Regret Bounds for Log-loss via Bayesian Algorithms

Changlong Wu, Mohsen Heidari, Ananth Grama and Wojciech Szpankowski

Abstract

We study sequential probability assignment in the context of online learning under logarithmic loss and obtain tight lower and upper bounds for sequential minimax regret. Sequential minimax regret is defined as the minimum excess loss over data horizon T that a predictor incurs over the best expert in a class, when the samples are presented sequentially and *adversarially*. Our upper bounds are established by applying Bayesian averaging over a novel "smooth truncated covering" of the expert class. This allows us to obtain tight (minimax) upper bounds that subsume the best known non-constructive bounds in an algorithmic fashion. For lower bounds, we reduce the problem to analyzing the fixed design regret via a novel application of Shtarkov sum adapted to online learning. We demonstrate the effectiveness of our approach by establishing tight regret bounds for a wide range of expert classes. In particular, we fully characterize the regret of generalized linear function with *worst* Lipschitz transform functions when the parameters are restricted to a unit norm ℓ_s ($s \geq 2$) ball of dimension d. We show that the regret grows as $\Theta(d \log T)$ when $d \leq O(T^{s/(s+1)-\epsilon})$ for all $\epsilon > 0$ (with precise constant 1 when $d \leq e^{o(\log T)}$) and $\tilde{O}(T^{s/(s+1)})$ when $d \geq \Omega(T^{s/(s+1)})$. Finally, we show that the Bayesian approach may not always be optimal if the support of the prior is included in the reference class itself.

Index Terms

Sequential probability assignment, Online Regression, Logarithmic Loss, Bayesian Algorithm, Shtarkov Sum

I. Introduction

Sequential probability assignment is a fundamental problem in information theory, portfolio optimization, and machine learning. The problem is formulated as a game between Nature and predictor. At each time step t, the predictor attempts to predict a distribution μ_t over a set $\mathcal{Y}$ of Nature's next possible outcome (i.e., label). Nature reveals the true label y_t after the predictor has made a prediction. The predictor incurs a loss that depends on the prediction μ_t and Nature's outcome y_t. The goal of the predictor is to minimize its cumulative loss over a finite time-step horizon T. A commonly used measure of loss is *logarithmic* loss, defined as $-\log(\mu_t(y_t))$, i.e., the negative logarithm of the probability assigned to the true label. Logarithmic loss is both a natural metric for real machine learning tasks [1], [2], [3], [4] and also deeply connected to universal compression in information theory [5], [6], [7], [8], [9], [10], [11].

In practice, minimization of cumulative loss by itself is not very informative without knowing the generating mechanism underlying Nature. Hence, one must compare it to some constant comparator to assess relative loss. In this context, information theory and machine learning communities introduced the notion of *regret* defined as the difference between the cumulative loss incurred by the predictor and the minimal cumulative loss incurred by the best *expert* in an expert class $\mathcal{H}$, without making any assumption regarding the generating mechanism underlying Nature. Furthermore, to obtain fundamental limits for individual sequences one rather considers *minimax* regret in which the regret is estimated for the worst sequence and the best prediction rule. In this paper, we study sequential (agnostic) minimax regret for a broad class of experts $\mathcal{H}$. We address the challenging question of how minimax regret depends on the structure of the class of experts $\mathcal{H}$. Observe that when the class of experts is interpreted as a

This paper was presented in part at ISIT2022.

All authors are with the Center for Science of Information (CSoI). C. Wu (email:wuchangl@hawaii.edu), A. Grama (email: ayg@cs.purdue.edu), and W. Szpankowski (email: szpan@purdue.edu) are with Purdue University. M. Heidari (email: mohsenhd@umich.edu) is with Indiana University.

This work was partially supported by the NSF Center for Science of Information (CSoI) Grant CCF-0939370, by NSF Grants CCF-2006440, CCF-2007238, and CCF- 2211423, and in addition by Google Research Grant and by Rolls Royce.

class of sources (distributions), we fall under universal compression studied extensively in information theory [5], [6], [7], [8], [9], [10], [11].

In online learning studied in this paper, the situation is more complicated, since one must also to deal with the *features*, known in information theory as *side information*. Let $\mathcal{X}$ be the set of features, we model experts as functions $\mathcal{X} \to \mathcal{D}(\mathcal{Y})$, where $\mathcal{D}(\mathcal{Y})$ is the set of all probability measures over $\mathcal{Y}$. For the clarity of presentation, throughout the paper, we assume that $\mathcal{Y} = \{0, 1\}$ and the distributions over $\mathcal{Y}$ is specified as a real number in $\hat{\mathcal{Y}} = [0, 1]$, which is interpreted as the probability assigned on label 1. However, we should emphasis that our results also work for a general finite set $\mathcal{Y}$.

More precisely, experts are modeled as a set of functions $\mathcal{H} \subset [0,1]^{\mathcal{X}}$, the predictor at each time t is a function $\phi_t : \mathcal{X}^t \times \mathcal{Y}^{t-1} \to [0, 1]$ and the logarithmic loss can be written as:

$$\ell(\hat{y}_t, y_t) = -y_t \log(\hat{y}_t) - (1 - y_t) \log(1 - \hat{y}_t). \tag{1}$$

The *pointwise regret* is defined as:

$$R(\hat{y}^T, y^T, \mathcal{H}|\mathbf{x}^T) = \sum_{t=1}^{T} \ell(\hat{y}_t, y_t) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t), \tag{2}$$

where $\hat{y}_t = \phi_t(y^{t-1}, \mathbf{x}^t)$ is the predicted outcome. Observe that the first term in (2) represents the accumulated loss incurred by the learning algorithm (the predictor), while the second summation deals with the best experts within $\mathcal{H}$ in the hindsight after observing $\mathbf{x}^T$ and y^T. Depending to how sequences $\mathbf{x}^T, y^T$ are generated by Nature, we define the following two notions of *minimax* regrets:

Fixed Design: This point of view studies minimal regret for the worst realization of the label with the feature vector $\mathbf{x}^T$ known in advance (also known as *transduction online learning* in the literature). Let $\phi_t, t > 0$ be the strategies of the predictor. The *fixed design minimax regret* for a given $\mathbf{x}^T$ is defined as [12]:

$$r_T^*(\mathcal{H}|\mathbf{x}^T) = \inf_{\phi^T} \sup_{y^T} R(\phi^T, y^T, \mathcal{H}|\mathbf{x}^T). \tag{3}$$

Further, the fixed design *maximal* minimax regret is:

$$r_T^*(\mathcal{H}) = \sup_{\mathbf{x}^T} \inf_{\phi^T} \sup_{y^T} R(\phi^T, y^T, \mathcal{H}|\mathbf{x}^T). \tag{4}$$

When the side information/features are known in advance, we can view the set of experts as a set of sources (i.e., probability distributions over $\mathcal{Y}^T$). In this case, we can show [13] that minimax regret $r_T^*(\mathcal{H}|\mathbf{x}^T)$ is completely characterized by the so called *Shtarkov sum* introduced in Shtarkov's seminal work [14]. However, we notice that Shtarkov's characterization only works for the *static* or *simulatable* [1] experts, i.e., the prediction made by the experts can only depend on the past labels. A more challenging (and more realistic) scenario is when the features are revealed sequentially as well, as defined in the sequel.

Sequential Design: In this paper we primarily focus on *sequential* or *agnostic* minimax regret in which the optimization on regret is performed at each time t without knowing in advance $\mathbf{x}^T$ or y^T. Then the *sequential (maximal) minimax regret* is: [15]

$$r_T^a(\mathcal{H}) = \sup_{\mathbf{x}_1} \inf_{\hat{y}_1} \sup_{y_1} \cdots \sup_{\mathbf{x}_T} \inf_{\hat{y}_T} \sup_{y_T} R(\hat{y}^T, y^T, \mathcal{H}|\mathbf{x}^T). \tag{5}$$

It is not hard to show that $r_T^a(\mathcal{H}) \geq r_T^*(\mathcal{H})$ for all $\mathcal{H}$, moreover, $r_T^a(\mathcal{H})$ can be exponentially larger (i.e., T vs $\log T$) than $r_T^*(\mathcal{H})$ for certain $\mathcal{H}$, see [12] and Proposition 1. However, we will show in this paper that $r_T^*(\mathcal{H})$ can actually match $r_T^a(\mathcal{H})$ for many *natural* classes, arguing that $r_T^*(\mathcal{H})$, and its Shtarkov sum representation, can be a powerful technique for deriving tight lower bounds for $r_T^a(\mathcal{H})$.

Our main goal in this paper is to gain insights into the growth of sequential regret $r_T^a(\mathcal{H})$ for various classes $\mathcal{H}$, and to show how the structure of $\mathcal{H}$ impacts the precise growth of regret. We summarize our main contributions as follows:

A. Summary of Results

We make two major contributions in this paper. We provide a general *algorithmic* framework for obtaining tight regret upper bounds for $r_T^a(\mathcal{H})$ using the concept of (global) *sequential covering* (that was implicitly used in [16], [15] for general convex and Lipschitz losses, e.g., absolute loss) together with a novel *Smooth Truncated Bayesian Algorithm* (Algorithm 2). The main technical contribution of this part is a smooth truncation approach (see Lemma 4) that allows us to resolve the difficulty of unboundedness and non-Lipschitz of log-loss as in [2]. Our first result (Theorem 1) establishes a general tight upper bound on $r_T^a(\mathcal{H})$ that subsumes the state-of-the-art bounds in [2], [17] established only *non-constructively* (i.e., via the minimax theorem). Moreover, our upper bound in Theorem 1 provides optimal constants that improve universally the bound in [17]. The optimality of constants also allows us to obtain the tightest bounds with optimal leading constants for special classes. Our second general contribution is a lower bounding technique through fixed design regret $r_T^*(\mathcal{H})$ and its Shtarkov sum, which extends the results in [13], [18], [4] that were only established for Logistic regression.

Beyond general results, we study a broad set of concrete expert classes and derive tight regret (lower and upper) bounds for $r_T^a(\mathcal{H})$ with provable optimal constants. In Section IV, we study a general Lipschitz parametric class $\mathcal{H}$ with parameters restricted to a norm ℓ_s ball in $\mathbb{R}^d$. This includes, e.g., the logistic regression, normal location model and neural networks with Lipschitz activation functions. In particular, we show that regret grows as

$$d\log(T/d) + O(d)$$

for such a class $\mathcal{H}$, where the constant 1 in front of $d\log(T/d)$ is optimal for $d \ll T/\log T$ (see Theorem 2 and 3). Section IV-B extends results to the case when the function $\log f$ has bounded Hessian for any $f \in \mathcal{H}$, showing that the regret is upper bounded as

$$\frac{d}{2}\log(T/d) + O(d)$$

under ℓ_2 ball (see Theorem 4). In Theorem 5 we derive a matching lower bound of the form

$$\frac{d}{2}\log(T/d^{(s+2)/s}) + O(d)$$

for generalized *linear* functions of the form $f(\langle \mathbf{w}, \mathbf{x}\rangle)$ with $f'(0) \neq 0$, $\mathbf{w}, \mathbf{x}$ in a unit ℓ_s ball and $d \ll T^{s/(s+2)}$, where $\mathbf{w}$ represents a d dimensional weight vector. This result recovers all the lower bounds in [4] obtained for logistic regression [1] but with full range of $s \geq 1$. In Section IV-C, we show that the leading constant $\frac{1}{2}$ can still be achieved for certain classes with non-Lipschitz $\log f$ (therefore with unbounded Hessian). In particular, this holds for the linear function $|\langle \mathbf{w}, \mathbf{x}\rangle|$ with $\mathbf{w}$ in a ℓ_1 ball. This follows from a general result for classes generated by convex combinations of simple classes established in Theorem 6. In Section V, we study the situation in which data dimension d grows faster than T through the *sequential fat-shattering* number introduced in [15]. In particular, we show that for the linear function $f(\mathbf{w}, \mathbf{x}) = |\langle \mathbf{w}, \mathbf{x}\rangle|$ with $\mathbf{w}$ in a unit ℓ_s ball and $\mathbf{x}$ in a unit $\ell_{s/(s-1)}$ ball (of dimension $d \geq T$), the regret satisfies

$$\Omega(T^{s/(s+1)}) \leq r_T^a(\mathcal{H}_f) \leq \tilde{O}(T^{s/(s+1)}),$$

where $\tilde{O}$ hides poly-logarithmic factors and $s \geq 2$. Section V also provides lower bounds for general Lipschitz type classes. In particular our result (Theorem 9) recovers the lower bounds of [17] for the non-parametric Lipschitz function classes, but with general ℓ_s norms and simpler proof, which may be of independent interest. Finally, in Section VI, we discuss the issue of optimality of Bayesian algorithms and show that there exist finite classes $\mathcal{H}$ such that the best achievable sequential regret is of order $O(\sqrt{\log|\mathcal{H}|})$ but the Bayesian algorithm with *any* prior over $\mathcal{H}$ can only achieve a sequential regret of order $\Omega(\log|\mathcal{H}|)$.

[1]The technique of [4] based on the redundancy capacity theorem also works for other functions with bounded second derivatives, like the probit function.

B. Related Work

a) Regrets in online leanring: A large body of existing work in online regression deals with logistic regression. The work of [19] studies pointwise regret of logistic regression in the *proper* setting. Unlike *improper* learning, investigated in our work, where feature $\mathbf{x}_t$ at time t is also available to the learner, [19] showed that pointwise regret is $\Theta(T^{1/3})$ for $d = 1$ and $O(\sqrt{T})$ for $d > 1$. Furthermore, [20] demonstrates results that the regret for logistic regression grows as $O(d \log T/d)$, and was further generalized in [3]. These results were strengthened in [4], which also provides matching lower bounds. Precise asymptotics for the fixed design minimax regret were recently presented in [13], [18] for $d = o(T^{1/3})$.

Regret bounds under logarithmic loss for general expert class $\mathcal{H}$ was first investigated by Vovk under the framework of mixable losses [21], [22]. In particular, Vovk showed that for finite class $\mathcal{H}$, the regret growth is $\log |\mathcal{H}|$ via the *aggregating algorithm* (i.e., the Bayesian algorithm that we will discuss below). We refer the reader to [1, Chapter 3.5, 3.6] and the references therein for more results on this topic. Cesa-Bianchi and Lugosi [23] were the first to investigate log-loss under general (infinite) expert class $\mathcal{H}$ [1, Chapter 9.10, 9.11], where they derived a general upper bound using the concept of covering number and a two-stage prediction scheme. In particular, Cesa-Bianchi and Lugosi showed that for Lipschitz parametric classes with values bounded away from $\{0, 1\}$, one can achieve a regret bound of the form $d/2 \log(T/d)$. When the values are close to $\{0, 1\}$, they used a *hard* truncation approach that gives a sub-optimal bound of the form $(3/2)d \log(T/d)$ (this is not explicitly shown in [23] but can be derived using their approach). Moreover, the approach of [23] only works for the fixed design regret (or *simulatable* in their context). In [2], the authors extended the result of [23] to the sequential case via the machinery of sequential covering that was established in [24]. However, [2] also used the same *hard* truncation as in [23] resulting in suboptimal upper bounds. In [17], the authors obtained an upper bound similar to the upper bound presented in Theorem 1 using the observation that the $\log$ function is self-concordant. In particular, this allows them to resolve the tight bounds for non-parametric Lipschitz functions mapping $[0, 1]^s \to [0, 1]$. However, their bounds are proved *non-constructively*, i.e., the proof does not provide an algorithm that achieves such bounds. We note also a recent result [25] that uses a similar idea of smoothing for controlling the unboundedness of log-loss, however, the result assumes that features $\mathbf{x}^T$ are presented $i.i.d.$. More importantly, the results in [25] only hold for *average case* regret.

b) Regrets in information theory: The *fixed design* regret is closely related to the *universal compression* and *universal prediction* studied in [5], [6], [7], [8], [9], [10], [11], [26]. Here, there is no feature vector $\mathbf{x}^t$ and the dimension $d = 1$. A sequence y^T is generated by a source P that belongs to a class of sources $\mathcal{S}$, which can be viewed as the reference class $\mathcal{H}$ in online learning. The minimax regret for logarithmic loss is given by [27], [14], [5]

$$r_T^*(\mathcal{S}) = \min_Q \max_{y^T} [-\log Q(y^T) + \log \sup_{P \in \mathcal{S}} P(y^T)],$$

where Q is the universal probability assignment approximating the unknown $P \in \mathcal{S}$. The main question is how the structure of $\mathcal{S}$ impacts the growth of minimax regret. Let m denote the alphabet size (in online learning, we only consider $m = 2$). It is known [5], [6], [7], [8], [9], [10], [11] that for Markov sources of order r, regret grows as $\frac{m^r(m-1)}{2} \log T$ for fixed m [7], [6], [8], [28]. In [28], minimax regret was analyzed for all ranges of m and T. For non-Markovian sources, the growth is super logarithmic. For example, for renewal sources of order r, regret is $\Theta(T^{r/(r+1)})$ [29], and the precise constant in front of the leading term is known for $r = 1$ [30]. We note that [31], [32] studied the general classes of densities smoothly parameterized by d-dimensional data to obtain general results for minimax regret that can be phrased as online regret. Notably, [33] studied the case with ℓ_1 controls on the parameters and derived a $\sqrt{T \log(1 + d/T)}$ type regret bound. However, such results are still under the regime of fixed design case. In [34] the authors studied smooth parametric sources in the sequential case using a similar Bayesian averaging approach as in our work. However, the result of [34] only holds when the dimension of the parameter is a constant (i.e., does not scale as T). Sequential prediction under log-loss with side information has also being investigated in [35], [36], [37], [38], notably [36], [35], [37] consider the scenario where the features $\mathbf{x}^T$ are generate by an $i.i.d.$ source. However, such results heavily rely on the underlying statistical mechanism for generating data, while in our paper the selection of $\mathbf{x}^T, y^T$ are completely general and can even be adversarial.

We note that [39] considers an intermediate scenario between full adversary and well-specified case, but without considering the side information. Furthermore, in [40] the authors provide justifications of selecting the log-loss in statistical inference problems. The optimality of Bayesian algorithms has also being studied in [41], but only for the *simulatable* case.

The current work is a substantial extension of our preliminary results [42].

