Precise Minimax Regret for Logistic Regression with Categorical Feature Values

Philippe Jacquet PHILIPPE.JACQUT@INRIA.FR
INRIA
Paris, France

Gil I. Shamir GSHAMIR@IEEE.ORG
Google Inc
Pittsburgh, PA, USA

Wojciech Szpankowski SZPAN@PURDUE.EDU
Department of Computer Science
Purdue University
West Lafayette, IN, USA

Editors: Vitaly Feldman, Katrina Ligett and Sivan Sabato

Abstract

We study logistic regression with binary labels and categorical (discrete) feature values. Our goal is to evaluate precisely the (maximal) minimax regret. We express it as the so called Shtarkov sum known in information theory. To the best of our knowledge such a sum was never computed in the context of logistic regression.

To be more precise, the pointwise regret of an online algorithm is defined as the (excess) loss it incurs over some value of a constant comparator (weight vector) that is used for prediction. 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 possible learning algorithms/ distributions, therefore it constitutes a lower bound for the pointwise regret. For finite dimension d and N distinct feature vectors we show that the maximal minimax regret grows as

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

where T is the number of rounds of running a training algorithm and C_d is explicitly computable constant that depends on the feature values and dimension d. We also extend these results to non-binary labels. The *precise* maximal minimax regret presented here is the first result of this kind. Our findings are obtained using tools of analytic combinatorics and information theory.

1. Introduction

Logistic regression has been important in theory and practice of modern machine learning. It has been used for tasks, such as, category classification, click-through-rate prediction, and risk assessment. A model consists of a set of features, whose parameters represent their effect on some outcome. In an online supervised setup, such a model is trained to learn these parameters from examples whose outcomes are already labeled. The training algorithm consumes data in rounds, where at each round $t = 1, \ldots, T$, it is allowed to predict the label based only on the labels it observed in the past $t - 1$

rounds. In each round, the prediction algorithm incurs some *loss* and updates its belief of the model parameters. The *pointwise regret* (for all sequences) of an online algorithm is defined as the (excess) loss it incurs over some value of a constant *comparator* (weight vector) that is used for prediction for the complete sequence. The pointwise regret for logistic regression with a Bayesian learning algorithm has been studied in Foster et al. (2018); Hazan et al. (2014); Kakade and Ng (2005); McMahan and Streeter (2012); Shamir (2020).

In this paper, we introduce the *maximal minimax regret* that for a given feature sequence maximizes pointwise regret over label sequences and minimizes over all learning distributions that best approximate the label sequence (see also Shamir and Szpankowski (2021)). We express it as the so called Shtarkov sum, as in Shtarkov (1987), that we evaluate asymptotically. We study the minimax regret using methods outside traditional machine learning toolbox, namely analytic combinatorics (see Szpankowski (2001); Flajolet and Sedgewick (2008)) and universal compression (see Shtarkov (1987); Drmota and Szpankowski (2004); Szpankowski (1998); Szpankowski and Weinberger (2012); Xie and Barron (1997, 2000)).

For a start, we review various notions of regret and redundancy from information theory that we adopt for the performance evaluation of logistic regression. The *pointwise redundancy* $R_T(P; y^T)$ and the *average redundancy* $\bar{R}_T(P)$ for a *given* source P (distribution) and source (label) sequence $y^T = (y_1, \dots, y_T)$ of length T (over alphabet of size m or in ML language over sequences with m distinct label values) are defined as

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

where $H_T(P)$ is the entropy of P, $\mathbf{E}$ denotes the expectation, and $L(y^T)$ is the code length of some code $L(\cdot)$. In the online learning – and indeed in information theory – one ignores the integer nature of the length (however, see Drmota and Szpankowski (2004)) and replace it by $L(y^T) = -\log Q(y^T)$ for some distribution Q that best approximates P. The above definitions imply a probabilistic setting, in which there is some source that generated the data. A non-probabilistic setting considers *individual sequences* (see, e.g., Shtarkov (1987)), where the *maximal* redundancy is defined as

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

which somewhat decouples it from modeling assumptions, as pointed out by Rissanen (1978, 1996).

In universal learning and compression, it is assumed that we have some knowledge about a family of sources $\mathcal{S}$ generating real data. Following Davisson (1973), in information theory the average minimax redundancy $\bar{R}_T(\mathcal{S})$ and the maximal minimax redundancy $R_T^*(\mathcal{S})$ for family $\mathcal{S}$ are defined as follows

$$\bar{R}_T(\mathcal{S}) = \min_{Q} \sup_{P \in \mathcal{S}} \sum_{y^T} P(y^T) \log[P(y^T)/Q(y^T)],$$

$$R_T^*(\mathcal{S}) = \min_{Q} \sup_{P \in \mathcal{S}} \max_{y^T} \left[\log(P(y^T)/Q(y^T))\right].$$

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

There are other measures of optimality for learning, coding, gambling, and prediction that are used in universal modeling and machine learning. We refer here to minimax *regrets* defined as follows (cf. Drmota and Szpankowski (2004); Xie and Barron (1997, 2000)):

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

and to the maxmin regret

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

We call $\bar{r}_T(\mathcal{S})$ the *average* minimax regret, $r^*_T(\mathcal{S})$ the maximal minimax regret and $\underline{r}_T(\mathcal{S})$ the maxmin regret. It is easy to see that $\bar{R}_T(\mathcal{S}) \le \bar{r}_T(\mathcal{S})$, and, $r^*_T(\mathcal{S}) = R^*_T(\mathcal{S})$. For more sophisticated relation between various regrets and redundancy see Drmota and Szpankowski (2004).

In this paper we focus on analyzing the maximal minimax regret for logistic regression minimized over all learning distribution/ algorithms with categorical (discrete) feature values, that is, we assume there are N distinct feature vectors with each feature taking finite number of values over a finite alphabet. In Theorem 1 we show that the maximal minimax regret of dimension $d = O(1)$ for categorical feature values grows asymptotically as

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

where C_d is a constant that depends on the feature values. For example, for $d = 1$ with features values $\{a_1, \ldots, a_N\}$ we find in Corollary 2

$$C_1 = \log\left(\int_{-\infty}^{\infty} \sqrt{\sum_{j=1}^{N} a_j^2 \alpha_j (1+e^{-a_j w})^{-1}(1+e^{a_j w})^{-1} dw}\right)$$

where α_j is the fraction of T rounds that feature value a_j is applied. This seems to be the first *precise* result of this kind in the area of logistic regret. In Theorem 3 we extend these results to non-binary labels. We should point out that the maximal minimax regret constitutes a lower bound of the pointwise regret for any specific algorithm and label sequences.

We now briefly review relevant literature of information theory and machine learning. We start with information theory assuming that the size of the underlying alphabet is m (and effectively assuming $d = 1$). In Drmota and Szpankowski (2004); Orlitsky and Santhanam (2004); Rissanen (1996); Shamir (2006); Szpankowski (1998); Xie and Barron (1997, 2000) it was proved that for a large class of sources (up to Markovian but not for non-Markovian as shown in Csiszar and Shields (1995); Flajolet and Szpankowski (2002); Drmota and Szpankowski (2004)) the redundancy grows as $\frac{m}{2}\log T + O(1)$ when m is fixed and $\frac{m}{2}\log(T/m)$ for $m = o(T)$ (see also Orlitsky and Santhanam (2004); Shamir (2006)). A full asymptotic expansion for the regret and redundancy for the whole range of m are derived in Szpankowski and Weinberger (2012).

