Minimax Regret with Unbounded Weights

Michael Drmota
TU Wien
Vienna, Austria
drmota@tuwien.ac.at

Philippe Jacquet
INRIA
Paris, France
philippe.jacquet@inria.fr

Changlong Wu
Purdue University
W. Lafayette, USA
wuchangl@hawaii.edu

Wojciech Szpankowski
Purdue University
W. Lafayette, USA
szpan@purdue.edu

***Abstract*—In online learning, a learner receives data in rounds $1 \le t \le T$ and at each round predicts a label which is then compared to the true label resulting in a loss. The total loss over T rounds, when compared to a loss over the best expert from a class of experts, is called the regret. This paper focuses on *logarithmic loss* over a class of experts $\mathcal{H}_{p,\mathbf{w}}$, represented by a probability distribution p and parameterized by a d-dimensional weight vector $\mathbf{w}$. Unlike previous work that studied bounded weight, we assume that the norm of the weight can be *unbounded*. This unboundedness poses a challenging problem that leads to unexpected results. For such a class of weighted experts we analyze the (fixed design) minimax regret for the best predictor and worst label sequence. Such a minimax regret turns out to be a universal lower bound for most regrets analyzed in the literature. For bounded weights it is known that the minimax regret can grow like $(d/2)\log(TR^2/d)$ where R is an upper bound on the weight norm. In contrast, we show in this paper that for unbounded norm with $R = \infty$ the minimax regret is asymptotically $(d-1)\log(T/d)$ for a logistic-like expert class which we also extend to $R = \Omega(\sqrt{T})$. We prove it by introducing the so called splittable label sequences that partition the weight space into T^{d-1} regions with maximum sequence probability equal to 1. Finally, for a general class of monotone experts we present an upper bound $2d\log T$ for the minimax regret.**

I. Introduction

We study the problem of online learning and online regret minimization. Formally, we phrase the online learning problem in terms of a game between nature/environment and a learner/predictor. Broadly, the objective of the learner is to process past observations to predict the next realization of the nature's labeling sequence. At each round $t \in \mathbb{N}$, let y_t be the true label that is yet to be revealed. The learner obtains a d dimensional data/feature vector $\mathbf{x}_t \in \mathbb{R}^d$ to make a prediction $\hat{y}_t = g_t(y^{t-1}, \mathbf{x}^t)$, where g_t represents the strategy/algorithm of the learner to obtain its prediction based on the past and current observations. Once the prediction is made, 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 $\hat{\mathcal{Y}} \in \mathbb{R}$ and $\mathcal{Y} \in \{-1, 1\}$ 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 (forecasters or experts) denoted as $\mathcal{H}$. Here, $\mathcal{H}$ is a collection of predicting functions $h : \mathbb{R}^d \mapsto \mathbb{R}$, with input being $\mathbf{x}_t$ at each time t. After T rounds, the *pointwise regret* is defined as

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

where $\hat{y}_t = g_t(y^{t-1}, \mathbf{x}^t)$ where $y^{t-1} = (y_1, \ldots, y_{t-1})$ and $\mathbf{x}^t = (\mathbf{x}_1, \ldots, \mathbf{x}_t)$ for all $t \in [T]$. The first summation of $\mathcal{R}(g^T, y^T, \mathcal{H}|\mathbf{x}^T)$ represents the accumulated loss of the learner (algorithmic loss) while the second summation represents the loss of the best expert within the class $\mathcal{H}$.

In this paper, we consider only the logarithmic loss $\ell : \hat{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}^+$ defined as

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

Moreover, we restrict our study to a specific class of experts defined as

$$\mathcal{H}_{p,\mathbf{w}} = \{h : \mathbb{R}^d \to \mathbb{R} : \ h(\mathbf{x}) = p(\langle \mathbf{w}|\mathbf{x}\rangle) : \mathbf{w}, \mathbf{x} \in \mathbb{R}^d\}, \tag{2}$$

where $\mathbf{w}$ is a d dimensional weight vector, $\langle \mathbf{w}|\mathbf{x}\rangle$ is the scalar product of $\mathbf{x}_t$ and $\mathbf{w}$, and $p(w)$ with $w = \langle \mathbf{w}|\mathbf{x}\rangle$ is a function $\mathbb{R} \to [0, 1]$. Often $p(w)$ is either the logistic function $p(w) = (1 + \exp(-w))^{-1}$ (see e.g., [13], [17]) or the probit function $p(w) = \Phi(-w)$ where $\Phi(w)$ is CDF (cumulative distribution function) of the normal distribution (see [5]). Throughout, we assume that $\mathbf{x}_t$ lies on a compact manifold $\mathcal{M}_d$ such as $\mathcal{M}_d = [-1, 1]^d$ or $\mathcal{M}_d = \mathcal{S}_d$ where $\mathcal{S}_d$ is a d dimension sphere. We do *not* restrict the weights $\mathbf{w} \in \mathbb{R}^d$ and this seems never been analyzed in the past, to the best of our knowledge. To be precise, we assume that $\|\mathbf{w}\|_2 \le R \le \infty$ and study two cases: (i) $R = \infty$ and (ii) $R = \Omega(\sqrt{T})$ when $T \to \infty$.

There are several perspectives on studying regret that decouples it from the prediction strategy, sequences of labels, and data. In the *fixed design* setting analyzed here, we study the minimal regret for the worst realization of the label with the feature vector $\mathbf{x}^T$ known in advance. The *fixed design minimax regret* is then defined as

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

Note that this notion was also studied in [18], [14], and in [7] under the name *transductive online learning*. To further decouple it from the feature vector $\mathbf{x}_t$ one either maximizes over all $\mathbf{x}_t$ or take the average over the features, that is, $\bar{r}^*_T(\mathcal{H}) := \mathbf{E}_{\mathbf{x}}[r^*_T(\mathcal{H}|\mathbf{x}^T)]$ where the feature vector $\mathbf{x}_t$ is generated *i.i.d.* from some distribution μ over $\mathcal{M}_d$.

