On the Sample Complexity of Statistical Learning: Estimation, Prediction and Decision Making

by

Zeyu Jia

B.S., Peking University (2020).
S.M., Massachusetts Institute of Technology (2022).

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2026

Authored by: Zeyu Jia
Department of Electrical Engineering and Computer Science
May 5, 2026

Certified by: Yury Polyanskiy
Professor of Electrical Engineering and Computer Science, Thesis Supervisor

Certified by: Alexander Rakhlin
Professor of Brain and Cognitive Sciences, Thesis Supervisor

Accepted by: Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

THESIS COMMITTEE

Thesis Supervisors

Yury Polyanskiy
Professor of Electrical Engineering and Computer Science

Alexander Rakhlin
Professor of Brain and Cognitive Sciences

Thesis Readers

Yury Polyanskiy
Professor of Electrical Engineering and Computer Science

Alexander Rakhlin
Professor of Brain and Cognitive Sciences

Guy Bresler
Associate Professor of Electrical Engineering and Computer Science

On the Sample Complexity of Statistical Learning: Estimation, Prediction and Decision Making

by

Zeyu Jia

Submitted to the Department of Electrical Engineering and Computer Science
on May 5, 2026 in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

ABSTRACT

This thesis studies the sample complexity and data efficiency of statistical learning across three foundational settings: Estimation, Prediction, and Decision Making.

In the first part, we investigate density estimation for Gaussian models. We characterize the exact convergence rates of smoothed Wasserstein distances for empirical measures of Gaussian mixture distributions with subgaussian components, completely closing the gap between the parametric and non-parametric regimes. We then establish a Kullback–Leibler versus squared Hellinger comparison for Gaussian mixtures that shows minimax density estimation is tightly governed by Hellinger entropy, resolving a longstanding open problem. Finally, we study the relationship between goodness-of-fit testing and density estimation in the Gaussian sequence model, obtaining tight sample complexity reductions between the two tasks, and extend our analysis to likelihood-free hypothesis testing over ℓ_p bodies.

In the second part, we study sequential prediction and regression. For sequential probability assignment under logarithmic loss, we develop new techniques for controlling Rademacher processes with unbounded increments and introduce a notion of sequential square-root entropy that tightly captures the complexity of general function classes. We also study structural lower bounds for online regression by formulating a "gapped" scale-sensitive dimension, which yields non-vacuous lower bounds on offset Rademacher complexity for any bounded function class.

In the third part, we address reinforcement learning with outcome-based feedback. We first show, via a change of trajectory measure argument, that offline reinforcement learning from trajectory-level rewards can achieve the same statistical efficiency as step-by-step process supervision, providing theoretical justification for using advantage functions as optimal process reward models. We then study online exploration under general function approximation with only outcome-level feedback, establishing matching upper and lower bounds on the sample complexity of active exploration in this challenging setting.

Thesis supervisor: Yury Polyanskiy
Title: Professor of Electrical Engineering and Computer Science

Thesis supervisor: Alexander Rakhlin
Title: Professor of Brain and Cognitive Sciences

Acknowledgments

I would like to express my deepest gratitude to my advisors, Yury Polyanskiy and Sasha Rakhlin. Yury introduced me to the world of information theory and topics in statistics, including density estimation and goodness-of-fit testing, expertly guiding me through the scientific process. When we wrote our first paper together, I struggled with academic writing; Yury patiently taught me how to convey my ideas clearly and effectively. His mentorship profoundly shaped my writing style and made me a fundamentally better researcher. Sasha has been equally central to my success, continuously providing invaluable guidance. Over the first five years of my PhD, I worked on a long-term project concerning sequential prediction and encountered numerous technical hurdles. Through it all, Sasha consistently encouraged me to push past these difficulties, offering wisdom and relentless support.

I also wish to warmly thank my thesis committee member, Guy Bresler. I first had the pleasure of interacting with Guy during my research qualification exam, and he has been incredibly supportive of my research ever since.

I am deeply grateful to all of my phenomenal collaborators: Adam Block, Alex Ayoub, Ayush Sekhari, Chen-Yu Wei, Dhruv Madeka, Csaba Szepesvari, Dean Foster, Fan Chen, Gene Li, Jian Qian, Lin F. Yang, Mengdi Wang, Nathan Srebro, Randy Jia, Sasha Rakhlin, Tengyang Xie, Yifei Wang, Yihong Wu, Yaqi Duan, Yinyu Ye, Yurun Yuan, Yury Polyanskiy, and Zaiwen Wen. Collaborating alongside them has been an absolute privilege, and I have learned an immense amount from each of them.

My time at MIT was profoundly enriched by the friendships I forged, both within and outside my immediate research group. I am deeply grateful for the camaraderie and day-to-day support of my peers, including but not limited to Adam Block, Ayush Sekhari, Chenghao Guo, Feng Zhu, Haochuan Li, Jian Qian, Tiancheng Yu, Yunbei Xu, and Yuzhou Gu. The countless insightful discussions and memorable moments we shared made this graduate journey truly unforgettable.

Last but certainly not least, I would like to thank my family for their unconditional love and support throughout this journey. In particular, my deepest gratitude goes to my wife, Xiwen Hu. Because of her, Washington, D.C., became my second home during my PhD, and much of the research presented in this thesis was carried out during our time there. Her steadfast encouragement, unwavering support for my career, and endless care have been truly invaluable.

Contents

Chapter 1

Introduction

Statistical learning theory seeks to understand the fundamental limits of how much information can be extracted from finite data, establishing mathematical guarantees for modern algorithms. This thesis explores the sample complexity and data efficiency of three foundational pillars in the field: Estimation, Prediction, and Decision Making.

The first pillar, Estimation, examines the classical statistical problem of recovering an underlying probability distribution or signal from noisy observational data [1, 2]. The second pillar, Prediction, shifts focus to sequential environments where an algorithm must process streaming data and make accurate forecasts on the fly, typically aiming to minimize regret against optimal hindsight strategies [3, 4]. The final pillar, Decision Making, addresses the interactive learning paradigm where an agent actively explores its environment to maximize rewards, a hallmark of modern reinforcement learning [5]. Across these three settings, we leverage structured tools from information theory, empirical process theory, and concentration inequalities to establish tight characterizations of minimax rates of convergence, algorithmic sample efficiency, and the fundamental capacity limits of learning models.

1.1 Density Estimation for Gaussian Models

The first part of this thesis focuses on Density Estimation for Gaussian Models. We start with the fundamental problem of how fast empirical measures converge for Gaussian mixture models, and analyze the convergence under the Wasserstein distance after smoothing. While optimal transport metrics like the Wasserstein distance are deeply important for modern statistics and machine learning [6], un-smoothed empirical measures typically suffer from the severe curse of dimensionality [7]. Recent work has shown that smoothing with Gaussian noise can drastically improve these rates [8]. In Chapter 2, we completely close the theoretical gap between parametric and non-parametric regimes, characterizing the exact convergence rates for Gaussian mixture distributions with subgaussian components under smoothed Wasserstein distances.

Next, we explore density estimation through the lens of metric entropy for Gaussian mixture models. A massive literature has sought to characterize minimax density estimation rates via local or global metric entropy bounds, originating from classical nonparametric theories [1, 9–11]. For Gaussian mixtures, we prove an upper bound to the Kullback-Leibler

divergence by the squared Hellinger distance up to polynomial factors in the dimension. This result implies that minimax estimation risks for Gaussian mixtures with bounded or subgaussian components are tightly governed by the Hellinger entropy, effectively resolving a longstanding bottleneck in the density estimation literature (Chapter 3).

Finally, we study the interplay between the sample complexities of goodness-of-fit testing and density estimation in the Gaussian sequence model. Historically, nonparametric statistics has treated estimation [2, 12] and testing [13, 14] as largely separate problems. In this work, we establish upper and lower bounds on the minimax sample complexity of goodness-of-fit testing in terms of the sample complexity of density estimation over convex, compact, and orthosymmetric sets. We further extend this analysis to the likelihood-free hypothesis testing setting [15], where we characterize the optimal testing region for ℓ_p bodies (Chapter 4).

1.2 Sequential Prediction and Regression

The second part of this thesis turns to the sequential prediction and regression framework, focusing on the problem of optimizing predictions over time. We begin with sequential probability assignment under logarithmic loss, a setting that has been widely studied in online learning, sequence compression, and modern next-token prediction [4]. Prior work has struggled with the unbounded nature of the logarithmic loss [16–18]. In contrast, we develop new techniques for controlling the expectation of Rademacher processes with unbounded coefficients and analyze the minimax regret via a novel notion of sequential square-root entropy. We show that this quantity succinctly captures the intrinsic complexity of function classes, leading to tight upper bounds for contextual prediction tasks (Chapter 5).

Next, we shift our focus to proving structural lower bounds for both classical and sequential regression. The classical scale-sensitive dimension and its sequential analogues are central to the study of uniform martingale laws and online prediction [3, 19, 20]. However, standard definitions often fail to reliably control offset Rademacher averages. To resolve this, we formulate a "gapped" version of the scale-sensitive dimension, building upon older multi-class variants [21]. We demonstrate the utility of this gapped dimension in proving non-vacuous, stronger lower bounds for offset Rademacher complexity for any bounded class, establishing tighter performance limits for online regression and transductive learning (Chapter 6).

1.3 Reinforcement Learning with Outcome-Based Feedback

We next move to the part of the thesis that addresses Decision Making through the study of Reinforcement Learning with Outcome-Based Feedback. Traditionally, reinforcement learning relies on granular, step-by-step reward signals to facilitate credit assignment [5]. However, in modern AI applications, most notably in training large language models (LLMs) to perform complex reasoning tasks [22–24], collecting detailed step-by-step human evaluations (process supervision) is often prohibitively expensive [e.g., 25, 26]. Instead, supervision is predominantly outcome-based, where feedback or preferences are only provided over the entire generated response. While outcome supervision offers clear practical advantages for

scalable data collection, it presents a significantly harder temporal credit assignment problem, raising fundamental questions about its baseline statistical efficiency. First, we investigate the statistical guarantees of offline reinforcement learning from total trajectory-level rewards versus step-by-step process supervision. By developing a novel change of trajectory measure analysis, we demonstrate the surprising result that outcome-based feedback can often achieve equivalent statistical efficiency to process supervision without suffering an exponential blow-up in sample complexity. Furthermore, we provide a theoretical justification for using advantage functions as optimal process reward models to impute missing rewards, fundamentally challenging the conventional wisdom that explicit step-by-step verification is strictly necessary (Chapter 7).

While outcome-based feedback is sufficient for offline learning in certain conditions, the feasibility of efficient *online* exploration with only trajectory-level feedback presents a unique challenge. Online learning, where an agent actively explores to gather new data, is central to adaptive systems that must continually navigate dynamic environments without relying solely on pre-collected datasets. Previous studies on online exploration with outcome or preference feedback have primarily assumed well-structured, linear reward functions [27–31] or focused narrowly on specific RLHF paradigms [32–35]. In Chapter 8, we generalize beyond these constraints by establishing the fundamental statistical limits and developing sample-efficient algorithms for active online RL exploration under general function approximation when only outcome rewards are available.

1.4 Bibliographical Remarks

This thesis is based on the following works. Chapter 2 is based on Block, Jia, Polyanskiy, and Rakhlin. Chapter 3 is based on Jia, Polyanskiy, and Wu. Chapter 4 is based on Jia and Polyanskiy. Chapter 5 is based on Jia, Polyanskiy, and Rakhlin. Chapter 6 is based on Jia, Polyanskiy, and Rakhlin. Chapter 7 is based on Jia, Rakhlin, and Xie. Chapter 8 is based on Chen, Jia, Rakhlin, and Xie.

Part I

Density Estimation for Gaussian Models

Chapter 2

Rate of convergence of the smoothed empirical Wasserstein distance

This chapter is based on the paper "Rate of convergence of the smoothed empirical Wasserstein distance" by Adam Block, Zeyu Jia, Yury Polyanskiy, and Alexander Rakhlin, published at the Annales de l'Institut Henri Poincaré (AIHP).

2.1 Introduction and main results

Given n iid samples $X_1, \dots, X_n$ from a probability measure $\mathbb{P}$ on $\mathbb{R}^d$ let us denote by $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ the empirical distribution. As $n \to \infty$ it is well known that $\mathbb{P}_n \to \mathbb{P}$ according to many different notions of convergence. The literature on the topic is voluminous. Here we are interested in convergence in Wasserstein W_p-distances, cf. Villani [6], defined for $p \geq 1$ as

$$W_p(\mathbb{P}, \mathbb{Q})^p = \inf_{P_{X,Y}} \{\mathbb{E}[\|X - Y\|^p] : P_X = \mathbb{P}, P_Y = \mathbb{Q}\},$$

where $\|\cdot\|$ is Euclidean norm. Already in Dudley [7] it was shown that

$$W_1(\mathbb{P}_n, \mathbb{P}) = \Theta(n^{-1/d}),$$

for $d \geq 3$ and compactly supported $\mathbb{P}$ absolutely continuous with respect to Lebesgue measure. Dudley's technique relied on the characterization (special to $p = 1$) of W_1 as the supremum over expectations of Lipschitz functions. His idea of recursive partitioning was cleverly adapted to the realm of couplings in Boissard and Le Gouic [36], recovering Dudley's convergence rate of $n^{-1/d}$ also for $p > 1$. For standard Gaussian distributions, Bobkov and Ledoux [37] settles the rate of $\Theta(\frac{\log \log n}{n})$ for one-dimensional distributions, Berthet and Fort [38] proves that the constant in the Θ is 1, i.e. $\frac{\mathbb{E}[W_1(\mathbb{P}_n, \mathbb{P})]}{(\log \log n)/n} = 1 + o(1)$. Ledoux [39] and Talagrand [40] settles the rate of $\Theta(\frac{\log^2 n}{n})$ for two-dimensional distributions, and Bobkov and Ledoux [37] settles the rate of $\Theta(\frac{1}{n(\log n)^{p/2}})$ for d dimensional distributions with $d \geq 3$. See Dereich, Scheutzow, and Schottstedt [41], Fournier and Guillin [42], and Weed and Bach [43] for more on this line of work, and also for a thorough survey of the recent literature.

Somewhat surprisingly, it was discovered in Goldfeld et al. [8] that the rate of convergence improves all the way to (dimension-independent) $n^{-1/2}$ if one merely regularizes both $\mathbb{P}_n$ and

$\mathbb{P}$ by convolving with the Gaussian density.[1] More precisely, let $\varphi_{\sigma^2}(x) \triangleq (2\pi\sigma^2)^{-d/2} e^{-\frac{\|x\|^2}{2\sigma^2}}$ be the density of $\mathcal{N}(0, \sigma^2 I_d)$, and for any probability measure $\mathbb{P}$ on $\mathbb{R}^d$ we define the convolved measure via

$$\mathbb{P} * \mathcal{N}(0, \sigma^2 I_d)(E) = \int_E dz \mathbb{E}\left[\varphi_{\sigma^2}(X - z)\right], \quad X \sim \mathbb{P},$$

where E is any Borel set. Then Goldfeld et al. [8, Prop. 6] shows

$$\mathbb{E}[W_2^2(\mathbb{P}_n * \mathcal{N}(0, \sigma^2 I_d), \mathbb{P} * \mathcal{N}(0, \sigma^2 I_d))] \le \frac{C(d, \sigma, K)}{n}, \tag{2.1}$$

whenever $\mathbb{P}$ is K-subgaussian and $K < \frac{\sigma}{2}$. We recall that $X \sim \mathbb{P}$ is K-subgaussian if

$$\mathbb{E}[e^{(\lambda, X - \mathbb{E}[X])}] \le e^{\frac{1}{2}K^2\|\lambda\|^2} \qquad \forall \lambda \in \mathbb{R}^d. \tag{2.2}$$

Note that in (2.1) constant C does not depend on $\mathbb{P}$. Estimate (2.1) is most exciting for large d, but even for $d = 1$ and $\mathbb{P} = \mathcal{N}(0, 1)$ it is non-trivial as $\mathbb{E}[W_2^2(\mathbb{P}_n, \mathbb{P})] = \Theta(\frac{\log\log n}{n})$ (see Bobkov and Ledoux [37] and Berthet and Fort [38]). Another surprising feature is Goldfeld et al. [8, Corollary 2]: for $K \ge \sqrt{2}\sigma$ there exists a K-subgaussian distribution $\mathbb{P}$ in $\mathbb{R}^1$ such that

$$\lim_{n\to\infty} n\mathbb{E}[W_2^2(\mathbb{P}_n * \mathcal{N}(0, \sigma^2 I_d), \mathbb{P} * \mathcal{N}(0, \sigma^2 I_d))] = \infty, \tag{2.3}$$

where the expectation is with respect to n samples according to $\mathbb{P}$. We say that the rate of convergence is "parametric" if (2.1) holds and otherwise call it "non-parametric". Thus, the results of Goldfeld et al. [8] show that parametric rate for smoothed-W_2 is only attained by sufficiently light-tailed distributions $\mathbb{P}$, e.g. the subgaussian constant of distribution $\mathbb{P}$ is less than the scale of noise over two.

In this chapter we prove three principal results:

1. Theorem 2.1 resolves the gap between the location of the parametric and non-parametric region: it turns out that for $K < \sigma$ we always have (2.1), while for $K > \sigma$ we have (2.3) for some K-subgaussian distribution $\mathbb{P}$ in $\mathbb{R}^1$. (We remark that for W_1 we always have parametric rate $n^{-1/2}$ for all $K, \sigma > 0$, cf Goldfeld et al. [8, Proposition 1].)

2. In the region of non-parametric rates ($K > \sigma$) a natural question arises: what rates of convergence are possible? In other words, what is the value of

$$\alpha = \alpha(K, \sigma, d) \triangleq \inf_{\mathbb{P}-K\text{-subgaussian}} \lim_{n\to\infty} -\frac{\log \mathbb{E}[W_2(\mathbb{P}_n 1 * \mathcal{N}(0, \sigma^2 I_d), \mathbb{P} * \mathcal{N}(0, \sigma^2 I_d))]}{\log n}. \tag{2.4}$$

   Previously, it was only known that $\frac{1}{4} \le \alpha \le \frac{1}{2}$ for all $K > \sigma$. The lower bound $1/4$ of α follows from Goldfeld et al. [8]. Theorem 2.4 shows that for $d = 1$ we have

$$\alpha(K, \sigma, d = 1) = \frac{(\sigma^2 + K^2)^2}{4(\sigma^4 + K^4)}.$$

[1] Of course, the price to pay for this fast rate is a constant in front of $n^{-1/2}$, which can be exponential in d for certain $\mathbb{P}$, cf Goldfeld et al. [8].

3. We can see that for a class of K-subgaussian distributions, the convergence rate of $W_2(\mathbb{P}_n * \mathcal{N}(0, \sigma^2 I_d), \mathbb{P} * \mathcal{N}(0, \sigma^2 I_d))$ changes from $n^{-1/4}$ to $n^{-1/2}$ as σ increases from 0 to K, after which the rate remains $n^{-1/2}$. Our final result (Theorem 2.5) shows that, despite being intimately related to W_2, the Kullback-Leibler (KL) divergence behaves rather differently: For all K-subgaussian $\mathbb{P}$ we have

$$\mathbb{E}[D_{KL}(\mathbb{P}_n * \mathcal{N}(0, \sigma^2 I_d) \| \mathbb{P} * \mathcal{N}(0, \sigma^2 I_d))] \leq \begin{cases} \frac{C(\sigma, K, d) \log^d n}{n}, & K > \sigma \\ \frac{C(\sigma, K, d)}{n}, & K < \sigma \end{cases}, \tag{2.5}$$

where $D_{KL}(\mu \| \nu) = \int d\nu f(x) \log f(x)$, $f \triangleq \frac{d\mu}{d\nu}$ whenever μ is absolutely continuous with respect to ν.

2.1.1 Main results and proof ideas

Our first result is the following:

Theorem 2.1. *If $K < \sigma$, then for any K-subgaussian distribution $\mathbb{P}$, we have*

$$\mathbb{E}\left[W_2^2(\mathbb{P}_n * \mathcal{N}(0, \sigma^2 I_d), \mathbb{P} * \mathcal{N}(0, \sigma^2 I_d))\right] = \mathcal{O}\left(\frac{1}{n}\right),$$

where $\mathbb{P}_n$ is the empirical measure of $\mathbb{P}$ with n samples, and the expectation is over these n samples. If $K > \sigma$, then there exists a K-subgaussian distribution $\mathbb{P}$ such that

$$\mathbb{E}\left[W_2^2(\mathbb{P}_n * \mathcal{N}(0, \sigma^2 I_d), \mathbb{P} * \mathcal{N}(0, \sigma^2 I_d))\right] = \omega\left(\frac{1}{n}\right).$$

Previous results. Goldfeld et al. [8] shows when $K < \sigma/2$, $\mathbb{E}\left[W_2^2(\mathbb{P}_n * \mathcal{N}(0, \sigma^2 I_d), \mathbb{P} * \mathcal{N}(0, \sigma^2 I_d))\right]$ converges with rate $\mathcal{O}\left(\frac{1}{n}\right)$; when $K > \sqrt{2}\sigma$, $\mathbb{E}\left[W_2^2(\mathbb{P}_n * \mathcal{N}(0, \sigma^2 I_d), \mathbb{P} * \mathcal{N}(0, \sigma^2 I_d))\right]$ converges with rate $\omega\left(\frac{1}{n}\right)$. Here is an obvious gap between $K < \sigma/2$ and $K > \sqrt{2}\sigma$, and our results close this gap.

Proof Idea. Let us introduce the χ^2-mutual information for a pair of random variables S, Y as

$$I_{\chi^2}(S; Y) \triangleq \chi^2(P_{S,Y} \| P_S \otimes P_Y),$$

where $\chi^2(P\|Q) = \int \left(\frac{dP}{dQ}\right)^2 dQ - 1$.

We will consider the case where $S \sim \mathbb{P}$, $Y = S + Z$ with $Z \sim \mathcal{N}(0, \sigma^2)$ independent to S. According to Goldfeld et al. [8], the convergence rate of smoothed empirical measure under W_2, KL-divergence and the χ^2-divergence is closely related to $I_{\chi^2}(S; Y)$:

- (**Prop. 6** in Goldfeld et al. [8]): If $\mathbb{P}$ is K-subgaussian with $K < \sigma$ and $I_{\chi^2}(S; Y) < \infty$, then $\mathbb{E}\left[W_2^2(\mathbb{P}_n * \mathcal{N}(0, \sigma^2 I_d), \mathbb{P} * \mathcal{N}(0, \sigma^2 I_d))\right] = \mathcal{O}(\frac{1}{n})$.
- (**Corr. 2** in Goldfeld et al. [8]): If $I_{\chi^2}(S; Y) = \infty$, then for any $\tau < \sigma$, $\mathbb{E}\left[W_2^2(\mathbb{P}_n * \mathcal{N}(0, \tau^2 I_d), \mathbb{P} * \mathcal{N}(0, \tau$ $\omega(\frac{1}{n})$.

Hence our results follow from the main technical propositions 2.2 and 2.3. □

Proposition 2.2. *When $K < \sigma$, for any K-subgaussian d-dimensional distribution $\mathbb{P}$, we have $I_{\chi^2}(S;Y) < \infty$, where $S \sim \mathbb{P}, Z \sim \mathcal{N}(0, \sigma^2 I_d), S \perp\!\!\!\perp Z$ and $Y = S + Z$.*

Proposition 2.3. *When $K > \sigma$, there exists some 1-dimensional K-subgaussian distribution $\mathbb{P}$ such that $I_{\chi^2}(S;Y) = \infty$ for $S \sim \mathbb{P}, Z \sim \mathcal{N}(0, \sigma^2), S \perp\!\!\!\perp Z$ and $Y = S + Z$.*

Next, we give a tight characterization on the W_2-convergence rate in dimension $d = 1$.

Theorem 2.4. *[Improved bounds for dimension-1] Fix $K > \sigma > 0$ and let*

$$\alpha = \frac{(\sigma^2 + K^2)^2}{4(\sigma^4 + K^4)}. \tag{2.6}$$

1. *(Lower Bound) There exist a K-subgaussian distribution $\mathbb{P}$ over $\mathbb{R}$ and $0 < \delta_n \to 0$ such that*
$$\limsup_{n\to\infty} n^{\alpha+\delta_n} \mathbb{E}\left[W_2(\mathbb{P}_n * \mathcal{N}(0, \sigma^2 I_d), \mathbb{P} * \mathcal{N}(0, \sigma^2 I_d))\right] > 0.$$
2. *(Upper Bound) There exists a sequence $0 < \delta_n \to 0$ such that for any 1-dimensional K-subgaussian $\mathbb{P}$ over $\mathbb{R}$ and $n \geq 2$, we have*
$$\mathbb{E}\left[W_2^2(\mathbb{P} * \mathcal{N}(0, \sigma^2), \mathbb{P}_n * \mathcal{N}(0, \sigma^2))\right] \leq n^{-2\alpha+\delta_n} \tag{2.7}$$

Finally we provide an upper bound on the convergence of smoothed empirical measures under KL divergence:

Theorem 2.5. *Suppose $\mathbb{P}$ is a d-dimensional K-subgaussian distribution, then for any $\sigma > 0$, we have*

$$\mathbb{E}\left[D_{KL}\left(\mathbb{P}_n * \mathcal{N}(0, \sigma^2 I_d) \middle\| \mathbb{P} * \mathcal{N}(0, \sigma^2 I_d)\right)\right] = \mathcal{O}\left(\frac{(\log n)^d}{n}\right).$$

2.2 Proof of Proposition 2.2

In this section, we provide a proof of Proposition 2.2. Detailed technical lemmas used in the following proof are provided in appendix A.

Proof. By decomposing $\mathbb{R}^d$ into cubes c_i of unit diameter and using the K-subgaussianity of S, we show that for $K < \sigma$, the χ^2-divergence remains bounded. Specifically, we use the property that for such distributions, $\mathbb{E}[\exp(\lambda\|S\|^2)] < \infty$ for sufficiently small λ. The proof proceeds by bounding the density ratio $\rho(\mathbf{y}) = \mathbb{E}[\exp(-\|\mathbf{y} - S\|^2/(2\sigma^2))]$ and demonstrating that the integral $\int \varphi_{\sigma^2}(\mathbf{y} - S)^2/\rho(\mathbf{y})d\mathbf{y}$ is finite when averaged over S. □

2.3 Proof of Proposition 2.3

The main idea of this proof is to construct a hard example $\mathbb{P} = \sum_{k=0}^{\infty} p_k \delta_{r_k}$ where points are spaced far apart. Specifically, we choose r_k as a geometric sequence. This construction ensures that the smoothed empirical measure converges slower than the parametric rate.

Proof Sketch. We construct $\mathbb{P} = \sum p_k \delta_{r_k}$ with $p_k \sim \exp(-r_k^2/2K^2)$. By ensuring r_{k+1}/r_k is sufficiently large, we show that the $I_{\chi^2}(S;Y)$ term diverges, which by Goldfeld et al. [8] implies the non-parametric rate. □

2.4 Summary of Non-Parametric Rates

In the regime where $K > \sigma$, we establish that the convergence rate is precisely given by $n^{-\alpha}$ (up to log factors) where α is defined in (2.6). The proof involves a novel conditioning and truncation argument for the 2-Wasserstein distance, leveraging the explicit transport map formula in 1D. Detailed proofs for the lower and upper bounds of Theorem 2.4 can be found in appendix A.1.

Chapter 3

Entropic characterization of optimal rates for learning Gaussian mixtures

This chapter is based on the paper "Entropic characterization of optimal rates for learning Gaussian mixtures" by Zeyu Jia, Yury Polyanskiy, and Yihong Wu, published at the Conference on Learning Theory (COLT) 2023.

3.1 Introduction

Gaussian mixtures are among the most popular and useful classes of distributions for modeling real data with heterogeneity. Specifically, each d-dimensional *mixing distribution* π induces a Gaussian *mixture* f_π, which is the convolution of π with the d-dimensional standard Gaussian distribution $\mathcal{N}(0, I_d)$, namely

$$f_\pi(x) = (\pi * \varphi)(x) = \int_{\mathbb{R}^d} \varphi(x-z)\pi(dz),$$

where $\varphi(z) = \frac{1}{\sqrt{2\pi}^d} \exp\left(-\|z\|_2^2/2\right)$ is the standard normal density.

There is a vast literature in statistics and machine learning on various aspects of mixture models such as parameter estimation and clustering. In this chapter we focus on learning the mixture model in the sense of density estimation. To this end, it is necessary to impose tail conditions on the mixing distribution. Specifically, we consider two classes of Gaussian mixtures classes, wherein the mixing distribution is either compactly supported or subgaussian.

To measure the density estimation error, it is common to use f-divergences, notably, Kullback-Leibler (KL) divergence $D_{\mathrm{KL}}(f\|g) = \int f \log \frac{f}{g}$, the squared Hellinger distance $H^2(f,g) = \int(\sqrt{f} - \sqrt{g})^2$, and the total variation distance $D_{\mathrm{TV}}(f,g) = \frac{1}{2}\int |f-g|$. In this chapter, we are chiefly concerned with the estimation rates under the KL-divergence and the squared Hellinger distance, as opposed to the L_2 distance due to its lack of operational meaning.

Estimating Gaussian mixture densities is a classical topic in nonparametric statistics. Under the squared Hellinger loss, the minimax lower bound $\Omega((\log n)^d/n)$ was proved in [44, 45] for subgaussian mixing distributions. On the constructive side, nonparametric maximum likelihood estimator (NPMLE) and sieve MLE have been analyzed in [46–52]. In particular,

the NPMLE, which offers a practical algorithm for properly learning the mixture model, is shown to achieve a near-parametric rate of $O((\log n)^2/n)$ for the subgaussian case [52], which is subsequently generalized to $O(\log^{d+1} n/n)$ in d dimensions [53]. Similar results for compactly supported mixing distribution are also obtained in [54, Theorem 20]. Despite these advances, determining the optimal rate remains a long-standing open question even in one dimension.

Departing from maximum likelihood, there is a long line of work that aims at characterizing density estimation rates in terms of metric entropy of the class. These general entropic upper bounds originate from the seminal work of [1, 9, 10] for the Hellinger loss, [55] for the D_{TV} loss, and [56] for the KL loss. On the other hand, entropic lower bounds for KL and Hellinger losses were established for both the batch [57] and sequential estimation [56]. However, these entropy-based upper and lower bounds in general do not match unless extra conditions are imposed on the behavior on the model class. Notably, a simple condition that ensures a sharp entropic determination of the minimax rate is the comparability of the Hellinger and KL divergence, namely, for any density f and g in the model class:

$$D_{\mathrm{KL}}(f\|g) \asymp H^2(f,g) \tag{3.1}$$

where $\asymp$ denotes equality within constant multiplicative factors. Note that the one-sided inequality $D_{\mathrm{KL}}(f\|g) \geq H^2(f,g)$ is always true cf. e.g. [58]. As such, whenever D_{KL} is dominated by H^2, the sharp minimax rate is determined by the local Hellinger entropy of the model class.

Indeed, this entropy-based approach has been successfully taken in [59] to determine the sharp rate for the special case of *finite-component* Gaussian mixtures in general dimensions. Specifically, for the class of k-component GMs, [59, Theorem 4.2] shows that

$$D_{\mathrm{KL}}(f_\pi\|f_\eta) \asymp_k H^2(f_\pi, f_\eta), \tag{3.2}$$

where π and η are k-atomic distributions supported on a Euclidean ball of bounded radius in $\mathbb{R}^d$. The crucial part of (3.2) is that it does not depend on the ambient dimension d. As such, this allows the optimal squared Hellinger rate to be determined by the local entropy, which, in turn, can be tightly estimated via the low rank of the moment tensor, leading to the sharp rate of $\Theta_k(\frac{d}{n})$ that holds even in high dimensions.

In this chapter we resolve the question of KL to Hellinger comparison and show that with a constant factor that depends (at most linearly) on the dimension, by proving that

$$D_{\mathrm{KL}}(f_\pi\|f_\eta) \asymp_d H^2(f_\pi, f_\eta), \tag{3.3}$$

where π and η are arbitrary distributions supported on a bounded ball in $\mathbb{R}^d$; furthermore, this result can be made dimension-free with an extra logarithmic factor. In addition, we show that (3.3) holds for $(1-\varepsilon)$-subgaussian mixing distributions but fails for $(1+\varepsilon)$-subgaussian distributions.

The new comparison result has various statistical consequences, of which we report here one (see Corollary 3.9). To estimate the GM density with compactly supported or $(1-\varepsilon)$-subgaussian mixing distributions based on an iid sample of size n, the minimax proper or improper density estimation risks under KL divergence or squared Hellinger distance are

tightly characterized by the *local* Hellinger entropy of the density class. Furthermore, the minimax risks in the sequential version (as opposed to the batch setting above) of this problem are tightly characterized by the *global* Hellinger entropy of the class.

Notation Let $B_2(r) = \left\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 \le r\right\}$ denote the Euclidean ball of radius r centered at 0. Denote by $\mathtt{supp}(\pi)$ the support of a probability measure π. We say a distribution π on $\mathbb{R}^d$ is K-subgaussian if for $X \sim \pi$,

$$\mathbf{P}[\|X\|_2 > t] \le \exp\left(-\frac{t^2}{2K^2}\right), \quad \forall t \ge 0.$$

3.2 Main Results

Before discussing their statistical consequences, we first state the main comparison results that control the KL divergence between Gaussian mixtures using their Hellinger distance.

For compactly supported mixing distributions, our main result is as follows:

Theorem 3.1. *Let π and η be supported on $B_2(M)$ in $\mathbb{R}^d$ where $M \ge 2$. Then*

$$D_{\mathrm{KL}}(f_\pi \| f_\eta) \le 5154(M^2 \vee d) H^2(f_\pi, f_\eta).$$

Complementing Theorem 3.1, we also have the following dimension-free upper bound at the price of a mere logarithmic factor.

Theorem 3.2. *Let π and η be supported on $B_2(M)$ in $\mathbb{R}^d$ where $M \ge 1$. Then*

$$D_{\mathrm{KL}}(f_\pi \| f_\eta) \le 200 M^2 H^2(f_\pi, f_\eta) + 16 H^2(f_\pi, f_\eta) \log \frac{1}{H^2(f_\pi, f_\eta)}.$$

Next we consider the class of subgaussian mixing distributions. When $K < 1$, the KL divergence is indeed proportional to the squared Hellinger distance. When $K > 1$, such upper bound does not exist.

Theorem 3.3. *Let π, η be two d-dimensional K-subgaussian distributions where $K < 1$. Then*

$$D_{\mathrm{KL}}(f_\pi \| f_\eta) \le 1660056 \left(\frac{1}{(1-K)^3} \vee 8d^3\right) H^2(f_\pi, f_\eta).$$

Theorem 3.4. *Fix $K > 1$. For any $C > 0$, there exists a 1-dimensional K-subgaussian distribution π such that*

$$D_{\mathrm{KL}}(f_\pi \| \mathcal{N}(0,1)) \ge C \cdot H^2(f_\pi, \mathcal{N}(0,1)).$$

We further have the following dimension-free upper bound that holds for all $K > 0$.

Theorem 3.5. *Let π and η be K-subgaussian distributions on $\mathbb{R}^d$. Then*

$$D_{\mathrm{KL}}(f_\pi \| f_\eta) \le (10240K^4 + 652) H^2(f_\pi, f_\eta) \log \frac{4}{H^2(f_\pi, f_\eta)}.$$

Definition 3.6 (Covering Number and Local Covering Number)**.** *Let $\mathcal{P}$ be a set of distributions over some measurable space $\mathcal{X}$. The Hellinger covering number of $\mathcal{P}$ is*

$$\mathcal{N}_H(\mathcal{P}, \varepsilon) \triangleq \min\left\{N : \exists Q_1, \cdots, Q_n \in \Delta(\mathcal{X}), \ \sup_{P\in\mathcal{P}} \inf_{1\le i\le N} H(P, Q_i) \le \varepsilon\right\},$$

where $\Delta(\mathcal{X})$ denotes the collection of all probability distributions on $\mathcal{X}$. The local Hellinger covering number of $\mathcal{P}$ is

$$\mathcal{N}_{loc,H}(\mathcal{P}, \varepsilon) \triangleq \sup_{P\in\mathcal{P}, \eta\ge\varepsilon} \mathcal{N}_H(B_H(P, \eta) \cap \mathcal{P}, \eta/2),$$

where $B_H(P, \eta)$ is the Hellinger ball of radius η centered at P.

Definition 3.7 (Proper and Improper Density Estimation Minimax Risk)**.** *For a given class $\mathcal{P}$ of distributions over $\mathcal{X}$, we define the improper minimax risk $R_{H^2,n}, R_{KL,n}$ and the proper minimax risk $\tilde{R}_{KL,n}$ with sample size n as follows: for $d \in \{H^2, D_{\mathrm{KL}}\}$, we define*

$$R_{d,n}(\mathcal{P}) \triangleq \inf_{\hat{f}_n} \sup_{f\in\mathcal{P}} \mathbb{E}_f\left[d(f, \hat{f}_n)\right],$$

and also

$$\tilde{R}_{KL,n}(\mathcal{P}) \triangleq \inf_{\hat{f}_n\in\mathcal{P}} \sup_{f\in\mathcal{P}} \mathbb{E}_f\left[D_{\mathrm{KL}}(f\|\hat{f}_n)\right],$$

where $\hat{f}_n(\cdot) = \hat{f}_n(\cdot; X_1, \ldots, X_n)$ is a density estimator based on $X_1, \ldots, X_n$ drawn iid from P.

Definition 3.8 (Sequential Density Estimation Minimax Risk)**.** *For a given class $\mathcal{P}$ of distributions over $\mathcal{X}$, the sequential minimax risks $C_{H^2,n}$ and $C_{KL,n}$ are defined as: for $d \in \{H^2, D_{\mathrm{KL}}\}$,*

$$C_{d,n}(\mathcal{P}) = \inf_{\hat{f}_1,\cdots,\hat{f}_n} \sup_{f\in\mathcal{P}} \sum_{t=1}^{n} \mathbb{E}[d(f, \hat{f}_t(\cdot|X_1, \cdots, X_{t-1}))]$$

where $\hat{f}_t : \mathcal{X}^{t-1} \to \Delta(\mathcal{X})$ denotes the density estimator at time t based on observations $X_1, \cdots, X_{t-1}$.

Corollary 3.9. *Let $\mathcal{P}_{com}(M)$ and $\mathcal{P}_{sub}(K)$ denote the collection of d-dimensional Gaussian mixtures where the mixing distribution is supported on $B_2(M)$ and K-subgaussian, respectively. Then for any compact (under Hellinger) subset $\mathcal{P}$ where $\mathcal{P} \subset \mathcal{P}_{com}(M)$ or $\mathcal{P} \subset \mathcal{P}_{sub}(K)$ with $K < 1$, we have the following characterization on the proper or improper minimax risk:*

$$R_{H^2,n}(\mathcal{P}) \asymp R_{KL,n}(\mathcal{P}) \asymp \tilde{R}_{KL,n}(\mathcal{P}) \asymp \inf_{\varepsilon>0} \varepsilon^2 + \frac{1}{n}\log \mathcal{N}_{loc,H}(\mathcal{P}, \varepsilon),$$

and also for sequential minimax risk:

$$C_{H^2,n}(\mathcal{P}) \asymp C_{KL,n}(\mathcal{P}) \asymp \inf_{\varepsilon>0} n\varepsilon^2 + \log \mathcal{N}_H(\mathcal{P}, \varepsilon).$$

Here $\asymp$ hides constants that may depend on M, K, or d but not on n.

3.3 Proof of Theorem 3.1 in one dimension

Theorem 3.10 (One-dimensional version of Theorem 3.1)**.** *Let π and η be supported on $B_2(M)$ in $\mathbb{R}$ where $M \geq 2$. Then*

$$D_{\mathrm{KL}}(f_\pi \| f_\eta) \leq 1563 M^2 H^2(f_\pi, f_\eta).$$

For simplicity we abbreviate $f_\pi(\cdot)$ and $f_\eta(\cdot)$ as $p(\cdot), q(\cdot)$. Then we have

$$D_{\mathrm{KL}}(p\|q) = \int_{-\infty}^{\infty} q(x) \cdot \frac{p(x)}{q(x)} \log \frac{p(x)}{q(x)} dx, \quad H^2(p,q) = \int_{-\infty}^{\infty} q(x) \cdot \left(\sqrt{\frac{p(x)}{q(x)}} - 1 \right)^2 dx.$$

Lemma 3.11. *Let $p = \pi * \mathcal{N}(0,1)$ where $\mathtt{supp}(\pi) \subset [-M, M]$. Then we have $\forall y \geq r \geq x \geq M$,*

$$p(y) \exp\left(\frac{(y-M)^2 - (r-M)^2}{2} \right) \leq p(r) \leq p(x).$$

Lemma 3.12. *Let $p = \pi * \mathcal{N}(0,1)$ where $\mathtt{supp}(\pi) \subset [-M, M]$. Then*

$$|\nabla \log p(r)| \leq 3|r| + 4M, \quad \forall r \in \mathbb{R}.$$

Lemma 3.13. *For every $0 \leq t \leq \exp(8M^2)$ with $M \geq 1$, we have*

$$t \log t - t + 1 \leq 9M^2 \left(\sqrt{t} - 1 \right)^2.$$

Proof of Theorem 3.10. It is easy to see that for every $x \in [-2M, 2M]$, we have

$$\frac{1}{\sqrt{2\pi}} \exp\left(-8M^2\right) \leq p(x), q(x) \leq \frac{1}{\sqrt{2\pi}}.$$

Let r be the smallest positive number (possibly infinite) such that $\log \frac{p(r)}{q(r)} \geq 8M^2$, then we have $r \geq 2M$. Without loss of generality we assume $r < \infty$. Since $\log \frac{p(\cdot)}{q(\cdot)}$ is a continuous function, we have $\log \frac{p(r)}{q(r)} = 8M^2$. According to Lemma 3.11, we have for every $x \geq r$,

$$p(x) \leq p(r) \exp\left(-\frac{(x-M)^2 - (r-M)^2}{2} \right), \quad q(x) \leq q(r) \exp\left(-\frac{(x-M)^2 - (r-M)^2}{2} \right)$$

and according to Lemma 3.12 we have

$$\left| \log \frac{p(x)}{q(x)} \right| \leq \left| \log \frac{p(r)}{q(r)} \right| + \int_r^x (3|t| + 4M) dt = 8M^2 + (x-r)(3x + 3r + 8M), \quad \forall x \geq r \geq 0.$$

Therefore, we obtain that

$$\int_r^\infty p(x) \log \frac{p(x)}{q(x)} - p(x) + q(x) dx \leq \int_r^\infty p(x) \log \frac{p(x)}{q(x)} + q(x) dx$$

$$
\begin{aligned}
&\le p(r)\exp\left(\frac{(r-M)^2}{2}\right)\int_r^\infty (8M^2+(x-r)(3x+3r+8M))\exp\left(-\frac{(x-M)^2}{2}\right)dx\\
&\quad + q(r)\exp\left(\frac{(r-M)^2}{2}\right)\int_r^\infty \exp\left(-\frac{(x-M)^2}{2}\right)dx\\
&\le p(r)\cdot\left(\frac{6}{(r-M)^3}+\frac{6r+8M}{(r-M)^2}+\frac{8M^2}{r-M}\right)+\frac{q(r)}{r-M} \le \frac{36M^2}{r-M}p(r)+\frac{q(r)}{r-M}\\
&\le \frac{37M^2}{r-M}p(r),
\end{aligned}
$$

where the last inequality uses the fact $M \ge 1$ and $\log\frac{p(r)}{q(r)} = 8M^2 \ge 0$ hence $p(r) \ge q(r)$.

Moreover, according to Lemma 3.12 we also notice that for $0 \le x \le r$ we have $\left|\nabla \log\frac{p(x)}{q(x)}\right| \le 6x+8M \le 6r+8M$. Hence noticing that $\log\frac{p(r)}{q(r)} = 8M^2$ and also $r \ge 2M$, we have for every $r - \frac{M^2}{r+M} \le x \le r$,

$$
\log\frac{p(x)}{q(x)} \ge 8M^2 - \frac{M^2}{r+M}\cdot(6r+8M) \ge M^2
$$

and also $p(x) \ge p(r)$ according to Lemma 3.11. Therefore, noticing $M \ge 1$, we have

$$
H^2(p,q) \ge \int_{r-\frac{M^2}{r+M}}^r p(x)\cdot\left(\sqrt{\frac{q(x)}{p(x)}}-1\right)^2 dx \ge \frac{p(r)M^2}{(r+M)}\cdot\left(1-\frac{1}{e^{M^2/2}}\right)^2 \ge \frac{p(r)M^2}{7(r+M)}.
$$

Since $r \ge 2M$, we obtain that

$$
\int_r^\infty p(x)\log\frac{p(x)}{q(x)} - p(x) + q(x)dx \le 777H^2(p,q).
$$

Similarly, if we let s to be the largest negative number (possibly negative infinite) such that $\log\frac{p(s)}{q(s)} \ge 8M^2$, then we will also have

$$
\int_{-\infty}^s p(x)\log\frac{p(x)}{q(x)} - p(x) + q(x)dx \le 777H^2(p,q).
$$

Next we consider those $s \le x \le r$. For those x we have $\log\frac{p(x)}{q(x)} \le 8M^2$. Hence according to Lemma 3.13, we have

$$
\frac{p(x)}{q(x)}\log\frac{p(x)}{q(x)} - \frac{p(x)}{q(x)} + 1 \le 9M^2\left(\sqrt{\frac{p(x)}{q(x)}}-1\right)^2.
$$

Therefore,

$$
\begin{aligned}
&\int_s^r p(x)\log\frac{p(x)}{q(x)} - p(x) + q(x)dx = \int_s^r q(x)\cdot\left(\frac{p(x)}{q(x)}\log\frac{p(x)}{q(x)} - \frac{p(x)}{q(x)} + 1\right)dx\\
&\le 9M^2\int_s^r q(x)\cdot\left(\sqrt{\frac{p(x)}{q(x)}}-1\right)^2 dx \le 9M^2H^2(p,q).
\end{aligned}
$$

Overall, we have shown that

$$D_{\mathrm{KL}}(p\|q) \leq (777 + 777 + 9M^2)H^2(p,q) \leq 1563M^2H^2(p,q),$$

which finishes the proof of Theorem 3.10. □

Chapter 4

Testing and Estimation in Orthosymmetric Gaussian Sequence Model

This chapter is based on the paper "Testing and estimation in orthosymmetric Gaussian sequence model" by Zeyu Jia and Yury Polyanskiy.

4.1 Introduction

Two flagship problems in non-parametric statistics are establishing minimax rates (or, equivalently, minimax sample complexities) of density estimation and testing. In the case of density estimation, statistician is given an apriori fixed large class Γ of probability distributions and n iid samples $X_1, \dots, X_n$ from $P \in \Gamma$ and is tasked with producing an estimate $\hat{P}$ of the distribution such that $\mathbb{E}[d(P, \hat{P})] \le \varepsilon$, where d is a distance and ε is the target accuracy. The minimal number n of samples required for the worst-case choice of P is the sample complexity $n_{\text{est}}(\Gamma, \varepsilon)$. For example, if Γ consists of all β-smooth densities on a compact set in $\mathbb{R}^d$ and if accuracy metric is $d = \|\cdot\|_2$ (the ℓ_2 distance), then the classical work of Ibragimov and Khasminskii [2] showed that $n_{\text{est}}(\Gamma, \varepsilon) \asymp \varepsilon^{-(2\beta+d)/\beta}$ upto universal constants, cf. [12].

In the case of testing (or goodness-of-fit), statistician is given an apriori fixed large class Γ of probability distributions, a member $P_0 \in \Gamma$ and n iid samples $X_1, \dots, X_n$ from some $P \in \Gamma$, and is tasked with testing hypothesis $P = P_0$ against $d(P, P_0) > \varepsilon$ with a fixed small probability of error under either of the alternatives. The minimal number of samples needed for accomplishing this task (worst case over P and P_0) is the complexity of testing $n_{\text{gof}}(\Gamma, \varepsilon)$. This way of looking at non-parametric testing, as well as many foundational results, was proposed by Ingster [13, 14, 60]. In particular, for the Lipschitz class and D_{TV} metric his result shows that $n_{\text{gof}}(\Gamma, \varepsilon) \asymp \varepsilon^{-(d/2+2)}$. The fact that testing can be done with significantly fewer samples was both surprising and impactful for the development of theoretical statistics in 20th century.

Despite the similarities of two problems, our level of understanding of n_{est} and n_{gof} is dramatically different. Specifically, the work of LeCam [61], Birge [9, 10], Yatracos [55], Yang and Barron [56] established a direct characterization of n_{est} in terms of metric entropy

of the class Γ, thus reducing the problem to that of approximation theory. At the same time, while several powerful tools [62] were developed for bounding n_{gof}, its evaluation for each class Γ is largely still ad hoc. Thus, our longer term goal is to obtain metric entropy type characterization of n_{gof}. This work is a step in this direction: by establishing general inequalities relating n_{gof} and n_{est}, implicitly we also obtain entropic bounds on n_{gof}.

The origin of these kind of relations is the work [15], which studied yet another statistical problem of *likelihood-free hypothesis testing* (LFHT), see (4.10) below, which "interpolates" between estimation and testing. By inspecting the minimax region for LFHT at two end-points [15] observed that for rather different non-parametric types of models (smooth densities, discrete distributions, Gaussian sequence model over ellipsoids) one has a general relationship

$$\frac{n_{\text{est}}(\Gamma,\varepsilon)}{\varepsilon^2} \asymp n^2_{\text{gof}}(\Gamma,\varepsilon)\,. \tag{4.1}$$

This relation can informally be "derived" as follows. Up to precision ε the non-parametric model Γ can be thought of as a model of finite dimension $D(\Gamma,\varepsilon)$, rigorously defined below as a Kolmogorov dimension. Then, classical *parametric* estimation and testing rates allow us to guess

$$n_{\text{est}}(\Gamma,\varepsilon) \asymp \frac{D(\Gamma,\varepsilon)}{\varepsilon^2}, \qquad n_{\text{gof}}(\Gamma,\varepsilon) \asymp \frac{\sqrt{D(\Gamma,\varepsilon)}}{\varepsilon^2}\,, \tag{4.2}$$

which clearly imply (4.1). In addition to examples from [15], the relationship (4.1) also holds for a special class of Gaussian sequence models with a QCO (compact , convex, quadratically convex and orthosymmetric) constraint set, as follows from [63] and [64] (see Prop. Theorem 4.14 below).

While these considerations support the validity of (4.1), unfortunately neither (4.2) nor (4.1) hold in general. Indeed, a counter-example is implicitly contained in [65], see (4.4) below. Nevertheless, in this work we show that the $\lesssim$ direction of (4.1) holds under the mere assumption of orthosymmetry, see (4.3).

This paper is written solely in the context of *Gaussian sequence model*, for which Γ consists of infinite-dimensional Gaussian densities $\mathcal{N}(\theta, I_\infty)$ with mean θ constrained to belong to a compact convex set in ℓ_2. We will abuse notation and simply write $\theta \in \Gamma$ identifying densities with their means. Such models are both simplest to study and are universal: any non-parametric class Γ of densities can be shown to be Le Cam equivalent (in the limit of $n \to \infty$) to a certain Gaussian sequence model, cf. [66, 67].

4.1.1 Our Contribution

Our contributions are two-fold. First, we derive a collection of results bounding n_{gof} in terms of n_{est}. Second, we also completely characterize LFHT region of general ℓ_p-bodies.

More specifically, for any convex, compact and orthosymmetric set Γ (see definitions below), we show in Theorem 4.8 that the minimax sample complexity $n_{\text{gof}}(\Gamma,\varepsilon)$ of goodness-of-fit testing and the minimax sample complexity $n_{\text{est}}(\Gamma,\varepsilon)$ of density estimation satisfy

$$n_{\text{est}}(\Gamma,\varepsilon) \lesssim \frac{n^2_{\text{gof}}(\Gamma,\varepsilon)}{\varepsilon^2} \cdot \text{polylog}(D(\Gamma,\varepsilon))\,, \tag{4.3}$$

where $D(\Gamma, \varepsilon)$ is the Kolmogorov dimension of Γ (see also Prop. Theorem C.2). Hence, under rather general conditions half of the relationship (4.1) holds up to polylog factors. The unusual aspect of our proof is that a lower bound on testing is shown by extracting a hard to test mixture distribution from the analysis of a soft-thresholding estimator.

As we mentioned above, if in addition to orthosymmetry one also assumes quadratic-convexity of Γ (i.e. if Γ is QCO), then results of [63] on estimation and [64] on testing together imply validity of (4.1) for such sets (see Prop. Theorem 4.14), thus showing that our lower bound is generally tight. We recall that ℓ_p-bodies [63, 65] with $p \geq 2$ are QCO sets.

Are there any counter-examples to (4.1)? The answer is positive as in fact already implicitly shown in [65]. Specifically, let us define the following

$$\Gamma = \left\{ \boldsymbol{\theta} = (\theta_{1:\infty}) : \sum_{i \geq 1} i \cdot |\theta_i| \leq 1 \right\}, \tag{4.4}$$

which is an example of a more general class of an ℓ_p-body with $p = 1$. For this specific class, we find out in Theorem 4.5 and Theorem 4.6 that

$$n_{\mathrm{gof}}(\Gamma, \varepsilon) = \tilde{\Theta}\left(\varepsilon^{-\frac{12}{5}}\right) \quad \text{and} \quad n_{\mathrm{est}}(\Gamma, \varepsilon) = \tilde{\Theta}\left(\varepsilon^{-\frac{8}{3}}\right),$$

thus clearly violating (4.1) and showing that one can indeed have $n_{\mathrm{est}} \ll n_{\mathrm{gof}}^2/\varepsilon^2$.

Our second contribution is in establishing minimax rates (regions) for the LFHT problem defined in (4.10). Recall that in [15], it was shown that the optimal testing region of LFHT for Gaussian sequence model over an ellipsoid Γ satisfies

$$\left\{ m \geq \frac{1}{\varepsilon^2}, \quad n \gtrsim n_{\mathrm{gof}}(\Gamma, \varepsilon), \quad \text{and} \quad mn \gtrsim n_{\mathrm{gof}}^2(\Gamma, \varepsilon) \right\}. \tag{4.5}$$

It is natural to ask whether this relation is in some sense universal. In this work, we show that for Γ, which are orthosymmetric, convex and quadratically convex (e.g. ℓ_p-bodies with $p \geq 2$), the LFHT region remains the same. In particular, this is true for ℓ_p bodies with $p \geq 2$. For $p < 2$, the general form of the LFHT region is given in terms of an "effective dimension" $d(\Gamma, n, \varepsilon)$ depending on n. Specifically, we show that the testing region is given by

$$\left\{ m \geq \frac{1}{\varepsilon^2}, \quad n \gtrsim \frac{\sqrt{d(\Gamma, n, \varepsilon)}}{\varepsilon^2}, \quad \text{and} \quad mn \gtrsim \frac{d(\Gamma, n, \varepsilon)}{\varepsilon^4} \right\}.$$

For example, for the set eq. (4.4), we have $d(\Gamma, n, \varepsilon) \asymp \frac{1}{\sqrt{n} \cdot \varepsilon^2}$ and the LFHT region is given by

$$\left\{ m \geq \varepsilon^{-2}, n \gtrsim \varepsilon^{-\frac{12}{5}}, \quad \text{and} \quad m \cdot n^{\frac{3}{2}} \gtrsim \varepsilon^{-6} \right\}.$$

In particular, this shows that the "regular LFHT" region, where the boundary is defined in terms of the product mn as in (4.5), is specific to ℓ_p-bodies with $p \geq 2$, while for $p < 2$ the region is rather different.

4.1.2 Related Works

We review some related literatures in this section.

Non-parametric Density Estimation: As we already discussed in the introduction, there has been a long line of work studying the density estimation. For fairly general non-parametric classes and distances between distributions Le Cam [61] (also in [68]) and Birge [9, 10] characterized the minimax rate in terms of the (local) Hellinger metric entropy. Other general estimators were proposed by Yang-Barron [56], Yatracos [55] and others. In the context of estimating smooth densities, the study of kernel density estimators is a rich subject [12, Chapter 1], as is the method of wavelet-based techniques [69]. In the context of Gaussian sequence models, density estimation corresponds to parameter (mean) estimation, with state of the art beautifully summarized in [70].

Gaussian sequence model: In the context of Gaussian sequence model density estimation is equivalent to parameter (θ) estimation. This question received significant attention, see [71, Chapter 4]. We specifically mention pioneering work of Pinsker [72], who demonstrated optimality of linear estimators for the case of certain ellipsoids, and [63], who significantly extended this idea by showing optimality (upto a universal factor) of projection estimators for all quadratically convex sets – a notion which also found application in stochastic optimization [73].

Goodness-of-Fit Testing: The sample complexity of goodness-of-fit testing has been pioneered by the already mentioned works of Ingster, whose book [74] surveys many of the classical developments. Subsequently, [75] obtained tight goodness-of-fit testing rate for Besov bodies $B_{s,p,q}(R)$ where $p \in (0, 2)$.

For Gaussian sequence model, we mention results of [76] and, very relevant for our work, a comprehensive paper of Baraud [65]. Specifically, for ℓ_p bodies $\{\theta_{1:D} : \sum_{i=1}^{D} |\theta_i|^p/a_i^p \le 1\}$ where $a_1 \ge a_2 \ge \cdots \ge a_D$, they proposed the following dimension

$$\rho(n) = \sup_{d \in [D]} [(\sqrt{d}/n^2 \wedge a_d^2 \lceil \sqrt{d} \rceil^{1-2/p})],$$

and they show that the minimax sample complexity of goodness-of-fit testing is the smallest n such that $\rho(n) \le \varepsilon$. More recently, [77] refined sample complexity to make it depend on a specific (rather than worst-case) choice of the mean in the null-hypothesis. In the special case of QCO sets Γ Neykov [64] characterizes (within a constant factor) the goodness-of-fit testing sample complexity in terms ofcritical radius, which in turn is derived from Kolmogorov widths. The case of Γ being a d-dimensional convex cone was studied in [78], in particular demonstrating that in the case of "ice-cream cones" one can get $n_{\text{gof}} \asymp \frac{1}{\varepsilon^2}$ but $n_{\text{est}} \asymp \frac{d}{\varepsilon^2}$, due to null-case being at the apex of the cone.

A large amount of work has also been done on the topic in computer science literature under different names of *identity testing* or *uniformity testing*, see [79–85]. A recent lower bound for robust testing was proposed in [86]. [87, 88] proposed minimax optimal testing scheme for cases with unknown variances. An excellent monograph [89] surveys this line of work.

Likelihood Free Hypothesis Testing: The form of likelihood free hypothesis testing was firstly introduced in [90, 91]. They studied the problem in fixed finite alphabet. [92] showed that the testing scheme introduced in [90] is second-order optimal. This problem is extended

into sequential and distributional setting in [93–96]. In this work, we will focus on the setting introduced in [15].

4.1.3 Organization

In section 4.2 we review the basic concepts of goodness-of-fit testing, density estimation and likelihood-free hypothesis testing. In section 4.3, we study the relationship between goodness-of-fit testing and density estimation. In section 4.4 we study the feasible region of likelihood-free hypothesis testing. Specifically, in section 4.4.1 we build up relationship between the LFHT feasible region and density estimation, and in section 4.4.2 we calculate the LFHT feasible region for ℓ_p bodies with $p \leq 2$.

4.1.4 Notations

For $\boldsymbol{\theta} \in \mathbb{R}^d$, we use $\mathcal{N}(\boldsymbol{\theta}, I_d)$ to denote the multivariate Gaussian distribution with mean $\boldsymbol{\theta}$ and variance matrix to be the identity matrix. We use $\mathbf{0}$ to denote the all-zero vector. We use $a_n = \mathcal{O}(b_n)$ or $a_n \lesssim b_n$ ($a_n = \Omega(b_n)$ or $a_n \gtrsim b_n$) to denote the inequality $a_n \leq c \cdot b_n$ ($a_n \geq c \cdot b_n$) for all n for some fixed positive constant c.

4.2 Preliminaries and notation

We review some basic concepts of the Gaussian sequence model [74], goodness-of-fit testing [74], density estimation [70] and likelihood-free hypothesis testing [15] in this section.

Gaussian Sequence Model. We focus on unit-variance multivariate Gaussian location model, which is specified by an integer $D \in [1, \infty]$ and a subset $\Gamma \subseteq \mathbb{R}^D$, so that the model consists of all distributions $\mathcal{P}(\Gamma) \triangleq \{\mathcal{N}(\boldsymbol{\theta}, I_D) : \boldsymbol{\theta} \in \Gamma\}$. When $D = \infty$ we also refer to this model as Gaussian sequence model. In the following, we often use subset Γ to denote the model $\mathcal{P}(\Gamma)$ itself.

Goodness of Fit Testing. Given an integer $D \in [1, \infty]$, we conduct the goodness of fit testing in $\mathbb{R}^D$. In this task, the statistician is given an class of parameters $\Gamma \subseteq \mathbb{R}^D$, and has the knowledge that the true parameter of the model $\boldsymbol{\theta} \in \Gamma$. However, $\boldsymbol{\theta}$ is unknown and the statistician can only obtain information of $\boldsymbol{\theta}$ through n samples $\mathbf{X} = (\mathbf{X}_{1:n}) \sim \mathbb{P}_{\mathtt{X}}^{\otimes n}$ where $\mathbb{P}_{\mathtt{X}} = \mathcal{N}(\boldsymbol{\theta}, I_D)$. The statistician conducts the following hypothesis testing problem

$$H_0 : \boldsymbol{\theta} = \mathbf{0} \quad \text{versus} \quad H_1 : \|\boldsymbol{\theta}\|_2 \geq \varepsilon, \tag{4.6}$$

i.e., based on the samples in $\mathbf{X}$, the statistician makes choice $\psi(\mathbf{X}) \in \{0, 1\}$, where $\psi(\mathbf{X}) = 0$ denotes the statistician accepts hypothesis H_0 and $\psi(\mathbf{X}) = 1$ denotes the statistician rejects H_0. We focus on the smallest number of n such that

$$\max_{i \in \{0,1\}} \sup_{P \in H_i} \mathbf{P}(\psi(\mathbf{X}) \neq i) \leq \frac{1}{4}, \tag{4.7}$$

where $\sup_{P \in H_0}$ denotes the case $\boldsymbol{\theta} = \mathbf{0}$, and $\sup_{P \in H_1}$ denotes the supreme over $\boldsymbol{\theta}$ with $\|\boldsymbol{\theta}\|_2 \geq \varepsilon$. We let $n_{\text{gof}}(\Gamma, \varepsilon)$ to denote the smallest number of samples such that eq. (4.7) holds.

Density Estimation. Given an integer $D \in [1, \infty]$, we conduct the density estimation task in $\mathbb{R}^D$. In this task, the statistician is given an class of parameters $\Gamma \subseteq \mathbb{R}^D$, and has the knowledge that the parameter of the groundtruth distribution $\boldsymbol{\theta} \in \Gamma$. However, $\boldsymbol{\theta}$ is unknown and the statistician can only obtain information of $\boldsymbol{\theta}$ through n samples $\mathbf{X} = (\mathbf{X}_{1:n}) \sim \mathbb{P}_{\mathtt{X}}^{\otimes n}$ where $\mathbb{P}_{\mathtt{X}} = \mathcal{N}(\boldsymbol{\theta}, I_D)$. Based on the samples in $\mathbf{X}$, the statistician proposes an estimator $\widehat{\boldsymbol{\theta}}(\mathbf{X}) \in \mathbb{R}^D$. We focus on the smallest number of samples n such that there exists an estimator which achieves expected ℓ_2 estimation error no more than ε^2 for any distribution in the distribution class, i.e.

$$\inf_{\widehat{\boldsymbol{\theta}}} \sup_{\boldsymbol{\theta} \in \Gamma} \mathbb{E}\left[\left\|\widehat{\boldsymbol{\theta}}(\mathbf{X}) - \boldsymbol{\theta}\right\|_2^2\right] \leq \varepsilon^2. \tag{4.8}$$

We use $n_{\text{est}}(\Gamma, \varepsilon)$ to denote the smallest number of samples such that eq. (4.8) holds.

Likelihood Free Hypothesis Testing. The task of likelihood free hypothesis testing was first introduced in [15]. The statistician conducts the testing in $\mathbb{R}^D$, and the statistician is given the class Γ of the Gaussian sequence model. Suppose there are two artificial densities $\mathbb{P}_{\mathtt{X}} = \mathcal{N}(\boldsymbol{\theta}^1, I_D)$, $\mathbb{P}_{\mathtt{Y}} = (\boldsymbol{\theta}^2, I_D)$ and a true density $\mathbb{P}_{\mathtt{Z}} = \mathcal{N}(\boldsymbol{\theta}, I_D)$ where $\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \boldsymbol{\theta} \in \Gamma$ are unknown to the statistician. After collecting n samples from $\mathbb{P}_{\mathtt{X}}$ and $\mathbb{P}_{\mathtt{Y}}$ each to form artificial datasets $\mathbf{X} \sim \mathbb{P}_{\mathtt{X}}^{\otimes n}$ and $\mathbf{Y} \sim \mathbb{P}_{\mathtt{Y}}^{\otimes n}$, and collecting m samples from $\mathbb{P}_{\mathtt{Z}}$ to form the ground truth dataset $\mathbf{Z} \sim \mathbb{P}_{\mathtt{Z}}^{\otimes m}$, the statistician conducts the following hypothesis testing problem:

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}^1 \quad \text{versus} \quad H_1 : \boldsymbol{\theta} = \boldsymbol{\theta}^2. \tag{4.9}$$

To characterize the minimax sample complexity of the above hypothesis testing problem, we suppose $\boldsymbol{\theta}^1$ and $\boldsymbol{\theta}^2$ satisfies $\|\boldsymbol{\theta}^1 - \boldsymbol{\theta}^2\|_2 \geq \varepsilon$ for some $\varepsilon > 0$, and this piece of information is given to the statistician. After receiving data $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \in (\mathbb{R}^D)^n \times (\mathbb{R}^D)^n \times (\mathbb{R}^D)^m$, the statistician makes choice $\psi(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \in \{0, 1\}$, where $\psi(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = 0$ denotes the statistician accepts hypothesis H_0 and $\psi(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = 1$ denotes the statistician rejects H_0. We focus on the conditions of (m, n) such that

$$\max_{i \in \{0,1\}} \sup_{P \in H_i} \mathbb{P}(\psi(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \neq i) \leq \frac{1}{4}, \tag{4.10}$$

where $\sup_{P \in H_0}$ denotes the supreme over $\boldsymbol{\theta}^1$ and $\boldsymbol{\theta}^2$ with $\|\boldsymbol{\theta}^1 - \boldsymbol{\theta}^2\|_2 \geq \varepsilon$ and $\boldsymbol{\theta} = \boldsymbol{\theta}^1$, and $\sup_{P \in H_1}$ denotes the supreme over $\boldsymbol{\theta}^1$ and $\boldsymbol{\theta}^2$ with $\|\boldsymbol{\theta}^1 - \boldsymbol{\theta}^2\|_2 \geq \varepsilon$ and $\boldsymbol{\theta} = \boldsymbol{\theta}^2$.

For any pair (m, n), we say that (m, n) lies in the feasible region of the likelihood-free hypothesis test if and only if there exists some test scheme ψ such that eq. (4.10) holds.

ℓ_p Bodies. Given an integer $D \in [1, \infty]$, the ℓ_p bodies in $\mathbb{R}^D$ is characterized by a non-increasing nonnegative sequence $a_1 \geq a_2 \geq \cdots \geq a_D \geq 0$. The ℓ_p body (also appears in [65]) characterized by $a_{1:D}$ is given by

$$\Gamma = \left\{ \theta = (\theta_1, \theta_2, \cdots, \theta_D) \;\middle|\; \sum_{t=1}^{D} \frac{|\theta_t|^p}{a_t^p} \leq 1 \right\} \subseteq \mathbb{R}^D. \tag{4.11}$$

When $D = \infty$, in order to guarantee the compactness of this infinite-dimensional set, we only consider the ℓ_p bodies characterized by sequence a_t goes to zero, i.e. $\lim_{t \to \infty} a_t = 0$.

Orthosymmetric Sets. We recall the definition of orthosymmetric sets, first introduced in [63]: Given an integer $D \in [1, \infty]$, we say $\Gamma \subseteq \mathbb{R}^D$ is an orthosymmetric set if for any element $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_D) \in \Gamma$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_D) \in \{-1, 1\}^D$, we have

$$\boldsymbol{\theta}_{\boldsymbol{\varepsilon}} = (\varepsilon_1\theta_1, \varepsilon_2\theta_2, \cdots, \varepsilon_D\theta_D) \in \Gamma.$$

We notice that many sets satisfy the orthosymmetric property. It is easy to verify that the ℓ_p bodies introduced in eq. (4.11) are orthosymmetric. Additionally, if there exists a function $f : (\mathbb{R}_+ \cup \{0\})^D \to \mathbb{R}$ such that Γ can be represented as

$$\Gamma = \{\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_D) : f(|\theta_1|, |\theta_2|, \cdots, |\theta_D|) \geq 0\}, \tag{4.12}$$

then Γ is an orthosymmetric set. It is easy to see that all ℓ_p bodies can be written in the above form for some function f.

Kolmogorov Dimension and Coordinate-wise Kolmogorov Dimension. Given an integer $D \in [1, \infty]$, the Kolmogorov dimension of a set $\Gamma \subseteq \mathbb{R}^D$ is a notion to measure the minimal dimension of an affine space so that the distance between any point in Γ and the affine space is bounded by some tolerance. This notion is given formally as in the following definition.

Definition 4.1 (Kolmogorov Dimension)**.** *Suppose $\Gamma \subseteq \mathbb{R}^D$. For any $\varepsilon > 0$, we define the Kolmogorov dimension $D(\Gamma, \varepsilon)$ of Γ at scale ε to be the largest integer $d \leq D$ such that*

$$\inf_{\Pi_d} \sup_{\boldsymbol{\theta} \in \Gamma} \|\boldsymbol{\theta} - \Pi_d \boldsymbol{\theta}\|_2 \leq \varepsilon,$$

where Π_d denotes a d-dimensional linear projection, and the infimum is over all possible d-dimensional linear projections.

Notice that in the above definition, the projection operator can be chosen to be any linear projections. However, in some cases, we have to restrict ourselves to project only along the coordinates. In this way, we define the following coordinate-wise Kolmogorov dimension.

Definition 4.2 (Coordinate-wise Kolmogorov Dimension)**.** *Suppose $\Gamma \subseteq \mathbb{R}^D$. For any $\varepsilon > 0$, we define the coordinate-wise Kolmogorov dimension $D_{\mathsf{c}}(\Gamma, \varepsilon)$ of Γ at scale ε to be the largest integer $d \leq D$ such that*

$$\inf_{A : |A| = d} \sup_{\boldsymbol{\theta} \in \Gamma} \sum_{i \in \mathbb{Z}_+ \setminus A} (\theta_i)^2 \leq \varepsilon^2, \qquad \text{where } \boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_D),$$

where the infimum is over all size-d subsets A of $\{1, 2, \cdots, D\}$.

It is clear that for any set Γ and $\varepsilon > 0$, $D_{\mathsf{c}}(\Gamma, \varepsilon) \geq D(\Gamma, \varepsilon)$. However, it is possible to have $D_{\mathsf{c}} \gg D$. Fortunately, our bounds depend on D_{c} only logarithmically, and hence the following pair of results can be used to roughly relate D_{c} to D and ε.

Proposition 4.3. *For any orthosymmetric, convex, compact set Γ, we have*

$$D\left(\Gamma, \frac{\varepsilon}{D_{\mathsf{c}}(\Gamma, \varepsilon)}\right) \geq \frac{D_{\mathsf{c}}(\Gamma, \varepsilon)}{2} - 2.$$

Corollary 4.4. *Suppose orthosymmetric, convex, compact set Γ satisfies $D(\Gamma, \varepsilon) \lesssim \varepsilon^{-p}$ for some $0 < p < 1$. Then the coordinate-wise Kolmogorov dimension satisfies*

$$D_{\mathsf{c}}(\Gamma, \varepsilon) \lesssim \varepsilon^{-p/(1-p)}.$$

4.3 Goodness-of-Fit and Density Estimation

We return to the testing-estimation relation (4.1) that was shown in [15] to hold for a variety of models, including Gaussian sequence models with Γ being an ellipsoid. Our first goal (section 4.3.1) is to exhibit a natural Gaussian sequence model with convex Γ, for which the sample complexity of goodness-of-fit testing is $\tilde{\Theta}(\varepsilon^{-12/5})$, while the sample complexity of density estimation is $\tilde{\Theta}(\varepsilon^{-8/3})$, thus showing that (4.1) fails. Next, in section 4.3.2 we show that nevertheless, for orthosymmetric convex Γ a one-sided comparison always holds: $\varepsilon^2 n_{\text{gof}}^2 \gtrsim n_{\text{est}}$. Finally, in appendix C.4 we show that two-sided relationship (4.1) in fact holds for all orthosymmetric, compact, convex and quadratically-convex Γ.

4.3.1 A Counter Example

We consider the set defined in eq. (4.4): for $D = \infty$, and

$$\Gamma = \left\{ \boldsymbol{\theta} = (\theta_{1:\infty}) : \quad \sum_{i \geq 1} i \cdot |\theta_i| \leq 1 \right\} \subseteq \mathbb{R}^D.$$

Then we have the following characterization in the goodness-of-fit testing sample complexity and density estimation sample complexity.

Proposition 4.5. *For set Γ defined in eq. (4.4), the minimax density estimation sample complexity $n_{\text{est}}(\Gamma, \varepsilon)$ satisfies*

$$n_{\text{est}}(\Gamma, \varepsilon) = \tilde{\Theta}\left(\varepsilon^{-8/3}\right).$$

Proposition 4.6. *For set Γ defined in eq. (4.4), the minimax goodness-of-fit testing sample complexity $n_{\text{gof}}(\Gamma, \varepsilon)$ satisfies*

$$n_{\text{gof}}(\Gamma, \varepsilon) = \tilde{\Theta}\left(\varepsilon^{-12/5}\right).$$

The proof of Theorem 4.5 and Theorem 4.6 are deferred to appendix C.1. From Theorem 4.5 and Theorem 4.6, we learn that for this specific set Γ, eq. (4.1) cannot hold when ε goes to zero, even up to log factors this conjecture still fails.

4.3.2 One-side inequality for Orthosymmetric Convex Sets

Even though there exists a set Γ such that eq. (4.1) fails, we show that a one-side inequality between the minimax sample complexity of goodness-of-fit testing and the minimax sample complexity of density estimation holds up to log factors, if we assume orthosymmetric property, compactness and convexity of the set Γ. This result is summarized in the following theorem.

Theorem 4.7. *For positive integer D, suppose $\Gamma \in \mathbb{R}^D$ is an orthosymmetric, compact and convex set (orthosymmetric sets defined in section 4.2). Then we have*

$$n_{\text{gof}}\left(\Gamma, \frac{\varepsilon}{\sqrt{2}}\right)^2 \geq \frac{1}{64 \log(2D)} \cdot \frac{n_{\text{est}}(\Gamma, \varepsilon)}{\varepsilon^2}.$$

The proof of Theorem 4.7 is deferred to appendix C.2. Theorem 4.7 has the following implication on the infinite-dimensional orthosymmetric convex sets. Then we have the following corollary, which generalize Theorem 4.7 into infinite dimensional sets. The following corollary lower bounds n_{gof} in terms of the coordinate-wise Kolmogorov dimension of the set Γ (defined in Theorem 4.2).

Corollary 4.8. *Suppose the coordinate-wise Kolmogorov dimension of Γ at scale ε to be $D_{\mathsf{c}}(\Gamma, \varepsilon)$. Then if Γ is compact, convex and orthosymmetric, we have for any $\varepsilon > 0$,*

$$n_{\text{gof}}\left(\Gamma, \frac{\varepsilon}{\sqrt{2}}\right)^2 \geq \frac{1}{64\log(2D_{\mathsf{c}}(\Gamma,\varepsilon))} \cdot \frac{n_{\text{est}}(\Gamma, \sqrt{3}\varepsilon)}{\varepsilon^2}.$$

The proof of Theorem 4.8 is deferred to appendix C.2.

Remark 4.9. *It is easy to see that all ℓ_p bodies defined in section 4.2 are orthosymmetric, compact and convex sets. Hence Theorem 4.7 (for finite dimensional sets), and Theorem 4.8 (for infinite dimensional sets) holds for all ℓ_p bodies.*

Remark 4.10. *The notion of coordinate-wise Kolmogorov dimension $D_{\mathsf{c}}(\Gamma, \varepsilon)$ (defined in Theorem 4.2) differs from the traditional Kolmogorov dimension $D(\Gamma, \varepsilon)$ (defined in Theorem 4.1), as the projection spaces in D_{c} are restricted to be parallel or perpendicular to the coordinate axes. It is clear that $D_{\mathsf{c}}(\Gamma, \varepsilon) \geq D(\Gamma, \varepsilon)$. However, in some cases, $D_{\mathsf{c}}(\Gamma, \varepsilon)$ can also be upper-bounded in terms of $D(\Gamma, \varepsilon)$. For further discussion, see appendix C.3.*

Remark 4.11. *We notice that without assuming the orthosymmetric property, Theorem 4.7 can fail. For example, [78] shows that for the d-dimensional circular cone with null-hypothesis at the apex of the cone the goodness-of-fit testing complexity is $n_{\text{gof}} \asymp 1/\varepsilon^2$ (independent to the dimension d), while the estimation complexity is the usual $n_{\text{est}} \asymp d/\varepsilon^2$. Based on this idea, a non-orthosymmetric counter-example to the lower bound in the Theorem can be constructed by taking*

$$\Gamma = \left\{\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots) : \sum_{i=2}^{\infty} \theta_i^2 \leq \theta_1^2, \theta_1 \geq 0\right\} \cap \left\{\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots) : \sum_{i=1}^{\infty} i^2 \cdot \theta_i^2 \leq 1\right\}.$$

Indeed, it can be shown similarly to [78] that $n_{\text{gof}} \asymp \varepsilon^{-2}$, while $n_{\text{est}} \gtrsim \varepsilon^{-3}$.

4.3.3 Testing-Estimation Equivalence for Quadratically Convex Set

In the above, we already verify that the lower bound side of eq. (4.1) holds, if we assume orthosymmetric property, compactness and also convexity of the set Γ. We wonder in what circumstances can the opposite side of inequality eq. (4.1) hold. In the following, we show that if we additionally assume that quadratically convexity of set Γ, then the opposite side of eq. (4.1) also holds. This result, together with Theorem 4.7 (or Theorem 4.8), shows that the equivalence relation eq. (4.1) holds up to log factors, if assuming that the set is orthosymmetric, compact, convex and quadratically convex. Before presenting this result, we first recap the definition of quadratically convex sets (first introduced in [63]).

Definition 4.12 (Quadratically Convex Set)**.** *We say a set $\Gamma \subseteq \mathbb{R}^D$ is quadratically convex, if the following set is convex:*

$$\left\{\boldsymbol{\theta}^2 : \boldsymbol{\theta} \in \Gamma\right\},$$

where $\boldsymbol{\theta}^2 = (\theta_1^2, \cdots, \theta_D^2)$ for $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_D) \in \Gamma$.

Remark 4.13. *The main result in [63] was the proof of optimality (upto a universal constant factor 1.25) of linear estimators (also known as projection estimators) for the minimax estimation rate in the Gaussian location model over an orthosymmetric, compact, convex and quadratically convex class.*

We have the following theorem for quadratically convex sets.

Proposition 4.14 ([63, 64])**.** *There exists universal constants $c, C > 0$ such that for any QCO (compact, convex, quadratically convex and orthosymmetric) set Γ,*

$$n_{\text{gof}}(\Gamma, c\varepsilon)^2 \lesssim \frac{n_{\text{est}}(\Gamma, \varepsilon)}{\varepsilon^2} \lesssim n_{\text{gof}}(\Gamma, C\varepsilon)^2, \qquad \forall \varepsilon > 0.$$

Proof. Seminal work [63] shows that for QCO sets, $D(\Gamma, c\varepsilon)/\varepsilon^2 \lesssim n_{\text{est}}(\Gamma, \varepsilon) \lesssim D(\Gamma, C\varepsilon)/\varepsilon^2$. In [64] it is shown that $n_{\text{gof}}(\Gamma, c\varepsilon) \lesssim \sqrt{D(\Gamma, \varepsilon)}/\varepsilon^2 \lesssim n_{\text{gof}}(\Gamma, C\varepsilon)$ for QCO sets. These results together imply the proposition. □

In appendix C.4 we also include a slightly more general result showing that the second inequality in the Proposition holds under the weaker assumption on Γ (namely, only requiring that projection estimators be order-optimal).

Remark 4.15. *As shown in [63], ℓ_p-bodies with $p \geq 2$ are quadratically convex sets.*

4.4 Likelihood Free Hypothesis Testing

In this section, we study the feasible region of likelihood free hypothesis testing (LFHT) problems, which is introduced in [15]. We already reviewed the basic of LFHT in section 4.2.

Earlier in section 4.3.2 we established a lower bound for the goodness-of-fit testing sample complexity in terms of the density estimation sample complexity (for compact, convex and orthosymmetric sets). Under additional assumption of quadratic convexity (Theorem 4.12), we have shown (section 4.3.3) an upper bound matching the lower bound up to log factors.

4.4.1 LFHT and Density Estimation

In this section we generalize both of these bounds to the setting of LFHT. Theorem 4.16 establishes the lower bound (for compact, convex and orthosymmetric sets), while Theorem 4.17 shows a matching upper bound (under additional assumption of quadratic convexity). As a corollary, this resolves (up to log factors) characterization of the LFHT region of ℓ_p bodies with $p \geq 2$.

