Regret for Online Regression with General Log-Type Losses

Mohsen Heidari MHEIDARI@PURDUE.EDU
Purdue University
W. Lafayette, IN 47907

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

Wojciech Szpankowski SZPAN@PURDUE.EDU
Purdue University
West Lafayette, IN, USA

Abstract

In online learning a learner receives data in rounds $1 \leq t \leq T$ and at each round predicts a binary label which is then compared to that of the best forecaster incurring a loss. The total loss over T rounds, when compared to a loss over a *concept class* or experts/ forecasters, is called the regret.

In this paper we precisely analyze the minimax regret which is the regret for the best predictor (learning) distribution and the worst label sequences for a fixed data sequence. We show that for the logarithmic loss over a concept class parametrized by a given probability $p(\cdot)$ (e.g., logistic regression and normal distribution) the minimax regret grows like $(d/2)\log T + C_d + O(d^{3/2}/\sqrt{T})$ where d is data dimension and C_d is a "constant" that depends on the probability p, dimension, and data. We then turn to quantum data that requires us to consider minimax regret on manifolds (sphere) and prove an upper bound for the regret in this case. To establish such a precise finding we use a combination of analytic combinatorics, information theory, discrete probability, and topology.

1. Introduction

In online learning data becomes available in a sequential order and is used to update the best predictor for future data, unlike in the batch learning which generate the best predictor by learning on the entire training data set. Various methods for predictions are used: recursive least squares, stochastic gradient descent, kernel methods, online convex optimization such as "follow the leader", and others. In this paper we will not introduce any new prediction algorithm, but rather evaluate the quality of the best predictor by analyzing a regret. The (pointwise) *regret* of an online algorithm is defined as the (excess) loss it incurs over some value of a constant *comparator* (e.g., prediction provided by a class of experts/ forecasters). This will give us a universal lower limit which every good online algorithm should asymptotically match (i.e., the first or the first two leading terms of the underlying regret).

We phrase our online learning problem in terms of a game between nature/ environment and a learner/predictor. Broadly, the objective of the learner is to process past observations to predict the next realization of the nature's labeling sequence. At each round $t \in \mathbb{N}$, let y_t be the true label that is yet to be revealed. At time t, the learner obtains a d dimensional input/ feature vector $\mathbf{x}_t \in \mathbb{R}^d$.

In addition to $\mathbf{x}_t$, the learner may use the past observations $(\mathbf{x}_r, y_r)$, $r < t$ to make a (probabilistic) prediction $\hat{y}_t \in \mathbb{R}$ of the true label. Therefore, the prediction can be written as $\hat{y}_t = g_t(y^{t-1}, \mathbf{x}^t)$, where g_t represents the strategy/ algorithm of the learner to obtain its prediction based on the past and current observations. Hence, the learner is modeled by the sequence of predicting actions $g_t, t > 0$. Once a prediction is made, the nature reveals the true label y_t and the learner incurs some *loss* evaluated based on a predefined function $\ell : \hat{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}^+$, where $\hat{\mathcal{Y}} \subseteq \mathbb{R}$ and $\mathcal{Y} = \{-1, 1\}$ are the prediction and label domains respectively. In regret analysis, we are interested in comparing the accumulated loss of the learner with that of the best strategy within a predefined class of predictors (forecasters or experts) denoted as $\mathcal{C}$. In fact, $\mathcal{C}$ is a collection of predicting functions $h : \mathbb{R}^d \mapsto \mathbb{R}$, with input being $\mathbf{x}_t$ at each time t. Therefore, given a learner prediction function $g_t : \mathbb{R}^d \to \mathbb{R}$ and after T rounds with the realizations $(y_t, \mathbf{x}_t)_{t=1}^T$, the *pointwise regret* is defined as

$$R(g^T, y^T, \mathcal{C}|\mathbf{x}^T) = \sum_{t=1}^{T} \ell(\hat{y}_t, y_t) - \inf_{h \in \mathcal{C}} \sum_{t=1}^{T} \ell(h(x_t), y_t),$$

where $\hat{y}_t = g_t(y^{t-1}, \mathbf{x}^t)$ (e.g., *binary* prediction may be $\text{sign}(g_t)$). Throughout we write $y^T = (y_1, \ldots, y_T)$ and $\mathbf{x}^T = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$ and condition on fixed $\mathbf{x}^T$. The first and the second summations above represent the accumulated loss of the learner (algorithmic loss) and the best forecaster within the class $\mathcal{C}$, respectively. There are at least two main perspectives on analyzing the regret.

Fixed Design: This point of view studies the minimal regret for the worst realization of the label with the feature vector $\mathbf{x}^T$ known in advance. Suppose that the learner has a fixed strategy $g_t, t > 0$. Then, the *fixed design minimax regret* is defined as

$$r_T^*(\mathcal{C}|\mathbf{x}^T) = \inf_{g^T} \sup_{y^T} R(g^T, y^T, \mathcal{C}|\mathbf{x}^T). \tag{1}$$

Note that this notion was also studied in Shamir and Szpankowski (2021); Jacquet et al. (2021), and in Cesa-Bianchi and Shamir (2011) under the name *transductive online learning*.

Sequential Design: In this approach, the optimization on regret is performed at every time t without knowing y^T. Then the *sequential minimax regret* with fixed $\mathbf{x}^T$ is defined as in Rakhlin and Sridharan (2014)

$$r_T^a(\mathcal{C}|\mathbf{x}^T) = \inf_{\hat{y}_1} \sup_{y_1} \cdots \inf_{\hat{y}_T} \inf_{y_T} R(\hat{y}^T, y^T, \mathcal{C}|\mathbf{x}^T). \tag{2}$$

However, as recently proved in Wu et al. (2022) these two formulations of minimax regrets are equal, that is, $r_T^*(\mathcal{C}|\mathbf{x}^T) = r_T^a(\mathcal{C}|\mathbf{x}^T)$. This is actually important since best techniques known for analyzing the sequential regret $r_T^a(\mathcal{C}|\mathbf{x}^T)$ with logarithmic loss give rather unimpressive upper bound of $O(\sqrt{T})$ which is far away from the achievable regret as discuss in this paper (see also Shamir and Szpankowski (2021); Jacquet et al. (2021)).

More specifically, in this paper we consider a hypothesis (expert) class that generalizes previously studies on linear predictors with logarithmic losses:

$$\mathcal{C}_{p,\mathbf{w}} = \{h : \mathbb{R}^d \to \mathbb{R} : \ h(\mathbf{x}) = p(\langle \mathbf{w}|\mathbf{x} \rangle) : \mathbf{w}, \mathbf{x} \in \mathcal{M} \subset \mathbb{R}^d\}, \tag{3}$$

where $\mathbf{w}$ is a d dimensional weight vector and $\langle \mathbf{w}|\mathbf{x} \rangle$ is the scalar product, $\mathcal{M}$ is a manifold in $\mathbb{R}^d$, while $p(w)$ with $w = \langle \mathbf{w}|\mathbf{x} \rangle$ is a probability function. Often $p(w)$ is either the logistic function $p(w) = (1 + \exp(-w))^{-1}$ (see Bach (2010); McMahan and Streeter (2012); Hazan et al. (2014);

Shamir (2020) or the probit function $p(w) = \Phi(-w)$ where $\Phi(w)$ is CDF (cumulative distribution function) of the normal distribution (see Aldrich et al. (1984); Bishop (2006)).

Furthermore, in this paper we study only logarithmic loss. More specifically, for any $h \in \mathcal{C}$, we interpret $h(\mathbf{x}) \in [0,1]$ as the probability assigned to $Y = -1$, that is, $P(Y_t = -1|\mathbf{x}^t, \mathbf{w}) = p(\langle \mathbf{w}|\mathbf{x}_t \rangle)$. Therefore, the logarithmic loss function corresponding to $h(\mathbf{x})$ is equal to $\ell(h(\mathbf{x}), y) = -\log|(1+y)/2 - h(\mathbf{x})|$, for all $y \in \{-1, 1\}$. We prove in Lemma 8 of Appendix A that this loss function is equivalent to the logarithmic loss function of the model probability, that is, $\ell(h(\mathbf{x}), y) = -\log P(Y = y|\mathbf{x})$. Furthermore, it is easy to show that $\ell(\hat{y}, y) = -\log P(\hat{Y} = y|\mathbf{x})$. In order to simplify our presentation, we denote that latter probability as $Q(y|\mathbf{x}) := P(\hat{Y} = y|\mathbf{x})$.

In view of the above definitions, we can now specifically define the pointwise regret as

$$R_T(Q, y^T|\mathbf{x}^T) = -\sum_{t=1}^{T} \log Q(y_t|\mathbf{x}_t) - \inf_{\mathbf{w}} \sum_{t=1}^{T} \log(1/P(y_t|\mathbf{x}_t, \mathbf{w})) \tag{4}$$

and then the (maximal) minimax regret studied here is defined as

$$r_T^*(\mathbf{x}^T) = \inf_{Q} \max_{y^T} R_T(Q, y^T|\mathbf{x}^T). \tag{5}$$

In this paper we provide a precise asymptotic expansion of the maximal minimax regret for the class $\mathcal{C}_{p,\mathbf{w}}$, a result that had been wanting for some time.

Main Contribution. Our contribution is two-fold. First, we present precise asymptotic expansions for the maximal minimax regret (5) through the so called Shtarkov sum (see Shtarkov (1987); Drmota and Szpankowski (2004)). Second, we apply new methodology using tools of analytic combinatorics such as complex asymptotics and Fourier transforms (see Flajolet and Sedgewick (2008); Szpankowski (2001)) to handle the Shtarkov sum for the concept class (3). The concept class $\mathcal{C}_{p,\mathbf{w}}$ was never analyzed for arbitrary $p(\cdot)$: most work on regret are restricted to the logistic regression which has a unique property making the analysis easier (see Section 3).