In this paper, we focus on precise asymptotics of the (fixed design) minimax r^*_T for large d, T, and *unbounded* weights $\mathbf{w}$. Note that, the fixed design minimax regret forms a universal lower bound for a large class of regrets analyzed in literature [3], [4], [16], [18], [14], [24]. To see this, we consider the following general notion of minimax regret as in [25]. We assume that $\mathbf{x}^T$ is sampled from some distribution (random process) μ in a class $\mathcal{P}$ of distributions over $\mathbf{x}^T$ while the true labels are still presented adversarially. The *expected worst case minimax regret* is defined in [25] as

$$\tilde{r}_T(\mathcal{H},\mathcal{P}) = \inf_{g^T}\sup_{\mu\in\mathcal{P}} \mathbb{E}_{\mathbf{x}^T\sim\mu}\left[\sup_{y^T}\mathcal{R}(\hat{y}^T, y^T, \mathcal{H} \mid \mathbf{x}^T)\right],$$

where $g_t : \mathcal{X}^t \times \mathcal{Y}^{t-1} \to \hat{\mathcal{Y}}$ runs over all possible prediction rules, and $\hat{y}_t = g_t(\mathbf{x}^t, y^{t-1})$. This regret is quite general and recovers previously known minimax regrets including the fixed design minimax regret. Indeed, $r_T^*(\mathcal{H}|\mathbf{x}^T)$ is equal to $\tilde{r}_T(\mathcal{H},\mathcal{P})$ when $\mathcal{P}$ is the singleton distribution that assigns probability 1 to $\mathbf{x}^T$. More importantly, if $\mathcal{P}$ is a class of all singleton distributions over $\mathcal{X}^T$, then $\tilde{r}_T(\mathcal{H},\mathcal{P}) = r_T^a(\mathcal{H})$ where the *sequential minimax regret* $r_T^a(\mathcal{H})$ is defined as

$$r_T^a(\mathcal{H}) = \inf_{g^T}\sup_{\mathbf{x}^T, y^T}\left(\sum_{t=1}^T \ell(\hat{y}_t, y_t) - \inf_{h\in\mathcal{H}}\sum_{t=1}^T \ell(h(\mathbf{x}_t), y_t)\right) \tag{4}$$

and it is equivalent to

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

as discussed e.g., in [16]. Clearly, the *maximum* fixed design minimax regret $r_T^*(\mathcal{H}) \overset{\text{def}}{=} \sup_{\mathbf{x}^T} r_T^*(\mathcal{H}|\mathbf{x}^T)$ is a lower bound for $r_T^a(\mathcal{H})$. Moreover, if $\mathcal{P}$ is the class of all $i.i.d.$ distributions over $\mathbf{x}^T$, then $r_T^*(\mathcal{H})$ is also a lower bound of $\tilde{r}_T(\mathcal{H},\mathcal{P})$ up to a $\log T$ factor (see Theorem 21 in [25]).

a) Our Contribution: In this paper, we present for the first time tight bounds for the fixed design minimax regret for the class $\mathcal{H}_{p,\mathbf{w}}$ with *unbounded* $\mathbf{w}$ and logistic-like as well as general monotone p. We accomplish it by analyzing Shtarkov's sum using tools from *analytic combinatorics* such as complex asymptotics as discussed in [10], [21]. See also [3], [4], [26].

More precisely, we first represent the minimax regret as the logarithm of the Shtarkov sum over all label sequences of the maximum label probability which turns out to be the maximum-likelihood distribution. Shtarkov's sum arose already in the universal compression as witnessed by [8], [9], [22]. In this paper we consider the minimax regret with unbounded weights which leads to a challenging problem with unexpected results. It is known by [11], [19], [26] that, for bounded weights, the regret grows asymptotically like $(d/2)\log(R^2T/d)$ where $\|\mathbf{w}\| \le R$. However, this does not scale well with large R. Our first main result demonstrates that for $R = \infty$ the fixed design regret satisfies $(d-1)\log(T/d) \le r_T^*(\mathcal{H}_{p,\mathbf{w}}) \le 2d\log T$ for any *monotone* function p. This establishes a fundamental distinction with the *sequential* minimax regret as in (4), for which a $\Omega(\log R)$ lower bound is necessary [11]. We then provide tighter bounds for the logistic function, showing that for $i.i.d.$ generated features $r_T^*(\mathcal{H}_{p,\mathbf{w}}) \le (d-1)\log(T/d) + O(1)$. Finally, we demonstrate that for logistic function, an $(d-3)\log(T/d)$ lower bound is achievable even with $R = \Omega(\sqrt{T})$. These together provide a *precise* characterization of the fixed design minimax regret for the logistic function.

We establish these bounds by introducing a set of "splittable" label sequences y^T that partition the $\mathbf{w}$ space into $O((T/d)^{d-1})$ regions where the optimal label probability is 1. To obtain an upper bound for the logistic functions, we apply Mellin transform and precise counting of label sequences that are k Hamming distance away from a splittable sequence.

b) Related Work: Online learning can be viewed as a universal compression (source coding) with side information as discussed in [2]. In particular, the online regret in information theory language is the excess of a universal code length over the shortest description achieved by a class of sources (experts) with extra side information available to the learner. There are a number of useful and precise *bounds* on various regrets. We mention here only a few: [11], [13], [15], [17], [26] Our findings are closest to [18], [14], [26]. For example, in [14] a precise maximal minimax regret is analyzed but only for *finite* number of feature values and fixed dimension d while in [26] only bounded $\mathbf{w}$ was studied. We should also mention that the general form of the minimax regret, is related to Fisher information and was already known in information theory [23] but only for fixed d and bounded weights.