Theorem 4.16. *Suppose $\Gamma \subseteq \mathbb{R}^D$ is an orthosymmetric, compact and convex set. Then if a pair of integers (m,n) lies in the LFHT region, i.e. there exists testing scheme $\psi : (\mathbb{R}^D)^n \times (\mathbb{R}^D)^n \times (\mathbb{R}^D)^m$ such that eq. (4.10), then (m,n) satisfies*

$$m \geq \frac{1}{\varepsilon^2}, \qquad n \geq \frac{\sqrt{n_{\text{est}}(\Gamma, \sqrt{2}\varepsilon)}}{8\varepsilon \cdot \sqrt{\log(4D)}}$$
$$\text{and} \quad mn \geq \frac{n_{\text{est}}(\Gamma, \sqrt{2}\varepsilon) - 1}{3072\varepsilon^2 \cdot \log(4D)}.$$

The proof of Theorem 4.16 can be found in appendix C.5.1. Next, we provide an upper bound to the feasible region of LFHT, in terms of the Kolmogorov dimension. We introduce the following testing scheme: For $\varepsilon > 0$, let Π_d to be the d-dimensional linear projection satisfying

$$\sup_{\boldsymbol{\theta} \in \Gamma} \|\boldsymbol{\theta} - \Pi_d \boldsymbol{\theta}\|_2 \leq \frac{\varepsilon}{3}, \tag{4.13}$$

where $d = D(\Gamma, \varepsilon/3)$ denotes the Kolmogorov dimension of set Γ at scale $\varepsilon/3$ (Theorem 4.1). Consider samples $\mathbf{X} = (\mathbf{X}_1, \cdots, \mathbf{X}_n) \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{\mathtt{X}}$, $\mathbf{Y} = (\mathbf{Y}_1, \cdots, \mathbf{Y}_n) \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{\mathtt{Y}}$ and $\mathbf{Z} = (\mathbf{Z}_1, \cdots, \mathbf{Z}_n) \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{\mathtt{Z}}$. Letting

$$\hat{\theta}_{\mathtt{X}} = \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i, \quad \hat{\theta}_{\mathtt{Y}} = \frac{1}{n}\sum_{t=1}^n \mathbf{Y}_i \quad \text{and} \quad \hat{\theta}_{\mathtt{Z}} = \frac{1}{n}\sum_{t=1}^n \mathbf{Z}_i,$$

we define the testing function

$$T_{\mathtt{LF}}(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = \left\|\Pi_d[\hat{\theta}_{\mathtt{X}} - \hat{\theta}_{\mathtt{Z}}]\right\|^2 - \left\|\Pi_d[\hat{\theta}_{\mathtt{Y}} - \hat{\theta}_{\mathtt{Z}}]\right\|^2, \tag{4.14}$$

and consider the testing scheme: $\psi(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = \mathbb{I}[T_{\mathtt{LF}} \geq 0]$. The following theorem indicates that that this testing scheme satisfies eq. (4.10), as long as m, n lower bounded by some functions of the Kolmogorov dimension $D(\Gamma, \varepsilon/3)$.

Theorem 4.17. *If the pair of integers (m,n) satisfies*

$$m \geq \frac{96}{\varepsilon^2}, \quad n \geq \frac{96\sqrt{D(\Gamma, \varepsilon/3)}}{\varepsilon^2} \quad \text{and} \quad mn \geq \frac{768 D(\Gamma, \varepsilon/3)}{\varepsilon^4}, \tag{4.15}$$

the testing scheme $\psi(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = \mathbb{I}\{T_{\mathtt{LF}} \geq 0\}$ with $T_{\mathtt{LF}}$ defined in eq. (4.14) satisfies eq. (4.10).

The proof of Theorem 4.17 is deferred to appendix C.5.2. The above upper bound characterization is with respect to the Kolmogorov dimension. We wonder the relationship between the feasible region of the above testing scheme and the minimax sample complexity of goodness-of-fit testing or density estimation with Γ. With the help of the relations in [63] between the minimax density estimation sample complexity, and the Kolmogorov dimension, we have the following direct corollary.

Corollary 4.18. *Suppose Γ is an uncondintional, compact, convex and quadratically convex set, then there exists a universal positive constant c_0 such that if the pair of integers (m, n) satisfies*

$$m \geq \frac{c_0}{\varepsilon^2}, \quad n \geq \frac{c_0\sqrt{n_{\mathrm{est}}(\Gamma, \varepsilon/9)}}{\varepsilon} \quad \textit{and} \quad mn \geq c_0^2 \cdot n_{\mathrm{est}}(\Gamma, \varepsilon/9)^2,$$

the testing scheme $\psi(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = \mathbb{I}\{T_{\mathtt{LF}} \geq 0\}$ with $T_{\mathtt{LF}}$ defined in eq. (4.14) satisfies eq. (4.10).

This corollary directly follows from Theorem 4.17 and the lower bound of the minimax density estimation sample complexity from the Kolmogorov dimension for sets which are orthosymmetric, compact, convex and quadratically convex. We next apply Theorem 4.16 and Theorem 4.18 to ℓ_p bodies with $p \geq 2$ (recall eq. (4.11)), which are orthosymmetric, compact, convex and quadratically convex from [63]. Hence we have the following characterization.

Corollary 4.19 (LFHT for ℓ_p-body with $p \geq 2$)**.** *For ℓ_p bodies given in eq. (4.11), we define*

$$\tilde{D}(\varepsilon) = \min\left\{ n \in \mathbb{Z}_+ : \forall\, \boldsymbol{\theta} \in \Gamma,\ \sum_{j\in[D]} (\theta_j)^2 \wedge \frac{1}{n} \leq \varepsilon^2 \right\}.$$

Then if there exists a testing scheme ψ such that eq. (4.10) holds, then (m, n) satisfies

$$m \gtrsim \frac{1}{\varepsilon^2}, \quad n \gtrsim \sqrt{\frac{\tilde{D}(\sqrt{2}\varepsilon)}{\varepsilon^2 \log(4D)}} \quad \textit{and} \quad mn \gtrsim \frac{\tilde{D}(\sqrt{2}\varepsilon)}{\varepsilon^2 \log(4D)},$$

and if (m, n) satisfies

$$m \gtrsim \frac{1}{\varepsilon^2}, \quad n \gtrsim \sqrt{\frac{\tilde{D}(\varepsilon/3)}{\varepsilon}} \quad \textit{and} \quad mn \gtrsim \frac{\tilde{D}(\varepsilon/3)}{\varepsilon^2},$$

then there exists a testing scheme ψ such that eq. (4.10) holds.

The proof amounts to applying Theorem 4.16 and Theorem 4.18 after noticing that $n_{\mathrm{est}}(\varepsilon) \asymp \tilde{D}(\varepsilon)$ for all all ℓ_p-bodies, cf. [70, Chapter 4].

4.4.2 Tight characterization of LFHT for ℓ_p Bodies with $p \leq 2$

From the last section, we have both sufficient and necessary conditions of the feasible region of LFHT, in terms of the density estimation rate. These characterization are tight for ℓ_p bodies where $p \geq 2$. However, when $p \leq 2$, the set Γ is no longer quadratically convex, hence Theorem 4.18 fails. In this section, we provide tight characterization for ℓ_p bodies where $p \leq 2$.

Firstly, we recall the form of the testing region in [15]:

$$\left\{ (m, n) : \quad m \gtrsim \frac{1}{\varepsilon^2}, \quad n \gtrsim n_{\mathrm{gof}}(\varepsilon) \quad \text{and} \quad mn \gtrsim n_{\mathrm{gof}}(\varepsilon)^2 \right\},$$

and this work proposed a conjecture that for any set the testing region is always in the above form. While the results in the previous section does not violates this conjecture, we wonder whether this conjecture holds for ℓ_p bodies where $p \leq 2$. Surprisingly, it turns out the testing region for ℓ_p bodies is not in the above form. This can be seen from the following theorem.

Theorem 4.20 (Informal). *For given positive integer n and set of parameters Γ given in eq. (4.11), we define function $d(\Gamma, n, \varepsilon)$ as*

$$d(\Gamma, n, \varepsilon) = \max\left\{d : (a_d)^p n^{\frac{p-2}{2}} \gtrsim \varepsilon^2\right\}.$$

Then the LFHT region for Γ at scale ε is given by

$$\left\{(m, n) : \quad m \gtrsim \frac{1}{\varepsilon^2}, \quad n \gtrsim \frac{\sqrt{d(\Gamma, n, \varepsilon)}}{\varepsilon^2} \quad \text{and} \quad mn \gtrsim \frac{d(\Gamma, n, \varepsilon)}{\varepsilon^4}\right\}. \tag{4.16}$$

In the following, we will first present the results for finite dimensions (in section 4.4.2 and section 4.4.2), and later we will generalize the results to infinite dimensional sets in section 4.4.2. At the end of this section, as an example, we revisit the set Γ defined in eq. (4.4), and we present the closed form of the feasible testing region of this set.

Testing Scheme and Upper Bounds

We first let $\delta = 1/32$ and define

$$d_u(\Gamma, n, \varepsilon) = \max\left\{d : (a_d)^p n^{\frac{p-2}{2}} \gtrsim \frac{\varepsilon^2}{576 \log(4D/\delta)}\right\}. \tag{4.17}$$

Suppose (m, n) satisfies

$$\left\{(m, n) : \quad m \geq \frac{32}{\varepsilon^2}, \quad n \geq \frac{32\sqrt{d_u(\Gamma, n, \varepsilon)}}{\varepsilon^2} \quad \text{and} \quad mn \geq \frac{512 d_u(\Gamma, n, \varepsilon)}{\varepsilon^4}\right\}. \tag{4.18}$$

In this section we will come up with a testing scheme Ψ such that eq. (4.10) is achieved. First of all, for cases where $m > n$, if we let $m_0 = n$, then (m_0, n) is also in the above region, and also $m_0 \leq m$. Hence without loss of generality, we only need to construct a testing scheme for $m \leq n$.

Next, when $mn \geq 1024 d_u(\Gamma, n, \varepsilon)/\varepsilon^2$, if $m \geq 32\sqrt{d_u(\Gamma, n, \varepsilon)}/\varepsilon^2$, we let $m_0 = n_0 = 32\sqrt{d_u(\Gamma, n, \varepsilon)}/\varepsilon^2$, then $m_0 \leq m$ and $n_0 \leq n$, and (m_0, n_0) is in the region defined in eq. (4.18), hence we only need to construct testing for (m_0, n_0). If $m \leq 32\sqrt{d_u(\Gamma, n, \varepsilon)}/\varepsilon^2$, by letting $n_0 = \lceil 512 d_u(\Gamma, n, \varepsilon)/(\varepsilon^2 m)\rceil$ we will have $m \leq n_0 \leq n$, and (m, n_0) satisfies $mn_0 \leq 1024 d_u(\Gamma, n, \varepsilon)/(\varepsilon^2 m)$. We only need to construct testing for (m, n_0). Therefore, without loss of generality, we assume

$$mn \leq \frac{1024 d_u(\Gamma, n, \varepsilon)}{\varepsilon^4}. \tag{4.19}$$

We introduce a testing scheme for ℓ_p bodies in this section. Inspired by the soft-thresholding algorithms for density estimation, which are known to be minimax optimal for ℓ_p balls where $p \in [1, 2]$ [70] and also the goodness of fit testing algorithms for ℓ_p bodies [65], we design an algorithm for the setting of likelihood-free hypothesis testing. To describe the testing regime, we use $\mathbf{X} = (\mathbf{X}_1, \cdots, \mathbf{X}_n)$ to denote the n i.i.d. samples collected from $\mathbb{P}_{\mathtt{X}}$, $\mathbf{Y} = (\mathbf{Y}_1, \cdots, \mathbf{Y}_n)$

to denote the n i.i.d. samples collected from $\mathbb{P}_{\mathtt{Y}}$ and $\mathbf{Z} = (\mathbf{X}_1, \cdots, \mathbf{Z}_m)$ to denote the m i.i.d. samples collected from $\mathbb{P}_{\mathtt{Z}}$.

To begin with, we divide those n samples $(\mathbf{X}_1, \cdots, \mathbf{X}_n)$ from $\mathbb{P}_{\mathtt{X}}$ and samples $\mathbf{Y}_1, \cdots, \mathbf{Y}_n$ from $\mathbb{P}_{\mathtt{Y}}$ each into two parts: let $n_0 = \lfloor n/2 \rfloor$, and

$$\begin{aligned} &\mathbf{X}^1 = (\mathbf{X}_1, \cdots, \mathbf{X}_{n_0}), \quad \mathbf{X}^2 = (\mathbf{X}_{n_0+1}, \cdots, \mathbf{X}_n), \\ \text{and} \quad &\mathbf{Y}^1 = (\mathbf{Y}_1, \cdots, \mathbf{Y}_{n_0}), \quad \mathbf{Y}^2 = (\mathbf{Y}_{n_0+1}, \cdots, \mathbf{Y}_n). \end{aligned}$$

We use $\hat{\theta}^1_{\mathtt{X}}, \hat{\theta}^2_{\mathtt{X}}, \hat{\theta}^1_{\mathtt{Y}}, \hat{\theta}^2_{\mathtt{Y}}$ and $\hat{\theta}_{\mathtt{Z}}$ to denote the average of samples within $\mathbf{X}^1, \mathbf{X}^2, \mathbf{Y}^1, \mathbf{Y}^2$ and $\mathbf{Z}$, i.e.

$$\begin{aligned} &\hat{\theta}^1_{\mathtt{X}} = \frac{1}{n_0}\sum_{i=1}^{n_0} \mathbf{X}_i, \quad \hat{\theta}^2_{\mathtt{X}} = \frac{1}{n_0}\sum_{i=n_0+1}^{n} \mathbf{X}_i, \quad \hat{\theta}^1_{\mathtt{Y}} = \frac{1}{n_0}\sum_{i=1}^{n_0} \mathbf{Y}_i, \\ &\hat{\theta}^2_{\mathtt{Y}} = \frac{1}{n_0}\sum_{i=n_0+1}^{n} \mathbf{Y}_i, \quad \text{and} \quad \hat{\theta}_{\mathtt{Z}} = \frac{1}{m}\sum_{j=1}^{m} \mathbf{Z}_j. \end{aligned}$$

The testing involves the following two steps:

(i) Calculating a subspace of small dimension such that $\mathbb{P}_{\mathtt{X}}$ and $\mathbb{P}_{\mathtt{Y}}$ are different within the subspace;

(ii) Adopting the likelihood free hypothesis testing scheme in [15, (2.2)] within the subspace.

We will illustrate these two steps in detail:

(a) **Calculating the Subspace:** Our goal in this step is to use samples from $\mathbf{X}^1$ and $\mathbf{Y}^1$ to calculate a subset $T \subseteq [D]$ such that

(i) Limit size: $|T| \lesssim d_u(\Gamma, n, \varepsilon)$;

(ii) Enough Separation: $\sum_{t\in T} \left((p_{\mathtt{X}})_t - (p_{\mathtt{Y}})_t\right)^2 \gtrsim \varepsilon^2$.

To achive this, we construct the following set $T_1, T_2 \subseteq [D]$ which satisfies $T_1 \cap T_2 = \emptyset$ and $T_1 \cup T_2 = [D]$:

$$T_1 = \{1, 2, \ldots, d_u(\Gamma, n, \varepsilon)\} \quad \text{and} \quad T_2 = \{d_u(\Gamma, n, \varepsilon) + 1, d_u(\Gamma, n, \varepsilon) + 2, \ldots, D\}, \tag{4.20}$$

where $d_u(\Gamma, n, \varepsilon)$ is defined in eq. (4.17). For $t \in [D]$, we suppose the t-th coordinate of $\hat{\theta}^1_{\mathtt{X}}$ and $\hat{\theta}^1_{\mathtt{Y}}$ to be $(\hat{\theta}^1_{\mathtt{X}})_t$ and $(\hat{\theta}^1_{\mathtt{Y}})_t$. We further define

$$T_3 = \left\{ t: \quad t \in T_2 \ \text{ and } \ |(\hat{\theta}^1_{\mathtt{X}})_t - (\hat{\theta}^1_{\mathtt{Y}})_t| \geq 4\sqrt{\frac{2\log(2D/\delta)}{n}} \right\}. \tag{4.21}$$

for some parameter $\delta > 0$ to be defined later. And we construct $T = T_1 \cup T_3$.

(b) **Testing in the Subspace:** After obtaining the set T of coordinates the last step, we use the samples within $\mathbf{X}^2, \mathbf{Y}^2$ and $\mathbf{Z}$ to do likelihood free hypothesis testing within the subspace formed through coordinates in T. The testing scheme is similar to [15, (2.2)],

i.e. if we use $(\hat{\theta}^2_{\mathtt{X}})_t$, $(\hat{\theta}^2_{\mathtt{Y}})_t$ and $(\hat{\theta}_{\mathtt{Z}})_t$ to denote t-th coordinates of $\hat{\theta}^2_{\mathtt{X}}$, $\hat{\theta}^2_{\mathtt{Y}}$ and $\hat{\theta}_{\mathtt{Z}}$, then we define

$$T_{\mathtt{LF}} = \sum_{t\in T}\left[\left((\hat{\theta}^2_{\mathtt{X}})_t - (\hat{\theta}_{\mathtt{Z}})_t\right)^2 - \left((\hat{\theta}^2_{\mathtt{Y}})_t - (\hat{\theta}_{\mathtt{Z}})_t\right)^2\right]. \tag{4.22}$$

And we adopt the testing scheme $\psi(\mathbf{X},\mathbf{Y},\mathbf{Z}) = \mathbb{I}\{T_{\mathtt{LF}} \geq 0\}$.

We argue that the testing scheme obtained in the above steps is minimax optimal up to constants and log factors. This result is summarized in the following theorem.

Theorem 4.21. *For $1 \leq p \leq 2$, if (m,n) lies in the region eq. (4.18), the test $\psi(\mathbf{X},\mathbf{Y},\mathbf{Z}) = \mathbb{I}\{T_{\mathtt{LF}} \geq 0\}$ with $T_{\mathtt{LF}}$ defined in eq. (4.22) satisfies eq. (4.10).*

The proof of Theorem 4.21 is deferred to appendix C.6.1.

Lower Bounds

In this section, we show that the region in eq. (4.16) is also necessary for the testing. We first define

$$d_l(\Gamma,n,\varepsilon) = \max\left\{d : (a_d)^p n^{\frac{p-2}{2}} \geq 192\varepsilon^2\right\}. \tag{4.23}$$

This is summarized in the following theorem.

Theorem 4.22. *If there exists a testing scheme ψ, which takes n samples from $p_{\mathtt{X}}$ and $p_{\mathtt{Y}}$ and m samples from $p_{\mathtt{Z}}$, and ψ satisfies eq. (4.10), then (m,n) satisfies*

$$m \geq \frac{1}{\varepsilon^2}, \quad n \geq \frac{\sqrt{d_l(\Gamma,n,\varepsilon)}}{2\varepsilon^2}, \quad \text{and} \quad mn \geq \frac{d_l(\Gamma,n,\varepsilon)}{96\varepsilon^4}. \tag{4.24}$$

The proof of Theorem 4.22 is deferred to appendix C.6.2.

Comparison with results of Baraud

In [65], the minimax sample complexity of goodness-of-fit testing for ℓ_p bodies is given. According to [65, Section 4.1] and [65, Section 4.2], for fixed positive n and ℓ_p bodies Γ defined in eq. (4.4), if we let

$$d = \arg\max_d \left\{\frac{\sqrt{d}}{n} \wedge a_d^2 \cdot d^{1-2/p}\right\},$$

then the goodness-of-fit testing can be done if and only if $\varepsilon^2 \gtrsim \frac{\sqrt{d}}{n}$. Hence the goodness-of-fit testing region at scale ε satisfies

$$n \gtrsim \frac{\sqrt{d(\Gamma,n,\varepsilon)}}{\varepsilon^2},$$

where $d(\Gamma,n,\varepsilon)$ is given by

$$d(\Gamma,n,\varepsilon) = \max\left\{d : (a_d)^p n^{\frac{p-2}{2}} \geq \varepsilon^2\right\}.$$

We notice that this form of $d(\Gamma,n,\varepsilon)$ is similar to the form $d_u(\Gamma,n,\varepsilon)$ and $d_l(\Gamma,n,\varepsilon)$ defined in eq. (4.17) and eq. (4.23).

Generalization to Infinite Dimensional Sets

In this section, we generalize Theorem 4.21 and Theorem 4.22 to infinite dimensional ℓ_p bodies. For a given positive non-increasing sequence $a_1 \geq a_2 \geq \cdots \geq a_n$, we consider set

$$\Gamma = \left\{ \boldsymbol{\theta} = (\theta_{1:\infty}) : \sum_{t=1}^{\infty} \frac{|\theta_t|^p}{a_t^p} \leq 1 \right\}. \tag{4.25}$$

To guarantee the compactness of the set Γ, we assume that $\lim_{t\to\infty} a_t = 0$. To begin with, we recall the definition of the coordinate-wise Kolmogorov dimension in Theorem 4.2. For this set, we can easily check that

$$D_{\mathsf{c}}(\Gamma, \varepsilon) = \min \{ D \geq 1 : a_D \leq \varepsilon \}. \tag{4.26}$$

Then we can characterize the feasible region of likelihood-free hypothesis testing in the following theorems:

Theorem 4.23 (Upper Bounds for Infinite Dimensional Sets)**.** *For set Γ in the form eq. (4.25), we let the Kolmogorov dimension of Γ at scale ε to be $D(\Gamma, \varepsilon)$. We further define*

$$d_u(\Gamma, n, \varepsilon) = \max \left\{ d : (a_d)^p n^{\frac{p-2}{2}} \geq \frac{\varepsilon^2/9}{576 \log(4 D_{\mathsf{c}}(\Gamma, \varepsilon/3)/\delta)} \right\}.$$

Then if (m, n) satisfies

$$\left\{ (m,n) : \quad m \geq \frac{32}{\varepsilon^2}, \quad n \geq \frac{32\sqrt{d_u(\Gamma, n, \varepsilon)}}{\varepsilon^2} \quad \text{and} \quad mn \geq \frac{512 d_u(\Gamma, n, \varepsilon)}{\varepsilon^4} \right\},$$

then if we adopt the testing scheme ψ defined in section 4.4.2 with $D = D_{\mathsf{c}}(\Gamma, \varepsilon/3)$, then ψ satisfies eq. (4.10).

Theorem 4.24 (Lower Bounds for Infinite Dimensional Sets)**.** *For given Γ, we define*

$$d_l(\Gamma, n, \varepsilon) = \max \left\{ d : (a_d)^p n^{\frac{p-2}{2}} \geq 192 \varepsilon^2 \right\}.$$

Then any (m, n) such that there exists a test ψ which satisfies eq. (4.10) must satisfy

$$m \geq \frac{1}{\varepsilon^2}, \quad n \geq \frac{\sqrt{d_l(\Gamma, n, \varepsilon)}}{2\varepsilon^2}, \quad \text{and} \quad mn \geq \frac{d_l(\Gamma, n, \varepsilon)}{96\varepsilon^4}.$$

The proof of Theorem 4.23 and Theorem 4.24 are deferred to appendix C.6.3.

Examples

We revisit the set defined in eq. (4.4). In this section, we will calculate the feasible region of likelihood-free hypothesis testing for set Γ, which is summarized in the following proposition.

Proposition 4.25. *The feasible region of* (m, n) *of likelihood-free hypothesis testing for set* Γ *defined in eq.* (4.4) *contains the following set:*

$$\left\{(m,n): \quad m \geq \varepsilon^{-2},\ n \gtrsim \varepsilon^{-\frac{12}{5}} \log^{2/5}(1/\varepsilon),\ m \cdot n^{\frac{3}{2}} \gtrsim \varepsilon^{-6} \log(1/\varepsilon)\right\},$$

and is contained by the following set:

$$\left\{(m,n): \quad m \geq \varepsilon^{-2},\ n \gtrsim \varepsilon^{-\frac{12}{5}},\ m \cdot n^{\frac{3}{2}} \gtrsim \varepsilon^{-6}\right\},$$

where we use $\gtrsim$ *to hide universal constant factors.*

The proof of Theorem 4.25 directly follows from Theorem 4.23 and Theorem 4.24 after noticing that $D(\Gamma, \varepsilon) = 1/\varepsilon$ and

$$d_u(\Gamma, n, \varepsilon) \lesssim \frac{\log(1/\varepsilon)}{\sqrt{n} \cdot \varepsilon^2}, \quad \text{and} \quad d_l(\Gamma, n, \varepsilon) \gtrsim \frac{1}{\sqrt{n} \cdot \varepsilon^2}.$$

Note that this proposition also provide a hard case where the conditions of likelihood-free hypothesis testing region do not hold in [15, Open Problem 4].

Part II

Sequential Prediction and Regression

Chapter 5

On the Minimax Regret of Sequential Probability Assignment via Square-Root Entropy

This chapter is based on the paper "On the Minimax Regret of Sequential Probability Assignment via Square-Root Entropy" by Zeyu Jia, Yury Polyanskiy, and Alexander Rakhlin, published published at the Conference on Learning Theory (COLT) 2025.

5.1 Introduction

We consider the problem of sequential probability assignment under logarithmic loss. This framework has been studied extensively over the decades in fields such as information theory—where it relates to sequence compression—in gambling and sequential investment—where it is linked to wealth growth—and in online learning [4]. In its more recent incarnation, next-token prediction has emerged as a central challenge in training large language models, where the goal is to minimize the logarithmic loss (commonly referred to as cross-entropy loss) on nearly all available data.

Let us now describe the formal setup. On each round $t = 1, \dots, n$, the forecaster chooses a distribution $\widehat{p}_t$ over the finite alphabet $\mathcal{Y}$, observes $y_t \in \mathcal{Y}$, and incurs a loss of $-\log \widehat{p}_t(y_t)$. Over the n rounds, the cumulative cost is $\sum_{t=1}^{n} -\log \widehat{p}_t(y_t)$. Since the distribution $\widehat{p}_t$ is chosen based on the previous outcomes $y_1, \dots, y_{t-1}$, we associate $\widehat{p}_t$ with a conditional distribution $\widehat{p}(\cdot|y_1, \dots, y_{t-1})$ and write the cumulative loss succinctly as $-\log \widehat{\mathbf{p}}(\mathbf{y})$, where $\widehat{\mathbf{p}}$ is the corresponding joint distribution over sequences $\mathbf{y} = (y_1, \dots, y_n)$.

The cumulative loss of the forecaster can be compared to that of the best "expert" in a class $\mathcal{Q} \subseteq \Delta(\mathcal{Y}^n)$, each identified with a joint probability distribution $\mathbf{q} \in \mathcal{Q}$. The forecaster aims to minimize regret

$$\sum_{t=1}^{n} -\log \widehat{p}_t(y_t) - \inf_{\mathbf{q}\in\mathcal{Q}} \sum_{t=1}^{n} -\log q_t(y_t|y_1, \dots, y_{n-1}) = \sup_{\mathbf{q}\in\mathcal{Q}} \log\left(\frac{\mathbf{q}(\mathbf{y})}{\widehat{\mathbf{p}}(\mathbf{y})}\right) \tag{5.1}$$

for *any* sequence $y_1, \dots, y_n$. As such, the problem falls under the umbrella of worst-case prediction (also known as individual sequence prediction).

In the more general problem of *prediction with side information* (or, contextual prediction), the forecaster observes additional covariates $x_t \in \mathcal{X}$ prior to making the probabilistic forecast $\widehat{p}_t \in \Delta(\mathcal{Y})$ on round t. In this case, the regret expression becomes

$$\sum_{t=1}^{n} -\log \widehat{p}_t(y_t) - \inf_{\mathbf{q}\in\mathcal{Q}} \sum_{t=1}^{n} -\log q_t(y_t|x_t) \tag{5.2}$$

and $\mathbf{q} = (q_1, \ldots, q_n)$ is a sequence of conditional distributions $q_t : \mathcal{X} \to \Delta(\mathcal{Y})$. Of course, if $x_t = (y_1, \ldots, y_{t-1})$ and $\mathcal{X} = \mathcal{Y}^*$, the problem reduces to the non-contextual version in (5.1).

The intrinsic difficulty of the prediction problem in the non-contextual case is

$$\mathcal{R}_n(\mathcal{Q}) \coloneqq \inf_{\widehat{\mathbf{p}}\in\Delta(\mathcal{Y}^n)} \sup_{\mathbf{y}\in\mathcal{Y}^n} \sup_{\mathbf{q}\in\mathcal{Q}} \log \left(\frac{\mathbf{q}(\mathbf{y})}{\widehat{\mathbf{p}}(\mathbf{y})} \right), \tag{5.3}$$

a quantity referred to as the *worst-case redundancy*, or *minimax regret*. A similar notion can be defined for the contextual version of the problem, when $(x_1, \ldots, x_n)$ also form an individual (i.e. arbitrary) sequence; however, for brevity, we defer this definition to section 5.3.

The goal of this paper is to analyze the behavior of $\mathcal{R}_n(\mathcal{Q})$, for both contextual and non-contextual cases, in terms of geometric concepts—such as covering numbers (or entropy) and scale-sensitive dimensions—analogous to how sample complexity is quantified in statistical learning through the complexity measures of the function class. This objective is not new; over the past several decades, numerous seminal ideas have been developed to address this question [16, 97–102], and more recently in [17, 18], among many others.

In particular, the classical result of [99] states that in the non-contextual case, $\mathcal{R}_n(\mathcal{Q})$ has the following closed form:

$$\mathcal{R}_n(\mathcal{Q}) = \log \sum_{\mathbf{y}\in\mathcal{Y}^n} \sup_{\mathbf{q}\in\mathcal{Q}} \mathbf{q}(\mathbf{y}), \tag{5.4}$$

and the optimal strategy in eq. (5.3) is attained by the Shtarkov distribution $\mathbf{p}^*(\mathbf{y}) \propto \sup_{\mathbf{q}\in\mathcal{Q}} \mathbf{q}(\mathbf{y})$, also known as the normalized maximum likelihood. While (5.4) is more succinct than (5.3), it is still not amenable to analysis with standard tools, except for special cases [4].

5.1.1 Towards a General Result

To the best of our knowledge, the first analysis of minimax regret for non-parametric (but iid) class $\mathcal{Q}$ was proposed in [103], who presented an upper bound involving a Dudley integral. This work was extended to a general $\mathcal{Q}$ in [16], who observed that, owing to the equalizing property of the optimal strategy $\mathbf{p}^*$, the Shtarkov sum (5.4) can be expressed as $\mathcal{R}_n(\mathcal{Q}) = \mathbb{E}_{\mathbf{y}\sim\mathbf{p}^*}\left[\sup_{\mathbf{q}\in\mathcal{Q}} \log \frac{\mathbf{q}(\mathbf{y})}{\mathbf{p}^*(\mathbf{y})}\right]$ and further upper bounded by the expected supremum of a subgaussian process indexed by the collection $\mathcal{Q}$, much in the spirit of the empirical process theory approach in statistics and learning theory [104]. Notably, the process was shown to be subgaussian with respect to a pseudometric $d(f, g) = \left(\sum_{t=1}^{n} \max_{y_{1:t}} (\log f(y_t|y_{1:t-1}) - \log g(y_t|y_{1:t-1}))^2\right)^{1/2}$, where $y_{1:t} := (y_1, \ldots, y_t)$. [16] subsequently developed a Dudley-integral-style bound for the Shtarkov sum; however, the induced

covering numbers are difficult to control due to the unbounded nature of the logarithm for small values, ultimately leading to generally suboptimal upper bounds on $\mathcal{R}_n(\mathcal{Q})$ as a consequence of clipping probabilities away from 0. Remarkably, retaining the logarithm in the definition of the pseudometric yielded an interesting consequence: the main upper bound in [16, Theorem 3] assumes a form typical of bounds obtained via localized or *offset Rademacher* complexities, as encountered in square loss regression [104–106].

Several subsequent attempts have been made to derive tighter upper bounds on minimax regret, with the focus shifting toward the contextual case. In an effort to obtain, as an upper bound on minimax regret, a stochastic process that is subgaussian with respect to a pseudometric on the *values* of the distributions q_t rather than on their logarithms, [17] employed a first-order expansion of the logarithmic loss. This approach upper bounded the minimax regret by a version of sequential offset Rademacher complexity. Unfortunately, despite their efforts to tame the explosive nature of the derivatives, the authors were unable to derive upper bounds on the offset process that were independent of the clipping range, even in the finite case [17, Lemma 2].

The important work of [18] leveraged the self-concordance properties of the logarithm to upper bound the minimax value in the contextual case by an offset-like process that offered clear advantages over earlier approaches. In particular, for a finite collection, the resulting process could be controlled without resorting to clipping. However, the process did not exhibit a subgaussian nature, which prevented the authors from employing chaining arguments. This issue arises from the presence of linear terms of the form $q_t(y_t)/p_t(y_t)$, which are small in expectation over $y_t \sim p_t$ (thus permitting single-scale discretization) but become uncontrolled when the ratio is squared.

The Hellinger distance has long been recognized as a convenient metric on the space of distributions [56, 57, 104, 107, 108]. In particular, as an ℓ_2-distance between the square roots of distributions, it offers the possibility of combining the benefits of offset-based analysis with those of multi-scale chaining. This is the approach we adopt in this paper. Specifically, we employ an approximation $\log x \le \zeta(x) - \frac{1}{4\log(n|\mathcal{Y}|)} \cdot \zeta(x)^2$, which holds over an appropriate range of x, and where $\zeta(x)$ behaves as $2(\sqrt{x}-1)$ for $x \le 1$. Applied, roughly speaking, to $x = q_t(y_t)/p_t(y_t)$, this inequality allows us to leverage symmetrization and chaining techniques while also capitalizing on the fast rates provided by the offset sequential Rademacher process. Our approach, therefore, appears to resolve the technical issues encountered by the various techniques, starting with [16], at least in the so-called Donsker regime (with respect to our entropy definition), where chaining provides an advantage.

To demonstrate the sharpness of our results—again, in the Donsker regime—for the contextual version of the problem, we develop new lower bound techniques that build upon [109, Lemma 10]. In particular, we introduce a novel sequential scale-sensitive dimension, prove a combinatorial result that controls the size of the sequential cover in terms of this dimension, and employ this new notion to derive nearly matching lower bounds for any $\mathcal{Q}$ (in the contextual case). This approach significantly strengthens the earlier work, which only guaranteed lower bounds for a modified function class. Our techniques will be presented in full detail in the companion paper [110].

We now summarize our contributions.

5.1.2 Summary of Main Results

We study minimax regret in both non-contextual (section 5.2) and contextual (section 5.3) settings. Our results below are stated with respect to sequential square-root entropy, $\mathcal{H}_{\mathsf{sq}}(\mathcal{Q}, \alpha, n)$, defined formally in section 5.1.3.

An upper bound on minimax regret for the non-contextual case: For any class of distributions $\mathcal{Q} \subseteq \Delta(\mathcal{Y}^n)$, with sequential square-root entropy $\mathcal{H}_{\mathsf{sq}}(\mathcal{Q}, \alpha, n)$ at scale α, the minimax regret (5.3) (and, hence, the Shtarkov sum (5.4)) has the following upper bound:

$$\mathcal{R}_n(\mathcal{Q}) \lesssim 1 + \inf_{\gamma > \delta > 0} \left\{ n\delta\sqrt{|\mathcal{Y}|} + \sqrt{n|\mathcal{Y}|} \int_\delta^\gamma \sqrt{\mathcal{H}_{\mathsf{sq}}(\mathcal{Q}, \alpha, n)} d\alpha + \mathcal{H}_{\mathsf{sq}}(\mathcal{Q}, \gamma, n) \right\},$$

where we use $\lesssim$ to hide constants and $\log(n|\mathcal{Y}|)$ factors.

Tight characterization of contextual sequential probability assignment: Focusing on the binary alphabets for simplicity, we provide both an upper bound and a lower bound for the minimax regret, defined below in (5.12) and again denoted here as $\mathcal{R}_n(\mathcal{Q})$. The following upper bound holds in terms of sequential square-root entropy:

$$\mathcal{R}_n(\mathcal{Q}) \lesssim 1 + \inf_{\gamma > \delta > 0} \left\{ n\delta + \sqrt{n} \int_\delta^\gamma \sqrt{\mathcal{H}_{\mathsf{sq}}(\mathcal{Q}, \alpha, n)} d\alpha + \mathcal{H}_{\mathsf{sq}}(\mathcal{Q}, \gamma, n) \right\}.$$

According to this upper bound, for any nonparametric function class $\mathcal{Q}$ which satisfies $\mathcal{H}_{\mathsf{sq}}(\mathcal{Q}, \alpha, n) = \mathcal{O}\left(\alpha^{-p}\right)$, the minimax regret is upper bounded as

$$\mathcal{R}_n(\mathcal{Q}) = \begin{cases} \tilde{\mathcal{O}}\left(n^{\frac{p}{p+2}}\right) & \text{if } 0 < p \leq 2, \\ \tilde{\mathcal{O}}\left(n^{\frac{p-1}{p}}\right) & \text{if } p > 2. \end{cases} \tag{5.5}$$

In addition, we establish a lower bound demonstrating the tightness of (5.5) for $0 \leq p \leq 2$. Hence, for nonparametric classes with parameter $p \leq 2$, our results offer a tight characterization of the minimax regret in terms of the sequential square-root entropy. Our upper bound further yields an $\tilde{\mathcal{O}}\left(\sqrt{n}\right)$ bound for the Hilbert ball problem, thereby answering a question posed in [17].

Our contributions are also technical. The proof of the upper bound introduces a novel approach to analyzing the expectation of the offset Rademacher process, enabling us to handle cases with unbounded coefficients. We adopt a chaining argument alongside the analysis of offset Rademacher processes in our proof. On the lower bound side, as mentioned, our techniques involve a new definition of a scale-sensitive dimension and a novel argument for lower-bounding the sequential offset Rademacher complexity that is applicable beyond this paper.

Overall, our results largely resolve the open problem stated in [17] by tightly characterizing the minimax regret of contextual probability assignment for any class of conditional probability distributions in terms of entropic quantities, at least in the Donsker regime (according to the our definition of entropy).

5.1.3 Notation

Given $\mathbf{q} \in \Delta(\mathcal{Y}^n)$, we write $q_t(y_t \mid \mathbf{y}) = q_t(y_t \mid y_{1:t-1})$ to denote the conditional probability for any length-n sequence $\mathbf{y} = (y_1, \cdots, y_n) \in \mathcal{Y}^n$. A $\{0,1\}$-path $\mathbf{w}$ of depth n is a tuple $(w_1, \ldots, w_n) \in \{0,1\}^n$. For any set $\mathcal{X}$, a depth-n $\mathcal{X}$-valued binary tree (or, simply, 'a tree') $\mathbf{x}$ has $2^n - 1$ nodes, where each node takes value in $\mathcal{X}$. Formally, $\mathbf{x} = (x_1, \ldots, x_n)$ with $x_i : \{0,1\}^{i-1} \to \mathcal{X}$. We write $x_i(\mathbf{w}) = x_i(w_{1:i-1})$ for brevity. For a depth-n $\mathcal{X}$-valued tree $\mathbf{x}$ and function $f : \mathcal{X} \to [0,1]$, we use $f \circ \mathbf{x}$ to denote the depth-n $[0,1]$-valued tree whose value at depth t on path $\mathbf{w}$ equals to $f(x_t(\mathbf{w}))$. We write $\mathcal{F} \circ \mathbf{x} = \{f \circ \mathbf{x} : f \in \mathcal{F}\}$.

Additionally, we use the following asymptotic notation: for positive sequence $\{a_n\}$ and $\{b_n\}$ (or functions $f(\alpha), g(\alpha) : (0,1) \to \mathbb{R}_+$, we use $a_n = \mathcal{O}(b_n)$ (or $f(\alpha) = \mathcal{O}(g(\alpha))$) if there exists a positive constant c such that $a_n \le c \cdot b_n$ for any n (or $f(\alpha) \le c \cdot g(\alpha)$ for any α), and we use $a_n = \tilde{\mathcal{O}}(b_n)$ if there exists a positive constant c and positive integer r such that $a_n \le c \cdot (\log n)^r \cdot b_n$ (or $f(\alpha) \le c \cdot (\log(1/\alpha))^r \cdot g(\alpha)$). We use notation $a_n = \Omega(b_n)$ (or $f(\alpha) = \Omega(g(\alpha))$ if and only if $b_n = \mathcal{O}(a_n)$ (or $g(\alpha) = \mathcal{O}(f(\alpha))$, and $\tilde{\Omega}$ is defined similarly. The notation $a_n = \Theta(b_n)$ (or $f(\alpha) = \Theta(g(\alpha))$) is used if and only if both $a_n = \mathcal{O}(b_n)$ and $a_n = \Omega(b_n)$ hold (or both $f(\alpha) = \mathcal{O}(g(\alpha))$ and $f(\alpha) = \Omega(g(\alpha))$ hold). The notation $\tilde{\Theta}$ is defined similarly.

5.1.4 Organization

In section 5.2, we present our results in upper bounding the Shtarkov sum using sequential square-root entropy. In section 5.3, we revisit the problem of contextual sequential probability assignment, and provide upper and lower bounds for the minimax regret in terms of sequential square-root entropy. Finally, in section 5.4 we provide a proof sketch of our main result, Theorem 5.2. All the technical proofs are deferred to the appendix.

5.2 Upper Bound for Shtarkov Sum through Sequential Square-Root Covering

In this section, we upper bound the minimax regret eq. (5.3) or Shtarkov sum eq. (5.4) in terms of the ℓ_∞ sequential square-root covering defined as follows.

Definition 5.1 (sequential square-root cover and entropy)**.** *Let $\mathcal{Y}$ be a finite alphabet. For a class of joint distributions $\mathcal{Q}$ over $\mathcal{Y}^n$, we say that a finite class $\mathcal{V}$ of joint distributions over $\mathcal{Y}^n$ is a sequential square-root cover (in the ℓ_∞ sense) of $\mathcal{Q}$ at scale α if*

$$\sup_{\mathbf{q} \in \mathcal{Q}} \max_{\mathbf{w} \in \mathcal{Y}^n} \min_{\mathbf{v} \in \mathcal{V}} \max_{t \in [n]} \max_{y \in \mathcal{Y}} \left| \sqrt{q_t(y \mid \mathbf{w})} - \sqrt{v_t(y \mid \mathbf{w})} \right| \le \alpha. \tag{5.6}$$

We use $\mathcal{N}_{\mathsf{sq}}(\mathcal{Q}, \alpha, n)$ to denote the size of the smallest cover of class $\mathcal{Q}$, and we use $\mathcal{H}_{\mathsf{sq}}(\mathcal{Q}, \alpha, n) = \log \mathcal{N}_{\mathsf{sq}}(\mathcal{Q}, \alpha, n)$ to denote the sequential square-root entropy of $\mathcal{Q}$.

In words, the requirement placed on $\mathcal{V}$ is that for any joint distribution $\mathbf{q} \in \mathcal{Q}$ and any sequence $\mathbf{w} \in \mathcal{Y}^n$, there exists a "representative" joint distribution $\mathbf{v}$ in $\mathcal{V}$ that is close to $\mathbf{q}$ in

terms of the difference of square roots of the conditional probabilities $\mathbf{q}$ and $\mathbf{v}$ assign to any outcome y, uniformly for all time steps.

In this definition, ℓ_∞ refers to the maximum over $t \in [n]$, which is consistent with prior uses of such sequential and empirical notions of a cover. We also remark that $\max_{y\in\mathcal{Y}} \left|\sqrt{q_t(y \mid \mathbf{w})} - \sqrt{v_t(y \mid \mathbf{w})}\right|$ is within $\sqrt{|\mathcal{Y}|}$ of the Hellinger distance between these two conditional distributions (which is the ℓ_2 version with respect to the $y \in \mathcal{Y}$). If scaling with $|\mathcal{Y}|$ is not of interest, we can instead think of the sequential square-root cover as a *sequential Hellinger cover*.

Theorem 5.2. *For any $n \geq 7$ and class $\mathcal{Q} \subseteq \Delta(\mathcal{Y}^n)$, we have*

$$\mathcal{R}_n(\mathcal{Q}) = \tilde{\mathcal{O}}\left(1 + \inf_{\gamma>\delta>0}\left\{n\delta\sqrt{|\mathcal{Y}|} + \sqrt{n|\mathcal{Y}|}\int_\delta^\gamma \sqrt{\mathcal{H}_{\mathsf{sq}}(\mathcal{Q},\alpha,n)}d\alpha + \mathcal{H}_{\mathsf{sq}}(\mathcal{Q},\gamma,n)\right\}\right),$$

where $\tilde{\mathcal{O}}$ hides constants and logarithmic factors of n and $|\mathcal{Y}|$.

The proof of Theorem 5.2 is deferred to appendix D.2, and we provide a sketch of the proof in section 5.4. The theorem immediately implies an upper bound on $\mathcal{R}_n(\mathcal{Q})$ whenever the sequential square-root entropy scales with α^{-p}, as shown in the following corollary.

Corollary 5.3. *When $\mathcal{H}_{\mathsf{sq}}(\mathcal{Q},\alpha,n) = \tilde{\mathcal{O}}\left(\alpha^{-p}\right)$ for some $p \geq 0$, it holds that*

$$\mathcal{R}_n(\mathcal{Q}) = \begin{cases} \tilde{\mathcal{O}}\left(n^{\frac{p}{p+2}}\right) & \text{if } p \leq 2, \\ \tilde{\mathcal{O}}\left(n^{\frac{p-1}{p}}\right) & \text{if } p > 2. \end{cases}$$

5.2.1 Comparison with Previous Results

We compare our results with [4, 16], which also provide an upper bound on the minimax regret using entropy. Taking

$$d(f,g) = \left(\frac{1}{n}\sum_{t=1}^n \max_{y_{1:n}} (\log f(y_t|y_{1:t-1}) - \log g(y_t|y_{1:t-1}))^2\right)^{1/2}, \tag{5.7}$$

as the (pseudo)metric, the authors define a notion of entropy $\mathcal{H}_{\mathsf{log}}(\mathcal{Q},\alpha,n)$ as the logarithm of the size of the smallest covering at scale α under d. [16] establish that

$$\mathcal{R}_n(\mathcal{Q}) \lesssim \inf_{\gamma>0}\left\{\sqrt{n}\int_0^\gamma \sqrt{\mathcal{H}_{\mathsf{log}}(\mathcal{F},\varepsilon,n)}d\varepsilon + \mathcal{H}_{\mathsf{log}}(\mathcal{F},\gamma,n)\right\}. \tag{5.8}$$

The form of the bound appears frequently in the literature on prediction with square loss, in both fixed design regression and online regression. Writing the definition of the above covering notion in the form of (5.6), we have

$$\sup_{\mathbf{q}\in\mathcal{Q}} \min_{\mathbf{v}\in\mathcal{V}} \max_{\mathbf{w}\in\mathcal{Y}^n} \max_{t\in[n]} \max_{y\in\mathcal{Y}} |\log q_t(y \mid \mathbf{w}) - \log v_t(y \mid \mathbf{w})| \leq \alpha. \tag{5.9}$$

with the only difference that we opted for $\max_{t\in[n]}$ instead of the ℓ_2 version employed above. Modulo this difference, the requirement (5.9) is clearly more stringent than (5.6) as the representative $\mathbf{v}$ has to be chosen irrespective of the data $\mathbf{w}$, making the notion of the cover similar to the (often prohibitively large) sup-norm cover. The line of work on sequential complexities addresses this shortcoming via symmetrization, an approach we also take in this paper. Finally, we note that $\sup_{y\in\mathcal{Y}} |\sqrt{p(y)} - \sqrt{q(y)}| \le \sup_{y\in\mathcal{Y}} |\log p(y) - \log q(y)|$ and, thus, we expect $\mathcal{H}_{\log}$ to be larger (and often much larger) than $\mathcal{H}_{\mathsf{sq}}$.

Note that while the sequential square-root entropy is an improvement over the entropy in [16], there are still interesting distribution classes where it does not yield the correct bound. For example, consider the renewal process class (definition is included in appendix D.6). It is known from [111] that minimax regret is $\Theta(\sqrt{n})$. In appendix D.6, however, we show that the sequential square-root entropy is always lower bounded by $\Omega(n)$.

5.3 Binary Contextual Sequential Probability Assignment

In this section, we connect the problem of contextual sequential probability assignment to the non-contextual case discussed in the previous section. Application of the general bound of Theorem 5.2 will then lead to the main results of our paper.

For simplicity of presentation, and to make our results more directly comparable to prior work, we focus on the binary alphabet $\mathcal{Y} = \{0, 1\}$. With some abuse of notation we let $\widehat{p}_t \in [0, 1]$ denote the probability of the outcome 1. The loss incurred on round t after making the prediction $\widehat{p}_t$ can thus be written as

$$\ell(\widehat{p}_t, y_t) := -y_t \log \widehat{p}_t - (1 - y_t) \log(1 - \widehat{p}_t). \tag{5.10}$$

Similarly, we re-parametrize conditional distributions $\mathcal{Q}$ by instead working with a class $\mathcal{F}$ of experts, mapping $\mathcal{X}$ to $[0, 1]$. This re-parametrization is consistent with other prior works. With this notation, the cumulative loss of an expert f is $\sum_{t=1}^{n} \ell(f(x_t), y_t)$, and regret is defined (in a form that is more explicit than (5.2)) as

$$\mathcal{R}_n(\mathcal{F}, \widehat{p}_{1:n}, x_{1:n}, y_{1:n}) := \sum_{t=1}^{n} \ell(\widehat{p}_t, y_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t). \tag{5.11}$$

Recall that x_t may depend arbitrarily on the history

$$\mathcal{H}_t = \{x_1, \widehat{p}_1, y_1, \ldots, x_{t-1}, \widehat{p}_{t-1}, y_{t-1}\},$$

and y_t may depend arbitrarily on $\mathcal{H}_t, x_t$ and $\widehat{p}_t$. Based on the order of making predictions and observing outcomes, we define the minimax regret as

$$\mathcal{R}_n(\mathcal{F}) = \sup_{x_1\in\mathcal{X}} \inf_{\widehat{p}_1\in[0,1]} \sup_{y_1\in\{0,1\}} \cdots \sup_{x_n\in\mathcal{X}} \inf_{\widehat{p}_n\in[0,1]} \sup_{y_n\in\{0,1\}} \mathcal{R}(\mathcal{F}, \widehat{p}_{1:n}, x_{1:n}, y_{1:n}), \tag{5.12}$$

or, more succinctly, as

$$\mathcal{R}_n(\mathcal{F}) = \left\{ \sup_{x_t\in\mathcal{X}} \inf_{\widehat{p}_t\in[0,1]} \sup_{y_t\in\{0,1\}} \right\}_{t=1}^{n} \mathcal{R}(\mathcal{F}, \widehat{p}_{1:n}, x_{1:n}, y_{1:n}).$$

Here, the curly braces indicated a repeated application of the operators.

The above expression indeed matches the aforementioned dependencies. To make the connection to the previous section, we start with the following observation. Consider an adversary that is not allowed to adapt the sequence of x's to the past predictions made by the forecaster and instead has to fix ahead of time a strategy for choosing x's based only on the outcomes y's; this is equivalent to fixing an $\mathcal{X}$-valued tree $\mathbf{x} = (x_1, \dots, x_n)$ and presenting $x_t(y_{1:t-1})$ to the forecaster at the beginning of round t. Notably, the sequence of y_t's can still be adapted to the predictions of the forecaster. The following lemma states that such an adversary is just as powerful as the fully-adaptive one, even if the forecaster knows the strategy $\mathbf{x}$:

Lemma 5.4. *For any $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ which is convex in its first argument, and for any $\phi : \mathcal{Y}^n \times \mathcal{X}^n \to \mathbb{R}$,*

$$\begin{aligned}
&\left\{ \sup_{x_t} \inf_{\widehat{p}_t} \sup_{y_t} \right\}_{t=1}^{n} \left[\sum_{t=1}^{n} \ell(\widehat{p}_t, y_t) - \phi(y_{1:n}, x_{1:n}) \right] \\
&= \sup_{\mathbf{x}} \left\{ \inf_{\widehat{p}_t} \sup_{y_t} \right\}_{t=1}^{n} \left[\sum_{t=1}^{n} \ell(\widehat{p}_t, y_t) - \phi(y_{1:n}, x_1, x_2(y_1), \dots, x_n(y_{1:n-1})) \right]
\end{aligned}$$

where the supremum in the last expression is over all depth-n $\mathcal{X}$-valued trees $\mathbf{x}$. In particular, for $\phi(y_{1:n}, x_{1:n}) = \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t)$ and logarithmic loss discussed in this paper,

$$\mathcal{R}_n(\mathcal{F}) = \sup_{\mathbf{x}} \mathcal{R}_n(\mathcal{Q}_{\mathbf{x}})$$

for the set $\mathcal{Q}_{\mathbf{x}} = \mathcal{F} \circ \mathbf{x} = \{f \circ \mathbf{x} : f \in \mathcal{F}\}$ with $(f \circ \mathbf{x})(\mathbf{y}) = \prod_{t=1}^{n} \{\mathbb{I}[y_t = 1] f(x_t(\mathbf{y})) + \mathbb{I}[y_t = 0](1 - f(x_t(\mathbf{y}))\}$ for any $\mathbf{y} \in \{0,1\}^n$.

For the logarithmic loss, this result was proved in [112, Theorem 3.2]. The proof of Theorem 5.4 is deferred to appendix D.3.1.

The importance of this proposition is two-fold. First, it shows a possibly counter-intuitive property that regret is unchanged if the adversary's x's are not allowed to depend on the actions of the forecaster, but only on the y's. In other words, there exists a best possible adversarial tree $\mathbf{x}$ that saturates regret for all possible strategies of the forecaster.[1] Second, note that when the tree $\mathbf{x}$ is fixed ahead of time, the resulting problem corresponds to the problem discussed in Theorem 5.2 with $\mathcal{Q}_{\mathbf{x}} = \mathcal{F} \circ \mathbf{x}$.

5.3.1 Upper Bound with Sequential Square-Root Entropy

We now repeat the definition Theorem 5.1, adapting it to the case of binary alphabet and the real-valued re-parametrization of probabilities:

[1] Note that the optimal learning algorithm for this $\mathbf{x}$ tree is not guaranteed to be optimal for the actual problem (5.12).

Definition 5.5 (sequential square-root cover and entropy)**.** *Suppose $\mathcal{V}$ and $\mathcal{A}$ are two sets of $[0,1]$-valued binary trees of depth n. We say $\mathcal{V}$ is a sequential square-root cover (in the ℓ_∞ sense) of $\mathcal{A}$ at scale α if*

$$\max_{\mathbf{y}\in\{0,1\}^n} \sup_{\mathbf{a}\in\mathcal{A}} \inf_{\mathbf{v}\in\mathcal{V}} \max_{t\in[n]} \max \left\{ \left| \sqrt{a_t(\mathbf{y})} - \sqrt{v_t(\mathbf{y})} \right|, \left| \sqrt{1-a_t(\mathbf{y})} - \sqrt{1-v_t(\mathbf{y})} \right| \right\} \le \alpha.$$

We use $\mathcal{N}_{\mathsf{sq}}(\mathcal{A},\alpha,n)$ to denote the size of the smallest sequential square-root cover at scale α. For any set $\mathcal{X}$ and function class $\mathcal{F}\subseteq\{f:\mathcal{X}\to[0,1]\}$, the sequential square-root entropy of function class $\mathcal{F}$ on an $\mathcal{X}$-valued tree $\mathbf{x}$ (of depth n) at scale α is defined as

$$\mathcal{H}_{\mathsf{sq}}(\mathcal{F},\alpha,n,\mathbf{x}) = \log \mathcal{N}_{\mathsf{sq}}(\mathcal{F}\circ\mathbf{x},\alpha,n).$$

With the above definition of sequential square-root entropy, we have the following theorem:

Theorem 5.6. *For any $\mathcal{X}$ and function class $\mathcal{F}\in[0,1]^{\mathcal{X}}$, we have*

$$\mathcal{R}_n(\mathcal{F}) = \tilde{\mathcal{O}}\left(\sup_{\mathbf{x}} \left\{ 1 + \inf_{\gamma>\delta>0} \left\{ n\delta + \sqrt{n}\int_\delta^\gamma \sqrt{\mathcal{H}_{\mathsf{sq}}(\mathcal{F},\alpha,n,\mathbf{x})}d\alpha + \mathcal{H}_{\mathsf{sq}}(\mathcal{F},\gamma,n,\mathbf{x}) \right\} \right\} \right),$$

where the supremum is over all depth-n $\mathcal{X}$-valued trees $\mathbf{x}$.

This theorem has the following direct corollary, which provides explicit upper bounds on the minimax regret whenever the growth of sequential square-root entropy at scale α is bounded by $\tilde{\mathcal{O}}(\alpha^{-p})$ for some $p\ge 0$.

Corollary 5.7. *For any function class $\mathcal{F}\subseteq\{f:\mathcal{X}\to[0,1]\}$, suppose the sequential square-root entropy $\mathcal{H}_{\mathsf{sq}}(\mathcal{F},\alpha,n,\mathbf{x})$ at scale α satisfies $\sup_{\mathbf{x}}\mathcal{H}_{\mathsf{sq}}(\mathcal{F},\alpha,n,\mathbf{x}) = \tilde{\mathcal{O}}\left(\alpha^{-p}\right)$ for some $p\ge 0$. The minimax regret $\mathcal{R}_n(\mathcal{F})$ is upper bounded by*

$$\mathcal{R}_n(\mathcal{F}) = \begin{cases} \tilde{\mathcal{O}}\left(n^{\frac{p}{p+2}}\right) & \text{if } 0\le p\le 2, \\ \tilde{\mathcal{O}}\left(n^{\frac{p-1}{p}}\right) & \text{if } p>2. \end{cases}$$

The proofs of Theorem 5.6 and Theorem 5.7 are deferred to appendix D.3.1.

5.3.2 Comparison with Previous Upper Bounds

We compare our results to those of [17] and [18]. These two works use the sequential entropy $\mathcal{H}_\infty(\mathcal{F},\alpha,n,\mathbf{x})$, defined as

$$\mathcal{H}_\infty(\mathcal{F},\alpha,n,\mathbf{x}) = \log \mathcal{N}_\infty(\mathcal{F}\circ\mathbf{x},\alpha,n), \tag{5.13}$$

with $\mathcal{N}_\infty(\mathcal{F}\circ\mathbf{x},\alpha,n)$ being the size of smallest cover V such that

$$\max_{\mathbf{y}\in\{0,1\}^n} \sup_{f\in\mathcal{F}} \min_{\mathbf{v}\in V} \max_{t\in[n]} |f(x_t(\mathbf{y})) - v_t(\mathbf{y})| \le \alpha.$$

With proper choice of parameters, our results can recover the upper bounds in [17, Theorem 1 and Theorem 4]. This follows from the next result relating sequential square-root entropy and the sequential entropy defined above.

Proposition 5.8. *Suppose $\delta > 0$. For any $\mathcal{F} \subseteq \{f : \mathcal{X} \to [\delta, 1-\delta]\}$, $\alpha > 0$, and depth-n $\mathcal{X}$-valued tree $\mathbf{x}$,*

$$\mathcal{H}_{\mathsf{sq}}(\mathcal{F}, \alpha/\sqrt{\delta}, n, \mathbf{x}) \le \mathcal{H}_\infty(\mathcal{F}, \alpha, n, \mathbf{x}).$$

For nonparametric class $\mathcal{F}$ which satisfies $\sup_{\mathbf{x}} \mathcal{H}_\infty(\mathcal{F}, \alpha, n, \mathbf{x}) \asymp \alpha^{-q}$ for some $q > 0$, [18, Theorem 2] proves the following upper bound for the minimax regret:

$$\mathcal{R}_n(\mathcal{F}) = \mathcal{O}\left(n^{\frac{q}{q+1}}\right). \tag{5.14}$$

Theorem 5.7 recovers this result for $0 < q \le 1$, up to logarithmic factors, via the following proposition.

Proposition 5.9. *For any function class $\mathcal{F} \subseteq \{f : \mathcal{X} \to [0,1]\}$, $\alpha > 0$, and depth-n $\mathcal{X}$-valued tree $\mathbf{x}$, we have*

$$\mathcal{H}_{\mathsf{sq}}(\mathcal{F}, 2\alpha, n, \mathbf{x}) \le \mathcal{H}_\infty(\mathcal{F}, \alpha^2, n, \mathbf{x}).$$

The proofs of Theorem 5.8 and Theorem 5.9 are deferred to appendix D.3.2.

5.3.3 Lower Bound with Sequential Square-Root Entropy

In this section, we provide lower bounds on the minimax regret $\mathcal{R}_n(\mathcal{F})$ defined in eq. (5.12) via sequential square-root entropy.

Theorem 5.10. *Suppose function class $\mathcal{F} \subseteq [0,1]^{\mathcal{X}}$ satisfies*

$$\sup_{\mathbf{x}} \mathcal{H}_{\mathsf{sq}}(\mathcal{F}, \alpha, n, \mathbf{x}) = \tilde{\Omega}\left(\alpha^{-p}\right), \tag{5.15}$$

where the supremum is over all depth-n $\mathcal{X}$-valued trees. Then we have the following lower bound on the minimax regret:

$$\mathcal{R}_n(\mathcal{F}) = \tilde{\Omega}\left(n^{\frac{p}{p+2}}\right).$$

The proof of Theorem 5.10 rests on a definition of a new type of sequential scale-sensitive dimension of the function class $\mathcal{F}$. We further relate the sequential square-root entropy and the minimax regret to this dimension. The details are deferred to appendix D.4. Notice that according to Theorem 5.7 and Theorem 5.10, if the sequential square-root entropy of a function class $\mathcal{F}$ satisfies

$$\sup_{\mathbf{x}} \mathcal{H}_{\mathsf{sq}}(\mathcal{F}, \alpha, n, \mathbf{x}) = \tilde{\Theta}(\alpha^{-p}),$$

for some $0 \le p \le 2$, then we have the following tight characterization of the minimax regret up to log factors:

$$\mathcal{R}_n(\mathcal{F}) = \tilde{\Theta}\left(n^{\frac{p}{p+2}}\right).$$

However, when $p > 2$, there exists a gap between the lower bound in Theorem 5.10 and the upper bound in Theorem 5.7. Indeed, the following result shows that the upper bound $\tilde{\mathcal{O}}\left(n^{\frac{p-1}{p}}\right)$ is not improvable in general.

Theorem 5.11. *Suppose the function class $\mathcal{F} \subseteq \{f : \mathcal{X} \to [7/16, 9/16]\}$ satisfies*

$$\sup_{\mathbf{x}} \mathcal{H}_{\mathsf{sq}}(\mathcal{F}, \alpha, n, \mathbf{x}) = \tilde{\Omega}\left(\alpha^{-p}\right). \tag{5.16}$$

Then we have the following lower bound on the minimax regret

$$\mathcal{R}_n(\mathcal{F}) = \Omega\left(n^{\frac{p-1}{p}}\right).$$

Additionally, for any integer $p > 2$, there exists a class $\mathcal{F} \subseteq \{f : \mathcal{X} \to [7/16, 9/16]\}$ such that eq. (5.16) holds.

The proof of Theorem 5.11 is deferred to appendix D.4. Comparing Theorem 5.11 and Theorem 5.7, we see a dichotomy between the regime of $0 \leq p \leq 2$ and the regime of $p > 2$, where the rates of minimax regret $\mathcal{R}_n(\mathcal{F})$ have different behaviors. Such a dichotomy is analogous to the one for online regression [109] and to misspecified regression with i.i.d. data [113].

5.3.4 Examples

In this section, we provide several examples to illustrate Theorem 5.6 and Theorem 5.7. We consider the example of linear class (Hilbert ball class) and the class of one-dimensional Lipschitz Functions.

Hilbert Ball Consider $\mathcal{X} = B_2(1)$ to be the infinite dimensional unit ball, and function class $\mathcal{F} \subseteq [0,1]^{\mathcal{X}}$ defined as

$$\mathcal{F} = \left\{ f : f(x) = \frac{1 + \langle w, x \rangle}{2} \text{ for some } w \in B_2(1) \right\}. \tag{5.17}$$

This class is generally viewed as 'hard case' in existing literature. [17] proposed an ad hoc follow-the-regularized-leader (FTRL) algorithm with log-barrier regularizer, which achieves the optimal regret $\tilde{\mathcal{O}}(\sqrt{n})$. In terms of entropy characterizations, the same paper provided a loose upper bound of $\mathcal{O}(n^{3/4})$ and [18, 114] provided an upper bound of $\mathcal{O}(n^{2/3})$ using their versions of sequential entropies. The present work is the first to define an appropriate version of sequential entropy (and a corresponding regret bound) to derive a matching $\tilde{O}(\sqrt{n})$ regret bound.

We first truncate the function class $\mathcal{F}$ as follows:

$$\mathcal{F}_{1/n} = \left\{ f : f(x) = \frac{1 + \langle w, x \rangle}{2} \text{ for some } w \in B_2(1 - 1/n) \right\}. \tag{5.18}$$

The following lemma indicates that minimax regret of $\mathcal{F}_{1/n}$ is similar to that of $\mathcal{F}$.

Lemma 5.12. *For $\mathcal{F}$ and $\mathcal{F}_{1/n}$ defined in eq. (5.17) and eq. (5.18), the minimax regret satisfies*

$$\mathcal{R}_n(\mathcal{F}) \leq \mathcal{R}_n(\mathcal{F}_{1/n}) + 2.$$

The proof of Theorem 5.12 is deferred to appendix D.5. Equipped with this lemma, we only need to bound the sequential square-root entropy of function class $\mathcal{F}_{1/n}$.

Proposition 5.13. *It holds that*

$$\sup_{\mathbf{x}} \mathcal{H}_{\mathsf{sq}}(\mathcal{F}_{1/n}, \alpha, n, \mathbf{x}) = \mathcal{O}\left(\frac{\log n}{\alpha^2} \cdot \log\left(\frac{n}{\alpha}\right)\right).$$

The proof of Theorem 5.13 is deferred to appendix D.5. As a consequence, in view of Theorem 5.7, we conclude:

Corollary 5.14. *The minimax regret $\mathcal{R}_n(\mathcal{F})$ of Hilbert ball class $\mathcal{F}$ satisfies*

$$\mathcal{R}_n(\mathcal{F}) = \tilde{\mathcal{O}}\left(\sqrt{n}\right).$$

One-Dimensional Lipschitz Function Class We consider the example of one-dimensional Lipschitz function class, which has been studied in [18, 114, 115], among others. In this case, the context set is $\mathcal{X} = [0, 1]$, and the function class $\mathcal{F}$ is defined to be

$$\mathcal{F} = \{f : [0, 1] \to [0, 1], f \text{ is 1-Lipschitz}\}. \tag{5.19}$$

In [18], the minimax regret is shown to be upper bounded by $\tilde{\mathcal{O}}(\sqrt{n})$, which matches the lower bound. We now recover this rate using sequential square-root entropy. According to Theorem 5.9, the characterization $\sup_{\mathbf{x}} \mathcal{H}_\infty(\mathcal{F}, \alpha, n, \mathbf{x}) = \Theta(\alpha^{-1})$ for one-dimensional Lipschitz function class in [18, Theorem 3] directly indicates that for square-root entropy,

$$\sup_{\mathbf{x}} \mathcal{H}_{\mathsf{sq}}(\mathcal{F}, \alpha, n, \mathbf{x}) = \mathcal{O}\left(\alpha^{-2}\right).$$

Similarly to the Hilbert ball example, we conclude:

Corollary 5.15. *For $\mathcal{X} = [0, 1]$ and Lipschitz function class $\mathcal{F}$ defined in eq. (5.19),*

$$\mathcal{R}_n(\mathcal{F}) = \tilde{\mathcal{O}}(\sqrt{n}).$$

5.4 Proof Sketch of Theorem 5.2

In this section, we sketch the proof of Theorem 5.2. The detailed proof is deferred to appendix D.2. We break up the proof into the following key steps:

Transform minimax regret into the dual form. Our first step of analyzing the minimax regret $\mathcal{R}_n(\mathcal{Q})$ is to transform it into the dual form [17, 18]:

$$\mathcal{R}_n(\mathcal{Q}) = \sup_{\mathbf{p}} \mathbb{E}_{\mathbf{y}\sim\mathbf{p}} \mathcal{R}_n(\mathcal{Q}, \mathbf{p}, \mathbf{y}), \quad \text{where} \quad \mathcal{R}_n(\mathcal{Q}, \mathbf{p}, \mathbf{y}) := \sup_{\mathbf{q}\in\mathcal{Q}} \log\left(\frac{\mathbf{q}(\mathbf{y})}{\mathbf{p}(\mathbf{y})}\right).$$

where the first supremum is over all joint distributions $\mathbf{p} \in \Delta(\mathcal{Y}^n)$. In the following, we upper bound $\mathbb{E}_{\mathbf{y}\sim\mathbf{p}} \mathcal{R}_n(\mathcal{Q}, \mathbf{p}, \mathbf{y})$ for any $\mathbf{p}$.

Truncating the distributions. We first show that by truncating the distribution $\mathbf{p}$ and every $\mathbf{q} \in \mathcal{Q}$ so that all conditional probabilities $p_t(y_t \mid \mathbf{w})$ and $q_t(y_t \mid \mathbf{w})$ take values in the interval $[\delta, 1-\delta]$, for an appropriate δ, we pay an additional constant factor in regret.

Construct offset Rademacher processes. After truncation, we proceed to introduce the offset Rademacher processes through a symmetrization argument. To do this, we define $\zeta : (0, \infty) \to \mathbb{R}$ satisfying the following three properties: for some appropriately chosen positive number c,

(i) **Transformation of logarithm:** $\log x \le \zeta(x) - c \cdot \zeta(x)^2$ for any $\delta \le x \le 1/\delta$, where δ is the truncation scale. This property is inspired by the transformation in [18].

(ii) **Nonnegativity of divergence:** $\mathbb{E}_{y\sim p}\{-\zeta(f(y)/p(y)) - c \cdot \zeta(f(y)/p(y))^2\} \ge 0$ for any $f, p \in \Delta(\mathcal{Y})$. This property is inspired by the proof of [16] where nonnegativity of KL was used. Here we ensure that $-\zeta(x) - c \cdot \zeta(x)^2$ is convex with respect to x and takes on the value 0 at $x = 1$, inducing an f-divergence.

(iii) **Lipschitz property:** for any $p, q \in [0, \infty)$, $|\zeta(p) - \zeta(q)| \le 2|\sqrt{p} - \sqrt{q}|$.

The explicit form of ζ with these three conditions is given in appendix D.2.1. As a consequence, we obtain the following inequality after a sequential symmetrization argument:

$$\begin{aligned}
\mathbb{E}_{\mathbf{y}\sim\mathbf{p}}\mathcal{R}_n(\mathcal{Q}, \mathbf{p}, \mathbf{y}) &= \mathbb{E}_{\mathbf{w}\sim\mathbf{p}}\left[\sup_{\mathbf{q}\in\mathcal{Q}} \sum_{t=1}^{n} \left(\log q_t(w_t \mid \mathbf{w}) - \log p_t(w_t \mid \mathbf{w})\right)\right] \\
&\le \mathbb{E}_{\mathbf{w}\sim\mathbf{p}}\left[\sup_{\mathbf{q}\in\mathcal{Q}} \sum_{t=1}^{n} \left\{\zeta\left(\frac{q_t(w_t \mid \mathbf{w})}{p_t(w_t \mid \mathbf{w})}\right) - c \cdot \zeta\left(\frac{q_t(w_t \mid \mathbf{w})}{p_t(w_t \mid \mathbf{w})}\right)^2\right\}\right] \\
&\le \mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}} \sum_{t=1}^{n} \varepsilon_t \zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - c \cdot \zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)^2\right] \\
&\quad + \mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}} \sum_{t=1}^{n} (-\varepsilon_t) \zeta\left(\frac{q_t(z_t \mid \mathbf{w})}{p_t(z_t \mid \mathbf{w})}\right) - c \cdot \zeta\left(\frac{q_t(z_t \mid \mathbf{w})}{p_t(z_t \mid \mathbf{w})}\right)^2\right], \qquad (5.20)
\end{aligned}$$

where ε_t are Rademacher random variables, i.e. $\varepsilon_{1:n} \overset{\text{i.i.d.}}{\sim} \text{Unif}\{-1, 1\}$, and $\mathbf{y} = (y_{1:n}), \mathbf{z} = (z_{1:n}), \mathbf{w} = (z_{1:n})$ have a specific coupling: $y_t, z_t \overset{\text{i.i.d.}}{\sim} p_t(\cdot \mid w_{1:t-1})$, and $w_t = y_t$ if $\varepsilon_t = 1$ or $w_t = z_t$ if $\varepsilon_t = -1$. The scheme that w_t chooses y_t or z_t based on the value of ε_t is a variant of the "selectors" approach of [116].

Analysis through chaining technique Finally, to upper bound the right hand side of eq. (5.20), we adopt the chaining technique [17, 109, 117, 118]. We sketch the beginning of the argument. The first term (and, analogously, the second term) in (5.20) can be decomposed through a chain of N approximating representatives as

$$\mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}} \sum_{t=1}^{n} \varepsilon_t \left\{\zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - c \cdot \zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)^2\right\}\right] \qquad (5.21)$$

$$
\begin{aligned}
&\leq \mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\sum_{t=1}^{n}\varepsilon_t\left\{\zeta\left(\frac{q_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)-\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_1](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)\right\}\right]\\
&\quad+\sum_{i=1}^{N-1}\mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\sum_{t=1}^{n}\varepsilon_t\left\{\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_i](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)-\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_{i+1}](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)\right\}\right]\\
&\quad+\mathbb{E}\left[\sup_{\mathbf{v}\in\mathcal{V}(\alpha_N)\cup\{\mathbf{p}\}}\sum_{t=1}^{n}\left\{\varepsilon_t\zeta\left(\frac{v_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)-\frac{c}{4}\cdot\zeta\left(\frac{v_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)^2\right\}\right].
\end{aligned}
$$

where $\mathbf{v}[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_i]$ is an element of an α_i-cover $\mathcal{V}(\alpha_i)$ of $\mathcal{Q}$. The three terms in the above decomposition give rise to the corresponding three terms in the bound of Theorem 5.2: the approximation at the finest scale (term 1), the Dudley-style term (term 2), and the finite cover at the coarsest scale (term 3).

We will use Cauchy-Schwarz inequality to bound the first term. The second term is a form of sequential Rademacher process. The third term is an offset sequential Rademacher process. However, the key difficulty in dealing with the second and third terms is that the coefficients of the Rademacher random variables are not uniformly bounded by a constant, and directly applying prior techniques does not provide the desired upper bounds. To overcome this issue, we establish upper bounds on offset and non-offset sequential Rademacher processes with unbounded coefficients (Theorem D.1, Theorem D.2), heavily relying on the properties of the function ζ. Since the latter has $\sqrt{p_t}$-type terms in the denominator in the relevant range of behavior, the squared increments of the process, *under the expectation over* p_t, are controlled. The formal proofs are deferred to the appendix.

Chapter 6

A Gapped Scale-Sensitive Dimension and Lower Bounds for Offset Rademacher Complexity

This chapter is based on the paper "A Gapped Scale-Sensitive Dimension and Lower Bounds for Offset Rademacher Complexity" by Zeyu Jia, Yury Polyanskiy, and Alexander Rakhlin.

6.1 Introduction

The celebrated Vapnik-Chervonenkis dimension $\mathsf{vc}(\mathcal{F})$ of a binary-valued function class $\mathcal{F}$ and the scale-sensitive dimension $\mathsf{vc}(\mathcal{F}, \alpha)$ of a real-valued function class $\mathcal{F}$ are central notions in the study of empirical processes and convergence of statistical learning methods [19, 119, 120]. Sequential analogues of these notions—the Littlestone dimension $\mathsf{ldim}(\mathcal{F})$ and the sequential scale-sensitive dimension $\mathrm{fat}(\mathcal{F}, \alpha)$—have been shown to play an analogously central role in the study of uniform martingale laws and online prediction [3, 20, 121].

In this paper, we study "gapped" versions of $\mathsf{vc}(\mathcal{F}, \alpha)$ and $\mathrm{fat}(\mathcal{F}, \alpha)$. The modification yields a dimension that is no larger than the original one, yet can still be shown to control covering numbers in both sequential and non-sequential cases. More importantly, the new notion gives us a more precise control on the functions involved in "shattering" and thus yields non-vacuous lower bounds for offset Rademacher complexities for *any* uniformly bounded class—both in the classical and sequential cases—and, as a consequence, tighter lower bounds for online prediction problems, such as online regression or transductive learning. Our definition in the non-sequential case can also be seen as a modification of the Natarajan dimension [122, 123], and was, in fact, introduced in [21].

We first motivate the development in this paper on the simpler case of non-sequential data. We start by recalling the definition of the Vapnik-Chervonenkis dimension and its scale-sensitive version. Given a class $\mathcal{F} \subseteq \{f : \mathcal{X} \to \{-1, 1\}\}$ of binary-valued functions on some set $\mathcal{X}$, consider the projection $\mathcal{F}|_{x_1,\dots,x_d} = \{(f(x_1), \dots, f(x_d)) : f \in \mathcal{F}\}$ onto d elements $x_1, \dots, x_d \in \mathcal{X}$. The VC-dimension $\mathsf{vc}(\mathcal{F})$ is the largest d such that there exist $\{x_1, \dots, x_d\}$ with $\mathcal{F}|_{x_1,\dots,x_d} = \{-1, 1\}^d$. For a real-valued class $\mathcal{F} \subseteq \{f : \mathcal{X} \to [-1, 1]\}$ and a scale $\alpha \geq 0$,

the scale-sensitive dimension $\mathsf{vc}(\mathcal{F}, \alpha)$ is defined to be the largest d such that there exist $x_1, \ldots, x_d \in \mathcal{X}$ and $\mathbf{s} \in [-1, 1]^d$ with the following property: for any $\boldsymbol{\varepsilon} \in \{-1, 1\}^d$ there exists $f^{\boldsymbol{\varepsilon}} \in \mathcal{F}$ with $\varepsilon_i(f^{\boldsymbol{\varepsilon}}(x_i) - s_i) \geq \alpha/2$ for all $i \in \{1, \ldots, d\}$. We say that $\mathcal{F}$ *shatters* $x_1, \ldots, x_n$ at scale α if the aforementioned property holds. Shattering can also be visualized as a property that $\mathcal{F}|_{x_1,\ldots,x_d}$ "contains" a cube $\mathbf{s} + (\alpha/2)\{-1, +1\}^d$ at scale α with a center at $\mathbf{s}$, in the sense that for each direction $\boldsymbol{\varepsilon}$, there is a function $f^{\boldsymbol{\varepsilon}} \in \mathcal{F}$ whose projection onto the data lies in the quadrant outside the vertex $\mathbf{s} + (\alpha/2)\boldsymbol{\varepsilon}$.

Note that the definition of shattering does not tell us whether $f^{\boldsymbol{\varepsilon}}$ is close to the corresponding vertex, i.e.

$$f^{\varepsilon}(x_i) \approx s_i + \varepsilon_i \alpha/2 \tag{6.1}$$

for every $i \in \{1, \ldots, d\}$. We can see that such a requirement (in the non-sequential case described here) can be satisfied under the assumption of convexity of $\mathcal{F}$. Unfortunately, for the sequential case described in section 6.3, the nature of the restriction that (6.1) imposes on $\mathcal{F}$ is less clear.

We finish this introductory section by motivating the utility of the additional requirement (6.1). Once again, we only discuss the non-sequential case here. Consider the following lower bound on Rademacher averages of a class $\mathcal{F}$ in terms of $\mathsf{vc}(\mathcal{F}, \alpha)$:

$$\mathbb{E}_\varepsilon \sup_{f \in \mathcal{F}} \sum_{i=1}^{d} \varepsilon_i f(x_i) \geq \mathbb{E}_\varepsilon \sum_{i=1}^{d} \varepsilon_i (f^{\boldsymbol{\varepsilon}}(x_i) - s_i) \geq n\alpha/2 \tag{6.2}$$

where $d = \mathsf{vc}(\mathcal{F}, \alpha)$ and $\{x_1, \ldots, x_d\}$ is a set whose existence is guaranteed by the definition. Here the expectation is with respect to independent Rademacher random variables ε_i (taking values ± 1 with equal probability). The lower bound argument can be extended to $d > \mathsf{vc}(\mathcal{F}, \alpha)$ by considering repeated blocks of points and appealing to Khintchine's inequality. Such an argument leads to lower bounds of order $\Omega(\alpha\sqrt{\mathsf{vc}(\mathcal{F}, \alpha)n})$, implying that the scale-sensitive dimension is an inherent barrier for Rademacher averages to be small, and, as a consequence, a barrier for certain learning problems. Indeed, the notion of Radmacher averages in (6.2) is known to be a key object in the study of prediction with i.id. data and "non-curved" losses. On the other hand, for loss functions such as square loss, it is the *offset Rademacher averages*—or closely related local Rademacher averages [124]—that in many situations correctly quantify the rates of convergence. The (non-sequential) offset Rademacher averages are defined as:[1]

$$\mathbb{E}_\varepsilon \sup_{f \in \mathcal{F} - \mathcal{F}} \sum_{i=1}^{n} \varepsilon_i f(x_i) - c \cdot f(x_i)^2. \tag{6.3}$$

Unfortunately, the lack of control on the magnitudes of departures of $f^{\boldsymbol{\varepsilon}}(x_i)$ from $s_i \pm \alpha/2$ prevents us from obtaining sufficiently strong lower bounds when considering a shattered set, as the negative term in (6.3) may render the lower bound vacuous. This question motivates the study of gapped scale-sensitive dimensions, as presented in the next sections for both the non-sequential and sequential cases.

[1] We replaced $\mathcal{F}$ with $\mathcal{F} - \mathcal{F}$ to simplify the centering issues.

We briefly mention that a number of other versions of combinatorial dimensions have been proposed over the last few decades (see [125–127] and references therein). To the best of our knowledge, these notions are different from those proposed in the present paper, and do not immediately imply the lower bounds we seek.

Organization We start the technical part of the paper with the non-sequential version of the gapped dimension in section 6.2. We first introduce the gapped dimension for *integer-valued* classes in section 6.2.1 and state a version of the Sauer-Shelah-Vapnik-Chervonenkis lemma for this combinatorial definition due to [21],[2] yielding control of covering numbers. We then extend the definition to the real-valued case in section 6.2.2 and prove that this scale-sensitive dimension controls covering numbers, similarly to the development in [128] for the standard definition. We then prove that offset Rademacher averages for any uniformly bounded class are lower bounded according to the behavior of the gapped scale-sensitive dimension of this class (section 6.2.4), and present the ensuing lower bound for the problem of online transductive regression. section 6.3 mirrors the development in section 6.2 for the sequential case, with an application to online regression.

