Phase Transition of Regret for Logistic Regression with Large Weights

Michael Drmota MICHAEL.DRMOTA@TUWIEN.AC.AT
TU Wien, Vienna, Austria

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

Changlong Wu WUCHANGL@HAWAII.EDU
Purdue University, West Lafayette, IN, USA

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

Editors: *Matus Telgarsky and Jonathan Ullman*

Abstract

In online learning, a learner receives data in rounds $1 \le t \le T$ and at each round predicts a label which is then compared to the true label resulting in a loss. The total loss over T rounds, when compared to a loss over the best expert from a class of experts, is called the regret. We study minimax regret for the best predictor and worst label sequence. This paper focuses on *logarithmic loss* over a class of experts $\mathcal{H}_{\mathbf{w}}$ parameterized by a d-dimensional weight vector $\mathbf{w}$ which can be unbounded and increase with T. For bounded weights it is known that the minimax regret can grow no faster than $(d/2)\log(TR^2/d)$, hence the leading coefficient in front of $\log T$ can raise untamed with growing R. However, in this paper, we demonstrate a phase transition showing that the minimax regret for $R \ge T$ and large d asymptotically becomes $(d \pm 1)\log T + O(\log\log T)$ for a logistic-like expert class, which can naturally be generalized to a larger class of experts. We prove our findings by introducing the so-called splitable label sequences that partition the weight space into T^{d-1} regions (of equal sign for the scalar product of weights and features), powered with analytic combinatorics (e.g., Mellin Transform, saddle point method) and discrete geometry.

Keywords: Online learning, minimax regret, discrete geometry, analytic combinatorics

1. Introduction

The problem of online learning under logarithmic loss and its regret analysis has been extensively studied over the past decade Grunwald (2007); Rakhlin and Sridharan (2015); Foster et al. (2018); Wu et al. (2022, 2023a). Nevertheless, even in the case of logistic regression, precise second-order asymptotics for unbounded weights remain largely unexplored, with the notable exceptions of Jacquet et al. (2021) and the recent works Drmota et al. (2024, 2025), which analyze scenarios where the weights grow either sublinearly with the number of rounds or without bound (see also Qian et al. (2024)). In this paper, we address the regime in which the weight growth rate exceeds the number of rounds, revealing a sharp phase transition. Our analysis builds on several novel techniques, drawing from analytic combinatorics and discrete geometry.

To set the stage of our discussion, we briefly recall that the online learning problem is often phrased in terms of a game between nature/ environment and a learner/predictor. At each round

$t \in \mathbb{N}$, the learner obtains a d dimensional data/ feature vector $\mathbf{x}_t \in \mathbb{R}^d$ to make a prediction $\hat{y}_t$ of the true label $y_t \in \{-1, 1\}$. The learner makes the prediction $\hat{y}_t$. 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 $\hat{\mathcal{Y}} \in \mathbb{R}$ and $\mathcal{Y} \in \{-1, 1\}$ are the prediction and label domains, respectively. In regret analysis, one compares the accumulated loss of the learner to the best strategy within a predefined class of predictor functions $h \in \mathcal{H} : \mathbb{R}^d \mapsto \mathbb{R}$ with bounded input $\|\mathbf{x}_t\| \leq 1$. After T rounds, the *pointwise regret* is

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

Throughout we write $y^T = (y_1, \ldots, y_T)$ and $\mathbf{x}^T = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$.

In this paper we focus on logarithmic loss $\ell : \hat{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}^+$ that we write as $\ell(\hat{y}_t, y_t) = -\log \hat{y}_t$, if $y_t > 0$, and $\ell(\hat{y}_t, y_t) = -\log(1 - \hat{y}_t)$ otherwise. Furthermore, we restrict our study to the class of experts:

$$\mathcal{H}_{p,\mathbf{w}} = \{h : \mathbb{R}^d \to \mathbb{R} : \ h(\mathbf{x}) = p(\langle \mathbf{w}|\mathbf{x} \rangle) : \mathbf{w}, \mathbf{x} \in \mathbb{R}^d\}, \tag{1}$$

where $\mathbf{w}$ is a d dimensional weight vector, $\langle \mathbf{w}|\mathbf{x} \rangle$ is the scalar product of $\mathbf{x}_t$ and $\mathbf{w}$, and $p(w)$ with $w = \langle \mathbf{w}|\mathbf{x} \rangle$ is a probability function. To ease the presentation, we mostly discuss the logistic regression with $p(w) = (1 + \exp(-w))^{-1}$ (see e.g., Hazan et al. (2014); Shamir (2020), and we leave the extension to other functions to a forthcoming paper. Throughout, assume that $\mathbf{x}_t$ lies on a d-dimensional sphere S_d while the weights $\|\mathbf{w}\| \leq R \leq \infty$ with R possibly increasing with T.

In the *fixed design* approach (also known as transductive online learning) analyzed here the minimal regret for the worst realization of the label with the feature vector $\mathbf{x}^T$ known in advance is defined as

$$r(\mathbf{x}^T) := r(\mathcal{H}|\mathbf{x}^T) = \inf_{g^T} \sup_{y^T} \mathcal{R}(g^T, y^T, \mathcal{H}|\mathbf{x}^T). \tag{2}$$

To decouple it from the feature vector $\mathbf{x}_t$ one either maximizes over all $\mathbf{x}_t$ or takes the average over the features. We study here the *average* fixed design minimax regret $\bar{r}_T(\mathcal{H}) := \mathbf{E}_{\mathbf{x}^T}[r(\mathcal{H}|\mathbf{x}^T)]$ as well as $r_T^*(\mathcal{H}) = \max_{\mathbf{x}^T} r(\mathcal{H}|\mathbf{x}^T)]$ where feature vector $\mathbf{x}_t$ is generated by an i.i.d distribution. The importance of fixed-design regret lies in the fact that it is a universal lower bound for various regrets discussed in the literature Wu et al. (2022, 2023b).

As discussed in Jacquet et al. (2021); Shamir (2020) the minimax regret $r(\mathbf{x}^T)$ can be studied through the so called Shtarkov sum which for $||\mathbf{w}|| \leq R$ becomes

$$S(\mathbf{x}^T) = \sum_{y^T} \sup_{||\mathbf{w}|| \leq R} P(y^T|\mathbf{x}^T, \mathbf{w}) \tag{3}$$

where $P(y^T|\mathbf{x}^T, \mathbf{w}) = \prod_{y_t > 0} p(\langle \mathbf{x}_t, \mathbf{w} \rangle) \prod_{y_t \leq 0} (1 - p(\langle \mathbf{x}_t, \mathbf{w} \rangle))$, and the regret is then $r(\mathbf{x}^T) = \log S(\mathbf{x}^T)$. In this paper, we investigate the Shtarkov sum and the minimax regret in the regime of large $R \geq T$, revealing a phase transition as illustrated in Figure 1. Throughout, we denote the Shtarkov sum and the minimax regret by $S^R(\mathbf{x}^T)$ and $r^R(\mathbf{x}^T)$, respectively, to emphasize their dependence on R.

1.1. Related Work

Online learning under logarithmic loss can be viewed as universal compression (source coding) with side information, as discussed in Barron et al. (1998); Takeuchi and Barron (2006) and Drmota and

Szpankowski (2023); Jacquet and Szpankowski (2004); Szpankowski and Weinberger (2012); Xie and Barron (1997, 2000). The logistic-type class of experts, as in (1), has been studied extensively in Foster et al. (2018); Hazan et al. (2014); Rakhlin and Sridharan (2015); Shamir (2020); Wu et al. (2022) under various formulations of regret. In particular, it is known that for any range R of the weight vector $\mathbf{w}$, the minimax regret can be upper bounded by $(d/2)\log(TR^2/d)$ for the *sequential* regret setting, where both $\mathbf{x}^T$ and y^T are selected sequentially Foster et al. (2018); Shamir (2020); Wu et al. (2022).

For the *fixed-design* regret studied here, the precise dependence on the weight norm R is largely unexplored. Several prior results, such as Shamir (2020); Wu et al. (2022); Mayo et al. (2022), show that the regret lower bound grows as $(d/2)\log(T/d^2)$ (with no dependence on R), which can deviate arbitrarily from the generic $(d/2)\log(TR^2/d)$ upper bound for large R. Recently, Drmota et al. (2024) proved that for *fixed-design* regret, the upper bound can be improved to $2d\log T$ for a general *monotone class*, even with $R=\infty$. But this bound is too rough, our aim is to *precisely* characterize fixed-design regret in the transition region, which turns out to be between $(d/2)\log T$ and $d\log T$ as R increases as a function of T. Very recently, Drmota et al. (2025) partially addressed this challenge by showing that when $R=o(T^{1/4})$, the regret grows linearly from $0.5d\log T$ to $0.625d\log T$ (see Figure 1). This paper fills the gap by focusing on the regime $R\geq\Omega(T)$. In passing we also disprove a conjecture in Drmota et al. (2024) which heuristically argued that the regret scales as $d\log T$ for $R>\sqrt{T}$. Our analysis shows that this behavior is guaranteed only when $R\geq T$, while the precise behavior in the intermediate region $O(\sqrt{T})<R<\Omega(T)$ remains open. To the best of our knowledge, Drmota et al. (2024, 2025) are the only works that provide a precise characterization of fixed-design regret with unbounded weights (although Qian et al. (2024) also studied unbounded weights, they did not address its precise behavior). We emphasize that analyzing the transition region poses substantial technical challenges when a precise characterization is required (i.e., precise up to the *second-order* asymptotics).

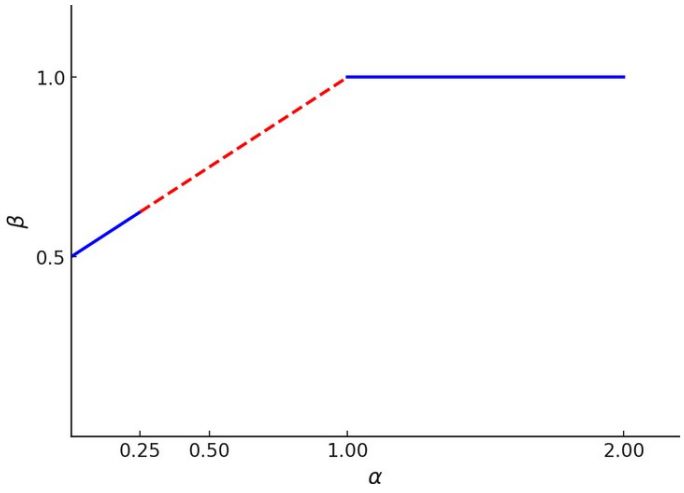

Figure 1: Phase transition of the regret growth constant β in $r_T\sim\beta d\log T$ for $R=T^\alpha$. The lower blue segment, given by $\beta=\alpha/2+1/2$, was proved in Drmota et al. (2025). The upper blue segment is established in this paper, while the red dashed segment remains conjectural.

1.2. Summary of Contributions

In this paper, we present for the first time a precise second-order asymptotic characterization of the transition region $R \geq T$. In Theorems 5 and 6, we establish the following results, which we summarize informally below:

Theorem 1 (Informal) (i) *For $T \leq R \leq \infty$ the following bounds hold for* **Shtarkov sum***:*

$$c_1 T^{d-1} (\log T)^{-(d-1)} \leq \mathbf{E}_{\mathbf{x}^T}[S^R(\mathbf{x}^T)] \leq c_2 T^{d+1} (\log T)^{3(d-1)}$$

for sufficiently large T and positive constants c_1, c_2.

(ii) *Uniformly for $c_3 T \log T \leq R \leq \infty$ the following bounds hold for the average* **minimax regret**

$$(d-1)\log T - O(\log T/\log\log T) \leq \mathbf{E}_{\mathbf{x}^T}[r^R(\mathbf{x}^T)] \leq (d+1)\log T + O(\log\log T)$$

for large T and a positive constants c_3.

