D(\bar{x}_n \| \hat{\mu}_n),$$

for all $\mu \in \Xi_0$, which is closely related to the generalized Pythagorean inequality for Bregman divergences (Cesa-Bianchi and Lugosi, 2006). Another expression which we are going to use is:

$$D(\mu_1 \| \mu_2) - D(\mu_1 \| \mu_3) = D(\mu_3 \| \mu_2) + (\theta_2 - \theta_3)^T (\mu_3 - \mu_1), \tag{4}$$

for all $\mu_1, \mu_2, \mu_3 \in \Xi_0$, where $\theta_i = \theta(\mu_i)$, $i = 1, 2, 3$. This can be derived by writing KL-divergences on both sides according to (1).

2.2 Sequential Prediction

At every iteration $n = 1, 2, \ldots$, the prediction $P(\,\cdot \mid x^{n-1})$ depends on the past outcomes x^{n-1} and has the form of a probability distribution on $\mathcal{X}$, and therefore can be considered as a conditional of the joint distribution of outcomes in $\mathcal{X}^n$, which is $P(x^n) = \prod_{i=1}^n P(x_i|x^{i-1})$. Conversely, any probability distribution P on the set $\mathcal{X}^n$ defines a prediction strategy induced by its conditional distributions $P(\,\cdot \mid x^i)$ for $0 \leq i < n$ (Cesa-Bianchi and Lugosi, 2006, Grünwald, 2007). The performance of the strategy P on the outcome sequence x^n is measured relative to the best strategy in the model (reference set of strategies) $\mathcal{P}$ by the *regret*, defined as:

$$\mathcal{R}(P; x^n) = \sum_{i=1}^n -\log P(x_i|x^{i-1}) - \inf_{P_\theta \in \mathcal{P}} \sum_{i=1}^n -\log P_\theta(x_i) = -\log P(x^n) - \inf_{P_\theta \in \mathcal{P}} -\log P_\theta(x^n). \tag{5}$$

The regret is the difference in cumulative losses incurred so far by the prediction strategy and the best strategy (distribution) in the model. We are usually interested in the worst-case regret, $\mathcal{R}(P; n) = \sup_{x^n} \mathcal{R}(P; x^n)$. Unfortunately, for most common exponential families $\mathcal{P} = \{P_\theta \colon \theta \in \Theta\}$ (extended to sequences by the i.i.d. assumption), $\mathcal{R}(P; n)$ cannot be made finite even for $n = 1$, whatever P is. Indeed, let $\mathcal{P}$ be a one-dimensional normal family $N(\theta, 1)$ with fixed unit variance, so that $\Theta = \mathbb{R}$. For every strategy P, we must have $P(x_1) \to 0$ as $x_1 \to \infty$ (otherwise P would not be normalizable), and therefore $-\log P(x_1) \to \infty$. On the other hand, $\inf_{P_\theta \in \mathcal{P}} -\log P_\theta(x_1) = -\log P_{x_1}(x_1) = \frac{1}{2} \log 2\pi$, so that by increasing x_1, we can make the regret as large as we want.

Therefore, to obtain non-trivial regret bounds, we choose a compact convex subset $\Theta_0 \subset \Theta$ and define $\mathcal{P} = \{P_\theta \colon \theta \in \Theta_0\}$; equivalently, we can choose $\Xi_0 = \mu(\Theta_0)$ and define $\mathcal{P} = \{P_\mu \colon \mu \in \Xi_0\}$ (we will use both interchangeably). Then,

$$\mathcal{R}(P; x^n) = -\log P(x^n) + \log P_{\hat{\mu}_n}(x^n),$$

where $\hat{\mu}_n$ is the ML estimator relative to Ξ_0.

Let P be a prediction strategy. If for every n, $P(x_n|x^{n-1}) \in \mathcal{P}$, i.e. $P(x_n|x^{n-1}) = P_{\bar{\mu}_{n-1}}(x_n)$ for some $\bar{\mu}_{n-1} = \bar{\mu}_{n-1}(x^{n-1})$, we call such P a *plug-in strategy.* In other words, a plug-in strategy always predicts with one of the distributions from the model. An example of a plug-in strategy is the *maximum likelihood* (or *follow the leader*) strategy defined as $P_{\mathrm{ML}}(x_n|x^{n-1}) = P_{\hat{\mu}_{n-1}}(x_n)$.

There is, however, an advantage in using strategies which are not in the model. An important out-model strategy is the *normalized maximum likelihood (NML)* strategy, defined as:

$$P_{\mathrm{NML}}(x^n) = \frac{\sup_{\theta \in \Theta_0} P_\theta(x^n)}{\int_{\mathcal{X}^n} \sup_{\theta \in \Theta_0} P_\theta(z^n) \, \mathrm{d}z^n} = \frac{\sup_{\mu \in \Xi_0} P_\mu(x^n)}{\int_{\mathcal{X}^n} \sup_{\mu \in \Xi_0} P_\mu(z^n) \, \mathrm{d}z^n}. \tag{6}$$

NML is known to be the *minimax* prediction strategy for the log-loss game: it can be shown (see, e.g. Cesa-Bianchi and Lugosi (2001)) that:

$$\inf_P \sup_{x^n} \mathcal{R}(P; x^n) = \mathcal{R}(P_{\mathrm{NML}}; n),$$

where the infimum is over all, both in-model (plug-in) and out-model, prediction strategies. The value of $\mathcal{R}(P_{\mathrm{NML}}; n)$ is also known: if $\mathcal{P}$ is a k-dimensional exponential family and Ξ_0 is a closed convex subset of Ξ with non-empty interior, then

$$\mathcal{R}(P_{\mathrm{NML}}, x^n) = \frac{k}{2} \log n + O(1). \tag{7}$$

For a proof, see e.g. (Grünwald, 2007). (7) is the famous 'k over 2 log n formula', refinements of which lie at the basis of practical approximations to MDL and Bayesian learning (Grünwald, 2007). Since the NML strategy is minimax, a worst-case regret of $\frac{k}{2} \log n + O(1)$ is optimal.

3 Regret Bounds for ML Strategy

In this section, we analyze the performance of ML strategy and show, that under an additional boundedness assumption on the data sequence, one can get a non-trivial regret bound. The bound, however, reveals suboptimal behavior and dependence on the range of the data sequence. Later, we show the bound is essentially unimprovable for any plug-in strategy.

Before we prove the main theorem, we need several propositions, which will also be useful in the next section, while proving the regret bound for SNML.

