Sequential prediction under log-loss and misspecification

Meir Feder MEIR@TAUEX.TAU.AC.IL
School of Electrical Engineering, Tel-Aviv University, Israel

Yury Polyanskiy YP@MIT.EDU
Dept. of EECS, MIT, USA

Abstract

We consider the question of sequential prediction under the log-loss in terms of cumulative regret. Namely, given a hypothesis class of distributions, learner sequentially predicts the (distribution of the) next letter in sequence and its performance is compared to the baseline of the best constant predictor from the hypothesis class. The well-specified case corresponds to an additional assumption that the data-generating distribution belongs to the hypothesis class as well. Here we present results in the more general misspecified case. Due to special properties of the log-loss, the same problem arises in the context of competitive-optimality in density estimation and model selection. For the d-dimensional Gaussian location hypothesis class, we show that cumulative regrets in the well-specified and misspecified cases asymptotically coincide. In other words, we provide an $o(1)$ characterization of the distribution-free (or PAC) regret in this case – the first such result as far as we know. We recall that the worst-case (or individual-sequence) regret in this case is larger by an additive constant $\frac{d}{2} + o(1)$. Surprisingly, neither the traditional Bayesian estimators, nor the Shtarkov's normalized maximum likelihood achieve the PAC regret and our estimator requires special "robustification" against heavy-tailed data. In addition, we show two general results for misspecified regret: the existence and uniqueness of the optimal estimator, and the bound sandwiching the misspecified regret between well-specified regrets with (asymptotically) close hypotheses classes.

Keywords: Online learning, distribution-free PAC learning, log-loss, agnostic learning, sequential probability assignment, misspecified models

1. Introduction

This paper considers problems of the following type:

$$\min_{Q} \max_{P\in\Phi, P^*\in\Theta} \mathbb{E}_{Y\sim P}\left[\log\frac{P^*(Y)}{Q(Y)}\right],$$

where Θ and Φ are some collections of distributions. The goal is to find the (approximate) value of $\min\max$ and the (approximate) minimizer Q^*. There are several ways in which this abstract problem can arise (see Section 1.2). The problem has been studied in information theory, statistics and machine learning predominantly in the following two cases: when $\Phi = \Theta$ (well-specified or "stochastic" case) and when Φ consists of all distributions (worst-case or "individual-sequence"). However, the natural intermediate case of when Φ consists of all iid distributions (a case we designate by the name "PAC") has not been studied as much. We report new results pertaining to the cases of $\Phi \neq \Theta$. Such a setting has been known under the names of model-mismatch, misspecified regret, agnostic learning or distribution-free PAC (in case Φ is all iid distributions). Our paper can be filed under either of these.

Notation. We use $P \ll Q$ for absolute continuity of measures, $\frac{dP}{dQ}$ for Radon-Nikodym derivatives, $D(P\|Q) = \mathbb{E}_P[\log \frac{dP}{dQ}]$ for Kullback-Leibler (KL) divergence, $I(X;Y) = D(P_{X,Y}\|P_X \otimes P_Y)$ for mutual information, $P \otimes Q$ for a product measure, $P^{\otimes n}$ for an n-fold product of P with itself, $y^n = (y_1, \ldots, y_n)$ for an n-vector, $\mathcal{Y}^n$ for a measurable space of n-vectors, $\mathcal{P}(\mathcal{Y})$ for a set of probability measures on a measurable space $\mathcal{Y}$, $\mathcal{P}_{iid}(\mathcal{Y}^n) = \{P^{\otimes n} : P \in \mathcal{Y}(Y)\}$, $[k] = \{1, \ldots, k\}$ for $k \in \mathbb{Z}_+$, Leb denotes Lebesgue measure.

1.1. Defining regret under model misspecification

Fix a measurable space $\mathcal{Y}$ and a collection of hypotheses $P_\theta, \theta \in \Theta$ of measures on it, which we will call model class Θ. Suppose an iid sequence $Y_i \sim P$ is observed and our goal is to provide an estimate of its distribution that is (almost) as good as the best possible hypothesis P_{θ^*}. More specifically, suppose that having observed $y_1, \ldots, y_{t-1}$ we output our estimate distribution $Q_t(\cdot) = Q_t(\cdot|y^{t-1})$, and then upon observing Y_t experience a (relative) regret of $\log \frac{dP_{\theta^*}}{dQ_t}(Y_t)$. Our goal is to minimize

$$\sup_{\theta^*, P} \mathbb{E}\left[\sum_{t=1}^n \log \frac{dP_{\theta^*}}{dQ_t}(Y_t)\right], \tag{1}$$

where supremum over θ^* corresponds to chosing the best in-model match and supremum over P corresponds to the worst-case choice of the data generating distribution. (Non-iid models and/or generating distributions, e.g. Markov processes, can be handled by taking $n = 1$ and extending $\mathcal{Y}$.)

The most studied case of this problem is the *well-specified* case, when in addition we restrict supremum over P to $P = P_\theta$ for some $\theta \in \Theta$. In this case, it is clear that the optimal choice of $\theta^* = \theta$ and we get the well-known definition of the the optimal minimax (cumulative) regret, called the capacity of Θ:

$$C_n(\{P_\theta, \theta \in \Theta\}) = C_n(\Theta) = \inf_{Q_{Y^n}} \sup_{\theta \in \Theta} \mathbb{E}^\theta\left[\log \frac{dP_\theta^{\otimes n}}{dQ_{Y^n}}(Y^n)\right].$$

A simple observation shows that

$$C_n(\Theta) = \inf_{Q_{Y^n}} \sup_{\theta \in \Theta} D(P_\theta^{\otimes n}\|Q_{Y^n}). \tag{2}$$

A fundamental theorem of Kemperman Kemperman (1974) states that whenever $C_n(\Theta) < \infty$ there exists a unique $Q^*_{Y^n}$ such that

$$C_n(\Theta) = \sup_{\theta \in \Theta} D(P_\theta^{\otimes n}\|Q^*_{Y^n}),$$

and, furthermore,

$$C_n(\Theta) = \sup_\pi I(\theta; Y^n), \tag{3}$$

where supremum is over all (finitely supported) priors on Θ. In application to sequential prediction, this result is also known as the capacity-redundancy theorem Gallager (1979); Ryabko (1979); Davisson and Leon-Garcia (1980) and its strong version is given in Merhav and Feder (1995).

Notice that, in particular, whenever $C_n(\Theta) < \infty$ there must exist a measure μ such that $P_\theta \ll \mu$ for all θ (e.g. one can take $\mu = Q^*_1$). Thus, in the sequel we fix an auxiliary measure μ on $\mathcal{Y}$ and assume that

$$P_\theta(dy) = f_\theta(y)\mu(dy),$$

that is the family P_θ is given by its relative densities f_θ.

In this paper we study the misspecified case where the supremum over data-generating distributions does not have to come from the model class Θ (and in fact is not even required to be iid).

Definition 1 *For a given $\mathcal{Y}$, μ, a collection of densities $\{f_\theta, \theta \in \Theta\}$ and a collection of distributions Φ_n on $\mathcal{Y}^n$ we define*

$$F_n(\{f_\theta, \theta \in \Theta\}, \mu, \Phi_n) \triangleq \inf_q \sup_{P \in \Phi_n} \sup_{\theta \in \Theta} \mathbb{E}_{Y^n \sim P}\left[\log \frac{\prod_{t=1}^n f_\theta(Y_t)}{q(Y^n)}\right], \qquad (4)$$

where infimum is over all $q : \mathcal{Y}^n \to \mathbb{R}_+$ with $\int q d\mu^{\otimes n} = 1$. When $n = 1$ and $\Phi \subset \mathcal{P}(\mathcal{Y})$, we shorten $F_1(\{f_\theta, \theta \in \Theta\}, \mu, \Phi)$ to just $F(\Theta, \Phi)$.

There are three subtleties (discussed in detail in Appendix A): zeros under $\log$, non-existence of $\mathbb{E}$ and the fact that quantity F_n *may* depend on a choice of densities f_θ for representing $\{P_\theta\}$.

In the most extreme case, we take $\Phi_n = \mathcal{P}(\mathcal{Y}^n) \triangleq \{\text{all distributions on } \mathcal{Y}^n\}$. The resulting quantity is known as the individual-sequence regret:

$$\Gamma_n(\{f_\theta, \theta \in \Theta\}, \mu) = \inf_q \sup_{y^n \in \mathcal{Y}^n} \sup_{\theta \in \Theta} \log \frac{\prod_{t=1}^n f_\theta(y_t)}{q(y^n)}, \qquad (5)$$

A result of Shtarkov Shtarkov (1988) shows that that infimum in the definition is achieved by

$$q(y^n) = e^{-\Gamma_n} \bar{f}(y^n), \quad \bar{f}(y^n) \triangleq \sup_\theta \prod_{t=1}^n f_\theta(y_t), \qquad (6)$$

assuming that (a) $\bar{f}(y^n)$ is measurable; and (b) that $\int \bar{f} d\mu^{\otimes n} = e^{\Gamma_n} < \infty$.

From the learning point of view, the most interesting case is perhaps $\Phi_n = \mathcal{P}_{iid}(\mathcal{Y}^n) \triangleq \{\text{all iid distributions}\}$, which corresponds to the fully distribution-free regret (or agnostic learning). We denote this special case by $F_n^{(PAC)}(\{f_\theta\}, \mu)$. Note that we always have

$$C_n \le F_n^{(PAC)} \le \Gamma_n .$$

The main motivation for this work was to understand whether $F_n^{(PAC)}$ is closer to C_n or Γ_n. All of the results in this paper suggest the former, thus providing certain justification for the classical focus on the well-specified case.

Our first such result is the following.

Theorem 2 *Let $\mathcal{Y} = \mathbb{R}^d$, $\mu = \mathrm{Leb}$ and $f_\theta(y) = (2\pi)^{-\frac{d}{2}} e^{-\frac{1}{2}\|y-\theta\|^2}$, Θ – a compact subset of $\mathbb{R}^d$ with $\mathrm{Leb}(\Theta) > 0$. Then we have*

$$F_n^{(PAC)} = C_n(\Theta) + o(1),$$

whereas $\Gamma_n = C_n(\Theta) + \frac{d}{2} + o(1)$. The estimator we construct simultaneously achieves $C_n(\Theta) + o(1)$ in the PAC setting and $\Gamma_n + o(1)$ in the individual-sequence setting.

There are several surprises about this result (see Section 2.1 for details). First, the Shtarkov distribution (6) only achieves a suboptimal regret of Γ_n. Thus, this means that there exist an online predictor which is able to exploit the special structure of the iid generated data and therefore reduce the regret compared to the fully adversarial case of Γ_n. Second, the distribution Q_{Y^n} that is (asymptotically) optimal for C_n, namely, the Bayes average over the Jeffrey's prior[1] similarly does not achieve $F_n^{(PAC)}$. Third, our proof strongly suggests that the optimal predictor should provide some robustification against the cases when Y is heavy tailed.

Our next result concerns collections Φ_n slightly smaller than $\mathcal{P}_{iid}(\mathcal{Y}^n)$. We only need to state the $n = 1$ result.

Theorem 3 *Suppose that $\Phi \subset \mathcal{P}(\mathcal{Y})$ is such that for every $P \in \Phi$ we have*

$$D(P\|\Theta) \triangleq \inf_{\theta\in\Theta} D(P\|P_\theta) < \infty\,. \tag{7}$$

Suppose $F(\{P_\theta\}, \mu, \Phi) < \infty$, then there exists a unique distribution $Q^* \ll \mu$ with density q^* such that

$$F(\{f_\theta\}, \mu, \Phi) = \sup_{P\in\Phi} \sup_{\theta\in\Theta} \mathbb{E}_{Y\sim P}\left[\log \frac{f_\theta(Y)}{q^*(Y)}\right]\,.$$

Furthermore, for every $P \in \Phi$ we have $D(P\|Q^*) \le F + D(P\|\Theta) < \infty$.

To appreciate weakness of the condition in the preceding theorem, notice that $D(P\|\Theta) = \infty$ means that average loss of the oracle estimator is infinite, since $\inf_\theta D(P\|P_\theta) = \infty$. However, as Theorem 2 shows even when both losses are infinite in expectation, their difference may still be bounded (just notice that $D(P\|\Theta) = \infty \Leftarrow \mathbb{E}_P[\|X\|^2] = \infty$ and thus there are plenty of such P).

Our final result is about a further smaller collections Φ.

Theorem 4 *Suppose that $\Phi \subset \mathcal{P}(\mathcal{Y})$ is such that a) $P \ll \mu$ for every $P \in \Phi$; b) $C_n(\Phi) = \tau_n n$, $\tau_n \to 0$. Then for every $\epsilon_n \gg \tau_n$ we have*

$$F_n(\Theta, \Phi^{\otimes n}) \le F_n(\Theta, \Theta_{\epsilon_n}^{\otimes n}) + o(1)\,,$$

where $\Theta_\epsilon = \{P \in \Phi : D(P\|\Theta) \le \epsilon\}$, $\Phi^{\otimes n} = \{P^{\otimes n} : P \in \Phi\}$ and similarly for $\Theta_\epsilon^{\otimes n}$.