This indicates that a phase transition occurs around $R \sim T$, as illustrated in Figure 1. The behavior of the regret in the intermediate regime $T^{1/4} < R < T$ remains unknown; however, we conjecture that it grows linearly with α for $R = T^\alpha$ (see the red dashed line in Figure 1). Our main proof technique builds on the notion of splittable sequences, first employed by Drmota et al. (2024), combined with novel geometric covering lemmas for high-dimensional spheres (Lemma 3 and 4), which we believe are of independent interest. Moreover, our analysis of the upper bound leverages powerful analytic tools from analytic combinatorics—such as Mellin transforms, generating functions, and complex asymptotics.

2. Main Results

In this section, we first introduce the notion of *splittable sequences*, followed by key covering lemmas for the unit sphere that enable us to derive general lower and upper bounds for the Shtarkov sum and the minimax regret. We then present a concise overview of our two main results on the precise asymptotics of the Shtarkov sum and regret, with full proofs deferred to the next section and the Appendix.

Splitable Sequences. Given a feature sequence $\mathbf{x}^T = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$ (with $\mathbf{x}_t \in S^d$) a *chamber* c is a cone of vectors $\mathbf{w} \in \mathbb{R}^d$ where for each t the quantity $\langle \mathbf{w}, \mathbf{x}_t \rangle$ keeps the same sign over c. Equivalently, we consider the T hyperplanes $H_t = \{\mathbf{w} \in \mathbb{R}^d : \langle \mathbf{w}, \mathbf{x}_t \rangle = 0\}$, $1 \leq t \leq T$, and then a chamber is just a connected component of the complement $\mathbb{R}^d \setminus \bigcup_{t=1}^T H_t$. Suppose that the set of vectors $\mathbf{x}^T$ is in general linear position; that is, every choice of d vectors are linearly independent. Then the number $M_d(T)$ of chambers is given by

$$M_d(T) = 2\sum_{i=0}^{d-1} \binom{T-1}{i} = 2\binom{T-1}{d-1}(1 + O(1/T)) \sim 2\frac{T^{d-1}}{(d-1)!}.$$

This follows, for example, from an application of Zavlavsky's theorem (Greene and Zaslavsky, 1983). In all the other cases, $M_d(T)$ is an upper bound for the number of chambers.

A label sequence y^T is called splitable if there exists $\mathbf{w} \in \mathbb{R}^d$ such that $y_t\langle \mathbf{w}, \mathbf{x}_t\rangle > 0$ for $t = 1, \dots, T$. We denote the set of splitable sequences as $\mathcal{SP}(\mathbf{x}^T)$, that is

$$\mathcal{SP}(\mathbf{x}^T) := \{y^T : \exists \mathbf{w} \in \mathbb{R}^d \, \forall t \in [T], \; y_t\langle \mathbf{x}_t, \mathbf{w}\rangle > 0\}. \tag{4}$$

Such sequences partition $\mathbb{R}^d \setminus \bigcup_{t=1}^T H_t$ into chambers as illustrated in Figure 2 (for $d = 2$). Clearly there is a bijection between chambers and splitable label sequences. Thus when the set $\mathbf{x}^T$ is in general linear position, the number of splitable sequences is precisely $M_d(T)$.

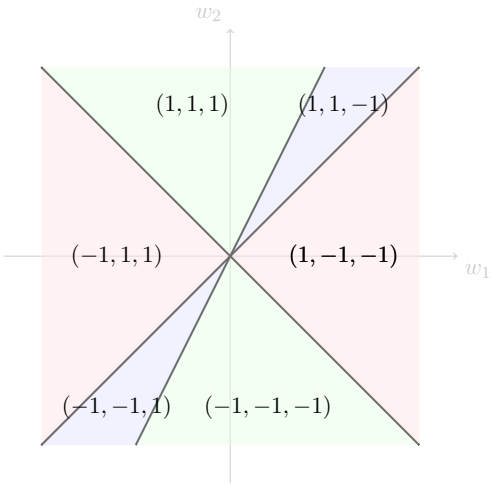

Figure 2: Illustration to splitable sequences for $T = 3$ and $d = 2$ where the boundary of the chamber are solutions (of $\mathbf{w}$) to $\langle \mathbf{x}_t, \mathbf{w}\rangle = 0$ where $\mathbf{x}_1 = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}), \mathbf{x}_2 = (-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}), \mathbf{x}_3 = (-\frac{2}{\sqrt{5}}, \frac{1}{\sqrt{5}})$). There are six splitable sequences y^3 out of eight possible label sequences (e.g., $(-1, 1, -1)$ is not splitable).

Now we fix a chamber c or equivalently a splitable sequence y^T. Then a sequence y'^T is called k-splitable in c if exactly k indices from y^T and y'^T differ. Equivalently, this means that there are exactly k indices t such that y'_t and $\langle \mathbf{w}, \mathbf{x}_t\rangle$ have different signs, where $\mathbf{w} \in c$.

A very simple observation is that for every chamber c and the corresponding label sequence y^T we have $\sup_{\mathbf{w}\in c} P(y^T|\mathbf{x}^T, \mathbf{w}) = \sup_{\mathbf{w}\in c} \prod_{t=1}^T p(|\langle \mathbf{w}, \mathbf{x}_t\rangle|) = 1$ if $R = \infty$. Thus, if the feature sequence $\mathbf{x}^T$ is in general linear position we find for $R = \infty$

$$S^\infty(\mathbf{x}^T) \geq M_d(T) \sim 2\frac{T^{d-1}}{(d-1)!}$$

and consequently $r^\infty(\mathbf{x}^T) \geq (d-1)\log T + O(1)$ which was already proved in Drmota et al. (2024). We now extend this idea to the regime $R \sim T$, which will require several technical covering lemmas, as shown below.

Geometrization: Cover Lemmas. We call a discrete set $U = \{\mathbf{u}_i \in S^d : 1 \leq i \leq N\}$ on the sphere S^d a δ-cover if for every $\mathbf{u} \in S^d$ there exists $i \in \{1, \dots, N\}$ with $\|\mathbf{u} - \mathbf{u}_i\| \leq \delta$. (Equivalently, the balls $B(\mathbf{u}_i, \delta)$, $1 \leq i \leq N$, cover the sphere S^d.) Similarly we call a discrete set $V = \{\mathbf{v}_i \in S^d : 1 \leq i \leq N\}$ on the sphere S^d a δ-packing if the balls $B(\mathbf{v}_i, \delta)$, $1 \leq i \leq N$, are disjoint. In particular, the following can be found in Wainwright (2019).

Lemma 2 *For every $\delta > 0$ there exists a δ-cover $U = \{\mathbf{u}_i \in S^d : 1 \le i \le N\}$ on the sphere S^d with $N \le c_d \delta^{-(d-1)}$, where $c_d > 0$ is a constant depending on the dimension d.*

Furthermore, for every $\delta > 0$ there exists a δ-packing $V = \{\mathbf{v}_i \in S^d : 1 \le i \le N\}$ on the sphere S^d with $N \ge c'_d \delta^{-(d-1)}$, where $c'_d > 0$ is a constant depending on the dimension d.

With the help of Lemma 2 we construct a subset A_T of the ball $B(\mathbf{0}, T) = \{\mathbf{w} \in \mathbb{R}^d : \|\mathbf{w}\| \le T\}$ of size $|A_T| \le T^{d+1+\alpha}$ such that

$$A_T = \left\{ \frac{j}{T} \mathbf{u}_i : 1 \le j \le T^2,\ 1 \le i \le T^{d-1+\alpha} \right\},$$

where the set $U = \{\mathbf{u}_i : 1 \le i \le T^{d-1+\alpha}\} \subset S^d$ is a δ-cover with

$$\delta = c_d^{-\frac{1}{d-1}} T^{-1-\frac{\alpha}{d-1}}. \tag{5}$$

We now consider the set of chambers $C(\mathbf{x}^T)$ that corresponds to a feature sequence $\mathbf{x}^T$. We call *cell* the intersection of a chamber $c \in C(\mathbf{x}^T)$ with S^d and we denote by $\tilde{C} = \tilde{C}(\mathbf{x}^T)$ the set of all cells (related to $\mathbf{x}^T$). The next two covering lemmas are crucial to establish upper and lower bounds for the Shtarkov sum and regret.

Lemma 3 *Suppose that the vectors $\mathbf{x}_t$, $1 \le t \le T$, of a feature sequence are uniformly and independently distributed on S^d. Suppose further that $0 \le \alpha \le d$ and that $U = \{\mathbf{u}_i \in S^d : 1 \le i \le T^{d-1+\alpha}\}$ is a δ-cover on S^d with δ as in (5). Then with probability $> 1 - O(T^{-2d})$*

$$\max_{1 \le i \le T^{d-1+\alpha}} \left| B(\mathbf{u}_i, \delta) \cap C(\mathbf{x}^T) \right| = O\left((\log T)^{d-1} \right),$$

where all estimates are uniform for $0 \le \alpha \le d$.

Proof We first consider a ball $B(\mathbf{u}, \delta) = \{\mathbf{v} \in S^d : \|\mathbf{u} - \mathbf{v}\| \le \delta\}$ on S^d and want to estimate the number of cells (or chambers) of $C(\mathbf{x}^T)$ that intersect with $B(\mathbf{u}, \delta)$. Note that δ is related to the maximal angle θ between $\mathbf{u}$ and $\mathbf{v}$ by $\delta = 2 \sin \frac{\theta}{2}$. For this purpose, we first observe that $H_t = \{\mathbf{w} \in \mathbb{R}^d : \langle \mathbf{w}, \mathbf{x}_t \rangle = 0\}$ intersects $B(\mathbf{u}, 2 \sin \frac{\theta}{2})$ if and only if $|\langle \mathbf{u}, \mathbf{x}_t \rangle| \le \sin \theta$. Hence, the number of cells (or chambers) that intersect with $B(\mathbf{u}, 2 \sin \frac{\theta}{2})$ depends solely on the number

$$m_T(\mathbf{u}) = \left| \{1 \le t \le T : |\langle \mathbf{u}, \mathbf{x}_t \rangle| \le \sin \theta\} \right|,$$

In particular, by Zavlavsky's theorem (Greene and Zaslavsky, 1983), we have

$$\left| B(\mathbf{u}, 2 \sin(\theta/2)) \cap C(\mathbf{x}^T) \right| = O(1 + m_T(\mathbf{u})^{d-1})$$

Note that

$$m_T(\mathbf{u}) = \sum_{t=1}^{T} \mathbf{1}_{[|\langle \mathbf{u}, \mathbf{x}_t \rangle| \le \sin \theta]}$$

is a sum of iid indicator function and has a binomial distribution $Bi(T, p)$ with parameter

$$p \quad = \quad \mathbf{P}(|\langle \mathbf{u}, \mathbf{x}_t \rangle| \le \sin \theta) = \frac{1}{s_d} \int_0^\theta (\cos t)^{d-1} \, dt \le c''_d T^{-1-\frac{\alpha}{d-1}}$$

for some constant $c''_d > 0$. It's moment generating function is then given by

$$\mathbf{E}[e^{\xi m_T(\mathbf{u})}] = (1 + (e^{\xi} - 1)p)^T \leq e^{(e^{\xi}-1)c''_d}$$

and by an application of Chernov's bound we obtain (by choosing $\xi = 4d$ and $s = \log T$)

$$\mathbf{P}\left(m_T(\mathbf{u}) \geq \log T\right) \leq e^{-\xi \log T}(1 + (e^{\xi} - 1)p)^T = O(T^{-4d}).$$

Finally by applying the union bound it follows that

$$\mathbf{P}\left(\exists i \in \{1, \ldots, T^{d-1+\alpha}\} : m_T(\mathbf{u}_i) \geq \log T\right) \leq C\, T^{d-1+\alpha} T^{-4d} \leq T^{-2d}.$$