More precisely, we first represent the minimax regret (5) as the logarithm of the Shtarkov sum over all label sequences of the optimal label probability which turns out to be the maximum-likelihood distribution (which we equate to Q). Shtarkov's sum arose already in the universal compression as witnessed by Drmota and Szpankowski (2004); Szpankowski and Weinberger (2012). We show in Theorem 2 that for $d = o(T^{1/3})$ the minimax regret grows as $\frac{d}{2}\log(2T/\pi) + C_d(p, \mathbf{x}^T) + O(d^{3/2}/\sqrt{T})$ where the "constant" $C_d(p, \mathbf{x}^T)$ depends on the function $p(\cdot)$ describing the model class, dimension d and data $\mathbf{x}^T$. Then in Theorem 4 we extend it to the case when the data and the weight belong to a manifold $\mathcal{M}$. This brings differential topology into the play. We mention that even if both theorems look alike, the proofs are significantly different since in the latter case we need to take into account the curvature of the manifold $\mathcal{M}$. Lastly, we extend our results to online learning of observations obeying laws of quantum mechanics. In this problem, at each round a quantum state is observed, however, here we only consider a plarization of a photon. The objective of the quantum learner is to arrive at a probabilistic prediction of the classical labels associated with the quantum states. In this problem, the vectors $\mathbf{w}$ belong to a sphere $\mathcal{S}$ hence Theorem 4 can be used to derive in Theorem 5 bounds on such a regret.

Related Work In this paper we study a machine learning problem (see, e.g., Cesa-Bianchi and Lugosi (2006); Shalev-Schwartz and Ben-David (2014)) that of minimax regret for a general online

regression with the logarithmic loss using a combined methodology of analytic combinatorics (see, e.g., Flajolet and Sedgewick (2008); Jacquet and Szpankowski (2015); Szpankowski (2001)) and information theory, in particular the universal source coding (see, e.g., Barron et al. (1998); Drmota and Szpankowski (2004); Krichevsky and Trofimov (1981); Orlitsky and Santhanam (2004); Rissanen (1984, 1996); Shamir (2006); Xie and Barron (1997)).

With respect to information theory, the minimax online regret is similar to the redundancy of universal coding with dimensional $d = 1$ studied extensively in information theory. For example, for general label alphabet of size m, it is known 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 (see Drmota and Szpankowski (2004); Rissanen (1996); Shamir (2006); Szpankowski (1998); Xie and Barron (1997, 2000)) and $\frac{m-1}{2}\log(T/m)$ for $m = o(T)$ (see also Orlitsky and Santhanam (2004); Shamir (2006); Szpankowski and Weinberger (2012)).

With respect to machine learning, most existing work on online regression deal with logistic regression. We first mention work of Hazan et al. (2014) who studied the pointwise regret of the logistic regression for the *proper* setting, that is, when at time t the decision regarding $\mathbf{w}_t$ is based on knowledge available to the learner up to time $t-1$. Unlike the *improper* learning, studied in this paper, where feature $\mathbf{x}_t$ at time is also available to the learner and Hazan et al. (2014) showed that the pointwise regret is $\Theta(T^{1/3})$ for $d = 1$ and $O(\sqrt{T})$ for $d > 1$. Furthermore, Kakade and Ng (2005) were first to demonstrate results showing that *pointwise regret* for logistic regression grows like $O(d\log T/d)$ where for fixed dimension d and $m = 2$. This was further generalized in Foster et al. (2018) to all m showing that the regret can grow as $O(d\log(T/d))$. These results were strengthened in Shamir (2020), which also provided matching lower bounds. Finally, we should add that the worst case *adversarial* minimax regret is studied in a series of papers by Rakhlin and Sridharan (2014, 2015) using sequential Rademacher complexity, however, for logarithm loss function this technique gives only unimpressive bound $O(\sqrt{T})$ which is far away from the correct growth $(d/2)\log T$ (see also Wu et al. (2022)).

Our results are closest to Shamir and Szpankowski (2021); Jacquet et al. (2021), however, those authors only study logistic regression. Furthermore, in Jacquet et al. (2021) a precise maximal minimax regret is analyzed but only for *finite* number of feature values and fixed dimension d. More precisely, Jacquet et al. (2021) addresses a relaxed problem with at most a finite number $N = o(\sqrt{T})$ of distinct feature vector values and regret is analyzed only for a fixed dimension $d = O(1)$. In this paper we (i) consider a *general online regression* with logarithmic loss, (ii) feature vectors can take any values and can lie on a manifold; (iii) the dimension d can grow with T as $d = o(T^{1/3})$; and (iv) a different methodology based on a multidimensional Laplace is used while the analysis of Jacquet et al. (2021) is based on multidimensional Gaussian approximation, which fails in our setting Recently, in Jacquet et al. (2022) some preliminary results were presented for the logistic regression, however, as we shall see in Section 3 the logistic regression is a very special case and extending it to other (e.g., Gaussian) regressions is very challenging. We point out that even our analysis of the logistic regression in this paper is unique as seen in the proof of Lemma 6.

2. Main Results

In this section we precisely formulate our problem, and then provide general solution followed by specific results for logistic and probit regression as well as for quantum data.

2.1. Problem Formulation

We denote by $\mathbf{x}_t = (x_{1,t}, \ldots, x_{d,t}), t > 0$, a d-dimensional bounded feature vector in $\mathbb{R}^d$ such that $||\mathbf{x}|| \leq 1$. Furthermore, after T rounds, the features and labels are denoted as $\mathbf{x}^T$ and y^T with $y_t \in \{-1, 1\}$. At last, we introduce a weight vector $\mathbf{w}_t \in \mathbb{R}^d$ defining the class $\mathcal{C}_{p,\mathbf{w}}$. We mostly study the case when $\mathbf{w} \in \mathbb{R}^d$, however, we also consider cases when $\mathbf{w}$ belongs a manifold $\mathcal{M} \subseteq \mathbb{R}^d$.

In this paper we derive precise asymptotics of the (maximal) minimax regret defined in (5). We recall that $R_T(Q, y^T|\mathbf{x}^T)$ is the pointwise regret that for the concept class $\mathcal{C}_{p,\mathbf{w}}$ defined in (3) can be expressed as

$$R_T(Q, y^T|\mathbf{x}^T) = -\sum_{t=1}^{T} \log Q(y_t|\mathbf{x}_t) + \sup_{\mathbf{w}\in\mathcal{M}} \sum_{t=1}^{T} \log P(y_t|\mathbf{x}_t, \mathbf{w}) = \log \frac{\sup_{\mathbf{w}\in\mathcal{M}} P(y^T|\mathbf{x}^T, \mathbf{w})}{Q(y^T|\mathbf{x}^T)}$$

where

$$P(Y_t = -1|\mathbf{x}, \mathbf{w}) = p(\langle \mathbf{w}|\mathbf{x}\rangle), \quad \text{and} \quad P(Y_t = +1|\mathbf{x}, \mathbf{w}) = 1 - p(\langle \mathbf{w}|\mathbf{x}\rangle). \tag{6}$$

We postulate the following

(H) The functions $-\log p(w)$ and $-\log(1 - p(w))$ are convex with bounded first and second derivatives.

Note that $Q(y_t|\mathbf{x}_t)$ is a distribution, denoting the algorithm's actions that approximates the forecaster's probability $P(y_t|\mathbf{x}_t, \mathbf{w})$. In this formulation we do not predict explicitly $\hat{y}_t$ of the label y_t but rather we use the learning distribution $Q(\cdot)$ to calculate $\hat{y}_t$.

In order to study precisely the minimax regret $r_T^*(\mathbf{x}^T)$ we need a more succinct and computationally manageable representation of it. Following Shtarkov (1987); Drmota and Szpankowski (2004); Cesa-Bianchi and Lugosi (2006) we add and subtract from (5) the logarithm of the Shtarkov sum

$$S_T(\mathbf{x}^T) := \sum_{y^T} \sup_{\mathbf{w}\in\mathcal{M}} P(y^T|\mathbf{x}^T, \mathbf{w}) \tag{7}$$

resulting in

$$\begin{aligned} r_T^*(\mathbf{x}^T) &= \min_Q \sup_{\mathbf{w}\in\mathcal{M}} \max_{y^T} (-\log Q(y^T|\mathbf{x}^T) + \log P^*(y^T|\mathbf{x}^T)) + \log \sum_{y^T} \sup_{\mathbf{w}\in\mathcal{M}} P(y^T|\mathbf{x}^T, \mathbf{w}) \\ &= \log \sum_{y^T} \sup_{\mathbf{w}\in\mathcal{M}} P(y^T|\mathbf{x}^T, \mathbf{w}) = \log S_T(\mathbf{x}^T) \end{aligned} \tag{8}$$

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

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

being the *maximum-likelihood distribution*. Indeed, since Q and P^* are distributions, there is at least one y^T such that the first term in (8) is nonnegative, so that $Q = P^*$ minimizes it (see also Cesa-Bianchi and Lugosi (2006)).

