Precise Minimax Regret for Logistic Regression

Philippe Jacquet
INRIA
France
Email: philippe.jacquet@inria.edu

Gil I. Shamir
Google
Pittsburgh, PA, USA
Email: gshamir@ieee.org

Wojciech Szpankowski
Dept. Computer Science
Purdue University
Email: szpan@purdue.edu

***Abstract*—We study online logistic regression with binary labels and general feature values in which a learner tries to predict an outcome/ label based on data/ features received in rounds. Our goal is to evaluate precisely the (maximal) *minimax regret* which we analyze using a unique and novel combination of information-theoretic and analytic combinatorics tools such as Fourier transform, saddle point method, and Mellin transform in the multi-dimensional settings. To be more precise, the pointwise regret of an online algorithm is defined as the (excess) loss it incurs over a constant comparator which is used for prediction. In the minimax scenario we seek the best learning distribution for the worst label sequence. For dimension $d = o(T^{1/3})$ we show that the maximal minimax regret grows as $d/2 \cdot \log(2T/\pi) + C_d + O(d^{3/2}/\sqrt{T})$ where T is the number of rounds of running a training algorithm and C_d is explicitly computable constant that depends on dimension d and feature values. We compute explicitly the constant C_d for features uniformly distributed on a d-dimensional sphere or ball.**

I. Introduction

In online learning sequentially received data must be used to update the predictor for subsequent data, that is, a model is trained to learn parameters from examples/ samples 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. The prediction algorithm incurs for each round some *loss* and updates its belief of the model parameters. In this paper we study a more specific setting of online *logistic regression* for binary classification. Logistic regression has received a lot of attention in machine learning ([3], [24]) due to several important applications from category classification to risk assessment.

More precisely, we phrase our learning problem in terms of a game between nature/ environment and a learner. At each round the learner obtains a d dimensional input/ feature vector $\mathbf{x}_t$ and makes prediction $\hat{y}_t$. Then the nature reveals the true output/ label y_t. Throughout we assume binary labels $y_t \in \{-1, 1\}$ (however, see Section III-B for extension to non-binary labels) and bounded features $\mathbf{x}_t$ living in a space of dimension d. Thus at round t the learner incurs some *loss* which we denote as $\ell(\hat{y}_t, y_t)$. For $t \in \{1, \ldots, T\}$ we write $y^T = (y_1, \ldots, y_T)$ and $\mathbf{x}^T = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$. Then the cumulative relative loss or better *pointwise regret* is defined as in [11], [9], [27]

$$R_T(\hat{y}^T, y^T|\mathbf{x}^T) = \sum_{t=1}^{T} \ell(\hat{y}_t, y_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^{T} \ell(f(\mathbf{x}_t), y_t)$$

where $\mathcal{F}$ is a reference class of functions called also forecasters or experts. In this paper, we consider the (maximal) *minimax regret* defined as

$$r_T^*(\mathbf{x}^T) = \inf_{\hat{y}^T} \max_{y^T} R_T(\hat{y}^T, y^T|\mathbf{x}^T) \tag{1}$$

which constitutes a fundamental/ universal lower bound over a class of learning algorithms. This regret could be viewed as "fixed-design" (or conditional) minimax regret with $\mathbf{x}^T$ known in advance. Another minimax regret called here the adversarial minimax regret with known $\mathbf{x}^T$ can be defined as

$$\begin{aligned} r_T^a(\mathbf{x}^T) &\triangleq \langle\!\langle \min_{\hat{y}_t} \max_{y_t} \rangle\!\rangle_{t=1}^{T} R(\mathbf{x}^T, y^T, \hat{y}^T) \\ &\triangleq \min_{\hat{y}_1} \max_{y_1} \min_{\hat{y}_2} \max_{y_2} \cdots \min_{\hat{y}_T} \max_{y_T} R(\hat{y}^T, y^T|\mathbf{x}^T). \end{aligned} \tag{2}$$

It was recently proved that $r_T^*(\mathbf{x}^T) = r_T^a(\mathbf{x}^T)$ [32]. Hereafter, we focus in $r_T(\mathbf{x}^T)$ and give precise asymptotic expansion of this regret.

In this paper we consider a more specific model, namely *logistic regression* with logarithmic loss function and linear reference class. More specifically, we restrict the reference class to $\mathcal{F} = \{\mathbf{x}_t \to f(\mathbf{x}_t) = \text{Sigmoid}(\langle \mathbf{x}_t, \mathbf{w}\rangle)\}$ where $\text{Sigmoid}(\langle \mathbf{x}_t, \mathbf{w}\rangle) = (1 + \exp(-\langle \mathbf{x}_t, \mathbf{w}\rangle))^{-1}$ is the logistic function and $\langle \mathbf{x}_t, \mathbf{w}\rangle = \sum_{i=1}^{d} x_{i,t} w_i$ is the scalar product for some weight vector $\mathbf{w} = (w_1, \ldots, w_d)$. We also define $\ell(y_t|\mathbf{x}_t, \mathbf{w}) := \ell(f(\mathbf{x}_t), y_t) := \log(1 + \exp(-y_t \langle \mathbf{x}_t, \mathbf{w}\rangle))$ as the loss function for the logistic regression.