Proposition 1 *Let* $\bar{y}_n = \frac{(n-1)\hat{\mu}_{n-1}+x_n}{n}$, *and let* $\tilde{\mu}_n = \arg\min_{\mu \in \Xi_0} D(\bar{y}_n \| \mu)$. *Then:*

$$-\log P_{\hat{\mu}_{n-1}}(x^n) + \log P_{\hat{\mu}_n}(x^n) \leq nD(\bar{y}_n \| \hat{\mu}_{n-1}) - nD(\bar{y}_n \| \tilde{\mu}_n). \tag{8}$$

Proof: From the definition of $\tilde{\mu}_n$, we have $D(\bar{y}_n \| \tilde{\mu}_n) \leq D(\bar{y}_n \| \hat{\mu}_n)$, so that:

$$\begin{aligned}
D(\bar{y}_n \| \hat{\mu}_{n-1}) - D(\bar{y}_n \| \tilde{\mu}_n) &\geq D(\bar{y}_n \| \hat{\mu}_{n-1}) - D(\bar{y}_n \| \hat{\mu}_n) && \text{(from definition of } \tilde{\mu}_n\text{)} \\
&= D(\hat{\mu}_n \| \hat{\mu}_{n-1}) + (\hat{\theta}_{n-1} - \hat{\theta}_n)^T (\hat{\mu}_n - \bar{y}_n) && \text{(from (4))} \\
&= D(\hat{\mu}_n \| \hat{\mu}_{n-1}) + (\hat{\theta}_{n-1} - \hat{\theta}_n)^T (\hat{\mu}_n - \bar{x}_n) + (\hat{\theta}_{n-1} - \hat{\theta}_n)^T (\bar{x}_n - \bar{y}_n) \\
&= D(\hat{\mu}_n \| \hat{\mu}_{n-1}) + (\hat{\theta}_{n-1} - \hat{\theta}_n)^T (\hat{\mu}_n - \bar{x}_n) + \frac{n}{n-1} (\hat{\theta}_{n-1} - \hat{\theta}_n)^T (\bar{x}_{n-1} - \hat{\mu}_{n-1}) \\
&\geq D(\hat{\mu}_n \| \hat{\mu}_{n-1}) + (\hat{\theta}_{n-1} - \hat{\theta}_n)^T (\hat{\mu}_n - \bar{x}_n) && \text{(from (3))} \\
&= -(\hat{\theta}_{n-1} - \hat{\theta}_n)^T \bar{x}_n + \psi(\hat{\theta}_{n-1}) - \psi(\hat{\theta}_n) \\
&= \frac{1}{n} \left(-\log P_{\hat{\mu}_{n-1}}(x^n) + \log P_{\hat{\mu}_n}(x^n) \right).
\end{aligned}$$

■

Proposition 1 states that if we pretend $\hat{\mu}_{n-1}$ (rather than $\bar{x}_{n-1}$) is the sufficient statistic in the previous iteration, and do all the updates accordingly, then the drop of the KL divergence from the data to the ML estimator per iteration will decrease. Thus, if we imagine the data is generated by an adversary trying to maximize the regret, it does not pay for him/her to choose $\bar{x}_{n-1}$ outside Ξ_0 (since then $\hat{\mu}_{n-1} \neq \bar{x}_{n-1}$).

We now show that within a compact set Ξ_0, the KL-divergence behaves approximately as a quadratic form:

Proposition 2 *Let* Ξ_1 *be a compact subset of* Ξ. *Then, for all* $\mu, \mu' \in \Xi_1$,

$$D(\mu \| \mu') \leq \frac{1}{2} (\mu - \mu')^T I(\mu') (\mu - \mu') + C \| \mu - \mu' \|^3,$$

where $C < \infty$ *depends on* Ξ_1.

Proof: We need two standard results regarding the properties of KL divergence (see, e.g. Barndorff-Nielsen (1978), Grünwald (2007)): for any $\mu, \mu' \in \Xi$, it holds:

1. $D(\mu \| \mu') \geq 0$ and the equality only holds for $\mu = \mu'$,
2. For exponential families, $\nabla^2_{\mu} D(\mu \| \mu') = I(\mu)$.

By Taylor expanding $D(\mu \| \mu')$ around μ' up to the second order, we get:

$$D(\mu \| \mu') = D(\mu' \| \mu') + \nabla_{\mu} D(\mu \| \mu')^T \big|_{\mu = \mu'} (\mu - \mu') + \frac{1}{2} (\mu - \mu')^T I(\bar{\mu}) (\mu - \mu'),$$

for some $\bar{\mu}$ between μ and μ'. Due to the first property the zeroth order term disappears; the second order term also disappears because the gradient vanishes at the minimum, so we have:

$$\begin{aligned}
D(\mu \| \mu') = \frac{1}{2} (\mu - \mu')^T I(\bar{\mu}) (\mu - \mu') &= \frac{1}{2} (\mu - \mu')^T I(\mu') (\mu - \mu') + \frac{1}{2} (\mu - \mu')^T \left(I(\bar{\mu}) - I(\mu') \right) (\mu - \mu') \\
&\leq \frac{1}{2} (\mu - \mu')^T I(\mu') (\mu - \mu') + \frac{1}{2} \| I(\bar{\mu}) - I(\mu') \| \| \mu - \mu' \|^2,
\end{aligned} \tag{9}$$

where $\| \cdot \|$ denotes vector or matrix norm, depending on the context. Taylor expanding $I(\bar{\mu})$ around μ' up to the first order gives $I(\bar{\mu}) = I(\mu') + \nabla I(\tilde{\mu})^T (\bar{\mu} - \mu')$, for some $\tilde{\mu}$ between $\bar{\mu}$ and μ'. From that we get:

$$\| I(\bar{\mu}) - I(\mu') \| \leq \| \nabla I(\tilde{\mu}) \| \| \bar{\mu} - \mu' \| \leq C \| \bar{\mu} - \mu' \|, \tag{10}$$

where $C = \sup_{\mu \in \Xi_1} \|\nabla I(\mu)\|$ is finite due to compactness of Ξ_1 and continuity of all derivatives of the information matrix. It follows from the definition of $\bar{\mu}$ that $\|\bar{\mu} - \mu'\| \leq \|\mu - \mu'\|$; using this in (10) and plugging the result into (9) finishes the proof. ■

Proposition 3 *Let the data sequence $x_1, x_2, \ldots$ be such that $\|x_n\| \leq B$ for all n. Then, for all large n,*

$$\log P_{\hat{\mu}_n}(x^n) - \log P_{\hat{\mu}_{n-1}}(x^n) \leq \frac{1}{2n}(\hat{\mu}_{n-1} - x_n)^T I(\hat{\mu}_{n-1})(\hat{\mu}_{n-1} - x_n) + \frac{C}{n^2}, \tag{11}$$

where C depends on both Ξ_0 and B.

