



HAL
open science

Precise Minimax Regret for Logistic Regression

Philippe Jacquet, Gil I. Shamir, Wojciech Szpankowski

► **To cite this version:**

Philippe Jacquet, Gil I. Shamir, Wojciech Szpankowski. Precise Minimax Regret for Logistic Regression. ISIT 2022 - IEEE International Symposium on Information Theory, Jun 2022, Espoo, Finland. hal-03279834

HAL Id: hal-03279834

<https://hal.science/hal-03279834>

Submitted on 6 Jul 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Precise Minimax Regret for Logistic Regression

Philippe Jacquet

INRIA

Paris, France

PHILIPPE.JACQUT@INRIA.FR

Gil I. Shamir

Google Inc

Pittsburgh, PA, USA

GSHAMIR@IEEE.ORG

Wojciech Szpankowski

Department of Computer Science

West Lafayette, IN, USA

SZPAN@PURDUE.EDU

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

We study online logistic regression with binary labels and general feature values in which a learner sequentially 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 combinatoric 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 (weight vector) that is used for prediction. It depends on the feature values, label sequence, and the learning algorithm. In the maximal minimax scenario we seek the best weights for the worst label sequence over all label distributions. For dimension $d = o(T^{1/3})$ we show that the maximal minimax regret grows as

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

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 data. For features uniformly distributed on a d -dimensional sphere or ball we estimate precisely the constant C_d showing that $C_d \sim -(d/2) \log(d/\sqrt{2\pi})$ leading to the minimax regret growing for large d as $(d/2) \log(T/d) - (d/2) \log(\sqrt{8\pi}) + O(1)$. We also extend these results to non-binary labels. The precise maximal minimax regret presented here is the first result of this kind for any feature values and wide range of d . This provides a precise answer to the challenge posed in [McMahan and Streeter \(2010\)](#).

1. Introduction

In online learning sequentially received data must be used to update the predictor for subsequent data. In a supervised online setup, 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 $t \in \{1, 2, \dots, T\}$, 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 recently received a lot of attention in machine learning (Cesa-Bianchi and Lugosi (2006); Shalev-Schwartz and Ben-David (2014)) 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 3.2 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 $t \in \{1, \dots, T\}$ we write $y^T = (y_1, \dots, y_T)$ and $\mathbf{x}^T = (\mathbf{x}_1, \dots, \mathbf{x}_T)$. Then the cumulative relative loss or better *pointwise regret* is defined as in Hazan et al. (2014); Foster et al. (2018); Shamir (2020)

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

where \mathcal{F} is a reference class of functions. More interestingly, it is more rewarding to consider the maximal *minimax regret* defined as

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

In Rakhlin and Sridharan (2014) the worst case minimax regret is studied for all feature vector \mathbf{x}^T , that is, $\max_{\mathbf{x}^T} r_T(\mathbf{x}^T)$.

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 \mathcal{F} to linear functions, that is, $f(\mathbf{x}_t) = \langle \mathbf{x}_t, \mathbf{w} \rangle$ where $\langle \mathbf{x}_t, \mathbf{w} \rangle = \sum_{i=1}^d x_{i,t} w_i$ for some weight vector $\mathbf{w} = (w_1, \dots, w_d)$. Furthermore, as the loss function we take the logistic regression function defined as $\ell(y_t | \mathbf{x}_t, \mathbf{w}) := \ell(f(\mathbf{x}_t), y_t) := \log(1 + \exp(-y_t \langle \mathbf{x}_t, \mathbf{w} \rangle))$.

Finally, we need to choose a class of learning algorithms that predict \hat{y}_t . First, we consider only *improper learning* in which the prediction \hat{y}_t depends on data and labels $(\mathbf{x}_i, y_i)_{i=1}^{t-1}$ seen up to time $t-1$ and data \mathbf{x}_t received at time t . We then postulate that the prediction is based on a learning distribution $Q(y_t | \mathbf{x}_t)$. The most popular class of learning algorithms are Bayesian (cf. Foster et al. (2018); Kakade and Ng (2005); Shamir (2020)), however, we do *not* make such assumption here. For such a setting the pointwise regret for a *given* learning distribution Q is then defined as

$$R_T(Q, y^T | \mathbf{x}^T) = - \sum_{t=1}^T \log Q(y_t | \mathbf{x}_t) - \inf_{\mathbf{w}} \sum_{t=1}^T \ell(\langle \mathbf{x}_t, \mathbf{w} \rangle, y_t) \quad (2)$$

while the *maximal minimax regret* studied here is

$$r_T(\mathbf{x}^T) = \inf_Q \max_{y^T} R_T(Q, y^T | \mathbf{x}^T). \quad (3)$$

Observe that

$$r_T(\mathbf{x}^T) \leq \max_{y^T} R_T(Q, y^T | \mathbf{x}^T) \quad (4)$$

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.

Table 1: Comparison of results

| Problem Setting | Previous Results | This Paper |
|---|---|---|
| Binary Labels, $d = 1$ | : $r_T^* = \frac{1}{2} \log T$ [Davisson (1973)] [Krichevsky and Trofimov (1981)] [McMahan and Streeter (2012)] | |
| Multi Labels; $d = 1$ $\omega(1) = m = o(T)$ | : $r_T^* = \frac{m-1}{2} \log(T/m)$ [Krichevsky and Trofimov] [Orlitsky and Santhanam (2004)] [Shamir (2006a)] | |
| Multi Labels; $d = 1$ $m = \alpha T$ $T = o(m)$ | : $\bar{r}_T^* = T \log B_\alpha + O(1)$ [explicit constant B_α] [Szpankowski and Weinberger (2012)] $r_T^* = T \log \frac{m}{T} + O(T^2/m)$ [Orlitsky and Santhanam (2004); Shamir (2006a); Szpankowski and Weinberger (2012)] | |
| Binary Labels Multi Dimensions d | : Pointwise regret for Bayesian algorithms Q : $L_2 : \ \mathbf{w}^*\ _2 \leq B$ $O(B\sqrt{dT})$ [Xiao (2010)] $R_T \leq \frac{d}{2} \log(1+T)$ [Kakade and Ng (2005)] $R_T \leq \frac{d}{2} \log\left(\frac{B^2 T}{d} + e\right)$ [Foster et al.] $\frac{d}{2} \log \frac{T}{d^3} \leq R_T \leq \frac{d}{2} \log \frac{B^2 T}{d^2}$ [Shamir (2020)] Minimax regret: $\max_{\mathbf{x}^T} r_T^*(\mathbf{x}^T) = O(\text{polylog}(T))$ [Rakhlin and Sridharan (2014, 2015)] For <i>finite feature values</i> and $d = O(1)$: $r_T^* = \frac{d}{2} \log(2T/\pi) + C + O(1/\sqrt{T})$ where C explicit constant Jacquet et al. (2020). | Minimax regret: $r_T^* \leq \max_{y^T} R_T(y^T, Q), \forall Q.$ For any $\mathbf{x} \in [0, 1]$ and $d = o(T^{1/3})$: $r_T^* = \frac{d}{2} \log(2T/\pi) + C_d + O(d^{3/2}/\sqrt{T})$ where $C_d = \int_{\mathbf{R}^d} \sqrt{\det(\mathbf{B}(\mathbf{w}))} d\mathbf{w}.$ Features in d -dimensional sphere: $C_d \sim -\frac{d}{2} \log(d/2\pi)$, hence: $r_T^* = \frac{d}{2} \log \frac{T}{d} - \frac{d}{2} \log \sqrt{8\pi} + O(1).$ |
| Multi Labels m Multi Dimensions d | : Pointwise regret for Bayesian algorithm: L_2 constraints: $R_T \leq 5md \log\left(\frac{BT}{dm} + e\right)$ [Foster et al. (2018)] $\ \mathbf{w}^{*(m)}\ _\infty \leq B$: $R_T \geq \frac{d(m-1)}{2} \log\left(\frac{T}{d \cdot m}\right)$ [Shamir (2020)] Minimax regret: for <i>finite feature values</i> for $d, m = O(1)$: $r_T^* = \frac{d(m-1)}{2} \log(2T/\pi) + C_m + O(1/\sqrt{T})$ where C_m explicit constant Jacquet et al. (2020). | Minimax regret: $\mathbf{x} \in [0, 1]$ and $md = o(T^{1/3})$: $r_T^* = \frac{d(m-1)}{2} \log(2T/\pi) + C_{m,d} + O((dm)^{3/2}/\sqrt{T})$ where $C_{m,d}$ explicit constant. |