Regarding the online convex optimization literature, logarithmic regret has been shown for strongly convex loss functions. Logistic regression, however, fell in the category of weakly convex loss functions, for which $O(\sqrt{T})$ regret bounds have been shown. In most machine learning literature, the feature values are assumed to belong either to the interval $[0, 1]$ or are binary $\{0, 1\}$ (active or passive). To the best of our knowledge, Kakade and Ng (2005) are first to demonstrate results that suggest $O(d\log T/d)$ regret for logistic regression, using *Bayesian model averaging*. As mention before, the redundancy results we described from the information theory literature apply to the single dimensional binary labels logistic regression problem. Similar $O(\log T)$ 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 problem 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), who show that the pointwise regret is $d/2\log(T/d)+\log\log T$ for $d=o(\sqrt{T})$, again with Bayesian averaging. The worst case minimax regret was studied in a series of papers by Rakhlin and Sridharan (2014) using Rademacher complexity rather than Shtarkov sum approach. Here, we analyze precisely the maximal minimax regret for individual sequences and discrete feature values over a class of learning algorithms/ distributions (not necessary Bayesian).

2. Problem Formulation and Notation

We denote by $\mathbf{x}_t=(x_{1,t},\ldots,x_{d,t})^\tau$ a d-dimensional column feature vector where τ denotes transpose operator. Notice that $\mathbf{x}^T$ is a $T\times d$ matrix with $\mathbf{x}_t=(x_{1,t},\ldots,x_{d,t})$ as a row. The label binary vector is denoted as $y^T=(y_1,\ldots,y_T)$ with $y_t\in\{-1,1\}$. The vector $\mathbf{w}_t=(w_{1,t},\ldots,w_{d,t})^\tau$ representing d-dimensional weights is used to design a prediction algorithm, which we will not discuss here. Furthermore, we assume that the feature vector $\mathbf{x}_t$ takes only finite number of (vector) values, that is, we set $\mathbf{x}_t=\mathbf{a}_j$ for $j=1,\ldots,N$ where $\mathbf{a}_j=(a_{1,j},\ldots,a_{d,j})^\tau$ with $a_{ij}\in\mathcal{A}$ for $i=1,\ldots,d$, $j=1,\ldots,N$, and some finite alphabet $\mathcal{A}$. For example, for $d=1$ we simply have $x_t\in\mathcal{A}=\{a_1,\ldots,a_N\}$ for all t. Finally, by T_j we denote the number of t such that $\mathbf{x}_t=\mathbf{a}_j$ where

$$T_1+\cdots+T_N=T.$$

For $T_j>0$ we also write $\alpha_j=T_j/T$.

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

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

where $\langle\mathbf{x}_t,\mathbf{w}_t\rangle=\sum_{i=1}^{d}x_{i,t}w_{i,t}$. In our case, we can re-write $L(y^T|\mathbf{x}^T,\mathbf{w}^T)$ as

$$L(y^T|\mathbf{x}^T,\mathbf{w}^T)=\sum_{j=1}^{N}\log\prod_{i=1}^{T_j}\left[1+\exp(-y_{t_{j_i}}\langle\mathbf{a}_j,\mathbf{w}_{t_{j_i}}\rangle)\right]$$

where t_{j_i} is a subsequence of $t=1,\ldots,T$ such that $\mathbf{x}_{t_{j_i}}=\mathbf{a}_j$.

It is convenient to write $\ell(y_t|\mathbf{x}_t,\mathbf{w}_t):=\log\left[1+\exp(-y_t\langle\mathbf{x}_t,\mathbf{w}_t\rangle)\right]$. Both $\ell(y_t|\mathbf{x}_t,\mathbf{w}_t)$ and $L((y^T|\mathbf{x}^T,\mathbf{w}^T)$ depend on $\mathbf{x}_t$ and $\mathbf{w}_t$ only through the product $\langle\mathbf{x}_t,\mathbf{w}_t\rangle$. Notice that for binary labels, the probability of a label is given by

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

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

The goal of a learning algorithm is to find the best approximation $Q(y_t|\mathbf{x}_t)$ of the unknown distribution $P(y_t|\mathbf{x}_t, \mathbf{w}_t)$. Hence, we also denote $\ell(Q, y_t|\mathbf{x}_t) = -\log Q(y_t|\mathbf{x}_t)$. The *pointwise regret* for all sequences $(y_t, \mathbf{x}_t)$ is defined as in Hazan (2012); Foster et al. (2018); Shamir (2020)

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

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

$$R_T(Q, y^T|\mathbf{x}^T) = \log \frac{\sup_{\mathbf{w}} P(y^T|\mathbf{x}^T, \mathbf{w})}{Q(y^T|\mathbf{x}^T)}. \tag{3}$$

In our setting we have

$$P(y^T|\mathbf{x}^T, \mathbf{w}) = \prod_{j=1}^{N} \left(\frac{1}{1 + \exp(-\langle \mathbf{a}_j, \mathbf{w} \rangle)} \right)^{k_j} \cdot \left(\frac{1}{1 + \exp(\langle \mathbf{a}_j, \mathbf{w} \rangle)} \right)^{T_j - k_j} \tag{4}$$

where, we recall, T_j is the number of rounds with $\mathbf{a}_j$ feature vector, and k_j is the number of $y_t = 1$ among T_j rounds. Expression (4) is a consequence of the discrete nature of feature values.

The pointwise regret $R_T(Q, y^T|\mathbf{x}^T)$ is a function of learning algorithm Q, y_t and $\mathbf{x}_t$, so it depends on *individual* sequences (e.g., see Kakade and Ng (2005); Shamir (2020)). A better measure of the learning algorithm performance should decouple the regret from the learning distribution Q (so it can provide a universal lower bound for all algorithms) and the fluctuations of y^T (but may still depend on the feature vector $\mathbf{x}^T$). Following information-theoretic view as in Shtarkov (1987); Drmota and Szpankowski (2004), we define the *maximal minimax regret* (conditioned on $\mathbf{x}^T$) as

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

Notice that this definition is over all possible learning distributions represented by Q. Therefore, we have

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

and it represents a general lower bound for the pointwise regret predominately studied in the machine learning community. For example, in the Bayesian framework one sets

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

where $\rho(\mathbf{w})$ represents a prior. In this paper, we do not restrict ourselves to Bayesian algorithms.

We first find a more succinct representation of the maximal minimax regret. Then, noting that $\ell(y_t|\mathbf{x}_t, \mathbf{w}) = -\log P(y_t|\mathbf{x}_t, \mathbf{w})$, and following Shtarkov (1987); Drmota and Szpankowski (2004) we find

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

where

$$P^*(y^T|\mathbf{x}^T) := \frac{\sup_{\mathbf{w}} P(y^T|\mathbf{x}^T, \mathbf{w})}{\sum_{v^T} \sup_{\mathbf{w}} P(v^T|\mathbf{x}^T, \mathbf{w})} \tag{5}$$

is the *maximum-likelihood distribution* and we set $Q(y^T|\mathbf{x}^T) = P^*(y^T|\mathbf{x}^T)$. In summary

$$r_T^*(\mathbf{x}^T) = \log \sum_{y^T} \sup_{\mathbf{w}} P(y^T|\mathbf{x}^T, \mathbf{w}) =: \log d_T(\mathbf{x}^T). \tag{6}$$

Observe that for not optimal $Q \neq P^*$ there will be extra $O(1)$ term in the maximal minimax regret as discussed in Drmota and Szpankowski (2004). In passing, we point out that Rakhlin and Sridharan (2014) proposed a slightly different approach to the (worst case) minimax regret (i.e., $\max_{\mathbf{x}^T} r_T^*(\mathbf{x}^T)$) and studied it using the Rademacher complexity.