The meaning of this last result is the following. Since $F_n(\Theta, \Theta_\epsilon^{\otimes n}) \le C_n(\Theta_\epsilon)$, we conclude that for any $\epsilon > 0$ we have

$$C_n(\Theta) \le F_n(\Theta, \Phi^{\otimes n}) \le C_n(\Theta_\epsilon) + o(1)\,. \tag{8}$$

Since Θ_ϵ is a very small enlargement of the model class Θ, in many cases we will have $C_n(\Theta_\epsilon) \to C_n(\Theta)$ as $\epsilon \to 0$ (but not always – see example in Section F.1). In such cases, taking $\epsilon \to 0$ sufficiently slowly will recover $F_n(\Theta, \Phi^{\otimes n}) = C_n(\Theta) + o(1)$ – the same result we have shown in Theorem 2 but for a much larger misspecfication (Φ used there certainly has $C_n(\Phi) = \infty$).

The practical implication of Theorem 4, thus, is that the optimal misspecified regret equals to, and (almost) optimal estimators can be constructed by reducing to the well-specified case with a slightly enlarged model class Θ_ϵ (to which one is free to apply Shtarkov or Jeffreys prior estimators).

[1] Explicitly, $Q_{Y^n}(\cdot) = \int_\Theta \mathrm{Leb}(d\theta) P_\theta^{\otimes n}(\cdot)$.

This last message can be demonstrated heuristically via the following chain (we ignore all rigor and appeal to the intuition here). Define $c(P) \triangleq \inf_\theta D(P\|P_\theta)$ and consider

$$\begin{aligned}
&F_n(\Theta, \Phi^{\otimes n}) \\
&= \inf_{Q_{Y^n}} \sup_{P,\theta} \mathbb{E}_{Y^n \overset{iid}{\sim} P}\left[\log \frac{P_\theta^{\otimes n}(Y^n)}{Q_{Y^n}(Y^n)}\right] = \inf_{Q_{Y^n}} \sup_{P,\theta} \mathbb{E}_{Y^n \overset{iid}{\sim} P}\left[\log \frac{P_\theta^{\otimes n}(Y^n)}{Q_{Y^n}(Y^n)} \frac{P^{\otimes n}(Y^n)}{P^{\otimes n}(Y^n)}\right] \\
&= \inf_{Q_{Y^n}} \sup_{P} D(P\|Q_{Y^n}) - n \inf_\theta D(P\|P_\theta) = \inf_{Q_{Y^n}} \sup_{P} D(P\|Q_{Y^n}) - nc(P) \\
&= \inf_{Q_{Y^n}} \sup_{\pi} \mathbb{E}_{\phi\sim\pi}\left[D(P_\phi\|Q_{Y^n}) - nc(P_\phi)\right] = \sup_{\pi} \inf_{Q_{Y^n}} \mathbb{E}_{\phi\sim\pi}\left[D(P_\phi\|Q_{Y^n}) - nc(P_\phi)\right] \\
&= \sup_\pi I(\phi; Y^n) - n\mathbb{E}_\pi[c(P)]\,, \qquad (9)
\end{aligned}$$

where for clarity we introduced index ϕ indexing over all distribution $P_\phi \in \Phi$, and a prior π on the random variable ϕ on Φ and used the fact that min-max equal max-min for convex-concave functions. Suppose, furthermore, that $F_n(\Theta, \Phi^{\otimes n}) \le \Gamma_n(\Theta) = o(n)$. Then, this implies that the least-favorable prior π in the final equation above must satisfy $\mathbb{E}_\pi[c(P)] \le \frac{1}{n}\Gamma_n(\Theta) \to 0$. This means that while the misspecified setting permits $Y^n \overset{iid}{\sim} P$ for any $P \in \Phi$, in reality restricting P to those with $D(P\|\Theta) = o(1)$ results in a vanishing influence on regret. This last statement can be taken as a summary of our work.

Structure of the paper. In the next Section 1.2 we present motivation for studying $F(\Theta, \Phi)$. Section 2 proves Theorem 2 and discusses the surprises mentioned above. Appendices are devoted to the proofs of the two other Theorems. Finally, Appendix G lists some of the open questions we consider interesting.

1.2. Motivation and literature

Motivation. Why would one consider quantity like $F(\Theta, \Phi)$? The distinguishing property of the log-loss is that the same quantity appears simultaneously in three conceptually very different tasks: sequential prediction, online density estimation and model selection.

First, consider the *sequential prediction*, we think of Θ as hypothesis class, and the learner's goal is to predict $Y^n \overset{iid}{\sim} P$ as good as the best hypothesis in the class, which is given by $P_{\theta^*} = \arg\min_{\theta\in\Theta} \mathbb{E}_P \log \frac{1}{P_\theta}$. The cumulative regret of the learner $Q_t(\cdot|Y^{t-1})$ with respect to the hypothesis class Θ becomes

$$\mathrm{Reg}(\{Q_t\}, \Theta, \Phi) \triangleq \sup_{P\in\Phi} \sum_{t=1}^n \mathbb{E}_{Y^n \overset{iid}{\sim} P}[\log \frac{1}{Q_t(Y_t|Y^{t-1})} - \log \frac{1}{P_{\theta^*}(Y_t)}]\,. \qquad (10)$$

And, clearly, the problem $F_n(\Theta, \Phi^{\otimes n})$ corresponds to searching for a learner that minimizes this regret. See (Cesa-Bianchi and Lugosi, 2006, Chapter 9) and Merhav and Feder (1998) for more.

Second, let us replace both numerators in (10) with $P(Y_t)$ to get

$$\mathrm{Reg}(\{Q_t\}, \Theta, \Phi) = \sup_{P\in\Phi} \sum_{t=1}^n \left\{\mathbb{E}_{Y^n \overset{iid}{\sim} P}[D(P\|Q_t)] - \min_{\theta\in\Theta} D(P\|P_\theta)\right\} \qquad (11)$$

We see that now regret of the learner Q_t corresponds to a problem of *density estimation*. Indeed, consider first the case of $\Phi = \Theta$, in which case the last term is zero and the regret becomes simply the cumulative KL-divergence loss. Thus, the problem $F_n(\Theta, \Phi^{\otimes n}) = C_n(\Theta)$ is merely a cumulative version of the (improper) density estimation of the class Θ. This observation leads to sharp results in statistics, as pioneered by Haussler et al. (1997); Yang and Barron (1999).

The misspecified case of $\Phi \neq \Theta$ corresponds, then, to the competitive optimality variation of the density estimation. Importance of this problem was highlighted by the influential Orlitsky and Suresh (2015), who considered estimating large-alphabet discrete distributions. They noted that estimators achieving $\min_{\{Q\}} \max_{P\in\Phi} \mathbb{E}[D(P\|Q)]$ are empirically rather uninteresting. However, by selecting a natural class Θ and seeking to minimize (11) one does discover interesting estimators.

Third, in *model selection*, one seeks to compare two models Θ_1 vs Θ_2 given observations $Y^n \overset{iid}{\sim} P$. In the spirit of maximum likelihood (or minimal KL-divergence), a natural way to decide which model fits the data better would be to compare

$$\sup_{\theta\in\Theta_1} \mathbb{E}_{Y\sim P}[\log P_\theta(Y)] \lessgtr \sup_{\theta\in\Theta_2} \mathbb{E}_{Y\sim P}[\log P_\theta(Y)]\,.$$

However, this requires computing population averages w.r.t. P. Attempt to fix this issue by replacing $\mathbb{E}_{Y\sim P}[\log P_\theta(Y)]$ with $\frac{1}{n}\sum_{i=1}^n \log P_\theta(y_i)$ results in well-known significant biases for large models Θ. The idea behind the minimum description length (MDL) principle Grünwald and Grunwald (2007) is to associate with each model Θ_i a certain "composite likelihood" $\log Q_i(y^n)$, where each Q_i is chosen to satisfy for all $P \in \Phi$ and $Y^n \overset{iid}{\sim} P$

$$\frac{1}{n}\mathbb{E}[\log Q_i(Y^n)] \approx \sup_{\theta\in\Theta_i} \mathbb{E}[\log P_\theta(Y)]\,.$$

Clearly, the Q_i that makes this $\approx$ the tightest is the one to minimize $\mathrm{Reg}(\{Q_t\}, \Theta_i, \Phi)$. We note that in the classical incarnation of the MDL, one uses either $\Phi = \Theta_i$ or $\Phi = \{\text{all distributions}\}$ for defining Q_i. Following the results in this paper, we propose that taking Φ to be all i.i.d. distributions would result in model selection criteria much more robust to outliers and deviations.

In all, we suggest that a robustified MDL should be implemented as follows: To compare two model classes Θ_1 and Θ_2, one (a) finds good predictors (closely) attaining $F_n^{(PAC)}(\Theta_i)$ for $i = \{1, 2\}$; (b) runs each predictor against the sequence $y_1, \ldots, y_n$; and (c) the winning model is the one whose predictor incurred smaller log-loss.

Literature. We do not survey the enormous literature available on well-specified C_n and individual-sequence Γ_n problems and refer to Cesa-Bianchi and Lugosi (2006); Merhav and Feder (1998).

For the misspecified case, the most directly relevant paper we are aware of is Takeuchi and Barron (1998), which focuses precisely on $F_n(\Theta, \Phi_n)$. However, the methods there are purely applicable to finite-dimensional parametric models and restricted classes Φ_n, which certainly do not include the case of $F_n^{(PAC)}$.[2]

Another relevant work Grünwald and De Rooij (2005) studied the plugin maximum-likelihood (ML) predictors $Q_t(\cdot|Y^{t-1}) = P_{\widehat{\theta}_{ML}(Y^{t-1})}$, where $\widehat{\theta}_{ML}(Y^{t-1})$ is a (slightly modified) ML estimate of θ. They showed that in a 1-parameter exponential families the model misspecification leads to a regret $\frac{c}{2}\log n + O(1)$, where $c = \frac{\mathrm{Var}_P[Y]}{\mathrm{Var}_{P_{\theta^*}}[Y]}$. So compared to the optimal scaling (of e.g.

[2] Note that Takeuchi and Barron (1998) only contains proof sketches, and furthermore different versions we found online state different conditions on Φ. One of the authors of Takeuchi and Barron (1998) confirmed to us that the more complete version is still in preparation.

Shtarkov's predictor) plugin ML estimator can be suboptimal by an arbitrarily bad factor. Subsequently, Grünwald and Kotlowski (2010) showed that replacing $\widehat{\theta}_{ML}$ above with any other function $\widehat{\theta}(Y^{t-1})$ (different for all t), results in exactly the same scaling of regret.

A notable alternative to a plug-in ML estimator is the sequential normalized ML (SNML) estimator, which is simply a conditional Shtarkov distribution Rissanen and Roos (2007), and a close relative of the last-step-minimax algorithm of Takimoto and Warmuth (2000). Regarding these estimators, it was shown Kotłowski and Grünwald (2011) that the SNML does achieve the correct $\frac{d}{2}\log n + O(1)$ regret in the well-specified case. It turns out that SNML and the Bayes estimators with Jeffreys prior coincide whenever they are optimal Bartlett et al. (2013).

The space of questions and amount of literature dramatically expands once we incorporate regressors X_i into the picture, so that prediction of Y_{t+1} is done on the basis of $(X_1, \ldots, X_{t+1}, Y_1, \ldots, Y_t)$. Here, the gap between the PAC-misspecified and worst-case is very easy to demonstrate (take $Y_i = 1\{X_i \le \theta\}$ with $\theta \in [0,1]$ – the 1D-barrier – which cannot be predicted, $\Gamma_n = \infty$, in the worst case, but is easy in the iid case). For increasingly more general losses, Rakhlin et al. (2010); Rakhlin and Sridharan (2014, 2015a) show that regret can be sharply characterized by the metric-entropy type quantites (sequential Rademacher complexity). However, for the log-loss it turns out that the entropic characterization is not possible, cf. Bilodeau et al. (2020) building on the predictor from Rakhlin and Sridharan (2015b). Other recent results about non-parametric models under log-loss can be found in Fogel and Feder (2018); Grünwald and Mehta (2020), which study non-cumulative (batch) regret under the misspecification.

The SNML idea was extended to the case of regressors in Fogel and Feder (2018) under the name of predictive NML (pNML). Subsequently, Rosas et al. (2020) demonstrated that the role of training is to focus attention to a smaller subclass of Θ, on which one may perform the NML.