2.2. General Result

We now adopt the concept class $\mathcal{C}_{p,\mathbf{w}}$ as defined in (3) with $\mathbf{w} \in \mathbb{R}^d$ and bounded $\mathbf{x}_t$, i.e., without any loss of generality $||\mathbf{x}_t|| \leq 1$. Setting $w = \langle \mathbf{w}|\mathbf{x}_t\rangle$ we simply write $p(w) = p(\langle \mathbf{w}|\mathbf{x}_t\rangle)$ that satisfies (6). We also assume that the hypothesis (H) holds. Observe that for any $\mathbf{x}^T$ and y^T, $P(y^T|\mathbf{x}^T, \mathbf{w}) = \prod_t P(y_t|\mathbf{x}_t, \mathbf{w})$, with the probability terms as in (6). Hence,

$$P(y^T|\mathbf{x}^T, \mathbf{w}) = \prod_{t:y_t=-1} p(\langle \mathbf{w}|\mathbf{x}_t\rangle) \prod_{t:y_t=1} (1 - p(\langle \mathbf{w}|\mathbf{x}_t\rangle)) = \exp\big(-L(y^T|\mathbf{x}^T, \mathbf{w})\big) \tag{10}$$

so that $L(y^T|\mathbf{x}^T, \mathbf{w}) = -\log P(y^T|\mathbf{x}^T, \mathbf{w})$. Our objective is to estimate asymptotically the logarithms of the Shtarkov's sum defined in (7). However, this representation of the Shtarkov sum is not convenient for an asymptotic evaluation. The next lemma presents our main technical tool used in this paper. The proof is delayed till the next section.

Lemma 1 *Let hypothesis (H) hold and $L(y^T|\mathbf{x}^T, \mathbf{w}) = -\log P(y^T|\mathbf{x}^T, \mathbf{w})$ with $P(y^T|\mathbf{x}^T, \mathbf{w})$ as in (10). Then*

$$S(\mathbf{x}^T) = \frac{1}{(2\pi)^d} \sum_{y^T} \int_{\mathbb{R}^d} \exp\big(-L(y^T|\mathbf{w})\big) \det(\nabla^2 L(y^T|\mathbf{w}))\delta\mathbf{w} \int_{\mathbb{R}^d} \exp\big(-i\langle \nabla L(y^T|\mathbf{w})|\mathbf{z}\rangle\big) d\mathbf{z} \tag{11}$$

where $i = \sqrt{-1}$, and $\nabla L(y^T|\mathbf{w})$ and $\nabla^2 L(y^T|\mathbf{w})$ are the gradient and the Laplacian of $L(y^T|\mathbf{w})$.

In the next section and Appendix we prove our first main result.

Theorem 2 *Let $\mathbf{w} \in \mathbb{R}^d$. We assume the sequences $\mathbf{x}_t$ spans $\mathbb{R}^d$ and $||\mathbf{x}_t|| \leq 1$. We also write $p(w) = p(\langle \mathbf{w}|\mathbf{x}\rangle)$ satisfying (6) and hypothesis (H). Then asymptotically for $d = o(T^{1/3})$*

$$r^*(\mathbf{x}^T) = \frac{d}{2}\log T - \frac{d}{2}\log 2\pi + C_d(p, \mathbf{x}^T) + O(d^{3/2}/\sqrt{T}) \tag{12}$$

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

$$C_d(p, \mathbf{x}^T) = \log\left(\int_{\mathbf{R}^d} \sqrt{\det\left(\frac{1}{T}\mathbf{B}_d(\mathbf{w}, \mathbf{x}^T)\right)} dw_1 \cdots dw_d\right) \tag{13}$$

with

$$\mathbf{B}(\mathbf{w}, \mathbf{x}^T) = \sum_{t=1}^{T} \frac{p'(\langle \mathbf{w}|\mathbf{x}^t\rangle))^2}{(1 - p(\langle \mathbf{w}|\mathbf{x}^t\rangle))p(\langle \mathbf{w}|\mathbf{x}^t\rangle)} \mathbf{x}_t \otimes \mathbf{x}_t \tag{14}$$

and $\mathbf{x}_t \otimes \mathbf{x}_t = \mathbf{x}_t\mathbf{x}_t^\tau$ being the tensor product of $\mathbf{x}_t$ (i.e., $d \times d$ matrix) with τ as the transpose.

We present two examples with $p(w)$ being either the logistic function or probit function.

Example 1 (Logistic Function) *Let now* $\mathrm{Sigmoid}(w) = (1 + e^{-w})^{-1}$, *that is,* $\mathrm{Sigmoid}(\langle \mathbf{w}|\mathbf{x}_t\rangle)$ *is the class of logistic functions. Then* $\mathbf{B}(\mathbf{w})$ *becomes (see Shamir (2020); Jacquet et al. (2022))*

$$\mathbf{B}(\mathbf{w}) = \sum_{t=1}^{T} \mathrm{Sigmoid}(\langle \mathbf{w}|\mathbf{x}_t\rangle)(1 - \mathrm{Sigmoid}(\langle \mathbf{w}|\mathbf{x}_t\rangle)\mathbf{x}_t \otimes \mathbf{x}_t.$$

Example 2 (Probit Function) *We now assume that* $p(\langle \mathbf{w}|\mathbf{x}_t\rangle) = \Phi(-\langle \mathbf{w}|\mathbf{x}_t\rangle)$ *where* $\Phi(w)$ *is the CDF of the standard normal distribution. Noting that* $\phi(w) = \Phi'(w) = (2\pi)^{-1/2}\exp\big(-w^2/2\big)$ *we find*

$$\mathbf{B}(\mathbf{w}) = \sum_{t=1}^{T} \frac{\phi^2(\langle \mathbf{w}|\mathbf{x}_t\rangle)}{\Phi(\langle \mathbf{w}|\mathbf{x}_t\rangle)\Phi(-\langle \mathbf{w}|\mathbf{x}_t\rangle)} \mathbf{x}_t \otimes \mathbf{x}_t.$$

We should add that we are not aware of any precise results for the probit and that precise analysis of the Gaussian regression is an order of magnitude harder than the logistic regression, as can be seen in the proof of Theorem 2 in the next section.

In passing we observe that if data $\mathbf{X}_t$ is generated by a stationary ergodic source, then by the *ergodic theorem* we conclude that with high probability (whp) $\frac{1}{T}\mathbf{B}(\mathbf{w}, \mathbf{X}^T) \to \mathbf{E}_X[\mathbf{B}(\mathbf{w}, \mathbf{X})] := \bar{\mathbf{B}}(\mathbf{w})$ when $T \to \infty$. Therefore the discrepancy $C_d(p, \mathbf{x}^T)$ satisfies in probability

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

See Jacquet et al. (2022) for results on the regret for the logistic regression when $\mathbf{x}_t$ lie on a sphere.

2.3. Minimax Regret on Manifolds

In this section we consider the minimax regret on manifolds $\mathcal{M} \subset \mathbb{R}^d$ of dimension $d-1$. This is required in quantum setup where quantum data and weights must lie on a unit sphere $\mathcal{S}_d$. As a consequence, hypothesis (H) cannot hold as shown in below lemma with the proof in the Appendix.

Lemma 3 *For* $\mathbf{w} \in \mathcal{M}$ *with non zero curvature (e.g., sphere) the functions* $-\log p(\langle \mathbf{w}|\mathbf{x}\rangle)$ *and* $-\log(1 - p(\langle \mathbf{w}|\mathbf{x}\rangle))$ *cannot be both convex.*

As a result of Lemma 3 we may have multiple $\mathbf{w}^*$ maximizing $P(y^T|\mathbf{w})$, all of them satisfying $\nabla L(y^T|\mathbf{w}) = 0$. Therefore, Lemma 1 can only give us an upper bound. In the Appendix we prove our second main result.

Theorem 4 *Consider* $\mathbf{w} \in \mathcal{M}$ *of dimension* $(d-1)$*, and bounded* $\mathbf{x}_t$ *spanning* $\mathbb{R}^d$*. Then asymptotically for* $d = o(T^{1/3})$

$$r^*(\mathbf{x}^T) \le \frac{d-1}{2}\log T - \frac{d-1}{2}\log 2\pi + C_d(p, \mathbf{x}^T) + O(d^{3/2}/\sqrt{T}) \tag{15}$$

where $C_d(p, \mathbf{x}^T)$ *is (below determinant is to be understood as* $(d-1)\times(d-1)$ *determinant)*

$$C_d(p, \mathcal{M}, \mathbf{x}^T) = \log\left(\int_{\mathcal{M}} \sqrt{\left|\det\left(\frac{1}{T}\Pi_{\mathcal{M}}(\mathbf{w})\mathbf{B}_d(\mathbf{w}, \mathbf{x}^T)\Pi_{\mathcal{M}}(\mathbf{w})\right)\right|}d\mathbf{w}\right) \tag{16}$$

where $\Pi_{\mathcal{M}}(\mathbf{w})$ *is the orthogonal projection at* $\mathbf{w}$ *on the tangent hyperplane to* $\mathcal{M}$ *and* $\mathbf{B}_d(\mathbf{w}, \mathbf{x}^T)$ *is given in (14). We also have as* $\Pi_{\mathcal{M}}(\mathbf{w}) = \mathbf{I}_d - \mathbf{u}\otimes\mathbf{u}$ *where* $\mathbf{u}$ *is the orthonormal vector to the tangent hyperplane where* $\mathbf{I}_d$ *is the identity operator.*

Example 3 (Regret analysis on the Sphere) *Consider* $p(\langle \mathbf{w}, \mathbf{x}\rangle) = (1 + e^{(2|\langle \mathbf{w},\mathbf{x}\rangle|^2 - 1)})^{-1}$ *for* $\mathbf{w}$ *and* $\mathbf{x}$ *on the unite sphere in* $\mathbb{R}^d$ *(motivated by quantum logistic regression). Then, the theorem gives an upper bound on the minimax regret with* $\mathcal{M} = \mathcal{S}_d$*, the unite sphere in* $\mathbb{R}^d$*, and*

$$\mathbf{B}_d(\mathbf{w}, \mathbf{x}^T) = 16\sum_{t=1}^{T} |\langle \mathbf{x}_t|\mathbf{w}\rangle|^2 e^{(2|\langle \mathbf{x}_t|\mathbf{w}\rangle|^2 - 1)}\mathbf{x}_t \otimes \mathbf{x}_t.$$

