Precise Minimax Regret for Logistic Regression

Philippe Jacquet
INRIA
Paris, France

Gil I. Shamir
Google Inc
Pittsburgh, PA, USA

Wojciech Szpankowski
Department of Computer Science
West Lafayette, IN, USA

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 \((d/2)\log(T/d) - (d/2)\log(8\pi) + O(1)\). 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 \( t \in \{1, 2, \ldots, T\} \), 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 recently received a lot of attention in machine learning (Cesa-Bianchi and Lugosi (2006); Shalev-Shwartz 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 \(x_t\) 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 \(x_t\) 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 \(x^T = (x_1, \ldots, x_T)\). Then the cumulative relative loss or better pointwise regret is defined as in Hazan et al. (2014); Foster et al. (2018); Shamir (2020)

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

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

\[
r_T(x^T) = \inf_{\hat{y}^T} \max_{y^T} R_T(\hat{y}^T, y^T | x^T).
\]

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

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, \(f(x_t) = \langle x_t, w \rangle\) where \(\langle x_t, w \rangle = \sum_{i=1}^{d} x_{i,t}w_i\) for some weight vector \(w = (w_1, \ldots, w_d)\). Furthermore, as the loss function we take the logistic regression function defined as \(\ell(y_t | x_t, w) := \ell(f(x_t), y_t) := \log(1 + \exp(-y_t \langle x_t, w \rangle))\).

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 \((x_i, y_i)_{i=1}^{t-1}\) seen up to time \(t-1\) and data \(x_t\) received at time \(t\). We then postulate that the prediction is based on a learning distribution \(Q(y_i | x_i)\). 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

\[
R_T(Q, y^T | x^T) = -\sum_{t=1}^{T} \log Q(y_t | x_t) - \inf_{w} \sum_{t=1}^{T} \ell(\langle x_t, w \rangle, y_t)
\]

while the maximal minimax regret studied here is

\[
r_T(x^T) = \inf_{Q} \max_{y^T} R_T(Q, y^T | x^T).
\]

Observe that

\[
r_T(x^T) \leq \max_{y^T} R_T(Q, y^T | x^T)
\]

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

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

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

where the constant \( C_d(x^T) \) depends on the dimension \( d \) and data \( x^T \). 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 \( C_d(x^T) \) when data \( x^T \) are distributed uniformly on a \( d \)-dimensional sphere \( S_d \) and a ball \( B_d \). This allows us to show in Corollary 5 that for large \( d \) the minimax regret grows as \( \frac{d}{2} \log T - \frac{d}{2} \log \sqrt{8\pi} + O(1) \). 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 mentioned 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 \( C_d(x^T) \) when the feature \( x^T \) 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); Orlitsky 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); Or litsky 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 Orlitsky 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 $w_t$ is based on knowledge available to the learner up to time $t - 1$. Unlike the improper learning, studied in this paper, where feature $x_t$ 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 $x_t = (x_{1,t}, \ldots, x_{d,t})$ a $d$-dimensional feature vector such that $||x|| \leq 1$ for some norm $|| \cdot ||$. The label binary vector denoted as $y^T = (y_1, \ldots, y_T)$ with $y_t \in \{-1, 1\}$ (however, we also present in Section 3.2 some results for non-binary labels). Finally, $w_t = (w_{1,t}, \ldots, w_{d,t})$ 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).

The cumulative logistic loss of an algorithm that plays $w_t$ at round $t$ is

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

(5)

where $\langle x_t, w_t \rangle = \sum_{i=1}^{d} x_{i,t} w_{i,t}$ is the scalar product of $x_t$ and $w_t$. To simplify we also write $\ell(y_t | x_t, w_t) := \log [1 + \exp(-y_t \langle x_t, w_t \rangle)]$. Both $\ell(y_t | x_t, w_t)$ and $L(y^T | x^T, w^T)$ depend on $x_t$ and $w_t$ only through the product $\langle x_t, w_t \rangle$. 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

$$P(y_t | x_t, w_t) = \frac{1}{1 + \exp(-y_t \langle x_t, w_t \rangle)}$$

(6)

and clearly $\ell(y_t | x_t, w_t) = -\log P(y_t | x_t, w_t)$.

