Sequential vs. Fixed Design Regrets in Online Learning

Changlong Wu
CSoI
Purdue University
Email: wuchangl@hawaii.edu

Mohsen Heidari
CSoI
Purdue University
Email: mheidari@purdue.edu

Ananth Grama
CSoI
Purdue University
Email: ayg@cs.purdue.edu

Wojciech Szpankowski
CSoI
Purdue University
Email: szpan@purdue.edu

***Abstract*—In source coding since Davisson's seminal paper [1] various redundancy and regrets were thoroughly analyzed, from pointwise redundancy, to average and maximal minimax and maxmin regrets. Similarly, in online learning, there are various formulations of regrets that are grouped into *fixed-design* (when data is known in advance) and *sequential*. This position paper gives a brief overview of current formulations of regrets, and provides a thorough comparison of the sequential and fixed design formulations. Moreover, inspired by the source coding literature, new classes of regrets, from average to worst case minimax, are introduced. In particular, it is shown that the fixed design and sequential regrets are equal in the worst case and average sense when data is known in advance; but, in maximal sense (when maximizing over data), the former can be significantly smaller than the latter. Specifically, this paper proves that under *logarithmic* loss (i) for linear predictors the two maximal formulations are of the same order; and (ii) for linear threshold predictors, fixed design maximal regret is logarithmically smaller than the sequential one.**

I. Introduction

In universal source coding of information theory the goal is to find the best (shortest) description of a sequence generated by a source from a class of sources $\mathcal{S}$. Since the seminal paper of Davisson [1] the quality of universal compression is measured by one of many minimax redundancies which are defined as the excess of the real compression over the optimal one, either on average or for the worst case. This program led to several deep results attesting the performance of universal compression algorithms. Prior work in information theory such as [1]–[8] established numerous techniques and result. Furthermore, relationships between different minimax redundancies/ regrets were comprehensively analyzed in [9].

The situation in learning is more complex since we must deal not only with the label sequence $y^T = (y_1, \ldots, y_T)$ but also data $\mathbf{x}^T = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$ [10], [11]. In the online learning, the training algorithm consumes d dimensional data $\mathbf{x}^T$ in rounds where at each round $t \in \{1, 2, \ldots, T\}$, it is allowed to predict the label $\hat{y}_t$ based only on the labels it observed in the past $t-1$ rounds. Then the true label y_t is revealed and the loss $\ell(y_t, \hat{y}_t)$ is incurred. The (pointwise) *regret* of an online algorithm is defined as the (excess) loss it incurs over some value of a constant *comparator* that is used for prediction of the complete sequence. Several minimax regrets can be defined in the fixed design (when the whole feature data is known in advance) or in the sequential scenario where optimizations are performed online. Furthermore, in machine learning we also optimize over a concept class that represents possible predictors. This unattainable situation begs for better understanding the relationships between various regrets leading to a generalization of [9] to the machine learning situation.

a) Regrets in Information Theory: We now briefly review various notions of regret and redundancy from universal source coding of information theory. The *pointwise redundancy* $R_T(P; y^T)$ and the *average redundancy* $\bar{R}_T(P)$ for a *given* source P and source (label) sequence $y^T = (y_1, \ldots, y_T)$ of length T over alphabet of size m are defined as $R_T(P; y^T) = L(y^T) + \log P(y^T)$ and $\bar{R}_T(P) = \mathbf{E}[L(Y^T)] - H_T(P)$ where $H_T(P)$ is the entropy for a block of length T, $\mathbf{E}$ denotes the expectation, and $L(y^T)$ is the loss of some algorithm that predicts y^T. The loss can be thought of $L(y^T) = -\log Q(y^T)$ for some unknown distribution Q that approximates P. In the worst case setting of *individual sequences* [4], we define the *maximal or worst case regret* as

$$R_T^*(Q, P) = \max_{y^T} [-\log Q(y^T) + \log P(y^T)].$$

In universal sources coding we often do not know P but only a class of sources $\mathcal{S}$ to which P belongs. Following Davisson [1], we analyze the minimax *regrets* defined as follows (cf. [7]–[9]):

$$\bar{r}_T(\mathcal{S}) = \min_Q \sup_{P \in \mathcal{S}} \mathbf{E}_P[-\log Q(y^T) + \log \sup_{P \in \mathcal{S}} P(y^T)],$$
$$r_T^*(\mathcal{S}) = \min_Q \max_{y^T} [-\log Q(y^T) + \log \sup_{P \in \mathcal{S}} P(y^T)].$$

The big question is how these different regrets are related for different classes of sources $\mathcal{S}$. In [9] it is shown that if the maximum likelihood distribution belongs to the convex hull of $\mathcal{S}$, then $r_T^*(\mathcal{S}) - \bar{r}_T(\mathcal{S}) = O(c_T(\mathcal{S}))$ where

$$c_T(\mathcal{S}) = \sum_{y^T} P(y^T) \log \left(\sup_{P \in \mathcal{S}} P(y^T)/P(y^T) \right).$$

It was also proved that $r_T^*(\mathcal{S}) = \bar{r}_T(\mathcal{S}) + O(1)$ provided the maximum likelihood distribution belongs to the convex hull of $\mathcal{S}$ (e.g., for sources with finite memory such as Markovian sources). It is known [3], [5], [7]–[9], [12], [13] that for a large class of sources the redundancy grows as $\frac{m-1}{2} \log T$ when the alphabet size m is fixed and $\frac{m-1}{2} \log(T/m)$ for $m = o(T)$ (see [3], [6], [12]).