In the minimax scenario studied in this paper, we postulate that the prediction is based on a learning distribution $Q(y_t|\mathbf{x}_t) = Q(y_t|\hat{y}_t)$ which best approximates the expert prediction $P(y_t|\mathbf{x}_t, \mathbf{w}_t)$. For example, the most popular class of learning algorithms are Bayesian (cf. [9], [16], [27]), but we do *not* make such assumption here. For such a setting the pointwise regret for a *given* learning distribution Q is then defined as

$$R_T(Q, y^T|\mathbf{x}^T) = -\sum_{t=1}^{T} \log Q(y_t|\mathbf{x}_t) - \inf_{\mathbf{w}} \sum_{t=1}^{T} \ell(\langle \mathbf{x}_t, \mathbf{w}\rangle, y_t). \tag{3}$$

The (maximal) minimax regret studied here is defined as

$$r_T^*(\mathbf{x}^T) = \inf_{Q} \max_{y^T} R_T(Q, y^T|\mathbf{x}^T). \tag{4}$$

Observe that $r_T^*(\mathbf{x}^T) \le \max_{y^T} R_T(Q, y^T|\mathbf{x}^T)$ for any learning algorithm and all label sequences. In this paper we provide

a precise asymptotic expansion of the maximal minimax regret, a result that had been wanting for some time.

a) Our Contributions and Methods.: Our contribution is two-fold. First, we present precise asymptotic expansions for the maximal minimax regret (4) through the so called Shtarkov sum (cf. [28], [6]). Second, we apply new methodology using tools of analytic combinatorics such as complex asymptotics and Fourier as well as Mellin transforms (cf [7], [30]) to handle Shtarkov sum for the logistic regression.

More precisely, we first represent the minimax regret (4) as the logarithm of the Shtarkov sum over all label sequences of the optimal label probability which turns out to be the maximum-likelihood distribution as presented in (12). Such a sum arose already in the universal compression as witnessed by [28], [6], [31]. In Theorem 1 we show that for $d = o(T^{1/3})$ the minimax regret grows as

$$\frac{d}{2}\log(2T/\pi) + C_d(\mathbf{x}^T) + O(d^{3/2}/\sqrt{T})$$

where the constant $C_d(\mathbf{x}^T)$ depends on the dimension d and data $\mathbf{x}^T$. We explicitly express this constant as the logarithm of a multi-dimensional integral over the determinant of a matrix that depends on data and the logistic function (cf. (18)). We generalize it to non-binary labels in Theorem 2. Then we find in Theorem 3 an asymptotic expression for $C_d(\mathbf{x}^T)$ when data $\mathbf{x}^T$ are distributed uniformly on a d-dimensional sphere $\mathcal{S}_d$ and a ball $\mathcal{B}_d$. This allows us to show in Theorem 3 that for large d the minimax regret grows with high probability as $\frac{d}{2}\log\frac{T}{d} - \frac{d}{2}\log\sqrt{8\pi} + O(1)$.

Our second technical contribution lies in unique and novel methodology based on analytic combinatorics. As mention above, we represent the maximal minimax regret as a Shtarkov sum. Only recently [25] and [13] introduced Shtarkov sum in the context of logistic regression. To analyze asymptotically the Shtarkov sum, we translate the Shtarkov sum into a d-dimensional integral that we evaluate using a multi-dimensional Laplace/ saddle point method. Further embellishments, including Mellin transform, discrete geometry and spectral representation of matrices, are required to study the constant $C_d(\mathbf{x}^T)$ when the feature $\mathbf{x}^T$ lie on a d-dimensional sphere.

b) Related Work.: In this paper we combine methodology of analytic combinatorics (see, e.g., [7], [15], [30]) and information theory (see, e.g., [1], [6], [17], [19], [22], [23], [26], [33]) to study a machine learning problem (see, e.g., [3], [24]), namely, the regret of logistic regression.

The set up of the logistic regression is similar to the redundancy of universal coding studied extensively in information theory. It corresponds to a single dimensional (i.e., $d = 1$) regret problem for logistic regression. In this case, with m being the alphabet size or the number of labels, it is known that for a large class of sources (up to Markovian but not for non-Markovian as discussed in [4], [8]) the redundancy grows as $\frac{m-1}{2}\log T$ when the alphabet size m is fixed (see [6], [23], [26], [29], [33], [34]) and $\frac{m-1}{2}\log(T/m)$ for $m = o(T)$ (see also [19], [26], [31]). In fact in [31] full asymptotic expansions were derived for all ranges of m.

