A General Lower Bound for Regret in Logistic Regression

Gil I. Shamir and Wojciech Szpankowski, *Fellow, IEEE,*

***Abstract*—We study logistic regression with binary features in which the number (or degree) of occurring features determines the label probability. This model fits one of social networks, where the number of friends determines the likelihood of outcomes instead of the identity of the friends, or more generally, a graph model, where the degree of a node can determine its behavior. It includes the case in which weights can be viewed as i.i.d. (e.g., in Bayesian modeling). For such a model, we introduce the *maximal minimax regret* that we analyze using a unique combination of analytic combinatorics and information theory. More importantly, the resulting regret is a general lower bound for the pointwise regret of a general logistic regression over all algorithms (learning distributions). We show that the introduced worst case (maximum over feature sequences) maximal minimax regret grows asymptotically as $(d/2)\log T + (d/2)\log(\pi/2) + O(d/\sqrt{T})$ for dimensionality $d = o(\sqrt{T})$, which is a lower bound for a regret of a general logistic regression. We extend our results to loss functions other than logistic loss and non-binary labels. Finally, if label probabilities are restricted to be monotonic with the degree of the example, we provide precise results for the minimax regret showing that the leading term decreases to $(d/2)\log(T/d^3)$ for large d.**

I. Introduction

Logistic regression has recently received a lot of attention in machine learning (see [24]) due to several important applications from category classification to risk assessment. It consists of a set of features, whose parameters represent their effect on some outcome. In a supervised online setup, a model is trained to learn these parameters from examples whose outcomes are already labeled. The training algorithm consumes data in rounds, where at each round $t \in \{1, 2, \ldots, T\}$, it is allowed to predict the label based only on the labels it observed in the past $t-1$ rounds and features up to time t. The prediction algorithm incurs for each round some *loss* and updates its belief of the model parameters. The (pointwise) *regret* of an online algorithm is defined as the (excess) loss it incurs over some value of a constant *comparator* (weight vector) that is used for prediction of the complete sequence.

Here we introduce and analyze the *maximal minimax regret* that for a given feature sequence maximizes the pointwise regret over label sequences and minimizes over learned distributions. Such a minimax regret was analyzed in information theory in the context of universal compression as discussed in [29], [7] while (pointwise) regret for logistic regression has been studied in [11], [14], [18], [20], [17], [27].

In this paper, instead of learning an individual parameter per feature, we study the problem where the outcome or *label* observed for an example is a function of the number (degree) of nonzero binary features present in the example (see assumption (16) in Section III). This setup can characterize social networks, in which the likelihood of some outcome depends on the number of friends a person has instead of on who those friends are. More generally, this setup can be representative of graphs, where an outcome in some node depends on the degree of the node, and not on which nodes are its neighbors (e.g., see graph structural compression in [2] and [19]). Given that the class of parameters for which we precisely compute the regret for this problem is a subset of the class of parameters in the standard logistic regression setup, this regret is a "universal" lower bound for a general pointwise regret.

We first focus on *binary* labels and consider the case of binary features $\mathbf{x}_t$ (i.e., the examples or feature values vector at any round t is a sparse vector in $\{0,1\}^d$, where d is the dimension – the number of features). For such a set up, we introduce the *maximal minimax regret* that we analyze using a unique combination of analytic combinatorics and information theory. In Theorem 1 we rigorously show that the introduced minimax regret serves as a general lower bound for a general pointwise regret over any algorithm/ learning distribution. In this in mind, we turn our attention to

Parts of this paper were presented at the 2011 IEEE International Symposium on Information Theory, St. Petersburg, Russia.

Gil I. Shamir is with Google Inc. Pittsburgh, PA, USA (email: gshamir@ieee.org).

W. Szpankowski is with the Department of Computer Science, Purdue University, IN 47907, USA (e-mail: spa@cs.purdue.edu). This work was partially supported by the NSF Center for Science of Information (CSoI) Grant CCF-0939370, and in addition by NSF Grants CCF-1524312, CCF-2006440, and CCF-2007238.

precise analysis of the introduced minimax regret. In Theorem 2 we show that the *average* maximal minimax regret grows asymptotically like $\frac{d}{2}\log T + \frac{1}{2}\log(\alpha_1\cdots\alpha_d) + \frac{d}{2}\log(\pi/2) + O(d/\sqrt{T})$ where T denotes the number of rounds, and α_i is the fraction of feature vector $\mathbf{x}_t$ with exactly j active features for $t=1,\ldots,T$. We also show that the worst case (maximum over all feature sequences $\mathbf{x}^T$) maximal minimax behaves asymptotically like $\frac{d}{2}\log\frac{T}{d} + \frac{d}{2}\log(\pi/2) + O(d!/\sqrt{T})$ for large T. We also consider monotone label distributions and show in Theorem 4 that the worst case minimax is reduced to $\frac{d}{2}\log\left(\frac{T}{d^3}\right) + \frac{d}{2}\log(\pi e^2/2) - \frac{1}{2}\log(2\pi d) + O(d/\sqrt{T})$. Finally, in Theorem 5 we extend it to non-binary labels of size m such that $m^{3/2}d = o(\sqrt{T})$. Furthermore, in Corollary 1 we summarize our findings regarding the lower bound on the standard regret. At last, in Theorem 3 we briefly discuss Bayesian setting and present results for the pointwise and average regrets.

Studying this problem, we adapt techniques from the universal compression literature (see [29], [32], [33], [34]) and analytic combinatorics (see [9], [31]) that apply complex asymptotics to solve discrete problems. We first review various notions of regret and redundancy from information theory that we adopt to the performance evaluation of logistic regression. Here, we assume that $d = 1$ and alphabet is of size m. 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 $\mathcal{A}$ of size m are defined as

$$\begin{aligned} R_T(P;y^T) &= L(y^T) + \log P(y^T), \\ \bar{R}_T(P) &= \mathbf{E}[L(Y^T)] - H_T(P), \end{aligned}$$

where $H_T(P)$ is the entropy for a block of length T, $\mathbf{E}$ denotes the expectation, and $L(y^T)$ is the code length of some code $L(\cdot)$ (both redundancy quantities are implicitly functions of the code). In the online learning – and indeed in information theory – one ignores the integer nature of the length and replace it by $L(y^T) = -\log Q(y^T)$ for some unknown distribution Q that approximates P. The definitions above imply a probabilistic setting, in which there is some source that generated the data. A non-probabilistic setting considers *individual sequences* (see, e.g., [29]), where we define the *maximal* or *worst case* redundancy as

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

In practice, one can only hope to have some knowledge about a family of sources $\mathcal{S}$ that generates real data. Following Davisson [5], we define the average minimax redundancy $\bar{R}_T(\mathcal{S})$ and the worst case (maximal) minimax redundancy $R_T^*(\mathcal{S})$ for family $\mathcal{S}$ as follows

$$\begin{aligned} \bar{R}_T(\mathcal{S}) &= \min_Q \sup_{P\in\mathcal{S}} \sum_{y^T} P(y^T)\log[P(y^T)/Q(y^T)], \\ R_T^*(\mathcal{S}) &= \min_Q \sup_{P\in\mathcal{S}} \max_{y^T} \left\{\log\left[P(y^T)/Q(y^T)\right]\right\}. \end{aligned}$$

In words, we search for the best code or distribution Q for the worst source P on average and for the worst label sequence y^T for individual sequences.