b) Regret in Online Learning: We phrase our online learning problem in terms of a game between nature/ environment and a learner/predictor. Broadly, the learner's objective is to process past observations to predict the next realization of nature's labeling sequence. At each round $t \in \mathbb{N}$, let y_t be the true label yet to be revealed. At each time t, the learner obtains a d dimensional input/ feature vector $\mathbf{x}_t \in \mathbb{R}^d$. In addition to $\mathbf{x}_t$, the learner may use the past observations $(\mathbf{x}_r, y_r)$, $r < t$ to make a prediction $\hat{y}_t$ of the true label. Therefore, the prediction can be written as $\hat{y}_t = g_t(y^{t-1}, \mathbf{x}^t)$, where g_t represents the strategy of the learner to obtain its prediction based on the past and current observations. Hence, the learner is modeled by the sequence of predicting actions $g_t, t > 0$. Once a prediction is made, the nature reveals the true label y_t and the learner incurs some *loss* evaluated based on a predefined function $\ell : \hat{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}$, where $\hat{\mathcal{Y}}$ and $\mathcal{Y}$ are the prediction and label domains respectively.

In regret analysis, we are interested in comparing the accumulated loss of the learner with that of the best strategy within a predefined class of predictors (experts) denoted as $\mathcal{H}$. More precisely, $\mathcal{H}$ is a collection of predicting functions $h : \mathbb{R}^d \mapsto \hat{\mathcal{Y}}$, with input being $\mathbf{x}_t$ at each time t. Therefore, given a learner $g_t, t > 0$ and after T rounds with the realizations $(y_t, \mathbf{x}_t)_{t=1}^T$, the *pointwise regret* is defined as

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

where $\hat{y}_t = g_t(y^{t-1}, \mathbf{x}^t)$, and the first and the second summations represent the accumulated loss of the learner and the best predictor in $\mathcal{H}$, respectively. There are two main perspectives on analyzing the regret, highlighted next.

Fixed Design: This point of view studies the minimal regret for the worst realization of the label with the feature vector $\mathbf{x}^T$ known in advance. Let $g_t, t > 0$ be the strategy of the predictor. Then, the *fixed design minimax regret* is defined as

$$r_T^*(\mathcal{H}|\mathbf{x}^T) = \inf_{g^T} \sup_{y^T} R(g^T, y^T, \mathcal{H}|\mathbf{x}^T). \tag{1}$$

Further, the fixed design *maximal* minimax regret is given by

$$r_T^*(\mathcal{H}) = \sup_{\mathbf{x}^T} \inf_{g^T} \sup_{y^T} R(g^T, y^T, \mathcal{H}|\mathbf{x}^T). \tag{2}$$

Note that this notion was also discussed in the literature under the name of *Transductive Online Learning*, see [14].

Sequential Design: In this point of view, the optimization on regret is performed at every time t without knowing in advance $\mathbf{x}^T$ or y^T. Then the *sequential minimax regret* is defined as

$$r_T^a(\mathcal{H}|\mathbf{x}^T) = \inf_{\hat{y}_1} \sup_{y_1} \cdots \inf_{\hat{y}_T} \sup_{y_T} R(\hat{y}^T, y^T, \mathcal{H}|\mathbf{x}^T).^{1} \tag{3}$$

Moreover, the *sequential maximal minimax regret* is [15]

$$r_T^a(\mathcal{H}) = \sup_{\mathbf{x}_1} \inf_{\hat{y}_1} \sup_{y_1} \cdots \sup_{\mathbf{x}_T} \inf_{\hat{y}_T} \sup_{y_T} R(\hat{y}^T, y^T, \mathcal{H}|\mathbf{x}^T). \tag{4}$$

The sequential regret should be interpreted as follows. At round $t = 1$, the adversarial chooses a $\mathbf{x}_1 \in \mathbb{R}^d$ that gives the worst regret for the best choice of $\hat{y}_1$ for the worst choice of y_1 and for the interleaved worst, best, worst choices of $\mathbf{x}_t$, $h(\cdot)$ to compute $\hat{y}_t$ and y_t, respectively. Then, for the given $\mathbf{x}_1$, and subsequent worst/best choices, the player chooses $\hat{y}_1$, and the adversary chooses the worst choice of y_1.

Average Regrets: We also introduce the averaged variants of regrets. Let D be a joint probability distribution on $(\mathbb{R}^d, \mathcal{Y})^T$. Then the fixed design average minimax regret is

$$\bar{r}_T(\mathcal{H}, D|\mathbf{x}^T) = \inf_{g^T} \mathbb{E}_{Y^T \sim D_{Y^T|\mathbf{x}^T}} R(g^T, Y^T, \mathcal{H}|\mathbf{x}^T).$$