Thus, with probability $> 1 - T^{-2d}$ we have

$$\max_{1 \leq i \leq T^{d-1+\alpha}} \left|B(\mathbf{u}_i, \delta) \cap C(\mathbf{x}^T)\right| = O\left((\log T)^{d-1}\right)$$

as desired. ■

Lemma 4 *Suppose that the vectors $\mathbf{x}_t$, $1 \leq t \leq T$, of a feature sequence are uniformly and independently distributed on S^d. Suppose further that*

$$\log(c'''_d \log T)/\log T \leq \alpha \leq d - 1$$

for a properly chosen constant $c'''_d > 0$ (depending on d) and that $V = \{\mathbf{v}_i \in S^d : 1 \leq i \leq T^{d-1-\alpha}\}$ is a δ-packing on S^d with

$$\delta = (c'_d)^{-\frac{1}{d-1}} T^{-1+\frac{\alpha}{d-1}}.$$

Then with probability $> 1 - T^{-2d}$

$$\max_{c \in C(\mathbf{x}^T)} |V \cap c| \leq 1.$$

Proof For $\mathbf{v}, \mathbf{w} \in S^d$ we denote by $\theta = \theta(\mathbf{v}, \mathbf{w}) = \arccos(\langle \mathbf{v}, \mathbf{w} \rangle)$ the angle between $\mathbf{v}$ and $\mathbf{w}$. Then a random hyperplane H_t separates $\mathbf{v}$ and $\mathbf{w}$ with probability

$$\mathbf{P}\left(H_t \text{ separates } \mathbf{v} \text{ and } \mathbf{w}\right) = \frac{\theta(\mathbf{v}, \mathbf{w})}{\pi}.$$

Hence, the probability $q_T(\theta)$ that none of the T hyperplanes H_t ($1 \leq t \leq T$) separates the pair $\mathbf{v}$ and $\mathbf{w}$ (with angle θ) is given by

$$q_T(\theta) = \left(1 - \frac{\theta}{\pi}\right)^T \leq e^{-T\theta/\pi}.$$

Actually, this is precisely the probability that $\mathbf{v}$ and $\mathbf{w}$ are contained in the same chamber (or cell). In particular, if $\theta \geq 2\delta$ (for some $\delta > 0$) then $q_T(\theta) \leq e^{-2T\delta/\pi}$. Suppose now that $V = \{\mathbf{v}_i : 1 \leq i \leq T^{d-1-\alpha}\}$ is a δ-packing with

$$\delta = (c'_d)^{-\frac{1}{d-1}} T^{-1+\frac{\alpha}{d-1}}.$$

Since $\|\mathbf{v}_i - \mathbf{v}_j\| \geq 2\delta$ for $i \neq j$ we find

$$\begin{aligned}\mathbf{P}\left(\exists i \neq j:\ \mathbf{v}_i \text{ and } \mathbf{v}_j \text{ lie in the same chamber }\right) &\leq T^{2(d-1-\alpha)} e^{-2T\delta/\pi} \\ &\leq T^{2(d-1-\alpha)} e^{-2(c'_d)^{-\frac{1}{d-1}} T^{\frac{\alpha}{d-1}}/\pi} \\ &\leq T^{-2d}\end{aligned}$$

for $(d-1)\log(c'''_d \log T)/\log T \leq \alpha \leq d-1$ and a properly chosen constant

$$c'''_d \geq \frac{\pi(4d+1)}{2}(c'_d)^{\frac{1}{d-1}}.$$

Therefore, with probability $> 1 - T^{-2d}$ each chamber contains at most one element of V. ■

Intuitively, Lemma 3 shows that, with high probability (over the randomness of $\mathbf{x}^T$), the ball $B(\mathbf{u}_i, \delta)$ centered at any covering point $\mathbf{u}_i$ with radius δ cannot intersect many chambers induced by $\mathbf{x}^T$. On the other hand, Lemma 4 shows that, with high probability, no chamber can intersect with many *packing* points as well. These two lemmas allow us to concentrate our analysis on (fixed) *finite* packing and covering sets of the parameter space, independent of $\mathbf{x}^T$, which will then be used in deriving our lower and upper bounds, respectively.

Lower Bound for Average Shtarkov Sum. We now derive a lower bound for the Shtarkov sum $\mathbf{E}[S^R(\mathbf{x}^T)]$ for $R \geq T$. By the monotonicity of Shtarkov sum, we only need to prove a lower bound for $R = T$. The idea is to lower bound the Shtarkov sum only over the splittable sequences and to choose $\mathbf{w}$ in an appropriate way (any choice of $\mathbf{w}$ yields a lower bound). We fix $\alpha = (d-1)\log(c'''_d \log T)/(\log T)$ and let $V = \{\mathbf{v}_i : 1 \leq i \leq T^{d-1-\alpha}\}$ be a δ-packing of S^d with $\delta = (c'_d)^{-1/(d-1)} T^{-1+\alpha/(d-1)}$.

Let $y(c)^T$ be a splitable sequence associated with chamber c. By Lemma 4 we have $|V \cap c| \leq 1$ for all $c \in C(\mathbf{x}^T)$ with probability $\geq 1 - T^{-2d}$. Thus,

$$\begin{aligned}S^T(\mathbf{x}^T) &\geq \sum_{c \in C(\mathbf{x}^T)} \max_{\mathbf{w} \in c,\, |\mathbf{w}| \leq T} P(y(c)^T | \mathbf{x}^T, \mathbf{w}) \\ &\overset{(\star)}{\geq} \sum_{c \in C(\mathbf{x}^T)} \sum_{\mathbf{v}_i \in V \cap c} P(y(c)^T | \mathbf{x}^T, T\mathbf{v}_i) \\ &= \sum_{1 \leq i \leq T^{d-1-\alpha}} P(y(c)^T | \mathbf{x}^T, T\mathbf{v}_i) \\ &= \sum_{1 \leq i \leq T^{d-1-\alpha}} \prod_{t=1}^{T} \frac{1}{1 + e^{-T|\langle \mathbf{x}_t, \mathbf{v}_i \rangle|}}, \end{aligned} \tag{6}$$

holds with probability $\geq 1 - T^{-2d}$, where $(\star)$ follows by Lemma 4. This implies that

$$\mathbf{E}[S^T(\mathbf{x}^T)] = \Omega\left(T^{d-1-\alpha}\left(\mathbf{E}\left[\frac{1}{1 + e^{-T|\langle \mathbf{x}, \mathbf{v} \rangle|}}\right]\right)^T\right).$$

Since $\mathbf{E}\left[\frac{1}{1+e^{-T|\langle \mathbf{x}, \mathbf{v} \rangle|}}\right] = 1 - \frac{c}{T}(1 + o(1))$, it follows that

$$\mathbf{E}[S^T(\mathbf{x}^T)] = \Omega\left(T^{d-1-\alpha}\right) = \Omega\left(T^{d-1}(\log T)^{-d+1}\right).$$

Lower Bound for Average Minimax Regret. Note that the averaged minimax regret equals $\mathbf{E}[\log S^R(\mathbf{x}^T)]$. Unfortunately, this cannot be directly lower bounded by $\mathbf{E}[S^R(\mathbf{x}^T)]$, since Jensen's inequality goes in the opposite direction. To overcome this difficulty, we instead establish a high-probability lower bound on $S^R(\mathbf{x}^T)$ for a slightly larger radius $R = \Omega\left(\sqrt{d}\,T\log T\right)$. From (6) we have, wp $\geq 1 - T^{-2d}$ for all $R \geq T$, that

$$S^R(\mathbf{x}^T) \geq \sum_{1\leq i\leq T^{d-1-\alpha}} \prod_{t=1}^{T} \frac{1}{1+e^{-R|\langle \mathbf{x}_t, \mathbf{v}_i\rangle|}}.$$

For every $\mathbf{v}_i \in V$, the distribution of the inner product $\langle \mathbf{x}_t, \mathbf{v}_i\rangle$ has the density (see, e.g., White (1953); Costa (2015))

$$h(s) = \frac{\Gamma(d/2)}{\sqrt{\pi}\,\Gamma((d-1)/2)}(1-s^2)^{(d-3)/2}, \qquad |s| \leq 1.$$

Note that

$$\frac{\Gamma(d/2)}{\sqrt{\pi}\,\Gamma((d-1)/2)} \sim \sqrt{\frac{d}{2\pi}}.$$

Thus, for $\delta \leq \frac{1}{2}$,

$$\mathbf{P}\left(|\langle \mathbf{x}_t, \mathbf{v}_i\rangle| \leq \delta\right) \leq C_1\sqrt{d}\,\delta$$

for some (absolute) constant $C_1 > 0$. We now consider the random variable

$$Y_i = \left|\{t \in \{1,\ldots,T\} : \log(1+e^{-R(|\langle \mathbf{x}_t, \mathbf{v}_i\rangle|}) \geq 1/T\}\right|$$

that is the sum of T independent Bernoulli random variables with

$$\begin{aligned}
\mu &= \mathbf{E}[Y_i] \\
&= T\,\mathbf{P}\left(\log(1+e^{-R|\langle \mathbf{x}_1, \mathbf{v}_i\rangle|}) \geq 1/T\right) \\
&\leq T\,\mathbf{P}\left(|\langle \mathbf{x}_1, \mathbf{v}_i\rangle| \leq (\log T)/R\right) \\
&\leq C_1\frac{\sqrt{d}\,T\log T}{R}.
\end{aligned}$$

Note that $\log(1+e^{-R(|\langle \mathbf{x}_1, \mathbf{v}_i\rangle|}) \geq 1/T$ implies $|\langle \mathbf{x}_1, \mathbf{v}_i\rangle| \leq (\log T)/R$. Consequently, by setting

$$R = C_1 e\sqrt{d}\,T\log T$$

and by applying Lemma 13 of Appendix A we obtain for $y \geq 1$

$$\mathbf{P}\left(Y_i \geq y\right) \leq \left(\frac{eC_1\sqrt{d}\,T\log T}{Ry}\right)^y \leq e^{-y\log y}.$$

Hence, by setting $y = 3d\log T/(\log\log T)$ we arrive at

$$e^{-y\log y} = e^{-3d\log T(1-(\log\log T)/(\log\log\log T))} \leq \frac{1}{T^{2d}}$$

for sufficiently large T. Thus, with probability $> 1 - T^{-2d}$ we obtain

$$\prod_{t=1}^{T} \frac{1}{1+e^{-R|\langle \mathbf{x}_t, \mathbf{v}_i\rangle|}} \geq e^{-(T-y)/T - y\log 2} \geq \frac{1}{e} T^{-(3\log 2)d/(\log\log T)}.$$

Applying the union bound, it follows that with probability $\geq 1 - T^{-d}$,

$$S^R(\mathbf{x}^T) = \Omega\left((\log T)^{-(d-1)} T^{d-1}\, T^{-(3\log 2)d/(\log\log T)}\right).$$

Consequently, with probability $\geq 1 - T^{-d}$, we have

$$r^R(\mathbf{x}^T) \geq (d-1)\log T - O\left(\frac{d\log T}{\log\log T}\right) \tag{7}$$

for sufficiently large T, where the O-constant does not depend on $\mathbf{x}^T$. This further implies that the expected regret is of the same order, since the regret is nonnegative.

Upper Bound for Shtarkov Sum. Here we only give general ideas for the upper bound on the Shtarkov sum, delaying the detailed proof to Section 3 and the Appendix. We first recall that the following upper bound, proved in Drmota et al. (2024) (see also Appendix B for complete proof) holds

$$S^\infty(\mathbf{x}^T) = O(T^{2d}), \tag{8}$$

and will be used several times in this paper.

By the monotonicity of the Shtarkov sum, it suffices to derive an upper bound for $S^\infty(\mathbf{x}^T)$, which then automatically applies to $S^R(\mathbf{x}^T)$ for any R.