Our Contributions and Methods. Our contribution is two-fold. First, we present precise asymptotic expansions for the maximal minimax regret (3) through the so called Shtarkov sum (cf. Shtarkov (1987); Drmota and Szpankowski (2004)). Second, we apply new methodology based using tools of analytic combinatorics such as complex asymptotics and Fourier as well as Mellin transforms (cf Flajolet and Sedgewick (2008); Szpankowski (2001)) to handle Shtarkov sum for the logistic regression.

More precisely, we first represent the minimax regret (3) as the logarithm of the so called Shtarkov sum over all label sequences of the optimal label probability. Such a sum arose already in the universal compression as witnessed by Shtarkov (1987); Drmota and Szpankowski (2004); Szpankowski and Weinberger (2012). 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. (17)). After generalizing it to non-binary labels in Theorem 2, we find in Theorem 4 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 Corollary 5 that for large d the minimax regret grows as $\frac{d}{2} \log \frac{T}{d} - \frac{d}{2} \log \sqrt{8\pi} + O(1)$. In Table 1 we compare our precise findings to known results for the pointwise regret and minimax regret.

Our second technical contribution is in unique and novel methodology based on analytic combinatorics. As mention above, we represent the maximal minimax regret as a Shtarkov sum. Only recently Jacquet et al. (2020) introduced Shtarkov sum in the context of logistic regression (see also Shamir and Szpankowski (2020)). However, as discussed below, Jacquet et al. (2020) studied the minimax regret for *finite* number of distinct feature values, which requires a different method that is used in Jacquet et al. (2020). To analyze asymptotically the Shtarkov sum, we first found the optimal weights \mathbf{w}^* that happen to lie on T -dimensional hyperplane. Then, we translate the Shtarkov sum into a d -dimensional integral that we evaluate using a multi-dimensional saddle point method. Further embellishments, including 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.

Related Work. In this paper we combine methodology of analytic combinatorics (see, e.g., Flajolet and Sedgewick (2008); Jacquet and Szpankowski (2015); Szpankowski (2001)) and information theory (see, e.g., Barron et al. (1998); Drmota and Szpankowski (2004); Krichevsky and Trofimov (1981); Orłitsky and Santhanam (2004); Rissanen (1984, 1996); Shamir (2006b); Xie and Barron (1997)) to study a machine learning problem (see, e.g., Cesa-Bianchi and Lugosi (2006); Shalev-Schwartz and Ben-David (2014)), 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 Drmota and Szpankowski (2004); Orłitsky and Santhanam (2004); Rissanen (1996); Shamir (2006b); Szpankowski (1998); Xie and Barron (1997, 2000) that for a large class of sources (up to Markovian but not for non-Markovian as discussed in Csiszar and Shields (1995); Flajolet and Szpankowski (2002)) 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 also Orłitsky and Santhanam (2004); Shamir (2006b); Szpankowski and Weinberger (2012)). In fact in Szpankowski and Weinberger (2012) 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 the exception of Rakhlin and Sridharan (2014)) with logarithmic regret in the strongly and weakly convex setting. We note that logistic regression seems to fall under weakly

convex setting. A general minimax regret for a wide variety of loss function and references classes are discussed in a series of papers by [Rakhlin and Sridharan \(2014, 2015\)](#).

We first mention work of [Hazan et al. \(2014\)](#) 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 [Hazan et al. \(2014\)](#) showed that the pointwise regret is $\Theta(T^{1/3})$ for $d = 1$ and $O(\sqrt{T})$ for $d > 1$.

For improper learning a more precise results are known. To the best of our knowledge, [Kakade and Ng \(2005\)](#) 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 [Foster et al. \(2018\)](#) to all m . The authors of [Kakade and Ng \(2005\)](#) 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 [McMahan and Streeter \(2012\)](#). The authors of [McMahan and Streeter \(2012\)](#) then posed the question of what happens for larger dimensions. Subsequently, [Foster et al. \(2018\)](#) demonstrated how to achieve regret bounds of $O(d \log(T/d))$ with Bayesian model averaging. These results were strengthened in [Shamir \(2020\)](#), which also provided matching lower bounds. Recently, [Jacquet et al. \(2020\)](#) analyze a precise maximal minimax regret but only for *finite* number of feature values and fixed dimension d . To the best of our knowledge here we present the first precise results for minimax regret.

2. Problem Formulation and Notation

We denote by $\mathbf{x}_t = (x_{1,t}, \dots, x_{d,t})$ a d -dimensional feature vector such that $\|x\| \leq 1$ for some norm $\|\cdot\|$. The label binary vector is denoted as $\mathbf{y}^T = (y_1, \dots, y_T)$ with $y_t \in \{-1, 1\}$ (however, we also present in Section 3.2 some results for non-binary labels). Finally, $\mathbf{w}_t = (w_{1,t}, \dots, w_{d,t})$ 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).

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

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

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 $\ell(y_t | \mathbf{x}_t, \mathbf{w}_t) := \log [1 + \exp(-y_t \langle \mathbf{x}_t, \mathbf{w}_t \rangle)]$. Both $\ell(y_t | \mathbf{x}_t, \mathbf{w}_t)$ and $L(\mathbf{y}^T | \mathbf{x}^T, \mathbf{w}^T)$ 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

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

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 of a learning algorithm (in our probabilistic setting) is to find the best approximation $Q(y_t | \mathbf{x}_t)$ of the unknown distribution $P(y_t | \mathbf{x}_t, \mathbf{w}_t)$. We notice that Q represents an algorithm that predicts y_t . For example, in Bayesian setting, as in [Hazan et al. \(2014\)](#); [Foster et al. \(2018\)](#); [Shamir \(2020\)](#), the label probability $Q(y_t | \mathbf{x}_t)$ is a mixture over \mathbf{w} with some prior $\rho(\mathbf{w})$, that

is,

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

In this paper we do *not* restrict Q to Bayesian learning algorithms.