Notation We denote $x_{1:d} = \{x_1, \ldots, x_d\}$ and $[M] = \{1, \ldots, M\}$ for integer $M > 1$.For functions $A(\alpha, n)$ and $B(\alpha, n)$, we use $A(\alpha, n) = \Omega(B(\alpha, n))$ or $B(\alpha, n) = \mathcal{O}(A(\alpha, n))$ to denote $A(\alpha, n) \geq c \cdot B(\alpha, n)$ for any $\alpha > 0$ and positive integer n, with some fixed positive constant c. We use $A(\alpha, n) = \tilde{\Omega}(B(\alpha, n))$ or $B(\alpha, n) = \tilde{\mathcal{O}}(A(\alpha, n))$ to denote $A(\alpha, n) \geq c \cdot B(\alpha, n)/\log^r(n/\alpha)$ holds for any $\alpha > 0$ and positive integer n, with some fixed positive constants c, r.

6.2 Non-Sequential Gapped Dimensions

6.2.1 Integer-Valued Functions

Suppose M is a positive integer. Let $\mathfrak{c} : [M] \times [M] \to \mathbb{R}_+ \cup \{0\}$ be a distance metric. Consider the following combinatorial parameter [21]:

Definition 6.1 ((Non-Sequential) Gapped Dimension)**.** *Let $\mathcal{F} \subseteq \{f : \mathcal{X} \to [M]\}$ be a function class and fix $\alpha \geq 0$. We sat that $\mathcal{F}$ shatters the set $\{x_1, \ldots, x_d\} \subseteq \mathcal{X}$ at scale α if there exists $s_1 = (s_1[-1], s_1[1]), s_2 = (s_2[-1], s_2[1]), \ldots, s_d = (s_d[-1], s_d[1]) \in [M] \times [M]$ with the following properties:*

1. *For any $t \in [d]$, $\mathfrak{c}(s_t[1], s_t[-1]) \geq \alpha$;*
2. *For any $\boldsymbol{\varepsilon} \in \{-1, 1\}^d$, there exists $f^{\boldsymbol{\varepsilon}} \in \mathcal{F}$ such that $f^{\boldsymbol{\varepsilon}}(x_t) = s_t[\varepsilon_t]$ for any $t \in [d]$.*

We define the (non-sequential) gapped scale-sensitive dimension $d_{\mathfrak{c}}(\mathcal{F}, \alpha)$ (or $d(\mathcal{F}, \alpha)$ when $\mathfrak{c}$ is clear from context) of $\mathcal{F}$ as the largest d such that there exists $\{x_1, \ldots, x_d\}$ which is shattered by $\mathcal{F}$.

[2] After this paper was completed, we were informed by Peter Bartlett that the definition of the "gapped" dimension and Lemma 6.3 are contained in [21].

The gapped dimension in Theorem 6.1 was introduced in [21]. The definition is similar to that of the Natarajan dimension [122, 123] for multi-class learning, with the important difference that the two "labels" (denoted by the choice $s_t[1]$ and $s_t[-1]$) are α-separated (hence the name *gapped* dimension); unlike multi-class problems where the the labels are treated as a categorical variable, in our case they are ordinal.

We now recall the standard definition of a covering number, which we state here with respect to the distance $\mathfrak{c}$ on each coordinate.

Definition 6.2 ((Non-Sequential) Covering Number for Integer-Valued Functions)**.** *Fix $x_1, \dots, x_n \in \mathcal{X}$. We say that the set $\mathcal{V} = \{\mathbf{v} = (v_1, \dots, v_n) \in [M]^n\}$, is a (non-sequential) cover of $\mathcal{F}$ on $x_{1:n}$ at scale α if for any function $f \in \mathcal{F}$, there exists $\mathbf{v} \in \mathcal{V}$ such that*

$$\max_{t \in [n]} \mathfrak{c}(f(x_t), v_t) \leq \alpha.$$

We use $\mathcal{N}_{\mathfrak{c},\infty}(\mathcal{F}, x_{1:n}, \alpha)$ (or $\mathcal{N}_\infty(\mathcal{F}, x_{1:n}, \alpha)$ when $\mathfrak{c}$ is clear from context) to denote the size of the smallest non-sequential cover of $\mathcal{F}$ on $x_{1:n}$ at scale α.

The following lemma upper bounds the covering number of a integer-valued class by the gapped dimension.

Lemma 6.3 ([21])**.** *Let $\mathcal{F} \subseteq \{f : \mathcal{X} \to [M]\}$ and $\{x_1, \dots, x_n\} \subseteq \mathcal{X}$. Then we have*

$$\log \mathcal{N}_\infty(\mathcal{F}, x_{1:n}, \alpha) \leq 16 d(\mathcal{F}, \alpha) \log^2(enM)$$

The proof of Theorem 6.3 is deferred to appendix E.1.1. This result is similar to that of [128], which provides an upper bound on the covering number in terms of the (original) scale-sensitive dimension. Since the gapped dimension can be smaller than the original definition, Theorem 6.3 is an improvement over the corresponding result in [128].

6.2.2 Real-Valued Functions

We now turn our attention to the case of real-valued function classes. Let $\mathfrak{c} : [-1,1] \times [-1,1] \to \mathbb{R}_+ \cup \{0\}$ be a distance metric; for example, we may choose $\mathfrak{c}(a,b) = |a-b|$. We define the following notion of non-sequential gapped scale-sensitive dimension:

Definition 6.4 ((Non-Sequential) Gapped Scale-Sensitive Dimension)**.** *Let $\mathcal{F} \subseteq \{f : \mathcal{X} \to [-1,1]\}$ and fix $\alpha, \beta > 0$. We say that $\mathcal{F}$ shatters $\{x_1, \dots, x_d\} \subseteq \mathcal{X}$ at scale (α, β) if there exist $s_1 = (s_1[-1], s_1[1]), s_2 = (s_2[-1], s_2[1]), \dots, s_d = (s_d[-1], s_d[1]) \in [-1,1] \times [-1,1]$ with the following properties:*

1. *For any $t \in [d]$, $\mathfrak{c}(s_t[1], s_t[-1]) \geq \alpha$;*
2. *For any $\boldsymbol{\varepsilon} \in \{-1,1\}^d$, there exists $f^{\boldsymbol{\varepsilon}} \in \mathcal{F}$ such that $\mathfrak{c}(f^{\boldsymbol{\varepsilon}}(x_t), s_t[\varepsilon_t]) \leq \beta$ for any $t \in [d]$.*

We define the (non-sequential) gapped scale-sensitive dimension $d(\mathcal{F}, \alpha, \beta)$ of $\mathcal{F}$ as the largest d such that there exist $x_{1:d} \in \mathcal{X}$ shattered by $\mathcal{F}$ at scale (α, β).

To illustrate the geometric requirement of the gapped scale-sensitive dimension, we refer to fig. 6.1. The standard definition of scale-sensitive dimension asks for a cube of side length α to be inscribed in the set, where "inscribed" means that there is an element of the set $\mathcal{F}|_{x_1,\ldots,x_d}$ in any of the 2^d quadrants, for each vertex of the hypercube (formally, for any $\boldsymbol{\varepsilon} \in \{-1,1\}^d$ there exists $f^{\boldsymbol{\varepsilon}} \in \mathcal{F}$ with $\varepsilon_i(f^{\boldsymbol{\varepsilon}}(x_i) - s_i) \geq \alpha/2$ for all $i \in \{1,\ldots,d\}$). In contrast, the gapped scale-sensitive dimension asks for a hypercube of side-length at least α to be inscribed in the set, where "inscribed" means that each of the 2^d vertices are β-close coordinate-wise to some element of $\mathcal{F}|_{x_1,\ldots,x_d}$. While it is immediate that $d(\mathcal{F},\alpha,\beta) \leq \mathsf{vc}(\mathcal{F},\alpha-2\beta)$ for $\beta < \alpha/2$, we show that this gap cannot be too large. We prove this fact by first establishing a relation between covering numbers and the gapped dimension.

Figure 6.1: Grey area depicts a set $\mathcal{F}|_{x_1,\ldots,x_d}$. The hypercube on the left is "inscribed" in the classical sense, while the hyperrectangle on the right is "inscribed" according to the proposed definition.

Definition 6.5 ((Non-Sequential) Covering Number for Real-Valued Functions)**.** *We say that a set $\mathcal{V} \subseteq [-1,1]^n$ is a cover of $\mathcal{F}$ on $\{x_1,\ldots,x_n\} \subseteq \mathcal{X}$ at scale α if for any function $f \in \mathcal{F}$, there exists $v = (v_1,\ldots,v_n) \in \mathcal{V}$ such that*

$$\max_{t\in[n]} \mathfrak{c}(f(x_t), v_t) \leq \alpha.$$

We use $\mathcal{N}_{\mathfrak{c},\infty}(\mathcal{F}, x_{1:n}, \alpha)$ (or $\mathcal{N}_\infty(\mathcal{F}, x_{1:n}, \alpha)$ when $\mathfrak{c}$ is clear from context) to denote the size of smallest sequential cover of $\mathcal{F}$ on $x_{1:n}$ at scale α.

Proposition 6.6. *For $\alpha, \beta > 0$, suppose there exists a positive integer M and M real numbers $-1 \leq u_1 < u_2 < \ldots < u_M \leq 1$ such that for any $u \in [-1,1]$, there exists some $i \in [M]$ such that $\mathfrak{c}(u, u_i) \leq \beta$. Then for any function class $\mathcal{F} \subseteq \{f : \mathcal{X} \to [-1,1]\}$ and $\{x_1,\ldots,x_n\} \subseteq \mathcal{X}$,*

$$\log \mathcal{N}_\infty(\mathcal{F}, x_{1:n}, \alpha+\beta) \leq 16 d(\mathcal{F},\alpha,\beta) \log^2(enM).$$

When $\mathfrak{c}(a,b) = |a-b|$, a feasible value of M is $\lfloor 2/\beta \rfloor$, which implies that

$$\log \mathcal{N}_\infty(\mathcal{F}, x_{1:n}, \alpha+\beta) \leq 16 d(\mathcal{F},\alpha,\beta) \log^2\left(\frac{2en}{\beta}\right).$$

A version of this result already appears as Lemma 4.3 in [21], and we present the proof in appendix E.1.2 for completeness. We remark that the logarithmic dependence on n is unavoidable. The question of reducing the power from 2 to 1 (with the classical definition of scale-sensitive dimension) has been studied in [129]; we did not attempt to answer this question for the gapped dimension.

6.2.3 Comparison to Scale-Sensitive Dimension

In this section, we compare the classic scale-sensitive dimension for real-valued function class and the non-sequential scale-sensitive dimension defined in Theorem 6.4. We first recall the definition of the scale-sensitive dimension [19, 120].

Definition 6.7. *Given a function class $\mathcal{F} \subseteq \{f : \mathcal{X} \to [-1,1]\}$, we say that $\mathcal{F}$ shatters $\{x_1, \dots, x_d\} \subseteq \mathcal{X}$ at scale $\alpha > 0$ if there exists witnesses $s_1, \dots, s_d \in [-1,1]$ such that for any $\boldsymbol{\varepsilon} = (\varepsilon_{1:d}) \in \{-1,1\}^d$, there exists $f^{\boldsymbol{\varepsilon}} \in \mathcal{F}$ such that for all $t \in [n]$, $\varepsilon_t \cdot (f^{\boldsymbol{\varepsilon}}(x_t) - s_t) \geq \alpha/2$. The scale-sensitive dimension $\mathsf{vc}(\mathcal{F}, \alpha)$ is defined to be the largest d such that there exists $\{x_1, \dots, x_d\} \subseteq \mathcal{X}$ shattered by $\mathcal{F}$ at scale α.*

We have the following relations between the non-sequential scale-sensitive dimension $\mathsf{vc}(\mathcal{F}, \alpha)$ and the gapped dimension $d(\mathcal{F}, \alpha)$ with $\mathfrak{c}(a,b) = |a-b|$.

Proposition 6.8. *Given a function class $\mathcal{F} \subseteq \{\mathcal{X} \to [-1,1]\}$, for any $0 < 2\beta < \alpha$, we have*

$$d(\mathcal{F}, \alpha, \beta) \leq \mathsf{vc}(\mathcal{F}, \alpha - 2\beta).$$

Proposition 6.9. *Given a function class $\mathcal{F} \subseteq \{\mathcal{X} \to [-1,1]\}$, for any $\alpha, \beta > 0$, we have*

$$\mathsf{vc}(\mathcal{F}, 3(\alpha+\beta)) \leq 288 d(\mathcal{F}, \alpha, \beta) \cdot \log^2\left(\frac{384 d(\mathcal{F}, \alpha, \beta)}{\beta}\right).$$

Furthermore, if $\mathcal{F}$ is convex, then for any $\alpha, \beta > 0$ with $2\beta < \alpha$,

$$\mathsf{vc}(\mathcal{F}, \alpha) \leq d(\mathcal{F}, \alpha, \beta).$$

The proofs of Theorem 6.8 and Theorem 6.9 are almost identical to the proof of [21, Theorem 4.2] and [21, Theorem 4.3]. We defer both proofs to appendix E.1.3. We remark that Theorem 6.9 is proved *using* Theorem 6.6. We are not aware of a direct proof of Theorem 6.9, which would, of course, allow us to re-use existing estimates of covering numbers via non-gapped dimensions.

There is an extra squared logarithmic factor in Theorem 6.9 compared to Theorem 6.8. The following proposition indicates that at least one logarithmic factor in Theorem 6.9 is necessary.

Proposition 6.10. *There exists a class of contexts $\mathcal{X}$, and a class of functions $\mathcal{F}$, such that for any $0 < 2\beta < \alpha < 1$, we have*

$$d(\mathcal{F}, \alpha, \beta) = 1, \quad \text{and} \quad \mathsf{vc}(\mathcal{F}, \alpha) \geq \left\lfloor \log_2\left(\frac{1}{\alpha}\right) \right\rfloor.$$

The proof of Theorem 6.10 is deferred to appendix E.1.3.

6.2.4 Lower Bounds for Offset Rademacher Complexity and Transductive Regression

We fix some positive constant $C > 0$ and set of contexts $\mathcal{X}$. For function class $\mathcal{F} \subseteq \{\mathcal{X} \to [-1,1]\}$, $x_{1:n} \in \mathcal{X}^n$, $\mu_{1:n} \in [-1,1]^n$ and $\boldsymbol{\varepsilon} \in \{-1,1\}^n$, we define the supremum of the offset Rademacher process as

$$\mathcal{R}_n(\mathcal{F}, \mu_{1:n}, x_{1:n}, \boldsymbol{\varepsilon}) = \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} C \cdot \varepsilon_t (f(x_t) - \mu_t) - (f(x_t) - \mu_t)^2. \tag{6.4}$$

The expectation $\mathbb{E}[\mathcal{R}_n(\mathcal{F}, \mu_{1:n}, x_{1:n}, \boldsymbol{\varepsilon})]$ with respect to i.i.d. Rademacher random variables $\boldsymbol{\varepsilon}$ is termed the offset Rademacher complexity, and it is known to quantify the performance of Least Squares and related methods in regression.

Theorem 6.11. *Suppose $C \geq 2$. Let $\mathfrak{c}(a,b) = |a-b|$. If there exists $p \geq 0$ such that $d(\mathcal{F}, \alpha, \alpha/(20nC)) = \tilde{\Omega}\left(\alpha^{-p}\right)$ holds for any $\alpha > 0$ and positive integer n, then for any positive integer n,*

$$\sup_{\mu_{1:n}, x_{1:n}} \mathbb{E}[\mathcal{R}_n(\mathcal{F}, \mu_{1:n}, x_{1:n}, \boldsymbol{\varepsilon})] = \tilde{\Omega}\left(n^{\frac{p}{p+2}}\right), \tag{6.5}$$

where the expectation is with respect to i.i.d. Rademacher random variables $\boldsymbol{\varepsilon} = (\varepsilon_{1:n}) \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(\{-1,1\})$.

The proof of Theorem 6.11 is deferred to appendix E.1.5.

Application: Transductive Regression In an n-round online transductive prediction problem, the forecaster is given a function class $\mathcal{F} \subseteq \{\mathcal{X} \to [-1,1]\}$ and contexts $\{x_1, \ldots, x_n\} \subseteq \mathcal{X}$ before the interaction starts. On each round $t = 1, \ldots, n$, the forecaster makes prediction $\widehat{y}_t \in [-2,2]$. Nature (or, adversary) then reveals the label $y_t \in [-2,2]$. The forecaster's objective is to minimize the regret with respect to the performance of the best forecaster within the class $\mathcal{F}$ in hindsight. Considering the square loss, we can write the optimal regret in this game as the following minimax value:

$$\mathcal{V}_n(\mathcal{F}) := \sup_{x_{1:n} \in \mathcal{X}^n} \left\{ \inf_{\widehat{y}_t} \sup_{y_t} \right\}_{t=1}^{n} \left[\sum_{t=1}^{n} (\widehat{y}_t - y_t)^2 - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} (f(x_t) - y_t)^2 \right], \tag{6.6}$$

where $\{\cdot\}_{t=1}^{n}$ denotes repeated application of the operators. We remark that a typical assumption in transductive regression is that the set $\{x_1, \ldots, x_n\}$ is known, but not the order of appearance of its elements [130]. Such a setting is more difficult than the minimax game in eq. (6.6), as the forecaster has less information about contexts throughout the game. Hence, the lower bound for Theorem 6.12 also applies to this more widely studied setting.

The following theorem, a transductive analogue of [109], lower bounds the regret in eq. (6.6) by the (non-sequential) offset Rademacher process (6.4).

Lemma 6.12. *For any function class $\mathcal{F} \subseteq \{\mathcal{X} \to [-1,1]\}$, we have the following lower bound on the transductive regression objective:*

$$\mathcal{V}_n(\mathcal{F}) \geq \sup_{x_{1:n} \in \mathcal{X}} \sup_{\mu_{1:n} \in [-1,1]} \mathbb{E}_{\varepsilon_t} \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^{n} 2\varepsilon_t (f(x_t) - \mu_t) - (f(x_t) - \mu_t)^2 \right],$$

where $\varepsilon_{1:n} \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(\{-1,1\})$.

We defer the proof of Theorem 6.12 to appendix E.1.5. Theorem 6.12 together with Theorem 6.11 and Theorem 6.6 implies the following lower bound for transductive online regression in terms of the non-sequential covering number defined in Theorem 6.5. Notably, the lower bound holds for any $\mathcal{F} \subseteq \{f : \mathcal{X} \to [-1,1]\}$. On the downside, the $[-2,2]$ range of predictions and outcomes does not correspond to the $[-1,1]$ range of the functions in $\mathcal{F}$, making the problem misspecified in an atypical manner; this issue also appears in [109, Lemma 4], and we are not aware of other general proof techniques that circumvent this.

Corollary 6.13. *Fix a function class $\mathcal{F} \subseteq \{\mathcal{X} \to [-1,1]\}$ over context space $\mathcal{X}$. Suppose there exists real number $p \geq 0$ such that the ℓ_∞ covering number under distance $\mathfrak{c}(a,b) = |a-b|$ satisfies $\sup_{x_{1:n}} \log \mathcal{N}_\infty(\mathcal{F}, x_{1:n}, \alpha) = \tilde{\Omega}\left(\alpha^{-p}\right)$ for any $\alpha > 0$ and positive integer n. Then*

$$\mathcal{V}_n(\mathcal{F}) = \tilde{\Omega}\left(n^{\frac{p}{p+2}}\right).$$

In a manner similar to [109, Lemma 9], we can also show that $\mathcal{V}_n(\mathcal{F})$ is lower bounded by $n^{\frac{p-1}{p}}$ up to constants, under the same setting of Theorem 6.13. Hence, we obtain a complete picture of lower bounds for transductive regression in terms of the non-sequential covering numbers, modulo the misspecification issue discussed above.

6.3 Sequential Gapped Dimensions

We recall that the sequential versions of aforementioned complexities are defined in terms of trees (or, equivalently, predictable processes). An $\mathcal{X}$-valued tree $\mathbf{x}$ of depth n is a sequence of maps $x_1, \dots, x_n$, with $x_t : \{-1,1\}^{t-1} \to \mathcal{X}$ and $x_1 \in \mathcal{X}$ a constant. We refer to $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n) \in \{-1,1\}^n$ as a path of length n. Slightly abusing the notation, we write $x_t(\boldsymbol{\varepsilon})$ for $x_t(\boldsymbol{\varepsilon}_{1:t-1})$, where, henceforth, $\boldsymbol{\varepsilon}_{1:t-1} = (\varepsilon_1, \dots, \varepsilon_{t-1})$. It is convenient to think of $\mathbf{x}$ as a full binary tree labeled by elements of $\mathcal{X}$. Similarly, we can define a tree labeled by $\mathbb{R}$ or any other set.

We recall that constant-level trees (those with $x_t(\boldsymbol{\varepsilon}) = x_t \in \mathcal{X}$ for all $t \in [n]$ and $\boldsymbol{\varepsilon}$) correspond to a "tuple" of points $(x_1, \dots, x_n)$. Likewise, sequential generalizations—such as sequential cover—reduce to the classical notions when the trees are constant-level [131]. However, we remark that the relations between the various sequential quantities do not imply the analogous relations in the non-sequential case. For this reason, in this paper we developed both sequential and non-sequential results separately.

6.3.1 Integer-Valued Functions

As before, let $\mathfrak{c} : [M] \times [M] \to \mathbb{R}_+ \cup \{0\}$ be a distance metric. For a class of $[M]$-valued functions, we define the following notion of a sequential dimension:

Definition 6.14 (Sequential Gapped Dimension)**.** *Let $\mathcal{F} \subseteq \{f : \mathcal{X} \to [M]\}$ be a function class and fix $\alpha \geq 0$. We say that $\mathcal{F}$ shatters an $\mathcal{X}$-valued tree $\mathbf{x}$ of depth d at scale α if there exists an $[M] \times [M]$-valued tree $\mathbf{s}$ of depth d with the following properties:*

1. *For any* $\boldsymbol{\varepsilon} \in \{-1,1\}^d$, $s_t(\boldsymbol{\varepsilon}) = (s_t(\boldsymbol{\varepsilon})[-1], s_t(\boldsymbol{\varepsilon})[1])$ *with* $\mathfrak{c}(s_t(\boldsymbol{\varepsilon})[1], s_t(\boldsymbol{\varepsilon})[-1]) \geq \alpha$.

2. *For any* $\boldsymbol{\varepsilon} \in \{-1,1\}^d$, *there exists* $f^{\boldsymbol{\varepsilon}} \in \mathcal{F}$ *such that* $f^{\boldsymbol{\varepsilon}}(x_t(\boldsymbol{\varepsilon})) = s_t(\boldsymbol{\varepsilon})[\varepsilon_t]$ *for all* $t \in [d]$.

We define the gapped sequential scale-sensitive dimension $d^{\mathsf{seq}}(\mathcal{F}, \alpha)$ *of* $\mathcal{F}$ *as the largest* d *such that there exists an* α*-shattered tree* $\mathbf{x}$ *of depth* d.

Definition 6.15 (Sequential Covering Number for Integer-Valued Functions)**.** *Given an* $\mathcal{X}$*-valued tree* $\mathbf{x}$ *of depth* n*, we say that a set* $\mathcal{V}$ *of* $[M]$*-valued trees of depth* n *is a sequential cover of* $\mathcal{F}$ *on* $\mathbf{x}$ *at scale* α*, if for any function* $f \in \mathcal{F}$ *and* $\boldsymbol{\varepsilon} \in \{-1,1\}^n$*, there exists* $\mathbf{v} \in \mathcal{V}$ *such that for any* $t \in [n]$, $\mathfrak{c}(f(x_t(\boldsymbol{\varepsilon})), v_t(\boldsymbol{\varepsilon})) \leq \alpha$.

We use $\mathcal{N}^{\mathsf{seq}}_{\mathfrak{c},\infty}(\mathcal{F}, \mathbf{x}, \alpha)$ *(or* $\mathcal{N}^{\mathsf{seq}}_{\infty}(\mathcal{F}, \mathbf{x}, \alpha)$ *when* $\mathfrak{c}$ *is clear from context) to denote the size of smallest sequential cover of* $\mathcal{F}$ *on* $\mathbf{x}$ *at scale* α.

The following lemma upper bounds the sequential covering number for an integer-valued function class in terms of the sequential gapped dimension of the class, an analogue of Theorem 6.3.

Lemma 6.16. *Let* $\mathcal{F} \subseteq \{f : \mathcal{X} \to [M]\}$ *and let* $\mathbf{x}$ *be an* $\mathcal{X}$*-valued tree of depth* n*. We have*

$$\log \mathcal{N}^{\mathsf{seq}}_{\infty}(\mathcal{F}, \mathbf{x}, \alpha) \leq d^{\mathsf{seq}}(\mathcal{F}, \alpha) \log(enM)$$

The proof of Theorem 6.16 is deferred to appendix E.2.

6.3.2 Real-Valued Functions

Let $\mathfrak{c} : [-1,1] \times [-1,1] \to \mathbb{R}_+ \cup \{0\}$ be a distance metric, as in section 6.2.2. We define the following notion of complexity of $\mathcal{F}$, with respect to this metric:

Definition 6.17 (Sequential Gapped Scale-sensitive Dimension)**.** *Let* $\mathcal{F} \subseteq \{f : \mathcal{X} \to [-1,1]\}$ *be a function class and fix* $\alpha, \beta > 0$*. We say that* $\mathcal{F}$ *shatters an* $\mathcal{X}$*-valued tree* $\mathbf{x}$ *of depth* d *at scale* (α, β) *if there exists a* $([-1,1] \times [-1,1])$*-valued tree* $\mathbf{s}$ *of depth* d *with the following properties:*

1. *For any* $\boldsymbol{\varepsilon} \in \{-1,1\}^d$, $s_t(\boldsymbol{\varepsilon}) = (s_t(\boldsymbol{\varepsilon})[-1], s_t(\boldsymbol{\varepsilon})[1])$ *with* $\mathfrak{c}(s_t(\boldsymbol{\varepsilon})[1], s_t(\boldsymbol{\varepsilon})[-1]) \geq \alpha$.

2. *For any* $\boldsymbol{\varepsilon} \in \{-1,1\}^d$, *there exists* $f^{\boldsymbol{\varepsilon}} \in \mathcal{F}$ *such that* $\mathfrak{c}(f^{\boldsymbol{\varepsilon}}(x_t(\boldsymbol{\varepsilon})), s_t(\boldsymbol{\varepsilon})[\varepsilon_t]) \leq \beta$.

We define the gapped sequential scale-sensitive dimension $d^{\mathsf{seq}}(\mathcal{F}, \alpha, \beta)$ *of* $\mathcal{F}$ *as the largest* d *such that there exists an* (α, β)*-shattered tree* $\mathbf{x}$ *of depth* d.

We now define sequential covering numbers and prove that their growth is controlled by the behavior of d^{seq}.

Definition 6.18 (Sequential Covering Number for Real-Valued Functions)**.** *Given an* $\mathcal{X}$*-valued tree* $\mathbf{x}$ *of depth* n*, we say that a set* $\mathcal{V}$ *of* $\mathbb{R}$*-valued trees of depth* n *is a sequential cover of* $\mathcal{F}$ *on* $\mathbf{x}$ *at scale* α*, if for any function* $f \in \mathcal{F}$ *and* $\boldsymbol{\varepsilon} \in \{-1,1\}^n$*, there exists* $\mathbf{v} \in \mathcal{V}$ *such that for any* $t \in [n]$, $\mathfrak{c}(f(x_t(\boldsymbol{\varepsilon})), v_t(\boldsymbol{\varepsilon})) \leq \alpha$.

We use $\mathcal{N}^{\mathsf{seq}}_{\mathfrak{c},\infty}(\mathcal{F}, \mathbf{x}, \alpha)$ *(or* $\mathcal{N}^{\mathsf{seq}}_{\infty}(\mathcal{F}, \mathbf{x}, \alpha)$ *when* $\mathfrak{c}$ *is clear from context) to denote the size of smallest sequential cover of* $\mathcal{F}$ *on* $\mathbf{x}$ *at scale* α.

Proposition 6.19. *For given real numbers $\alpha, \beta > 0$, suppose there exists a positive integer M and M real numbers $-1 \le u_1 < u_2 < \ldots < u_M \le 1$ such that for any $u \in [-1,1]$, there exists some $i \in [M]$ such that $\mathfrak{c}(u, u_i) \le \beta$. Then for any function class $\mathcal{F} \subseteq \{f : \mathcal{X} \to [-1,1]\}$, positive integer n and $\alpha, \beta > 0$, for any depth-n $\mathcal{X}$-valued tree $\mathbf{x}$ we have*

$$\log \mathcal{N}_\infty^{\mathsf{seq}}(\mathcal{F}, \mathbf{x}, \alpha + \beta) \le d^{\mathsf{seq}}(\mathcal{F}, \alpha, \beta) \log(enM).$$

When $\mathfrak{c}(a,b) = |a-b|$, a feasible value of M is $\lfloor 2/\beta \rfloor$, which implies that

$$\log \mathcal{N}_\infty^{\mathsf{seq}}(\mathcal{F}, x_{1:n}, \alpha + \beta) \le d^{\mathsf{seq}}(\mathcal{F}, \alpha, \beta) \log\left(\frac{2en}{\beta}\right).$$

The proof of Theorem 6.19 is deferred to appendix E.2. The structure of the proof follows that in [20]; see also [132] for a different sequential generalization of Natarajan and Steele dimensions.

As in the previous works, Theorem 6.19 relies (via discretization) on covering numbers and combinatorial dimensions for integer-valued function classes, as developed in section 6.3.1.

6.3.3 Comparison to Sequential Scale-Sensitive Dimension

We recall the definition of sequential scale-sensitive dimension in [118].

Definition 6.20 (Sequential Scale-sensitive Dimension [118])**.** *Given a set $\mathcal{X}$ and a function class $\mathcal{F} \subseteq \{\mathcal{X} \to [-1,1]\}$, a depth-$d$ $\mathcal{X}$-valued tree $\mathbf{x}$ is shattered by $\mathcal{F}$ at scale $\alpha > 0$ if and only if there exists a depth-d $[-1,1]$-valued tree $\mathbf{v}$ such that for any $\boldsymbol{\varepsilon} \in \{-1,1\}^n$, there exists $f^{\boldsymbol{\varepsilon}} \in \mathcal{F}$ which satisfies*

$$\varepsilon_t \cdot (f^{\boldsymbol{\varepsilon}}(x_t(\boldsymbol{\varepsilon})) - v_t(\boldsymbol{\varepsilon})) \ge \frac{\alpha}{2}.$$

The tree $\mathbf{v}$ is called the witness of the shattering. The sequential scale-sensitive dimension $\mathrm{fat}(\mathcal{F}, \alpha)$ is defined to be the largest d such that there exists depth-d $\mathcal{X}$-valued tree $\mathbf{x}$ shattered by $\mathcal{F}$ at scale α.

We have the following relations between the sequential scale-sensitive dimension $\mathrm{fat}(\mathcal{F}, \alpha)$ in Theorem 6.20 and the gapped sequential scale-sensitive dimension $d^{\mathsf{seq}}(\mathcal{F}, \alpha)$ with the distance function $\mathfrak{c}(a,b) = |a-b|$ in Theorem 6.17. First, observe that a tree that is (α, β)-shattered in the above sense for $\beta \le \alpha/4$ is also $\alpha/2$-shattered in the sense of the sequential scale-sensitive dimension fat. Indeed, we may choose $\bar{\mathbf{s}}$ as a $[-1,1]$-valued tree defined by $\bar{s}_t(\boldsymbol{\varepsilon}) = (s_t(\boldsymbol{\varepsilon})[1] + s_t(\boldsymbol{\varepsilon})[-1])/2$. We then have that for any $\boldsymbol{\varepsilon} \in \{-1,1\}^d$, there exists $f^{\boldsymbol{\varepsilon}} \in \mathcal{F}$ such that $\varepsilon_t(f^{\boldsymbol{\varepsilon}}(x_t(\boldsymbol{\varepsilon})) - \bar{s}_t(\boldsymbol{\varepsilon})) \ge \alpha/4$. More precisely, we have the following:

Proposition 6.21. *Given a set $\mathcal{X}$ and a function class $\mathcal{F} \subseteq \{\mathcal{X} \to [-1,1]\}$, for any $0 < 2\beta < \alpha$, we have*

$$d^{\mathsf{seq}}(\mathcal{F}, \alpha, \beta) \le \mathrm{fat}(\mathcal{F}, \alpha - 2\beta)$$

For the reverse direction, the following holds:

Proposition 6.22. *Given set $\mathcal{X}$ and function class $\mathcal{F} \subseteq \{\mathcal{X} \to [-1,1]\}$, for any $\alpha, \beta > 0$, we have*

$$\mathrm{fat}(\mathcal{F}, 3(\alpha + \beta)) \le 4 d^{\mathsf{seq}}(\mathcal{F}, \alpha, \beta) \log\left(\frac{12 d^{\mathsf{seq}}(\mathcal{F}, \alpha, \beta)}{\beta}\right).$$

Just as in the non-sequential case, we remark that Theorem 6.22 is proved (in appendix E.2.3) *using* Theorem 6.19, not the other way around. Furthermore, the following result says that the $\log(1/\beta)$ factor in Theorem 6.22 is indeed necessary.

Proposition 6.23. *Consider the case where $\mathcal{X} = \{x\}$, and function class $\mathcal{F} = \{f : \mathcal{X} \to [-1,1] : f(x) \in [-1,1]\}$. Then for any $0 < 2\beta < \alpha < 1$, we have*

$$d^{\mathsf{seq}}(\mathcal{F}, \alpha, \beta) = 1, \quad \textit{and} \quad \mathrm{fat}(\mathcal{F}, \alpha) \geq \left\lfloor \log_2 \left(\frac{1}{\alpha} \right) \right\rfloor .$$

The proof of Theorem 6.23 is deferred to appendix E.2.3.

6.3.4 Lower Bounds for Sequential Offset Rademacher Complexity and Online Regression

We fix some positive constant $C > 0$. For any set of contexts $\mathcal{X}$, function class $\mathcal{F} \subseteq \{\mathcal{X} \to [-1,1]\}$, depth-$n$ $\mathcal{X}$-valued tree $\mathbf{x}$, depth-n $[-1,1]$-valued tree $\boldsymbol{\mu}$ and $\{-1,1\}$-path $\boldsymbol{\varepsilon} \in \{-1,1\}^n$, we define

$$\mathcal{R}_n(\mathcal{F}, \boldsymbol{\mu}, \mathbf{x}, \boldsymbol{\varepsilon}) = \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \left\{ C \cdot \varepsilon_t(f(x_t(\boldsymbol{\varepsilon})) - \mu_t(\boldsymbol{\varepsilon})) - (f(x_t(\boldsymbol{\varepsilon})) - \mu_t(\boldsymbol{\varepsilon}))^2 \right\} .$$

The expected value $\mathbb{E}[\mathcal{R}_n(\mathcal{F}, \boldsymbol{\mu}, \mathbf{x}, \boldsymbol{\varepsilon})]$ is the sequential offset Rademacher complexity, known to govern the rates of online regression with squared and other strongly convex and smooth losses. We now show that this complexity is lower bounded, for any class $\mathcal{F}$, by the scaling behavior of sequential scale-sensitive dimensions.

Theorem 6.24. *Let $\mathfrak{c}(a,b) = |a-b|$. If there exists a constant $p \geq 0$ such that $d^{\mathsf{seq}}(\mathcal{F}, \alpha, \alpha/20) = \tilde{\Omega}\left(\alpha^{-p}\right)$ for any $\alpha > 0$, then for $C \geq 2$, we have*

$$\sup_{\boldsymbol{\mu}, \mathbf{x}} \mathbb{E}[\mathcal{R}_n(\mathcal{F}, \boldsymbol{\mu}, \mathbf{x}, \boldsymbol{\varepsilon})] = \tilde{\Omega}\left(n^{\frac{p}{p+2}} \right),$$

where the supremum is over all depth-n $\mathcal{X}$-valued trees $\mathbf{x}$ and depth-n $[-1,1]$-valued trees $\boldsymbol{\mu}$, and the expectation is with respect to i.i.d. Rademacher random variables $\boldsymbol{\varepsilon} = (\varepsilon_{1:n}) \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(\{-1,1\})$.

The proof of Theorem 6.24 is deferred to appendix E.2.

Application: Online Regression In parallel to section 6.2.4, Theorem 6.24 implies the following lower bound (Theorem 6.25) for online regression for any $\mathcal{F}$, significantly improving upon the lower bound in [109]; the latter result only guaranteed *existence* of $\mathcal{F}$ with such lower bound properties. This improvement was the main motivation for this paper.

To formally state the lower bound, define the minimax value of the online prediction problem $\mathcal{V}_n^{\mathsf{seq}}(\mathcal{F})$ as

$$\mathcal{V}_n^{\mathsf{seq}}(\mathcal{F}) := \left\{ \sup_{x_t \in \mathcal{X}} \inf_{\widehat{y}_t} \sup_{y_t} \right\}_{t=1}^{n} \left[\sum_{t=1}^{n} (\widehat{y}_t - y_t)^2 - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} (f(x_t) - y_t)^2 \right] .$$

Compared to the transductive setting in eq. (6.6), here the forecaster does not have access to the context x_t until the beginning of round t.

Corollary 6.25. *Fix a function class $\mathcal{F} \subseteq \{\mathcal{X} \to [-1,1]\}$ over context space $\mathcal{X}$. Suppose there exists a real number $p \geq 0$ such that the sequential covering number (defined in Theorem 6.18) under distance $\mathfrak{c}(a,b) = |a-b|$ satisfies $\sup_{\mathbf{x}} \log \mathcal{N}_\infty^{\mathsf{seq}}(\mathcal{F}, \mathbf{x}, \alpha) = \tilde{\Omega}\left(\alpha^{-p}\right)$, where the supremum is over all depth-n $\mathcal{X}$-valued trees. Then we have the following lower bound for the minimax regret:*

$$\mathcal{V}_n^{\mathsf{seq}}(\mathcal{F}) = \tilde{\Omega}\left(n^{\frac{p}{p+2}}\right).$$

Part III

Reinforcement Learning with Outcome-Based Feedback

Chapter 7

Do We Need to Verify Step by Step? Rethinking Process Supervision from a Theoretical Perspective

This chapter is based on the paper "Do We Need to Verify Step by Step? Rethinking Process Supervision from a Theoretical Perspective" by Zeyu Jia, Alexander Rakhlin, and Tengyang Xie, published at the International Conference on Machine Learning (ICML) 2025.

7.1 Introduction

Reward signals play a central role in reinforcement learning, and it has been hypothesized that intelligence and its associated abilities could emerge naturally from the simple principle of reward maximization [133]. Over the past decade, this idea has been powerfully demonstrated across diverse AI systems. In specialized domains like AlphaGo Zero [134], superhuman performance has been achieved by maximizing simple, well-defined environmental reward signals. The paradigm has also proven transformative for general-purpose AI systems, particularly in training large language models (LLMs) using reinforcement learning [22–24]. However, for these more open-ended systems, the challenge of reward specification is significantly more complex, requiring reward signals to be learned from human-annotated data through reward modeling rather than being manually specified.

This challenge of reward specification has led to the emergence of two fundamental supervision paradigms in reinforcement learning [e.g., 25, 26]:

- *Outcome supervision:* Reward feedback is provided only after the final output, based on the final outcomes or—in the case of LLMs—overall quality of the model's chain-of-thought (CoT).
- *Process supervision:* Fine-grained reward feedback is provided based on the quality of each intermediate step (e.g., correctness of each step in the CoT in the case of LLMs).

The choice between these paradigms represents a fundamental trade-off in reinforcement learning system design. Process supervision offers several intuitive advantages: it provides

more granular feedback, enables better interpretability of model decisions, and potentially allows for more efficient credit assignment in long reasoning chains. These benefits have led to significant investment in collecting step-by-step feedback data, despite the substantial human effort required [26]. The granular nature of process supervision also aligns with how humans often learn and teach—through step-by-step guidance rather than just final outcomes.

However, the high cost of collecting process supervision data raises important questions about its necessity. Outcome supervision, while providing less detailed feedback, offers practical advantages in terms of data collection efficiency and scalability. It also reflects various natural settings of human learning where detailed step-by-step feedback may not be available (e.g., game-playing). Recent empirical advances in automated process supervision derived from outcome supervision [135–139] or in directly learning from outcome supervision [140] suggest that the statistical benefits of process supervision might not be as fundamental as previously thought.

This tension between process and outcome supervision touches on deeper questions in machine learning and cognitive science: How much explicit supervision is truly necessary for effective learning? Can systems learn optimal behavior from sparse reward signals, or is step-by-step guidance fundamentally necessary? Understanding these questions has important implications not just for practical system design, but also for our broader understanding of learning and intelligence.

7.1.1 Our Results

This paper examines the statistical performance of reinforcement learning under outcome supervision—an emerging paradigm that has garnered significant attention in large language model research. Our findings challenge conventional wisdom that outcome supervision is inherently more difficult than process supervision due to its coarser feedback.

(1) Our main results (section 7.3) demonstrate that given a dataset of trajectories with only cumulative rewards (as in outcome supervision), we can transform the data into trajectories with per-step rewards, while only paying an additive error that scales with the *state-action concentrability*. As a result, we can transform any algorithm that takes trajectories with per-step rewards as input into an algorithm that takes trajectories with total rewards as input, with essentially no loss in statistical performance up to polynomial factors of the horizon.

(2) We also provide (section 7.4) a theoretical analysis of using Q-functions or advantage functions as reward functions, a popular approach in practice for mimicking process supervision from outcome supervision data [e.g., 135, 136, 139]. We prove that the advantage function of *any policy* can serve as an optimal process reward model; in contradistinction, using Q-functions could lead to sub-optimal results.

Beyond these two main messages, we make the following contributions:

(1) Our key technical contribution—*Change of Trajectory Measure Lemma* (Theorem 7.3)—is applicable beyond our main results. The change of measure is a fundamental operation in analyzing off-policy evaluation [e.g., 141], offline reinforcement learning [e.g., 142],

and out-of-domain generalization [143]. To our knowledge, Theorem 7.3 presents the first result concerning trajectory-level change of measure via step-level distribution shift.

(2) We also extend our main results to the setting of preference-based reinforcement learning (section 7.3.4). Namely, we transform preference-based trajectory data, generated according to the Bradley-Terry model, into a dataset of trajectories with per-step reward. In particular, for direct preference optimization [144], we improve the previous analyses and show that its sample complexity only scales with the state-action concentrability coefficients instead of trajectory concentrability coefficients—potentially, an exponential improvement.

7.1.2 Notation

We use $a_n \lesssim b_n$ or $a_n = \mathcal{O}(b_n)$ whenever there exists some universal positive constant c such that $a_n \leq c \cdot b_n$. We use $\sigma : \mathbb{R} \to [0,1]$ to denote the sigmoid function $x \mapsto \sigma(x) = \exp(x)/(1+\exp(x))$.

7.2 Background

In this section, we introduce key prerequisite concepts. We begin with the basics of Markov Decision Processes (section 7.2.1). We then discuss the two aforementioned supervision paradigms in reinforcement learning (section 7.2.2). Finally, we review the concepts of state-action and trajectory concentrability (section 7.2.3).

7.2.1 Markov Decision Processes

An MDP M consists of a tuple $(\mathcal{S}, \mathcal{A}, P, r^\star, H)$. Here $H \in \mathbb{Z}_+$ denotes the horizon, $\mathcal{S} = \cup_{h=1}^H \mathcal{S}_h$ denotes the layered state space, $\mathcal{A}$ denotes the action space, $P = (P_1, \cdots, P_H)$ with $P_h : \mathcal{S}_h \times \mathcal{A} \to \Delta(\mathcal{S}_{h+1})$ denoting the transition model, and $r^\star : \mathcal{S} \times \mathcal{A} \to [0,1]$ is the deterministic ground truth reward model. For simplicity, we let $s_1 \in \mathcal{S}_1$ be a fixed initial state.

For a trajectory $\tau = (s_1, a_1, s_2, a_2, \cdots, s_H, a_H)$, we write $r(\tau) \coloneqq \sum_{h=1}^H r(s_h, a_h)$ to denote the total reward accumulated along the trajectory under deterministic reward model $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ (which can be the ground truth reward model $r^\star$ or any learned reward model $\widehat{r}$). A (Markov) policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ is a mapping from the state space $\mathcal{S}$ to a distribution on the action space $\mathcal{A}$. The notation $\mathbb{P}_{\tau\sim\pi}$ and $\mathbb{E}_{\tau\sim\pi}$ stands for the probability and expectation with respect to trajectories τ sampled according to policy π within the transition model given by P, starting from a fixed state s_1. For any given policy π, the occupancy measure of any state $s_h \in \mathcal{S}_h$ in layer h and any state-action pair $(s_h, a_h) \in \mathcal{S}_h \times \mathcal{A}$ are defined, respectively, as $d^\pi(s_h) \coloneqq \mathbb{P}_{\tau\sim\pi}(s_h \in \tau)$ and $d^\pi(s_h, a_h) \coloneqq \mathbb{P}_{\tau\sim\pi}((s_h, a_h) \in \tau)$. Additionally, we define the trajectory occupancy measure for any trajectory τ, $d^\pi(\tau) \coloneqq \mathbb{P}_{\tau'\sim\pi}(\tau = \tau')$, as well as $\pi(\tau) \coloneqq \prod_{(s_h,a_h)\in\tau} \pi(a_h \mid s_h)$ (note that $d^\pi(\tau)$ and $\pi(\tau)$ are different when transitions is stochastic). For any policy π, we use $J(\pi)$ to denote the expected total reward of trajectories

collected by π under the ground truth reward $r^\star$, i.e., $J(\pi) \coloneqq \mathbb{E}_{\tau\sim\pi}[r^\star(\tau)]$. For a specific reward model r, we use $J_r(\pi)$ to denote the expected total reward of trajectories collected by π under reward r, i.e., $J_r(\pi) \coloneqq \mathbb{E}_{\tau\sim\pi}[r(\tau)]$. We assume that $r^\star(\tau) \in [0, 1]$ for any trajectory τ.

7.2.2 Outcome Supervision and Process Supervision

This paper focuses on two basic supervision paradigms in reinforcement learning: *process supervision* and *outcome supervision*. We analyze these approaches through the lens of *statistical complexity* rather than algorithmic implementation details.

The distinction lies in the temporal resolution of available reward signals:

- Process supervision provides step-wise rewards during trajectory collection. More precisely, the offline data has the form

$$\mathcal{D}_P \coloneqq \{(s_1, a_1, r_1^\star, s_2, a_2, r_2^\star, \cdots, s_H, a_H, r_H^\star)\}, \tag{7.1}$$

  where $r_h = r^\star(s_h, a_h)$ denotes the ground truth reward value at step h. This setting is compelling compared to outcome supervision, especially for complex multi-step reasoning problems, as it provides more precise feedback that can pinpoint the location of suboptimal actions and allows to correct for cases where the agent makes mistakes in the middle of the reasoning path but reaches the correct final answer [25, 26].

- Outcome supervision reveals only the cumulative rewards for complete trajectories. The data for this setting has the form

$$\mathcal{D}_O \coloneqq \{(s_1, a_1, s_2, a_2, \cdots, s_H, a_H, R)\}, \tag{7.2}$$

  where $R = \sum_{h=1}^{H} r^\star(s_h, a_h)$ denotes the total reward along the trajectory $(s_1, a_1, ..., s_H, a_H)$.

  We also consider preference learning, where the learner only has access to trajectory-level pairwise preferences, as an extension of outcome supervision. Our main results are applicable to both cases, and are discussed in section 7.3. The problem of reinforcement learning from human preferences [e.g., 145] or human feedback [e.g., 22] typically falls under this paradigm, where only the trajectory-level supervision signal is available.

Our analysis reveals the temporal resolution distinction described above does not inherently create statistical complexity gaps when proper coverage conditions hold. Our results formalize this insight in the following two settings: (1) Offline RL with different reward signal resolutions (section 7.3), and (2) Online RL with verifier/rollout access (section 7.4).

Outcome-Supervised and Process-Supervised Reward Models Two other commonly used terms related to this paper are outcome-supervised reward models (ORMs) and process-supervised reward models (PRMs). ORMs are learned to evaluate the final outcomes of

whole trajectories, corresponding to the value R in eq. (7.2). PRMs are learned to evaluate the intermediate rewards of each state-action pair, corresponding to the values $r^\star(s_h, a_h)$ in eq. (7.1) for all $h \in [H]$.

In this paper, we use ORMs and PRMs to refer specifically to different algorithmic approaches for learning reward models, rather than the underlying supervision paradigms discussed earlier, as learning ORMs or PRMs does not necessarily require the underlying data to be collected under the same supervision paradigm.

7.2.3 State-Action Coverage and Trajectory Coverage

The coverage condition—typically referred to as a bounded *concentrability coefficient* [142, 146–152]—has played a central role in the theory of offline (or, batch) reinforcement learning, and has recently gained growing attention in online reinforcement learning through a related concept of a *coverability coefficient* [153–156].

In this paper, we use coverage conditions to capture the statistical complexity of different supervision paradigms. To motivate the importance of coverage notions, consider the following approach for imputing the missing rewards in outcome supervision. Suppose we minimize the trajectory-level regression objective over a class of reward functions $\mathcal{R} = \{r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}\}$,

$$\widehat{r} = \arg\min_{r \in \mathcal{R}} \sum_{(\tau, R) \in \mathcal{D}_O} \left(r(\tau) - R\right)^2, \tag{7.3}$$

where the outcome supervision data $\mathcal{D}_O$ are collected by some reference policy π_{off}. With the learned reward model $\widehat{r}$, we can now employ an offline RL method of our choice. However, we have a (standard in the literature) mismatch: while $\left|\sum_{h=1}^H \widehat{r}(s_h, a_h) - \sum_{h=1}^H r^\star(s_h, a_h)\right|$ is small for trajectories collected by π_{off}, we care about the error $\left|J(\pi) - J_{\widehat{r}}(\pi)\right|$ for some policy π that differs from π_{off}, where $J(\pi)$ and $J_{\widehat{r}}(\pi)$ are defined in section 7.2.1. A naive approach for capturing such a change of trajectory measure is to use the *trajectory concentrability coefficient*,

$$C_{\mathsf{traj}}(\pi, \pi_{\text{off}}) := \sup_{\tau} \frac{d^\pi(\tau)}{d^{\pi_{\text{off}}}(\tau)}, \tag{7.4}$$

where the supremum is over all possible trajectories.

This trajectory concentrability coefficient is usually considered to be prohibitively large, and its limitation has been widely studied in the literature on off-policy evaluation [e.g., 141, 157–159]. An alterative approach is to express upper bounds (if possible) in terms of the *state-action concentrability coefficient*, commonly used in offline policy learning literature [e.g., 146–149] and defined as follows:

$$C_{\mathsf{sa}}(\pi, \pi_{\text{off}}) := \max_{h \in [H]} \sup_{s_h \in \mathcal{S}_h, a \in \mathcal{A}} \frac{d^\pi(s_h, a_h)}{d^{\pi_{\text{off}}}(s_h, a_h)}. \tag{7.5}$$

The state-action concentrability coefficient is always smaller than the trajectory concentrability coefficient, and the difference can be exponential. This is because $d^\pi(s_h, a_h)$ aggregates over all trajectories that pass through (s_h, a_h), and thus

$$\frac{d^\pi(s_h, a_h)}{d^{\pi_{\text{off}}}(s_h, a_h)} = \frac{\sum_{\tau : (s_h, a_h) \in \tau} d^\pi(\tau)}{\sum_{\tau : (s_h, a_h) \in \tau} d^{\pi_{\text{off}}}(\tau)} \le \sup_{\tau} \frac{d^\pi(\tau)}{d^{\pi_{\text{off}}}(\tau)},$$

for all $h \in [H]$ and $(s_h, a_h) \in \mathcal{S}_h \times \mathcal{A}$.

It is straightforward to see that using the state-action concentrability coefficient is usually sufficient for the process supervision paradigm, the setting prevalent in the offline RL literature. The intuition is that process supervision allows us to achieve small estimation error at the state-action level, which only poses a state-action level change of measure during the subsequent policy learning step. *Perhaps surprisingly, we show that state-action concentrability is also sufficient for the outcome supervision case.*

Single-Policy and All-Policy Coverage Another important concept in the offline RL literature is the distinction between single-policy and all-policy concentrability. Using the state-action concentrability coefficient as an example, single-policy concentrability refers to the coverage of a specific target policy (such as the optimal policy $\pi^\star$), defined as $C_{\mathsf{sa}}(\pi^\star, \pi_{\text{off}})$. In contrast, all-policy concentrability considers the coverage of all policies in a policy class Π, defined as $C_{\mathsf{sa}}(\Pi, \pi_{\text{off}}) \coloneqq \sup_{\pi \in \Pi} C_{\mathsf{sa}}(\pi, \pi_{\text{off}})$.

Single-policy versus all-policy coverage is one of the central questions in the offline RL literature. For more details, we refer the reader to Xie et al. [142], Chen and Jiang [149], and Jiang and Xie [160], but this is not the focus of this paper. In the following sections, we will present results using all-policy coverage conditions for simplicity, and defer the single-policy case to the appendix.

7.3 Outcome Supervision: Statistical Guarantees

7.3.1 Learning a Reward Model from Total Reward

We present a simple approach to estimate rewards in an outcome supervision dataset of the form eq. (7.2) using least squares regression, assuming that the learner has access to a class of reward models $\mathcal{R}$. This transformation allows the learner to use outcome supervision data with methods designed for process reward data, as detailed below. We have the following theorem for the least squares estimate of the rewards:

Theorem 7.1. *Suppose the dataset $\mathcal{D}_O$ is collected i.i.d. according to policy π_{off} in the MDP $M = (\mathcal{S}, \mathcal{A}, P, r^\star, H)$ with the ground truth reward model $r^\star \in \mathcal{R}$. Then, with probability at least $1 - \delta$, for any policy π, the PRM reward model $\widehat{r}$ computed by eq. (7.3) satisfies*

$$|J_{\widehat{r}}(\pi) - J(\pi)| \lesssim H^{3/2} \cdot \sqrt{\frac{C_{\mathsf{sa}}(\pi, \pi_{\text{off}}) \cdot \log(|\mathcal{R}|/\delta)}{|\mathcal{D}_O|}},$$

where $C_{\mathsf{sa}}(\pi, \pi_{\text{off}})$ is the state-action concentrability coefficient defined in eq. (7.5).

The proof of Theorem 7.1 is deferred to appendix F.1.2. Theorem 7.1 yields an approach for transforming any offline RL algorithm which takes trajectories with per-step reward into an offline RL algorithm which takes trajectories with total reward as input. More precisely, we split the outcome supervised data, use the first part to estimate the reward function via least squares, and then use this estimate to impute the missing rewards on the second part of the data. We summarize this basic transformation in the following algorithm.

Algorithm 1 Offline Outcome-to-Process Transformation

1: **Input:** Offline dataset with total rewards $\mathcal{D}_O = \{(s_1, a_1, \cdots, s_H, a_H, R)\}$, Reward model class $\mathcal{R}$, Offline RL Algorithm $\mathfrak{A}$
2: Split $\mathcal{D}_O$ into two datasets $\mathcal{D}_O^1$ and $\mathcal{D}_O^2$ of equal size.
3: Compute $\widehat{r}$ by solving eq. (7.3) with dataset $\mathcal{D}_O^1$ and the reward class $\mathcal{R}$.
4: Construct dataset $\mathcal{D}_P = \{(s_1, a_1, r_1, \cdots, s_H, a_H, r_H)$ from $\mathcal{D}_O^2$, where $\forall (s_1, a_1, \cdots, s_H, a_H, R) \in \mathcal{D}_O^2$,

$$r_h = \widehat{r}(s_h, a_h), \qquad \forall h \in [H].$$

5: Call algorithm $\mathfrak{A}$ with dataset $\mathcal{D}_P$ and output learned policy $\widehat{\pi}$.
6: **Output:** Learned policy $\widehat{\pi}$.

Corollary 7.2. *Fix a policy set Π which contains the optimal policy $\pi^\star$. Suppose the offline RL algorithm $\mathfrak{A}$ always outputs a policy $\widehat{\pi} \in \Pi$ with error at most $\varepsilon_{\mathsf{alg}}$, i.e., $J(\pi^\star) - J(\widehat{\pi}) \le \varepsilon_{\mathsf{alg}}$, with probability at least $1-\delta$. Then the policy output by Algorithm 1 satisfies with probability at least $1-2\delta$,*

$$\max_{\pi} J(\pi) - J(\widehat{\pi}) \lesssim \varepsilon_{\mathsf{alg}} + H^{3/2} \cdot \sqrt{\frac{C_{\mathsf{sa}}(\Pi, \pi_{\mathrm{off}}) \cdot \log(|\mathcal{R}|/\delta)}{|\mathcal{D}_O|}},$$

where $C_{\mathsf{sa}}(\Pi, \pi_{\mathrm{off}}) := \sup_{\pi \in \Pi} C_{\mathsf{sa}}(\pi, \pi_{\mathrm{off}})$.

Notice that the bound in the above theorem suffers from all-policy concentrability, regardless of which algorithm $\mathfrak{A}$ is used with the transformed data. This occurs because the transformation fixes the learned reward model, requiring us to account for the distribution shift between π_{off} and the data-dependent policy $\widehat{\pi}$ which can be any policy in the policy class Π. This is a common issue in classical offline RL without specific methods like pessimism, particularly for the case of partial coverage. However, the concept of pessimism can also be applied to the outcome supervision setting in our paper, where we learn a specific reward model for each policy, as commonly done in the (process-supervision) offline RL literature [e.g., 152, 161–163]. Following this approach, we can transform model-based offline RL algorithms that use pessimism [152, 161] into algorithms that employ outcome supervision data. The sample complexity of these transformed algorithms scales with the single-policy concentrability $C_{\mathsf{sa}}(\pi^\star, \pi_{\mathrm{off}})$, which depends only on the optimal policy $\pi^\star$. We defer further details to appendix F.2.

Statistical Efficiency We now argue that there is no significant statistical edge for process supervision paradigm compared to the outcome supervision paradigm in the offline setting. The latter corresponds to standard offline RL problems [160, 164], for which a rich body of work exists analyzing sample complexity. Our "equivalence" argument primarily focus on coverage conditions, since different coverage notions (e.g., state-action-level vs. trajectory-level) can lead to exponential differences, as discussed in section 7.2.3. While our results establish equivalence with respect to coverage conditions, we acknowledge they may still be subject to polynomial factors of H; removing such factors is an avenue for further research.

If we consider the worst-case scenario, it is easy to see that any algorithm which outputs an ε-optimal policy requires at least $\sup_\pi C_{\mathsf{sa}}(\pi, \pi_{\text{off}})/\varepsilon^2$ number of samples. To see this, we may consider the two-armed bandit with action a_1 and a_2, and $r(a_1) = \pm\varepsilon, r(a_2) = 0$, and policy $\pi = \text{Unif}(\{a_1, a_2\})$. In the meantime, many classical offline RL algorithms, such as Fitted Q-Iteration [147, 165], the theoretical backbone of Deep Q-Network (DQN), require sample complexity that scales with $C_{\mathsf{sa}}(\Pi, \pi_{\text{off}})/\varepsilon^2$ for obtaining an ε-optimal policy [149, 166]. Hence Theorem 7.2 provides a transformation with the same sample complexity as in these works (up to polynomial in horizon factors), when encountering outcome supervision reward data.

As for the instance-dependent case (corresponding to the single-policy coverage discussed in section 7.2.3), the lower bound result in Xie et al. [167] shows that any algorithm requires at least $C_{\mathsf{sa}}(\pi^\star, \pi_{\text{off}})/\varepsilon^2$ number of samples to output an ε-optimal policy, which only depends on the coverage of optimal policy $\pi^\star$. Recent offline RL algorithms [142, 152, 162, 163] indeed reach that sample complexity in terms of the single-policy concentrability $C_{\mathsf{sa}}(\pi^\star, \pi_{\text{off}})$. Our results presented in appendix F.2 match the upper bound of these offline RL algorithms for the process supervision case and also enjoy the same sample complexity depending on the single-policy concentrability $C_{\mathsf{sa}}(\pi^\star, \pi_{\text{off}})$.

7.3.2 Change of Trajectory Measure

The proof of Theorem 7.1 relies on the following key change of trajectory measure lemma, which states that changing the measure of trajectory returns can be done at the price of state-action concentrability, up to logarithmic and polynomial-in-horizon factors. This lemma will be used for bounding the error between the true reward model $r^\star$ and the learned reward model $\widehat{r}$. Thus, by setting $f = r^\star - \widehat{r}$, we only need to show that the expectation of the absolute value of $f(\tau) \coloneqq \sum_{(s_h,a_h)\sim\tau} f(s_h, a_h)$ is small for $\tau \sim \pi$ when controlling under π_{off}.

Lemma 7.3. *(Change of Trajectory Measure Lemma) For MDP $M = (\mathcal{S}, \mathcal{A}, T, f, H)$ with any function $f : \mathcal{S} \times \mathcal{A} \to [-1, 1]$, for any two policies π and π_{off}, we have*

$$\frac{\mathbb{E}_{\tau\sim\pi}[f(\tau)^2]}{\mathbb{E}_{\tau\sim\pi_{\text{off}}}[f(\tau)^2]} \lesssim H^3 C_{\mathsf{sa}}(\pi, \pi_{\text{off}}),$$

where $C_{\mathsf{sa}}(\pi, \pi_{\text{off}})$ is defined in eq. (7.5). Additionally, the following holds without the extraneous log factors:

$$\mathbb{E}_{\tau\sim\pi}\left[|f(\tau)|\right] \lesssim \sqrt{H^3 C_{\mathsf{sa}}(\pi, \pi_{\text{off}}) \cdot \mathbb{E}_{\tau\sim\pi_{\text{off}}}[f(\tau)^2]}.$$

This lemma reveals a perhaps surprising insight: when the squared sum of some state-action value functions $[\sum_{(s_h,a_h)\sim\tau} f(s_h, a_h)]^2$ is small under the off-policy trajectory distribution $\tau \sim \pi_{\text{off}}$, we only need to account for state-action-level distribution shifts between π_{off} and π to bound the same squared sum under $\tau \sim \pi$. This holds true even though controlling such trajectory sums theoretically cannot prevent cases where individual terms have equal and large magnitude but opposite signs (i.e., where $|a| = |b| > 0$ but $a + b = 0$). We provide a proof sketch of Theorem 7.3 in the following. The detailed proof is deferred to appendix F.1.1.

7.3.3 Proof Sketch of Theorem 7.3

In this section, we outline the key insights behind the proof of Theorem 7.3. The central observation is that controlling the trajectory-level variance of f under a reference policy π_{off} implies an automatic control of variance on prefixes and suffixes over the entire trajectory, with only a polynomial overhead in the horizon length H. This seemingly simple fact leads to perhaps surprisingly strong guarantees.

Insight I: Trajectory-level bound controls the second moment on prefixes and suffixes At first glance, small value $\big|f(\tau)\big|$ over the entire trajectory τ does *not* obviously guarantee that the value of f on either (i) every prefix $\tau_{1:h}$ or (ii) every suffix $\tau_{h+1:H}$ is small. In principle, large positive and large negative portions of a single trajectory could "cancel" each other out, resulting in a small overall sum $|f(\tau)| = |f(\tau_{1:h}) + f(\tau_{h+1:H})|$.

Crucially, however, thanks to the *Markov property*, we can argue that if f has small second moment (under π_{off}) and if a state s_h is visited sufficiently often by π_{off}, f cannot have high variance on the prefix (leading up to s_h) and suffix (following s_h). Indeed, if the value of f on the prefix (or suffix) has large variance, then conditioned on passing through s_h that is visited sufficiently often by π_{off}, the value of f on the entire trajectory also has large variance, which directly implies the large variance (hence, large second moment) of $f(\tau)$. Hence, even though the trajectory-level bound looks coarse, it forces each state s_h to have relatively stable partial sums in both the prefix and suffix directions under π_{off}.

Insight II: Layer-by-layer "locking" with only state-action coverage Next, we want to argue that if *all* states in a trajectory satisfy the above low-variance property (we call such states "good" states), then the reward of the entire trajectory cannot have large absolute value. We call this the "locking in" property here for brevity. In the following, we argue that "locking in" happens with high probability, even under policy π.

According to the earlier argument, "bad" states (opposite of "good" states) cannot have large visitation probability under π_{off}. Then, by the definition of $C_{\mathsf{sa}}(\pi, \pi_{\text{off}})$, which upper bounds the probability ratio between π and π_{off} at any state, we conclude that such bad states also have low probability under π, up to a factor of $C_{\mathsf{sa}}(\pi, \pi_{\text{off}})$. Thus, we avoid exponential blow-up over the horizon because we only "pay" for distribution shift at each individual (s, a), rather than for entire trajectories: $\Pr_{\tau\sim\pi}(\text{bad}) \le C_{\mathsf{sa}}(\pi, \pi_{\text{off}}) \cdot \mathbb{P}_{\tau\sim\pi_{\text{off}}}(\text{bad})$.

Hence, with high probability, π visits only "good" states throughout its trajectory, ensuring that each layer h "locks in" a small partial sum (as both its prefix and suffix have low variance).[1] When we stitch these layers from $h = 1$ to $h = H$, the entire sum $f(\tau)$ is guaranteed to have small absolute values.

7.3.4 Extension to Preference-Based Reinforcement Learning

In the previous section, we studied the statistical complexity of outcome supervision under the data format of eq. (7.2), where the outcome reward is provided at the end of each trajectory.

[1] Our formal proof also needs to consider the "good" state-action-state tuples, which are similar to "good" states but involve the (s_h, a_h, s_{h+1}) tuple. We omit the details here for brevity, and readers can refer to the full proof in appendix F.1.1.

Preference-based reinforcement learning [e.g., 168–170] represents another well-established paradigm that extends outcome supervision and is commonly employed for learning from human preferences [e.g., 22, 145, 171, 172].

Recent work on implicit reward modeling through single-step contrastive learning approaches [DPO; 144] aims to eliminate the need for explicit reward modeling. While extensive prior work [e.g., 173–176] has focused on the sample complexity of these implicit reward modeling approaches, most existing bounds rely on trajectory-level change of measure. These results are also considered to depend on the trajectory concentrability under naive simplifications, which can grow exponentially with the horizon length (see detailed discussion in section 7.2.3 and Section 3.2 of [177]).

In this section, we extend our main results to the preference-based reinforcement learning setting. As a direct application of our Change of Trajectory Measure Lemma (Theorem 7.3), we first provide a sample complexity bound of preference based RL which only scales with state-action concentrability instead of trajectory concentrability, a result applicable to standard explicit reward modeling approaches as well as implicit reward modeling approaches (i.e., DPO).

In preference-based RL, we suppose that for any trajectory τ, the total reward along τ satisfies $r(\tau) \in [0, V_{\max}]$. To form the dataset $\mathcal{D}$ of preferences, the learner collects two reward-free trajectories $\tau = (s_1, a_1, s_2, a_2, \cdots, s_H, a_H)$ and $\tau' = (s'_1, a'_1, s'_2, a'_2, \cdots, s'_H, a'_H)$ according to policy π_{ref}, and receives the information about the order (τ_+, τ_-) of τ, τ', based on the preference $y \sim \mathbb{P}(\tau \succ \tau')$. We adopt the Bradley-Terry model [178]:

$$\mathbb{P}(\tau \succ \tau') = \frac{\exp(r(\tau))}{\exp(r(\tau)) + \exp(r(\tau'))}. \tag{7.6}$$

The labeled preference dataset $\mathcal{D}$ consists of ordered samples (τ_+, τ_-), where both trajectories are collected according to policy π_{ref} and labeled according to the Bradley-Terry model eq. (7.6).

Improved Analysis of Preference-Based RL with Explicit Reward Modeling

We first provide the analysis for the case where an explicit reward modeling procedure is used for preference-based RL. Suppose the learner is given the preference-based dataset $\mathcal{D}_{\text{pref}}$ and a reward class $\mathcal{R}$, where for every reward model $r \in \mathcal{R}$ and any trajectory τ, $r(\tau) \in [0, V_{\max}]$. In the following result, an analogue of Theorem 7.1, we transform the preference-based dataset into the reward model via maximum likelihood rather than the method of least squares.

Theorem 7.4. *Suppose $\mathcal{D}_{\text{pref}} = \{(\tau_+, \tau_-)\}$ contains i.i.d. pairs of sequences collected according to π_{ref} and ordered according to eq. (7.6) with $r^\star \in \mathcal{R}$. Let*

$$\widehat{r} = \arg\min_{r \in \mathcal{R}} \sum_{(\tau+,\tau-) \in \mathcal{D}_{\text{pref}}} \log \sigma(r(\tau_+) - r(\tau_-))$$

be the maximum likelihood estimate. With probability at least $1 - \delta$, for any two policies $\pi, \pi' \in \Pi$, it holds that

$$\mathbb{E}_{\tau\sim\pi,\tau'\sim\pi'}[|r^\star(\tau)-r^\star(\tau')-\widehat{r}(\tau)+\widehat{r}(\tau')|] \lesssim H^{3/2}V_{\max}e^{2V_{\max}}\cdot\sqrt{\frac{(C_{\mathsf{sa}}(\pi,\pi_{\mathsf{ref}})\vee C_{\mathsf{sa}}(\pi',\pi_{\mathsf{ref}}))\log(|\Pi|/\delta)}{|\mathcal{D}|}}. \tag{7.7}$$

The proof of Theorem 7.4 is deferred to appendix F.1.3. Notice that with eq. (7.7), if we could obtain $\widehat{\pi}$ as the optimal policy under reward model $\widehat{r}$ (otherwise we can just pay for the sub-optimality as in Theorem 7.2), then with high probability $\widehat{\pi}$ is a near-optimal policy. To see this, we choose $\pi = \widehat{\pi} = \arg\max_\pi J_{\widehat{r}}(\widehat{\pi})$ and $\pi' = \pi^\star = \arg\max_\pi J(\widehat{\pi})$, then with probability at least $1-\delta$,

$$\begin{aligned} J(\pi^\star) - J(\widehat{\pi}) &\le J(\pi^\star) - J(\widehat{\pi}) - J_{\widehat{r}}(\pi^\star) + J_{\widehat{r}}(\widehat{\pi}) = \mathbb{E}_{\tau\sim\widehat{\pi},\tau'\sim\pi^\star}[r^\star(\tau') - r^\star(\tau) - \widehat{r}(\tau') + \widehat{r}(\tau)] \\ &\le \mathbb{E}_{\tau\sim\widehat{\pi},\tau'\sim\pi^\star}[|r^\star(\tau') - r^\star(\tau) - \widehat{r}(\tau') + \widehat{r}(\tau)|] \\ &\lesssim H^{3/2}V_{\max}e^{2V_{\max}}\cdot\sqrt{\frac{(C_{\mathsf{sa}}(\widehat{\pi},\pi_{\mathsf{ref}})\vee C_{\mathsf{sa}}(\pi^\star,\pi_{\mathsf{ref}}))\log(|\Pi|/\delta)}{|\mathcal{D}|}}, \end{aligned}$$

Therefore, as we acquire enough samples, $J(\pi^\star) - J(\widehat{\pi})$ converges to zero with high probability, implying that $\widehat{\pi}$ is a near-optimal policy. This convergence requires the same condition of bounded state-action concentrability as in Theorem 7.2.

Improved Analysis of DPO Algorithm

We now extend our main results to the implicit reward modeling setting and analyze the sample complexity of the DPO algorithm [144]. DPO is a popular implicit reward algorithm that converts the two-step process of reward modeling and policy optimization into a single-step contrastive learning problem. DPO is commonly used in the token-level setup of LLMs, where actions (tokens) are directly appended to states (contexts) [144, 179]. In this case, the state-action concentrability coefficient essentially reduces to trajectory-level concentrability, as the last state is contains the trajectory. However, recent work indicates that DPO-style algorithms are applicable beyond the token-level setup, e.g., in environments with deterministic transition dynamics but still Markovian states, e.g., in robotics [177, 180]. In these settings, our bounds with only state-action concentrability can be substantially tighter than existing trajectory-level ones.

Following Xie et al. [177], we assume deterministic ground-truth transition dynamics and consider the following KL-regularized objective [177, 181, 182]: for some positive number β,

$$\mathcal{J}_\beta(\pi) \coloneqq J(\pi) - \beta D_{\mathrm{KL}}(\pi(\tau) \,\|\, \pi_{\mathsf{ref}}(\tau)) = \mathbb{E}_{\tau\sim\pi}\left[r^\star(\tau) - \beta\log\frac{\pi(\tau)}{\pi_{\mathsf{ref}}(\tau)}\right]. \tag{7.8}$$

The policy $\pi^\star_\beta$ which maximizes $\mathcal{J}_\beta(\pi)$ in eq. (7.8) satisfies $\pi^\star_\beta(\tau) \propto \pi_{\mathsf{ref}}(\tau)\exp(r^\star(\tau)/\beta)$ for any trajectory τ. It is easy to verify that $\pi^\star_\beta$ is a Markov policy. We assume the learner has access to a Markov policy class $\Pi \subset \mathcal{S}^{\mathcal{A}}$, and aims to find a policy $\widehat{\pi}$ that is nearly optimal with respect to the policy class Π, i.e.

$$\max_{\pi\in\Pi}\mathcal{J}_\beta(\pi) - \mathcal{J}_\beta(\widehat{\pi}) \le \varepsilon.$$

The DPO algorithm [144] takes the dataset $\mathcal{D} = \{(\tau_+, \tau_-)\}$ as input, and outputs the policy $\widehat{\pi} \in \Pi$ which maximizes the log likelihood, i.e.,

$$\widehat{\pi} \leftarrow \arg\min_{\pi \in \Pi} \left\{ \sum_{(\tau_+, \tau_-) \in \mathcal{D}} \log \left[\sigma \left(\beta \log \frac{\pi(\tau_+)}{\pi_{\mathsf{ref}}(\tau_+)} - \beta \log \frac{\pi(\tau_-)}{\pi_{\mathsf{ref}}(\tau_-)} \right) \right] \right\}. \tag{7.9}$$

In the following, we provide a refined analysis of the sample complexity of this algorithm. We first make the following assumptions.

Assumption 7.5 (Policy Realizability)**.** *The optimal policy is in the policy class* Π*, i.e.,* $\pi^\star_\beta \in \Pi$.

Policy realizability is a minimal assumption for sample-efficient reinforcement learning and is necessary for establishing many standard results [183–185].

In addition, we make the following assumptions of bounded trajectory concentrability.

Assumption 7.6. *For any policy* $\pi \in \Pi$ *and trajectory* τ, $\left| \log \frac{\pi(\tau)}{\pi_{\mathsf{ref}}(\tau)} \right| \leq \frac{V_{\max}}{\beta}$.

This assumption is commonly used in the literature on implicit reward modeling approaches for preference-based reinforcement learning [e.g., 177, 186]. The intuition behind this assumption is that we treat $\log \frac{\pi(\tau)}{\pi_{\mathsf{ref}}(\tau)}$ as an implicit value function,[2] so that this assumption essentially plays the same role as bounded value functions in the analysis.

If the above assumptions hold, we have the following upper bound on the sample complexity of the DPO algorithm in terms of the state-action concentrability.

Theorem 7.7. *Suppose Theorems 7.5 and 7.6 hold, with probability at least* $1 - \delta$*, the output* $\widehat{\pi}$ *of the DPO algorithm in eq. (7.9) satisfies*

$$\mathcal{J}_\beta(\pi^\star_\beta) - \mathcal{J}_\beta(\widehat{\pi}) \lesssim H^{3/2} V_{\max} e^{2V_{\max}} \cdot \sqrt{\frac{\sup_{\pi \in \Pi} C_{\mathsf{sa}}(\pi, \pi_{\mathrm{off}}) \log(|\Pi|/\delta)}{|\mathcal{D}|}}.$$

The proof of Theorem 7.7 is deferred to appendix F.1.3. Note that, the exponential dependence on the reward range $V_{\max}$ is a fundamental characteristic of the Bradley-Terry model, not an artifact of our analysis. Indeed, this exponential dependence appears consistently across the theoretical literature on RLHF [e.g., 173, 177, 186, 188, 189].

7.4 Advantage Function Learning with Rollouts

In this section, we analyze a common empirical strategy for converting outcome-supervised data into process supervision by leveraging online rollouts. The central observation is that, given access to an environment that returns final outcomes, one can initiate rollouts from individual state-action pairs and use the resulting outcomes to approximate their "quality."

[2] In fact, $\log \pi(\tau)/\pi_{\mathsf{ref}}(\tau)$ corresponds to a more complex combination of value functions and rewards. We refer readers to Xie et al. [177], Rafailov et al. [179], and Watson, Huang, and Heess [187] for further details.

Multiple works have adopted variations of this idea, relying on Q-functions [e.g., 135], advantage functions [e.g., 139], or other specialized value estimators [e.g., 136].

Although these methods have demonstrated empirical promise, their theoretical properties remain relatively unexplored. Establishing a theoretical foundation could reveal the assumptions and conditions under which these methods are effective and enable principled comparisons to alternative reward modeling approaches. In what follows, we present (to our knowledge) the first theoretical study of advantage-based reward learning with online rollouts. We show that the advantage function of *any* policy can serve as a valid process-based reward model, recovering the same optimal policy as the original environment. By contrast, we also prove a lower bound indicating that simply using the Q-function can fail: in certain cases, the Q-function-based reward model produces suboptimal or undesired policies.

7.4.1 Algorithm and Upper Bounds

For MDP is $M = (\mathcal{S}, \mathcal{A}, P, r, H)$, suppose the transition model P is known to the learner, but the reward model r is unknown. For any given policy μ, we define the advantage function $A^\mu : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ as

$$A^\mu(s,a) = Q_h^\mu(s,a) - V_h^\mu(s), \qquad \forall h \in [H], s \in \mathcal{S}_h, a \in \mathcal{A}, \tag{7.10}$$

where Q_h^μ and V_h^μ denote the Q-function and value function of policy μ.

Setlur et al. [139] find an approximation $\widehat{r}$ to A^μ and then invoke a policy gradient algorithm for the MDP with transition model P and reward function $\widehat{r}$, yielding a policy $\widehat{\pi}$. We provide a theoretical guarantee for this approach by showing that the performance gap of $\widehat{\pi}$ to the optimal policy can be upper bounded in terms of the error $\varepsilon_{\mathsf{stat}}$ of approximating the advantage function and the error $\varepsilon_{\mathsf{alg}}$ of optimizing the policy. Before stating the result, we define the concentrability coefficient with respect to distribution ν,

$$C_{\mathsf{sa}}(\nu) \coloneqq \sup_{\pi} \sup_{s \in \mathcal{S}, a \in \mathcal{A}} \frac{d^\pi(s,a)}{\nu(s,a)}, \tag{7.11}$$

where the outer supremum is over all possible policies.

Theorem 7.8. *Suppose there exists some policy μ and distribution $\nu \in \Delta(\mathcal{S} \times \mathcal{A})$ such that*

$$\mathbb{E}_{(s,a)\sim\nu}\left[|\widehat{r}(s,a) - A^\mu(s,a)|^2\right] \le \varepsilon_{\mathsf{stat}}. \tag{7.12}$$

If policy $\widehat{\pi}$ satisfies

$$\max_\pi J_{\widehat{r}}(\pi) - J_{\widehat{r}}(\widehat{\pi}) \le \varepsilon_{\mathsf{alg}}, \tag{7.13}$$

then it also satisfies

$$\max_\pi J(\pi) - J(\widehat{\pi}) \le 2H\sqrt{C_{\mathsf{sa}}(\nu)} \cdot \varepsilon_{\mathsf{stat}} + \varepsilon_{\mathsf{alg}}.$$

The proof of Theorem 7.8 is deferred to appendix F.3. Intuitively, the proof follows from the performance difference lemma [190], which states that for any policies π and μ: $J(\pi) - J(\mu) = H \cdot \mathbb{E}_{(s_h,a_h)\sim d^\pi}[A^\mu(s_h,a_h)]$. This implies that maximizing J_{A^μ} (treating A^μ as

the reward function) is equivalent to maximizing J, since they differ only by the constant term $J(\mu)$. Therefore, both optimization problems yield the same optimal policy.

There are several ways of obtaining an estimate $\widehat{r}$ of the advantage function A^μ that satisfies eq. (7.12). One commonly used approach is Monte-Carlo sampling, for instance as in Setlur et al. [139]. In detail, this approach first collects a dataset $\mathcal{D}$ of data $(s, a, \widehat{A})$, where $(s, a) \in \mathcal{S} \times \mathcal{A}$ and $\widehat{A}$ is calculated via rollout from (s, a) under policy μ, serving as an approximation to the advantage function of policy μ at (s, a). Next, we fit a reward function $\widehat{r}$ in a reward class to the dataset $\mathcal{D}$. Then, as long as the reward class realizes the ground truth advantage function A^μ, the reward function $\widehat{r}$ satisfies eq. (7.12) with high probability.

7.4.2 Lower Bound on Failure of Using Q-Functions

Theorem 7.8 indicates that the MDP with reward function set to be the advantage function of any policy μ has the same best policy as the original MDPs. One may wonder whether the same holds for the Q-function as well. In this section, we disprove this by providing a hard MDP with best policy $\pi^\star$, and a policy μ, so that the best policy of the MDP with reward function Q^μ is not $\pi^\star$.

Theorem 7.9. *There exists an MDP $M = (\mathcal{S}, \mathcal{A}, P, r, H)$, and a policy $\mu \in \mathcal{A}^\mathcal{S}$, such that*

$$\max_\pi J_r(\pi) - J_r(\widehat{\pi}) \geq \frac{1}{3},$$

for $\widehat{\pi} = \arg\max_\pi J_{Q^\mu}(\pi)$. Here $Q^\mu : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the Q-function of MDP M.

The proof of Theorem 7.9 is deferred to appendix F.3.

Theorem 7.9 indicates that the widely used approach in large language model training, whereby an approximate Q-function is learned in place of the reward model, is theoretically incorrect and possibly outputs undesired policies. In contrast, using the advantage function as the reward model is theoretically justified.

7.5 Related Work

Process vs. Outcome Supervision Our work is motivated by the empirical effectiveness of process supervision over outcome supervision, particularly in language model reasoning tasks [25, 26, 191]. To address the challenges of cost and scalability in obtaining human-annotated process labels, recent approaches [135, 136, 139] have developed automated methods to generate process supervision from outcome-based signals, leveraging Q-functions and advantage functions under specific policies. When data is provided in the form of preferences, outcome supervision is sometimes conducted with implicit rewards, as seen in works such as Direct Preference Optimization [137, 138, 144, 192].