2. Proof of Theorem 2

Let us now restrict attention to the special case of a compact set $\Theta \subset \mathbb{R}^d$ and the following model class (Gaussian Location Model, GLM):

$$p_\theta(y) = (2\pi)^{-d/2} e^{-\|y-\theta\|^2/2} . \tag{12}$$

Define also

$$I_n = \frac{d}{2} \ln \frac{n}{2\pi e} + \ln \mathrm{Leb}(\Theta) . \tag{13}$$

Standard results, e.g. Xie and Barron (2000), show that

$$C_n = I_n + o(1), \quad \Gamma_n = I_n + \frac{d}{2} + o(1) . \tag{14}$$

Proof [Proof of Theorem 2] Since it is clear that $F_n \ge C_n$ we only need to prove an upper bound. We will, thus, prove that for any $\delta > 0$ there exists an $n_0 = n_0(\delta)$ such that for all $n \ge n_0$ we have

$$F_n \le I_n + \delta .$$

For any set $A \subset \mathbb{R}^d$ we define $d(x, A) = \inf\{\|x-y\| : y \in A\}$. For each $\tau \ge 0$ we define compact sets $\Theta_\tau = \{x : d(x, \Theta) \le \tau\}$ and the projector on Θ_τ as

$$c_\tau(y) \triangleq \operatorname*{argmin}_{x \in \Theta_\tau} \|x - y\| ,$$

with ties resolved in such a way that the resulting function is measureable (that this is possible follows from the measurable selection theorem (due to Kuratowski and Ryll-Nardzewski): indeed for any open U the set $\{y : d(y, U \cap \Theta_\tau) = d(y, \Theta_\tau)|\}$ is measurable).

Fix $v \in \mathbb{R}_+$ and $\tau > 0$ and define the Shtarkov distribution P_S on Y^n with density

$$p_S(y^n) = \frac{1}{Z_{v,\tau}} \sup_{\theta \in \Theta_\tau} (2\pi v)^{-nd/2} e^{-\frac{1}{2v}\sum_{t=1}^n \|y_t - \theta\|^2}\,, \tag{15}$$

where $Z_{v,\tau}$ is the normalization constant, which satisfies $\ln Z_{v,\tau} \approx I_n + \frac{d}{2}$ (when $v \to 1, \tau \to 0$ and $n \to \infty$) as shown in Lemma 6 (Appendix B).

Remark 5 *An important point of our analysis is the following. The Shtarkov distribution (with $v = 1$ and $\tau = 0$) achieves both the C_n (up to $o(1)$) and Γ_n (exactly). However, even for $Y \overset{iid}{\sim} \mathcal{N}(\theta, \sigma^2 I_d)$ with $\sigma^2 < 1$ it yields a suboptimal regret. The choice $v = 1 - n^{-1}$ fixes this problem and makes Shtarkov optimal for a class of all $O(1)$-subgaussian Y. However, for heavy-tailed Y Shtarkov remains suboptimal (Section 2.1) and we need to incorporate some robustification into the estimator. This was surprising to us.*

We will freely use the following change of coordinates on y^n. Let $V_1 = \{y^n : y_1 = \ldots = y_n \in \mathbb{R}^d\}$ be a d-dimensional subspace and $V_\perp$ its orthogonal complement. We denote by $y_\perp$ the orthogonal projection of y^n onto $V_\perp$. We then have the following convenient orthogonal decomposition:

$$y^n = y_\perp + \bar{y}_n \otimes \mathbf{1}\,, \qquad \bar{y}_n = \frac{1}{n}\sum_{t=1}^n y_t$$

where $\mathbf{1}$ is the all-1 vector in $\mathbb{R}^n$ and for $a \in \mathbb{R}^d$, $b \in \mathbb{R}^n$ the $(a \otimes b)_m = a_{m \mod d} b_{\lceil m/d \rceil}$ is the standard Kronecker product. Note also that for any function $f(y^n) = \psi(y_\perp, \bar{y}_n)$ we have

$$\int_{\mathbb{R}^{nd}} f(y^n)\mathrm{Leb}(dy^n) = n^{d/2} \int_{\mathbb{R}^{(n-1)d}} \mathrm{Leb}(dx_1) \int_{\mathbb{R}^d} \mathrm{Leb}(dx_2)\psi(x_1, x_2)\,. \tag{16}$$

It is easy to see that we have

$$p_S(y^n) = \frac{1}{Z_{v,\tau}}(2\pi v)^{-nd/2} e^{-\frac{1}{2v}(\|y_\perp\|^2 + n\|\bar{y}_n - c_\tau(\bar{y}_n)\|^2)} \tag{17}$$

$$p_\theta(y^n) = (2\pi)^{-nd/2} e^{-\frac{1}{2}(\|y_\perp\|^2 + n\|\bar{y}_n - \theta\|^2)}\,. \tag{18}$$

Next fix $\alpha > 0$ and define distribution P_E on $\mathbb{R}^{nd}$ with density[3]

$$p_E(y^n) = (2\pi)^{-(n-1)d/2} e^{-\frac{1}{2}\|y_\perp\|^2} \times \left(\frac{\alpha}{2\sqrt{n}}\right)^d e^{-\alpha\|\bar{y}_n\|_1}\,,$$

where $\|v\|_1 = \sum_{j=1}^d |v_j|$ for any $v \in \mathbb{R}^d$. Using (16) we can check that this is indeed a valid probability density.

[3] Instead of $\|\bar{y}_n\|_1$ we could equally well use $\|\bar{y}_n\|$ but then normalization constant would be more complicated. This is the only reason for using the 1-norm.

Finally, fix $\lambda > 0$, set $\bar{\lambda} = 1 - \lambda$ and define the estimator's density:

$$q_{Y^n}(y^n) = \bar{\lambda} p_S(y^n) + \lambda p_E(y^n)\,. \tag{19}$$

We complete the proof by showing the following statement: For every $\delta > 0$ we can select sufficiently small $\tau, \lambda > 0$ and sufficiently large n_0 and $\beta > 0$ so that for all $n \ge n_0$ by setting

$$v = v_n = 1 - \frac{1}{n} + \frac{\beta}{n^2} \tag{20}$$

we have

$$\sup_{\theta \in \Theta} \mathbb{E}_{Y_t \overset{iid}{\sim} P_Y} \left[R(Y^n; \theta)\right] \le I_n + \delta\,, \qquad \forall P_Y \tag{21}$$

where $R(y^n; \theta) = \ln \frac{p_\theta(y^n)}{q(y^n)}$, which we will upper bound as

$$R(y^n; \theta) \le \min(R_1(y^n; \theta), R_2(y^n; \theta))\,, \tag{22}$$

where $R_1(y^n; \theta) \triangleq \ln \frac{p_\theta(y^n)}{\bar{\lambda} p_S(y^n)}$ and $R_2(y^n; \theta) \triangleq \ln \frac{p_\theta(y^n)}{\lambda p_E(y^n)}$.

First, we show that without loss of generality we may assume that $\mathbb{E}[\|Y\|^2] < \infty$. We have

$$R_2(y^n; \theta) = \widetilde{b}_n + \alpha \|\bar{y}_n\|_1 - \frac{n}{2}\|\bar{y}_n - \theta\|^2\,,$$

where $\widetilde{b}_n = \frac{d}{2} \ln \frac{4n}{2\pi\alpha^2} - \ln \lambda$. From (22) and $\|\bar{y}_n\|_1 \le \sqrt{d}\|\bar{y}_n\|$ we get

$$R(y^n; \theta) \le \widetilde{b}_n + \alpha\|\theta\|_1 + \alpha\sqrt{d}\|\bar{y}_n - \theta\| - \frac{n}{2}\|\bar{Y}_n - \theta\|^2\,.$$

Note that for any random vector X with $\mathbb{E}[\|X\|^2] = \infty$ and any constants $a, b > 0$ we must have $\mathbb{E}[a\|X\| - b\|X\|^2] = -\infty$. Thus, if $\mathbb{E}[\|\bar{Y}_n\|^2] = \infty$ then the expectation in (21) equals $-\infty$ and there is nothing to prove. Consequently, we assume $\mathbb{E}[\|\bar{Y}_n\|^2] < \infty$, which by Lemma 7 (Appendix B) implies $\mathbb{E}[\|Y\|^2] < \infty$, as claimed. We denote $\mu \triangleq \mathbb{E}[Y]$ and $V \triangleq \mathbb{E}[\|Y - \mu\|^2]$.

For Y with finite two moments we can decompose

$$\sup_{\theta} \mathbb{E}[R(Y^n; \theta)] = \text{const} + \sup_{\theta \in \Theta} \mathbb{E}[-\frac{n}{2}\|\bar{Y}_n - \theta\|^2]\,.$$

and thus the supremum over θ is attained at $\theta = c_0(\mu)$. Fixing $\theta = c_0(\mu)$ we have

$$R_1(y^n; \theta) = a_n + \frac{1}{2}\sum_t \frac{1}{v}\|y_t - c_\tau(\bar{y}_n)\|^2 - \|y_t - c_0(\mu)\|^2\,, \qquad a_n \triangleq \frac{nd}{2}\ln v + \ln \frac{Z_{v,\tau}}{\bar{\lambda}}$$

$$R_2(y^n; \theta) = a_n + b_n + \alpha\|\bar{y}_n\|_1 - \frac{n}{2}\|\bar{y}_n - c_0(\mu)\|^2\,, \qquad b_n \triangleq \widetilde{b}_n - a_n$$

We transform expression for R_1 using the following identities valid for arbitrary $x \in \mathbb{R}^d$:

$$\sum_t \|Y_t - x\|^2 = n\|\bar{Y}_n - x\|^2 + n\widehat{V}\,, \qquad \widehat{V} \triangleq \frac{1}{n}\sum_t \|Y_t - \bar{Y}_n\|^2$$

$$\mathbb{E}\left[\|\bar{Y}_n - x\|^2\right] = \frac{V}{n} + \|x - \mu\|^2, \qquad \mathbb{E}\left[\widehat{V}\right] = \frac{n-1}{n}V\,,$$

Applying these to R_1 we get

$$\begin{aligned}\min(R_1, R_2) &= a_n - \frac{n}{2}\|\bar{Y}_n - c_0(\mu)\|^2 + \min\left(\frac{n}{2v}\|\bar{Y}_n - c_\tau(\bar{Y}_n)\|^2 + \frac{n}{2}(v^{-1}-1)\widehat{V}, b_n + \alpha\|\bar{Y}_n\|_1\right) \\ &\le a_n - \frac{n}{2}\|\bar{Y}_n - c_0(\mu)\|^2 + \frac{n}{2}(v^{-1}-1)\widehat{V} + \widetilde{R}_n\,, \qquad (23) \\ \widetilde{R}_n &\triangleq \min\left(\frac{n}{2v}\|\bar{Y}_n - c_\tau(\bar{Y}_n)\|^2, b_n + \alpha\|\bar{Y}_n\|_1\right)\end{aligned}$$

We now consider separately $\mu \in \Theta_\tau$ and $\mu \notin \Theta_\tau$. Suppose the former and consider $\|\bar{y}_n - \mu\| \le \tau$. Then denoting $x = c_0(\mu) + \bar{y}_n - \mu \in \Theta_\tau$, we have from the definition of c_τ that

$$\|\bar{y}_n - c_\tau(\bar{y}_n)\| \le \|\bar{y}_n - x\| = \|\mu - c_0(\mu)\|\,. \qquad (24)$$

Also observe that $\beta_2 = \sup_n b_n$ and $\beta_3 = \sup_{\mu\in\Theta_\tau}\|\mu\|$ are both finite. Thus, for some $\beta_4 > 0$

$$\mathbb{E}[(b_n + \alpha\|\bar{Y}_n\|_1)1\{\|\bar{Y}_n - \mu\| > \tau\}] \le \beta_4\frac{V}{2n}\,. \qquad (25)$$

Indeed, notice that $\|\bar{Y}_n\|_1 \le \sqrt{d}\|\bar{Y}_n\|_2 \le \sqrt{d}\|\bar{Y}_n - \mu\|_2 + \sqrt{d}\beta_3$. From Chebyshev we have then $\mathbb{P}[\|\bar{Y}_n - \mu\| \ge \tau] \le \frac{V}{n\tau^2}$. On the other hand,

$$\mathbb{E}[\|\bar{Y}_n - \mu\|1\{\|\bar{Y}_n - \mu\| \ge \tau\} \le \frac{1}{\tau}\mathbb{E}[\|\bar{Y}_n - \mu\|^2] = \frac{V}{n\tau}\,.$$

Combining these two estimates yields (25).