The *pointwise regret* for a given algorithm/ distribution Q is defined for individual sequences (y_t, \mathbf{x}_t) as in [Kakade and Ng \(2005\)](#); [Hazan \(2012\)](#); [Foster et al. \(2018\)](#); [Shamir \(2020\)](#)

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

for some fixed comparator \mathbf{w} . Thus, in terms of the label distributions P and Q we find

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

where

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

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 [Davisson \(1973\)](#); [Drmota and Szpankowski \(2004\)](#); [Xie and Barron \(2000\)](#), we define the *maximal minimax regret* (conditioned on \mathbf{x}^T) as follows

$$r_T^*(\mathbf{x}^T) := \inf_Q \max_{y^T} [R_T(Q, y^T|\mathbf{x}^T)]. \quad (9)$$

Notice that this definition is over all possible learning algorithms represented by Q . Therefore, it constitutes a (universal) lower bound of the pointwise regret – as expressed in (4) – for all label sequences and for all learning distributions Q , including the Bayesian ones studied in [Kakade and Ng \(2005\)](#); [Foster et al. \(2018\)](#); [Shamir \(2020\)](#).

We study in this paper precise growth of the maximal 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 [Shtarkov \(1987\)](#); [Drmota and Szpankowski \(2004\)](#) we add and subtract from $R_T(Q, y^T|\mathbf{x}^T)$ of (9) the logarithm of the Shtarkov sum defined as

$$S_T(\mathbf{x}^T) := \sum_{y^T} \sup_{\mathbf{w}} P(y^T|\mathbf{x}^T, \mathbf{w}) \quad (10)$$

resulting in

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

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*. Hereafter, we shall study asymptotics of the Shtarkov sum $S_T(\mathbf{x}^T)$ for large T .

3. Main Results

In this section we present our main results. Throughout we write

$$p(w) := (1 + e^{-w})^{-1}, \quad \text{and} \quad q(w) = 1 - p(w) = p(-w).$$

We aim at estimating asymptotically the Shtarkov sum (10) for large T and wide range of d .

3.1. 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 another expression on $P(y^T | \mathbf{x}^T, \mathbf{w})$ as follows

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

Let now

$$L_T(\mathbf{w}) = L_T(\mathbf{w}, \mathbf{x}^T) = \sum_{t=1}^T \log(1 + e^{\langle \mathbf{w}, \mathbf{x}_t \rangle}), \quad \text{and} \quad \mathbf{A}_T = \mathbf{A}(y^T) = \frac{1}{2} \sum_{t=1}^T (1 + y_t) \mathbf{x}_t.$$

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 Appendix. 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 vector gradient of $L_T(\mathbf{w})$. It is easy to see that

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

due to the crucial property $p'(w) = p(w)q(w)$. The optimal probability $P^*(y^T|\mathbf{x}^T, \mathbf{w})$ is then

$$P^*(\mathbf{w}^*) = P^*(y^T|\mathbf{x}^T) = \exp(-L_T(\mathbf{w}^*) + \langle \mathbf{w}^*, \mathbf{G}_T(\mathbf{w}^*) \rangle). \quad (15)$$

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 saddle point. 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 can be formulated as follows that we prove in the next section.

Theorem 1 *Let $\mathbf{x}_t \in [0, 1]^d$, and $p(w) = (1 + e^{-w})^{-1}$ with $q(w) = 1 - p(w)$. The maximal minimax regret becomes asymptotically for $d = o(T^{1/3})$*

$$r^*(\mathbf{x}^T) = \frac{d}{2} \log T - \frac{d}{2} \log 2\pi + C_d(\mathbf{x}^T) + O(d^{3/2}/\sqrt{T}) \quad (16)$$

where

$$C_d(\mathbf{x}^T) = \log \left(\int_{\mathbf{R}^d} \sqrt{\det(\mathbf{B}_d(\mathbf{w}, \mathbf{x}))} dw_1 \cdots dw_d \right) \quad (17)$$

with

$$\mathbf{B}(\mathbf{w}, \mathbf{x}^T) = \frac{1}{T} \sum_{t=1}^T p(\langle \mathbf{x}_t, \mathbf{w} \rangle) q(\langle \mathbf{x}_t, \mathbf{w} \rangle) \mathbf{x}_t \otimes \mathbf{x}_t \quad (18)$$

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

$$\mathbf{B}(\mathbf{w}, \mathbf{X}^T) \rightarrow \mathbf{E}_X[\mathbf{B}(\mathbf{w}, \mathbf{X})] := \bar{\mathbf{B}}(\mathbf{w}) \quad (19)$$

when $T \rightarrow \infty$. We will use this expression in the next section to estimate precisely the constant $C_d(\mathbf{x}^T)$ for features \mathbf{x}^T distributed on a sphere and a ball.

3.2. Extension to Non-binary Labels

Let us now consider a non-binary label alphabet \mathcal{Y} of size m . We will follow [Foster et al. \(2018\)](#) and define a matrix $\mathbf{W} = [\mathbf{w}_1, \dots, \mathbf{w}_{m-1}]$ such that $\mathbf{w}_i = (w_{1,i}, \dots, w_{d,i})$. The multinomial logistic function known also as *softmax function* is then defined as

$$p_\ell(\mathbf{x}^\tau \mathbf{W}) = \frac{e^{\langle \mathbf{x}, \mathbf{w}_\ell \rangle}}{\sum_{k=1}^m e^{\langle \mathbf{x}, \mathbf{w}_k \rangle}} \quad \text{and} \quad q(\mathbf{x}^\tau \mathbf{W}) = 1 - \sum_{i=1}^{m-1} p_\ell(\mathbf{x}^\tau \mathbf{W}) \quad (20)$$

for $\ell = 1, \dots, m-1$. Let also $\mathbf{p} = (p_1, \dots, p_m)$.

Following our derivation for binary labels, we can prove the following result.

Theorem 2 Let $\mathbf{x}_t \in [0, 1]^d$ for the label alphabet \mathcal{Y} be of size m , and for $\mathbf{W} = [\mathbf{w}_1, \dots, \mathbf{w}_{m-1}]$ we define $p_\ell(\mathbf{x}^T \mathbf{W})$ for $\ell = 1, \dots, m-1$ as in (20). Then the maximal minimax regret becomes asymptotically for $md = o(T^{1/3})$

$$r^*(\mathbf{x}^T) = \frac{d(m-1)}{2} \log \frac{T}{2\pi} + \log \left(\int_{\mathbf{R}^{d(m-1)}} \sqrt{\det(\mathbf{B}(\mathbf{W}))} d\mathbf{w}_1 \cdots d\mathbf{w}_{m-1} \right) + O((md)^{3/2}/\sqrt{T}) \quad (21)$$

where $\mathbf{B}_{d,m}(\mathbf{W})$ is a $d(m-1) \times d(m-1)$ matrix defined as

$$\mathbf{B}(\mathbf{W}) = \frac{1}{T} \sum_{i=1}^T [\text{Diag}(\mathbf{p}(\mathbf{x}_i^T \mathbf{W})) - \mathbf{p}(\mathbf{x}_i^T \mathbf{W}) \otimes \mathbf{p}(\mathbf{x}_i^T \mathbf{W})] \otimes \mathbf{x}_i \otimes \mathbf{x}_i.$$

3.3. 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 . We explain our ideas on \mathbf{x}_t distributed uniformly on the sphere \mathcal{S}_d of radius 1. By (19) we know that

$$\mathbf{B}(\mathbf{w}, \mathcal{S}_d) \rightarrow \bar{\mathbf{B}}(\mathbf{w}) = \frac{1}{s_d} \int_{\mathcal{S}_d} p(\langle \mathbf{x}, \mathbf{w} \rangle) q(\langle \mathbf{x}, \mathbf{w} \rangle) \mathbf{x} \otimes \mathbf{x} d\mathbf{x}. \quad (22)$$

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 first present the following lemma that we prove in the Appendix using analytic tools of such as complex asymptotics and Mellin transform.