II. Notation and Preliminaries

We denote $\mathcal{X}$ as the input feature space and $\mathcal{H}$ as the expert class, which is a class of functions mapping $\mathcal{X} \to [0,1]$. Throughout the paper, we use $\hat{\mathcal{Y}} = [0,1]$ and $\mathcal{Y} = \{0,1\}$ to denote the prediction and label spaces respectively. We often use an auxiliary set $\mathcal{W}$ to index $\mathcal{H}$. If f is a function mapping $\mathcal{W} \times \mathcal{X} \to [0,1]$, we define the class induced by f as:

$$\mathcal{H}_f = \{f(\mathbf{w}, \cdot) \in [0,1]^{\mathcal{X}} : \mathbf{w} \in \mathcal{W}\}.$$

We say a function g is *sequential* if it maps $\mathcal{X}^* \to [0,1]$, where $\mathcal{X}^*$ is set of all finite sequences with elements in $\mathcal{X}$. We denote $\mathcal{G}$ as a class of *sequential* functions. Note that any function mapping $\mathcal{X} \to [0,1]$ can be viewed as a special case of sequential function. If T is a time horizon, then for any $t \in [T]$, we write $\mathbf{x}^t = \{\mathbf{x}_1, \cdots, \mathbf{x}_t\}$ and $y^t = \{y_1, \cdots, y_t\}$. We use standard asymptotic notation $f(t) = O(g(t))$ if there exists a constant C such that $f(t) \le Cg(t)$ for sufficient large t and $f(t) = \Omega(g(t))$ if $g(t) = O(f(t))$. We denote $f(t) \ll g(t)$ or $f(t) = o(g(t))$ if $\limsup_{t\to\infty} f(t)/g(t) = 0$ and $f(t) \gg g(t)$ if $g(t) \ll f(t)$. We also assume that the function $\log(x)$ is the natural logarithm (base e).

The main objective of this paper is to study the growth of sequential minimax regret $r_T^a(\mathcal{H})$ for a large class of experts $\mathcal{H}$. We accomplish it using two different techniques. For lower bound, we precisely estimate fixed design minimax regret $r_T^*(\mathcal{H}|\mathbf{x}^T)$ using the Shtarkov sum [14], discussed next. For the upper bound, we construct novel global covering sets $\mathcal{G}$ of $\mathcal{H}$, and design a new (truncated) Bayesian algorithm to find precise bounds with constants that are provably optimal.

a) Lower Bounds: We investigate the lower bound of the sequential regret $r_T^a(\mathcal{H})$ by considering its corresponding fixed design minimax regret $r_T^*(\mathcal{H}|\mathbf{x}^T)$ and $r_T^*(\mathcal{H}) = \sup_{\mathbf{x}^T} r_T^*(\mathcal{H}|\mathbf{x}^T)$. We are able to do this using the recent result [12], which we quote next:

Proposition 1 (Wu et al., 2022)**.** *Let $\mathcal{H}$ be any general hypothesis class and ℓ be any loss function. Then*

$$r_T^a(\mathcal{H}) \ge r_T^*(\mathcal{H}).$$

In particular, for logarithmic loss ℓ and threshold functions $\mathcal{H} = \{1\{w \ge x\} : w, x \in [0,1]\}$, we have $r_T^a(\mathcal{H}) = T$ but $r_T^*(\mathcal{H}) = \log(T+1)$.

We establish precise growth of $r_T^*(\mathcal{H})$ by estimating the Shtarkov sum that has been analyzed in information theory [14], [5] and recently applied to online learning [43], [13]. For logarithmic loss, the Shtarkov sum (conditioned on $\mathbf{x}^T$) is defined as follows [2]

$$S_T(\mathcal{H}|\mathbf{x}^T) \stackrel{\text{def}}{=} \sum_{y^T \in \{0,1\}^T} \sup_{h\in\mathcal{H}} P_h(y^T \mid \mathbf{x}^T), \tag{6}$$

where $P_h(y^T \mid \mathbf{x}^T) = \prod_{t=1}^T h(\mathbf{x}_t)^{y_t}(1-h(\mathbf{x}_t))^{1-y_t}$ and we *interpret* $h(\mathbf{x}_t) = P(y_t = 1|\mathbf{x}_t)$. The regret can be expressed in terms of the Shtarkov sum (see [13, Equation (6)] or [1, Theorem 9.1]) as:

$$r_T^*(\mathcal{H}) = \sup_{\mathbf{x}^T} \log S_T(\mathcal{H}|\mathbf{x}^T). \tag{7}$$

This allows us to reduce the lower bounds of $r_T^*(\mathcal{H})$ to the lower bounds of $S_T(\mathcal{H}|\mathbf{x}^T)$. For some cases, one can lower bound $S_T(\mathcal{H}|\mathbf{x}^T)$ analytically, see e.g., Theorem 5 and 7. However, in most of the cases, an analytic formula

[2]Note that the Starkov sum can be defined for any class of measures, however, here we only use the form for product measures.

is not available (or hard to derive). We deal with such cases using a non-parametric approach, as established in Lemma 1 below [3].

Lemma 1. *Let $\mathcal{P}$ be a finite class of distributions over the same domain Ω. Denote*

$$S = \sum_{\omega \in \Omega} \max_{p \in \mathcal{P}} p(\omega)$$

to be the Shtarkov sum. Then for any estimation rule $\Phi : \Omega \to \mathcal{P}$ we have:

$$S \geq |\mathcal{P}| \cdot \left(1 - \max_{p \in \mathcal{P}} p\left(\{\omega : \Phi(\omega) \neq p\}\right)\right)$$

Proof. Note that Φ partitions Ω into $|\mathcal{P}|$ disjoint parts. For any $p \in \mathcal{P}$, we denote $\Omega_p = \{\omega \in \Omega : \Phi(\omega) = p\}$ to be the partition corresponding to p. We have:

$$\sum_{p \in \mathcal{P}} p(\Omega_p) = \sum_{\omega \in \Omega} p_\omega(\omega) \leq \sum_{\omega \in \Omega} \max_{p \in \mathcal{P}} p(\omega) = S,$$

where $p_\omega \in \mathcal{P}$ is the distribution such that $\omega \in \Omega_{p_\omega}$. This implies

$$\min_{p \in \mathcal{P}} p(\Omega_p) \leq \frac{S}{|\mathcal{P}|}.$$

The result follows by taking the complements of Ω_p. □

We note that the crucial part of applying this approach is selection of features $\mathbf{x}^T$ that maximize $S_T(\mathcal{H}|\mathbf{x}^T)$, which is generally non-trivial, see Section V.

b) Upper Bounds: We now discuss our constructive approach to upper bounds. In the next section, we present our Smooth truncated Bayesian Algorithm (Algorithm 2) that provides a constructive and often achievable upper bound. Here we summarize some, mostly known, preliminaries.

Let $\mathcal{G}$ be any class of functions mapping $\mathcal{X}^* \to [0, 1]$. Let $\mathcal{W}$ be an index set of $\mathcal{G}$ and μ be an arbitrary finite measure over $\mathcal{W}$. The standard Bayesian predictor with prior μ is presented in Algorithm 1. Based on this algorithm, we can prove following two lemmas (see e.g., [1, Chapter 3.3]) that are used to establish most of the upper bounds in this paper. For completeness, we provide simple proofs in Appendix A.

Algorithm 1 Bayesian predictor

Input: Reference class $\mathcal{G} := \{g_w : w \in \mathcal{W}\}$ with index set $\mathcal{W}$ and prior μ over $\mathcal{W}$

1: Set $p_w(y^0 \mid \mathbf{x}^0) = 1$ for all $w \in \mathcal{W}$.
2: **for** $t = 1, \cdots, T$ **do**
3: Receive feature vector $\mathbf{x}_t$
4: Make prediction with the following equation:

$$\hat{y}_t = \frac{\int_{\mathcal{W}} g_w(\mathbf{x}^t) p_w(y^{t-1} \mid \mathbf{x}^{t-1}) \mathrm{d}\mu}{\int_{\mathcal{W}} p_w(y^{t-1} \mid \mathbf{x}^{t-1}) \mathrm{d}\mu}.$$

5: Receive label y_t
6: For all $w \in \mathcal{W}$, update: $p_w(y^t \mid \mathbf{x}^t) = e^{-\ell(g_w(\mathbf{x}^t), y_t)} p_w(y^{t-1} \mid \mathbf{x}^{t-1})$.
7: **end for**

Lemma 2. *Let $\mathcal{G}$ be a class of functions $g_w : \mathcal{X}^* \to [0, 1], w \in \mathcal{W}$. Let $\hat{y}_t$ be the Bayesian prediction rule as in Step 4 of Algorithm 1 with prior μ. Then, for any $\mathbf{x}^T$ and y^T we have*

$$\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) \leq -\log \frac{\int_{\mathcal{W}} p_w(y^T \mid \mathbf{x}^T) \mathrm{d}\mu}{\int_{\mathcal{W}} 1 \mathrm{d}\mu},$$

[3]Note that Lemma 1 is conceptually similar to the redundancy-capacity theorem based approach as in [4], however, we are bounding the Shtarkov sum here directly without going through the Fano's inequality.

where $p_w(y^T \mid \mathbf{x}^T) = e^{-\sum_{t=1}^T \ell(g_w(\mathbf{x}^t), y_t)}$ *and* ℓ *is the log-loss as in equation (1).*

The following lemma bounds the regret under log-loss of finite classes, which is well known.

Lemma 3. *For any finite class of functions* $\mathcal{G}$*, we have* $r_T^a(\mathcal{G}) \leq \log |\mathcal{G}|$.

We note that the crucial part of applying the Bayesian approach is to appropriately design a class $\mathcal{G}$ that *covers* $\mathcal{H}$. This will be the main theme of our work on deriving upper bounds on $r_T^a(\mathcal{H})$. We also discuss the strength of this approach in Section VI.

III. Main Result and Bayesian Averaging

We first introduce a notation of covering set called the *global sequential cover*, which was implicitly used in [15, Section 6.1] for deriving regret bounds under absolute loss and dates back to the ideas in [16].

Definition 1 (Global sequential covering)**.** *For any* $\mathcal{H} \subset [0,1]^{\mathcal{X}}$*, we say that class* $\mathcal{G}$ *of functions mapping* $\mathcal{X}^* \to [0,1]$ *is a global* sequential α*-covering of* $\mathcal{H}$ *at scale* α *if for any* $\mathbf{x}^T \in \mathcal{X}^T$ *and* $h \in \mathcal{H}$*, there exists* $g \in \mathcal{G}$ *such that* $\forall t \in [T]$,

$$|h(\mathbf{x}_t) - g(\mathbf{x}^t)| \leq \alpha.$$

Throughout we assume that $0 \leq \alpha \leq 1$.

Note that global sequential covering defined in Definition 1 is different from the (local) sequential covering used in [17] (and originally from [15]), since our covering functions *do not* depend on the underlying trees in the definition of [15]. This is crucial to apply our covering set directly in an algorithmic way (see Algorithm 2). Note that, global sequential covering subsumes the classical notion of uniform covering, where the covering function does not depend on past observations. However, global sequential covering can be substantially smaller than the uniform cover. We discuss this in Section V. Note that the definition of sequential covering can be extended to sequential experts as well (simply replace $h(\mathbf{x}_t)$ with $h(\mathbf{x}^t)$). More generally, one can also consider the case when the experts could depend on both the past features and labels. We refer to Appendix C for more discussions.

A. General Upper Bounds via Smoothing.

We are now in position to state our first main result:

Theorem 1. *If for any* $\alpha \geq 0$ *there exists a global sequential* α*-covering set* $\mathcal{G}_\alpha$ *of* $\mathcal{H}$*, then*

$$r_T^a(\mathcal{H}) \leq \inf_{0 \leq \alpha \leq 1} \left\{ T \log(1+2\alpha) + \log |\mathcal{G}_\alpha| \right\} \leq \inf_{0 \leq \alpha \leq 1} \left\{ 2\alpha T + \log |\mathcal{G}_\alpha| \right\}, \tag{8}$$

and this bound is achived by using Algorithm 2 over $\mathcal{G}_\alpha$ *with truncation parameter* α *and uniform prior.*

We point out that Theorem 1 also improves the upper bounds of [17] by obtaining better constants in front of both αT and $\log |\mathcal{G}_\alpha|$ (i.e., from $(4,4)$ to $(2,1)$), while the results of [17] are only proved non-constructively. The proof of Theorem 1 is based on the following key lemma.

Lemma 4. *Suppose* $\mathcal{H}$ *has a global sequential* α*-covering set* $\mathcal{G}$ *for some* $\alpha \in [0,1]$*. Then, there exists a truncated set* $\tilde{\mathcal{G}}$ *of* $\mathcal{G}$ *with* $|\tilde{\mathcal{G}}| = |\mathcal{G}|$ *such that for all* $\mathbf{x}^T, y^T$ *and* $h \in \mathcal{H}$ *there exists a* $\tilde{g} \in \tilde{\mathcal{G}}$ *satisfying*

$$\frac{p_h(y^T \mid \mathbf{x}^T)}{p_{\tilde{g}}(y^T \mid \mathbf{x}^T)} \leq (1+2\alpha)^T, \tag{9}$$

where

$$p_h(y^T \mid \mathbf{x}^T) = \prod_{t=1}^T h(\mathbf{x}_t)^{y_t}(1-h(\mathbf{x}_t))^{1-y_t} \quad \text{and} \quad p_{\tilde{g}}(y^T \mid \mathbf{x}^T) = \prod_{t=1}^T \tilde{g}(\mathbf{x}^t)^{y_t}(1-\tilde{g}(\mathbf{x}^t))^{1-y_t}.$$

Algorithm 2 Smooth truncated Bayesian predictor

Input: Reference class $\mathcal{G}$ with index set $\mathcal{W}$ and prior μ over $\mathcal{W}$, and truncation parameter α

1: Let $p_w(y^0 \mid \mathbf{x}^0) = 1$ for all $w \in \mathcal{W}$
2: **for** $t = 1, \cdots, T$ **do**
3: Receive feature $\mathbf{x}_t$
4: For all $w \in \mathcal{W}$, set

$$\tilde{g}_w(\mathbf{x}^t) = \frac{g_w(\mathbf{x}^t) + \alpha}{1 + 2\alpha}$$

5: Make prediction

$$\hat{y}_t = \frac{\int_{\mathcal{W}} \tilde{g}_w(\mathbf{x}^t) p_w(y^{t-1} \mid \mathbf{x}^{t-1}) \mathrm{d}\mu}{\int_{\mathcal{W}} p_w(y^{t-1} \mid \mathbf{x}^{t-1}) \mathrm{d}\mu}$$

6: Receive label y_t
7: For all $w \in \mathcal{W}$, update: $p_w(y^t \mid \mathbf{x}^t) = e^{-\ell(\tilde{g}_w(\mathbf{x}^t), y_t)} p_w(y^{t-1} \mid \mathbf{x}^{t-1})$.
8: **end for**

Proof. We construct the set $\tilde{\mathcal{G}}$ as in Algorithm 2. For any $g \in \mathcal{G}$, we define a smooth truncated function $\tilde{g}$ such that for any $\mathbf{x}^t \in \mathcal{X}^*$

$$\tilde{g}(\mathbf{x}^t) = \frac{g(\mathbf{x}^t) + \alpha}{1 + 2\alpha}.$$

We introduce the following short-hand notation; for any function f, we define $f(y_t) = f(\mathbf{x}^t)^{y_t}(1 - f(\mathbf{x}^t))^{1-y_t}$. For any $\mathbf{x}^T, y^T$, and $h \in \mathcal{H}$, let $g \in \mathcal{G}$ be a α-covering of h and $\tilde{g}$ be the truncated function as defined above. Now, the key observation is that for any $y_t \in \{0, 1\}$, we have $h(y_t) \le g(y_t) + \alpha$ since g α-covers h. This implies that

$$\frac{h(y_t)}{\tilde{g}(y_t)} \le \frac{g(y_t) + \alpha}{(g(y_t) + \alpha)/(1 + 2\alpha)} = 1 + 2\alpha.$$

Therefore, we have

$$\begin{aligned} \frac{p_h(y^T \mid \mathbf{x}^T)}{p_{\tilde{g}}(y^T \mid \mathbf{x}^T)} &= \prod_{t=1}^{T} \frac{h(y_t)}{\tilde{g}(y_t)} \\ &\le (1 + 2\alpha)^T. \end{aligned}$$

This completes the proof of Lemma 4. □

Remark 1. *A different way of understanding Lemma 4 is through the concept of sequential* dominance. *We say a function class $\mathcal{G}$ sequentially α-dominates a class $\mathcal{H}$ if for any $\mathbf{x}^T$, y^T and $h \in \mathcal{H}$ there exists $g \in \mathcal{G}$ such that $\log p_h(y^T \mid \mathbf{x}^T) - \log p_g(y^T \mid \mathbf{x}^T) \le \alpha T$. Lemma 4 proves that if a class $\mathcal{H}$ admits a sequential α-*covering *set, then one can convert the covering set into a sequential 2α-dominate set. The concept of* dominance *is the "right" way of characterizing the log-loss due to the multiplicative nature of log-loss.*

Proof of Theorem 1. We show that for any $0 \le \alpha \le 1$, if an α-covering set $\mathcal{G}_\alpha$ exists, then one can achieve the claimed bound for such an α. To do so, we run the Smooth truncated Bayesian Algorithm (Algorithm 2) on $\mathcal{G}_\alpha$ with uniform prior and truncation parameter α. We denote $\tilde{\mathcal{G}}_\alpha$ to be the truncated class of $\mathcal{G}_\alpha$ as in Lemma 4 (same as the step 4 of Algorithm 2). We now fix $\mathbf{x}^T, y^T$. By Lemma 3 (with $\mathcal{G}$ being $\tilde{\mathcal{G}}_\alpha$), we have:

$$\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) \le \inf_{\tilde{g} \in \tilde{\mathcal{G}}_\alpha} \sum_{t=1}^{T} \ell(\tilde{g}(\mathbf{x}^t), y_t) + \log |\tilde{\mathcal{G}}_\alpha| = \inf_{\tilde{g} \in \tilde{\mathcal{G}}_\alpha} \sum_{t=1}^{T} \ell(\tilde{g}(\mathbf{x}^t), y_t) + \log |\mathcal{G}_\alpha|,$$

the last equality follows from $|\mathcal{G}_\alpha| = |\tilde{\mathcal{G}}_\alpha|$. Since $\sum_{t=1}^{T} \ell(f(\mathbf{x}^t), y_t) = -\log p_f(y^T \mid \mathbf{x}^T)$ for any function f, then

by Lemma 4 we conclude that:

$$\inf_{h\in\mathcal{H}}\sum_{t=1}^{T}\ell(h(\mathbf{x}_t),y_t)\geq\inf_{\tilde{g}\in\tilde{\mathcal{G}}_\alpha}\sum_{t=1}^{T}\ell(\tilde{g}(\mathbf{x}^t),y_t)-T\log\left(1+2\alpha\right).$$

The result follows by combining the inequalities and noticing that $\log(1+x)\leq x$ for all $x\geq -1$. □

We further note that for any constants c_1, c_2 for which the bound $r_T^a(\mathcal{H})\leq c_1\alpha T + c_2\log|\mathcal{G}_\alpha|$ holds universally, we must have $c_1\geq 2$ and $c_2\geq 1$. Therefore, Theorem 1 achieves the optimal constants in the form $c_1\alpha T+c_2\log|\mathcal{G}_\alpha|$. To see this, we let $\mathcal{X}=[T]$ and define g to be the function that maps every $t\in[T]$ to $\frac{1}{2}$. Let $\mathcal{H}$ be the class of functions that maps to $[1/2-\alpha, 1/2+\alpha]$. Clearly, $\mathcal{H}$ is α-covered by g. By noting that the maximum probability is $(1/2+\alpha)^T=(1+2\alpha)^T(1/2)^T$, we compute the Shtarkov sum (6) to find:

$$r_T^a(\mathcal{H})\geq r_T^*(\mathcal{H})\geq\log(1+2\alpha)^T\sim 2\alpha T,$$

where $\sim$ holds when α is sufficiently small. This implies that we must have $c_1\geq 2$. The fact that $c_2\geq 1$ is due to the fact that the mixability constant of log-loss is 1, which also follows from Theorem 3 below.