In the machine learning literature a general online optimization is studied, and generally *pointwise regret* is analyzed with logarithmic regret in the strongly and weakly convex setting. We note that logistic regression seems to fall under weakly convex setting. We first mention work of [11] who studied the pointwise regret of the logistic regression for the *proper* setting, that is, when at time t the decision regarding $\mathbf{w}_t$ is based on knowledge available to the learner up to time $t-1$. Unlike the *improper* learning, studied in this paper, where feature $\mathbf{x}_t$ at time is also available to the learner and [11] showed that the pointwise regret is $\Theta(T^{1/3})$ for $d = 1$ and $O(\sqrt{T})$ for $d > 1$. Furthermore, the worst case adversarial minimax regret for a wide variety of loss function and references classes is discussed in a series of papers by [20], [21]. However, it should be pointed out that the authors of [1], [35] studied a general classes of densities smoothly parametrized by a d-dimensional data to obtain general results for the (average) minimax regret that can be phrased as an online regret.

For improper learning a more precise results are known. To the best of our knowledge, [16] were first to demonstrate results that suggest that *pointwise regret* for logistic regression grows like $O(d\log T/d)$ where for fixed dimension d and $m = 2$, which was further generalized in [9] to all m. The authors of [16] 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 [18]. The authors of [18] then posed the question of what happens for larger dimensions. Subsequently, [9] 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. Recently, [13] analyze a precise maximal minimax regret but only for *finite* number of feature values and fixed dimension d. More precisely, [13] addresses a relaxed problem with at most a finite number $N = o(\sqrt{T})$ of distinct feature vector values and regret is analyzed only for a fixed dimension $d = O(1)$. In this paper we consider: (i) no relaxation, where feature vectors can take any values (up to $N = T$ distinct ones); (ii) the dimension d can grow with T as $d = o(T^{1/3})$; and (iii) a different methodology based on a multidimensional Laplace is used while the analysis of [13] is based on multidimensional Gaussian approximation, which fails in our setting. To the best of our knowledge here we present the first precise results for the minimax regret.

II. Problem Formulation and Notation

We denote by $\mathbf{x}_t = (x_{1,t}, \ldots, x_{d,t})$ a d-dimensional feature vector such that $||x|| \leq 1$ for some norm $||\cdot||$. We also assume that the set $\mathbf{x}_t$ spans $\mathbf{R}^d$. The label binary vector is denoted as $y^T = (y_1, \ldots, y_T)$ with $y_t \in \{-1, 1\}$ (however, we also present in Section III-B some results for non-binary labels). Finally, $\mathbf{w}_t = (w_{1,t}, \ldots, w_{d,t})$ is a d-dimensional vector of weights. In this paper, we do not address the method used to learn the weights (e.g., gradient method or Bayesian mixing), however, see the end of this section for some remarks on a sequential algorithm.

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

$$L(y^T|\mathbf{x}^T,\mathbf{w}^T) := \sum_{t=1}^{T} \log\left[1+\exp(-y_t\langle \mathbf{x}_t,\mathbf{w}_t\rangle)\right] \quad (5)$$

where $\langle \mathbf{x}_t,\mathbf{w}_t\rangle = \sum_{i=1}^{d} x_{i,t}w_{i,t}$ is the scalar product of $\mathbf{x}_t$ and $\mathbf{w}_t$. To simplify we also 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 $\langle \mathbf{x}_t,\mathbf{w}_t\rangle$. As mention in the introduction, it is convenient to interpret the logistic function in probabilistic terms. The probability of a label is then given by

$$P(y_t|\mathbf{x}_t,\mathbf{w}_t) = \left(1+\exp(-y_t\langle \mathbf{x}_t,\mathbf{w}_t\rangle)\right)^{-1} \quad (6)$$

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

Finally, we observe that the goal is to find the best learning distribution $Q(y_t|\mathbf{x}_t)$ of the unknown distribution $P(y_t|\mathbf{x}_t,\mathbf{w}_t)$. We notice that Q can be used to design an algorithm that predicts y_t. The *pointwise regret* for a given algorithm/ distribution Q is defined in (3) for individual sequences $(y_t,\mathbf{x}_t)$ following [16], [10], [9], [27]. Thus

$$R_T(Q,y^T|\mathbf{x}^T) = \log\frac{\sup_{\mathbf{w}} P(y^T|\mathbf{x}^T,\mathbf{w})}{Q(y^T|\mathbf{x}^T)} \quad (7)$$

where

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

The pointwise regret $R_T(Q,y^T|\mathbf{x}^T)$ is a function of label sequence y^T, data/ feature vector $\mathbf{x}^T$, and algorithm/ label distribution Q. 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 minimize over a class of learning algorithms/ distributions Q. Following information-theoretic view, as in [5], [6], [34], we define the (maximal) minimax regret (conditioned on $\mathbf{x}^T$) as in (4). Notice that this definition is over all possible learning algorithms represented by Q. Therefore, it constitutes a (universal) lower bound of the pointwise regret for all label sequences and for all learning distributions Q, including the Bayesian ones studied in [16], [9], [27].