Finally, we observe that the goal of a learning algorithm (in our probabilistic setting) is to find the best approximation $Q(y_t | x_t)$ of the unknown distribution $P(y_t | x_t, w_t)$. 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 $Q(y_t | x_t)$ is a mixture over $w$ with some prior $\rho(w)$, that
is,
\[ Q(y^T|x^T) := \int_w \rho(w)P(y^T|x^T, w)dw. \]

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 \((y_t, x_t)\) as in Kakade and Ng (2005); Hazan (2012); Foster et al. (2018); Shamir (2020)

\[ R(Q, y^T|x^T) := -\sum_{t=1}^T \log Q(y_t|x_t) - \min_w \sum_{t=1}^T \ell(y_t|x_t, w) \]

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

\[ R_T(Q, y^T|x^T) = \log \sup_w \frac{P(y^T|x^T, w)}{Q(y^T|x^T)} \tag{7} \]

where

\[ P(y^T|x^T, w) = \prod_{t=1}^T \left(1 + \exp(-y_t\langle x_t, w \rangle)\right)^{-1}. \tag{8} \]

The pointwise regret \( R_T(Q, y^T|x^T) \) is a function of label sequence \( y^T \), data/ feature vector \( x^T \), 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 \( x^T \)) 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 \( x^T \)) as follows

\[ r^*_T(x^T) := \inf_Q \max_{y^T} [R_T(Q, y^T|x^T)]. \tag{9} \]

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 \( R_T(Q, y^T|x^T) \) of (9) the logarithm of the Shtarkov sum defined as

\[ S_T(x^T) := \sup_{y^T} \sup_w P(y^T|x^T, w) \tag{10} \]

resulting in

\[ r^*_T(x^T) = \inf_Q \max_{y^T} (-\log Q(y^T|x^T) + \log P^*(y^T|x^T)) + \log \sup_{y^T} \sup_w P(y^T|x^T, w) \]

\[ = \log \sup_{y^T} \sup_w P(y^T|x^T, w) = \log S_T(x^T) \tag{11} \]
where we set \( Q(y^T, x^T) = P^*(y^T|x^T) \) with

\[
P^*(y^T|x^T) := \frac{\sup_w P(y^T|x^T, w)}{\sum_{v^T} \sup_w P(v^T|x^T, w)}
\]

being the maximum-likelihood distribution. Hereafter, we shall study asymptotics of the Shtarkov sum \( S_T(x^T) \) 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 \( P(y^T|x^T, w) \) as given in (8). Noting that

\[
P(y_t = 1|x_t, w) = \frac{1}{1 + \exp(-\langle x_t, w \rangle)} = \exp(\langle x_t, w \rangle) \frac{1}{1 + \exp(\langle x_t, w \rangle)}
\]

we find another expression on \( P(y^T|x^T, w) \) as follows

\[
P(y^T|x^T, w) = \prod_{t=1}^{T} \frac{\exp \left( \frac{1+y_t}{2} \langle x_t, w \rangle \right)}{1 + \exp(\langle x_t, w \rangle)}
\]

\[
= \exp \left( - \sum_{t=1}^{T} \log(1 + e^{\langle x_t, w \rangle}) + \sum_{t=1}^{T} \frac{1}{2} (1 + y_t) \langle x_t, w \rangle \right).
\]

Let now

\[
L_T(w) = L_T(w, x^T) = \sum_{t=1}^{T} \log(1 + e^{\langle w, x_t \rangle}), \quad \text{and} \quad A_T = A(y^T) = \frac{1}{2} \sum_{t=1}^{T} (1 + y_t)x_t.
\]

Then \( P(y^T|x^T, w) \) becomes

\[
P(y^T|x^T, w) = \exp \left( -L_T(w, x^T) + \langle w, A_T(y^T) \rangle \right). \quad (13)
\]

Now we sketch the road map of our approach, leaving technical details to the next section and Appendix. The optimal value \( w^* \) that maximizes \( P(y^T|x^T, w) \) satisfies

\[
\nabla_w L_T(w^*) = A_T(y^T)
\]