IV. Lipschitz Parametric Classes

We now consider a Lipschitz parametric function class. Given a function $f:\mathcal{W}\times\mathcal{X}\to[0,1]$, define the following class

$$\mathcal{H}_f=\{f(\mathbf{w},\cdot)\in[0,1]^{\mathcal{X}}\ :\ \mathbf{w}\in\mathcal{W}\},$$

where $\mathbf{w}\in\mathcal{W}$ is often a d-dimensional vector in $\mathbb{R}^d$.

We will assume that $f(\mathbf{w},\mathbf{x})$ is L-Lipschitz on $\mathbf{w}$ for every $\mathbf{x}$, where $L\in\mathbb{R}^+$. More formally, $\forall\mathbf{w}_1,\mathbf{w}_2\in\mathcal{W}$ and $\mathbf{x}\in\mathcal{X}$, we require that L-Lipschitz function f satisfy:

$$|f(\mathbf{w}_1,\mathbf{x})-f(\mathbf{w}_2,\mathbf{x})|\leq L||\mathbf{w}_1-\mathbf{w}_2||,$$

where $||\cdot||$ is some norm on $\mathcal{W}$. For example, if we take $\mathcal{W}\subset\mathbb{R}^d$, then the norm can be ℓ_1, ℓ_2 or ℓ_∞ norm. For any specific norm $||\cdot||$, we write $\mathcal{B}(R)$ to be the ball under such norm with radius R in $\mathcal{W}$. In particular, we denote by $\mathcal{B}_s^d(R)$ the ball in $\mathbb{R}^d$ of radius R under ℓ_s norm centered at the origin.

A. Characterization of Lipschitz Class

We start with the following upper bound for *general* Lipschitz parametric classes. We provide here a simple covering based proof, but refer to Appendix E for a proof without relying on construction of a cover.

Theorem 2. *Let $f:\mathcal{B}_s^d(R)\times\mathbb{R}^d\to[0,1]$ be a L-Lipschitz function under ℓ_s norm. Then*

$$r_T^a(\mathcal{H}_f)\leq\min\left\{d\log\left(\frac{2RLT}{d}+1\right)+2d, T\right\}. \tag{10}$$

Proof. By L-Lipschitz condition, to find an α-covering in the sense of Definition 1, we only need to find a covering of $\mathcal{B}_s^d(R)$ with radius α/L under ℓ_s norm. By standard result (see e.g. Lemma 5.7 and Example 5.8 of [44]) we know that the covering size is upper bounded by

$$\left(\frac{2RL}{\alpha}+1\right)^d.$$

By Theorem 1, we find:

$$r_T^a(\mathcal{H}_f)\leq\inf_{0<\alpha<1}\left\{2\alpha T+d\log\left(\frac{2RL}{\alpha}+1\right)\right\}.$$

Taking $\alpha=d/T$, we conclude:

$$r_T^a(\mathcal{H}_f)\leq d\log\left(\frac{2RLT}{d}+1\right)+2d.$$

This completes the proof for $T \geq d$. The upper bound T is achieved by predicting $\frac{1}{2}$ every time. $\square$

Example 1. For logistic function $f(\mathbf{w}, \mathbf{x}) = (1 + e^{-\langle \mathbf{w}, \mathbf{x} \rangle})^{-1}$, and $\mathbf{w} \in \mathcal{B}_2^d(R)$ with $\mathbf{x} \in \mathcal{B}_2^d(1)$, our result recovers those of [3], but with a better leading constant (the bound in [3] has a constant 5). Note that, the result in [17] also provides a sub-optimal constant $c \sim 4$. Moreover, our bounds have a logarithmic dependency on Lipschitz constant L (note that [3] achieves a linear dependency on L, though their Lipschitz condition is for function $\log f$).

The question arises whether the factor in front of $\log T$ can be improved to $d/2$ instead of d as discussed in some recent papers [4], [13], [18]. In Theorem 3 below, we show that, in general, it cannot unless we further strengthen our assumption (see Theorem 4).

Theorem 3. *For any d, T, R, L such that $T \gg d \log(RLT)$, there exists L-Lipschitz function $f : \mathcal{B}_s^d(R) \times \mathbb{R}^d \to [0, 1]$ under ℓ_s norm such that*

$$r_T^a(\mathcal{H}_f) \geq d \log \left(\frac{RLT}{d} \right) - d \log 64 - d \log \log(RLT). \tag{11}$$

We remark that Theorem 3 also demonstrates the tightness of the upper bound in Theorem 2 for all the parameters d, T, R, L on the leading growth. To prove Theorem 3 we first introduce the following well known packing number estimates of Boolean cube, see [44, Example 5.3].

Lemma 5. *For any M with $M \leq e^{T/8}$, there exist M vectors $v_1, v_2, \cdots, v_M \in \{0, 1\}^T$ such that for any $i \neq j \in [M]$ we have:*

$$\sum_{t=1}^{T} 1\{v_i[t] \neq v_j[t]\} \geq T/4.$$

Proof of Theorem 3. Let $\mathbf{x}_1, \cdots, \mathbf{x}_T \in \mathbb{R}^d$ be any distinct points. We construct an L-Lipschitz function $f(\mathbf{w}, \mathbf{x})$ such that the regret restricted only on $\mathbf{x}^T$ is large, which will give us the desired lower bound. To do so, we consider a maximum packing M of the parameter space $\mathcal{B}_s^d(R)$ of radius $\alpha/L > 0$ under ℓ_s norm (where α is to be determined later). Standard volume argument (see Chapter 5 of [44]) yields:

$$|M| \geq \left(\frac{LR}{2\alpha} \right)^d.$$

Now, we define an L-Lipschitz functions $f(\mathbf{w}, \mathbf{x})$ only on $\mathbf{w} \in M$ and $\mathbf{x} \in \{\mathbf{x}_1, \cdots, \mathbf{x}_T\}$. By Lemma 5 (assume for now the conditions are satisfied), we can find $|M|$ binary vectors $V \subset \{0, 1\}^T$ such that any pair of vectors has Hamming distance lower bounded by $T/4$. For each of the vector $v \in V$, we define a vector $u \in [0, 1]^T$ in the following manner – for all $t \in [T]$:

1. If $v[t] = 0$ then set $u[t] = 0$;
2. If $v[t] = 1$ then set $u[t] = \alpha$.

Denote by U the set of all such vectors u. Note that $|U| = |M|$. For any $\mathbf{w} \in M$, we define the following function f: we associate a unique $u \in U$ such that for all $t \in [T]$:

$$f(\mathbf{w}, \mathbf{x}_t) = u[t].$$

We now show that f is indeed L-Lipschitz restricted on M for all $\mathbf{x}_t \in \{\mathbf{x}_1, \cdots, \mathbf{x}_T\}$. This is because for any $\mathbf{w}_1 \neq \mathbf{w}_2 \in M$, we have $|f(\mathbf{w}_1, \mathbf{x}_t) - f(\mathbf{w}_2, \mathbf{x}_t)| \leq \alpha$ by definition of U and $||\mathbf{w}_1 - \mathbf{w}_2||_2 \geq \alpha/L$, since M is a packing.

We now view vectors $u \in U$ as a product of Bernoulli distributions with each coordinate t independently sampled from $\mathrm{Bern}(u[t])$. We show that the sources in U are identifiable. To see this, we note that for any distinct pairs $u_1, u_2 \in U$, there exists a set $I \subset [T]$ such that u_1 and u_2 differ on I and $|I| \geq T/4$. This further implies that there exists a set $J \subset I$ with $|J| \geq T/8$ such that u_1 takes all 0 on J and u_2 takes all α on J (or vice versa). We can then distinguish u_1, u_2 by checking if the samples on J are all 0s or not. The probability of making an error is upper bounded by:

$$(1 - \alpha)^{T/8} \leq e^{-\alpha T/8}.$$

Since there are only $|M|^2$ such pairs, we have the probability of wrongly identifying the source upper bounded by

$$|M|^2 e^{-\alpha T/8}.$$

Taking $\alpha = \frac{16d\log(RLT)}{T}$, the error probability is upper bounded by:

$$\left(\frac{RLT}{32d\log(RLT)}\right)^{2d} e^{-2d\log(RLT)} \leq \left(\frac{1}{32d\log(RLT)}\right)^{2d} \leq \frac{1}{2},$$

for sufficiently large d, T, where we use the fact that $|M| \leq \left(\frac{RLT}{32d\log(RLT)}\right)^d$. Note that we only showed a lower bound on $|M|$, but this is not a problem since we can always remove some points from M to make the upper bound holds as well.

By Lemma 1, we know that the Shtarkov sum of sources in U is lower bounded by $|M|/2$. Therefore, we have:

$$r_T^a(\mathcal{H}_f) \geq r_T^*(\mathcal{H}_f) \geq \log(|M|/2) \geq d\log\left(RLT/d\right) - d\log 64 - d\log\log(RLT).$$

Now, we have to extend the function f to the whole set $\mathcal{B}_s^d(R)$ and keep the L-Lipschitz property. This follows from a classical result in real analysis (see [45, Theorem 1]) by defining for all $\mathbf{w} \in \mathcal{B}_s^d(R)$ and $\mathbf{x}_t \in \{\mathbf{x}_1, \cdots, \mathbf{x}_T\}$

$$f(\mathbf{w}, \mathbf{x}_t) = \sup_{\mathbf{w}' \in M} \{f(\mathbf{w}', \mathbf{x}_t) - L||\mathbf{w} - \mathbf{w}'||_s\}.$$

For $\mathbf{x} \notin \{\mathbf{x}_1, \cdots, \mathbf{x}_T\}$, we can simply let $f(\mathbf{w}, \mathbf{x}) = 0$ for all $\mathbf{w}$.

Finally, we need to check that the condition of Lemma 5 holds for our choice of α, but this is satisfied by our assumption $T \gg d\log(RLT)$. □

B. Lipschitz Class with Bounded Hessian.

As we have shown in Theorem 3, the leading constant, 1, of the regret for Lipschitz parametric classes can not be improved in general. We now show that for some special function f, one can improve the constant to $\frac{1}{2}$, as already noted in [4], [13], [18]. For any function $f : \mathbb{R}^d \times \mathbb{R}^d \to [0, 1]$, we say the Hessian of $\log f$ is uniformly bounded on $\mathcal{X} \subset \mathbb{R}^d$, if there exists constant C such that for any $\mathbf{w} \in \mathbb{R}^d$, $\mathbf{x} \in \mathcal{X}$, and $y \in \{0, 1\}$ we have:

$$\sup_{||\mathbf{u}||_2 \leq 1} |\mathbf{u}^\tau \nabla_{\mathbf{w}}^2 \log f(\mathbf{w}, \mathbf{x})^y (1 - f(\mathbf{w}, \mathbf{x}))^{1-y} \mathbf{u}| \leq C,$$

where $\nabla_{\mathbf{w}}^2$ is the Hessian at $\mathbf{w}$. We note that this condition is similar to the common assumption on the Fisher information matrix as in [7], [32] in the non-sequential case.

Theorem 4. *Let $f : \mathbb{R}^d \times \mathbb{R}^d \to [0, 1]$ be a function such that the Hessian of $\log f$ is uniformly bounded by C on $\mathcal{X}$. Let*

$$\mathcal{H}_f = \{f(\mathbf{w}, \mathbf{x}) : \mathbf{w} \in \mathcal{W}, \mathbf{x} \in \mathcal{X}\}$$

be such a class of f restricted to some compact set $\mathcal{W} \subset \mathbb{R}^d$. Then for any $\epsilon > 0$,

$$r_T^a(\mathcal{H}_f) \leq \log \frac{Vol(\mathcal{W}^*)}{Vol(\mathcal{B}_2^d(\epsilon))} + \frac{1}{2} CT\epsilon^2 + \log 2. \tag{12}$$

where $\mathcal{W}^* = \{\mathbf{w} + \boldsymbol{u} \mid \mathbf{w} \in \mathcal{W},\ \boldsymbol{u} \in \mathcal{B}_2^d(\epsilon)\}$, $Vol(\cdot)$ is volume under Lebesgue measure. In particular, for $\mathcal{W} = \mathcal{B}_2^d(R)$ and $\epsilon = \sqrt{d/CT}$, we have:

$$r_T^a(\mathcal{H}_f) \leq \frac{d}{2}\log\left(\frac{2CR^2T}{d} + 2\right) + d/2 + \log 2.$$

Note that Theorem 4 subsumes the results of [20], [4][4], where the authors considered functions of form $f(\langle \mathbf{w}, \mathbf{x} \rangle)$, and required that the second derivative of $\log f$ be bounded, see also [1, Chapter 11.10]. However, the KL-divergence-based argument of [20] can not be used directly in the setup of Theorem 4 since we *do not* assume that the function

[4]To get the upper bounds in [4] one only needs to estimate the volume of ℓ_s balls, which is well known [46].

f has a linear structure. Our main proof technique for Theorem 4 is a direct application of Lemma 2 and an estimation of the integrals via Taylor expansion.

Proof of Theorem 4. The proof resembles that of [3] but running the Bayesian predictor (Algorithm 1) over $\mathcal{W}^*$ instead of $\mathcal{W}$ with $\mathcal{G}$ being $\mathcal{H}_f$ and μ being Lebesgue measure. Let $\mathbf{x}^T$, y^T, and $\hat{y}^T$ be the feature, label, and predictions of the Bayesian predictor, respectively. By Lemma 2:

$$\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) \leq -\log \frac{\int_{\mathcal{W}^*} p_{\mathbf{w}}(y^T \mid \mathbf{x}^T) \mathrm{d}\mu}{\int_{\mathcal{W}^*} 1 \mathrm{d}\mu}, \tag{13}$$

where μ is the Lebesgue measure and

$$p_{\mathbf{w}}(y^T \mid \mathbf{x}^T) = \prod_{t=1}^{T} f(\mathbf{w}, \mathbf{x}_t)^{y_t} (1 - f(\mathbf{w}, \mathbf{x}_t))^{1-y_t}.$$

We now write $h_t(\mathbf{w}) \stackrel{\text{def}}{=} \log f(\mathbf{w}, \mathbf{x}_t)^{y_t} (1 - f(\mathbf{w}, \mathbf{x}_t))^{1-y_t}$ to simplify notation. It is easy to see that $\ell(f(\mathbf{w}, \mathbf{x}_t), y_t) = -h_t(\mathbf{w})$. Let $\mathbf{w}^*$ be the point in $\mathcal{W}$ that maximizes

$$h(\mathbf{w}) \stackrel{\text{def}}{=} \sum_{t=1}^{T} h_t(\mathbf{w}).$$

Let $\mathbf{u} = \nabla h(\mathbf{w}^*)$ be the gradient of h at $\mathbf{w}^*$. By Taylor theorem, we have for any $\mathbf{w} \in \mathcal{W}^*$:

$$h(\mathbf{w}) = h(\mathbf{w}^*) + \mathbf{u}^\tau (\mathbf{w} - \mathbf{w}^*) + \frac{1}{2} (\mathbf{w} - \mathbf{w}^*)^\tau \nabla^2_{\mathbf{w}'} h(\mathbf{w}') (\mathbf{w} - \mathbf{w}^*),$$

where $\mathbf{w}'$ is a convex combination of $\mathbf{w}$ and $\mathbf{w}^*$ and $\mathbf{u}^\tau$ is the transpose of $\mathbf{u}$.