There are other measures of optimality for coding, gambling, learning, and prediction that are used in universal modeling, learning, and coding. We refer here to minimax *regrets* defined as follows (cf. [7], [33], [34]):

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

and to the maxmin regret

$$\underline{r}_T(\mathcal{S}) = \sup_{P\in\mathcal{S}} \min_Q \mathbf{E}[-\log Q(y^T) + \log\sup_{P\in\mathcal{S}} P(y^T)].$$

We call $\bar{r}_T(\mathcal{S})$ the *average* minimax regret, $r_T^*(\mathcal{S})$ the worst case (maximal) minimax regret and $\underline{r}_T(\mathcal{S})$ the maxmin regret. Clearly, $\bar{R}_T(\mathcal{S}) \le \bar{r}_T(\mathcal{S})$, and, $r_T^*(\mathcal{S}) = R_T^*(\mathcal{S})$.

In [7] it is also 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

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

Furthermore, it is known [7], [21], [23], [25], [30], [33], [34] that for a large class of sources (up to Markovian but not for non-Markovian [10], [7]) 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 [21], [25], [32]). In fact in [32] full asymptotic expansions were derived for all ranges of m.

Finally, we review ML literature with respect to pointwise regret. To the best of our knowledge, in the ML literature [18] was first to demonstrate results that suggest that pointwise regret for logistic regression grows like $O(d\log T/d)$w where for fixed dimension d and $m = 2$, which was further generalized in [11] to all m. The authors of [18] used *Bayesian model averaging*. The $O(\log T)$ pointwise and individual sequence regret can be achieved for the single dimensional problem with gradient methods based approaches, as was demonstrated in [20]. The authors of [20] then posed the problem of what happens for larger dimensions. Subsequently,

[11] demonstrated how to achieve regret bounds of $O(d\log(T/d))$ with Bayesian model averaging. These results were strengthened in [27], which also provided matching lower bounds (see also [17], [15]). We finally should point out that our results on minimax regret are not restricted to Bayesian modeling.

The paper is organized as follows. In the next section we formally introduce the minimax regret and represent is as a Shtarkov-like sum [29], [7]. Then we present our main results which we prove in the last section.

II. Problem Formulation and Notation

We denote by $\mathbf{x}_t = (x_{1,t}, \ldots, x_{d,t})$ a d-dimensional *binary* feature vector where $x_{i,t} \in \{0,1\}$ for $t = 1,\ldots,T$. Notice that $\mathbf{x}^T$ is a $T \times d$ matrix with $\mathbf{x}_t = (x_{1,t}, \ldots, x_{d,t})$ as a row. The label binary vector is denoted as $y^T = (y_1, \ldots, y_T)$ with $y_t \in \{-1, 1\}$. Finally, $\mathbf{w}_t = (w_{1,t}, \ldots, w_{d,t})$ is a d-dimensional vector of feature weights. In this paper, we do not address the method used to learn the weights (e.g., gradient method or Bayesian mixing). Instead, we assume that the weight are *exchangeable* (see assumption (D) in the next section) leading to our model in which the number of features determines the label probability. Notice that a practical learning algorithm will be agnostic to the vector $\mathbf{w}$, and instead will learn “features”, which are a function of the degree of $\mathbf{x}_t$, i.e., the number of nonzero components in $\mathbf{x}_t$.

The *logistic loss* of an algorithm that *plays* $\mathbf{w}_t$ at round t is

$$L(y^T|\mathbf{x}^T, \mathbf{w}^T) := \sum_{t=1}^{T} \log\left[1 + \exp(-y_t\langle \mathbf{x}_t, \mathbf{w}_t\rangle)\right] \quad (1)$$

where $\langle \mathbf{x}_t, \mathbf{w}_t\rangle = \sum_{i=1}^d x_{i,t}w_{i,t}$. It is convenient to write $\ell(y_t|\mathbf{x}_t, \mathbf{w}_t) := \log\left[1 + \exp(-y_t\langle \mathbf{x}_t, \mathbf{w}_t\rangle)\right]$. Both $\ell(y_t|\mathbf{x}_t, \mathbf{w}_t)$ and $L(y^T|\mathbf{x}^T, \mathbf{w}^T)$ depend on $\mathbf{x}_t$ and $\mathbf{w}_t$ only through the product $\langle \mathbf{x}_t, \mathbf{w}_t\rangle$.

The probability of a label is given by

$$P(y_t|\mathbf{x}_t, \mathbf{w}_t) = \frac{1}{1 + \exp(-y_t\langle \mathbf{x}_t, \mathbf{w}_t\rangle)} \quad (2)$$

and then

$$\ell(y_t|\mathbf{x}_t, \mathbf{w}_t) = -\log P(y_t|\mathbf{x}_t, \mathbf{w}_t).$$

However, in many applications (e.g., when approximating the logistic function by a Gaussian distribution [28]) it is desirable to consider a larger class of loss function (e.g., see [28]). Since the logistic function depends only on the product $y_t\langle \mathbf{x}_t, \mathbf{w}_t\rangle$ we set $w_t = \langle \mathbf{x}_t, \mathbf{w}_t\rangle$ and define a function $F(w)$ that satisfies the following properties

$$F(-w) = 1 - F(w),\ F(0) = .5,\ F(-\infty) = 0, F(\infty) = 1. \quad (3)$$

For example, F could be the logistic function as defined in (2) or the Gaussian Cumulative Distribution Function (CDF) $F(y_t w_t) = \Phi(\sqrt{\pi/8}y_t w_t)$ as in [1]. Then, we can re-write (2) as

$$P(y_t|\mathbf{x}_t, \mathbf{w}_t) = F(y_t\langle \mathbf{x}_t, \mathbf{w}_t\rangle), \quad (4)$$
$$\ell(y_t|\mathbf{x}_t, \mathbf{w}_t) = -\log F(y_t\langle \mathbf{x}_t, \mathbf{w}_t\rangle). \quad (5)$$

Finally, we observe that the goal of a learning algorithm is to find the best approximation $Q(y_t|\mathbf{x}_t)$ of the unknown distribution $P(y_t|\mathbf{x}_t, \mathbf{w}_t)$. Therefore, we shall write

$$\ell_Q(y_t|\mathbf{x}_t) = -\log Q(y_t|\mathbf{x}_t).$$

Here, Q represents an algorithm that predicts y_t.

The *pointwise regret* is defined for individual sequences $(y_t, \mathbf{x}_t)$ as in [13], [11], [27]

$$r(y^T, Q|\mathbf{x}^T) := \sum_{t=1}^{T} \ell_Q(y_t|\mathbf{x}_t) - \min_{\mathbf{w}} \sum_{t=1}^{T} \ell(y_t|\mathbf{x}_t, \mathbf{w})$$

for some fixed $\mathbf{w}$. Thus

$$r(y^T, Q|\mathbf{x}^T) = \log \frac{\sup_{\mathbf{w}} P(y^T|\mathbf{x}^T, \mathbf{w})}{Q(y^T|\mathbf{x}^T)} \quad (6)$$
$$= \log \frac{\sup_{\mathbf{w}} F(y^T\langle \mathbf{x}^T, \mathbf{w}\rangle)}{Q(y^T|\mathbf{x}^T)}. \quad (7)$$

The pointwise regret $r(y^T, Q|\mathbf{x}^T)$ is a function of y^T and $\mathbf{x}^T$, so it depends on *individual* sequences. Furthermore, it depends on the algorithm, represented by $Q(y^T|\mathbf{x}^T)$. A better measure of online logistic regression performance should decouple the regret from the fluctuations of y^T (but may still depend on the feature vector $\mathbf{x}^T$) and the learning algorithm Q. Following information-theoretic view as in [5], [7], [34], we define the *maximal minimax regret* (conditioned on $\mathbf{x}^T$) as

$$r_T^*(\mathbf{x}^T) := \inf_{Q} \max_{y^T} [r(y^T, Q|\mathbf{x}^T)]. \quad (8)$$

Notice that this definition is over all possible learning algorithms represented by Q and therefore it constitutes a lower bound for a general regret over all algorithms.