We now bound $\mathbb{E}[\widetilde{R}_n]$ by retaining the first term of the minimum if $\|\bar{Y}_n - \mu\| \le \tau$ (and invoking (24)) and the second term otherwise. This results in a bound

$$\mathbb{E}[\widetilde{R}_n] \le \beta_4\frac{V}{n} + \frac{n}{2v}\|\mu - c_0(\mu)\|^2\,.$$

Plugging the latter into (23) we have shown that for all $\mu \in \Theta_\tau$

$$\mathbb{E}[\min(R_1(Y^n), R_2(Y^n))] \le a_n + \frac{n}{2}\|\mu - c_0(\mu)\|^2(v^{-1}-1) + \frac{1}{2}V\gamma_n\,,$$

where $\gamma_n = (v^{-1}-1)(n-1) + \frac{\beta_4}{n} - 1$. Recall that due to (20) we have

$$\frac{1-v}{v}(n-1) = \frac{n-1}{n}\frac{1-\beta/n}{1-\frac{1}{n}+\frac{\beta}{n^2}} \le \frac{n-1}{n}\frac{1-\beta/n}{1-\frac{1}{n}} = 1 - \frac{\beta}{n}\,. \qquad (26)$$

Consequently, we have $\gamma_n \le \frac{\beta_4-\beta}{n}$. By chosing $\beta = \beta_4$ we obtain

$$\mathbb{E}[\min(R_1(Y^n), R_2(Y^n))] \le a_n + \frac{n}{2}\|\mu - c_0(\mu)\|^2(v^{-1}-1)\,.$$

Again applying (26) we further upper bound $\frac{n}{2}(v^{-1}-1) \le 1$ for all sufficiently large n, and finally obtain

$$\mathbb{E}[\min(R_1(Y^n), R_2(Y^n))] \le a_n + \|\mu - c_0(\mu)\|^2 \le a_n + \tau^2\,,$$

since $\mu \in \Theta_\tau$. The proof of (21) in this case is completed after noticing that

$$a_n = -\frac{d}{2} + \ln Z_{v,\tau} - \ln \bar{\lambda} + o(1)$$

and that by Lemma 6 (Appendix B), sufficiently small τ and $\lambda > 0$ yield $a_n + \tau^2 \le I_n + \delta$.

Next consider $\mu \notin \Theta_\tau$. In this case

$$\mathbb{E}[\|\bar{Y}_n\|_1] \le \sqrt{d}(\mathbb{E}[\|\bar{Y}_n - \mu\|] + \|\mu - c_0(\mu)\| + \|c_0(\mu)\|) \tag{27}$$

$$\le \sqrt{d}(\sqrt{\frac{V}{n}} + \|\mu - c_0(\mu)\| + \beta_3) \tag{28}$$

Noticing that $\widetilde{b}_n = \frac{d}{2}\ln n + O(1)$ and denoting $\widetilde{\alpha} = \sqrt{d}\alpha$ we get for some large constant β_5: the bound

$$\mathbb{E}[R_2(Y^n;\theta)] \le \frac{d}{2}\ln n + \beta_5 - \frac{V}{2} + \widetilde{\alpha}\sqrt{\frac{V}{n}} - \frac{n}{2}\|\mu - c_0(\mu)\|^2 + \widetilde{\alpha}\|\mu - c_0(\mu)\|$$

Note that $\sup_{n,V>0} -V/2 + \widetilde{\alpha}\sqrt{V/n} < \infty$. Hence, the third and fourth terms can be absorbed into the second. For sufficiently large n the function $-\frac{n}{2}t^2 + \widetilde{\alpha}t$ is monotonically decreasing on $t \in [\tau, \infty)$ and thus we have shown

$$\mathbb{E}[R_2(Y^n;\theta)] \le \frac{d}{2}\ln n + \beta_5 - \frac{n}{2}\tau^2 + \widetilde{\alpha}\tau\,.$$

Clearly, for sufficiently large n the right-hand side of the last inequality is $\le I_n$. This concludes the proof of (21).

Finally, to show the claim about q_{Y^n} also achieving $\Gamma_n + o(1)$ we have to only notice that optimal λ in (19) tends to 0 as $n \to \infty$ and thus $\log \frac{p_S(y^n)}{q_Y(y^n)} = O(\lambda) \to 0$, implying that density q_{Y^n} also attains $\Gamma_n + o(1)$ regret in the individual sequence setting. ■

Interestingly, a similar technique can be used to show a certain curious robustness result for maximum entropy. See Appendix C for details.

2.1. Discussion

Suboptimality of Shtarkov's estimator. As we remarked above, the Shtarkov distribution (which simultaneously achieves $C_n + o(1)$ in the well-specified setting and Γ_n in the worst-case one), surprisingly, is suboptimal for the distribution-free misspecified case.

First, let us focus on the case of $d = 1$ and $\Theta = [-b, b]$. By dividing (18) and (17), we can derive that Shtarkov's distribution ($v = 1, \tau = 0$) achieves:

$$\sup_\theta \mathbb{E}[\log \frac{p_\theta(Y^n)}{p_S(Y^n)}] = \log Z_{1,0} + \frac{n}{2}\mathbb{E}[(\bar{Y}_n - c(\bar{Y}_n))^2 - (\bar{Y}_n - c(\mathbb{E}[Y])^2]\,,$$

where we defined $c(y) = c_0(y) = \mathrm{sign}(y)|y| \cap b$ and assumed that $Y^n \overset{iid}{\sim} P$ with $\mathbb{E}[|Y|] < \infty$. If we furthermore, assume $\mathbb{E}[Y] \in [-b, b]$ the we get

$$\max_\theta \mathbb{E}\log \frac{p_\theta(Y^n)}{p_S(Y^n)} = \Gamma_n - \frac{1}{2}\mathrm{Var}[Y] + \frac{n}{2}\mathbb{E}[(c(\bar{Y}_n) - \bar{Y}_n)^2]\,. \tag{29}$$

Notice that if $Y_j \overset{iid}{\sim} \mathcal{N}(\theta, 1)$ with $\theta \in (-b, b)$ then the third term is $o(1)$ and we conclude that Shtarkov's distribution attains $C_n(\Theta) + o(1)$ in the well-specified case[4]. However, in the misspecified case the situation is different. Consider, the following heavy-tailed distribution

$$Y = \begin{cases} 0, & \text{w.p. } 1 - \frac{1}{4b^2n^2} \\ \pm 2bn, & \text{w.p. } \frac{1}{8b^2n^2} \end{cases} \tag{30}$$

This has $\mathbb{E}[Y] = 0$ and $\mathrm{Var}[Y] = 1$, but notice the following issue. We have that with probability $\approx \frac{\text{const}}{n}$ out of n iid samples exactly one is going to take the value $2bn$, achieving $\bar{Y}_n = 2b$. Thus, while $\mathbb{E}[Y] = 0$ we have $\mathbb{P}[\bar{Y}_n - c(\bar{Y}_n) \geq b] \geq \frac{\text{const}}{n}$ and, therefore, $\frac{n}{2}\mathbb{E}[(c(\bar{Y}_n) - \bar{Y}_n)^2] \geq \text{const}$. Consequently,

$$\sup_{P \in \Phi} \max_{\theta} \mathbb{E} \log \frac{p_\theta(Y^n)}{p_S(Y^n)} \geq C_n(\Theta) + \text{const} + o(1) \,,$$

implying that Shtarkov distribution does not achieve the optimal value of $F_n^{(PAC)} = C_n + o(1)$.

Suboptimality of Bayes estimator based on Jeffreys prior. Next, we want to show that the Jeffreys prior based estimator also does not achieve $F_n^{(PAC)}$. We consider the following density

$$p_J(y^n) = \frac{1}{2b} \int_{-b}^{b} d\theta p_\theta(y^n) = \frac{1}{2b\sqrt{n}} (2\pi)^{-(n-1)/2} e^{-\frac{1}{2}\|y_\perp\|^2} q_2(\bar{y}_n) \,, \tag{31}$$

where we applied (18) and denoted the single-variate function

$$q_2(y) = \mathbb{P}[|y + G_n| < b], \qquad G_n \sim \mathcal{N}(0, 1/n) \,.$$

Dividing (18) by (31) we obtain :

$$\sup_{\theta} \mathbb{E}[\log \frac{p_\theta(Y^n)}{p_J(Y^n)}] = \log \frac{2b\sqrt{n}}{\sqrt{2\pi}} - \inf_{\theta} \mathbb{E}[\frac{n}{2}(\bar{Y}_n - \theta)^2 - \log q_2(\bar{Y}_n)] \,,$$

or assuming that $\mathbb{E}[Y] \in [-b, b]$ and $\mathrm{Var}[Y] = 1$ we get (cf. (13))

$$\sup_{\theta} \mathbb{E}[\log \frac{p_\theta(Y^n)}{p_J(Y^n)}] = I_n - \mathbb{E}[\log q_2(\bar{Y}_n)] \,. \tag{32}$$

When $Y_j \overset{iid}{\sim} \mathcal{N}(\theta, 1)$ (or any subgaussian distribution), we have $q_2(\bar{Y}_n) = 1 - e^{-\Omega(n)}$ with high probability and thus, the last term in (32) is $o(1)$ and we see that indeed Jeffreys prior estimator achieves $C_n + o(1)$ regret in the well-specified case.

However, when $Y_j \overset{iid}{\sim} P$ with distribution (30) the problem occurs. As we argued above, with probablity $\geq \frac{\text{const}}{n}$ we get $\bar{Y}_n = 2b$, which implies $-\log q_2(\bar{Y}_n) = \Omega(n)$, and in turn $\mathbb{E}[-\log q_2(\bar{Y}_n)] > c > 0$. Hence, the regret of p_J is suboptimal as well.

[4] To handle $\theta = \pm b$, we need to consider Shtarkov for an infinitesimally enlarged domain $[-b-\tau, b+\tau]$. In the case of mismatched variance, i.e. when $\mathrm{Var}[Y] < 1$, we need to set $v = 1 - \frac{1}{n}$ instead of $v = 1$.

The nature of the gap between $F_n^{(PAC)}$ and Γ_n. As a final remark, we discuss the meaning of $F_n^{(PAC)} < \Gamma_n$. First, it is easy to show that if we take $\Phi_n = \{\text{exchangeable distributions on } \mathcal{Y}^n\}$ in (4) then the resulting regret $F_n = \Gamma_n$. One may be tempted to infer from this that the PAC-optimal estimator is able to somehow exploit the inherent structure of the iid data (even adversarially generated) and improve prediction compared to the worst-case / exchangeable data. We argue, however, that it is not the quality of the estimator that deteriorates upon relaxing the iid assumption, but rather the quality of the oracle predictor increases.

More exactly, notice that when supremum in (4) is evaluated over all (or all exchangeable) distributions, then the maximum is attained at the extremal point, corresponding to a single known sequence y^n (or its permutations) as evidenced by (5). This, in turn, gives the oracle more freedom as it can adapt to the realization Y^n as opposed to just the distribution of it. This extra freedom is what results in the $\frac{d}{2}$ increase in regret, not an iid structure per se. Indeed, let us redefine the regret in a well-specified case as follows:

$$\bar{C}_n(\Theta) \triangleq \inf_Q \sup_\theta \mathbb{E}_{Y^n \overset{iid}{\sim} P_\theta} \left[\sup_{\theta^*} \log \frac{dP_{\theta^*}}{dQ}(Y^n) \right] ,$$

corresponding to oracle-estimator $P_{\theta^*}^{\otimes n}$ that is chosen given the knowledge of Y^n. In Appendix D we argue that in most cases (and certainly in the GLM model of Theorem 2) we have

$$\bar{C}_n(\Theta) = \Gamma_n + o(1) . \qquad (33)$$

This demonstrates that even in the well-specified case if we give oracle the power to adapt to realization we do get the same regret as Γ_n, thus clarifying the nature of the gap between $F_n^{(PAC)}$ and Γ_n.

Acknowledgments

We would like to thank Olivier Catoni for useful comments, Peter Grünwald for pointing out Takeuchi and Barron (1998), and Alexander Rakhlin for useful discussions. We thank anonymous reviewers for observations regarding exchangeable distributions in Section 2.1. The work of MF was partially supported by a grant from the Israeli Science Foundation (ISF) Grant number 819/20. The work of YP was supported in part by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-09-39370, NSF Grant ECCS-1808692, and the MIT-IBM Watson AI Lab.

References

Peter Bartlett, Peter Grünwald, Peter Harremoës, Fares Hedayati, and Wojciech Kotlowski. Horizon-independent optimal prediction with log-loss in exponential families. In *Conference on Learning Theory*, pages 639–661. PMLR, 2013.

Blair Bilodeau, Dylan Foster, and Daniel Roy. Tight bounds on minimax regret under logarithmic loss via self-concordance. In *International Conference on Machine Learning*, pages 919–929. PMLR, 2020.

Stéphane Boucheron, Aurélien Garivier, and Elisabeth Gassiat. Coding on countably infinite alphabets. *IEEE Transactions on Information Theory*, 55(1):358–373, 2008.

N. Cesa-Bianchi and G. Lugosi. *Prediction, learning and games*. Cambridge University Press, 2006.

Lee D. Davisson and Alberto Leon-Garcia. A source matching approach to finding minimax codes. *IEEE Transactions on Information Theory*, 26(2):166–174, 1980.

Yaniv Fogel and Meir Feder. Universal supervised learning for individual data. *arXiv preprint arXiv:1812.09520*, 2018.

Robert G. Gallager. Source coding with side information and universal coding. Technical Report LIDS Tech. Rep. LIDS-P-937, MIT, Cambridge, MA, 1979.

Peter Grünwald and Steven De Rooij. Asymptotic log-loss of prequential maximum likelihood codes. In *International Conference on Computational Learning Theory*, pages 652–667. Springer, 2005.