Proof: Proposition 1 states that the left hand side of (11) is upper bounded by $nD(\bar{y}_n \| \hat{\mu}_{n-1}) - nD(\bar{y}_n \| \tilde{\mu}_n)$. Let $\Xi_1 \subset \Xi$ be a compact set such that $\Xi_0 \subset \Xi_1$ and:

$$\inf_{\mu \in \Xi \setminus \Xi_1, \mu' \in \Xi_0} \|\mu - \mu'\| \geq \delta$$

for some $\delta > 0$. In other words, Ξ_0 and the outside of Ξ_1 never come arbitrarily close to each other. Such a set always exists, because Ξ is open, while Ξ_0 is compact. Compactness of Ξ_0 also imply that $\|\hat{\mu}_n\| \leq C_{\Xi_0}$ for some $C_{\Xi_0} < \infty$. Due to boundedness of x_n we have:

$$\|\bar{y}_n - \hat{\mu}_{n-1}\| = \left\| \frac{x_n - \hat{\mu}_{n-1}}{n} \right\| \leq \frac{B + C_{\Xi_0}}{n} \leq \delta,$$

for all sufficiently large n, which implies that $\bar{y}_n \in \Xi_1$. Using first Proposition 1, and then Proposition 2 with $\mu = \bar{y}_n$ and $\mu' = \hat{\mu}_{n-1}$ for compact set Ξ_1, we get:

$$\begin{aligned}
\log P_{\hat{\mu}_n}(x^n) - \log P_{\hat{\mu}_{n-1}}(x^n) &\leq nD(\bar{y}_n \| \hat{\mu}_{n-1}) - nD(\bar{y}_n \| \tilde{\mu}_n) \leq nD(\bar{y}_n \| \hat{\mu}_{n-1}) \\
&\leq \frac{n}{2}(\bar{y}_n - \hat{\mu}_{n-1})^T I(\hat{\mu}_{n-1})(\bar{y}_n - \hat{\mu}_{n-1}) + nC\|\bar{y}_n - \hat{\mu}_{n-1}\|^3 \\
&= \frac{1}{2n}(x_n - \hat{\mu}_{n-1})^T I(\hat{\mu}_{n-1})(x_n - \hat{\mu}_{n-1}) + \frac{C}{n^2}\|x_n - \hat{\mu}_{n-1}\|^3 \\
&\leq \frac{1}{2n}(x_n - \hat{\mu}_{n-1})^T I(\hat{\mu}_{n-1})(x_n - \hat{\mu}_{n-1}) + \frac{C(B + C_{\Xi_0})^3}{n^2}.
\end{aligned}$$

■

With the propositions stated above, we are able to prove the main theorem of this section:

Theorem 1 *Let the data sequence $x_1, x_2, \ldots$ be such that $\|x_n\| \leq B$. Then,*

$$\mathcal{R}(P_{\mathrm{ML}}; x^n) \leq \frac{I_{\Xi_0}(B + C_{\Xi_0})^2}{2} \log n + O(1),$$

where $C_{\Xi_0} = \max_{\mu \in \Xi_0} \|\mu\|$ and $I_{\Xi_0} = \max_{\mu \in \Xi_0} \|I(\mu)\|$.

Proof: Proposition 3 states that there exists n_0 such that for all $n \geq n_0$,

$$\log P_{\hat{\mu}_n}(x^n) - \log P_{\hat{\mu}_{n-1}}(x^n) \leq \frac{1}{2n}(\hat{\mu}_{n-1} - x_n)^T I(\hat{\mu}_{n-1})(\hat{\mu}_{n-1} - x_n) + \frac{C}{n^2} \leq \frac{I_{\Xi_0}(B + C_{\Xi_0})^2}{2n} + \frac{C}{n^2}$$

For $n < n_0$,

$$\log P_{\hat{\mu}_n}(x^n) - \log P_{\hat{\mu}_{n-1}}(x^n) \leq -\log P_{\hat{\mu}_{n-1}}(x^n) = \bar{x}_n \hat{\mu}_{n-1} - \psi(\hat{\theta}_{n-1}) \leq C < \infty,$$

due to compactness of Ξ_0 and boundedness of $\bar{x}_n$. Using those bounds, we get:

$$\begin{aligned}
\mathcal{R}(P_{\mathrm{ML}}; x^n) &= \sum_{i=1}^{n} - \log P_{\hat{\theta}_{i-1}}(x_i) + \log P_{\hat{\theta}_n}(x^n) = \sum_{i=1}^{n} - \log P_{\hat{\theta}_{i-1}}(x_i) + \log P_{\hat{\theta}_i}(x^i) - \log P_{\hat{\theta}_{i-1}}(x^{i-1}) \\
&= \sum_{i=1}^{n} - \log P_{\hat{\theta}_{i-1}}(x^i) + \log P_{\hat{\theta}_i}(x^i) \leq O(1) + \sum_{i=n_0}^{n} - \log P_{\hat{\theta}_{i-1}}(x^i) + \log P_{\hat{\theta}_i}(x^i) \\
&\leq O(1) + \frac{I_{\Xi_0}(B + C_{\Xi_0})^2}{2} \sum_{i=n_0}^{n} \frac{1}{i} + C \sum_{i=n_0}^{n} \frac{1}{n^2} = \frac{I_{\Xi_0}(B + C_{\Xi_0})^2}{2} \log n + O(1).
\end{aligned}$$

■

Theorem 1 states that when playing against distributions from a compact subset of the parameter space, the ML strategy achieves the logarithmic regret growth. However, the constant factor in front of the logarithm can be very large, especially when the bound on the data sequence B is large.

An important question is whether the bound in Theorem 1 is improvable or whether any of the assumptions of the theorem can be relaxed? The answer appears to be negative. First, without restricting the reference strategies to the compact subset Ξ_0, one cannot account for a logarithmic regret at all. Dasgupta and Hsu (2007) consider predicting with Gaussian distributions with unknown mean and covariance without restricting the parameter space, and show that the ML strategy will incur linear regret growth. Second, the following theorem shows that one cannot avoid the boundedness assumption and the bound in Theorem 1 is essentially unimprovable:

Theorem 2 *Let $k = 1$, i.e. the exponential family is one-dimensional. Let P be a plug-in prediction strategy, i.e. $P(x_n|x^{n-1}) = P_{\bar{\mu}_{n-1}}(x_n)$, for some $\bar{\mu}_{n-1} = \bar{\mu}_{n-1}(x^{n-1}) \in \Xi$, $n = 1, 2, \ldots$. Then, for Lebesgue almost all $\mu \in \Xi$ (all except Lebesgue measure zero set), there exists a data sequence $x_1, x_2, \ldots$, such that $\|x_n - \mu\| \leq B$, for which:*

$$\mathcal{R}(P; x^n) \geq \frac{I(\mu)B^2}{2} \log n + O(1).$$

Proof: We will use a theorem from Grünwald and Kotłowski (2010), which states for any plug-in strategy P and one dimensional exponential family $\mathcal{M}$, when the outcomes are generated i.i.d. by some distribution P^*, with $\mathbb{E}_{P^*}[x] = \mu^* \in \Xi$,

$$E_{P^*}[\mathcal{R}(P; x^n)] \geq \frac{1}{2} \frac{\mathrm{var}_{P^*}(x)}{\mathrm{var}_{P_{\mu^*}}(x)} \log n + O(1), \tag{12}$$

for Lebesgue almost all $\mu^* \in \Xi$, where var denotes the variance.