II. Problem Formulation

We denote by $\mathbf{x}_t = (x_{1,t},\ldots,x_{d,t})$ a d-dimensional bounded data/feature vector such that $\|\mathbf{x}\| \le 1$ for some norm $\|\cdot\|$, e.g., $\mathbf{x}_t \in \mathcal{S}_d \subset \mathbf{R}^d$ where $\mathcal{S}_d$ is a sphere of radius 1. The binary label vector is written as $y^T = (y_1,\ldots,y_T)$ with $y_t \in \{-1,1\}$. At last, the d dimensional weight vector $\mathbf{w}_t = (w_{1,t},\ldots,w_{d,t})$ is assumed to belong to a ball $\mathcal{B}(R)$ of radius R with $R \le \infty$. Thus $\mathbf{w}$ can be unbounded and this scenario has never been analyzed in depth before. We study first the case $R = \infty$ and then consider the transition region when $R; = R_T \to \infty$ with T but not slower than $\sqrt{T}$. In a forthcoming paper we study the region of $R = o(\sqrt{T})$ which is a very hard problem if one wants a precise statement.

In this paper we focus on the logarithmic loss (1) for the class of experts $\mathcal{H} := \mathcal{H}_{p,\mathbf{w}}$ as defined in (2). More specifically, for any $h \in \mathcal{H}$, we interpret $h(\mathbf{x}) \in [0,1]$ as the probability assigned to $Y = -1$, that is, $P(Y_t = -1|\mathbf{x}_t, \mathbf{w}) = p(\langle\mathbf{w}|\mathbf{x}_t\rangle)$ is the probability assigned to $Y = -1$, and $\ell(h(\mathbf{x}), y) = -\log P(Y = y|\mathbf{x})$. Likewise, for any probability distribution Q over $\{-1,1\}^T$ we can induce a predictor $\hat{y}_t = Q(-1|y^{t-1})$. Therefore, $\ell(\hat{y}, y) := -\log Q(y_t|y^{t-1})$ where Q represents a “universal” distribution that approximates the expert probability $P(y_t|\mathbf{x}_t, \mathbf{w})$. Thus, we can re-write the pointwise regret in terms of Q as

$$\mathcal{R}_T(Q, y^T|\mathbf{x}^T) =$$

$$-\sum_{t=1}^T \log Q(y_t|y^{t-1}) + \sup_{\mathbf{w}}\sum_{t=1}^T \log P(y_t|\mathbf{x}_t, \mathbf{w}) =$$

$$= \log\frac{\sup_{\mathbf{w}} P(y^T|\mathbf{x}^T, \mathbf{w})}{Q(y^T)}.$$

Then the fixed design minimax regret can be represented as

$$r_T^*(\mathcal{H}|\mathbf{x}^T) = \inf_Q \max_{y^T} \mathcal{R}_T(Q, y^T|\mathbf{x}^T), \tag{5}$$

and $\bar{r}_T^*(\mathcal{H}) := \mathbf{E}_{\mathbf{x}}[r_T^*(\mathcal{H}|\mathbf{x}^T)]$ where it is assumed that $\mathbf{x}_t$ is generated by a memoryless ($i.i.d.$) source.

In order to study precisely the the minimax regret $r_T^*(\mathcal{H}|\mathbf{x}^T)$ we first find a more manageable representation. Following [20], [8], [6] we add/subtract from (5) the logarithm of the Shtarkov sum defined as

$$S_T(\mathbf{x}^T) := \sum_{y^T} \sup_{\mathbf{w}} P(y^T|\mathbf{x}^T, \mathbf{w}) \tag{6}$$

resulting in

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

which we shall use throughout the rest of the paper.

III. Main Results

As discussed above, we assume that $\|\mathbf{w}\|_2 \leq R$ and mostly focus on the case $R = \infty$, however, we also consider the case when R is finite but grows w.r.t. T. Observe that for a sequence of labels y^T and a sequence of features $\mathbf{x}^T$ we have

$$P(y^T|\mathbf{x}^T, \mathbf{w}) = \prod_{t: y_t < 0} p(\langle \mathbf{w}|\mathbf{x}_t \rangle) \prod_{t: y_t > 0} (1 - p(\langle \mathbf{w}|\mathbf{x}_t \rangle)). \tag{7}$$

In particular, for the logistic function $p(w) = (1 + e^{-w})^{-1}$ this becomes

$$P(y^T|\mathbf{x}^T, \mathbf{w}) = \prod_{t=1}^{T} \frac{\exp\left(\frac{1+y_t}{2} \langle \mathbf{x}_t, \mathbf{w} \rangle\right)}{1 + \exp(\langle \mathbf{x}_t, \mathbf{w} \rangle)}. \tag{8}$$

We now define some important notation. Let

$$\mathbf{w}^*(y^T|\mathbf{x}^T) = \arg\sup_{\mathbf{w}} P(y^T|\mathbf{x}^T, \mathbf{w}),$$
$$P^*(y^T|\mathbf{x}^T) = \sup_{\mathbf{w}} P(y^T|\mathbf{x}^T, \mathbf{w}^*).$$

Observe that $\mathbf{w}^*(y^T|\mathbf{x}^T) \leq \infty$ and for $\mathbf{w}^*(y^T|\mathbf{x}^T) = \infty$ we may have $P^*(y^T|\mathbf{x}^T) = 1$. Throughout we often drop $\mathbf{x}^T$, especially when $\mathbf{x}^T$ is fixed.