Similarly, the sequential average minimax regret is

$$\bar{r}_T^a(\mathcal{H}, D|\mathbf{x}^T) = \inf_{\hat{y}_1} \mathbb{E}_{Y_1} \cdots \inf_{\hat{y}_T} \mathbb{E}_{Y_T|\mathbf{x}^{T-1}y^{T-1}} R(\hat{y}^T, Y^T, \mathcal{H}|\mathbf{x}^T).$$

c) Main Contributions: In the next section, we present our main results establishing relationships between fixed design and sequential minimax regrets for different hypothesis classes and loss functions. Throughout, we assume binary labels y_t and bounded features $\mathbf{x}_t$ living in $\mathbb{R}^d$. We prove in Theorem 1 that the sequential and fixed design regrets are equal in two cases when data is known in advance: (i) for the worst case formulation $r_T^a(\mathcal{H}|\mathbf{x}^T) = r_T^*(\mathcal{H}|\mathbf{x}^T)$; and (ii) for the average sense $\bar{r}_T(\mathcal{H}, D|\mathbf{x}^T) = \bar{r}_T^a(\mathcal{H}, D|\mathbf{x}^T)$. Moreover, maximizing over $\mathbf{x}^T$ in a fixed design scenario gives us a universal lower bound for the sequential minimax regret, that is, $r_T^a(\mathcal{H}) \geq r_T^*(\mathcal{H})$. Thus a question arises as to whether this lower bound is achievable or not. It turns out that the answer depends on the hypothesis classes. In Theorem 2 we prove that for the logarithmic loss and linear class functions these two regrets are asymptotically equivalent. But, we show that for linear threshold predictors, the fixed design maximal regret is logarithmically smaller than the sequential one (see Theorem 3) under log-loss. More precisely, $r_T^a(\mathcal{H}_a) \geq T$, while $r_T^*(\mathcal{H}_a) \leq \log(T+1)$. Our main results are presented in Section II with some proofs in Section III.

II. Main Results

In this section, we present our main results in two categories: First, we discuss general relationships between the maximal and average fixed design and sequential minimax regrets. Then we focus on certain hypothesis classes $\mathcal{H}$ for which we present some precise asymptotics indicating that in some cases the maximal minimax regret $r_T^*(\mathcal{H})$ and the sequential minimax regret $r_T^a(\mathcal{H})$ are asymptotically equal while for other classes one can only assert that $r_T^a(\mathcal{H}) > r_T^*(\mathcal{H})$.

We first briefly review main techniques used in this context. There are some standard techniques to derive bounds on these two regrets. For example, when the loss function is $\ell(\hat{y}, y) = |y - \hat{y}|$ (generalized error) for any class $\mathcal{H}$ of binary-valued functions, one can show [14], [16]

$$r_T^*(\mathcal{H}) = O(\sqrt{\mathrm{VC}(\mathcal{H})T}), \quad r_T^a(\mathcal{H}) = O(\sqrt{\mathrm{Ldim}(\mathcal{H})T}),$$

where $\mathrm{VC}(\mathcal{H})$ and $\mathrm{Ldim}(\mathcal{H})$ are the Vapnik–Chervonenkis (VC) and the Littlestone dimensions of $\mathcal{H}$, respectively.

In addition to such techniques, we are aware of two main methods to analyze the regrets: (i) the *Sequential Rademacher*

[1] Here we abused $\hat{y}_t$ to denote a number instead of a function.

complexity [15] for the sequential regrets; and (ii) *Shtarkov Sum* for the fixed design regrets. The first approach is typically used to analyze $r_T^a(\mathcal{H})$ and provides bounds on the regrets for general predictor classes. Namely, for $\mathcal{H}$ being a class of binary valued functions with the absolute loss $\ell(\hat{y}, y) = |y - \hat{y}|$, the sequential regret is bounded as

$$\mathscr{R}_T^a(\mathcal{H}) \leq r_T^a(\mathcal{H}) \leq 2\mathscr{R}_T^a(\mathcal{H})$$

where $\mathscr{R}_T^a(\mathcal{H})$ is the Sequential Rademacher complexity of $\mathcal{H}$, see [15, Definition 5]. While this approach is quite successful in characterizing the regrets under absolute loss (and similarly convex bounded losses), it can be quite loose in characterizing unbounded losses [17], e.g., logarithmic loss of the form $\ell(P, \mathbf{x}^T) = -\log P(y^T|\mathbf{x}^T, \mathcal{H})$.

The second approach (Shtarkov sum [4], [9]) typically leads to precise characterizations for logarithmic losses [18]. Particularly, for any class $\mathcal{H}$ of functions with image in $[0, 1]$ and with logarithmic loss, we have that

$$r_T^*(\mathcal{H}) = \sup_{\mathbf{x}^T} \log S_T(\mathcal{H}|\mathbf{x}^T) = \sup_{\mathbf{x}^T} \log \sum_{y^T} \sup_{h\in\mathcal{H}} P_h(y^T|\mathbf{x}^T),$$

where $S_T(\mathcal{H}|\mathbf{x}^T)$ is the Shtarkov sum of $\mathcal{H}$ conditioned on $\mathbf{x}^T$ [18]. However, the Shtarkov sum based approach can only address fixed design regrets. The main objective of this paper is to investigate the relationships between the two notions of the regrets under different losses, and develop general approaches for analyzing the regrets.

Now we present our main results. We start with relationships between the regrets for a general hypothesis class $\mathcal{H}$ and a general loss function ℓ. The following relations between different minimax regrets hold.