We distinguish between three cases, where the optimum is obtained for some $\mathbf{w}^*$ with $\|\mathbf{w}^*\| \leq W_0$, where $W_0 < \|\mathbf{w}^*\| \leq W_1$, and where $\|\mathbf{w}^*\| > W_1$, with

$$W_0 = T^{1-\frac{\alpha}{d-1}} \quad \text{and} \quad W_1 = T, \tag{9}$$

where

$$\alpha = (d-1)\frac{\log(c_d'''\log T)}{\log T}, \tag{10}$$

and c_d''' is sufficiently large (just depending on the dimension d). We set

$$S_0^\infty(\mathbf{x}^T) = \sum_{y^T\in\{-1,1\}^T} \max_{\|\mathbf{w}\|\leq W_0} P(y^T|\mathbf{x}^T,\mathbf{w}), \tag{11}$$

$$S_1^\infty(\mathbf{x}^T) = \sum_{y^T\in\{-1,1\}^T} \max_{W_0\leq\|\mathbf{w}\|\leq W_1} P(y^T|\mathbf{x}^T,\mathbf{w}) \tag{12}$$

$$S_2^\infty(\mathbf{x}^T) = \sum_{y^T\in\{-1,1\}^T} \sup_{\|\mathbf{w}\|> W_1} P(y^T|\mathbf{x}^T,\mathbf{w}). \tag{13}$$

Clearly, we have then

$$S^\infty(\mathbf{x}^T) \leq S_0^\infty(\mathbf{x}^T) + S_1^\infty(\mathbf{x}^T) + S_2^\infty(\mathbf{x}^T).$$

We will show that

$$c_1 T^{d-1}(\log T)^{-(d-1)} \leq \mathbf{E}[S^\infty(\mathbf{x}^T)] \leq c_2 T^{d+1}(\log T)^{3(d-1)}$$

uniformly holds for all sufficiently large T and suitable positive constants c_1, c_2.

Summary of Main Theorems. We are now in a position to state our main results.

Theorem 5 *Suppose that the vectors $\mathbf{x}_t$, $1 \le t \le T$, of a feature sequence are uniformly and independently distributed on S^d. Then, uniformly for $T \le R \le \infty$, we have*

$$c_1 T^{d-1} (\log T)^{-(d-1)} \le \mathbf{E}[S^R(\mathbf{x}^T)] \le c_2 T^{d+1} (\log T)^{3(d-1)}$$

for sufficiently large T and suitable positive constants c_1, c_2.

A corresponding result for the regret holds in a slightly smaller range. Note that we have already established the lower bound in (7). The upper bound follows from Theorem 5 together with Jensen's inequality:

$$\mathbf{E}[r^R(\mathbf{x}^T)] = \mathbf{E}\big[\log S^R(\mathbf{x}^T)\big] \le \log \mathbf{E}[S^R(\mathbf{x}^T)],$$

which holds for all $R > 0$.

Theorem 6 *Suppose that the vectors $\mathbf{x}_t$, $1 \le t \le T$, of a feature sequence are uniformly and independently distributed on S^d. Then, uniformly for $c_3 T \log T \le R \le \infty$, we have*

$$(d-1)\log T - O(\log T / \log\log T) \le \mathbf{E}[r^R(\mathbf{x}^T)] \le (d+1)\log T + O(\log\log T)$$

for sufficiently large T and suitable positive constants c_3.

3. Analysis

In this section, we focus on proving Theorem 5, in particular the upper bound given in (11)–(13). Before proceeding, we first present some auxiliary results concerning the average behavior of the logistic function and its concentration.

3.1. Concentration of the Logistic Function

We start with asymptotic results for the average (over $\mathbf{x}^T$) of the logistic function.

Lemma 7 *Suppose that $\mathbf{x}$ is a random variable that is uniformly distributed on S^d and $\mathbf{w} \in \mathbb{R}^d$. Then the expected values*

$$I = I(\|\mathbf{w}\|) = \mathbf{E}\left[\log(1 + e^{-|\langle \mathbf{w}, \mathbf{x}_t \rangle|})\right] \quad \textit{and} \quad J = J(\|\mathbf{w}\|) = \mathbf{E}\left[e^{-|\langle \mathbf{w}, \mathbf{x}_t \rangle|}\right]$$

only depend on the length $w = \|\mathbf{w}\|$ and are asymptotically given by

$$I(w) = \frac{\pi^2}{12 s_d w} + O(w^{-2}) \quad \textit{and} \quad J(w) = \frac{1}{s_d w} + O(w^{-2}),$$

where

$$s_d = \int_0^{\pi/2} (\cos\theta)^{d-1}\, d\theta = \frac{\sqrt{\pi}\,\Gamma(d/2)}{2\Gamma((d+1)/2)} \sim \frac{\sqrt{\pi}}{\sqrt{2d}}.$$

Proof We first write $I(w) = I(||\mathbf{w}||)$ explicitly as follows

$$I(w) = \frac{1}{s_d} \int_0^{\pi/2} \log(1 + e^{-w \sin\theta})(\cos\theta)^{d-1} d\theta$$

with the normalizing factor s_d as above. We now apply the powerful Mellin transform (Szpankowski, 2001) to analyze $I(w)$ for large w. We first recall definition the Mellin transform $I^*(w)$ of $I(w)$ which is $I^*(s) = \int_0^\infty I(w) w^{s-1} dw$ for some complex s. We find

$$\begin{aligned} I^*(s) &= \int_0^\infty I(w) w^{s-1} dw \\ &= \frac{1}{s_d} \int_0^{\pi/2} (\cos\theta)^{d-1} d\theta \int_0^\infty \log(1 + e^{-w\sin\theta}) w^{s-1} dw \\ &= \frac{1}{s_d} \int_0^{\pi/2} (\cos\theta)^{d-1} d\theta (\sin\theta)^{-s} \int_0^\infty \log(1 + e^{-v}) v^{s-1} dv. \end{aligned}$$

We now observe that the latter integral is the Mellin transform of $\log(1 + e^{-v})$ which becomes

$$\int_0^\infty \log(1 + e^{-v}) v^{s-1} dv = \Gamma(s)\eta(s+1)$$

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

$$\eta(s) = \sum_{n=1}^\infty \frac{(-1)^{n+1}}{n^s}, \quad \Re(s) > 0.$$

Furthermore,

$$\int_0^{\pi/2} (\sin\theta)^s (\cos\theta)^{d-1} d\theta = \frac{\Gamma(d/2)\Gamma(1/2 + s/2)}{2\Gamma(d/2 + s/2 + 1/2)}.$$

Putting everything together we find

$$I*(s) = \frac{1}{s_d} \Gamma(s)\eta(s+1) \frac{\Gamma(d/2)\Gamma(1/2 - s/2)}{2\Gamma(d/2 - s/2 + 1/2)}, \quad 0 < \Re(s) < 1.$$

The main singularity of $I^*(s)$ is at $s = 1$ with residue $\frac{\pi^2}{12 s_d}$ since $\eta(2) = \pi^2/12$ and $\Gamma((1-s)/2) \sim 2/(1-s)$. By inverse Mellin transform, we find for large w

$$I(w) = \frac{\pi^2}{12 s_d w} + O(\frac{1}{w^2}).$$

The asymptotics of $J(w)$ follows the same pattern except that the starting point is

$$J(w) = \frac{1}{s_d} \int_0^{\pi/2} e^{-w\sin\theta} (\cos\theta)^{d-1} d\theta$$

with the Mellin transform

$$J^*(s) = \frac{1}{s_d} \Gamma(s) \frac{\Gamma(d/2)\Gamma(1/2 - s/2)}{2\Gamma(d/2 - s/2 + 1/2)}$$

leading to

$$J(w) = \frac{1}{s_d w} + O(1/w^2)$$

for large w. ■

Next we prove some concentration results of the sum of logistic functions.

Lemma 8 *Set*

$$I_T(\mathbf{w}, \mathbf{x}^T) = \sum_{t=1}^{T} \log\left(1 + e^{-|\langle \mathbf{w}, \mathbf{x}_t \rangle|}\right),$$

where $\mathbf{x}^T = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$ *are iid on* S^d*. Then for sufficiently small* $\varepsilon > 0$ *there exist a constant* $\beta = \Omega\left(\varepsilon\sqrt{d}\right)$ *such that for every* $\mathbf{w} \in \mathbb{R}^d$

$$\mathbf{P}\left(I_T(\mathbf{w}, \mathbf{x}^T) \notin [(1-\varepsilon)TI(\|\mathbf{w}\|), (1+\varepsilon)TI(\|\mathbf{w}\|)]\right) = O\left(\exp\left(-\beta\frac{T}{\|\mathbf{w}\|}\right)\right).$$

Similarly, if we set

$$J_T(\mathbf{w}, \mathbf{x}^T) = \sum_{t=1}^{T} e^{-|\langle \mathbf{w}, \mathbf{x}_t \rangle|},$$

then we also have

$$\mathbf{P}\left(J_T(\mathbf{w}, \mathbf{x}^T) \notin [(1-\varepsilon)TJ(\|\mathbf{w}\|), (1+\varepsilon)TJ(\|\mathbf{w}\|)]\right) = O\left(\exp\left(-\beta\frac{T}{\|\mathbf{w}\|}\right)\right)$$

for $\beta = \Omega\left(\varepsilon\sqrt{d}\right)$.

Proof We first study the moment generating function

$$\mathbf{E}[e^{\xi I_T(\mathbf{w}, \mathbf{x}^T)}] = f(\|\mathbf{w}\|, \xi)^T$$

with

$$f(w, \xi) = \frac{1}{s_d}\int_0^{\pi/2} (1 + e^{-w\sin\theta})^{\xi} (\cos\theta)^{d-1} d\theta.$$

The Mellin transform of the mapping $w \mapsto f(w, \xi) - 1$ is defined for $\Re(s) > 0$ and is given by

$$M(f(w, \xi) - 1, s) = g^*(s, \xi)\frac{\Gamma(d/2)\Gamma(1/2 - s/2)}{2\Gamma(d/2 - s/2 + 1/2)},$$

where $g^*(s, \xi)$ is the Mellin transform of the mapping

$$x \mapsto (1 + e^{-x})^{\xi} - 1$$

that exists for all ξ and for all s with $\Re(s) > 0$. Thus, the smallest singularities of $M(f(w, \xi) - 1, s)$ are at $s = 1$ and $s = 3$ (assuming that $d > 2$). This implies that $f(w, \xi)$ behaves like

$$f(w, \xi) = 1 + I(w)\xi + \frac{h(\xi)}{w} + O\left(\frac{1}{w^3}\right)$$

with

$$h(\xi) = \int_0^\infty \Big((1+e^{-x})^\xi - 1 - \log(1+e^{-x}) \Big)\, dx = \alpha_2 \xi^2 + O(\xi^3),$$

where $\alpha_2 = \int_0^\infty (\log(1+e^{-x}))^2 dx$. Note that these expansions only hold for $\xi = O(1)$.

By Chernoff's inequality we have for any random variable

$$\mathbf{P}(X \geq A) \leq e^{-\xi A} \mathbf{E}[e^{\xi X}] \qquad (\xi > 0)$$

and

$$\mathbf{P}(X \leq B) \leq e^{-\xi B} \mathbf{E}[e^{\xi X}] \qquad (\xi < 0).$$

Thus, we get (for $\xi > 0$)

$$\mathbf{P}\left(I_T(\mathbf{w}, \mathbf{x}^T) \geq (1+\varepsilon) T\, I(w) \right) \leq e^{-\xi \varepsilon T\, I(w) + T \alpha_2 \xi^2 / w(1+O(\xi))}$$

and (for $\xi < 0$)

$$\mathbf{P}\left(I_T(\mathbf{w}, \mathbf{x}^T) \leq (1-\varepsilon) T\, I(w) \right) \leq e^{\xi \varepsilon T\, I(w) + T \alpha_2 \xi^2 / w(1+O(\xi))}.$$

We just choose

$$\xi = \pm c'' \varepsilon^{1/2}$$

for a sufficiently small constant $c'' > 0$ and have completed the first part of the proof.

Next we set (for $|\xi| \leq 1$)

$$\begin{aligned} g(w,\xi) &= \frac{1}{s_d} \int_0^{\pi/2} \exp\left(\xi e^{-w \sin\theta} \right) (\cos\theta)^{d-1} d\theta \\ &= 1 + \frac{1}{s_d} \int_0^{\pi/2} \xi e^{-w \sin\theta} (\cos\theta)^{d-1} d\theta + O\left(\frac{\xi^2}{s_d} \int_0^{\pi/2} e^{-2w \sin\theta} \cos\theta\, d\theta \right) \\ &= 1 + \frac{\xi}{s_d w} + O\left(\frac{\xi^2}{s_d w} \right). \end{aligned}$$

Thus, we are in the completely same situation as for $f(w,\xi)$ and thus obtain the same kind of estimates. ■

We end this subsection with some deterministic estimates on the sum of logistic functions.