We study in this paper the precise growth of the minimax regret for large T and wide range of d. However, to accomplish it we need a more succinct and computationally manageable representation of the maximal minimax regret. Following [28], [6], [3] we add and subtract from (4) the logarithm of the Shtarkov sum defined as

$$S_T(\mathbf{x}^T) \triangleq \sum_{y^T}\sup_{\mathbf{w}} P(y^T|\mathbf{x}^T,\mathbf{w}) \quad (9)$$

resulting in

$$\begin{aligned} r_T^*(\mathbf{x}^T) &= \min_Q \sup_{\mathbf{w}} \max_{y^T}(-\log Q(y^T|\mathbf{x}^T) + \log P^*(y^T|\mathbf{x}^T)) \\ &\quad + \log\sum_{y^T}\sup_{\mathbf{w}} P(y^T|\mathbf{x}^T,\mathbf{w}) \quad (10)\\ &= \log\sum_{y^T}\sup_{\mathbf{w}} P(y^T|\mathbf{x}^T,\mathbf{w}) = \log S_T(\mathbf{x}^T) \quad (11)\end{aligned}$$

where we set $Q(y^T,\mathbf{x}^T) = P^*(y^T|\mathbf{x}^T)$ with

$$P^*(y^T|\mathbf{x}^T) := \frac{\sup_{\mathbf{w}} P(y^T|\mathbf{x}^T,\mathbf{w})}{\sum_{v^T}\sup_{\mathbf{w}} P(v^T|\mathbf{x}^T,\mathbf{w})} \quad (12)$$

being the *maximum-likelihood distribution*. Indeed, since Q and P^* are distributions, there is at least one y^T such that the first term in (11) is nonnegative, so that $Q=P^*$ minimizes it (see also [3]). We also observe that the Shtarkov sum is invariant under scaling up data $\mathbf{x}^T$ (and scaling down weights).

III. Main Results

In this section we estimate asymptotically the Shtarkov sum (9) for large T and wide range of d Throughout we write $p(w) := (1+e^{-w})^{-1}$ and $q(w) = 1-p(w) = p(-w)$.

A. Minimax Regret for General Case

We start with a general expression for the probability $P(y^T|\mathbf{x}^T,\mathbf{w})$ as given in (8). Noting that

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

we find

$$\begin{aligned} P(y^T|\mathbf{x}^T,\mathbf{w}) &= \prod_{t=1}^{T}\frac{\exp\left(\frac{1+y_t}{2}\langle \mathbf{x}_t,\mathbf{w}\rangle\right)}{1+\exp(\langle \mathbf{x}_t,\mathbf{w}\rangle)} \\ &= \exp\left(-\sum_{t=1}^{T}\log(1+e^{\langle \mathbf{x}_t,\mathbf{w}\rangle}) + \sum_{t=1}^{T}\frac{1+y_t}{2}\langle \mathbf{x}_t,\mathbf{w}\rangle\right).\end{aligned}$$

Let now

$$L_T(\mathbf{w}) = L_T(\mathbf{w},\mathbf{x}^T) = \sum_{t=1}^{T}\log(1+e^{\langle \mathbf{w},\mathbf{x}_t\rangle}),$$

$$\mathbf{A}_T(y^T,\mathbf{x}^T) := \mathbf{A}(y^T) = \frac{1}{2}\sum_{t=1}^{T}(1+y_t)\mathbf{x}_t.$$

Then $P(y^T|\mathbf{x}^T,\mathbf{w})$ becomes

$$P(y^T|\mathbf{x}^T,\mathbf{w}) = \exp\left(-L_T(\mathbf{w},\mathbf{x}^T) + \langle \mathbf{w},\mathbf{A}_T(y^T)\rangle\right). \quad (13)$$

Now we sketch the road map of our approach, leaving technical details to the next section and full paper [14]. The optimal value $\mathbf{w}^*$ that maximizes $P(y^T|\mathbf{x}^T,\mathbf{w})$ satisfies

$$\nabla_{\mathbf{w}} L_T(\mathbf{w}^*) = \mathbf{A}_T(y^T) \quad (14)$$

where $\nabla L_T(\mathbf{w}^*)$ is the gradient vector of $L_T(\mathbf{w})$. It is easy to see that

$$\mathbf{G}_T(\mathbf{w}) := \nabla_{\mathbf{w}} L_T(\mathbf{w}) = \sum_{t=1}^{T} p(\langle \mathbf{w},\mathbf{x}_t\rangle)\mathbf{x}_t \quad (15)$$

due to the crucial property $p'(w) = p(w)q(w)$. In the next section (see discussion below (26)) we prove that $\mathbf{w}^*$ exists and is unique in $[-\infty, \infty]^d$.

In view of the above, the optimal probability $P^*(y^T|\mathbf{x}^T, \mathbf{w})$ is then

$$P^*(\mathbf{w}^*) = P^*(y^T|\mathbf{x}^T) = \exp(-L_T(\mathbf{w}^*) + \langle \mathbf{w}^*, \mathbf{G}_T(\mathbf{w}^*)\rangle). \tag{16}$$

In the next section, we apply Laplace/Fourier transform to represent the Shtarkov sum $S_T(\mathbf{x}^T)$ as a multidimensional integral that we evaluate using the multidimensional Laplace method. This will allow us to conclude that

$$S_T(\mathbf{x}^T) = \int_{\mathbf{R}^d} \sqrt{\det(\nabla \mathbf{G}(\mathbf{w}^*)/(2\pi)}d\mathbf{w}^* \cdot \left(1 + O\left(\frac{d^{3/2}}{\sqrt{T}}\right)\right)$$

where

$$\nabla \mathbf{G}(\mathbf{w}) = \sum_{t=1}^{T} p(\langle \mathbf{w}, \mathbf{x}_t\rangle)q(\langle \mathbf{w}, \mathbf{x}_t\rangle)\mathbf{x}_t \otimes \mathbf{x}_t.$$

In summary, our first main result proved in the next section can be formulated as follows.