RL with Trajectory Feedback A closely related line of theoretical work is reinforcement learning with trajectory feedback or aggregate bandit feedback (or, bandit and semi-bandit feedback) [27, 29–32, 193], where the learner only receives trajectory-level feedback at the

end of each episode. This line of work also includes preference-based RL [28, 34, 173, 188], which operates on trajectory-level pair preferences. While most existing works in this area focus on online exploration settings, our paper investigates offline learning and analyzes the statistical relationship between process (step-level) and outcome (trajectory-level) feedback.

Offline Reinforcement Learning Our work is most closely related to offline (batch) reinforcement learning in the classical reinforcement learning literature. The paradigm of reinforcement learning with process-supervised data is essentially an offline RL problem, where a rich body of existing theoretical results [e.g., 142, 146–152, 161–163, 166] can be applied to our paper, either directly for the process supervision case, or serving as subroutines in our Algorithm 1 for the outcome supervision case. Within these results, Chen and Jiang [149] and Xie and Jiang [166] develop model-free algorithms under all-policy coverage conditions, while Xie and Jiang [151] proposes a model-free approach requiring only realizability assumptions but with stronger coverage requirements. Xie et al. [142] and Cheng et al. [162] investigate model-free offline RL under partial coverage settings, and Bhardwaj et al. [152], Xie et al. [161], and Uehara and Sun [163] address model-based offline RL with partial coverage.

Off-Policy Evaluation Our work also connects to the rich literature on off-policy evaluation (OPE) in reinforcement learning. A central challenge in OPE is the change of measure problem, where extensive research [141, 157–159] has investigated the significant distinction between state-action coverage and trajectory coverage conditions. These findings highlight the significance of our main results, particularly our change of trajectory measure lemma.

Reward Shaping and Internal Rewards A related but distinct line of research focuses on augmenting sparse reward functions to improve learning efficiency. Reward shaping techniques have been extensively employed as a method for providing denser learning signals while preserving optimal policies [e.g., 194–196]. Similarly, intrinsic rewards based on prediction errors of environment dynamics [197] or random networks [198] have been proposed to tackle sparse reward settings. However, these approaches differ fundamentally from our work in their objectives – while reward shaping and intrinsic rewards aim to improve exploration in online RL by modifying the reward landscape, our analysis focuses on the statistical properties of learning in the offline setting where the data distribution is fixed.

7.6 Conclusion

In this paper, we present a way of transforming data in the setting of outcome supervision into the data in the setting of process supervision. This transformation enables us to design offline algorithms in the outcome supervision model from the large pool of algorithms that use process supervision. This transformation extends to preference-based algorithms such as DPO. Beyond this transformation, we also provide theoretical guarantees for algorithms using an approximate advantage function as the reward function.

While our transformation scheme works for most of the offline algorithms, the theoretical guarantees require that the outcome supervision data are collected offline. How to construct similar transformation for online data or online algorithms is left for future work.

Chapter 8

Outcome-Based Online Reinforcement Learning: Algorithms and Fundamental Limits

This chapter is based on the paper "Outcome-Based Online Reinforcement Learning: Algorithms and Fundamental Limits" by Fan Chen, Zeyu Jia, Alexander Rakhlin, and Tengyang Xie, published at the Conference on Neural Information Processing Systems (NeurIPS) 2025.

8.1 Introduction

Reinforcement learning with outcome-based feedback is a fundamental paradigm where agents receive rewards only at the end of complete trajectories rather than at individual steps. This feedback model naturally arises in many applications, from large language model training [22–24], where human preferences are provided for entire outputs rather than individual tokens, to clinical trials, where patient outcomes are only observable after a complete treatment regimen. Despite the prevalence of such settings, the statistical implications of outcome-based feedback for online exploration remain poorly understood.

In traditional reinforcement learning [5], agents observe rewards immediately after each action, providing a granular signal that directly links actions to their consequences. In contrast, outcome-based feedback presents a fundamental challenge: when rewards are only observed at the trajectory level, determining which specific actions contributed to the final outcome becomes significantly more difficult. This credit assignment problem is particularly acute in sequential decision-making tasks with long horizons, where many different action combinations could lead to the observed outcome.

While recent work [199] has shown that outcome-based feedback is sufficient for offline reinforcement learning under certain conditions, the feasibility of efficient online exploration with only trajectory-level feedback remains an open question. Online learning—where an agent actively explores to gather new data—is essential for adaptive systems that must learn in dynamic environments without pre-collected datasets. This leads to our central question:

When is online exploration with outcome-based reward statistically tractable?

This question has been studied in the setting where the reward function is assumed to be well-structured [27–31], with a primary focus on the *linear* reward functions. Similar reliance on the well-behaved reward structure also appears in the recent work on Reinforcement Learning from Human Feedback (RLHF) [32–35], where only *preference* feedback is available. However, well-behaved reward structure is dedicated and might fail to capture many real-world scenarios with general function approximation. In this paper, we address this question by providing a comprehensive theoretical analysis of outcome-based online reinforcement learning with general function approximation. We investigate when efficient exploration is possible with only trajectory-level feedback and characterize the fundamental statistical limits of learning in this setting. Our main results are as follows:

(1) We present a model-free algorithm for outcome-based online RL with general function approximation (Algorithm 2) that relies solely on trajectory-level reward feedback rather than per-step feedback. Our algorithm achieves a complexity bound of $\widetilde{O}(C_{\text{cov}}H^3/\varepsilon^2)$ under standard realizability and completeness assumptions, where C_{cov} is the coverability coefficient that measures an intrinsic complexity of the underlying MDP. This bound applies in the general function approximation setting where state spaces may be large or infinite, requiring only that value functions can be represented by an appropriate function class with bounded statistical complexity.

(2) For the special case of deterministic MDPs, we present a simpler algorithm based on Bellman residual minimization (Algorithm 3) that achieves similar theoretical guarantees with improved computational efficiency.

(3) As extension, we generalize our approach to preference-based reinforcement learning (section 8.4), where feedback comes in the form of binary preferences between trajectory pairs under the Bradley-Terry-Luce model. This extension bridges the gap to practical reinforcement learning from human feedback (RLHF) scenarios, where even outcome reward feedback is rare.

(4) We also identify a fundamental separation between outcome-based and per-step feedback (section 8.5). Specifically, there exists a MDP with known transition dynamics and horizon $H = 2$, and the reward being a d-dimensional generalized linear function, while in this problem $e^{\Omega(d)}$ samples are necessary to learn a near-optimal policy with only outcome reward. However, such a problem is known to be *easy* with per-step reward feedback, in the sense that existing algorithms can return an ε-optimal within $\widetilde{O}\left(d^2/\varepsilon^2\right)$ rounds with per-step feedback. This separation demonstrates that delicate analysis based on well-behaved reward structure can fail catastrophically when only outcome reward feedback is available.

Our results provide a theoretical foundation for understanding when outcome-based exploration is tractable and when it presents insurmountable statistical barriers. By characterizing these fundamental limits, we offer guidance for the development of efficient algorithms for learning from trajectory-level feedback in online settings and highlight the precise conditions under which outcome-based feedback is statistically equivalent to per-step feedback.

8.2 Preliminaries

Markov Decision Process An MDP M is specified by a tuple $(\mathcal{S}, \mathcal{A}, \mathbb{T}, \rho, R, H)$, with state space $\mathcal{S}$, action space $\mathcal{A}$, horizon H, transition kernel $\mathbb{T} = (\mathbb{T}_h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S}))_{h=1}^{H-1}$, is initial state distribution $\rho \in \Delta(\mathcal{S})$, and the mean reward function $R = (R_h : \mathcal{S} \times \mathcal{A} \to [0,1])_{h=1}^{H}$. At the start of each episode, the environment randomly draws an initial state $s_1 \sim \rho$, and then at each step $h \in [H]$, after the agent takes action a_h, the environment generates the next state $s_{h+1} \sim \mathbb{T}(\cdot|s_h, a_h)$. The episode terminates immediately after a_H is taken, and, for notational simplicity, we denote s_{H+1} to be the deterministic terminal state. We denote $\tau = (s_1, a_1, \cdots, s_H, a_H)$ to be the trajectory, and throughout this paper we always assume the reward function is normalized, i.e., $R(\tau) := \sum_{h=1}^{H} R_h(s_h, a_h) \in [0,1]$ almost surely.

In addition to the states, the learner may also observe the reward feedback after the episode terminates. In the *process reward* feedback setting, the learner receives a random reward vector $(r_1, \cdots, r_H) \in [0,1]^H$ such that $\mathbb{E}[r_h|\tau] = R_h(s_h, a_h)$ for each $h \in [H]$. In the *outcome reward* setting, the learner only receives a single reward value $r \in [H]$ such that $\mathbb{E}[r|\tau] = \sum_{h=1}^{H} R_h(s_h, a_h)$.

Policies, value functions, and the Bellman operator A (randomized) policy π is specified as $\{\pi_h : \mathcal{S} \to \Delta(\mathcal{A})\}$, and it induces a distribution $\mathbb{P}^\pi$ of trajectory $\tau = (s_1, a_1, \cdots, s_H, a_H)$ by $s_1 \sim \rho$, and for each $h \in [H]$, $a_h \sim \pi_h(s_h)$, $s_{h+1} \sim \mathbb{T}_h(s_h, a_h)$. We let $\mathbb{E}^\pi[\cdot]$ to be the corresponding expectation.

The expected cumulative reward of a policy π is given by $J(\pi) := \mathbb{E}^\pi \left[\sum_{h=1}^{H} R_h(s_h, a_h)\right]$. The value function and Q-function of π is defined as

$$V_h^\pi(s) := \mathbb{E}^\pi \left[\sum_{\ell=h}^{H} R_\ell(s_\ell, a_\ell) \middle| s_h = s\right], \qquad Q_h^\pi(s,a) := \mathbb{E}^\pi \left[\sum_{\ell=h}^{H} R_\ell(s_\ell, a_\ell) \middle| s_h = s, a_h = a\right].$$

Let $\pi^\star$ denote an optimal policy (i.e., $\pi^\star \in \arg\max_\pi J(\pi)$), and let $V^\star$ and $Q^\star$ be the corresponding value function and Q-function. It is well-known that $(V^\star, Q^\star)$ satisfies the following Bellman equation for each $s \in \mathcal{S}, a \in \mathcal{A}, h \in [H]$:

$$V_h^\star(s) = \max_{a \in \mathcal{A}} Q_h^\star(s,a), \qquad Q_h^\star(s,a) = R_h(s,a) + \mathbb{E}_{s' \sim \mathbb{T}_h(\cdot|s,a)} V_{h+1}^\star(s'), \tag{8.1}$$

with the convention that $V_{H+1}^\star = 0$. Therefore, we define the Bellman operator $\mathcal{T}_h$ as follows: for any $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, $\mathcal{T}_h f$ is defined as

$$[\mathcal{T}_h f](s,a) := R_h(s,a) + \mathbb{E}_{s' \sim \mathbb{T}_h(\cdot|s,a)} \max_{a' \in \mathcal{A}} f(s', a'). \tag{8.2}$$

Then, it is straightforward to verify that the Bellman equation reduces to $Q_h^\star = \mathcal{T}_h Q_{h+1}^\star$ for $h \in [H]$.

Complexity measure of the MDP Coverability is a natural notion for measuring the difficulty of learning in the underlying MDP [153].

Definition 8.1 (Coverability). *For a given MDP M and a policy class Π, the coverability C_{cov} is defined as*

$$C_{\mathrm{cov}}(\Pi; M) := \min_{\mu_1, \cdots, \mu_H \in \Delta(\mathcal{S}\times\mathcal{A})} \max_{h\in[H], \pi\in\Pi} \left\| \frac{d_h^\pi}{\mu_h} \right\|_\infty,$$

where $\left\| \frac{d_h^\pi}{\mu_h} \right\|_\infty := \max_{s\in\mathcal{A}, a\in\mathcal{A}} \frac{d_h^\pi(s,a)}{\mu_h(s,a)}$.

The coverability coefficient of an MDP is an inherent measure of the diversity of the state-action distributions. Our main upper bounds scale with the coverability of the underlying MDP $M^\star$, and in this case we abbreviate $C_{\mathrm{cov}}(\Pi) := C_{\mathrm{cov}}(\Pi; M^\star)$ for succinctness.

Function approximation In this paper, we work with (model-free) function approximation, where the learner have access to a *value function class* $\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_H$ and a *reward function class* $\mathcal{R} = \mathcal{R}_1 \times \cdots \times \mathcal{R}_H$ with each $\mathcal{F}_h, \mathcal{R}_h \subseteq (\mathcal{S}\times\mathcal{A} \to [0,1])$.

The function class $\mathcal{F}$ and $\mathcal{R}$ consist of candidate functions to approximate $Q^\star$ and the ground-truth reward function $R^\star$.[1] In the literature of RL with general function approximation, it is typically assumed that the function classes are *realizable*, i.e., $Q^\star \in \mathcal{F}$ and $R^\star \in \mathcal{R}$. In this paper, we adopt the following relaxed realizability condition with a fixed approximation error $\varepsilon_{\mathrm{app}} \geq 0$.

Assumption 8.2 (Realizability). *There exists $Q^\sharp \in \mathcal{F}$ and $R^\sharp \in \mathcal{R}$ such that $\max_{h\in[H]} \|Q_h^\sharp - Q_h^\star\|_\infty \leq \varepsilon_{\mathrm{app}}$, $\max_{h\in[H]} \|R_h^\sharp - R_h^\star\|_\infty \leq \varepsilon_{\mathrm{app}}$.*

For each value function $f \in \mathcal{F}$, it induces a greedy policy π_f given by $\pi_{f,h}(s) := \arg\max_{a\in\mathcal{A}} f(s,a)$. Therefore, the value function class $\mathcal{F}$ induces a policy class $\Pi_{\mathcal{F}} := \{\pi_f : f \in \mathcal{F}\}$, and we take our policy class $\Pi = \Pi_{\mathcal{F}}$ for the remaining part of this paper.

The complexity of the function class is measured by the covering number.

Definition 8.3 (Covering number). *For a function class $\mathcal{H} \subseteq (\mathcal{X} \to \mathbb{R})$ and parameter $\alpha \geq 0$, an α-covering of $\mathcal{H}$ (with respect to the sup norm) is a subset $\mathcal{H}' \subseteq \mathcal{H}$ such that for any $f \in \mathcal{H}$, there exists $f' \in \mathcal{H}'$ with $\sup_{x\in\mathcal{X}} |f(x) - f'(x)| \leq \alpha$. We define the α-covering number of $\mathcal{H}$ as $N(\mathcal{H}, \alpha) := \min\{|\mathcal{H}'| : \mathcal{H}' \text{ is a } \alpha\text{-covering of } \mathcal{H}\}$.*

Bellman completeness A reward function $R = (R_1, \cdots, R_H) \in \mathcal{R}$ induces a Bellman operator as

$$[\mathcal{T}_{R,h} f](s,a) := R_h(s,a) + \mathbb{E}_{s'\sim\mathbb{T}_h(\cdot|s,a)} \max_{a'\in\mathcal{A}} f_{h+1}(s',a'), \qquad \forall f = (f_1, \cdots, f_H) \in \mathcal{F},$$

where we also adopt the notation $f_{H+1} = 0$ for any $f \in \mathcal{F}$. Most literature on RL with general function approximation also makes use of a richer comparator function class $\mathcal{G} = \mathcal{G}_1 \times \cdots \times \mathcal{G}_H$ that satisfies the following *Bellman completeness* [142, 153, 200, etc.].

Assumption 8.4 (Bellman completeness). *For each $h \in [H]$, $\mathcal{F}_h \subseteq \mathcal{G}_h$. For any $f \in \mathcal{F}$ and $R \in \mathcal{R}$, it holds $\inf_{g_h\in\mathcal{G}_h} \|\mathcal{T}_{R,h} f_{h+1} - g_h\|_\infty \leq \varepsilon_{\mathrm{app}}$ for $h \in [H]$.*

[1] In the following, we always write $R^\star$ for the true reward function to avoid confusion.

Miscellaneous notation For any $p \in [0,1]$, we define $\mathrm{Bern}(p)$ to be the Bernoulli distribution with $\mathbb{P}(X=1)=p$. For functions f and $g \geq 0$, we use $f = O(g)$ to denote that there exists a universal constant C such that $f \leq C \cdot g$.

8.3 Sample-Efficient Online RL with Outcome Reward

In this section, we present a model-free RL algorithm with outcome reward, which achieves sample complexity guarantee scaling with the coverability coefficient and the log-covering number of the function classes.

8.3.1 Main Result

We present Algorithm 2, which is based on the principle of optimism. For simplicity of presentation, we assume that the initial state s_1 is fixed.

The crux of the proposed algorithm is a new method for performing Fitted-Q Iteration with only outcome reward, in contrast to most existing RL algorithms (with general function approximation) that make use the process reward $(r_1, \cdots, r_H)$ to fit the Q-function for each step [200, 201, etc.]. A natural first idea is to fit, given a dataset $\mathcal{D} = \{(\tau, r)\}$ consisting of previously observed (trajectory, outcome reward) pairs, a reward model from the reward function class $\mathcal{R}$ by optimizing the following reward model loss:

$$\mathcal{L}_{\mathcal{D}}^{\mathsf{RM}}(R) := \sum_{(\tau,r)\in\mathcal{D}} \left(\sum_{h=1}^{H} R_h(s_h, a_h) - r \right)^2. \tag{8.3}$$

As discussed below in Theorem 8.6, directly fitting an estimated reward model based on $\mathcal{L}_{\mathcal{D}}^{\mathsf{RM}}$ can lead to bad performance. Instead, our algorithm jointly optimizes over the value functions and reward models, as detailed below.

For any proxy reward model $R \in \mathcal{R}$ and a value function $f \in \mathcal{F}$, we define the Bellman error at step $h \in [H]$ as

$$\mathcal{E}_{\mathcal{D},h}(f_h, f_{h+1}; R) := \sum_{(\tau,r)\in\mathcal{D}} \Big(f_h(s_h, a_h) - R_h(s_h, a_h) - \max_{a'} f_{h+1}(s_{h+1}, a') \Big)^2, \tag{8.4}$$

a measure of violation of the Bellman equation (8.1) with the proxy reward model R. Then, we introduce the Bellman loss defined as

$$\mathcal{L}_{\mathcal{D}}^{\mathsf{BE}}(f; R) := \sum_{h=1}^{H} \mathcal{E}_{\mathcal{D},h}(f_h, f_{h+1}; R) - \inf_{g\in\mathcal{G}} \sum_{h=1}^{H} \mathcal{E}_{\mathcal{D},h}(g_h, f_{h+1}; R), \tag{8.5}$$

where we subtract the infimum of $g \in \mathcal{G}$ over the helper function class $\mathcal{G}$, a common approach to overcoming the double-sampling problem [147, 200, 202].

Algorithm 2 Outcome-Based Exploration with Optimism

input: Q-function class $\mathcal{F}$, reward function class $\mathcal{R}$, comparator class $\mathcal{G}$, parameter $\lambda > 0$, reference policy π_{ref}.

initialize: $\mathcal{D} \leftarrow \emptyset$.

1: **for** $t = 1, 2, \ldots, T$ **do**
2: Compute the optimistic estimates: $(f^{(t)}, R^{(t)}) = \max_{f \in \mathcal{F}, R \in \mathcal{R}} \lambda f_1(s_1) - \mathcal{L}_{\mathcal{D}}^{\mathsf{BE}}(f; R) - \mathcal{L}_{\mathcal{D}}^{\mathsf{RM}}(R)$.
3: Select policy $\pi^{(t)} \leftarrow \pi_{f^{(t)}}$.
4: **for** $h = 1, 2, \cdots, H$ **do**
5: Execute $\pi^{(t)} \circ_h \pi_{\text{ref}}$ for one episode and obtain $(\tau^{(t,h)}, r^{(t,h)})$
6: Update dataset: $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\tau^{(t,h)}, r^{(t,h)})\}$.
7: **end for**
8: **end for**
9: Output $\widehat{\pi} = \text{Unif}(\pi^{(1:T)})$.

Algorithm First fix an arbitrary policy π_{ref} (can be the policy which takes an arbitrary action a_0 at al states). The proposed algorithm takes in a value function class $\mathcal{F}$, a reward function class $\mathcal{R}$ and a comparator function class $\mathcal{G}$, and performs the following two steps for each iteration $t = 1, 2, \cdots, T$:

1. (Optimism) Compute optimistic estimates of $(Q^\star, R^\star)$ through solving the following joint maximization problem with the dataset $\mathcal{D}$ consisting of all previously observed (trajectory, outcome reward) pairs:

$$(f^{(t)}, R^{(t)}) = \max_{f \in \mathcal{F}, R \in \mathcal{R}} \lambda f_1(s_1) - \mathcal{L}_{\mathcal{D}}^{\mathsf{BE}}(f; R) - \mathcal{L}_{\mathcal{D}}^{\mathsf{RM}}(R), \tag{8.6}$$

   where for any $f \in \mathcal{F}$ we denote $f_1(s_1) := \max_{a \in \mathcal{A}} f_1(s_1, a)$ to be the value of f at the initial state. Therefore, the optimization problem (8.6) enforces optimism by balancing the estimated value $f_1(s_1)$ and the estimation error $\mathcal{L}_{\mathcal{D}}^{\mathsf{BE}}(f; R) + \mathcal{L}_{\mathcal{D}}^{\mathsf{RM}}(R)$ though a hyper-parameter $\lambda \geq 0$.

2. (Data collection) Based on the optimism estimate $f^{(t)}$, the algorithm selects $\pi^{(t)} := \pi_{f^{(t)}}$. To collect data, the algorithm then executes the exploration policies $\pi \circ_h \pi_{\text{ref}}$ for each $h \in [H]$, where for any policy π and π_{ref}, we let $\pi \circ_h \pi_{\text{ref}}$ be the policy that executes π for the first h steps, and then executes π_{ref} starting at the $(h+1)$-th step.

Theoretical analysis For Algorithm 2, we provide the following sample complexity guarantee, which scales with the coverability $C_{\text{cov}} = C_{\text{cov}}(\Pi_{\mathcal{F}})$, where $\Pi_{\mathcal{F}} = \{\pi_f : f \in \mathcal{F}\}$ is the policy class induced by $\mathcal{F}$. To simplify the presentation, we denote $\log N_T := \inf_{\alpha \geq 0} (\log N(\alpha) + T\alpha)$, where $N(\alpha)$ is defined as

$$N(\alpha) := \max_{h \in [H]} \{N(\mathcal{F}_h, \alpha), N(\mathcal{R}_h, \alpha), N(\mathcal{G}_h, \alpha)\}.$$

With the function classes being parametric, it is clear that $\log N_T \leq O(d \log(T))$.

Theorem 8.5. *Let $\delta \in (0,1)$. Suppose that Theorem 8.2 and Theorem 8.4 hold, and the parameters are chosen as*

$$\lambda = c_0 \max\left\{\frac{H\log(N_{TH^2}/\delta)}{\varepsilon}, TH\varepsilon_{\mathrm{app}}\right\}, \qquad T \geq c_1 \frac{C_{\mathrm{cov}} H^2 \log(T)}{\varepsilon^2} \cdot \log(N_{TH^2}/\delta), \tag{8.7}$$

where $c_0, c_1 > 0$ are absolute constants. Then with probability at least $1-\delta$, the output policy $\widehat{\pi}$ of Algorithm 2 satisfies $V^\star(s_1) - V^{\widehat{\pi}}(s_1) \leq \varepsilon + O\left(C_{\mathrm{cov}} H^2 \log(T) \cdot \varepsilon_{\mathrm{app}}\right)$.

The proof of Theorem 8.5 is deferred to appendix G.3. Particularly, we note that when the function classes satisfy $\log N_T \leq \widetilde{O}(d)$ and $\varepsilon_{\mathrm{app}} = 0$, Algorithm 2 outputs an ε-optimal policy with sample complexity

$$TH \leq \widetilde{O}\left(\frac{C_{\mathrm{cov}} d H^3}{\varepsilon^2}\right).$$

Notably, the coverability C_{cov} measures the inherent complexity of the underlying MDP $M^\star$ [153] and it is independent of the reward function class. As our result only depends on the coverability C_{cov} and the statistical complexity of the function classes, it does *not* rely on the structure of reward functions, while previous works assume the reward functions are either linear [27, 30] or admit low *trajectory* eluder dimension [32, 33].

Remark 8.6. *In Algorithm 2, the reward functions $R_{(t)}$ and Q-functions $f_{(t)}$ are jointly optimized (see eq. (8.6)). A natural question is whether these can be optimized separately—i.e., first learning a fitted reward model and then applying optimism to the Q-functions based on the learned reward model. We show that this decoupled approach can lead to failures: due to reward model mismatch, the algorithm may become 'trapped' in regions where the exploratory policy fails to gather informative data. As a result, the sample complexity can become infinite in the worst case. See appendix G.6.1 in the appendix for details.*

8.3.2 A Simpler Algorithm for Deterministic MDPs

A disadvantage of Algorithm 2 is that it requires solving a max-min optimization problem (8.6), as the Bellman loss $\mathcal{L}_{\mathcal{D}}^{\mathsf{BE}}$ involves a minimization problem over $\mathcal{G}$. While such computationally inefficient optimization problems are the common subroutines of existing function approximation RL algorithms [33, 200, 203, 204, etc.], it turns out that Algorithm 2 can be significantly simplified when the transition dynamics in underlying MDP are deterministic.

Assumption 8.7. *The transition kernel $\mathbb{T}$ is deterministic, i.e., for any $h \in [H]$ and $s_h \in \mathcal{S}, a_h \in \mathcal{A}$, there is a unique state $s_{h+1} \in \mathcal{S}$ such that $\mathbb{T}_h(s_{h+1} \mid s_h, a_h) = 1$.*

Note that in this setting, the initial state s_1 and the outcome reward r can still be random. This setting is also referred to as *Deterministic Contextual MDP* in Xie et al. [177].

Value difference as reward model A key observation is that, when the underlying MDP $M^\star$ is deterministic, the Bellman equation (8.1) trivially reduces to the following equality

$$Q_h^\star(s_h, a_h) = R_h^\star(s_h, a_h) + V_{h+1}^\star(s_{h+1}),$$

Algorithm 3 Outcome-Based Exploration with Optimism for Determinsitic MDP

input: Function class $\mathcal{F}$, parameter $\lambda > 0$.
initialize: $\mathcal{D} \leftarrow \emptyset$, initial estimate $f^{(1)} \in \mathcal{F}$.

1: **for** $t = 1, 2, \ldots, T$ **do**
2: Receive $s^{(t)}$ and compute the optimistic estimates: $f^{(t)} = \max_{f \in \mathcal{F}} \lambda f_1(s_1^{(t)}) - \mathcal{L}_{\mathcal{D}}^{\mathsf{BR}}(f)$.
3: Select policy $\pi^{(t)} \leftarrow \pi_{f^{(t)}}$.
4: Execute $\pi^{(t)}$ to obtain a trajectory $\tau^{(t)} = (s_1^{(t)}, a_1^{(t)}, \ldots, s_H^{(t)}, a_H^{(t)})$ with outcome reward $r^{(t)}$.
5: Update dataset: $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\tau^{(t)}, r^{(t)})\}$.
6: **end for**
7: Output $\widehat{\pi} = \mathrm{Unif}(\pi^{(1:T)})$.

which holds almost surely. Hence, for any trajectory τ, it holds that

$$R^\star(\tau) = \sum_{h=1}^{H} R_h^\star(s_h, a_h) = \sum_{h=1}^{H} \left[Q_h^\star(s_h, a_h) - V_{h+1}^\star(s_{h+1}) \right].$$

Therefore, any value function $f \in \mathcal{F}$ induces an outcome reward model $R^f : (\mathcal{S} \times \mathcal{A})^H \to \mathbb{R}$ defined as

$$R^f(\tau) = \sum_{h=1}^{H} \left[f_h(s_h, a_h) - f_{h+1}(s_{h+1}) \right],$$

where we adopt the notation $f_h(s) := \max_{a \in \mathcal{A}} f_h(s, a)$ for $h \in [H]$. This observation motivates the following Bellman Residual loss:

$$\mathcal{L}_{\mathcal{D}}^{\mathsf{BR}}(f) := \sum_{(\tau, r) \in \mathcal{D}} \left(\sum_{h=1}^{H} \left[f_h(s_h, a_h) - f_{h+1}(s_{h+1}) \right] - r \right)^2, \tag{8.8}$$

where $\mathcal{D} = \{(\tau, r)\}$ is any dataset consisting of (trajectory, outcome reward) pairs.

Bellman Residual Minimization (BRM) with Optimism For deterministic MDP, we propose Algorithm 3 as a simplification of our main algorithm. Similar to Algorithm 2, the proposed algorithm takes in the value function class $\mathcal{F}$ and alternates between the following two steps for each round $t = 1, 2, \cdots, T$:

1. (Optimism) Compute optimistic estimates of $Q^\star$ through solving the following maximization problem with the dataset $\mathcal{D}$ consisting of all previously observed (trajectory, outcome reward) pairs:

$$f^{(t)} = \max_{f \in \mathcal{F}} \lambda f_1(s_1) - \mathcal{L}_{\mathcal{D}}^{\mathsf{BR}}(f), \tag{8.9}$$

enforcing optimism by balancing the estimated value $f_1(s_1)$ and the Bellman residual loss $\mathcal{L}_{\mathcal{D}}^{\mathsf{BR}}(f)$.

2. (Data collection) Based on the optimistic estimate $f^{(t)}$, selects $\pi^{(t)} := \pi_{f^{(t)}}$ and collect a trajectory $\tau^{(t)} = (s_1^{(t)}, a_1^{(t)}, \ldots, s_H^{(t)}, a_H^{(t)})$ with outcome reward $r^{(t)}$.

Compared to Algorithm 2, Algorithm 3 has the several advantages. First, it does not rely on the reward function class $\mathcal{R}$ and the comparator function class $\mathcal{G}$, and the Bellman residual loss $\mathcal{L}_{\mathcal{D}}^{\mathsf{BR}}$ is much simpler than the Bellman loss $\mathcal{L}_{\mathcal{D}}^{\mathsf{BE}}$, thanks to the deterministic nature of the underlying MDP. Therefore, Algorithm 3 is more amenable to computationally efficient implementation, as it replaces the max-min optimization problem (8.6) in Algorithm 2 with a much simpler maximization problem (8.9). Further, for every round t, the algorithm only needs to collect one episode from the greedy policy $\pi^{(t)}$.

Theoretical analysis We present the upper bound of Algorithm 3 in terms of the following notion of coverability,

$$C'_{\mathrm{cov}}(\Pi) := \mathbb{E}_{s_1 \sim \rho} C_{\mathrm{cov}}(\Pi; M^\star_{s_1}),$$

where $M^\star$ is the underlying MDP, $M^\star_{s_1}$ is the MDP with deterministic initial state s_1 and the same transition dynamics as $M^\star$, and $\Pi = \Pi_{\mathcal{F}}$ is the policy class induced by $\mathcal{F}$. In general, $C'_{\mathrm{cov}}(\Pi)$ is always an upper bound on the coverability $C_{\mathrm{cov}}(\Pi)$, and the guarantee of Algorithm 3 scales with $C'_{\mathrm{cov}}(\Pi)$ as it avoids the layer-wise exploration strategy of Algorithm 2. We also denote

$$\log N_{\mathcal{F},T} := \inf_{\alpha \geq 0} \left(\max_{h \in [H]} N(\mathcal{F}_h, \alpha) + T\alpha \right).$$

Theorem 8.8. *Let $\delta \in (0,1)$. Suppose that Theorem 8.2 holds, and the parameters are chosen as*

$$\lambda = c_0 \max\left\{ \frac{H^3 \log(N_{\mathcal{F},T}/\delta)}{\varepsilon}, T\varepsilon_{\mathrm{app}} \right\}, \quad T \geq c_1 \frac{C'_{\mathrm{cov}}(\Pi) H^4 \log(T)}{\varepsilon^2} \cdot \log(N_{\mathcal{F},T}/\delta), \tag{8.10}$$

where $c_0, c_1 > 0$ are absolute constants. Then with probability at least $1-\delta$, Algorithm 3 achieves

$$\frac{1}{T} \sum_{t=1}^{T} \left(V^\star(s_1^{(t)}) - V^{\pi^{(t)}}(s_1^{(t)}) \right) \leq \varepsilon + O\left(C'_{\mathrm{cov}}(\Pi) H \log(T) \cdot \varepsilon_{\mathrm{app}} \right).$$

The above upper bound provides the PAC guarantee through the standard online-to-batch conversion, and its proof is deferred to appendix G.4. It is worth noting that Theorem 8.8 only relies on realizability assumption on the Q-function class $\mathcal{F}$, significantly relaxing the assumptions of realizablity (Theorem 8.2) and completeness (Theorem 8.4) in Theorem 8.5.

8.4 Preference-based Reinforcement Learning

The goal of preference-based learning is to find a near-optimal policy only through interacting with the environment that provides *preference feedback*. As an extension of our results presented in section 8.3, in this section we present a similar algorithm for preference-based RL with the same sample complexity guarantee.

Preference-based learning in MDP In preference-based RL, the interaction protocol of the learner with the environment is specified as follows. For each round $t = 1, 2, \cdots$,

- The learner selects policy $\pi^{(t,+)}$ and $\pi^{(t,-)}$.
- The learner receives trajectories $\tau^{(t,+)} \sim \pi^{(t,+)}$, $\tau^{(t,-)} \sim \pi^{(t,-)}$, and *preference feedback* $y^{(t)} \sim \mathrm{Bern}\left(\mathsf{C}(\tau^{(t,+)}, \tau^{(t,-)})\right)$, where C is a comparison function.

Intuitively, for any trajectory pair (τ^+, τ^-), the comparison function $\mathsf{C}(\tau^+, \tau^-) = \mathbb{P}(\tau^+ \succ \tau^-)$ measures the probability that τ^+ is more preferred. In this paper, we mainly focus on the Bradley-Terry-Luce (BTR) model [178], which is widely used on RLHF literature. We expect that our algorithm and analysis techniques apply to a broader class of preference models.

Definition 8.9 (BTR model)**.** *The comparison function* C *is specified as*

$$\mathsf{C}(\tau^+, \tau^-) = \frac{\exp\left(\beta R^\star(\tau^+)\right)}{\exp\left(\beta R^\star(\tau^+)\right) + \exp\left(\beta R^\star(\tau^-)\right)},$$

where $R^\star$ *is the ground-truth reward function,* $\beta > 0$ *is a parameter.*

Under BTR model, the preference feedback in fact contains information of the outcome rewards. Hence, in this sense, preference-based RL can be regarded as an extension of outcome-based RL with weaker feedback.

Algorithm for preference-based RL To extend Algorithm 2, we need to modify the reward model loss $\mathcal{L}_{\mathcal{D}}^{\mathsf{RM}}$ (defined in (8.3)) to incorporate preference feedback. For any dataset $\mathcal{D} = \{(\tau^+, \tau^-, y)\}$ consisting of (trajectories, preference) pair, we introduce the following preference-based reward model loss $\mathcal{L}_{\mathcal{D}}^{\mathsf{PbRM}}$:

$$\mathcal{L}_{\mathcal{D}}^{\mathsf{PbRM}}(R) := \sum_{(\tau^+, \tau^-, y) \in \mathcal{D}} L\left(R(\tau^+) - R(\tau^-), y\right), \tag{8.11}$$

where $L(w, y) := -\beta w y + \log(1 + e^{\beta w})$ is the logistic loss. It is well-known that under BTR model (Theorem 8.9), the ground-truth reward $R^\star$ is the population minimizer of $\mathcal{L}_{\mathcal{D}}^{\mathsf{PbRM}}$, and any approximate minimizer of $\mathcal{L}_{\mathcal{D}}^{\mathsf{PbRM}}$ can serve as a proxy for $R^\star$. Therefore, with the loss function $\mathcal{L}_{\mathcal{D}}^{\mathsf{PbRM}}$, we propose the following algorithm (Algorithm 5, detailed description in appendix G.5), which generalizes Algorithm 2 to handle preference feedback: For each iteration $t = 1, 2, \cdots, T$, the algorithm performs the following two steps.

1. (Optimism) Compute optimistic estimates of $(Q^\star, R^\star)$ through solving the following joint maximization problem with the dataset $\mathcal{D}$ consisting of all previously observed (trajectories, feedback) pairs:

$$(f^{(t)}, R^{(t)}) = \max_{f \in \mathcal{F}, R \in \mathcal{R}} \lambda \left[f_1(s_1) - \widehat{V}_{\mathcal{D},R}^{\mathrm{ref}} \right] - \mathcal{L}_{\mathcal{D}}^{\mathsf{BE}}(f; R) - \mathcal{L}_{\mathcal{D}}^{\mathsf{PbRM}}(R), \tag{8.12}$$

   where the Bellman loss $\mathcal{L}_{\mathcal{D}}^{\mathsf{BE}}$ is defined in (8.5), $\widehat{V}_{\mathcal{D},R}^{\mathrm{ref}}$ is the estimated value function of π_{ref} defined as

$$\widehat{V}_{\mathcal{D},R}^{\mathrm{ref}} := \frac{1}{|\mathcal{D}|} \sum_{(\tau^+, \tau^-, y) \in \mathcal{D}} R(\tau^-). \tag{8.13}$$

The term $f_1(s_1) - \widehat{V}^{\mathrm{ref}}_{\mathcal{D},R}$ can be regarded as an estimate of the advantage of π_f over π_{ref} under (f, R). It is introduced to avoid over-estimating the optimal value, as the preference feedback only provide information between the *difference* between two trajectories.

2. (Data collection) The algorithm selects the greedy policy $\pi^{(t)} := \pi_{f^{(t)}}$. For each $h \in [H]$, the algorithm sets $\pi^{(t,h,+)} := \pi \circ_h \pi_{\mathrm{ref}}$ and $\pi^{(t,h,-)} := \pi_{\mathrm{ref}}$, executes $(\pi^{(t,h,+)}, \pi^{(t,h,-)})$ to collects trajectories $(\tau^{(t,h,+)}, \tau^{(t,h,-)})$ and the preference feedback $y^{(t,h)}$.

We provide the following sample complexity guarantee of the algorithm above.

Theorem 8.10. *Let $\delta \in (0,1)$. Suppose that Theorem 8.2 and Theorem 8.4 hold, and the parameters of Algorithm 5 are chosen as*

$$\lambda = c_0 \max\left\{\frac{H\log(N_{TH^2}/\delta)}{\varepsilon}, TH\varepsilon_{\mathrm{app}}\right\}, \qquad T \geq \widetilde{O}\left(\frac{C_{\mathrm{cov}}H^2}{\varepsilon^2}\cdot\log(N_{TH^2}/\delta)\right), \tag{8.14}$$

where $c_0 > 0$ is an absolute constant, and $\widetilde{O}(\cdot)$ omits poly-logarithmic factors and constant depending on β. Then, with probability at least $1-\delta$, the output policy $\widehat{\pi}$ of Algorithm 5 satisfies

$$V^\star(s_1) - V^{\widehat{\pi}}(s_1) \leq \varepsilon + \widetilde{O}\left(C_{\mathrm{cov}}H^2\varepsilon_{\mathrm{app}}\right).$$

The proof of Theorem 8.10 is deferred to appendix G.5.1.

8.5 Lower Bounds

As shown by Theorem 8.5, with bounded coverability of the MDP and appropriate assumptions on the function classes, finding a near-optimal policy within a polynomial number of episodes with outcome rewards is possible. In this setting, our sample complexity bounds match the sample complexity of Algorithm GOLF [153, 200], up to a factor of $\widetilde{O}(H)$. This indicates that, under bounded coverability, learning with outcome-based rewards is *almost* statistically equivalent to learning with process rewards.

Additionally, in the setting of offline reinforcement learning with bounded *uniform concentrability*, the results of Jia, Rakhlin, and Xie [199] indicate that there is also a statistical equivalence between learning with outcome rewards and learning with process rewards. Therefore, it is natural to ask the following question:

Is learning with outcome rewards always statistically equivalent to learning with process rewards in RL?

However, it turns out that the answer is *negative* if the statistical complexity is measured with respect to the *structure* of the value (reward) function classes.

More specifically, we construct a class of MDPs with horizon $H = 2$, *known* transition $\mathbb{T}$, and d-dimensional *generalized linear* reward models (appendix G.6.2). With process reward feedback, such a problem is known to be *easy* as it admits low (Bellman) eluder dimension [200, 205, etc.], and existing algorithms can learn an ε-optimal policy using $\widetilde{O}(d^2/\varepsilon^2)$ episodes with process rewards. However, given only access to outcome rewards, we show that this

problem is *as hard as* learning ReLU linear bandits [206, 207], and hence it requires at least $e^{\Omega(d)}$ episodes to learn. Hence, in this setting, there is an exponential separation between learning with process rewards and learning with outcome-based rewards.

Theorem 8.11. *For any positive integer $d \geq 1$, there exists a class $\mathcal{M}$ of two-layer MDPs with a fixed transition kernel $\mathbb{T}$ and initial state s_1, such that the following holds:*

(a) There exists an algorithm that, for any MDP $M^\star \in \mathcal{M}$ and any $\varepsilon \in (0,1)$, given access to process reward *feedback, returns an ε-optimal policy with high probability using $\widetilde{O}\left(d^2/\varepsilon^2\right)$ episodes.*

(b) Suppose that there exists an algorithm that, for any MDP $M^\star \in \mathcal{M}$, given only access to outcome reward*, returns a 0.1-optimal policy with probability at least $\frac{3}{4}$ using T episodes. Then it must hold that $T = \Omega(e^{c_1 d})$, where c_1 is an absolute constant.*

This exponential separation demonstrates that the delicate analysis based on well-behaved Bellman errors [200, 201, 208, etc.] crucially relies on the process reward feedback, and the resulting guarantees might not be preserved in the setting where only outcome reward feedback is available.

8.6 Conclusion

In this work, we develop a model-free, sample-efficient algorithm for outcome-based reinforcement learning that relies solely on trajectory-level rewards and achieves theoretical guarantees bounded by coverability under function approximation. From the lower bound side, we show that joint optimization of reward and value functions is essential, and establish a fundamental exponential gap between outcome-based and per-step feedback. For deterministic MDPs, we propose a simpler, more efficient variant, and extend our approach to preference-based feedback, demonstrating that it preserves the same statistical efficiency.

Appendix A

Proofs for Chapter 2

This appendix contains the detailed proofs for the results presented in Chapter 2, including the lower and upper bounds for the 2-Wasserstein distance and the KL divergence bounds.

A.1 Proofs for Chapter 3 Main Theorems

A.1.1 Proof of the Lower Bound in Theorem 2.4

The main idea of the lower bound is that already simple binary Gaussian mixtures demonstrate why for $K > \sigma$ the rate of smoothed Wasserstein convergence should slow down.

Proposition A.1. *Fix $K > \sigma > 0$. For every $h > 0$ we consider a Bernoulli distribution $\mathbb{P}_h = (1 - p_h)\delta_0 + p_h\delta_h$, with $p = e^{-h^2/(2K^2)}$. Then for any $\varepsilon > 0$ we have*

$$\sup_h \mathbb{E}\left[W_2(\mathbb{P}_h * \mathcal{N}(0, \sigma^2), \mathbb{P}_{h,n} * \mathcal{N}(0, \sigma^2))\right] = \Omega\left(n^{-\alpha-\varepsilon}\right),$$

where $\mathbb{P}_{h,n}$ is the empirical measure constructed from n i.i.d. samples from $\mathbb{P}_h$, and α is defined in (2.6).

The proof of Proposition A.1 and the construction of the multi-point distribution that provides the lim sup result in Theorem 2.4 follow from careful analysis of the CDF differences and the use of the Berry-Esseen Theorem.

A.1.2 Proof of the Upper Bound of Theorem 2.4

We start the proof from the following observation Villani [6]: for two distributions $\mathbb{Q}_1, \mathbb{Q}_2$ on $\mathbb{R}$ with absolutely continuous CDFs $F_1(x), F_2(x)$ the optimal coupling for the 2-Wasserstein distance is given by $F_2^{-1}(F_1(x))$, implying an explicit formula:

$$W_2(\mathbb{Q}_1, \mathbb{Q}_2)^2 = \mathbb{E}[\left|F_2^{-1}(F_1(T_1)) - T_1\right|^2], \quad T_1 \sim \mathbb{Q}_1 .$$

In our case we set $\mathbb{Q}_1 = \mathbb{P} * \mathcal{N}(0, 1)$ and $\mathbb{Q}_2 = \mathbb{P}_n * \mathcal{N}(0, 1)$. The proof proceeds via conditioning on the subgaussianity of the empirical measure, truncating the evaluation range, and applying a concentration bound on $|F_n(t) - F(t)|$ relative to the PDF $\rho(t)$.

A.2 KL Divergence and Rényi Mutual Information

A.2.1 Proof of Theorem 2.5

To prove Theorem 2.5, we leverage Rényi divergence D_λ and the relationship established in Van Erven and Harremos [209] that $\lim_{\lambda \to 1} D_\lambda = D_{KL}$. We show that for K-subgaussian distributions, the order-λ Rényi mutual information $I_\lambda(X;Y)$ is bounded by $O(\log(1/(2-\lambda)))$. By choosing $\lambda = 2 - 1/\log n$, we recover the $O(\frac{(\log n)^d}{n})$ rate.

A.3 Supplemental Examples and Calculations

A.3.1 Non-Existence of Uniform Bound for LSI Constants

We prove that for the Bernoulli distribution class $\mathbb{P}_h = (1-p_h)\delta_0 + p_h\delta_h$, the constants in the log-Sobolev inequalities (LSI) do not have a uniform bound when $\sigma < K$. This corresponds to Corollary 1.1 in Chapter 2.

Proof Sketch. By constructing test functions f_h that transition between the two peaks of the smoothed Bernoulli measure, we show that the LSI constant C_h must grow with h to satisfy the inequality $\int f^2 \log f^2 d\mu \leq C_h \int |f'|^2 d\mu$. Specifically, as $h \to \infty$, the gap between the distribution's modes becomes too large for a fixed constant to hold. □

Appendix B

Proofs for Chapter 3

This appendix contains the detailed proofs for the results presented in Chapter 3.2, including dimensionality results, subgaussian dichotomies, and comparisons for other distance metrics.

B.1 Proof of Theorem 3.2

First of all, we notice that [57, Lemma 5] shows that for any $\delta, \lambda > 1$ such that

$$0 < \delta < \exp(-1/2) \quad \text{and} \quad \log\log(1/\delta)/\log(1/\delta) \leq (\lambda - 1)/2, \tag{B.1}$$

and any probability measures $\mathbb{P}, \mathbb{Q}, \mathbb{S}$ and $\mathbb{Q}' = (1-\delta)\mathbb{Q} + \delta\mathbb{S}$, we have

$$D_{\mathrm{KL}}(\mathbb{P}\|\mathbb{Q}') \leq \frac{2\log(1/\delta)}{(1-\delta)^2} H^2(\mathbb{P}, \mathbb{Q}) + \frac{4\delta\log(1/\delta)}{(1-\delta)^2} + \delta^{\frac{\lambda-1}{2}} \cdot \int_{\mathbb{R}^d} \frac{(d\mathbb{P})^\lambda}{(d\mathbb{S})^{\lambda-1}}. \tag{B.2}$$

Let $\lambda = 3$, then as long as $0 < \delta < 1/2$, (B.1) holds. Choose, $\mathbb{P} = f_\pi$ and $\mathbb{S} = \mathbb{Q} = f_\eta$, we get

$$D_{\mathrm{KL}}(f_\pi\|f_\eta) \leq \frac{2\log(1/\delta)}{(1-\delta)^2} H^2(f_\pi, f_\eta) + \frac{4\delta\log(1/\delta)}{(1-\delta)^2} + \delta \cdot \int_{\mathbb{R}^d} \frac{f_\pi(\mathbf{x})^3}{f_\eta(\mathbf{x})^2} d\mathbf{x}. \tag{B.3}$$

Notice that the last term is an f-divergence with $f = x^3$, which is a convex function, hence $\int_{\mathbb{R}^d} \frac{f_\pi(\mathbf{x})^3}{f_\eta(\mathbf{x})^2} d\mathbf{x}$ is convex in (f_π, f_η). Define the set $\mathscr{P}(M)$ of Gaussian mixtures with mixing distributions supported on the ball $B_2(M)$:

$$\mathscr{P}(M) = \{\pi * \mathcal{N}(0, I_d) : \mathtt{supp}(\pi) \subset B_2(M)\}$$

which is a convex set. Hence the maximum value of $\int_{\mathbb{R}^d} \frac{f_\pi(\mathbf{x})^3}{f_\eta(\mathbf{x})^2} d\mathbf{x}$ where $f_\pi, f_\eta \in \mathscr{P}(M)$ is attained when f_π, f_η are both at the boundary of $\mathscr{P}(M)$, i.e. $\exists \mathbf{u}, \mathbf{v}$ with $\|\mathbf{u}\|_2, \|\mathbf{v}\|_2 \leq M$ and we have $f_\pi = \delta_{\mathbf{u}} * \mathcal{N}(0, I_d)$, $f_\eta = \delta_{\mathbf{v}} * \mathcal{N}(0, I_d)$. Therefore, we have

$$\int_{\mathbb{R}^d} \frac{f_\pi(\mathbf{x})^3}{f_\eta(\mathbf{x})^2} d\mathbf{x} \leq \sup_{\mathbf{u},\mathbf{v}:\|\mathbf{u}\|,\|\mathbf{v}\|_2 \leq M} \int_{\mathbb{R}^d} \frac{\left(\frac{1}{\sqrt{2\pi}^d} \exp\left(-\frac{\|\mathbf{x}+\mathbf{u}\|_2^2}{2}\right)\right)^3}{\left(\frac{1}{\sqrt{2\pi}^d} \exp\left(-\frac{\|\mathbf{x}+\mathbf{v}\|_2^2}{2}\right)\right)^2} d\mathbf{x}$$

$$
\begin{aligned}
&= \sup_{\mathbf{u},\mathbf{v}:\|\mathbf{u}\|,\|\mathbf{v}\|_2\le M} \frac{1}{\sqrt{2\pi}^d} \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}\left(\mathbf{x}^T\mathbf{x} + 6\mathbf{x}^T\mathbf{u} - 4\mathbf{x}^T\mathbf{v} + 3\mathbf{u}^T\mathbf{u} - 2\mathbf{v}^T\mathbf{v}\right)\right) d\mathbf{x} \\
&= \sup_{\mathbf{u},\mathbf{v}:\|\mathbf{u}\|,\|\mathbf{v}\|_2\le M} \frac{1}{\sqrt{2\pi}^d} \exp\left(3\|\mathbf{u}-\mathbf{v}\|_2^2\right) \cdot \int_{\mathbb{R}^d} \exp\left(-\frac{\|\mathbf{x}+3\mathbf{u}-2\mathbf{v}\|_2^2}{2}\right) d\mathbf{x} \\
&= \sup_{\mathbf{u},\mathbf{v}:\|\mathbf{u}\|,\|\mathbf{v}\|_2\le M} \exp\left(3\|\mathbf{u}-\mathbf{v}\|_2^2\right) = \exp\left(12M^2\right).
\end{aligned}
$$

Therefore, according to (B.3), we have for any $\delta \in [0, 1/2]$,

$$
D_{\mathrm{KL}}(f_\pi \| f_\eta) \le \frac{2\log(1/\delta)}{(1-\delta)^2} H^2(f_\pi, f_\eta) + \frac{4\delta\log(1/\delta)}{(1-\delta)^2} + \exp(12M^2)\delta.
$$

Choosing $\delta = \exp(-12M^2)H^2(f_\pi, f_\eta) \in [0, 1/2]$ and noticing that $(1-\delta)^2 \ge \frac{1}{2}$, we get

$$
\begin{aligned}
D_{\mathrm{KL}}(f_\pi \| f_\eta) &\le H^2(f_\pi, f_\eta) + 96M^2H^2(f_\pi, f_\eta) + 16H^2(f_\pi, f_\eta)\log\frac{1}{H^2(f_\pi, f_\eta)} \\
&\le 97M^2H^2(f_\pi, f_\eta) + 16H^2(f_\pi, f_\eta)\log\frac{1}{H^2(f_\pi, f_\eta)}.
\end{aligned}
$$

This finishes the proof of Theorem 3.2.

B.2 Proof of Corollary 3.9

For convenience, denote by $\tilde{R}_{H^2,n}$ the minimax squared Hellinger risk for improper density estimation, similar to $\tilde{R}_{KL,n}$. First, notice that for $\mathcal{P} \subset \mathcal{P}_{com}(M)$ or $\mathcal{P} \subset \mathcal{P}_{sub}(K)$

$$
R_{H^2,n}(\mathcal{P}) \le \tilde{R}_{H^2,n}(\mathcal{P}), \quad R_{KL,n}(\mathcal{P}) \le \tilde{R}_{KL,n}(\mathcal{P}),
$$

and also

$$
R_{H^2,n}(\mathcal{P}) \lesssim R_{KL,n}(\mathcal{P})
$$

since $H^2(\mathbb{P}, \mathbb{Q}) \le D_{\mathrm{KL}}(\mathbb{P}\|\mathbb{Q})$ holds for all distributions $\mathbb{P}$ and $\mathbb{Q}$. Moreover, according to Theorems 3.1 and 3.3, for $\mathbb{P}, \mathbb{Q} \in \mathcal{P}$, we have $D_{\mathrm{KL}}(\mathbb{P}\|\mathbb{Q}) \lesssim H^2(\mathbb{P}, \mathbb{Q})$, which indicates that

$$
\tilde{R}_{KL,n}(\mathcal{P}) \lesssim \tilde{R}_{H^2,n}(\mathcal{P}).
$$

Therefore, we have

$$
R_{H^2,n}(\mathcal{P}) \lesssim R_{KL,n}(\mathcal{P}) \le \tilde{R}_{KL,n}(\mathcal{P}) \lesssim \tilde{R}_{H^2,n}(\mathcal{P}).
$$

Next, we notice that

$$
R_{H^2,n}(\mathcal{P}) = \inf_{\hat{f}_n} \sup_{f\in\mathcal{P}} \mathbb{E}_f\left[H^2(\hat{f}_n, f)\right].
$$

For one estimator $\hat{f}_n$, suppose $\tilde{f}_n$ is the projection of $\hat{f}_n$ into $\mathcal{P}$ under Hellinger distance. Then for every $f \in \mathcal{P}$, we have $H(\tilde{f}_n, f) \le 2H(\hat{f}, f)$. Therefore, we have $R_{H^2,n}(\mathcal{P}) \ge \frac{1}{4}\tilde{R}_{H^2,n}(\mathcal{P})$.

Hence we have proved that $R_{H^2,n}(\mathcal{P}) \asymp R_{KL,n}(\mathcal{P}) \asymp \tilde{R}_{KL,n}(\mathcal{P}) \asymp \tilde{R}_{H^2,n}(\mathcal{P})$. Similarly, for sequential density estimation minimax risks, we can also show that

$$C_{H^2,n}(\mathcal{P}) \asymp C_{KL,n}(\mathcal{P}) \asymp \tilde{C}_{KL,n}(\mathcal{P}) \asymp \tilde{C}_{H^2,n}(\mathcal{P}).$$

Therefore, to prove Corollary 3.9, we only need to show:

$$R_{H^2,n} \asymp \inf_{\varepsilon>0} \varepsilon^2 + \frac{1}{n} \log \mathcal{N}_{loc,H}(\mathcal{P}, \varepsilon),$$
$$C_{KL,n} \asymp \inf_{\varepsilon>0} n\varepsilon^2 + \log \mathcal{N}_H(\mathcal{P}, \varepsilon).$$

For the first inequality above, the upper bound follows from the Le Cam-Birgé construction. For the second inequality, it follows from results in [57].

B.3 Proof of Theorem 3.1 in General Dimensions

Without loss of generality, we assume $d \le M^2$. For simplicity we abbreviate $f_\pi(\cdot)$ and $f_\eta(\cdot)$ as $p(\cdot), q(\cdot)$.

Lemma B.1. *Suppose $p = \pi * \mathcal{N}(0, I_d)$, where $\mathtt{supp}(\pi) \subset B_2(M)$. Then for any $\omega \in \Omega$, we have:*

1. *$\forall r' \in [M, r]$, we have $p(\mathbf{x}(r', \omega)) \ge p(\mathbf{x}(r, \omega))$.*
2. *$\forall r' \ge r \ge M$, we have $p(\mathbf{x}(r', \omega)) \le p(\mathbf{x}(r, \omega)) \exp\left(-\frac{(r'-M)^2-(r-M)^2}{2}\right)$*

Lemma B.2. *Suppose $p = \pi * \mathcal{N}(0, I_d)$ where $\mathtt{supp}(\pi) \subset B_2(M)$. Then we have*

$$\|\nabla \log p(\mathbf{x})\|_2 \le 3\|\mathbf{x}\|_2 + 4M, \quad \forall \mathbf{x} \in \mathbb{R}^d.$$

The proof of Theorem 3.1 follows by integrating in polar coordinates and using the one-dimensional techniques established in Section 3.3.

B.4 Proof of Theorem 3.4

Given some constant $r > 1$, we consider the following distribution:

$$\pi_r = (1 - h_r)\delta_0 + h_r \delta_r,$$

where $h_r = \exp\left(-\frac{r^2}{2K^2}\right)$. Then for any $r > 1$, π_r is a K-subgaussian distribution.

We let $p_r = \pi_r * \mathcal{N}(0, 1)$. The proof follows by analyzing the KL divergence and Hellinger distance in three regions: $(-\infty, -r]$, $[-r, r]$ and $[r, \infty)$. Choosing r large enough shows that D_{KL}/H^2 can be arbitrarily large when $K > 1$.

B.5 Proof of Theorem 3.3

Lemma B.3. *Suppose $p = \pi * \mathcal{N}(0, I_d)$, where π is a K-subgaussian distribution with $K < 1$. Then for every $r \geq \frac{2\sqrt{K}}{1-\sqrt{K}}$ and any $\omega \in \Omega$, we have:*

1. *$\forall r' \in \left[\frac{2\sqrt{K}}{1-\sqrt{K}}, r\right]$, we have $p(\mathbf{x}(r', \omega)) \geq \frac{p(\mathbf{x}(r,\omega))}{7}$.*
2. *$\forall r' \geq r$, we have $p(\mathbf{x}(r', \omega)) \leq 7p(\mathbf{x}(r, \omega)) \cdot \exp\left(-\frac{1-K}{4} r(r' - r)\right)$.*

Lemma B.4. *Suppose $p = \pi * \mathcal{N}(0, I_d)$, where π is a K-subgaussian distribution with $K < 1$. For $\mathbf{x} \in \mathbb{R}^d$, we have:*

$$\|\nabla \log p(\mathbf{x})\|_2 \leq 3\|\mathbf{x}\|_2 + 12.$$

Integration in polar coordinates then establishes the linear comparison between KL and H^2 for $K < 1$.

B.6 Proof of Theorem 3.5

Using the f-divergence convexity and Jensen's inequality, we bound the integral $\int \frac{f_\pi^\lambda}{f_\eta^{\lambda-1}}$ for K-subgaussian distributions. By optimizing over λ and δ in the inequality of [57], we obtain the dimension-free bound with a logarithmic factor.

B.7 Comparison inequalities for other distances

In this section, we discuss comparison inequalities for high-order divergences and distances.

Theorem B.5. *For d-dimensional distributions π, η supported on $B_2(M)$ with $M \geq 2$, we have*

$$\chi^2(f_\pi \| f_\eta) \leq 2 \exp\left(50(M^2 \vee d)\right) H^2(f_\pi, f_\eta).$$

Theorem B.6. *Suppose π and η are one-dimensional distributions supported on $[-M, M]$ with $M \geq 1$. Then we have:*

$$D_{\mathrm{TV}}(f_\pi, f_\eta) \leq \left(8\sqrt{M} + 2\log^{1/4} \frac{1}{\|f_\pi - f_\eta\|_2}\right) \|f_\pi - f_\eta\|_2$$

Theorem B.7. *For any one-dimensional distributions π, η:*

$$\|f_\pi - f_\eta\|_2 \leq \left(\log^{1/4} \frac{1}{D_{\mathrm{TV}}(f_\pi, f_\eta)} \vee 3\right) D_{\mathrm{TV}}(f_\pi, f_\eta).$$

Appendix C

Proofs for Chapter 4

C.1 Proofs of Theorem 4.5 and Theorem 4.6

Proof of Theorem 4.5. Our proof is divided into two parts: the lower bound to the density estimation sample complexity and the upper bound to the density estimation sample complexity.

Lower Bound to Density Estimation: Without loss of generality we assume $(2\varepsilon)^{-2/3}$ is an integer (otherwise we replace ε by $\lfloor(2\varepsilon)^{-2/3}\rfloor^{-3/2}/2$ and the argument follows only up to constant). We first construct a infinite-dimensional rectangle which is a subset of Γ:

$$M = \left\{\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots) : |\theta_i| \le (2\varepsilon)^{4/3},\ \forall 1 \le i \le (2\varepsilon)^{-2/3}, \quad \text{and} \quad \theta_i = 0,\ \forall i \ge d\right\}.$$

We can verify that for any $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots) \in M$,

$$\sum_{i=1}^{\infty} i \cdot |\theta_i| \le \sum_{i=1}^{(2\varepsilon)^{-2/3}} i \cdot (2\varepsilon)^{4/3} \le 1.$$

Hence $\boldsymbol{\theta} \in \Gamma$. Therefore, $M \subseteq \Gamma$. We next lower bound the density estimation error of estimation using n samples: according to [70, Proposition 4.16], we have

$$\inf_{\widehat{\boldsymbol{\theta}}_n} \sup_{\boldsymbol{\theta} \in M} \mathbb{E}\left[\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\|^2\right] = (2\varepsilon)^{-2/3} \cdot \inf_{\widehat{\theta}_n} \sup_{|\theta| \le (2\varepsilon)^{4/3}} \mathbb{E}\left[(\widehat{\theta}_n - \theta)^2\right],$$

where $\widehat{\boldsymbol{\theta}}_n$ denotes an estimator with n i.i.d. samples coming from $\mathcal{N}(\boldsymbol{\theta}, I)$, and $\widehat{\theta}_n$ denotes an estimator with n i.i.d. samples coming from $\mathcal{N}(\theta, 1)$. Next, according to Van trees inequality [210] (also in [70, (4.9)]), we have

$$\inf_{\widehat{\theta}_n} \sup_{|\theta| \le (2\varepsilon)^{4/3}} \mathbb{E}\left[(\widehat{\theta}_n - \theta)^2\right] \ge \frac{1}{2n} \wedge \frac{1}{2} \cdot (2\varepsilon)^{8/3}.$$

Hence we obtain that

$$\inf_{\widehat{\boldsymbol{\theta}}_n} \sup_{\boldsymbol{\theta} \in M} \mathbb{E}\left[\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\|^2\right] \ge (2\varepsilon)^{-2/3} \cdot \left(\frac{1}{2n} \wedge (2\varepsilon)^{8/3}\right) = 2\varepsilon^2 \wedge \frac{1}{2n \cdot (2\varepsilon)^{2/3}}.$$

Hence in order to achieve no more than ε^2 density estimation error (i.e. eq. (4.8) holds), we require

$$n_{\text{est}}(\Gamma, \varepsilon) \geq \frac{\varepsilon^{-8/3}}{4}. \tag{C.1}$$

Upper Bound to Density Estimation: Next, we construct an estimator which will achieve ε^2 density estimation error (i.e. eq. (4.8) holds) with no more than $\widetilde{\mathcal{O}}(\varepsilon^{-8/3})$ samples. For n samples $\mathbf{X} = (\mathbf{X}^{1:n})$ where each $\mathbf{X}^i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{\theta}, I)$, we consider the following estimator: let $\widehat{\theta}(\mathbf{X}) = (\widehat{\theta}_1, \widehat{\theta}_2, \cdots) = \frac{1}{n}\sum_{i=1}^n X_i$, and further construct estimator $\widetilde{\theta}(\mathbf{X}) = (\widetilde{\theta}_1(\mathbf{X}), \widetilde{\theta}_2(\mathbf{X}), \cdots)$, where

$$\widetilde{\theta}_i = \begin{cases} \delta_\lambda(\hat{\theta}_i) & \forall 1 \leq i \leq D, \\ 0 & i > D, \end{cases}$$

where $\lambda > 0$ and $D \in \mathbb{Z}_+$ are parameters to be specified later. Here $\delta_\lambda(\cdot) : \mathbb{R} \to \mathbb{R}$ is the soft-thresholding function defined as:

$$\delta_\lambda(x) = \begin{cases} x - \lambda & x \geq \lambda, \\ 0 & -\lambda \leq x < \lambda, \\ x + \lambda & x < -\lambda. \end{cases}$$

We will show that with proper choices of λ and D, the density estimation error can be upper bounded by ε^2 with $n = \widetilde{\mathcal{O}}(1/\varepsilon^{8/3})$ number of samples. We write $\mathbf{X}^i = (X_1^i, X_2^i, \cdots)$. Then we have $X_j^1, X_j^2, \cdots X_j^n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta_j, 1)$ for any $j \geq 0$. And we only need to verify for some choice of λ and D,

$$\sum_{j=1}^{\infty} \mathbb{E}\left[(\widetilde{\theta}_j(\mathbf{X}) - \theta_j)^2\right] \leq \varepsilon^2.$$

When $j > D$, we have $\widetilde{\theta}_j(\mathbf{X}) = 0$, hence

$$\mathbb{E}\left[(\widetilde{\theta}_j(\mathbf{X}) - \theta_j)^2\right] = \theta_j^2.$$

And for $j \leq D$, according to [70, (8.12)], we have

$$\mathbb{E}\left[(\widetilde{\theta}_j(\mathbf{X}) - \theta_j)^2\right] \leq \frac{1}{n}\exp\left(-\frac{n\lambda^2}{2}\right) + \min\left\{\theta_j^2, \frac{1}{n} + \lambda^2\right\}.$$

Thus, we obtain

$$\begin{aligned} &\sum_{j=1}^{\infty} \mathbb{E}\left[(\widetilde{\theta}_j(\mathbf{X}) - \theta_j)^2\right] \\ &\leq \frac{D}{n}\exp\left(-\frac{n\lambda^2}{2}\right) + \sum_{j=1}^{D} \min\left\{\theta_j^2, \frac{1}{n} + \lambda^2\right\} + \sum_{j=D+1}^{\infty} \theta_j^2. \end{aligned}$$

Since $\boldsymbol{\theta} \in \Theta$, we have

$$\sum_{j=1}^{\infty} j \cdot |\theta_j| \leq 1.$$

If we choose $D = \lceil 2/\varepsilon \rceil$, for $\theta = (\theta_1, \theta_2, \cdots)$, we will have

$$\sum_{j=D+1}^{\infty} \theta_j^2 \le \frac{1}{D^2} \left(\sum_{j=D+1}^{\infty} j \cdot |\theta_j| \right)^2 \le \frac{1}{D^2} \le \frac{\varepsilon^2}{4}. \tag{C.2}$$

In the next, we will choose the value of λ and further upper bound

$$\sum_{j=1}^{D} \min \left\{ \theta_j^2, \frac{1}{n} + \lambda^2 \right\}. \tag{C.3}$$

Since set $\{(\theta_1, \cdots, \theta_D) : \sum_{j=1}^{D} j \cdot |\theta_j| \le 1\}$ is compact, there must exists some $(\theta_1, \cdots, \theta_D)$ which maximizes (C.3). Without loss of generality we assume $|\theta_j| \le \sqrt{1/n + \lambda^2}$ (otherwise truncate the value of all θ_j to interval $[-\sqrt{1/n + \lambda^2}, \sqrt{1/n + \lambda^2}]$ and it results in another maximizer). Further if there exists $j_1 \ne j_2$ which both satisfy $0 < |\theta_j| < \sqrt{1/n + \lambda^2}$, then by slightly modifying θ_{j_1} and θ_{j_2} we can make (C.3) larger, while keeping the parameter still in the set. This contradicts to the assumption that $(\theta_1, \cdots, \theta_D)$ is a maximizer. Therefore, there exists at most one of $1 \le j \le D$ which satisfies $0 < |\theta_j| < \sqrt{1/n + \lambda^2}$. We let integer N to be the smallest integer such that

$$\frac{N(N-1)}{2} \cdot \sqrt{\frac{1}{n} + \lambda^2} \ge 1.$$

Then the maximizer has at most N nonzero items, which implies

$$\sup_{\sum_{j=1}^{D} j \cdot |\theta_j| \le 1} \sum_{j=1}^{D} \min \left\{ \theta_j^2, \frac{1}{n} + \lambda^2 \right\} \le N \cdot \left(\frac{1}{n} + \lambda^2 \right).$$

Finally, we choose $\lambda = (\varepsilon/8)^{4/3}$. When $n \ge (\varepsilon/8)^{-8/3} \cdot 4 \log(1/\varepsilon)$ we have

$$N \le 2 \cdot (\varepsilon/8)^{-2/3} \quad \text{and} \quad \frac{1}{n} \le \lambda^2,$$

which implies

$$\sum_{j=1}^{D} \min \left\{ \theta_j^2, \frac{1}{n} + \lambda^2 \right\} \le 2 \cdot (\varepsilon/8)^{-2/3} \cdot 2\lambda^2 \le \frac{\varepsilon^2}{2}. \tag{C.4}$$

Further according to our choice of n we have

$$\frac{D}{n} \exp\left(-\frac{n\lambda^2}{2} \right) \le \frac{\varepsilon^2}{2}, \tag{C.5}$$

Combining eq. (C.2), eq. (C.4) and eq. (C.5), we obtain that

$$\frac{D}{n} \exp\left(-\frac{n\lambda^2}{2} \right) + \sum_{j=1}^{D} \min \left\{ \theta_j^2, \frac{1}{n} + \lambda^2 \right\} + \sum_{j=D+1}^{\infty} \theta_j^2 \le \frac{\varepsilon^2}{4} + \frac{\varepsilon^2}{2} + \frac{\varepsilon^2}{4} = \varepsilon^2.$$

Therefore, we obtain

$$n_{\text{est}}(\Gamma, \varepsilon) \leq \left(\frac{\varepsilon}{8}\right)^{-8/3} \cdot 4 \log\left(\frac{1}{\varepsilon}\right). \tag{C.6}$$

Above all, according to eq. (C.1) and eq. (C.6), for set Γ defined in eq. (4.4), we have

$$n_{\text{est}}(\Gamma, \varepsilon) = \tilde{\Theta}\left(\varepsilon^{-8/3}\right).$$

□

Proof of Theorem 4.6. Our proof is divided into two parts: the lower bound to the goodness-of-fit testing sample complexity and the upper bound to the goodness-of-fit testing sample complexity.

Lower Bound to Goodness-of-fit Testing: For any fixed distribution $\mu \in \Delta(\mathbb{R}^\infty)$, we define distribution $\mathbb{P}_{0,X}$ to be the distribution of $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ sampled according to the following way: first sample $\boldsymbol{\theta} \sim \mu$, then sample $\mathbf{X} = (\mathbf{X}^{1:n}) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{\theta}, I)$. And we define distribution $\mathbb{P}_{1,X}$ to be the distribution of $\mathbf{X} = (\mathbf{X}^{1:n}) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, I)$. According to [15, Lemma 5], we have the following lower bound on the testing error (left hand side of eq. (4.7))

$$\inf_{\psi} \max_{i \in \{0,1\}} \sup_{P \in H_i} \mathbb{P}(\psi(X) \neq i) \geq \frac{1}{2}(1 - \text{TV}(\mathbb{P}_{0,X}, \mathbb{P}_{1,X})) - \mu(\Gamma^c) - \mu(B_2(\varepsilon)), \tag{C.7}$$

where Γ^c is the complement of set Γ, and $B_2(\varepsilon)$ denotes the ℓ_2-ball of radius ε. We next construct such a prior μ. First of all, we choose μ to be supported in the following subset of Γ:

$$\Gamma_d = \{\theta = (\theta_1, \theta_2, \cdots) : \theta \in \Gamma, \theta_{d+1} = \theta_{d+2} = \cdots = 0\} \subseteq \Gamma,$$

where d is some positive integer to be specified later. Since for any $\mathbf{X}^i \sim \mathcal{N}(\boldsymbol{\theta}, I)$, the first d coordinate $X^i_{1:d}$ of $\mathbf{X}^i$ only depends on $\theta_1, \cdots, \theta_d$, and the rest coordinates are pure i.i.d. standard Gaussian noises, in order to lower bound the goodness-of-fit testing sample complexity of Γ_d, without loss of generality we only need to consider the first N coordinates, i.e. we can ignore coordinates larger than N and assume $\Gamma_d \subseteq \mathbb{R}^d$:

$$\Gamma_d = \left\{\theta = (\theta_1, \theta_2, \cdots, \theta_d) : \sum_{i=1}^{d} i \cdot |\theta_i| \leq 1\right\}. \tag{C.8}$$

Next, we construct a distribution $\mu \in \Delta(\Gamma_d)$ with Γ_d defined in eq. (C.8). For some one-dimensional distribution $q \in \Delta([-1, 1])$, in the following form:

$$q(\cdot) = (1 - h)\delta_0(\cdot) + \frac{h}{2}\delta_r(\cdot) + \frac{h}{2}\delta_{-r}(\cdot),$$

where $h \in (0, 1)$ and $r \geq 0$ are parameters to be specified later, we consider the following product distribution:

$$\mu = \bigotimes_{i=1}^{d} q \in \Delta(\mathbb{R}^d). \tag{C.9}$$

Next, we will lower bound the right hand side of eq. (C.7). First we notice that

$$\mathrm{TV}(\mathbb{P}_{0,X}, \mathbb{P}_{1,X})^2 \le \chi^2(\mathbb{P}_{1,X} \| \mathbb{P}_{0,X}) - 1.$$

We next adopt the Ingster's trick in [60] and further upper bound the χ^2-divergence: if we use $\varphi_{\boldsymbol{\theta}}(\cdot)$ to denote the probability density function of $\mathcal{N}(\boldsymbol{\theta}, I_D)$, we have

$$\chi^2(\mathbb{P}_{1,X} \| \mathbb{P}_{0,X}) = \mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\theta}' \overset{\text{i.i.d.}}{\sim} \mu} \left[\int_{(\mathbb{R}^D)^n} \prod_{t=1}^n \frac{\varphi_{\boldsymbol{\theta}}(\mathbf{z}^t)\varphi_{\boldsymbol{\theta}'}(\mathbf{z}^t)}{\varphi_{\mathbf{0}}(\mathbf{z}^t)} d\mathbf{z}^1 \cdots d\mathbf{z}^n \right] = \mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\theta}' \overset{\text{i.i.d.}}{\sim} \mu} [\exp(n \cdot \langle \boldsymbol{\theta}, \boldsymbol{\theta}' \rangle)]$$

Since μ is the product distribution defined in eq. (C.9), we have

$$\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\theta}' \overset{\text{i.i.d.}}{\sim} \mu} [\exp(n \cdot \langle \boldsymbol{\theta}, \boldsymbol{\theta}' \rangle)] = \prod_{i=1}^d \left(1 + \frac{h^2}{4} \cdot (\exp(nr^2) - 1) + \frac{h^2}{4} \cdot (\exp(-nr^2) - 1) \right).$$

When $nr^2 \le 1$, we have

$$\frac{\exp(nr^2) + \exp(-nr^2)}{2} - 1 \le n^2 r^4.$$

Therefore, when $dh^2 n^2 r^4 \le 1$, we have

$$\begin{aligned} &\prod_{i=1}^d \left(1 + \frac{h^2}{4} \cdot (\exp(nr^2) - 1) + \frac{h^2}{4} \cdot (\exp(-nr^2) - 1) \right) \\ &\quad \le \prod_{i=1}^d \left(1 + \frac{h^2}{2} \cdot n^2 r^2 \right) \le 1 + dh^2 n^2 r^4, \end{aligned}$$

which implies that

$$\mathrm{TV}(\mathbb{P}_{0,X}, \mathbb{P}_{1,X}) \le \sqrt{dh^2 n^2 r^4}.$$

According to Hoeffding inequality, for any $\delta > 0$, with probability at least $1 - \delta$ we have

$$\sum_{t=1}^d i \cdot |\theta_i| \le d \cdot dhr + d \cdot r \cdot \sqrt{2d \log(1/\delta)}.$$

Additionally, again according to Hoeffding inequality, with probability at least $1 - \delta$ we have

$$\sum_{t=1}^d |\theta_t|^2 = \sum_{t=1}^d |\theta_t|^2 \ge dhr^2 - r^2 \cdot \sqrt{2d \log(2/\delta)}.$$

As long as $d \ge 12/h^2$, we have with probability at least $4/5$,

$$\sum_{t=1}^d i \cdot |\theta_i| \le 2d^2 hr \quad \text{and} \quad \sum_{t=1}^d |\theta_t|^2 \ge \frac{1}{2} dhr^2.$$

Hence if $2d^2 hr \le 1$ and $dhr^2/2 \ge \varepsilon^2$, we have

$$\mu(\boldsymbol{\theta} \in \Gamma^c) + \mu(\boldsymbol{\theta} \in B_2(\varepsilon)) \le \frac{1}{5}.$$

Finally, we choose

$$d = \varepsilon^{-4/5}, \quad h = \frac{1}{16}\varepsilon^{2/5} \quad \text{and} \quad r = 8\varepsilon^{6/5}.$$

Then if $n \le \varepsilon^{-12/5}/64$ we have $nr^2 \le 1$ and also $dh^2n^2r^4 \le 1$. We can also verify that according to the above choices of (d, h, r), $d \ge 12/h^2$, $2d^2hr \le 1$ and $dhr^2/2 \ge \varepsilon^2$ always holds. Hence have

$$\mu(\boldsymbol{\theta} \in \Gamma^c) + \mu(\boldsymbol{\theta} \in B_2(\varepsilon)) \le \frac{1}{5}$$

and also

$$\mathrm{TV}(\mathbb{P}_{0,X}, \mathbb{P}_{1,X}) \le \sqrt{dh^2n^2r^4} \le \frac{1}{16}.$$

Bringing these two inequalities back to eq. (C.7), we obtain that if $n \le \varepsilon^{-12/5}/64$,

$$\inf_{\psi} \max_{i\in\{0,1\}} \sup_{P\in H_i} \mathbb{P}(\psi(X) \ne i) \ge \frac{1}{2} \cdot \left(1 - \frac{1}{16}\right) - \frac{1}{5} > \frac{1}{4}.$$

Upper Bound to Goodness-of-fit Testing: Given n samples $\mathbf{X} = (\mathbf{X}^{1:n}) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{\theta}, I_d)$, we design a test for the testing problem eq. (4.6). We let $D = 2/\varepsilon$, and $d = \varepsilon^{-4/5}$. For any $1 \le i \le D$, we let

$$\widehat{\theta}_i = \frac{1}{n}\sum_{t=1}^{n} X_i^t, \quad \text{where } X_i^t \text{ is the } i\text{-th coordinates of } \mathbf{X}^t.$$

Consider the following test:

(i) If $\sum_{i=1}^{d}(\widehat{\theta}_i)^2 \le \varepsilon^2/2$ and $\max_{d\le i\le D}|\widehat{\theta}_i| \le \varepsilon$, we accept the null hypothesis $\boldsymbol{\theta} = \mathbf{0}$;

(ii) If $\sum_{i=1}^{d}(\widehat{\theta}_i)^2 > \varepsilon^2/2$ or $\max_{d\le i\le D}|\widehat{\theta}_i| > \varepsilon$, we reject the null hypothesis $\boldsymbol{\theta} = \mathbf{0}$.

Suppose the above testing scheme is ψ. Next we will verify eq. (4.7) for the testing scheme ψ as long as the number of sample satisfies

$$n \ge \varepsilon^{-12/5} \log\left(\frac{16}{\varepsilon}\right). \tag{C.10}$$

First we verify the case where $i = 0$, i.e.

$$\sup_{P\in H_0} \mathbb{P}(\psi(\mathbf{X}) = 1) \le \frac{1}{4}. \tag{C.11}$$

Notice that when $\boldsymbol{\theta} = 0$, we have

$$\widehat{\theta}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1/n) \quad \forall 1 \le i \le D.$$

Hence we only need to prove that

$$\mathbb{P}\left(\sum_{i=1}^{d}(\widehat{\theta}_i)^2 > \frac{\varepsilon^2}{2}\right) \le \frac{1}{8} \quad \text{and} \quad \mathbb{P}\left(\max_{d+1\le i\le D}|\widehat{\theta}_i| \le \varepsilon\right) \ge \frac{7}{8}. \tag{C.12}$$

To verify the first inequality of eq. (C.12), we calculate

$$\mathbb{E}\left[\sum_{i=1}^{d}(\widehat{\theta}_i)^2\right]=\frac{d}{n}\quad\text{and}\quad\mathrm{Var}\left[\sum_{i=1}^{d}(\widehat{\theta}_i)^2\right]=\frac{2d}{n^2}.$$

Hence according to Chebyshev inequality we obtain that when n satisfies eq. (C.10), we have

$$\mathbb{P}\left(\sum_{i=1}^{d}(\widehat{\theta}_i)^2>\frac{\varepsilon^2}{8}\right)\leq\frac{d/n^2}{|\varepsilon^2/2-d/n|^2}\leq\frac{1}{8}.$$

Additionally, since $\widehat{\theta}_i\overset{\text{i.i.d.}}{\sim}\mathcal{N}(0,1/n)$, when n satisfies eq. (C.10), we have

$$\mathbb{P}\left(\max_{d+1\leq i\leq D}|\widehat{\theta}_i|\leq\varepsilon^{6/5}\right)=\left(1-\varphi_1(\varepsilon^{6/5}\sqrt{n})\right)^{D-d}\geq 1-D\cdot\varphi_1(\varepsilon^{6/5}\sqrt{n})\geq 1-D\cdot\exp\left(-n\varepsilon^{12/5}\right)\geq\frac{7}{8}.$$

where $\varphi_1(\cdot)$ is the probability density function of one-dimensional standard normal distribution. Hence both inequalities in eq. (C.12) are verified.