Following [29], [7] we first find a more succinct representation of the maximal minimax regret. Let

$$P^*(y^T|\mathbf{x}^T) := \frac{\sup_{\mathbf{w}} P(y^T|\mathbf{x}^T, \mathbf{w})}{\sum_{v^T} \sup_{\mathbf{w}} P(v^T|\mathbf{x}^T, \mathbf{w})} \quad (9)$$

be the *maximum-likelihood distribution*. Then

$$\begin{aligned} r_T^*(\mathbf{x}^T) &= \min_{Q} \sup_{\mathbf{w}} \max_{y^T} (-\log Q(y^T|\mathbf{x}^T) + \log P(y^T|\mathbf{x}^T, \mathbf{w})) \\ &= \min_{Q} \max_{y^T} [\log P^*(y^T|\mathbf{x}^T)/Q(y^T|\mathbf{x}^t)] \\ &\quad + \log \sum_{y^T} \sup_{\mathbf{w}} P(y^T|\mathbf{x}^T, \mathbf{w}) \\ &= \log \sum_{y^T} \sup_{\mathbf{w}} P(y^T|\mathbf{x}^T, \mathbf{w}) =: \log d_T(\mathbf{x}^T) \end{aligned} \quad (10)$$

if we chose $Q(y^T|\mathbf{x}^T) = P^*(y^T|\mathbf{x}^T)$. The above sum is often called the Shtarkov sum; see [29], [7], [12]. Observe that for not optimal Q (i.e., $Q \neq P^*$) there will be extra $O(1)$ term in the maximal minimax regret. We also write $\sup_P P(y^T|\mathbf{x}^T, \mathbf{w})$ for $\sup_{\mathbf{w}} P(y^T|\mathbf{x}^T, \mathbf{w})$.

In the next section in Theorem 2 (see Theorem 5 for non-binary features) we precisely evaluate the above Shtarkov sum under additional assumption (D) (see (16) below) in which we postulate that the label probability is a function of the number (degree) of active features. In Theorem 4 we present results for the regret when the label probabilities are monotone. But first in Theorem 1 below we show that such a regret with assumption (D) constitutes a lower bound for a general regret and arbitrary feature values.

Notice that $r_T^*(\mathbf{x}^T)$ is still a function of the feature vector $\mathbf{x}^T$. To bypass this dependency, we define the *worst case maximal minimax* r_T^* as

$$r_T^* = \max_{\mathbf{x}^T} \left[\log \sum_{y^T} \sup_{\mathbf{w}} P(y^T|\mathbf{x}^T, \mathbf{w}) \right], \qquad (11)$$

that is,

$$r_T^* = \max_{\mathbf{x}^T} \inf_{Q} \max_{y^T} [r(y^T|\mathbf{x}^T)].$$

This worst case minimax regret is the closest to the minimax formulation of [22].

We may also take a probabilistic view point and assume that the feature vector is a realization of a random sequence $\mathbf{X}^T$. This leads to the *average maximal minimax regret* defined as

$$\bar{r}_T^* = \mathbf{E}_{\mathbf{X}^T}[r_T^*(\mathbf{X}^T)] = \mathbf{E}_{\mathbf{X}^\mathbf{T}}[\log d_T(\mathbf{X}^T)]. \qquad (12)$$

In Theorem 2 below we summarize our findings regarding the average maximal minimax regret for any learning algorithm Q and the sequence $z^T = (y^T, \mathbf{x}^T)$.

Finally, in the Bayesian modeling, the learning distribution is a mixture over $\mathbf{w}$ with a prior $\rho(\mathbf{w})$ defined as

$$Q(y^T|\mathbf{x}^T) := \int_{\mathbf{w}} \rho(\mathbf{w}) P(y^T|\mathbf{x}^T, \mathbf{w}) d\mathbf{w}.$$

In this case, the pointwise regret becomes

$$r(y^T, \rho|\mathbf{x}^T) = \log \frac{\sup_{\mathbf{w}} P(y^T|\mathbf{x}^T, \mathbf{w})}{Q(y^T|\mathbf{x}^T)}. \qquad (13)$$

As in the worst case scenario, we can bypass dependency of $\mathbf{x}^T$ by taking the maximum over $\mathbf{x}^T$ or average over feature distribution. In Theorem 3 below we present some precise results for this Bayesian regret.

III. Main Results

Throughout this section, we assume that features are *binary*, that is either $x_{i,t} = 1$ (active) or $x_{i,t} = 0$, unless stated otherwise. We consider two scenarios: in the *deterministic* case we assume that $\mathbf{x}^T$ is given, while in the *stochastic* case we assume that $\mathbf{x}^T$ is stochastically generated by some distribution. In both cases, we define T_j as the number of rounds t for which $\sum_{i=1}^d x_{i,t} = j$, that is, the number of feature vectors with exactly j active features. More formally,

$$T_j = |\{t : \sum_{i=1}^d x_{i,t} = j\}|.$$

Furthermore, by $\alpha_j > 0$ we denote the fraction of $\mathbf{x}_t$ ($t = 1, \ldots, T$) that has exactly j active features. Hence in the deterministic scenario $\alpha_j = T_j/T > 0$ (i.e., $T_j > 0$) and $T = T_1 + \cdots + T_d$.

In the the stochastic scenario we have

$$\alpha_j := P\left(\sum_{i=1}^d x_{i,t} = j\right) \qquad (14)$$

as the probability that exactly j features are equal to 1. In particular, if we assume that $\mathbf{x}_t = (x_{1,t}, \ldots, x_{d,t})$ is distributed as the binomial(d, p) where $P(x_{i,t} = 1) = p$, then for all $i \in [d]$ and all t.

$$\alpha_j := P\left(\sum_{i=1}^d x_{i,t} = j\right) = \binom{d}{j} p^j (1-p)^{d-j}. \qquad (15)$$

Furthermore, in the stochastic scenario, $(T_1, \ldots, T_d)$ are random variables distributed as the multinomial $(T, \alpha_1, \ldots, \alpha_d)$, that is,

$$P(T_1, \ldots T_d) = \binom{T}{T_1, \ldots, T_d} \alpha_1^{T_1} \cdots \alpha_d^{T_d}$$

where $T = T_1 + \cdots + T_d$.

We now introduce the main assumption about $\mathbf{w}^T$ that converts the problem into one that depends only on the degree distribution (i.e., number of active features of $\mathbf{x}^T$). We will assume that $\mathbf{w}_t = (w_{1,t}, \ldots, w_{d,t})$ generated according to a prior distribution are *exchangeable*. This defines our model in which the number (or degree) of occurring features determines the label probability. More precisely,

(D) For every k tuple $(j_1, \ldots, j_k) \in \{1, \ldots, d\}$ and all $1 \le t < s \le T$ we have

$$w^{(k)} := w_{1,t} + \cdots + w_{k,t} \stackrel{d}{=} w_{j_1,s} + \cdots + w_{j_k,s} \qquad (16)$$

where $\stackrel{d}{=}$ means equal "in distribution". In other words, the sum of k weights (those weights that have the corresponding features $x_{i,t} = 1$) for any time t has the same distributions that we denote as $w^{(k)} := w_{1,t} + \cdots + w_{k,t}$.

Observe that assumption (D) holds if all weights are *identically and independently distributed* (i.i.d.). Indeed, if say a weight distribution is W, then $w_{1,t}+\cdots+w_{k,t} \overset{d}{=} W \star \cdots \star W$ for all t where $\star$ denotes convolution.