A. Unbounded Weights and Splittable Sequences

We now assume that $R = \infty$ so that $\mathbf{w}$ is unbounded. We shall find all y^T such that $\mathbf{w}^*(y^T) = \infty$, thus for those y^T we have $P^*(y^T|\mathbf{x}^T) = 1$. In this case, we present tight lower and upper bounds for the regret. We work with the following class

$$\mathcal{H}^{\mathsf{split}} = \{p(\langle \mathbf{w}|\mathbf{x} \rangle) : \ p(-\infty) = 0 \text{ and } p(\infty) = 1\}. \tag{9}$$

We call a sequence y^T *splittable* if there exists $\mathbf{w}$ such that $\forall t \in [T], \ y_t \langle \mathbf{x}_t, \mathbf{w} \rangle > 0$ for a given $\mathbf{x}^T$. We denote the set of such sequences as $\mathcal{SP}(\mathbf{x}^T)$, that is

$$\mathcal{SP}(\mathbf{x}^T) := \{y^T : \ \forall t \in [T], \ y_t \langle \mathbf{x}_t, \mathbf{w} \rangle > 0, \ \mathbf{w} \in \mathbb{R}^d\}. \tag{10}$$

Such sequences partition the $\mathbf{w}$-space into regions as illustrated in Figure 1 (for $d = 2$). The next lemma bounds the number of such splittable sequences.

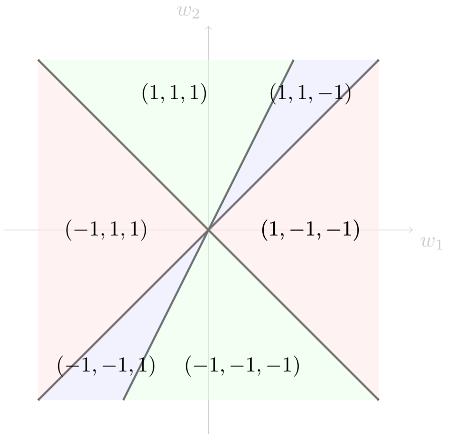

Fig. 1. Illustration to splittable sequences for $T = 3$ and $d = 2$ where the boundary of the regions are solutions (of $\mathbf{w}$) to $\langle \mathbf{x}_t, \mathbf{w} \rangle = 0$ where $\mathbf{x}_1 = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}), \mathbf{x}_2 = (-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}), \mathbf{x}_3 = (-\frac{2}{\sqrt{5}}, \frac{1}{\sqrt{5}}))$. There are six splittable sequences y^3 out of eight possible label sequences (*e.g* $(-1, 1, -1)$ is not splittable).

Lemma 1. *For any $\mathbf{x}^T$ in a general position (i.e., for all $k \leq d+1$ there is no k points of $\mathbf{x}^T$ that lie on a $k-2$ dimensional linear space) the following holds*

$$|\mathcal{SP}(\mathbf{x}^T)| = 2\sum_{i=0}^{d-1} \binom{T-1}{i} = 2\binom{T-1}{d-1}(1 + O(1/T))$$
$$\geq \left(\frac{T-1}{d-1}\right)^{d-1}. \tag{11}$$

Proof. This follows by computing the regions/chambers of d dimensional general position hyperplane arrangements [1, Theorem 3.1]. We illustrate this in Figure 1 for $d = 2$ and $T = 3$ leading to six out of eight splittable sequences in agreement with (11). □

a) Lower Bound: For $y^T \in \mathcal{SP}(\mathbf{x}^T)$ we have $\mathbf{w}^*(y^T) = \infty$ leading to the following lower bound for the regret.

Lemma 2. *Let $\mathcal{H}^{\mathsf{split}}$ be as in (9), we have*

$$r_T^*(\mathcal{H}^{\mathsf{split}}|\mathbf{x}^T) \geq \log|\mathcal{SP}(\mathbf{x}^T)| \geq (d-1)\log\left(\frac{T-1}{d-1}\right) + \log 2 \tag{12}$$

for all T, d and $\mathbf{x}^T$ in the general position.

Proof. The minimax regret $r_T^*(\mathcal{H}^{\mathsf{split}}|\mathbf{x}^T) = \log S_T(\mathbf{x}^T)$ and

$$S_T(\mathbf{x}^T) = \sum_{y^T} P^*(y^T|\mathbf{x}^T) \geq \sum_{y^T \in \mathcal{SP}(\mathbf{x}^T)} P^*(y^T|\mathbf{x}^T)$$
$$= |\mathcal{SP}(\mathbf{x}^T)|$$

since $P^*(y^T|\mathbf{x}^T) = 1$ for $y^T \in \mathcal{SP}(\mathbf{x}^T)$ by (9). □

b) Upper Bound for Logistic Regression: We now focus on the logistic regression, and provide a precise matching (including the leading constant) upper bound.

Theorem 1. *If $p(w) = (1 + e^{-w})^{-1}$ (logistic function), then for $\mathbf{x}^T$ uniformly distributed over a unit sphere, with high probability we have*

$$r_T^*(\mathcal{H}^{\mathsf{logistic}}|\mathbf{x}^T) \leq (d-1)\log\left(\frac{T-1}{d-1}\right) + O(1), \tag{13}$$

thus $\bar{r}^*_T(\mathcal{H}^{\mathsf{logistic}}) = (d-1)\log((T-1)/(d-1)) + O(1)$.