(14)

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

\[
G_T(w) := \nabla_w L_T(w) = \sum_{t=1}^{T} p(\langle w, x_t \rangle)x_t
\]
due to the crucial property \( p'(w) = p(w)q(w) \). The optimal probability \( P^*(y^T|x^T, w) \) is then

\[
P^*(w^*) = P^*(y^T|x^T) = \exp(-L_T(w^*) + \langle w^*, G_T(w^*) \rangle).
\]

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

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

where

\[
\nabla G(w) = \sum_{t=1}^{T} p(\langle w, x_t \rangle) q(\langle w, x_t \rangle) x_t \otimes x_t.
\]

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

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

\[
r^*(x^T) = \frac{d}{2} \log T - \frac{d}{2} \log 2\pi + C_d(x^T) + O(d^{3/2}/\sqrt{T})
\]

where

\[
C_d(x^T) = \log\left(\int_{\mathbb{R}^d} \sqrt{\det(B_d(w, x))dw_1 \cdots dw_d}\right)
\]

with

\[
B(w, x^T) = \frac{1}{T} \sum_{t=1}^{T} p(\langle x_t, w \rangle) q(\langle x_t, w \rangle) x_t \otimes x_t
\]

and \( x_t \otimes x_t = x_t x_t^T \) being the tensor product of \( x_t \) with \( \tau \) denoting the transpose.

In passing we should observe that if data \( X_t \) is generated by a stationary ergodic source, then by the ergodic theorem we conclude that

\[
B(w, X^T) \to E_X[B(w, X)] := \bar{B}(w)
\]

when \( T \to \infty \). We will use this expression in the next section to estimate precisely the constant \( C_d(x^T) \) for features \( x^T \) 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 \( W = [w_1, \ldots, w_{m-1}] \) such that \( w_i = (w_{1,i}, \ldots, w_{d,i}) \). The multinomial logistic function known also as softmax function is then defined as

\[
p_\ell(x^TW) = \frac{e^{\langle x, w_\ell \rangle}}{\sum_{k=1}^{m} e^{\langle x, w_k \rangle}} \quad \text{and} \quad q(x^TW) = 1 - \sum_{i=1}^{m-1} p_\ell(x^TW)
\]

for \( \ell = 1, \ldots, m - 1 \). Let also \( p = (p_1, \ldots, p_m) \).

Following our derivation for binary labels, we can prove the following result.
Theorem 2 Let \(x_t \in [0, 1]^d\) for the label alphabet \(\mathcal{Y}\) be of size \(m\), and for \(W = [w_1, \ldots, w_{m-1}]\) we define \(p_\ell(x^TW)\) for \(\ell = 1, \ldots, m - 1\) as in (20). Then the maximal minimax regret becomes asymptotically for \(md = o(T^{1/3})\)

\[
r^*(x^T) = \frac{d(m-1)}{2} \log \frac{T}{2\pi} + \log \left( \int_{\mathbb{R}^{d(m-1)}} \sqrt{\det(B(W))} dw_1 \cdots dw_{m-1} \right) + O((md)^{3/2}/\sqrt{T})
\]

where \(B_{d,m}(W)\) is a \(d(m-1) \times d(m-1)\) matrix defined as

\[
B(W) = \frac{1}{T} \sum_{i=1}^{T} \left[ \text{Diag}(p(x_i^TW)) - p(x_i^TW) \otimes p(x_i^TW) \right] \otimes x_t \otimes x_t.
\]

3.3. Spherical Features

Now we assume that the feature \(x^t\) are either uniformly distributed on a \(d\)-dimensional sphere \(S_d\) or inside a \(d\)-dimensional ball \(B_d\) for large \(d\). We explain our ideas on \(x_t\) distributed uniformly on the sphere \(S_d\) of radius 1. By (19) we know that

\[
B(w, S_d) = \frac{1}{s_d} \int_{S_d} p(\langle xw \rangle)q(\langle xw \rangle)x \otimes x dx.
\]

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

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 \(u = w/\|w\|\).