Assumption (D) reformulates the problem w.r.t. the degree distribution of the feature vector. Define $1(\mathbf{x}_t) := \sum_{i=1}^d x_{i,t}$ as the number of 1's in $\mathbf{x}_t$. Then, all probabilities $P(y_t|1(\mathbf{x}_t) = k)$ for a given k are equal, and we denote them as

$$\theta_k(y_t) := P(y_t|1(\mathbf{x}_t) = k) = F(y_t w^{(k)})$$

where F is defined in (3). We can view the above as a change of measure from $\mathbf{w}$ to $\boldsymbol{\theta}$. Finally let $\theta_k := \theta(y_t = 1) = F(w^{(k)})$. Then

$$P(y^T|\mathbf{x}^T, \boldsymbol{\theta}) = \prod_{j=1}^d \theta_j^{k_j}(1-\theta_j)^{T_j-k_j} \tag{17}$$

where $T_1 + \cdots + T_d = T$ and k_j is the number of $y_t = 1$ among T_j with j active features.

Next, to estimate the maximal minimax regret, we need to compute

$$\sup_{\boldsymbol{\theta}} P(y^T|\mathbf{x}^T, \boldsymbol{\theta}) = \sup_P P(y^T|\mathbf{x}^T, \boldsymbol{\theta})$$

which actually becomes

$$\sup_{\boldsymbol{\theta}} P(y^T|\mathbf{x}^T, \boldsymbol{\theta}) = \prod_{j=1}^d \left(\frac{k_j}{T_j}\right)^{k_j} \left(\frac{T_j - k_j}{T_j}\right)^{T_j-k_j}. \tag{18}$$

This leads to the following Shtarkov sum

$$d_T(\mathbf{x}^T) = \sum_{k_1=0}^{T_1} \binom{T_1}{k_1} \left(\frac{k_1}{T_1}\right)^{k_1} \left(\frac{T_1-k_1}{T_1}\right)^{T_1-k_1} \cdots \tag{19}$$

$$\cdots \sum_{k_d=0}^{T_d} \left(\frac{k_d}{T_d}\right)^{k_d} \left(\frac{T_d-k_d}{T_d}\right)^{T_d-k_d}.$$

This is a sophisticated sum to evaluate. In the next section (see in particular Lemma 1) we present a methodology that gives a precise asymptotic approximation of it.

A. Lower Bound

The model just introduced with assumption (D) has another interesting and useful property. It turns out that it constitutes a lower bound for the standard minimax regret $R_T^*(\mathbf{x}^T)$. We compare it to the minimax regret $r_T^*(\mathbf{x}^T)$ defined above under the assumption (D) with binary features $\mathbf{x}^T \in \{0,1\}^T$.

Theorem 1. *The minimax regret $r^*(\mathbf{x}^T)$ with binary features under assumption (D) constitutes a lower bound for a general maximal minimax regret $R_T^*(\mathbf{x}^T)$, that is*

$$R_T^*(\mathbf{x}^T) \geq r_T^*(\mathbf{x}^T). \tag{20}$$

The regret $r_T^*(\mathbf{x}^T)$ is precisely estimated in Theorems 2 and 5 below.

Proof: In a general case, the probability $P(y^T|\mathbf{x}^T, \mathbf{w}))$ is a complicated product of probabilities that depend not only of how many active features are there but also on their locations. Let us group all probabilities in which there are exactly one active feature under $\theta_{1,t}(y_t|\mathbf{x}^t)$, two active features under $\theta_{2,t}(y_t|\mathbf{x}^t)$, and so on until all active features under $\theta_{d,t}(y_t|\mathbf{x}^t)$. Now, to lower bound $\sup_P P(y^T|\mathbf{x}^T, \mathbf{w})$ we choose particular values for $\theta_{j,t}(y_t|\mathbf{x}^t)$ for each j. Namely, we set

$$\theta_{j,t}(y_t|\mathbf{x}^t) = \frac{k_j}{T_j}$$

where, as before, k_j is the number of $y_t = 1$ among T_j that have j active feature. But then the (log of) Shtarkov sum, and hence the maximal minimax regret becomes exactly $R_T^*(\mathbf{x}^T)$ under our model (D). This completes the proof. ■

B. Precise Regret for Binary Labels

We now deal with $r_T^*(\mathbf{x}^T)$ under assumption (D). Using analytic combinatorics [31] and binomial sum asymptotics [16], [8], we will prove in the next section the following result regarding the asymptotic expansion of the average and worst case minimax regret under assumption (D).

Theorem 2. *Under assumption (D) for any function F satisfying (3) the average maximal minimax regret for $d = o(\sqrt{T})$ is given by*

$$\bar{r}_T^* = \frac{d}{2}\log(T) + \frac{1}{2}\log(\alpha_1 \cdots \alpha_d) + \frac{d}{2}\log(\pi/2) + O(d/\sqrt{T}) \tag{21}$$

and its worst case minimax regret (maximum over $\mathbf{x}^T$ or in this case $(T_1, \ldots, T_d)$) is

$$r_T^* = \frac{d}{2}\log\left(\frac{T}{d}\right) + \frac{d}{2}\log(\pi/2) + O(d/\sqrt{T}) \tag{22}$$

for large T.

Let us now present results for the Bayesian pointwise regret as defined in (13) which becomes

$$r(y^T, \rho|\mathbf{x}^T) = \log \frac{\sup_{\theta_1,\ldots\theta_d} \prod_{j=1}^d \theta_j^{k_j}(1-\theta_j)^{T_j-k_j}}{\int_{\boldsymbol{\theta}} \rho(\boldsymbol{\theta}) \prod_{j=1}^d \theta_j^{k_j}(1-\theta_j)^{T_j-k_j}}. \tag{23}$$

By (18) we can re-write it as (with $T_i > 0$)

$$r(y^T, \rho|\mathbf{x}^T) = \log \frac{\prod_{j=1}^d \left(\frac{k_j}{T_j}\right)^{k_j} \left(\frac{T_j - k_j}{T_j}\right)^{T_j - k_j}}{\int_\theta \rho(\theta) \prod_{j=1}^d \theta_j^{k_j} (1-\theta_j)^{T_j - k_j}}. \tag{24}$$

In the next section, using Stirling's approximation, Dirichlet distribution, binomial sum asymptotics [3], [8], [16] and analytic combinatorics [31] we prove the following results.

Theorem 3. *For the Bayesian setting with Jeffrey's prior presented in (52) under assumption (D) for any function F satisfying (3), the pointwise regret in the deterministic setting becomes*

$$r(y^T|\mathbf{x}^T) = \frac{1}{2}\log(T_1 \cdots T_d) + \frac{d}{2}\log(\pi/2) + O(1/T) \tag{25}$$

for large T_j and k_j. The average pointwise regret is

$$\mathbf{E}_{\mathbf{x}}[r(y^T|\mathbf{x}^T] = \frac{d}{2}\log T + \frac{1}{2}\log(\alpha_1 \cdots \alpha_d) + \frac{d}{2}\log(\pi/2) + O(d/T) \tag{26}$$

for large T with $\alpha_i > 0$.

C. Monotone Label Distribution

We now study a natural extension of our previous results. We observe that a reasonable behavior that one could expect from the degree distribution logistic regression problem is one of monotonicity of the probability as function of the number of active features in an example. The probability of a positive label should either increase for every added feature (or at least not decrease), or if the features negatively affect this probability, every added active feature should decrease (or not increase) the probability of the positive label. Let us then consider a class of non-decreasing label probabilities [26], that is,

$$\mathcal{M} = \{\boldsymbol{\theta} : \theta_1 \le \theta_2 \le \cdots \le \theta_d,\ 0 \le \theta_i \le 1\}.$$

We denote by $r_T^*(\mathcal{M}) = \log d_T(\mathcal{M})$ the worst case minimax regret for this class.