Lemma 9 *We have*

$$\sum_{t=1}^T \Big| \log(1+e^{-|\langle \mathbf{w}_1, \mathbf{x}_t \rangle|}) - \log(1+e^{-|\langle \mathbf{w}_2, \mathbf{x}_t \rangle|}) \Big| \leq \|\mathbf{w}_1 - \mathbf{w}_2\| \sum_{t=1}^T e^{-\min(|\langle \mathbf{w}_1, \mathbf{x}_t \rangle|, |\langle \mathbf{w}_2, \mathbf{x}_t \rangle|)}. \quad (14)$$

Proof Since

$$\big| \log(1+e^{-x})' \big| = \frac{1}{1+e^x} \leq e^{-x} \qquad (x \geq 0)$$

we have

$$\big| \log(1+e^{-x}) - \log(1+e^{-y}) \big| \leq e^{-\min(x,y)} |x-y| \qquad (x, y \geq 0).$$

This implies

$$\left|\log(1+e^{-|\langle \mathbf{w}_1,\mathbf{x}_t\rangle|}) - \log(1+e^{-|\langle \mathbf{w}_2,\mathbf{x}_t\rangle|})\right| \;\leq\; \big||\langle \mathbf{w}_1,\mathbf{x}_t\rangle| - |\langle \mathbf{w}_2,\mathbf{x}_t\rangle|\big|\, e^{-\min(|\langle \mathbf{w}_1,\mathbf{x}_t\rangle|,|\langle \mathbf{w}_2,\mathbf{x}_t\rangle|)}.$$

However, we also have

$$\big||\langle \mathbf{w}_1,\mathbf{x}_t\rangle| - |\langle \mathbf{w}_2,\mathbf{x}_t\rangle|\big| \leq |\langle \mathbf{w}_1,\mathbf{x}_t\rangle - \langle \mathbf{w}_2,\mathbf{x}_t\rangle| \leq \|\mathbf{w}_1-\mathbf{w}_2\|.$$

So, (14) follows immediately. ■

In the sequel we will use Lemma 9 in a 2-step approximation. First, we apply a scaling, that is, consider $\mathbf{w}_1 = w_1\mathbf{u}$ and $\mathbf{w}_2 = w_2\mathbf{u}$ with $\mathbf{u} \in S^d$. Then

$$\sum_{t=1}^{T}\left|\log(1+e^{-|\langle \mathbf{w}_1,\mathbf{x}_t\rangle|}) - \log(1+e^{-|\langle \mathbf{w}_2,\mathbf{x}_t\rangle|})\right| \leq T\,|\mathbf{w}_1-\mathbf{w}_2|. \tag{15}$$

Secondly if $\|\mathbf{w}_1\| = \|\mathbf{w}_2\|$, that is, if $\mathbf{w}_1 = w\mathbf{u}_1$ and $\mathbf{w}_2 = w\mathbf{u}_2$ with $\mathbf{u}_1, \mathbf{u}_2 \in S^d$. Then

$$\sum_{t=1}^{T}\left|\log(1+e^{-|\langle \mathbf{w}_1,\mathbf{x}_t\rangle|}) - \log(1+e^{-|\langle \mathbf{w}_2,\mathbf{x}_t\rangle|})\right| \leq w\|\mathbf{u}_1-\mathbf{u}_2\|\sum_{t=1}^{T} e^{-w\min(|\langle \mathbf{u}_1,\mathbf{x}_t\rangle|,|\langle \mathbf{u}_2,\mathbf{x}_t\rangle|)} \tag{16}$$

for any w.

3.2. Proof of Upper Bound

We now prove the upper bound of Theorem 5 by considering three parts of the Shtarkov sum as expressed in (11)-(13).

3.2.1. UPPER BOUND FOR $S_0^\infty(\mathbf{x}^T)$

For given features $\mathbf{x}^T$ let $C(\mathbf{x}^T)$ be the set of chambers. Furthermore we denote by $y^T(c)$ the splitable label sequence that corresponds to $c \in C(\mathbf{x}^T)$. Moreover, if J_k is a subset of size k of $\{1,\ldots,T\}$ then $y^T(c,J_k)$ denotes the k-splitable label sequence that is obtained from $y^T(c)$ by changing the signs at the indices J_k. We denote by $\mathbf{w}^*(c,J_k)$ a vector that maximizes $P(y^T(c,J_k)|\mathbf{x}^T,\mathbf{w})$ in the closure $\bar{c}$ of c with $\|\mathbf{w}^*(c,J_k)\| \leq W_0$ where W_0 is defined in (9). Thus

$$\max_{\mathbf{w}\in\bar{c},\,\|\mathbf{w}\|\leq W_0} P(y^T(c,J_k)|\mathbf{x}^T,\mathbf{w}) = P(y^T(c,J_k)|\mathbf{x}^T,\mathbf{w}^*(c,J_k)).$$

We then clearly have

$$S_0^\infty(\mathbf{x}^T) \leq \sum_{c\in C(\mathbf{x}^T)}\sum_{k=0}^{T}\sum_{J_k} P(y^T(c,J_k)|\mathbf{x}^T,\mathbf{w}^*(c,J_k)).$$

By definition then

$$\begin{aligned}
&P(y^T(c,J_k)|\mathbf{x}^T,\mathbf{w}^*(c,J_k))\\
&\quad= \prod_{t=1}^{T}\left(1+e^{-|\langle \mathbf{w}^*(c,J_k),\mathbf{x}_t\rangle|}\right)^{-1}\exp\left(-\sum_{j\in J_k}|\langle \mathbf{w}^*(c,J_k),\mathbf{x}_j\rangle|\right)\\
&\quad= \exp\left(-\sum_{t=1}^{T}\log\left(1+e^{-|\langle \mathbf{w}^*(c,J_k),\mathbf{x}_t\rangle|}\right) - \sum_{j\in J_k}|\langle \mathbf{w}^*(c,J_k),\mathbf{x}_j\rangle|\right).
\end{aligned}$$

We now take the best approximation vector $\tilde{\mathbf{w}}^*(c,J_k) \in A_T$ so that (by Lemma 9)

$$\begin{aligned}
&P(y^T(c,J_k)|\mathbf{x}^T,\mathbf{w}^*(c,J_k)) \\
&= O\Bigg(\exp\left(-\sum_{t=1}^{T}\log\left(1+e^{-|\langle\tilde{\mathbf{w}}^*(c,J_k),\mathbf{x}_t\rangle|}\right)-\sum_{j\in J_k}|\langle\tilde{\mathbf{w}}^*(c,J_k),\mathbf{x}_j\rangle|\right) \\
&\quad\times\exp\left(O\left(1+T^{-\frac{2\alpha}{d-1}}\sum_{t=1}^{T}e^{-|\langle\tilde{\mathbf{w}}^*(c,J_k),\mathbf{x}_t\rangle|}+\frac{k}{T}+kT^{-\frac{2\alpha}{d-1}}\right)\right)\Bigg) \\
&= O\Bigg(\exp\left(-I_T(\tilde{\mathbf{w}}^*(c,J_k),\mathbf{x}^T)+O\left(T^{-\frac{\alpha}{d-1}}J_T(\tilde{\mathbf{w}}^*(c,J_k),\mathbf{x}^T)\right)\right) \\
&\quad\times\exp\left(-\|\tilde{\mathbf{w}}^*(c,J_k)\|\sum_{j\in J_k}|\sin(\theta_j)|+O\left(kT^{-\frac{2\alpha}{d-1}}\right)\right)\Bigg),
\end{aligned}$$

where α is given in (10) and θ_t denotes the co-angle between $\tilde{\mathbf{w}}^*(c,J_k) \in A_T$ and $\mathbf{x}_t$:

$$\sin\theta_t = \frac{\langle\tilde{\mathbf{w}}^*(c,J_k)),\mathbf{x}_t\rangle}{\|\tilde{\mathbf{w}}^*(c,J_k)\|}.$$

We now apply Lemma 8 and obtain that with probability

$$> 1-O\left(T^{d+1+\alpha}\exp\left(-\beta_0\frac{T}{W_0}\right)\right) = 1-O\left(T^{-2d}\right)$$

leading to

$$(1-\varepsilon)T\,I(\tilde{\mathbf{w}}) \le I_T(\tilde{\mathbf{w}},\mathbf{x}^T) \le (1+\varepsilon)T\,I(\tilde{\mathbf{w}})$$

and

$$(1-\varepsilon)T\,J(\tilde{\mathbf{w}}) \le J_T(\tilde{\mathbf{w}},\mathbf{x}^T) \le (1+\varepsilon)T\,J(\tilde{\mathbf{w}})$$

for all $\tilde{\mathbf{w}} \in A_T$. Note that we have used the assumption that c_d''' is sufficiently large so that

$$T^{d+1+\alpha}\exp\left(-\beta_0\frac{T}{W_0}\right) \le T^{-2d}.$$

Thus, we have with high probability

$$\begin{aligned}
&P(y^T(c,J_k)|\mathbf{x}^T,\mathbf{w}^*(c,J_k)) \\
&= O\Bigg(\exp\left(-(1-\varepsilon)T\,I(\tilde{\mathbf{w}}^*(c,J_k))+O\left(T^{1-\frac{\alpha}{d-1}}J(\tilde{\mathbf{w}}^*(c,J_k))\right)\right) \\
&\quad\times\exp\left(-\|\tilde{\mathbf{w}}^*(c,J_k)\|\sum_{j\in J_k}|\sin(\theta_j)|+O\left(kT^{\frac{-2\alpha}{d-1}}\right)\right)\Bigg) \\
&= O\left(\exp\left(-(1-2\varepsilon)T\pi^2/(6s_d\|\mathbf{w}^*(c,J_k)\|)-\|\tilde{\mathbf{w}}^*(c,J_k)\|\sum_{j\in J_k}|\sin(\theta_j)|+O\left(kT^{\frac{-2\alpha}{d-1}}\right)\right)\right).
\end{aligned}$$

By a (very) simple computation, we also find

$$(1-2\varepsilon)T\pi^2/(6s_d\tilde{w}^*(c,J_k))+\tilde{w}^*(c,J_k)\sum_{t\in J_k}|\sin\theta_t|\geq 2\sqrt{(1-2\varepsilon)\frac{T\pi^2}{6s_d}\sum_{t\in J_k}|\sin\theta_t|}.$$

Thus, we obtain (with high probability) the upper bound

$$\begin{aligned}&P(y^T(c,J_k)|\mathbf{x}^T,\mathbf{w}^*(c,J_k))\\&=O\left(\exp\left(-2(1-2\varepsilon)\sqrt{\frac{(1-2\varepsilon)T\pi^2}{6s_d}\sum_{t\in J_k}|\sin\theta_t|}+O\left(kT^{-\frac{2\alpha}{d-1}}\right)\right)\right).\end{aligned}$$

We now apply Lemma 3 which implies that with high probability $\tilde{\mathbf{w}}\in A_T$ is at most $O(1)$ times the best approximation of a vector $\mathbf{w}^*(c,J_k))$ (uniformly for every k and J_k). Thus,

$$\begin{aligned}S_0^\infty(\mathbf{x}^T)&=O\Bigg((\log T)^{d-1}\sum_{\tilde{\mathbf{w}}\in A_T}\sum_{k=0}^{T}\sum_{J_k}\\&\qquad\exp\left(-2\sqrt{\frac{(1-2\varepsilon)T\pi^2}{6s_d}\sum_{t\in J_k}|\sin\theta_t|}+O\left(kT^{\frac{-2\alpha}{d-1}}\right)\right)\Bigg),\end{aligned}$$

where θ_t denotes the co-angle between $\tilde{\mathbf{w}}^*\in A_T$ and $\mathbf{x}_t$.