(i) We have the following expression

\[
\bar{B}(w) = \mu(w)I_{d-1}(u) + \lambda(w)u \otimes u
\]

where \(I_d\) is the identity operator orthogonal to \(u\) and

\[
\lambda(w) = \frac{s_{d-1}}{s_d} \int_0^\pi \cos(\theta)^2 \sin(\theta)^{d-2} f(\cos(\theta)\|w\|) d\theta
\]

and

\[
\mu(w) = \frac{s_{d-1}}{s_d} \int_0^\pi \frac{\sin(\theta)^{d}}{d-1} f(\cos(\theta)\|w\|) d\theta
\]

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

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

\[
\det(\bar{B}(w)) = 2 \left( \frac{s_{d-1}}{3s_d} \pi^2 \|w\|^{-3} (1 + O(\|w\|^{-2})) \right)^d
\]

for large \(\|w\| \to \infty\).

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 $S_d$ or ball $B_d$. Then the corresponding minimax regrets satisfy (16) with the constants $C_d$

\[
C_d(S_d) = -\frac{d}{2} \log \frac{d}{4} + \frac{d}{4} \log(\pi/8) + \frac{3}{8} \log e + O(1/d) \quad (27)
\]

and

\[
C_d(B_d) = -\frac{d}{2} \log \frac{d}{4} + \frac{d}{4} \log(\pi/8) - \frac{1}{8} \log e + O(1/d) \quad (28)
\]

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 $S_d$

\[
r^*(S_d) = \frac{d}{2} \log \frac{T}{d} - \frac{d}{2} \log \sqrt{8\pi} + \frac{3}{8} \log e + O(d^{3/2}/\sqrt{T})
\]

and for the features inside a $d$ dimensional ball $B_d$ we find

\[
r^*(B_d) = \frac{d}{2} \log \frac{T}{d} - \frac{d}{2} \log \sqrt{8\pi} - \frac{1}{8} \log e + O(d^{3/2}/\sqrt{T})
\]

for large $d$.

4. Analysis

4.1. Proof of Theorem 1

Let $A_T$ be the set of achievable partial sums of the vectors $x_t$, i.e.,

\[
A_T := \{A : \exists y^T \in \{-1, 1\}^T : A_T(y^T) = A\}
\]

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

\[
F_T(w) = \sum_{y^T} e^{(w,A_T)} = \prod_t (1 + e^{(w,x_t)}) = \exp(L_T(w))
\]

which can also be written as

\[
F_T(w) = \int \rho_T(A)e^{(w,A)}dA \quad \text{with} \quad \rho_T(w) = \sum_{\mathbf{A} \in A_T} N(A) \delta_A,
\]

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

\[
S_T(x^T) = \sum_{\mathbf{A} \in A_T} N_T(A) \exp(-L(w^*(A)) + \langle w^*(A), A \rangle)
\]

that we evaluate asymptotically for large $T$. 

10
We now express (31) as
\[ S_T(x^T) = \int \rho_T(A)K(A)dx, \quad \text{where} \quad K(A) = \exp(-L(w^*(A)) + \langle w^*(A), A \rangle). \quad (32) \]

Since \( w^*(A) \) is the inverse of function \( G_T \), which is in \( C^\infty \), we conclude that \( K(A) \) is in \( C^\infty \) and has a finite support contained in \([ -T, T]^d \). Let
\[ K^*(w) = \int_{R^d} K(A)e^{i(A,w)}dA \]
be the Fourier transform of function \( K(A) \). Perceval theorem tells us that
\[ S_T(x^T) = \frac{1}{(2\pi)^d} \int_{c+R^d} F_T(iw)K^*(-w)dw. \quad (33) \]

Therefore,
\[ K^*(w) = \int_{R^d} \exp(-L_T(w^*(A)) + \langle (w^*(A) + iw), A \rangle)dA. \]