We will apply the theorem with P^* which distributes its mass equally on the two outcomes $x = \mu + B$ and $x = \mu - B$. Then, $\mathbb{E}_{P^*}[x] = \mu^* = \mu$ and $\mathrm{var}_{P^*}(x) = B^2$. Moreover, $\mathrm{var}_{P_{\mu^*}}(x) = I^{-1}(\mu)$. Since the bound (12) holds in expectation, it must also hold for some particular data sequence. ■

4 Regret Bounds for Sequential Normalized Maximum Likelihood Strategy

In the previous section, it was shown that the ML strategy behaves suboptimally and cannot achieve logarithmic regret without bounding the data sequence. In this section, we present a modification of the ML strategy, that achieves logarithmic regret without any boundedness assumptions on the data, and also achieves the optimal constant in front of the logarithm.

The modification is based on calculating the ML estimator of the data sequence *including the current observation.* In other words, the strategy predicts x_n with a distribution proportional to $P_{\hat{\mu}_n}(x_n)$: $P(x_n|x^{n-1}) \propto P_{\hat{\mu}_n}(x_n)$. The proportionality constant differs from unity since $P_{\hat{\mu}_n}(x_n)$ does not normalize properly anymore ($\hat{\mu}_n$ depends on x_n). If $\hat{\mu}_n^x$ denotes the ML estimator for the data sequence $x_1, \ldots, x_{n-1}, x$ (i.e. $\hat{\mu}_n^{x_n} = \hat{\mu}_n$), we can write the strategy as follows:

$$P(x_n|x^{n-1}) = \frac{P_{\hat{\mu}_n}(x^n)}{\int_{\mathcal{X}} P_{\hat{\mu}_n^x}(x^{n-1}, x) dx}. \tag{13}$$

Typically, the strategy will *not* predict with one of the distributions from the model. The strategy (13) is known as *sequential normalized maximum likelihood* (SNML) (Rissanen and Roos, 2007, Roos and Rissanen, 2008) and will be denoted P_{SNML}. It was arrived at from a different starting point: by noticing that (13) is the prediction of the NML distribution in the n-th iteration, with time horizon n. In other words, SNML predicts as NML, assuming that the current iteration is the last iteration. Therefore, contrary to NML, the time horizon of the game does not need to be known. As noted by Rissanen and Roos (2007), the prediction of the SNML strategy can be defined as the solution to the following minimax problem:

$$P_{\text{SNML}}(\cdot|x_{n-1}) = \operatorname*{arg\,min}_{P(\cdot|x_{n-1})} \max_{x_n} \mathcal{R}(P; x^n) = \operatorname*{arg\,min}_{P(\cdot|x_{n-1})} \max_{x_n} \Big\{ -\log P(x_n|x^{n-1}) + \log P_{\hat{\mu}_n}(x^n) \Big\}. \tag{14}$$

A similar idea, though restricted to strategies within the model, was introduced by Takimoto and Warmuth (2000) under the name *last-step minimax.* It is defined as (14), except that the argmin is only over the distributions from the model $\mathcal{P}$. However, the strategy obtained in such a way is a plug-in strategy, and plug-in strategies were already ruled out from having optimal regret bound by Theorem 2.

Surprisingly, the behavior of the worst-case regret for (13) for general exponential families has not been studied before. At the same time, this issue seems to be important, not only from a practical point of view (as SNML constitutes an effective algorithm for online density estimation and model selection), but also from a conceptual point of view. It raises the following question: how much do we loose if we base our decision in a given moment by looking only one step ahead instead of looking at the whole possible future up to a given time horizon. In this section, we show that we do not loose much, at least asymptotically: the regret of the SNML strategy is at most $\frac{k}{2}\log n + O(1)$ for all exponential families, which matches, up to the $O(1)$ term, the minimax regret bound (in Section 5 we will see that for some exponential families even the difference in the $O(1)$ terms is negligible).

We start by rewriting the regret of SNML strategy using (in the second line) telescoping:

$$\begin{aligned}\mathcal{R}(P;x^n) &= -\log P(x^n) + \log P_{\hat{\mu}_n}(x^n) \\ &= \sum_{i=1}^{n} -\log P(x_i|x^{i-1}) + \log P_{\hat{\mu}_i}(x^i) - \log P_{\hat{\mu}_{i-1}}(x^{i-1}) \\ &= \sum_{i=1}^{n} \log \int_{\mathcal{X}} P_{\hat{\mu}_i^x}(x^{i-1},x)dx - \log P_{\hat{\mu}_{i-1}}(x^{i-1}) \\ &= \sum_{i=1}^{n} \log \int_{\mathcal{X}} \exp\Big\{ \log P_{\hat{\mu}_i^x}(x^{i-1},x) - \log P_{\hat{\mu}_{i-1}}(x^{i-1}) \Big\} dx. \qquad (15)\end{aligned}$$

We will bound each term of the sum separately. To this end, we need the following lemma, which states that the integral is negligibly small outside a ball around $\hat{\mu}_{n-1}$, slowly growing with n.

Lemma 3 *Fix an arbitrary $\alpha > 0$ and let $\mathcal{B}_n(\alpha) = \{x \in \mathcal{X} \colon \|x - \hat{\mu}_{n-1}\| \leq n^{\alpha}\}$. Then, for every $\gamma > 0$, there exists a constant $C_\gamma > 0$ such that for all large n,*

$$\int_{\mathcal{X}\setminus\mathcal{B}_n(\alpha)} \exp\Big\{ \log P_{\hat{\mu}_n^x}(x^{n-1},x) - \log P_{\hat{\mu}_{n-1}}(x^{n-1}) \Big\} dx \leq C_\gamma n^{-\gamma}. \qquad (16)$$

Proof: Let us denote the left hand side of (16) as A_n. Rewriting the integral, we get:

$$A_n = \int_{\mathcal{X}\setminus\mathcal{B}_n(\alpha)} P_{\hat{\mu}_n^x}(x) \exp\Big\{ \log P_{\hat{\mu}_n^x}(x^{n-1}) - \log P_{\hat{\mu}_{n-1}}(x^{n-1}) \Big\} dx \leq \int_{\mathcal{X}\setminus\mathcal{B}_n(\alpha)} P_{\hat{\mu}_n^x}(x)dx,$$

because from the definition of $\hat{\mu}_{n-1}$, $P_{\hat{\mu}_{n-1}}(x^{n-1}) \geq P_{\hat{\mu}_n^x}(x^{n-1})$.