The main idea of our approach is as follows (details can be found in Section IV-C): Observe that the unconstrained set $\Theta = \{\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d) :\ 0 \le \theta_i \le 1\}$ can be divided into $d!$ subsets

$$\mathcal{M}_\pi = \{\boldsymbol{\theta} \in \Theta : \theta_{\pi(1)} \le \theta_{\pi(2)} \le ... \le \theta_{\pi(d)}\}$$

for any permutation π of $\{1, 2, \ldots, d\}$. Furthermore, for any symmetric functional f it is easy to prove (see [6])

$$\int_{\mathcal{M}} f(\theta_1, ..., \theta_d) d\theta_1 ... d\theta_d = \frac{1}{d!}\int_\Theta f(\theta_1, ..., \theta_d) d\theta_1 ... d\theta_d. \tag{27}$$

More precisely, let $d_T(\Theta)$ and $d_T(\mathcal{M})$ denote the Shtarkov sums for the the unconstrained distribution and the monotone distribution when $T_i = T/d$ (i.e, the worst case regret), respectively. Then, in the next section we prove that

$$d_T(\mathcal{M}) = \frac{1}{d!} d_T(\Theta)(1 + O(d!/\sqrt{T})). \tag{28}$$

This leads to our next main result with the proof delayed till the next section (see also [26], [6]).

Theorem 4. *Under assumption (D) for any function F satisfying (3), the worst case minimax regret $r_T(\mathcal{M})$ for the class $\mathcal{M}$ of monotone label distribution is*

$$r_T^*(\mathcal{M}) = r_T^* - \log d! + O(d!/\sqrt{T}), \tag{29}$$

where r_T^* is computed in (22). For large d such that $d = o(\log T / \log\log T)$ we have

$$r_T^*(\mathcal{M}) = \frac{d}{2}\log\left(\frac{T}{d^3}\right) + \frac{d}{2}\log(\pi e^2/2) - \frac{1}{2}\log(2\pi d) + O(d!/\sqrt{T}) \tag{30}$$

for large T.

We conclude that the restriction to monotone label probabilities decreases the regret by about $\log d!$, that is, the leading term decreases to $(d/2)\log(T/d^3)$, as observed in [26] for universal compression.

D. Regret for Non-binary Labels

Finally, we briefly discuss how to extend our results to non-binary labels, say, label alphabet $\mathcal{Y}$ of size m. Following [11] we need to extend the weight vector to the weight matrix $\mathbf{W} = [\mathbf{w}_1, \ldots,_{m-1}]$ such that $\mathbf{w}_i = (w_{1,i}, \ldots, w_{d,i})$. Then the multinomial logistic function known also as *softmax function* is defined as

$$P(y_t = \ell|\mathbf{x}_t, \mathbf{W}_t) = \frac{e^{\langle \mathbf{x}_t, \mathbf{w}_{t,\ell}\rangle}}{\sum_{k=1}^m e^{\langle \mathbf{x}_t, \mathbf{w}_{t,k}\rangle}} \tag{31}$$

for $\ell = 1, \ldots, m-1$. Also, as before we define

$$\theta_{k,\ell}(y_t) := P(y_t = \ell | 1(\mathbf{x}_t) = k, \mathbf{W})$$

leading to

$$P(y^T|\mathbf{x}^T, \boldsymbol{\theta}) = \prod_{j=1}^d \prod_{\ell=1}^m \theta_{j,\ell}^{k_{j,\ell}} \tag{32}$$

where $k_{j,\ell}$ is the number labels equal to ℓ among T_j.

Following the footsteps of our analysis for the binary labels and using general Lemma 1 we arrive at the following generalization of Theorem 2.

Theorem 5. *Under assumption (D) for any function F satisfying (3) the average maximal minimax regret for $m = O(1)$ and $d = o(\sqrt{T})$ becomes*

$$\bar{r}_T^* = \frac{d(m-1)}{2}\log(T/2) + \frac{(m-1)d}{2}\log(\alpha_1\cdots\alpha_d) + \\ + \frac{d}{2}\log(\pi/\Gamma^2(m/2)) + O(d/\sqrt{T}) \quad (33)$$

(with $\Gamma(x)$ being the Euler gamma function) while for $dm^{3/2} = o(\sqrt{T})$ we find

$$\bar{r}_T^* = \frac{d(m-1)}{2}\log(T/m) + \frac{(m-1)d}{2}\log(\alpha_1\cdots\alpha_d) + \\ + \frac{md}{2}\log e + O(dm^{3/2}/\sqrt{T}). \quad (34)$$

Furthermore, the worst case maximal minimax for $dm^{3/2} = o(\sqrt{T})$ is

$$r_T^* = \frac{d(m-1)}{2}\log(T/(dm)) + \frac{md}{2}\log e + O(dm^{3/2}/\sqrt{T}) \quad (35)$$

for large T.

Finally, we go back to the lower bound discussed in Theorem 1. We compare the general pointwise regret $r_T(y^T, Q|\mathbf{x}^T)$ to the worst case minimax regret r_T^* (maximized over $\mathbf{x}^T \in \{0,1\}^T$) just computed in Theorem 5. We conclude the following general lower bound.

Corollary 1. *Consider a general pointwise regret $r_T(y^T, Q|\mathbf{x}^T)$ for any algorithm/ distribution Q. The following holds for $dm^{3/2} = o(\sqrt{T})$*

$$\max_{(\mathbf{x}^T, y^T)} r_T(y^T, Q|\mathbf{x}^T) \geq \frac{d(m-1)}{2}\log(T/(dm)) \quad (36) \\ + \frac{md}{2}\log e + O(dm^{3/2}/\sqrt{T})$$

for large T.

IV. Analysis

In this this section we prove our main results Theorems 2– 5. To simplify notation we often drop conditioning on $\mathbf{x}_t$.

A. *Proof of Theorem 2 and 5*

We first prove Theorem 2. Recall that the maximal minimax regret $r_T^*(\mathbf{x}^T)$ is given by (10). To estimate it we need to find $\sup_{\boldsymbol{\theta}} P(y^T|\mathbf{x}^T, \theta)$. But as easy to see the optimal $\theta_j = k_j/T_j$ leading to (18) which we repeat below

$$\sup_{\boldsymbol{\theta}} P(y^T|\mathbf{x}^T) = \prod_{j=1}^{d} \left(\frac{k_j}{T_j}\right)^{k_j} \left(\frac{T_j - k_j}{T_j}\right)^{T_j - k_j}. \quad (37)$$

We start with the average minimax $\bar{r}_T^* = \log \bar{d}_T$ where

$$\bar{d}_T = \sum_{T_1+\cdots+T_d=T} \binom{T}{T_1,\ldots T_d} \prod_{j=1}^{d} \alpha_j^{T_j} \cdot \quad (38)$$
$$\cdot \sum_{k_1=0}^{T_1} \binom{T_1}{k_1} \left(\frac{k_1}{T_1}\right)^{k_1} \left(\frac{T_1-k_1}{T_1}\right)^{T_1-k_1} \cdots$$
$$\cdots \sum_{k_d=0}^{T_d} \left(\frac{k_d}{T_d}\right)^{k_d} \left(\frac{T_d-k_d}{T_d}\right)^{T_d-k_d}.$$