Theorem 1: Let $\mathbf{x}_t \in [-1, 1]^d$ and span $\mathbf{R}^d$ as well as $p(w) = (1 + e^{-w})^{-1}$ with $q(w) = 1 - p(w)$. Then asymptotically for $d = o(T^{1/3})$

$$r^*(\mathbf{x}^T) = \frac{d}{2}\log T - \frac{d}{2}\log 2\pi + C_d(\mathbf{x}^T) + O(d^{3/2}/\sqrt{T}) \tag{17}$$

where the "discrepancy" $C_d(\mathbf{x}^T)$ is

$$C_d(\mathbf{x}^T) = \log\left(\int_{\mathbf{R}^d} \sqrt{\det(\mathbf{B}_d(\mathbf{w}, \mathbf{x}^T))}dw_1 \cdots dw_d\right) \tag{18}$$

with

$$\mathbf{B}(\mathbf{w}, \mathbf{x}^T) = \frac{1}{T}\sum_{t=1}^{T} p(\langle \mathbf{x}_t, \mathbf{w}\rangle)q(\langle \mathbf{x}_t, \mathbf{w}\rangle)\mathbf{x}_t \otimes \mathbf{x}_t \tag{19}$$

and $\mathbf{x}_t \otimes \mathbf{x}_t = \mathbf{x}_t\mathbf{x}_t^\tau$ being the tensor product of $\mathbf{x}_t$ with τ denoting the transpose.

In passing we should observe that if data $\mathbf{X}_t$ is generated by a stationary ergodic source, then by the *ergodic theorem* we conclude that with high probability (whp)

$$\mathbf{B}(\mathbf{w}, \mathbf{X}^T) \to \mathbf{E}_X[\mathbf{B}(\mathbf{w}, \mathbf{X})] \stackrel{\triangle}{=} \bar{\mathbf{B}}(\mathbf{w}) \tag{20}$$

when $T \to \infty$. Therefore the discrepancy $C_d(\mathbf{x}^T)$ satisfies

$$C_d(\mathbf{X}^T) \to \log\left(\int_{\mathbf{R}^d} \sqrt{\det(\bar{\mathbf{B}}(\mathbf{w})}d\mathbf{w}\right)$$

in probability (see [2]).

B. Extension to Non-binary Labels

Let us now consider a non-binary label alphabet of size m. We will follow [9] and define a matrix $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_{m-1}]$ where each $\mathbf{w}_k$ is a vector of dimension d.The multinomial logistic function known also as *softmax function* is then defined as

$$p_\ell(\mathbf{x}^\tau\mathbf{W}) = \frac{e^{\langle \mathbf{x}, \mathbf{w}_\ell\rangle}}{1 + \sum_{k=1}^{m-1} e^{\langle \mathbf{x}, \mathbf{w}_k\rangle}} \tag{21}$$

and $q(\mathbf{x}^\tau\mathbf{W}) = 1 - p_\ell(\mathbf{x}^\tau\mathbf{W})$ for $\ell = 1, \ldots, m-1$. Alternatively, we can express $\langle \mathbf{x}, \mathbf{w}_\ell\rangle$ as $\langle \mathbf{x}, \mathbf{W}\mathbf{e}_\ell\rangle$ where $\mathbf{e}_\ell$ is a column vector of dimension $m-1$ with all coefficient equal to zero except the ℓth coefficient which is set at 1. We also have $\langle \mathbf{x}, \mathbf{w}_\ell\rangle = \mathrm{tr}(\mathbf{W}\mathbf{e}_\ell \otimes \mathbf{x}^\tau)$. Finally, we write $\mathbf{p} = (p_1, \ldots, p_m)$.

Following the footsteps of our derivations for the binary labels we present our second main result

Theorem 2: Let $\mathbf{x}_t \in [-1, 1]^d$ and span $\mathbf{R}^d$. The label alphabet is of size m, and for $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_{m-1}]$ we define $p_\ell(\mathbf{x}^\tau\mathbf{W})$ for $\ell = 1, \ldots, m-1$ as in (21). Then asymptotically for $md = o(T^{1/3})$

$$r^*(\mathbf{x}^T) = \frac{d(m-1)}{2}\log\frac{T}{2\pi} + O((md)^{3/2}/\sqrt{T}) + \log\left(\int_{\mathbf{R}^{dm}} + \sqrt{\det(\mathbf{B}(\mathbf{W}))}d\mathbf{w}_1 \cdots d\mathbf{w}_{m-1}\right)$$

where $\mathbf{B}(\mathbf{W})$ is equal to

$$\frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t \otimes [\mathrm{Diag}\,(\mathbf{p}(\mathbf{x}_i^\tau\mathbf{W})) - \mathbf{p}(\mathbf{x}_i^\tau\mathbf{W}) \otimes \mathbf{p}(\mathbf{x}_i^\tau\mathbf{W})] \otimes \mathbf{x}_t.$$

is a $d(m-1) \times (m-1)d$ matrix.