The sum $\log d_T(\mathbf{x}^T)$ in (6) is often called the *Shtarkov sum* as in Drmota and Szpankowski (2004); Grunwald (2007). To the best of our knowledge the Shtarkov sum was never evaluated in this context. The goal of this paper is exactly to do this asymptotically for categorical feature values, up to $o(1)$ term in order to show the impact of the feature values on the minimax regret. As we shall see the feature values emerge only in the second term of the asymptotic expansion (see Theorem 1 and Theorem 3).

3. Main Results

In this section we present our main results. We use the notation from the previous section, and in addition, we write

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

Our goal is the estimate asymptotically the Shtarkov sum

$$d_T(\mathbf{x}^T) = \sum_{y^T} \sup_{\mathbf{w}} P(y^T|\mathbf{x}^T, \mathbf{w}),$$

where, simplifying (4), we arrive at

$$\begin{aligned} P(y^T|\mathbf{x}^T, \mathbf{w}) &= \prod_{j=1}^{N} p(\langle \mathbf{a}_j \mathbf{w} \rangle)^{k_j} q(\langle \mathbf{a}_j \mathbf{w} \rangle)^{T_j - k_j} \\ &= \frac{1}{\prod_{j=1}^{N} (1 + e^{\langle \mathbf{a}_j \mathbf{w} \rangle})^{T_j}} \cdot \exp\left(\sum_{j=1}^{N} k_j \langle \mathbf{a}_j \mathbf{w} \rangle \right) \end{aligned}$$

with k_j being the number of $y_t = 1$ in T_j rounds that use feature vector $\mathbf{a}_j$.

Maximizing $P(y^T|, \mathbf{x}^T, \mathbf{w})$ with respect to $\mathbf{w}$ leads to $\mathbf{w}^* = \mathbf{w}$ satisfying

$$\sum_{j=1}^{N} \mathbf{a}_j p(\langle \mathbf{a}_j \mathbf{w} \rangle) T_j = \sum_{j=1}^{N} \mathbf{a}_j k_j \tag{7}$$

where we use the fact $p'(w) = p(w)q(w)$. Notice that the above is a system of d linear equations, thus the tuple that share the same optimal value $\mathbf{w}^*$ are in the intersection of d hyperplanes $\mathcal{H}_1(\mathbf{w}^*)$, $\mathcal{H}_2(\mathbf{w}^*), \dots, \mathcal{H}_d(\mathbf{w}^*)$ where

$$\mathcal{H}_i(\mathbf{w}) = \{k^N = (k_1, \dots, k_N) : \sum_{j=1}^{N} a_{i,j}(k_j - p(\langle \mathbf{a}_j \mathbf{w} \rangle)T_j) = 0\} \tag{8}$$

where $i = 1, \dots, d$. For convenience we denote by $\mathcal{H}^d(\mathbf{w}) = \mathcal{H}_1(\mathbf{w}) \cap \dots \cap \mathcal{H}_d(\mathbf{w})$ as a space vector of co-dimension d.

To estimate the Shtarkov sum we proceed as follows. Since the quantity $\mathbf{w}^*$ does not change when $k^N = (k_1, \dots, k_N)$ is in the hyperplane $\mathcal{H}^d(\mathbf{w}^*)$ the rule of the game will be to cut the set $[0, T_1] \times \dots \times [0, T_N]$ into parallel slices each representing the hyperplane $\mathcal{H}^d(\mathbf{w})$. This allows us to replace the summation of y^T by the following

$$d_T(\mathbf{x}^T) = \sum_{\mathbf{w}} \sum_{k^N \in \mathcal{H}^d(\mathbf{w})} B(k^N, \mathbf{w}) \tag{9}$$

where

$$B(k^N, \mathbf{w}) = \prod_{j=1}^{N} \binom{T_j}{k_j} p(\langle \mathbf{a}_, j \mathbf{w} \rangle)^{k_j} q(\langle \mathbf{a}_j, \mathbf{w} \rangle)^{T_j - k_j}. \tag{10}$$

Note that there are $\binom{T_j}{k_j}$ ways to set $y_t = 1$ among T_j positions. Since $B(k^N, \mathbf{w})$ is the product of binomial distributions it is maximized at

$$k^N(\mathbf{w}) = (p(\langle \mathbf{a}_1 \mathbf{w} \rangle)T_1, \dots, p(\langle \mathbf{a}_N \mathbf{w} \rangle)T_N)$$

that defines a manifold $\mathcal{L}$ of dimension d in the hyper-cube $[0, T_1] \times \ldots \times [0, T_N]$ of dimension N (thus we assume $d < N$).

Approximating the product of binomial distributions by normal approximations, we arrive at

$$P(y^T | \mathbf{w}^*) = \frac{\exp\left(-\sum_j \frac{(k_j - k_j(\mathbf{w}))^2}{2p(\langle \mathbf{a}_j \mathbf{w} \rangle)q(\langle \mathbf{a}_j \mathbf{w} \rangle)T_j}\right)}{\prod_j \sqrt{2\pi p(\langle \mathbf{a}_j \mathbf{w} \rangle)q(\langle \mathbf{a}_j \mathbf{w} \rangle)T_j}}(1 + O(N/\sqrt{T})),$$

where the error term comes from the rate of convergence in the central limit theorem. Written differently we have

$$P(y^T | \mathbf{a}^T, \mathbf{w}) \sim \sqrt{\det(\mathbf{A}(\mathbf{w})/(2\pi))} \exp\left(-\frac{\langle (k^N - k^N(\mathbf{w}))^\tau \mathbf{A}(\mathbf{w})(k^N - k^N(\mathbf{w})) \rangle}{2}\right)$$

where

$$\mathbf{A}(\mathbf{w}) = \mathrm{Diag}\left(\frac{1}{p(\langle \mathbf{a}_1 \mathbf{w} \rangle)q(\langle \mathbf{a}_1 \mathbf{w} \rangle)T_1}, \dots, \frac{1}{p(\langle \mathbf{a}_N \mathbf{w} \rangle)q(\langle \mathbf{a}_N \mathbf{w} \rangle)T_N}\right).$$

In order to evaluate the minimax regret $r^*(\mathbf{x}^T) = \log d_T(\mathbf{x}^T)$ we shall use the Euler-Maclaurin formula (e.g., Szpankowski (2001)) leading to

$$\begin{aligned} d_T(\mathbf{x}^T) &= \int_{[0,T_1] \times \dots \times [0,T_N]} P(y^T | \mathbf{w}^*) dk^N \\ &= \int_{\mathbf{R}^d} \delta(\mathbf{w}) dw_1 \cdots dw_d \int_{\mathcal{H}^d(\mathbf{w})} P(y^T | \mathbf{w}) dk^N, \end{aligned} \tag{11}$$

where $\delta(\mathbf{w})$ is a thickness factor that takes into account the variation of spacing between the parallel subspaces $\mathcal{H}^d(\mathbf{w})$ and counts the number of y^T between $\mathcal{H}(w)$ and $\mathcal{H}(w+dw)$. We compute it in the next section.