Now, the key observation is that for any point $\mathbf{w}$ such that $\mathbf{u}^\tau (\mathbf{w} - \mathbf{w}^*) \geq 0$, we have:

$$h(\mathbf{w}) \geq h(\mathbf{w}^*) + \frac{1}{2} (\mathbf{w} - \mathbf{w}^*)^\tau \nabla^2_{\mathbf{w}'} h(\mathbf{w}') (\mathbf{w} - \mathbf{w}^*) \geq h(\mathbf{w}^*) - \frac{1}{2} CT ||\mathbf{w} - \mathbf{w}^*||_2^2, \tag{14}$$

where the last inequality follows from our assumption about the bounded Hessian of $\log f$. Let B be the half ball of radius ϵ centered at $\mathbf{w}^*$ such that for all $\mathbf{w} \in B$ we have $\mathbf{u}^T (\mathbf{w} - \mathbf{w}^*) \geq 0$. By (14), for all $\mathbf{w} \in B$

$$h(\mathbf{w}) \geq h(\mathbf{w}^*) - \frac{1}{2} CT\epsilon^2. \tag{15}$$

Note that $B \subset \mathcal{W}^*$. Then, using above observations we arrive at:

$$\begin{aligned}
\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) &\leq -\log \frac{\int_{\mathcal{W}^*} p_{\mathbf{w}}(y^T \mid \mathbf{x}^T) \mathrm{d}\mu}{\int_{\mathcal{W}^*} 1 \mathrm{d}\mu} && (16)\\
&\leq -\log \frac{\int_{B} p_{\mathbf{w}}(y^T \mid \mathbf{x}^T) \mathrm{d}\mu}{\int_{\mathcal{W}^*} 1 \mathrm{d}\mu}, \text{ since } B \subset \mathcal{W}^* && (17)\\
&\leq -\log \frac{e^{-CT\epsilon^2/2} \int_{B} p_{\mathbf{w}^*}(y^T \mid \mathbf{x}^T) \mathrm{d}\mu}{\int_{\mathcal{W}^*} 1 \mathrm{d}\mu} && (18)\\
&= -\log p_{\mathbf{w}^*}(y^T \mid \mathbf{x}^T) + CT\epsilon^2/2 - \log \frac{\mathrm{Vol}(B)}{\mathrm{Vol}(\mathcal{W}^*)} && (19)\\
&= -\log p_{\mathbf{w}^*}(y^T \mid \mathbf{x}^T) + CT\epsilon^2/2 - \log \frac{\frac{1}{2}\mathrm{Vol}(\mathcal{B}_2^d(\epsilon))}{\mathrm{Vol}(\mathcal{W}^*)} && (20)\\
&= \sum_{t=1}^{T} \ell(f(\mathbf{w}^*, \mathbf{x}_t), y_t) + \log \frac{\mathrm{Vol}(\mathcal{W}^*)}{\mathrm{Vol}(\mathcal{B}_2^d(\epsilon))} + CT\epsilon^2/2 + \log 2. && (21)
\end{aligned}$$

This completes the proof of the general bound. The last part of the theorem follows from the fact that if $\mathcal{W} = \mathcal{B}_2^d(R)$,

$\mathcal{W}^* = \mathcal{B}_2^d(R+\epsilon)$, and noting that:

$$\frac{\text{Vol}(\mathcal{B}_2^d(R+\epsilon))}{\text{Vol}(\mathcal{B}_2^d(\epsilon))} \leq (R/\epsilon+1)^d.$$

□

Remark 2. *When compared to the technique in [34], Theorem 4 does not assume that the gradient critical point of the loss is zero (e.g., the minimum may occur on the boundary). This is why we need to restrict to the half ball B in order to discard the linear term of Taylor expansion in Equation (15). Moreover, in the proof, we work directly on the continuous space instead of a discretized cover, giving an efficient algorithm provided the posterior is efficiently samplable (by e.g., assuming some log-concavity of f as in [3]).*

We complete this part with the following lower bound for *generalized linear functions* under unit ℓ_s balls.

Theorem 5. *Let $f : \mathbb{R} \to [0,1]$ be an arbitrary function such that there exist $c_1, c_2 \in (0,1)$ and for all $r > 0$ we have $[c_1 - c_2 d^{-r}, c_1 + c_2 d^{-r}] \subset f([-d^{-r}, d^{-r}])$ for sufficiently large d. Let*

$$\mathcal{H}_f = \{f(\langle \mathbf{w}, \mathbf{x}\rangle) : \mathbf{w} \in \mathcal{B}_s^d(1), \mathbf{x} \in \mathcal{B}_s^d(1)\}$$

where $s > 0$. Then

$$r_T^a(\mathcal{H}_f) \geq \frac{d}{2}\log\left(\frac{T}{d^{(s+2)/s}}\right) - O(d) \tag{22}$$

where O hides some absolute constant that is independent of d and T.

Note that for the logistic function $f(x) = (1+e^{-x})^{-1}$ Theorem 5 holds with $c_1 = \frac{1}{2}$ and $c_2 = \frac{1}{5}$. Therefore,
1. If $s = 1$, then

$$r_T^a(\mathcal{H}_f) \geq \frac{d}{2}\log\left(\frac{T}{d^3}\right) - O(d).$$

2. If $s = 2$, then

$$r_T^a(\mathcal{H}_f) \geq \frac{d}{2}\log\left(\frac{T}{d^2}\right) - O(d).$$

3. If $s = \infty$, then

$$r_T^a(\mathcal{H}_f) \geq \frac{d}{2}\log\left(\frac{T}{d}\right) - O(d).$$

This recovers all the lower bounds from [4]. We note that a simple sufficient condition for Theorem 5 to hold is to require $f'(0) \neq 0$ if $f(x)$ is differentiable.

Remark 3. *We note that the leading constants in Theorems 4 and 5 match only if d grows sub-polynomially w.r.t. T (i.e., when $d = e^{o(\log T)}$). However, when d grows polynomailly w.r.t. T the leading constants will no longer match, though the bounds still have the asymptotic rate $\Theta(d \log T)$ for $d \ll T^{(s/s+2)-\epsilon}$ with $\epsilon > 0$. Moreover, for $s = 2$, the condition $d \ll \sqrt{T}$ for achieving a $\Omega(d)$ lower bound cannot be relaxed since for logistic function f one can achieve an $\tilde{O}(\sqrt{T})$ upper bound independent of dimension d, see [3, Example 2].*

The proof of Theorem 5 is based on the following technique lemma [5], see Appendix D for detailed proof.

Lemma 6. *The following inequality holds, for $r > 0$:*

$$\sum_{\boldsymbol{y} \in \{0,1\}^{T/d}} \sup_{w \in [c_1 - c_2 d^{-r}, c_1 + c_2 d^{-r}]} P(\boldsymbol{y} \mid w) \geq \Omega(\sqrt{T/d^{2r+1}}), \tag{23}$$

where $P(\boldsymbol{y} \mid w) = w^k (1-w)^{T/d-k}$ with k being the number of 1s in $\boldsymbol{y}$.

Proof of Theorem 5. We choose a particular $\mathbf{x}^T$: We split $\mathbf{x}^T$ into d blocks, each with length of T/d. With that, the ith parts of the input and the output are denoted by $\mathbf{x}^{(i)} = (\mathbf{x}_{(T/d)*(i-1)+1}, \cdots, \mathbf{x}_{(T/d)*i})$ and $\mathbf{y}^{(i)} =$

[5] A similar technique for the ℓ_2 case was used in [12] recently and developed independently by [47].

$(y_{(T/d)*(i-1)+1}, \cdots, y_{(T/d)*i})$, respectively. We define for any $\mathbf{x}_t$ in the ith block $\mathbf{x}^{(i)}$ equals $\mathbf{e}_i$ the standard d base of $\mathbb{R}^d$ with 1 in position i and 0s elsewhere. Note that, with this choice of $\mathbf{x}_t$s, we have $\langle \mathbf{w}, \mathbf{x}_t \rangle = w_i$, where w_i is the ith coordinate of $\mathbf{w}$ and $\mathbf{x}_t \in \mathbf{x}^{(i)}$.

We will lower bound $r_T^*(\mathcal{H}_f \mid \mathbf{x}^T)$, which will automatically give a lower bound on $r_T^a(\mathcal{H}_f)$. We only need to compute the following Shtarkov sum:

$$S_T(\mathcal{H}_f|\mathbf{x}^T) = \sum_{y^T \in \{0,1\}^T} \sup_{\mathbf{w} \in \mathcal{B}_s^d(1)} \prod_{i=1}^{d} P_f(\mathbf{y}^{(i)}|w_i), \tag{24}$$

where $P_f(\mathbf{y}^{(i)}|w_i) = f(w_i)^{k_i}(1-f(w_i))^{T/d-k_i}$ with k_i being the number of 1s in $\mathbf{y}^{(i)}$. We observe that:

$$\begin{aligned}
S_T(\mathcal{H}_f|\mathbf{x}^T) &\geq \sum_{y^T \in \{0,1\}^T} \prod_{i=1}^{d} \sup_{w_i \in [-d^{-1/s}, d^{-1/s}]} P_f(\mathbf{y}^{(i)}|w_i) \\
&= \prod_{i=1}^{d} \sum_{\mathbf{y}^{(i)} \in \{0,1\}^{T/d}} \sup_{w_i \in [-d^{-1/s}, d^{-1/s}]} P_f(\mathbf{y}^{(i)}|w_i) \\
&= \Big(\sum_{\mathbf{y} \in \{0,1\}^{T/d}} \sup_{w \in [-d^{-1/s}, d^{-1/s}]} P_f(\mathbf{y}|w) \Big)^d \\
&\geq \left(\sum_{\mathbf{y} \in \{T/d\}} \sup_{w \in [c_1 - c_2 d^{-1/s}, c_1 + c_2 d^{-1/s}]} P(\mathbf{y} \mid w) \right)^d
\end{aligned}$$

where $P(\mathbf{y} \mid w)$ is as in Lemma 6 and the last inequality holds since $[c_1 - c_2 d^{-1/s}, c_1 + c_2 d^{-1/s}] \subset f([d^{-1/s}, d^{-1/s}])$ by the assumption. Now, Lemma 6 implies that

$$S_T(\mathcal{H}_f \mid \mathbf{x}^T) \geq c^d \left(\frac{T}{d^{(s+2)/s}} \right)^{d/2},$$

where c is some absolute constant that is independent of d and T. We conclude that:

$$r_T^a(\mathcal{H}_f) \geq r_T^*(\mathcal{H}_f) \geq \log S_T(\mathcal{H}_f|\mathbf{x}^T) \geq \frac{d}{2} \log \left(\frac{T}{d^{(s+2)/s}} \right) - O(d),$$

which completes the proof. □

C. Tight Constants Beyond Bounded Hessian

One may observe that the main technique we use in the previous sections for upper bounds is to apply a Bayesian predictor over some *uniform* cover on the parameter space. In this section, we demonstrate that one can improve the upper bounds by performing the Bayesian averaging over some *non-uniform* cover for certain hypothesis classes. This phenomenon is well known in the information theory community for the fixed design case [31]. However, for the sequential case, the situation is more complicated, since we need to deal with $\mathbf{x}^T$ without knowing it in advance. We note that, this phenomenon was discussed for the sequential case by Shamir [4] for analyzing the regret of Logistic regression, where the author used a Gaussian prior to obtain a $\frac{1}{2}$ leading constant. However, as we demonstrated in the proof of Theorem 4, this can actually be achieved by a uniform prior as well. We now show in this section that the non-uniform cover can indeed provide tighter *sequential* regret bounds than the uniform cover for certain classes.

Lemma 7. *Let h_1, h_2 be arbitrary functions that map $\mathcal{X} \to [0,1]$, we define a function class $\mathcal{H} = \{h_w = wh_1 + (1-w)h_2 : w \in [0,1]\}$. Then*

$$r_T^a(\mathcal{H}) \leq \frac{1}{2} \log T + \log \log T + O(1).$$

Proof. For any given $\mathbf{x}^T, y^T$, we denote $\tilde{h}_i(\mathbf{x}_t) = (1-y_t)(1-h_i(\mathbf{x}_t)) + y_t h_i(\mathbf{x}_t)$ for $i \in \{1,2\}$. We also denote $\tilde{h}_w(\mathbf{x}_t) = w\tilde{h}_1(\mathbf{x}_t) + (1-w)\tilde{h}_2(\mathbf{x}_t)$. We have the log-loss over $\mathbf{x}^T, y^T$ against h_w as:

$$\sum_{t=1}^{T} \ell(h_w(\mathbf{x}_t), y_t) = \sum_{t=1}^{T} -\log \tilde{h}_w(\mathbf{x}_t).$$

Let $L_t(w) = -\log \tilde{h}_w(\mathbf{x}_t)$; we have:

$$L_t''(w) = \frac{(\tilde{h}_1(\mathbf{x}_t) - \tilde{h}_2(\mathbf{x}_t))^2}{\tilde{h}_w(\mathbf{x}_t)^2}.$$

We now consider two cases (using elementary algebra):

1. If $\tilde{h}_1(\mathbf{x}_t) \geq \tilde{h}_2(\mathbf{x}_t)$, we have:

$$|L_t''(w)| = \frac{1}{(w + \tilde{h}_2(\mathbf{x}_t)/(\tilde{h}_1(\mathbf{x}_t) - \tilde{h}_2(\mathbf{x}_t)))^2} \leq \frac{1}{w^2}.$$

2 If $\tilde{h}_1(\mathbf{x}_t) \leq \tilde{h}_2(\mathbf{x}_t)$, we have:

$$|L_t''(w)| = \frac{1}{(1 - w + \tilde{h}_1(\mathbf{x}_t)/(\tilde{h}_2(\mathbf{x}_t) - \tilde{h}_1(\mathbf{x}_t)))^2} \leq \frac{1}{(1-w)^2}.$$

Let E be a *non-uniform* cover that is constructed as follows. Let $e_0 = \frac{1}{T}$, we define recursively:

$$e_n = \left(1 + \sqrt{\frac{1}{T}}\right) e_{n-1}.$$

Let N be the maximum number for which $e_N \leq \frac{1}{2}$. It is easy to verify that $N \leq O(\sqrt{T}\log T)$. We define $E = \{e_n, 1-e_n : n \in \{0, \cdots, N\}\} \cup \{0,1\}$. By construction of E, we have for any $w \in [1/T, 1-1/T]$ there exists $e \in E$ such that:

$$|e - w| \leq \min\{w\sqrt{1/T}, (1-w)\sqrt{1/T}\}.$$

Let $w^* \in [0,1]$ be the point that achieves the minimal of $L(w) = \sum_{t=1}^{T} L_t(w)$. We have either $w^* = 0, 1$ or $L'(w^*) = 0$. For the latter case, we have by Taylor expansion, for all e (where w' is convex combination of e, w^*):

$$L(e) = L(w^*) + L''(w')(e - w^*)^2.$$

Taking $e \in E$ such that $|e - w^*| \leq \min\{w^*\sqrt{1/T}, (1-w^*)\sqrt{1/T}\}$ (assume for now $w^* \in [1/T, 1-1/T]$) and using the property for $L''(w)$ above, we have:

$$L(e) \leq L(w^*) + O(1).$$

By applying Algorithm 1 over $\{h_e : e \in E\}$ with uniform prior, we obtain the following regret bound:

$$r_T^a(\mathcal{H}) \leq \log|E| + O(1) = \frac{1}{2}\log T + \log\log T + O(1).$$

Finally, we observe that the case for $w^* \leq 1/T$ or $\geq 1 - 1/T$ does not affect the result, since by taking $e = 1/T$ or $1 - 1/T$ (respectively) guarantees that $L(e) \leq L(w^*) + O(1)$, see e.g., [12, Theorem 2]. □

Remark 4. *Note that a uniform cover as in Theorem 2 can only provide a* $\log T$ *upper bound for the class in Lemma 7. This follows by taking* $\forall \mathbf{x} \in \mathcal{X}$, $h_1(\mathbf{x}) = 0$ *and* $h_2(\mathbf{x}) = 1$, *i.e., the fixed design case with Bernoulli sources. To see this, we let* $y^T = 1 \cdots 1$ *be all* 1 *sequence and* $0, \delta, 2\delta, \cdots, 1$ *be a uniform cover of* $[0,1]$ *with step size* δ. *The probability assigned on* y^T *given by the Bayesian algorithm (with uniform prior over the cover) is*

$$\delta \sum_{n=1}^{1/\delta} (n\delta)^T \leq \delta^{T+1} \sum_{n=1}^{1/\delta} n^T \leq \delta^{T+1} \min\left\{\frac{(1/\delta+1)^{T+1}}{T+1}, \frac{1}{\delta^T}\sum_{n=1}^{T} e^{-i\delta T}\right\} \leq \max\left\{\frac{2}{T+1}, 2\delta\right\},$$

where the second inequality follows by that $\sum_{n=1}^{N} n^T \le (N+1)^{T+1}/(T+1)$ and $n^T/N^T \le e^{-(N-n)T/N}$, the last inequality follows by considering two cases, i.e., $\delta \ge 1/(2T)$ or $\delta < 1/(2T)$. Now, if $\delta \le \frac{\log^2 T}{T}$, then the regret is lower bounded by $(1-o(1))\log T$ (against the source of parameter 1). Else, we take y^T being any binary sequence with $T\cdot\delta/4$ ones. In this case the probability assigned by any source $\mathsf{Bernoulli}(n\delta)$ on y^T equals $e^{-T(\mathsf{KL}(n\delta||\delta/4)+H(\delta/4))} \le e^{-T(\delta/4+H(\delta/4))}$, where we used the fact that $\mathsf{KL}(n\delta||\delta/4) \ge \delta/4$ for sufficient small δ and $n \in \mathbb{N}$. Therefore, the Bayesian algorithm must assign probability $\le e^{-T(\delta/4+H(\delta/4))}$ on y^T. Since the assignment of source $\delta/4$ on y^T is exactly $e^{-TH(\delta/4)}$, the regret is lower bounded by $T\cdot\delta/4 \ge \Omega(\log^2 T) \ge (1-o(1))\log T$.