The above sum is quite complicated to estimate, but we will find asymptotics up to the $O(1)$. We write (38) as follows

$$\bar{d}_T = \sum_{T_1+\cdots+T_d=T} \binom{T}{T_1,\ldots T_d} \prod_{j=1}^{d} \alpha_j^{T_j} f(T_i)$$

where

$$f(T_i) = \sum_{k_i=0}^{T_i} \binom{T_i}{k_i} \left(\frac{k_i}{T_i}\right)^{k_i} \left(\frac{T_i-k_i}{T_i}\right)^{T_i-k_i}.$$

In Lemma 1 below we prove the following asymptotic expansion for $m = 2$

$$f(T_i) = \sqrt{\frac{\pi T_i}{2}}\left(1 + O(1/\sqrt{T_i}\right) \quad (39)$$

for large T_i. In general, for any fixed label alphabet size m, we prove in Lemma 1

$$f(T_i) = \left(\frac{T_i}{2}\right)^{\frac{m-1}{2}} \frac{\sqrt{\pi}}{\Gamma(m/2)}\left(1 + O(1/\sqrt{T_i}\right) \quad (40)$$

and for large $m = O(T^{1/3})$ we find

$$f(T_i) = \left(\frac{T_i}{m}\right)^{\frac{m-1}{2}} e^{m/2}\left(1 + O(m^{3/2}/\sqrt{T_i}\right).$$

These asymptotics are enough to establish (22) of Theorem 2 and (35) of Theorem 5.

To complete the proof for the average minimax regret as in (21) we appeal to the following binomial sum estimate [3]

$$\sum_{T_1+\cdots+T_d=T} \binom{T}{T_1,\ldots T_d} \prod_{j=1}^{d} \alpha_j^{T_j} f(T_j) =$$
$$= f(\alpha_1 T)\cdots f(\alpha_d T)(1 + O(dm^{3/2}/\sqrt{T})).$$

We provide some justification of the above binomial sum estimate in the following remark.

Remark 1. The binomial (better multinomial) sums will pop up again in this paper, so let us give an intuitive derivation. In general, a multinomial sum is defined as for any T

$$S_f(T) := \sum_{k_1+\cdots+k_d=T} \binom{T}{\mathbf{k}} \boldsymbol{\theta}^{\mathbf{k}} f(\mathbf{k}) \tag{41}$$

where $f(\mathbf{k}) : \mathbb{R}^d \to \mathbb{R}$ is a function of at most polynomial growth and $\boldsymbol{\theta}^{\mathbf{k}} = \prod_{i=1}^d \theta_i^{k_i}$ where $\theta_1 + \cdots + \theta_d = 1$. In [3], [8], [16] it is proved that such a sum grows asymptotically as $f(n\boldsymbol{\theta})$. For the reader's convenience we offer a streamlined justification; in particular when $f(\mathbf{k})$ has an analytic continuation to a complex cone around the real positive axis [16], [31]. In general, Taylor's expansion of f around $n\boldsymbol{\theta}$ is

$$f(\mathbf{x}) = f(n\boldsymbol{\theta}) + (\mathbf{x} - \boldsymbol{\theta})\nabla f(n\boldsymbol{\theta}) + \\ + \frac{1}{2}(\mathbf{x} - n\boldsymbol{\theta})\nabla^2 f(\mathbf{x}')(\mathbf{x} - n\boldsymbol{\theta})$$

for some $\mathbf{x}'$ in the vicinity of $n\boldsymbol{\theta}$, where we use the same simplified notations as before. Observe now that

$$\begin{aligned} S_f(T) &= \mathbf{E}[f(\mathbf{X})] && (42)\\ &= f(T\boldsymbol{\theta}) + O(T \max_{\mathbf{x}',ij} f''_{ij}(\mathbf{x}')) && (43)\\ &= f(T\boldsymbol{\theta}) + O(T\xi(n)), && (44)\end{aligned}$$

where $\mathbf{X}$ is a multinomial distribution with parameters T and $\boldsymbol{\theta}$ and $f''_{ij}(\mathbf{x})$ is the second derivative with respect to x_i and x_j which is order of magnitude smaller than the leading term when $f(x)$ is of a poly-logarithmic growth. Observe that we use the fact that variance of $\mathbf{X}$ is of order $O(T)$. The above asymptotic result is useful as long as the first term dominates the second term $O(T\xi(T))$, as is the case in our situation.

To be more precise, and get more terms, let us consider $d = 2$ and the following binomial sum

$$S_f(T) = \sum_{k=0}^{T} \binom{T}{k} \theta^k (1-\theta)^{T-k} f(k).$$

The Taylor expansion around the mean $\mathbf{E}[X] = T\theta$ of the binom(n, θ) up to three terms is

$$f(X) = f(\mathbf{E}[X]) + (X - \mathbf{E}[X])f'(\mathbf{E}[X] + \\ + \frac{1}{2}(X - \mathbf{E}X)^2 f''(\mathbf{E}[X]) + O\left((X - \mathbf{E}[X])^3 f'''(\xi)\right).$$

Taking expectation of the above with respect to $X \sim \text{binom}(T, \theta)$ we find

$$S_f(T) = f(T\theta) + \frac{1}{2}T\theta(1-\theta)f''(T\theta) + O(Tf'''(\xi).$$

In particular, if $f(x) = \log x$, then we obtain (55).

Thus to finish our derivation, we now need to prove (40) which we present below in a more general form.

Lemma 1. *Define for some finite positive integer m the following sum for any T*

$$D_{T,m} = \sum_{k_1+\cdots+k_m=T} \binom{T}{k_1, \ldots, k_m} \left(\frac{k_1}{T}\right)^{k_1} \cdots \left(\frac{k_m}{T}\right)^{k_m} \tag{45}$$

and then $d_{T,m} = \log D_{T,m}$ is

$$\begin{aligned} d_{T,m} =& \frac{m-1}{2} \log\left(\frac{T}{2}\right) + \log\left(\frac{\sqrt{\pi}}{\Gamma(\frac{m}{2})}\right) \\ &+ \frac{\Gamma(\frac{m}{2}) m \log e}{3\Gamma(\frac{m}{2} - \frac{1}{2})} \cdot \frac{\sqrt{2}}{\sqrt{T}} \\ &+ \left(\frac{3 + m(m-2)(2m+1)}{36} - \frac{\Gamma^2(\frac{m}{2}) m^2}{9\Gamma^2(\frac{m}{2} - \frac{1}{2})}\right) \cdot \frac{1}{T} \\ &+ O\left(\frac{1}{T^{3/2}}\right) \end{aligned}$$

for large T.

Proof: The asymptotic of the sequence of numbers $D_{T,m}$, (for m constant), is derived in [30] through the so-called *tree-like generating function*, defined as

$$D_m(z) = \sum_{T=0}^{\infty} \frac{T^T}{T!} D_{T,m} z^T.$$

Here, we will follow the same methodology, which we review next. The first step is to use (45) to define an appropriate recurrence on $D_{T,m}$, and to employ the convolution formula for generating functions (cf. [31]) to relate $D_m(z)$ to the tree-like generating function of the sequence $\langle 1, 1, \ldots \rangle$, namely

$$B(z) = \sum_{k=0}^{\infty} \frac{k^k}{k!} z^k.$$

This function, in turn, can be shown to satisfy

$$B(z) = \frac{1}{1 - \tau(z)} \tag{46}$$

for $|z| < e^{-1}$, where $\tau(z)$ is the well-known *tree function*, which is a solution to the implicit equation

$$\tau(z) = z e^{\tau(z)} \tag{47}$$

with $|\tau(z)| < 1$.[1] Now it is easy to see (see [30]) that the generating function $D_m(z)$ of $D_{T,m}$ satisfies, for $|z| < e^{-1}$,

$$D_m(z) = [B(z)]^m$$

and, consequently,

$$D_{T,m} = \frac{T!}{T^T} [z^T] [B(z)]^m \tag{48}$$

[1] In terms of the standard *Lambert-W* function, we have $\tau(z) = -W(-z)$.

where $[z^T]f(z)$ denotes the coefficient of z^T in $f(z)$.