Theorem 1 (General Relations)**.** *Let $\mathcal{H}$ be any general hypothesis class and ℓ be any loss function. Then $r_T^a(\mathcal{H}|\mathbf{x}^T) = r_T^*(\mathcal{H}|\mathbf{x}^T)$ and $\bar{r}_T^a(\mathcal{H}, D|\mathbf{x}^T) = \bar{r}_T(\mathcal{H}, D|\mathbf{x}^T)$ for any $\mathbf{x}^T \in \mathbb{R}^{dT}$ and distribution D. Furthermore, $r_T^a(\mathcal{H}) \geq r_T^*(\mathcal{H})$, and the inequality is strict for certain $\mathcal{H}$, and loss function ℓ.*

We now consider a few hypothesis classes and loss functions to assess the strength of the inequality in Theorem 1.

Logarithmic loss: Let $f : \mathbb{R} \to [0, 1]$ be a L-Lipschitz function such that there exist constant c, such that for all $B \in [-1, 1]$, $f(x)$ takes the full range of $[\frac{1}{2} - cB, \frac{1}{2} + cB] \subset [0, 1]$ with $x \in [-B, B]$. We denote $\mathcal{B}_2^d$ to be unit ball of dimension d with ℓ_2 norm. In this subsection we consider the following hypothesis class

$$\mathcal{H}_{f,\mathbf{w}} = \{h : \mathbb{R}^d \to \mathbb{R} : \ h(\mathbf{x}) = f(\langle \mathbf{w}, \mathbf{x}\rangle) : \mathbf{w}, \mathbf{x} \in \mathcal{B}_2^d\},$$

where $\mathbf{w}$ is a d dimensional weight vector and $\langle \mathbf{w}, \mathbf{x}\rangle$ is the scalar product. For any $h \in \mathcal{H}$, we interpret $h(\mathbf{x}) \in [0, 1]$ as the probability assigned to $Y = 1$. Therefore, the logarithmic loss function corresponding to $h(\mathbf{x})$ equals to $\ell(h(\mathbf{x}), y) = -\log|1 - h(\mathbf{x}) - y|$, for all $y \in \{0, 1\}$. We just observe that this loss function is equivalent to the logarithmic loss function of the correct classification probability, that is, $\ell(h(\mathbf{x}), y) = -\log P(Y = y|\mathbf{x})$.

Theorem 2. *If $d = O(T^{1/2-\varepsilon})$, then the regret for the hypothesis class $\mathcal{H}_{f,\mathbf{w}}$ and the log-loss function satisfies*

$$r_T^*(\mathcal{H}_{f,\mathbf{w}}) = \Theta(d \log T), \qquad r_T^a(\mathcal{H}_{f,\mathbf{w}}) = \Theta(d \log T).$$

In summary, $r_T^*(\mathcal{H}_{f,\mathbf{w}}) \asymp r_T^a(\mathcal{H}_{f,\mathbf{w}})$.

Theorem 2 shows that the $d \log T$ upper bound holds for any L-Lipschitz function f, while in [19] the author considers only the case when the second derivative of $\log(1/f(x))$ is bounded (see [19, Equation (19)])[2].

Lastly, we end this section with studying regrets for linear threshold predictors. In particular, we can prove our third main result in the following theorem. The proof is rather standard and left to the reader due to the lack of space.

Theorem 3. *Let $\mathcal{H}_a = \{h_a : h_a(x) = 1\{x \geq a\}, a, x \in \mathbb{R}\}$ be the class of all linear threshold functions. If $\ell(y_t, \hat{y}_t) = |y_t - \hat{y}_t|$, i.e., the absolute loss, then*

$$r_T^a(\mathcal{H}_a) \geq \frac{T}{2}, \text{ and } r_T^*(\mathcal{H}_a) \leq O(\sqrt{T \log T}).$$

If $\ell(y_t, \hat{y}_t) = -\log|1 - y_t - \hat{y}_t|$, i.e, the logarithmic loss, then $r_T^a(\mathcal{H}_a) = T$ and $r_T^*(\mathcal{H}_a) = \log(T + 1)$.

The above results lead us to the following open problem:

Problem 1 (Open Problem)**.** *Under what conditions on the class $\mathcal{H}$ do we have $r_T^a(\mathcal{H}) \asymp r_T^*(\mathcal{H})$ for the log-loss? In particular, suppose the fat-shattering number and sequential shattering number as in [15] are of the same order, do we have $r_T^a(\mathcal{H}) \leq \text{poly}\log(T) r_T^*(\mathcal{H})$?*

We should remark that without any restriction on $\mathcal{H}$, for any function $f(T) \leq T$, we can construct a class $\mathcal{H}$ such that $r_T^*(\mathcal{H}) \leq O(\log T)$ but $r_T^a(\mathcal{H}) \geq f(T)$ for infinitely many T and $r_T^a(\mathcal{H} \leq O(f(T))$. This is achieved by choosing $\mathcal{H}$ such that the sequential fat-shattering number scales much faster than the fat-shattering number (by e.g., embedding the linear threshold functions).

III. Proof of The Main Results

A. Proof of Theorem 1

We start with the following lemma.

Lemma 1. *Let A, B be two sets, and $f : A \times B \to \mathbb{R}$ be an arbitrary function, then*

$$\inf_{g\in\mathcal{G}} \sup_{b\in B} f(g(b), b) = \sup_{b\in B} \inf_{a\in A} f(a, b),$$

where $\mathcal{G} = A^B$ is the class of all functions from $B \to A$.