By change of variable \( A = G(w^*) \) we arrive at
\[ K^*(w) = \int_{R^d} \exp(-L_T(w^*) + \langle (w^* + iw), G(w^*) \rangle)\det(\nabla G(w^*))dw^* \quad (34) \]
leading to
\[ S_T(x^T) = \frac{1}{(2\pi)^d} \int_{R^d} \exp(-L_T(w^*) + \langle w^*, G_T(w^*) \rangle)\det(\nabla G_T(w^*))dw^*
\cdot \int_{-ic+R^d} \exp(L_T(iw) - i\langle w, G_T(w^*) \rangle)dw. \quad (35) \]

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 \(-ic + R^d\) to \(-iw^* + R^d\), finding
\[ S_T(x^T) = \frac{1}{(2\pi)^d} \int_{R^d} \exp(-L_T(w^*) + \langle w^*, G_T(w^*) \rangle)\det(\nabla G_T(w^*))dw^*
\cdot \int_{-iw^*+R^d} \exp(L_T(iw) - i\langle w, G_T(w^*) \rangle)dw. \quad (36) \]

Finally, we notice that on the segment \( \Im(w) = iw^* \) the quantity \( L_T(iw) - i\langle w, G_T(w^*) \rangle \) attains its maximum at \( w = -iw^* \), since
\[ \nabla(L_T(iw) - i\langle w, G_T(w^*) \rangle) = iG_T(iw) - iG_T(w^*) \]
is zero when \( iw = w^* \). Hence, for \( x \to 0 \) we conclude
\[ L_T(w^* + ix) - \langle (w + ix), G_T(w^*) \rangle = L_T(w^*) - \langle (w, G_T(w^*)) \rangle
\]
where \( R(w^*) \) is the third derivative of \( L_T(w) \) on \( w^* \). But \( \nabla^2 L_T(w) \) and \( R(w^*) \) 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

\[
\int_{\mathbb{R}^d} \exp \left( L_T(w) - \langle w, G_T(w^*) \rangle \right) dw = \frac{\exp(L_T(w^*) - \langle w^*, G_T(w^*) \rangle)}{\sqrt{\det(\nabla^2 L_T(w^*)/(2\pi))}} \times \left( 1 + O \left( \frac{d^{3/2}}{\sqrt{T}} \right) \right).
\]

(37)

(38)

The error term follows from Inglot and Majewski (2014) using

\[
\int_{\mathbb{R}^d} |x|^3 \exp(-a|x|^2) = \frac{1}{a^{(d+3)/2} \pi^{d/2} \Gamma(d/2)} \frac{1}{\Gamma(d/2)} \Gamma(d/2 + 3/2) \Gamma(d/2)
\]

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 \( \lambda(w) \) and \( \mu(w) \), the eigenvalues of matrix \( \bar{B}(w) \) for large \( d \). The main contribution of both integrals is for \( \theta \) around \( \pi/2 \). For \( \theta = \pi/2 + \sqrt{2} d^2 x \) we have

\[
\begin{align*}
\sin(\theta)^{d-2} & \sim \exp\left( -\frac{d-2}{2} x^2 \right) \\
\cos(\theta)^2 & \sim \frac{2}{\pi^2} x^2 \\
f(\cos(\theta)||w||) & \sim \frac{1}{4} \exp\left( -\frac{||w||^2}{2d} x^2 \right)
\end{align*}
\]

(40)

leading to

\[
\int_{0}^{\pi} \sin(\theta)^{d-2} \cos(\theta)^2 f(\cos(\theta)||w||) d\theta \sim (2/d)^{3/2} \int_{-\infty}^{+\infty} \exp\left( -\frac{1}{2} \frac{||w||^2 - 4}{2d} x^2 \right) x^2 dx
\]

\[
= (2/d)^{3/2} \frac{\sqrt{\pi}}{2} \left( 1 + \frac{||w||^2 - 4}{2d} \right)^{-3/2}
\]

(41)

and

\[
\int_{0}^{\pi} \sin(\theta)^d f(\cos(\theta)||w||) d\theta \sim (2/d)^{1/2} \int_{-\infty}^{+\infty} \exp\left( -\frac{1}{2} \frac{||w||^2}{2d} x^2 \right) dx
\]

\[
= (2/d)^{1/2} \sqrt{\pi} \left( 1 + \frac{||w||^2}{2d} \right)^{-1/2}.
\]