Next, we verify eq. (4.7) for the case $i=1$, i.e.

$$\sup_{P\in H_1}\mathbb{P}(\psi(\mathbf{X})=0)\leq\frac{1}{4}.\tag{C.13}$$

We only need to show that for any $\boldsymbol{\theta}\in\Gamma$ with $\|\boldsymbol{\theta}\|_2\geq\varepsilon$, we always have

$$\mathbb{E}_{\mathbf{X}\overset{\text{i.i.d.}}{\sim}\mathcal{N}(\boldsymbol{\theta},I)}\left[\psi(\mathbf{X})=1\right]\leq\frac{1}{4}.$$

Notice that when $\mathbf{X}\overset{\text{i.i.d.}}{\sim}\mathcal{N}(\boldsymbol{\theta},I)$, we have $\widehat{\theta}_i\sim\mathcal{N}(\theta_i,1/n)$ independently. For $\boldsymbol{\theta}\in\Gamma$ which satisfies $\|\boldsymbol{\theta}\|_2\geq\varepsilon$, we have

$$\sum_{i=D+1}^{\infty}(\theta_i)^2\leq\left(\sum_{i=D+1}^{\infty}|\theta_i|\right)\leq\frac{1}{D^2}\left(\sum_{i=D+1}^{\infty}i\cdot|\theta_i|\right)\leq\frac{\varepsilon^2}{4},$$

which implies that

$$\sum_{i=1}^{d}(\theta_i)^2+\sum_{i=d+1}^{D}(\theta_i)^2=\sum_{i=1}^{D}(\theta_i)^2\geq\frac{\varepsilon^2}{2}.$$

Therefore, we have either $\sum_{i=1}^{d}(\theta_i)^2\geq\varepsilon^2/4$, or $\sum_{i=d+1}^{D}(\theta_i)^2\geq\varepsilon^2/4$. If we have $\sum_{i=1}^{d}(\theta_i)^2\geq\varepsilon^2/4$, since $\widehat{\theta}_i\sim\mathcal{N}(\theta_i,1/n)$ independently, we can calculate

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{d}(\widehat{\theta}_i)^2\right]&=\sum_{i=1}^{d}(\theta_i)^2+\frac{d}{n}\\ \mathrm{Var}\left[\sum_{i=1}^{d}(\widehat{\theta}_i)^2\right]&=\frac{4}{n}\cdot\sum_{i=1}^{d}(\theta_i)^2+\frac{2d}{n^2}.\end{aligned}$$

Therefore, according to Chebyshev inequality we obtain that when n satisfies eq. (C.10)

$$\mathbb{P}\left(\sum_{i=1}^{d}(\widehat{\theta}_i)^2 \le \frac{\varepsilon^2}{8}\right) \le \frac{\frac{4}{n}\cdot\sum_{i=1}^{d}(\theta_i)^2 + \frac{2d}{n^2}}{|\sum_{i=1}^{d}(\theta_i)^2 + \frac{d}{n} - \frac{\varepsilon^2}{8}|^2} \le \frac{1}{4}, \quad \text{hence} \quad \mathbb{P}\left(\sum_{i=1}^{d}(\widehat{\theta}_i)^2 \ge \frac{\varepsilon^2}{8}\right) \ge \frac{3}{4}.$$

Next, we consider the case where $\sum_{i=d+1}^{D}(\theta_i)^2 \ge \varepsilon^2/4$. We notice that

$$\sum_{i=d+1}^{D}(\theta_i)^2 \le \left(\sum_{i=d+1}^{D}|\theta_i|\right)\cdot\max_{d+1\le i\le D}|\theta_i| \le \frac{1}{d}\cdot\left(\sum_{i=d+1}^{D} i\cdot|\theta_i|\right)\cdot\max_{d+1\le i\le D}|\theta_i| \le \frac{1}{d}\cdot\max_{d+1\le i\le D}|\theta_i|,$$

which implies that

$$\max_{d+1\le i\le D}|\theta_i| \ge d\cdot\frac{\varepsilon^2}{4} \ge \varepsilon^{6/5}.$$

Since $\widehat{\theta}_i \sim \mathcal{N}(\theta_i, 1/n)$ independently, when n satisfies eq. (C.10), we have

$$\begin{aligned}\mathbb{P}\left(\max_{d+1\le i\le D}|\widehat{\theta}_i - \theta_i| \le \varepsilon^{6/5}\right) &= (1-\varphi_1(\varepsilon^{6/5}\sqrt{n}))^{D-d} \\ &\ge 1 - D\cdot\varphi(\varepsilon^{6/5}\sqrt{n}) \ge 1 - D\cdot\exp(-n\varepsilon^{12/5}) \ge \frac{3}{4},\end{aligned}$$

which implies that

$$\mathbb{P}\left(\max_{d+1\le i\le D}|\widehat{\theta}_i| \ge \varepsilon^{6/5}\right) \ge \frac{3}{4}.$$

According to the form of the testing scheme ψ, we verified eq. (4.7) for the case $i = 1$. □

C.2 Proofs of Theorem 4.7 and Theorem 4.8

In order to prove Theorem 4.7, we first bring up the following property of orthosymmetric convex sets.

Proposition C.1. *Suppose $\Gamma \subseteq \mathbb{R}^d$ is an orthosymmetric set. If Γ is also convex, then for any $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_D) \in \Gamma$ and $\alpha = (\alpha_1, \cdots, \boldsymbol{\alpha}_D) \in [-1, 1]^D$, we have $\boldsymbol{\theta}\cdot\boldsymbol{\alpha} = (\alpha_1\theta_1, \alpha_2\theta_2, \cdots, \alpha_D\theta_D) \in \Gamma$.*

Now we are ready for the proof of Theorem 4.7.

Proof of Theorem 4.7. Without loss of generality, we assume that $n_{\text{est}}(\Gamma, \varepsilon) \ge 4$. Our proof proceeds as follows. First, we consider parameter estimation via soft-thresholding. We show that there must exist $\boldsymbol{\theta}^*$ in Γ such that $\|\boldsymbol{\theta}^*\|^2 \asymp \varepsilon^2$ and each entry $|\boldsymbol{\theta}_i^*| \le n_{\text{est}}(\Gamma, \varepsilon)^{-1/2}\text{polylog}(n, \varepsilon)$ (for otherwise, soft-thresholding estimator would beat the optimal sample complexity of estimation). Second, we use Ingster's method of simple ($\boldsymbol{\theta} = 0$) vs composite ($\boldsymbol{\theta} = (\pm\boldsymbol{\theta}_1^*, \ldots, \pm\boldsymbol{\theta}_D^*)$) hypothesis testing to lower bound the goodness of fit sample complexity $n_{\text{gof}}(\Gamma, \varepsilon)$.

First Part: Density Estimation Rate: We define the soft-thresholding function $\mathrm{sth} : \mathbb{R} \times (\mathbb{R}_+ \cup \{0\}) \to \mathbb{R}$:

$$\mathrm{sth}(x, \lambda) = \begin{cases} x - \lambda & \text{if } x \geq \lambda, \\ 0 & \text{if } -\lambda \leq x < \lambda, \\ x + \lambda & \text{if } x < -\lambda. \end{cases} \tag{C.14}$$

For some $\boldsymbol{\theta} \in \Gamma$, given n samples $X = X^{1:n} \sim \mathcal{N}(\boldsymbol{\theta}, I_D)^{\otimes n}$, we construct the soft-thresholding estimator $\widehat{\boldsymbol{\theta}}^n_{\mathrm{sth}} = (\widehat{\theta}^n_1, \cdots, \widehat{\theta}^n_D)$ as follows: for any $j \in [D]$, we choose

$$\widehat{\theta}^n_j = \mathrm{sth}\left(\frac{1}{n}\sum_{i=1}^n X^i_j, \lambda(n)\right),$$

where we denote $X^i = (X^i_1, X^i_2, \cdots, X^i_D)$, and we let

$$\lambda(n) \triangleq \sqrt{\frac{2\log\left({}^{2D}/_{(n\varepsilon^2)}\right) \vee 0}{n}}. \tag{C.15}$$

We let $n_{\mathrm{sth}}(\varepsilon)$ to denote the smallest n such that $\widehat{\boldsymbol{\theta}}^n_{\mathrm{sth}}$ induces expected estimation error ε^2 for all $\boldsymbol{\theta} \in \Gamma$, i.e.

$$n_{\mathrm{sth}}(\varepsilon) = \min\left\{n : \sup_{\boldsymbol{\theta}\in\Gamma} \mathbb{E}\left[\|\widehat{\boldsymbol{\theta}}^n_{\mathrm{sth}} - \boldsymbol{\theta}\|^2\right] \leq \varepsilon^2\right\}. \tag{C.16}$$

Since $n_{\mathrm{est}}(\Gamma, \varepsilon)$ is the minimal number of samples in order to reach expected estimation error ε^2, we have

$$n_{\mathrm{sth}}(\varepsilon) \geq n_{\mathrm{est}}(\Gamma, \varepsilon).$$

According to eq. (C.16), we have

$$\sup_{\boldsymbol{\theta}\in\Gamma} \mathbb{E}_{X_{1:n_{\mathrm{sth}}(\varepsilon)} \overset{\mathrm{iid}}{\sim} \mathcal{N}(\boldsymbol{\theta}, I_D)}\left[\left\|\widehat{\boldsymbol{\theta}}^{n_{\mathrm{sth}}(\varepsilon)}_{\mathrm{sth}} - \boldsymbol{\theta}\right\|_2^2\right] \leq \varepsilon^2 \tag{C.17}$$

and for any $n < n_{\mathrm{sth}}(\varepsilon)$,

$$\sup_{\boldsymbol{\theta}\in\Gamma} \mathbb{E}_{X_{1:n} \overset{\mathrm{iid}}{\sim} \mathcal{N}(\boldsymbol{\theta}, I_D)}\left[\left\|\widehat{\boldsymbol{\theta}}^{n}_{\mathrm{sth}} - \boldsymbol{\theta}\right\|_2^2\right] \geq \varepsilon^2. \tag{C.18}$$

Next, according to [70, (8.7), (8.12)], we have for any positive integer n and $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_D) \in \Gamma$ that

$$\mathbb{E}_{X_{1:n} \overset{\mathrm{iid}}{\sim} \mathcal{N}(\boldsymbol{\theta}, I_D)}\left[\left\|\widehat{\boldsymbol{\theta}}^{n}_{\mathrm{sth}} - \boldsymbol{\theta}\right\|_2^2\right] \leq \sum_{i=1}^D \frac{1}{n}\exp\left(-\frac{n\lambda(n)^2}{2}\right) + \min\left\{\theta_i^2, \frac{1}{n} + \lambda(n)^2\right\},$$

which implies

$$\sup_{\boldsymbol{\theta}\in\Gamma} \mathbb{E}_{X_{1:n} \sim \mathcal{N}(\boldsymbol{\theta}, I_D)}\left[\left\|\widehat{\boldsymbol{\theta}}^{n}_{\mathrm{sth}} - \boldsymbol{\theta}\right\|_2^2\right] \leq \sup_{\boldsymbol{\theta}\in\Gamma}\left\{\sum_{i=1}^D \frac{1}{n}\exp\left(-\frac{n\lambda(n)^2}{2}\right) + \min\left\{\theta_i^2, \frac{1}{n} + \lambda(n)^2\right\}\right\}.$$

With our choice of $\lambda(n)$ in (C.15), we have for any $1 \le i \le D$,

$$\sum_{i=1}^{D} \frac{1}{n} \exp\left(-\frac{n\lambda(n)^2}{2}\right) \le \frac{\varepsilon^2}{2} \quad \text{and} \quad \frac{1}{n} + \lambda(n)^2 \le \frac{1 + (2\log({}^{2D}/_{(n\varepsilon^2)}) \vee 0)}{n},$$

which implies

$$\sup_{\boldsymbol{\theta}\in\Gamma} \mathbb{E}_{X_{1:n}\sim\mathcal{N}(\boldsymbol{\theta},I_D)} \left[\left\|\widehat{\boldsymbol{\theta}}_{\text{sth}} - \boldsymbol{\theta}\right\|_2^2\right] \le \frac{\varepsilon^2}{2} + \sup_{\boldsymbol{\theta}\in\Gamma} \sum_{i=1}^{D} \min\left\{\theta_i^2, \frac{1 + \log({}^{2D}/_{(n\varepsilon^2)})}{n}\right\}.$$

We define set $L \subseteq \mathbb{R}^D$ as

$$L = \left\{\boldsymbol{\theta} = (\theta_1, \cdots, \theta_D) \;\middle|\; |\theta_i| \le \sqrt{\frac{2\log({}^{2D}/_{(n\varepsilon^2)}) \vee 0 + 1}{n}}\right\}$$

Note that according to Theorem C.1 we can replace $\sup_{\boldsymbol{\theta}\in\Gamma}$ with $\sup_{\boldsymbol{\theta}\in\Gamma\cap L}$ obtaining

$$\begin{aligned}
&\sup_{\boldsymbol{\theta}\in\Gamma} \sum_{i=1}^{D} \min\left\{\theta_i^2, \frac{2\log({}^{2D}/_{(n\varepsilon^2)}) \vee 0 + 1}{n}\right\} \\
&\quad = \sup_{\boldsymbol{\theta}\in\Gamma\cap L} \sum_{i=1}^{D} \min\left\{\theta_i^2, \frac{2\log({}^{2D}/_{(n\varepsilon^2)}) \vee 0 + 1}{n}\right\} = \sup_{\boldsymbol{\theta}\in\Gamma\cap L} \|\boldsymbol{\theta}\|^2.
\end{aligned}$$

We choose $n = n_{\text{sth}}(\varepsilon) - 1$. Since $\Gamma \cap L$ is compact, the above supreme is achieved at some $\boldsymbol{\theta}^* = (\theta_1^*, \cdots, \theta_D^*) \in \Gamma \cap L$, which has two properties. On one hand it satisfies

$$|\theta_i^*| \le \sqrt{\frac{2\log({}^{2D}/_{((n_{\text{est}}(\varepsilon)-1)\varepsilon^2)}) \vee 0 + 1}{n_{\text{sth}}(\varepsilon) - 1}}, \tag{C.19}$$

On the other hand, according to (C.18), we obtain that

$$\sum_{i=1}^{D} (\theta_i^*)^2 \ge \sup_{\boldsymbol{\theta}\in\Gamma} \mathbb{E}_{X_{1:n_{\text{sth}}(\varepsilon)-1} \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{\theta},I_D)} \left[\left\|\widehat{\boldsymbol{\theta}}_{\text{sth}}^{n_{\text{sth}}(\varepsilon)-1} - \boldsymbol{\theta}\right\|_2^2\right] - \frac{\varepsilon^2}{2} \ge \frac{\varepsilon^2}{2}.$$

Additionally, according to [70, Lemma 8.3], we have for any positive integer n and $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_D) \in \Gamma$,

$$\mathbb{E}_{X_{1:n} \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{\theta},I_D)} \left[\left\|\widehat{\boldsymbol{\theta}}_{\text{sth}}^{n} - \boldsymbol{\theta}\right\|_2^2\right] \ge \frac{1}{2} \cdot \sum_{i=1}^{D} \min\left\{\theta_i^2, \frac{1}{n} + \lambda(n)^2\right\}$$

We choose $n = n_{\text{sth}}(\varepsilon)$ and $\boldsymbol{\theta} = \boldsymbol{\theta}^* = (\theta_1^*, \cdots, \theta_D^*)$ defined above. (C.17) implies

$$\sum_{i=1}^{D} \min\left\{(\theta_i^*)^2, \frac{1}{n_{\text{sth}}(\varepsilon)} + \lambda(n_{\text{sth}}(\varepsilon))^2\right\} \le 2\sup_{\boldsymbol{\theta}\in\Gamma} \mathbb{E}_{X_{1:n_{\text{sth}}(\varepsilon)} \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{\theta},I_D)} \left[\left\|\widehat{\boldsymbol{\theta}}_{\text{sth}}^{n_{\text{sth}}(\varepsilon)} - \boldsymbol{\theta}\right\|_2^2\right] \le 2\varepsilon^2.$$

Further when $n_{\mathrm{sth}}(\varepsilon) \geq 4$, eq. (C.19) gives that for any $i \in [D]$,

$$\frac{1}{n_{\mathrm{sth}}(\varepsilon)} + \lambda(n_{\mathrm{sth}}(\varepsilon))^2 = \frac{1}{n_{\mathrm{sth}}(\varepsilon)} + \frac{2\log\left({}^{2D}/_{(n_{\mathrm{sth}}(\varepsilon)\varepsilon^2)}\right) \vee 0 + 1}{n_{\mathrm{sth}}(\varepsilon)}$$
$$\geq \frac{1}{2} \cdot \frac{2\log\left({}^{2D}/_{((n_{\mathrm{sth}}(\varepsilon)-1))\varepsilon^2)}\right) \vee 0 + 1}{n_{\mathrm{sth}}(\varepsilon) - 1} \geq \frac{1}{4}(\theta_i^*)^2,$$

Hence we obtain that

$$\sum_{i=1}^{D} (\theta_i^*)^2 \leq 4 \sum_{i=1}^{D} \min\left\{(\theta_i^*)^2, \frac{1}{n_{\mathrm{sth}}(\varepsilon)} + \lambda(n_{\mathrm{sth}}(\varepsilon))^2\right\} \leq 8\varepsilon^2.$$

Overall, we constructed $\boldsymbol{\theta}^* = (\theta_1^*, \cdots, \theta_D^*) \in \Gamma$ such that

(a) For any $1 \leq i \leq D$, $|\theta_i^*| \leq \sqrt{\frac{2\log\left({}^{2D}/_{((n_{\mathrm{est}}(\varepsilon)-1)\varepsilon^2)}\right)\vee 0+1}{n_{\mathrm{sth}}(\varepsilon)-1}}$.

(b) $\sum_{i=1}^{D} (\theta_i^*)^2 \geq {}^{\varepsilon^2}/_2$.

(c) $\sum_{i=1}^{D} (\theta_i^*)^2 \leq 8\varepsilon^2$.

Second Part: Goodness of Fit Rate: Suppose we already identified $\boldsymbol{\theta}^*$ which satisfies item (a), item (b) and item (c). We consider distribution $\nu = \nu_1 \otimes \nu_2 \otimes \cdots \otimes \nu_D$ where $\nu_i \in \Delta(\mathbb{R})$ is given by

$$\nu_i(\cdot) = \frac{\delta_{\theta_i^*}(\cdot)}{2} + \frac{\delta_{-\theta_i^*}(\cdot)}{2}.$$

Since $\boldsymbol{\theta}^* \in \Gamma$ and Γ is orthosymmetric, we have $\nu \in \Delta(\Gamma)$. And item (b) gives for any $\boldsymbol{\theta} \in \mathrm{supp}(\nu)$, we always have $\|\boldsymbol{\theta}\| \geq {}^{\varepsilon}/_{\sqrt{2}}$. Thus, according to [74, Proposition 2.11], for any testing scheme ψ with n samples for the following testing problem:

$$H_0 : \boldsymbol{\theta} = \mathbf{0} \quad \text{versus} \quad H_1 : \|\boldsymbol{\theta}\|_2 \geq \frac{\varepsilon}{\sqrt{2}},\ \boldsymbol{\theta} \in \Gamma$$

as long as $n \leq \sqrt{\frac{n_{\mathrm{est}}(\Gamma,\varepsilon)}{64\varepsilon^2\log(2D)}}$, we have

$$\sup_{P\in H_0} \mathbb{P}(\psi(X) \neq i) + \sup_{P\in H_1} \mathbb{P}(\psi(X) \neq i) \geq 1 - \frac{1}{2}\mathrm{TV}(\mathbb{E}_{\boldsymbol{\theta}\sim\nu}[P_{\boldsymbol{\theta}}^{\otimes n}], P_0^{\otimes n}),$$

where we denoted $P_{\boldsymbol{\theta}} = \mathcal{N}(\boldsymbol{\theta}, I_D)$. Hence, if we can show that the $D_{\mathrm{TV}} \geq \frac{1}{2}$ for a certain value of n it must imply that $n_{\mathrm{gof}}(\Gamma, {}^{\varepsilon}/_{\sqrt{2}}) \geq n$. We will indeed show that $n = \sqrt{\frac{n_{\mathrm{est}}(\Gamma,\varepsilon)}{64\varepsilon^2\log(2D)}}$ works.

To that end, we first use a standard bound [211, Proposition 7.15]

$$\mathrm{TV}(\mathbb{E}_{P\sim\mu(\nu)}[P^{\otimes n}], P_0^{\otimes n}) \leq \sqrt{\frac{1}{4}\chi^2\left(\mathbb{E}_{P\sim\mu(\nu)}[P^{\otimes n}] \| P_0^{\otimes n}\right)}.$$

Next, we have the standard bound of Ingster

$$\chi^2\left(\mathbb{E}_{P\sim\mu(\nu)}[P^{\otimes n}] \| P_0^{\otimes n}\right) \overset{(i)}{\leq} \prod_{i=1}^{D} \exp\left(\frac{1}{2}n^2(\theta_i^*)^4\right) - 1 = \exp\left(\frac{n^2}{2}\sum_{i=1}^{D}(\theta_i^*)^4\right) - 1,$$

where (i) uses [60, (3.68)] and the inequality $\frac{1}{2}\exp(x)+\frac{1}{2}\exp(-x)\le\exp\left(\frac{1}{2}x^2\right)$. Next according to item (a) and item (c), we obtain that when $n_{\mathrm{est}}(\Gamma,\varepsilon)\ge 4$,

$$\begin{aligned}\sum_{i=1}^{D}(\theta_i^*)^4 &\le \left(\max_{i\in[D]}|\theta_i^*|\right)^2\cdot\left(\sum_{i=1}^{D}(\theta_i^*)^2\right)\\ &\le 8\varepsilon^2\cdot\frac{2\log\left({}^{2D}/_{((n_{\mathrm{est}}(\varepsilon)-1)\varepsilon^2)}\right)\vee 0+1}{n_{\mathrm{sth}}(\varepsilon)-1}\le 64\varepsilon^2\cdot\frac{\log\left({}^{2D}/_{(n_{\mathrm{est}}(\Gamma,\varepsilon)\varepsilon^2)}\right)\vee 0+1}{n_{\mathrm{est}}(\Gamma,\varepsilon)}.\end{aligned}$$

Next, we notice that when the diameter $\mathrm{diam}(\Gamma)$ of Γ is less than ε, then the density estimation rate $n_{\mathrm{est}}(\Gamma,\varepsilon)$ is zero, hence the inequality in Theorem 4.7 obviously holds. When $\mathrm{diam}(\Gamma)\ge\varepsilon$, the density estimation complexity $n_{\mathrm{est}}(\Gamma,\varepsilon)$ satisfies $n_{\mathrm{est}}(\Gamma,\varepsilon)\ge\varepsilon^{-2}$. Hence we obtain that

$$\log\left({}^{2D}/_{(n_{\mathrm{est}}(\Gamma,\varepsilon)\varepsilon^2)}\right)\vee 0+1\le\log(2D)$$

Hence when n satisfies

$$n^2\le\frac{n_{\mathrm{est}}(\Gamma,\varepsilon)}{64\varepsilon^2\cdot\log(2D)},$$

we have

$$\chi^2\left(\mathbb{E}_{P\sim\mu(\nu)}[P^{\otimes n}]\|P_0^{\otimes n}\right)\le\exp(1/2)-1\le\frac{2}{3},$$

completing the proof. □

Proof of Theorem 4.8. We fix $\varepsilon>0$ and let $D=D_{\mathsf{c}}(\Gamma,\varepsilon)$. According to the definition of coordinate-wise Kolmogorov dimension Theorem 4.2, there exists a subset A of $\mathbb{Z}_+$ such that

$$\sup_{\boldsymbol{\theta}\in\Gamma}\sum_{i\in\mathbb{Z}_+\setminus A}(\theta_i)^2\le 2\varepsilon^2.$$

Without loss of generality that we assume $A=\{1,2,\cdots,D\}$, and we use Π_D to denote the projection operator onto the first D coordinates.

We construct $\Gamma_D=\{\Pi_D\boldsymbol{\theta}:\boldsymbol{\theta}\in\Gamma\}$. Then for any density estimator $\widehat{\boldsymbol{\theta}}_D$ for Γ_D, we can construct a density estimator $\widehat{\boldsymbol{\theta}}$: when taking X, we define estimator $\widehat{\boldsymbol{\theta}}$ as

$$\widehat{\boldsymbol{\theta}}(X)=\big(\underbrace{\widehat{\theta}_1(X),\widehat{\theta}_2(X),\cdots,\widehat{\theta}_k(X)}_{\text{first } D \text{ coordinates match } \widehat{\boldsymbol{\theta}}_D(\Pi_D X)},0,0,\cdots\big),$$

where $\Pi_D X$ denotes the data after taking projection Π_D for all elements in X. Then we have for any $\boldsymbol{\theta}\in\Gamma$,

$$\mathbb{E}\left[\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}(X)\|_2^2\right]=\mathbb{E}\left[\|\Pi_D\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_D(\Pi_D X)\|_2^2\right]+\|\Pi_D\boldsymbol{\theta}-\boldsymbol{\theta}\|^2\le\mathbb{E}\left[\|\Pi_D\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_D(\Pi_D X)\|_2^2\right]+2\varepsilon^2.$$

Notice that $\Pi_D X$ can be viewed as i.i.d. samples collected through $\mathcal{N}(\Pi_D\boldsymbol{\theta},I_D)$, we obtain that

$$n_{\mathrm{est}}(\Gamma,\sqrt{3}\varepsilon)\le n_{\mathrm{est}}(\Gamma_D,\varepsilon).\tag{C.20}$$

Additionally, since Γ is orthosymmetric and convex, we have $\Gamma_D\subseteq\Gamma$. When carrying the goodness of fit test for Γ_D, the data after D-th coordinate are generated according to $\mathcal{N}(0,1)$

and independent to the parameter $\boldsymbol{\theta} \in \Gamma_D$. Therefore, carrying goodness of fit testing to Γ is no easier than carrying goodness of fit testing to Γ_D, which implies

$$n_{\text{gof}}\left(\Gamma, \frac{\varepsilon}{\sqrt{2}}\right) \geq n_{\text{gof}}\left(\Gamma_D, \frac{\varepsilon}{\sqrt{2}}\right). \tag{C.21}$$

Since Γ is compact, orthosymmetric and convex, Γ_D is a D-dimensional compact, orthosymmetric and convex set as well. Hence Theorem 4.7 implies that

$$n_{\text{gof}}\left(\Gamma_D, \frac{\varepsilon}{\sqrt{2}}\right)^2 \geq \frac{1}{64\log(2D)} \cdot \frac{n_{\text{est}}(\Gamma_D, \varepsilon)}{\varepsilon^2}.$$

Bringing in eq. (C.20) and eq. (C.21), we obtain that

$$\begin{aligned} n_{\text{gof}}\left(\Gamma, \frac{\varepsilon}{\sqrt{2}}\right)^2 &\geq \frac{1}{64\log(2D)} \cdot \frac{n_{\text{est}}(\Gamma, \sqrt{3}\varepsilon)}{\varepsilon^2} \\ &= \frac{1}{64\log(2D)} \cdot \frac{n_{\text{est}}(\Gamma, \sqrt{3}\varepsilon)}{\varepsilon^2}. \end{aligned}$$

□

C.3 Discussion between Kolmogorov Dimension and Coordinate-wise Kolmogorov Dimension

For any orthosymmetric set Γ, we have the following relationship, which provides an upper bound of the coordinate-wise Kolmogorov dimension $D_{\mathsf{c}}(\Gamma, \varepsilon)$ in terms of the traditional Kolmogorov dimension $D(\Gamma, \varepsilon)$.

Proposition C.2. *For any orthosymmetric, convex, compact set Γ, we have*

$$D\left(\Gamma, \frac{\varepsilon}{D_{\mathsf{c}}(\Gamma, \varepsilon)}\right) \geq \frac{D_{\mathsf{c}}(\Gamma, \varepsilon)}{2} - 2.$$

The proof of Theorem C.2 requires the following lemmas.

Lemma C.3. *Suppose set $\Lambda = \{\mathbf{e}_1, \cdots, \mathbf{e}_d\}$ consists of orthogonal unit vectors. Then for any $(d-1)$-dimensional projection Π, we have*

$$\max_{i \in [d]} \|\mathbf{e}_i - \Pi[\mathbf{e}_i]\| \geq \frac{1}{\sqrt{d}}. \tag{C.22}$$

Proof of Theorem C.3. Suppose $\mathbf{v}_1, \cdots, \mathbf{v}_{d-1}$ to be a set of orthogonal basis of the image of projection Π. Then for any $i \in [d]$,

$$\|\mathbf{e}_i - \Pi[\mathbf{e}_i]\|^2 = \|\mathbf{e}_i\|^2 - \|\Pi[\mathbf{e}_i]\|^2 = 1 - \sum_{j=1}^{d-1} \langle \mathbf{e}_i, \mathbf{v}_j \rangle^2,$$

which implies that

$$\sum_{i=1}^{d} \|\mathbf{e}_i - \Pi[\mathbf{e}_i]\|^2 = d - \sum_{i=1}^{d}\sum_{j=1}^{d-1} \langle \mathbf{e}_i, \mathbf{v}_j\rangle^2 = d - \sum_{j=1}^{d-1}\left(\sum_{i=1}^{d} \langle \mathbf{e}_i, \mathbf{v}_j\rangle^2\right).$$

Since $\mathbf{e}_1, \cdots, \mathbf{e}_d$ are orthogonal unit vectors,

$$\sum_{i=1}^{d} \langle \mathbf{e}_i, \mathbf{v}_j\rangle^2 = \left\|\Pi_{\mathrm{span}(\mathbf{e}_1,\cdots,\mathbf{e}_d)}[\mathbf{v}_j]\right\| \le 1, \qquad \forall j \in [d-1],$$

where the above Π denotes the projection onto the space spanned by $\mathbf{e}_1, \cdots, \mathbf{e}_d$. Therefore,

$$\sum_{i=1}^{d} \|\mathbf{e}_i - \Pi[\mathbf{e}_i]\| \ge d - (d-1) = 1,$$

which implies eq. (C.22). □

Lemma C.4. *For orthosymmetric, convex, compact set Γ, suppose $d = D_{\mathsf{c}}(\Gamma, \varepsilon) - 1$, then there exists d orthogonal vectors $\mathbf{u}_1, \cdots, \mathbf{u}_{\lfloor d/2\rfloor} \in \Gamma$ such that for any $1 \le i \le \lfloor d/2\rfloor$, $\|\mathbf{u}_i\| \ge \varepsilon/\sqrt{d}$.*

Proof of Theorem C.4. Suppose $\mathbf{e}_1, \mathbf{e}_2, \cdots$ are unit vectors of each coordinates, and we let $v_i = \sup\{v \ge 0 : v\mathbf{e}_i \in \Gamma\}$. Then since Γ is closed, we have $\mathbf{v}_i = v_i\mathbf{e}_i \in \Gamma$. Without loss of generality we assume $v_1 \ge v_2 \ge \cdots$. Then according to the definition of coordinate-wise Kolmogorov dimension in Theorem 4.2, we have

$$\sum_{i=d+1}^{D} (v_i)^2 > \varepsilon^2.$$

If $v_d \ge \varepsilon/\sqrt{d}$, then by choosing $\mathbf{u}_i = \mathbf{v}_i$, these vectors satisfy the conditions. Next we assume $v_d < \varepsilon/\sqrt{d}$. To construct $\mathbf{u}_1, \cdots, \mathbf{u}_{\lfloor d/2\rfloor}$, we initiate the following process: letting $\mathbf{u}_1 = \sum_{i=d+1}^{k_1} (v_i)^2$, where k_1 is the smallest number such that $\|\mathbf{u}_1\| \ge \varepsilon/\sqrt{d}$. Then let $\mathbf{u}_2 = \sum_{i=k_1+1}^{k_2} (v_i)^2$ where k_2 is the smallest number such that $\|\mathbf{u}_2\| \ge \varepsilon/\sqrt{d}$, and so on. Since $v_i \le v_d < \varepsilon/\sqrt{d}$ holds for any $i \ge d+1$, the above construction gives that

$$\frac{\varepsilon}{\sqrt{d}} \le \|\mathbf{u}_i\| \le \frac{\sqrt{2}\varepsilon}{\sqrt{d}}.$$

Therefore, since $\sum_{i=d+1}^{D} (v_i)^2 > \varepsilon^2$, the above process can proceed at least $\lfloor d/2\rfloor$ times. Hence it is eligible to construct $\mathbf{u}_1, \cdots, \mathbf{u}_{\lfloor d/2\rfloor}$ so that $\|\mathbf{u}_i\| \ge \varepsilon/\sqrt{d}$ holds. It is easy to verify that $\mathbf{u}_1, \cdots, \mathbf{u}_{\lfloor d/2\rfloor}$ are orthogonal. And since Γ is orthosymmetric and convex, we have $\mathbf{u}_i \in \Gamma$ for any i. □

Now we are ready to prove Theorem C.2.

Proof of Theorem C.2. Let $d = D_{\mathsf{c}}(\Gamma, \varepsilon) - 1$. According to Theorem C.4, there exists orthogonal vectors $\mathbf{u}_1, \cdots, \mathbf{u}_{\lfloor d/2 \rfloor} \in \Gamma$ such that $\|\mathbf{u}_i\| \geq \varepsilon/\sqrt{d}$. By injecting orthogonal unit vectors $\sqrt{d}\mathbf{u}_i/\varepsilon$ into Theorem C.3, we have

$$\inf_{\Pi} \max_{i \in \lfloor d/2 \rfloor} \|\mathbf{u}_i - \Pi[\mathbf{u}_i]\| \geq \frac{\varepsilon}{\sqrt{d} \cdot \sqrt{\lfloor d/2 \rfloor}} > \frac{\varepsilon}{D_{\mathsf{c}}(\Gamma, \varepsilon)},$$

where the infimum is over all possible $\lfloor d/2 \rfloor - 1$ projections. Since $\mathbf{u}_i \in \Gamma$, we obtain that

$$\inf_{\Pi} \max_{\mathbf{u} \in \Gamma} \|\mathbf{u} - \Pi[\mathbf{u}]\| > \frac{\varepsilon}{D_{\mathsf{c}}(\Gamma, \varepsilon)}.$$

Hence we have the following lower bound to the traditional Kolmogorov dimension:

$$D\left(\Gamma, \frac{\varepsilon}{D_{\mathsf{c}}(\Gamma, \varepsilon)}\right) \geq \lfloor d/2 \rfloor - 1 \geq \frac{D_{\mathsf{c}}(\Gamma, \varepsilon)}{2} - 2.$$

□

Theorem C.2 has the following direct corollary.

Corollary C.5. *Suppose orthosymmetric, convex, compact set Γ satisfies $D(\Gamma, \varepsilon) \lesssim \varepsilon^{-p}$ for some $0 < p < 1$. Then the coordinate-wise Kolmogorov dimension satisfies*

$$D_{\mathsf{c}}(\Gamma, \varepsilon) \lesssim \varepsilon^{-p/(1-p)}.$$

C.4 Alternative proof of Theorem 4.14

We present a version of the second inequality in Theorem 4.14 in this section.

Proposition C.6. *Suppose Γ any set such that the projection estimator is minimax optimal up to constants, i.e. there exists a positive constant c such that*

$$\inf_{\Pi} \sup_{\boldsymbol{\theta} \in \Gamma} \mathbb{E}\left[\|\Pi(X) - \boldsymbol{\theta}\|_2^2\right] \leq c \cdot \inf_{\widehat{\boldsymbol{\theta}}} \sup_{\boldsymbol{\theta} \in \Gamma} \mathbb{E}\left[\left\|\widehat{\boldsymbol{\theta}}(X) - \boldsymbol{\theta}\right\|_2^2\right], \tag{C.23}$$

where Π denotes the class of projection estimators, and $\boldsymbol{\theta}$ denotes any estimators. Then we have

$$n_{\mathrm{gof}}(\Gamma, 2\sqrt{c}\varepsilon)^2 \leq \frac{81 n_{\mathrm{est}}(\Gamma, \varepsilon)}{\varepsilon^2}.$$

We remind that [63] famously showed that (**??**) holds for QCO sets. Thus, the result above strictly generalizes second inequality in Theorem 4.14.

Proof. We let $D(\Gamma, \sqrt{c}\varepsilon)$ to be the Kolmogorov dimension of set Γ at scale $\sqrt{c}\varepsilon$ (the definition of Kolmogorov dimension is given in Theorem 4.1). We use $\mathcal{S}^\star$ to denote the $D(\Gamma, 3\varepsilon)$-dimensional subspace which achieves the minimizer in the definition of Kolmogorov

dimension Theorem 4.1, and $\Pi_{\mathcal{S}^\star}$ is the projection into $\mathcal{S}^\star$. According to [65], as long as $n \geq 9\sqrt{D(\Gamma, \sqrt{c}\varepsilon)}/(c\varepsilon^2)$, there exists a constant c such that the testing scheme

$$\psi(X) = \mathbb{I}\left[\left\|\Pi_{\mathcal{S}^\star}\left[\frac{1}{n}\sum_{i=1}^n X_i\right]\right\|_2^2 \geq c\right],$$

satisfies

$$\mathbb{P}_{X \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, I)}(\psi(X) = 1) \leq \frac{1}{4} \quad \text{and} \quad \mathbb{P}_{X \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{\theta}, I)}(\psi(X) = 0) \leq \frac{1}{4} \quad \text{for any } \|\boldsymbol{\theta}\|_2 \geq 2\sqrt{c}.$$

This implies that

$$n_{\text{gof}}(\Gamma, 2\sqrt{c}\varepsilon) \leq \frac{9\sqrt{D(\Gamma, \sqrt{c}\varepsilon)}}{c\varepsilon^2}.$$

Next, according to eq. (C.23),

$$\begin{aligned}\inf_{\widehat{\boldsymbol{\theta}}} \sup_{\boldsymbol{\theta}\in\Gamma} \mathbb{E}\left[\left\|\widehat{\boldsymbol{\theta}}(X) - \boldsymbol{\theta}\right\|_2^2\right] &\geq \frac{1}{c} \cdot \inf_{\Pi} \sup_{\boldsymbol{\theta}\in\Gamma} \mathbb{E}\left[\|\Pi(X) - \boldsymbol{\theta}\|_2^2\right] \\ &\geq \frac{1}{c} \cdot \inf_{d} \left\{\inf_{\Pi_d} \sup_{\boldsymbol{\theta}\in\Gamma} \left\{\|\Pi_d(\boldsymbol{\theta}) - \boldsymbol{\theta}\|_2^2 + \frac{d}{n}\right\}\right\}.\end{aligned}$$

where $\inf_\Pi$ denotes infimum over all projections. Hence when $n < D(\Gamma, \sqrt{c}\varepsilon)/(c\varepsilon^2)$, for any integer $d < D(\Gamma, \sqrt{c}\varepsilon)$,

$$\inf_{\Pi_d} \sup_{\boldsymbol{\theta}\in\Gamma} \left\{\|\Pi_d(\boldsymbol{\theta}) - \boldsymbol{\theta}\|_2^2 + \frac{d}{n}\right\} > \inf_{\Pi_d} \sup_{\boldsymbol{\theta}\in\Gamma} \|\Pi_d(\boldsymbol{\theta}) - \boldsymbol{\theta}\|_2^2 \geq (\sqrt{c}\varepsilon)^2 = c\varepsilon^2,$$

and when $d \geq D(\Gamma, \sqrt{c}\varepsilon)$,

$$\inf_{\Pi_d} \sup_{\boldsymbol{\theta}\in\Gamma} \left\{\|\Pi_d(\boldsymbol{\theta}) - \boldsymbol{\theta}\|_2^2 + \frac{d}{n}\right\} \geq \frac{d}{n} > c\varepsilon^2.$$

Therefore, we obtain that

$$\inf_{\widehat{\boldsymbol{\theta}}} \sup_{\boldsymbol{\theta}\in\Gamma} \mathbb{E}\left[\left\|\widehat{\boldsymbol{\theta}}(X) - \boldsymbol{\theta}\right\|_2^2\right] > \frac{c\varepsilon^2}{c} = \varepsilon^2,$$

which implies that

$$81 n_{\text{est}}(\Gamma, \varepsilon) \geq \frac{81 D(\Gamma, \sqrt{c}\varepsilon)}{c\varepsilon^2} \geq \varepsilon^2 \cdot n_{\text{gof}}(\Gamma, 2\sqrt{c}\varepsilon)^2.$$

□

We note that we use [63] to show that a quadratically convex and orthosymmetric set Θ has a lower bound on the estimation sample complexity $n_{\text{est}} \gtrsim \frac{D(\varepsilon)}{\varepsilon^2}$. This result would automatically follow if we could show a geometric result that any such set necessarily contains $\gtrsim D(\varepsilon)$ orthogonal vectors of length ε (or, equivalently, contains a ball of dimension D and radius ε), since then the lower bound would follow from standard results on Gaussian location model, e.g. [211, Theorem 30.1]. We summarize this observation into the following conjecture. It is an interesting question in convex geometry to prove or disprove it.

Conjecture There exist universal positive constants c_1 and c_2 such that for any orthosymmetric, compact, convex and quadratically convex set Θ, there exist $D(\Theta, c_1\varepsilon)$ orthogonal vectors of length $(c_2\varepsilon)$ all contained within Θ.

C.5 Analysis of LFHT for General Convex Sets

C.5.1 Proof of Theorem 4.16

Proof of Theorem 4.16. Same as the proof of Theorem 4.7, there exists $\boldsymbol{\theta}^* = (\theta_1^*, \cdots, \theta_D^*) \in \Gamma$ such that

(a) For any $1 \le i \le D$, $|\theta_i^*| \le 2\sqrt{\frac{\log(2D)}{n_{\text{est}}(\Gamma, \sqrt{2}\varepsilon)-1}}$.

(b) $\sum_{i=1}^d (\theta_i^*)^2 \ge \varepsilon^2$.

(c) $\sum_{i=1}^d (\theta_i^*)^2 \le 16\varepsilon^2$.

Next, we use these three properties to bound the LFHT testing region of set Γ. First of all, according to [15, Proposition 1], if (m, n) lies in the LFHT testing region, we must have

$$m \ge \frac{1}{\varepsilon^2} \quad \text{and} \quad n \ge n_{\text{gof}}(\Gamma, \varepsilon) \ge \frac{1}{8\sqrt{\log(2D)}} \cdot \frac{\sqrt{n_{\text{est}}(\Gamma, \sqrt{2}\varepsilon)}}{\varepsilon}, \tag{C.24}$$

where the last inequality follows from Theorem 4.7. Hence we only need to verify that if (m, n) lies in the LFHT testing region, then

$$mn \ge \frac{n_{\text{est}}(\Gamma, \sqrt{2}\varepsilon) - 1}{3072\varepsilon^2 \cdot \log(2D)}. \tag{C.25}$$

Without loss of generality, we assume

$$m \le \frac{\sqrt{n_{\text{est}}(\Gamma, \sqrt{2}\varepsilon) - 1}}{192\sqrt{\log(2D)}}, \tag{C.26}$$

otherwise the inequality eq. (C.25) follows from the lower bound to n in eq. (C.24). Next we will verify that if (m, n) satisfies

$$mn \le \frac{n_{\text{est}}(\Gamma, \sqrt{2}\varepsilon) - 1}{3072\varepsilon^2 \cdot \log(2D)}, \tag{C.27}$$

then eq. (4.10) fails.

For any fixed distribution $\mu \in \Delta(\Gamma)$, we define distribution $\mathbb{P}_{0,XYZ}$ to be the distribution of $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ sampled according to the following way: first sample $\boldsymbol{\theta} \sim \mu$, then sample $\mathbf{X} = (\mathbf{X}^{1:n}) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{\theta}, I_D)$, $\mathbf{Y} = (\mathbf{Y}^{1:n}) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, I_D)$ and $\mathbf{Z} = (\mathbf{Z}^{1:m}) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{\theta}, I_D)$. And we define distribution $\mathbb{P}_{1,XYZ}$ to be the distribution of $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ sampled according to the following way: first sample $\boldsymbol{\theta} \sim \mu$, then sample $\mathbf{X} = (\mathbf{X}^{1:n}) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{\theta}, I_D)$, $\mathbf{Y} = (\mathbf{Y}^{1:n}) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, I_D)$ and $\mathbf{Z} = (\mathbf{Z}^{1:m}) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, I_D)$. Similarly, we can define distribution $\mathbb{P}_{0,XZ}, \mathbb{P}_{1,XZ}, \mathbb{P}_{0,X}, \mathbb{P}_{1,X}$, and also conditional distribution $\mathbb{P}_{0,Z|X}$ and $\mathbb{P}_{1,Z|X}$. Then [15, Lemma 5] gives that for any such μ supported in $\Gamma \backslash B_2(\varepsilon)$,

$$\inf_{\psi} \max_{i \in \{0,1\}} \sup_{P \in H_i} \mathbb{P}(\psi(X, Y, Z) \ne i) \ge \frac{1}{2}\left(1 - \text{TV}(\mathbb{P}_{0,XYZ}, \mathbb{P}_{1,XYZ})\right). \tag{C.28}$$

In the following proof, we choose distribution $\mu \in \Delta(\Gamma)$ to be the product of symmetric ternary distributions, i.e. for $\theta_i = (\theta_1, \cdots, \theta_D) \in \Gamma$,

$$\mu(\theta) = \prod_{j=1}^{D} \mu_j(\theta_j) \quad \text{where} \quad \mu_j = \frac{\delta_{\theta_j^\star}}{2} + \frac{\delta_{-\theta_j^\star}}{2}.$$

According to item (c) and the property of orthosymmetric, we have $\mu \in \Delta(\Gamma)$. We can also verify that for any $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_D)$ which belongs to the support of μ, we have $|\theta_j| = |\theta_j^\star|$ for any $j \in [D]$. Hence

$$\|\boldsymbol{\theta}\|_2^2 = \sum_{j=1}^{D} (\theta_j)^2 = \sum_{j=1}^{D} (\theta_j^\star)^2 \geq \varepsilon^2,$$

where the last inequality uses item (b). This verifies that μ is supported in $\Gamma \backslash B_2(\varepsilon)$.

Next, we will calculate the TV distance $\mathrm{TV}(\mathbb{P}_{0,XYZ}, \mathbb{P}_{1,XYZ})$:

$$\begin{aligned}
\mathrm{TV}(\mathbb{P}_{0,XYZ}, \mathbb{P}_{1,XYZ})^2 &= \mathrm{TV}(\mathbb{P}_{0,XZ}, \mathbb{P}_{1,XZ})^2 \leq D_{\mathrm{KL}}(\mathbb{P}_{0,XZ} \parallel \mathbb{P}_{1,XZ}) \\
&= D_{\mathrm{KL}}(\mathbb{P}_{0,Z|X} \parallel \mathbb{P}_{1,Z|X} \mid \mathbb{P}_{0,X}) + D_{\mathrm{KL}}(\mathbb{P}_{0,X} \parallel \mathbb{P}_{1,X}) \\
&= D_{\mathrm{KL}}(\mathbb{P}_{0,Z|X} \parallel \mathbb{P}_{1,Z|X} \mid \mathbb{P}_{0,X}) \leq \chi^2(\mathbb{P}_{0,Z|X} \parallel \mathbb{P}_{1,Z|X} \mid \mathbb{P}_{0,X}). \qquad \text{(C.29)}
\end{aligned}$$

Hence in order to bound the above TV distance, we only need to upper bound the conditional χ^2-divergence in the right hand side. Using $\varphi_{\boldsymbol{\theta}}(\cdot)$ to denote the density function of $\mathcal{N}(\boldsymbol{\theta}, I_D)$, according to Ingster's trick [60], we have

$$\begin{aligned}
&\chi^2(\mathbb{P}_{0,Z|X} \parallel \mathbb{P}_{1,Z|X} \mid \mathbb{P}_{0,X}) + 1 \\
&= \mathbb{E}_{\mathbf{X} \sim \mathbb{P}_{0,X}} \left[\mathbb{E}_{\boldsymbol{\theta}|\mathbf{X}, \boldsymbol{\theta}'|\mathbf{X}} \left[\int_{(\mathbb{R}^D)^m} \prod_{t=1}^{m} \frac{\varphi_{\boldsymbol{\theta}}(\mathbf{z}^t) \varphi_{\boldsymbol{\theta}'}(\mathbf{z}^t)}{\varphi_{\mathbf{0}}(\mathbf{z}^t)} d\mathbf{z}_1 \cdots d\mathbf{z}_m \mid \mathbf{X} \right] \right] \\
&= \mathbb{E}_{\mathbf{X} \sim \mathbb{P}_{0,X}} \left[\mathbb{E}_{\boldsymbol{\theta}|\mathbf{X}, \boldsymbol{\theta}'|\mathbf{X}} [\exp\left(m \langle \boldsymbol{\theta}, \boldsymbol{\theta}' \rangle\right) \mid \mathbf{X}] \right] \\
&= \mathbb{E}_{\mathbf{X} \sim \mathbb{P}_{0,X}} \left[\prod_{j=1}^{D} \left(\mathbb{P}(\theta_j = \theta_j' \mid \mathbf{X}) \cdot \exp(m(\theta_i^\star)^2) + \mathbb{P}(\theta_j \neq \theta_j' \mid \mathbf{X}) \cdot \exp(-m(\theta_i^\star)^2) \right) \right] \\
&= \prod_{j=1}^{D} \mathbb{E}_{\mathbf{X} \sim \mathbb{P}_{0,X}} \left[\mathbb{P}(\theta_j = \theta_j' \mid \mathbf{X}) \cdot \exp(m(\theta_i^\star)^2) + \mathbb{P}(\theta_j \neq \theta_j' \mid \mathbf{X}) \cdot \exp(-m(\theta_i^\star)^2) \right], \qquad \text{(C.30)}
\end{aligned}$$

where $\boldsymbol{\theta}, \boldsymbol{\theta}'$ are i.i.d. sampled according to $\mathbb{P}(\boldsymbol{\theta} \mid \mathbf{X})$, and the last equation uses the fact that conditioned on $\mathbf{X}$, we have (θ_j, θ_j') independent to each other for any $j \in [D]$. In the following, we write $\mathbf{X} = \mathbf{X}^{1:n}$ which denotes the n samples, and we further denote $\mathbf{X}^i = (X_1^i, \cdots, X_D^i)$, where X_j^i denotes the j-th coordinate of $\mathbf{X}^i$. We notice that θ_j only depends on $X_j^{1:n} = (X_j^1, \cdots, X_j^n)$. According to Bayes rule, we can calculate

$$\begin{aligned}
\mathbb{P}(\theta_j = 1 \mid \mathbf{X}) &= \frac{\Pr(X_j^{1:n}, \eta = 1)}{\Pr(X_j^{1:n}, \eta_j = -1) + \Pr(X_j^{1:n}, \eta_j = 1)} \\
&= \frac{\prod_{i=1}^{n} \exp\left(-(X_j^i - \theta_j^\star)^2/2\right)}{\prod_{i=1}^{n} \exp\left(-(X_j^i - \theta_j^\star)^2/2\right) + \prod_{i=1}^{n} \exp\left(-(X_j^i + \theta_j^\star)^2/2\right)}
\end{aligned}$$

$$= \frac{\exp\left(\theta_j^\star \cdot \sum_{i=1}^n X_j^i\right)}{\exp\left(\theta_j^\star \cdot \sum_{i=1}^n X_j^i\right) + \exp\left(-\theta_j^\star \cdot \sum_{i=1}^n X_j^i\right)}. \tag{C.31}$$

Similarly, we get

$$\mathbb{P}(\theta_j = -1 \mid \mathbf{X}) = \frac{\exp\left(-\theta_j^\star \cdot \sum_{i=1}^n X_j^i\right)}{\exp\left(\theta_j^\star \cdot \sum_{i=1}^n X_j^i\right) + \exp\left(-\theta_j^\star \cdot \sum_{i=1}^n X_j^i\right)}. \tag{C.32}$$

In the following, we use $[0,1]$-valued random variable $p_j(\mathbf{X})$ to denote

$$p_j(\mathbf{X}) = \mathbb{P}(\theta_j = 1 \mid \mathbf{X}), \qquad \forall j \in [D].$$

We further notice that θ'_j and θ_j are i.i.d. conditioned on $\mathbf{X}$. Hence we obtain

$$\begin{aligned}
&\mathbb{P}(\theta_j = \theta'_j \mid \mathbf{X}) \cdot \exp(m(\theta_i^\star)^2) + \mathbb{P}(\theta_j \neq \theta'_j \mid \mathbf{X}) \cdot \exp(-m(\theta_i^\star)^2) \\
&\quad = (p_j(\mathbf{X})^2 + (1 - p_j(\mathbf{X}))^2) \exp(m(\theta_j^\star)^2) + 2p_j(\mathbf{X})(1 - p_j(\mathbf{X})) \exp(-m(\theta_j^\star)^2) \\
&\quad = \exp(m(\theta_j^\star)^2) + \exp(-m(\theta_j^\star)^2) - 1 + \frac{(1 - 2p_j(\mathbf{X}))^2}{2} \cdot \left(\exp(m(\theta_j^\star)^2) - \exp(-m(\theta_j^\star)^2)\right).
\end{aligned} \tag{C.33}$$

Next according to eq. (C.26) and item (a), we have for any $j \in [D]$, $m(\theta_j^\star)^2 \leq 1$, which implies that

$$\begin{aligned}
&\exp(m(\theta_j^\star)^2) + \exp(-m(\theta_j^\star)^2) - 2 \leq 1 + 4m^2(\theta_j^\star)^4 \\
\text{and} \qquad &\exp(m(\theta_j^\star)^2) - \exp(-m(\theta_j^\star)^2) \leq 4m(\theta_j^\star)^2.
\end{aligned} \tag{C.34}$$

We further notice that

$$\begin{aligned}
\frac{(1 - 2p_j(\mathbf{X}))^2}{2} &= \frac{(\exp\left(\theta_j^\star \cdot \sum_{i=1}^n X_j^i\right) - \exp\left(-\theta_j^\star \cdot \sum_{i=1}^n X_j^i\right))^2}{2(\exp\left(\theta_j^\star \cdot \sum_{i=1}^n X_j^i\right) + \exp\left(-\theta_j^\star \cdot \sum_{i=1}^n X_j^i\right))^2} \\
&\leq \frac{(\exp\left(\theta_j^\star \cdot \sum_{i=1}^n X_j^i\right) - \exp\left(-\theta_j^\star \cdot \sum_{i=1}^n X_j^i\right))^2}{8},
\end{aligned}$$

which implies that

$$\begin{aligned}
\mathbb{E}_{\mathbf{X} \sim \mathbb{P}_{0,X}}\left[\frac{(1 - 2p_j(\mathbf{X}))^2}{2}\right] &\leq \mathbb{E}_{\mathbf{X} \sim \mathbb{P}_{0,X}}\left[\frac{(\exp\left(\theta_j^\star \cdot \sum_{i=1}^n X_j^i\right) - \exp\left(-\theta_j^\star \cdot \sum_{i=1}^n X_j^i\right))^2}{8}\right] \\
&= \frac{\exp(4n(\theta_j^\star)^2) - 1}{8},
\end{aligned}$$

where the last equation uses the equation $\mathbb{E}[\exp(\alpha X)] = \exp(\alpha\mu + \alpha^2\sigma^2/2)$ for $X \sim \mathcal{N}(\mu, \sigma)$, and also the way of sampling $\mathbf{X}$ from $\mathbb{P}_{0,X}$. For $j \in [D]$, if $n(\theta_j^\star)^2 \leq 1$, then according to the inequality $\exp(x) - 1 \leq 2x$ for $0 \leq x \leq 1$, we have $\mathbb{E}_{\mathbf{X} \sim \mathbb{P}_{0,X}}\left[(1 - 2p_j(\mathbf{X}))^2/2\right] \leq n(\theta_j^\star)^2$. If $n(\theta_j^\star)^2 \geq 1$, since $p_j(\mathbf{X}) \in [0,1]$, we have $\mathbb{E}_{\mathbf{X} \sim \mathbb{P}_{0,X}}\left[(1 - 2p_j(\mathbf{X}))^2/2\right] \leq 1 \leq n(\theta_j^\star)^2$. Hence we obtain that for any $j \in [D]$,

$$\mathbb{E}_{\mathbf{X} \sim \mathbb{P}_{0,X}}\left[\frac{(1 - 2p_j(\mathbf{X}))^2}{2}\right] \leq n(\theta_j^\star)^2.$$

Bring this inequality and eq. (C.34) back to eq. (C.30), we obtain that

$$\chi^2(\mathbb{P}_{0,Z|X} \parallel \mathbb{P}_{1,Z|X} \mid \mathbb{P}_{0,X}) + 1 \le \prod_{j=1}^{D} \left(1 + 4m^2(\theta_j^\star)^4 + 4mn(\theta_j^\star)^4\right)$$
$$\le \left(1 + \frac{4m^2 + 4mn}{D} \cdot \sum_{j=1}^{D} (\theta_j^\star)^4\right)^D,$$

where the last inequality uses the Jensen's inequality. Therefore, if eq. (C.27) holds, then together with eq. (C.26) and also item (a) and item (c), we have

$$(4m^2 + 4mn) \cdot \sum_{j=1}^{D} (\theta_j^\star)^4 \le (4m^2 + 4mn) \cdot \sum_{j=1}^{D} (\theta_j^\star)^2 \cdot \max_{j\in[D]} |\theta_j^\star|^2 \le \frac{1}{6},$$

which implies that

$$\mathrm{TV}(\mathbb{P}_{0,XYZ}, \mathbb{P}_{1,XYZ}) \le \sqrt{\chi^2(\mathbb{P}_{0,Z|X} \parallel \mathbb{P}_{1,Z|X} \mid \mathbb{P}_{0,X})} \le \sqrt{(1 + 1/(6D))^D - 1} \le \sqrt{e^{1/6} - 1} \le \frac{1}{2}$$

Hence according to eq. (C.28), eq. (4.10) fails. □

C.5.2 Proof of Theorem 4.17

Proof of Theorem 4.17. To verify eq. (4.10), we only need to show that

$$\sup_{P \in H_i} \mathbb{P}(\psi(X, Y, Z) \neq i) \le \frac{1}{4}, \quad \forall i \in \{0, 1\}.$$

Without loss of generality, we only prove the above inequality for $i = 0$, i.e. when $p_{\mathtt{Z}} = p_{\mathtt{X}}$, we always have

$$\mathbb{P}(T_{\mathtt{LF}} \ge 0) \le \frac{1}{4}. \tag{C.35}$$

The proof of cases where $i = 1$ follows similarly.

Without loss of generality, we assume the projection Π_d is onto the first d-coordinates. Assume

$$\hat{\theta}_{\mathtt{X}} = (\hat{\theta}_{\mathtt{X}})_{1:D}, \quad \hat{\theta}_{\mathtt{Y}} = (\hat{\theta}_{\mathtt{Y}})_{1:D}, \quad \hat{\theta}_{\mathtt{Z}} = (\hat{\theta}_{\mathtt{Z}})_{1:D}, \quad p_{\mathtt{X}} = (p_{\mathtt{X}})_{1:D}, \quad p_{\mathtt{Y}} = (p_{\mathtt{Y}})_{1:D} \quad \text{and} \quad p_{\mathtt{Z}} = (p_{\mathtt{Z}})_{1:D}$$

For every $t \in [d]$, we let

$$u_t = \left((\hat{\theta}_{\mathtt{X}})_t - (\hat{\theta}_{\mathtt{Z}})_t\right)^2 - \left((\hat{\theta}_{\mathtt{Y}})_t - (\hat{\theta}_{\mathtt{Z}})_t\right)^2.$$

Then we have

$$(\hat{\theta}_{\mathtt{X}})_t \sim \mathcal{N}\left((p_{\mathtt{X}})_t, \frac{1}{n}\right), \qquad (\hat{\theta}_{\mathtt{Y}})_t \sim \mathcal{N}\left((p_{\mathtt{Y}})_t, \frac{1}{n}\right) \quad \text{and} \quad (\hat{\theta}_{\mathtt{Z}})_t \sim \mathcal{N}\left((p_{\mathtt{Z}})_t, \frac{1}{n}\right).$$

Hence we get

$$\mathbb{E}[u_t] = [(p_{\mathtt{X}})_t]^2 - [(p_{\mathtt{Y}})_t]^2 - 2((p_{\mathtt{X}})_t - (p_{\mathtt{Y}})_t)(p_{\mathtt{Z}})_t$$

and

$$\begin{aligned}\mathrm{Var}(u_t) &= \mathbb{E}[(u_t)^2] - (\mathbb{E}[u_t])^2 \\ &= \frac{4}{n}((p_{\mathtt{X}})_t - (p_{\mathtt{Z}})_t)^2 + \frac{4}{n}((p_{\mathtt{Y}})_t - (p_{\mathtt{Z}})_t)^2 + \frac{4}{m}((p_{\mathtt{X}})_t - (p_{\mathtt{Y}})_t)^2 + \frac{4}{n^2} + \frac{8}{mn}.\end{aligned}$$

Notice that $p_{\mathtt{Z}} = p_{\mathtt{X}}$, we get

$$\mathbb{E}[u_t] = -((p_{\mathtt{X}})_t - (p_{\mathtt{Y}})_t)^2 \quad \text{and} \quad \mathrm{Var}(u_t) = \left(\frac{4}{m} + \frac{4}{n}\right)((p_{\mathtt{X}})_t - (p_{\mathtt{Y}})_t)^2 + \frac{4}{n^2} + \frac{8}{mn}.$$

Next notice that $T_{\mathtt{LF}} = \sum_{t\in[d]} u_t$, we obtain

$$\mathbb{E}[T_{\mathtt{LF}}] = -\sum_{t\in[d]}((p_{\mathtt{X}})_t - (p_{\mathtt{Y}})_t)^2 \quad \text{and} \quad \mathrm{Var}(T_{\mathtt{LF}}) = \left(\frac{4}{m} + \frac{4}{n}\right)\cdot\sum_{t\in[d]}((p_{\mathtt{X}})_t - (p_{\mathtt{Y}})_t)^2 + \frac{4d}{n^2} + \frac{8d}{mn}.$$

According to our choice of d in eq. (4.13), we have

$$\begin{aligned}\sum_{t\in[d]}((p_{\mathtt{X}})_t - (p_{\mathtt{Y}})_t)^2 &= \|p_{\mathtt{X}} - p_{\mathtt{Y}}\|_2^2 - \sum_{t\notin[d]}((p_{\mathtt{X}})_t - (p_{\mathtt{Y}})_t)^2 \\ &\geq \varepsilon^2 - 2\sum_{t\notin[d]}((p_{\mathtt{X}})_t)^2 - 2\sum_{t\notin[d]}((p_{\mathtt{X}})_t)^2 \geq \varepsilon^2 - \frac{2\varepsilon^2}{9} - \frac{2\varepsilon^2}{9} \geq \frac{\varepsilon^2}{2}.\end{aligned}$$

Therefore, if $m \geq 96/\varepsilon^2$, $n \geq 96\sqrt{d}/\varepsilon^2$ and $mn \geq 768d/\varepsilon^4$, we have

$$\mathrm{Var}(T_{\mathtt{LF}}) \leq \frac{1}{4}\cdot(\mathbb{E}[T_{\mathtt{LF}}])^2.$$

Therefore, according to Chebyshev's inequality, we obtain that

$$\Pr(T_{\mathtt{LF}} \geq 0) \leq \frac{\mathrm{Var}(T_{\mathtt{LF}})}{(\mathbb{E}[T_{\mathtt{LF}}])^2} \leq \frac{1}{4},$$

which verifies eq. (C.35).

□

C.6 Analysis of LFHT for ℓ_p Bodies

C.6.1 Proof of Theorem 4.21

The proof of theorem Theorem 4.21 requires the following lemma:

Lemma C.7. *Suppose that $1 \leq p \leq 2$. For $p_{\mathtt{X}}, p_{\mathtt{Y}} \in \Gamma$ with $\|p_{\mathtt{X}} - p_{\mathtt{Y}}\|_2 \geq \varepsilon$, with probability at least $1-\delta$, the set T calculated in item (a) satisfies*

(a) $\sum_{t\in T}\left((p_{\mathtt{X}})_t-(p_{\mathtt{Y}})_t\right)^2\geq\frac{\varepsilon^2}{2}$.

(b) $\operatorname{card}(T)\leq 2d_u(\Gamma,n,\varepsilon)$.

Proof of Theorem C.7. Since $(\hat{\theta}_{\mathtt{X}})_t$, $(\hat{\theta}_{\mathtt{Y}})_t$ are empirical estimation of $(p_{\mathtt{X}})_t$ and $(p_{\mathtt{Y}})_t$, with n samples each:

$$(\mathbf{X}_1)_t,\cdots,(\mathbf{X}_n)_t\overset{\text{i.i.d.}}{\sim}\mathcal{N}((p_{\mathtt{X}})_t,1),\quad\text{and}\quad(\mathbf{Y}_1)_t,\cdots,(\mathbf{Y}_n)_t\overset{\text{i.i.d.}}{\sim}\mathcal{N}((p_{\mathtt{Y}})_t,1),\qquad\forall t\in[D],$$

we have $(\hat{\theta}_{\mathtt{X}})_t\sim\mathcal{N}((p_{\mathtt{X}})_t,1/n)$ and $(\hat{\theta}_{\mathtt{Y}})_t\sim\mathcal{N}((p_{\mathtt{Y}})_t,1/n)$. Hence according to [212, Proposition 2.1.2], we have with probability at least $1-\delta$, for any $t\in[D]$, both of the following events hold:

$$\left|(\hat{\theta}_{\mathtt{X}})_t-(p_{\mathtt{X}})_t\right|\leq\sqrt{\frac{2\log(4D/\delta)}{n}}\quad\text{and}\quad\left|(\hat{\theta}_{\mathtt{Y}})_t-(p_{\mathtt{Y}})_t\right|\leq\sqrt{\frac{2\log(4D/\delta)}{n}}.\tag{C.36}$$

In the rest of the proof, we assume both events in eq. (C.36) holds for any $t\in[D]$, and we will prove the two conditions in Theorem C.7.

We first verify item (a). According to our construction of set T_2 in eq. (4.20) for any $t\in T_2$, we have $t>d_u(\Gamma,n,\varepsilon)$. Hence for any $t\in T_2$, the coefficient a_t in eq. (4.11) satisfies $a_t\leq a_{d_u(\Gamma,n,\varepsilon)+1}$. Since $p_{\mathtt{X}},p_{\mathtt{Y}}\in\Gamma$, we obtain

$$\sum_{t\in T_2}\frac{|(p_{\mathtt{X}})_t|^p}{(a_{d_u(\Gamma,n,\varepsilon)+1})^p}\leq\sum_{t\in T_2}\frac{|(p_{\mathtt{X}})_t|^p}{(a_t)^p}\leq\sum_{t\in[D]}\frac{|(p_{\mathtt{X}})_t|^p}{(a_t)^p}\leq 1,$$

and

$$\sum_{t\in T_2}\frac{|(p_{\mathtt{Y}})_t|^p}{(a_{d_u(\Gamma,n,\varepsilon)+1})^p}\leq\sum_{t\in T_2}\frac{|(p_{\mathtt{Y}})_t|^p}{(a_t)^p}\leq\sum_{t\in[D]}\frac{|(p_{\mathtt{Y}})_t|^p}{(a_t)^p}\leq 1,$$

which implies that

$$\sum_{t\in T_2}|(p_{\mathtt{X}})_t-(p_{\mathtt{Y}})_t|^p\leq 2\sum_{t\in T_2}\left[|(p_{\mathtt{X}})_t|^p+|(p_{\mathtt{Y}})_t|^p\right]\leq 4(a_{d_u(\Gamma,n,\varepsilon)+1})^p,\tag{C.37}$$

where the first inequality uses the fact that $|a-b|^p\leq 2(|a|^p+|b|^p)$ for any $a,b\in\mathbb{R}$ and $1\leq p\leq 2$. Therefore,

$$\begin{aligned}
&\sum_{t\in T_2}\left((p_{\mathtt{X}})_t-(p_{\mathtt{Y}})_t\right)^2\cdot\mathbb{I}\left[|(p_{\mathtt{X}})_t-(p_{\mathtt{Y}})_t|<6\sqrt{\frac{2\log(4D/\delta)}{n}}\right]\\
&\quad\leq\left[\sum_{t\in T_2}|(p_{\mathtt{X}})_t-(p_{\mathtt{Y}})_t|^p\right]\cdot\left(6\sqrt{\frac{2\log(4D/\delta)}{n}}\right)^{2-p}\\
&\quad\leq 4(a_{d_u(\Gamma,n,\varepsilon)+1})^p n^{\frac{p-2}{2}}\cdot\sqrt{72\log(4D/\delta)}^{2-p}\\
&\quad\leq 4(a_{d_u(\Gamma,n,\varepsilon)+1})^p n^{\frac{p-2}{2}}\cdot 72\log(4D/\delta)
\end{aligned}$$

According to the definition of $d_u(\Gamma,n,\varepsilon)$ in eq. (4.17), we have

$$(a_{d_u(\Gamma,n,\varepsilon)+1})^p n^{\frac{p-2}{2}}<\frac{\varepsilon^2}{576\log(4D/\delta)},\tag{C.38}$$

which implies that

$$\sum_{t\in T_2}((p_{\mathtt{X}})_t-(p_{\mathtt{Y}})_t)^2\cdot\mathbb{I}\left[|(p_{\mathtt{X}})_t-(p_{\mathtt{Y}})_t|<6\sqrt{\frac{2\log(4D/\delta)}{n}}\right]\le\frac{\varepsilon^2}{2}.$$

According to the construction of set T, we have for any $t\in[D]\backslash T$,

$$|(\hat{\theta}_{\mathtt{X}})_t-(\hat{\theta}_{\mathtt{Y}})_t|\le 4\sqrt{\frac{2\log(4D/\delta)}{n}}.$$

Hence the conditions in eq. (C.36) indicates that

$$|(p_{\mathtt{X}})_t-(p_{\mathtt{Y}})_t|\le|(\hat{\theta}_{\mathtt{X}})_t-(\hat{\theta}_{\mathtt{Y}})_t|+2\sqrt{\frac{2\log(4D/\delta)}{n}}\le 6\sqrt{\frac{2\log(4D/\delta)}{n}},$$

which implies that

$$\begin{aligned}
&\sum_{t\in T}((p_{\mathtt{X}})_t-(p_{\mathtt{Y}})_t)^2\\
&\quad\ge\sum_{t=1}^{D}((p_{\mathtt{X}})_t-(p_{\mathtt{Y}})_t)^2-\sum_{t\in T_2}((p_{\mathtt{X}})_t-(p_{\mathtt{Y}})_t)^2\cdot\mathbb{I}\left[|(p_{\mathtt{X}})_t-(p_{\mathtt{Y}})_t|<6\sqrt{\frac{2\log(4D/\delta)}{n}}\right]\\
&\quad\ge\frac{\varepsilon^2}{2}.
\end{aligned}$$

We next verify item (b). Since $\mathrm{card}(T_1)=d_u(\Gamma,n,\varepsilon)$ and $T=T_1\cup T_3$ according to its definition, we only need to verify $\mathrm{card}(T_3)\le d_u(\Gamma,n,\varepsilon)$. We further notice that the conditions in eq. (C.36) gives that for any $t\in T_3$,

$$|(p_{\mathtt{X}})_t-(p_{\mathtt{Y}})_t|\ge|(\hat{\theta}_{\mathtt{X}})_t-(\hat{\theta}_{\mathtt{Y}})_t|-2\sqrt{\frac{2\log(4D/\delta)}{n}}\ge 2\sqrt{\frac{2\log(4D/\delta)}{n}}.$$

Next, since $T_3\subseteq T_2$, eq. (C.37) indicates that

$$\sum_{t\in T_3}|(p_{\mathtt{X}})_t-(p_{\mathtt{Y}})_t|^p\le\sum_{t\in T_2}|(p_{\mathtt{X}})_t-(p_{\mathtt{Y}})_t|^p\le 4(a_{d_u(\Gamma,n,\varepsilon)+1})^p,$$

which implies that

$$\mathrm{card}(T_3)\le\frac{4(a_{d_u(\Gamma,n,\varepsilon)+1})^p}{\left(2\sqrt{\frac{2\log(4D/\delta)}{n}}\right)^p}\le 2(a_{d_u(\Gamma,n,\varepsilon)+1})^p n^{p/2}\overset{(i)}{\le}\frac{2n\varepsilon^2}{576\log(4D/\delta)^2}\le n\varepsilon^2,$$

where inequality (i) uses eq. (C.38). Notice from eq. (4.19) and also the condition that

$$mn\le\frac{1024d_u(\Gamma,n,\varepsilon)}{\varepsilon^4},$$

we have

$$n\varepsilon^2 \le d_u(\Gamma, n, \varepsilon),$$

which implies that

$$\operatorname{card}(T) = \operatorname{card}(T_1) + \operatorname{card}(T_3) \le d_u(\Gamma, n, \varepsilon) + d_u(\Gamma, n, \varepsilon) \le 2d_u(\Gamma, n, \varepsilon).$$

Hence item (b) is verified. □

Now equipped with Theorem C.7, we are ready to prove Theorem 4.21.

Proof of Theorem 4.21. To verify eq. (4.10), we only need to show that

$$\sup_{P \in H_i} \mathbb{P}(\psi(X, Y, Z) \neq i) \le \frac{1}{4}, \quad \forall i \in \{0, 1\}.$$

Without loss of generality, we only prove the above inequality for $i = 0$, i.e. when $p_{\mathtt{Z}} = p_{\mathtt{X}}$, we always have

$$\mathbb{P}(T_{\mathtt{LF}} \ge 0) \le \frac{1}{4}.$$

The proof of cases where $i = 1$ follows similarly.

For every $t \in [D]$, we let

$$u_t = \left((\hat{\theta}^2_{\mathtt{X}})_t - (\hat{\theta}_{\mathtt{Z}})_t\right)^2 - \left((\hat{\theta}^2_{\mathtt{Y}})_t - (\hat{\theta}_{\mathtt{Z}})_t\right)^2.$$

And for simplicity, we set $N = n - n_0 = n - \lfloor n/2 \rfloor$. Then we have

$$(\hat{\theta}^2_{\mathtt{X}})_t \sim \mathcal{N}\left((p_{\mathtt{X}})_t, \frac{1}{N}\right), \quad (\hat{\theta}^2_{\mathtt{Y}})_t \sim \mathcal{N}\left((p_{\mathtt{Y}})_t, \frac{1}{N}\right) \quad \text{and} \quad (\hat{\theta}_{\mathtt{Z}})_t \sim \mathcal{N}\left((p_{\mathtt{Z}})_t, \frac{1}{m}\right).$$

Therefore, since $(\mathbf{X}^2, \mathbf{Y}^2, \mathbf{Z}) \perp\!\!\!\perp (\mathbf{X}^1, \mathbf{Y}^1)$, we can calculate

$$\mathbb{E}[u_t \mid \mathbf{X}^1, \mathbf{Y}^1] = [(p_{\mathtt{X}})_t]^2 - [(p_{\mathtt{Y}})_t]^2 - 2((p_{\mathtt{X}})_t - (p_{\mathtt{Y}})_t)(p_{\mathtt{Z}})_t,$$

and further

$$\begin{aligned}\operatorname{Var}(u_t \mid \mathbf{X}^1, \mathbf{Y}^1) &= \mathbb{E}[u_t^2 \mid \mathbf{X}^1, \mathbf{Y}^1] - (\mathbb{E}[u_t \mid \mathbf{X}^1, \mathbf{Y}^1])^2 \\ &= \frac{4}{N}((p_{\mathtt{X}})_t - (p_{\mathtt{Z}})_t)^2 + \frac{4}{N}((p_{\mathtt{Y}})_t - (p_{\mathtt{Z}})_t)^2 + \frac{4}{m}((p_{\mathtt{X}})_t - (p_{\mathtt{Y}})_t)^2 + \frac{4}{N^2} + \frac{8}{mN}.\end{aligned}$$

When $p_{\mathtt{Z}} = p_{\mathtt{X}}$, we have

$$\begin{aligned}\mathbb{E}[u_t \mid \mathbf{X}^1, \mathbf{Y}^1] &= -((p_{\mathtt{X}})_t - (p_{\mathtt{Y}})_t)^2 \\ \operatorname{Var}(u_t \mid \mathbf{X}^1, \mathbf{Y}^1) &= \left(\frac{4}{m} + \frac{4}{N}\right) \cdot ((p_{\mathtt{X}})_t - (p_{\mathtt{Y}})_t)^2 + \frac{4}{N^2} + \frac{8}{mN}.\end{aligned}$$

According to our construction of set $T \subseteq [D]$, set T is deterministic when conditioned on samples $\mathbf{X}^1, \mathbf{Y}^1$. Therefore, we obtain that

$$\mathbb{E}[T_{\mathtt{LF}} \mid \mathbf{X}^1, \mathbf{Y}^1] = -\sum_{t \in T}((p_{\mathtt{X}})_t - (p_{\mathtt{Y}})_t)^2$$

$$\mathrm{Var}(T_{\mathtt{LF}} \mid \mathbf{X}^1, \mathbf{Y}^1) = \left(\frac{4}{m} + \frac{4}{N}\right) \cdot \left(\sum_{t \in T} ((p_{\mathtt{X}})_t - (p_{\mathtt{Y}})_t)^2\right) + \left(\frac{4}{N^2} + \frac{8}{mN}\right) \cdot |T|.$$

When the conditions item (a) and item (b) in Theorem C.7 holds, we have

$$\sum_{t \in T} ((p_{\mathtt{X}})_t - (p_{\mathtt{Y}})_t)^2 \geq \frac{\varepsilon^2}{2} \quad \text{and} \quad |T| \leq 2d_u(\Gamma, n, \varepsilon).$$

Next we notice that $N = n - \lfloor n/2 \rfloor \geq n/2$. Hence, when (m, n) satisfies

$$m \geq \frac{32}{\varepsilon^2}, \quad n \geq \frac{32\sqrt{d_u(\Gamma, n, \varepsilon)}}{\varepsilon^2} \quad \text{and} \quad mn \geq \frac{1024 d_u(\Gamma, n, \varepsilon)}{\varepsilon^2},$$

which implies that

$$\mathrm{Var}(T_{\mathtt{LF}} \mid \mathbf{X}^1, \mathbf{Y}^1) \leq \frac{7(\mathbb{E}[T_{\mathtt{LF}} \mid \mathbf{X}^1, \mathbf{Y}^1])^2}{32}.$$

Therefore, according to Chebyshev's inequality, we obtain that

$$\Pr(T_{\mathtt{LF}} \geq 0 \mid \mathbf{X}^1, \mathbf{Y}^1) \leq \frac{\mathrm{Var}(T_{\mathtt{LF}} \mid \mathbf{X}^1, \mathbf{Y}^1)}{(\mathbb{E}[T_{\mathtt{LF}} \mid \mathbf{X}^1, \mathbf{Y}^1])^2} \leq \frac{7}{32}.$$

Finally, we notice that according to Theorem C.7 with $\delta = 1/32$, with probability at least $31/32$, conditions item (a) and item (b) both holds, which implies that

$$\Pr(T_{\mathtt{LF}} \geq 0) \leq \Pr(T_{\mathtt{LF}} \geq 0 \mid \mathbf{X}^1, \mathbf{Y}^1) + \frac{1}{32} \leq \frac{7}{32} + \frac{1}{32} = \frac{1}{4}.$$

□

C.6.2 Proof of Theorem 4.22

We present the proof of Theorem 4.22 in this section.

Proof of Theorem 4.22. According to [65, Proposition 3], the condition of goodness-of-fit test (the condition where eq. (4.7) holds) is

$$n \geq n_{\mathrm{gof}} \triangleq \frac{\sqrt{d_l(\Gamma, n, \varepsilon)}}{2\varepsilon^2}.$$

According to [15, Proposition 1], if there exists a test which satisfies eq. (4.10), then (m, n) has to satisfies

$$m \geq \frac{1}{\varepsilon^2}, \quad \text{and} \quad n \geq n_{\mathrm{gof}} = \frac{\sqrt{d_l(\Gamma, n, \varepsilon)}}{2\varepsilon^2}.$$

Therefore, we only need to verify the third condition, i.e.

$$mn \gtrsim \frac{d_l(\Gamma, n, \varepsilon)}{\varepsilon^4}.$$

If $m \geq n$, since n satisfies $n \geq \sqrt{d(\Gamma, n, \varepsilon)}/(2\varepsilon^2)$, we have

$$mn \geq n^2 \geq \frac{d_l(\Gamma, n, \varepsilon)}{4\varepsilon^2}$$

and the third condition of eq. (4.24) is automatically satisfied. In the following, we assume $m < n$, and we will verify the third condition of eq. (4.24). Above all, in the following we assume

$$n > m \geq \frac{1}{\varepsilon^2}, \quad \text{and} \quad n \geq \frac{\sqrt{d_l(\Gamma, n, \varepsilon)}}{2\varepsilon^2}, \tag{C.39}$$

and we will show that if

$$mn < \frac{d_l(\Gamma, n, \varepsilon)}{96\varepsilon^4}, \tag{C.40}$$

then eq. (4.10) fails.