Proof: The proof is similar to the idea in the proof of [15, Lemma 3]. For any b, we define function $\hat{g}(b) = \arg\min_{a\in A} f(a, b)$. We have $\sup_{b\in B} f(\hat{g}(b), b) = \sup_{b\in B} \inf_{a\in A} f(a, b)$. Thus, we have LHS $\leq$ RHS. To see the converse, let $g^* = \arg\min_{g\in G}(\sup_{b\in B} f(g(b), b))$. We have $\sup_{b\in B} f(g^*(b), b) \geq \sup_{b\in B} \inf_{a\in A} f(a, b)$ since in the

[2]Note that the result in [20] also implies a similar upper bound, but has a worse leading constant.

former term the first argument in $f(a,b)$ is fixed while the second term has an infimum. This implies LHS $\geq$ RHS. ■

The first part of Theorem 1 follows from the next lemma.

Lemma 2. *For any function $f : \hat{\mathcal{Y}}^T \times \mathcal{Y}^T \to \mathbb{R}$, we have*

$$\inf_{g_1 \in \mathcal{G}_1, \cdots, g_T \in \mathcal{G}_T} \sup_{y^T \in \mathcal{Y}^T} f(g_1(\emptyset), g_2(y_1), \cdots, g_T(y^{T-1}), y^T)$$
$$= \inf_{\hat{y}_1 \in \hat{\mathcal{Y}}} \sup_{y_1 \in \mathcal{Y}} \cdots \inf_{\hat{y}_T \in \hat{\mathcal{Y}}} \sup_{y_T \in \mathcal{Y}} f(\hat{y}^T, y^T),$$

where $\emptyset$ is empty string.

Proof: We prove the lemma by induction on T, the case for $T = 0$ is trivial since g_1 is simply an element in $\hat{\mathcal{Y}}$. Note that $\inf_{\hat{y}_T} \sup_{y_T} f(\hat{y}^T, y^T)$ can be viewed as a function over $\hat{\mathcal{Y}}^{T-1} \times \mathcal{Y}^{T-1}$. By induction hypothesis for $T-1$, we have

$$\inf_{\hat{y}_1 \in \hat{\mathcal{Y}}} \sup_{y_1 \in \mathcal{Y}} \cdots \inf_{\hat{y}_T \in \hat{\mathcal{Y}}} \sup_{y_T \in \mathcal{Y}} f(\hat{y}^T, y^T) =$$
$$\inf_{g^{T-1}} \sup_{y^{T-1}} \inf_{\hat{y}^T} \sup_{y_T} f(g_1(\emptyset), \cdots, g_{T-1}(y_1^{T-2}), y_1^{T-1}, \hat{y}_T, y_T).$$

By Lemma 1, for any g^{T-1}, we have

$$\sup_{y^{T-1}} \inf_{\hat{y}^T} \sup_{y_T} f(g_1(\emptyset), \cdots, g_{T-1}(y^{T-2}), y^{T-1}, \hat{y}_T, y_T) =$$
$$\inf_{g_T \in \mathcal{G}_T} \sup_{y^T} f(g_1(\emptyset), \cdots, g_T(y_1^{T-1}), y_1^T).$$

Putting $\inf_{g_1^{T-1}}$ on both side one will finish the proof. ■

Note that the average case part in Theorem 1 follows from similar argument as above by notice that $\inf_g \mathbb{E}_X f(g(X), X) = \mathbb{E}_X \inf_a f(a, X)$. The final part follows from the minimax inequality $\inf_a \sup_b f(a,b) \geq \sup_b \inf_a f(a,b)$ since we can iteratively interchange $\sup_{\mathbf{x}_t}$ and $\inf_{\hat{y}_t}$ in the expression of r_T^a without increase the value.

B. Proof of Theorem 2

We first prove the lower bound of $r_T^*(\mathcal{H}_{f,\mathbf{w}})$ by choosing a particular $\mathbf{x}^T$. We partition the inputs $\mathbf{x}^T$ and outputs $\mathbf{y}^T$ into d parts where each part has length of T/d. With that, the ith part of the inputs and the outputs are denoted by $\mathbf{x}^{(i)} = (\mathbf{x}_{(T/d)*(i-1)+1}, \cdots, \mathbf{x}_{(T/d)*i})$ and $\mathbf{y}^{(i)} = (y_{(T/d)*(i-1)+1}, \cdots, y_{(T/d)*i})$, respectively. Note that each element of $\mathbf{x}^{(i)}$ is a vector in $\mathbb{R}^d$. We choose any element of $\mathbf{x}^{(i)}$ equal to the standard basis vector $_i \in \mathbb{R}^d$, that is 1 at the coordinate i and 0 elsewhere.

For each $\mathbf{x}_i, \mathbf{w} \in \mathbb{R}^d$, we interpret $\Pr[y_i = 1|\mathbf{x}_i] = f(\langle \mathbf{w}, \mathbf{x}_i \rangle)$ at the prediction of y_i. Which further defines a product measure $p_{\mathbf{w}}(y^T|\mathbf{x}^T) = \prod_{i=1}^T \Pr[y_i|\mathbf{x}_i]$ over $y^T \in \{0,1\}^n$.