Peter Grünwald and Wojciech Kotlowski. Prequential plug-in codes that achieve optimal redundancy rates even if the model is wrong. In *2010 IEEE International Symposium on Information Theory*, pages 1383–1387. IEEE, 2010.

Peter D Grünwald and Abhijit Grunwald. *The minimum description length principle*. MIT press, 2007.

Peter D. Grünwald and Nishant A. Mehta. Fast rates for general unbounded loss functions: From erm to generalized bayes. *Journal of Machine Learning Research*, 21(56):1–80, 2020. URL http://jmlr.org/papers/v21/18-488.html.

David Haussler, Manfred Opper, et al. Mutual information, metric entropy and cumulative relative entropy risk. *The Annals of Statistics*, 25(6):2451–2492, 1997.

Zeyu Jia, Yury Polyanskiy, and Alexander Rakhlin. Work in progress. 2021.

JHB Kemperman. On the Shannon capacity of an arbitrary channel. In *Indagationes Mathematicae (Proceedings)*, volume 77, pages 101–115. North-Holland, 1974.

Wojciech Kotłowski and Peter Grünwald. Maximum likelihood vs. sequential normalized maximum likelihood in on-line density estimation. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 457–476, 2011.

Neri Merhav and Meir Feder. A strong version of the redundancy-capacity theorem of universal coding. *IEEE Transactions on Information Theory*, 41(3):714–722, 1995.

Neri Merhav and Meir Feder. Universal prediction. *IEEE Transactions on Information Theory*, 44 (6):2124–2147, 1998.

Peter Meyer-Nieberg. *Banach lattices*. Springer Science & Business Media, 2012.

Alon Orlitsky and Ananda Theertha Suresh. Competitive distribution estimation: Why is Good-Turing good. *Advances in Neural Information Processing Systems*, 28:2143–2151, 2015.

Yury Polyanskiy and Yihong Wu. Lecture notes on information theory. *Lecture Notes for ECE563 (UIUC) and 6.441 (MIT)*, 2012-2016. URL http://people.lids.mit.edu/yp/homepage/data/itlectures_v5.pdf.

Alexander Rakhlin and Karthik Sridharan. Online non-parametric regression. In *Conference on Learning Theory*, pages 1232–1264. PMLR, 2014.

Alexander Rakhlin and Karthik Sridharan. Online nonparametric regression with general loss functions. *arXiv preprint arXiv:1501.06598*, 2015a.

Alexander Rakhlin and Karthik Sridharan. Sequential probability assignment with binary alphabets and large classes of experts. *arXiv preprint arXiv:1501.07340*, 2015b.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning: Random averages, combinatorial parameters, and learnability. *Advances in Neural Information Processing Systems*, 23:1984–1992, 2010.

Jorma Rissanen and Teemu Roos. Conditional nml universal models. In *2007 Information Theory and Applications Workshop*, pages 337–341. IEEE, 2007.

Fernando Rosas, Pedro Mediano, and Michael Gastpar. Learning, compression, and leakage: Minimizing classification error via meta-universal compression principles. *arXiv preprint arXiv:2010.07382*, 2020.

Boris Yakovlevich Ryabko. Coding of a source with unknown but ordered probabilities. *Problems of Information Transmission*, 15(2):134–138, 1979.

Yuri M. Shtarkov. Universal sequential coding of single messages. *Probl. Inform. Transm.*, 23, Jul. 1988.

Jun-ichi Takeuchi and Andrew R Barron. Robustly minimax codes for universal data compression. In *Proceedings of the Twenty-First Symposium on Information Theory and Its Applications (SITA'98)*, 1998.

Eiji Takimoto and Manfred K Warmuth. The last-step minimax algorithm. In *International Conference on Algorithmic Learning Theory*, pages 279–290. Springer, 2000.

Qun Xie and A.R. Barron. Asymptotic minimax regret for data compression, gambling, and prediction. *IEEE Trans. Inf. Theory*, 46(2):431 –445, Mar. 2000.

Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, pages 1564–1599, 1999.

Appendix A. Technical remarks on the definition of $F(\Theta, \Phi)$

Unlike the well-specified case, where definition (2) is elegant and rigorous, the regret in the misspecified case, the F_n and Γ_n, is more subtle. We list some of the issues in this section.

In (4) and below we adopt the following rules for evaluting $\log$ of the ratio: For any $c > 0$ we set

$$\log \frac{0}{c} = -\infty, \quad \log \frac{c}{0} = +\infty, \quad \log \frac{0}{0} = 0\,. \tag{34}$$

In the infimum over densities q in (4) we only consider *admissible* q, i.e. those q such that for every $P \in \Phi$ and every $\theta \in \Theta$ the expectation of the log-likelihood ratio is well-defined (but possibly infinite)[5] If no such q exist then we take $F = \infty$.

A crucial subtlety concerning quantities F_n and Γ_n is that they are not necessarily functions of the distributions $\{P_\theta\}$. Rather they depend on the particular chosen representative densities f_θ. In some cases, corresponding to the same family P_θ one can choose densities f_θ such that Γ_n is increased by an arbitrary amount. (One example is to take $P_\theta = \mathrm{Uniform}[\theta, 1+\theta]$ and compare densities $f_\theta(x) = 1\{\theta \le x \le 1+\theta\}$ with $f_\theta(x) = 1\{\theta < x \le 1+\theta\} + 1\{x = \theta\}$.) For this reason, we write $C_n(\{P_\theta\})$ but $\Gamma_n(\{f_\theta\}, \mu)$.[6]

Similarly, whereas for the well-specified case we can think of predictor as submitting at each step a distribution $Q_t(\cdot|Y^{t-1})$, in the misspecified case we insist that the predictor submits a density $q_t(\cdot|y^{t-1})$.

Finally, note that none of these difficulties apply to the case when $\mathcal{Y}$ is countable.

Appendix B. Technical results for Section 2

Lemma 6 *For any compact Θ with $\mathrm{Leb}(\Theta) > 0$ we have*

$$\ln Z_{v,\tau} = \frac{d}{2} \ln \frac{n}{2\pi v} + \ln \mathrm{Leb}(\Theta_\tau) + o(1)\,,$$

where $Z_{v,\tau}$ is the normalization constant from (15). Furthermore, as $\tau \to 0$ we have $\mathrm{Leb}(\Theta_\tau) \to \mathrm{Leb}(\Theta)$.

Proof Let us define a function $d(z, B) = \inf_{z' \in B} \|z - z'\|$ which is continuous for any set B. We have

$$\mathrm{Leb}(\Theta_\tau) = \int_{\mathbb{R}^d} 1\{d(z, \Theta) \le \tau\} dz\,.$$

Notice that as $\tau \to 0$ the sequence of functions $1\{d(z, \Theta) \le \tau\}$ converges pointwise to $1\{z \in \Theta\}$. By compactness of Θ, the set Θ_τ is bounded and hence we have from the dominated convergence $\mathrm{Leb}(\Theta_\tau) \to \mathrm{Leb}(\Theta)$ as $\tau \to 0$. This proves the second assertion.

To prove the first assertion we apply identity (16) to (17) to get

$$Z_{v,\tau} = \left(\frac{2\pi v}{n}\right)^{-d/2} \int_{\mathbb{R}^d} e^{-\alpha d(z, \Theta_\tau)^2} dz\,, \quad \alpha \triangleq \frac{n}{2v}\,. \tag{35}$$

[5] Recall that in Lebesgue integration theory we set $\mathbb{E}[X] = \mathbb{E}[\min(X, 0)] + \mathbb{E}[\max(X, 0)]$ unless the two summands are $-\infty$ and $+\infty$, in which case the expectation is undefined.

[6] This suggests that perhaps a more sensible definition of Γ_n would be to replace $\sup_\theta f_\theta(y^n)$ with the definition of supremum common in the theory of Banach lattices (Meyer-Nieberg, 2012, Section 2.6), namely $\mathrm{bsup}_\theta f_\theta(y^n)$ is defined as any function g such that $f_\theta \le g$ (μ-a.e. for all θ) and if $f_\theta \le h$ (μ-a.e. for all θ), then $g \le h$ (μ-a.e.). The advantages of bsup are that a) Γ_n becomes insensitive to μ-negligible modifications of f_θ's, and b) $\mathrm{bsup}_\theta f_\theta(y^n)$ is automatically measurable if μ is σ-finite, cf (Meyer-Nieberg, 2012, Lemma 2.6.1).

From here the result follows since for any bounded set B we have

$$\int e^{-\alpha d(z,B)^2} dz \to \mathrm{Leb}(B), \qquad \alpha \to \infty\,.$$

Indeed, since B is bounded we can include it into the ball of radius r and thus $d(z,B) \ge (\|z\| - r)_+$ for some finite $r > 0$. Since $e^{-(\|z\|-r)_+^2}$ is integrable over Leb in $\mathbb{R}^d$, it dominates all integrands for $\alpha \ge 1$ and we have from the dominated convergence theorem

$$\lim_{\alpha\to\infty} \int e^{-\alpha d(z,B)^2} dz = \int \lim_{\alpha\to\infty} e^{-\alpha d(z,B)^2} dz = \mathrm{Leb}(B)\,.$$

■

Lemma 7 *Let X_i be iid random variables and let $\bar{X}_n = \frac{1}{n}\sum_{t=1}^n X_t$. Then $\mathbb{E}[\bar{X}_n^2] < \infty$ iff $\mathbb{E}[X_1^2] < \infty$.*

Proof We only need to prove that $\mathbb{E}[(\bar{X}_n)^2] < \infty$ implies $\mathbb{E}[(X_1)^2] < \infty$. To that end, let $\Psi(t) = \mathbb{E}[e^{itX_1}]$ be the characteristic function of X_1. From the conditions we know that the complex-valued function $f(t) = \Psi(t/n)^n$ is twice continuously differentiable for all $t \in \mathbb{R}$ and $f(0) = 1$. In a small neighborhood of $z = 1$ on the complex plane there exists an analytic function $f_1(z)$ satisfying $f_1(z)^n = z$. We have then $\Psi(t) = f_1(f(nt))$ in a small neighborhood of $t = 0$. In particular, $\Psi(t)$ is differentiable at 0 and thus X_1 has a finite second moment. ■

Appendix C. Robustness of maximum differential entropy

For any random variable X with pdf f_X let us define $h(X) = -\mathbb{E}[\log f_X(X)]$ to be its differential entropy. It is well known that

$$\sup\{h(X) : X \in [-b,b]\} = \log(2b)\,,$$

showing that uniform distribution has maximal entropy among all distributions with a given support. The following shows a certain kind of robustness result.

Lemma 8 *As $\epsilon \to 0+$ we have*

$$\sup\{h(M+Z) : M \in [-b,b], \mathbb{E}[Z] = 0, \mathrm{Var}[Z] \le \epsilon\} = \log(2b) + O(\epsilon^{1/3} \log \frac{1}{\epsilon})\,.$$

where in the supremization we do not require M and Z be independent, but do require the distribution of $M + Z$ to have density.

Proof Since for any density q and $dQ = q d\mathrm{Leb}$ we have $D(P_{M+Z}\|Q) \ge 0$, or in other words

$$h(M+Z) \le -\mathbb{E}[\log q(M+Z)]\,, \tag{36}$$

Similar to the proof of Theorem 2 we make the following choice ($\tau, \lambda, \alpha > 0$ are to be chosen later):

$$q(x) = \bar{\lambda}\frac{1}{2(b+\tau)} 1\{|x| \le b+\tau\} + \frac{\lambda\alpha}{2} e^{-\alpha|x|}\,.$$

The analysis of the RHS of (36) proceeds similar to the proof of Theorem 2: we consider two cases separately: $|Z| \le \tau$ and $|Z| > \tau$ and bounding $\log \frac{1}{q(M+Z)}$ accordingly on each. Indeed, we have

$$\mathbb{E}[1\{|Z| \le \tau\} \log \frac{1}{q(M+Z)}] \le \log \frac{2(b+\tau)}{1-\lambda}\,.$$

For the other case, denote $p = \mathbb{P}[|Z| > \tau]$ and notice that $\mathbb{E}[|M+Z| 1\{|Z| > \tau\}] \le bp + \frac{1}{\tau}\mathbb{E}[Z^2] \le bp + \frac{\epsilon}{\tau}$ yielding

$$\mathbb{E}[1\{|Z| > \tau\} \log \frac{1}{q(M+Z)}] \le p(\log \frac{2}{\lambda\alpha} + \alpha b) + \frac{\alpha\epsilon}{\tau}\,.$$

Taking here $\alpha = 1$, $\lambda = \tau = \epsilon^{1/3}$ we obtain the result. ■

Appendix D. Justification of (33)

Recall that we are interested in checking

$$\bar{C}_n(\Theta) \triangleq \inf_Q \sup_\theta \mathbb{E}_{Y^n \overset{iid}{\sim} P_\theta}\left[\sup_{\theta^*} \frac{dP_{\theta^*}}{dQ}(Y^n)\right] \overset{?}{=} \Gamma_n + o(1)$$

under some regularity assumptions (say smooth finite-parameter families). To that end note that the inner optimization can be solved explicitly yielding the Shtarkov density $P_S = \sup_\theta p_\theta(y^n) \exp -\Gamma_n$. With this in mind we obtain:

$$\begin{aligned}\bar{C}_n(\Theta) &= \Gamma_n + \inf_Q \sup_\theta \mathbb{E}_{Y^n \overset{iid}{\sim} P_\theta}\left[\log \frac{p_S(y^n)}{q(Y^n)}\right] \\ &= \Gamma_n + \inf_Q \sup_\theta D(P_\theta^{\otimes^n} \| Q_{Y^n}) - D(P_\theta^{\otimes n} \| P_S)\,.\end{aligned}$$

Now extending the inner supremum to supremum over priors on θ and lower-bounding $\inf\sup \ge \sup\inf$[7], we get

$$\bar{C}_n(\Theta) \ge \Gamma_n + \sup_\pi I(\theta; Y^n) - \mathbb{E}_{\theta\sim\pi}[D(P_\theta \| P_S)]\,.$$

Notice also that $I(\theta; Y^n) - \mathbb{E}_{\theta\sim\pi}[D(P_\theta \| P_S)] = -D(P_\pi \| P_S)$, where we denoted $P_\pi = \int \pi(d\theta) P_\theta^{\otimes n}$. In other words, we have shown that

$$0 \le \Gamma_n - \bar{C}_n(\Theta) \le \inf_\pi D(P_\pi \| P_S)\,,$$

i.e. the gap between the two corresponds to how well the Shtarkov distribution can be approximated by a Bayes mixture. In the GLM case, a simple explicit computation taking π to be uniform on $[-b, b]$ (so that $P_\pi = P_J$ in (31)) yields

$$\Gamma_n - \bar{C}_n(\Theta) = O(1/\sqrt{n})\,.$$

(See also (Grünwald and Grunwald, 2007, Appendix to Chapter 8).)