Our goal is to get upper bounds for the expected value $\mathbf{E}[S_0^\infty(\mathbf{x}^T)]$. We distinguish between two cases. In the first case, where Lemma 3 cannot be applied, we use the general upper bound $S_0^\infty(\mathbf{x}^T)=O(T^{2d})$, but this occurs in the worst case with probability $O(T^{-2d})$. In the second case we just apply the above upper bound and obtain

$$\begin{aligned}&\mathbf{E}[S_0^\infty(\mathbf{x}^T)]=O\Bigg(1\\&+(\log T)^{d-1}T^{\alpha+d+1}\sum_{k=0}^{T}\binom{T}{k}\mathbf{E}\left[\exp\left(-2(1-2\varepsilon)\sqrt{\frac{(1-2\varepsilon)T\pi^2}{6s_d}\sum_{t\in J_k}|\sin\theta_t|}+O\left(kT^{\frac{-2\alpha}{d-1}}\right)\right)\right]\Bigg).\end{aligned}$$

Hence, by applying Lemma 10 we finally arrive at

$$\begin{aligned}\mathbf{E}[S_0^\infty(\mathbf{x}^T)]&=O\left(1+(\log T)^{d-1}T^{\alpha+d+1}\sum_{k=0}^{T}\binom{T}{k}\frac{1}{T^k}\frac{(2k)!}{k!}\left(\frac{3e^{O(T^{-\frac{2\alpha}{d-1}})}}{2\pi^2(1-2\varepsilon)}\right)^k\right)\\&=O\left(T^{d+1}(\log T)^{2(d-1)}\right)\end{aligned}$$

as desired.

3.2.2. UPPER BOUND FOR $S_1^\infty(\mathbf{x}^T)$

We start in the same way as for $S_0^\infty(\mathbf{x}^T)$, However, we denote now by $\mathbf{w}^*(c,J_k)\in A_T$ a vector that maximizes $P(y^T(c,J_k)|\mathbf{x}^T,\mathbf{w})$ in the closure $\bar{c}$ of c when $W_0\leq\|\mathbf{w}\|\leq W_1$ where W_1 is defined in (9):

$$\max_{\mathbf{w}\in\bar{c},\,W_0\leq\|\mathbf{w}\|\leq W_1}P(y^T(c,J_k)|\mathbf{x}^T,\mathbf{w})=P(y^T(c,J_k)|\mathbf{x}^T,\mathbf{w}^*(c,J_k)).$$

As above we then have

$$S_1^\infty(\mathbf{x}^T) \leq \sum_{c\in C(\mathbf{x}^T)} \sum_{k=0}^{T} \sum_{J_k} P(y^T(c,J_k)|\mathbf{x}^T, \mathbf{w}^*(c,J_k)).$$

We recall that

$$\begin{aligned}
&P(y^T(c,J_k)|\mathbf{x}^T, \mathbf{w}^*(c,J_k)) \\
&\quad = \prod_{t=1}^{T} \left(1+e^{-|\langle \mathbf{w}^*(c,J_k),\mathbf{x}_t\rangle|}\right)^{-1} \exp\left(-\sum_{j\in J_k} |\langle \mathbf{w}^*(c,J_k),\mathbf{x}_j\rangle|\right) \\
&\quad = \exp\left(-\sum_{t=1}^{T} \log\left(1+e^{-|\langle \mathbf{w}^*(c,J_k),\mathbf{x}_t\rangle|}\right) - \sum_{j\in J_k} |\langle \mathbf{w}^*(c,J_k),\mathbf{x}_j\rangle|\right).
\end{aligned}$$

As above we now take the best approximation vector $\tilde{\mathbf{w}}^*(c,J_k) \in A_T$ so that

$$\begin{aligned}
&P(y^T(c,J_k)|\mathbf{x}^T, \mathbf{w}^*(c,J_k)) \\
&\quad = O\Bigg(\exp\left(-\sum_{t=1}^{T} \log\left(1+e^{-|\langle \tilde{\mathbf{w}}^*(c,J_k),\mathbf{x}_t\rangle|}\right) - \sum_{j\in J_k} |\langle \tilde{\mathbf{w}}^*(c,J_k),\mathbf{x}_j\rangle|\right) \\
&\qquad \times \exp\left(O\left(1+T^{-\frac{\alpha}{d-1}} \sum_{t=1}^{T} e^{-|\langle \tilde{\mathbf{w}}^*(c,J_k),\mathbf{x}_t\rangle|} + \frac{k}{T} + kT^{-\frac{\alpha}{d-1}}\right)\right)\Bigg) \\
&\quad = O\Bigg(\prod_{t\notin J_k} \frac{1+O\left(T^{-\frac{\alpha}{d-1}} e^{-\|\tilde{\mathbf{w}}^*(c,J_k)\||\sin(\theta_t)|}\right)}{1+e^{-\|\tilde{\mathbf{w}}^*(c,J_k)\||\sin(\theta_t)|}} \prod_{j\in J_k} \frac{e^{-\|\tilde{\mathbf{w}}^*(c,J_k)\||\sin(\theta_j)|}}{1+e^{-\|\tilde{\mathbf{w}}^*(c,J_k)\||\sin(\theta_j)|}} e^{O(kT^{-\alpha/(d-1)})}\Bigg),
\end{aligned}$$

where θ_t denotes the co-angle between $\tilde{\mathbf{w}}^*(c,J_k) \in A_T$ and $\mathbf{x}_t$. We then apply Lemma 3 and obtain

$$\begin{aligned}
S_1^\infty(\mathbf{x}^T) \quad = \quad & O\Bigg((\log T)^{d-1} \sum_{\tilde{\mathbf{w}}\in A_T} \sum_{k=0}^{T} \sum_{J_k} \\
& \prod_{t\notin J_k} \frac{1+O\left(T^{-\frac{\alpha}{d-1}} e^{-\|\tilde{\mathbf{w}}^*\||\sin(\theta_t)|}\right)}{1+e^{-\|\tilde{\mathbf{w}}^*\||\sin(\theta_t)|}} \prod_{j\in J_k} \frac{e^{-\|\tilde{\mathbf{w}}^*\||\sin(\theta_j)|}}{1+e^{-\|\tilde{\mathbf{w}}^*\||\sin(\theta_j)|}} e^{O(kT^{-\alpha/(d-1)})}\Bigg).
\end{aligned}$$

Again we distinguish two cases. In the first case, where Lemma 3 cannot be applied, we use the general upper bound $S_1^\infty(\mathbf{x}^T) = O(T^{2d})$ (as proved in Appendix B), but this occurs in the worst

case with probability $O(T^{-2d})$. In the second case we just apply the above upper bound and obtain

$$\begin{aligned}\mathbf{E}[S_1^\infty(\mathbf{x}^T)] &= O\Bigg(1+(\log T)^{d-1}T^{\alpha+d+1}\sum_{k=0}^{T}\binom{T}{k}\\ &\qquad \max_{W_0\le\tilde{w}\le W_1}\mathbf{E}\left[\prod_{t\notin J_k}\frac{1+O\left(T^{-\frac{\alpha}{d-1}}e^{-\tilde{w}|\sin(\theta_t)|}\right)}{1+e^{-\tilde{w}|\sin(\theta_t)|}}\prod_{j\in J_k}\frac{e^{-\tilde{w}|\sin(\theta_j)|}}{1+e^{-\tilde{w}|\sin(\theta_j)|}}\right]e^{O(kT^{-\alpha/(d-1)})}\Bigg)\\ &= O\Bigg(1+(\log T)^{d-1}T^{\alpha+d+1}\sum_{k=0}^{T}\binom{T}{k}\\ &\qquad \max_{W_0\le\tilde{w}\le W_1}\prod_{t\notin J_k}\mathbf{E}\left[\frac{1+O\left(T^{-\frac{\alpha}{d-1}}e^{-\tilde{w}|\sin(\theta_t)|}\right)}{1+e^{-\tilde{w}|\sin(\theta_t)|}}\right]\prod_{j\in J_k}\mathbf{E}\left[\frac{e^{-\tilde{w}|\sin(\theta_j)|}}{1+e^{-\tilde{w}|\sin(\theta_j)|}}\right]e^{O(kT^{-\alpha/(d-1)})}\Bigg).\end{aligned}$$

By Lemma 11 and by Lemma 12 (applied to $w_1 = w$) we have

$$\mathbf{E}\left[\frac{1+O\left(T^{-\frac{\alpha}{d-1}}e^{-\tilde{w}|\sin(\theta_t)|}\right)}{1+e^{-\tilde{w}|\sin(\theta_t)|}}\right] = 1-\frac{\log 2}{s_d\tilde{w}}\left(1+O\left(\frac{1}{\tilde{w}^2}\right)+O\left(T^{-\frac{\alpha}{d-1}}\right)\right).$$

Another application of Lemma 12 leads to

$$\mathbf{E}\left[\frac{e^{-\tilde{w}|\sin(\theta_j)|}}{1+e^{-\tilde{w}|\sin(\theta_j)|}}\right] = \frac{\log 2}{s_d\tilde{w}}\left(1+O\left(\frac{1}{\tilde{w}^2}\right)\right).$$

Thus, we have to deal with

$$\max_{W_0\le\tilde{w}\le W_1}\left(1-\frac{\log 2}{s_d\tilde{w}}\left(1+O\left(\frac{1}{\tilde{w}^2}\right)+O\left(T^{-\frac{\alpha}{d-1}}\right)\right)\right)^{T-k}\left(\frac{\log 2}{s_d\tilde{w}}\left(1+O\left(\frac{1}{\tilde{w}^2}\right)+O\left(T^{-\frac{\alpha}{d-1}}\right)\right)\right)^{k}.$$

It is an easy exercise that the maximum is obtained for $\tilde{w} = W_0$ if $k \ge \frac{\log 2}{s_d}T^{\frac{\alpha}{d-1}}$. We just have to consider the derivative with respect of $\tilde{w}$ of the logarithm and to observe that it is asymptotically given by

$$\left(T\frac{\log 2}{s_d\tilde{w}}-k+O(1)\right)\frac{\log 2}{s_d\tilde{w}^2}\left(1+O\left(T^{-\frac{\alpha}{d-1}}\right)\right).$$

Clearly if $\tilde{w} \ge W_0 = T^{1-\frac{\alpha}{d-1}}$ this expression is negative. Thus, the maximum is obtained for $\tilde{w} = W_0$

$$\begin{aligned}&\sum_{k\ge\frac{\log 2}{s_d}T^{\frac{\alpha}{d-1}}}\binom{T}{k}\left(1-\frac{\log 2}{s_dW_0}\left(1+O\left(\frac{1}{W_0^2}\right)+O\left(T^{-\frac{\alpha}{d-1}}\right)\right)\right)^{T-k}\\ &\qquad\qquad\left(\frac{\log 2}{s_dW_0}\left(1+O\left(\frac{1}{W_0^2}\right)+O\left(T^{-\frac{\alpha}{d-1}}\right)\right)\right)^{k}\\ &= O\left(\left(1+O\left(\frac{1}{s_dW_0^3}\right)+O\left(\frac{T^{-\frac{\alpha}{d-1}}}{W_0}\right)\right)^{T}\right)\\ &= O\left(\exp\left(T/(s_dW_0^3)+T^{1-\frac{\alpha}{d-1}}/W_0\right)\right)\\ &= O(1).\end{aligned}$$

On the other hand, if $0 \le k < \frac{\log 2}{s_d} T^{\frac{\alpha}{d-1}}$ we have that the maximum is obtained for $\tilde{w}_0 = T \log 2/(s_d k)$ so that we are left with the sum

$$\sum_{k \le \frac{\log 2}{s_d} T^{\frac{\alpha}{d-1}}} \binom{T}{k} \left(1 - \frac{k}{T}\left(1 + O\left(T^{-\frac{\alpha}{d-1}}\right)\right)\right)^{T-k} \left(\frac{k}{T}\left(1 + O\left(T^{-\frac{\alpha}{d-1}}\right)\right)\right)^{k}$$

$$= O\left(\sum_{k \le \frac{\log 2}{s_d} T^{\frac{\alpha}{d-1}}} \frac{T^k}{k!} e^{-k + O(k^2/T)} \frac{(k + O(1))^k}{T^k} \right)$$

$$= O\left(\sum_{k \le \frac{\log 2}{s_d} T^{\frac{\alpha}{d-1}}} e^{O(k^2/T)} \right) = O\left(T^{\frac{\alpha}{d-1}}\right).$$

Summing up, this implies

$$\begin{aligned} \mathbf{E}[S_1^\infty(\mathbf{x}^T)] &= O\left(1 + (\log T)^{d-1} T^{\alpha+d+1}\left(1 + T^{\frac{\alpha}{d-1}}\right)\right) \\ &= O\left(T^{d+1} (\log T)^{3(d-1)}\right) \end{aligned}$$

as desired.