Proof. Let $I_T(\mathbf{w}, \mathbf{x}^T) = \sum_t \log(1 + e^{-|\langle \mathbf{w}|\mathbf{x}_t\rangle|})$ and define $\mathbf{w} = w\mathbf{u}$ where $\mathbf{u}$ is a unit vector. We know that if y^T is a splittable sequence, then $P(y^T|\mathbf{x}^T, \mathbf{w}) = \exp(-I(\mathbf{w}, \mathbf{x}^T))$. Let $I_T(w) = \mathbf{E}_{\mathbf{x}}[I_T(\mathbf{w}, \mathbf{x}^T)]$ where $\mathbf{x}_t$s are uniform on a unit sphere. Clearly $I_T(w) = TI(w)$. By expanding the expectation

$$I(w) = \frac{1}{s_d}\int_0^{\pi/2} \log(1 + e^{-w\sin\theta})(\cos\theta)^{d-1}d\theta$$

with the normalizing factor

$$s_d = \int_0^{\pi/2}(\cos\theta)^{d-1}d\theta = \frac{\sqrt{\pi}\Gamma(d/2)}{2\Gamma((d+1)/2)}.$$

We now apply the powerful Mellin transform to estimate $I(w)$. Recall that, $I^*(s)$ is the Mellin transform of $I(w)$ if $I^*(s) = \int_0^\infty I(w)w^{s-1}dw$ for some complex s (see [21]). Using properties of the Mellin transform we obtain

$$\begin{aligned} I^*(s) &= \int_0^\infty I(w)dw \\ &= \frac{1}{s_d}\int_0^{\pi/2}(\cos\theta)^{d-1}(\sin\theta)^{-s}d\theta(1-2^{-s})\Gamma(s)\zeta(s+1) \\ = & \quad \frac{2(1-2^{-s})}{s_d}\zeta(s+1)\frac{\Gamma((-s+1)/2)\Gamma(d/2)\Gamma(s)}{\Gamma((-s+d+1)/2)}. \end{aligned}$$

The main singularity of $I^*(s)$ is at $s = 1$ with residue $\frac{\alpha}{s_d}$ where $\alpha = \zeta(2) = \frac{\pi^2}{6}$. By inverse Mellin transform, we find

$$I(w) = \frac{\alpha}{s_d w} + O(\frac{1}{w^2}).$$

Let's now consider sequences that are within Hamming distance 1 from a splittable sequence, that is, the inequality $y_t\langle \mathbf{w}|\mathbf{x}_t\rangle > 0$ holds for all t except one. Without loss of generality, let's assume that the switched index is y_1. We have for all $t \neq 1$, $y_t\langle \mathbf{w}|\mathbf{x}_t\rangle = w|\sin\theta_t|$, where $w = \|\mathbf{w}\|$ and θ_t is the co-angle between $\mathbf{w}$ and $\mathbf{x}_t$. Therefore, we have w.h.p. over $\mathbf{x}^T$ that for all $\mathbf{w}$

$$-\log P(y^T|\mathbf{x}^T, \mathbf{w}) = w|\sin\theta_1| + I_T(w).$$

This is justified by: (1) Using Chernoff bound we have, for any $\varepsilon, \mathbf{w}$, $P\left(I_T(w\mathbf{u}, \mathbf{x}^T) \notin [(1-\varepsilon)I_T(w), (1+\varepsilon)I_T(w)]\right) < \exp(-T\frac{\beta}{w})$ for some $\beta > 0$; (2) This can be extended to hold for all $\mathbf{w}$ via a covering argument. Since $-w|\sin\theta_1| - I_T(w\mathbf{u}, \mathbf{x}^T) \sim -w|\sin\theta_1| - \frac{T\alpha}{s_d w}$, the maximum value of $P(y^T|\mathbf{x}^T, w\mathbf{u})$ with respect to w is $-2\sqrt{\frac{\alpha}{s_d}T|\sin\theta_1|}$. Therefore, by summing over all T switch positions, we have

$$\begin{aligned} S^1_T(\mathbf{x}^T)/|\mathcal{SP}(\mathbf{x}^T)| &= \sum_t e^{-2\sqrt{\alpha/s_d T\sin\theta_t}} \\ &\approx \frac{T}{s_d}\int_0^{\pi/2} e^{-2\sqrt{\alpha/s_d T\sin\theta}}(\cos\theta)^{d-1}d\theta \\ &= \frac{1}{\alpha}\int_0^{\alpha T/s_d} e^{-2\sqrt{x}}\left(1 - \frac{x^2 s_d^2}{\alpha^2T^2}\right)^{(d-1)/2} dx \\ &= \frac{1}{\alpha}\int_0^\infty e^{-2\sqrt{x}}dx = \frac{1}{2\alpha} + O(\exp(-\sqrt{T})). \end{aligned} \quad (14)$$

Here, $S^k_T(\mathbf{x}^T)$ is the Shtarkov sum that counting only sequences y^T with distance k to a splittable sequences. Now consider general k, that is, sequences that are within Hamming distance k from a splittable sequence. Assuming without loss of generalities that the mismatches are on the k first features, we have

$$-\log P(y^T|\mathbf{x}^T, w\mathbf{u}) = w\sin\theta_1 + \cdots + w\sin\theta_k + TI(w).$$