This allows us to formulate our main result with the detailed proof delayed till the next section

Theorem 1 *Let $\mathbf{x}_t = \mathbf{a}_j$ for $j = 1, \dots, N$ where $\mathbf{a}_j = (a_{1,j}, \dots, a_{d,j})^\tau$ with $a_{ij} \in \mathcal{A}$ for some finite alphabet $\mathcal{A}$. Define also $p(w) = (1+e^{-w})^{-1}$ with $q(w) = 1 - p(w)$. Then the maximal minimax regret becomes asymptotically for $N = o(\sqrt{T})$ and $d = O(1)$ ($d \le N$)*

$$r^*(\mathbf{x}^T) = \frac{d}{2}\log T - \frac{d}{2}\log 2\pi + \log\left(\int_{\mathbf{R}^d}\sqrt{\det(\tilde{\mathbf{B}}_d(\mathbf{w}))}dw_1\cdots dw_d\right) + O(N/\sqrt{T}) \tag{12}$$

where $\tilde{\mathbf{B}}_d(\mathbf{w})$ is a $d \times d$ matrix computed as follows

$$\tilde{\mathbf{B}}_d(\mathbf{w}) = \sum_{i=1}^{N} \alpha_i p(\langle \mathbf{a}_i \mathbf{w}\rangle) q(\langle \mathbf{a}_i \mathbf{w}\rangle)\mathbf{a}_i \mathbf{a}_i^\tau$$

that is, i, j element of $\tilde{\mathbf{B}}_d(\mathbf{w})$ is $\langle \mathbf{u}_i \tilde{\mathbf{A}}_d^{-1}(\mathbf{w})\mathbf{u}_j\rangle$ with $\mathbf{u}_i = (a_{i,1}, \cdots, a_{i,N})$ and

$$\tilde{\mathbf{A}}_d(\mathbf{w}) = \mathrm{Diag}\left(\frac{1}{p(\langle \mathbf{a}_1 \mathbf{w}\rangle)q(\langle \mathbf{a}_1 \mathbf{w}\rangle)\alpha_1}, \dots, \frac{1}{p(\langle \mathbf{a}_N \mathbf{w}\rangle)q(\langle \mathbf{a}_N \mathbf{w}\rangle)\alpha_N}\right)$$

where $0 < \alpha_i = T_i/T < 1$ and $\sum_j \alpha_j = 1$

Large d. To understand better the impact of large d on the regret, we shall use the following well known fact

$$\frac{d}{\mathrm{tr}(\tilde{\mathbf{B}}^{-1})} \le \det{}^{1/d}(\tilde{\mathbf{B}}) \le \frac{\mathrm{tr}(\tilde{\mathbf{B}})}{d}$$

where $\mathrm{tr}(\tilde{\mathbf{B}})$ is the trace of $\tilde{\mathbf{B}}$. Therefore, we find

$$r^*(\mathbf{x}^T) \le \frac{d}{2}\log\frac{T}{d} - \frac{d}{2}\log 2\pi + \log\left(\int_{\mathbf{R}^d}\sqrt{[\mathrm{tr}(\tilde{\mathbf{B}}(w)]^d}dw_1\cdots dw_d\right) + O(N/\sqrt{T})$$

which seems to be asymptotically correct on the leading term. However, it is still an open problem to find precise asymptotic regret for other ranges of d, N and T. Some recent results in this direction are reported in Shamir (2020).

Special Cases: $d = 1$. In the special case when $d = 1$ we find a simpler expression as in the corollary below.

Corollary 2 *Let $x_i \in \{a_1, \dots, a_N\}$ and $d = 1$. Then the maximal minimax regret becomes*

$$\begin{aligned} r_T^*(x^T) = \log d(x^T) &= \frac{1}{2}\log T - \frac{1}{2}\log(2\pi) + \\ &+ \log\left(\int_{-\infty}^{\infty}\sqrt{\sum_j a_j^2 p(a_j w) q(a_j w) T_j/T dw}\right) + O(N/\sqrt{T}) \end{aligned} \tag{13}$$

for large T.

In particular, when $N = 1$ and $a_1 = 1$ (or all a_i are the same) we find

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{(1+e^{-w})(1+e^{w})}} dw = \int_{-\infty}^{\infty} \frac{e^{w/2}}{1+e^{w}} dw = \pi$$

and therefore

$$r_T^*(x^T) = \frac{1}{2}\log T + \frac{1}{2}\log(\pi/2) + o(1)$$

as in Drmota and Szpankowski (2004).

Minimization of redundancy. The part of $r_T^*(x^T)$ that depends on the feature values, say for $d = 1$, is

$$\int_{-\infty}^{\infty} \sqrt{\sum_j a_j^2 p(a_j w) q(a_j w) T_j / T} dw.$$

Thanks to the concavity property of the square root function we hence have

$$\begin{aligned} \int_{-\infty}^{\infty} \sqrt{\sum_j a_j^2 p(a_j w) q(a_j w) T_j / T} dw \quad &\geq \quad \sum_j \frac{T_j}{T} \int_{-\infty}^{\infty} \sqrt{a_j^2 p(a_j w) q(a_j w)} dw \\ &= \quad \sum_j \frac{T_j}{T} \int_{-\infty}^{\infty} \sqrt{p(w)q(w)} dw = \pi. \end{aligned} \tag{14}$$

This minimum value is obtained when all a_i are the same.

3.1. Extension to non-binary labels

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

$$p_\ell(\mathbf{a}^\tau \mathbf{W}) = \frac{e^{\langle \mathbf{a}, \mathbf{w}_\ell \rangle}}{\sum_{k=1}^{m} e^{\langle \mathbf{a}, \mathbf{w}_k \rangle}} \tag{15}$$

for $\ell = 1, \ldots, m-1$ and

$$q(\mathbf{a}^\tau \mathbf{W}) = 1 - \sum_{i=1}^{m-1} p_\ell(\mathbf{a}^\tau \mathbf{W}).$$

We now only briefly describe steps needed to extend our previous analysis to the non-binary case. Recall that $\mathbf{a}_j = (a_{1,j}, \ldots, a_{d,j})^\tau$ is the j-th feature values column vector. We first observe that the binomial distribution will become the multinomial distribution, and in particular (10) is

$$B_n(\mathbf{k}^N, \mathbf{W}) = \prod_{i=1}^{N} \binom{T_i}{\mathbf{k}_i} \prod_{\ell=1}^{m-1} p_\ell(\mathbf{a}_i^\tau \mathbf{W})^{k_{j,\ell}} q(\mathbf{a}_j^\tau \mathbf{W})^{T_j - \sum_\ell k_{j,\ell}}$$

which we replace by the $(m-1)$-dimensional normal distribution with the covariance matrix

$$\Sigma_i(\mathbf{W}) = \mathrm{Diag}(\mathbf{p}(\mathbf{a}_i^\tau \mathbf{W})) - \mathbf{p}(\mathbf{a}_i^\tau \mathbf{W})\mathbf{p}^\tau(\mathbf{a}_i^\tau \mathbf{W}),$$

where $\mathbf{p}(\mathbf{a}) = [p_1(\mathbf{a}), \ldots, \mathbf{p}_{m-1}(\mathbf{a})]^\tau$ is the probability column vector. Also, as before we denote by $\mathbf{A}_i(\mathbf{W})$ the inverse of the above covariance $(m-1) \times (m-1)$ matrix.