[7] In fact, there is always equality $\inf\sup = \sup\inf$, argued same as for (3), but we do not need this extension.

Appendix E. Proof of Theorem 3

The idea of the proof is best illustrated by looking at the heuristic derivation (9), which concluded with

$$F(\Theta, \Phi) = \sup_{\pi} I(P; Y) - \mathbb{E}_{\pi}[c(P)]\,, \tag{37}$$

where supremum is taken over all prior distributions π on Φ, and the joint distribution of P, Y is given by

$$\mathbb{P}[P = a, Y = b] = \pi(a)P(b)\,.$$

The expression (37) is just a (Langrangian version of the) cost-constrained channel capacity calculation.[8] Thus, a theorem of Kemperman Kemperman (1974) implies that if we take any sequence π_k of priors attaining supremum in (37), the sequence of corresponding induced distributions P_{Y_k} converges, in the sense that $D(P_{Y_k} \| Q^*) \to 0$ (and hence in total variation), with limit Q^* independent of the sequence π_k. This unique Q^* is what also achieves optimality of $F(\Phi, \Theta)$.

Unfortunately, the argument above is very informal. The function $P \mapsto c(P)$ may not be measurable, the stochastic transformation taking element $P \in \Phi$ and outputting a random element $Y \sim P$ may not be a Markov kernel, etc. We proceed, thus, in a rather different way.

Lemma 9 *Let $P, Q, R \ll \mu$ and f_P, f_Q, f_R denote their densities. Then (with* log *of the ratio evaluated according to* (34)*)*

$$\mathbb{E}_P\left[\log \frac{f_R}{f_Q}\right] = D(P\|Q) - D(P\|R)\,, \tag{38}$$

whenever not both divergences are infinite.

Proof First, suppose $D(P\|Q) = \infty$ and $D(P\|R) < \infty$. Then that $P[f_R(Y) = 0] = 0$, and hence in computation of the expectation in (38) only the second part of convention (34) can possibly apply. Since also $f_P > 0$ P-almost surely, we have

$$\log \frac{f_R}{f_Q} = \log \frac{f_R}{f_P} + \log \frac{f_P}{f_Q}\,, \tag{39}$$

with both log's evaluated according to (34). Taking expectation over P we see that the first term, equal to $-D(P\|R)$, is finite, whereas the second term is infinite. Thus, the expectation in (38) is well-defined and equal to $+\infty$, as is the LHS of (38).

Now consider $D(P\|Q) < \infty$. This implies that $P[q(Y) = 0] = 0$ and this time in (38) only the first part of convention (34) can apply. Thus, again we have identity (39). Since the P-expectation of the second term is finite, and of the first term non-negative, we again conclude that expectation in (38) is well-defined, equals the LHS of (38) (and both sides are possibly equal to $-\infty$). ■

Lemma 10 *Let $P, Q \ll \mu$ and f_P, f_Q be their relative densities. Then (with convention* (34) *for the log) we have*

$$\mathbb{E}_P\left[\max(\log \frac{f_P}{f_Q}, 0)\right] \ge -\frac{\log e}{e}\,. \tag{40}$$

Consequently, the expectation $\mathbb{E}_P\left[\log \frac{f_P}{f_Q}\right]$ is well-defined and non-negative (but could be $+\infty$).

[8] Incidentally, this point of view also suggests a non-trivial method for finding the optimal Q^*: the Blahut-Arimoto algorithm.

Proof Let $g(x) = \max(x \log x, 0)$. It is clear $-\frac{\log e}{e} \le g(x) \le 0$ for all x. Since $f_P(Y) > 0$ for P-almost all Y, in convention (34) only the $\frac{1}{0}$ case is possible, which is excluded by the $\max(\cdot, 0)$ from the expectation in (40). Thus, the LHS in (40) equals

$$\begin{aligned}\int_{\{f_P > f_Q > 0\}} f_P(y) \log \frac{f_P(y)}{f_Q(y)} d\mu &= \int_{\{f_P > f_Q > 0\}} f_Q(y) \frac{f_P(y)}{f_Q(y)} \log \frac{f_P(y)}{f_Q(y)} d\mu \\ &= \int_{\{f_Q > 0\}} g\left(\frac{f_P(y)}{f_Q(y)}\right) d\mu \\ &\ge -\frac{\log e}{e}\,.\end{aligned}$$

Since the negative part of $\mathbb{E}_P\left[\log \frac{f_P}{f_Q}\right]$ is bounded, the expectation is well-defined. If $P[f_Q = 0] > 0$ then it is clearly $+\infty$. Otherwise, the said expectation equals $D(P\|Q) \ge 0$. ■

Below we will freely use two facts about well-defined integrals (and expectations). If $f = g + h$ and $\int |h| d\mu < \infty$ the $\int f d\mu$ and $\int g d\mu$ are defined or undefined simultaneously. If $\mu = \mu_0 + \mu_1$ and $\int f d\mu$ is well-defined, then so are $\int f d\mu_0$ and $\int f d\mu_1$ (the opposite is note true, since we can have $\int f d\mu_0 = +\infty$ and $\int f d\mu_1 = -\infty$).

Let us introduce a collection of distributions Π on $\Theta \times \mathcal{Y}$ as follows:

$$\Pi = \left\{ \sum_{k=1}^m p_k \delta_{\theta_k} \times P_k : p_k \ge 0, \sum p_k = 1, \theta_k \in \Theta, P_k \in \Phi \right\}\,.$$

Note that Π is convex. We give Π any topology under which linear operations are continuous (e.g. topology of total variation). For each element $\pi \in \Pi$ we denote by $\pi_Y = \sum_{k=1}^m p_k P_k$ the marginal induced on the second coordinate. By constraint (7) we have $\pi_Y \ll \mu$ and thus we denote $f_\pi = \frac{d\pi_Y}{d\mu}$ the relative density of π_Y. On Π, let us define the following functional:

$$\tilde{J}(\pi) = \begin{cases} \mathbb{E}_\pi\left[\log \frac{f_\theta(Y)}{f_\pi(Y)}\right]\,, & \mathbb{E}[\cdot] \text{ is well-defined} \\ -\infty, & \text{o/w.} \end{cases} \tag{41}$$

We remind of the convention (34) for the log, and observe that $f_\pi(Y) > 0$ almost surely, implying that only the first clause of the convention can possibly apply.

Let $\operatorname{dom} \tilde{J} = \{\pi : \tilde{J}(\pi) > -\infty\}$.

Lemma 11 (Properties of $\tilde{J}$) *The set* $\operatorname{dom} \tilde{J}$ *is convex. The functional* $\pi \mapsto \tilde{J}(\pi)$ *is concave. Furthermore,* $\tilde{J}$ *satisfies for all* $\pi_0, \pi_1 \in \operatorname{dom} \tilde{J}$ *and* $\lambda \in [0, 1]$ *the bound*

$$\tilde{J}(\lambda \pi_1 + (1-\lambda)\pi_0) \le \lambda \tilde{J}(\pi_1) + (1-\lambda)\tilde{J}(\pi_0) + h(\lambda)\,, \tag{42}$$

where $h(x) = -x \log x - (1-x)\log(1-x)$ *is entropy of Ber*(x) *random variable.*

Proof

Indeed, consider $\pi = \lambda\pi_1 + (1-\lambda)\pi_0$. Then π_0-almost surely we have $f_{\pi_0}(Y) > 0$ and $f_\pi(Y) > 0$. Thus, even under convention (34) (first clause) we have π_0-almost surely:

$$\log \frac{f_\theta(Y)}{f_\pi(Y)} = \log \frac{f_\theta(Y)}{f_{\pi_0}(Y)} + \log \frac{f_{\pi_0}(Y)}{f_\pi(Y)}\,.$$

Consequently, we get

$$\mathbb{E}_{\pi_0}\left[\log\frac{f_\theta(Y)}{f_\pi(Y)}\right] = \mathbb{E}_{\pi_0}\left[\log\frac{f_\theta(Y)}{f_{\pi_0}(Y)} + \log\frac{f_{\pi_0}(Y)}{f_\pi}\right] = \widetilde{J}(\pi_0) + D(\pi_{0,Y}\|\pi_Y)\,. \tag{43}$$

Since $D(\pi_{0,Y}\|\pi_Y) \le \log\frac{1}{1-\lambda}$ we conclude that the expectation in the LHS of the last display is well-defined and $> -\infty$. Similarly, the expectation over π_1 is also well-defined and $> -\infty$. Since $\mathbb{E}_\pi = \lambda\mathbb{E}_{\pi_1} + (1-\lambda)\mathbb{E}_{\pi_0}$, we conclude that $\pi \in \mathrm{dom}\,\widetilde{J}$.

Next we prove concavity of $\widetilde{J}(\cdot)$. Indeed, from non-negativity of KL divergence and identity (43) we conclude

$$\widetilde{J}(\lambda\pi_1 + (1-\lambda)\pi_0) \ge \lambda\widetilde{J}(\pi_1) + (1-\lambda)\widetilde{J}(\pi_0)\,.$$

To prove the last claim, consider $\pi = \lambda\pi_1 + (1-\lambda)\pi_0$ and, explicitly, $\pi = \sum_{k=1}^m p_k\delta_{\theta_k} \times P_k$. This implies that

$$\pi_u = \sum_k p_{u,k}P_k \times \delta_{\theta_k}\,, \qquad u \in \{0,1\}$$

where $\lambda p_{1,k} + (1-\lambda)p_{0,k} = p_k$ for all $k \in [m]$. Next, define a joint distribution on four random variables: for all $k \in [m], b \in \{0,1\}, y \in \mathcal{Y}$ set

$$\mathbb{P}[B = b, \theta = \theta_k, \phi = k, Y \in dy] = (1 - \lambda - (1-2\lambda)b)p_{b,k}P_k(dy)\,. \tag{44}$$

If $\pi \in \mathrm{dom}\,\widetilde{J}$ then we have

$$\widetilde{J}(\pi) = I(\phi; Y) - \mathbb{E}[c(\phi,\theta)]\,, \tag{45}$$

where $c(\phi,\theta) = D(P_\phi\|P_\theta) \in [0,+\infty]$ and both terms are finite. Indeed, we have

$$\widetilde{J}(\pi) = \sum_k p_k\mathbb{E}_{Y\sim P_k}\left[\log\frac{f_{\theta_k}(Y)}{f_\pi(Y)}\right]\,.$$

Now, let $f_k = \frac{dP_k}{d\mu}$. Clearly, $P_k[f_k(Y) = 0] = P_k[f_\pi(Y) = 0] = 0$ and thus, we have (under convention (34) for all logs) P_k-almost surely

$$\log\frac{f_{\theta_k}(Y)}{f_\pi(Y)} = \log\frac{f_{\theta_k}(Y)}{f_k(Y)} + \log\frac{f_k(Y)}{f_\pi(Y)}\,. \tag{46}$$

From Lemma 10 the P_k-expectation of each term is well-defined, and since $\log\frac{f_k}{f_\pi} \le \log\frac{1}{p_k}$, the second expectation is finite. Thus, overall we can take P_k-expectation of (46) and conclude

$$\mathbb{E}_{Y\sim P_k}\left[\log\frac{f_{\theta_k}(Y)}{f_\pi(Y)}\right] = D(P_k\|\pi_Y) - D(P_k\|P_{\theta_k})\,.$$

Summing over k we obtain (45). Note that $I(\phi;Y) \le \log m$ and thus for any $\pi \in \mathrm{dom}\,\widetilde{J}$ we must have $D(P_k\|P_{\theta_k}) < \infty$ for any k with $p_k > 0$.