The maximum is $e^{-2\sqrt{\alpha/s_d T(\sin\theta_1+\cdots+\sin\theta_k)}}$. Since there are $\binom{T}{k}$ possible choices of the k switches, we have:

$$S^k_T(\mathbf{x}^T)/|\mathcal{SP}(\mathbf{x}^T)| = \binom{T}{k}\left(\frac{1}{s_d}\right)^k \times$$

$$\int_{[0,\pi/2]^k} e^{-2\sqrt{\alpha/s_d T(\sin\theta_1+\cdots+\sin\theta_k)}}\left(\cos\theta_1\cdots\cos\theta_k\right)^{d-1}d\theta_1\cdots d\theta_k$$

$$= \binom{T}{k}\left(\frac{1}{\alpha T}\right)^k \int_{[0,\alpha T/s_d]^k} e^{-2\sqrt{x_1+\cdots+x_k}}\left(1 - \frac{x_1^2 s_d^2}{\alpha^2T^2}\right)^{(d-1)/2} \times$$

$$\cdots\left(1 - \frac{x_k^2 s_d^2}{\alpha^2T^2}\right)^{(d-1)/2} dx_1\cdots dx_k$$

$$\begin{aligned} &\sim \binom{T}{k}\left(\frac{1}{\alpha T}\right)^k \int_{[0,\infty[^k} e^{-2\sqrt{x_1+\cdots+x_k}}dx_1\cdots dx_k \\ &= \binom{T}{k}\left(\frac{1}{\alpha T}\right)^k \frac{1}{(k-1)!}\int_0^\infty e^{-2\sqrt{y}}y^{k-1}dy \\ &= \binom{T}{k}\left(\frac{1}{4\alpha T}\right)^k \frac{(2k)!}{(k!)} \le \binom{2k}{k}\left(\frac{3}{2\pi^2}\right)^k. \end{aligned}$$

We now notice that $\sum_{k=1}^\infty \binom{2k}{k}x^k = (1-4x)^{-1/2}$ and conclude that the Shtarkov sum is upper bounded (with high probability) by

$$S_T(\mathbf{x}^T) \le 2\binom{T-1}{d-1}\frac{1}{\sqrt{1-\frac{6}{\pi^2}}}$$

as desired. Note that, there are many overlaps among these sequences, our result provide only an upper bound. □

B. A General Upper Bound

We now consider general features $\mathbf{x}^T$ and a general monotone class $\mathcal{H}$ and derive a general upper bound on the minimax regret. More specifically, let $\mathcal{X}$ be the feature space and $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ be the expert class. We recall now some standard facts: We say a sequence $\mathbf{x}^d \in \mathcal{X}^d$ is *pseudo-shattered* by $\mathcal{H}$ witnessed by $\mathbf{s}^d \in \mathbb{R}^d$ if for any $\epsilon^d \in \{0,1\}^d$ there exists $h \in \mathcal{H}$ such that for all $t \in [d]$: (1) If $\epsilon_t = 1$, then $h(\mathbf{x}_t) \ge s_t$; (2) If $\epsilon_t = 0$, then $h(\mathbf{x}_t) < s_t$. The pseudo-dimension of $\mathcal{H}$ is defined to be the maximum number d such that there exist $\mathbf{x}^d$ that can be pseudo-shattered by $\mathcal{H}$, denoted as $\mathsf{P}(\mathcal{H})$.

Lemma 3. *Let* $\mathcal{H}^{\mathsf{lin}} = \{p(\langle \mathbf{x}|\mathbf{w}\rangle) = \langle \mathbf{w}|\mathbf{x}\rangle : \mathbf{w}, \mathbf{x} \in \mathbb{R}^d\}$, *then*

$$\mathsf{P}(\mathcal{H}^{\mathsf{lin}}) \le d.$$

Proof. Suppose otherwise, there exists a sequence $\mathbf{x}^{d+1}$ that can be pseudo-shattered by $\mathcal{H}$ witnessed by $\mathbf{s}^{d+1}$. For any

$\epsilon^{d+1} \in \{0,1\}^{d+1}$, we denote $\bar{\epsilon}^{d+1}$ to be the complement of ϵ^{d+1} such that for all $t \in [d+1]$, $\epsilon_t = 1 - \bar{\epsilon}_t$. Let $\mathbf{w}_1, \mathbf{w}_2$ be the parameters corresponding to the functions in $\mathcal{H}^{\mathsf{lin}}$ that witness ϵ^{d+1} and $\bar{\epsilon}^{d+1}$ respectively. Taking $\mathbf{w}' = \mathbf{w}_1 - \mathbf{w}_2$, we have for all $t \in [d+1]$: (1) If $\epsilon_t = 1$, then $h_{\mathbf{w}'}(\mathbf{x}_t) \geq 0$; (2) If $\epsilon_t = 0$, then $h_{\mathbf{w}'}(\mathbf{x}_t) < 0$. To see this, assume w.o.l.g. $\epsilon_t = 1$, we have $h_{\mathbf{w}_1}(\mathbf{x}_t) \geq s_t$ while $h_{\mathbf{w}_2}(\mathbf{x}_t) < s_t$. Therefore, by linearity, we have $h_{\mathbf{w}'}(\mathbf{x}_t) = h_{\mathbf{w}_1}(\mathbf{x}_t) - h_{\mathbf{w}_2}(\mathbf{x}_t) \geq 0$. This implies that the sequence $\mathbf{x}^{d+1}$ is VC-shattered by $\mathcal{H}^{\mathsf{thres}} = \{1\{\langle \mathbf{w}|\mathbf{x}\rangle \geq 0\} : \mathbf{w}, \mathbf{x} \in \mathbb{R}^d\}$. This contradicts to the fact that VC-dimension of $\mathcal{H}^{\mathsf{thres}}$ is upper bounded by d. □

We now define a new general monotone hypothesis class:

$$\mathcal{H}^{\mathsf{mono}} := \{p(\langle \mathbf{x}|\mathbf{w}\rangle) : \ p \text{ is monotone increasing } \}.$$

The next lemma follows directly from the definition.