C. Spherical Features

Now we assume that the feature $\mathbf{x}_t$ are either uniformly distributed on a d-dimensional sphere $\mathcal{S}_d$ or inside a d-dimensional ball $\mathcal{B}_d$ for large d. By (20) we know that with high probability (whp)

$$\mathbf{B}(\mathbf{w}, \mathcal{S}_d) \to \bar{\mathbf{B}}(\mathbf{w}) = \frac{1}{s_d}\int_{\mathcal{S}_d} p(\langle \mathbf{x}\mathbf{w}\rangle)q(\langle \mathbf{x}\mathbf{w}\rangle)\mathbf{x} \otimes \mathbf{x}\, d\mathbf{x}$$

where s_d is the area of the hypersphere of dimension d and radius 1, that is, $s_d = 2\pi^{(d+1)/2}/\Gamma(\frac{d+1}{2})$.

We now write $C_d(\mathcal{S}_d)$ and $C_d(\mathcal{B}_d)$ for $C_d(\mathbf{x}^T)$ when $\mathbf{x}^T$ lies on the sphere $\mathcal{S}_d$ and on the ball $\mathcal{B}_d$, respectively.

Theorem 3: Under assumptions of Theorem 1 let the feature vector lies on the d-dimensional sphere $\mathcal{S}_d$ or ball $\mathcal{B}_d$. Then with high probability

$$C_d(\mathcal{S}_d) = -\frac{d}{2}\log\frac{d}{4} + \frac{d}{4}\log(\pi/8) + \frac{3}{8}\log e + O(1/d) \tag{22}$$

and

$$C_d(\mathcal{B}_d) = -\frac{d}{2}\log\frac{d}{4} + \frac{d}{4}\log(\pi/8) - \frac{1}{8}\log e + O(1/d), \tag{23}$$

respectively. Furthermore, the minimax regrets are:

$$r_T^*(\mathcal{S}_d) = \frac{d}{2}\log\frac{T}{d} - \frac{d}{2}\log\sqrt{8\pi} + \frac{3}{8}\log e + O(d^{3/2}/\sqrt{T})$$

and for the features inside a d dimensional ball $\mathcal{B}_d$ we find

$$r_T^*(\mathcal{B}_d) = \frac{d}{2}\log\frac{T}{d} - \frac{d}{2}\log\sqrt{8\pi} - \frac{1}{8}\log e + O(d^{3/2}/\sqrt{T}).$$

IV. Proof of Theorem 1

Let $\mathcal{A}_T(\mathbf{x}^T)$ be the set of achievable partial sums of the vectors $\mathbf{x}_t$, i.e.,

$$\mathcal{A}_T := \mathcal{A}_T(\mathbf{x}^T) = \{\mathbf{a} \in \mathbf{R}^d : \exists y^T \in \{-1,1\}^T : \mathbf{A}_T(y^T) = \mathbf{a}\}$$

and let $N(\mathbf{A})$ be the number of y^T tuples such that $\mathbf{A}_T(y^T) = \mathbf{A}$. The enumeration Laplace-like function of $e^{\langle \mathbf{w}, \mathbf{A}\rangle}$ then satisfies

$$F_T(\mathbf{w}) = \sum_{y^T} e^{\langle \mathbf{w}, \mathbf{A}\rangle} = \prod_t (1 + e^{\langle \mathbf{w}, \mathbf{x}_t\rangle}) = \exp(L_T(\mathbf{w})) \tag{24}$$

which can also be written as

$$F_T(\mathbf{w}) = \int e^{\langle \mathbf{w}, \mathbf{A}\rangle} d\rho_T(\mathbf{A}) \quad \text{with} \quad \rho_T = \sum_{\mathbf{a} \in \mathcal{A}_T} N(\mathbf{a}) \delta_{\mathbf{a}}$$

where $\delta_{\mathbf{a}}$ is the Dirac function on vector $\mathbf{a}$ and ρ_T is viewed as a measure. Using (16) and above we can re-write the Shtarkov sum as

$$S_T(\mathbf{x}^T) = \sum_{\mathbf{A} \in \mathcal{A}_T(\mathbf{x}^T)} N_T(\mathbf{A}) \exp(-L(\mathbf{w}^*(\mathbf{A})) + \langle \mathbf{w}^*(\mathbf{A}), \mathbf{A}\rangle) \tag{25}$$

that we evaluate asymptotically for large T.