Lemma 3 Let $f(x) = p(x)q(x) = [(1 + e^{-x})(1 + e^x)]^{-1}$ and $\mathbf{u} = \mathbf{w}/\|\mathbf{w}\|$.

(i) We have the following expression

$$\bar{\mathbf{B}}(\mathbf{w}) = \mu(\mathbf{w}) \mathbf{I}_{d-1}(\mathbf{u}) + \lambda(\mathbf{w}) \mathbf{u} \otimes \mathbf{u} \quad (23)$$

where \mathbf{I}_d is the identity operator orthogonal to \mathbf{u} and

$$\lambda(\mathbf{w}) = \frac{s_{d-1}}{s_d} \int_0^\pi \cos(\theta)^2 \sin(\theta)^{d-2} f(\cos(\theta)\|\mathbf{w}\|) d\theta \quad (24)$$

and

$$\mu(\mathbf{w}) = \frac{s_{d-1}}{s_d} \int_0^\pi \frac{\sin(\theta)^d}{d-1} f(\cos(\theta)\|\mathbf{w}\|) d\theta \quad (25)$$

are the eigenvalues of $\bar{\mathbf{B}}(\mathbf{w})$ with multiplicity 1 and $d-1$, respectively.

(ii) Furthermore, $\det \bar{\mathbf{B}}(\mathbf{w}) = \lambda(\mathbf{w}) \cdot \mu^{d-1}(\mathbf{w})$ and both $\lambda(\mathbf{w})$ and $\mu(\mathbf{w})$ are of order $O(\|\mathbf{w}\|^{-3})$ and $\det(\mathbf{B}(\mathbf{w}))$ is $O(\|\mathbf{w}\|^{-3d})$. More precisely,

$$\det(\mathbf{B}(\mathbf{w})) = 2 \left(\frac{s_{d-1}}{3s_d} \pi^2 \|\mathbf{w}\|^{-3} (1 + O(\|\mathbf{w}\|^{-2})) \right)^d \quad (26)$$

for large $\|\mathbf{w}\| \rightarrow \infty$.

Using Lemma 3 we prove in the next section the following theorem.

Theorem 4 *Under assumptions of Theorem 1 let us now postulate that the feature vector lies on the d -dimensional sphere \mathcal{S}_d or ball \mathcal{B}_d . Then the corresponding minimax regrets satisfy (16) with the constants C_d*

$$C_d(\mathcal{S}_d) = -\frac{d}{2} \log \frac{d}{4} + \frac{d}{4} \log(\pi/8) + \frac{3}{8} \log e + O(1/d) \quad (27)$$

and

$$C_d(\mathcal{B}_d) = -\frac{d}{2} \log \frac{d}{4} + \frac{d}{4} \log(\pi/8) - \frac{1}{8} \log e + O(1/d) \quad (28)$$

respectively.

In conclusion we notice that for large d the regret grows as $d/2 \log(T/d)$. More precisely, we end this section with the following corollary.

Corollary 5 *Under assumptions of Theorem 1 the minimax regret becomes for features lying uniformly on the sphere \mathcal{S}_d*

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

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

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

for large d .

4. Analysis

4.1. Proof of Theorem 1

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

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

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 $e^{\langle \mathbf{w}, \mathbf{A}_T \rangle}$ then satisfies

$$F_T(\mathbf{w}) = \sum_{y^T} e^{\langle \mathbf{w}, \mathbf{A}_T \rangle} = \prod_t (1 + e^{\langle \mathbf{w}, \mathbf{x}_t \rangle}) = \exp(L_T(\mathbf{w})) \quad (29)$$

which can also be written as

$$F_T(\mathbf{w}) = \int \rho_T(\mathbf{A}) e^{\langle \mathbf{w}, \mathbf{A} \rangle} d\mathbf{A} \quad \text{with} \quad \rho_T(\mathbf{w}) = \sum_{\mathbf{A} \in \mathcal{A}_T} N(\mathbf{A}) \delta_{\mathbf{A}}, \quad (30)$$

where $\delta_{\mathbf{A}}$ is the Dirac function on vector \mathbf{A} . Using (15) and above we can re-write the Shtarkov sum as

$$S_T(\mathbf{x}^T) = \sum_{\mathbf{A} \in \mathcal{A}_T} N_T(\mathbf{A}) \exp(-L(\mathbf{w}^*(\mathbf{A})) + \langle \mathbf{w}^*(\mathbf{A}), \mathbf{A} \rangle) \quad (31)$$

that we evaluate asymptotically for large T .

We now express (31) as

$$S_T(\mathbf{x}^T) = \int \rho_T(\mathbf{A})K(\mathbf{A})d\mathbf{x}, \quad \text{where} \quad K(\mathbf{A}) = \exp(-L(\mathbf{w}^*(\mathbf{A})) + \langle \mathbf{w}^*(\mathbf{A}), \mathbf{A} \rangle). \quad (32)$$

Since $w^*(\mathbf{A})$ is the inverse of function \mathbf{G}_T , which is in C^∞ , we conclude that $K(\mathbf{A})$ is in C^∞ and has a finite support contained in $[-T, T]^d$. Let

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

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

$$S_T(\mathbf{x}^T) = \frac{1}{(2\pi)^d} \int_{\mathbf{c}+\mathbf{R}^d} F_T(i\mathbf{w})K^*(-\mathbf{w})d\mathbf{w}. \quad (33)$$

Therefore,

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

By change of variable $\mathbf{A} = \mathbf{G}(\mathbf{w}^*)$ we arrive at

$$K^*(\mathbf{w}) = \int_{\mathbf{R}^d} \exp(-L_T(\mathbf{w}^*) + \langle (\mathbf{w}^* + i\mathbf{w}), \mathbf{G}(\mathbf{w}^*) \rangle) \det(\nabla \mathbf{G}(\mathbf{w}^*)) d\mathbf{w}^* \quad (34)$$

leading to

$$\begin{aligned} S_T(\mathbf{x}^T) &= \frac{1}{(2\pi)^d} \int_{\mathbf{R}^d} \exp(-L_T(\mathbf{w}^*) + \langle \mathbf{w}^*, \mathbf{G}_T(\mathbf{w}^*) \rangle) \det(\nabla \mathbf{G}_T(\mathbf{w}^*)) d\mathbf{w}^* \\ &\quad \cdot \int_{-i\mathbf{c}+\mathbf{R}^d} \exp(L_T(i\mathbf{w}) - i\langle \mathbf{w}, \mathbf{G}_T(\mathbf{w}^*) \rangle) d\mathbf{w}. \end{aligned} \quad (35)$$

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

$$\begin{aligned} S_T(\mathbf{x}^T) &= \frac{1}{(2\pi)^d} \int_{\mathbf{R}^d} \exp(-L_T(\mathbf{w}^*) + \langle \mathbf{w}^*, \mathbf{G}_T(\mathbf{w}^*) \rangle) \det(\nabla \mathbf{G}_T(\mathbf{w}^*)) d\mathbf{w}^* \\ &\quad \cdot \int_{-i\mathbf{w}^*+\mathbf{R}^d} \exp(L_T(i\mathbf{w}) - i\langle \mathbf{w}, \mathbf{G}_T(\mathbf{w}^*) \rangle) d\mathbf{w}. \end{aligned} \quad (36)$$