Lemma 4. *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be an arbitrary monotone increasing function. For any class $\mathcal{H}$, we denote by $\mathcal{H}^\sigma = \{\sigma(h) : h \in \mathcal{H}\}$. Then*

$$\mathsf{P}(\mathcal{H}^\sigma) \leq \mathsf{P}(\mathcal{H}).$$

Let $\mathcal{H} \subset [0,1]^{\mathcal{X}}$ be an arbitrary class with values in $[0,1]$. Let $J \subset [0,1]$ be a discretization of $[0,1]$ with step size $2\alpha > 0$, i.e. , $|J| \leq 1/2\alpha$. We define the discretized class

$$\tilde{\mathcal{H}} = \{\tilde{h}(\mathbf{x}) = \arg\min_{a \in J}\{|a - h(x)|\} : h \in \mathcal{H}\},$$

where we break ties by choosing the *larger* one.

Lemma 5. *For any class $\mathcal{H} \subset [0,1]^{\mathcal{X}}$, we have*

$$\mathsf{P}(\tilde{\mathcal{H}}) \leq \mathsf{P}(\mathcal{H}).$$

Proof. Let $\mathbf{x}^d$ be the sequence that can be pseudo-shattered by $\tilde{\mathcal{H}}$ witnessed by $\mathbf{s}^d$. We enumerate $J = \{0, 2\alpha, \cdots, 1\}$. For any $t \in [d]$, there must be some i such that $2i\alpha < s_t \leq 2(i+1)\alpha$. We take $s'_t = (2i+1)\alpha$. We now claim that $\mathbf{x}^d$ is pseudo-shattered by $\mathcal{H}$ witnessed by $\mathbf{s}'^d$. For any $\epsilon^d \in \{0,1\}^d$, we take $\tilde{h} \in \tilde{\mathcal{H}}$ such that if $\epsilon_t = 1$ then $\tilde{h}(\mathbf{x}_t) \geq s_t$ and $\tilde{h}(\mathbf{x}_t) < s_t$ otherwise. Let $h \in \mathcal{H}$ be any function with discretization $\tilde{h}$. We have if $\tilde{h}(\mathbf{x}_t) \geq s_t$, then $\tilde{h}(\mathbf{x}_t) \geq s'_t + \alpha$, meaning that $h(\mathbf{x}_t) \geq \tilde{h}(\mathbf{x}_t) - \alpha \geq s'_t$. If $\tilde{h}(\mathbf{x}_t) < s_t$, then $\tilde{h}(\mathbf{x}_t) \leq s_t - \alpha$. This implies $h(\mathbf{x}_t) < \tilde{h}(\mathbf{x}_t) + \alpha \leq s'_t$, where we used the fact that the discretization resolves ties by choosing the larger one. This completes the proof. □

The following lemma bounds the size of discretized class w.r.t. pseudo-dimension, which is due to [12, Corollary 3].

Lemma 6. *For any discretized class $\tilde{\mathcal{H}} \subset J^{\mathcal{X}}$ and $\mathbf{x}^T \in \mathcal{X}^T$ we denote by $\tilde{\mathcal{H}}_{\mathbf{x}^T}$ the class of $\tilde{\mathcal{H}}$ restricted on $\mathbf{x}^T$. Then*

$$|\tilde{\mathcal{H}}_{\mathbf{x}^T}| \leq (T|J|)^{\mathsf{P}(\tilde{\mathcal{H}})}.$$

We now state our main upper bound:

Lemma 7. *For any $\mathbf{x}^T \in \mathbb{R}^{dT}$, we have*

$$r^*_T(\mathcal{H}^{\mathsf{mono}} \mid \mathbf{x}^T) \leq 2d\log T + 1.$$

Proof. Let $\tilde{\mathcal{H}}$ be the discretization of $\mathcal{H}^{\mathsf{mono}}$ with step size $2/T$. Note that $\tilde{\mathcal{H}}$ is an $1/T$-cover of $\mathcal{H}^{\mathsf{mono}}$, meaning that

$$r^*_T(\mathcal{H}^{\mathsf{mono}} \mid \mathbf{x}^T) \leq r^*_T(\tilde{\mathcal{H}} \mid \mathbf{x}^T) + 1.$$

By the above Lemmas 4–6 , we have $\mathsf{P}(\tilde{\mathcal{H}}) \leq d$. By Lemma 4, we conclude that

$$\begin{aligned} r^*_T(\tilde{\mathcal{H}} \mid \mathbf{x}^T) &\leq \log|\tilde{\mathcal{H}}_{\mathbf{x}^T}| \\ &\leq d\log(T|J|) \leq d\log(T^2/2) \leq 2d\log T. \end{aligned}$$

This completes the proof. □

In conclusion, we arrive at our next main result.

Theorem 2. *For any monotone class $\mathcal{H}^{\mathsf{mono}}$ and almost all $\mathbf{x}^T$ (under Lebesgue measure) the following holds*

$$(d-1)\log\left(\frac{T-1}{d-1}\right) + \log 2 \leq r^*_T(\mathcal{H}^{\mathsf{mono}}|\mathbf{x}^T) \leq 2d\log T + 1.$$

C. Logistic Regression: Lower Bound for $R = \Omega(\sqrt{T})$

We now consider the case when $\mathbf{w}$ is finite but may grow with T. We know from [26] that the growth $(d/2)\log T$ holds for $R = \exp(o(\log T))$. We also know from Lemma 2 and Theorem 2 that the minimax regret for unbounded $\mathbf{w}$ (i.e., $R = \infty$) is

$$r^*_T(\mathcal{H}^{\mathsf{mono}}|\mathbf{x}^T) = \Theta(d\log(T/d))$$

for the class $\mathcal{H}^{\mathsf{mono}}$. Two questions naturally arise: (i) For what $R(T)$ the growth of the minimax regret transits from $d\log T$ to $(d/2)\log T$; (ii) Is there a smooth transition from $(d/2)\log T$ to $d\log T$, and can the regret grows like $D\log T$ where $d/2 \leq D \leq d$ when $R \to \infty$? Here we only partially answer question (i) for logistic regression.