Online Learning of Polarized Photons. We explore applications of our analyses to online learning of quantum states, however, here we consider only a single photon. This problem is the online variant of recent models for classification of quantum states Heidari et al. (2021, 2022). Specifically, in optical systems (or spin systems), the observation at each round t is a polarized photon accompanied with a classical label $y_t \in \{-1, 1\}$ which yet to be revealed. The label y_t contains one bit of information about the photon (e.g., being "single" or "non-single" (Kudyshev et al., 2020)). The polarized photon is represented by a quantum state $|\mathbf{x}_t\rangle, t > 0$ in an underlying two-level Hilbert space. Without loss of generality, this state is in the superposition form in the computational basis as

$$|\mathbf{x}_t\rangle = x_{t,0}\,|0\rangle + x_{t,1}\,|1\rangle\,,$$

where $(x_{t,0}, x_{t,1}) \in \mathbb{C}^2$ with $\|\mathbf{x}_t\| = 1$. The objective is to seek a quantum procedure for determining the best direction $\mathbf{w}$ in the Bloch sphere for polarization measurement to recover the unknown label y_t (see Figure 1). This is done by processing the past observations up to time t as in classical online learning. However, the main difference is that the photons $|\mathbf{x}_t\rangle, t > 0$ are observable only through quantum measurements — hence abiding the quantum postulates. Therefore, the predictors in this model are quantum measurements that act on the photons and produce a prediction. Unlike classical setting, the outcome is inherently random because of the quantum measurements indeterminism.

The POVM[1] representation of a measurement in the direction of $\mathbf{w}$ is given as $M_{\mathbf{w}} := \{|\mathbf{w}\rangle\langle\mathbf{w}|, \mathbf{I} - |\mathbf{w}\rangle\langle\mathbf{w}|\}$, where $|\mathbf{w}\rangle = w_0\,|0\rangle + w_1\,|1\rangle$ with $\mathbf{w} = (w_0, w_1) \in \mathbb{C}^2$ and $\|\mathbf{w}\| = 1$. Moreover, the class of predictors is the collection of all such measurements that is denoted as $\mathcal{C} := \{M_{\mathbf{w}} : \mathbf{w} \in \mathbb{C}^2\}$.

The probability that a measurement $M_{\mathbf{w}}$ is correct is $|\langle\mathbf{w}|\mathbf{x}_t\rangle|^2$ when $y_t = 1$ and $1 - |\langle\mathbf{w}|\mathbf{x}_t\rangle|^2$ when $y_t = -1$. Thus the system is equivalent to a generalized regression with probability function $p(x) = |x|^2$ (notice that the variable is now complex). Furthermore, note that the quantum state formulation necessitates that vectors $\mathbf{w}$ and $\mathbf{x}_t$ have unit norm. Hence, the regret analysis operates on the Bloch sphere as a manifold which is the complex unit circle. We now present the regret for this setup with the log of the error probability as the measure of the loss.

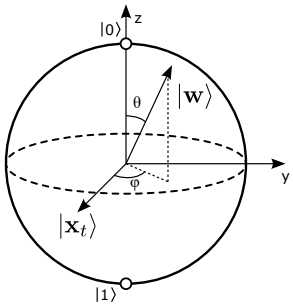

Figure 1: Bloch sphere and measurement of $\mathbf{w}$.

Corollary 5 *When the loss function is* $\ell(y_t, \mathbf{w}|\mathbf{x}_t) = -\log P(\hat{Y}_t \neq y_t|\mathbf{x}_t, \mathbf{w}_t) = -\log p(\langle\mathbf{w}, \mathbf{x}\rangle) - \log(1 - p(\langle\mathbf{w}, \mathbf{x}\rangle))$ *with* $p(\langle\mathbf{w}, \mathbf{x}\rangle) = |\langle\mathbf{x}_t|\mathbf{w}\rangle|^2$ *the quantum minimax regret becomes*

$$r^*(|\mathbf{x}\rangle^T) \leq \log T + \log\left(\int_{\mathcal{S}_b} \sqrt{\left|\det\left(\frac{1}{T}\Pi_{\mathbf{w}}\mathbf{B}(\mathbf{w}, \mathbf{x}^T)\Pi_{\mathbf{w}}\right)\right|}\, d\mathbf{w}\right) + O(1) \tag{17}$$

where $\mathcal{S}_b$ *is the Bloch sphere,* $\Pi_{\mathbf{w}} = \mathbf{I} - |\mathbf{w}\rangle\langle\mathbf{w}|$ *and* $\mathbf{B}(\mathbf{w}, \mathbf{x}^T) = \sum_t \frac{4}{1-|\langle\mathbf{x}_t|\mathbf{w}\rangle|^2}\,|\mathbf{x}_t\rangle\,\langle\mathbf{x}_t|\,.$

The theorem is proved by noting that $\mathbf{w}$ belongs to the Bloch sphere which is a two dimensional manifold inside a three dimensional space (Figure 1). Hence, we can apply Theorem 4 with $d = 3$.

1. For detailed formulation of quantum measurements see (Michael A. Nielsen, 2010).

3. Proof of Theorem 2

The proof of Theorem 2 is quite complex. It uses several analytic tools such as multidimensional Fourier transform and Laplace method combined with probabilistic and combinatorial analyzes. We split the proof in three parts: First we prove Lemma 1, and then we consider the special case of the logistic regression that entails simpler analysis, and finally we consider the general case.

Proof of Lemma 1. To estimate the Shtarkov sum (7) we need $P^*(y^T|\mathbf{x}^T) := \sup_{\mathbf{w}} P(y^T|\mathbf{x}^T, \mathbf{w})$ as expressed in (10). Let the maximum of $P(y^T|\mathbf{x}^T, \mathbf{w})$ be attained at $\mathbf{w}^* := \mathbf{w}^*(y^T, \mathbf{x}^T)$, hence $\nabla L(y^T|\mathbf{x}^T, \mathbf{w}^*) = 0$. Under hypothesis (H) such a solution exists and unique, even if some components of $\mathbf{w}^*$ might be infinite leading to $P^*(y^T|\mathbf{x}^T) = 1$ for some y^T that do not change asymptotically the Shtarkov sum. Clearly, we have from (10) for every $\mathbf{x}^T$ and $\mathbf{w}$

$$\nabla L(y^T|\mathbf{x}^T, \mathbf{w}) = -\sum_{t:y_t<0} \frac{p'(\langle \mathbf{w}|\mathbf{x}\rangle)}{p(\langle \mathbf{w}|\mathbf{x}\rangle)}\mathbf{x}_t + \sum_{t:y_t>0} \frac{p'(\langle \mathbf{w}|\mathbf{x}\rangle)}{1-p(\langle \mathbf{w}|\mathbf{x}\rangle)}\mathbf{x}_t. \tag{18}$$

For fixed $\mathbf{x}^T$, as assumed in this paper, we often omit $\mathbf{x}^T$; e.g., we write $L(y^T|\mathbf{w})$ and $\mathbf{B}(\mathbf{w})$ for $\mathbf{B}_d(\mathbf{w}, \mathbf{x}^T)$. We also define $\mathbf{G}_{y^T}(\mathbf{w}) = \nabla L(y^T|\mathbf{w})$. Observe that the optimal $\mathbf{w}^*$ equals to $\mathbf{w}^* = \mathbf{G}^{-1}_{y^T}(0)$. Let $h_{y^T}(\mathbf{a}) = \exp(L(y^T|\mathbf{G}^{-1}_{y^T}(\mathbf{a}))$ and $\widetilde{h}_{y^T}(z)$ be its Fourier transform, then

$$\widetilde{h}_{y^T}(z) = \int_{\mathbb{R}^d} h_{y^T}(\mathbf{a}) e^{-i\langle \mathbf{a}|\mathbf{z}\rangle} d\mathbf{a} = \int_{\mathbb{R}^d} \exp\big(-L(y^T|\mathbf{w})\big)\det(\nabla \mathbf{G}_{y^T}(\mathbf{w}))e^{-i\langle \mathbf{G}_{y^T}(\mathbf{w})|\mathbf{z}\rangle} d\mathbf{w}.$$

The inverse Fourier is then

$$h_{y^T}(\mathbf{a}) = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \widetilde{h}_{y^T}(\mathbf{z}) e^{\langle \mathbf{a}|\mathbf{z}\rangle} d\mathbf{z} \tag{19}$$

which allows us to compute $P(y^T|\mathbf{w}^*)$ as

$$\begin{aligned} P(y^T|\mathbf{w}^*) =& \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \widetilde{h}_{y^T}(\mathbf{z}) d\mathbf{z} \\ =& \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d\times\mathbb{R}^d} \exp\big(-L(y^T|\mathbf{w})\big)\det(\nabla \mathbf{G}_{y^T}(\mathbf{w}))\exp\big(-i\langle \mathbf{G}_{y^T}(\mathbf{w})|\mathbf{z}\rangle\big) d\mathbf{w} d\mathbf{z}. \end{aligned}$$

We finish the proof by noting that $\nabla L(y^T|\mathbf{w}) = \mathbf{G}_{y^T}(\mathbf{w})$ and $\nabla^2 L(y^T|\mathbf{w}) = \nabla \mathbf{G}_{y^T}(\mathbf{w})$.