Finally, observe that

$$P(y^T|\mathbf{a}, \mathbf{W}) = \frac{1}{\prod_{j=1}^N q(\mathbf{a}_j^\tau \mathbf{W})^{T_j}} \exp\left(\sum_{\ell=1}^{m-1} k_{j\ell}\langle \mathbf{a}_j \mathbf{w}_\ell\rangle\right).$$

From this we can obtain the system of linear equations as in (8) for every ℓ. Thus the set of optimal $\mathbf{W}$ is a hyperplane $\mathcal{H}^{(m-1)d}$ that are parallel subspaces of codimension $(m-1)d$. Therefore $\mathcal{H}^{(m-1)d}$ is orthogonal to the vectors $\mathbf{u}_{i,\ell}$ belonging to $\mathbf{R}^{(m-1)N}$. The (k,j)-th coefficient of vector $\mathbf{u}_{i,\ell}$ is $\delta_{(k=\ell)} a_{ji}$. It is convenient to represent $\mathbf{u}_{i,\ell}$ as vector in $\mathbf{R}^{(m-1)} \times \mathbf{R}^N$.

Following the footsteps of our previous derivations we arrive at the following final result.

Theorem 3 *Let $\mathbf{x}_t = \mathbf{a}_j$ for $j = 1, \ldots, N$ where $\mathbf{a}_j = (a_{1,j}, \ldots, a_{d,j})$ with $a_{ij} \in \mathcal{A}$ for some finite set $\mathcal{A}$. Furthermore, let the label alphabet $\mathcal{Y}$ be of size m, and $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_{m-1}]$. Finally, $p_\ell(\mathbf{a}^\tau \mathbf{W})$ for $\ell = 1, \ldots, m-1$ are defined in (15). Then the maximal minimax regret becomes asymptotically for $N = o(\sqrt{T})$ and $m, d = O(1)$*

$$r^*(\mathbf{x}^T) = \frac{d(m-1)}{2} \log \frac{T}{2\pi} + \log\left(\int_{\mathbf{R}^{d(m-1)}} \sqrt{\det(\tilde{\mathbf{B}}_{d,m}(\mathbf{W}))} d\mathbf{w}_1 \cdots d\mathbf{w}_{m-1}\right) + O(N/\sqrt{T}) \tag{16}$$

where $\tilde{\mathbf{B}}_{d,m}(\mathbf{W})$ is a $d(m-1) \times d(m-1)$ matrix whose $ik, j\ell$ coefficient is $\langle \mathbf{u}_{ik} \tilde{\mathbf{A}}_{d,m}^{-1}(\mathbf{W}) \mathbf{u}_{j\ell}\rangle$ with the k, j coefficient of $\mathbf{u}_{i\ell} \in \mathbf{R}^{(m-1)N}$ being $a_{ji}\delta_{k\neq\ell}$ and

$$\tilde{\mathbf{A}}_{d,m}^{-1}(\mathbf{W}) = \sum_{i=1}^N \alpha_i \left(\mathrm{Diag}\left(\mathbf{p}(\mathbf{a}_i^\tau \mathbf{W})\right) - \mathbf{p}(\mathbf{a}_i^\tau \mathbf{W})\mathbf{p}^\tau(\mathbf{a}_i^\tau \mathbf{W})\right)$$

where $0 < \alpha_j = T_j/T < 1$ and $\sum_j \alpha_j = 1$

4. Analysis

In this this section we prove our main result Theorem 1. However, we split the analysis in two parts. First we prove Corollary 2 for $d = 1$ to illustrate our methods. Then we provide missing parts of the proof of Theorem 1.

4.1. Proof of Corollary 2

Assume now that $d = 1$, We notice that the quantity we want to maximize is the product of the binomial coefficients

$$B_j(k_j, w) = \binom{T_j}{k_j} p(a_j w)^{k_j} q(a_j w)^{T_j - k_j}.$$

The aim is to maximize each coefficients separately. The maximum (w and the T_j fixed) is attained by k_j which is the closest to

$$k_j(w) = p(a_j w) T_j.$$

Thanks to the asymptotic properties of the binomial distribution the maximum is attained asymptotically at

$$\max B_j(k_j, w) = (2\pi p(a_j w) q(a_j w) T_j)^{-1/2}(1 + O(1/\sqrt{T_j})).$$

Furthermore if k_j is close to $k_j(w) \in \mathcal{H}(w^*)$ (e.g., typically $k_j = k_j(w) + O(\sqrt{T_j} \log T_j)$), we have by virtue of the normal limit of the binomial distribution:

$$\binom{T_j}{k_j} p(a_j w)^{k_j} q(a_j w)^{T_j - k_j} = \frac{1}{\sqrt{2\pi p(a_j w) q(a_j w) T_j}} \exp\left(-\frac{(k_j - k_j(w))^2}{2p(a_j w)q(a_j w)}\right)(1 + O(1/\sqrt{T_j}). \tag{17}$$

We shall use the following known lemma to justify (17) (e.g., see Szpankowski (2001)).

Lemma 4 *Let $p_n(k) = \binom{n}{k} p^k q^{n-k}$ where $q = 1 - p$ be the binomial distribution. Then for $|k - pn| \leq n^{1/2+\varepsilon}$ we have*

$$p_n(k) = \frac{1}{\sqrt{2\pi p(1-p)n}} \exp\left(-\frac{(k-pn)^2}{2p(1-p)n}\right) + O(n^{-\delta}) \tag{18}$$

uniformly as $n \to \infty$. Furthermore

$$\sum_{|k-np| > \sqrt{p(1-p)} n^{1/2+\varepsilon}} p_n(k) < 2n^{-\varepsilon} e^{-n^{2\varepsilon}/2} \tag{19}$$

for large n.

Let us now evaluate the Shtarkov sum. Using the Euler-Maclaurin formula we have

$$d(x^T) = \int_{-\infty}^{\infty} \delta(w) \int_{\mathcal{H}(w)} B(k^N, w) dk^N dw$$

where $\delta(w)$ is the thickness factor, that is, the volume between $\mathcal{H}(w)$ and $\mathcal{H}(w + dw)$ and $B(k^N, w)$ is the product of $B_j(k^N, w)$. Assuming that only the tuples within $O(\sqrt{T} \log T)$ of the mean (see Lemma 4) significantly contribute to $d_T(x^T)$, we can substitute $B(k^N, w)$ by

$$B(k^N(w), w) = \frac{1}{\prod_j \sqrt{2\pi p(a_j w) q(a_j w) T_j}} \exp\left(-\sum_j \frac{(k_j - k_j(w))^2}{2p(a_j w)q(a_j w)T_j}\right)(1 + O(N/\sqrt{T}))$$

or written differently

$$B(k^N(w), w) = \sqrt{\det(\mathbf{A}(w)/(2\pi))} \exp\left(-\frac{\langle (k^N)^\tau \mathbf{A}(w) k^N \rangle}{2}\right)$$

where

$$\mathbf{A}(w) = \mathrm{Diag}\left(\frac{1}{p(a_1 w)q(a_1 w)T_1}, \ldots, \frac{1}{p(a_N w)q(a_N w)T_N}\right).$$