Using Lemma 7, we prove the following lemma:

Lemma 8. *Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two finite classes, we define:*

$$\mathcal{H} = \{wh_1 + (1-w)h_2 : h_1 \in \mathcal{H}_2,\ h_2 \in \mathcal{H}_2,\ w \in [0,1]\}.$$

Then there exists a covering class $\mathcal{H}'$ such that for all $\mathbf{x}^T, y^T$ and $h \in \mathcal{H}$ there exists $h' \in \mathcal{H}'$ such that:

$$\log p_h(y^T \mid \mathbf{x}^T) \le \log p_{h'}(y^T \mid \mathbf{x}^T) + C,$$

for some absolute constant $C > 0$ and there exists absolute constant $c > 0$:

$$|\mathcal{H}'| \le |\mathcal{H}_1||\mathcal{H}_2| * (c\sqrt{T}\log T).$$

Proof. Let E be the covering set constructed in the proof of Lemma 7, we define:

$$\mathcal{H}' = \{eh_1 + (1-e)h_2 : h_1 \in \mathcal{H}_1,\ h_2 \in \mathcal{H}_2 \text{ and } e \in E\}.$$

We show that $\mathcal{H}'$ is the desired class. Let $\mathbf{x}^T, y^T$ be arbitrary feature and label sequences. Let $h^* \in \mathcal{H}$ be the function that achieves the maximum $p_{h^*}(y^T \mid \mathbf{x}^T)$, where $p_{h^*}(y^T \mid \mathbf{x}^T)$ is defined in Lemma 4. There must be functions $h_1 \in \mathcal{H}_1$, $h_2 \in \mathcal{H}_2$ and $w^* \in [0,1]$ such that $h^* = w^*h_1 + (1-w^*)h_2$. Since classes $\mathcal{H}_1$ and $\mathcal{H}_2$ are discrete, we must have either $w^* = 0, 1$ or $L'(w^*) = 0$, where $L(w) = -\log p_{h_w}(y^T \mid \mathbf{x}^T)$ is the function in the proof of Lemma 7 and $h_w = wh_1 + (1-w)h_2$. Using the same argument as in the proof of Lemma 7, we have $\log p_{h^*}(y^T \mid \mathbf{x}^T) \le \log p_{h_e}(y^T \mid \mathbf{x}^T) + O(1)$, where $h_e = eh_1 + (1-e)h_2$ with $e \in E$. Since h^* achieves the maximum of $p_h(\mathbf{x}^T \mid y^T)$, we have for any $h \in \mathcal{H}$, $\log p_h(y^T \mid \mathbf{x}^T) \le \log p_{h^*}(y^T \mid \mathbf{x}^T) \le \log p_{h_e}(y^T \mid \mathbf{x}^T) + O(1)$. The result follows by counting the size of $\mathcal{H}'$. □

Theorem 6. *Let $h_1, h_2, \cdots, h_s$ be s functions and*

$$\mathcal{H} = \left\{ \sum_{j=1}^{s} w_j h_j : \sum_{j=1}^{s} w_s = 1 \text{ and } w_j \in [0,1] \right\}.$$

Then

$$r_T^a(\mathcal{H}) \le \frac{s-1}{2}\log T + (s-1)\log\log T + O(s).$$

Proof. Let τ be any full binary tree with s leaves and root v_0 (which has $s-1$ internal nodes and s leaves). We will associate each node in the tree with a *set* of functions in a bottom-up fashion with each of leaves associated to one of the sets $\{h_j\}$ with $j \in [s]$ respectively. For each internal node v, we denote $\mathcal{H}_{v_L}$ and $\mathcal{H}_{v_R}$ to be the sets corresponding to its left and right child respectively. We then associate the covering set of $\mathcal{H}_{v_L}$ and $\mathcal{H}_{v_R}$ as in Lemma 8 to the node v. By Lemma 8, the set $\mathcal{H}_{v_0}$ associated to the root v_0 has size at most $(c * \sqrt{T}\log T)^{s-1}$ for some absolute constant c since there are only $s-1$ internal nodes and each internal nodes contribute one merge in the above process. Moreover, by Lemma 8 we will incur an additive constant C for the error of covering after each merge. Therefore for any $\mathbf{x}^T, y^T$ and $h \in \mathcal{H}$ there exist $h' \in \mathcal{H}_{v_0}$ such that $\log p_h(y^T \mid \mathbf{x}^T) \le \log p_{h'}(y^T \mid \mathbf{x}^T) + (s-1)C$. The result follows applying Algorithm 1 over $\mathcal{H}_{v_0}$ with uniform prior. □

Corollary 1. *Let $f = \langle \mathbf{w}, \mathbf{x} \rangle$ with $\mathbf{w} \in \mathcal{B}_1^d(1) \cap [0,1]^d$ and $\mathbf{x} \in [0,1]^d$. Then we have*

$$r_T^a(\mathcal{H}_f) \leq \frac{d}{2} \log T + d \log \log T + O(d).$$

Proof. Note that $\mathcal{H}_f$ is a convex combination of $\langle \mathbf{e}_j, \mathbf{x} \rangle$ with $j \in [d]$ and the all-zero valued function, where $\mathbf{e}_j$ is the standard base of $\mathbb{R}^d$ with value 1 at position j and zeros elsewhere. The result follows by Theorem 6 by taking $s = d + 1$. □

Remark 5. *Note that the function $\log f$ in Corollary 1 is neither Lipschitz nor has bounded Hessian, yet we can still achieve a $\frac{1}{2}$ leading constant. It is worth noting that using the result in [1, Chapter 9.10] and the expected majorizing martingale characterization for r_T^a under Lipschitz loss as established in [48], one can show that the $\frac{1}{2}$ constant can be achieved (non-constructively) if we only assume $\log f$ is Lipschitz (i.e., no requirement on Hessian). This does not apply to the case in Corollary 1, since in our case the function $\log f$ is not even Lipschitz.*

V. Large Growth

We now present some results for large d growing even faster than T. We will show that the size of *global* sequential covering (Definition 1) of a class $\mathcal{H}$ can be bounded by the sequential fat-shattering number of $\mathcal{H}$ in a similar fashion as in [15]. We first introduce the notion of sequential fat-shattering number as in [15].

We denote $\{0,1\}^{\leq d}$ to be the set of all binary sequences of length less than or equal to d. A binary tree of depth d with labels in $\mathcal{X}$ is defined to be a map $\tau : \{0,1\}^{\leq d} \to \mathcal{X}$. For any function class $\mathcal{H} \subset [0,1]^{\mathcal{X}}$, we say $\mathcal{H}$ α-fat shatters tree τ if there exists $[0,1]$-value tree $\mathbf{s} : \{0,1\}^{\leq d} \to [0,1]$ such that for any binary sequence $\epsilon_1^d \in \{0,1\}^d$ there exists $h \in \mathcal{H}$ such that for all $t \in [d]$:

1. If $\epsilon_t = 0$, then $h(\tau(\epsilon_1^{t-1})) \leq \mathbf{s}(\epsilon_1^{t-1}) - \alpha$;
2. If $\epsilon_t = 1$, then $h(\tau(\epsilon_1^{t-1})) \geq \mathbf{s}(\epsilon_1^{t-1}) + \alpha$.

Definition 2. *The sequential α-fat shattering number of $\mathcal{H}$ is defined to be the maximum number $d(\alpha)$ such that $\mathcal{H}$ α-fat shatters a tree τ of depth $d := d(\alpha)$.*

In the lemma below, we present an upper bound for the cardinality of the global sequential covering set w.r.t. the sequential fat-shattering number with algorithmically constructed cover set $\mathcal{G}_\alpha$, see [15, Section 6.1].

Lemma 9. *Let $\mathcal{H} \subset [0,1]^{\mathcal{X}}$ be any class of functions and $d(\alpha)$ be the sequential α-fat shattering number of $\mathcal{H}$. Then there exists a global sequential α-covering set $\mathcal{G}_\alpha$ of $\mathcal{H}$ as in Definition 1 such that:*

$$|\mathcal{G}_\alpha| \leq \sum_{t=0}^{d(\alpha/2)} \binom{T}{t} \left\lceil \frac{1}{\alpha} \right\rceil^t \leq \left\lceil \frac{T}{\alpha} \right\rceil^{d(\alpha/2)+1}. \tag{25}$$

Proof. This follows by combining the Lemma 14 and Lemma 15 of [15] and noting that they consider interval $[-1,1]$ and define α-fat shattering with $\alpha/2$ margin, instead of $[0,1]$ and α we use here. See also [42]. □

We also prove the following simple Lipschitz composition lemma for global sequential covering.

Lemma 10. *Let $\mathcal{H} \subset [0,1]^{\mathcal{X}}$ be a function class, $f : [0,1] \to [0,1]$ be a L-Lipschitz function and $\mathcal{H}' = \{f \circ h : h \in \mathcal{H}\}$. If there exists a global sequential α-covering set $\mathcal{G}_\alpha$ of $\mathcal{H}$ then there exists a global sequential $L\alpha$-covering set $\mathcal{G}'_\alpha$ of $\mathcal{H}'$ such that $|\mathcal{G}_\alpha| = |\mathcal{G}'_\alpha|$.*

Proof. Let $\mathcal{G}'_\alpha = \{f \circ g : g \in \mathcal{G}_\alpha\}$; one can verify, by L-Lipschitz property of f, that $\mathcal{G}'_\alpha$ is the desired covering set. □

Example 2. By [49, Proposition 16] (taking $\Psi(\mathbf{w}) = ||\mathbf{w}||_s^s$ in the proposition and observing that $\Psi(\mathbf{w})$ is s-uniform convex for all $s \geq 2$, see e.g., [50, Section 3]) we know that the sequential α-fat shattering number of linear functions $\langle \mathbf{w}, \mathbf{x} \rangle$ with $\mathbf{w} \in \mathcal{B}_s^d(1)$, $\mathbf{x} \in \mathcal{B}_{s/(s-1)}^d(1)$ and $s \geq 2$ is of order $\tilde{O}(\alpha^{-s})$, where in $\tilde{O}$ we hide a polylog

factor. Let $f(\mathbf{w}, \mathbf{x}) = |\langle \mathbf{w}, \mathbf{x} \rangle|$, we have by Lemma 9 and 10 (since $|\cdot|$ is 1-Lipschitz) that the global sequential α-covering number of $\mathcal{H}_f = \{|\langle \mathbf{w}, \mathbf{x} \rangle| : \mathbf{w} \in \mathcal{B}_2^d(1),\ \mathbf{x} \in \mathcal{B}_{s/(s-1)}^d(1)\}$ is upper bounded by:

$$\left\lceil \frac{T}{\alpha} \right\rceil^{\tilde{O}(\alpha^{-s})}.$$

Applying Theorem 1, we have:

$$r_T^a(\mathcal{H}_f) \leq \inf_{0<\alpha<1} \left\{ 2\alpha T + \tilde{O}\left(\frac{1}{\alpha^s}\right) \right\} \leq \tilde{O}(T^{s/(s+1)}),$$

by taking $\alpha = T^{-1/(s+1)}$. This bound is *independent* of the data dimension d.

Remark 6. *Observe that for any class $\mathcal{H}$ with sequential fat-shattering number of order α^{-s}, one can achieve a regret upper bound of order $\tilde{O}(T^{s/s+1})$ by Theorem 1. We refer to [24], [49] for estimations of sequential fat-shattering number of a variety of classes.*

We complete this part with the following lower bound:

Theorem 7. *For any $s \geq 1$, we define:*

$$\mathcal{D}_s = \left\{ \boldsymbol{p} \in [0,1]^T : \sum_{t=1}^T p_t^s \leq 1 \right\}.$$

We can view the vectors in $\mathcal{D}_s$ as functions mapping $[T] \to [0,1]$. Then

$$r_T^a(\mathcal{D}_s) \geq r_T^*(\mathcal{D}_s) \geq \Omega(T^{s/s+1}). \tag{26}$$

Proof. It is sufficient to compute the Shtarkov sum as in (6). For any $y^T \in \{0,1\}^T$ with k 1s, we claim that:

$$\sup_{\mathbf{p} \in \mathcal{D}_s} p(y^T) = \frac{1}{k^{k/s}},$$

where

$$p(y^T) = \prod_{t=1}^T p_t^{y_t} (1-p_t)^{1-y_t}.$$

To see this, we use a *perturbation* argument. Denote I to be the positions in y^T that take value 1, and let $|I| = k$. For any $\mathbf{p}$ such that $p(y^T)$ is maximum, we must have $p_j = 0$ for all $j \notin I$. Suppose otherwise, we then can move some probability mass on p_j to some $p_i < 1$ with $i \in I$, which will increase the value of $p(y^T)$, thus a contradiction. Now, we need to show that:

$$\prod_{i \in I} p_i \leq \frac{1}{k^{k/s}}$$

This follows easily by AM-GM (i.e., arithmetic mean vs geometric mean) inequality since $\sum_{i \in I} p_i^s \leq 1$ and it is equality when $p_i = \frac{1}{k^{1/s}}$ for all $i \in I$. Now, the Shtarkov sum can be written as:

$$\sum_{k=0}^T \binom{T}{k} \frac{1}{k^{k/s}}. \tag{27}$$

To find a lower bound, we only need to estimate the maximum term in the summation. We have:

$$\max_k \binom{T}{k} \frac{1}{k^{k/s}} \geq \max_k \frac{T^k}{k^{(1+1/s)k}} \geq e^{\frac{s+1}{s \cdot e} T^{s/s+1}},$$

where the last inequality follows by taking $k = \frac{1}{e}T^{s/s+1}$, and we also use the fact that:

$$\binom{T}{k} \geq \frac{T^k}{k^k}.$$

Therefore, we have:

$$r_T^*(\mathcal{D}_s) \geq \frac{s+1}{s \cdot e} T^{s/s+1} = \Omega(T^{s/s+1}),$$

which completes the proof. □

To see why Theorem 7 implies a lower bound for $f(\mathbf{w}, \mathbf{x}) = |\langle \mathbf{w}, \mathbf{x} \rangle|$ with $d \geq T$, as in Example 2, we take $\mathbf{w} \in \mathcal{B}_s^T(1)$ (i.e., with $d = T$) and define $\mathbf{x}_t = \mathbf{e}_t$ with $\mathbf{e}_t$ being the standard base of $\mathbb{R}^T$ that takes value 1 at position t and zero otherwise. Note that the functions of $\mathcal{H}_f$ with $f(\mathbf{w}, \mathbf{x}) = |\langle \mathbf{w}, \mathbf{x} \rangle|$ restricted on $\mathbf{x}^T$ is exactly the functions in $\mathcal{D}_s$. Then:

$$r_T^a(\mathcal{H}_f) \geq r_T^*(\mathcal{H}_f) \geq r_T^*(\mathcal{D}_s) \geq \Omega(T^{s/(s+1)})$$

and this is a matching lower bound of Example 2. Note that, for $s = 2$ we will have a regret bound for $|\langle \mathbf{w}, \mathbf{x} \rangle|$ of order $T^{2/3}$. It is proved in [2] that for function $f(\mathbf{w}, \mathbf{x}) = \frac{\langle \mathbf{w}, \mathbf{x} \rangle + 1}{2}$, one can achieve regret of the form $\tilde{O}(\sqrt{T})$[6]. Example 2 implies that the generalized linear functions of form $f(\langle \mathbf{w}, \mathbf{x} \rangle)$ can have different regrets with polynomial gap even with a simple shift on the value (though they have the same covering number).

A. Tighter Lower Bounds for Generalized Linear Functions.

We now provide additional lower bounds through the fixed design regret. As we have demonstrated in Theorems 5 and 7 that lower bounds can be derived by selecting some appropriate $\mathbf{x}^T$ that maximizes $r_T^*(\mathcal{H} \mid \mathbf{x}^T)$. In these theorems, we only choose $\mathbf{x}^T$ to be some combinations of the standard base of $\mathbb{R}^d$. In this section, we present examples where a more sophisticated selection of $\mathbf{x}^T$ leads to better lower bounds.

The following theorem shows that the leading constant 1 in Theorem 3 holds even for generalized *linear* functions with Lipschitz transform function:

Theorem 8. *For any $s \geq 1$ and $d \ll T^{s/(s+1)}/\log T$, there exists a 1-Lipschitz function $f : [0,1] \to [0,1]$ such that for hypothesis class:*

$$\mathcal{H}_f = \{f(\langle \mathbf{w}, \mathbf{x} \rangle) : \mathbf{w} \in \mathcal{B}_s^d(1), \mathbf{x} \in \mathcal{B}_{s/(s-1)}^d(1)\},$$

we have

$$r_T^*(\mathcal{H}_f) \geq d \log \left(\frac{T}{d^{(s+1)/s}} \right) - O(d \log \log T).$$

Note that, Theorem 3 applies to the *general* Lipschitz parametric class, while Theorem 8 applies specifically to the generalized *linear* functions with Lipschitz transform function (which is a subset of Lipschitz parametric class). That is why we have different exponents on the denominator inside the leading log term, i.e., d v.s. $d^{(s+1)/s}$. By Theorem 2 and Example 2 we know that both of the dependencies (for large d) are tight. Moreover, the lower bound in Theorem 8 (as with Theorem 3) only holds for some *hard* function f, while the lower bounds in Theorems 5 and 7 are proved for specific functions f.

The proof of Theorem 8 is based on the following technical lemma, which can be viewed as a correlated version of Lemma 5, see Appendix D for detailed proof.

Lemma 11. *There exists a binary sequence $b_1, b_2, \cdots, b_{2T} \in \{0,1\}$ such that for any $i \neq j \in [T]$ we have*

$$\sum_{t=1}^{T} 1\{b_{i+t} \neq b_{j+t}\} \geq \frac{T}{8}.$$

[6] A $\tilde{\Omega}(\sqrt{T})$ lower bound for $d \geq \sqrt{T}$ can be derived from Theorem 5, recovering [2, Lemma 8].

Proof of Theorem 8. We partition the $\mathbf{x}^T$ into d parts each of length T/d (assume w.l.o.g. that d divides T [7]), and denote $K = T/d$. For the ith part in the partition, we will select the K features of form $x_1\mathbf{e}_i, \cdots, x_K\mathbf{e}_i$, where $x_j \in [0,1]$ and $\mathbf{e}_i$ is the standard base of $\mathbb{R}^d$ with position i being 1 and zeros elsewhere. Note that all of the parts in the partition have the *same* scalars x_js but different vector $\mathbf{e}_i$. Clearly, the value of $f(\langle \mathbf{w}, \mathbf{x}\rangle)$ with $\mathbf{x}$ in the ith part depends only on w_i, where w_i is the ith coordinate of $\mathbf{w}$. Since $\mathbf{w} \in \mathcal{B}_s^d(1)$, we may assume that each w_i takes values in the full range $[-d^{-1/s}, d^{-1/s}]$ and selected independently. Therefore, it is sufficient to construct a function $f : [0,1] \to [0,1]$ such that the class

$$\mathcal{Q} = \{f(w \cdot x) : w \in [-d^{-1/s}, d^{-1/s}],\ x \in [0,1]\}$$

has large regret.