We now express (25) as

$$S_T(\mathbf{x}^T) = \int K(\mathbf{A}) d\rho_T(\mathbf{A}), \tag{26}$$

where $K(\mathbf{A}) = \exp(-L(\mathbf{w}^*(\mathbf{A})) + \langle \mathbf{w}^*(\mathbf{A}), \mathbf{A}\rangle)$. Here $\mathbf{w}^*(\mathbf{A})$ is the inverse of function $\mathbf{G}_T(\mathbf{w})$. This inverse exists because the function $L(\mathbf{w})$ is in C^∞ and it is convex since its second derivative $\nabla^2 L(\mathbf{w}) = \sum_t p(\langle \mathbf{x}_t p(\mathbf{w})\rangle) q(\langle \mathbf{x}_t p(\mathbf{w})\rangle) \mathbf{x}_t \otimes \mathbf{x}_t$ is strictly positive as soon as $\mathbf{R}^d$ is spanned by $\mathbf{x}_t$. Therefore, it has a unique minimum although it can have some infinite coordinates (which will make the optimal probability $P^* = 1$ but not changing in any substantial way the asymptotics of the Shtarkov sum). In some cases the minimum can be at infinity. Indeed, consider the case when the pairs $(\mathbf{x}_t, y_t)$ form a separable set, *i.e* for binary labels there is an hyperplane that separates the +1 labelled features from the -1 labelled features. In this case let $\mathbf{w}_0$ be such that $\langle \mathbf{w}_0, \mathbf{x}_t\rangle > 0$ for all $y_t = 1$ and $\langle \mathbf{w}_0, \mathbf{x}_t\rangle < 0$ otherwise. The minimum is reached at $\mathbf{w}_0 = +\infty$ since $\lim \nabla L(\mathbf{w}) = \mathbf{A}$ when $\mathbf{w} = x\mathbf{w}_0$ with $x \to +\infty$. There can be hybrid cases where the minimum is attained over $\pm\infty$ for some coordinates of $\mathbf{w}$ while the other components remain finite. The example is when the pair label features are separable on some components but are not on the other components. But these are very "trivial" cases. We conclude that $K(\mathbf{A})$ is in C^∞ and has a finite support contained in $[-T, T]^d$.

Let now

$$K^*(\mathbf{w}) = \int_{\mathbf{R}^d} K(\mathbf{A}) e^{i\langle \mathbf{w}, \mathbf{A}\rangle} d\mathbf{A}$$

be the Fourier transform of function $K(\mathbf{A})$. Parseval theorem for multidimensional Fourier transform tells us that

$$S_T(\mathbf{x}^T) = \frac{1}{(2\pi)^d} \int_{\mathbf{c} + \mathbf{R}^d} F_T(i\mathbf{w}) K^*(-\mathbf{w}) d\mathbf{w} \tag{27}$$

where $\mathbf{c}$ is any constant in the domain of the Fourier existence. Notice that

$$K^*(\mathbf{w}) = \int_{\mathbf{R}^d} \exp(-L_T(\mathbf{w}^*(\mathbf{A})) + \langle (\mathbf{w}^*(\mathbf{A}) + i\mathbf{w}), \mathbf{A}\rangle) d\mathbf{A}.$$

By change of variable $\mathbf{A} = \mathbf{G}(\mathbf{w}^*)$ we find that $K^*(\mathbf{w})$ is equal to

$$\int_{\mathbf{R}^d} \exp(-L_T(\mathbf{w}^*) + \langle (\mathbf{w}^* + i\mathbf{w}), \mathbf{G}(\mathbf{w}^*)\rangle) \det(\nabla \mathbf{G}(\mathbf{w}^*)) d\mathbf{w}^* \tag{28}$$

leading to $(2\pi)^d \cdot S_T(\mathbf{x}^T)$ being equal to

$$\int_{\mathbf{R}^d} \exp(-L_T(\mathbf{w}^*) + \langle \mathbf{w}^*, \mathbf{G}_T(\mathbf{w}^*)\rangle) \det(\nabla \mathbf{G}_T(\mathbf{w}^*)) d\mathbf{w}^* \cdot \int_{-i\mathbf{c} + \mathbf{R}^d} \exp(L_T(i\mathbf{w}) - i\langle \mathbf{w}, \mathbf{G}_T(\mathbf{w}^*)\rangle) d\mathbf{w}.$$

We now take the advantage of the fact that the functions under the integrals are analytic functions so that we can move the path of integration of the second integral from $-i\mathbf{c} + \mathbf{R}^d$ to $-i\mathbf{w}^* + \mathbf{R}^d$, finding that $(2\pi)^d \cdot S_T(\mathbf{x}^T)$ is equal to

$$\int_{\mathbf{R}^d} \exp(-L_T(\mathbf{w}^*) + \langle \mathbf{w}^*, \mathbf{G}_T(\mathbf{w}^*)\rangle) \det(\nabla \mathbf{G}_T(\mathbf{w}^*)) d\mathbf{w}^* \cdot \int_{-i\mathbf{w}^* + \mathbf{R}^d} \exp(L_T(i\mathbf{w}) - i\langle \mathbf{w}, \mathbf{G}_T(\mathbf{w}^*)\rangle) d\mathbf{w}. \tag{29}$$