Note that for any given $\mathbf{x}^T$

$$r_T^*(\mathcal{H}_{f,\mathbf{w}}) \geq r_T^*(\mathcal{H}_{f,\mathbf{w}}|\mathbf{x}^T) = \log S_T(\mathcal{H}_{f,\mathbf{w}}|\mathbf{x}^T),$$

where $S_T(\mathcal{H}_{f,\mathbf{w}}|\mathbf{x}^T)$ is the Shtarkov sum of the measures $p_{\mathbf{w}}(\cdot|\mathbf{x}^T)$. Therefore, it is sufficient to lower bound $S_T(\mathcal{H}_{f,\mathbf{w}}|\mathbf{x}^T)$ which, with the above notation, is written as

$$S_T(\mathcal{H}_{f,\mathbf{w}}|\mathbf{x}^T) = \sum_{y^T \in \{0,1\}^n} \sup_{\mathbf{w} \in \mathcal{B}_2^d} p_{\mathbf{w}}(y^T|\mathbf{x}^T).$$

By the selection of $\mathbf{x}^T$ in the first paragraph and since $\langle \mathbf{w}, _i \rangle = w_i$, we have

$$S_T(\mathcal{H}_{f,\mathbf{w}}|\mathbf{x}^T) = \sum_{y^T \in \{0,1\}^T} \sup_{\mathbf{w} \in \mathcal{B}_2^d} \prod_{i=1}^d P_f(\mathbf{y}^{(i)}|w_i), \quad (7)$$

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 $\mathbf{y}^{(i)}$. Next, we derive a lower bound on S_T by limiting the range of $\mathbf{w}$ so that $|w_i| \leq 1/\sqrt{d}$ for all $i \in [d]$. Therefore, we have that

$$S_T(\mathcal{H}_{f,\mathbf{w}}|\mathbf{x}^T) \geq \sum_{y^T \in \{0,1\}^T} \prod_{i=1}^d \sup_{w_i \in [-1/\sqrt{d}, 1/\sqrt{d}]} P_f(\mathbf{y}^{(i)}|w_i)$$
$$= \prod_{i=1}^d \sum_{\mathbf{y}^{(i)} \in \{0,1\}^{T/d}} \sup_{w_i \in [-1/\sqrt{d}, 1/\sqrt{d}]} P_f(\mathbf{y}^{(i)}|w_i)$$
$$= \Big(\sum_{\mathbf{y} \in \{0,1\}^{T/d}} \sup_{w \in [-1/\sqrt{d}, 1/\sqrt{d}]} P_f(\mathbf{y}|w) \Big)^d, \quad (8)$$

where the last equality holds due to the symmetry over i. We proceed with the following lemma.

Lemma 3. *The following inequality holds:*

$$\sum_{\mathbf{y} \in \{0,1\}^{T/d}} \sup_{w \in [\frac{-1}{\sqrt{d}}, \frac{1}{\sqrt{d}}]} P_f(\mathbf{y}|w) \geq \Omega(\sqrt{T/d^2}). \quad (9)$$

Proof: Note that for all $k \in [T/d]$, there exists a constant $C \in \mathbb{R}^+$ such that

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

Note that

$$\sum_{\mathbf{y} \in \{0,1\}^{T/d}} \sup_{w \in [\frac{-1}{\sqrt{d}}, \frac{1}{\sqrt{d}}]} P_f(\mathbf{y}|w) \geq \sum_{k=T/2d-cT/d\sqrt{d}}^{T/2d+cT/d\sqrt{d}} B(k, T/d).$$

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

$$\frac{1}{\sqrt{k(T/d-k)}} \geq \frac{d/T}{\sqrt{\frac{1}{1/4-c^2/d}}} = \sqrt{(1/4-c^2/d)}d/T.$$

Therefore, the LHS of (9) is lower bounded by

$$C\sqrt{(1/4-c^2/d)}\sqrt{\frac{T}{d}}\frac{2c}{\sqrt{d}} = \Omega(\sqrt{T/d^2}),$$

for $d > 4c^2$. ■

With this lemma and (8), we find the following lower bound

$$S_T(\mathcal{H}_{f,\mathbf{w}}|\mathbf{x}^T) \geq c_1^d \left(\frac{T}{d^2}\right)^{d/2},$$

for some constant c_1 independent of T, d. As a result,

$$r_T^*(\mathcal{H}_{f,\mathbf{w}}) \geq \log S_T(\mathcal{H}_{f,\mathbf{w}}|\mathbf{x}^T) \geq d/2 \log T - d \log d + d \log c_1,$$

which completes the proof of the lower bound in the theorem.

We now prove the upper bound for $r_T^*(\mathcal{H}_{f,\mathbf{w}})$. For that, we choose a covering set $\mathcal{G}$ of $\mathcal{B}_2^d$ such that any point in $\mathcal{B}_2^d$ is d/LT^2 close to some point in $\mathcal{G}$ under ℓ_2 distance. Clearly, we have $|\mathcal{G}| \leq \left(3LT^2/d\right)^d$ since the covering number is upper bounded by packing number of half radius. For any given $\mathbf{x}_1, \cdots, \mathbf{x}_T$ and $\mathbf{w} \in \mathcal{B}_2^d$, we can define a probability measure $p_{\mathbf{w}}$ on $\{0,1\}^T$ such that

$$p_{\mathbf{w}}(y^T|\mathbf{x}^T) = \prod_{i=1}^{T} f(\langle \mathbf{w}, \mathbf{x}_i \rangle)^{y_i} (1 - f(\langle \mathbf{w}, \mathbf{x}_i \rangle))^{1-y_i}.$$