Finally, we notice that on the segment $\mathfrak{S}(\mathbf{w}) = i\mathbf{w}^*$ 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 $\mathbf{x} \rightarrow 0$ we conclude

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

where $R(\mathbf{w}^*)$ is the third derivative of $L_T(\mathbf{w})$ on \mathbf{w}^* . But $\nabla^2 L_T(\mathbf{w})$ and $R(\mathbf{w}^*)$ are of order $O(T)$, hence we can apply the multidimensional saddle point method (in fact, Laplace method; cf. [Pemantle and Wilson \(2013\)](#)) to find

$$\int_{\mathfrak{S}(\mathbf{w})=i\mathbf{w}^*} \exp(L_T(\mathbf{w}) - \langle \mathbf{w}, \mathbf{G}_T(\mathbf{w}^*) \rangle) d\mathbf{w} = \frac{\exp(L_T(\mathbf{w}^*) - \langle \mathbf{w}^*, \mathbf{G}_T(\mathbf{w}^*) \rangle)}{\sqrt{\det(\nabla^2 L_T(\mathbf{w}^*)/(2\pi))}} \quad (37)$$

$$\times \left(1 + O\left(\frac{d^{3/2}}{\sqrt{T}}\right) \right). \quad (38)$$

The error term follows from [Inglot and Majewski \(2014\)](#) using

$$\int_{\mathbf{R}^d} |x|^3 \exp(-a|x|^2) = 1/a^{(d+3)/2} \pi^{d/2} \frac{\Gamma(d/2 + 3/2)}{\Gamma(d/2)} \quad (39)$$

and applying the Stirling formula to the gamma function for large d . After substituting in (36), we complete the proof of Theorem 1. Details are discussed in the appendix.

4.2. Proof of Theorem 4

In order to prove asymptotic results of Theorem 4, we must evaluate the quantities $\lambda(\mathbf{w})$ and $\mu(\mathbf{w})$, the eigenvalues of matrix $\bar{\mathbf{B}}(\mathbf{w})$ for large d . The main contribution of both integrals is for θ around $\frac{\pi}{2}$. For $\theta = \frac{\pi}{2} + \sqrt{\frac{2}{d}}x$ we have

$$\begin{cases} \sin(\theta)^{d-2} \sim \exp\left(-\frac{d-2}{d}x^2\right) \\ \cos(\theta)^2 \sim \frac{2}{d}x^2 \\ f(\cos(\theta)\|\mathbf{w}\|) \sim \frac{1}{4} \exp\left(-\frac{\|\mathbf{w}\|^2}{2d}x^2\right) \end{cases} \quad (40)$$

leading to

$$\begin{aligned} \int_0^\pi \sin(\theta)^{d-2} \cos(\theta)^2 f(\cos(\theta)\|\mathbf{w}\|) d\theta &\sim (2/d)^{3/2} \int_{-\infty}^{+\infty} \exp\left(-\left(1 + \frac{\|\mathbf{w}\|^2 - 4}{2d}\right)x^2\right) x^2 dx \\ &= (2/d)^{3/2} \frac{\sqrt{\pi}}{2} \left(1 + \frac{\|\mathbf{w}\|^2 - 4}{2d}\right)^{-3/2} \end{aligned} \quad (41)$$

and

$$\begin{aligned} \int_0^\pi \sin(\theta)^d f(\cos(\theta)\|\mathbf{w}\|) d\theta &\sim (2/d)^{1/2} \int_{-\infty}^{+\infty} \exp\left(-\left(1 + \frac{\|\mathbf{w}\|^2}{2d}\right)x^2\right) dx \\ &= (2/d)^{1/2} \sqrt{\pi} \left(1 + \frac{\|\mathbf{w}\|^2}{2d}\right)^{-1/2}. \end{aligned} \quad (42)$$

In summary

$$\det(\mathbf{B}(\mathbf{w})) \sim \left(\frac{s_{d-1}}{s_d}\right)^d (2/d)^{d/2} \frac{1}{d(d-1)^{d-1}} \pi^{d/2} \exp\left(-\frac{\|\mathbf{w}\|^2}{4}\right). \quad (43)$$

To complete the derivation, we need to integrate $\sqrt{\det(\mathbf{B}(\mathbf{w}))}$ over the vectors \mathbf{w} . This leads to

$$\int \sqrt{\det(\mathbf{B}(\mathbf{w}))} d\mathbf{w} \sim \frac{1}{2^d} \left(\frac{s_{d-1}}{s_d} \right)^{d/2} (2/d)^{d/4} \frac{1}{\sqrt{d}} \frac{1}{\sqrt{d-1}^{d-1}} \pi^{d/4} (2\sqrt{2\pi})^d. \quad (44)$$

The final touch is to get an estimate of the ratio $\frac{s_{d-1}}{s_d}$. But

$$\begin{aligned} \frac{s_{d-1}}{s_d} &= \sqrt{\pi} \frac{\Gamma(d/2)}{\Gamma((d-1)/2)}, \\ \frac{\Gamma(d/2)}{\Gamma((d-1))} &= \sqrt{\frac{d}{2} - \frac{1}{4}} + O(1/d) \end{aligned}$$

so that

$$\int \sqrt{\det(\mathbf{B}(\mathbf{w}))} d\mathbf{w} \sim (\pi/8)^{d/4} e^{3/8} (d/4)^{-d/2}. \quad (45)$$

which completes the derivation of (27).

We prove (28) in a similar manner where the extra factor $\left(\frac{d}{d-1}\right)^{d/2} \left(\frac{d-1}{d+1}\right)^{\frac{d-1}{2}}$ comes from the volume of a ball. At the end, for the unit ball features we find

$$\int \sqrt{\det(\mathbf{B}(\mathbf{w}))} d\mathbf{w} \sim (\pi/8)^{d/4} e^{-1/8} (d/4)^{-d/2}. \quad (46)$$

More details can be found in the Appendix.

Acknowledgments

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, CCF-2007238, and Google Research Award.

References

- 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.
- N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning and Games*. Cambridge University Press, 2006.
- I. Csiszar and P. Shields. Redundancy rates for renewal and other processes. *IEEE Trans. Inf. Theory*, 42:2065–2072, 1995.
- L. D. Davisson. Universal noiseless coding. *IEEE Trans. Inf. Theory*, IT-19(6):783–795, Nov. 1973.
- M. Drmota and W. Szpankowski. Precise minimax redundancy and regrets. *IEEE Trans. Inf. Theory*, IT-50:2686–2707, 2004.
- P. Flajolet and R. Sedgewick. *Analytic Combinatorics*. Cambridge University Press, 2008.