We will use the fact that exponential families have all central moments finite (Barndorff-Nielsen, 1978). This means, that $\int_{\mathcal{X}} \|x - \mu\|^{\beta} P_\mu(x)dx < \infty$, for all $\beta = 1, 2, \ldots$, so if $x \to \infty$, $P_\mu(x)$ must converge to 0 faster than any monomial $\|x - \mu\|^{-\beta}$ for the integral to be finite. This implies that for any $\beta > 0$, $P_\mu(x) \leq C_{\mu,\beta}\|x - \mu\|^{-\beta}$ for some $C_{\mu,\beta} < \infty$. Moreover, $C_\beta := \sup_{\mu\in\Xi_0} C_{\mu,\beta}$ is finite. Otherwise, we could form a monotonic sequence μ_i with $C_{\mu_i,\beta} > C_{\mu_{i-1},\beta} + 1$, $i = 1, 2, \ldots$; due to compactness of Ξ_0, this sequence has a subsequence converging to $\mu^* \in \Xi_0$ and due to monotonicity, $C_{\mu^*,\beta}$ cannot be finite, which is a contradiction. Therefore, we can write:

$$A_n \leq \int_{\mathcal{X}\setminus\mathcal{B}_n(\alpha)} P_{\hat{\mu}_n^x}(x)dx \leq C_\beta \int_{\mathcal{X}\setminus\mathcal{B}_n(\alpha)} \|x - \hat{\mu}_n^x\|^{-\beta}.$$

From the triangle inequality, $\|x - \hat{\mu}_n^x\| \geq \|x - \hat{\mu}_{n-1}\| - \|\hat{\mu}_n^x - \hat{\mu}_{n-1}\|$. Since the former is at least n^α for $x \notin \mathcal{B}(\alpha)$, while the latter is bounded, for n large enough, $\|x - \hat{\mu}_n^x\| \geq \frac{1}{2}\|x - \hat{\mu}_{n-1}\|$ and therefore:

$$A_n \leq C_\beta 2^\beta \int_{\|x\|\geq n^\alpha} \|x\|^{-\beta} dx \leq C' n^{\alpha(k-\beta)}.$$

Setting $C_\gamma = C'$ and $\gamma = \alpha(\beta - k)$ finishes the proof. ■

We are now ready to prove the main theorem:

Theorem 4 *Assume the setting of Section 2.1; in particular, let the ML distribution be defined as in (2) where Θ_0 (and hence Ξ_0) is compact. Let P be the SNML strategy. Then,*

$$\mathcal{R}(P;x^n) \leq \frac{k}{2}\log n + O(1),$$

where the constant in $O(1)$ depends on Ξ_0, but does not depend on the data sequence x^n.

Proof: Using the simple fact that $\log(a+b) \leq \max\{0, \log(a)\} + b$ for $a, b \geq 0$, and applying Lemma 3, we can bound:

$$\log \int_{\mathcal{X}} e^{\log P_{\hat{\mu}_n^x}(x^{n-1},x) - \log P_{\hat{\mu}_{n-1}}(x^{n-1})} dx \leq \max\Big\{0, \log \int_{\mathcal{B}_n(\alpha)} e^{\log P_{\hat{\mu}_n^x}(x^{n-1},x) - \log P_{\hat{\mu}_{n-1}}(x^{n-1})} dx \Big\} + C_\gamma n^{-\gamma}. \tag{17}$$

Let us denote the integral over $\mathcal{B}_n(\alpha)$ on the right hand side of (17) as S_n. We have:

$$\begin{aligned}
S_n &= \log \int_{\mathcal{B}_n(\alpha)} \exp\Big\{ \log P_{\hat{\mu}_n^x}(x^{n-1}, x) - \log P_{\hat{\mu}_{n-1}}(x^{n-1}, x) + \log P_{\hat{\mu}_{n-1}}(x) \Big\} dx \\
&= \log \int_{\mathcal{B}_n(\alpha)} P_{\hat{\mu}_{n-1}}(x) \exp\Big\{ \log P_{\hat{\mu}_n^x}(x^{n-1}, x) - \log P_{\hat{\mu}_{n-1}}(x^{n-1}, x) \Big\} dx \\
&\leq \log \int_{\mathcal{B}_n(\alpha)} P_{\hat{\mu}_{n-1}}(x) \exp\Big\{ nD(\bar{y}_n^x \| \hat{\mu}_{n-1}) - nD(\bar{y}_n^x \| \tilde{\mu}_n^x) \Big\} dx \qquad \text{(Proposition 1)} \\
&\leq \log \int_{\mathcal{B}_n(\alpha)} P_{\hat{\mu}_{n-1}}(x) \exp\Big\{ nD(\bar{y}_n^x \| \hat{\mu}_{n-1}) \Big\} dx,
\end{aligned}$$

where $\bar{y}_n^x = \frac{(n-1)\hat{\mu}_{n-1} + x}{n}$ and $\tilde{\mu}_n^x = \arg\min_{\mu \in \Xi_0} D(\bar{y}_n^x \| \mu)$.

Let $\Xi_1 \subset \Xi$ be a compact set such that $\Xi_0 \subset \Xi_1$ and: $\inf_{\mu \in \Xi \setminus \Xi_1, \mu' \in \Xi_0} \|\mu - \mu'\| \geq \delta$ for some $\delta > 0$, as in the proof of Proposition 3. Let us choose $\alpha < 1/4$ in the definition of $\mathcal{B}_n(\alpha)$. Then, for n large enough,

$$\|\bar{y}_n^x - \hat{\mu}_{n-1}\| = \frac{\|x - \hat{\mu}_{n-1}\|}{n} \leq n^{\alpha - 1} < \delta$$

for sufficiently large n, which means that $\bar{y}_n^x \in \Xi_1$. Using Proposition 2 with $\mu = \bar{y}_n^x$ and $\mu' = \hat{\mu}_{n-1}$ for compact set Ξ_1 and $x \in \mathcal{B}_n(\alpha)$, we get:

$$\begin{aligned}
nD(\bar{y}_n^x \| \hat{\mu}_{n-1}) &\leq \frac{n}{2}(\bar{y}_n^x - \hat{\mu}_{n-1})^T I(\hat{\mu}_{n-1})(\bar{y}_n^x - \hat{\mu}_{n-1}) + nC\|\bar{y}_n^x - \hat{\mu}_{n-1}\|^3 \\
&= \frac{1}{2n}(x - \hat{\mu}_{n-1})^T I(\hat{\mu}_{n-1})(x - \hat{\mu}_{n-1}) + \frac{C}{n^2}\|x - \hat{\mu}_{n-1}\|^3 \\
&\leq \frac{1}{2n}(x - \hat{\mu}_{n-1})^T I(\hat{\mu}_{n-1})(x - \hat{\mu}_{n-1}) + C' n^{3\alpha - 2}
\end{aligned}$$

Let $B(x)$ be a function, equal to the right hand side of the above inequality, if $x \in \mathcal{B}_n(\alpha)$, and $B(x) = 0$ for $x \notin \mathcal{B}_n(\alpha)$. Since $I(\mu)$ is continuous on Ξ, and Ξ_0 is compact, $\sup_{\mu \in \Xi_0} \|I(\mu)\| = I_{\Xi_0} < \infty$, and thus $B(x)$ is bounded by:

$$B(x) \leq \frac{1}{2n} I_{\Xi_0} n^{2\alpha} + C' n^{3\alpha - 2} = C'' n^{2\alpha - 1},$$

(because $\alpha < 1/4$). Note that:

$$\log \int_{\mathcal{B}_n(\alpha)} P_{\hat{\mu}_{n-1}}(x) e^{B(x)} dx \leq \log \int_{\mathcal{X}} P_{\hat{\mu}_{n-1}}(x) e^{B(x)} dx = \log \mathbb{E}\left[e^{B(x)}\right].$$