Logistic Regression. We continue with the proof of Theorem 2 but it is now convenient to split it in two parts. In the first part, which we discuss here, we assume that $p(w) = (1+e^{-w})^{-1}$, that is, $p(\langle \mathbf{w}|\mathbf{x}\rangle)$ is the logistic regression. We first notice that hypothesis (H) holds. Indeed $\frac{d^2}{dw^2}\log p(w) = \frac{d^2}{dw^2}\log(1-p(w)) = p(w)(1-p(w)) < 0$. More importantly, the Laplacian $\nabla^2 L(y^T|\mathbf{w})$ does not depend on the label sequence y^T and is a function of only $\mathbf{w}$. Indeed

$$\nabla^2 \log p(\langle \mathbf{w}|\mathbf{x}\rangle) = \nabla^2 \log(1-p(\langle \mathbf{w}|\mathbf{x}\rangle)) = -\frac{1}{(1+e^{-\langle \mathbf{w}|\mathbf{x}\rangle})(1+e^{\langle \mathbf{w}|\mathbf{x}\rangle})}\mathbf{x}_t \otimes \mathbf{x}_t, \tag{20}$$

that is, $\nabla^2 \log p(\langle \mathbf{w}|\mathbf{x}\rangle) = -\frac{(p'(\langle \mathbf{w}|\mathbf{x}\rangle))^2}{(1-p(\langle \mathbf{w}|\mathbf{x}\rangle))p(\langle \mathbf{w}|\mathbf{x}\rangle)}$ leading to

$$\nabla^2 L(y^T|\mathbf{w}) = \mathbf{B}(\mathbf{w}) \tag{21}$$

where $\mathbf{B}(\mathbf{w})$ is defined in (14). This is crucial to analyze the Shtarkov sum as expressed in (7). To see this, let us define

$$S(\mathbf{x}^T|\mathbf{w}) = \sum_{y^T} \det(\nabla^2 L(y^T|\mathbf{w})) \int_{\mathbb{R}^d} P(y^T|\mathbf{w}) \exp\big(-i\langle \nabla L(y^T|\mathbf{w})|\mathbf{z}\rangle\big) d\mathbf{z}$$

where we note that $P(y^T|\mathbf{w}) = \exp\big(L(y^T|\mathbf{x})\big)$. Observe that by Lemma 1 we have

$$S(\mathbf{x}^T) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} S(\mathbf{x}^T|\mathbf{w}) d\mathbf{w}.$$

To compute $S(\mathbf{x}^T|\mathbf{w})$ we proceed as follows. We first notice that due to property (21) of the logistic regression, we can write

$$S(\mathbf{x}^T|\mathbf{w}) = \det(B(\mathbf{w})) \int_{\mathbb{R}^d} \sum_{y^T} P(y^T|\mathbf{w}) \exp\big(-i\langle \nabla L(y^T|\mathbf{w})|\mathbf{z}\rangle\big) d\mathbf{z}. \tag{22}$$

But

$$P(y^T|\mathbf{w}) \exp\big(-i\langle \nabla L(y^T|\mathbf{w})|\mathbf{z}\big) = \prod_{t:y_t<0} p(\langle \mathbf{w}|\mathbf{x}_t\rangle) e^{-i\frac{p'(\langle \mathbf{w}|\mathbf{x}_t\rangle)}{p(\langle \mathbf{w}|\mathbf{x}_t\rangle)}\langle \mathbf{x}_t|\mathbf{z}\rangle} \prod_{t:y_t>0} (1-p(\langle \mathbf{w}|\mathbf{x}_t\rangle)) e^{i\frac{p'(\langle \mathbf{w}|\mathbf{x}_t\rangle)}{1-p(\langle \mathbf{w}|\mathbf{x}_t\rangle)}\langle \mathbf{x}_t|\mathbf{z}\rangle}$$

leading to

$$\sum_{y^T} P(y^T|\mathbf{w}) \exp\big(-i\langle \nabla L(y^T|\mathbf{w})|\mathbf{z}\big) = \\ = \prod_t \left(p(\langle \mathbf{w}|\mathbf{x}_t\rangle) e^{-i\frac{p'(\langle \mathbf{w}|\mathbf{x}_t\rangle)}{p(\langle \mathbf{w}|\mathbf{x}_t\rangle)}\langle \mathbf{x}_t|\mathbf{z}\rangle} + (1-p(\langle \mathbf{w}|\mathbf{x}_t\rangle)) e^{i\frac{p'(\langle \mathbf{w}|\mathbf{x}_t\rangle}{1-p(\langle \mathbf{w}|\mathbf{x}_t\rangle)}\langle \mathbf{x}_t|\mathbf{z}\rangle} \right).$$

In summary, the following lemma completes the proof of Theorem 2 for the logistic regression.

Lemma 6 *Under the assumptions of Theorem 2 when* $p(w) = (1+e^{-w})^{-1}$ *we find for* $d = o(T^{1/3})$

$$S(\mathbf{x}^T|\mathbf{w}) = (2\pi)^{d/2} \sqrt{\det(\mathbf{B}(w)} \left(1 + \frac{d^{3/2}}{\sqrt{T}}\right), \tag{23}$$

$$S(\mathbf{x}^T) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} S(\mathbf{x}^T|\mathbf{w}) d\mathbf{w} = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \sqrt{\det(\mathbf{B}(\mathbf{w}))} d\mathbf{w} \left(1 + \frac{d^{3/2}}{\sqrt{T}}\right). \tag{24}$$

Proof To evaluate $S(\mathbf{x}^T|\mathbf{w})$ we focus on estimating the following

$$S_1(\mathbf{x}^T|\mathbf{w}) = \int_{\mathbb{R}^d} \prod_t f(\mathbf{x}_t, \mathbf{z}) d\mathbf{z} = \int_{\mathbb{R}^d} \exp\left(\sum_t \log f(\mathbf{x}_t, \mathbf{w})\right) d\mathbf{z}$$

where

$$f(\mathbf{x}, \mathbf{z}) = p(\langle \mathbf{w}|\mathbf{x}\rangle) e^{-i\frac{p'(\langle \mathbf{w}|\mathbf{x}\rangle)}{p(\langle \mathbf{w}|\mathbf{x}\rangle)}\langle \mathbf{x}|\mathbf{z}\rangle} + (1-p(\langle \mathbf{w}|\mathbf{x}\rangle)) e^{i\frac{p'(\langle \mathbf{w}|\mathbf{x}\rangle)}{1-p(\langle \mathbf{w}|\mathbf{x}\rangle)}\langle \mathbf{x}|\mathbf{z}\rangle}. \tag{25}$$

To evaluate $S_1(\mathbf{x}^T|\mathbf{w})$ we apply the multidimensional Laplace method. We observe the function $\log f(\mathbf{x},\mathbf{z})$ of $\mathbf{z}$ attains its maximal value at $\mathbf{z}=0$, with maximum value 1, hence maximum of $\log f(\mathbf{x},\mathbf{z})$ is 0. The Laplacian at $\mathbf{z}=0$ is $-\frac{(p'(\langle\mathbf{w}|\mathbf{x}\rangle))^2}{(1-p(\langle\mathbf{w}|\mathbf{x}\rangle)p(\langle\mathbf{w}|\mathbf{x}\rangle)}\mathbf{x}\otimes\mathbf{x}$ which after summing over the $\mathbf{x}_t$ is exactly the opposite of $\nabla^2 L(y^T|\mathbf{w})$, that is, $-\mathbf{B}(\mathbf{w})$. Now the Taylor expansion of $\sum_t \log f(\mathbf{x}_t,\mathbf{z})$ is

$$\sum_{t=1}^{T}\log f(\mathbf{x}_t,\mathbf{z}) = -\frac{1}{2}\langle \mathbf{z}\mathbf{B}(\mathbf{w})|\mathbf{z}\rangle + O(T\|\mathbf{z}\|^3) \tag{26}$$

as long as $\mathbf{x}_t$ is uniformly bounded. Therefore, by Inglot and Majewski (2014) we arrive at

$$S_1(\mathbf{x}^T|\mathbf{w}) = \int_{\mathbb{R}^d}\exp\left(\sum_t \log f(\mathbf{x}_t,\mathbf{w})\right)d\mathbf{z} = \frac{(2\pi)^{d/2}}{\sqrt{\det\mathbf{B}(\mathbf{w})}}\left(1+\frac{d^{3/2}}{\sqrt{T}}\right) \tag{27}$$

proving (23) and (24) after using (22). The error term estimation is discussed in Appendix D. ■

General Case. In the logistic case for all y^T we have $\nabla^2 L(y^T|\mathbf{w}) = \mathbf{B}(\mathbf{w})$. In general, this is not the case. In fact, $\nabla^2 L(y^T|\mathbf{w})$ varies with y^T, and when it is viewed as a random variable we have $\mathbb{E}[Y_t] = 1-2p(w)$ and $\mathbb{E}[\nabla^2 L(y^T|\mathbf{w})] = \mathbf{B}(\mathbf{w})$. Indeed,

$$\mathbb{E}[\nabla^2 L(y^T|\mathbf{w})] = \sum_t p(\langle\mathbf{w}|\mathbf{x}_t\rangle)\left((p'(\langle\mathbf{w}|\mathbf{x}_t\rangle))^2\frac{1}{(p(\langle\mathbf{w}|\mathbf{x}_t\rangle))^2} - p''(\langle\mathbf{w}|\mathbf{x}_t\rangle)\frac{1}{p(\langle\mathbf{w}|\mathbf{x}_t\rangle)}\right)\mathbf{x}_t\otimes\mathbf{x}_t$$
$$+\sum_t(1-p(\langle\mathbf{w}|\mathbf{x}_t\rangle))\left((p'(\langle\mathbf{w}|\mathbf{x}_t\rangle))^2\frac{1}{(1-p(\langle\mathbf{w}|\mathbf{x}_t\rangle))^2} + p''(\langle\mathbf{w}|\mathbf{x}_t\rangle)\frac{1}{1-p(\langle\mathbf{w}|\mathbf{x}_t\rangle)}\right)\mathbf{x}_t\otimes\mathbf{x}_t$$
$$=\sum_t\frac{(p'(\langle\mathbf{w}|\mathbf{x}_t\rangle)^2}{(1-p(\langle\mathbf{w}|\mathbf{x}_t\rangle))p(\langle\mathbf{w}|\mathbf{x}_t\rangle)} = \mathbf{B}(\mathbf{w}).$$