Similarly, we show that $\widetilde{J}(\pi_b) = I(\phi; Y|B = b) - \mathbb{E}[c(\phi,\theta)|B = b]$ for $b = 0, 1$. Finally, since $\phi \perp\!\!\!\perp B|Y$ and the chain rule for mutual information we obtain

$$I(\phi; Y) = I(\phi, B; Y) = I(\phi; Y|B) + I(B; Y) \le I(\phi; Y|B) + h(\lambda)\,,$$

where in the last step we used the fact that $B \sim \text{Ber}(\lambda)$. ■

We next define the following functional: For every $\pi \in \Pi$ and $q : \mathcal{Y} \to \mathbb{R}_+$ with $\int q d\mu = 1$ we define

$$J_1(\pi, q) = \begin{cases} \mathbb{E}_\pi \left[\log \frac{f_\theta(Y)}{q(Y)}\right] , & \mathbb{E}[\cdot] \text{ is well-defined} \\ +\infty, & \text{o/w.} \end{cases}$$

Again, we remind of the convention (34). We claim that

$$F(\Theta, \Phi) = \inf_q \sup_{\pi \in \operatorname{dom} \widetilde{J}} J_1(\pi, q)\,. \tag{47}$$

Indeed, by agreement made in Appendix A for any inadmissible q we have that its regret is declared to be $+\infty$. Recall that q is inadmissible if for any pair $\theta \in \Theta$ and $P \in \Phi$ the expectation $\mathbb{E}_{Y \sim P} \log \frac{f_\theta(Y)}{q(Y)}$ is undefined. But then taking $\pi = \delta_\theta \times P$, shows that for such a q we must have $\sup_{\pi \in \operatorname{dom} \widetilde{J}} J_1(\pi, q) = +\infty$ as well. For admissible q's it is clear that only the regular case in the definition of $J_1(\pi, q)$ applies and thus $\pi \mapsto J_1(\pi, q)$ is affine on Π. Consequently, supremum over $\pi \in \Pi$ coincides with the supremum over special $\pi = \delta_\theta \times P$. Note that $\pi = \delta_\theta \times P \in \operatorname{dom} \widetilde{J}$ iff $D(P\|P_\theta) < \infty$. Therefore, proof of (47) is completed once we can show

$$\sup_{\theta \in \Theta, P \in \Phi} \mathbb{E}_P[\log \frac{f_\theta}{q}] = \sup_{\theta \in \Theta, P \in \Phi : D(P\|P_\theta) < \infty} \mathbb{E}_P[\log \frac{f_\theta}{q}]\,. \tag{48}$$

To show (48), we first prove the following lemma:

Equipped with Lemma 9, we can argue (48) as follows. Suppose for some $P \in \Phi$ we have $D(P\|Q) = \infty$, where $dQ = q d\mu$. Bu then, by assumption (7) there exists θ_0 such that $D(P\|P_{\theta_0}) < \infty$. Thus (38) implies both sides of (48) evaluate to $+\infty$. Next, suppose for every $P \in \Phi$ we have $D(P\|Q) < \infty$. Then again from (38) we see that pair (P, θ_0) with $D(P\|P_{\theta_0}) = \infty$ yield $-\infty$ values in the RHS of (48) and can be excluded.

Lemma 12 *For any $\pi \in \operatorname{dom} \widetilde{J}$ and any q we have*

$$J_1(\pi, q) = \widetilde{J}(\pi) + D(\pi_Y \| Q) > -\infty\,. \tag{49}$$

In particular, for any $\pi \in \operatorname{dom} \widetilde{J}$ we have

$$\widetilde{J}(\pi) = \min_q J_1(\pi, q) = J_1(\pi, f_\pi)\,.$$

Proof As before, under π measure on (θ, Y) we have that $f_\pi(Y) > 0$ almost surely. Since $\widetilde{J}(\pi) > -\infty$, we see from (41) that $\pi[f_\theta(Y) = 0] = 0$ and thus $f_\theta(Y) > 0$ almost surely as well. Thus, under convention (34) for all logs we have almost surely

$$X \triangleq \log \frac{f_\theta(Y)}{q(Y)} = \log \frac{f_\theta(Y)}{f_\pi(Y)} + \log \frac{f_\pi(Y)}{q(Y)}\,.$$

Denoting the two terms as A and B. By assumption $\mathbb{E}_\pi[A] = \widetilde{J}(\pi) > -\infty$. On the other hand, from Lemma 10 we know $\mathbb{E}_\pi[\max(B, 0)] > -\infty$ and thus the expectation $\mathbb{E}_\pi[X]$ is well-defined (and $> -\infty$). Consequently, the value $J_1(\pi, q) = \mathbb{E}_\pi[X] = \mathbb{E}_\pi[A] + \mathbb{E}_\pi[B]$, completing the proof of (49). ■

We next establish the saddle-point property of $J_1(\pi, q)$ on finite-dimensional subsets of Π.

Lemma 13 (Saddle point) *Let $\pi_1, \dots, \pi_k$ be any elements in $\mathrm{dom}\,\widetilde{J}$ with $\max_i \widetilde{J}(\pi_i) < \infty$. Let $\Pi_k = \mathrm{co}(\pi_1, \dots, \pi_k)$. Then the function $\pi \mapsto \widetilde{J}(\pi)$ is continuous on Π_k and achieves its maximum $F = \max_{\pi \in \Pi_k} \widetilde{J}(\pi)$. For any maximizer π^*, set $q^* = f_{\pi^*}$. Then the pair (π^*, q^*) is the saddle point: For all $\pi \in \Pi_k$ and all $q \ge 0$, $\int q d\mu = 1$ we have*

$$J_1(\pi, q^*) \le F = J_1(\pi^*, q^*) \le J_1(\pi^*, q) \tag{50}$$

The density q^* with the property $F = \max_\pi J_1(\pi, q^*)$ is unique.

Proof Let us take $q_0 = \frac{1}{k}\sum_{i=1}^k f_{\pi_i}$ and denote $dQ_0 = q_0 d\mu$. Then from (49) we have

$$\widetilde{J}(\pi) = J_1(\pi, q_0) - D(\pi_Y \| Q_0)\,.$$

Notice that the map $\pi \mapsto D(\pi_Y \| Q_0)$ is continuous on Π_k (since $0 \le \frac{d\pi_Y}{dQ_0} \le k$, so bounded-convergence theorem holds). On the other hand, $\pi \to J_1(\pi, q_0)$ is affine and continuous on Π_k (since it is finite at extremal points $\pi_1, \dots \pi_k$). This proves continuity of $\widetilde{J}(\pi)$ on Π_k. (Note that continuity on the interior of Π_k automatically follows from concavity.)

The right-hand inequality in (50) follows from (49) and non-negativity of divergence. For the (key) left-most inequality fix $\pi \in \Pi_k$ and define for each $\lambda \in [0, 1)$ the density $q_\lambda = (1-\lambda)q^* + \lambda f_\pi$ and $\pi_\lambda = (1-\lambda)\pi^* + \lambda\pi$. Then, we have

$$F \ge \widetilde{J}(\pi_\lambda) = J_1(\pi_\lambda, q_\lambda) \ge (1-\lambda)J_1(\pi^*, q_\lambda) + \lambda J_1(\pi, q_\lambda) \ge (1-\lambda)F + \lambda J_1(\pi, q_\lambda)\,.$$

This implies, $J_1(\pi, q_\lambda) \le F$ for all $\lambda > 0$. From (49) we have:

$$F \ge J_1(\pi, q_\lambda) = \widetilde{J}(\pi) + D(\pi_Y \| Q_\lambda)\,.$$

Taking limit as $\lambda \to 0$ and using lower-semicontinuity of divergence

$$\lim_{\lambda \to 0} D(\pi_Y \| Q_\lambda) \ge D(\pi_Y \| \pi_Y^*)$$

results in

$$F \ge \widetilde{J}(\pi) + D(\pi_Y \| \pi_Y^*) = J_1(\pi, q^*)\,,$$

where the last step is by (49). This completes the proof of (50).

To prove uniqueness of q^* suppose there is $\widetilde{q}^*$ (density) and $\widetilde{Q}^*$ (measure) such that $\sup_\pi J_1(\pi, \widetilde{q}^*) = F$. Plugin $\pi = \pi^*$ in this identity and observe:

$$F \ge J_1(\pi^*, \widetilde{q}^*) = \widetilde{J}(\pi^*) + D(\pi_Y^* \| \widetilde{Q}^*) = F + D(\pi_Y^* \| \widetilde{Q}^*)\,,$$

implying $\pi_Y^* = \widetilde{Q}^*$. ■

With these preparations we proceed to the main subject of this section.

Proof [Proof of Theorem 3] Identity (47) implies that

$$F(\Theta, \Phi) = \inf_q \sup_{\pi \in \mathrm{dom}\,\widetilde{J}} J_1(\pi, q)\,.$$

From (49) we have then

$$\widetilde{F} \triangleq \sup_{\pi \in \mathrm{dom}\,\widetilde{J}} \widetilde{J}(\pi) = \sup_\pi \inf_q J_1(\pi, q) \le \inf_q \sup_\pi J_1(\pi, q) = F(\Theta, \Phi) < \infty\,. \tag{51}$$

Thus, consider any sequence π'_k such that $\widetilde{J}(\pi'_k) \nearrow \widetilde{F}$. Let us now denote $\Pi_k = \mathrm{co}(\pi'_1, \dots, \pi'_k)$ and choose

$$\pi_k \in \operatorname*{argmax}_{\pi \in \Pi_k} \widetilde{J}(\pi)\,.$$

Denote the sequence of induced densities $q_k \triangleq f_{\pi_k}$. (Lemma 13 shows such π_k's exist and q_k only depends on Π_k but not π_k.) We will prove the following facts:

1. The sequence of densities q_k converges to a density q^* in the sense of $D(Q_k \| Q^*) \to 0$ (and, thus, in total variation).
2. For every $\pi \in \Pi_k$ we have $J_1(\pi, q^*) \le \widetilde{F}$.
3. $\widetilde{F} = F(\Theta, \Phi)$ and, furthermore,

$$J_1(\pi, q^*) \le F(\Theta, \Phi) \qquad \forall \pi \in \operatorname{dom} \widetilde{J}\,. \tag{52}$$

This proves a convenient characterization (analog of (3))

$$F(\Theta, \Phi) = \sup_{\pi} \widetilde{J}(\pi) \tag{53}$$

and completes the proof of the Theorem. Indeed, from (52) taking supremum over π we obtain optimality of q^*. Had there existed another $\widetilde{q}^*$ with the property (52) then we would have from (49)

$$\widetilde{J}(\pi_k) \le \widetilde{J}(\pi_k) + D(Q_k \| \widetilde{Q}^*) = J_1(\pi, \widetilde{q}^*) \le F(\Theta, \Phi)\,.$$

Since $\widetilde{J}(\pi_k) \nearrow F(\Theta, \Phi)$ we conclude that $D(Q_k \| \widetilde{Q}^*) \to 0$, and thus Q_k converges to $\widetilde{Q}^*$ in total variation. But Q_k converges to Q^* as well, so $\widetilde{Q}^* = Q^*$ and $\widetilde{q}^* = q^*$ μ-almost everywhere.

To prove the first statement, we apply Lemma 13 as follows. Let $F_m = \widetilde{J}(\pi_m)$ and notice since $\pi_k \in \Pi_{k+m}$ for $m \ge 0$ that

$$F_k + D(Q_k \| Q_{k+m}) = J_1(\pi_k, q^*_{k+m}) \le F_{k+m} \le \widetilde{F}\,.$$

Thus, $D(Q_k \| Q_{k+m}) \le \widetilde{F} - F_k$ and $\sup_m D(Q_k \| Q_{k+m}) \to 0$ as $k \to \infty$. This implies that Q_k form a Cauchy sequence in total variation and thus have a limit point Q^*. From lower semicontinuity of divergence we also have $D(Q_k \| Q^*) \le \lim_{m\to\infty} D(Q_k \| Q_{k+m}) \le \widetilde{F} - F_k$ and thus $D(Q_k \| Q^*) \to 0$ as $k \to \infty$.