3.2.3. UPPER BOUND FOR $S_2^\infty(\mathbf{x}^T)$

As in the previous case, we consider a chamber c, an integer $k \ge 0$, a subset J_k of $\{1, \dots, T\}$ of size k, and suppose that

$$\sup_{\mathbf{w} \in \overline{c},\, \|\mathbf{w}\| \ge W_1} P(y^T(c, J_k) | \mathbf{x}^T, \mathbf{w}) = P(y^T(c, J_k) | \mathbf{x}^T, \mathbf{w}^*(c, J_k)),$$

where $\|\mathbf{w}^*(c, J_k)\| \ge W_1$. If $k = 0$ then we formally have to include the case $\|\mathbf{w}^*(c, J_k)\| = \infty$ which leads to the trivial upper bound $P(y^T(c, J_k) | \mathbf{x}^T, \mathbf{w}^*(c, J_k)) = 1$. We, thus, have

$$S_2^\infty(\mathbf{x}^T) \le |C(\mathbf{x}^T)| + \sum_{c \in C(\mathbf{x}^T)} \sum_{k=1}^{T} \sum_{J_k} P(y^T(c, J_k) | \mathbf{x}^T, \mathbf{w}^*(c, J_k))$$

that we will analyze in a similar way as above. Note that $|C(\mathbf{x}^T)| = O(T^{d-1})$. So we only have to deal with the case $k \ge 1$.

In this case, we certainly have $\|\mathbf{w}^*(c, J_k)\| < \infty$. Furthermore, we can use the trivial inequality

$$\begin{aligned}
&P(y^T(c, J_k)|\mathbf{x}^T, \mathbf{w}^*(c, J_k)) \\
&\quad = \prod_{t=1}^{T} \left(1 + e^{-|\langle \mathbf{w}^*(c,J_k), \mathbf{x}_t\rangle|}\right)^{-1} \exp\left(-\sum_{j\in J_k} |\langle \mathbf{w}^*(c, J_k), \mathbf{x}_j\rangle|\right) \\
&\quad \leq \exp\left(-\sum_{j\in J_k} |\langle \mathbf{w}^*(c, J_k), \mathbf{x}_j\rangle|\right) \\
&\quad \leq \exp\left(-W_1 \sum_{j\in J_k} |\langle \mathbf{u}^*(c, J_k), \mathbf{x}_j\rangle|\right),
\end{aligned}$$

where

$$\mathbf{u}^*(c, J_k) = \frac{\mathbf{w}^*(c, J_k)}{\|\mathbf{w}^*(c, J_k)\|}.$$

We now take a δ-cover $U_T = \{\mathbf{u}_i : 1 \leq i \leq T^{d-1+\alpha}\}$ of S^d with $\delta = c_d^{-1/(d-1)} T^{-1-\frac{\alpha}{d-1}}$ (as in Lemma 3), where $\alpha = (d-1)\log(c_d''' \log T)/(\log T)$ with a properly chosen constant $c_d''' > 0$. We proceed now very similarly as above to obtain

$$S_2^\infty(\mathbf{x}^T) \quad O\left(T^{d-1} + (\log T)^{d-1} \sum_{\mathbf{u}_i\in U_T} \sum_{k=1}^{T} \sum_{J_k} \exp\left(-W_1 \sum_{j\in J_k} |\sin(\theta_j)| + O(kT^{-\frac{\alpha}{d-1}})\right)\right),$$

where θ_t denotes the co-angle between $\tilde{\mathbf{u}}_i \in U_T$ and $\mathbf{x}_t$. Recall that $W_1 = T$. Since

$$\mathbf{E}\left[e^{-T\sin\theta}\right] = \frac{1}{s_d} \int_0^{\pi/2} e^{-T\sin\theta} (\cos\theta)^{d-1} d\theta \leq \frac{1}{T s_d}$$

we, thus, obtain

$$\begin{aligned}
\mathbf{E}\left[S_2^\infty(\mathbf{x}^T)\right] &= O\left(T^{d-1} + (\log T)^{d-1} \sum_{\mathbf{u}_i\in U_T} \sum_{k=1}^{T} \binom{T}{k} \exp\left(O(kT^{-\frac{\alpha}{d-1}})\right) \frac{1}{(Ts_d)^k}\right) \\
&= O\left(T^{d-1} + (\log T)^{d-1} T^{d-1+\alpha} \sum_{k\geq 1} \frac{\exp\left(O(kT^{-\frac{\alpha}{d-1}})\right)}{k!}\right) \\
&= O\left(T^{d-1} (\log T)^{2(d-1)}\right)
\end{aligned}$$

which completes the proof of the upper bound of Theorem 5.

References

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

M.H.M. Costa. Geometric probability and random vectors. In *Lecture Notes on Geometric Probability*. 2015. Available as lecture notes.

M. Drmota and W. Szpankowski. *Analytic Information Theory: From Compression to Learning*. Cambridge University Press, Cambridge, 2023.

M. Drmota, P. Jacquet, C. Wu, and W. Szpankowski. Minimax regret with unbounded weights. In *ISIT*, 2024.

M. Drmota, P. Jacquet, C. Wu, and W. Szpankowski. Precise regularized minimax regret with unbounded weights. *IEEE Transaction on Information Theory*, 2025.

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.

Curtis Greene and Thomas Zaslavsky. On the interpretation of whitney numbers through arrangements of hyperplanes, zonotopes, non-Radon partitions, and orientations of graphs. *Transactions of the American Mathematical Society*, 280:97–126, 1983.

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

David Haussler and Philip M. Long. A generalization of Sauer's lemma. *Journal of Combinatorial Theory, Series A*, 71(2):219–240, August 1995. doi: 10.1016/0097-3165(95)90042-X.

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*, pages 197–209. MIT press, 2014.

P. Jacquet and W. Szpankowski. Markov types and minimax redundancy for markov sources. *IEEE Trans. on Information Theory*, 5o:1393–1402, 2004.

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.

J. Mayo, H. Hadiji, and T. van Erven. Scale-free unconstrained online learning for curved losses. In *COLT*, volume 178, pages 1–34, 2022.

Jian Qian, Alexander Rakhlin, and Nikita Zhivotovskiy. Refined risk bounds for unbounded losses via transductive priors. *arXiv preprint arXiv:2410.21621*, 2024.

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

Gil I Shamir. Logistic regression regret: What's the catch? In *Conference on Learning Theory*, pages 3296–3319. PMLR, 2020.

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.

J. Takeuchi and A. Barron. Asymptotically minimax regret by bayes mixtures for non-exponential families. In *IEEE Information Theory Workshop*, 2006.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

M.E.A.D. White. On the distribution of the cosine of the angle between two random vectors. *The Annals of Mathematical Statistics*, 24(3):429–440, 1953.

Changlong Wu, Mohsen Heidari, Ananth Grama, and Wojciech Szpankowski. Precise regret bounds for log-loss via a truncated bayesian algorithm. In *Advances in Neural Information Processing Systems*, 2022.

Changlong Wu, Mohsen Heidari, Ananth Grama, and Wojciech Szpankowski. Regret bounds for log-loss via bayesian algorithms. *Trans. Information Theory*, 69(9):5971–5989, 2023a.

Changlong Wu, Mohsen Heidari, Ananth Grama, and Wojciech Szpankowski. Expected worst case regret via stochastic sequential covering. *Transactions on Machine Learning Research*, 2023b.

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. Some Technical Lemma

We will need some of the following integral relations.

Lemma 10 *For every constants $c > 0$ and for every integer $k \geq 0$ we have, as $T \to \infty$,*

$$\left(\frac{T}{s_d}\right)^k \int\limits_{[0,\pi/2]^k} \exp\left(-\sqrt{\frac{c}{s_d} T \sum_{t=1}^{k} \sin\theta_t}\right) (\cos\theta_1 \cdots \cos\theta_k)^{d-1} d\theta_1 \cdots d\theta_k$$
$$= O\left(\frac{(2k)!}{k!} c^{-k}\right),$$

where the O-constant depends on c.

Proof We first use the substitution

$$x_t = \frac{T}{s_d} \sin\theta_t \qquad (t = 1, \ldots, T)$$

to find the following integral that we upper bound by

$$\begin{aligned}
&\int_{[0,T/s_d]^k} \exp\left(-\sqrt{c\sum_{t=1}^k x_t}\right) \prod_{t=1}^k \left(1-\frac{x_j^2 s_d^2}{T^2}\right)^{\frac{d-2}{2}} dx_1\cdots dx_k \\
&\le \int_{[0,\infty)^k} \exp\left(-\sqrt{c\sum_{t=1}^k x_t}\right) dx_1\cdots dx_k \\
&= \frac{1}{(k-1)!}\int_0^\infty y^{k-1}\exp\left(-\sqrt{cy}\right) dy \\
&= \frac{1}{(k-1)!}\int_0^\infty y^{k-1}\exp\left(-\sqrt{cy}\right) dy.
\end{aligned}$$

The last integral is easy to compute:

$$\begin{aligned}
\frac{1}{(k-1)!}\int_0^\infty y^{k-1}\exp\left(-\sqrt{cy}\right)\, dy &= \frac{1}{2(k-1)!}c^{-k}(2k-1)! \\
&= \frac{c^{-k}}{2}\frac{(2k)!}{k!}.
\end{aligned}$$

This proves the lemma. ■

Lemma 11 *We have* $w \to \infty$

$$\frac{1}{s_d}\int_0^{\pi/2} \frac{(\cos\theta)^{d-1}}{1+e^{-w\sin\theta}} d\theta = 1 - \frac{\log 2}{s_d w} + O\left(\frac{1}{s_d w^3}\right).$$

Proof By using the relations

$$\frac{1}{1+e^{-w\sin\theta}} = 1 - \frac{e^{-w\sin\theta}}{1+e^{-w\sin\theta}}$$

and

$$\int_0^{\pi/2} (\cos\theta)^{d-1}\, d\theta = s_d$$

and by applying the subsitution $x = w\sin\theta$, we have

$$\begin{aligned}
\frac{1}{s_d}\int_0^{\pi/2} \frac{(\cos\theta)^{d-1}}{1+e^{-w\sin\theta}} d\theta &= 1 - \frac{1}{s_d}\int_0^{\pi/2} \frac{e^{-w\sin\theta}}{1+e^{-w\sin\theta}} (\cos\theta)^{d-1} d\theta \\
&= 1 - \frac{1}{s_d w}\int_0^{w} \frac{e^{-x}}{1+e^{-x}} \left(1-\frac{x^2}{w^2}\right)^{\frac{d-2}{2}} dx \\
&= 1 - \frac{1}{s_d w}\int_0^{\infty} \frac{e^{-x}}{1+e^{-x}}\, dx + O\left(\frac{1}{s_d w^3}\right).
\end{aligned}$$

This proves the lemma. ■

Lemma 12 *Suppose that $w \to \infty$. Then we have uniformly for $\frac{1}{2}w \le w_1 \le 2w$*

$$\frac{1}{s_d}\int_0^{\pi/2} \frac{e^{-w_1 \sin\theta}}{1+e^{-w\sin\theta}} (\cos\theta)^{d-1} d\theta = \frac{\log 2}{s_d w} + O\left(\frac{1}{s_d w^3}\right) + O\left(\frac{|w_1 - w|}{s_d}\right)$$

Proof Again we use the substitution $x = w\sin\theta$ and obtain as in the proof of Lemma 11

$$\frac{1}{s_d}\int_0^{\pi/2} \frac{e^{-w_1 \sin\theta}}{1+e^{-w\sin\theta}} (\cos\theta)^{d-1} d\theta = \frac{1}{s_d w}\int_0^{\infty} \frac{e^{-\frac{w_1}{w}x}}{1+e^{-x}}\, dx + O\left(\frac{1}{s_d w^3}\right).$$

Next we observe that the mapping

$$u \mapsto \int_0^{\infty} \frac{e^{-ux}}{1+e^{-x}}\, dx$$

is continuously differentiable for $\frac{1}{2} \le u \le 2$ (even in a larger interval) which gives

$$\int_0^{\infty} \frac{e^{-ux}}{1+e^{-x}}\, dx = \int_0^{\infty} \frac{e^{-x}}{1+e^{-x}}\, dx + O(|u-1|) = \log 2 + O(|u-1|).$$

Clearly we have

$$\frac{1}{w}\left|\frac{w_1}{w} - 1\right| = O(|w_1 - w|)$$

for $\frac{1}{2}w \le w_1 \le 2w$. This completes the proof of the lemma. ∎

We conclude this section with a (known) concentration lemma (whose proof we include for completeness) that we have used in the proof of the lower bound.