We now split $\nabla^2 L(y^T|\mathbf{w})$ as $\nabla^2 L(y^T|\mathbf{w}) = \mathbf{A}(\mathbf{w}) + \mathbf{F}(y^T)$ with $\mathbf{F}(y^T) = \sum_{t=1}^T y_t\mathbf{F}_t$ where

$$\mathbf{A}(\mathbf{w}) = \frac{1}{2}\sum_t\left(\frac{(p'(\langle\mathbf{w}|\mathbf{x}_t\rangle))^2}{(p(\langle\mathbf{w}|\mathbf{x}_t\rangle))^2} + \frac{(p'(\langle\mathbf{w}|\mathbf{x}_t\rangle))^2}{(1-p(\langle\mathbf{w}|\mathbf{x}_t\rangle))^2} - p''(\langle\mathbf{w}|\mathbf{x}_t\rangle)\left(\frac{1}{p(\langle\mathbf{w}|\mathbf{x}_t\rangle)} - \frac{1}{1-p(\langle\mathbf{w}|\mathbf{x}_t\rangle)}\right)\right)\mathbf{x}_t\otimes\mathbf{x}_t$$
$$\mathbf{F}_t = \frac{1}{2}\left(\frac{-(p'(\langle\mathbf{w}|\mathbf{x}_t\rangle))^2}{(p(\langle\mathbf{w}|\mathbf{x}_t\rangle))^2} + \frac{(p'(\langle\mathbf{w}|\mathbf{x}_t\rangle))^2}{(1-p(\langle\mathbf{w}|\mathbf{x}_t\rangle))^2} + p''(\langle\mathbf{w}|\mathbf{x}_t\rangle)\left(\frac{1}{p(\langle\mathbf{w}|\mathbf{x}_t\rangle)} + \frac{1}{1-p(\langle\mathbf{w}|\mathbf{x}_t\rangle)}\right)\right)\mathbf{x}_t\otimes\mathbf{x}_t.$$

We also have the following identity $\mathbb{E}[\mathbf{F}(y^T)] = \sum_{t=1}^T(1-2p(\langle\mathbf{w}|\mathbf{x}_t\rangle))\mathbf{F}_t$. We now need a large deviation result for $\nabla^2 L(y^T|\mathbf{w})$ or $\mathbf{F}(y^T)$ proved in the next lemma.

Lemma 7 *Under hypothesis (H), $\mathbf{F}^T = \{\mathbf{F}_t\}_{t\le T}$ is a uniformly bounded sequence for all T, and for any $A>0$ and $\alpha>1/2$ there exists B such that*

$$P\left(\|\mathbf{F}(y^T) - \mathbb{E}[\mathbf{F}(y^T)]\| > AdT^\alpha\right) \le 2d^2 e^{-BT^{2\alpha-1}} \tag{28}$$

where $\|.\|$ is an arbitrary metric.

Proof Let $\mathbf{\Theta}$ be a complex $d\times d$ matrix, and let $\widetilde{\mathbf{F}}_T(\mathbf{\Theta})$ be the Laplace transform of $\mathbf{F}(y^T)$, that is, $\widetilde{\mathbf{F}}_T(\mathbf{\Theta}) = \mathbb{E}[e^{\mathrm{Tr}(\mathbf{\Theta}\mathbf{F}(y^T))}]$, where the trace $\mathrm{Tr}(\cdot)$ is the classical expression for the dot product of matrices. We have the expression when y^T is viewed as a random variable with probability $P(y^T|\mathbf{w})$

$$\widetilde{\mathbf{F}}_T(\mathbf{\Theta}) = \prod_t\left(p(\langle\mathbf{w}|\mathbf{x}_t\rangle)e^{-\mathrm{Tr}(\mathbf{\Theta}\mathbf{F}_t)} + (1-p(\langle\mathbf{w}|\mathbf{x}_t\rangle))e^{\mathrm{Tr}(\mathbf{\Theta}\mathbf{F}_t)}\right). \tag{29}$$

We will show that there exists a simply connected complex neighborhood $\mathcal{U}$ of the null matrix, such that for all T the following $\boldsymbol{\Theta} \in \mathcal{U}$ implies that $\log \widetilde{\mathbf{F}}_T(\boldsymbol{\Theta})$ exists and is uniformly $O(T)$. Indeed by rewriting (29)

$$p(\langle \mathbf{w}|\mathbf{x}_t\rangle)e^{-\mathrm{Tr}(\boldsymbol{\Theta}\mathbf{F}_t)} + (1-p(\langle \mathbf{w}|\mathbf{x}_t\rangle))e^{\mathrm{Tr}(\boldsymbol{\Theta}\mathbf{F}_t)} = e^{\mathrm{Tr}(\boldsymbol{\Theta}\mathbf{F}_t)}\left(1+p(\langle \mathbf{w}|\mathbf{x}_t\rangle)(e^{-2\mathrm{Tr}(\boldsymbol{\Theta}\mathbf{F}_t)}-1)\right)$$

and notice that $\|\mathbf{F}_t\| < F$ for some $F > 0$ (here we take $\|\mathbf{F}_t\| = \sqrt{\mathrm{Tr}(\mathbf{F}_t^2)}$ with the classic matrix dot product expression). Then taking $\|\boldsymbol{\Theta}\| < \frac{1}{20F}$ we have

$$\left|p(\langle \mathbf{w}|\mathbf{x}_t\rangle)(e^{-2\mathrm{Tr}(\boldsymbol{\Theta}\mathbf{F}_t)}-1)\right| < 1/2 \quad \text{and} \quad p(\langle \mathbf{w}|\mathbf{x}_t\rangle)(e^{-2\mathrm{Tr}(\boldsymbol{\Theta}\mathbf{F}_t)}-1) \neq 0.$$

Since $\mathcal{U}$ is simply connected, the logarithm of $\widetilde{\mathbf{F}}_T(\boldsymbol{\Theta})$ exists, and it turns out that the logarithm is always bounded by some $C > 0$, hence the logarithm of the product satisfies: $|\log \widetilde{\mathbf{F}}(\boldsymbol{\Theta})| \leq TC$. As a consequence, $\log \widetilde{\mathbf{F}}(\boldsymbol{\Theta})$ is an analytic function on $\mathcal{U}$ and its derivatives are also $O(T)$, in particular the second derivative. Since the first derivative of $\log \widetilde{\mathbf{F}}_T(\boldsymbol{\Theta})$ at $\boldsymbol{\Theta} = 0$ is exactly $\mathbb{E}[\mathbf{F}(y^T)]$, we have the following Taylor expansion

$$\log \widetilde{\mathbf{F}}(\boldsymbol{\Theta}) = \mathrm{Tr}\left(\boldsymbol{\Theta}\mathbb{E}[\mathbf{F}(y^T)]\right) + O(T\|\boldsymbol{\Theta}\|^2) \tag{30}$$

with $O(T\|\boldsymbol{\Theta}\|^2) \leq RT\|\boldsymbol{\Theta}\|^2$ for some $R > 0$. We will use this estimate via the Chebychev inequality. Having $\|\mathbf{F}(y^T) - \mathbb{E}[\mathbf{F}(y^T)]\| > AdT^\alpha$ implies to have one of d^2 component of $\mathbf{F}(y^T) - \mathbb{E}[\mathbf{F}(y^T)]$ greater than AT^α or smaller than $-AT^\alpha$. For $(i,j) \in \{1,\ldots,d\}^2$ let $\mathbf{F}(y^T)_{ij}$ denote the ij component of $\mathbf{F}(y^T)$. We look at $P\left(\mathbf{F}(y^T)_{ij} > \mathbb{E}[\mathbf{F}(y^T)]_{ij} + AT^\alpha\right)$. If $\mathbf{e}_{ij}$ is the matrix with all components equal to zero, except the (i,j)-th element which is equal to 1, then $\mathbf{F}(y^T)_{ij} = \mathrm{Tr}(\mathbf{e}_{ij}\mathbf{F}(y^T))$ and for all $\theta > 0$ by Chebychev inequality

$$P\left(\mathbf{F}(y^T)_{ij} > \mathbb{E}[\mathbf{F}(y^T)]_{ij} + AT^\alpha\right) \leq \frac{\mathbb{E}[e^{\theta \mathrm{Tr}(\mathbf{e}_{ij}\mathbf{F}(y^T))}]}{\exp(\theta \mathrm{Tr}(\mathbf{e}_{ij}\mathbf{F}(y^T)) + \theta AT^\alpha)}.$$

Since $\mathbb{E}[e^{\theta \mathrm{Tr}(\mathbf{e}_{ij}\mathbf{F}(y^T))}] = \widetilde{\mathbf{F}}(\theta \mathbf{e}_{ij})$, and thanks to the estimate (30) with $\|\mathbf{e}_{ij}\|^2 = 1$, the right-hand side is upper bounded by $\exp\left(RT\theta^2 - \theta AT^\alpha\right)$ with the minimum value $\exp\left(-\frac{A^2T^{2\alpha-1}}{4R}\right)$. We also have $P\left(\mathbf{F}(y^T)_{ij} < \mathbb{E}[\mathbf{F}(y^T)]_{ij} - AT^\alpha\right)$ but with $\theta < 0$. This concludes the proof with $B = \frac{A}{4R}$. ■