Let $\gamma \in [0,1]$ be a small real number depending on K that is to be determined later. For any number $t \le 1/\gamma$, we define $z_t = (1-\gamma)^{t-1}$. Let $\mathcal{Z} = \{z_t : t \le 1/\gamma\}$. We now select the $x_1, \cdots, x_K$ to be elements in $\mathcal{Z}$ such that each repeats at least $\lfloor \gamma K \rfloor$ times (there can be some elements in $\mathcal{Z}$ that repeat more than $\lfloor \gamma K \rfloor$ times). We also define a sequence z'_t with $t \le 2/\gamma$ such that $z'_t = d^{-1/s}(1-\gamma)^{t-1}$. We observe that $|z'_t - z'_{t+1}| \ge d^{-1/s} \cdot e^{-3} \cdot \gamma$ for all $t \le 2/\gamma$, since $(1-\gamma)^{2/\gamma} \sim e^{-2}$ for γ small enough. For any number $t \le 1/\gamma$, we define $w_t = d^{-1/s}(1-\gamma)^{t-1} \in [-d^{-1/s}, d^{-1/s}]$.

Let $M = \lfloor 1/\gamma \rfloor$ and b^{2M} be the binary sequence as in Lemma 11 with $T = M$. Define f to be the function over z'_ts with $t \le 2/\gamma$, such that $f(z'_t) = 0$ if $b_t = 0$ and $f(z'_t) = d^{-1/s} \cdot e^{-3} \cdot \gamma$ if $b_t = 1$. Clearly, f is 1-Lipschitz over the z'_ts. By Lemma 11, for any w_i and w_j as defined above with $i \ne j \le 1/\gamma$, there are at least $M/16$ positions $t \le 1/\gamma$ such that $f(w_i \cdot z_t)$ take all 0s and $f(w_j \cdot z_t)$ take all values $d^{-1/s} \cdot e^{-3} \cdot \gamma$ (or vice versa). Note that, any of the w_i defines a product distribution over $\{0,1\}^K$ such that each coordinate $t \le K$ is an independent Bernoulli random variable with parameter $f(w_i \cdot x_t)$. Since each z_t appears least $\lfloor \gamma K \rfloor$ times in the x_ts, we have the probability of wrongly identifying the source (of the product distribution) with parameters w_i and w_j being upper bounded by

$$(1 - d^{-1/s} \cdot e^{-3} \cdot \gamma)^{(M/16)\cdot\lfloor \gamma K \rfloor} \le e^{-d^{-1/s}\cdot e^{-3}\cdot\gamma(K - 1/\gamma - \gamma K)/16}.$$

Taking $\gamma = 64 \cdot e^3 \cdot (d^{1/s} \cdot \log K)/K$ and applying a union bound on all the pairs w_i, w_j (there are at most K^2 such pairs), we can make error probability upper bounded by $1/2$ for sufficiently large K. We now verify that $\gamma \to 0$ as $K \to \infty$, this is guaranteed by our assumption that $d \ll T^{s/(s+1)}/\log T$. By Lemma 1, we have

$$r_T^*(\mathcal{Q}) \ge \log(K/(d^{1/s}\log K)) - O(1).$$

Since $K = T/d$ and each of the parts of the partition of $\mathbf{x}^T$ are independent, we have

$$r_T^*(\mathcal{H}) \ge d \cdot r_T^*(\mathcal{Q}) \ge d \log\left(\frac{T}{d^{(s+1)/s}}\right) - O(d \log\log T).$$

Finally, by [45, Theorem 1], we can extend f to the whole set $[0,1]$ while keep the Lipshitz condition. □

Remark 7. *Note that the condition $d \ll T^{s/(s+1)}$ cannot be relaxed (upto poly-log factors in general) by Example 2. Our results in Theorem 8, Theorem 2, and Example 2 imply an interesting threshold phenomenon (for the generalized linear functions with worst case Lipschitz transform function), i.e., when $d \ll T^{(s/s+1)-\epsilon}$ with $\epsilon > 0$, the regret grows as $\Theta(d \log T)$, while for $d \gg T^{s/s+1}$, the regret grows as $\tilde{O}(T^{s/s+1})$. Moreover the leading constant is exactly 1 for the $d \log T$ term if d is sub-polynomial w.r.t. T (e.g., $d = e^{\sqrt{\log T}}$).*

B. Additional Large Classes.

In this section we consider a general (including non-parametric) class $\mathcal{H} = \{h \in [0,1]^{\mathcal{B}_s^d(1)} : \forall \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{B}_s^d(1),\ |h(\mathbf{x}_1) - h(\mathbf{x}_2)| \le ||x_1 - x_2||_s\}$ of all Lipschitz functions mapping from a unit ℓ_s ball to $[0,1]$. We also assume that the Lipschitz condition is under the norm ℓ_s as well [8]. The following theorem establishes a lower bound

[7] Otherwise, we round T to be some $T' \le T$ that is divisible by d, this only incurs a $O(d)$ regret loss.

[8] Note that our technique can be generalized to cases when the Lipschitz conditions are defined in a different norm.

for such function classes (a matching upper bound – up to poly-log factors – can be derived using Theorem 1 and the classical uniform covering numbers as in [51, Lemma 5.2], see also [17]).

Theorem 9. *For any $d, T, s \geq 1$ such that $d \ll \log T$, we let $\mathcal{H} = \{h \in [0,1]^{\mathcal{B}_s^d(1)} : \forall \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{B}_s^d(1),\ |h(\mathbf{x}_1) - h(\mathbf{x}_2)| \leq ||x_1 - x_2||_s\}$. Then*

$$r_T^*(\mathcal{H}) \geq \Omega(T^{d/d+1}).$$

Moreover, if $d \geq \Omega(\log T)$, then $r_T^*(\mathcal{H}) \geq \Omega(T)$.

Proof. Let M be a maximum packing of $\mathcal{B}_s^d(1)$ under ℓ_s norm with radius T^{-r}, where $r = \frac{1}{d+1}$. Standard volume argument yields that:

$$|M| \geq T^{dr}.$$

Note that the packing number is independent of s, since we are packing a ℓ_s ball under the same ℓ_s norm. Assume w.l.o.g. that $|M| = \lfloor T^{dr} \rfloor$. We now select $\mathbf{x}^T$ to be all the elements in M such that each of them is repeated at least $\lfloor T^{1-dr} \rfloor$ times in $\mathbf{x}^T$. We now select a class $\mathcal{F} \subset \mathcal{H}$ of functions that map $M \to \{0, T^{-r}\}$ such that any two functions differ by at least $\lfloor T^{dr}/4 \rfloor$ elements in M. By Lemma 5, we know that there are at least $2^{T^{dr}/8}$ such functions. By removing some functions, we may assume that there are $\lfloor 2^{T^{dr}/16} \rfloor$ functions in $\mathcal{F}$. We now view the function $f \in \mathcal{F}$ as a product of Bernoulli process over $\{0,1\}^T$ such that each coordinate t distributed as Bernoulli$(f(\mathbf{x}_t))$. Now, for any distinct pair of functions $f_1, f_2 \in \mathcal{F}$, we have the probability of wrongly identifying the source $f \in \{f_1, f_2\}$ by observing samples generated by f is upper bounded by (using the same identification rule as in Theorem 3 by checking that the samples at distinct positions of f_1, f_2 are all 0s or not):

$$\left(1 - T^{-r}\right)^{\lfloor T^{1-dr} \rfloor \cdot \lfloor T^{dr}/4 \rfloor / 2} \leq e^{-T^{-r} \cdot \lfloor T^{1-dr} \rfloor \cdot \lfloor T^{dr}/4 \rfloor / 2} \leq e^{-(1/8 - o(1))T^{1-r}},$$

where we used the fact that each element in M repeats at least $\lfloor T^{1-dr} \rfloor$ times in $\mathbf{x}^T$ and $d \ll \log T$. In order for a union bound over all pairs in $\mathcal{F}$ to work (there are at most $2^{T^{dr}/8}$ such pairs), it is sufficient to have (since $2 < e$):

$$T^{dr}/8 \leq T^{1-r}/8.$$

This holds when $r = \frac{1}{1+d}$. Therefore, by union bound, we have for sufficient large T, one can identify the sources in $\mathcal{F}$ with error probability upper bounded by $\frac{1}{2}$. Invoking Lemma 1, we conclude that

$$r_T^*(\mathcal{H}) \geq r_T^*(\mathcal{F}) \geq \frac{1}{16 \log 2} T^{d/d+1} - \log(2).$$

By definition, the functions in $\mathcal{F}$ restricted on M are 1-Lipschitz under ℓ_s norm. By [45, Theorem 1], we can extend them to the whole set of $\mathcal{B}_s^d(1)$ while keeping the Lipschitz property. The last part of the theorem follows from the fact that when $d = c \log T$ for any constant c, we have $T^{-1/(1+d)} \sim e^{-1/c}$. We can therefore let c be small enough so that $e^{-1/c} \leq \frac{1}{16}$, which will allow our argument above to go thorough (by reducing the size of $\mathcal{F}$ by a constant on the exponent) and result in a regret bound $\Omega(T^{d/(d+1)}) \geq \Omega(T)$. Note that regret is monotone increasing according to d, thus the result follows. □

Remark 8. *Note that Theorem 9 recovers the lower bound in [17, Theorem 3] established only for the ℓ_∞ ball. The main technique of [17] is to construct a distribution over some grid of $[0,1]^d$ and show that regret is large for the samples generated $i.i.d.$ from this distribution, which we believe is not easily generalizable to the ℓ_s balls as we established in Theorem 9. Moreover, our proof is simpler and more accessible for exposition. It is worth noting that in the proof of Theorem 9 we construct $\mathcal{F}$ by packing the* feature *space, which differs from the proof Theorem 3 since the packing is performed on the* parameter *space.*

VI. Optimality of Bayesian Averaging

So far, all the upper bounds established in previous sections are achieved by applying Bayesian averaging over some well designed finite cover (e.g., uniform, non-uniform and sequential covers). A natural question is whether the best bound can always be achieved by a Bayesian algorithm with prior over some finite cover. While we are

unable to answer this question in its generality, we will show in this section that the Bayesian algorithm can be arbitrary loose if the support of the prior is included in the hypothesis class itself.

We first prove the following lemma, which establishes a non-constructive upper bound for $r_T^a(\mathcal{H})$ based on the expected majorizing martingale characterization of [48]. However, we emphasize that this approach only works for the Lipschitz convex losses and does not apply to log-loss in general. We are able to apply this approach here because we have restricted the value of functions in $\mathcal{H}$ to be around $\frac{1}{2}$, which effectively reduced the log-loss to a Lipschitz convex loss.

Lemma 12. *Let $\mathcal{H}$ be a finite set of functions mapping $\mathcal{X} \to \left[\frac{1-1/\sqrt{T}}{2}, \frac{1+1/\sqrt{T}}{2}\right]$. The following general upper bound holds:*

$$r_T^a(\mathcal{H}) \leq O(\sqrt{\log|\mathcal{H}|}).$$

Proof. Denote $\mathbf{z}_t = (\mathbf{x}_t, y_t)$ and $\ell(h, \mathbf{z}_t) = \ell(h(\mathbf{x}_t), y_t)$, we have by [48, Theorem 7] that

$$r_T^a(\mathcal{H}) \leq \sup_D \mathbb{E} \sup_{h\in\mathcal{H}} \left[\sum_{t=1}^T \mathbb{E}_t[\ell(h, \mathbf{z}_t)] - \ell(h, \mathbf{z}_t)\right],$$

where D runs over all joint distributions over $\mathbf{z}^T$, and $\mathbb{E}_t$ is the conditional expectation conditioning on $\mathbf{z}^{t-1}$. Note that for any given h, the terms $\mathbb{E}_t[\ell(h, \mathbf{z}_t)] - \ell(h, \mathbf{z}_t)$ form martingale differences. Since h takes values in $\left[(1-\sqrt{1/T})/2, (1+\sqrt{1/T})/2\right]$, by cancelling the $1/2$ factor, we have $\mathbb{E}_t[\ell(h, \mathbf{z}_t)] - \ell(h, \mathbf{z}_t) \in \left[-2\sqrt{1/T}, 2\sqrt{1/T}\right]$ for sufficiently large T, where we have used the fact that $\log(1+\alpha) \sim \alpha$ when α is small. Denote $X_t^h = \mathbb{E}_t[\ell(h, \mathbf{z}_t)] - \ell(h, \mathbf{z}_t)$, we have by Azuma's inequality [1, Lemma A.7] and $\sum_{t=1}^T (X_t^h)^2 \leq 4$ that:

$$\Pr\left[\sum_{t=1}^T X_t^h \geq \epsilon\right] \leq e^{-\epsilon^2/2}.$$

By union bound on $\mathcal{H}$, we have:

$$\begin{aligned}
r_T^a(\mathcal{H}) &\leq \sup_D \mathbb{E}\left[\sup_{h\in\mathcal{H}} \sum_{t=1}^T X_1^h\right] \\
&\leq \sup_D \int_{\epsilon>0} \Pr\left[\sup_{h\in\mathcal{H}} \sum_{t=1}^T X_1^h \geq \epsilon\right] \\
&\leq \int_{0\leq\epsilon\leq\sqrt{2\log|\mathcal{H}|}} 1 + \int_{\epsilon>\sqrt{2\log|\mathcal{H}|}} |\mathcal{H}| e^{-\epsilon^2/2} \\
&\leq O(\sqrt{\log|\mathcal{H}|})
\end{aligned}$$

which completes the proof. □

Note that Lemma 12 only shows that a $O(\sqrt{\log|\mathcal{H}|})$ sequential regret can be achieved by *some* algorithm, though we are unaware how the algorithm can be implemented. We will show in the following discussion that such a bound cannot be achieved by the Bayesian algorithm with *any* prior over $\mathcal{H}$. We need the following characterization of Bayesian algorithms under Log-loss, which is due to Kakade and Ng [52]. However, we will use the lemma below differently than [52], since we will be deriving *lower bounds* for Bayesian algorithm instead of upper bounds as in [52]. To our knowledge this lower bounding technique is new in literature.

Lemma 13. *Let $\mathcal{G} \subset [0,1]^{\mathcal{X}^*}$ be any class with index set $\mathcal{W}$ and μ, ν be probability distributions over $\mathcal{W}$. For any given $\mathbf{x}^T, y^T$, we denote*

$$L_\nu(\mathbf{x}^T, y^T) = \int_{\mathcal{W}} \sum_{t=1}^T \ell(g_w(\mathbf{x}^t), y_t) d\nu.$$

Let $\hat{y}_t$ be the Baysian predictor as in Algorithm 1 with prior μ, then

$$\sum_{t=1}^{T} \ell(\hat{y}_t, y) - L_\nu(\mathbf{x}^T, y^T) = \mathsf{KL}(\nu||\mu) - \mathsf{KL}(\nu||\mu_T),$$

where $\mu_T(u) = \frac{p_{g_u}(y^T|\mathbf{x}^T)\mu(u)}{\int_{\mathcal{W}} p_{g_w}(y^T|\mathbf{x}^T)d\mu}$ is the posterior distribution of μ conditioning on $\mathbf{x}^T, y^T$, $p_f(y^T \mid \mathbf{x}^T)$ is defined in Lemma 4 and KL is the KL-divergence.

Proof. The proof is identical to the proof in [52, Lemma 2.1] except that we *do not* lower bound $\mathsf{KL}(\nu||\mu_T)$ being 0. □

We have the following theorem, which is the main contribution of this section:

Theorem 10. *For any infinite domain $\mathcal{X}$, there exists a finite class $\mathcal{H}$ of functions mapping $\mathcal{X} \to \left[\frac{1-1/\sqrt{T}}{2}, \frac{1+1/\sqrt{T}}{2}\right]$ with $|\mathcal{H}| \leq e^{\sqrt{T}}$ such that:*

$$r_T^a(\mathcal{H}) \leq O(\sqrt{\log|\mathcal{H}|}),$$

but for any prior μ over $\mathcal{H}$, the Bayesian algorithm (Algorithm 1) with prior μ can only achieve the sequential regret of order:

$$\Omega(\log|\mathcal{H}|).$$

In particular, by taking $|\mathcal{H}| = e^{\sqrt{T}}$ we established a $T^{1/4}$ factor between the best achievable regret and the regret achievable by a Bayesian algorithm with best prior over $\mathcal{H}$.

Proof. The general upper bound follows directly from Lemma 12. We now show that there exists a class $\mathcal{H}$ of functions mapping $\mathcal{X} \to \left[\frac{1-1/\sqrt{T}}{2}, \frac{1+1/\sqrt{T}}{2}\right]$ such that for any prior μ on $\mathcal{H}$, the Bayesian predictor (Algorithm 1) with prior μ can only achieve the sequential regret lower bounded by:

$$\Omega(\log|\mathcal{H}|),$$

thus establishing a $\sqrt{\log|\mathcal{H}|}$ gap compared to the best achievable upper bound established in Lemma 12. To do so, we define $\mathcal{H}$ to be the class that for any $i \in [|\mathcal{H}|]$ there exists some $\mathbf{x}$ such that $h_i(\mathbf{x}) = \frac{1+1/\sqrt{T}}{2}$ and $h_j(\mathbf{x}) = \frac{1-1/\sqrt{T}}{2}$ for all $i \neq j \in [|\mathcal{H}|]$. Such a class clearly exists when $|\mathcal{X}| \geq |\mathcal{H}|$.