Lemma 13 *Let $X_1, \dots, X_T$ be independent Bernoulli random variables, i.e., $X_i \in \{0,1\}$, and set*

$$S = \sum_{i=1}^{T} X_i, \quad \textit{and} \quad \mu = \mathbf{E}[S] = \sum_{i=1}^{T} \mathbf{E}[X_i].$$

Then for any $s \ge e\mu$, the following bound holds:

$$\mathbf{P}(S \ge s) \le \left(\frac{e\mu}{s}\right)^s.$$

Proof By Chernoff's inequality we have for any $\lambda > 0$,

$$\mathbf{P}(S \ge s) \le \frac{\mathbf{E}[e^{\lambda S}]}{e^{\lambda s}}.$$

Since the X_i are independent, we have

$$\mathbf{E}[e^{\lambda S}] = \prod_{i=1}^{T} \mathbf{E}[e^{\lambda X_i}] = \prod_{i=1}^{T} \left((1-p_i) + p_i e^{\lambda}\right),$$

where $p_i = \mathbf{E}[X_i]$. Using the inequality $1 - p_i + p_i e^{\lambda} \le \exp(p_i(e^{\lambda} - 1))$, we, thus, obtain

$$\mathbf{E}[e^{\lambda S}] \le \exp\left(\mu(e^{\lambda} - 1)\right).$$

and consequently

$$\mathbf{P}(S \geq s) \leq \exp\left(\mu(e^{\lambda} - 1) - \lambda s\right).$$

In order to optimize the bound, we choose λ to minimize the exponent. Setting the derivative of the exponent to zero

$$\frac{d}{d\lambda}\left[\mu(e^{\lambda} - 1) - \lambda s\right] = \mu e^{\lambda} - s = 0,$$

leads to $e^{\lambda} = \frac{s}{\mu}$ and thus $\lambda = \log\left(\frac{s}{\mu}\right)$. Note that by assumption $s \geq e\mu$ so that $\lambda \geq 1$. This leads to

$$\mathbf{P}(S \geq s) \leq \exp\left(\mu\left(\frac{s}{\mu} - 1\right) - s\log\left(\frac{s}{\mu}\right)\right) = \exp\left(s - \mu - s\log\left(\frac{s}{\mu}\right)\right).$$

and by dropping the factor $e^{-\mu} \leq 1$ it simplifies to

$$\mathbf{P}(S \geq s) \leq \left(\frac{e\mu}{s}\right)^{s},$$

which in only significant if $s \geq e\mu$. ■

Appendix B. Upper Bound for Monotone Classes

Let $\mathcal{X}$ be the feature space, we denote $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ as the expert cl ass. We say a sequence $\mathbf{x}^d \in \mathcal{X}^d$ is *pseudo-shattered* by $\mathcal{H}$ witnessed by $\mathbf{s}^d \in \mathbb{R}^d$ if for any $\epsilon^d \in \{0,1\}^d$ there exists $h \in \mathcal{H}$ such that for all $t \in [d]$:

1. If $\epsilon_t = 1$, then $h(\mathbf{x}_t) \geq s_t$;
2. If $\epsilon_t = 0$, then $h(\mathbf{x}_t) < s_t$.

The pseudo-dimension of $\mathcal{H}$ is defined to be the maximum number d such that there exist $\mathbf{x}^d$ that can be pseudo-shattered by $\mathcal{H}$, denoted as $\mathsf{P}(\mathcal{H})$.

Lemma 14 *Let* $\mathcal{H}^{\mathsf{lin}} = \{h_{\mathbf{w}}(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x}\rangle : \mathbf{w}, \mathbf{x} \in \mathbb{R}^d\}$*, then*

$$\mathsf{P}(\mathcal{H}^{\mathsf{lin}}) \leq d.$$

Proof Suppose otherwise, there exists sequence $\mathbf{x}^{d+1}$ that can be pseudo-shattered by $\mathcal{H}$ witnessed by $\mathbf{s}^{d+1}$. For any $\epsilon^{d+1} \in \{0,1\}^{d+1}$, we denote $\bar{\epsilon}^{d+1}$ to be the complement of ϵ^{d+1} such that for all $t \in [d+1]$, $\epsilon_t = 1 - \bar{\epsilon}_t$. Let $\mathbf{w}_1, \mathbf{w}_2$ be the parameters corresponding to the functions in $\mathcal{H}^{\mathsf{lin}}$ that witness ϵ^{d+1} and $\bar{\epsilon}^{d+1}$ respectively. Taking $\mathbf{w}' = \mathbf{w}_1 - \mathbf{w}_2$, we have for all $t \in [d+1]$

1. If $\epsilon_t = 1$, then $h_{\mathbf{w}'}(\mathbf{x}_t) \geq 0$;
2. If $\epsilon_t = 0$, then $h_{\mathbf{w}'}(\mathbf{x}_t) < 0$;

To see this, assume w.o.l.g. $\epsilon_t = 1$, we have $h_{\mathbf{w}_1}(\mathbf{x}_t) \geq s_t$ while $h_{\mathbf{w}_2}(\mathbf{x}_t) < s_t$. Therefore, by linearity, we have $h_{\mathbf{w}'}(\mathbf{x}_t) = h_{\mathbf{w}_1}(\mathbf{x}_t) - h_{\mathbf{w}_2}(\mathbf{x}_t) \geq 0$. This implies that the sequence $\mathbf{x}^{d+1}$ is VC-shattered by $\mathcal{H}^{\mathsf{thres}} = \{1\{\langle \mathbf{w}, \mathbf{x}\rangle \geq 0\} : \mathbf{w}, \mathbf{x} \in \mathbb{R}^d\}$. This contradicts to the fact that VC-dimension of $\mathcal{H}^{\mathsf{thres}}$ is upper bounded by d. ■

We now define a new general monotone hypothesis class:

$$\mathcal{H}^{\mathsf{mono}} := \{p(\langle \mathbf{x}|\mathbf{w}\rangle) : \ p \text{ is monotone increasing } \}.$$

The next lemma follows directly from the definition.

Lemma 15 *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be an arbitrary monotone increasing function. For any class $\mathcal{H}$, we denote by $\mathcal{H}^\sigma = \{\sigma(h) : h \in \mathcal{H}\}$. Then*

$$\mathsf{P}(\mathcal{H}^\sigma) \leq \mathsf{P}(\mathcal{H}).$$

Let $\mathcal{H} \subset [0,1]^{\mathcal{X}}$ be an arbitrary class with values in $[0,1]$. Let $J \subset [0,1]$ be a discretization of $[0,1]$ with step size $2\alpha > 0$, i.e. , $|J| \leq 1/2\alpha$. We define the discretized class

$$\tilde{\mathcal{H}} = \{\tilde{h}(\mathbf{x}) = \arg\min_{a \in J}\{|a - h(x)|\} : h \in \mathcal{H}\},$$

where we break ties by choosing the *larger* one.

Lemma 16 *For any class $\mathcal{H} \subset [0,1]^{\mathcal{X}}$, we have $\mathsf{P}(\tilde{\mathcal{H}}) \leq \mathsf{P}(\mathcal{H})$.*

Proof Let $\mathbf{x}^d$ be the sequence that can be pseudo-shattered by $\tilde{\mathcal{H}}$ witnessed by $\mathbf{s}^d$. We enumerate $J = \{0, 2\alpha, \cdots, 1\}$. For any $t \in [d]$, there must be some i such that $2i\alpha < s_t \leq 2(i+1)\alpha$. We take $s'_t = (2i+1)\alpha$. We now claim that $\mathbf{x}^d$ is pseudo-shattered by $\mathcal{H}$ witnessed by $\mathbf{s}'^d$. For any $\epsilon^d \in \{0,1\}^d$, we take $\tilde{h} \in \tilde{\mathcal{H}}$ such that if $\epsilon_t = 1$ then $\tilde{h}(\mathbf{x}_t) \geq s_t$ and $\tilde{h}(\mathbf{x}_t) < s_t$ otherwise. Let $h \in \mathcal{H}$ be any function with discretization $\tilde{h}$. We have if $\tilde{h}(\mathbf{x}_t) \geq s_t$, then $\tilde{h}(\mathbf{x}_t) \geq s'_t + \alpha$, meaning that $h(\mathbf{x}_t) \geq \tilde{h}(\mathbf{x}_t) - \alpha \geq s'_t$. If $\tilde{h}(\mathbf{x}_t) < s_t$, then $\tilde{h}(\mathbf{x}_t) \leq s_t - \alpha$. This implies $h(\mathbf{x}_t) < \tilde{h}(\mathbf{x}_t) + \alpha \leq s'_t$, where we used the fact that the discretization resolves ties by choosing the larger one. This completes the proof. ■

The following lemma bounds the size of the discretized class w.r.t. pseudo-dimension, which is due to (Haussler and Long, 1995, Corollary 3).

Lemma 17 *For any discretized class $\tilde{\mathcal{H}} \subset J^{\mathcal{X}}$ and $\mathbf{x}^T \in \mathcal{X}^T$ we denote by $\tilde{\mathcal{H}}_{\mathbf{x}^T}$ the class of $\tilde{\mathcal{H}}$ restricted on $\mathbf{x}^T$. Then*

$$|\tilde{\mathcal{H}}_{\mathbf{x}^T}| \leq (T|J|)^{\mathsf{P}(\tilde{\mathcal{H}})}.$$

We now state our main upper bound:

Lemma 18 *For any $\mathbf{x}^T \in \mathbb{R}^{dT}$, we have*

$$r_T^*(\mathcal{H}^{\mathsf{mono}} \mid \mathbf{x}^T) \leq 2d \log T + 1.$$

Proof Let $\tilde{\mathcal{H}}$ be the discretization of $\mathcal{H}^{\mathsf{mono}}$ with step size $2/T$. Note that $\tilde{\mathcal{H}}$ is an $1/T$-cover of $\mathcal{H}^{\mathsf{mono}}$, meaning that

$$r_T^*(\mathcal{H}^{\mathsf{mono}} \mid \mathbf{x}^T) \le r_T^*(\tilde{\mathcal{H}} \mid \mathbf{x}^T) + 1.$$

By the above Lemmas 15–17 , we have $\mathsf{P}(\tilde{\mathcal{H}}) \le d$. By Lemma 15, we conclude that

$$\begin{aligned} r_T^*(\tilde{\mathcal{H}} \mid \mathbf{x}^T) &\le \log |\tilde{\mathcal{H}}_{\mathbf{x}^T}| \\ &\le d \log(T|J|) \le d \log(T^2/2) \le 2d \log T. \end{aligned}$$

This completes the proof. ■