In the following, we assume eq. (C.40) holds, and we let $d = d_l(\Gamma, n, \varepsilon)$. For any fixed distribution $\mu \in \Delta(\mathbb{R}^D)$, we define distribution $\mathbb{P}_{0,XYZ}$ to be the distribution of $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ sampled according to the following way: first sample $\boldsymbol{\theta} \sim \mu$, then sample $\mathbf{X} = (\mathbf{X}^{1:n}) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{\theta}, I_D)$, $\mathbf{Y} = (\mathbf{Y}^{1:n}) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, I_D)$ and $\mathbf{Z} = (\mathbf{Z}^{1:m}) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{\theta}, I_D)$. And we define distribution $\mathbb{P}_{1,XYZ}$ to be the distribution of $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ sampled according to the following way: first sample $\boldsymbol{\theta} \sim \mu$, then sample $\mathbf{X} = (\mathbf{X}^{1:n}) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{\theta}, I_D)$, $\mathbf{Y} = (\mathbf{Y}^{1:n}) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, I_D)$ and $\mathbf{Z} = (\mathbf{Z}^{1:m}) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, I_D)$. Similarly, we can define distribution $\mathbb{P}_{0,XZ}, \mathbb{P}_{1,XZ}, \mathbb{P}_{0,X}, \mathbb{P}_{1,X}$, and also conditional distribution $\mathbb{P}_{0,Z|X}$ and $\mathbb{P}_{1,Z|X}$. Then [15, Lemma 5] gives that for any such μ,

$$\inf_{\psi} \max_{i \in \{0,1\}} \sup_{P \in H_i} \mathbb{P}(\psi(X, Y, Z) \neq i) \geq \frac{1}{2}\left(1 - \mathrm{TV}(\mathbb{P}_{0,XYZ}, \mathbb{P}_{1,XYZ})\right) - \mu(\Gamma^c) - \mu(B_2(\varepsilon)), \tag{C.41}$$

where Γ^c denotes the complement of set Γ, i.e. $\Gamma^c = \{\Gamma \in \mathbb{R}^D \mid \theta \notin \Gamma\}$, and $B_2(\varepsilon)$ denotes the ℓ_2-ball of radius ε, i.e. $B_2(\varepsilon) = \{\theta \in \mathbb{R}^D : \|\theta\|_2 \leq \varepsilon\}$. In the following proof, we choose distribution μ to be the following product of symmetric ternary distributions, i.e. for $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_D) \in \Gamma$,

$$\mu(\boldsymbol{\theta}) = \prod_{j=1}^{D} \mu_j(\theta_j) \quad \text{where } \mu_j = \begin{cases} (1-h) \cdot \delta_0 + \frac{h}{2} \cdot \delta_r + \frac{h}{2} \cdot \delta_{-r} & \text{if } 1 \leq j \leq d, \\ \delta_0 & \text{if } d+1 \leq j \leq D, \end{cases}$$

where δ_r denotes the point distribution at $r \in \mathbb{R}$, and parameters $d \in [D]$, $h \in [0, 1]$ and $r > 0$ will be specified later. Then we have

$$\begin{aligned} \mathrm{TV}(\mathbb{P}_{0,XYZ}, \mathbb{P}_{1,XYZ})^2 &= \mathrm{TV}(\mathbb{P}_{0,XZ}, \mathbb{P}_{1,XZ})^2 \leq D_{\mathrm{KL}}(\mathbb{P}_{0,XZ} \parallel \mathbb{P}_{1,XZ}) \\ &= D_{\mathrm{KL}}(\mathbb{P}_{0,Z|X} \parallel \mathbb{P}_{1,Z|X} \mid \mathbb{P}_{0,X}) + D_{\mathrm{KL}}(\mathbb{P}_{0,X} \parallel \mathbb{P}_{1,X}) \\ &= D_{\mathrm{KL}}(\mathbb{P}_{0,Z|X} \parallel \mathbb{P}_{1,Z|X} \mid \mathbb{P}_{0,X}) \leq \chi^2(\mathbb{P}_{0,Z|X} \parallel \mathbb{P}_{1,Z|X} \mid \mathbb{P}_{0,X}). \end{aligned} \tag{C.42}$$

If we use $\varphi_{\boldsymbol{\theta}}(\cdot)$ to denote the density function of $\mathcal{N}(\boldsymbol{\theta}, I_D)$, according to Ingster's trick [60] we have

$\chi^2(\mathbb{P}_{0,Z|X} \parallel \mathbb{P}_{1,Z|X} \mid \mathbb{P}_{0,X}) + 1$

$$
\begin{aligned}
&= \mathbb{E}_{\mathbf{X}\sim\mathbb{P}_{0,X}}\left[\mathbb{E}_{\boldsymbol{\theta}|\mathbf{X},\boldsymbol{\theta}'|\mathbf{X}}\left[\int_{(\mathbb{R}^D)^m}\prod_{t=1}^{m}\frac{\varphi_{\boldsymbol{\theta}}(\mathbf{z}^t)\varphi_{\boldsymbol{\theta}'}(\mathbf{z}^t)}{\varphi_{\mathbf{0}}(\mathbf{z}^t)}d\mathbf{z}_1\cdots d\mathbf{z}_m \mid \mathbf{X}\right]\right] \\
&= \mathbb{E}_{\mathbf{X}\sim\mathbb{P}_{0,X}}\left[\mathbb{E}_{\boldsymbol{\theta}|\mathbf{X},\boldsymbol{\theta}'|\mathbf{X}}[\exp\left(m\langle\boldsymbol{\theta},\boldsymbol{\theta}'\rangle\right) \mid \mathbf{X}]\right] \\
&= \mathbb{E}_{\mathbf{X}\sim\mathbb{P}_{0,X}}\left[\prod_{j=1}^{d}\left(\mathbb{P}(\theta_j=\theta'_j\neq 0 \mid \mathbf{X})(\exp(mr^2)-1)+\mathbb{P}(\theta_j=-\theta'_j\neq 0 \mid X)(\exp(-mr^2)-1)+1\right)\right] \\
&= \prod_{j=1}^{d}\mathbb{E}_{\mathbf{X}\sim\mathbb{P}_{0,X}}\left[\mathbb{P}(\theta_j=\theta'_j\neq 0 \mid \mathbf{X})(\exp(mr^2)-1)+\mathbb{P}(\theta_j=-\theta'_j\neq 0 \mid X)(\exp(-mr^2)-1)+1\right],
\end{aligned}
\tag{C.43}
$$

where $\boldsymbol{\theta},\boldsymbol{\theta}'$ are i.i.d. sampled according to $\mathbb{P}(\boldsymbol{\theta} \mid \mathbf{X})$, and the last equation uses the fact that conditioned on $\mathbf{X}$, we have (θ_j,θ'_j) independent to each other for any $j\in[D]$. In the following, we write $\mathbf{X}=\mathbf{X}^{1:n}$ which denotes the n samples, and we further denote $\mathbf{X}^i=(X_1^i,\cdots,X_D^i)$, where X_j^i denotes the j-th coordinate of $\mathbf{X}^i$. We notice that θ_j only depends on $X_j^{1:n}=(X_j^1,\cdots,X_j^n)$. According to Bayes rule, we can calculate

$$
\begin{aligned}
&\mathbb{P}(\theta_j=1 \mid \mathbf{X}) \\
&= \frac{\Pr(X_j^{1:n},\eta=1)}{\Pr(X_j^{1:n},\eta_j=0)+\Pr(X_j^{1:n},\eta_j=-1)+\Pr(X_j^{1:n},\eta_j=1)} \\
&= \frac{h/2\cdot\prod_{i=1}^n\exp\left(-(X_j^i-r)^2/2\right)}{(1-h)\cdot\prod_{i=1}^n\exp\left(-(X_j^i)^2/2\right)+h/2\cdot\prod_{i=1}^n\exp\left(-(X_j^i-r)^2/2\right)+h/2\cdot\prod_{i=1}^n\exp\left(-(X_j^i+r)^2/2\right)} \\
&= \frac{h/2\cdot\exp\left(r\cdot\sum_{i=1}^n X_j^i\right)}{(1-h)\cdot\exp\left(r^2/2\right)+h/2\cdot\exp\left(r\cdot\sum_{i=1}^n X_j^i\right)+h/2\cdot\exp\left(-r\cdot\sum_{i=1}^n X_j^i\right)}.
\end{aligned}
\tag{C.44}
$$

Similarly, we get

$$
\mathbb{P}(\theta_j=-1 \mid \mathbf{X})=\frac{h/2\cdot\exp\left(-r\cdot\sum_{i=1}^n X_j^i\right)}{(1-h)\cdot\exp\left(r^2/2\right)+h/2\cdot\exp\left(r\cdot\sum_{i=1}^n X_j^i\right)+h/2\cdot\exp\left(-r\cdot\sum_{i=1}^n X_j^i\right)}.
\tag{C.45}
$$

In the following, we use $[0,1]$-valued random variables $p_j(\mathbf{X})$ and $q_j(\mathbf{X})$ to denote

$$
p_j(\mathbf{X})=\mathbb{P}(\theta_j=1 \mid \mathbf{X}) \quad \text{and} \quad q_j(\mathbf{X})=\mathbb{P}(\theta_j=-1 \mid \mathbf{X}), \qquad \forall j\in[D].
$$

We further notice that θ'_j and θ_j are i.i.d. conditioned on $\mathbf{X}$. Hence we obtain

$$
\begin{aligned}
&\mathbb{P}(\theta_j=\theta'_j\neq 0 \mid X)\left(\exp(mr^2)-1\right)+\mathbb{P}(\theta_j=-\theta'_j\neq 0 \mid X)\left(\exp(-mr^2)-1\right)+1 \\
&= 1+\left(p_j(\mathbf{X})^2+q_j(\mathbf{X})^2\right)\left(\exp(mr^2)-1\right)+2p_j(\mathbf{X})q_j(\mathbf{X})\left(\exp(-mr^2)-1\right) \\
&= 1+\frac{\left(p_j(\mathbf{X})+q_j(\mathbf{X})\right)^2}{2}\cdot\left(\exp(mr^2)+\exp(-mr^2)-2\right)+\frac{\left(p_j(\mathbf{X})-q_j(\mathbf{X})\right)^2}{2}\cdot\left(\exp(mr^2)-\exp(-mr^2)\right).
\end{aligned}
\tag{C.46}
$$

Next using the AM-GM inequality we obtain that

$$
(1-h)\cdot\exp\left(\frac{r^2}{2}\right)+\frac{h}{2}\cdot\exp\left(r\cdot\sum_{i=1}^n X_j^i\right)+\frac{h}{2}\cdot\exp\left(-r\cdot\sum_{i=1}^n X_j^i\right)\geq 1.
$$

Bringing this back to eq. (C.44) and eq. (C.45), we obtain that

$$\begin{aligned}
&(p_j(\mathbf{X}) + q_j(\mathbf{X}))^2 \le \frac{h^2}{4} \cdot \left(\exp\left(r \cdot \sum_{i=1}^{n} X_j^i \right) + \exp\left(-r \cdot \sum_{i=1}^{n} X_j^i \right) \right)^2 \\
\text{and} \quad &(p_j(\mathbf{X}) - q_j(\mathbf{X}))^2 \le \frac{h^2}{4} \cdot \left(\exp\left(r \cdot \sum_{i=1}^{n} X_j^i \right) - \exp\left(-r \cdot \sum_{i=1}^{n} X_j^i \right) \right)^2 .
\end{aligned} \tag{C.47}$$

We notice that according to the method of collecting samples,

$$X_j^i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta_j, 1) \quad \text{and} \quad \theta_j \sim (1-h) \cdot \delta_0 + \frac{\eta}{2} \cdot \delta_r + \frac{\eta}{2} \cdot \delta_{-r},$$

which implies

$$\sum_{i=1}^{n} X_j^i \sim (1-h) \cdot \mathcal{N}(0, n) + \frac{h}{2} \cdot \mathcal{N}(nr, n) + \frac{h}{2} \cdot \mathcal{N}(-nr, n).$$

Next, we notice that for Gaussian random variable $X \sim \mathcal{N}(\mu, \sigma)$, we have

$$\mathbb{E}[\exp(\alpha X)] = \exp\left(\alpha\mu + \frac{\alpha^2\sigma^2}{2} \right).$$

Bringing this back to eq. (C.47) we obtain that

$$\begin{aligned}
\mathbb{E}\left[(p_j(\mathbf{X}) + q_j(\mathbf{X}))^2 \right] &\le \frac{h^2}{4} \cdot \mathbb{E}\left[\exp\left(2r \cdot \sum_{i=1}^{n} X_j^i \right) + 2 + \exp\left(-2r \cdot \sum_{i=1}^{n} X_j^i \right) \right] \\
&= \frac{h^2}{2} \exp\left(2r^2 n\right) \cdot \left(1 - h + \frac{h}{2} \exp\left(2r^2 n\right) + \frac{h}{2} \exp\left(-2r^2 n\right) \right) + \frac{h^2}{2} \\
\mathbb{E}\left[(p_j(\mathbf{X}) - q_j(\mathbf{X}))^2 \right] &\le \frac{h^2}{4} \cdot \mathbb{E}\left[\exp\left(2r \cdot \sum_{i=1}^{n} X_j^i \right) - 2 + \exp\left(-2r \cdot \sum_{i=1}^{n} X_j^i \right) \right] \\
&= \frac{h^2}{2} \exp\left(2r^2 n\right) \cdot \left(1 - h + \frac{h}{2} \exp\left(2r^2 n\right) + \frac{h}{2} \exp\left(-2r^2 n\right) \right) - \frac{h^2}{2}.
\end{aligned}$$

Hence if conditions

$$mr^2 \le 1 \quad \text{and} \quad nr^2 \le 1 \tag{C.48}$$

both hold, we have the following inequalities:

$$\begin{aligned}
&\mathbb{E}\left[(p_j(\mathbf{X}) + q_j(\mathbf{X}))^2 \right] \le 30h^2, \qquad \mathbb{E}\left[(p_j(\mathbf{X}) - q_j(\mathbf{X}))^2 \right] \le 30h^2 r^2 n, \\
\text{and} \quad &\exp(mr^2) + \exp(-mr^2) - 2 \le 2m^2 r^4, \qquad \exp(mr^2) - \exp(-mr^2) \le 3mr^2.
\end{aligned}$$

Bringing them back to eq. (C.46), we obtain that

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{P}(\theta_j = \theta_j' \neq 0 \mid X) \left(\exp(mr^2) - 1 \right) + \mathbb{P}(\theta_j = -\theta_j' \neq 0 \mid X) \left(\exp(-mr^2) - 1 \right) + 1 \right] \\
&\quad \le 1 + 30h^2 m^2 r^4 + 45h^2 mnr^4 \le 1 + 75h^2 mnr^4,
\end{aligned}$$

where the last inequality uses the assumption $m < n$. Bring back to eq. (C.43), we obtain that

$$\begin{aligned}
&\chi^2(\mathbb{P}_{0,Z|X} \parallel \mathbb{P}_{1,Z|X} \mid \mathbb{P}_{0,X})] \\
&\le \prod_{j=1}^{d} \mathbb{E}\left[\mathbb{P}(\theta_j = \theta'_j \neq 0 \mid X)\left(\exp(mr^2) - 1\right) + \mathbb{P}(\theta_j = -\theta'_j \neq 0 \mid X)\left(\exp(-mr^2) - 1\right) + 1\right] - 1 \\
&\le (1 + 75h^2 mnr^4)^d - 1.
\end{aligned}$$

Hence according to eq. (C.42), this implies that

$$\mathrm{TV}(\mathbb{P}_{0,XYZ}, \mathbb{P}_{1,XYZ})^2 \le \sqrt{(1 + 75h^2 mnr^4)^d - 1}. \tag{C.49}$$

Finally, we calculate the probability $\mu(\Gamma^c)$ and $\mu(B_2(\varepsilon))$. According to our choice $d = d(\Gamma, n, \varepsilon)$, when sampling $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_D) \sim \mu$, with probability at least $1 - \delta$ we have

$$\sum_{t=1}^{D} \frac{|\theta_t|^p}{a_t^p} = \sum_{t=1}^{d} \frac{|\theta_t|^p}{a_d^p} \le \frac{dhr^p}{a_d^p} + \frac{r^p \cdot \sqrt{2d\log(2/\delta)}}{a_d^p},$$

where the last inequality uses Hoeffding inequality. Additionally, when sampling $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_D) \sim \mu$, with probability at least $1 - \delta$ we have

$$\sum_{t=1}^{D} |\theta_t|^2 = \sum_{t=1}^{d} |\theta_t|^2 \ge dhr^2 - r^2 \cdot \sqrt{2d\log(2/\delta)}.$$

As long as $h^2 d \ge 24$, with probability at least $9/10$ we have both

$$\sum_{t=1}^{D} \frac{|\theta_t|^p}{a_t^p} \le \frac{2dhr^p}{a_d^p} \quad \text{and} \quad \sum_{t=1}^{D} |\theta_t|^2 \ge \frac{1}{2} \cdot dhr^2.$$

We choose $d = d_l(\Gamma, n, \varepsilon)$ (here $d_l(\Gamma, n, \varepsilon)$ is defined in eq. (4.23)), then we have

$$(a_d)^p n^{\frac{p-2}{2}} \ge 192\varepsilon^2.$$

We further let

$$h = \frac{96\varepsilon^2 n}{d}, \quad \text{and} \quad r = \frac{1}{\sqrt{n}}.$$

According to the first inequality in eq. (C.39), and also eq. (C.40), we have $n \le d/(96\varepsilon^2)$. This implies that $h \le 1$. Additionally, according to eq. (C.39) we have

$$h^2 = \frac{96\varepsilon^4 n^2}{d^2} \ge \frac{96\varepsilon^4}{d^2} \cdot \frac{d}{4\varepsilon^4} \ge \frac{24}{d}.$$

Hence with probability at least $9/10$ we have both

$$\sum_{t=1}^{D} \frac{|\theta_t|^p}{a_t^p} \le \frac{2dhr^p}{a_d^p} \le 192\varepsilon^2 \cdot \frac{n^{\frac{2-p}{2}}}{a_d^p} \le 1, \quad \text{and} \quad \sum_{t=1}^{D} |\theta_t|^2 \ge \frac{1}{2} dhr^2 \ge \varepsilon^2,$$

which implies that

$$\mu(\Gamma^c) + \mu(B_2(\varepsilon)) \le \frac{1}{10}.$$

Additionally, by our choice of r and also eq. (C.39), eq. (C.48) always holds. Hence by eq. (C.49),

$$\mathrm{TV}(\mathbb{P}_{0,XYZ}, \mathbb{P}_{1,XYZ})^2 \le \sqrt{(1+75h^2mnr^4)^d - 1} \le \frac{1}{10}.$$

Therefore, according to eq. (C.41), we obtain that

$$\inf_{\psi} \max_{i\in\{0,1\}} \sup_{P\in H_i} \mathbb{P}(\psi(X,Y,Z) \ne i) > \frac{1}{4},$$

which implies that eq. (4.10) fails.

Above all, we have verify that we must have

$$mn \ge \frac{d_l(\Gamma, n, \varepsilon)}{96\varepsilon^4}$$

in order to let eq. (4.10) satisfied. □

C.6.3 Proof of Theorem 4.23

We present the proof of Theorem 4.23 in this section.

Proof of Theorem 4.23. Suppose $p_{\mathtt{X}} = \mathcal{N}(\boldsymbol{\theta}^x, I)$, $p_{\mathtt{Y}} = \mathcal{N}(\boldsymbol{\theta}^y, I)$ and $p_{\mathtt{Z}} = \mathcal{N}(\boldsymbol{\theta}^z, I)$, and we let

$$\boldsymbol{\theta}^x = (\theta^x_{1:\infty}), \quad \boldsymbol{\theta}^y = (\theta^y_{1:\infty}), \quad \text{and} \quad \boldsymbol{\theta}^z = (\theta^z_{1:\infty}).$$

We let $D = D(\Gamma, \varepsilon)$, and we define $\widetilde{\boldsymbol{\theta}}^x, \widetilde{\boldsymbol{\theta}}^y, \widetilde{\boldsymbol{\theta}}^z \in \mathbb{R}^D$ as

$$\widetilde{\boldsymbol{\theta}}^x = (\theta^x_1, \cdots, \theta^x_D), \quad \widetilde{\boldsymbol{\theta}}^y = (\theta^y_1, \cdots, \theta^y_D), \quad \text{and} \quad \widetilde{\boldsymbol{\theta}}^z = (\theta^z_1, \cdots, \theta^z_D),$$

and we further define D-dimensional distributions

$$\widetilde{p}_{\mathtt{X}} = \mathcal{N}(\widetilde{\boldsymbol{\theta}}^x, I_D), \quad \widetilde{p}_{\mathtt{Y}} = \mathcal{N}(\widetilde{\boldsymbol{\theta}}^y, I_D), \quad \text{and} \quad \widetilde{p}_{\mathtt{Z}} = \mathcal{N}(\widetilde{\boldsymbol{\theta}}^z, I_D).$$

According to the definition of D, we have for any $\boldsymbol{\theta} = (\theta_{1:\infty}) \in \Theta$,

$$\sum_{t=D+1}^{\infty} (\theta_t)^2 \le (a_D)^2 \cdot \sum_{t=1}^{D} \frac{|\theta_t|^2}{(a_t)^2} \overset{(i)}{\le} (a_D)^2 \cdot \sum_{t=1}^{D} \frac{|\theta_t|^p}{(a_t)^p} \le (a_D)^2 \overset{(ii)}{\le} \frac{\varepsilon^2}{9},$$

where (i) uses the fact that $|\theta_t|/a_t \le 1$ for any t, and (ii) uses eq. (4.26). Therefore, since

$$\|\boldsymbol{\theta}_x - \boldsymbol{\theta}_y\|_2 \ge \varepsilon,$$

we have

$$\|\widetilde{\boldsymbol{\theta}}_x - \widetilde{\boldsymbol{\theta}}_y\|_2 \ge \|\boldsymbol{\theta}_x - \boldsymbol{\theta}_y\|_2 - \|\boldsymbol{\theta}_x - \widetilde{\boldsymbol{\theta}}_x\|_2 - \|\boldsymbol{\theta}_y - \widetilde{\boldsymbol{\theta}}_y\|_2 \ge \varepsilon - \frac{\varepsilon}{3} - \frac{\varepsilon}{3} = \frac{\varepsilon}{3}.$$

Hence via taking the testing scheme in section 4.4.2 for the first D coordinates, and also replacing ε with $\varepsilon/3$, we have that as long as (m, n) satisfies

$$\left\{(m,n): \qquad m \gtrsim \frac{1}{\varepsilon^2}, \quad n \gtrsim \frac{\sqrt{d_u(\Gamma,n,\varepsilon)}}{\varepsilon^2} \quad \text{and} \quad mn \gtrsim \frac{d_u(\Gamma,n,\varepsilon)}{\varepsilon^4}\right\},$$

then ψ is a feasible testing scheme which satisfies eq. (4.10). □

Proof of Theorem 4.24. We set $d = d_l(\Gamma, n, \varepsilon)$. We consider the following subset of Γ:

$$\Gamma_d = \left\{\boldsymbol{\theta} = (\theta_{1:\infty}) : \sum_{t=1}^{d} \frac{|\theta_t|^p}{(a_t)^p} \le 1, \quad \text{and} \quad \theta_t = 0, \quad \forall t \ge d+1\right\} \subseteq \Gamma.$$

Then if we can do likelihood-free hypothesis testing with (m, n) samples for Γ, then we can also do likelihood-free hypothesis testing with (m, n) samples for Γ_d. Notice that to do likelihood-free hypothesis testing, the data from coordinates greater than d are completely independent to the $p_{\mathtt{X}}, p_{\mathtt{Y}}$ and $p_{\mathtt{Z}}$. Hence without loss of generality we can assume that the model consists of dimension-d distributions. Therefore, according to Theorem 4.22, we get the desired result. □

C.6.4 Missing Proofs in section 4.4.2

Proof of Theorem 4.25. We notice that the Γ defined in eq. (4.4) is an infinite dimensional ℓ_1 body with $a_t = 1/t$. Therefore, we can calculate that

$$D(\Gamma, \varepsilon) = \frac{1}{\varepsilon}, \quad \text{and} \quad d_l(\Gamma, n, \varepsilon) \asymp d_u(\Gamma, n, \varepsilon) \asymp \frac{1}{\sqrt{n}\varepsilon^2},$$

where in the above equations, we use $\asymp$ to hide constants and log factors of n and ε as well. Therefore, according to Theorem 4.24 and Theorem 4.23, we obtain the feasible region of likelihood-free hypothesis testing:

$$\left\{(m,n): \quad m \gtrsim \varepsilon^{-2}, \quad n \gtrsim \varepsilon^{-12/5}, \quad mn^{3/2} \gtrsim \varepsilon^{-6}\right\}.$$

□

Appendix D

Proofs for Chapter 5

D.1 Finite Class Lemmas

We first provide a version of [109, Lemma 10].

Lemma D.1. *Suppose $\varepsilon_{1:n}$ are n i.i.d. Rademacher random variables, i.e. $\varepsilon_{1:n} \overset{i.i.d.}{\sim} \mathrm{Unif}(\{-1,1\})$, and $\mathcal{G}_{1:n}$ is a filtration which satisfies that $\mathbb{E}[\varepsilon_t \mid \mathcal{G}_t] = 0$ for any $t \in [n]$. Given n sets $\mathcal{S}_1, \ldots, \mathcal{S}_n$, we suppose $s_1, s_2, \ldots, s_n$ are $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_n$-valued random variables such that s_t is $\mathcal{G}_t$-measurable, i.e. σ-algebra $\sigma(s_t) \subseteq \mathcal{G}_t$. For class $\mathcal{A}$ of tuples $\mathbf{a} = (a_1, a_2, \ldots, a_n)$ with $a_t : \mathcal{S}_t \to \mathbb{R}$ for all $t \in [n]$, we have for any $\lambda > 0$,*

$$\{\mathbb{E}_{s_t}\mathbb{E}_{\varepsilon_t}\}_{t=1}^n \left[\sup_{\mathbf{a}\in\mathcal{A}} \sum_{t=1}^n a_t(s_t)\varepsilon_t - \lambda a_t(s_t)^2\right] \le \frac{\log|\mathcal{A}|}{2\lambda},$$

where we denote $\mathbf{a} = (a_1, a_2, \ldots, a_n)$.

Proof. We observe that

$$\begin{aligned}
&\{\mathbb{E}_{s_t}\mathbb{E}_{\varepsilon_t}\}_{t=1}^n \left[\sup_{\mathbf{a}\in\mathcal{A}} \sum_{t=1}^n a_t(s_t)\varepsilon_t - \lambda a_t(s_t)^2\right] \\
&\overset{(i)}{\le} \frac{1}{2\lambda}\log \{\mathbb{E}_{s_t}\mathbb{E}_{\varepsilon_t}\}_{t=1}^n \sup_{\mathbf{a}\in\mathcal{A}} \left[\exp\left(2\lambda\sum_{t=1}^n a_t(s_t)\varepsilon_t - 2\lambda^2 a_t(s_t)^2\right)\right] \\
&\overset{(ii)}{\le} \frac{1}{2\lambda}\log \sum_{\mathbf{a}\in\mathcal{A}} \{\mathbb{E}_{s_t}\mathbb{E}_{\varepsilon_t}\}_{t=1}^n \exp\left(2\lambda\sum_{t=1}^n a_t(s_t)\varepsilon_t - 2\lambda^2 a_t(s_t)^2\right) \\
&= \frac{1}{2\lambda}\log \sum_{\mathbf{a}\in\mathcal{A}} \{\mathbb{E}_{s_t}\mathbb{E}_{\varepsilon_t}\}_{t=1}^{n-1} \left[\exp\left(2\lambda\sum_{t=1}^{n-1} a_t(s_t)\varepsilon_t - 2\lambda^2 a_t(s_t)^2\right)\right. \\
&\qquad \left.\cdot\, \mathbb{E}_{s_n}\left[\exp\left(-2\lambda^2 a_n(s_n)^2\right)\left(\frac{\exp(2\lambda a_n(s_n))}{2} + \frac{\exp(-2\lambda a_n(s_n))}{2}\right) \mid \mathcal{G}_n\right]\right] \\
&\overset{(iii)}{\le} \frac{1}{2\lambda}\log \sum_{\mathbf{a}\in\mathcal{A}} \{\mathbb{E}_{s_t}\mathbb{E}_{\varepsilon_t}\}_{t=1}^{n-1} \exp\left(2\lambda\sum_{t=1}^{n-1} a_t(s_t)\varepsilon_t - 2\lambda^2 a_t(s_t)^2\right),
\end{aligned}$$

where in (i) we use the Jensen's inequality, in (ii) we use replace the sup by the sum since the terms inside sup are always positive, and in (iii) we use the inequality $\exp(x^2/2) \geq \exp(x)/2 + \exp(-x)/2$ for any $x \in \mathbb{R}$. By repeating the argument n times we obtain that

$$\{\mathbb{E}_{s_t}\mathbb{E}_{\varepsilon_t}\}_{t=1}^n \left[\sup_{\mathbf{a}\in\mathcal{A}} \sum_{t=1}^n a_t(s_t)\varepsilon_t - \lambda a_t(s_t)^2\right] \leq \frac{1}{2\lambda}\log\sum_{\mathbf{a}\in\mathcal{A}} 1 = \frac{\log|\mathcal{A}|}{2\lambda}.$$

□

Theorem D.1 implies the following upper bound for non-offset Rademacher processes, which enables us to bound the Rademacher process with random coefficients that are only small on average.

Lemma D.2. *Suppose $\varepsilon_{1:n}$ are n i.i.d. Rademacher random variables, i.e. $\varepsilon_{1:n} \overset{i.i.d.}{\sim} \mathrm{Unif}(\{-1,1\})$, and $\mathcal{G}_{1:n}$ is a filtration which satisfies that $\mathbb{E}[\varepsilon_t \mid \mathcal{G}_t] = 0$ for any $t \in [n]$. Given n sets $\mathcal{S}_1, \ldots, \mathcal{S}_n$, we suppose $s_1, s_2, \ldots, s_n$ are $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_n$-valued random variables such that s_t is $\mathcal{G}_t$-measurable, i.e. σ-algebra $\sigma(s_t) \subset \mathcal{G}_t$. For class $\mathcal{A}$ of tuples $\mathbf{a} = (a_1, a_2, \ldots, a_n)$ with $a_t : \mathcal{S}_t \to \mathbb{R}$ for all $t \in [n]$, we have*

$$\{\mathbb{E}_{s_t}\mathbb{E}_{\varepsilon_t}\}_{t=1}^n \left[\sup_{\mathbf{a}\in\mathcal{A}} \sum_{t=1}^n a_t(s_t)\varepsilon_t\right] \leq \sqrt{2\log|\mathcal{A}|}\cdot\sqrt{\mathbb{E}\left[\sup_{\mathbf{a}\in\mathcal{A}} \sum_{t=1}^n a_t(s_t)^2\right]}.$$

In particular, for $\mathcal{S}_t = \{\pm 1\}^{t-1}$ and $s_t = (\varepsilon_{1:t-1}) \in \mathcal{S}_t$,

$$\mathbb{E}_{\varepsilon_{1:n}}\left[\sup_{\mathbf{a}\in\mathcal{A}} \sum_{t=1}^n a_t(\varepsilon_{1:t-1})\varepsilon_t\right] \leq \sqrt{2\log|\mathcal{A}|}\cdot\sqrt{\mathbb{E}\left[\sup_{\mathbf{a}\in\mathcal{A}} \sum_{t=1}^n a_t(\varepsilon_{1:t-1})^2\right]}. \tag{D.1}$$

Proof. According to Theorem D.1, we have for any $\lambda > 0$,

$$\{\mathbb{E}_{s_t}\mathbb{E}_{\varepsilon_t}\}_{t=1}^n \left[\sup_{\mathbf{a}\in\mathcal{A}} \sum_{t=1}^n a_t(s_t)\varepsilon_t - \lambda a_t(s_t)^2\right] \leq \frac{\log|\mathcal{A}|}{2\lambda}.$$

We let

$$\beta = \{\mathbb{E}_{s_t}\mathbb{E}_{\varepsilon_t}\}_{t=1}^n \left[\sup_{\mathbf{a}\in\mathcal{A}} \sum_{t=1}^n a_t(s_t)^2\right] = \mathbb{E}\left[\sup_{\mathbf{a}\in\mathcal{A}} \sum_{t=1}^n a_t(s_t)^2\right].$$

By choosing $\lambda = \sqrt{\frac{\log|\mathcal{A}|}{2\beta}} > 0$, we obtain that

$$\begin{aligned}
&\{\mathbb{E}_{s_t}\mathbb{E}_{\varepsilon_t}\}_{t=1}^n \left[\sup_{\mathbf{a}\in\mathcal{A}} \sum_{t=1}^n a_t(s_t)\varepsilon_t\right] \\
&\leq \{\mathbb{E}_{s_t}\mathbb{E}_{\varepsilon_t}\}_{t=1}^n \left[\sup_{\mathbf{a}\in\mathcal{A}} \sum_{t=1}^n a_t(s_t)\varepsilon_t - \lambda a_t(s_t)^2\right] + \lambda\cdot\{\mathbb{E}_{s_t}\mathbb{E}_{\varepsilon_t}\}_{t=1}^n \left[\sup_{\mathbf{a}\in\mathcal{A}} \sum_{t=1}^n a_t(s_t)^2\right] \\
&\leq \frac{\log|\mathcal{A}|}{2\lambda} + \lambda\beta = \sqrt{2\beta\log|\mathcal{A}|} = \sqrt{2\log|\mathcal{A}|}\cdot\sqrt{\mathbb{E}\left[\sup_{\mathbf{a}\in\mathcal{A}} \sum_{t=1}^n a_t(s_t)^2\right]}.
\end{aligned}$$

□

Theorem D.2 is an improvement on the finite class lemma in [118, Lemma 1]; the latter result was proved with the supremum (rather than the expected value) over $\varepsilon_{1:n}$ under the square root in (D.1).

D.2 Proof of Theorem 5.2

D.2.1 Proof Outline

The proof has the following structure. Our first step is to write the minimax regret in the dual form using the minimax theorem. This technique is widely used in the analysis of minimax regret of online learning [17, 18, 109, 213–215]. The next step is to truncate the functions and forecaster's strategies away from 0. This analysis technique is also used in [16, 214]. Our main steps in the proof include constructing an offset Rademacher process using a symmetrization argument [216] and using chaining techniques [104, 217] to analyze the offset Rademacher process. The analysis of the chaining steps involves complex dependence of the Rademacher variables and the coefficients, and this is one of the technical hurdles.

Conversion to Dual Form Game

We have the following standard result (see e.g. [18, Lemma 6] or [17, Eq. 27])):

Lemma D.3. *For any $\mathcal{Q} \in \Delta(\mathcal{Y}^n)$, the minimax regret $\mathcal{R}_n(\mathcal{Q})$ has the following dual form representation*

$$\mathcal{R}_n(\mathcal{Q}) = \sup_{\mathbf{p}} \mathbb{E}_{\mathbf{y}\sim\mathbf{p}} \mathcal{R}_n(\mathcal{Q}, \mathbf{p}, \mathbf{y}),$$

where the supremum is over all joint distributions $\mathbf{p} \in \Delta(\mathcal{Y}^n)$ and

$$\mathcal{R}_n(\mathcal{Q}, \mathbf{p}, \mathbf{y}) \coloneqq \sup_{\mathbf{q}\in\mathcal{Q}} \log\left(\frac{\mathbf{q}(\mathbf{y})}{\mathbf{p}(\mathbf{y})}\right). \tag{D.2}$$

Proof of Theorem D.3. We notice that

$$\mathcal{R}_n(\mathcal{Q}) = \inf_{\widehat{\mathbf{p}}} \sup_{\mathbf{y}} \mathcal{R}_n(\mathcal{Q}, \widehat{\mathbf{p}}, \mathbf{y}) = \inf_{\widehat{\mathbf{p}}} \sup_{\mathbf{p}} \mathbb{E}_{\mathbf{y}\sim\mathbf{p}}[\mathcal{R}_n(\mathcal{Q}, \widehat{\mathbf{p}}, \mathbf{y})].$$

Since $\Delta(\mathcal{Y}^n)$ is compact, and $\mathbb{E}_{\mathbf{y}\sim\mathbf{p}}[\mathcal{R}_n(\mathcal{Q}, \widehat{\mathbf{p}}, \mathbf{y})]$ is convex with respect to $\widehat{\mathbf{p}}$ and concave with respect to $\mathbf{p}$, von Neumann minimax theorem [218] gives

$$\mathcal{R}_n(\mathcal{Q}) = \sup_{\mathbf{p}} \inf_{\widehat{\mathbf{p}}} \mathbb{E}_{\mathbf{y}\sim\mathbf{p}}[\mathcal{R}_n(\mathcal{Q}, \widehat{\mathbf{p}}, \mathbf{y})] = \sup_{\mathbf{p}} \mathbb{E}_{\mathbf{y}\sim\mathbf{p}}[\mathcal{R}_n(\mathcal{Q}, \mathbf{p}, \mathbf{y})],$$

where the last inequality uses the fact that the infimum of $\inf_{\widehat{\mathbf{p}}} \mathbb{E}_{\mathbf{y}\sim\mathbf{p}}[\mathcal{R}_n(\mathcal{Q}, \widehat{\mathbf{p}}, \mathbf{y})]$ is attained when $\widehat{\mathbf{p}} = \mathbf{p}$. □

In the remainder, we upper bound the minimax regret for $\mathbb{E}_{\mathbf{y}\sim\mathbf{p}} \mathcal{R}_n(\mathcal{Q}, \mathbf{p}, \mathbf{y})$ for any fixed $\mathbf{p} \in \Delta(\mathcal{Y}^n)$.

Truncation of Functions and Probabilities

For $\mathbf{p} \in \Delta(\mathcal{Y}^n)$ with $\mathbf{p}(\mathbf{y}) = \prod_{t=1}^n p_t(y_t \mid \mathbf{y})$ for every $\mathbf{y} \in \mathcal{Y}^n$, and $\delta < 1/(4|\mathcal{Y}|)$, we define distribution $\mathbf{p}^\delta \in \Delta(\mathcal{Y}^n)$ with $\mathbf{p}^\delta(\mathbf{y}) = \prod_{t=1}^n p_t^\delta(y_t \mid \mathbf{y})$ for every $\mathbf{y} \in \mathcal{Y}^n$, where

$$p_t^\delta(y_t \mid \mathbf{y}) = \begin{cases} \delta & \text{if } p_t(y_t \mid \mathbf{y}) < \delta \\ p_t(y_t \mid \mathbf{y}) & \text{if } \delta \le p_t(y_t \mid \mathbf{y}) \le 2\delta, \\ p_t(y_t \mid \mathbf{y}) \cdot \frac{1-\sum_{y\in\mathcal{Y}} p_t(y|\mathbf{y})\mathbb{I}[\delta \le p_t(y|\mathbf{y}) < 2\delta] - \delta \sum_{y\in\mathcal{Y}} \mathbb{I}[p_t(y|\mathbf{y}) < \delta]}{1-\sum_{y\in\mathcal{Y}} p_t(y|\mathbf{y})\mathbb{I}[p_t(y|\mathbf{y}) < 2\delta]} & \text{if } p_t(y \mid \mathbf{y}) \ge 2\delta. \end{cases} \tag{D.3}$$

It holds that $p_t^\delta(\cdot \mid \mathbf{y}) \in \Delta(\mathcal{Y})$ for any $\mathbf{y} \in \mathcal{Y}^n$ and $t \in [n]$. Additionally, we notice that

$$1 - \sum_{y\in\mathcal{Y}} p_t(y \mid \mathbf{y})\mathbb{I}[\delta \le p_t(y_t \mid \mathbf{y}) < 2\delta] - \delta \sum_{y\in\mathcal{Y}} \mathbb{I}[p_t(y \mid \mathbf{y}) < \delta] \ge 1 - |\mathcal{Y}| \cdot 2\delta \ge \frac{1}{2},$$

which implies that $p_t^\delta(y_t \mid \mathbf{y}) \ge \frac{1}{2} p_t(y_t \mid \mathbf{y})$ if $p_t(y_t \mid \mathbf{y}) \ge 2\delta$. Hence, for any $\mathbf{y} \in \mathcal{Y}^n$, we always have

$$p_t^\delta(y_t \mid \mathbf{y}) > \delta.$$

For class $\mathcal{Q} \subseteq \Delta(\mathcal{Y}^n)$, we define $\mathcal{Q}^\delta = \{\mathbf{q}^\delta \mid \mathbf{q} \in \mathcal{Q}\}$. Then we have the following lemmas:

Lemma D.4. *Suppose $\delta \le \frac{1}{4|\mathcal{Y}|}$. For any $\mathbf{p} \in \Delta(\mathcal{Y}^n)$, $\mathbf{y} \in \mathcal{Y}^n$ and $\mathcal{Q} \subseteq \Delta(\mathcal{Y}^n)$, we have*

$$\mathcal{R}_n(\mathcal{Q}, \mathbf{p}, \mathbf{y}) \le \mathcal{R}_n(\mathcal{Q}^\delta, \mathbf{p}, \mathbf{y}) + 4n\delta \cdot |\mathcal{Y}|.$$

Lemma D.5. *Suppose $\delta \le \frac{1}{4|\mathcal{Y}|}$. For any $\mathcal{Q} \subseteq \Delta(\mathcal{Y}^n)$ and $\mathbf{p} \in \Delta(\mathcal{Y}^n)$, we have*

$$\mathbb{E}_{\mathbf{y}\sim\mathbf{p}}\left[\mathcal{R}_n(\mathcal{Q}^\delta, \mathbf{p}, \mathbf{y})\right] \le \mathbb{E}_{\mathbf{y}\sim\mathbf{p}^\delta}\left[\mathcal{R}_n(\mathcal{Q}^\delta, \mathbf{p}^\delta, \mathbf{y})\right] + 2n^2|\mathcal{Y}|\delta \log\frac{1}{\delta}.$$

The proofs of Theorem D.4 and Theorem D.5 are deferred to appendix D.2.2. In the following, we use $\Delta_n(\mathcal{Y}^n)$ to denote the following joint distribution set:

$$\Delta_n(\mathcal{Y}^n) \coloneqq \left\{\mathbf{q} \in \Delta(\mathcal{Y}^n) : q_t(y_t \mid \mathbf{y}) \ge 1/(n^2|\mathcal{Y}|), \forall \mathbf{y} \in \mathcal{Y}^n\right\}. \tag{D.4}$$

Symmetrization and Construction of Offset Rademacher Process

To facilitate the symmetrization argument, we define the following function $\zeta : \mathbb{R}^+ \to \mathbb{R}$: for any $t \ge 0$,

$$\zeta(t) = \begin{cases} 2\left(\sqrt{t} - 1\right), & t \le 1, \\ 2\log\left(\frac{t+1}{2}\right), & t > 1. \end{cases} \tag{D.5}$$

For the justification of this choice of ζ see section 5.4. We next introduce the following three properties of the function ζ, whose proofs are deferred to appendix D.2.3.

Proposition D.6. *For every $0 < x \le n^2|\mathcal{Y}|$,*

$$\log x \le \zeta(x) - \frac{1}{4\log(n|\mathcal{Y}|)} \cdot \zeta(x)^2. \tag{D.6}$$

Proposition D.7. *For any distribution $f, p \in \Delta(\mathcal{Y})$, we have*

$$\mathbb{E}_{y\sim p}\left[-\zeta\left(\frac{f(y)}{p(y)}\right) - \frac{1}{4\log(n|\mathcal{Y}|)}\cdot\zeta\left(\frac{f(y)}{p(y)}\right)^2\right] \geq 0.$$

The above proposition can also be obtained by noticing that function $-\zeta(x) - \frac{1}{4\log(n|\mathcal{Y}|)}\cdot\zeta(x)^2$ is convex in x, and the result follows from the property of f-divergences [219, Theorem 7.5].

Proposition D.8. *For any $p, q \in [0, \infty)$, we have*

$$|\zeta(p) - \zeta(q)| \leq 2\left|\sqrt{p} - \sqrt{q}\right|.$$

Next, we state the symmetrization argument. The symmetrization argument will use the following circle-dot product distributions.

Definition D.9 (Circle-dot Product Distributions)**.** *For label set $\mathcal{Y}$ and any distribution $\mathbf{p} \in \Delta(\mathcal{Y}^n)$, we define the Circle-dot product distribution $\odot\mathbf{p} \in \Delta(\{-1,1\}^n) \times \Delta(\mathcal{Y}^n) \times \Delta(\mathcal{Y}^n) \times \Delta(\mathcal{Y}^n)$ such that $(\boldsymbol{\varepsilon}, \mathbf{w}, \mathbf{y}, \mathbf{z}) \sim \odot\mathbf{p}$ are sampled according to the following process: first sample $\boldsymbol{\varepsilon} = (\varepsilon_{1:n}) \overset{\text{i.i.d.}}{\sim} \text{Unif}\{-1,1\}$, then repeat the following process for sampling $\mathbf{w} = (w_{1:t}), \mathbf{y} = (y_{1:t})$ and $\mathbf{z} = (z_{1:t})$ from $t = 1$ to n: sample $y_t, z_t \overset{\text{i.i.d.}}{\sim} p_t(\cdot \mid w_{1:t-1})$, and set $w_t = y_t$ if $\varepsilon_t = 1$ or $w_t = z_t$ if $\varepsilon_t = -1$.*

Lemma D.10 (Symmetrization)**.** *For any joint distribution $\mathbf{p} \in \Delta_n(\mathcal{Y}^n)$ where $\Delta_n(\mathcal{Y}^n)$ is defined in eq. (D.4), suppose $(\boldsymbol{\varepsilon}, \mathbf{w}, \mathbf{y}, \mathbf{z}) \sim \odot\mathbf{p}$. Then for any joint distribution class $\mathcal{Q} \subseteq \Delta_n(\mathcal{Y}^n)$, we have the following upper bound:*

$$\begin{aligned}
&\mathbb{E}_{\mathbf{y}\sim\mathbf{p}}\mathcal{R}_n(\mathcal{Q}, \mathbf{p}, \mathbf{y}) \\
&\leq \mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\sum_{t=1}^n \varepsilon_t\zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \frac{1}{4\log(n|\mathcal{Y}|)}\zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)^2\right] \\
&\quad + \mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\sum_{t=1}^n (-\varepsilon_t)\zeta\left(\frac{q_t(z_t \mid \mathbf{w})}{p_t(z_t \mid \mathbf{w})}\right) - \frac{1}{4\log(n|\mathcal{Y}|)}\zeta\left(\frac{q_t(z_t \mid \mathbf{w})}{p_t(z_t \mid \mathbf{w})}\right)^2\right],
\end{aligned} \tag{D.7}$$

where the expectation is with respect to $(\boldsymbol{\varepsilon}, \mathbf{w}, \mathbf{y}, \mathbf{z}) \sim \odot\mathbf{p}$.

The proof of Theorem D.10 is deferred to appendix D.2.3. It is based on the aforementioned properties of function ζ, and the symmetrization technique in [116].

Chaining

We next analyze the right hand side of eq. (D.7) using a chaining argument. For simplicity we only upper bound the first term, and the bound on the second term is similar.

To adopt the chaining argument to the Rademacher process defined in the right hand side of eq. (D.7) while keeping the offset term, we need to establish certain properties of the sequential cover of the function class. Specifically, for any joint distribution $\mathbf{q} \in \mathcal{Q}$, we are required to have some instance $\mathbf{v}$ in the cover, such that the ℓ_2-norm of the coefficients with $\mathbf{q}$ is lower bounded by the ℓ_2-norm of the coefficients with $\mathbf{v}$, as is the following lemma, whose proof is deferred to appendix D.2.4:

Lemma D.11. *Fix joint distribution* $\mathbf{p} \in \Delta(\mathcal{Y}^n)$ *and class* $\mathcal{Q} \subseteq \Delta(\mathcal{Y}^n)$. *Let* $\mathcal{V}(\alpha)$ *be a sequential square-root cover of* $\mathcal{Q}$ *at scale* $\alpha > 0$. *Then for any* $\mathbf{q} \in \mathcal{Q}$, $\mathbf{v}' \in \mathcal{V}(\alpha)$ *and* $\mathbf{w}, \mathbf{y} \in \mathcal{Y}^n$, *there exists* $\mathbf{v} \in \mathcal{V}(\alpha) \cup \{\mathbf{p}\}$ *such that*

$$\sum_{t=1}^{n} \left(\zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \zeta\left(\frac{v_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)\right)^2 \le \sum_{t=1}^{n} \left(\zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \zeta\left(\frac{v'_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)\right)^2, \tag{D.8}$$

and

$$\sum_{t=1}^{n} \zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)^2 \ge \frac{1}{4} \sum_{t=1}^{n} \zeta\left(\frac{v_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)^2. \tag{D.9}$$

Remark D.12. *The above lemma is similar to [17, Eq. (40)], [109, Eq. (46)]. The additional atom* $\mathbf{p}$ *serves as the 'zero' element in [17, Eq. (40)].*

This lemma enables us to keep the offset terms during the chaining process. We now detail these steps. We fix N scales $0 < \alpha_1 < \alpha_2 < \cdots < \alpha_N$, and let $\mathcal{V}(\alpha_i)$ to be the smallest cover of $\mathcal{Q}$ at scale α_i under Theorem 5.1. Then we have the following lemma.

Lemma D.13. *For any* $i \in [N-1]$, *we fix* $\mathbf{v}[\mathbf{q}, \mathbf{p}, \mathbf{w}, \mathbf{y}, \alpha_i] \in \mathcal{V}(\alpha_i)$. *Suppose* $\mathbf{v}[\mathbf{q}, \mathbf{p}, \mathbf{w}, \mathbf{y}, \alpha_N] \in \mathcal{V}(\alpha_N) \cup \{\mathbf{p}\}$ *satisfies eq. (D.9) with* $\mathbf{v} = \mathbf{v}[\mathbf{q}, \mathbf{p}, \mathbf{w}, \mathbf{y}, \alpha_N]$. *We then hwave*

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}} \sum_{t=1}^{n} \left\{ \varepsilon_t \zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \frac{1}{4\log(n|\mathcal{Y}|)} \zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)^2 \right\}\right] \\
&\le \mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}} \sum_{t=1}^{n} \varepsilon_t \left\{ \zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \zeta\left(\frac{v_t[\mathbf{q}, \mathbf{p}, \mathbf{w}, \mathbf{y}, \alpha_1](y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)\right\}\right] \\
&\quad + \sum_{i=1}^{N-1} \mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}} \sum_{t=1}^{n} \varepsilon_t \left\{ \zeta\left(\frac{v_t[\mathbf{q}, \mathbf{p}, \mathbf{w}, \mathbf{y}, \alpha_i](y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \zeta\left(\frac{v_t[\mathbf{q}, \mathbf{p}, \mathbf{w}, \mathbf{y}, \alpha_{i+1}](y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)\right\}\right] \\
&\quad + \mathbb{E}\left[\sup_{\mathbf{v}\in\mathcal{V}(\alpha_N)\cup\{\mathbf{p}\}} \sum_{t=1}^{n} \left\{ \varepsilon_t \zeta\left(\frac{v_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \frac{1}{16\log(n|\mathcal{Y}|)} \cdot \zeta\left(\frac{v_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)^2 \right\}\right],
\end{aligned} \tag{D.10}$$

where the expectation is with respect to $(\boldsymbol{\varepsilon}, \mathbf{w}, \mathbf{y}, \mathbf{z}) \sim \odot\mathbf{p}$.

The proof of Theorem D.13 is deferred to appendix D.2.4. Next, we further upper bound the three terms in eq. (D.10). Specifically, we will use Cauchy-Schwarz inequality to bound the first term. The second term is a form of sequential Rademacher process. The third term is an offset Rademacher process. However, the key difficulty in dealing with the second and third terms is that the coefficients of the Rademacher random variables are not uniformly bounded by a constant, and directly applying prior techniques does not provide the desired upper bounds. To overcome this issue, we employ a technique that uses offset complexities instead, as in the proof of Theorem D.2 (see Theorem D.14 in the proof of Theorem 5.2 for a discussion). The formal proof of these arguments, together with the full proof of Theorem 5.2, is deferred to appendix D.2.5.

D.2.2 Missing Proofs in appendix D.2.1

Proof of Theorem D.4. Given the formula of $\mathcal{R}_n(\mathcal{Q}, \mathbf{p}, \mathbf{y})$ in eq. (D.2), we only need to verify

$$\sup_{\mathbf{q}\in\mathcal{Q}} \sum_{t=1}^{n} \frac{1}{q_t(y_t \mid \mathbf{y})} \geq \sup_{\mathbf{q}\in\mathcal{Q}^\delta} \sum_{t=1}^{n} \frac{1}{q_t(y_t \mid \mathbf{y})} - 4n|\mathcal{Y}|\delta. \tag{D.11}$$

Notice that according to our construction of truncation in eq. (D.3), we have for any $y \in \mathcal{Y}$ and $p \in \Delta(\mathcal{Y})$,

$$\log p(y) - \log p^\delta(y) = \log \frac{p(y)}{p^\delta(y)} \leq -\log\left(1 - 2\delta \cdot |\mathcal{Y}|\right) \leq 4\delta \cdot |\mathcal{Y}|, \tag{D.12}$$

where the last inequality uses the fact that $\delta \leq \frac{1}{4|\mathcal{Y}|}$ and $-\log(1-t) \leq 2t$ for any $t \leq 1/2$. Hence after noticing that $\mathcal{Q}^\delta = \{\mathbf{q}^\delta : \mathbf{q} \in \mathcal{Q}\}$, we obtain eq. (D.11). □

Proof of Theorem D.5. For any $s \in [n+1]$, we use notation $\mathbf{p}^{s,\delta} = (p_1^{s,\delta}, \ldots, p_n^{s,\delta})$ to denote a joint distribution such that

$$p_t^{s,\delta}(y_t \mid \mathbf{y}) = \begin{cases} p_t(y_t \mid \mathbf{y}) & \text{if } t < s, \\ p_t^\delta(y_t \mid \mathbf{y}) & \text{if } t \geq s. \end{cases} \qquad \forall \mathbf{y} \in \mathcal{Y}^n.$$

Then we have $\mathbf{p}^{1,\delta} = \mathbf{p}^\delta$ and $\mathbf{p}^{n+1,\delta} = \mathbf{p}$, and we can decompose

$$\begin{aligned} &\mathbb{E}_{\mathbf{y}\sim\mathbf{p}} \mathcal{R}_n(\mathcal{Q}^\delta, \mathbf{p}, \mathbf{y}) - \mathbb{E}_{\mathbf{y}\sim\mathbf{p}^\delta} \mathcal{R}_n(\mathcal{Q}^\delta, \mathbf{p}^\delta, \mathbf{y}) \\ &= \sum_{s=1}^{n} \left[\mathbb{E}_{\mathbf{y}\sim\mathbf{p}^{s+1,\delta}} \mathcal{R}_n(\mathcal{Q}^\delta, \mathbf{p}^{s+1,\delta}, \mathbf{y}) - \mathbb{E}_{\mathbf{y}\sim\mathbf{p}^{s,\delta}} \mathcal{R}_n(\mathcal{Q}^\delta, \mathbf{p}^{s,\delta}, \mathbf{y})\right]. \end{aligned} \tag{D.13}$$

We expand the right hand side of eq. (D.13) for each $s \in [n]$:

$$\begin{aligned} &\mathbb{E}_{\mathbf{y}\sim\mathbf{p}^{s+1,\delta}} \mathcal{R}_n(\mathcal{Q}^\delta, \mathbf{p}^{s+1,\delta}, \mathbf{y}) - \mathbb{E}_{\mathbf{y}\sim\mathbf{p}^{s,\delta}} \mathcal{R}_n(\mathcal{Q}^\delta, \mathbf{p}^{s,\delta}, \mathbf{y}) \\ &= \left\{\mathbb{E}_{y_t\sim p_t(\cdot|\mathbf{y})}\right\}_{t=1}^{s-1} \left[\mathbb{E}_{y_s\sim p_s(\cdot|\mathbf{y})} \left\{\mathbb{E}_{y_t\sim p_t^\delta(\cdot|\mathbf{y})}\right\}_{t=s+1}^{n} \mathcal{R}_n(\mathcal{Q}^\delta, \mathbf{p}^{s+1,\delta}, \mathbf{y}) \right. \\ &\quad \left. - \mathbb{E}_{y_s\sim p_s^\delta(\cdot|\mathbf{y})} \left\{\mathbb{E}_{y_t\sim p_t^\delta(\cdot|\mathbf{y})}\right\}_{t=s+1}^{n} \mathcal{R}_n(\mathcal{Q}^\delta, \mathbf{p}^{s,\delta}, \mathbf{y})\right]. \end{aligned} \tag{D.14}$$

Next, we fix $y_{1:s-1}$ and upper bound the expression inside the expectation:

$$\begin{aligned} &\mathbb{E}_{y_s\sim p_s(\cdot|\mathbf{y})} \left\{\mathbb{E}_{y_t\sim p_t^\delta(\cdot|\mathbf{y})}\right\}_{t=s+1}^{n} \mathcal{R}_n(\mathcal{Q}^\delta, \mathbf{p}^{s+1,\delta}, \mathbf{y}) - \mathbb{E}_{y_s\sim p_s^\delta(\cdot|\mathbf{y})} \left\{\mathbb{E}_{y_t\sim p_t^\delta(\cdot|\mathbf{y})}\right\}_{t=s+1}^{n} \mathcal{R}_n(\mathcal{Q}^\delta, \mathbf{p}^{s,\delta}, \mathbf{y}) \\ &= \mathbb{E}_{y_s\sim p_s(\cdot|\mathbf{y})} \left\{\mathbb{E}_{y_t\sim p_t^\delta(\cdot|\mathbf{y})}\right\}_{t=s+1}^{n} \left[\mathcal{R}_n(\mathcal{Q}^\delta, \mathbf{p}^{s+1,\delta}, \mathbf{y}) - \mathcal{R}_n(\mathcal{Q}^\delta, \mathbf{p}^{s,\delta}, \mathbf{y})\right] \\ &\quad + \left[\mathbb{E}_{y_s\sim p_s(\cdot|\mathbf{y})} \left\{\mathbb{E}_{y_t\sim p_t^\delta(\cdot|\mathbf{y})}\right\}_{t=s+1}^{n} \mathcal{R}_n(\mathcal{Q}^\delta, \mathbf{p}^{s,\delta}, \mathbf{y}) - \mathbb{E}_{y_s\sim p_s^\delta(\cdot|\mathbf{y})} \left\{\mathbb{E}_{y_t\sim p_t^\delta(\cdot|\mathbf{y})}\right\}_{t=s+1}^{n} \mathcal{R}_n(\mathcal{Q}^\delta, \mathbf{p}^{s,\delta}, \mathbf{y})\right] \end{aligned} \tag{D.15}$$

For the first term in the right hand side of eq. (D.15), when fixing $\mathbf{y} \in \mathcal{Y}^n$, we have

$$\mathcal{R}_n(\mathcal{Q}^\delta, \mathbf{p}^{s+1,\delta}, \mathbf{y}) - \mathcal{R}_n(\mathcal{Q}^\delta, \mathbf{p}^{s,\delta}, \mathbf{y}) = \sum_{t=1}^{n} \log\left(\frac{p_t^{s,\delta}(y_t \mid \mathbf{y})}{p_t^{s+1,\delta}(y_t \mid \mathbf{y})}\right) = \log\left(\frac{p_s^\delta(y_s \mid \mathbf{y})}{p_s(y_s \mid \mathbf{y})}\right),$$

which implies that

$$\mathbb{E}_{y_s\sim p_s(\cdot|\mathbf{y})}\{\mathbb{E}_{y_t\sim p_t(\cdot|\mathbf{y})}\}_{t=s+1}^{n}\left[\mathcal{R}_n(\mathcal{Q}^\delta,\mathbf{p}^{s+1,\delta},\mathbf{y})-\mathcal{R}_n(\mathcal{Q}^\delta,\mathbf{p}^{s,\delta},\mathbf{y})\right]$$
$$=\mathbb{E}_{y_s\sim p_s(\cdot|\mathbf{y})}\left[\log\left(\frac{p_s^\delta(y_s\mid\mathbf{y})}{p_s(y_s\mid\mathbf{y})}\right)\right]=-D_{\mathrm{KL}}(p_s(y_s\mid\mathbf{y})\|p_s^\delta(y_s\mid\mathbf{y}))\le 0.$$

For the second term in eq. (D.15), when fixing $y_{1:s-1}$, we have

$$\begin{aligned}
&\mathbb{E}_{y_s\sim p_s(\cdot|\mathbf{y})}\{\mathbb{E}_{y_t\sim p_t^\delta(\cdot|\mathbf{y})}\}_{t=s+1}^{n}\mathcal{R}_n(\mathcal{Q}^\delta,\mathbf{p}^{s,\delta},\mathbf{y})-\mathbb{E}_{y_s\sim p_s^\delta(\cdot|\mathbf{y})}\{\mathbb{E}_{y_t\sim p_t^\delta(\cdot|\mathbf{y})}\}_{t=s+1}^{n}\mathcal{R}_n(\mathcal{Q}^\delta,\mathbf{p}^{s,\delta},\mathbf{y})\\
&\overset{(i)}{=}\mathbb{E}_{y_s\sim p_s(\cdot|\mathbf{y})}\{\mathbb{E}_{y_t\sim p_t^\delta(\cdot|\mathbf{y})}\}_{t=s+1}^{n}\left[\sup_{\mathbf{q}\in\mathcal{Q}^\delta}\left\{\sum_{t=1}^{s-1}\log\frac{q_t(y_t\mid\mathbf{y})}{p_t(y_t\mid\mathbf{y})}+\sum_{t=s}^{n}\log\frac{q_t(y_t\mid\mathbf{y})}{p_t^\delta(y_t\mid\mathbf{y})}\right\}\right]\\
&\quad-\mathbb{E}_{y_s\sim p_s^\delta(\cdot|\mathbf{y})}\{\mathbb{E}_{y_t\sim p_t^\delta(\cdot|\mathbf{y})}\}_{t=s+1}^{n}\left[\sup_{\mathbf{q}\in\mathcal{Q}^\delta}\left\{\sum_{t=1}^{s-1}\log\frac{q_t(y_t\mid\mathbf{y})}{p_t(y_t\mid\mathbf{y})}+\sum_{t=s}^{n}\log\frac{q_t(y_t\mid\mathbf{y})}{p_t^\delta(y_t\mid\mathbf{y})}\right\}\right]\\
&\overset{(ii)}{=}\mathbb{E}_{y_s\sim p_s(\cdot|\mathbf{y})}\{\mathbb{E}_{y_t\sim p_t^\delta(\cdot|\mathbf{y})}\}_{t=s+1}^{n}\left[\sup_{\mathbf{q}\in\mathcal{Q}^\delta}\sum_{t=1}^{n}\log\frac{q_t(y_t\mid\mathbf{y})}{p_t^\delta(y_t\mid\mathbf{y})}\right]\\
&\quad-\mathbb{E}_{y_s\sim p_s^\delta(\cdot|\mathbf{y})}\{\mathbb{E}_{y_t\sim p_t^\delta(\cdot|\mathbf{y})}\}_{t=s+1}^{n}\left[\sup_{\mathbf{q}\in\mathcal{Q}^\delta}\sum_{t=1}^{n}\log\frac{q_t(y_t\mid\mathbf{y})}{p_t^\delta(y_t\mid\mathbf{y})}\right],
\end{aligned}\tag{D.16}$$

where (i) uses the formula of $\mathcal{R}_n(\mathcal{Q},\mathbf{p},\mathbf{y})$ in eq. (D.2) and the form of $\mathbf{p}^{s,\delta}$, and (ii) uses the fact that for $t\le s-1$, $p_t(y_t\mid\mathbf{y})$ cancels out in both terms, hence we can replace them by $p_t^\delta(y_t\mid\mathbf{y})$ at no additional cost. Notice that $\delta\le p_t^\delta(y_t\mid\mathbf{y})\le 1$ and $\delta\le q_t^\delta(y_t\mid\mathbf{y})\le 1$ hold for any $\mathbf{q}\in\mathcal{Q}^\delta$, $\mathbf{y}\in\mathcal{Y}^n$ and $t\in[n]$. Hence,

$$\left|\sup_{\mathbf{q}\in\mathcal{Q}^\delta}\sum_{t=1}^{n}\log\frac{q_t(y_t\mid\mathbf{y})}{p_t^\delta(y_t\mid\mathbf{y})}\right|\le n\log\frac{1}{\delta},\quad\forall\mathbf{y}\in\mathcal{Y}^n$$

which implies that when fixed $y_{1:s-1}$,

$$\text{RHS of eq. (D.16)}\le 2\mathrm{TV}\left(p_s(\cdot\mid\mathbf{y}),p_s^\delta(\cdot\mid\mathbf{y})\right)\cdot n\log\frac{1}{\delta}.$$

Based on eq. (D.3), we can calculate that

$$\mathrm{TV}\left(p_s(\cdot\mid\mathbf{y}),p_s^\delta(\cdot\mid\mathbf{y})\right)=\sum_{y\in\mathcal{Y}}(\delta-p_s(y\mid\mathbf{y}))\vee 0\le|\mathcal{Y}|\delta,$$

which implies that

$$\mathbb{E}_{y_s\sim p_s(\cdot|\mathbf{y})}\{\mathbb{E}_{y_t\sim p_t^\delta(\cdot|\mathbf{y})}\}_{t=s+1}^{n}\mathcal{R}_n(\mathcal{Q}^\delta,\mathbf{p}^{s,\delta},\mathbf{y})-\mathbb{E}_{y_s\sim p_s^\delta(\cdot|\mathbf{y})}\{\mathbb{E}_{y_t\sim p_t^\delta(\cdot|\mathbf{y})}\}_{t=s+1}^{n}\mathcal{R}_n(\mathcal{Q}^\delta,\mathbf{p}^{s,\delta},\mathbf{y})\le 2n|\mathcal{Y}|\delta\log\frac{1}{\delta}.$$

Bringing this upper bound back to eq. (D.15) and then further back to eq. (D.14), we obtain that

$$\mathbb{E}_{\mathbf{y}\sim\mathbf{p}^{s+1,\delta}}\mathcal{R}_n(\mathcal{Q}^\delta,\mathbf{p}^{s+1,\delta},\mathbf{y})-\mathbb{E}_{\mathbf{y}\sim\mathbf{p}^{s,\delta}}\mathcal{R}_n(\mathcal{Q}^\delta,\mathbf{p}^{s,\delta},\mathbf{y})\le 2n|\mathcal{Y}|\delta\log\frac{1}{\delta}.$$

Hence, according to eq. (D.13), we have

$$\mathbb{E}_{\mathbf{y}\sim\mathbf{p}}\mathcal{R}_n(\mathcal{Q}^\delta,\mathbf{p},\mathbf{y})\le\mathbb{E}_{\mathbf{y}\sim\mathbf{p}^\delta}\mathcal{R}_n(\mathcal{Q}^\delta,\mathbf{p}^\delta,\mathbf{y})+2n^2|\mathcal{Y}|\delta\log\frac{1}{\delta}.$$

□

D.2.3 Missing Proofs in appendix D.2.1

Proof of Theorem D.6. We first verify the upper bounds part in eq. (D.6). When $0 < x \le 1$, using the inequality $\log(1+t) \le t - t^2/2$ which holds for any $-1 < t \le 0$, we have for $n \ge 7$,

$$\log x = 2\log\left(\sqrt{x}\right) \le 2\left(\sqrt{x}-1\right) - \left(\sqrt{x}-1\right)^2 = \zeta(x) - \frac{1}{4}\zeta(x)^2 \le \zeta(x) - \frac{1}{2\log(n|\mathcal{Y}|)}\zeta(x)^2.$$

For $x > 1$, we first notice that function

$$\xi(x) = \frac{2\log((x+1)/2) - \log(x)}{\log^2((x+1)/2)}.$$

is a monotonically decreasing function on $[0,\infty)$, and for every $n \ge 7$ we have

$$\xi(n^2|\mathcal{Y}|) = \frac{2\log((n^2|\mathcal{Y}|+1)/2) - \log(n^2|\mathcal{Y}|)}{\log^2((n^2|\mathcal{Y}|+1)/2)} \ge \frac{1}{2\log(n^2|\mathcal{Y}|)} \ge \frac{1}{4\log(n|\mathcal{Y}|)},$$

which implies

$$\xi(x) \ge \xi(n^2|\mathcal{Y}|) \ge \frac{1}{4\log(n|\mathcal{Y}|)}, \quad \forall x \le n^2|\mathcal{Y}|.$$

Hence we obtain for any $0 < x \le n^2|\mathcal{Y}|$,

$$\log x \le \zeta(x) - \frac{1}{4\log(n|\mathcal{Y}|)}\zeta(x)^2.$$

□

Proof of Theorem D.7. First notice that for any $x \ge 1$ we have $\zeta(x) \ge 0$, and

$$\log\left(\frac{x+1}{2}\right) \le \sqrt{x} - 1.$$

Hence, we only need to verify

$$\mathbb{E}_{y\sim p}\left[-2\cdot\left(\sqrt{\frac{f(y)}{p(y)}}-1\right) - \frac{1}{4\log(n|\mathcal{Y}|)}\cdot\left(2\cdot\left(\sqrt{\frac{f(y)}{p(y)}}-1\right)\right)^2\right] \ge 0.$$

This can be verified by

$$\begin{aligned}
&\mathbb{E}_{y\sim p}\left[-\sqrt{\frac{f(y)}{p(y)}} + 1 - \frac{1}{2\log(n|\mathcal{Y}|)}\cdot\left(\sqrt{\frac{f(y)}{p(y)}}-1\right)^2\right] \\
&\ge \mathbb{E}_{y\sim p}\left[-\sqrt{\frac{f(y)}{p(y)}} + 1 - \frac{1}{2}\cdot\left(\sqrt{\frac{f(y)}{p(y)}}-1\right)^2\right] \\
&= \sum_{y\in\mathcal{Y}}\left[-\sqrt{f(y)p(y)} + p(y) - \frac{1}{2}f(y) + \sqrt{f(y)p(y)} - \frac{1}{2}p(y)\right] \\
&= 0.
\end{aligned}$$

□

Proof of Theorem D.8. We only need to verify that the function

$$h(t) = \begin{cases} 2(t-1) & \text{if } 0 < t \le 1 \\ 2\log\left(\frac{1+t^2}{2}\right) & \text{if } t > 1 \end{cases}$$

is a Lipschitz function with Lipschitz constant 2. This can be seen from

$$\frac{dh(t)}{dt} = \begin{cases} 2 & \text{if } 0 < t \le 1, \\ \frac{4t}{1+t^2} & \text{if } t > 1, \end{cases}$$

which satisfies $\left|\frac{dh(t)}{dt}\right| \le 2$ for any $t > 0$. □

Proof of Theorem D.10. Fix distribution $\mathbf{p} \in \Delta_n(\mathcal{Y}^n)$, and suppose random variables $(\boldsymbol{\varepsilon}, \mathbf{w}, \mathbf{y}, \mathbf{z}) \sim \odot\mathbf{p}$. Noticing that the marginal distribution of $\mathbf{w}$ is $\mathbf{p}$, we have

$$\begin{aligned} \mathbb{E}_{\mathbf{y}\sim\mathbf{p}}\mathcal{R}_n(\mathcal{Q}, \mathbf{p}, \mathbf{y}) &= \mathbb{E}_{\mathbf{w}\sim\mathbf{p}}\left[\sup_{\mathbf{q}\in\mathcal{Q}} \sum_{t=1}^{n} \left(\log q_t(w_t \mid \mathbf{w}) - \log p_t(w_t \mid \mathbf{w})\right)\right] \\ &\le \mathbb{E}_{\mathbf{w}\sim\mathbf{p}}\left[\sup_{\mathbf{q}\in\mathcal{Q}} \sum_{t=1}^{n} \left\{\zeta\left(\frac{q_t(w_t \mid \mathbf{w})}{p_t(w_t \mid \mathbf{w})}\right) - \frac{1}{4\log(n|\mathcal{Y}|)}\zeta\left(\frac{q_t(w_t \mid \mathbf{w})}{p_t(w_t \mid \mathbf{w})}\right)^2\right\}\right], \end{aligned} \tag{D.17}$$

where the last steps follows from Theorem D.6, and the fact that $\mathbf{p} \in \Delta_n(\mathcal{Y}^n)$ and $\mathcal{Q} \subseteq \Delta(\mathcal{Y}^n)$. Next, we define random variables $\mathbf{v} = (v_{1:n})$ coupled with random variables $(\boldsymbol{\varepsilon}, \mathbf{w}, \mathbf{y}, \mathbf{z}) \sim \odot\mathbf{p}$, in the way that

$$v_t = z_t \text{ if } \varepsilon_t = 1 \quad \text{and} \quad v_t = y_t \text{ if } \varepsilon_t = -1.$$

Then the marginal distribution of v_t conditioned on $w_{1:t-1}$ is $p_t(\cdot \mid \mathbf{w})$. Hence Theorem D.7 gives that for any $\mathbf{q} \in \mathcal{Q}$ and $t \in [n]$,

$$\mathbb{E}\left[-\zeta\left(\frac{q_t(v_t \mid \mathbf{w})}{p_t(v_t \mid \mathbf{w})}\right) - \frac{1}{4\log(n|\mathcal{Y}|)}\zeta\left(\frac{q_t(v_t \mid \mathbf{w})}{p_t(v_t \mid \mathbf{w})}\right)^2 \;\middle|\; w_{1:t-1}\right] \ge 0. \tag{D.18}$$

Hence we can further upper bound

$$\begin{aligned} &\mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}} \sum_{t=1}^{n} \left\{\zeta\left(\frac{q_t(w_t \mid \mathbf{w})}{p_t(w_t \mid \mathbf{w})}\right) - \frac{1}{4\log(n|\mathcal{Y}|)}\zeta\left(\frac{q_t(w_t \mid \mathbf{w})}{p_t(w_t \mid \mathbf{w})}\right)^2\right\}\right] \\ &\overset{(i)}{\le} \mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}} \sum_{t=1}^{n} \left\{\zeta\left(\frac{q_t(w_t \mid \mathbf{w})}{p_t(w_t \mid \mathbf{w})}\right) - \frac{1}{4\log(n|\mathcal{Y}|)}\zeta\left(\frac{q_t(w_t \mid \mathbf{w})}{p_t(w_t \mid \mathbf{w})}\right)^2\right\}\right] \\ &\quad + \sum_{t=1}^{n} \mathbb{E}\left[-\zeta\left(\frac{q_t(v_t \mid \mathbf{w})}{p_t(v_t \mid \mathbf{w})}\right) - \frac{1}{4\log(n|\mathcal{Y}|)}\zeta\left(\frac{q_t(v_t \mid \mathbf{w})}{p_t(v_t \mid \mathbf{w})}\right)^2 \;\middle|\; w_{1:t-1}\right]\Bigg] \\ &\overset{(ii)}{\le} \mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}} \sum_{t=1}^{n} \left\{\zeta\left(\frac{q_t(w_t \mid \mathbf{w})}{p_t(w_t \mid \mathbf{w})}\right) - \zeta\left(\frac{q_t(v_t \mid \mathbf{w})}{p_t(v_t \mid \mathbf{w})}\right)\right.\right. \end{aligned}$$

$$- \frac{1}{4\log(n|\mathcal{Y}|)}\zeta\left(\frac{q_t(w_t \mid \mathbf{w})}{p_t(w_t \mid \mathbf{w})}\right)^2 - \frac{1}{4\log(n|\mathcal{Y}|)}\zeta\left(\frac{q_t(v_t \mid \mathbf{w})}{p_t(v_t \mid \mathbf{w})}\right)^2 \Bigg\}\Bigg], \tag{D.19}$$

where in (i) we use eq. (D.18), in (ii) we use the Jensen's inequality. According to the construction of random variables $\boldsymbol{\varepsilon}, \mathbf{y}, \mathbf{z}, \mathbf{w}, \mathbf{v}$, we have

$$\zeta\left(\frac{q_t(w_t \mid \mathbf{w})}{p_t(w_t \mid \mathbf{w})}\right) - \zeta\left(\frac{q_t(v_t \mid \mathbf{w})}{p_t(v_t \mid \mathbf{w})}\right) = \varepsilon_t\zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \varepsilon_t\zeta\left(\frac{q_t(z_t \mid \mathbf{w})}{p_t(z_t \mid \mathbf{w})}\right)$$

and

$$\zeta\left(\frac{q_t(w_t \mid \mathbf{w})}{p_t(w_t \mid \mathbf{w})}\right)^2 + \zeta\left(\frac{q_t(v_t \mid \mathbf{w})}{p_t(v_t \mid \mathbf{w})}\right)^2 = \zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)^2 + \zeta\left(\frac{q_t(z_t \mid \mathbf{w})}{p_t(z_t \mid \mathbf{w})}\right)^2.$$

Bringing this back to eq. (D.19) and further back to eq. (D.17), we obtain that

$$\begin{aligned}
&\mathbb{E}_{\mathbf{y}\sim\mathbf{p}}\mathcal{R}_n(\mathcal{Q}, \mathbf{p}, \mathbf{y}) \\
&\quad\le \mathbb{E}\Bigg[\sup_{\mathbf{q}\in\mathcal{Q}}\sum_{t=1}^n \Bigg\{\varepsilon_t\zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \varepsilon_t\zeta\left(\frac{q_t(z_t \mid \mathbf{w})}{p_t(z_t \mid \mathbf{w})}\right) \\
&\qquad - \frac{1}{4\log(n|\mathcal{Y}|)}\zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)^2 - \frac{1}{4\log(n|\mathcal{Y}|)}\zeta\left(\frac{q_t(z_t \mid \mathbf{w})}{p_t(z_t \mid \mathbf{w})}\right)^2 \Bigg\}\Bigg] \\
&\quad\le \mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\sum_{t=1}^n \left\{\varepsilon_t\zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \frac{1}{4\log(n|\mathcal{Y}|)}\zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)^2\right\}\right] \\
&\qquad + \mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\sum_{t=1}^n \left\{(-\varepsilon_t)\zeta\left(\frac{q_t(z_t \mid \mathbf{w})}{p_t(z_t \mid \mathbf{w})}\right) - \frac{1}{4\log(n|\mathcal{Y}|)}\zeta\left(\frac{q_t(z_t \mid \mathbf{w})}{p_t(z_t \mid \mathbf{w})}\right)^2\right\}\right],
\end{aligned}$$

where the last inequality uses Jensen's inequality. □

D.2.4 Missing Proofs in appendix D.2.1

Proof of Theorem D.11. We fix $\mathbf{p} \in \Delta(\mathcal{Y}^n)$. For $\mathbf{q} \in \mathcal{Q}$, $\mathbf{v}' \in \mathcal{V}(\alpha)$ and $\mathbf{w}, \mathbf{y} \in \mathcal{Y}^n$, if we have

$$\sum_{t=1}^n \zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)^2 \ge \frac{1}{4}\sum_{t=1}^n \zeta\left(\frac{v'_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)^2$$

then we let $\mathbf{v} = \mathbf{v}'$ and it is easy to see that eq. (D.8) and eq. (D.9) both hold. Next we assume

$$\sum_{t=1}^n \zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)^2 < \frac{1}{4}\sum_{t=1}^n \zeta\left(\frac{v'_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)^2. \tag{D.20}$$

With $\mathbf{v} = \mathbf{p} \in \mathcal{V}(\alpha) \cup \{\mathbf{p}\}$ we will verify eq. (D.8) and eq. (D.9). First, since $\zeta(1) = 0$, we have

$$\sum_{t=1}^n \zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)^2 \ge 0 = \frac{1}{4}\sum_{t=1}^n \zeta\left(\frac{v_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)^2,$$

hence eq. (D.9) holds. Next, according to eq. (D.20) and Cauchy-Schwarz inequality,

$$\left(\sum_{t=1}^{n}\zeta\left(\frac{q_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)^2\right)\left(\sum_{t=1}^{n}\zeta\left(\frac{v'_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)^2\right)\geq\left(\sum_{t=1}^{n}\zeta\left(\frac{q_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)\zeta\left(\frac{v'_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)\right)^2,$$

we have

$$\sum_{t=1}^{n}\zeta\left(\frac{v'_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)^2\geq 2\sum_{t=1}^{n}\zeta\left(\frac{q_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)\zeta\left(\frac{v'_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right),$$

which implies that

$$\begin{aligned}\sum_{t=1}^{n}\left(\zeta\left(\frac{q_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)-\zeta\left(\frac{v'_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)\right)^2&\geq\sum_{t=1}^{n}\zeta\left(\frac{q_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)^2\\&=\sum_{t=1}^{n}\left(\zeta\left(\frac{q_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)-\zeta\left(\frac{v_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)\right)^2,\end{aligned}$$

hence eq. (D.8) holds. □

Proof of Theorem D.13. According to our choice of $\mathbf{v}[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_N]\in\mathcal{V}(\alpha_N)\cup\{\mathbf{p}\}$, eq. (D.9) holds with $\mathbf{v}=\mathbf{v}[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_N]$. Hence, we can upper bound the left hand side of eq. (D.10) as follows:

$$\begin{aligned}&\sup_{\mathbf{q}\in\mathcal{Q}}\sum_{t=1}^{n}\left\{\varepsilon_t\zeta\left(\frac{q_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)-\frac{1}{4\log(n|\mathcal{Y}|)}\zeta\left(\frac{q_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)^2\right\}\\&\quad=\sup_{\mathbf{q}\in\mathcal{Q}}\left\{\sum_{t=1}^{n}\varepsilon_t\left\{\zeta\left(\frac{q_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)-\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_N](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)\right\}\right.\\&\qquad+\sum_{t=1}^{n}\left\{\varepsilon_t\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_N](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)-\frac{1}{16\log(n|\mathcal{Y}|)}\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_N](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)^2\right\}\\&\qquad\left.+\frac{1}{16\log(n|\mathcal{Y}|)}\sum_{t=1}^{n}\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_N](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)^2-\frac{1}{4\log(n|\mathcal{Y}|)}\sum_{t=1}^{n}\zeta\left(\frac{q_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)^2\right\}\\&\overset{(i)}{\leq}\sup_{\mathbf{q}\in\mathcal{Q}}\left\{\sum_{t=1}^{n}\varepsilon_t\left\{\zeta\left(\frac{q_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)-\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_N](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)\right\}\right.\\&\qquad\left.+\sum_{t=1}^{n}\left\{\varepsilon_t\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_N](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)-\frac{1}{16\log(n|\mathcal{Y}|)}\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_N](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)^2\right\}\right\}\\&\overset{(ii)}{\leq}\sup_{\mathbf{q}\in\mathcal{Q}}\left\{\sum_{t=1}^{n}\varepsilon_t\left\{\zeta\left(\frac{q_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)-\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_N](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)\right\}\right\}\\&\qquad+\sup_{\mathbf{v}\in\mathcal{V}(\alpha_N)\cup\{\mathbf{p}\}}\left\{\sum_{t=1}^{n}\left\{\varepsilon_t\zeta\left(\frac{v_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)-\frac{1}{16\log(n|\mathcal{Y}|)}\zeta\left(\frac{v_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)^2\right\}\right\},\end{aligned}\tag{D.21}$$

where (i) uses the condition eq. (D.9), and (ii) uses the Jensen's inequality and the chioce $\mathbf{v}[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_N] \in \mathcal{V}(\alpha_N) \cup \{\mathbf{p}\}$. Next, we introduce $\mathbf{v}[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_i]$, and further upper bound the first term above via telescoping:

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\sum_{t=1}^{n}\varepsilon_t\left\{\zeta\left(\frac{q_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)-\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_N](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)\right\}\right]\\
&\quad=\mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\left\{\sum_{t=1}^{n}\varepsilon_t\left\{\zeta\left(\frac{q_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)-\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_1](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)\right\}\right.\right.\\
&\qquad\left.\left.+\sum_{i=1}^{N-1}\sum_{t=1}^{n}\varepsilon_t\left\{\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_i](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)-\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_{i+1}](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)\right\}\right\}\right]\\
&\quad\le\mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\sum_{t=1}^{n}\varepsilon_t\left\{\zeta\left(\frac{q_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)-\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_1](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)\right\}\right]\\
&\qquad+\sum_{i=1}^{N-1}\mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\sum_{t=1}^{n}\varepsilon_t\left\{\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_i](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)-\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_{i+1}](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)\right\}\right],
\end{aligned}
$$

where the last inequality is due to Jensen's inequality. Bringing this back to eq. (D.21), we obtain eq. (D.10). □

D.2.5 Proof of Theorem 5.2

Proof of Theorem 5.2. First of all, according to Theorem D.3, for any joint distribution class $\mathcal{Q}\subseteq\Delta(\mathcal{Y}^n)$, we have

$$\mathcal{R}_n(\mathcal{Q})=\sup_{\mathbf{p}}\mathbb{E}_{\mathbf{y}\sim\mathbf{p}}\mathcal{R}_n(\mathcal{Q},\mathbf{p},\mathbf{y}),$$

where the supreme is taken over all $\mathbf{p}\in\Delta(\mathcal{Y}^n)$. According to Theorem D.4 and Theorem D.5, we have

$$\mathbb{E}_{\mathbf{y}\sim\mathbf{p}}\left[\mathcal{R}_n(\mathcal{Q},\mathbf{p},\mathbf{y})\right]\le\mathbb{E}_{\mathbf{y}\sim\mathbf{p}^\delta}\left[\mathcal{R}_n(\mathcal{Q}^\delta,\mathbf{p}^\delta,\mathbf{y})\right]+6n^2|\mathcal{Y}|\delta\log\frac{1}{\delta}.$$

Choosing $\delta=1/(n^2|\mathcal{Y}|)$ for $n\ge2$, we conclude that

$$\mathbb{E}\left[\mathcal{R}_n(\mathcal{Q},\mathbf{p},\mathbf{y})\right]\le\mathbb{E}\left[\mathcal{R}_n(\mathcal{Q}^\delta,\mathbf{p}^\delta,\mathbf{y})\right]+12\log(n|\mathcal{Y}|).$$

Hence in order to prove Theorem 5.2, we only need to prove that

$$\mathbb{E}_{\mathbf{y}\sim\mathbf{p}}\left[\mathcal{R}_n(\mathcal{Q},\mathbf{p},\mathbf{y})\right]=\tilde{\mathcal{O}}\left(\inf_{\gamma>\delta>0}\left\{n\delta\sqrt{|\mathcal{Y}|}+\sqrt{n|\mathcal{Y}|}\int_\delta^\gamma\sqrt{\mathcal{H}_{\mathsf{sq}}(\mathcal{Q},\alpha,n)}d\alpha+\mathcal{H}_{\mathsf{sq}}(\mathcal{Q},\gamma,n)\right\}\right)\tag{D.22}$$

holds for any $\mathbf{p}\in\Delta_n(\mathcal{Y}^n)$ and $\mathcal{Q}\subseteq\Delta_n(\mathcal{Y}^n)$, where $\Delta_n(\mathcal{Y}^n)$ is defined in eq. (D.4). To prove this, we first notice that according to Theorem D.10, for $\mathbf{p}\in\Delta_n(\mathcal{Y}^n)$ and $\mathcal{Q}\subseteq\Delta_n(\mathcal{Y}^n)$, we have

$$\mathbb{E}_{\mathbf{y}\sim\mathbf{p}}\mathcal{R}_n(\mathcal{Q},\mathbf{p},\mathbf{y})$$

$$\leq \mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\sum_{t=1}^{n}\varepsilon_t\zeta\left(\frac{q_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)-\frac{1}{4\log(n|\mathcal{Y}|)}\zeta\left(\frac{q_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)^2\right]$$
$$+\mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\sum_{t=1}^{n}(-\varepsilon_t)\zeta\left(\frac{q_t(z_t\mid\mathbf{w})}{p_t(z_t\mid\mathbf{w})}\right)-\frac{1}{4\log(n|\mathcal{Y}|)}\zeta\left(\frac{q_t(z_t\mid\mathbf{w})}{p_t(z_t\mid\mathbf{w})}\right)^2\right],$$

In the following, we will upper bound the right hand side in the above formula. For convenience, we only provide upper bounds to the first term in the right hand side. The upper bound to the second term in the right hand side can be obtained similarly.