Let $\lambda_1, \lambda_2, \cdots, \lambda_{|\mathcal{H}|}$ be an arbitrary prior over $\mathcal{H}$ at the beginning of the Bayesian predictor, where λ_i is the probability mass assigned on the ith function in $\mathcal{H}$. We know that there must be some $\lambda_i \leq \frac{1}{|\mathcal{H}|}$. We now construct the following adversary for choosing $\mathbf{x}^T, y^T$: For all time steps t, we assign value $h_i(\mathbf{x}_t) = \frac{1+1/\sqrt{T}}{2}$ for h_i and $h_j(\mathbf{x}_t) = \frac{1-1/\sqrt{T}}{2}$ for all other functions $h_j \in \mathcal{H}$ (such $\mathbf{x}_t$ exists by definition of $\mathcal{H}$) and let $y_t = 1$. We now analyze the regret. By Lemma 13, we have by taking ν to be the distribution that assigns probability 1 on h_i that the pointwise regret (over the constructed $\mathbf{x}^T, y^T$) of Bayesian algorithm with prior μ equals:

$$\log 1/\lambda_i + \log\left(\frac{\lambda_i(1+\sqrt{1/T})^T}{\lambda_i(1+\sqrt{1/T})^T + (1-\lambda_i)(1-\sqrt{1/T})^T)}\right) \geq \Omega(\log|\mathcal{H}|),$$

where the first equality follows since $\mathsf{KL}(\nu||\mu) = -\log(\mu(h_i))$ for any μ if ν assign probability 1 on h_i, the last inequality holds since $\lambda_i \leq \frac{1}{|\mathcal{H}|}$ and $(1 \pm 1/\sqrt{T})^T \sim e^{\pm\sqrt{T}}$ for T large enough. We note that the regret lower bound established above can not be larger than $2\sqrt{T}$, which is achieved when $\lambda_i = 0$. Taking $|\mathcal{H}| = e^{\sqrt{T}}$, one will establish a $T^{1/4}$ factor on the discrepancy between the regret achieved by Bayesian algorithm and best achievable regret. □

Remark 9. *We note that a similar separation was discussed in [23, Section 5] but only holds for the simulatable experts and uniform prior. Our Theorem 10 is stronger since it holds for sequential regret and it rules out the possibility of optimality of Bayesian algorithms with* any *prior over $\mathcal{H}$ (not just uniform prior). We leave it as an open problem to determine if such a separation as in Theorem 10 exists if we allow the covering functions to be outside of $\mathcal{H}$ and sequential.*

VII. Conclusion

In this work we derived the best known lower and upper (often matching) bounds on sequential online regret for a large class of experts. We accomplish it by designing a new smooth truncated Bayesian algorithm, together with the concept of global sequential covering, that achieves these upper bounds. For the lower bound, we used a novel information-theoretic approach based on the Shtarkov sum and fixed designed minimax regret. These techniques can be successfully used for a broader set of problems, e.g., when the features $\mathbf{x}^T$ present stochastically [53]. There are also several open problems left unsolved in this paper: (1). it is known by [17] (as well as by Theorem 7) that the sequential covering number cannot completely characterize the sequential regret under log-loss, a natural question is if the sequential *dominate* set (see Remark 1) is sufficient to characterize the regret (note that all of the upper bounds established in this paper can be rephrased as constructing the *dominate* sets); (2). if not, can the Bayesian algorithm with prior supported over some well designed class of sequential functions achieves the optimal regret for any non-sequential expert class $\mathcal{H}$?

Appendix A
Proofs of Lemma 2 and Lemma 3

We prove here Lemma 2 and Lemma 3. For the reader's convenience we repeat both lemmas.

Lemma 2 *Let $\mathcal{G}$ be a class of functions $g_w : \mathcal{X}^* \to [0,1], w \in \mathcal{W}$. Let $\hat{y}_t$ be the Bayesian prediction rule as in Step 4 of Algorithm 1 with prior μ. Then, for any $\mathbf{x}^T$ and y^T we have*

$$\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) \le -\log \frac{\int_{\mathcal{W}} p_w(y^T \mid \mathbf{x}^T) d\mu}{\int_{\mathcal{W}} 1 d\mu},$$

where $p_w(y^T \mid \mathbf{x}^T) = e^{-\sum_{t=1}^{T} \ell(g_w(\mathbf{x}^t), y_t)}$ and ℓ is the log-loss.

Proof. We first observe that for any $y \in \{0,1\}$ we have $e^{-\ell(\cdot, y)}$ is concave over $[0,1]$. Let

$$\lambda_{t-1}(w) = \frac{p_w(y^{t-1} \mid \mathbf{x}^{t-1})}{\int_{\mathcal{W}} p_w(y^{t-1} \mid \mathbf{x}^{t-1}) \mathrm{d}\mu}.$$

Note that $\lambda_{t-1}(w)$ forms a probability density over $\mathcal{W}$ w.r.t. μ. By definition of $\hat{y}_t$, we have $\hat{y}_t = \mathbb{E}_{\lambda_{t-1}}[g_w(\mathbf{x}^t)]$, where the expectation is over the density of $\lambda_{t-1}(w)$. Therefore, by Jensen's inequality and the update procedure as in item 6 of Algorithm 1, we have:

$$e^{-\ell(\hat{y}_t, y_t)} = e^{-\ell(\mathbb{E}[g_w(\mathbf{x}^t)], y_t)} \ge \mathbb{E}[e^{-\ell(g_w(\mathbf{x}^t), y_t)}] = \frac{\int_{\mathcal{W}} p_w(y^t \mid \mathbf{x}^t) \mathrm{d}\mu}{\int_{\mathcal{W}} p_w(y^{t-1} \mid \mathbf{x}^{t-1}) \mathrm{d}\mu}.$$

By telescoping the sum, we find:

$$e^{-\sum_{t=1}^{T} \ell(\hat{y}_t, y_t)} \ge \frac{\int_{\mathcal{W}} p_w(y^T \mid \mathbf{x}^T) \mathrm{d}\mu}{\int_{\mathcal{W}} 1 \mathrm{d}\mu}.$$

This implies:

$$\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) \le -\log \frac{\int_{\mathcal{W}} p_w(y^T \mid \mathbf{x}^T) \mathrm{d}\mu}{\int_{\mathcal{W}} 1 \mathrm{d}\mu}$$

and completes the proof. □

Lemma 3 *For any finite class of experts $\mathcal{G}$*

$$r_T^a(\mathcal{G}) \le \log |\mathcal{G}|.$$

Proof. Let $\mu(w) = \frac{1}{|\mathcal{W}|}$ as in Lemma 2 and $\hat{y}_t$ be the Bayesian predictor with input $\mathcal{G}$ and μ. Then

$$\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) \leq -\log \frac{\int_{\mathcal{W}} p_w(y^T \mid \mathbf{x}^T) \mathrm{d}\mu}{\int_{\mathcal{W}} 1 \mathrm{d}\mu} \tag{28}$$

$$= -\log \int_{\mathcal{W}} p_w(y^T \mid \mathbf{x}^T) \mathrm{d}\mu + \log 1 \tag{29}$$

$$= -\log \int_{\mathcal{W}} p_w(y^T \mid \mathbf{x}^T) \mathrm{d}\mu \tag{30}$$

$$\leq -\log p_{w^*}(y^T \mid \mathbf{x}^T) + \log |\mathcal{W}|, \text{ where } w^* \text{ maximizes } p_w(y^T \mid \mathbf{x}^T) \tag{31}$$

$$= \sum_{t=1}^{T} \ell(g_{w^*}(\mathbf{x}^t), y_t) + \log |\mathcal{G}|, \quad \text{since } |\mathcal{W}| = |\mathcal{G}|. \tag{32}$$

This concludes the proof. □

APPENDIX B
FAT-SHATTERING BASE LOWER BOUNDS

In this appendix we prove a lower bound based on the sequential fat-shattering number in contrast to the offset sequential Rademacher complexity bound as in [2, Lemma 10]. We first prove the following simple lemma:

Lemma 14. *For any $x, \delta \in [0, 1]$ we have*

$$|\log(x + \delta) - \log(x)| \geq \frac{\delta}{1 + \delta} \geq \delta/2$$

Proof. Observe that $\log(x + \delta) - \log(x) = \log(1 + \delta/x) \geq \log(1 + \delta)$ since $x \leq 1$. The result follows by the elementary inequality $\log(1 + x) \geq \frac{x}{1+x}$ for all $x > -1$. □

Proposition 2. *Let $\mathcal{H} \subset [0, 1]^{\mathcal{X}}$ be a class with sequential α-fat shattering number of order $\Omega(\alpha^{-s})$. Then*

$$r_T^a(\mathcal{H}) \geq \Omega(T^{(s-1)/s}).$$

Proof. Let τ and $\mathbf{s}$ be the trees as in Definition 2 of sequential fat-shattering with $\alpha = T^{-1/s}$. We construct the following adversary by tracing the trees. At each time step t, we will be at some node v_t of τ (respectively in $\mathbf{s}$), initially with v_1 being the root. For any prediction $\hat{y}_t$ made by the predictor, if $\hat{y}_t \geq \mathbf{s}(v_t)$ we set $y_t = 0$ and v_{t+1} being the left child of v_t, else we set $y_t = 1$ and v_{t+1} being the right child of v_t. By definition of fat-shattering, we can find a function $h \in \mathcal{H}$ such that for any $\mathbf{x}_t$ we have $\ell(h(\mathbf{x}_t, y_t)) \leq \ell(\hat{y}_t, y_t)$ and $|h(\mathbf{x}_t) - \hat{y}_t| \geq \alpha$. By Lemma 14, we have $\ell(\hat{y}_t, y_t) - \ell(h(\mathbf{x}_t, y_t)) \geq \alpha/2$ for all $t \in [T]$. Since the fat-shattering number is of order α^{-s}, we have the depth of the tree is T. This implies a regret lower bound of order $\alpha T = T^{-1/s} \cdot T = T^{(s-1)/s}$. □

By Theorem 7, we know that the lower bound in Proposition 2 is not tight in general. However, it provides a *universal* lower bound that holds for any class with large sequential fat-shattering number.

APPENDIX C
EXTENSION TO SEQUENTIAL EXPERTS

In this appendix, we briefly discuss how our results can be extended to sequential experts that make predictions which depend on past features and labels. We say a function h a *full* sequential function if it maps $(\mathcal{X} \times \{0, 1\})^* \times \mathcal{X} \to [0, 1]$. Let $\mathcal{H}$ be a class of fully sequential functions. The sequential regret can be expressed as

$$r_T^a(\mathcal{H}) = \sup_{\mathbf{x}_1} \inf_{\hat{y}_1} \sup_{y_1} \cdots \sup_{\mathbf{x}_T} \inf_{\hat{y}_T} \sup_{y_T} \sum_{t=1}^{T} \ell(\hat{y}_t, y_t) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}^t, y^{t-1}), y_t).$$

We define the following concept of the *sequential dominance* as in Remark 1.

Definition 3. *For any fully sequential expert class $\mathcal{H}$, we say a fully sequential function class $\mathcal{G}$ sequentially α-dominate $\mathcal{H}$ if for all $\mathbf{x}^T, y^T$ and $h \in \mathcal{H}$ there exists $g \in \mathcal{G}$ such that*

$$\log p_h(y^T \mid \mathbf{x}^T) - \log p_g(y^T \mid \mathbf{x}^T) \le \alpha T,$$

where for $f \in \{h, g\}$

$$p_f(y^T \mid \mathbf{x}^T) = \prod_{t=1}^{T} f(\mathbf{x}^t, y^{t-1})^{y_t}(1 - f(\mathbf{x}^t, y^{t-1}))^{1-y_t}.$$

The following result is straightforward and extends Theorem 1.

Proposition 3. *Suppose a fully sequential expert class $\mathcal{H}$ admits a fully sequential α-dominate class $\mathcal{G}_\alpha$ for all $\alpha \ge 0$, then*

$$r_T^a(\mathcal{H}) \le \inf_{\alpha \ge 0}\{\alpha T + \log|\mathcal{G}_\alpha|\}.$$

Proof. We show that the claimed bound can be achieved for any given $\alpha \ge 0$. To do so, we run Algorithm 1 over $\mathcal{G}_\alpha$ with uniform prior and denote by $\hat{y}_t$ such predictor. By inspecting the proof of Lemma 3 (and Lemma 2), we know that it holds for the fully sequential functions as well. Using the same argument as in the proof of Theorem 1, we have $r_T^a(\mathcal{H}) \le \alpha T + \log|\mathcal{G}_\alpha|$. □

We now bound the size of sequential dominant sets through an extended notion of the sequential fat-shattering. Let $\tau : \{0,1\}^{\le d} \to (\mathcal{X} \times \{0,1\})^* \times \mathcal{X}$ be a $(\mathcal{X} \times \{0,1\})^* \times \mathcal{X}$-valued full binary tree such that for any $\epsilon^d \in \{0,1\}^d$ and $t \in [d]$ if we denote $\tau(\epsilon^{t-1}) = (\mathbf{x}^i, y^{i-1})$ then $\tau(\epsilon^t) = (\mathbf{x}^i\mathbf{x}'^j, y^{i-1}\epsilon_t y'^{j-1})$, i.e., $\tau(\epsilon^{t-1})$ must be a *prefix* of $\tau(\epsilon^t)$. We also denote by $\mathbf{s} : \{0,1\}^{\le d} \to [0,1]$ a $[0,1]$-valued full binary tree. We say a class of fully sequential functions $\mathcal{H}$ sequential α fat-shatters τ witnessed by $\mathbf{s}$ if for any path $\epsilon^d \in \{0,1\}^d$ there exists $h \in \mathcal{H}$ for all $t \in [T]$:

1. If $\epsilon_t = 0$, then $h(\tau(\epsilon^{t-1})) \le \mathbf{s}(\epsilon^{t-1}) - \alpha$;
2. If $\epsilon_t = 1$, then $h(\tau(\epsilon^{t-1})) \ge \mathbf{s}(\epsilon^{t-1}) + \alpha$.

The extended sequential α fat-shattering number of $\mathcal{H}$ is defined as the maximum number $d(\alpha)$ such that there exist trees τ and $\mathbf{s}$ with depth $d(\alpha)$ that can be α fat-shattered by $\mathcal{H}$. The following result extends Lemma 9:

Proposition 4. *Let $\mathcal{H}$ be a class of fully sequential functions with extended sequential α-fat shattering number α^{-s}, then it admits a sequential 4α-dominate set $\mathcal{G}$ such that*

$$\log|\mathcal{G}| \le \tilde{O}(\alpha^{-s}).$$

Proof. The proof is similar to the proof of Lemma 9 (see [15, Section 6]) but needs additionally to deal with the y^T. Let $J \subset [0,1]$ be a discretization of $[0,1]$ with step size 2α, where $|J| \le 1/(2\alpha)$. For any $h \in \mathcal{H}$, we define the discretized function h' of h by $h'(\mathbf{x}^t, y^{t-1}) = \arg\min_{a \in J}|a - h(\mathbf{x}^t, y^{t-1})|$ for all $\mathbf{x}^t, y^{t-1}$. W.l.o.g., we will assume the functions in $\mathcal{H}$ take values in J. Now, for any given $\mathbf{x}^T, y^T$ and $h^* \in \mathcal{H}$, we construct a predictor Φ in the following manner. It maintains a running class of functions $\mathcal{H}^t$, with $\mathcal{H}^1 = \mathcal{H}$. At each time step t, for any $a \in J$ and $y \in \{0,1\}$, we denote

$$\mathcal{H}^t_{(a,y)} = \{f_h(\mathbf{x}'^j, y'^{j-1}) = h(\mathbf{x}^t\mathbf{x}'^j, y^{t-1}yy'^{j-1}) : h(\mathbf{x}^t, y^{t-1}) = a,\ h \in \mathcal{H}^t\}$$

i.e., the restriction of $\mathcal{H}^t$ with history $\mathbf{x}^t, y^{t-1}y$. We denote $\mathsf{FAT}(\mathcal{H})$ as the extended sequential α fat-shattering number of $\mathcal{H}$. By definition of sequential fat shattering, we have for all $a \in J$, $\sup_{y\in\{0,1\}} \mathsf{FAT}(\mathcal{H}^t_{(a,y)}) \le \mathsf{FAT}(\mathcal{H}^t)$ and there cannot be two elements $a_1 \le a_2 \in J$ with $a_2 - a_1 > 2\alpha$ such that $\mathsf{FAT}(\mathcal{H}^t_{(a_1,0)}) = \mathsf{FAT}(\mathcal{H}^t)$ and $\mathsf{FAT}(\mathcal{H}^t_{(a_2,1)}) = \mathsf{FAT}(\mathcal{H}^t)$. There can be only the following cases: (1). for all (a,y), $\mathsf{FAT}(\mathcal{H}^t_{(a,y)}) \le \mathsf{FAT}(\mathcal{H}_t) - 1$; (2). there exists $b \in J$ such that for all $a < b - 2\alpha$ we have $\mathsf{FAT}(\mathcal{H}^t_{(a,0)}) \le \mathsf{FAT}(\mathcal{H}_t) - 1$ and for all $a > b + 2\alpha$ we have $\mathsf{FAT}(\mathcal{H}^t_{(a,1)}) \le \mathsf{FAT}(\mathcal{H}_t) - 1$ (we argue by cases). At time step t, we update $\mathcal{H}^{t+1} = \mathcal{H}^t_{(h^*(\mathbf{x}^t, y^{t-1}), y_t)}$ only if $\mathsf{FAT}(\mathcal{H}^{t+1}) \le \mathsf{FAT}(\mathcal{H}^t) - 1$ and predict $\Phi(\mathbf{x}^t, y^{t-1}) = h^*(\mathbf{x}^t, y^{t-1})$, else, we retain $\mathcal{H}^{t+1} = \mathcal{H}^t$ and predict $\Phi(\mathbf{x}^t, y^{t-1}) = \frac{b+2\alpha}{1+4\alpha}$. Since $\mathsf{FAT}(\mathcal{H}) \le \alpha^{-s}$, the class $\mathcal{H}^t$ can be updated at most α^{-s} times. Now, for any time step we update $\mathcal{H}_t$, we have $\Phi(\mathbf{x}^t, y^{t-1}) = h^*(\mathbf{x}^t, y^{t-1})$. For the non-update steps t, we have

$\ell(h^*(\mathbf{x}^t, y^{t-1}), y_t) \geq \ell(\Phi(\mathbf{x}^t, y^{t-1}), y_t) - 4\alpha$. Clearly this is true if $|h^*(\mathbf{x}^t, y^{t-1}) - b| \leq 2\alpha$ by Lemma 4. Else, by Case (2), we have either $h^*(\mathbf{x}^t, y^{t-1}) \leq b$ but $y_t = 1$ or $h^*(\mathbf{x}^t, y^{t-1}) \geq b$ but $y_t = 0$. In both cases, we have $\ell(h^*(\mathbf{x}_t, y^{t-1}), y_t) \geq \ell(\Phi(\mathbf{x}^t, y^{t-1}), y_t)$. Therefore, Φ 4α-dominate h^* on $\mathbf{x}^T, y^T$. Since the running of Φ depends only on the values of h^* at updates, we have there exists $I \subset [T]$ and $\{k_t\}_{t\in I} \in J^{|I|}$ such that the running of Φ by update at step t being k_t have the same predictions as running on h^*. For any such I and $\{k_t\}_{t\in I}$, we construct the function g_I by running Φ by update at step $t \in I$ with k_t. Since the functions $h^* \in \mathcal{H}$ and $\mathbf{x}^T, y^T$ are arbitrary, the class $\mathcal{G}$ of all such g_Is is sequential 4α-dominate $\mathcal{H}$ for all $\mathbf{x}^T, y^T$. The size of $\mathcal{G}$ now follows by counting the number of pairs $I, \{k_t\}_{t\in I}$ as [15, Section 6]. □

APPENDIX D
OMITTED PROOFS OF TECHNICAL LEMMAS

In this appendix, we provide the proofs for some technical lemmas that are omitted in the main text.