Defining $\beta(z) = B(z/e)$, $|z| < 1$, noticing that $[z^T]\beta(z) = e^{-T}[z^T]B(z)$, and applying Stirling's formula, (48) yields

$$D_{T,m} = \sqrt{2\pi T}\left(1 + O(T^{-1})\right)[z^T]\,[\beta(z)]^m\,. \tag{49}$$

Thus, it suffices to extract asymptotics of the coefficient at z^T of $[\beta(z)]^m$, for which a standard tool is Cauchy's coefficient formula [9], [31], that is,

$$[z^T][\beta(z)]^m = \frac{1}{2\pi i}\oint \frac{\beta^m(z)}{z^{T+1}}dz \tag{50}$$

where the integration is around a closed path containing $z = 0$ inside which $\beta^m(z)$ is analytic.

Now, for constant m we can use the Flajolet and Odlyzko *singularity analysis* [9], [31], which applies because $[\beta(z)]^m$ has algebraic singularities. Using (46) and (47), the singular expansion of $\beta(z)$ around its singularity $z = 1$ takes the form [4]

$$\beta(z) = \frac{1}{\sqrt{2(1-z)}} + \frac{1}{3} - \frac{\sqrt{2}}{24}\sqrt{(1-z)} + O(1-z).$$

Finally, we shall use [9]

$$\begin{aligned}[z^T](1-z)^{-\alpha} &= \binom{T+\alpha-1}{T} = \frac{\Gamma(n+\alpha)}{\Gamma(\alpha)\Gamma(T+1)} \\ &= \frac{T^{\alpha-1}}{\Gamma(\alpha)}\left(1 + \frac{\alpha(\alpha-1)}{2T} + O\left(\frac{1}{T^2}\right)\right)\end{aligned}$$

provided $\alpha \notin \{0, -1, -2, \ldots\}$.

For large m we need to use saddle point method as explained in [32]. This completes the proof. ■

B. Proof of Theorem 3

Our starting point is (24). We shall deal now with the denominator that we repeat below

$$Q(y^T) = \int_{\boldsymbol{\theta}} \rho(\boldsymbol{\theta}) \prod_{j=1}^{d} \theta_j^{k_j}(1-\theta_j)^{T_j-k_j}. \tag{51}$$

We need to choose the prior. We pick up Jeffrey's prior that is known to be asymptotically optimal for the average minimax regret [7], [6], [33], [34]. Thus, for each dimension we set

$$\rho(\boldsymbol{\theta}) = \prod_{i=1}^{d}\rho(\theta_i), \qquad \rho(\theta_i) = \frac{1}{\pi\sqrt{\theta_i(1-\theta_i)}}. \tag{52}$$

To handle the denominator, we introduce now the Dirichlet density as

$$\mathrm{Dir}(x_1,\ldots,x_m;\alpha_1,\ldots\alpha_m) = \frac{1}{B(\alpha_1,\ldots,\alpha_m)}\prod_{i=1}^{m} x_i^{\alpha_i-1}, \tag{53}$$

where $\sum_{i=1}^m x_i = 1$ and

$$B(\alpha_1,\ldots,\alpha_m) = \frac{\Gamma(\alpha_1)\cdots\Gamma(\alpha_m)}{\Gamma(\alpha_1+\cdots\alpha_m)}$$

is the Euler beta function. In the sequel we only use it for $m = 2$.

We apply the Dirichlet density to re-write $Q(y^T)$ as

$$\begin{aligned}Q(y^T) &= \prod_{i=1}^{d}\frac{1}{\pi}\int_{\theta_i}\theta_i^{k_i-1/2}(1-\theta_i)^{T_i-k_i-1/2} \\ &= \prod_{i=1}^{d}\frac{B(k_i+1/2, T_i-k_i+1/2)}{\pi}\end{aligned}$$

$$\cdot\frac{1}{B(k_i+1/2, T_i-k_i+1/2)}\int_{\theta_i}\theta_i^{k_i-1/2}(1-\theta_i)^{T_i-k_i-1/2}$$

$$= \prod_{i=1}^{d}\frac{B(k_i+1/2, T_i-k_i+1/2)}{\pi},$$

where the last equality comes from the fact that the normalized integral in the third line above is just the Dirichlet distribution, and hence is equal to 1. This leads to the following expression for the pointwise regret

$$r(y^T|\mathbf{x}^T) = \log\prod_{j=1}^{d}\frac{\pi\left(\frac{k_j}{T_j}\right)^{k_j}\left(\frac{T_j-k_j}{T_j}\right)^{T_j-k_j}}{B(k_j+1/2, T_j-k_j+1/2)}. \tag{54}$$

But

$$B(k_j+\frac{1}{2}, T_j-k_j+\frac{1}{2}) = \frac{\Gamma(k_j+1/2)\Gamma(T_j-k_j+1/2)}{\Gamma(T_j+1)}.$$

Therefore, using Stirling's formula, we find for large T_j and typical k_j

$$B(k_j+\frac{1}{2}, T_j-k_j+\frac{1}{2}) \sim \sqrt{\frac{2\pi}{T_j}}\left(\frac{k_j}{T_j}\right)^{k_j}\left(\frac{T_j-k_j}{T_j}\right)^{T_j-k_j}.$$

This proves (25) of Theorem 3.

To prove (26) we need to apply the (multi)(bi)nomial sum asymptotics as discussed in Remark 1. In our case it can be reduced to the following binomial sums as in [3], [8], [16]

$$\sum_{i=0}^{T}\binom{T}{k}\theta^k(1-\theta)^{T-k}\log k = \log(T\theta) - \frac{1-\theta}{2\theta T} + O(T^{-2}). \tag{55}$$

Notice that this asymptotic expansion is *not* uniform with respect to θ. Using the above, we complete the proof of Theorem 3.

C. Proof of Theorem 4

We first evaluate the conditional regret $r_T^{\mathcal{M}}(\mathbf{x}^T)$ and the corresponding Shtarkov sum $d_T^{\mathcal{M}}(\mathbf{x}^T)$ for the the monotone case. We shall follow the ideas of [6] adopted to our case. To proceed we need to understand the optimization

$$\sup_{\theta_1 \le \cdots \le \theta_d} \prod_{i=1}^{d} \theta_i^{k_i}(1-\theta_i)^{T_i-k_i}.$$

Using Kuhn-Ticker criteria, it is easy to see that the optimal $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$ in this case is either $\theta_i = k_i/T_i$ or for some $1 \le i < j \le d$ we have $\theta_i = \theta_j$. Let us denote the corresponding Shtarkov sums as $d_T^{\mathcal{M}^<}(\mathbf{x}^T)$ and $d_T^{\Theta\setminus\mathcal{M}^<}(\mathbf{x}^T)$, respectively. Therefore,

$$d_T^{\mathcal{M}}(\mathbf{x}^T) = d_T^{\mathcal{M}^<}(\mathbf{x}^T) + d_T^{\Theta\setminus\mathcal{M}^<}(\mathbf{x}^T).$$

We shall prove below that the second term above grows slower than the leading term, that is, we show that

$$d_T^{\Theta\setminus\mathcal{M}^<} = O(T^{\frac{d-1}{2}}) \tag{56}$$

when we set $T_i = T/d$ for the worst case Shtarkov sum $d_T^{\Theta\setminus\mathcal{M}^<}$. With this and noting that unconstrained Shtarkov sum $d_T(\mathbf{x}^T)$ is a symmetric function, we shall conclude, using (27), that

$$d_T(\mathbf{x}^T) = d! \cdot d_T^{\mathcal{M}^<}(\mathbf{x}^T) + O(d(T^{\frac{d-1}{2}})$$

establishing (28) and Theorem 4.