Next, we choose N positive real numbers $\alpha_1<\cdots<\alpha_N$ (values to be specified later). We let $\mathcal{V}(\alpha_i)$ be a smallest sequential square-root cover (as per Theorem 5.1) at scale α_i. For any $\mathbf{q}\in\mathcal{Q}$, $\mathbf{w},\mathbf{y}\in\mathcal{Y}^n$ and $t\in[n]$, we let

$$\mathbf{v}'[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_N]=\operatorname*{arg\,min}_{\mathbf{v}\in\mathcal{V}(\alpha_N)}\left\{\sum_{t=1}^{n}\left(\zeta\left(\frac{q_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)-\zeta\left(\frac{v_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)\right)^2\right\}. \tag{D.23}$$

According to Theorem D.11, there exists some $\mathbf{v}[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_N]\in\mathcal{V}(\alpha_N)\cup\{\mathbf{p}\}$ such that

$$\sum_{t=1}^{n}\left(\zeta\left(\frac{q_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)-\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_N](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)\right)^2$$
$$\leq\sum_{t=1}^{n}\left(\zeta\left(\frac{q_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)-\zeta\left(\frac{v'_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_N](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)\right)^2, \tag{D.24}$$

and

$$\sum_{t=1}^{n}\zeta\left(\frac{q_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)^2\geq\frac{1}{4}\sum_{t=1}^{n}\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_N](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)^2. \tag{D.25}$$

both hold. For $1\leq i\leq N-1$, we use $\mathcal{V}(\alpha_i)$ to denote the sequential cover of $\mathcal{Q}$ at scale α_i. For every $\mathbf{q}\in\mathcal{Q}$ and $\mathbf{w},\mathbf{y}\in\mathcal{Y}^n$, we let

$$\mathbf{v}[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_i]=\operatorname*{arg\,min}_{\mathbf{v}\in\mathcal{V}(\alpha_i)}\left\{\sum_{t=1}^{n}\left(\zeta\left(\frac{q_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)-\zeta\left(\frac{v_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)\right)^2\right\}. \tag{D.26}$$

Then according to Theorem D.13, we have

$$\mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\sum_{t=1}^{n}\varepsilon_t\left\{\zeta\left(\frac{q_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)-\frac{1}{4\log(n|\mathcal{Y}|)}\zeta\left(\frac{q_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)^2\right\}\right]$$
$$\leq\mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\left\{\sum_{t=1}^{n}\varepsilon_t\left\{\zeta\left(\frac{q_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)-\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_1](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)\right\}\right\}\right]$$
$$+\mathbb{E}\left[\sum_{i=1}^{N-1}\sup_{\mathbf{q}\in\mathcal{Q}}\left\{\sum_{t=1}^{n}\varepsilon_t\left\{\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_i](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)-\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_{i+1}](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)\right\}\right\}\right]$$
$$+\mathbb{E}\left[\sup_{\mathbf{v}\in\mathcal{V}(\alpha_N)\cup\{\mathbf{p}\}}\left\{\sum_{t=1}^{n}\left\{\varepsilon_t\zeta\left(\frac{v_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)-\frac{1}{16\log(n|\mathcal{Y}|)}\zeta\left(\frac{v_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)^2\right\}\right\}\right]. \tag{D.27}$$

In the following, we upper bound the three terms in eq. (D.27) respectively.

We let

$$\bar{\mathbf{v}}[\mathbf{q},\mathbf{p},\mathbf{w},\alpha_N] = \arg\min_{\mathbf{v}\in\mathcal{V}(\alpha_N)} \left\{ \max_{t\in[n]} \max_{y\in\mathcal{Y}} \left| \sqrt{q_t(y \mid \mathbf{w})} - \sqrt{v_t(y \mid \mathbf{w})} \right| \right\}. \tag{D.28}$$

Then for $(\boldsymbol{\varepsilon},\mathbf{w},\mathbf{y},\mathbf{z}) \sim \odot\mathbf{p}$, we have

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}} \sum_{t=1}^{n} \left(\zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_N](y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)\right)^2\right] \\
&\overset{(i)}{\leq} \mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}} \sum_{t=1}^{n} \left(\zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \zeta\left(\frac{v'_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_N](y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)\right)^2\right] \\
&\overset{(ii)}{\leq} \mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}} \sum_{t=1}^{n} \left(\zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \zeta\left(\frac{\bar{v}_t[\mathbf{q},\mathbf{p},\mathbf{w},\alpha_N](y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)\right)^2\right] \\
&\overset{(iii)}{\leq} 4\mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}} \sum_{t=1}^{n} \left(\sqrt{\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}} - \sqrt{\frac{\bar{v}_t[\mathbf{q},\mathbf{p},\mathbf{w},\alpha_N](y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}}\right)^2\right] \\
&\overset{(iv)}{\leq} 4\mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}} \sum_{t=1}^{n} \frac{\alpha_N^2}{p_t(y_t \mid \mathbf{w})}\right] = 4\alpha_N^2 \cdot \mathbb{E}\left[\sum_{t=1}^{n} \frac{1}{p_t(y_t \mid \mathbf{w})}\right] \\
&\overset{(v)}{\leq} 4n\alpha_N^2|\mathcal{Y}|,
\end{aligned} \tag{D.29}$$

where (i) uses eq. (D.24), (ii) uses the definition of $\mathbf{v}'[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_N]$ in eq. (D.23), (iii) uses the Lipschitz property of ζ function in Theorem D.8, (iv) uses the definition of $\bar{\mathbf{v}}$ in eq. (D.28) and Theorem 5.1: for fixed $\mathbf{p},\mathbf{q}$ and $\mathbf{w}$, for any $t \in [n]$ and $y_t \in \mathcal{Y}$,

$$\begin{aligned}
&\left|\sqrt{q_t(y_t \mid \mathbf{w})} - \sqrt{\bar{v}_t[\mathbf{q},\mathbf{p},\mathbf{w},\alpha_N](y_t \mid \mathbf{w})}\right| \\
&\quad= \min_{\mathbf{v}\in\mathcal{V}(\alpha_N)} \max_{t\in[n]} \max_{y\in\mathcal{Y}} \left|\sqrt{q_t(y_t \mid \mathbf{w})} - \sqrt{v_t(y_t \mid \mathbf{w})}\right| \\
&\quad\leq \sup_{\mathbf{q}\in\mathcal{Q}} \max_{\mathbf{w}\in\mathcal{Y}^n} \min_{\mathbf{v}\in\mathcal{V}(\alpha_N)} \max_{t\in[n]} \max_{y\in\mathcal{Y}} \left|\sqrt{q_t(y_t \mid \mathbf{w})} - \sqrt{v_t(y_t \mid \mathbf{w})}\right| \leq \alpha_N,
\end{aligned}$$

and finally (v) uses the fact that for any $t \in [n]$, conditionally on $w_{1:t-1}$, the distribution of y_t is $p_t(\cdot \mid \mathbf{w})$, hence

$$\mathbb{E}\left[\frac{1}{p_t(y_t \mid \mathbf{w})}\right] = \mathbb{E}\left[\sum_{y\in\mathcal{Y}} p_t(y \mid \mathbf{w}) \cdot \frac{1}{p_t(y \mid \mathbf{w})}\right] = |\mathcal{Y}|.$$

Then similar to eq. (D.29) (the definition of $\mathbf{v}[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_i]$ for $1 \leq i \leq N-1$ is similar to the definition of $\mathbf{v}'[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_N]$, hence the following inequality can be obtained by starting from the second line of eq. (D.29)), we can show that with $(\boldsymbol{\varepsilon},\mathbf{w},\mathbf{y},\mathbf{z}) \sim \odot\mathbf{p}$, for any $i \in [N-1]$,

$$\mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}} \sum_{t=1}^{n} \left(\zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_i](y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)\right)^2\right] \leq 4n\alpha_i^2|\mathcal{Y}|. \tag{D.30}$$

We are now ready to provide upper bounds for the three terms in eq. (D.27). For the first term in eq. (D.27), we have

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\left\{\sum_{t=1}^{n}\varepsilon_t\left\{\zeta\left(\frac{q_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)-\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_1](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)\right\}\right\}\right]\\
&\overset{(i)}{\leq}\sqrt{n}\cdot\mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\left\{\sqrt{\sum_{t=1}^{n}\left(\zeta\left(\frac{q_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)-\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_1](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)\right)^2}\right\}\right]\\
&\overset{(ii)}{\leq}\sqrt{n}\cdot\sqrt{\mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\sum_{t=1}^{n}\left(\zeta\left(\frac{q_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)-\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_1](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)\right)^2\right]}\\
&\overset{(iii)}{\leq}2n\alpha_1\sqrt{|\mathcal{Y}|},
\end{aligned}\tag{D.31}$$

where (i) uses Cauchy-Schwarz inequality, (ii) uses Jensen's inequality and (iii) uses eq. (D.30).

We next analyze the second term in eq. (D.27). Notice that for any $i\in[N-1]$ and $\lambda>0$, we can decompose the second term in eq. (D.27) as

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\left\{\sum_{t=1}^{n}\varepsilon_t\left\{\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_i](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)-\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_{i+1}](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)\right\}\right\}\right]\\
&\leq\mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\left\{\sum_{t=1}^{n}\varepsilon_t\left\{\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_i](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)-\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_{i+1}](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)\right\}\right.\right.\\
&\qquad\left.\left.-\lambda\cdot\left\{\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_i](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)-\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_{i+1}](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)\right\}^2\right\}\right]\\
&\qquad+\lambda\cdot\mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\sum_{t=1}^{n}\left\{\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_i](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)-\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_{i+1}](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)\right\}^2\right]
\end{aligned}\tag{D.32}$$

For fixed $\mathbf{p}$ and $i\in[N-1]$, we define set $\mathcal{U}_i$ as

$$\mathcal{U}_i\coloneqq\{\mathbf{u}:\mathbf{u}=\mathbf{u}[\mathbf{v}^i,\mathbf{v}^{i+1}]\text{ for some }\mathbf{v}^i\in\mathcal{V}(\alpha_i)\text{ and }\mathbf{v}^{i+1}\in\mathcal{V}(\alpha_{i+1})\cup\{\mathbf{p}\}\},$$

where for $\mathbf{v}^i\in\mathcal{V}(\alpha_i)$ and $\mathbf{v}^{i+1}\in\mathcal{V}(\alpha_{i+1})\cup\{\mathbf{p}\}$, the element $\mathbf{u}[\mathbf{v}^i,\mathbf{v}^{i+1}]$ is defined as $(u_{1:n})$ with $u_t=u_t(\cdot\mid\cdot):\mathcal{Y}\times\mathcal{Y}^{t-1}\to\mathbb{R}$ defined as

$$u_t(y_t\mid\mathbf{w})=\zeta\left(\frac{v_t^i(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)-\zeta\left(\frac{v_t^{i+1}(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right),\qquad\forall\ \mathbf{w},\mathbf{y}\in\mathcal{Y}^n\text{ and }t\in[n].$$

Then we have $|\mathcal{U}|=|\mathcal{V}(\alpha_i)|\cdot(|\mathcal{V}(\alpha_{i+1})|+1)$, and we can further upper bound eq. (D.32) by

$$\mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\left\{\sum_{t=1}^{n}\varepsilon_t\left\{\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_i](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)-\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_{i+1}](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)\right\}\right\}\right]$$

$$\leq \mathbb{E}\left[\sup_{\mathbf{u}\in\mathcal{U}}\sum_{t=1}^{n}\left\{\varepsilon_t u_t(y_t \mid \mathbf{w}) - \lambda\cdot u_t(y_t \mid \mathbf{w})^2\right\}\right]$$
$$+\lambda\cdot\mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\sum_{t=1}^{n}\left\{\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_i](y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_{i+1}](y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)\right\}^2\right] \tag{D.33}$$

For the first term in eq. (D.33), we adopt Theorem D.1 with

$$s_t = (w_{1:t-1}, y_t), \quad \mathcal{G}_t = \sigma(y_{1:t}, z_{1:t}, w_{1:t-1}) \quad \text{and} \quad a_t(s_t) = u_t(y_t \mid \mathbf{w}),$$

it is easy to see that $\mathbb{E}[\varepsilon_t \mid \mathcal{G}_t] = 0$, and $\sigma(s_t) \subseteq \mathcal{G}_t$. Hence we have

$$\mathbb{E}\left[\sup_{\mathbf{u}\in\mathcal{U}}\sum_{t=1}^{n}\left\{\varepsilon_t u_t(y_t \mid \mathbf{w}) - \lambda\cdot u_t(y_t \mid \mathbf{w})^2\right\}\right]$$
$$\leq \frac{\log|\mathcal{U}|}{2\lambda} \leq \frac{\log(|\mathcal{V}(\alpha_i)|(|\mathcal{V}(\alpha_{i+1})|+1))}{2\lambda} \leq \frac{2\log|\mathcal{V}(\alpha_i)|}{\lambda}, \tag{D.34}$$

where the last inequality uses the fact that $\alpha_i \leq \alpha_{i+1}$ hence $|\mathcal{V}(\alpha_{i+1})| \leq |\mathcal{V}(\alpha_i)|$. For the second term in eq. (D.33), we have

$$\lambda\cdot\mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\sum_{t=1}^{n}\left\{\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_i](y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_{i+1}](y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)\right\}^2\right]$$
$$\overset{(i)}{\leq} 2\lambda\cdot\mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\sum_{t=1}^{n}\left(\zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_i](y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)\right)^2\right]$$
$$+ 2\lambda\cdot\mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\sum_{t=1}^{n}\left(\zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_{i+1}](y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)\right)^2\right]$$
$$\overset{(ii)}{\leq} 8\lambda n\alpha_i^2|\mathcal{Y}| + 8\lambda n\alpha_{i+1}^2|\mathcal{Y}| \overset{(iii)}{\leq} 16\lambda n\alpha_{i+1}^2|\mathcal{Y}|, \tag{D.35}$$

where (i) we uses Cauchy-Schwarz inequality, (ii) we uses eq. (D.29) and eq. (D.30), and (iii) uses $\alpha_i \leq \alpha_{i+1}$. Finally we choose $\lambda = \sqrt{\frac{\log|\mathcal{V}(\alpha_i)|}{8n\alpha_{i+1}^2|\mathcal{Y}|}}$. Then bringing eq. (D.35) and eq. (D.34) into eq. (D.33), we obtain

$$\mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\left\{\sum_{t=1}^{n}\varepsilon_t\left\{\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_i](y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_{i+1}](y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)\right\}\right\}\right]$$
$$\leq \frac{2\log|\mathcal{V}(\alpha_i)|}{\lambda} + 16\lambda n\alpha_{i+1}^2|\mathcal{Y}| \leq 8\alpha_{i+1}\sqrt{2n|\mathcal{Y}|\log|\mathcal{V}(\alpha_i)|}. \tag{D.36}$$

Remark D.14. *Let us briefly discuss the novel and key aspect of the approach used to upper bound the left-hand side of* (D.36)*. In the classical case when the coefficients are non-random, one simply observes that the supremum is a maximum over a finite collection. Unfortunately, here the squared increments are themselves random and only small in expectation, due to the*

presence of the $p_t(y_t \mid \mathbf{w})$ term in the denominator. The key technical observation here is that one can alternatively work with the offset process (D.34), which can be controlled for any predictable coefficients irrespective of their magnitude*, as well as the expected squared distance between the coefficients in (D.35). We believe that this technique, which is summarized in Theorem D.2, will be useful beyond this paper.*

Finally, we analyze the last term (offset term) in eq. (D.27). We let the filtration $\mathcal{G}_t = \sigma(y_{1:t}, z_{1:t}, \varepsilon_{1:t-1})$ for $t \in [n]$. Then we have $\mathbb{E}[\varepsilon_t \mid \mathcal{G}_t] = 0$, and according to the process of getting $w_{1:n}$, $\sigma(w_{1:t-1}) \subseteq \mathcal{G}_t$. We let s_t in Theorem D.1 to be $(w_{1:t-1}, y_t)$, and

$$\mathcal{A} = \left\{ \mathbf{a} = (a_1, \cdots, a_n) \;\middle|\; a_t(w_{1:t-1}, y_t) = \zeta\left(\frac{v_t(y_t \mid w_{1:t-1})}{p_t(y_t \mid w_{1:t-1})}\right) \text{ for any } t \in [n] \text{ for some } \mathbf{v} \in \mathcal{V}(\alpha_N) \cup \{\mathbf{p}\} \right\}$$

Applying Theorem D.1 with $\lambda = \frac{1}{16\log(n|\mathcal{Y}|)}$, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\mathbf{v}\in\mathcal{V}(\alpha_N)\cup\{\mathbf{p}\}} \left\{\sum_{t=1}^n \left\{\varepsilon_t \zeta\left(\frac{v_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \frac{1}{16\log(n|\mathcal{Y}|)}\zeta\left(\frac{v_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)^2\right\}\right\}\right] \\
&\quad \le 8\log(n|\mathcal{Y}|)\cdot\log(|\mathcal{V}(\alpha_N)|+1). \qquad \text{(D.37)}
\end{aligned}$$

Finally, we specify the scales $\alpha_i = 2^{i-l}$ for some positive integer $l \ge N$. Bringing eq. (D.31), eq. (D.36) and eq. (D.37) into eq. (D.27), we obtain that

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}} \sum_{t=1}^n \varepsilon_t \left\{\zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \frac{1}{4\log(n|\mathcal{Y}|)}\zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)^2\right\}\right] \\
&\le 2n\alpha_1\sqrt{|\mathcal{Y}|} + \sum_{i=1}^{N-1} 8\alpha_{i+1}\sqrt{2n|\mathcal{Y}|}\cdot\sqrt{\log|\mathcal{V}(\alpha_i)|} + 8\log(n|\mathcal{Y}|)\cdot(\log|\mathcal{V}(\alpha_N)|+1) \\
&\overset{(i)}{\lesssim} n\alpha_1\sqrt{|\mathcal{Y}|} + \sum_{i=1}^{N-1}\alpha_{i+1}\sqrt{n|\mathcal{Y}|}\cdot\sqrt{\mathcal{H}_{\mathsf{sq}}(\mathcal{Q},\alpha_i,n)} + \log(n|\mathcal{Y}|)\cdot\mathcal{H}_{\mathsf{sq}}(\mathcal{Q},\alpha_N,n) \\
&\overset{(ii)}{\lesssim} n\alpha_1\sqrt{|\mathcal{Y}|} + \sum_{i=1}^{N-1}(\alpha_i-\alpha_{i-1})\sqrt{n|\mathcal{Y}|}\cdot\sqrt{\mathcal{H}_{\mathsf{sq}}(\mathcal{Q},\alpha_i,n)} + \log(n|\mathcal{Y}|)\cdot\mathcal{H}_{\mathsf{sq}}(\mathcal{Q},\alpha_N,n) \\
&\overset{(iii)}{\le} \frac{n\sqrt{|\mathcal{Y}|}}{2^{-l}} + \sqrt{n|\mathcal{Y}|}\int_{2^{-l}}^{2^{N-l}} \sqrt{\mathcal{H}_{\mathsf{sq}}(\mathcal{Q},\alpha,n)}d\alpha + \log(n|\mathcal{Y}|)\cdot\mathcal{H}_{\mathsf{sq}}(\mathcal{Q},2^{N-l},n),
\end{aligned}$$

where (i) uses the definition of the covering $\mathcal{V}(\alpha_i)$ we have $\log(|\mathcal{V}(\alpha_i)|+1) \lesssim \mathcal{H}_{\mathsf{sq}}(\mathcal{Q},\alpha,n)$ for any $\mathbf{p} \in \Delta(\mathcal{Y}^n)$, (ii) uses $\alpha_{i+1} = 2\alpha_i = 4\alpha_{i-1}$, and (iii) uses the fact that for any $\alpha_{i-1} \le \alpha \le \alpha_i$,

$$\mathcal{H}_{\mathsf{sq}}(\mathcal{Q},\alpha_i,n) \le \mathcal{H}_{\mathsf{sq}}(\mathcal{Q},\alpha,n).$$

For $0 < \delta < \gamma \le 1$, letting $l = \log_2(1/\delta)$ and $N = l + \log_2(\gamma)$, and according to Theorem D.10, we obtain that for any $\mathbf{p} \in \Delta_n(\mathcal{Y}^n)$,

$$\mathbb{E}_{\mathbf{y}\sim\mathbf{p}}\left[\mathcal{R}_n(\mathcal{Q},\mathbf{p},\mathbf{y})\right] \lesssim n\delta\sqrt{|\mathcal{Y}|} + \sqrt{n|\mathcal{Y}|}\int_\delta^\gamma \sqrt{2\mathcal{H}_{\mathsf{sq}}(\mathcal{Q},\alpha,n)}d\alpha + \log(n|\mathcal{Y}|)\cdot\mathcal{H}_{\mathsf{sq}}(\mathcal{Q},\gamma,n),$$

hence eq. (D.22) is verified. □

D.3 Missing Proofs in section 5.3

D.3.1 Missing Proofs in section 5.3.1

Proof of Theorem 5.4. We write the proof for

$$\phi(y_{1:n}, x_{1:n}) = \inf_{f\in\mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t),$$

but it will be clear from the following that this particular structure is not used. When ℓ is convex with respect to its first argument, we can write

$$\begin{aligned}
\mathcal{R}_n(\mathcal{F}) &= \left\{ \sup_{x_t} \inf_{\widehat{p}_t} \sup_{y_t} \right\}_{t=1}^{n} \left[\sum_{t=1}^{n} \ell(\widehat{p}_t, y_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) \right] \\
&= \left\{ \sup_{x_t} \inf_{\widehat{p}_t} \sup_{p_t} \mathbb{E}_{y_t\sim p_t} \right\}_{t=1}^{n} \left[\sum_{t=1}^{n} \ell(\widehat{p}_t, y_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) \right] \\
&\overset{(i)}{=} \left\{ \sup_{x_t} \sup_{p_t} \inf_{\widehat{p}_t} \mathbb{E}_{y_t\sim p_t} \right\}_{t=1}^{n} \left[\sum_{t=1}^{n} \ell(\widehat{p}_t, y_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) \right] \\
&= \left\{ \sup_{x_t} \sup_{p_t} \mathbb{E}_{y_t\sim p_t} \right\}_{t=1}^{n} \left[\sum_{t=1}^{n} \inf_{\widehat{p}_t} \mathbb{E}_{y_t\sim p_t} \ell(\widehat{p}_t, y_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) \right] \\
&\overset{(ii)}{=} \sup_{\mathbf{x}} \left\{ \sup_{p_t} \mathbb{E}_{y_t\sim p_t} \right\}_{t=1}^{n} \left[\sum_{t=1}^{n} \inf_{\widehat{p}_t} \mathbb{E}_{y_t\sim p_t} \ell(\widehat{p}_t, y_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t(\mathbf{y})), y_t) \right] \\
&= \sup_{\mathbf{x}} \left\{ \sup_{p_t} \inf_{\widehat{p}_t} \mathbb{E}_{y_t\sim p_t} \right\}_{t=1}^{n} \left[\sum_{t=1}^{n} \ell(\widehat{p}_t, y_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t(\mathbf{y})), y_t) \right] \\
&\overset{(iii)}{=} \sup_{\mathbf{x}} \left\{ \inf_{\widehat{p}_t} \sup_{y_t} \right\}_{t=1}^{n} \left[\sum_{t=1}^{n} \ell(\widehat{p}_t, y_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t(\mathbf{y})), y_t) \right] \\
&= \sup_{\mathbf{x}} \mathcal{R}_n(\mathcal{F}\circ\mathbf{x}),
\end{aligned}$$

where (i) and (iii) use the minimax theorem for convex-concave functions [220, Theorem 1.(ii)], and also the fact that ℓ is convex with respect to its first argument, and (ii) uses the fact that interleaving the supremum of x_t and expectation over y_t is equivalent to taking the supremum over trees $\mathbf{x}$ first (or, skolemization). □

Proof of Theorem 5.6. For a given function $\mathcal{F} \subseteq [0,1]^{\mathcal{X}}$ and depth-n $\mathcal{X}$-valued tree $\mathbf{x}$, we define a class $\mathcal{F}(\mathbf{x}) = \{f(\mathbf{x}) : f \in \mathcal{F}\} \subseteq \Delta(\{0,1\}^n)$ of joint distributions over $\{0,1\}^n$ as follows: for $\mathbf{y} = (y_{1:n}) \in \{0,1\}^n$, the probability of joint distribution $f(\mathbf{x})$ takes value $\mathbf{y}$ equals to

$$f(\mathbf{x})(\mathbf{y}) = \prod_{t=1}^{n} f(\mathbf{x})_t(y_t \mid \mathbf{y}),$$

and

$$f(\mathbf{x})_t(y_t \mid \mathbf{y}) = \begin{cases} f(x_t(\mathbf{y})) & \text{if } y_t = 1, \\ 1 - f(x_t(\mathbf{y})) & \text{if } y_t = 0. \end{cases} \tag{D.38}$$

According to Theorem 5.4 (see also [18, Lemma 6], [112] and [17, Eq. 27]), we can write

$$\mathcal{R}_n(\mathcal{F}) = \sup_{\mathbf{x},\mathbf{p}} \mathbb{E}_{\mathbf{y}\sim\mathbf{p}} \mathcal{R}_n(\mathcal{F}(\mathbf{x}), \mathbf{p}, \mathbf{y}) = \sup_{\mathbf{x}} \left[\mathcal{R}_n(\mathcal{F}(\mathbf{x}))\right],$$

where $\mathcal{R}_n(\cdot, \mathbf{p}, \mathbf{y})$ is defined in eq. (D.2), and $\mathcal{R}_n(\mathcal{F}(\mathbf{x}))$ is the Shtarkov sum eq. (5.3) for joint distribution class $\mathcal{F}(\mathbf{x})$.

In order to prove Theorem 5.6, we only need to verify that $\mathcal{H}_{\mathsf{sq}}(\mathcal{F}, \alpha, n, \mathbf{x}) \le \mathcal{H}_{\mathsf{sq}}(\mathcal{F}(\mathbf{x}), n, \alpha)$ for any tree $\mathbf{x}$. In the following, we verify this by showing that the sequential square-root covering of $\mathcal{F} \circ \mathbf{x}$ defined in Theorem 5.5 can form a sequential square-root covering of $\mathcal{F}(\mathbf{x})$ defined in Theorem 5.1 of the same size.

Suppose $\mathcal{V}(\alpha)$ is the sequential square-root covering of $\mathcal{F} \circ \mathbf{x}$ at scale α defined in Theorem 5.5. We define $\mathcal{U}(\alpha) \subseteq \Delta(\{0,1\}^n)$ from $\mathcal{V}$:

$$\mathcal{U}(\alpha) = \left\{ \mathbf{u}[\mathbf{v}] \in \Delta(\{0,1\}^n), \mathbf{v} \in \mathcal{V}(\alpha) : \mathbf{u}(\mathbf{y}) = \prod_{t=1}^{n} u_t[\mathbf{v}](y_t \mid \mathbf{y}), \forall \mathbf{y} \in \{0,1\}^n \right\},$$

where

$$u_t[\mathbf{v}](y_t \mid \mathbf{y}) \coloneqq \begin{cases} v_t(\mathbf{y}) & \text{if } y_t = 1, \\ 1 - v_t(\mathbf{y}) & \text{if } y_t = 0. \end{cases}$$

Then we have $|\mathcal{U}(\alpha)| = |\mathcal{V}(\alpha)|$. And according to Theorem 5.5 we have for any $f \in \mathcal{F}$, and $\mathbf{w} \in \{0,1\}^n$, there exists $\mathbf{u} \in \mathcal{U}$ such that for any $t \in [n]$,

$$\max_{y\in\{0,1\}} \left| \sqrt{u_t(y \mid \mathbf{w})} - \sqrt{f(\mathbf{x})_t(y \mid \mathbf{w})} \right| \le \alpha.$$

Therefore, $\mathcal{U}(\alpha)$ is a sequential square-root covering of $\mathcal{F}(\mathbf{x})$ according to Theorem 5.1. Noticing that $|\mathcal{U}(\alpha)| = |\mathcal{V}(\alpha)|$, Theorem 5.6 directly follows from Theorem 5.2. □

Proof of Theorem 5.7. Theorem 5.7 follows from Theorem 5.6 after replacing $\mathcal{H}_{\mathsf{sq}}(\mathcal{F}, \alpha, n, \mathbf{x})$ with $\tilde{\mathcal{O}}(\alpha^{-p})$ and the following choices of γ and δ:

$$(\gamma, \delta) = \begin{cases} \left(n^{-\frac{1}{p+2}}, n^{-1}\right) & \text{if } 0 \le p \le 2, \\ \left(1, n^{-\frac{1}{p}}\right) & \text{if } p > 2. \end{cases} \tag{D.39}$$

□

D.3.2 Missing Proofs in section 5.3.2

Proof of Theorem 5.8. This proposition directly follows from the following inequality: for any $p, q \in [0,1]$ with $p \in [\delta, 1-\delta]$,

$$\max\left\{ \left|\sqrt{p} - \sqrt{q}\right|, \left|\sqrt{1-p} - \sqrt{1-q}\right| \right\} \le \frac{|p-q|}{\sqrt{\delta}}.$$

In fact, we have

$$\begin{aligned}&\max\left\{\left|\sqrt{p}-\sqrt{q}\right|,\left|\sqrt{1-p}-\sqrt{1-q}\right|\right\}\\&=|p-q|\cdot\max\left\{\frac{1}{\sqrt{p}+\sqrt{q}},\frac{1}{\sqrt{1-p}+\sqrt{1-q}}\right\}\le\frac{|p-q|}{\sqrt{\delta}}.\end{aligned}$$

□

Proof of Theorem 5.9. This proposition is a direct corollary of the standard inequality, e.g. [221, (7.22)], which shows that the squared Hellinger distance and TV distance satisfy the bound:

$$H(p,q)^2\le 2D_{\mathrm{TV}}(p,q).$$

□

D.4 Missing Proofs in section 5.3.3

In this section, we will present the formal proof to Theorem 5.10 and Theorem 5.11.

D.4.1 Proof of Theorem 5.10

To prove Theorem 5.10, we define a dimension of function classes which characterizes the difficulty of sequential learning with the class. We will relate both the sequential square-root entropy and minimax regret to this dimension of the function class.

First, we define distance h between two real numbers in $[0,1]$:

$$h(a,b)=\max\left\{\left|\sqrt{a}-\sqrt{b}\right|,\left|\sqrt{1-a}-\sqrt{1-b}\right|\right\},\qquad\forall a,b\in[0,1].\tag{D.40}$$

Definition D.15. *Suppose $\mathcal{F}\in[0,1]^{\mathcal{X}}$ is a function class and $0<\beta<\alpha$. An $\mathcal{X}$-valued depth-d tree $\mathbf{x}$ is said to be (α,β)-shattered by $\mathcal{F}$ distance if there exists a $[\beta,1-\beta]\times[\beta,1-\beta]$-valued depth-$d$ tree $\mathbf{s}$ such that: for any path $\mathbf{y}=(y_{1:d})\in\{0,1\}^d$, $s_t(\mathbf{y})=(s_t(\mathbf{y})[0],s_t(\mathbf{y})[1])$ with $s_t(\mathbf{y})[0]<s_t(\mathbf{y})[1]$, and also there exists $f^{\mathbf{y}}\in\mathcal{F}$ such that*

$$|f^{\mathbf{y}}(x_t(\mathbf{y}))-s_t(\mathbf{y})[y_t]|<\beta\quad\text{and}\quad h\left(s_t(\mathbf{y})[0],s_t(\mathbf{y})[1]\right)>\alpha,\qquad\forall t\in[d].$$

The dimension $\mathfrak{D}(\mathcal{F},\alpha,\beta)$ is defined to be the largest d such that there exists a depth-d tree $\mathbf{x}$ which is (α,β)-shattered by $\mathcal{F}$.

The following proposition relates the dimension $\mathfrak{D}(\mathcal{F},\alpha,\beta)$ to the sequential square-root entropy.

Proposition D.16. *For any class $\mathcal{X}$, function class $\mathcal{F}\in[0,1]^{\mathcal{X}}$, positive integer n and $\alpha>0$, we have*

$$\sup_{\mathbf{x}}\mathcal{H}_{\mathsf{sq}}(\mathcal{F},\alpha+\sqrt{2\beta},n,\mathbf{x})\le\mathfrak{D}(\mathcal{F},\alpha,\beta)\log\left(\frac{en}{\beta}\right),$$

where the supremum is over all depth-n $\mathcal{X}$-valued tree $\mathbf{x}$.

The following proposition relates the dimension $\mathfrak{D}(\mathcal{F}, \alpha, \beta)$ to the minimax regret $\mathcal{R}_n(\mathcal{F})$.

Proposition D.17. *Suppose the function class $\mathcal{F} \subseteq [0,1]^{\mathcal{X}}$ satisfies $\mathfrak{D}(\mathcal{F}, \alpha, \alpha^4/16) = \tilde{\Omega}\left(\alpha^{-p}\right)$. Then for any positive integer n,*

$$\mathcal{R}_n(\mathcal{F}) = \tilde{\Omega}\left(n^{\frac{p}{p+2}}\right).$$

With the above two propositions of the dimension $\mathfrak{D}(\mathcal{F}, \alpha, \beta)$, we are ready to prove Theorem 5.10.

Proof of Theorem 5.10. Suppose class $\mathcal{F}$ satisfies $\sup_{\mathbf{x}} \mathcal{H}_{\mathsf{sq}}(\mathcal{F}, \alpha, n, \mathbf{x}) = \tilde{\Omega}\left(\alpha^{-p}\right)$. Then according to Theorem D.16, we have

$$\mathfrak{D}(F, \alpha, \alpha^4/16) \geq \sup_{\mathbf{x}} \mathcal{H}_{\mathsf{sq}}(\mathcal{F}, \alpha + \alpha^2/(2\sqrt{2}), n, \mathbf{x}) \cdot \log^{-1}\left(\frac{16en}{\alpha^4}\right) = \tilde{\Omega}\left(\alpha^{-p}\right).$$

Hence according to Theorem D.17, we have

$$\mathcal{R}_n(\mathcal{F}) = \tilde{\Omega}\left(n^{\frac{p}{p+2}}\right).$$

□

The following two subsections will be devoted to the proof of Theorem D.16 and Theorem D.17. A more general treatment of this approach, including the non-sequential analogue, will appear in the companion paper [110].

Proof of Theorem D.16

Before proving Theorem D.16, we first prove a similar results for sets of discrete-valued function classes. Suppose $\beta \in (0,1)$ satisfies $M = 1/(2\beta)$ is an positive integer. We define set:

$$U_\beta = \{\beta, 3\beta, 5\beta, \cdots, (2M-1)\cdot\beta\} \subseteq [0,1].$$

And we further define the dimension $\mathfrak{D}(\mathcal{F}, \alpha)$ for the discrete-valued function class $\mathcal{F}$ which contains functions mapping $\mathcal{X}$ into the set U_β.

Definition D.18. *Fix real number $\beta \in (0,1)$ which satisfies $1/(2\beta)$ is a positive integer. For function $\mathcal{F} \subseteq (U_\beta)^{\mathcal{X}}$ and real number $\alpha > 0$, a depth-d $\mathcal{X}$-valued $\mathbf{x}$ is said to be shattered by $\mathcal{F}$ at scale α, if there exists a depth-d $(\mathcal{U}_\beta \times \mathcal{U}_\beta)$-valued tree $\mathbf{s}$ such that: for any $\mathbf{y} \in \{0,1\}^d$, $s_t(\mathbf{y}) = (s_t(\mathbf{y})[0], s_t(\mathbf{y})[1])$ satisfies $s_t(\mathbf{y})[0] < s_t(\mathbf{y})[1]$, and for any $t \in [d]$, $h(s_t(\mathbf{y})[0], s_t(\mathbf{y})[1]) > \alpha$ (where h is defined in (D.40)), and for any $\mathbf{y} \in \{0,1\}^d$, there exists $f^{\mathbf{y}} \in \mathcal{F}$ such that $f^{\mathbf{y}}(x_t(\mathbf{y})) = s_t(\mathbf{y})[y_t]$ holds for all $t \in [d]$.*

The dimension $\mathfrak{D}(\mathcal{F}, \alpha)$ of $\mathcal{F}$ is defined to be the largest d such that there exists a depth-d $\mathcal{X}$-valued tree $\mathbf{x}$ shattered by $\mathcal{F}$ at scale α.

The following lemma indicates that for discrete-valued function set $\mathcal{F}$, the sequential square-root covering number of class $\mathcal{F}$ can be bounded by the dimension $\mathfrak{D}(\mathcal{F}, \alpha)$ of class $\mathcal{F}$.

Lemma D.19. *For function class $\mathcal{F} : \mathcal{X} \to U_\beta$, then for any depth-$n$ $\mathcal{X}$-valued tree $\mathbf{x}$, we have*

$$\mathcal{N}_{\mathsf{sq}}(\mathcal{F} \circ \mathbf{x}, \alpha, n) \le \left(\frac{en}{\beta}\right)^{\mathfrak{D}(\mathcal{F},\alpha)}$$

Proof of Theorem D.19. For any $\beta > 0$ such that $1/(2\beta)$ is an integer, we define

$$g_\beta(n, d) = \sum_{i=0}^{d} \binom{n}{i} \cdot (M-1)^i,$$

which satisfies [132]

$$g_\beta(n, d) = g_\beta(n-1, d) + (M-1) \cdot g_\beta(n-1, d-1). \tag{D.41}$$

We will prove this result by induction with the following induction argument:
$\mathfrak{G}(n, d)$: For any function class $\mathcal{F} \subseteq (U_\beta)^{\mathcal{X}}$ with $\mathfrak{D}(\mathcal{F}, \alpha) \le d$, and any depth-$n$ $\mathcal{X}$-valued tree $\mathbf{x}$, $\mathcal{N}_{\mathsf{sq}}(\mathcal{F} \circ \mathbf{x}, \alpha, n) \le g_\beta(d, n)$.

Base: There are two base case: $n \le d$ and $d = 0$.

When $n \le d$, we let

$$\mathcal{V} = \{\mathbf{v}[l_1, l_2, \cdots, l_n] : l_1, \cdots, l_n \in U_\beta\},$$

where $\mathbf{v}[l_1, \cdots, l_n]$ denotes the tree which takes value l_t at depth t along any path. Then it is easy to see that for any $f \in \mathcal{F}$, depth-n $\mathcal{X}$-valued tree $\mathbf{x}$, and any path $\mathbf{y} \in \{0,1\}^n$, there exists some $\mathbf{v} \in \mathcal{V}$ such that $f(x_t(\mathbf{y})) = v_t(\mathbf{y})$ for all $t \in [n]$. Hence $\mathcal{V}$ is a 0-sequential covering of $\mathcal{F} \circ \mathbf{x}$, hence $\mathcal{V}$ is also an α-sequential covering of $\mathcal{F} \circ \mathbf{x}$ as well. Hence we have

$$\mathcal{N}_{\mathsf{sq}}(\mathcal{F} \circ \mathbf{x}, \alpha, n) \le |\mathcal{V}| = |U_\beta|^n = M^n = \sum_{i=0}^{d} \binom{n}{i} \cdot (M-1)^i = g_\beta(n, d).$$

When $d = 0$, there is no depth-1 $\mathcal{X}$-valued tree which shatters $\mathcal{F}$ at scale α. This implies for any two $x, x' \in \mathcal{X}$, we always have $h(f(x), h(x')) \le \alpha$ (otherwise we can construct depth-1 tree $\mathbf{x}$ with $x_1(0) = x$ and $x_1(1) = x'$, then this tree is shattered by $\mathcal{F}$). For any $x_0 \in \mathcal{X}$, we construct depth-n $[0,1]$-valued tree $\mathbf{v}$ which always takes value $f(x_0)$ no matter the path and depth. Then for any $f \in \mathcal{F}$, depth-n $\mathcal{X}$-valued tree $\mathbf{x}$ and any path $\mathbf{y} \in \{0,1\}$, we always have $h(f(x_t(\mathbf{y})), v_t(\mathbf{y})) = h(f(x_t(\mathbf{y})), f(x_0)) \le \alpha$. Hence $\mathcal{V}$ is an α-sequential covering of $\mathcal{F} \circ \mathbf{x}$, and it satisfies $|\mathcal{V}| = 1 = g_\beta(n, 0)$.

Induction: Suppose the induction hypothesis $\mathfrak{G}(n-1, d-1)$ and $\mathfrak{G}(n-1, d)$ both holds. We will prove induction statement $\mathfrak{G}(n, d)$. For fixed function class $\mathcal{F}$ with $\mathfrak{D}(\mathcal{F}, \alpha) = d$ and depth-n $\mathcal{X}$-valued tree $\mathbf{x}$, we will construct a α-sequential covering to $\mathcal{F} \circ \mathbf{x}$ whose size is no more than $g_\beta(n, d)$. Suppose the root of tree $\mathbf{x}$ is x_1, the left subtree of x_1 is $\mathbf{x}^l$, and the right subtree of x_1 is $\mathbf{x}^r$. We partition the function class $\mathcal{F}$ as:

$$\mathcal{F} = \mathcal{F}_1 \cup \mathcal{F}_2 \cup \cdots \cup \mathcal{F}_{1/2\beta} \quad \text{where} \quad \mathcal{F}_k = \{f \in \mathcal{F} : f(x_1) = (2k-1)\beta\}, \forall 1 \le k \le M.$$

Then we have $\mathfrak{D}(\mathcal{F}_k, \alpha) \le \mathfrak{D}(\mathcal{F}, \alpha) = d$ for all $k \in [M]$. We let $\mathcal{K} = \{k \in [M] : \mathfrak{D}(\mathcal{F}_k, \alpha) = d\}$. Then for any $a, b \in \mathcal{K}$ and $a < b$, there exist two depth-d $\mathcal{X}$-valued trees $\mathbf{x}^a$ and $\mathbf{x}^b$, and also two depth-d $(U_\beta \times U_\beta)$-valued trees $\mathbf{s}^a$ and $\mathbf{s}^b$ such that for any $\mathbf{y} \in \{0,1\}^d$ and $t \in [d]$,

$$h(s_t^a(\mathbf{y})[0], s_t^a(\mathbf{y})[1]) > \alpha \quad \text{and} \quad h(s_t^b(\mathbf{y})[0], s_t^b(\mathbf{y})[1]) > \alpha,$$

and further for any $\mathbf{y} \in \{0,1\}^d$, there exists $f_a^{\mathbf{y}} \in \mathcal{F}_a$ and $f_b^{\mathbf{y}} \in \mathcal{F}_b$ such that for any $t \in [d]$,

$$f_a^{\mathbf{y}}(x_t^a(\mathbf{y})) = s_t^a(\mathbf{y})[y_t] \quad \text{and} \quad f_b^{\mathbf{y}}(x_t^b(\mathbf{y})) = s_t^b(\mathbf{y})[y_t],$$

If we further have $h((2a-1)\beta, (2b-1)\beta) > \alpha$, we construct a depth-$(d+1)$ $\mathcal{X}$-valued tree $\mathbf{x}$ with root x_0, left subtree of the root to be $\mathbf{x}^a$, and right subtree of the root to be $\mathbf{x}^b$, and also a depth-$(d+1)$ $U_\beta \times U_\beta$-valued tree $\mathbf{s}$ with root $((2a-1)\beta, (2b-1)\beta)$, left subtree of the root to be $\mathbf{s}^a$, and right subtree of the root to be $\mathbf{s}^b$. Then we can verify that for any $\mathbf{y} \in \{0,1\}^{d+1}$, and any $t \in [d+1]$, we have $s_t^b(\mathbf{y})[0] < s_t^b(\mathbf{y})[1]$, and $h(s_t(\mathbf{y})[0], s_t(\mathbf{y})[1]) > \alpha$. Further we let $\mathbf{y}' = (y_2, y_3, \cdots, y_{d+1}) \in \{0,1\}^d$, and if $y_1 = 0$, then by letting $f^{\mathbf{y}} = f_a^{\mathbf{y}'}$ we can verify that $f^{\mathbf{y}}(x_t(\mathbf{y})) = s_t(\mathbf{y})[y_t]$ for any $t \in [d+1]$, and if $y_1 = 1$, then by letting $f^{\mathbf{y}} = f_b^{\mathbf{y}'}$ we can verify that $f^{\mathbf{y}}(x_t(\mathbf{y})) = s_t(\mathbf{y})[y_t]$ for any $t \in [d+1]$. Hence, $\mathcal{F}$ is shattered by tree $\mathbf{x}$ of depth-$(d+1)$, leading to contradiction. Therefore, we have

$$h((2a-1)\beta, (2b-1)\beta) \le \alpha, \qquad \forall a, b \in \mathcal{K} \tag{D.42}$$

Next, for any $k \in [M]$ with $\mathfrak{D}(\mathcal{F}_k, \alpha) \le d-1$, according to induction hypothesis $\mathfrak{G}(n-1, d-1)$, there exists a sequential cover $\mathcal{V}_k^l$ of size $g_\beta(n-1, d-1)$ for the depth-$(n-1)$ $\mathcal{X}$-valued tree $\mathbf{x}^l$, and also a sequential cover $\mathcal{V}_k^r$ of size $g_\beta(n-1, d-1)$ for the depth-$(n-1)$ $\mathcal{X}$-valued tree $\mathbf{x}^r$. We then combine the elements in $\mathcal{V}_k^l$ and $\mathcal{V}_k^r$ into a set $\mathcal{V}_k$ of depth-n U_β-valued trees. We let $v_1 = (2k-1)\beta \in U_\beta$. Then according to the construction of $\mathcal{F}_k$ we have for any $f \in \mathcal{F}$, $f(x_1) = v_1$ hence $h(f(x_1), v_1) \le \alpha$. For $\mathbf{v}^l \in \mathcal{V}_k^l$ and $\mathbf{v}^r \in \mathcal{V}_k^r$, we define depth-$n$ U_β-valued tree $\mathbf{v}[\mathbf{v}^l, \mathbf{v}^r]$ as: for any path $\mathbf{y} \in \{0,1\}^n$, we let $\mathbf{y}' = (y_{2:n}) \in \{0,1\}^{n-1}$, and let $v_1[\mathbf{v}^l, \mathbf{v}^r](\mathbf{y}) = v_1$. If $y_1 = 0$, then let $v_t[\mathbf{v}^l, \mathbf{v}^r](\mathbf{y}) = v_{t-1}^l(\mathbf{y}')$, and if $y_1 = 1$, then let $v_t[\mathbf{v}^l, \mathbf{v}^r](\mathbf{y}) = v_{t-1}^r(\mathbf{y}')$. We construct $\mathcal{V}_k = \{\mathbf{v}[\mathbf{v}^l, \mathbf{v}^r]\}$ with $|\mathcal{V}_k| \le \max\{|\mathcal{V}_k^l|, |\mathcal{V}_k^r|\}$ to make sure that every element in $\mathcal{V}_k^l$ and $\mathcal{V}_k^r$ at least appear once in the construction of $\mathcal{V}_k$. Next, we will argue that $\mathcal{V}_k$ is a α-sequential cover of $\mathcal{F}_k \circ \mathbf{x}$. For any $f \in \mathcal{F}_k$ and $\mathbf{y} \in \{0,1\}^n$, if $y_1 = 0$, then since $\mathcal{V}_k^l$ is a α-sequential cover of $\mathcal{F}_k$, there exists $\mathbf{v}^l \in \mathcal{V}_k^l$ such that for any $2 \le t \le n$, $h(f(x_t(\mathbf{y})), v_t^l(\mathbf{y})) \le \alpha$. Suppose $\mathbf{v} = \mathbf{v}[\mathbf{v}^l, \mathbf{v}^r] \in \mathcal{V}_k$ for some $\mathbf{v}^r \in \mathcal{V}_k^r$, and we also have $h(f(x_1(\mathbf{y})), v_1(\mathbf{y})) \le \alpha$ according to the construction of $\mathcal{F}_k$. Hence for any $t \in [n]$, we always $h(f(x_t(\mathbf{y})), v_t(\mathbf{y})) \le \alpha$. Therefore, $\mathcal{V}'$ is a cover of $\mathcal{F}_k$. Further by induction hypothesis we have $\max\{|\mathcal{V}_k^l|, |\mathcal{V}_k^r|\} \le g_\beta(n-1, d-1)$. Hence $|\mathcal{V}_k| \le g_\beta(n-1, d-1)$.

If $\mathcal{K} = \emptyset$, then by letting $\mathcal{V} = \cup_{k \in [M]} \mathcal{V}_k$, $\mathcal{V}$ will be a α-sequential cover of $\mathcal{F} \circ \mathbf{x}$, and also

$$|\mathcal{V}| \le M \cdot g_\beta(n-1, d-1) \le g_\beta(n-1, d) + (M-1) g_\beta(n-1, d-1) = g_\beta(n, d),$$

where the inequality follows from the fact that $g_\beta(n-1, d-1) \le g_\beta(n-1, d)$ for any n, d, and the last equation follows from eq. (D.41).

Next, we consider cases where $|\mathcal{K}| \ge 1$ We construct $\mathcal{F}' = \cup_{k \in \mathcal{K}} \mathcal{F}_k$, then we have $\mathfrak{D}(\mathcal{F}', \alpha) \le \mathfrak{D}(\mathcal{F}, \alpha) = d$. According to the induction hypothesis $\mathfrak{G}(n-1, d)$, there exists a

sequential cover $\mathcal{V}^l$ of size $g_\beta(n-1,d)$ for the depth-$(n-1)$ $\mathcal{X}$-valued tree $\mathbf{x}^l$, and also a sequential cover $\mathcal{V}^l$ of size $g_\beta(n-1,d)$ for the depth-$(n-1)$ $\mathcal{X}$-valued tree $\mathbf{x}^l$. We then combine the elements in $\mathcal{V}^l$ and $\mathcal{V}^r$ into a $\mathcal{V}'$ of depth-n U_β-valued trees. We let $v_1 = f(x_1) \in U_\beta$ for some $f \in \mathcal{F}'$. Then according to the construction of $\mathcal{F}'$ we have for any $f \in \mathcal{F}'$, $h(f(x_1), v_1) \le \alpha$. For $\mathbf{v}^l \in \mathcal{V}^l$ and $\mathbf{v}^r \in \mathcal{V}^r$, we define depth-$n$ U_β-valued tree $\mathbf{v}[\mathbf{v}^l, \mathbf{v}^r]$ as: for any path $\mathbf{y} \in \{0,1\}^n$, we let $\mathbf{y}' = (y_{2:n}) \in \{0,1\}^{n-1}$, and let $v_1[\mathbf{v}^l, \mathbf{v}^r](\mathbf{y}) = v_1$. If $y_1 = 0$, then let $v_t[\mathbf{v}^l, \mathbf{v}^r](\mathbf{y}) = v^l_{t-1}(\mathbf{y}')$, and if $y_1 = 1$, then let $v_t[\mathbf{v}^l, \mathbf{v}^r](\mathbf{y}) = v^r_{t-1}(\mathbf{y}')$. We construct $\mathcal{V}' = \{\mathbf{v}[\mathbf{v}^l, \mathbf{v}^r]\}$ with $|\mathcal{V}'| \le \max\{|\mathcal{V}^l|, |\mathcal{V}^r|\}$ to make sure that every element in $\mathcal{V}^l$ and $\mathcal{V}^r$ at least appear once in the construction of $\mathcal{V}'$. Next, we will argue that $\mathcal{V}'$ is a α-sequential cover of $\mathcal{F}' \circ \mathbf{x}$. For any $f \in \mathcal{F}'$ and $\mathbf{y} \in \{0,1\}^n$, if $y_1 = 0$, then since $\mathcal{V}^l$ is a α-sequential cover of $\mathcal{F}'$, there exists $\mathbf{v}^l \in \mathcal{V}^l$ such that for any $2 \le t \le n$, $h(f(x_t(\mathbf{y})), v^l_t(\mathbf{y})) \le \alpha$. Suppose $\mathbf{v} = \mathbf{v}[\mathbf{v}^l, \mathbf{v}^r] \in \mathcal{V}'$ for some $\mathbf{v}^r \in \mathcal{V}^r$, and we also have $h(f(x_1(\mathbf{y})), v_1(\mathbf{y})) \le \alpha$ according to eq. (D.42) and the construction of $\mathcal{F}'$. Hence for any $t \in [n]$, we always $h(f(x_t(\mathbf{y})), v_t(\mathbf{y})) \le \alpha$. Therefore, $\mathcal{V}'$ is a cover of $\mathcal{F}'$. Further by induction hypothesis we have $\max\{|\mathcal{V}^l|, |\mathcal{V}^r|\} \le g_\beta(n-1,d)$. Hence $|\mathcal{V}'| \le g_\beta(n-1,d)$.

We further let $\mathcal{V} = \mathcal{V}' \cup (\cup_{k \notin \mathcal{K}} \mathcal{V}_k)$, and we have

$$|\mathcal{V}| \le (M-1) \cdot g_\beta(n-1, d-1) + g_\beta(n-1,d) = g_\beta(n,d),$$

where the last equation follows from eq. (D.41). Above all, we finish the proof of induction statement $\mathfrak{G}(n,d)$.

Hence by induction, we have

$$\mathcal{N}_{\mathsf{sq}}(\mathcal{F} \circ \mathbf{x}, \alpha, n) \le g_\beta(n, \mathfrak{D}(\mathcal{F}, \alpha)) = \sum_{i=0}^{\mathfrak{D}(\mathcal{F},\alpha)} \binom{n}{i} \cdot (M-1)^i \le \left(\frac{en}{\beta \mathfrak{D}(\mathcal{F},\alpha)}\right)^{\mathfrak{D}(\mathcal{F},\alpha)} \le \left(\frac{en}{\beta}\right)^{\mathfrak{D}(\mathcal{F},\alpha)}$$

□

Equipped with Theorem D.19, we are ready to prove Theorem D.16 that works for real-valued function classes.

Proof of Theorem D.16. For $\beta > 0$ where $1/(2\beta)$ is an integer, we let $M = 1/(2\beta)$, and we define

$$U_\beta = \{\beta, 3\beta, 5\beta, \cdots, (2M-1)\beta\}.$$

And for every $u \in [0,1]$, we define $\lfloor u \rfloor_\beta = \arg\min_{r \in U_\beta} |u - r|$. For any function $f \in \mathcal{F}$, we define $\lfloor f \rfloor_\beta : \mathcal{X} \to U_\beta$ as

$$\lfloor f \rfloor_\beta(x) \coloneqq \lfloor f(x) \rfloor_\beta.$$

We further let $\lfloor \mathcal{F} \rfloor_\beta = \{\lfloor f \rfloor_\beta : f \in \mathcal{F}\}$. According to Theorem D.15 and Theorem D.18 we know that if $\lfloor \mathcal{F} \rfloor$ is shattered by some $\mathcal{X}$-valued tree $\mathbf{x}$ at scale α, then $\mathbf{x}$ also also (α, β)-shattered by $\mathcal{F}$. Hence we have

$$\mathfrak{D}(\mathcal{F}, \alpha, \beta) \ge \mathfrak{D}(\lfloor \mathcal{F} \rfloor_\beta, \alpha).$$

Hence Theorem D.19 gives that for any depth-n $\mathcal{X}$-valued tree $\mathbf{x}$,

$$\mathcal{N}(\lfloor \mathcal{F} \rfloor_\beta \circ \mathbf{x}, \alpha, n) \le \left(\frac{en}{\beta}\right)^{\mathfrak{D}(\lfloor \mathcal{F} \rfloor_\beta, \alpha)} \le \left(\frac{en}{\beta}\right)^{\mathfrak{D}(\mathcal{F}, \alpha, \beta)}.$$

We let $\mathcal{V}$ to be the covering of $\lfloor \mathcal{F} \rfloor_\beta$ at scale α with size no more than $(en/\beta)^{\mathfrak{D}(\mathcal{F},\alpha,\beta)}$. Hence for any $f \in \mathcal{F}$, $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \{0,1\}^n$, there exists $\mathbf{v} \in \mathcal{V}$ such that for any $t \in [n]$,

$$h\left(\lfloor f \rfloor_\beta(x_t(\mathbf{y})), v_t(\mathbf{y})\right) \le \alpha.$$

According to the construction of $\lfloor f \rfloor_\beta$, we have $|\lfloor f \rfloor_\beta(x_t(\mathbf{y})) - f(x_t(\mathbf{y}))| \le \beta$, which implies that

$$\begin{aligned}
&h(\lfloor f \rfloor_\beta(x_t(\mathbf{y})), f(x_t(\mathbf{y})))^2 \\
&\quad \le \left(\sqrt{\lfloor f \rfloor_\beta(x_t(\mathbf{y}))} - \sqrt{f(x_t(\mathbf{y}))}\right)^2 + \left(\sqrt{1 - \lfloor f \rfloor_\beta(x_t(\mathbf{y}))} - \sqrt{1 - f(x_t(\mathbf{y}))}\right)^2 \\
&\quad \le |\lfloor f \rfloor_\beta(x_t(\mathbf{y})) - f(x_t(\mathbf{y}))| + |1 - \lfloor f \rfloor_\beta(x_t(\mathbf{y})) - (1 - f(x_t(\mathbf{y})))| \le 2\beta,
\end{aligned}$$

where the first inequality uses the fact that $(a-b)^2 \le (a-b)(a+b)$ for $a, b \ge 0$. Therefore, for any $t \in [t]$, we have

$$h(f(x_t(\mathbf{y})), v_t(\mathbf{y})) \le h(\lfloor f \rfloor_\beta(x_t(\mathbf{y})), f(x_t(\mathbf{y}))) + h\left(\lfloor f \rfloor_\beta(x_t(\mathbf{y})), v_t(\mathbf{y})\right) \le \alpha + \sqrt{2\beta},$$

which implies that $\mathcal{V}$ is an $(\alpha + \sqrt{2\beta})$ sequential covering of $\mathcal{F}$, i.e.

$$\mathcal{N}(\mathcal{F} \circ \mathbf{x}, \alpha + \sqrt{2\beta}, n) \le |\mathcal{V}| \le \left(\frac{en}{\beta}\right)^{\mathfrak{D}(\mathcal{F},\alpha,\beta)}.$$

Taking supremum over $\mathbf{x}$, we obtain that

$$\sup_{\mathbf{x}} \mathcal{H}_{\mathsf{sq}}(\mathcal{F}, \alpha + \sqrt{2\beta}, n, \mathbf{x}) = \sup_{\mathbf{x}} \log \mathcal{N}(\mathcal{F} \circ \mathbf{x}, \alpha + \sqrt{2\beta}, n) \le \mathfrak{D}(\mathcal{F}, \alpha, \beta) \log\left(\frac{en}{\beta}\right).$$

□

Proof of Theorem D.17

Before proving Theorem D.17, we present the following helper lemma.

Lemma D.20. *Suppose real number $p, \alpha, \beta \in [0,1]$ and integer n satisfies $\beta < \alpha^2$ and*

$$\alpha + \beta < p < 1 - \alpha - \beta, \quad \text{and} \quad n \le \frac{p(1-p)}{324\alpha^2} \vee 1. \tag{D.43}$$

We collect n samples from $\mathrm{Ber}(p)$ to form an empirical estimation $\widehat{p} \in [0,1]$, and we define

$$\varepsilon = \begin{cases} 1 & \text{if } \widehat{p} \ge p \\ -1 & \text{if } \widehat{p} < p. \end{cases} \tag{D.44}$$

Then we have

$$\mathbb{E}\left[\widehat{p} \log \frac{p + \varepsilon\alpha - \beta}{p} + (1 - \widehat{p}) \log \frac{1 - p - \varepsilon\alpha - \beta}{1 - p}\right] \ge \frac{\alpha^2}{8p(1-p)}, \tag{D.45}$$

where the expectation is over $\widehat{p}$ and ε. Additionally, when choosing $n = \lfloor \frac{p(1-p)}{324\alpha^2} \rfloor \vee 1$, we have

$$n \cdot \mathbb{E}\left[\widehat{p} \log \frac{p + \varepsilon\alpha - \beta}{p} + (1 - \widehat{p}) \log \frac{1 - p - \varepsilon\alpha - \beta}{1 - p}\right] \ge \frac{1}{5184}. \tag{D.46}$$

Proof of Theorem D.20. Without loss of generality, we assume $p \le 1/2$ (otherwise we replace p with $1-p$ and the argument follows similarly). Then we have $\alpha < 1/2$ and $\beta < 1/4$. Consider the following three cases:

(i) $1/36 < \alpha \le 1/2$,

(ii) $\alpha + \beta \ge p/2$ and $\alpha < 1/36$,

(iii) $\alpha + \beta < p/2$ and $\alpha < 1/36$.

When $1/36 < \alpha < 1/2$, since $p(1-p) \le 1/4$ and $\alpha^2 \ge 1/1296$, we always have $n = 1$, which implies

$$\begin{aligned}
&\mathbb{E}\left[\widehat{p}\log\frac{p+\varepsilon\alpha-\beta}{p} + (1-\widehat{p})\log\frac{1-p-\varepsilon\alpha-\beta}{1-p}\right] \\
&\quad = p\cdot\log\left(1+\frac{\alpha-\beta}{p}\right) + (1-p)\cdot\log\left(1+\frac{\alpha-\beta}{1-p}\right) \overset{(i)}{\ge} \frac{\alpha-\beta}{2} \overset{(ii)}{\ge} \frac{\alpha}{4} \overset{(iii)}{\ge} \frac{\alpha^2}{8p(1-p)},
\end{aligned}$$

where (i) uses $\alpha - \beta > 0$ and either $(\alpha-\beta)/p > 1/2$ or $(\alpha-\beta)/(1-p) > 1/2$ and $\log(1+x) \ge x/2$ for $0 \le x \le 2$, (ii) uses the fact that $\alpha \le 1/2$ and $\beta \le \alpha^2$, and (iii) uses the fact $\alpha \le p$ and $1-p \ge 1/2$.

When $\alpha + \beta > p/2$ and $\alpha < 1/36$, since $\beta < \alpha^2$ we have $\alpha > p/3$, which implies that

$$n \le \frac{p(1-p)}{324\alpha^2} \vee 1 < 1/p.$$

Hence $\widehat{p} < p$, i.e. $\varepsilon = -1$ if and only if $\widehat{p} = 0$. Therefore,

$$\begin{aligned}
&\mathbb{E}\left[\widehat{p}\log\frac{p+\varepsilon\alpha-\beta}{p} + (1-\widehat{p})\log\frac{1-p-\varepsilon\alpha-\beta}{1-p}\right] \\
&\quad = \Pr(\widehat{p}=0)\cdot\log\left(1+\frac{\alpha-\beta}{1-p}\right) + \mathbb{E}\left[\left(\widehat{p}\log\frac{p+\alpha-\beta}{p} + (1-\widehat{p})\log\frac{1-p-\alpha-\beta}{1-p}\right)\cdot\mathbb{I}[\widehat{p}>0]\right].
\end{aligned} \tag{D.47}$$

Since $p < 1/2$, we have $\alpha+\beta < 1/36+1/36^2 < (1-p)/2$, which implies $(\alpha+\beta)/(1-p) \le 1/2$. After noticing that $\log(1+x) \ge x - x^2$ holds for all $x \ge -1/2$ and also $\alpha - \beta > 0$, we can further upper bound eq. (D.47) by

$$\begin{aligned}
&\Pr(\widehat{p}=0)\cdot\left(\left(\frac{\alpha-\beta}{1-p}\right) - \left(\frac{\alpha-\beta}{1-p}\right)^2\right) \\
&\qquad + \mathbb{E}\left[\left(\widehat{p}\cdot\left(\left(\frac{\alpha-\beta}{p}\right) - \left(\frac{\alpha-\beta}{p}\right)^2\right) + (1-\widehat{p})\cdot\left(\left(\frac{-\alpha-\beta}{1-p}\right) - \left(\frac{-\alpha-\beta}{1-p}\right)^2\right)\right)\cdot\mathbb{I}[\widehat{p}>0]\right] \\
&\quad = \mathbb{E}\left[\widehat{p}\cdot\left(\left(\frac{\varepsilon\alpha-\beta}{p}\right) - \left(\frac{\varepsilon\alpha-\beta}{p}\right)^2\right) + (1-\widehat{p})\cdot\left(\left(\frac{-\varepsilon\alpha-\beta}{1-p}\right) - \left(\frac{-\varepsilon\alpha-\beta}{1-p}\right)^2\right)\right] \\
&\quad \overset{(i)}{\ge} \mathbb{E}\left[\frac{\varepsilon\widehat{p}\alpha}{p} + \frac{-\varepsilon(1-\widehat{p})\alpha}{1-p}\right] - \beta\cdot\mathbb{E}\left[\frac{\widehat{p}}{p} + \frac{1-\widehat{p}}{1-p}\right] - (\alpha+\beta)^2\cdot\mathbb{E}\left[\frac{\widehat{p}}{p^2} + \frac{1-\widehat{p}}{(1-p)^2}\right]
\end{aligned}$$

$$
\begin{aligned}
&\overset{(ii)}{=} \mathbb{E}\left[\frac{\varepsilon(\widehat{p}-p)\alpha}{p}+\frac{-\varepsilon(p-\widehat{p})\alpha}{1-p}\right]-2\beta-(\alpha+\beta)^2\left(\frac{1}{p}+\frac{1}{1-p}\right)\\
&=\alpha\cdot\mathbb{E}\left[\frac{|\widehat{p}-p|}{p(1-p)}\right]-2\beta-\frac{(\alpha+\beta)^2}{p(1-p)}\\
&\overset{(iii)}{\leq}\alpha\cdot\mathbb{E}\left[\frac{|\widehat{p}-p|}{p(1-p)}\right]-\frac{2\alpha^2}{p(1-p)}
\end{aligned}
$$

where (i) uses $|\varepsilon\alpha-\beta|\leq\alpha+\beta$, $|-\varepsilon\alpha-\beta|\leq\alpha+\beta$ and $\mathbb{E}[\widehat{p}]=p$, (ii) uses the definition of ε in eq. (D.44) and (iii) uses the fact that $0<\beta\leq\alpha^2\leq 1/1296$. According to Khintchine inequality [222], we have

$$
\mathbb{E}[|\widehat{p}-p|]\geq\sqrt{\frac{p(1-p)}{2n}},
$$

which implies

$$
\text{LHS of } eq.\ (D.45)\geq\frac{\alpha}{\sqrt{2np(1-p)}}-\frac{2\alpha^2}{p(1-p)}\geq\frac{\alpha^2}{p(1-p)},
$$

where the last inequality uses the fact that

$$
n\leq\frac{p(1-p)}{324\alpha^2}\vee 1\quad\text{and}\quad\frac{\alpha}{\sqrt{p(1-p)}}\leq\sqrt{\alpha}\cdot\sqrt{\frac{p}{p(1-p)}}\leq\frac{1}{3\sqrt{2}}.
$$

When $\alpha+\beta<p/2$ and $\alpha<1/36$, we have $(p-\alpha-\beta)/p\geq 1/2$ and also $(1-p-\alpha-\beta)/(1-p)\geq 1/2$, then for any $\varepsilon\in\{-1,1\}$,

$$
\frac{p+\varepsilon\alpha-\beta}{p}\geq\frac{1}{2},\quad\text{and}\quad\frac{1-p-\varepsilon\alpha-\beta}{1-p}\geq\frac{1}{2}.
$$

Notice that $\log(1+x)\geq x-x^2$ holds for all $x\geq-1/2$, which implies

$$
\begin{aligned}
&\text{LHS of } eq.\ (D.45)\\
&\geq\mathbb{E}\left[\widehat{p}\cdot\left(\left(\frac{\varepsilon\alpha-\beta}{p}\right)-\left(\frac{\varepsilon\alpha-\beta}{p}\right)^2\right)+(1-\widehat{p})\cdot\left(\left(\frac{-\varepsilon\alpha-\beta}{1-p}\right)-\left(\frac{-\varepsilon\alpha-\beta}{1-p}\right)^2\right)\right]\\
&=\mathbb{E}\left[\widehat{p}\left(\frac{\varepsilon\alpha-\beta}{p}\right)+(1-\widehat{p})\left(\frac{-\varepsilon\alpha-\beta}{1-p}\right)-\widehat{p}\left(\frac{\varepsilon\alpha-\beta}{p}\right)^2-(1-\widehat{p})\left(\frac{-\varepsilon\alpha-\beta}{1-p}\right)^2\right]\\
&\overset{(i)}{\geq}\mathbb{E}\left[\frac{\varepsilon\widehat{p}\alpha}{p}+\frac{-\varepsilon(1-\widehat{p})\alpha}{1-p}\right]-\beta\cdot\mathbb{E}\left[\frac{\widehat{p}}{p}+\frac{1-\widehat{p}}{1-p}\right]-(\alpha+\beta)^2\cdot\mathbb{E}\left[\frac{\widehat{p}}{p^2}+\frac{1-\widehat{p}}{(1-p)^2}\right]\\
&\overset{(ii)}{=}\mathbb{E}\left[\frac{\varepsilon(\widehat{p}-p)\alpha}{p}+\frac{-\varepsilon(p-\widehat{p})\alpha}{1-p}\right]-2\beta-(\alpha+\beta)^2\left(\frac{1}{p}+\frac{1}{1-p}\right)\\
&=\alpha\cdot\mathbb{E}\left[\frac{|\widehat{p}-p|}{p(1-p)}\right]-2\beta-\frac{(\alpha+\beta)^2}{p(1-p)}\\
&\overset{(iii)}{\leq}\alpha\cdot\mathbb{E}\left[\frac{|\widehat{p}-p|}{p(1-p)}\right]-\frac{2\alpha^2}{p(1-p)}
\end{aligned}
$$

where (i) uses $|\varepsilon\alpha - \beta| \le \alpha + \beta$, $|-\varepsilon\alpha - \beta| \le \alpha + \beta$ and $\mathbb{E}[\widehat{p}] = p$, (ii) uses the definition of ε in eq. (D.44) and (iii) uses the fact that $0 < \beta \le \alpha^2 \le 1$. According to [223, Theorem 1], we have

$$\mathbb{E}[|\widehat{p} - p|] \ge \sqrt{\frac{p(1-p)}{2n}},$$

which implies

$$\text{LHS of } eq.\ (D.45) \ge \frac{\alpha}{\sqrt{2np(1-p)}} - \frac{2\alpha^2}{p(1-p)} \ge \frac{\alpha^2}{p(1-p)},$$

where the last inequality uses the fact that

$$n \le \frac{p(1-p)}{324\alpha^2} \vee 1 \quad \text{and} \quad \frac{\alpha}{\sqrt{p(1-p)}} \le \sqrt{\alpha} \cdot \sqrt{\frac{p}{p(1-p)}} \le \frac{1}{3\sqrt{2}}.$$

Above all, we have verified eq. (D.45), and eq. (D.46) follows from eq. (D.45) directly. □

Now we are ready to prove Theorem D.17.

Proof of Theorem D.17. Let $x_0 \in \mathcal{X}$ be an arbitrary context. For fixed positive integer n, we let

$$\alpha_n = \arg\max_{\alpha > 0} \left\{ \mathfrak{D}(\mathcal{F}, \alpha, \alpha^4/16) \cdot \left(\left\lceil \frac{1}{162\alpha^2} \right\rceil \vee 1 \right) \le n \right\}.$$

Since $\mathfrak{D}(\mathcal{F}, \alpha, \alpha^4/16) = \tilde{\Omega}\left(\alpha^{-p}\right)$ for every $\alpha \ge 0$, we have

$$\mathfrak{D}(\mathcal{F}, \alpha_n, \alpha_n^4/16) = \tilde{\Omega}\left(n^{\frac{p}{p+2}}\right).$$

In the following, we will prove that for any positive integer n, we have

$$\mathcal{R}_n(\mathcal{F}) \ge \frac{\mathfrak{D}(\mathcal{F}, \alpha_n, \alpha_n^4/16)}{5184}.$$

We fix n, and let $\alpha = \alpha_n$, $d = \mathfrak{D}(\mathcal{F}, \alpha_n, \alpha_n^4/16)$. We let $\tilde{\mathbf{x}}$ to be the depth-d $\mathcal{X}$-valued tree shattered by $\mathcal{F}$ at scale $(\alpha_n, \alpha_n^4/16)$. Then according to Theorem D.15, there exists a depth-d $[0,1] \times [0,1]$-valued tree $\mathbf{s}$ such that for any path $\tilde{\mathbf{y}} = (\tilde{y}_{1:d}) \in \{0,1\}^d$, there exists $f^{\mathbf{y}} \in \mathcal{F}$ such that

$$\left| f^{\tilde{\mathbf{y}}}(t(\tilde{\mathbf{y}})) - s_t(\tilde{\mathbf{y}})[\tilde{y}_t] \right| < \frac{\alpha^4}{16} \quad \text{and} \quad h(s_t(\tilde{\mathbf{y}})[0], s_t(\tilde{\mathbf{y}})[1]) \ge \alpha \qquad \forall t \in [d], \tag{D.48}$$

After noticing that $h(u,v)^2/2 \le |u - v|$ for any $u, v \in [0,1]$, we have

$$\left| f^{\tilde{\mathbf{y}}}(t(\tilde{\mathbf{y}})) - s_t(\tilde{\mathbf{y}})[\tilde{y}_t] \right| \le \frac{(s_t(\tilde{\mathbf{y}})[0] - s_t(\tilde{\mathbf{y}})[1])^2}{4}$$

We define depth-d $[0,1]$-valued $\mathbf{v}$ as

$$v_t(\tilde{\mathbf{y}}) = \frac{s_t(\tilde{\mathbf{y}})[0] + s_t(\tilde{\mathbf{y}})[1]}{2}, \quad \forall \mathbf{y} \in \{0,1\}^n. \tag{D.49}$$

We can further verify that

$$\begin{aligned}
&h(s_t(\tilde{\mathbf{y}})[0], s_t(\tilde{\mathbf{y}})[1])^2 \\
&\quad = (s_t(\tilde{\mathbf{y}})[0] - s_t(\tilde{\mathbf{y}})[1]))^2 \\
&\qquad \cdot \max\left\{ \frac{1}{(\sqrt{s_t(\tilde{\mathbf{y}})[0]} + \sqrt{s_t(\tilde{\mathbf{y}})[1]})^2}, \frac{1}{(\sqrt{1-s_t(\tilde{\mathbf{y}})[0]} + \sqrt{1-s_t(\tilde{\mathbf{y}})[1]})^2} \right\} \\
&\quad \le (s_t(\tilde{\mathbf{y}})[0] - s_t(\tilde{\mathbf{y}})[1]))^2 \cdot \left(\frac{1}{s_t(\tilde{\mathbf{y}})[0] + s_t(\tilde{\mathbf{y}})[1]} + \frac{1}{1 - s_t(\tilde{\mathbf{y}})[0] + 1 - s_t(\tilde{\mathbf{y}})[1]} \right) \\
&\quad = \frac{2\left(s_t(\tilde{\mathbf{y}})[0] - s_t(\tilde{\mathbf{y}})[1])\right)^2}{v_t(\tilde{y})(1 - v_t(\tilde{y}))}
\end{aligned} \tag{D.50}$$

Next, we will show $\mathcal{R}_n(\mathcal{F}) \ge d$. According to Theorem D.3, for any $\mathbf{p} \in \Delta(\{0,1\}^n)$ and depth-n $\mathcal{X}$-valued tree $\mathbf{x}$, we have

$$\mathcal{R}_n(\mathcal{F}) \ge \mathbb{E}_{\mathbf{y}\sim\mathbf{p}} \mathcal{R}_n(\mathcal{F}(\mathbf{x}), \mathbf{p}, \mathbf{y}). \tag{D.51}$$

In the following, we will construct specific $\mathbf{p}$ and $\mathbf{x}$ so that the right-hand side in the above inequality is lower bounded by d. For a fixed a path $\mathbf{y} \in \{0,1\}^n$, we first define an auxillary $\{0,1\}$-path $\tilde{\mathbf{y}} = (\tilde{y}_{1:d}) \in \{0,1\}^d$ of length-d and also d integers: $k_1, k_2, \cdots, k_d$ in the following way: calculate $\tilde{y}_{1:d}$ and $k_{1:d}$ by turn:

$$k_t = \left\lfloor \frac{v_t(\tilde{y})(1 - v_t(\tilde{y}))}{324\left(s_t(\tilde{\mathbf{y}})[1] - s_t(\tilde{\mathbf{y}})[0]\right)^2} \right\rfloor \vee 1, \qquad \forall t \in [d]. \tag{D.52}$$

and

$$\tilde{y}_t = \mathbb{I}\left\{ \sum_{j=1}^{k_t} y_{k_1+\cdots+k_{t-1}+j} \ge k_t \cdot v_t(\tilde{\mathbf{y}}) \right\}, \qquad \forall t \in [d], \tag{D.53}$$

where $v_t(\tilde{\mathbf{y}})$ is defined in eq. (D.49). Notice that according to the above definition, k_t only depends on $y_1, \cdots, y_{k_1+\cdots+k_{t-1}}$, and $\tilde{y}_t$ depends on $y_1, \cdots, y_{k_1+\cdots+k_t}$. Additionally, according to eq. (D.50) and eq. (D.48), we have

$$k_t \le \frac{1}{162\alpha^2} \vee 1, \qquad \forall t \in [d]$$

which implies $k_1 + \cdots + k_d \le n$ always holds according to our choice of $\alpha = \alpha_n$. Hence $k_{1:d}$ and $\tilde{y}_{1:d}$ are all well-defined.