In Appendix A we prove

$$\sqrt{\det(\mathbf{A}(w)/(2\pi))} \int_{\mathcal{H}(w)} \exp\left(-\frac{1}{2}\langle \mathbf{z}^\tau \mathbf{A}(w) \mathbf{z}\right) d\mathbf{z} = \sqrt{\frac{1}{2\pi \langle \mathbf{u}^\tau \mathbf{A}^{-1}(w) \mathbf{u} \rangle}} \tag{20}$$

where $\mathbf{u} = (a_1, \ldots, a_N)^\tau$ is the unitary orthogonal vector to $\mathcal{H}(w)$. It is the same for all values of w since the hyperplanes are parallel and

$$\mathbf{u}^\tau = \frac{1}{\sqrt{\sum_j a_j^2}}(a_1, \cdots, a_N). \tag{21}$$

Thus

$$\langle \mathbf{u}^\tau \mathbf{A}^{-1}\mathbf{u}\rangle = \frac{\sum_j a_j^2 p(a_j w) q(a_j w) T_j}{\sum_j a_j^2}. \tag{22}$$

To finalize we need to find the thickness factor $\delta(w)$ that counts the number of y^T between $\mathcal{H}(w)$ and $\mathcal{H}(w+dw)$. As discussed, we cut the space $[0, T_1] \times \cdots \times [0, T_N]$ into parallel slices $\mathcal{H}(w)$. The hyperplane $\mathcal{H}(w)$ is the hyperplane orthogonal to $\mathbf{u}$ which contains the point $k^N(w)$. To reflect the full integral in the Cartesian metric $\delta(w)dw$ we must restrict thickness to slices between $\mathcal{H}(w)$ and $\mathcal{H}(w+dw)$. Since the hyperplane $\mathcal{H}(w+dw)$ is obtained by a translation of the hyperplane $\mathcal{H}(w)$ over the vector $(k^N)'(w)dw$. To compute $(k^N)'(w)dw$, we recall that $k^N(w^*) = (p(a_1 w)T_1, \ldots, p(a_N w)T_N)$ satisfies the following equation

$$\sum_j a_j T_j p(a_j w) = \sum_j a_j k_j(w). \tag{23}$$

Observe now that taking derivative of $k_j(w)$ with respect to w we obtain

$$\sum_j a_j^2 T_j p'(a_j w) = \sum_j a_j k_j'(w). \tag{24}$$

A simple by crucial observation here is that $p'(w) = p(w)(1-p(w)) = p(w)q(w)$ leading to the thickness $\delta(w)dw$ which is the component of the vector being orthogonal to $\mathcal{H}(w)$. We find

$$\delta(w) = \langle \mathbf{u}^\tau (k^N)'(w)\rangle = \frac{\sum_j a_j^2 p(a_j w) q(a_j w) T_j}{\sqrt{\sum_j a_j^2}}.$$

Putting everything together we find

$$d(x^T) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \sqrt{\sum_j a_j^2 p(a_j w) q(a_j w) T_j}\, dw \left(1 + O(N/\sqrt{T})\right). \tag{25}$$

This completes the proof of Corollary 2 for $d = 1$.

4.2. Finishing the Proof of Theorem 1

Following the same line of reasoning as in the previous subsection for general d we have

$$k^N \in \mathcal{H}_1(\mathbf{w}) \cap \cdots \cap \mathcal{H}_d(\mathbf{w}).$$

By approximating the underlying binomial distribution by the normal distribution, we arrive at

$$P(y^T|\mathbf{x}^T, \mathbf{w}) = \frac{\exp\left(-\sum_j \frac{(k_j - k_j(\mathbf{w}))^2}{2p(\langle \mathbf{a}_j \mathbf{w}\rangle) q(\langle \mathbf{a}_j \mathbf{w}\rangle) T_j}\right)}{\prod_j \sqrt{2\pi p(\langle \mathbf{a}_j \mathbf{w}\rangle) q(\langle \mathbf{a}_j \mathbf{w}\rangle) T_j}} (1 + O(N/\sqrt{T}))$$

or written differently

$$\sqrt{\det(\mathbf{A}(\mathbf{w})/(2\pi))}\exp\left(-\frac{\langle (k^N)^\tau \mathbf{A}(\mathbf{w})k^N\rangle}{2}\right)$$

where

$$\mathbf{A}(\mathbf{w}) = \mathrm{Diag}\left(\frac{1}{p(\langle \mathbf{a}_1\mathbf{w}\rangle)q(\langle \mathbf{a}_1\mathbf{w}\rangle)T_1},\ldots,\frac{1}{p(\langle \mathbf{a}_N\mathbf{w}\rangle)q(\langle \mathbf{a}_N\mathbf{w}\rangle)T_N}\right).$$

We now evaluate the minimax regret $r^*(\mathbf{x}^T) = \log d_T(\mathbf{x}^T)$ expressed in (11) which we repeat here

$$d_T(\mathbf{x}^T) = \int_{\mathbf{R}^d}\delta(\mathbf{w})dw_1\cdots dw_d\int_{\mathcal{H}^d(\mathbf{w})}P(y^T|\mathbf{w})dk^N \tag{26}$$

where $\delta(\mathbf{w})$ is a thickness indicator factor that takes into account the variation of spacing between the parallel subspaces $\mathcal{H}^d(\mathbf{w})$. In Appendix B we prove in (33) that

$$\sqrt{\det(\mathbf{A}(\mathbf{w})/(2\pi))}\int_{\mathcal{H}_1(\mathbf{w})\cap\cdots\cap\mathcal{H}_d(\mathbf{w})}\exp\left(-\frac{1}{2}\langle \mathbf{z}^\tau\mathbf{A}(\mathbf{w})\mathbf{z}\right)d\mathbf{z}^{N-d} = \sqrt{\frac{\det(\mathbf{U})}{\det(2\pi\mathbf{B}(\mathbf{w}))}} \tag{27}$$

where $\mathbf{U}$ is the $d\times d$ matrix whose i,j coefficient is $\langle \mathbf{u}_i\mathbf{u}_j\rangle$ and $\mathbf{B}(\mathbf{w})$ is the $d\times d$ matrix whose i,j coefficient is $\langle \mathbf{u}_i\mathbf{A}^{-1}(\mathbf{w})\mathbf{u}_j\rangle$. Thus

$$d_T(\mathbf{x}^T) = \int_{\mathbf{R}^d}\delta(\mathbf{w})\frac{\sqrt{\det(\mathbf{U})}}{\sqrt{\det(2\pi\mathbf{B}(\mathbf{w}))}}dw_1\cdots dw_d.$$

To finalize we need to express the thickness factor $\delta(\mathbf{w})$. As before, in the integral (26) we cut the space $[0,T_1]\times\cdots\times[0,T_N]$ into parallel slices $\mathcal{H}^d(w)$. The area between $\mathcal{H}(w_1,\ldots,w_d)$, and each of the $\mathcal{H}^d(w_1+dw_1,w_2,\ldots,w_d)$, $\mathcal{H}^d(w_1,w_2+dw_2,\ldots,w_d)\ldots$, and $\mathcal{H}^d(w_1,w_2,\ldots,w_d+dw_d)$ is equivalent to

$$\left|\det\left(\frac{\partial p_G(k^N(\mathbf{w}))}{\partial w_1},\ldots,\frac{\partial p_G(k^N(\mathbf{w}))}{\partial w_d}\right)\right|dw_1\cdots dw_d \tag{28}$$

where p_G is the orthogonal projection on the subspace $\mathcal{G}^d$ generated by the $\mathbf{u}_i$'s. We use here the known fact that the volume cut off by edge vectors $\mathbf{a}_1,\ldots\mathbf{a}_N$ is equal to $|\det(\mathbf{a}_1,\ldots\mathbf{a}_N)|$.