- P. Flajolet and W. Szpankowski. Analytic variations on redundancy rates of renewal processes. *IEEE Trans. Information Theory*, 48:2911–2921, 2002.
- Dylan J Foster, Satyen Kale, Haipeng Luo and Mehryar Mohri, and Karthik Sridharan. Logistic regression: The importance of being improper. In *COLT - Conference on Learning Theory*, 2018.
- 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.
- 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.
- T. Inglot and P. Majewski. Simple upper and lower bounds for the multivariate laplace approximation. *J. Approximation Theory*, 186:1–11, 2014.
- P. Jacquet and W. Szpankowski. *Analytic Pattern Matching: From DNA to Twitter*. Cambridge University Press, Cambridge, 2015.
- P. Jacquet, G. Shamir, and W. Szpankowski. Precise minimax regret for logistic regression with categorical feature values. In *ALT'21*, 2020.
- 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.
- R. E. Krichevsky and V. K. Trofimov. The performance of universal encoding. *IEEE Trans. Inform. Theory*, IT-27(2):199–207, Mar. 1981.
- H. B. McMahan and M. Streeter. Adaptive bound optimization for online convex optimization. In *COLT*, 2010.
- 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.
- A. Orłitsky and N. P. Santhanam. Speaking of infinity. *IEEE Trans. Inf. Theory*, 50(10):2215–2230, Oct. 2004.
- R. Pemantle and M. Wilson. *Analytic Combinatorics in Several Variables*. Cambridge University Press, 2013.
- A. Rakhlin and K. Sridharan. Online nonparametric regression. In *COLT*, pages 1232–1264, 2014.
- A. Rakhlin and K. Sridharan. Sdquential probability assignment with binary alphabet and large class of experts, 2015.
- J. Rissanen. Universal coding, information, prediction, and estimation. *IEEE Trans. Inform. Theory*, IT-30(4):629–636, Jul. 1984.
- J. Rissanen. Fisher information and stochastic complexity. *IEEE Trans. Information Theory*, 42: 40–47, 1996.

- S. Shalev-Schwartz and S. Ben-David. *Understanding Machine learning*. Cambridge University Press, 2014.
- G. Shamir and W. Szpankowski. A general lower bound for regret in logistic regression, 2020.
- G. I. Shamir. Universal lossless compression with unknown alphabets - the average case. *IEEE Trans. Inform. Theory*, 52(11):4915–4944, Nov. 2006a.
- G. I. Shamir. On the MDL principle for i.i.d. sources with large alphabets. *IEEE Trans. Inform. Theory*, 52(5):1939–1955, May 2006b.
- Gil I. Shamir. Logistic regression regret: What’s the catch? In *COLT*, pages 1–24, 2020.
- Y. M. Shtarkov. Universal sequential coding of single messages. *Problems of Information Transmission*, 23(3):3–17, Jul.-Sep. 1987.
- W. Szpankowski. On asymptotics of certain recurrences arising in universal coding. *Problems of Information Transmission*, 34:55–61, 1998.
- W. Szpankowski. *Average Case Analysis of Algorithms on Sequences*. Wiley, New York, 2001.
- W. Szpankowski and M. Weinberger. Minimax pointwise redundancy for memoryless models over large alphabets. *IEEE Trans. Information Theory*, 58:4094–4104, 2012.
- Lin Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *J. Machine Learning Research*, 11:2543–2596, 2010.
- Q. Xie and A. Barron. Minimax redundancy for the class of memoryless sources. *IEEE Trans. Information Theory*, pages 647–657, 1997.
- Q. Xie and A. Barron. Asymptotic minimax regret for data compression, gambling, and prediction. *IEEE Trans. Information Theory*, 46:431–445, 2000.

Appendix A. Proof of Lemma 3

We prove here Lemma 3 that we repeat below for the reader convenience.

Lemma 6 Let $f(x) = p(x)q(x) = [(1 + e^{-x})(1 + e^x)]^{-1}$ and $\mathbf{u} = \mathbf{w}/\|\mathbf{w}\|$.

(i) We have the following expression

$$\bar{\mathbf{B}}(\mathbf{w}) = \mu(\mathbf{w})\mathbf{I}_{d-1}(\mathbf{u}) + \lambda(\mathbf{w})\mathbf{u} \otimes \mathbf{u} \quad (47)$$

where \mathbf{I}_d is the identity operator orthogonal to \mathbf{u} and

$$\lambda(\mathbf{w}) = \frac{s_{d-1}}{s_d} \int_0^\pi \cos(\theta)^2 \sin(\theta)^{d-2} f(\cos(\theta)\|\mathbf{w}\|) d\theta \quad (48)$$

and

$$\mu(\mathbf{w}) = \frac{s_{d-1}}{s_d} \int_0^\pi \frac{\sin(\theta)^d}{d-1} f(\cos(\theta)\|\mathbf{w}\|) d\theta \quad (49)$$

are the eigenvalues of $\bar{\mathbf{B}}(\mathbf{w})$ with cardinality 1 and $d-1$, respectively.

(ii) Furthermore, $\det \bar{\mathbf{B}}(\mathbf{w}) = \lambda(\mathbf{w}) \cdot \mu^{d-1}(\mathbf{w})$ and both $\lambda(\mathbf{w})$ and $\mu(\mathbf{w})$ are of order $O(\|\mathbf{w}\|^{-3})$ and $\det(\mathbf{B}(\mathbf{w}))$ is $O(\|\mathbf{w}\|^{-3d})$. More precisely,

$$\det(\mathbf{B}(\mathbf{w}^*)) = 2 \left(\frac{s_{d-1}}{3s_d} \pi^2 \|\mathbf{w}\|^{-3} (1 + O(\|\mathbf{w}\|^{-2})) \right)^d \quad (50)$$

for large $\|\mathbf{w}\| \rightarrow \infty$.

We start with part (i). Let θ be the angle between \mathbf{x} and \mathbf{u} . We have the decomposition $\mathbf{x} = \cos(\theta)\mathbf{u} + \mathbf{b}$ with $\mathbf{b} \in \sin \theta \mathcal{S}_{d-1}(\mathbf{u})$ where $\mathcal{S}_{d-1}(\mathbf{u})$ is the unit hypersphere orthogonal to \mathbf{u} . Since \mathbf{x} 's have a spheric symmetry in its distribution, so it is the case for the \mathbf{b} 's in $\sin \theta \mathcal{S}_{d-1}(\mathbf{u})$ for any given angle θ . Thus

$$\bar{\mathbf{B}}(\mathbf{w}) = \frac{1}{s_d} \int_0^\pi f(\|\mathbf{w}\| \cos \theta) d\theta \int_{\sin \theta \mathcal{S}_{d-1}(\mathbf{u})} (\mathbf{b} + \cos \theta \mathbf{u}) \otimes (\mathbf{b} + \cos \theta \mathbf{u}) d\mathbf{b} \quad (51)$$

$$\begin{aligned} &= \frac{1}{s_d} \int_0^\pi f(\|\mathbf{w}\| \cos \theta) d\theta \int_{\sin \theta \mathcal{S}_{d-1}(\mathbf{u})} (\mathbf{b} \otimes \mathbf{b} + (\cos \theta)^2 \mathbf{u} \otimes \mathbf{u}) d\mathbf{b} \\ &+ \frac{1}{s_d} \int_0^\pi f(\|\mathbf{w}\| \cos \theta) d\theta \int_{\sin \theta \mathcal{S}_{d-1}(\mathbf{u})} \cos \theta (\mathbf{b} \otimes \mathbf{u} + \mathbf{u} \otimes \mathbf{b}) d\mathbf{b}. \end{aligned} \quad (52)$$

Again due to the spheric symmetry of \mathbf{b} we also have $\int_{\sin \theta \mathcal{S}_{d-1}(\mathbf{u})} \mathbf{b} = 0$ leading to

$$\begin{aligned} \bar{\mathbf{B}}(\mathbf{w}) &= \frac{1}{s_d} \int_0^\pi f(\|\mathbf{w}\| \cos \theta) d\theta \int_{\sin \theta \mathcal{S}_{d-1}(\mathbf{u})} (\mathbf{b} \otimes \mathbf{b} + (\cos \theta)^2 \mathbf{u} \otimes \mathbf{u}) d\mathbf{b} \\ &= \frac{1}{s_d} \int_0^\pi f(\|\mathbf{w}\| \cos \theta) (\sin \theta)^{d-1} d\theta \\ &\quad \int_{\mathcal{S}_{d-1}(\mathbf{u})} ((\sin \theta)^2 \mathbf{b} \otimes \mathbf{b} + (\cos \theta)^2 \mathbf{u} \otimes \mathbf{u}) d\mathbf{b}. \end{aligned} \quad (53)$$

The $(\sin \theta)^{d-1}$ factor arises from the change of integration domain from $\sin \theta \mathcal{S}_{d-1}(\mathbf{u})$ to $\mathcal{S}_{d-1}(\mathbf{u})$.