Theorem 3. *For $R = \Omega(\sqrt{T})$ the average minimax regret for the logistic regression grows as $d\log T$, more precisely*

$$\bar{r}^*_T(\mathcal{H}^{\mathsf{logistic}}) \geq (d-3)\log\frac{T-1}{d-1} + \log(\frac{6}{\pi^2}) + o(1)$$

for $d > 3$ and $\mathbf{x}^T$ distributed $i.i.d.$ over a unit sphere.

Proof. We consider now sequences that are exactly at Hamming distance 1 from a splittable sequence $\mathcal{SP}(\mathbf{x}^T)$ as defined in (10). To *avoid overlaps* between them we only consider sequences that are within distance one from a unique splitable sequence. In other words, we consider about $T^{d-1}/T^2 = T^{d-3}$ splittable sequences for which the distance one Hamming balls are disjoint. Then we follow the derivation from the proof of Theorem 1. In particular, we conclude from (14) that for $R = \Omega(\sqrt{T})$ (recall that $R = \|\mathbf{w}\|_2 = w$):

$$S_T(\mathbf{x}^T)/|\mathcal{SP}(\mathbf{x}^T)| \geq \frac{3}{\pi^2} + O(\exp(-\sqrt{T})).$$

Since instead of using all splittable sequences we only consider those that are at least of distance two, that is $O(T^{d-3})$ to avoid the overlap, and this gives us the desired lower bound for all $R = \Omega(\sqrt{T})$. □

References

[1] Martin Anthony, Peter L Bartlett, Peter L Bartlett, et al. *Neural network learning: Theoretical foundations*, volume 9. cambridge university press Cambridge, 1999.

[2] A. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Trans. Inf. Theory*, 44(6):2743–2760, Oct. 1998.

[3] Alankrita Bhatt and Young-Han Kim. Sequential prediction under log-loss with side information. In *Algorithmic Learning Theory*, pages 340–344. PMLR, 2021.

[4] Blair Bilodeau, Dylan J Foster, and Daniel M Roy. Minimax rates for conditional density estimation via empirical entropy. *arXiv preprint arXiv:2109.10461*, 2021.

[5] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006.

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

[7] Nicolo Cesa-Bianchi and Ohad Shamir. Efficient transductive online learning via randomized rounding. *arXiv preprint arXiv:1106.2429*, 2011.

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

[9] M. Drmota and W. Szpankowski. *Analytic Information Theory: From Compression to Learning*. Cambridge University Press, Cambridge, 2023.

[10] P. Flajolet and R. Sedgewick. *Analytic Combinatorics*. Cambridge University Press, 2008.

[11] Dylan J Foster, Satyen Kale, Haipeng Luo andMehryar Mohri, and Karthik Sridharan. Logistic regression: The importance of being improper. In *COLT - Conference on Learning Theory*, 2018.

[12] David Haussler and Philip M Long. A generalization of sauer's lemma. *Journal of Combinatorial Theory, Series A*, 71(2):219–240, 1995.

[13] E. Hazan, T. Koren, and K. Y. Levy. Logistic regression: Tight bounds for stochastic and online optimization. In *The 27th Conference on Learning Theory, COLT 2014*, pages 197–209. MIT press, 2014.

[14] P. Jacquet, G. I. Shamir, and W. Szpankowski. Precise minimax regret for logistic regression with categorical feature values. In *PRML: ALT'21*, volume 132, pages 755–771, 2021.

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

[16] Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning: Random averages, combinatorial parameters, and learnability. *NIPS*, 2010.

[17] G. I. Shamir. Logistic regression regret: What's the catch? In *COLT*, 2020.

[18] G. I. Shamir and W. Szpankowski. Low complexity approximate bayesian logistic regression for sparse online learning. In *ISIT21 and ArXiv: http://arxiv.org/abs/2101.12113*, 2021.

[19] Gil I Shamir. Logistic regression regret: What's the catch? In *Conference on Learning Theory*, pages 3296–3319. PMLR, 2020.

[20] Y. M. Shtarkov. Universal sequential coding of single messages. *Problems of Information Transmission*, 23(3):3–17, Jul.-Sep. 1987.

[21] W. Szpankowski. *Average Case Analysis of Algorithms on Sequences*. Wiley, New York, 2001.

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

[23] J. Takeuchi and A. Barron. Asymptotically minimax regret by bayes mixtures for non-exponential families. In *IEEE Information Theory Workshop*, 2006.

[24] C. Wu, M. Heidari, A. Grama, and W. Szpankowski. Sequential vs fixed design minimax regrets for learning. In *ISIT*, 2022.

[25] Changlong Wu, Mohsen Heidari, Ananth Grama, and Wojciech Szpankowski. Expected worst case regret via stochastic sequential covering. *arXiv preprint arxiv.org/abs/2209.04417*, 2022.

[26] Changlong Wu, Mohsen Heidari, Ananth Grama, and Wojciech Szpankowski. Precise regret bounds for log-loss via a truncated bayesian algorithm. In *Advances in Neural Information Processing Systems*, 2022.