To better understand (28) we notice that the $d\times d$ matrix with ij coefficient is $\langle \mathbf{u}_i^\tau\frac{\partial k^N(\mathbf{w})}{\partial w_j}\rangle$ is nothing less than matrix $\mathbf{B}(\mathbf{w})$. But to express the determinant we need its orthonormal base of $\mathcal{G}^d$ which we denote as $(\mathbf{e}_1,\ldots,\mathbf{e}_d)$. The determinant we are looking for is the determinant of the matrix $\mathbf{D}(\mathbf{w})$ whose ij coefficient is $\langle \mathbf{e}_i^\tau\frac{\partial k^N(\mathbf{w})}{\partial w_j}\rangle$. We can create an orthonormal base of $\mathcal{G}^d$ just by setting $\mathbf{e}_i = \sum_j e_{ij}\mathbf{u}_j$ as long as the matrix $\mathbf{E}$ with coefficients e_{ij} satisfies:

$$\mathbf{E}^t\mathbf{U}\mathbf{E} = \mathbf{I}_d \tag{29}$$

where $\mathbf{I}_d$ is the $d\times d$ identity matrix. Thus $\mathbf{D}(\mathbf{w}) = \mathbf{E}\mathbf{B}(\mathbf{w})$ and

$$\delta(\mathbf{w}) = |\det(\mathbf{D}(\mathbf{w}))| = |\det(\mathbf{E}\mathbf{B}(\mathbf{w}))| = \frac{\det(\mathbf{B}(\mathbf{w}))}{\sqrt{\det(\mathbf{U})}}. \tag{30}$$

Putting everything together:

$$d_T(\mathbf{x}^T) = \frac{1}{(2\pi)^{d/2}}\int_{\mathbf{R}^d}\sqrt{\det(\mathbf{B}(\mathbf{w})}dw_1\cdots dw_d\left(1+O(N/\sqrt{T})\right). \tag{31}$$

This proves Theorem 1, after some simple final calculations.

Appendix A: Special One-Dimensional Case

Let $\mathbf{A}$ be a self adjoint matrix which is definite positive. Let $\mathcal{H}$ be a hyperplane orthogonal to the unitary vector $\mathbf{u}$, not necessarily an eigenvector of $\mathbf{A}$. We want to compute the integral

$$I(\mathcal{H},\mathbf{A})=\sqrt{\det(\mathbf{A}/2\pi)}\int_{\mathcal{H}}\exp\left(-\frac{\langle\mathbf{z}^{\tau}\mathbf{A}\mathbf{z}\rangle}{2}\right)d\mathbf{z}$$

with $\mathbf{z}^{\tau}$ being the transpose of $\mathbf{z}=(z_1,\dots,z_N)$. We know that the integral on the whole space, since the integrand is a Gaussian density with $\mathbf{A}^{-1}$ as the covariance matrix. We will make the use of the following identity obtained by slicing the whole space into a folio of hyperplanes parallel to $\mathcal{H}$

$$\sqrt{\det(\mathbf{A}/2\pi)}\int_{-\infty}^{+\infty}dt\int_{\mathcal{H}}\exp\left(-\frac{\langle(\mathbf{z}+t\mathbf{u})^{\tau}\mathbf{A}(\mathbf{z}+t\mathbf{u})\rangle}{2}\right)d\mathbf{z}=1.$$

Let $\mathbf{u}=p(\mathbf{u})+\mathbf{v}$ where $p(\mathbf{u})$ is the projection of $\mathbf{u}$ on $\mathcal{H}$ according to the metric induced by $\mathbf{A}$. Thus in the integrand we have

$$\langle(\mathbf{z}+t\mathbf{u})^{\tau}\mathbf{A}(\mathbf{z}+t\mathbf{u})\rangle=\langle(\mathbf{z}+tp(\mathbf{u}))^{\tau}\mathbf{A}(\mathbf{z}+tp(\mathbf{u}))\rangle+t^2\langle\mathbf{v}^{\tau}\mathbf{A}\mathbf{v}\rangle.$$

For a given t we have by a simple change of variable

$$\int_{\mathcal{H}}\exp\left(-\frac{\langle(\mathbf{z}+tp(\mathbf{u}))^{\tau}\mathbf{A}(\mathbf{z}+tp(\mathbf{u}))\rangle}{2}\right)d\mathbf{z}=\int_{\mathcal{H}}\exp\left(-\frac{\langle\mathbf{z}^{\tau}\mathbf{A}\mathbf{z}\rangle}{2}\right)d\mathbf{z}.$$

Thus

$$I(\mathcal{H},\mathbf{A})\int_{-\infty}^{\infty}\exp(-t^2\langle\mathbf{v}^{\tau}\mathbf{A}\mathbf{v}\rangle/2)dt=1$$

and therefore,

$$I(\mathcal{H},\mathbf{A})=\sqrt{\frac{\langle\mathbf{v}^{\tau}\mathbf{A}\mathbf{v}\rangle}{2\pi}}.$$

In order to determine $\mathbf{v}$ we notice that if $\mathbf{v}$ is orthogonal to $\mathcal{H}$ with metric $\mathbf{A}$ then $\mathbf{A}\mathbf{v}$ is orthogonal to $\mathcal{H}$ with classic metric. Thus $\mathbf{A}\mathbf{v}$ is colinear with $\mathbf{u}$, or equivalently $\mathbf{A}^{-1}\mathbf{u}$ is colinear with $\mathbf{v}$. Since $\mathbf{u}-\mathbf{v}$ must belong to $\mathcal{H}$ then $\langle\mathbf{u}^{\tau}(\mathbf{u}-\mathbf{v})\rangle=0$ and

$$\mathbf{v}=\frac{1}{\langle\mathbf{u}^{\tau}\mathbf{A}^{-1}\mathbf{u}\rangle}\mathbf{A}^{-1}\mathbf{u},$$

and consequently $\langle\mathbf{v}^{\tau}\mathbf{A}\mathbf{v}\rangle=\frac{1}{\langle\mathbf{u}^{\tau}\mathbf{A}^{-1}\mathbf{u}\rangle}$. Finally, we arrive at

$$I(\mathcal{H},\mathbf{A})=\sqrt{\frac{1}{2\pi\langle\mathbf{u}^{\tau}\mathbf{A}^{-1}\mathbf{u}\rangle}}$$

which proves

$$\delta(w)=\langle\mathbf{u}^{\tau}(k^N)'(w)\rangle=\frac{\sum_{j=1}^{N}a_j^2p(a_jw)q(a_jw)T_j}{\sqrt{\sum_{j=1}^{N}a_j^2}}\tag{32}$$

in the $d=1$ case.