To establish (56), we first find an integral representation of $d_T(\mathbf{x}^T)$ defined in (19). Observe that using Stirling approximation followed by Euler-Maclaurin summation formula we have for any $1 \le i \le d$ (see also [6])

$$\sum_{k_i=0}^{T_i} \binom{T_i}{k_i} \left(\frac{k_i}{T_i}\right)^{k_i} \left(\frac{T_i-k_i}{T_i}\right)^{T_i-k_i} =$$

$$= \sqrt{\frac{T_i\pi}{2}}(1+O(1/\sqrt{T_i}))\frac{1}{B(1/2,1/2)}\int_0^1 \frac{dx_i}{\sqrt{x_i(1-x_i)}}$$

$$= \sqrt{\frac{T_i\pi}{2}}(1+O(1/\sqrt{T_i}))$$

since the integral in the second line is equal to 1 due to (53). Thus

$$d_T(\mathbf{x}^T) = \prod_{i=1}^{d} \sqrt{\frac{T_i\pi}{2}}(1+O(1/\sqrt{T_i}))$$

$$\cdot \frac{1}{B^d(1/2/1/2)} \int_{x_1=0}^{1} \cdots \int_{x_d=0}^{1} \prod_{i=1}^{d} \frac{dx_i}{\sqrt{x_i(1-x_i)}} =$$

$$= \prod_{i=1}^{d} \sqrt{\frac{T_i\pi}{2}}(1+O(1/\sqrt{T_i})).$$

However, for monotone label probabilities the above Shtarkov sum will change to

$$d_T^{\mathcal{M}}(\mathbf{x}^T) = \prod_{i=1}^{d} \sqrt{\frac{T_i\pi}{2}}(1+O(1/\sqrt{T_i})) \tag{57}$$

$$\cdot \frac{1}{B^d(1/2/1/2)} \int_{x_1 \le x_2 \le \cdots x_d} \prod_{i=1}^{d} \frac{dx_i}{\sqrt{x_i(1-x_i)}}. \tag{58}$$

To complete establishing (56), we simply our presentation by assuming that $\theta_i = \theta_j$ for some $1 \le i < j \le d$. We also write $k_{ij} = k_i + k_j$ and $T_{ij} = T_i + T_j$. Observe that then

$$d_T^{\Theta\setminus\mathcal{M}}(\mathbf{x}^T) = \sum_{k_1=0}^{T_1} \binom{T_1}{k_1} \left(\frac{k_1}{T_1}\right)^{k_1} \left(\frac{T_1-k_1}{T_1}\right)^{T_1-k_1} \cdots$$

$$\sum_{k_{ij}=0}^{T_{ij}} \binom{T_{ij}}{k_{ij}} \left(\frac{k_{ij}}{T_{ij}}\right)^{k_{ij}} \left(\frac{T_{ij}-k_1}{T_{ij}}\right)^{T_{ij}-k_{ij}} \cdots$$

$$\cdots \sum_{k_d=0}^{T_d} \left(\frac{k_d}{T_d}\right)^{k_d} \left(\frac{T_d-k_d}{T_d}\right)^{T_d-k_d}.$$

In summary, the above sum is similar of the general Shtarkov sum $d_T(\mathbf{x}^T)$ but of dimension $d-1$. Hence by the above integral representation, we easily show that it grows like $O(\sqrt{T_1, \ldots T_{ij} \ldots T_d}) = O(T^{(d-1)/2})$, as needed.

V. Conclusion and Further Extensions

We proposed a logistic regression in which the label probability is determined by the degree of existence of features in an example, instead of by the actual features. This regret, while interesting by itself for applications such as graph models, and specifically social network models, can also be used to bridge well established results from the universal compression literature to the study of regret in machine learning. We demonstrated that for this novel problem we can precisely compute various variants of the regret, showing logarithmic regret for this problem, which linearly increases with the dimensionality. The precise regret for this problem by itself serves as a *general lower bound* for the regret of standard logistic regression. This connection opens up a large range of possibilities to apply established theory in the study of universal compression redundancy to studying regret for online learning problems (see also [15]). Finally, if a monotonicity constraint is imposed on the parameters of the degree based probability, an additional second order reduction term of the regret was demonstrated.

References

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

[2] Y. Choi and W. Szpankowski. Compression of graphical structures: Fundamental limits, algorithms, and experiments. *IEEE Trans. Information Theory*, 58:620–638, 2012.

[3] J. Cichon and Z. Golebiewski. On the bernoulli sums and bernstein polynomials. In *AofA*, pages 179–190, 2012.

[4] R. Corless, G. Gonnet, D. Hare, D. Jeffrey, and D. Knuth. On the lambert w function. *Adv. Computational Mathematics*, 5:329–359, 1996.

[5] L. D. Davisson. Universal noiseless coding. *IEEE Trans. Inf. Theory*, IT-19(6):783–795, Nov. 1973.

[6] M. Drmota, G. I. Shamir, and W. Szpankowski. Sequential universal compression for non-binary sequences with constrained distributions. In *ArXiv: http://arxiv.org/abs/2005.10320*, 2021.

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

[8] P. Flajolet. Singularity analysis and asymptotics of bernoulli sums. *Theoretical Computer Science*, 215:371–381, 1999.

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

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

[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] P. D. Grunwald. *The Minimum Description Length Principle*. MIT Press, 2007.

[13] E. Hazan. The convex optimization approach to regret minimization. In *S. Sra, S. Nowozin, and S. Wright, editors, Optimization for Machine Learning*, pages 287–303. MIT press, 2012.

[14] 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*, page 197–209. MIT press, 2014.

[15] P. Jacquet, G. I. Shamir, and W. Szpankowski. Precise minimax regret for logistic regression with categorical feature values. In *PRML: ALT'21*, 2021.

[16] P. Jacquet and W. Szpankowski. Entropy computations via analytic depoissonization. *IEEE Trans. on Information Theory*, 45:1072–1081, 1999.

[17] R. Jezequel, P. Gillard, and A. Rudi. Efficient improper learning for online logistic regression. In *PMLR, COLT'20*, volume 125, 2020.

[18] Sham M Kakade and Andrew Y. Ng. Online bounds for bayesian algorithms. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 641–648. MIT Press, 2005.

[19] Tomasz Łuczak, Abram Magner, and Wojciech Szpankowski. Asymmetry and structural information in preferential attachment graphs. *Random Structures &Algorithms*, 55(3):696–718, 2019.

[20] H. B. McMahan and M. J. Streeter. Open problem: Better bounds for online logistic regression. In *Journal of Machine Learning Research-Proceedings Track, 23*, 2012.

[21] A. Orlitsky and N. P. Santhanam. Speaking of infinity. *IEEE Trans. Inf. Theory*, 50(10):2215–2230, Oct. 2004.

[22] A. Rakhlin and K. Sridharan. Online nonparametric regression with general los function. In *COLT*, 2014.

[23] J. Rissanen. Fisher information and stochastic complexity. *IEEE Trans. Information Theory*, 42:40–47, 1996.

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

[25] G. I. Shamir. On the MDL principle for i.i.d. sources with large alphabets. *IEEE Trans. Inform. Theory*, 52(5):1939–1955, May 2006.

[26] G. I. Shamir. Universal source coding for monotonic and fast decaying monotonic distributions. *IEEE Trans. Inform. Theory*, 59(11):7194–7211, Nov. 2013.

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

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

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

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

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

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

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

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