To prove the second claim, note that by Lemma 13 for any $\pi \in \Pi_k$ we have

$$J_1(\pi, q_m) \le F_m \le \widetilde{F} \qquad \forall m \ge k\,.$$

On the other hand, $J_1(\pi, q_m) = \widetilde{J}(\pi) + D(\pi_Y \| Q_m)$ and taking $m \to \infty$ and applying lower semicontinuity yet again, we get

$$J_1(\pi, q^*) = J(\pi) + D(\pi_Y \| Q^*) \le \widetilde{F}\,. \tag{54}$$

Finally, to prove (52) for an arbitrary π (not necessarily $\in \cup_k \Pi_k$), we can simply reapply the previous argument with $\widehat{\Pi}_k \triangleq \mathrm{co}(\pi, \pi_1, \dots, \pi_k) \supset \Pi_k$, to obtain sequence $\widehat{q}_k \to \widehat{q}^*$. Since $\pi_k \in \Pi_k$, for this new density we have

$$J_1(\pi_k, \widehat{q}^*) \le \widetilde{F}\,.$$

But $J_1(\pi_k, \widehat{q}^*) = \widetilde{J}(\pi_k) + D(Q_k \| \widehat{Q}^*) = F_k + D(Q_k \| \widehat{Q}^*)$, implying $D(Q_k \| \widehat{Q}^*) \to 0$ and, thus, $Q^* = \widehat{Q}^*$, and in particular, (54) holds. Taking supremum over π in (54) and comparing with (51) we get $\widetilde{F} = F(\Theta, \Phi)$. This establishes (52). ■

Appendix F. Proof of Theorem 4

First, we show the following result.

Lemma 14 *Let Φ be such that $C_1(\Phi) < \infty$ and every $P \in \Phi$ satisfies $P \ll \mu$. Let $\Phi_0 = \{P \in \Phi : D(P\|\Theta) < \infty\}$. Then*

$$F(\Theta, \Phi) = F(\Theta, \Phi_0)$$

Proof We only need to show $F(\Theta, \Phi) \le F(\Theta, \Phi_0)$. To that end, fix $\epsilon > 0$ and consider any q such that

$$\sup_{P\in\Phi}\sup_{\theta\in\Theta}\mathbb{E}_P[\log\frac{f_\theta}{q}] \le F(\Theta, \Phi_0) + \epsilon\,.$$

In addition, denote $q_1 = \frac{dQ_1}{d\mu}$ – density of the (unique) distribution Q_1 attaining the minimum

$$\min_{Q_1}\sup_{P\in\Phi} D(P\|Q_1) = C_1(\Phi) < \infty\,.$$

Similarly, for any $P \in \Phi$ we denote by $f_P = \frac{dP}{d\mu}$. For any θ and P we have

$$\mathbb{E}_P[\log\frac{f_\theta}{q_1}] = D(P\|Q_1) - D(P\|P_\theta)\,.$$

Indeed, almost surely (with convention (34)) we have $\log\frac{f_\theta}{q_1} = \log\frac{f_P}{q_1} - \log\frac{f_P}{f_\theta}$. Denoting $q_\lambda = \lambda q_1 + (1-\lambda)q$, we have

$$\mathbb{E}_P[\log\frac{f_\theta}{q_\lambda}] \le \log\frac{1}{\lambda} + \mathbb{E}_P[\log\frac{f_\theta}{q_1}] = \log\frac{1}{\lambda}D(P\|Q_1) - D(P\|P_\theta)\,.$$

Thus, for any $P \notin \Phi_0$ the above evaluates to $-\infty$. On the other hand, we have

$$\mathbb{E}_P[\log\frac{f_\theta}{q_\lambda}] \le \log\frac{1}{1-\lambda} + \mathbb{E}_P[\log\frac{f_\theta}{q}]$$

And thus taking supremum over $P \in \Phi$ and $\theta \in \Theta$ we get

$$\sup_{P\in\Phi}\sup_{\theta}\mathbb{E}_P[\log\frac{f_\theta}{q_\lambda}] \le \begin{cases} -\infty, & P \notin \Phi_0 \\ F(\Theta, \Phi_0) + \epsilon + \log\frac{1}{1-\lambda}, & P \in \Phi_0 \end{cases}$$

Taking $\lambda, \epsilon \to 0$ completes the proof. ■

The Theorem follows as a special case of the following result.

Lemma 15 *Let Φ be such that a) $P_\theta \in \Phi$ for all $\theta \in \Theta$, b) $P \ll \mu$ for every $P \in \Phi$, c) $C_1(\Phi) < \infty$. Then, we have for any $\epsilon > 0$ such that $\lambda_0 = \frac{C_1(\Phi)}{\epsilon} < 1$*

$$F(\Theta, \Phi) \le F(\Theta, \Theta_\epsilon) + \frac{h(\lambda_0)}{1-\lambda_0}\,.$$

Proof By preceding Lemma, we can assume further that $D(P\|\Theta) < \infty$ for all $P \in \Phi$. Therefore, Theorem 3 applies, and in particular (see (53)):

$$F(\Theta, \Phi) = \sup_{\pi} \widetilde{J}(\pi) \,.$$

Since $\Theta \subset \Phi$, we have that $F(\Theta, \Phi) \ge C(\Theta) \ge 0$. Thus, there exist π such that $\widetilde{J}(\pi) \ge 0$, and for any such π we have (see (45))

$$0 \le \widetilde{J}(\pi) = I(\phi; Y) - \mathbb{E}[c(\phi, \theta)] \le C_1(\Phi) - \mathbb{E}[c(\phi, \theta)] \,.$$

This implies via Markov inequality that

$$\mathbb{P}[c(\phi, \theta) > \epsilon] \le \frac{C_1(\Phi)}{\epsilon} \,.$$

This means that we can represent $\pi = \lambda \pi_1 + (1-\lambda)\pi_0$ with $\pi_0 = \sum_k p_{0,k} P_k \times \delta_{\theta_k}$ and $P_k \in \Theta_\epsilon$ whenever $p_{0,k} > 0$. Furthermore, $\lambda \le \frac{C_1(\Phi)}{\epsilon}$. By the bound (42) we have

$$\widetilde{J}(\pi) \le h(\lambda) + \lambda \widetilde{J}(\pi_1) + (1-\lambda)\widetilde{J}(\pi_0) \le h(\lambda) + \lambda F(\Theta, \Phi) + (1-\lambda) F(\Theta, \Theta_\epsilon) \,.$$

Taking supremum over π we obtain (after rearranging terms)

$$F(\Theta, \Phi) \le F(\Theta, \Theta_\epsilon) + \frac{h(\lambda)}{1-\lambda} \,.$$

The proof is completed by noticing that $x \mapsto \frac{h(x)}{1-x}$ is increasing on $[0, 1)$. ■

Proof [Proof of Theorem 4] If we apply previous theorem with Φ replaced by $\Phi^{\otimes n}$ and ϵ replaced by $n\epsilon$ we obtain:

$$F_n(\Theta, \Phi^{\otimes n}) \le F_n(\Theta, \Theta_{\epsilon_n}^{\otimes n}) + \frac{h(\lambda_0)}{1-\lambda_0}$$

where $\lambda_0 = \frac{C_n(\Phi)}{n\epsilon} = \frac{\tau_n}{\epsilon}$. If $\epsilon = \epsilon_n \gg \tau_n$ then $\lambda_0 \to 0$ and the proof is complete. ■

F.1. On difference between $C_n(\Theta)$ and $C_n(\Theta_\epsilon)$

As we discussed in (8), the meaning of Theorem 4 is to sandwich the misspecified regret between two well-specified ones: $C_n(\Theta)$ and $C_n(\Theta_\epsilon)$. In this section, we demonstrate by a simple example that the growth rates of $C_n(\Theta)$ and $C_n(\Theta_\epsilon)$ could be very different.

We consider an extension of the Gaussian location model, where a single measurement Y_1 is produced from $\phi = (\phi_0, \ldots)$ as

$$Y_1 = \phi + Z, \qquad Z \sim \mathcal{N}(0, I_\infty) \,.$$

(i.e. each of $Y_1, Y_2, \ldots, Y_n$ is itself an infinite sequence). We will take as Φ the Hilbert brick:

$$\Phi = \{\phi : 0 \le \phi_j \le 2^{-j}, j = 0, \ldots\} = \prod_{j=0}^{\infty} [0, 2^{-j}] \,.$$

(It is known that Φ is compact in ℓ_2, for example). Let $\Theta = \{\phi : \phi_1 = \phi_2 = \cdots = 0\}$.

Denote by $C_n([0, a])$ capacity of the 1D-GLM model. Then, from (14) we have

$$C_n(\Theta) = C_n([0, 1]) = \frac{1}{2} \ln(n/(2\pi)) + o(1)$$

For Φ we have

$$C_n(\Phi) = \sum_{j=0}^{\infty} C_n([0, 2^{-j}])\,.$$

Indeed, since the conditional transformation $\theta \mapsto Y$ has structure of a parallel memoryless channel, the optimization in (3) can be reduced to π that are independent across coordinates of θ, cf. (Polyanskiy and Wu, 2012-2016, Theorem 5.1). (The fact that there are countably-infinite number of coordinates does not cause any complications due to a certain continuity of mutual information: $I(A_1, \ldots ; B_1, \ldots) = \lim_{n\to\infty} I(A^n; B^n)$.)

To compute $C_n(\Phi)$, we note that from a special case of (35) (with $\Theta = [0, a] \subset \mathbb{R}$, $\tau = 0$, $v = 1$) we get

$$\Gamma_n([0, a]) = \log\left(1 + a\sqrt{\frac{n}{2\pi}}\right)\,. \tag{55}$$

Shtarkov (55) we get

$$C_n([0, a]) \le \Gamma_n([0, a]) = \log(a\sqrt{\frac{n}{2\pi}} + 1)\,.$$

We can also show that $\sum_{j=0}^{\infty} \log(a_0 2^{-j} + 1) \asymp (\log a_0)^2$ within absolute constants. Thus, we have

$$C_n(\Phi) = O(\log^2 n)\,.$$

(In fact, the argument below also shows $C_n(\Phi) = \Theta(\log^2 n)$.)

Next, note that $C_n([0, a]) = C_1([0, \sqrt{n}a])$ and from the asymptotics we know that $C_1([0, a]) = \ln(a/(2\pi)) + o(1)$, implying that we always have

$$C_1([0, a]) \ge \ln a - c_1$$

for some constant $c_1 > 0$. Thus, we conclude

$$C_n([0, a]) \ge \ln(\sqrt{nc_2}a)\,,$$

for some $0 < c_2 < 1$.

Observe that

$$\sum_{j\ge k} (2^{-j})^2 = \frac{3}{4} 4^{-k}\,.$$

Therefore, any Θ_ϵ always contains a sub-brick:

$$[0, 1] \times \{0\} \cdots \times \{0\} \times [0, 2^{-k_1}] \times [0, 2^{-k_1-1}] \cdots\,,$$

where k_1 is minimal such that

$$\frac{3}{4} 4^{-k_1} \le \frac{1}{2}\epsilon\,.$$

That is $k_1 = \frac{1}{2}\log_2 \frac{1}{\epsilon} + O(1)$. Thus, we get that

$$C_n(\Theta_\epsilon) \ge C_n(\Theta) + \sum_{i \ge k_1} C_n([0, 2^{-i}]) \tag{56}$$

$$\ge C_n(\Theta) + \sum_{j \ge 0} \max(\ln(\sqrt{nc_2}2^{-k_1-j}), 0) \tag{57}$$

$$\ge C_n(\Theta) + c(\ln(n\epsilon))^2\,, \tag{58}$$

for some constant $c > 0$. Overall, we see that for any $\epsilon > 0$ the order of $C_n(\Theta) \asymp \log n$ whereas $C_n(\Theta_\epsilon) \asymp \log^2 n$.

Appendix G. Open questions

- Are there examples where $C_n \ll F_n^{(PAC)} \asymp \Gamma_n < \infty$? What about $C_n \ll F_n^{(PAC)} \ll \Gamma_n$?
- Assumptions in our Theorem 3 unfortunately rule out the PAC case of $\Phi = \mathcal{P}_{iid}(\mathcal{Y}^n)$. Can we extend it to this case? Even in the special case of the setting of Theorem 2, can we prove existence and uniqueness of the minimizer?
- In Theorem 2 what is the order of the difference between $F_n^{(PAC)}$ and C_n?
- In the context of Theorem 2 can it be shown that no Bayes mixture is able to achieve optimal $F_n^{(PAC)} + o(1)$ regret? (Perhaps it can even be shown that no estimator with Gaussian tails can do so.)
- Extend Theorem 2 to (a) exponential families, (b) general smooth families.
- **(closed)** Consider $\mathcal{Y} = \mathbb{Z}_+$ and $\{P_\theta\}$ to be the class of all distributions on $\mathcal{Y}$ with first moment bounded by 1. It is easy to show that $C_n < \infty$, while $\Gamma_n = F_n^{(PAC)} = \infty$ (Shtarkov sum is unbounded). We note, cf. Jia et al. (2021), that restricting Φ_n to a subset $\Phi_n = \{P^{\otimes n} : D(P\|\Theta) < \infty\}$, still results in $F_n(\Theta, \Phi_n) = \infty$ (note that with this restriction on the data generating distribution the oracle loss is always finite, although unbounded). We also mention that for this model class the results of Boucheron et al. (2008) show $C_n = \omega(n^\alpha)$ for any $\alpha < 1/2$, and a more detailed analysis Jia et al. (2021) shows $C_n = \widetilde{\Theta}(\sqrt{n})$.