The quantity $\int_{\mathcal{S}_{d-1}(\mathbf{u})} \mathbf{b} \otimes \mathbf{b}$ is the $(d-1) \times (d-1)$ matrix whose (i, j) coefficient is $\int_{\mathcal{S}_{d-1}} b_i b_j d\mathbf{b}$. Clearly, by spheric symmetry of the \mathbf{b} vectors $\int_{\mathcal{S}_{d-1}} b_i b_j d\mathbf{b} = 0$ when $i \neq j$. We also have for all $i \neq j$:

$$\int_{\mathcal{S}_{d-1}} (b_i)^2 d\mathbf{b} = \int_{\mathcal{S}_{d-1}} (b_j)^2 d\mathbf{b} = \frac{1}{d-1} \int_{\mathcal{S}_{d-1}} \|\mathbf{b}\|^2 d\mathbf{b} = \frac{s_{d-1}}{d-1}. \quad (54)$$

Thus

$$\int_{\mathcal{S}_{d-1}(\mathbf{u})} \mathbf{b} \otimes \mathbf{b} d\mathbf{b} = \frac{s_{d-1}}{d-1} \mathbf{I}_{d-1}(\mathbf{u}) \quad (55)$$

and similarly

$$\int_{\mathcal{S}_{d-1}(\mathbf{u})} \mathbf{u} \otimes \mathbf{u} d\mathbf{b} = s_{d-1} \mathbf{u} \otimes \mathbf{u} \quad (56)$$

which completes the proof of part (i) of the lemma.

Now we move to part (ii) of Lemma 3. Both $\lambda(\mathbf{w})$ and $\mu(\mathbf{w})$ are functions of $w = \|\mathbf{w}\|$. We write $\lambda(w) = \lambda(\|\mathbf{w}\|)$ and $\mu(w) = \mu(\|\mathbf{w}\|)$. To capture the asymptotics of these functions we apply Mellin transform which is an effective tool of analytic combinatorics for complex asymptotics. The reader is referred to [Flajolet and Sedgewick \(2008\)](#) and [Szpankowski \(2001\)](#) for detailed discussions.

The Mellin transforms $\lambda^*(s)$ and $\mu^*(s)$ of $\lambda(w)$ and $\mu(w)$ are defined, respectively, as

$$\begin{cases} \lambda^*(s) &= \int_0^\infty \lambda(w) w^{s-1} dw, \\ \mu^*(s) &= \int_0^\infty \mu(w) w^{s-1} dw. \end{cases} \quad (57)$$

Observe now that

$$\lambda(w) = 2 \frac{s_{d-1}}{s_d} \int_0^{\pi/2} f(\cos(\theta)w) \cos^2(\theta) \sin^{d-2}(\theta) d\theta \quad (58)$$

$$= \frac{2s_{d-1}}{s_d} \int_0^1 y^2 (1-y^2)^{(d-3)/2} f(yx) dy \quad (59)$$

via the change of variable $y = \cos(\theta)$. Thus we find

$$\lambda^*(s) = \frac{2s_{d-1}}{s_d} \int_0^1 (1-y^2)^{(d-3)/2} y^2 dy \int_0^\infty f(yx) x^{s-1} dx \quad (60)$$

$$= \frac{2s_{d-1}}{s_d} f^*(s) \int_0^1 (1-y^2)^{(d-3)/2} y^{2-s} dy \quad (61)$$

$$= f^*(s) \beta_1^*(3-s) \quad (62)$$

where $f^*(s)$ is the Mellin transform of function $f(x) = p(x)q(x)$ and $\beta_1(s)$ is the Mellin transform of the function $(1-y^2)^{(d-3)/2} y$ defined over $[0, 1]$.

The Mellin transform $\beta_1^*(s)$ is defined for $\Re(s) > 0$ and being locally analytical it has poles on the negative even integers, corresponding to the Taylor expansion of $(1-y^2)^{(d-3)/2}$. The Mellin transform $f^*(s)$ of function $f(x)$ is

$$f^*(s) = (s-1)^2 (s-2) h^*(s-2)$$

where $h^*(s)$ is the Mellin transform of function $h(x) = \log(1 + e^{-x})$ defined for $\Re(s) > 0$. The Mellin transform $f^*(s)$ is defined for $\Re(s) > 2$ but the simple poles at $s = 1$ and $s = 2$ are canceled by the factor $(s - 1)(s - 2)$ thus is finally defined for $\Re(s) > 0$. More precisely, we have

$$h^*(s) = (1 - 2^{-s})\zeta(s + 1)\Gamma(s)$$

where $\Gamma(s)$ is the Euler gamma function and $\zeta(s)$ is the Riemann zeta function.

The product $\beta(2 - s)(s - 1)^2(s - 2)h^*(s - 2)$ is defined for $\Re(s) \in [0, 1]$. But the simple pole at $s = 1$ is canceled by the additional factor $(s - 1)$. The next pole is at $s = 3$ which has the residue $-\zeta(2)\Gamma(3) = -\pi^2/3$, thus $\lambda^*(s)$ is defined for $\Re(s) \in]0, 3[$.

We can make a similar analysis for $\mu^*(s)$ and we arrive at

$$\begin{cases} \lambda^*(s) &= 2^{\frac{s_d-1}{s_d}}\beta_1^*(3-s)f^*(s), \\ \mu^*(s) &= 2^{\frac{s_d-1}{(d-1)s_d}}\beta_2^*(1-s)f^*(s), \end{cases} \quad (63)$$

where $\beta_1^*(s)$ and $\beta_2^*(s)$ are respectively the Mellin transform of function $(1 - y^2)^{(d-3)/2}$ and function $(1 - y^2)^{(d-1)/2}$. The Mellin transform $\mu^*(s)$ is also defined on $\Re(s) \in]0, 3[$ and has a simple pole at $s = 3$ with residue $-\zeta(2)\Gamma(3)^{\frac{d-1}{2}} = -\pi^2/6(d - 1)$. For both $\lambda^*(s)$ and $\mu^*(s)$ the next pole is at $s = 5$.