(42)

In summary

\[
\det(\mathbf{B}(w)) \sim \left( \frac{8d-1}{8d} \right)^d (2/d)^{d/2} \frac{1}{d} \frac{1}{(d-1)^{d-1}} \pi^{d/2} \exp\left( -\frac{||w||^2}{4} \right).
\]

(43)
To complete the derivation, we need to integrate $\sqrt{\det(B(w))}$ over the vectors $w$. This leads to

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

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

$$\frac{s_{d-1}}{s_d} = \sqrt{\pi} \frac{\Gamma(d/2)}{\Gamma((d-1)/2)},$$

$$\frac{\Gamma(d/2)}{\Gamma((d-1)/2)} = \sqrt{\frac{d}{2}} - \frac{1}{4} + O(1/d)$$

so that

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

which completes the derivation of (27).

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

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

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


MINIMAX REGRET FOR LOGISTIC REGRESSION


A. Rakhlin and K. Sridharan. Sequential probability assignment with binary alphabet and large class of experts, 2015.


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 \( u = w/|w| \).

(i) We have the following expression

\[
\tilde{B}(w) = \mu(w)I_{d-1}(u) + \lambda(w)u \otimes u
\]

where \( I_d \) is the identity operator orthogonal to \( u \) and

\[
\lambda(w) = \frac{s_{d-1}}{s_d} \int_0^\pi \cos(\theta)^2 \sin(\theta)^{d-2} f(\cos(\theta)|w|)d\theta
\]

and

\[
\mu(w) = \frac{s_{d-1}}{s_d} \int_0^\pi \sin(\theta)^d f(\cos(\theta)|w|)d\theta
\]

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

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

\[
\det(\tilde{B}(w^*)) = 2 \left( \frac{s_{d-1}}{3s_d} \pi^2 |w|^{-3} (1 + O(|w|^{-2})) \right)^d
\]

for large \( |w| \to \infty \).

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

\[
\tilde{B}(w) = \frac{1}{s_d} \int_0^\pi f(|w| \cos(\theta))d\theta \int_{\sin \theta S_{d-1}(u)} (b + \cos \theta u) \otimes (b + \cos \theta u)db
\]

\[
= \frac{1}{s_d} \int_0^\pi f(|w| \cos(\theta))d\theta \int_{\sin \theta S_{d-1}(u)} (b \otimes b + (\cos \theta)^2 u \otimes u)db
\]

\[
+ \frac{1}{s_d} \int_0^\pi f(|w| \cos(\theta))d\theta \int_{\sin \theta S_{d-1}(u)} \cos \theta (b \otimes u + u \otimes b)db.
\]

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

\[
\tilde{B}(w) = \frac{1}{s_d} \int_0^\pi f(|w| \cos(\theta))d\theta \int_{\sin \theta S_{d-1}(u)} (b \otimes b + (\cos \theta)^2 u \otimes u)db
\]

\[
= \frac{1}{s_d} \int_0^\pi f(|w| \cos(\theta))(\sin \theta)^{d-1}d\theta
\]

\[
\int_{S_{d-1}(u)} ((\sin \theta)^2 b \otimes b + (\cos \theta)^2 u \otimes u)db.
\]
The \((\sin \theta)^{d-1}\) factor arises from the change of integration domain from \(\sin \theta S_{d-1}(u)\) to \(S_{d-1}(u)\).

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

\[
\int_{S_{d-1}} (b_i)^2 db = \int_{S_{d-1}} (b_j)^2 db = \frac{1}{d-1} \int_{S_{d-1}} \|b\|^2 db = \frac{s_{d-1}}{d-1}.
\]

Thus

\[
\int_{S_{d-1}(u)} b \otimes b db = \frac{s_{d-1}}{d-1} I_{d-1}(u)
\]

and similarly

\[
\int_{S_{d-1}(u)} u \otimes u db = s_{d-1} u \otimes u
\]