We can therefore use Hoeffding's lemma, $\log \mathbb{E}\left[e^{B(x)}\right] \leq \mathbb{E}\left[B(x)\right] + \frac{(C'')^2}{8} n^{2(2\alpha - 1)}$ (Cesa-Bianchi and Lugosi, 2006), and bound $\mathbb{E}\left[B(x)\right]$:

$$\begin{aligned}
\mathbb{E}\left[B(x)\right] &\leq \frac{1}{2n}\mathbb{E}\left[(x - \hat{\mu}_{n-1})^T I(\hat{\mu}_{n-1})(x - \hat{\mu}_{n-1})\right] + C' n^{3\alpha - 2} \\
&= \frac{1}{2n}\mathbb{E}\left[\mathrm{Tr}\Big\{(x - \hat{\mu}_{n-1})(x - \hat{\mu}_{n-1})^T I(\hat{\mu}_{n-1})\Big\}\right] + C' n^{3\alpha - 2} \\
&= \frac{1}{2n}\mathrm{Tr}\left\{\mathrm{Cov}_{\hat{\mu}_{n-1}}(x) I(\hat{\mu}_{n-1})\right\} + C' n^{3\alpha - 2} = \frac{k}{2n} + C' n^{3\alpha - 2}.
\end{aligned}$$

Thus, we finally get:

$$\log \int_{\mathcal{X}} e^{\log P_{\hat{\mu}_n^x}(x^{n-1},x) - \log P_{\hat{\mu}_{n-1}}(x^{n-1})} dx \leq \frac{k}{2n} + C' n^{3\alpha - 2} + \frac{(C'')^2}{8} n^{2(2\alpha - 1)} + C_\gamma n^{-\gamma} = \frac{k}{2n} + O(n^{-2}),$$

for sufficiently large γ, $\alpha < 1/4$, for all large n. Since:

$$\log \int_{\mathcal{X}} P_{\hat{\mu}_i^x}(x) \exp\Big\{ \log P_{\hat{\mu}_i^x}(x^{i-1}) - \log P_{\hat{\mu}_{i-1}}(x^{i-1}) \Big\} dx \leq \log \int_{\mathcal{X}} P_{\hat{\mu}_i^x}(x) dx \leq \log \int_{\mathcal{X}} P_{\hat{\mu}_1^x}(x) dx,$$

which is the minimax (NML) regret for $n = 1$ and thus finite (Section 2.2) , the terms in (15) are finite and well-controlled for small n. Thus we conclude that $\mathcal{R}(P; x^n) \leq \frac{k}{2}\log n + O(1)$. ■

5 Examples

In this section, we calculate and analyze SNML for few examples of commonly used exponential families.

5.1 Bernoulli Distribution

The SNML for Bernoulli was first considered by Takimoto and Warmuth (2000) as the *last-step minimax* algorithm. The simplest derivation is in the mean value parametrization, in which the Bernoulli distribution looks like:

$$P_\mu(x) = \mu^x (1-\mu)^{1-x},$$

where $x \in \{0,1\}$ and $\Xi = (0,1)$; note that we need to exclude from Ξ two extreme points 0 and 1 to be consistent with the assumptions made in this paper. We set $\Xi_0 = [\epsilon, 1-\epsilon]$. The ML estimator $\hat{\mu}_n$ to relative Ξ_0 equals:

$$\hat{\mu}_n = \min\{1-\epsilon, \max\{\epsilon, \bar{x}_n\}\},$$

i.e. $\hat{\mu}_n$ is $\bar{x}_n$ truncated to the range $[\epsilon, 1-\epsilon]$. Then, the SNML strategy can easily be computed from:

$$P_{\text{SNML}}(x_n = 1 | x^{n-1}) = \frac{(\hat{\mu}_n^1)^{k+1}(1-\hat{\mu}_n^1)^{n-k}}{(\hat{\mu}_n^1)^{k+1}(1-\hat{\mu}_n^1)^{n-k} + (\hat{\mu}_n^0)^{k}(1-\hat{\mu}_n^0)^{n-k+1}},$$

where $k = (n-1)\bar{x}_{n-1}$ is the number of ones in the past, and $\hat{\mu}_n^x$ for $x \in \{0,1\}$ is defined as usual. One can also show that even for the case $\Xi = [0,1]$ and $\epsilon = 0$, although not covered by our theorem, the regret of the strategy is bounded by $\frac{1}{2}\log(n+1) + \frac{1}{2}$ (Takimoto and Warmuth, 2000), and is superior even to the celebrated Krichevsky-Trofimov estimator (Cesa-Bianchi and Lugosi, 2006).

5.2 Exponential Distribution

The distribution has the form:

$$P_\theta(x) = \frac{1}{\mu} e^{-x/\mu},$$

with $\Xi = (0, \infty)$. The strategy becomes particularly simple if we take $\Xi_0 = \Xi$ (although this case is not covered by Theorem 4). Like in the Bernoulli case, the ML estimator is equal to the truncation of $\bar{x}_n$ into the range Ξ_0, so that for $\Xi_0 = (0,\infty)$, $\hat{\mu}_n = \bar{x}_n$. Thus:

$$P_{\text{SNML}}(x_n | x^{n-1}) \propto \sup_{\mu \in \Xi_0} P_\mu(x^n) = \bar{x}_n^{-n} e^{-\frac{\sum_{i=1}^n x_n}{\bar{x}_n}} = \frac{e^{-n} n^n}{((n-1)\bar{x}_{n-1} + x_n)^n},$$

which, after proper normalization, becomes:

$$P_{\text{SNML}}(x_n | x^{n-1}) = \frac{(n-1)^n \bar{x}_{n-1}^{n-1}}{((n-1)\bar{x}_{n-1} + x_n)^n}.$$

One can directly show that the per-round regret increase $\mathcal{R}(P_{\text{SNML}}; x^n) - \mathcal{R}(P_{\text{SNML}}; x^{n-1})$ equals $n \log \frac{n}{n-1} - 1$, which is upper bounded by $\frac{1}{2(n-1)}$ except the first iteration. Indeed, choosing $\Xi_0 = \Xi$ implies infinite regret in the first trial. That being said, in the rest of the game such a choice of Ξ_0 does not seem to be harmful anymore.