We now apply the inverse Mellin transform defined as (cf. [Szpankowski \(2001\)](#))

$$\begin{aligned} \lambda(w) &= \int_{\Re(s)=1} \lambda^*(s)w^{-s}ds, \\ \mu(w) &= \int_{\Re(s)=1} \mu^*(s)w^{-s}ds \end{aligned} \quad (64)$$

to extract asymptotics of $\lambda(w)$ and $\mu(w)$ for $w \rightarrow \infty$. By moving the integration path over the simple poles at $s = 3$ and $s = 5$ and catching the residues we finally obtain

$$\begin{aligned} \lambda(w) &= 2^{\frac{s_d-1}{3s_d}}\pi^2w^{-3} + O(w^{-5}), \\ \mu(w) &= \frac{s_d-1}{3s_d}\pi^2w^{-3} + O(w^{-5}) \end{aligned} \quad (65)$$

for $w \rightarrow \infty$. In conclusion

$$\det(\bar{\mathbf{B}}(\mathbf{w})) = 2 \left(\frac{s_d-1}{3s_d}\pi^2\|\mathbf{w}\|^{-3}(1 + O(\|\mathbf{w}\|^{-2})) \right)^d \quad (66)$$

when $\|\mathbf{w}\| \rightarrow \infty$. This completes the proof of Lemma 3.

Asymptotics for the ball \mathcal{B}_d . Here, we complete the proof of the asymptotics for the ball. We study the case where feature \mathbf{x}_t is uniformly distributed inside the unit ball \mathcal{B}_d of dimension d . In this case the vector \mathbf{x} such that $\langle \mathbf{u}, \mathbf{x} \rangle = \cos(\theta)$ satisfies the decomposition $\mathbf{x} = \cos(\theta)\mathbf{u} + \sin(\theta)\mathbf{b}$ where \mathbf{b} is uniformly distributed inside the unit ball of dimension $d - 1$ orthogonal to \mathbf{u} . Thus

$$\lambda(\mathbf{w}) = \frac{1}{v_d} \int_0^\pi \cos(\theta)^2 \sin(\theta)^d v_d f(\cos(\theta)\|\mathbf{w}\|) d\theta \quad (67)$$

where v_d is the volume of the unit ball in dimension d (in fact $v_d = s_d/d$). Therefore,

$$\mu(\mathbf{w}) = \frac{1}{v_d} \int_0^\pi \sin(\theta)^d f(\cos(\theta)\|\mathbf{w}\|) d\theta \int_{\|\mathbf{b}\|=1} \langle \mathbf{v}, \mathbf{b} \rangle^2 d\mathbf{b}. \quad (68)$$

$$= \frac{1}{v_d} \int_0^\pi \sin(\theta)^d f(\cos(\theta)\|\mathbf{w}\|) d\theta \int_0^1 s_{d-1} \frac{r^2}{d-1} r^{d-2} dr \quad (69)$$

$$= \frac{v_{d-1}}{v_d} \int_0^\pi \frac{\sin(\theta)^d}{d+1} f(\cos(\theta)\|\mathbf{w}\|) d\theta. \quad (70)$$

Using $\sqrt{\det(\bar{\mathbf{B}}(\mathbf{w}))} = \sqrt{\lambda(\mathbf{w})(\mu(\mathbf{w}))^{d-1}}$, all computations done with an extra factor $\left(\frac{d}{d-1}\right)^{d/2} \left(\frac{d-1}{d+1}\right)^{\frac{d-1}{2}}$, we find at the end

$$\int \sqrt{\det(\bar{\mathbf{B}}(\mathbf{w}))} d\mathbf{w} \sim (\pi/8)^{d/4} e^{-1/8} (d/4)^{-d/2} \quad (71)$$

which establishes (28).

A.1. Error of the Saddle Point Method

We provide here more details of the error term of the saddle point method as expressed in (38).

Theorem 7 (Error of (38)) *The error term of the Shtarkov sum (36) is*

$$O\left(\frac{d^{3/2}}{\sqrt{T\lambda(\mathbf{w}^*)^3}}\right)$$

where $\lambda(\mathbf{w}^*)$ is the main eigenvalue of $\bar{\mathbf{B}}(\mathbf{w}^*)$ at \mathbf{w}^* .

Proof The integral (37) is asymptotically approximated by the saddle point method which will also lead to the error term estimation. We use the following by the change of variable $\mathbf{x} = \sqrt{\nabla^2 L_T(\mathbf{w}^*)} \mathbf{w}$

$$\begin{aligned} \int_{\mathfrak{S}(\mathbf{w})=-\mathbf{w}^*} \exp(L_T(\mathbf{w}) - \langle \mathbf{w}, \mathbf{G}_T(\mathbf{w}^*) \rangle) d\mathbf{w} &= \frac{\exp(L_T(\mathbf{w}^*) - \langle \mathbf{w}^*, \mathbf{G}_T(\mathbf{w}^*) \rangle)}{(2\pi)^d \sqrt{\det(\nabla^2 L_T(\mathbf{w}^*))}} \\ &\times \int_{\mathbf{R}^d} \exp\left(-\frac{1}{2} \|\mathbf{x}\|^2 + O(L_T^{(3)}) \|(\sqrt{\nabla^2 L_T(\mathbf{w}^*)}^{-1} \mathbf{x})\|^3\right) \end{aligned}$$

where $L_T^{(3)}$ is an estimate of the norm the third derivative applied to $L_T(\mathbf{w}^*)$, and is of order $O(T)$. Using the fact that

$$\|(\sqrt{\nabla^2 L_T(\mathbf{w}^*)}^{-1} \mathbf{x})\| \leq \frac{\|\mathbf{x}\|}{\sqrt{T\lambda(\mathbf{w}^*)}}$$

where $\lambda(\mathbf{w}^*)$ is the main eigenvalue of $\nabla^2 L_T(\mathbf{w}^*)$, we find

$$\begin{aligned} \int_{\mathfrak{S}(\mathbf{w})=-\mathbf{w}^*} \exp(L_T(\mathbf{w}) - \langle \mathbf{w}, \mathbf{G}_T(\mathbf{w}^*) \rangle) d\mathbf{w} &= \frac{\exp(L_T(\mathbf{w}^*) - \langle \mathbf{w}^*, \mathbf{G}_T(\mathbf{w}^*) \rangle)}{(2\pi)^d \sqrt{\det(\nabla^2 L_T(\mathbf{w}^*))}} \\ &\times \int_{\mathbf{R}^d} \exp\left(-\frac{1}{2} \|\mathbf{x}\|^2\right) \left(1 + O\left(\frac{\|\mathbf{x}\|^3}{\sqrt{T\lambda(\mathbf{w}^*)^3}}\right)\right). \end{aligned}$$

But we know that for $a > 0$

$$\int_{\mathbf{R}^d} \|x\|^3 \exp(-a\|x\|^2) = 1/a^{(d+3)/2} \pi^{d/2} \frac{\Gamma(d/2 + 3/2)}{\Gamma(d/2)}, \quad (72)$$

thus we conclude that

$$\int_{\mathbf{R}^d} \exp\left(-\frac{1}{2} \|\mathbf{x}\|^2\right) O\left(\frac{\|\mathbf{x}\|^3}{\sqrt{T}}\right) d\mathbf{x} = O\left(\frac{\Gamma(d/2 + 3/2)}{\Gamma(d/2) \sqrt{T}}\right). \quad (73)$$

To complete, we observe that $\frac{\Gamma(d/2+3/2)}{\Gamma(d/2)} \sim (d/2)^{3/2}$ when $d \rightarrow \infty$. ■