We drop the dependence of $p_{\mathbf{w}}$ on $\mathbf{x}^T$ for notation convenience. Define $\mathcal{P}$ be the class of probability measures $p_{\mathbf{w}}$ induced by parameters $\mathbf{w} \in \mathcal{G}$. We define another class $\mathcal{P}'$ such that we replace any $p_{\mathbf{w}} \in \mathcal{P}$ with $p_{\mathbf{w}}(y_i) < 1/T$ or $> 1 - 1/T$ with $p_{\mathbf{w}}(y_i) = 1/T$ or $p_{\mathbf{w}}(y_i) = 1 - 1/T$ respectively. We claim that for any $\mathbf{w} \in \mathcal{B}_2^d$ there exist $p \in \mathcal{P}'$ such that

$$\sup_{y^T \in \{0,1\}} p_{\mathbf{w}}(y^T)/p(y^T) \leq e^d. \tag{10}$$

To see this, we choose $\mathbf{w}' \in \mathcal{G}$ that is d/LT^2 close to $\mathbf{w}$ under ℓ_2 distance. Let p be the probability measure in $\mathcal{P}'$ that is associated with $p_{\mathbf{w}'}$, i.e, p is the truncated distribution of $p_{\mathbf{w}'}$. We show that p is the desired distribution in (10). For any $i \in [T]$, if $p_{\mathbf{w}}(y_i) > 1 - \frac{1}{T}$, we have $p(y_i) \geq 1 - \frac{1}{T} - d/T^2$ (since by Lipschitz condition if $p_{\mathbf{w}}(y_i) \geq 1 - 1/T + d/T^2$ we have $p(y_i) = 1 - 1/T$ else we have $p(y_i) \geq p_{\mathbf{w}}(y_i) - d/T^2 \geq 1 - 1/T - d/T^2$.), which implies that (for $d \leq T$) we have $p_{\mathbf{w}}(y_i)/p(y_i) \leq 1 + 1/(T-2)$. As $p(y_i) \geq 1/T$ and $|p(y_i) - p_{\mathbf{w}}(y_i)| \leq d/T^2$ (by Lipschitz condition and definition of $\mathcal{G}, \mathcal{P}'$), for any $p_{\mathbf{w}}(y_i) \leq 1 - \frac{1}{T}$, we have

$$\frac{p_{\mathbf{w}}(y_i)}{p(y_i)} \leq \frac{p(y_i) + d/T^2}{p(y_i)} \leq 1 + d/T. \tag{11}$$

Putting everything together, we obtain the inequality $\frac{p_{\mathbf{w}}(y^T)}{p(y^T)} \leq (1 + d/T)^T \leq e^d$. Now, we observe that

$$\sum_{y^T} \sup_{\mathbf{w} \in \mathcal{B}_2^d} p_{\mathbf{w}}(y^T) \leq \sum_{y^T} e^d * p(y^T) \leq e^d \sum_{y^T} \sup_{p \in \mathcal{P}'} p(y^T).$$

Since the Shtarkov sum of a finite set is upper bounded by its size. We have

$$r_T^*(\mathcal{H}_{f,\mathbf{w}}) \leq \log |\mathcal{G}| + d = 2d \log T - d \log d + O(d),$$

where the constant in BigOh is independent of T, d.

Next, we prove the bounds for r_T^a as well. The lower bound is straightforward as $r_T^a \geq r_T^*$ (Theorem 1). In what follows, we prove the upper bound on r_T^a via the achievability argument. For that, we use the Bayesian scheme introduced in [21], and analyze its regret. We start with the description of this approach. Let $\mathcal{P}$ be a class of functions that map $\mathcal{X}^t \times \mathcal{Y}^{t-1} \to \Delta(\mathcal{Y})$ for all $t \in \mathbb{N}$, where $\Delta(\mathcal{Y})$ is the set of all probability measures on $\mathcal{Y}$. We assume that the functions in $\mathcal{P}$ are indexed by Θ. For any $\theta \in \Theta$, we denote

$$p(y^T|\mathbf{x}^T, \theta) = \prod_{t=1}^{T} q_\theta(y_t|\mathbf{x}^t, y^{t-1}), \tag{12}$$

where q_θ is the probability measure in $\mathcal{P}$ indexed by θ with input $(\mathbf{x}^t, y^{t-1})$. Let p_0 be some probability measure on Θ. We denote

$$p(y^T|\mathbf{x}^T) = \int p(y^T|\mathbf{x}^T, \theta) p_0(\theta) \mathrm{d}\theta. \tag{13}$$

The Bayesian prediction rule at time t is given by

$$\hat{p}(y_t|\mathbf{x}^t, y^{t-1}) = \frac{p(y^t|\mathbf{x}^t)}{p(y^{t-1}|\mathbf{x}^{t-1})}. \tag{14}$$

Clearly, this is a valid prediction rule since $p(y^t|\mathbf{x}^t)$ can be evaluated with only the observation of $(\mathbf{x}^t, y^{t-1})$. Moreover, for any $\mathbf{x}^T$, the probability assigned on y^T by the Bayesian rule is exactly $p(y^T|\mathbf{x}^T)$. The following lemma was proved in [21, Lemma 2.1] (see also [22, Theorem 1]). (Note that the lemma we present below is slightly stronger than [21] since we allow the functions in $\mathcal{P}$ to be dependent on past $\mathbf{x}_t, y_t$s, while the proof is exactly the same.)