Finally, we notice that on the segment $\Im(\mathbf{w}) = i\mathbf{w}^*$ the quantity $|L_T(i\mathbf{w}) - i\langle \mathbf{w}, \mathbf{G}_T(\mathbf{w}^*)\rangle|$ attains its maximum at $\mathbf{w} = -i\mathbf{w}^*$, since

$$\nabla(L_T(i\mathbf{w}) - i\langle \mathbf{w}, \mathbf{G}_T(\mathbf{w}^*)\rangle) = i\mathbf{G}_T(i\mathbf{w}) - i\mathbf{G}_T(\mathbf{w}^*)$$

is zero when $i\mathbf{w} = \mathbf{w}^*$. Hence, for $\mathbf{x} \to 0$ we conclude

$$\begin{aligned} &L_T(\mathbf{w}^* + i\mathbf{x}) - \langle (\mathbf{w} + i\mathbf{x}), \mathbf{G}_T(\mathbf{w}^*)\rangle \\ &= L_T(\mathbf{w}^*) - \langle (\mathbf{w}, \mathbf{G}_T(\mathbf{w}^*)\rangle \\ &- \frac{1}{2}\langle \mathbf{x}, \nabla^2, L_T(\mathbf{w}^*)\mathbf{x}\rangle + O(L_T^{(3)}(\mathbf{w}^*)\|\mathbf{x}\|^3) \end{aligned}$$

where $L_T^{(3)}(\mathbf{w}^*)$ is the third derivative of $L_T(\mathbf{w})$ at $\mathbf{w}^*$. But $\nabla^2 L_T(\mathbf{w})$ and $L_T^{(3)}(\mathbf{w}^*)$ are of order $O(T)$, hence we can apply the multidimensional Laplace method [12] to find

$$\int_{\Im(\mathbf{w}) = i\mathbf{w}^*} \exp\left(L_T(\mathbf{w}) - \langle \mathbf{w}, \mathbf{G}_T(\mathbf{w}^*)\rangle\right) d\mathbf{w} = \tag{30}$$

$$= \frac{\exp(L_T(\mathbf{w}^*) - \langle \mathbf{w}^*, \mathbf{G}_T(\mathbf{w}^*)\rangle)}{\sqrt{\det(\nabla^2 L_T(\mathbf{w}^*)/(2\pi))}} \times \left(1 + O\left(\frac{d^{3/2}}{\sqrt{T}}\right)\right). \tag{31}$$

Since $\exp(L_T(\mathbf{w}^*) - \langle \mathbf{w}^*, \mathbf{G}_T(\mathbf{w}^*)\rangle)$ cancels factor $\exp(-L_T(\mathbf{w}^*) + \langle \mathbf{w}^*, \mathbf{G}_T(\mathbf{w}^*)\rangle)$ in (29) we find the leading expression for $S_T(\mathbf{x}^T)$.

V. Acknowledgement

This work was partially supported by he 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] 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.

[2] P. Billingsley. *Convergence of Probability Measures*. John Wiley, 1968.

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

[4] I. Csiszar and P. Shields. Redundancy rates for renewal and other processes. *IEEE Trans. Inf. Theory*, 42:2065–2072, 1995.

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

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

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

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

[9] 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.

[10] 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.

[11] 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.

[12] T. Inglot and P. Majewski. Simple upper and lower bounds for the multivariate laplace approximation. *J. Approximation Theory*, 186:1–11, 2014.

[13] P. Jacquet, G. Shamir, and W. Szpankowski. Precise minimax regret for logistic regression with categorical feature values. In *ALT'21: PMLR*, volume 128, 2021.

[14] P. Jacquet, G. Shamor, and W. Szpankowski. Precise minimax regret for logistic regression, 2022.

[15] P. Jacquet and W. Szpankowski. *Analytic Pattern Matching: From DNA to Twitter*. Cambridge University Press, Cambridge, 2015.

[16] 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.

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

[18] 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.

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

[20] A. Rakhlin and K. Sridharan. Online nonparametric regression. In *COLT*, pages 1232–1264, 2014.

[21] A. Rakhlin and K. Sridharan. Sdquential probability assignment with binary alphabet and large class of experts, 2015.

[22] J. Rissanen. Universal coding, information, prediction, and estimation. *IEEE Trans. Inform. Theory*, IT-30(4):629–636, Jul. 1984.

[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. Shamir and W. Szpankowski. A general lower bound for regret in logistic regression. In *ISIT*, 2021.

[26] 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.

[27] Gil I. Shamir. Logistic regression regret: What's the catch? In *COLT*, pages 1–24, 2020.

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

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

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

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

[32] C. Wu, M. Heidari, A. Grama, and W. Szpankowski. Sequential vs fixed design regrets in online learning. In *ISIT*, 2022.

[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.

[35] K. Yamanishi. Minimax relative loss analysis for sequential prediction algorithms using parametric hypotheses. In *COLT*, pages 32–43, 1988.