Appendix B: General d dimensional case

Now let $\mathcal{H}^d$ be the intersection of d hyperplanes, respectively orthogonal to $\mathbf{u}_1, \mathbf{u}_2, \dots \mathbf{u}_d$ not necessarily orthonormal. We write

$$I(\mathcal{H}^d, \mathbf{A}) = \sqrt{\det(\mathbf{A}/(2\pi))} \int_{\mathcal{H}^d} \exp\left(-\frac{\mathbf{z}^\tau \mathbf{A}\mathbf{z}}{2}\right) d\mathbf{z}^{N-d}.$$

We know that

$$\sqrt{\det(\mathbf{A}/(2\pi))} \int \exp\left(-\frac{\mathbf{z}^\tau \mathbf{A}\mathbf{z}}{2}\right) d\mathbf{z}^{N-d} = 1.$$

Let $\mathcal{G}^d$ be the sub vector space orthogonal to $\mathcal{H}^d$. We have

$$\int \exp\left(-\frac{\mathbf{z}^\tau \mathbf{A}\mathbf{z}}{2}\right) d\mathbf{z}^N = \int_{\mathbf{z}\in\mathcal{H}^d} \int_{\mathbf{x}\in\mathcal{G}^d} \exp\left(\frac{\langle(\mathbf{z}+\mathbf{x})^\tau \mathbf{A}(\mathbf{z}+\mathbf{x})\rangle}{2}\right) d\mathbf{z}^{N-d} d\mathbf{x}^d.$$

The vector space is generated by the vectors $\mathbf{u}_i$. Let $\mathbf{t} = (t_1, \dots, t_d)$ and we denote by $\mathbf{x}(t) = \sum_i t_i \mathbf{u}_i$. Thus the change of variable leads to

$$\int \exp\left(-\frac{\mathbf{z}^\tau \mathbf{A}\mathbf{z}}{2}\right) d\mathbf{z}^N = \int_{\mathbf{z}\in\mathcal{H}^d} \int_{t\in\mathbf{R}^d} \exp\left(\frac{\langle(\mathbf{z}+\mathbf{x}(t))^\tau \mathbf{A}(\mathbf{z}+\mathbf{x}(t))\rangle}{2}\right) d\mathbf{z}^{N-d} \sqrt{\det(\mathbf{U})} d\mathbf{t}^d.$$

Let $p_H(\mathbf{x})$ be the projection of $\mathbf{x}$ on $\mathcal{H}^d$ according to metric A. We denote by $p_A(\mathbf{x}) = \mathbf{x} - p_H(\mathbf{x})$. Thus

$$\begin{aligned}
\int \exp\left(-\frac{\mathbf{z}^\tau \mathbf{A}\mathbf{z}}{2}\right) d\mathbf{z}^N &= \int_{\mathbf{z}\in\mathcal{H}^d} \int_{t^d} \exp\left(\frac{\langle(\mathbf{z}+p_H(\mathbf{x}(t)))^\tau \mathbf{A}(\mathbf{z}+p_H(\mathbf{x}(t)))\rangle}{2}\right) \sqrt{\det(\mathbf{U})} \\
&\quad \times \exp\left(-\frac{\langle p_A(\mathbf{x}(t))^\tau \mathbf{A} p_A(\mathbf{x}(t))\rangle}{2}\right) d\mathbf{z}^{N-d} dt^d \\
&= \int_{\mathbf{z}\in\mathcal{H}^d} \exp\left(-\frac{\mathbf{z}^\tau \mathbf{A}\mathbf{z}}{2}\right) d\mathbf{z}^{N-d} \sqrt{\det(\mathbf{U})} \\
&\quad \times \int_{t^d} \exp\left(-\frac{\langle p_A(\mathbf{x}(t))^\tau \mathbf{A} p_A(\mathbf{x}(t))\rangle}{2}\right) dt^d \\
&= I(\mathcal{H}^d \mathbf{A}) \sqrt{\det(\mathbf{U})} \int_{t^d} \exp\left(-\frac{\langle p_A(\mathbf{x}(t))^\tau \mathbf{A} p_A(\mathbf{x}(t))\rangle}{2}\right) dt^d.
\end{aligned}$$

The $p_A(\mathbf{x}(t))$ belongs to the vector space orthogonal to $\mathcal{H}^d$ according to metric $\mathbf{A}$. It is the image of $\mathcal{G}^d$ by the operator $\mathbf{A}^{-1}$ and is generated by the vectors $\mathbf{A}^{-1}\mathbf{u}_i$. Let c_{ij} be such that

$$p_A(\mathbf{u}_j) = \sum_{i=1}^N c_{ij} \mathbf{A}^{-1} \mathbf{u}_i.$$

We denote by $\mathbf{C} = [c_{ij}]_{i,j=1}^N$ the matrix whose ij coefficient is c_{ij}. To determine the matrix $\mathbf{C}$ we use the fact that for all j the vector $\mathbf{u}_j - p_A(\mathbf{u}_j)$ belongs to $\mathcal{H}^d$, *i.e* for all k: $\langle(\mathbf{u}_j - p_A(\mathbf{u}_j))^\tau \mathbf{u}_k\rangle = 0$, thus

$$\langle \mathbf{u}_j^\tau \mathbf{u}_k \rangle = \sum_{i=1}^N c_{ij} \langle \mathbf{u}_i^\tau \mathbf{A}^{-1} \mathbf{u}_k \rangle.$$

In other words, we have the matrix identity: $\mathbf{U} = \mathbf{BC}$ with $\mathbf{B}$ the $d \times d$ matrix whose ij coefficient is $\langle \mathbf{u}_i^\tau \mathbf{A}^{-1} \mathbf{u}_j \rangle$. Thus $\mathbf{C} = \mathbf{B}^{-1}\mathbf{U}$. We therefore have

$$\begin{aligned} p_A(\mathbf{x}(t))^\tau \mathbf{A} p_A(\mathbf{x}(t)) &= \sum_{ij=1}^{N} t_i t_j c_{ik} \langle \mathbf{u}_k^\tau \mathbf{A}^{-1} \mathbf{A} \mathbf{A}^{-1} \mathbf{u}_\ell \rangle c_{\ell j} \\ &= \langle \mathbf{t}^\tau \mathbf{C}^\tau \mathbf{B} \mathbf{C} \mathbf{t} \rangle = \langle \mathbf{t}^\tau \mathbf{U} \mathbf{B}^{-1} \mathbf{U} \mathbf{t} \rangle. \end{aligned}$$

Finally

$$\sqrt{\det(\mathbf{U})} \int_{t^d} \exp\left(-\frac{\langle p_A(\mathbf{x}(t))^\tau \mathbf{A} p_A(\mathbf{x}(t)) \rangle}{2}\right) dt^d =$$

$$\sqrt{\det(\mathbf{U})} \int_{t^d} \exp\left(-\frac{\langle t^\tau \mathbf{U} \mathbf{B}^{-1} \mathbf{U} t \rangle}{2}\right) d\theta^d =$$

$$= \sqrt{\frac{\det(\mathbf{U})}{\det(\mathbf{U}\mathbf{B}^{-1}\mathbf{U}/(2\pi))}} = \sqrt{\frac{\det(2\pi\mathbf{B})}{\det(\mathbf{U})}}.$$

In summary,

$$I(\mathcal{H}^d, \mathbf{A}) = \sqrt{\frac{\det(\mathbf{U})}{\det(2\pi\mathbf{B})}} \tag{33}$$

proving (27).

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

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 andMehryar Mohri, and Karthik Sridharan. Logistic regression: The importance of being improper. In *COLT - Conference on Learning Theory*, 2018.

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

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.

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.

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

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

J. Rissanen. Minimax codes for finite alphabets. *EEE Trans. Inform. Theory*, IT-24(3):389–392, 1978.

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

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

G. I. Shamir and W. Szpankowski. A general lower bound for regret in logistic regression, 2021.

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.

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.