Lemma 4. *Let Q be any distribution over Θ, and define*

$$L_Q(y^T|\mathbf{x}^T) = \int \log(1/p(y^T|\mathbf{x}^T, \theta)) Q(\theta) \mathrm{d}\theta.$$

Then for any $\mathbf{x}^T, y^T$, we have

$$\log(1/p(y^T|\mathbf{x}^T)) \leq L_Q(y^T|\mathbf{x}^T) + \mathrm{KL}(Q||p_0).$$

Corollary 1. *Suppose Θ is finite, and p_0 is the uniform distribution over Θ. Then for any $\mathbf{x}^T, y^T$*

$$\log(1/p(y^T|\mathbf{x}^T)) - \min_{\theta \in \Theta} \log(1/p(y^T|\mathbf{x}^T, \theta)) \leq \log |\Theta|.$$

This implies $r_T^a(\mathcal{P}) \leq \log |\Theta|$ under log-loss.

Proof: Let θ^* be the minimizer of $\min_\theta \log(1/p(y^T|x^T, \theta))$. Let Q be the distribution that assigns probability 1 on θ^*. By Lemma 4, we only need to show that $\mathrm{KL}(Q||p_0) \leq \log |\Theta|$ which follows trivially. ∎

Now, we are ready to prove the upper bound on r_T^a. We consider the same covering set $\mathcal{G}$ as in the proof of Theorem 2 for r_T^*. Note that, the selection of the $\mathcal{G}$ does not depend on the realization of $\mathbf{x}^T, y^T$. Moreover, the truncation for probabilities $f(\langle \mathbf{w}, \mathbf{x}_i \rangle) > 1 - 1/T$ or $< 1/T$ does not depend on future samples. We can run the Bayesian approach with uniform prior over $\mathcal{G}$. By Corollary 1, we have $r_T^a(\mathcal{G}) \leq \log |\mathcal{G}|$. Moreover, since for any samples $\mathbf{x}^T, y^T$ and $\mathbf{w} \in \mathcal{B}_2^d$ there exists some $p \in \mathcal{P}'$ (where $\mathcal{P}'s$ is defined as in the proof of Theorem 2 for r_T^*) so that $p_{\mathbf{w}}(y^T|\mathbf{x}^T)/p(y^T|\mathbf{x}^T) \leq e^d$. We have

$$\inf_{\mathbf{w} \in \mathcal{B}_2^d} \log(1/p_{\mathbf{w}}(y^T|\mathbf{x}^T)) \geq \inf_{p \in \mathcal{P}'} \log(1/p(y^T|\mathbf{x}^T)) - d.$$

Therefore, $r_T^a(\mathcal{H}_{f,\mathbf{w}}) \leq 2d \log T - d \log d + O(d)$.

Acknowledgment

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

REFERENCES

[1] L. D. Davisson, "Universal noiseless coding," *IEEE Trans. Inf. Theory*, vol. IT-19, pp. 783–795, Nov. 1973.

[2] R. E. Krichevsky and V. K. Trofimov, "The performance of universal encoding," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 199–207, Mar. 1981.

[3] G. I. Shamir, "On the MDL principle for i.i.d. sources with large alphabets," *IEEE Trans. Inform. Theory*, vol. 52, pp. 1939–1955, May 2006.

[4] Y. M. Shtarkov, "Universal sequential coding of single messages," *Problems of Information Transmission*, vol. 23, pp. 3–17, Jul.-Sep. 1987.

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

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

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

[8] Q. Xie and A. Barron, "Asymptotic minimax regret for data compression, gambling, and prediction," *IEEE Trans. Information Theory*, vol. 46, pp. 431–445, 2000.

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

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

[11] S. Shalev-Schwartz and S. Ben-David, *Understanding Machine learning*. Cambridge University Press, 2014.

[12] A. Orlitsky and N. P. Santhanam, "Speaking of infinity," *IEEE Trans. Inf. Theory*, vol. 50, pp. 2215–2230, Oct. 2004.

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

[14] N. Cesa-Bianchi and O. Shamir, "Efficient transductive online learning via randomized rounding," *arXiv preprint arXiv:1106.2429*, 2011.

[15] A. Rakhlin, K. Sridharan, and A. Tewari, "Online learning: Random averages, combinatorial parameters, and learnability," 2010.

[16] N. Alon, O. Ben-Eliezer, Y. Dagan, S. Moran, M. Naor, and E. Yogev, "Adversarial laws of large numbers and optimal regret in online classification," in *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 447–455, 2021.

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

[18] P. Jacquet, G. Shamir, and W. Szpankowski, "Precise minimax regret for logistic regression with categorical feature values," in *Algorithmic Learning Theory*, pp. 755–771, PMLR, 2021.

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

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

[21] S. Kakade and A. Y. Ng, "Online bounds for bayesian algorithms," 2008.

[22] Y. Freund, R. E. Schapire, Y. Singer, and M. K. Warmuth, "Using and combining predictors that specialize," in *Proceedings of the twenty-ninth annual ACM symposium on Theory of computing*, pp. 334–343, 1997.