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

Now we move to part (ii) of Lemma 3. Both \(\lambda(w)\) and \(\mu(w)\) are functions of \(w = \|w\|\). We write \(\lambda(w) = \lambda(\|w\|)\) and \(\mu(w) = \mu(\|w\|)\). 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 refereed 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

\[
\begin{align*}
\lambda^*(s) &= \int_0^\infty \lambda(w) w^{s-1} dw, \\
\mu^*(s) &= \int_0^\infty \mu(w) w^{s-1} dw.
\end{align*}
\]

Observe now that

\[
\lambda(w) = 2^{s_{d-1}} \frac{s_{d}}{d} \int_0^{\pi/2} f(\cos(\theta)w) \cos^2(\theta) \sin^{d-2}(\theta) d\theta
\]

\[
= 2^{s_{d-1}} \frac{s_{d}}{d} \int_0^1 y^2 (1 - y^2)^{(d-3)/2} f(yx) dy
\]

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

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

\[
= 2^{s_{d-1}} \frac{s_{d}}{d} f^*(s) \int_0^1 (1 - y^2)^{(d-3)/2} y^{3-s} dy
\]

\[
= f^*(s) \beta_1^*(3 - s)
\]

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

\[
f^*(s) = (s - 1)^2(s - 2) h^*(s - 2)
\]

17
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

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

(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 $-\zeta(2)\Gamma(3)\frac{3-1}{2} = -\pi^2/6(d - 1)$. 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))

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

(64)

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

$$\begin{align*}
\lambda(w) &= 2\frac{s-1}{3s-1}\pi^2 w^{-3} + O(w^{-5}), \\
\mu(w) &= \frac{s-1}{3s-1}\pi^2 w^{-3} + O(w^{-5})
\end{align*}$$

(65)

for $w \to \infty$. In conclusion

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

(66)

when $\|w\| \to \infty$. This completes the proof of Lemma 3.

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

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

(67)

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

$$\begin{align*}
\mu(w) &= \frac{1}{v_d} \int_0^\pi \sin(\theta)^d f(\cos(\theta)\|w\|) d\theta \int_{\|b\|=1} \langle vb \rangle^2 db, \\
&= \frac{1}{v_d} \int_0^\pi \sin(\theta)^d f(\cos(\theta)\|w\|) d\theta \int_1^{s_d-1} \frac{r^2}{d-1} r^{d-2} dr \\
&= \frac{v_d-1}{v_d} \int_0^\pi \sin(\theta)^d \left(\frac{r^2}{d-1} r^{d-2} \right) d\theta.
\end{align*}$$

(68-70)
Using $\sqrt{\det(B(w))} = \sqrt{\lambda(w)(\mu(w))^{d-1}}$, all computations done with an extra factor \( \left( \frac{d}{d-1} \right)^{d/2} \left( \frac{d-1}{d+1} \right)^{d-1/2} \), we find at the end

$$
\int \sqrt{\det(B(w))} \, dw \sim (\pi/8)^{d/4} e^{-1/8 (d/4) - d/2}
$$

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

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

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

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

$$
\int_{\Omega(w)=-w^*} \exp \left( L_T(w) - \langle w, G_T(w^*) \rangle \right) \, dw = \frac{\exp(L_T(w^*) - \langle w^*, G_T(w^*) \rangle)}{(2\pi)^d \sqrt{\det(\nabla^2 L_T(w^*))}} \times \int_{\mathbb{R}^d} \exp \left( -\frac{1}{2} \|x\|^2 + O(L_T^{(3)}(\|\sqrt{\nabla^2 L_T(w^*)}^{-1}x\|^3)) \right)
$$

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

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

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

$$
\int_{\Omega(w)=-w^*} \exp \left( L_T(w) - \langle w, G_T(w^*) \rangle \right) \, dw = \frac{\exp(L_T(w^*) - \langle w^*, G_T(w^*) \rangle)}{(2\pi)^d \sqrt{\det(\nabla^2 L_T(w^*))}} \times \int_{\mathbb{R}^d} \exp \left( -\frac{1}{2} \|x\|^2 \right) \left( 1 + O\left( \frac{\|x\|^3}{\sqrt{T \lambda(w^*)^{3}}} \right) \right).
$$

But we know that for $a > 0$

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

thus we conclude that

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

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

---

19