To complete the proof of Theorem 2 we notice that $\mathbb{E}[\mathbf{F}(y^T)] = \mathbf{B}(\mathbf{w}) - \mathbf{A}(\mathbf{w})$ and by Lemma 7

$$P\left(\|\nabla^2 L(y^T|\mathbf{w}) - \mathbf{B}(\mathbf{w}))\| > AdT^\alpha\right) \leq 2d^2 e^{-BT^{2\alpha-1}}.$$

Now $\nabla^2 L(y^T|\mathbf{w})$ is of order T and $\det(\nabla^2 L(y^T|\mathbf{w}))$ of order T^d. With probability greater than $1 - 2d^2 e^{-BT^{2\alpha-1}}$ we find that $\|\nabla^2 L(y^T|\mathbf{w}) - \mathbf{B}(\mathbf{w})\|$ is of order dT^α (with $\alpha < 1$) which implies that

$$\det(\nabla^2 L(y^T|\mathbf{w})) = \det(\mathbf{B}(\mathbf{w})(1 + O(dT^{\alpha-1})). \tag{31}$$

Let now $\mathcal{J}_\alpha$ be the set of sequences y^T satisfying (31). We already know that

$$\sum_{y^T \notin \mathcal{J}_\alpha} P(y^T|\mathbf{w}) < 2d^2 e^{-BT^{2\alpha-1}}$$

which exponentially decays since $\alpha > 1/2$. Therefore, the difference between $\det(\nabla^2 L(y^T|\mathbf{w}))$ and $\det(\mathbf{B}(\mathbf{w}))$ is of order $2d^2 e^{-BT^{2\alpha-1}}$. This leads to the following

$$\begin{aligned} S(\mathbf{x}^T|\mathbf{w}) &= \frac{1}{(2\pi)^d} \sum_{y^T} \det(\nabla^2 L(y^T|\mathbf{w})) \int_{\mathbb{R}^d} \exp\big(-L(y^T|\mathbf{w}) - i\langle \nabla L(y^T|\mathbf{w})|\mathbf{z}\rangle\big) d\mathbf{z} \\ &= \frac{\det(\mathbf{B}(\mathbf{w}))}{(2\pi)^d} \sum_{y^T} \int_{\mathbb{R}^d} \exp\big(-L(y^T|\mathbf{w}) - i\langle \nabla L(y^T|\mathbf{w})|\mathbf{z}\rangle\big) d\mathbf{z} \left(1 + O(d^2 e^{-BT^{2\alpha-1}})\right). \end{aligned}$$

Then as in the proof of Lemma 6 (see also Appendix D) the right-hand side of above is equal to $\frac{\sqrt{\det(\mathbf{B}(\mathbf{w}))}}{\sqrt{2\pi}^d}\left(1 + O(\frac{d^{3/2}}{\sqrt{T}})\right)$ which concludes the proof of Theorem 2.

References

J. Aldrich, F. Nelson, and S. Adler. Linear probability, logit, and probit models. *Sage*, 1984.

F. Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414, 2010.

A. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Trans. Inf. Theory*, 44(6):2743–2760, Oct. 1998.

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

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

Nicolo Cesa-Bianchi and Ohad Shamir. Efficient transductive online learning via randomized rounding. *arXiv preprint arXiv:1106.2429*, 2011.

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

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.

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.

Mohsen Heidari, Arun Padakandla, and Wojciech Szpankowski. A theoretical framework for learning from quantum data. In *IEEE International Symposium on Information Theory (ISIT)*, 2021.

Mohsen Heidari, Ananth Grama, and Wojciech Szpankowski. Toward physically realizable quantum neural networks. In *AAAI*, 2022.

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

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

P. Jacquet, G. I. Shamir, and W. Szpankowski. Precise minimax regret for logistic regression with categorical feature values. In *PRML: ALT'21*, volume 132, pages 755–771, 2021.

P. Jacquet, G. Shamir, and W. Szpankowski. Precise minimax regret for logistic regression, 2022.

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

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

Zhaxylyk A. Kudyshev, Simeon I. Bogdanov, Theodor Isacsson, Alexander V. Kildishev, Alexandra Boltasseva, and Vladimir M. Shalaev. Rapid classification of quantum sources enabled by machine learning. *Advanced Quantum Technologies*, 3(10):2000067, sep 2020. doi: 10.1002/qute.202000067.

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.

Isaac L. Chuang Michael A. Nielsen. *Quantum Computation and Quantum Information*. Cambridge University Pr., December 2010. ISBN 1107002176.

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 with general los function. In *COLT*, 2014.

Alexander Rakhlin and Karthik Sridharan. Sequential probability assignment with binary alphabets and large classes of experts. *arXiv preprint arXiv:1501.07340*, 2015.

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

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

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

G. 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. Logistic regression regret: What's the catch? In *COLT*, 2020.

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

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

D.J. Struik. *Lectures on Classical Differential Geometry*. Dover, 1968.

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.

C. Wu, M. Heidari, A. Grama, and W. Szpankowski. Sequential vs fixed design regrets in online learning, 2022.

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

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

Appendix A. A Representation of the Logarithmic Loss

In this Appendix we show the equivalence between the loss $\ell(p,y) = -\log|(1+y)/2 - p|$ for $y \in \{-1,1\}$ (often used in ML) and $-\log P(Y=y)$ that is used to define the minimax regret in this paper.

Lemma 8 *Let $\ell : [0,1] \times \{-1,1\} \to \mathbb{R}^+$ be the logarithmic loss function defined as*

$$\ell(p,y) = -\log|\frac{1+y}{2} - p|$$

for all $p \in [0,1]$ and $y \in \{-1,1\}$. Then,

$$\ell(p,y) = -\log P(Y=y),$$

where Y is a random variable with bias $P(Y=-1) = p$.

Proof Let Y be the random variable in the statement. Note that the PMF of Y can be written as

$$P(Y=y) = \mathbb{1}\{y=1\}(1-p) + \mathbb{1}\{y=-1\}p, \qquad \forall y \in \{-1,1\}.$$

On the other hand, the logarithmic loss equals to the following

$$\begin{aligned} -\log|\frac{1+y}{2} - p| &= -\mathbb{1}\{y=1\}\log(1-p) - \mathbb{1}\{y=-1\}\log(p) \\ &= -\log\Big(\mathbb{1}\{y=1\}(1-p) + \mathbb{1}\{y=-1\}p\Big) \\ &= -\log P(Y=y), \end{aligned}$$

where the last equality follows from the expression of the PMF of Y. ■

Appendix B. Proof of Lemma 3

We prove here Lemma 3 which repeat below.

Lemma 3: *For $\mathbf{w} \in \mathcal{M}$ with non zero curvature (e.g., sphere) the functions $-\log p(\langle \mathbf{w}|\mathbf{x}\rangle)$ and $-\log(1-p(\langle \mathbf{w}|\mathbf{x}\rangle))$ cannot be both convex.*

Proof We take the particular case with $d=1$ with $\mathcal{M}$ being the unit circle. Let $\mathbf{w} = \mathbf{u}(\theta) = [\cos\theta, \sin\theta]$ and let $f(\langle \mathbf{w}|\mathbf{x}\rangle)$ be a function. We have $\frac{\partial}{\partial\theta} f(\langle \mathbf{w}|\mathbf{x}\rangle) = \langle \mathbf{u}(\theta+\pi/2)|\mathbf{x}\rangle f'(\langle \mathbf{w}|\mathbf{x}\rangle)$ and the second derivative

$$\begin{aligned} \frac{\partial^2}{\partial\theta^2} f(\langle \mathbf{w}|\mathbf{x}\rangle) &= \langle \mathbf{u}(\theta+\pi/2)|\mathbf{x}\rangle^2 f''(\langle \mathbf{w}|\mathbf{x}\rangle) + \langle \mathbf{u}(\theta+\pi)|\mathbf{x}\rangle f'(\langle \mathbf{w}|\mathbf{x}\rangle) \\ &= (\|\mathbf{x}\|^2 - \langle \mathbf{u}(\theta)|\mathbf{x}\rangle^2) f''(\langle \mathbf{w}|\mathbf{x}\rangle) - \langle \mathbf{u}(\theta)|\mathbf{x}\rangle f'(\langle \mathbf{w}|\mathbf{x}\rangle). \end{aligned}$$

If we take $\mathbf{u}(\theta)$ aligned with $\mathbf{x}$ we have $\frac{\partial^2}{\partial\theta^2} f(\langle \mathbf{w}|\mathbf{x}\rangle) = -\|\mathbf{x}\| f'(\langle \mathbf{w}|\mathbf{x}\rangle)$ which must be strictly positive or negative. But for any real z we have $f'(z) = -\frac{p'(z)}{p(z)}$ when $f(z) = -\log p(z)$ and $f'(z) = \frac{p'(z)}{1-p(z)}$ when $f(z) = -\log(1-p(z))$ that are of the opposite signs, and therefore cannot be simultaneously convex/ concave functions. ■

Appendix C. Proof of Theorem 4

We observed in Lemma 7 that for $\mathbf{w} \in \mathcal{M}$ the functions $-\log p(\langle \mathbf{w}|\mathbf{x}\rangle)$ and $-\log(1-p(\langle \mathbf{w}|\mathbf{x}\rangle))$ are not both convex, therefore we cannot claim uniqueness of $\mathbf{w}^*$. We observe that all $\mathbf{w}^*$ satisfy $\nabla L(y^T|\mathbf{w}) = 0$, hence in our integral representation of Lemma 1 we will have a sum of local maxima of $P(Y^T|\mathbf{w})$ and this will lead to an upper bound for the minimax regret.