5.3 Gaussian Distribution, Fixed Variance

The k-dimensional Gaussian distribution $N(x|\mu, \Sigma)$ reads:

$$P_\mu(x) = (2\pi)^{-k/2} |\Sigma|^{-1/2} \exp\left\{ -\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu) \right\},$$

and without loss of generality we can assume $\Sigma = I$, the identity matrix (we can always rotate the coordinate system so that it matches with the principal axes of Σ, and rescale it by the inverses of the eigenvalues). We set $\Xi = \mathbb{R}^k$ and we will use $\Xi_0 = \{\mu \colon \|\mu\| \leq R\}$. The ML estimator $\hat{\mu}_n$ relative to Ξ_0 equals:

$$\hat{\mu}_n = \begin{cases} \bar{x}_n & \text{if } \|\bar{x}_n\| \leq R, \\ \frac{R}{\|\bar{x}_n\|}\bar{x}_n & \text{if } \|\bar{x}_n\| > R. \end{cases}$$

The derivation of the SNML strategy simplifies a lot if we choose $R \to \infty$. Then, $\hat{\mu}_n = \bar{x}_n$ and:

$$P_{\text{SNML}}(x_n | x^{n-1}) \propto e^{-\frac{1}{2}\sum_{i=1}^n \|x_i - \bar{x}_n\|^2}.$$

A bit of algebra reveals that:

$$\sum_{i=1}^{n} \|x_i - \bar{x}_n\|^2 = \sum_{i=1}^{n-1} \|x_i - \bar{x}_{n-1}\|^2 + \frac{n-1}{n}\|x_n - \bar{x}_{n-1}\|^2,$$

so that $P_{\text{SNML}}(x_n|x^{n-1}) \propto e^{-\frac{1}{2}\frac{n-1}{n}\|x_n - \bar{x}_{n-1}\|^2}$, which means that SNML predicts with $N(x|\bar{x}_{n-1}, \frac{n}{n-1}I)$. Note, that although SNML predict with a Gaussian distribution, it is *not* a plug-in type strategy, as the predictive distribution is *outside* the model (the variance is not equal to I). Although Theorem 4 does not apply for $R = \infty$, one can directly show that the per-round regret increase equals $\frac{k}{2}\log\frac{n}{n-1}$, which gives the regret $\frac{k}{2}\log n$ if we do *not* count the regret in the first iteration (which is, again, infinite).

5.4 Gaussian Distribution, Unknown Mean and Variance

As a final example, consider the family of one-dimensional Gaussian distributions with unknown mean μ and variance σ^2. Writing down the distribution in the natural parameterization:

$$P_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2} + \log\sigma \right\},$$

shows that the exponential family is two-dimensional with sufficient statistic (x, x^2). Similarly as in the previous cases, setting $\Xi_0 = \Xi$ significantly simplifies all the derivations. The ML estimator becomes $\hat{\mu}_n = \bar{x}_n$ and $\hat{\sigma}_n^2 = n^{-1}\sum(x_i - \bar{x}_n)^2$. The SNML strategy predicts with:

$$P_{\text{SNML}}(x_n|x^{n-1}) \propto (2\pi\hat{\sigma}_n^2)^{-n/2} \exp\left\{ -\frac{\sum_{i=1}^n (x_i - \bar{x}_n)^2}{2\hat{\sigma}_n^2} \right\} = \frac{e^{-\frac{n}{2}} n^{n/2}}{(2\pi(n-1))^{n/2}} \frac{1}{(\hat{\sigma}_{n-1}^2 + \frac{1}{n}(x_n - \bar{x}_{n-1})^2)^{n/2}},$$

which after normalization gives:

$$P_{\text{SNML}}(x_n|x^{n-1}) = \frac{\Gamma(n/2)}{\Gamma(1/2)\Gamma((n-1)/2)} (n\hat{\sigma}_{n-1}^2)^{-1/2} \left(1 + \frac{(x_n - \bar{x}_{n-1})^2}{n\hat{\sigma}_{n-1}^2} \right)^{-n/2}.$$

One can directly show that the per-round regret increase

$$\mathcal{R}(P_{\text{SNML}}; x^n) - \mathcal{R}(P_{\text{SNML}}; x^{n-1}) = \frac{n+1}{2}\log n - \frac{n}{2}\log(n-1) - \frac{1}{2}\log 2e + \log\frac{\Gamma((n-1)/2)}{\Gamma(n/2)},$$

and grows as $\frac{1}{n} + O(n^{-2})$ (except the first iteration), which is optimal for $k = 2$.

5.5 Open Problem

In the paper, the SNML strategy always plays against the exponential family $\mathcal{M}$ with parameter space restricted to a compact subset Ξ_0. As shown in Section 2, this is necessary, otherwise the strategy would suffer infinite regret already in the first iteration. However, as we saw in all the above examples, starting from the second iteration, choosing $\Xi_0 = \Xi$ does not hurt anymore, so that the only place in which restricting the power of the reference strategies to Ξ_0 matters, is the beginning of the game. A similar phenomenon was already found in the MDL community (Grünwald, 2007), in which it was noticed that some models with infinite complexity (i.e. with infinite worst-case regret) can be made finitely complex when conditioned on the first outcome(s).

Therefore we pose the following problem. Assume that the loss suffered in the first $n_0 \geq 1$ iterations is not included in the cumulative regret of the predicting strategy. Is it then possible to achieve the worst-case regret $\frac{k}{2}\log n + O(1)$ without restricting the parameter space of the exponential family?

6 SNML and Bayes with Jeffreys' Prior

Given a prior distribution $\pi(\mu)$ over Ξ, the *Bayes* prediction strategy P_{BAYES} is defined as:

$$P_{\text{BAYES}}(x^n) = \int_{\Xi} P_\mu(x^n)\pi(\mu)d\mu, \tag{18}$$

a π-mixture of distributions from Ξ. Let Ξ_0 be a compact subset of Ξ. It is known (Grünwald, 2007) that under the assumption, that the sequence $x_1, x_2, \ldots$ satisfies $\bar{x}_n \in \Xi_0$ for all large n, the Bayes strategy achieves asymptotically optimal regret $\mathcal{R}(P_{\text{BAYES}}; x^n) = \frac{k}{2} \log n + O(1)$. Moreover, using the *Jeffreys' prior* $\pi(\mu) \propto \sqrt{\det I(\mu)}$ minimizes the $O(1)$ term in the worst case, up to an $o(1)$ term, so that the Jeffreys' prior is in some sense a minimax prior, achieving asymptotically the same regret as NML.

Let us consider instead the case when $\Xi_0 = \Xi$. Although the joint distribution for Bayes with Jeffreys' prior $P_{\text{JEF}}(x^n)$ is often undefined in this case (the prior cannot be normalized), the conditionals $P_{\text{JEF}}(x_n|x^{n-1})$ for all $n \geq m$, for some m, can still be properly defined (Grünwald, 2007). This is also the case of SNML (examples in Section 5 show that SNML is properly defined for $n \geq 2$). Moreover, the same story will apply to NML (minimax) strategy if, instead of using definition (6), we will define NML through the conditionals:

$$P_{\text{NML}}(x_i|x^{i-1}) = \frac{P_{\text{NML}}(x^i)}{P_{\text{NML}}(x^{i-1})} = \frac{\int_{\mathcal{X}^{n-i}} \sup_{\mu \in \Xi} P_\mu(x^i, z^{n-i}) dz^{n-i}}{\int_{\mathcal{X}^{n-i+1}} \sup_{\mu \in \Xi} P_\mu(x^{i-1}, z^{n-i+1}) dz^{n-i+1}},$$

where n is the time horizon for NML. Note that the definition coincides with the concept of *conditional NML-2* in (Grünwald, 2007).