Proof of Lemma 6. By Stirling approximation, for all $k \in [T/d]$, there exists a constant $C \in \mathbb{R}^+$ such that

$$B(k, T/d) \stackrel{\text{def}}{=} \binom{T/d}{k} \left(\frac{k}{T/d}\right)^k \left(1 - \frac{k}{T/d}\right)^{T/d-k} \geq C\sqrt{\frac{T/d}{k(T/d-k)}}.$$

Since $P(\mathbf{y} \mid w)$ achieves maximum at $w = k * d/T$, we have:

$$\sum_{\mathbf{y}\in\{0,1\}^{T/d}} \sup_{w\in[c_1-c_2d^{-r}, c_1+c_2d^{-r}]} p(\mathbf{y} \mid w) \geq \sum_{k=c_1T/d-c_2T/d^{r+1}}^{c_1T/d+c_2T/d^{r+1}} B(k, T/d).$$

Therefore, for each k in the above summation, we have:

$$\frac{1}{\sqrt{k(T/d-k)}} \geq \sqrt{(c_1 + c_2d^{-r})(1 - c_1 - c_2d^{-r})}d/T.$$

Therefore, the LHS of (23) is lower bounded by:

$$C\sqrt{(c_1 + c_2d^{-r})(1 - c_1 - c_2d^{-r})}\sqrt{\frac{T}{d}}\frac{2c_2}{d^r} = \Omega(\sqrt{T/d^{2r+1}})$$

for sufficiently large d. □

Proof of Lemma 11. We use the probabilistic method to construct sequence b^{2T}. To do so, we select B^{2T} uniformly at random from $\{0,1\}^{2T}$ and show that the event of the lemma happens with positive probability. For any $i < j \in [T]$, we construct an $i.i.d.$ sequence $X_0, \cdots, X_{T/2-1}$ with uniform distribution over $\{0,1\}$ such that:

$$\sum_{t=0}^{T-1} 1\{B_{i+t} \neq B_{j+t}\} \geq \sum_{t=0}^{T/2-1} X_t.$$

To do so, we maintain an index set I and a set $\mathbf{X}$ of random variables, initial $I, \mathbf{X} = \emptyset$ and $t = 0$. For each $t \in [T-1]$, if the index $i + t \in I$, we remove $i + t$ from I and continue to $t + 1$; else, we add the indicator $1\{B_{i+t} \neq B_{j+t}\}$ to $\mathbf{X}$ and $j + t$ to I and continue to $t + 1$. Clearly all the random variables in $\mathbf{X}$ are mutually independent and distributed uniformly over $\{0,1\}$, since we add the indicator to $\mathbf{X}$ only when there is no overlap on the indexes and B^{2T} are i.i.d. random variables (notice that if $i + t \notin I$ then $j + t$ also does not appear in the previous indexes, since $i < j$). We claim that $|I| \geq T/2$. This follows from the fact that the conflict period must not be more than the non-conflict period. The lemma now follows by Chernoff bound, to show that $\sum_{t=0}^{T/2-1} X_t$ tightly concentrates on $T/4$ and a union bound on all pairs (i, j). □

APPENDIX E
ALTERNATIVE PROOF OF THEOREM 2

We now provide an alternative proof of Theorem 2 without relying on the construction of a cover of $\mathcal{B}_s^d(R)$. The proof follows similar path as in the proof of Theorem 4. Let μ be the Lebesgue measure over $\mathcal{B}_s^d(R+\alpha/L)$, where α is to be determined latter. Let f be the L-Lipschitz function as in Theorem 2 over $\mathcal{B}_s^d(R)$ for all $\mathbf{x} \in \mathbb{R}^d$. By [45, Theorem 1], we can extend f to $\mathcal{B}_s^d(R+\alpha/L)$ while keeping the L-Lipschitz property by defining for all $\mathbf{w} \in \mathcal{B}_s^d(R+\alpha/L)$ and $\mathbf{x} \in \mathbb{R}^d$

$$f(\mathbf{w}, \mathbf{x}) = \sup_{\mathbf{w}' \in \mathcal{B}_s^d(R)} \{f(\mathbf{w}', \mathbf{x}) - L||\mathbf{w} - \mathbf{w}'||_s\}.$$

Denote by $\hat{y}_t$ the predictor by running Algorithm 2 with prior μ over $\mathcal{H}_f$ (with the extended function f) and truncation parameter α. By Lemma 3, we have for any $\mathbf{x}^T, y^T$ that

$$\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) \le -\log \frac{\int_{\mathcal{B}_s^d(R+\alpha/L)} \tilde{p}_{\mathbf{w}}(y^T \mid \mathbf{x}^T) \mathrm{d}\mu}{\int_{\mathcal{B}_s^d(R+\alpha/L)} 1 \mathrm{d}\mu},$$

where

$$\tilde{p}_{\mathbf{w}}(y^T \mid \mathbf{x}^T) = \prod_{t=1}^{T} \tilde{f}(\mathbf{w}, \mathbf{x}_t)^{y_t} (1 - \tilde{f}(\mathbf{w}, \mathbf{x}_t))^{1-y_t}$$

and $\tilde{f} = \frac{f+\alpha}{1+2\alpha}$. Let $\mathbf{w}^* = \arg\max_{\mathbf{w} \in \mathcal{B}_s^d(R)} \{p_{\mathbf{w}}(y^T \mid \mathbf{x}^T)\}$, where $p_{\mathbf{w}}(y^T \mid \mathbf{x}^T) = \prod_{t=1}^{T} f(\mathbf{w}, \mathbf{x}_t)^{y_t} (1 - f(\mathbf{w}, \mathbf{x}_t))^{1-y_t}$. Denote by B the ℓ_s ball centered at $\mathbf{w}^*$ with radius α/L. Note that $B \subset \mathcal{B}_s^d(R+\alpha/L)$. By L-Lipschitz property, for any $\mathbf{w} \in B$ and $\mathbf{x} \in \mathbb{R}^d$ we have $|f(\mathbf{w}, \mathbf{x}) - f(\mathbf{w}^*, \mathbf{x})| \le \alpha$. By Lemma 4, this implies that for any $\mathbf{w} \in B$ we have $p_{\mathbf{w}^*}(y^T \mid \mathbf{x}^T) \le (1+2\alpha)^T \tilde{p}_{\mathbf{w}}(y^T \mid \mathbf{x}^T)$. Therefore,

$$\begin{aligned}
\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) &\le -\log \frac{\int_{\mathcal{B}_s^d(R+\alpha/L)} \tilde{p}_{\mathbf{w}}(y^T \mid \mathbf{x}^T) \mathrm{d}\mu}{\int_{\mathcal{B}_s^d(R+\alpha/L)} 1 \mathrm{d}\mu} \\
&\le -\log \frac{\int_B \tilde{p}_{\mathbf{w}}(y^T \mid \mathbf{x}^T) \mathrm{d}\mu}{\int_{\mathcal{B}_s^d(R+\alpha/L)} 1 \mathrm{d}\mu}, \quad \text{since } B \subset \mathcal{B}_s^d(R+\alpha/L) \\
&\le -\log \frac{e^{-2\alpha T} \int_B p_{\mathbf{w}^*}(y^T \mid \mathbf{x}^T) \mathrm{d}\mu}{\int_{\mathcal{B}_s^d(R+\alpha/L)} 1 \mathrm{d}\mu}, \\
&\le -\log p_{\mathbf{w}^*}(y^T \mid \mathbf{x}^T) + 2\alpha T + \log \frac{\mathsf{Vol}(\mathcal{B}_s^d(R+\alpha/L))}{\mathsf{Vol}(\mathcal{B}_s^d(\alpha/L))} \\
&\le \sum_{t=1}^{T} \ell(f(\mathbf{w}^*, \mathbf{x}), y_t) + 2\alpha T + d \log\left(\frac{RL}{\alpha} + 1\right).
\end{aligned}$$

Taking $\alpha = d/T$, we find

$$r_T^a(\mathcal{H}_f) \le d \log\left(\frac{RLT}{d} + 1\right) + 2d$$

as needed.

REFERENCES

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

[2] A. Rakhlin and K. Sridharan, "Sequential probability assignment with binary alphabets and large classes of experts," *arXiv preprint arXiv:1501.07340*, 2015.

[3] D. J. Foster, S. Kale, H. Luo, M. Mohri, and K. Sridharan, "Logistic regression: The importance of being improper," in *Conference on Learning Theory (COLT)*, 2018.

[4] G. I. Shamir, "Logistic regression regret: What's the catch?" in *Conference on Learning Theory (COLT)*. PMLR, 2020, pp. 3296–3319.

[5] M. Drmota and W. Szpankowski, "Precise minimax redundancy and regrets." *IEEE Trans. Information Theory*, vol. IT-50, pp. 2686–2707, 2004.

[6] A. Orlitsky and N. P. Santhanam, "Speaking of infinity," *IEEE Trans. Information Theory*, vol. 50, no. 10, pp. 2215–2230, Oct. 2004.

[7] J. Rissanen, "Fisher information and stochastic complexity," *IEEE Trans. Information Theory*, vol. 42, pp. 40–47, 1996.

[8] G. I. Shamir, "On the MDL principle for i.i.d. sources with large alphabets," *IEEE Trans. Information Theory*, vol. 52, no. 5, pp. 1939–1955, May 2006.

[9] W. Szpankowski, "On asymptotics of certain recurrences arising in universal coding," *Problems of Information Transmission*, vol. 34, pp. 55–61, 1998.

[10] Q. Xie and A. Barron, "Minimax redundancy for the class of memoryless sources," *IEEE Trans. Information Theory*, pp. 647–657, 1997.

[11] ——, "Asymptotic minimax regret for data compression, gambling, and prediction," *IEEE Trans. Information Theory*, vol. 46, pp. 431–445, 2000.

[12] C. Wu, M. Heidari, A. Grama, and W. Szpankowski, "Sequential vs. fixed design regrets in online learning," in *IEEE International Symposium on Information Theory (ISIT)*, 2022, pp. 438–443.

[13] P. Jacquet, G. I. Shamir, and W. Szpankowski, "Precise minimax regret for logistic regression with categorical feature values," in *Algorithmic Learning Theory (ALT)*, vol. 132, 2021, pp. 755–771.

[14] Y. M. Shtarkov, "Universal sequential coding of single messages," *Problems of Information Transmission*, vol. 23, no. 3, pp. 3–17, Jul.-Sep. 1987.

[15] A. Rakhlin, K. Sridharan, and A. Tewari, "Online learning: Random averages, combinatorial parameters, and learnability," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 23, 2010.

[16] S. Ben-David, D. Pál, and S. Shalev-Shwartz, "Agnostic online learning." in *Conference on Learning Theory (COLT)*, vol. 3, 2009.

[17] B. Bilodeau, D. Foster, and D. Roy, "Tight bounds on minimax regret under logarithmic loss via self-concordance," in *International Conference on Machine Learning (ICML)*. PMLR, 2020, pp. 919–929.

[18] P. Jacquet, G. I. Shamir, and W. Szpankowski, "Precise minimax regret for logistic regression," in *IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2022, pp. 444–449.

[19] E. Hazan, T. Koren, and K. Y. Levy, "Logistic regression: Tight bounds for stochastic and online optimization," in *Conference on Learning Theory (COLT)*. MIT press, 2014, pp. 197–209.

[20] S. M. Kakade and A. Y. Ng, "Online bounds for bayesian algorithms," in *Advances in Neural Information Processing Systems (NeurIPS)*, L. K. Saul, Y. Weiss, and L. Bottou, Eds. MIT Press, 2005, pp. 641–648.

[21] R. Jézéquel, P. Gaillard, and A. Rudi, "Mixability made efficient: Fast online multiclass logistic regression," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021.

[22] V. Vovk, "Competitive on-line statistics," *International Statistical Review*, vol. 69, no. 2, pp. 213–248, 2001.

[23] N. Cesa-Bianchi and G. Lugosi, "Worst-case bounds for the logarithmic loss of predictors," *Machine Learning*, vol. 43, no. 3, pp. 247–264, 2001.

[24] A. Rakhlin and K. Sridharan, "Online nonparametric regression with general loss function," in *Conference on Learning Theory (COLT)*, 2014.

[25] B. Bilodeau, D. J. Foster, and D. M. Roy, "Minimax rates for conditional density estimation via empirical entropy," *arXiv preprint arXiv:2109.10461*, 2021.

[26] N. Merhav and M. Feder, "Universal prediction," *IEEE Trans. on Information Theory*, vol. 44, no. 6, pp. 2124–2147, 1998.

[27] L. D. Davisson, "Universal noiseless coding," *IEEE Trans. Information Theory*, vol. IT-19, no. 6, pp. 783–795, Nov. 1973.

[28] W. Szpankowski and M. Weinberger, "Minimax pointwise redundancy for memoryless models over large alphabets," *IEEE Trans. Information Theory*, vol. 58, pp. 4094–4104, 2012.

[29] I. Csiszar and P. Shields, "Redundancy rates for renewal and other processes," *IEEE Trans. Information Theory*, vol. 42, pp. 2065–2072, 1995.

[30] P. Flajolet and W. Szpankowski, "Analytic variations on redundancy rates of renewal processes," *IEEE Trans. Information Theory*, vol. 48, pp. 2911–2921, 2002.

[31] B. Clarke and A. Barron, "Jeffreys' prior is asymptotically least favorable under entropy of risk," *J. Statistical Planning and Inference*, pp. 453 – 471, 1994.

[32] A. Barron, J. Rissanen, and B. Yu, "The minimum description length principle in coding and modeling," *IEEE Trans. Information Theory*, vol. 44, no. 6, pp. 2743–2760, Oct. 1998.

[33] A. Barron and X. Luo, "Mdl procedures with l1 penalty and their statistical risk," in *Proceedings of the 2008 workshop on information theoretic methods in science and engineering*, 2008.

[34] K. Yamanishi, "Minimax relative loss analysis for sequential prediction algorithms using parametric hypotheses," in *Conference on Learning Theory (COLT)*, 1998, pp. 32–43.

[35] Y. Fogel and M. Feder, "Universal supervised learning for individual data," *arXiv preprint arXiv:1812.09520*, 2018.

[36] ——, "Universal learning of individual data," in *IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2019, pp. 2289–2293.

[37] A. Bhatt and Y.-H. Kim, "Sequential prediction under log-loss with side information," in *Algorithmic Learning Theory (ALT)*. PMLR, 2021, pp. 340–344.

[38] Y. Shkel, M. Raginsky, and S. Verdú, "Sequential prediction with coded side information under logarithmic loss," in *Algorithmic Learning Theory (ALT)*. PMLR, 2018, pp. 753–769.

[39] M. Feder and Y. Polyanskiy, "Sequential prediction under log-loss and misspecification," in *Conference on Learning Theory (COLT)*. PMLR, 2021, pp. 1937–1964.

[40] J. Jiao, T. A. Courtade, K. Venkat, and T. Weissman, "Justification of logarithmic loss via the benefit of side information," *IEEE Trans. on Information Theory*, vol. 61, no. 10, pp. 5357–5365, 2015.
[41] F. Hedayati and P. L. Bartlett, "Exchangeability characterizes optimality of sequential normalized maximum likelihood and bayesian prediction," *IEEE Trans. on Information Theory*, vol. 63, no. 10, pp. 6767–6773, 2017.
[42] C. Wu, M. Heidari, A. Grama, and W. Szpankowski, "Precise regret bounds for log-loss via a truncated bayesian algorithm," *arXiv preprint arXiv:2205.03728; To appear at NeurIPS 2022*, 2022.
[43] G. I. Shamir and W. Szpankowski, "Low complexity approximate bayesian logistic regression for sparse online learning," *arXiv preprint arXiv:2101.12113*, 2021.
[44] M. J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, 2019, vol. 48.
[45] E. J. McShane, "Extension of range of functions," *Bulletin of the American Mathematical Society*, vol. 40, no. 12, pp. 837–842, 1934.
[46] X. Wang, "Volumes of generalized unit balls," *Mathematics Magazine*, vol. 78, no. 5, pp. 390–395, 2005.
[47] J. J. Mayo, H. Hadiji, and T. van Erven, "Scale-free unconstrained online learning for curved losses," *arXiv preprint arXiv:2202.05630*, 2022.
[48] A. Rakhlin, K. Sridharan, and A. Tewari, "Online learning via sequential complexities." *J. Machine Learning Research*, vol. 16, no. 1, pp. 155–186, 2015.
[49] ——, "Sequential complexities and uniform martingale laws of large numbers," *Probability Theory and Related Fields*, vol. 161, no. 1, pp. 111–153, 2015.
[50] T. Kerdreux, A. d'Aspremont, and S. Pokutta, "Local and global uniform convexity conditions," *arXiv preprint arXiv:2102.05134*, 2021.
[51] L.-A. Gottlieb, A. Kontorovich, and R. Krauthgamer, "Efficient regression in metric spaces via approximate lipschitz extension," *IEEE Trans. on Information Theory*, vol. 63, no. 8, pp. 4838–4849, 2017.
[52] S. M. Kakade and A. Ng, "Online bounds for bayesian algorithms," *Advances in neural information processing systems (NeurIPS)*, vol. 17, 2004.
[53] C. Wu, M. Heidari, A. Grama, and W. Szpankowski, "Expected worst case regret via stochastic sequential covering," *arXiv preprint arXiv:2209.04417*, 2022.