To be more precise, let $M(y^T\ \mathbf{x}^T)$ denote the sum of local maxmima stasifying $\nabla L(y^T|\mathbf{w}) = 0$. Then the next lemma is easy to prove following the footsteps of the proof of Lemma 1.

Lemma 9 *We have the following representation*

$$M(y^T|\mathbf{x}^T) = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \exp(-L(y^T|\mathbf{w})\det(\nabla^2 L(y^T|\mathbf{w})d\mathbf{w}\int_{\mathbb{R}^d}\exp\big(-i\langle \nabla L(y^T|\mathbf{w})|\mathbf{z}\rangle\big)d\mathbf{z}. \tag{32}$$

Now we defined generalized Shtarkov sum as

$$M(\mathbf{x}^T) = \sum_{y^T} M(y^T|\mathbf{x}^T)$$

and observe that $S(\mathbf{x}^T) \le M(\mathbf{x}^T)$ because of the possible multiple $\mathbf{w}^*$ giving as an upper bound for the minimax regret.

We are now in the position to prove Theorem 4 which can be reduced to show the following

$$M(\mathbf{x}^T) = \sum_{y^T} M(y^T|\mathbf{x}^T) = \int_{\mathcal{M}} \frac{\sqrt{\det(\tilde{\mathbf{B}}(\mathbf{w})}}{\sqrt{2\pi}^d} d\mathbf{w}\left(1 + O(\frac{d^{3/2}}{\sqrt{T}})\right)$$

where $\tilde{\mathbf{B}}(w) = \Pi(w)\mathbf{B}(w)\Pi(w)$ where $\mathbf{B}(w)$ is defined in (14) and $\Pi(w)$ is the orthogonal projection on $\mathcal{M}$.

The expression looks very similar to (11) but in fact the proof is very different. We first evaluate the following from which $M(\mathbf{x}^T)$ is easy to compute

$$M(\mathbf{x}^T|\mathbf{w}) = \frac{1}{(2\pi)^d}\sum_{y^T}|\det(\text{Hess}(L(y^T|\mathbf{w})))|\int_{\mathbb{R}^d}\exp\left((L(y^T|\mathbf{w}) - i\langle \text{grad}L(y^T|\mathbf{w})|\mathbf{z}\rangle\right)d\mathbf{z} \tag{33}$$

where Hess and grad respectively denote the Hessian and the gradient of $L(y^T|\mathbf{w})$ on the manifold $\mathcal{M}$. We simply have $\text{grad}L(y^T|\mathbf{w}) = \Pi(\mathbf{w})]\nabla L(y^T|\mathbf{w})$ but $\text{Hess}L(y^T|\mathbf{w})$ has a more complicated form because of the existence of non zero curvatures in the manifold. Indeed assuming that the quantities $L(y^T|\mathbf{w})$ can be extended beyond $\mathcal{M}$ we have

$$\text{Hess}L(y^T|\mathbf{w}) = \Pi(\mathbf{w})\nabla^2 L(y^T|\mathbf{w}) - \Gamma(\Pi(\mathbf{w})\nabla L(y^T|\mathbf{w})) \tag{34}$$

where Γ is the Christoffel operator which maps a vector on the tangent plane into a tensor in this plane. The Christoffel operators consists of the Christoffel symbols which have a complicated form (see Struik (1968)). But we don't need this level of details.

First, we can still use Lemma 7 to prove

$$P\left(\|\text{Hess}L(y^T|\mathbf{w}) - \tilde{\mathbf{B}}(\mathbf{w}))\| > AdT^{\alpha}\right) \le 2d^2 e^{-BT^{2\alpha-1}}. \tag{35}$$

Thus

$$M(\mathbf{x}^T|\mathbf{w}) = \frac{\det(\mathbf{B}^*(\mathbf{w}))}{(2\pi)^d} \sum_{y^T} \int_{\mathbb{R}^d} \exp\big(L(y^T|\mathbf{w}) - i\langle \text{grad} L(y^T|\mathbf{w})|\mathbf{z}\rangle\big) d\mathbf{z} \left(1 + O(d^2 e^{-BT^{2\alpha-1}})\right) \tag{36}$$

with $\mathbf{B}^*(\mathbf{w}) = \sum_{y^T} P(y^T|\mathbf{w}) \text{Hess} L(y^T|\mathbf{w})$. Second we notice that

$$\sum_{y^T} P(y^T|\mathbf{w}) \nabla L(y^T|\mathbf{w}) = 0$$

thus the Christoffel operator contribution cancels and

$$\sum_{y^T} P(y^T|\mathbf{w}) \text{Hess} L(y^T|\mathbf{w}) = \tilde{\mathbf{B}}(\mathbf{w}). \tag{37}$$

Finally we have

$$\begin{aligned} M(\mathbf{x}^T|\mathbf{w}) =& \frac{1}{(2\pi)^d} \det(\tilde{\mathbf{B}}(\mathbf{w})) \sum_{y^T} \int_{\mathbb{R}^d} \exp\left(\left(L(y^T|\mathbf{w}) - i\langle \Pi(\mathbf{w}) \nabla L(y^T|\mathbf{w})|\mathbf{z}\rangle\right) d\mathbf{z} + O(T^d e^{-BT^{2\alpha-1}}) \right. \\ =& \frac{1}{(2\pi)^d} \det(\tilde{\mathbf{B}}(\mathbf{w})) \int_{\mathbb{R}^d} \prod_t f(\mathbf{x}_t, \mathbf{z}) d\mathbf{z} \end{aligned}$$

with

$$f(\mathbf{x}, \mathbf{z}) = p(\langle \mathbf{w}|\mathbf{x}\rangle) e^{-i \frac{p'(\langle \mathbf{w}|\mathbf{x}\rangle)}{p(\langle \mathbf{w}|\mathbf{x}\rangle)} \langle \Pi(\mathbf{w})\mathbf{x}|\mathbf{z}\rangle} + (1 - p(\langle \mathbf{w}|\mathbf{x}\rangle)) e^{i \frac{p'(\langle \mathbf{w}|\mathbf{x}\rangle)}{1-p(\langle \mathbf{w}|\mathbf{x}\rangle)} \langle \Pi(\mathbf{w})\mathbf{x}|\mathbf{z}\rangle}$$

in (25).

To complete the proof of Theorem 4 we observe that

$$\sum_t \log f(\mathbf{x}_t, \mathbf{z}) = -\frac{1}{2} \langle \mathbf{z} \Pi(\mathbf{w}) \mathbf{B}(\mathbf{w}) \Pi(\mathbf{w}) \mathbf{z}\rangle + O(T\|\mathbf{z}\|^3)$$

and an application of the multi-dimensional Laplace methods, as discussed above, leads to

$$M(\mathbf{x}^T|\mathbf{w}) = \frac{\sqrt{\det(\tilde{\mathbf{B}}(\mathbf{w})}}{\sqrt{2\pi}^d} d\mathbf{w} \left(1 + O(\frac{d^{3/2}}{\sqrt{T}})\right)$$

which concludes the proof.

Appendix D. Error Term (27)

We provide more details how to estimate the error term in (27).

Lemma 10 (Error of (27)) *The error term of the Shtarkov sum (23) is*

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

for large T.

Proof Let us define

$$K(\mathbf{z}|\mathbf{w}) = \sum_{t=1}^{T} \log f(\mathbf{x}_t, \mathbf{z})$$

where $f(\mathbf{x}_t, \mathbf{z})$ is expressed as in (25). Note that $f(\mathbf{x}_t, \mathbf{z})$ depends on $\mathbf{w}$ but we do not explicitly write it down. It is also easy to see that $K(\mathbf{z}|\mathbf{w})$ and its derivative are of order $O(T)$. For $\mathbf{w}^* = \arg\max K(\mathbf{x}_t, \mathbf{z})$ we have the following Taylor expansion

$$\sum_{t=1}^{T} \log f(\mathbf{x}_t, \mathbf{z}) = -\frac{1}{2}\langle \mathbf{z}\mathbf{B}(\mathbf{w})|\mathbf{w})\mathbf{z}\rangle + O(K^{(3)}(\mathbf{z})\|\mathbf{z}\|^3)$$

where $K^{(3)}(\mathbf{z}) = O(T)$.

Let now $\mathbf{u} = \sqrt{\mathbf{B}(\mathbf{w})}\mathbf{z}$. Then (27) becomes

$$\int_{\mathbb{R}^d} \exp\left(\sum_t \log f(\mathbf{x}_t, \mathbf{z})\right) d\mathbf{z} = \frac{(2\pi)^{d/2}}{\sqrt{\det \mathbf{B}(\mathbf{w})}} \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}\|\mathbf{u}\|^2\right)\left(1 + O(\frac{\|\mathbf{u}\|^3}{\sqrt{T}})\right) d\mathbf{u}.$$

But we know that for $a > 0$

$$\int_{R^d} \|x\|^3 \exp\left(-a\|x\|^2\right) = 1/a^{(d+3)/2}\pi^{d/2}\frac{\Gamma(d/2+3/2)}{\Gamma(d/2)}, \tag{38}$$

thus we conclude that

$$\int_{\mathbf{R}^d} \exp\left(-\frac{1}{2}\|\mathbf{u}\|^2\right) O(\frac{\|\mathbf{u}\|^3}{\sqrt{T}}) d\mathbf{u} = O\left(\frac{\Gamma(d/2+3/2)}{\Gamma(d/2)\sqrt{T}}\right). \tag{39}$$

To complete, we observe that $\frac{\Gamma(d/2+3/2)}{\Gamma(d/2)} \sim (d/2)^{3/2}$ when $d \to \infty$. For more details see Inglot and Majewski (2014).

■