Interestingly, in all but one of the examples from Section 5, the SNML strategy and Jeffreys' strategy coincide. Only in the case of Bernoulli, the strategies differ, with an advantage for SNML, whose worst-case regret is smaller by a constant than the one achieved by Jeffreys' strategy[1]. The two open questions we pose are: (1) Under what conditions are SNML and Bayes with Jeffreys' prior the same? (2) If SNML and Jeffreys' differ, is there any general relationship between the worst-case regrets of the two?

To shed some light on the questions above, we prove the following fact, which shows that the equivalence of SNML and Jeffreys' implies that both strategies are equal to the NML strategy. This is what actually happens in three out of four examples shown in Section 5.

Theorem 5 *Let $\mathcal{P}$ be an exponential family, such that, starting from some m, the conditional distributions for SNML and Jeffreys' strategies are properly defined and coincide, $P_{\text{SNML}}(x_n|x^{n-1}) = P_{\text{JEFF}}(x_n|x^{n-1})$ for all x^n, $n \geq m$. Then, both strategies are equal to the minimax (NML) strategy $P_{\text{NML}}(x_n|x^{n-1})$, for $n \geq m$, and the NML strategy does not depend on the time horizon.*

Proof: Since the strategies might only be defined for $n \geq m$, let us focus on the *conditional regret* defined as $\mathcal{R}(P; x^{m:n}|x^{m-1}) = -\log P(x^{n:m}|x^{m-1}) + \log P_{\hat{\mu}_n}(x^n)$, where $x^{m:n} = x_m, x_{m+1}, \ldots, x_n$. From the definition (13), the conditional regret of the SNML strategy does not depend on the last outcome:

$$\mathcal{R}(P_{\text{SNML}}; x^{m:n}|x^{m-1}) = -\log P_{\text{SNML}}(x^{m:n-1}|x^{m-1}) + \log \int_{\mathcal{X}} P_{\hat{\mu}_n^x}(x^{n-1}, x) dx.$$

Since $P_{\text{JEF}}(x^{m:n}|x^{m-1}) = P_{\text{SNML}}(x^{m:n}|x^{m-1})$, Jeffreys' strategy inherits this property. On the other hand, Jeffreys' is a particular case of the Bayes strategy, which is known to be *exchangeable*, i.e. $P_{\text{BAYES}}(x^{m:n}|x^{m-1})$ does not depend on the order of the outcomes in $x^{m:n}$; this property follows directly from the definition (18). Since the comparator $P_{\hat{\mu}_n}(x^n)$ does not depend on the order either, the same property holds for the conditional regret. But then, if the conditional regret is invariant under changing the last outcome and under changing the order of the outcomes, it is also invariant under changing all the outcomes in $x^{m:n}$. This means that the strategy P_{JEF} gives equal conditional regret for all possible data sequences $x^{m:n}$ (i.e. the strategy is an *equalizer* of conditional regret), which implies $P_{\text{JEF}}(x^{m:n}|x^{m-1})$ must be equal $P_{\text{NML}}(x^{m:n}|x^{m-1})$ for all x^n (Grünwald, 2007). ■

Theorem (5) would also hold if we replace Jeffreys' strategy with a general Bayes strategy. However, due to the known relationship between the Jeffreys' strategy and the minimax regret, we do not expect the conditions of the theorem to be satisfied by Bayes with any other prior than Jeffreys'.

7 Conclusions and Further Work

We analyzed the regret of the ML ("follow the leader") strategy for general exponential families. The lower and upper bounds show that the ML strategy requires boundedness of the data sequence to obtain logarithmic regret, and moreover, the constant in front of the logarithm is suboptimal and can be very large. Those two drawbacks are essentially unavoidable, not only for the ML strategy, but for any plug-in strategy. However, we also showed that both problems are solved by adding the currently predicted outcome to the calculation of the maximum likelihood, followed by

[1] The Jeffreys' strategy for Bernoulli is the Krichevsky-Trofimov strategy.

normalization, which leads to the SNML strategy. We proved that SNML achieves asymptotically optimal regret. We also noted an interesting relationship to the Bayes strategy with Jeffreys' prior.

In future work, we plan to work on the two open questions posed in the paper: (1) Is is possible to relax the condition that the model is constrained to the compact subset of the parameter space by conditioning the regret on the outcome from the first iteration? (2) When is SNML equal to Bayes with Jeffreys' prior and is there any general relationship between the worst-case regrets of the two?

References

K. Azoury and M. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Journal of Machine Learning*, 43(3):211–246, 2001.

O.E. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory.* Wiley, Chichester, UK, 1978.

A. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, 44(6):2743–2760, 1998.

S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.

N. Cesa-Bianchi and G. Lugosi. Worst-case bounds for the logarithmic loss of predictors. *Journal of Machine Learning*, 43(3):247–264, 2001.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning and Games.* Cambridge University Press, 2006.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory.* John Wiley, 1991.

Sanjoy Dasgupta and Daniel Hsu. On-line estimation with the multivariate gaussian distribution. In *Conference on Learning Theory (COLT '07)*, 2007.

A.P. Dawid. Present position and potential developments: Some personal views, statistical theory, the prequential approach. *J. Royal Stat.Soc., Ser. A*, 147(2):278–292, 1984.

Y. Freund. Predicting a binary sequence almost as well as the optimal biased coin. In *Computational Learning Theory (COLT' 96)*, pages 89–98, 1996.

P. Grünwald and W. Kotłowski. Prequential plug-in codes that achieve optimal redundancy rates even if the model is wrong. In *The IEEE International Symposium on Information Theory (ISIT '10)*, 2010.

P. D. Grünwald. *The Minimum Description Length Principle.* MIT Press, Cambridge, MA, 2007.

P. D. Grünwald and S. de Rooij. Asymptotic log-loss of prequential maximum likelihood codes. In *Conference on Learning Theory (COLT 2005)*, pages 652–667, 2005.

E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2–3):169–192, 2007.

W. Kotłowski, P. Grünwald, and S. de Rooij. Following the flattened leader. In *Conference on Learning Theory (COLT '10)*, 2010.

M. Raginsky, R. F. Marcia, S. Jorge, and R. Willett. Sequential probability assignment via online onvex programming using exponential families. In *IEEE International Symposium on Information Theory*, 2009.

J. Rissanen. Fisher information and stochastic complexity. *IEEE Trans. Information Theory*, IT-42 (1):40–47, 1996.

J. Rissanen and T. Roos. Conditional NML universal models. In *Information Theory and Applications Workshop (ITA-07)*, pages 337–341, 2007.

T. Roos and J. Rissanen. On sequentially normalized maximum likelihood models. In *Workshop on Information Theoretic Methods in Science and Engineering (WITMSE-08)*, 2008.

Y. Shtarkov. Universal sequential coding of single messages. *Problems of Information Transmission*, 23(3):175–186, 1987.

E. Takimoto and M. Warmuth. The last-step minimax algorithm. In *Conference on Algorithmic Learning Theory (ALT '00)*, 2000.