The value of $(x_1(\mathbf{y}), x_t(\mathbf{y}), \cdots, x_n(\mathbf{y}))$ are in the following form:

$$(\underbrace{1(\tilde{\mathbf{y}}), 1(\tilde{\mathbf{y}}), \cdots, 1(\tilde{\mathbf{y}})}_{k_1 \text{ pieces}}, \underbrace{2(\tilde{\mathbf{y}}), 2(\tilde{\mathbf{y}}), \cdots, 2(\tilde{\mathbf{y}})}_{k_2 \text{ pieces}}, \cdots, \underbrace{d(\tilde{\mathbf{y}}), d(\tilde{\mathbf{y}}), \cdots, d(\tilde{\mathbf{y}})}_{k_d \text{ pieces}}, \underbrace{x_0, x_0, \cdots, x_0}_{(n-k_1-k_2-\cdots-k_d) \text{ pieces}}),$$

Similarly, the value of $(p_1(\mathbf{y}), p_t(\mathbf{y}), \cdots, p_n(\mathbf{y}))$ are in the following form:

$$(\underbrace{v_1(\tilde{\mathbf{y}}), v_1(\tilde{\mathbf{y}}), \cdots, v_1(\tilde{\mathbf{y}})}_{k_1 \text{ pieces}}, \underbrace{v_2(\tilde{\mathbf{y}}), v_2(\tilde{\mathbf{y}}), \cdots, v_2(\tilde{\mathbf{y}})}_{k_2 \text{ pieces}}, \cdots, \underbrace{v_d(\tilde{\mathbf{y}}), v_d(\tilde{\mathbf{y}}), \cdots, v_d(\tilde{\mathbf{y}})}_{k_d \text{ pieces}}, \underbrace{f^{\tilde{\mathbf{y}}}(x_0), f^{\tilde{\mathbf{y}}}(x_0) \cdots f^{\tilde{\mathbf{y}}}(x_0)}_{(n-k_1-k_2-\cdots-k_d) \text{ pieces}}),$$

where x_0 is the state we fixed at the beginning of the proof.

Then we write $\mathcal{R}_n(\mathcal{F}(\mathbf{x}), \mathbf{p}, \mathbf{y})$ in terms of segments 1 to d, and after noticing that $f^{\mathbf{y}} \in \mathcal{F}$ for any depth-d path $\mathbf{y} \in \{-1, 1\}^d$, we obtain

$$\begin{aligned}
&\mathcal{R}_n(\mathcal{F}(\mathbf{x}), \mathbf{p}, \mathbf{y}) \\
&\quad= \sup_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{d} \sum_{j=1}^{k_d} \log \frac{f(y_{k_1+\cdots+k_{t-1}+j} \mid x_{k_1+\cdots+k_{t-1}+j}(\mathbf{y}))}{p_{k_1+\cdots+k_{t-1}+j}(y_{k_1+\cdots+k_{t-1}+j} \mid \mathbf{y})} \right. \\
&\qquad \left. + \sum_{j=1}^{n-k_1-\cdots-k_d} \log \frac{f(y_{k_1+\cdots+k_d+j} \mid x_{k_1+\cdots+k_d+j}(\mathbf{y}))}{p_{k_1+\cdots+k_d+j}(y_{k_1+\cdots+k_d+j} \mid \mathbf{y})} \right\} \\
&\quad= \sup_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{d} \sum_{j=1}^{k_d} \log \frac{f(y_{k_1+\cdots+k_{t-1}+j} \mid t(\tilde{\mathbf{y}}))}{v_t(y_{k_1+\cdots+k_{t-1}+j} \mid \tilde{\mathbf{y}})} + \sum_{j=1}^{n-k_1-\cdots-k_d} \log \frac{f(y_{k_1+\cdots+k_d+j} \mid x)}{f^{\tilde{\mathbf{y}}}(y_{k_1+\cdots+k_d+j} \mid x_0)} \right\} \\
&\quad\geq \sum_{t=1}^{d} \sum_{j=1}^{k_d} \log \frac{f^{\tilde{\mathbf{y}}}(y_{k_1+\cdots+k_{t-1}+j} \mid t(\tilde{\mathbf{y}}))}{v_t(y_{k_1+\cdots+k_{t-1}+j} \mid \tilde{\mathbf{y}})}, \qquad \text{(D.54)}
\end{aligned}$$

where the last step takes $f = f^{\mathbf{y}} \in \mathcal{F}$. Next, for fixed $\mathbf{y}$, we define

$$\hat{v}_t = \frac{1}{k_t} \sum_{j=1}^{k_t} y_{k_1+\cdots+k_{t-1}+j},$$

and let

$$\gamma_t(\tilde{\mathbf{y}}) = \frac{s_t(\tilde{\mathbf{y}})[1] - s_t(\tilde{\mathbf{y}})[0]}{2}. \qquad \text{(D.55)}$$

Then using inequality $h(u, v)^2/2 \leq |u - v|$ for any $u, v \in [0, 1]$, we have

$$\gamma_t(\tilde{\mathbf{y}}) \geq \frac{h(s_t(\tilde{\mathbf{y}})[1], s_t(\tilde{\mathbf{y}})[0])^2}{4} \geq \frac{\alpha^2}{4}. \qquad \text{(D.56)}$$

Notice that $t(\tilde{\mathbf{y}})$ and $v_t(\tilde{\mathbf{y}})$ is independent to $y_{k_1+\cdots+k_{t-1}+1}, \cdots, y_{k_1+\cdots+k_{t-1}+k_t}$, we can rewrite eq. (D.54) as

$$\mathcal{R}_n(\mathcal{F}(\mathbf{x}), \mathbf{p}, \mathbf{y}) \geq \sum_{t=1}^{d} k_t \cdot \left(\hat{v}_t \log \frac{f^{\tilde{\mathbf{y}}}(t(\tilde{\mathbf{y}}))}{v_t(\tilde{\mathbf{y}})} + (1 - \hat{v}_t) \log \frac{1 - f^{\tilde{\mathbf{y}}}(t(\tilde{\mathbf{y}}))}{1 - v_t(\tilde{\mathbf{y}})} \right).$$

Next, we will calculate $\mathbb{E}_{\mathbf{y} \sim \mathbf{p}} [\mathcal{R}_n(\mathcal{F}(\mathbf{x}), \mathbf{p}, \mathbf{y})]$. According to the above equation, we can separate the expectation into the sum of d expectations as follows:

$$\begin{aligned}
&\mathbb{E}_{\mathbf{y} \sim \mathbf{p}} [\mathcal{R}_n(\mathcal{F}(\mathbf{x}), \mathbf{p}, \mathbf{y})] \\
&\quad= \sum_{t=1}^{d} \mathbb{E} \left[k_t \cdot \left(\hat{v}_t \log \frac{f^{\tilde{\mathbf{y}}}(t(\tilde{\mathbf{y}}))}{v_t(\tilde{\mathbf{y}})} + (1 - \hat{v}_t) \log \frac{1 - f^{\tilde{\mathbf{y}}}(t(\tilde{\mathbf{y}}))}{1 - v_t(\tilde{\mathbf{y}})} \right) \right] \\
&\quad= \sum_{t=1}^{d} \mathbb{E} \left[\mathbb{E} \left[k_t \cdot \left(\hat{v}_t \log \frac{f^{\tilde{\mathbf{y}}}(t(\tilde{\mathbf{y}}))}{v_t(\tilde{\mathbf{y}})} + (1 - \hat{v}_t) \log \frac{1 - f^{\tilde{\mathbf{y}}}(t(\tilde{\mathbf{y}}))}{1 - v_t(\tilde{\mathbf{y}})} \right) \;\middle|\; y_{1:(k_1+\cdots k_{t-1})} \right] \right]
\end{aligned}$$

$$\overset{(i)}{\geq} \sum_{t=1}^{d} \mathbb{E}\left[\mathbb{E}\left[k_t \cdot \left(\hat{v}_t \log \frac{s_t(\tilde{\mathbf{y}})[\tilde{y}_t] - \alpha^4/16}{v_t(\tilde{\mathbf{y}})} + (1-\hat{v}_t)\log\frac{1-s_t(\tilde{\mathbf{y}})[\tilde{y}_t]-\alpha^4/16}{1-v_t(\tilde{\mathbf{y}})}\right) \,\middle|\, y_{1:(k_1+\cdots k_{t-1})}\right]\right]$$

$$\overset{(ii)}{\geq} \sum_{t=1}^{d} \mathbb{E}\Bigg[\mathbb{E}\Bigg[k_t \cdot \Bigg(\hat{v}_t \log \frac{v_t(\tilde{\mathbf{y}}) + \varepsilon(\hat{v}_t)\gamma_t(\tilde{\mathbf{y}}) - \gamma_t(\tilde{\mathbf{y}})^2}{v_t(\tilde{\mathbf{y}})}$$

$$+ (1-\hat{v}_t)\log\frac{1-v_t(\tilde{\mathbf{y}}) - \varepsilon(\hat{v}_t)v_t(\tilde{\mathbf{y}}) - \gamma_t(\tilde{\mathbf{y}})^2}{1-v_t(\tilde{\mathbf{y}})}\Bigg) \,\Bigg|\, y_{1:(k_1+\cdots k_{t-1})}\Bigg]\Bigg] \tag{D.57}$$

where (i) uses the choice of $f^{\tilde{\mathbf{y}}}$ in eq. (D.48), and in (ii) we define

$$\varepsilon(\hat{v}_t) = \begin{cases} 1 & \text{if } \hat{v}_t \geq v_t(\tilde{\mathbf{y}}), \\ -1 & \text{if } \hat{v}_t < v_t(\tilde{\mathbf{y}}). \end{cases}$$

and it follows from our construciton of $\tilde{y}_t$ in eq. (D.53) and also $\alpha^2/4 \leq \gamma_t(\tilde{\mathbf{y}})$ from eq. (D.56). In eq. (D.57), conditioned on $y_{1:(k_1+\cdots+k_{t-1})}$, there is no randomness on k_t, $v_t(\tilde{\mathbf{y}})$ and also $t(\tilde{\mathbf{y}})$, hence all the randomness of the inner expectation comes from $\hat{v}_t$ and also $s_t(\tilde{\mathbf{y}})[\tilde{y}_t]$. With our choice of $\gamma_t(\tilde{\mathbf{y}})$ in eq. (D.55), we can further verify that $\gamma_t(\tilde{\mathbf{y}})+\gamma_t(\tilde{\mathbf{y}})^2 \leq s_t(\tilde{\mathbf{y}})[0]+\gamma_t(\tilde{\mathbf{y}}) \leq v_t(\tilde{\mathbf{y}})$. Hence, after noticing eq. (D.52), we can verify that the conditions in Theorem D.20 hold with $\alpha = \gamma_t(\tilde{\mathbf{y}}), \beta = \gamma_t(\tilde{\mathbf{y}})^2, p = v_t(\tilde{\mathbf{y}})$ and $n = k_t$. Additionally, according to our construction of $p_t(\mathbf{y})$, when conditioned on $y_{1:(k_1+\cdots+k_t)}$, we have

$$y_{k_1+\cdots k_{t-1}+1}, \cdots, y_{k_1+\cdots k_{t-1}+k_t} \overset{\text{i.i.d.}}{\sim} v_t(\cdot \mid \tilde{\mathbf{y}}).$$

Hence eq. (D.46) in Theorem D.20 implies that

$$\mathbb{E}\left[k_t \cdot \left(\hat{v}_t \log \frac{v_t(\tilde{\mathbf{y}}) + \varepsilon(\hat{v}_t)\gamma_t(\tilde{\mathbf{y}}) - \gamma_t(\tilde{\mathbf{y}})^2}{v_t(\tilde{\mathbf{y}})} + (1-\hat{v}_t)\log\frac{1-v_t(\tilde{\mathbf{y}}) - \varepsilon(\hat{v}_t)v_t(\tilde{\mathbf{y}}) - \gamma_t(\tilde{\mathbf{y}})^2}{1-v_t(\tilde{\mathbf{y}})}\right) \,\middle|\, y_{1:(k_1+\cdots k_{t-1})}\right] \geq$$

Bringing this back to eq. (D.57) and then further back to eq. (D.51), we obtain that

$$\mathcal{R}_n(\mathcal{F}) \geq \mathcal{R}_n(\mathcal{F}, \alpha, \alpha^4)/16 \geq d \cdot \frac{1}{5184} = \frac{\mathfrak{D}(\mathcal{F}, \alpha, \alpha^4/16)}{5184} = \Omega(n^{\frac{p}{p+2}}).$$

□

D.4.2 Proof of Theorem 5.11

We present the proof of Theorem 5.11 in this section. We first present a lemma showing that when $f, p \in [c, 1-c]$, we have $h(f,p) \asymp |f-p|$.

Lemma D.21. *Suppose c to be a positive constant in $(0, 1/2)$, then for any $f, p \in [c, 1-c]$, we have*

$$\frac{|f-p|}{\sqrt{2}} \leq h(f,p) \leq \frac{|f-p|}{2\sqrt{c}}.$$

Proof. As for the lower bound part, we notice that

$$\begin{aligned} h(f,p)^2 &\geq \frac{1}{2}\cdot\left(\left(\sqrt{f}-\sqrt{p}\right)^2+\left(\sqrt{1-f}-\sqrt{1-p}\right)^2\right) \\ &= \frac{1}{2}\cdot(f-p)^2\cdot\left(\frac{1}{(\sqrt{f}+\sqrt{p})^2}+\frac{1}{(\sqrt{1-f}+\sqrt{1-p})^2}\right) \\ &\geq \frac{1}{2}\cdot(f-p)^2\cdot\left(\frac{1}{2(f+p)}+\frac{1}{2(2-f-p)}\right) \\ &\geq \frac{1}{2}(f-p)^2, \end{aligned}$$

where the second and third inequalities both use Cauchy-Schwarz inequality.

Next we prove the upper bound part. Since $f,p\in[c,1-c]$, we have

$$(\sqrt{f}+\sqrt{p})^2\geq 4c \quad \text{and} \quad (\sqrt{1-f}+\sqrt{1-p})^2\geq 4c,$$

which implies that

$$h(f,p)\leq (f-p)^2\cdot\frac{1}{4c}=\frac{(f-p)^2}{4c}.$$

□

We are now ready to prove Theorem 5.11.

Proof of Theorem 5.11. Since $\sup_{\mathbf{x}}\mathcal{H}_{\mathsf{sq}}(\mathcal{F},\alpha,n,\mathbf{x})=\tilde{\Omega}(\alpha^{-p})$, by choosing $\beta=\alpha^2/2$ in Theorem D.15, we obtain

$$\mathfrak{D}(\mathcal{F},\alpha,\alpha^2/2)=\tilde{\Omega}\left(\alpha^{-p}\right),\quad \forall\alpha>0.$$

Hence if we choose β such that

$$\beta=\sup\left\{\beta\in(0,1/16):\mathfrak{D}(\mathcal{F},\beta,\beta^2/2)\geq n\right\},$$

we have $\beta=\tilde{\Omega}\left(n^{-1/p}\right)$. And according to Theorem D.15, there exists a depth-n $\mathcal{X}$-valued tree $\mathbf{x}$ shattered by $\mathcal{F}$ at scale $(\beta,\beta^2/2)$, i.e. there exists a depth-n $[0,1]\times[0,1]$-valued tree $\mathbf{s}$ such that for any path $\mathbf{y}=(y_{1:d})\in\{0,1\}^d$, $s_t(\mathbf{y})=(s_t(\mathbf{y})[0],s_t(\mathbf{y})[1])$ with $s_t(\mathbf{y})[0]<s_t(\mathbf{y})[1]$, and also there exists $f^{\mathbf{y}}\in\mathcal{F}$ such that

$$|f^{\mathbf{y}}(x_t(\mathbf{y}))-s_t(\mathbf{y})[y_t]|<\frac{\beta^2}{2}\quad\text{and}\quad h\left(s_t(\mathbf{y})[0],s_t(\mathbf{y})[1]\right)>\beta,\qquad\forall t\in[n]. \tag{D.58}$$

We further define $[0,1]$-valued tree $\mathbf{u}$ where for any $\mathbf{y}\in\{0,1\}^d$ and $t\in[n]$,

$$u_t(\mathbf{y})=\frac{s_t(\mathbf{y})[0]+s_t(\mathbf{y})[1]}{2}.$$

Since $\mathcal{F}\subseteq[7/16,9/16]^{\mathcal{X}}$, according to Theorem D.21 we have for any $\mathbf{y}\in\{0,1\}^n$, $3/8\leq s_t(\mathbf{y})[0]<s_t(\mathbf{y})[1]<5/8$, which implies that

$$s_t(\mathbf{y})[1]-s_t(\mathbf{y})[0]\geq 2\sqrt{3/8}\cdot h\left(s_t(\mathbf{y})[0],s_t(\mathbf{y})[1]\right)>\beta.$$

Hence we have $u_t(\mathbf{y}) \in [7/16, 9/16]$, and for any $\mathbf{y} \in \{0,1\}^n$,

$$f^{\mathbf{y}}(y_t \mid x_t(\mathbf{y})) - u_t(y_t \mid \mathbf{y}) \geq \beta - \frac{\beta^2}{2} \geq \frac{\beta}{2}.$$

We next notice that for any $f \in \mathcal{F}$, $\mathbf{y} \in \{0,1\}^n$ and $t \in [n]$,

$$\frac{f(y_t \mid x_t(\mathbf{y}))}{u_t(y_t \mid \mathbf{y})} \geq f(y_t \mid x_t(\mathbf{y})) \geq 7/16,$$

hence according to inequality $\log(1+x) \geq x - 3x^2/2$ for any $x \geq -9/16$, we have

$$\sum_{t=1}^{n} \log \frac{f(y_t \mid x_t(\mathbf{y}))}{u_t(y_t \mid \mathbf{y})} \geq \sum_{t=1}^{n} \left\{ \left(\frac{f(y_t \mid x_t(\mathbf{y}))}{u_t(y_t \mid \mathbf{y})} - 1 \right) - \frac{3}{2} \cdot \left(\frac{f(y_t \mid x_t(\mathbf{y}))}{u_t(y_t \mid \mathbf{y})} - 1 \right)^2 \right\}. \tag{D.59}$$

We choose $f = f^{\mathbf{y}}$ in the above inequality. After noticing that

$$f^{\mathbf{y}}(y_t \mid x_t(\mathbf{y})) \in \left[\frac{7}{16}, \frac{9}{16} \right], \qquad u_t(y_t \mid \mathbf{y}) \in \left[\frac{7}{16}, \frac{9}{16} \right] \quad \text{and} \quad f^{\mathbf{y}}(y_t \mid x_t(\mathbf{y})) \geq \frac{\beta}{2} + u_t(y_t \mid \mathbf{y}),$$

we have

$$0 < \frac{f^{\mathbf{y}}(y_t \mid x_t(\mathbf{y}))}{u_t(y_t \mid \mathbf{y})} - 1 \leq \frac{1 - 7/16}{7/16} - 1 = \frac{2}{7}.$$

Therefore, we have

$$\begin{aligned}
&\left(\frac{f^{\mathbf{y}}(y_t \mid x_t(\mathbf{y}))}{u_t(y_t \mid \mathbf{y})} - 1 \right) - \frac{3}{2} \cdot \left(\frac{f^{\mathbf{y}}(y_t \mid x_t(\mathbf{y}))}{u_t(y_t \mid \mathbf{y})} - 1 \right)^2 \\
&\geq \frac{4}{7} \left(\frac{f^{\mathbf{y}}(y_t \mid x_t(\mathbf{y}))}{u_t(y_t \mid \mathbf{y})} - 1 \right) \\
&\overset{(i)}{\geq} f^{\mathbf{y}}(y_t \mid x_t(\mathbf{y})) - u_t(y_t \mid \mathbf{y}) \geq \frac{\beta}{2},
\end{aligned}$$

where inequality (i) uses the fact that $u_t(y_t \mid \mathbf{y}) \leq 9/16 < 4/7$. Bringing back to eq. (D.59), we obtain that

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_{\mathbf{y} \sim \mathbf{p}} \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \log \frac{f^{\mathbf{y}}(y_t \mid x_t(\mathbf{y}))}{u_t(y_t \mid \mathbf{y})} \right] \geq \frac{n\beta}{2} = \tilde{\Omega}\left(n^{\frac{p-1}{p}} \right).$$

□

D.5 Missing Proofs in section 5.3.4

We first prove Theorem 5.12.

Proof of Theorem 5.12. Recall from Theorem D.3 we have

$$\mathcal{R}_n(\mathcal{F}) = \sup_{\mathbf{x}} \sup_{\mathbf{y}\sim\mathbf{p}} \left[\sum_{t=1}^{n} \log \frac{1}{p_t(y_t \mid \mathbf{y})} - \inf_{f\in\mathcal{F}} \sum_{t=1}^{n} \log \frac{1}{f(y_t \mid x_t(\mathbf{y}))} \right].$$

Hence we only need to prove that for any path $\mathbf{y} = (y_{1:n}) \in \{0,1\}^n$,

$$\sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \log f(y_t \mid x_t(\mathbf{y})) \le \sup_{f\in\mathcal{F}_{1/n}} \sum_{t=1}^{n} \log f(y_t \mid x_t(\mathbf{y})) + 2. \tag{D.60}$$

According to the definition of Hilbert ball class eq. (5.17), for any $f \in \mathcal{F}$, there exists $w \in B_2(1)$ such that

$$f(y_t \mid x_t(\mathbf{y})) = \frac{1 + (-1)^{y_t}\langle x_t(\mathbf{y}), w\rangle}{2}.$$

Next, we notice that for any real number $a \in (-1,1)$, we have

$$\begin{aligned}\log \frac{a+1}{2} &\le \log \frac{a + n/(n-1)}{2} \le \log \frac{(1-1/n)a+1}{2} + \log \frac{n}{n-1} \\ &\le \log \frac{(1-1/n)a+1}{2} + \frac{1}{n-1}.\end{aligned}$$

Therefore, we obtain

$$\frac{1 + (-1)^{y_t}\langle x_t(\mathbf{y}), w\rangle}{2} \le \frac{1 + (-1)^{y_t}\langle x_t(\mathbf{y}), (1-1/n)w\rangle}{2} + \frac{1}{n-1},$$

which implies that

$$\sum_{t=1}^{n} \frac{1 + (-1)^{y_t}\langle x_t(\mathbf{y}), w\rangle}{2} \le \sum_{t=1}^{n} \frac{1 + (-1)^{y_t}\langle x_t(\mathbf{y}), (1-1/n)w\rangle}{2} + \frac{n}{n-1}.$$

Since for any $w \in B_2(1)$, we always have $(1-1/n)w \in B_2(1-1/n)$, according to the definition of function class $\mathcal{F}_{1/n}$ we have

$$\begin{aligned}&\sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \log f(y_t \mid x_t(\mathbf{y})) \\ &\quad\le \sup_{f\in\mathcal{F}_{1/n}} \sum_{t=1}^{n} \log f(y_t \mid x_t(\mathbf{y})) + \frac{n}{n-1} \\ &\quad\le \sup_{f\in\mathcal{F}_{1/n}} \sum_{t=1}^{n} \log f(y_t \mid x_t(\mathbf{y})) + 2,\end{aligned}$$

which proves eq. (D.60). □

We next prove Theorem 5.13. Our proof requires the following definition of skipping binary tree.

Definition D.22. *For a given binary tree* $\mathbf{x}$*, we say a binary tree* $\mathbf{y}$ *is a skipping tree of* $\mathbf{x}$ *if*

1. *The set of vertices of* $\mathbf{y}$ *is a subset of the set of vertices of* $\mathbf{x}$*.*
2. *For two vertices* a, b *of* $\mathbf{y}$*, if* a *is* b*'s left child in* $\mathbf{y}$*, then* a *is a descendant of* b*'s left child in* $\mathbf{x}$*; and if* a *is* b*'s right child in* $\mathbf{y}$*, then* a *is a descendant of* b*'s right child in* $\mathbf{x}$*.*

We have the following properties of coloring over binary trees.

Lemma D.23. *We consider* k*-coloring over the vertices of a depth-*n *binary tree* $\mathbf{x}$*, where each vertices has been colored in one of* k *colors. Then when* $n \geq k(d-1)+1$*,* $\mathbf{x}$ *has a skipping tree of depth* d *whose nodes are of the same color.*

Proof. We prove a stronger result: For integers $d_1, d_2, \ldots, d_k \geq 0$, if binary tree $\mathbf{x}$ has depth at least $d_1 + d_2 + \cdots + d_k + 1$, and is colored in $1, \ldots, k$. Then there exists $1 \leq i \leq k$, such that $\mathbf{x}$ has a skipping tree of depth $d_i + 1$ whose vertices are all in color i.

We will prove this result by induction on $d_1 + \cdots + d_k$. When $d_1 + \cdots + d_k = 0$, we have $d_1 = \cdots = d_k = 0$. In this case, assuming the root is colored in i, then the root itself is a skipping tree with depth $d_i + 1$.

Next we assume that this result holds when $d_1 + \cdots + d_k = m$. When $d_1 + \cdots + d_k = m+1$, we assume the root a is colored in j $(1 \leq j \leq k)$. If $d_j = 0$, then we already have a skipping tree with only one vertex a in color j at depth $d_j + 1$. Next, we assume that $d_j \geq 1$. We consider the left binary tree $\mathbf{x}_1$ rooted at the left child of a, and also the right binary tree $\mathbf{x}_2$ rooted at the right child of a. Then both $\mathbf{x}_1$ and $\mathbf{x}_2$ are binary trees with depth

$$m = d_1 + \cdots + d_{j-1} + (d_i - 1) + d_{j+1} + \cdots + d_k = \hat{d}_1 + \cdots + \hat{d}_k.$$

where we let $\hat{d}_i = d_i$ if $i \neq j$ and $\hat{d}_i = d_i - 1$ if $i = j$. According to induction, $\exists 1 \leq i_1, i_2 \leq k$ such that there exists $\mathbf{x}_1$'s skipping tree $\mathbf{u}_1$ of depth $\hat{d}_{i_1} + 1$ whose vertices are all in color i_1 and also $\mathbf{x}_2$'s skipping tree $\mathbf{u}_2$ of depth $\hat{d}_{i_2} + 1$ whose vertices are all in color i_2. If $i_1 \neq j$, then we have $\hat{d}_{i_1} = d_{i_1}$. Hence the skipping tree $\mathbf{u}_1$ is also a skipping tree of $\mathbf{x}$ with depth $d_{i_1} + 1$ whose vertices are all in color i_1. This skipping tree is desirable. If $i_2 \neq j$, similarly we can also find a desirable skipping tree of $\mathbf{x}$. Finally if $i_1 = i_2 = j$, we consider the tree $\mathbf{y}$ with root j, and two subtrees of j's left child and right child to be $\mathbf{u}_1$ and $\mathbf{u}_2$. Then since the root of $\mathbf{u}_1$ is a descendant of j's left child the root of $\mathbf{u}_2$ is a descendant of j's right child, $\mathbf{y}$ is a skipping tree of $\mathbf{x}$. And we further know that vertices of $\mathbf{y}$ are all in color j and the depth of $\mathbf{y}$ is $d_j + 1$. Therefore, $\mathbf{y}$ is a desirable skipping tree. And we have finished proving the result for $d_1 + \cdots + d_k = m + 1$.

According to induction, this result holds for any $d_1, \ldots, d_n \geq 0$. □

Now we are ready to prove Theorem 5.13.

Proof of Theorem 5.13. First by choosing $\beta = \alpha^2/2$ in Theorem D.16 and replacing α by 4α, we obtain

$$\sup_{\mathbf{x}} \mathcal{H}_{\mathsf{sq}}(\mathcal{F}_{1/n}, 5\alpha, n, \mathbf{x}) \leq \mathfrak{D}(\mathcal{F}, 4\alpha, \alpha^2/2) \cdot \log\left(\frac{2en}{\alpha^2}\right).$$

Therefore, in order to prove Theorem 5.13, we only need to verify

$$\mathfrak{D}(\mathcal{F}_{1/n}, 4\alpha, \alpha^2/2) = \mathcal{O}\left(\frac{\log n}{\alpha^2}\right). \tag{D.61}$$

We let tree $\mathbf{x}'$ of depth d' to be the largest tree shattered by $\mathcal{F}_{1/n}$ at scale $(4\alpha, \alpha^2/2)$. Without loss of generality we assume d' is an odd number and $d' = 2d+1$. Then there exists a depth-d' $([0,1]\times[0,1])$-valud tree $\mathbf{s}'$ such that for any path $\mathbf{y}' \in \{0,1\}^{d'}$, $s'_t(\mathbf{y}') = (s'_t(\mathbf{y}')[0], s'_t(\mathbf{y}')[1])$ with $s'_t(\mathbf{y}')[0] < s'_t(\mathbf{y}')[1]$, and for any $\mathbf{y}' \in \{0,1\}^{d'}$, there exists $w^{\mathbf{y}'} \in B_2(1-1/n)$ such that

$$\left|\frac{1+\langle w^{\mathbf{y}'}, x'_t(\mathbf{y}')\rangle}{2} - s'_t(\mathbf{y}')[y_t]\right| < \frac{\alpha^2}{2} \quad \text{and} \quad h\left(s'_t(\mathbf{y}')[0], s'_t(\mathbf{y}')[1]\right) > 4\alpha, \qquad \forall t \in [d']. \tag{D.62}$$

We further construct a depth-d' $[0,1]$-valued tree $\mathbf{u}'$ where for any $\mathbf{y}' \in \{0,1\}^{d'}$ and $t \in [d']$,

$$s'_t(\mathbf{y}')[0] < u'_t(\mathbf{y}') < s'_t(\mathbf{y}')[1], \qquad h(u'_t(\mathbf{y}'), s'_t(\mathbf{y}')[0]) \geq 2\alpha \quad \text{and} \quad h(u'_t(\mathbf{y}'), s'_t(\mathbf{y}')[1]) \geq 2\alpha.$$

According to eq. (D.62), we have for any $\mathbf{y}' \in \{0,1\}^{d'}$ and $t \in [d']$,

$$(2y_t - 1)\cdot\left(\frac{1+\langle w^{\mathbf{y}'}, x'_t(\mathbf{y}')\rangle}{2} - u'_t(\mathbf{y}')\right) \geq 0 \quad \text{and} \quad h\left(\frac{1+\langle w^{\mathbf{y}'}, x'_t(\mathbf{y}')\rangle}{2}, u'_t(\mathbf{y}')\right) \geq \alpha.$$

We color the tree $\mathbf{x}'$ with two colors according to $u_t(\mathbf{y}')$: for each node $x'_t(\mathbf{y}')$, if $u'_t(\mathbf{y}') \leq 1/2$ we color it with color 1, otherwise we color it with color 0. According to Theorem D.23, there exists a skipping tree of depth $d = (d'-1)/2$ such that every node in this tree are of the same color. Without loss of generality, we assume that the skipping tree is in color 1. In the following, we only consider nodes of $\mathbf{x}'$ in this skipping tree, and also the corresponding nodes (along the same path) of $\mathbf{u}'$. And we obtain a tree $\mathbf{x}$ of depth d and an $[0,1/2]$-valued tree $\bar{\mathbf{u}}$ of depth d such that for any $\mathbf{y} \in \{0,1\}^d$, there exists $w \in B_2(1-1/n)$ such that for any $t \in [d]$,

$$(2y_t - 1)\cdot\left(\frac{1+\langle w, x_t(\mathbf{y})\rangle}{2} - \bar{u}_t(\mathbf{y})\right) \geq 0 \quad \text{and} \quad h\left(\frac{1+\langle w, x_t(\mathbf{y})\rangle}{2}, \bar{u}_t(\mathbf{y})\right) \geq \alpha.$$

We next notice that since function $2\sqrt{x}$ and $\log x - 2\sqrt{x}$ are both monotonically increasing function over $[0,1]$, for any $p, f \in [0,1]$ we have

$$|\log p - \log f| \geq 2|\sqrt{p} - \sqrt{f}|.$$

Additionally, since for any $p \in [0,1/2]$ and $f \in [0,1]$, we always have

$$|\sqrt{p} - \sqrt{f}| = \frac{|p-f|}{\sqrt{p}+\sqrt{f}} \geq \frac{(\sqrt{2}+1)|p-f|}{\sqrt{1-p}+\sqrt{1-f}} = (\sqrt{2}+1)\cdot\left|\sqrt{1-p} - \sqrt{1-f}\right|,$$

which implies that $|\sqrt{p} - \sqrt{f}| \geq h(f,p)/4$. Hence we obtain

$$\log f - \log p \geq \frac{h(f,p)}{2}.$$

By letting depth-d $\mathbb{R}$-valued tree $\mathbf{s}$ to be $s_t(\mathbf{y}) = \log \bar{u}_t(\mathbf{y})$ for any path $\mathbf{y} \in \{0,1\}^d$, we have that for $\mathbf{y} \in \{0,1\}^d$, there exists $w \in B_2(1-1/n)$ such that for any $t \in [d]$,

$$(2y_t - 1) \cdot \left(\log \frac{1 + \langle w, x_t(\mathbf{y}) \rangle}{2} - s_t(\mathbf{y}) \right) \geq \frac{\alpha}{2}$$

Since $x_t(\mathbf{y})$ and $s_t(\mathbf{y})$ only depend on $y_{1:t-1}$, by choosing $y_t = 1$, we obtain for some $w \in B_2(1-1/n)$,

$$\log \frac{1 + \langle w, x_t(\mathbf{y}) \rangle}{2} - s_t(\mathbf{y}) \geq \frac{\alpha}{2} > 0,$$

which implies that

$$s_t(\mathbf{y}) < \log \frac{1 + \langle w, x_t(\mathbf{y}) \rangle}{2} \leq \log \frac{1 + \|w\|_2 \|x_t(\mathbf{y})\|_2}{2} \leq \log \frac{1+1}{2} = 0.$$

Similarly by choosing $y_t = 0$, we obtain for some $w \in B_2(1-1/n)$,

$$\log \frac{1 + \langle w, x_t(\mathbf{y}) \rangle}{2} - s_t(\mathbf{y}) \leq -\frac{\alpha}{2} < 0,$$

which implies that

$$s_t(\mathbf{y}) > \log \frac{1 + \langle w, x_t(\mathbf{y}) \rangle}{2} \geq \log \frac{1 - \|w\|_2 \|x_t(\mathbf{y})\|_2}{2} \geq \log \frac{1 - (1-1/n)}{2} = -\log(2n).$$

Hence we obtain that for any t, we always have $s_t(\mathbf{y}) \in (-\log(2n), 0)$.

Next, we will color the binary tree $\mathbf{x}$ with $\lceil \log(2n) \rceil$ number of colors $0, 1, \ldots, \lfloor \log(2n) \rfloor$: for $\mathbf{y} \in \{0,1\}^d$, if $s_t(\mathbf{y}) \in [-k-1, -k)$, we will color vertex $x(\mathbf{y})$ in color k. According to Lemma D.23, there exists some i such that there exists a skipping tree $\mathbf{v}$ of depth $\bar{d} \geq \frac{d-1}{\lceil \log(2n) \rceil}$ whose nodes are all colored in k.

We consider a sequence of nodes $v_1(\mathbf{y}) \to v_t(\mathbf{y}) \to \cdots \to v_{\bar{d}}(\mathbf{y})$ in the skipping tree $\mathbf{v}$. Here we assume $v_{i+1}(\mathbf{y})$ is the left child or descendant of the left child of v_i if $y_i = 0$, or the right child or the descendant of the right child of v_i if $y_i = 1$. We let

$$v_i(\mathbf{y}) = x_{i,1} \to \cdots \to x_{i,l_i} = v_{i+1}(\mathbf{y}) \tag{D.63}$$

to be the sequence of nodes in tree $\mathbf{x}$ from $v_i(\mathbf{y})$ to $v_{i+1}(\mathbf{y})$, where $x_{1,2}$ is the right child of $x_{1,1}$, and $x_{i,j}$ is a child of $x_{i,j-1}$ (since $v_{i+1}(\mathbf{y})$ is a descendant of $v_i(\mathbf{y})$, there must exist such a path). We consider the following sequence of nodes in tree $\mathbf{x}$:

$$x_{1,1} \to x_{1,2} \cdots \to x_{1,l_1} \to x_{2,2} \to \cdots \to x_{2,l_2} \to x_{3,2} \to \cdots \to x_{3,l_3} \to \cdots x_{\bar{d},2} \to \cdots x_{\bar{d},l_{\bar{d}}}. \tag{D.64}$$

We define length-d $\{0,1\}$-valued path

$$\tilde{\mathbf{y}} = (\tilde{y}_{1,1}, \tilde{y}_{1,2}, \cdots, \tilde{y}_{1,l_1}, \tilde{y}_{2,1}, \cdots \tilde{y}_{2,l_2-1}, \tilde{y}_{3,1}, \cdots, \tilde{y}_{3,l_3-1}, \cdots, \tilde{y}_{\bar{d},1}, \cdots, \tilde{y}_{\bar{d},l_{\bar{d}}-1}, y_{\tilde{y}+1,1}, \cdots, y_{\bar{d}+1,l_{\bar{d}+1}-1}),$$

where $\tilde{y}_{i,j}$ is chosen to be 1 if $x_{i,j}$ is the right child of $x_{i,j-1}$ and be 0 if $x_{i,j}$ is the left child of $x_{i,j-1}$, and $y_{\bar{d}+1,1}, \cdots, y_{\bar{d}+1,l_{\bar{d}+1}-1}$ can be arbitrarily chosen with $l_{\bar{d}+1} - 1 = d - l_1 - \cdots - l_{\bar{d}} + \bar{d}$. Then according to the construction of this path we have

$$\tilde{y}_{i,1} = y_i, \quad \forall 1 \leq i \leq \bar{d}.$$

Suppose the vertices we meet in tree $\mathbf{s}$ along path $\tilde{y}$ at the same depth as $x_{i,j}$ to be $s_{i,j}(\tilde{\mathbf{y}})$. Then according to our assumption, there exists some $w \in B_2(1-1/n)$ such that

$$(2\tilde{y}_{i,j}-1)\cdot\left(\log\frac{1+\langle w, x_{i,j}\rangle}{2} - s_{i,j}(\tilde{\mathbf{y}})\right) \geq \frac{\alpha}{2}, \qquad \forall 1\leq i\leq \bar{d}, 1\leq j\leq l_i. \tag{D.65}$$

We further define depth-$\bar{d}$ $\mathbb{R}$-tree $\mathbf{u}=(u_1,\ldots,u_{\bar{d}})$ as

$$u_i(\mathbf{y}) = s_{i,1}(\tilde{\mathbf{y}}).$$

Choosing $j=1$ in eq. (D.65) and notice that $v_i(\mathbf{y})=x_{i,1}$ from eq. (D.63) and $y_i=\tilde{y}_{i,1}$ from eq. (D.64), we obtain

$$(2y_i-1)\left(\log\frac{1+\langle w, v_i(\mathbf{y})\rangle}{2} - u_i(\mathbf{y})\right) \geq \frac{\alpha}{2}, \quad \forall 1\leq i\leq \bar{d}.$$

According to our coloring and the definition of $s_{i,j}(\tilde{\mathbf{y}})$, we know that $u_i(\mathbf{y})=s_{i,1}(\tilde{\mathbf{y}})\in[-k-1,-k)$ holds for any $1\leq i\leq\bar{d}$. Therefore, for any $\mathbf{y}\in\{0,1\}^{\bar{d}}$, there exists $w\in B_2(1-1/n)$ such that for any $i\in[\bar{d}]$,

$$(2y_i-1)\cdot\left(\log\frac{1+\langle w, v_i(\mathbf{y})\rangle}{2} - u_i(\mathbf{y})\right) \geq \frac{\alpha}{2} \quad \text{and} \quad u_i(\mathbf{y})\in[-k-1,-k). \tag{D.66}$$

The above inequality is equivalent to for any $\mathbf{y}\in\{0,1\}^{\bar{d}}$, there exists $w\in B_2(1-1/n)$ such that for any $i\in[\bar{d}]$,

$$\langle w, v_t(\mathbf{y})\rangle \geq 2e^{u_t(\mathbf{y})}e^{\alpha/2}-1 \quad \text{if } y_t=1, \quad \text{and} \quad \langle w, v_t(\mathbf{y})\rangle \leq 2e^{u_t(\mathbf{y})}e^{-\alpha/2}-1 \quad \text{if } y_t=0. \tag{D.67}$$

For any given path $\mathbf{y}\in\{0,1\}^{\bar{d}}$, we call the $w\in B_2(1-1/n)$ which satisfies the above inequalities as $w[\mathbf{y}]$. We use $v_0=v_1(\mathbf{y})$ to denote the root of the tree $\mathbf{y}$. Then for any path $\mathbf{y}=(y_1,\ldots,y_{\bar{d}})$ with $y_1=0$, i.e. turn to left subtree in the first step, according to eq. (D.67) we have

$$\langle w[\mathbf{y}], v_0\rangle = \langle w[\mathbf{y}], v_1(\mathbf{y})\rangle \leq 2e^{u_1(\mathbf{y})}e^{-\alpha/2}-1 \leq 2e^{-k}-1,$$

where in the last inequality we uses the second inequality in eq. (D.66).

In the following, for every vector v, we decompose it into the parallel component and perpendicular component with respect to vector v_0: $v=v^{\perp}+v^{\parallel}$, where $v^{\parallel}\parallel v_0$ and $v^{\perp}\perp v_0$. Then we have

$$\|w[\mathbf{y}]^{\parallel}\|_2\|v_0\|_2 = |\langle w[\mathbf{y}], v_0\rangle| = 1-2e^{-k}.$$

Noticing that $\|v_0\|_2, \|w[\mathbf{y}]^{\parallel}\|_2\leq 1$, we will have $\|v_0\|_2, \|w[\mathbf{y}]^{\parallel}\|_2\geq 1-2e^{-k}$, hence

$$\|w[\mathbf{y}]^{\parallel}+v_0\|_2 \leq 1-(1-2e^{-k}) = 2e^{-k} \quad \text{and} \quad \|w[\mathbf{y}]^{\perp}\|_2 \leq \sqrt{1-(1-2e^{-k})^2} \leq 2e^{-k/2}. \tag{D.68}$$

Next, we consider any node $v_t(\mathbf{y})$ on the left subtree of $\mathbf{y}$, where we require the path $\mathbf{y}$ to the node satisfies $y_1=0$. By letting $y_t=0$ (since $v_t(\mathbf{y})$ does not depends on y_t so we can assign arbitrary value of y_t to obtain some properties of $v_t(\mathbf{y})$), according to (D.67) and the second inequality of eq. (D.66), we obtain

$$\langle w[\mathbf{y}], v_t(\mathbf{y})\rangle \leq 2e^{u_t(\mathbf{y})}e^{-\alpha/2}-1 \leq 2e^{-k}-1,$$

which implies

$$\|w[\mathbf{y}] + v_t(\mathbf{y})\|_2 \leq \sqrt{\|w[\mathbf{y}]\|_2^2 + \|v_t(\mathbf{y})\|_2^2 + 2\langle w[\mathbf{y}], v_t(\mathbf{y})\rangle} \leq \sqrt{2 + 2(2e^{-k} - 1)} = 2e^{-k/2},$$

Choosing $t = 1$ in the above inequality we obtain $\|w[\mathbf{y}]+v_0\|_2 \leq 2e^{-k/2}$. These two inequalities together indicates that

$$\|v_t(\mathbf{y}) - v_0\|_2 \leq 4e^{-k/2}.$$

Hence we have,

$$\|v_t(\mathbf{y})^\perp\|_2 \leq \|v_t(\mathbf{y}) - v_0\|_2 \leq 4e^{-k/2}. \tag{D.69}$$

Next, we decompose the inner product into the sum of inner product of parallel components and perpendicular components:

$$\langle w[\mathbf{y}], v_t(\mathbf{y})\rangle = \langle w[\mathbf{y}]^\parallel, v_t(\mathbf{y})^\parallel\rangle + \langle w[\mathbf{y}]^\perp, v_t(\mathbf{y})^\perp\rangle.$$

Noticing $\|w[\mathbf{y}]^\perp\|_2 \leq 2e^{-k/2}$ and $\|v_t(\mathbf{y})^\perp\|_2 \leq 4e^{-k/2}$ from eq. (D.68) and eq. (D.69), we obtain that

$$\langle w[\mathbf{y}]^\parallel, v_t(\mathbf{y})^\parallel\rangle = \langle w[\mathbf{y}], v_t(\mathbf{y})\rangle - \langle w[\mathbf{y}]^\perp, v_t(\mathbf{y})^\perp\rangle \leq 2e^{-k} - 1 + 2e^{-k/2} \cdot 4e^{-k/2} = 10e^{-k} - 1.$$

Since $\|w[\mathbf{y}]^\parallel\|_2 \leq 1$ and $\|v_t(\mathbf{y})^\parallel\|_2 \leq 1$, we have $\|w[\mathbf{y}]^\parallel\|_2, \|v_t(\mathbf{y})^\parallel\|_2 \geq 1 - 10e^{-k}$. Hence,

$$\|w[\mathbf{y}]^\parallel + v_t(\mathbf{y})^\parallel\|_2 \leq 1 - (1 - 10e^{-k}) = 10e^{-k},$$

This inequality together with the first inequality of eq. (D.68) indicates that

$$\|v_2(\mathbf{y})^\parallel - v_0\|_2 \leq 12e^{-k}.$$

Finally, we construct a tree $\mathbf{z}$ of depth $\bar{d} - 1$ shattered by the following function class $\mathcal{G}$ at scale $1/(20e)$ (definition of the shattering in the sense of [214]), hence according to [224, Page 67-68] we have an upper bound on $\bar{d}$.

$$\mathcal{G} = \{f | f(x) = \langle w, x\rangle, w, x \in B_2(1)\} \tag{D.70}$$

For any $\mathbf{y} \in \{0,1\}^{\bar{d}-1}$, we let

$$z_t(\mathbf{y}) = \frac{1}{5}e^{k/2}v_{t+1}((0,\mathbf{y}))^\perp + \frac{1}{5}v_{t+1}((0,\mathbf{y}))^\parallel, \quad \forall 1 \leq t \leq \bar{d} - 1, \tag{D.71}$$

where we use $(0,\mathbf{y})$ to denote the path of length $\bar{d}$ whose t-th element equals to y_{t-1} for $t \geq 2$ and the first element equals to 0. Then according to eq. (D.69), for any path $\mathbf{y}$ we have

$$\|z_t(\mathbf{y})\|_2 \leq \frac{1}{5}e^{k/2}\|v_{t+1}((0,\mathbf{y}))^\perp\|_2 + \frac{1}{5}\|v_{t+1}((0,\mathbf{y}))^\parallel\|_2 \leq \frac{4}{5} + \frac{1}{5} = 1,$$

which implies that $z_t(\mathbf{y}) \in B_2(1)$. We further notice from eq. (D.68) that

$$\|w((0,\mathbf{y}))^\parallel + v_0\|_2 \leq 2e^{-k} \quad \text{and} \quad \|w((0,\mathbf{y}))^\perp\|_2 \leq 2e^{-k/2}.$$

Hence by choosing

$$\bar{w}(\mathbf{y}) = \frac{1}{4}e^k \left(w((0,\mathbf{y}))^{\|} + v_0\right) + \frac{1}{4}e^{k/2} w((0,\mathbf{y}))^{\perp}, \tag{D.72}$$

we have

$$\|\bar{w}(\mathbf{y})\|_2 \le \frac{1}{4}e^k \left\|(w((0,\mathbf{y}))^{\|} + v_0\right\|_2 + \frac{1}{4}e^{k/2} \left\|w((0,\mathbf{y}))^{\perp}\right\|_2 \le \frac{2}{4} + \frac{2}{4} = 1,$$

which implies that $\bar{w}(\mathbf{y}) \in B_2(1)$. According to our choice of $\bar{w}(\mathbf{y})$ in eq. (D.72) and $z_t(\mathbf{y})$ in eq. (D.71), we have

$$\begin{aligned}
\langle \bar{w}(\mathbf{y}), z_t(\mathbf{y})\rangle &= \langle \bar{w}(\mathbf{y})^{\|}, z_t(\mathbf{y})^{\|}\rangle + \langle \bar{w}(\mathbf{y})^{\perp}, z_t(\mathbf{y})^{\perp}\rangle \\
&= \frac{1}{20}e^k \left\langle w((0,\mathbf{y}))^{\|} + v_0, v_{t+1}((0,\mathbf{y}))^{\|}\right\rangle + \frac{1}{20}e^k \left\langle w((0,\mathbf{y}))^{\perp}, v_{t+1}((0,\mathbf{y}))^{\perp}\right\rangle \\
&= \frac{1}{20}e^k \cdot \left(\langle w((0,\mathbf{y})), v_{t+1}((0,\mathbf{y}))\rangle + \langle v_0, v_{t+1}((0,\mathbf{y}))\rangle\right).
\end{aligned}$$

We construct another $(\bar{d}-1)$-depth $\mathbb{R}$-valued tree $\bar{\mathbf{s}}$ as: for any path $\mathbf{y} \in \{0,1\}^{\bar{d}-1}$,

$$\bar{s}_t(\mathbf{y}) = \frac{1}{20}e^k e^{u_{t+1}((0,\mathbf{y}))}\left(e^{\alpha/2} + e^{-\alpha/2}\right) + \langle v_0, v_{t+1}((0,\mathbf{y}))\rangle - \frac{1}{20}e^k.$$

The above defined $\bar{\mathbf{s}}$ is a tree since $\mathbf{u}$ and $\mathbf{v}$ are both trees. When $y_t = 1$, according to the first inequality of eq. (D.67) and also $u_{t+1}((0,\mathbf{y})) \ge -k-1$ according to the second inequality of eq. (D.66), we have

$$\begin{aligned}
\langle \bar{w}(\mathbf{y}), z_t(\mathbf{y})\rangle - \bar{s}_t(\mathbf{y}) &= \frac{1}{20}e^k \cdot \left(\langle w((0,\mathbf{y})), v_{t+1}((0,\mathbf{y}))\rangle - e^{u_{t+1}((0,\mathbf{y}))}\left(e^{\alpha/2} + e^{-\alpha/2}\right) + 1\right) \\
&\ge \frac{1}{20}e^k \cdot \left(2e^{u_{t+1}((0,\mathbf{y}))}e^{\alpha/2} - 1 - e^{u_{t+1}((0,\mathbf{y}))}\left(e^{\alpha/2} + e^{-\alpha/2}\right) + 1\right) \\
&= \frac{1}{20}e^k e^{u_{t+1}((0,\mathbf{y}))}(e^{\alpha/2} - e^{-\alpha/2}) \\
&\ge \frac{1}{20}e^k e^{u_{t+1}((0,\mathbf{y}))}\alpha \ge \frac{1}{20}e^k e^{-k-1}\alpha = \frac{1}{20e}\alpha,
\end{aligned}$$

where in the second inequality we use the fact that $e^{\alpha/2} - e^{-\alpha/2} \ge \alpha$. And when $y_t = 0$, we have

$$\begin{aligned}
\langle \bar{w}(\mathbf{y}), z_t(\mathbf{y})\rangle - \bar{s}_t(\mathbf{y}) &= \frac{1}{20}e^k \cdot \left(\langle w((0,\mathbf{y})), v_{t+1}((0,\mathbf{y}))\rangle - e^{u_{t+1}((0,\mathbf{y}))}\left(e^{\alpha/2} + e^{-\alpha/2}\right) + 1\right) \\
&\le \frac{1}{20}e^k \cdot \left(2e^{u_{t+1}((0,\mathbf{y}))}e^{-\alpha/2} - 1 - e^{u_{t+1}((0,\mathbf{y}))}\left(e^{\alpha/2} + e^{-\alpha/2}\right) + 1\right) \\
&= -\frac{1}{20}e^k e^{u_{t+1}((0,\mathbf{y}))}(e^{\alpha/2} - e^{-\alpha/2}) \\
&\le -\frac{1}{20}e^k e^{u_{t+1}((0,\mathbf{y}))}\alpha \le -\frac{1}{20}e^k e^{-k-1}\alpha = -\frac{1}{20e}\alpha.
\end{aligned}$$

Therefore, tree $\mathbf{z} \in B_2(1)$ is shattered by function class $\mathcal{G}$ (defined in eq. (D.70)) at scale $1/(20e)\alpha$.

According to [224, Page 67-68], the sequential fat shattering dimension of $\mathcal{G}$ at scale α is upper bounded by $16/\alpha^2$. Hence we have

$$\bar{d} - 1 \le \frac{16}{(1/(20e)\alpha)^2} = \frac{6400e^2}{\alpha^2}.$$

This inequality, together with the fact that $\bar{d} \ge \frac{d-1}{\lceil \log(2n) \rceil}$, implies that

$$d \le 1 + \lceil \log(2n) \rceil \cdot \left(1 + \frac{6400e^2}{\alpha^2}\right) = \mathcal{O}\left(\frac{\log n}{\alpha^2}\right).$$

Therefore, we have

$$\mathfrak{D}(\mathcal{F}_{1/n}, 4\alpha, \alpha^2/2) = \mathcal{O}\left(\frac{\log n}{\alpha^2}\right),$$

which verifies eq. (D.61). □

D.6 Renewal Process and Hardness through Sequential Square-root Entropy

We consider the following class of renewal process, originally introduced in [111].

Definition D.24 (Renewal Process Class [111])**.** *This class $\mathcal{Q}$ is defined over the alphabet $\mathcal{Y} = \{0, 1\}$ and parameterized by a distribution $p \in \Delta(\mathbb{Z}_+)$. Given p, we sample $T_i \overset{iid}{\sim} p$ and set $y_t = 1$ if $t = T_1 + \cdots + T_i$ for some $i \ge 1$ and otherwise $y_t = 0$.*

For this class $\mathcal{Q}$ the work [111] established that log-loss regret is $\Theta(\sqrt{n})$. Their proof leveraged sophisticated estimates on the partition number by Hardy and Ramanujan. Unfortunately, as we show in this appendix, the entropic bounds that we developed in this work, as well as those that were proposed before, are not able to yield correct upper bound on regret.

Specifically, we will verify that the sequential square-root entropy (defined in Theorem 5.1), and also the sequential log entropy defined in [4, 16] are both $\Omega(n)$, no matter what scale we choose. Therefore, by simply applying Theorem 5.2 or the entropy bound in [4, 16] will only give a vacuous bound $O(n)$ on regret.

Proposition D.25. *For any $0 < \alpha < 1/6$, we have $\mathcal{H}_{\mathsf{sq}}(\mathcal{Q}, \alpha, n) \ge n$. As for the log entropy (entropy with respect to distance eq. (5.7)) defined in [4, 16], we have $\mathcal{H}_{\mathsf{log}}(\mathcal{Q}, \alpha, n) \ge (1 - \log 2)n - o(n)$.*

Proof. For any $\boldsymbol{\varepsilon} \in \{-1, 1\}^n$, we construct a distribution $p^{\boldsymbol{\varepsilon}} \in \Delta(\mathbb{Z}_+)$ as

$$p^{\boldsymbol{\varepsilon}}(t) = \prod_{i=1}^{t-1}\left(\frac{1}{2} + 3\varepsilon_t \cdot \alpha\right) - \prod_{i=1}^{t}\left(\frac{1}{2} + 3\varepsilon_t \cdot \alpha\right).$$

It is easy to see that $p^{\boldsymbol{\varepsilon}}$ is a distribution on $\mathbb{Z}_+$. We let $q^{\boldsymbol{\varepsilon}}$ to be the distribution in $\mathcal{Q}$ which is parametrized by $p^{\boldsymbol{\varepsilon}}$. Then we can calculate that with $\mathbf{y}^0 = (0, 0, \cdots, 0) \in \{0, 1\}^n$,

$$q_t^{\boldsymbol{\varepsilon}}(0 \mid \mathbf{y}^0) = \frac{1}{2} + 3\varepsilon_t \cdot \alpha, \quad \forall t \in [n].$$

We first lower bound the sequential square-root entropy (defined in Theorem 5.1). Suppoose $\mathcal{V}$ is a finite cover of $\mathcal{Q}$ at scale α. Then for $\boldsymbol{\varepsilon} \in \{-1,1\}^n$, there exists $\mathbf{v}^{\boldsymbol{\varepsilon}} \in \mathcal{V}$ such that

$$\max_{\mathbf{w}} \max_{y\in\{0,1\}} \max_{t\in[n]} \left| \sqrt{v_t^{\boldsymbol{\varepsilon}}(y_t \mid \mathbf{w})} - \sqrt{q_t^{\boldsymbol{\varepsilon}}(y_t \mid \mathbf{w})} \right| \le \alpha,$$

which implies that

$$\max_{t\in[n]} \left| \sqrt{v_t^{\boldsymbol{\varepsilon}}(0 \mid \mathbf{y}^0)} - \sqrt{q_t^{\boldsymbol{\varepsilon}}(0 \mid \mathbf{y}^0)} \right| \le \alpha.$$

If there exists $\boldsymbol{\varepsilon}$ and $\boldsymbol{\varepsilon}'$ such that $\mathbf{v}^{\boldsymbol{\varepsilon}} = \mathbf{v}^{\boldsymbol{\varepsilon}'}$, then

$$\max_{t\in[n]} \left| \sqrt{q_t^{\boldsymbol{\varepsilon}'}(0 \mid \mathbf{y}^0)} - \sqrt{q_t^{\boldsymbol{\varepsilon}}(0 \mid \mathbf{y}^0)} \right| \le 2\alpha.$$

However, if $\varepsilon_t \neq \varepsilon_t'$, then

$$\left| \sqrt{q_t^{\boldsymbol{\varepsilon}'}(0 \mid \mathbf{y}^0)} - \sqrt{q_t^{\boldsymbol{\varepsilon}}(0 \mid \mathbf{y}^0)} \right| = \sqrt{\frac{1}{2} + 3\alpha} - \sqrt{\frac{1}{2} - 3\alpha} > 2\alpha,$$

leading to contradiction. Therefore, for any $\boldsymbol{\varepsilon} \neq \boldsymbol{\varepsilon}$, we have $v^{\boldsymbol{\varepsilon}} \neq v^{\boldsymbol{\varepsilon}'}$. This implies that $|\mathcal{V}| \ge 2^n$, hence

$$\mathcal{H}(\mathcal{Q}, \alpha, n) \ge n.$$

Next we lower bound the log-entropy $\mathcal{H}_{\log}$, which is the entropy with respect to the distance defined in eq. (5.7). Suppose $\mathcal{V}$ is a finite cover of $\mathcal{Q}$ at scale α. We first define set $E \subseteq \{-1,1\}^n$ such that for any two distinct items $\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}' \in E$, we have

$$\sum_{t=1}^{n} \mathbb{I}[\varepsilon_t \neq \varepsilon_t'] \ge \frac{n}{4}. \tag{D.73}$$

According to the lower bound of packing number under Hamming distances (see [221, Theorem 27.5]), there exists such set E which satisfies

$$\log |E| \ge (1 - \log 2)n - o(n).$$

Next, since $\mathcal{V}$ is a covering of $\mathcal{Q}$, for any $\boldsymbol{\varepsilon} \in E$, there exists $\mathbf{v}^{\boldsymbol{\varepsilon}} \in \mathcal{V}$ such that

$$\sum_{t=1}^{n} \sup_{\mathbf{y}} (\log v_t^{\boldsymbol{\varepsilon}}(y_t \mid \mathbf{y}) - \log q_t^{\boldsymbol{\varepsilon}}(y_t \mid \mathbf{y}))^2 \le n\alpha^2,$$

which implies that

$$\sum_{t=1}^{n} (\log v_t^{\boldsymbol{\varepsilon}}(0 \mid \mathbf{y}^0) - \log q_t^{\boldsymbol{\varepsilon}}(0 \mid \mathbf{y}^0))^2 \le n\alpha^2.$$

If there exists $\boldsymbol{\varepsilon}$ and $\boldsymbol{\varepsilon}'$ such that $\mathbf{v}^{\boldsymbol{\varepsilon}} = \mathbf{v}^{\boldsymbol{\varepsilon}'}$, then

$$\sum_{t=1}^{n} (\log q_t^{\boldsymbol{\varepsilon}}(0 \mid \mathbf{y}^0) - \log q_t^{\boldsymbol{\varepsilon}'}(0 \mid \mathbf{y}^0))^2 \le 4n\alpha^2. \tag{D.74}$$

However, if $\varepsilon_t \neq \varepsilon'_t$, then

$$|\log q_t^{\boldsymbol{\varepsilon}}(0 \mid \mathbf{y}^0) - \log q_t^{\boldsymbol{\varepsilon}'}(0 \mid \mathbf{y}^0)| \geq \left|\log \frac{1-6\alpha}{1+6\alpha}\right| > 6\alpha,$$

which implies that

$$\sum_{t=1}^{n} (\log q_t^{\boldsymbol{\varepsilon}}(0 \mid \mathbf{y}^0) - \log q_t^{\boldsymbol{\varepsilon}'}(0 \mid \mathbf{y}^0))^2 \geq 36\alpha^2 \cdot \sum_{t=1}^{n} \mathbb{I}[\varepsilon_t \neq \varepsilon'_t] > 9n\alpha^2,$$

where the last inequality follows from the construction of set E, i.e. eq. (D.73). This contradicts to eq. (D.74). Hence for any $\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}' \in E$, we have $\mathbf{v}^{\boldsymbol{\varepsilon}} \neq \mathbf{v}^{\boldsymbol{\varepsilon}'}$. This implies that $|\mathcal{V}| \geq |E|$, hence

$$\mathcal{H}(\mathcal{Q}, \alpha, n) \geq \log |E| \geq (1 - \log 2)n - o(n).$$

□

We see that the root cause of entropies being $\Omega(n)$ is the same: both definitions of $\mathcal{H}_{\mathsf{sq}}$ and $\mathcal{H}_{\mathsf{log}}$ in Theorem 5.1 and (5.7) take supremum over the "true path" $\mathbf{y}$ on the tree. In the example above, this corresponds to simply taking a path on the very left of the tree. The process class is so rich that already on this left-most path the entropy is $\Omega(n)$. However, this should not concern log-loss prediction as this left-most path would not happen too-often, unless $\mathbf{p}$ in (D.2) places all mass on all-0 input, in which case the $\mathcal{R}_n(\mathcal{Q}, \mathbf{p}) = \mathbf{0}$. Searching for the correct definition of entropy to handle this class is left to future work.

Appendix E

Proofs for Chapter 6

E.1 Missing Proofs in section 6.2

E.1.1 Proofs of Theorem 6.3

Proof of Theorem 6.3. Without loss of generality we assume $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$, since the largest subset of $\{x_1, \ldots, x_n\}$ shattered by $\mathcal{F}$ is also a subset of $\mathcal{X}$ shattered by $\mathcal{F}$. In the following, we only need to consider cases where

$$4d(\mathcal{F}, \alpha) \log(eMn) \leq n. \tag{E.1}$$

In cases where eq. (E.1) fails, by taking the covering of all functions which map $\{x_{1:n}\}$ to $[M]$, we have the covering number estimate

$$\mathcal{N}_\infty(\mathcal{F}, x_{1:n}, \alpha) \leq M^n = \exp(n \cdot \log M) \leq \exp(4d(\mathcal{F}, \alpha) \log(eMn) \cdot \log M) \leq \exp(16 d(\mathcal{F}, \alpha) \log^2(eMn)).$$

In the following, we let $d = d(\mathcal{F}, \alpha)$. We will use an approach similar to that in [128] to prove Theorem 6.3. We define the packing number $\mathcal{M}_\infty(\mathcal{F}, x_{1:n}, \alpha)$ of class $\mathcal{F}$ under design $x_{1:n}$ at scale α: we say $\widetilde{\mathcal{F}} \subseteq \mathcal{F}$ is a packing if for any $f, f' \in \widetilde{\mathcal{F}}$,

$$\max_{t \in [n]} \mathfrak{c}(f(x_t), f'(x_t)) \geq \alpha,$$

and we let $\mathcal{M}_\infty(\mathcal{F}, x_{1:n}, \alpha)$ be the size of the largest packing of class $\mathcal{F}$. Then according to covering-packing inequality we have

$$\mathcal{N}_\infty(\mathcal{F}, x_{1:n}, \alpha) \leq \mathcal{M}_\infty(\mathcal{F}, x_{1:n}, \alpha).$$

Hence, it suffices to prove

$$\log \mathcal{M}_\infty(\mathcal{F}, x_{1:n}, \alpha) \leq 16 d \log^2(eMn). \tag{E.2}$$

For set $X = (x_1, \ldots, x_l) \subseteq \{x_{1:l}\}$, and tuple $\mathbf{s} = (s_1, s_2, \ldots, s_l)$ where $s_t = (s_t[-1], s_t[1]) \in [M] \times [M]$ for any $t \in [l]$, we say the pair $(X, \mathbf{s})$ is shattered by $\mathcal{F}$ if the following two properties both hold:

(a) for any $t \in [l]$, $\mathfrak{c}(s_t[-1], s_t[-1]) \geq \alpha$;

(b) for any $\boldsymbol{\varepsilon} = (\varepsilon_{1:l}) \in \{-1, 1\}^l$, there exists $f^{\boldsymbol{\varepsilon}} \in \mathcal{F}$ such that $f^{\boldsymbol{\varepsilon}}(x_t) = s_t[\varepsilon_t]$ for any $t \in [l]$.

We define function

$$t(h, l) \coloneqq \min_{\tilde{\mathcal{X}} \subseteq \mathcal{X}, |\tilde{\mathcal{X}}| = l} \max\{k : \forall\ F \subseteq \mathcal{F} \text{ such that } |F| = h \text{ and } \forall f, f' \in F,\ \max_{x \in \tilde{\mathcal{X}}} \mathfrak{c}(f(x), f'(x)) \geq \alpha,$$
$$F \text{ shatters at least } k \text{ pairs } (X, \mathbf{s}) \text{ where } X \subseteq \tilde{\mathcal{X}}\}. \tag{E.3}$$

(if no packing of size h exists, $t(h, l)$ is defined to be infinity). According to the definition of $d = d(\mathcal{F}, \alpha)$, if $(X, \mathbf{s})$ is shattered by $\mathcal{F}$, then we have $|X| \leq d$. Additionally, for fixed X, there can be at most $(M^2)^{|X|}$ choices of $\mathbf{s}$ such that $(X, \mathbf{s})$ is shattered by $\mathcal{F}$. Therefore, by choosing h to be the size of largest packing, i.e. $\mathcal{M}_\infty(\mathcal{F}, x_{1:n}, \alpha)$, the number of pairs $(X, \mathbf{s})$ shattered by the largest packing is no more than the number of pairs $(X, \mathbf{s})$ shattered by $\mathcal{F}$, which implies that

$$t(\mathcal{M}_\infty(\mathcal{F}, x_{1:n}, \alpha), n) \leq \sum_{i=0}^{d} \binom{n}{i} \cdot M^{2i}. \tag{E.4}$$

It is sufficient to argue that for any $r + 1 \leq l \leq n$,

$$t(2(2lM^2)^r, l) > 2^r. \tag{E.5}$$

Indeed, with eq. (E.5) holding, it is easy to see from the definition of function t in eq. (E.3) that for $h_1 \leq h_2$ and any l, $t(h_1, l) \leq t(h_2, l)$. Notice that according to eq. (E.1),

$$n \geq 4d \log(eMn) \geq \log_2 \left\lfloor \sum_{i=1}^{d} \binom{n}{i} M^{2i} \right\rfloor + 2.$$

Hence, if

$$\mathcal{M}_\infty(\mathcal{F}, x_{1:n}, \alpha) > 2(2nM^2)^{\log_2 \lfloor \sum_{i=1}^{d} \binom{n}{i} M^{2i} \rfloor + 1},$$

we have

$$\begin{aligned} t(\mathcal{M}_\infty(\mathcal{F}, x_{1:n}, \alpha), n) &\geq t\left(2(2nM^2)^{\log_2 \lfloor \sum_{i=1}^{d} \binom{n}{i} M^{2i} \rfloor + 1}, n\right) \\ &\geq 2^{\log_2 \lfloor \sum_{i=1}^{d} \binom{n}{i} M^{2i} \rfloor + 1} > \sum_{i=1}^{d} \binom{n}{i} M^{2i}, \end{aligned}$$

which contradicts eq. (E.4). Therefore,

$$\log \mathcal{N}_\infty(\mathcal{F}, x_{1:n}, \alpha) \leq \log \mathcal{M}_\infty(\mathcal{F}, x_{1:n}, \alpha) \leq \log \left(2(2nM^2)^{\log_2 \lfloor \sum_{i=1}^{d} \binom{n}{i} M^{2i} \rfloor + 1}\right) \leq 16d \log^2(eMn).$$

In the remainder of the proof, we argue that eq. (E.5) holds by induction. We will verify the following two properties:

$$t(2, l) \geq 1 \qquad \forall l \geq 1, \tag{E.6}$$
$$t(2lM^2 \cdot 2m, l) \geq 2 \cdot t(2m, l - 1) \qquad \forall m \geq 1, l \geq 2. \tag{E.7}$$

We first verify eq. (E.6). For $\tilde{\mathcal{X}} \subseteq \mathcal{X}$ and any packing $F \subseteq \mathcal{F}$ with $|F| = 2$, there exists $f_1, f_2 \in F$ such that

$$\max_{x \in \tilde{\mathcal{X}}} \mathfrak{c}(f_1(x), f_2(x)) \geq \alpha.$$

We let $\tilde{\mathcal{X}}$ to be the one which takes the maximum in the above inequality, and let $X = (\subseteq \mathcal{X}$ and $\mathbf{s} = \{s_1\}$ with $s_1 = (f_1(x), f_2(x)) \in [M] \times [M]$, then $(X, \mathbf{s})$ is shattered by $\mathcal{F}$. Therefore, $t(2, l) \geq 1$.

We next verify eq. (E.7). Without loss of generality we assume the minimum over $\tilde{\mathcal{X}} \subseteq \mathcal{X}$ with $|\tilde{\mathcal{X}}| = l$ in eq. (E.3) is achieved by $\tilde{\mathcal{X}} = \{x_{1:l}\}$. Suppose packing $F \subseteq \mathcal{F}$ satisfies $|F| \geq 4lM^2 \cdot m$. We arbitrarily pair up functions in F to form:

$$F = \bigcup_{t=1}^{2lM^2 m} \{f_t, g_t\}.$$

For any $t \in [2lM^2 m]$, since F is a packing, there exists $x[t] \in \{x_{1:l}\}$ such that

$$\mathfrak{c}(f_t(x[t]), g_t(x[t])) \geq \alpha.$$

Next, for any $x \in \{x_{1:l}\}$ and $i, j \in [M]$ with $\mathfrak{c}(i, j) \geq \alpha$, we define the set

$$\mathcal{T}(x, i, j) = \{t \in [2lM^2 m] : x[t] = x \text{ and } f_t(x[t]) = i,\ g_t(x[t]) = j\}.$$

Then we have

$$\bigcup_{\substack{x \in \{x_{1:l}\} \\ i,j \in M, \mathfrak{c}(i,j) \geq \alpha}} \mathcal{T}(x, i, j) = [2lM^2 m].$$

According to the pigeonhole principle, there exists some $x_{t^\star} \in \{x_{1:l}\}$ and $i^\star, j^\star \in [M]$ with $\mathfrak{c}(i^\star, j^\star) \geq \alpha$ such that $|\mathcal{T}(x_{t^\star}, i^\star, j^\star)| \geq 2m$. We define two function classes $F_1, F_2 \subseteq F$ as

$$F_1 = \{f_t : t \in \mathcal{T}(x_t^\star, i^\star, j^\star)\} \quad \text{and} \quad F_2 = \{g_t : t \in \mathcal{T}(x_t^\star, i^\star, j^\star)\}.$$

Then we have $|F_1| = |F_2| \geq 2m$. Since functions in F_1 all take value $i^\star$ at $x_{t^\star}$, F_1 is a packing under design $\{x_{1:l}\} \backslash \{x_{t^\star}\}$. Hence, there exists a set V consisting of pairs $(X, \mathbf{s})$ shattered by class F_1, and $|V| \geq t(2m, l-1)$. Since functions in F_1 takes the same value at $x_{t^\star}$, for any $(X, \mathbf{s}) \in V$, $x_{t^\star} \notin X$. Similarly, there exists a set U shattered by F_2 with $|U| \geq t(2m, l-1)$, and for any $(X, \mathbf{s}) \in U$, $x_{t^\star} \notin X$. Since $F_1 \subseteq F$ and $F_2 \subseteq F$, any pairs $(X, \mathbf{s})$ in $U \cup V$ are also shattered by F. Additionally, for any pair $(X, \mathbf{s}) \in U \cap V$, we construct a new pair $(X \cup (x_{t^\star}), \mathbf{s} \cup (i^\star, j^\star)) \notin U \cup V$. Since $\mathfrak{c}(i^\star, j^\star) \geq \alpha$, this pair is also shattered by F. Hence F shatters at least

$$|U \cap V| + |U \cup V| = |U| + |V| = 2t(2m, l-1)$$

pairs, which implies that $t(4lM^2 m, l) \geq 2t(2m, l-1)$. This completes the proof of eq. (E.7).

With eq. (E.6) and eq. (E.7), when $r \leq l - 1$ we have

$$t(2(2lM^2)^r, l) \geq 2^r t(2, l-r) \geq 2^r.$$

which verifies eq. (E.5). □

E.1.2 Proof of Theorem 6.6

Proof of Theorem 6.6. We define distance $\mathfrak{c}' : [M] \times [M] \to \mathbb{R}_+ \cup \{0\}$:

$$\mathfrak{c}'(a, b) = \mathfrak{c}(u_a, u_b). \tag{E.8}$$

For any $f \in \mathcal{F}$, since f maps $\mathcal{X}$ into $[-1, 1]$, we define $\bar{f} : \mathcal{X} \to [M]$ to be an arbitrary function mapping $\mathcal{X}$ into $[M]$, which satisfies that for any $x \in \mathcal{X}$, $\mathfrak{c}(f(x), u_{\bar{f}(x)}) \leq \beta$. We construct function class $\bar{\mathcal{F}} = \{\bar{f} : f \in \mathcal{F}\} \subseteq \{\mathcal{X} \to [M]\}$. We use $d_{\mathfrak{c}'}(\bar{\mathcal{F}}, \alpha)$ to denote the non-sequential gapped dimension (defined in Theorem 6.1) of integer-valued function class $\bar{\mathcal{F}}$ at scale α under distance $\mathfrak{c}'$, and for simplicity we let $d = d_{\mathfrak{c}'}(\bar{\mathcal{F}}, \alpha)$. Suppose $x_{1:d} \in \mathcal{X}^d$ is shattered by $\mathcal{F}$. Then there exists $\bar{s}_1 = (\bar{s}_1[-1], \bar{s}_1[1]), \ldots, \bar{s}_d = (\bar{s}_d[-1], \bar{s}_d[1])$ such that for any $\boldsymbol{\varepsilon} \in \{-1, 1\}^d$ and $t \in [d]$, $\mathfrak{c}'(\bar{s}_t[-1], \bar{s}_t[1]) \geq \alpha$, and also for any $\boldsymbol{\varepsilon} \in \{-1, 1\}^d$, there exists $\bar{f}^{\boldsymbol{\varepsilon}} \in \bar{\mathcal{F}}$ such that $\bar{f}^{\boldsymbol{\varepsilon}}(x_t) = \bar{s}_t[\varepsilon_t]$ for any $t \in [d]$. Notice from the definition of $\bar{\mathcal{F}}$ that there exists some $f^{\boldsymbol{\varepsilon}} \in \mathcal{F}$ such that $\bar{f}^{\boldsymbol{\varepsilon}} = \overline{(f^{\boldsymbol{\varepsilon}})}$.

We next verify that $x_{1:n}$ is also shattered by $\mathcal{F}$ at scale (α, β), according to the definition Theorem 6.4. We construct $s_1 = (s_1[-1], s_1[1]), \ldots, s_d = (s_d[-1], s_d[1])$ according to $\bar{s}_{1:d}$ as follows:

$$s_t[-1] = u_{\bar{s}_t[-1]} \quad \text{and} \quad s_t[1] = u_{\bar{s}_t[1]}.$$

Then according to the definition of $\mathfrak{c}'$ in eq. (E.8), we have for any $t \in [d]$,

$$\mathfrak{c}(s_t[-1], s_t[1]) = \mathfrak{c}'(\bar{s}_t[-1], \bar{s}_t[1]) \geq \alpha.$$

Additionally, for any $\boldsymbol{\varepsilon} \in \{-1, 1\}^d$, there exists some $f^{\boldsymbol{\varepsilon}} \in \mathcal{F}$ such that $\overline{(f^{\boldsymbol{\varepsilon}})}(x_t) = \bar{s}_t[\varepsilon_t]$ for any $t \in [d]$, which implies that

$$\mathfrak{c}(f(x_t), s_t[\varepsilon_t]) = \mathfrak{c}(f(x_t), u_{\bar{s}_t[\varepsilon_t]}) = \mathfrak{c}(f(x_t), u_{\overline{(f^{\boldsymbol{\varepsilon}})}(x_t)}) \leq \beta, \quad \forall t \in [d],$$

where the last inequality follows from the definiton of $\bar{f}$. Therefore, $x_{1:d}$ is shattered by $\mathcal{F}$ at scale (α, β), and, according to Theorem 6.4, we have $d(\mathcal{F}, \alpha, \beta) \geq d$.

Next, we will upper bound the non-sequential covering number of $\mathcal{F}$ in terms of d. According to Theorem 6.3, for any $x_{1:n} \in \mathcal{X}^n$ there exists a non-sequential covering $\bar{\mathcal{V}}$ of $\bar{\mathcal{F}}$ with size no more than $\exp\left(16d \log^2(enM)\right)$. Hence, for any $f \in \mathcal{F}$, there exists some $\bar{\mathbf{v}} = (\bar{v}_{1:n}) \in \bar{\mathcal{V}}$ which satisfies

$$\mathfrak{c}'(\bar{f}(x_t), \bar{v}_t) \leq \alpha \qquad \forall t \in [n],$$

which implies that

$$\mathfrak{c}(f(x_t), u_{\bar{v}_t}) \leq \mathfrak{c}(f(x_t), u_{\bar{f}(x_t)}) + \mathfrak{c}(u_{\bar{f}(x_t)}, u_{\bar{v}_t}) \leq \beta + \alpha \qquad \forall t \in [n],$$

where the first inequality uses the triangle inequality of $\mathfrak{c}$, and the second inequality uses the definition of function $\bar{f}$. For every $\bar{\mathbf{v}} \in \bar{\mathcal{V}}$, we construct $\mathbf{v}^{\bar{\mathbf{v}}} = (v_{1:n}^{\bar{\mathbf{v}}}) \in [-1, 1]^n$: for any $t \in [n]$, $v_t^{\bar{\mathbf{v}}} = u_{\bar{v}_t}$. We further let $\mathcal{V} = \{\mathbf{v}^{\bar{\mathbf{v}}} : \bar{\mathbf{v}} \in \bar{\mathcal{V}}\}$. Then $\mathcal{V}$ is an $(\alpha + \beta)$-cover of $\mathcal{F}$ on $x_{1:n}$, which implies

$$\log \mathcal{N}_\infty(\mathcal{F}, \mathbf{x}, \alpha + \beta) \leq \log |\mathcal{V}| = \log |\bar{\mathcal{V}}| \leq 16d \log^2(enM) \leq 16d(\mathcal{F}, \alpha, \beta) \log^2(enM).$$

□

E.1.3 Missing Proofs in section 6.2.3

Proof of Theorem 6.8. We only need to prove that if $x_{1:d} \in \mathcal{X}^d$ is shattered by $\mathcal{F}$ at scale (α, β) according to Theorem 6.4, then $x_{1:d}$ is shattered by $\mathcal{F}$ at scale $\alpha - 2\beta$ according to the traditional notion of shattering as in Theorem 6.7.

If $x_{1:d}$ is shattered by $\mathcal{F}$ at scale (α, β) according to Theorem 6.4, then there exists $s_{1:d}$ where $s_t = (s_t[-1], s_t[1]) \in [-1, 1] \times [-1, 1]$ for any $t \in [d]$, such that $s_t[1] - s_t[-1] \geq \alpha$ for any $t \in [d]$, and also for any $\boldsymbol{\varepsilon} \in \{-1, 1\}^d$, there exists a function $f^{\boldsymbol{\varepsilon}} \in \mathcal{F}$ which satisfies $|f^{\boldsymbol{\varepsilon}}(x_t) - s_t[\varepsilon_t]| \leq \beta$ for any $t \in [d]$. We let

$$v_t = \frac{s_t[-1] + s_t[1]}{2}, \quad \forall t \in [n].$$

Since $\alpha > 2\beta$, we have for any $\boldsymbol{\varepsilon} \in \{-1, 1\}^d$,

$$\varepsilon_t \cdot (f^{\boldsymbol{\varepsilon}}(x_t) - v_t) \geq \frac{\alpha}{2} - \beta,$$

which implies that $x_{1:d}$ is shattered by $\mathcal{F}$ at scale $\alpha - 2\beta$, according to Theorem 6.7. □

Proof of Theorem 6.9. Let $d = \mathsf{vc}(\mathcal{F}, \alpha)$, then there exists $x_{1:d}$ shattered by $\mathcal{F}$ according to Theorem 6.7. Hence there exists $v_{1:d} \in [-1, 1]^d$ and also function $f^{\boldsymbol{\varepsilon}} \in \mathcal{F}$ for each $\boldsymbol{\varepsilon} \in \{-1, 1\}^d$, such that for any $\boldsymbol{\varepsilon} \in \{-1, 1\}^d$,

$$\varepsilon_t \cdot (f^{\boldsymbol{\varepsilon}}(x_t) - v_t) \geq \frac{\alpha}{2}.$$

Further notice for every $\boldsymbol{\varepsilon} \neq \boldsymbol{\varepsilon}'$, there exists $t \in [d]$ such that $\varepsilon_t \neq \varepsilon'_t$, which implies that for this specific t,

$$|f^{\boldsymbol{\varepsilon}}(x_t) - f^{\boldsymbol{\varepsilon}'}(x_t)| \geq \alpha.$$

Hence $\{f^{\boldsymbol{\varepsilon}} : \boldsymbol{\varepsilon} \in \{-1, 1\}^d\}$ forms a α-packing of $\mathcal{F}$ under ℓ_∞-norm under design $x_{1:d}$. Therefore, the covering-packing duality indicates that

$$\mathcal{N}_\infty(\mathcal{F}, x_{1:\mathsf{vc}(\mathcal{F},\alpha)}, \alpha/3) \geq 2^d = 2^{\mathsf{vc}(\mathcal{F},\alpha)}. \tag{E.9}$$

Next according to Theorem 6.6, we have for any n and $x_{1:n} \in \mathcal{X}^n$,

$$\mathcal{N}_\infty(\mathcal{F}, x_{1:n}, \alpha + \beta) \leq 16 d(\mathcal{F}, \alpha, \beta) \log^2\left(\frac{2en}{\beta}\right).$$

Hence by replacing α in eq. (E.9) with $3(\alpha + \beta)$, and choosing n to be $\mathsf{vc}(\mathcal{F}, 3(\alpha + \beta))$, we obtain that

$$\begin{aligned}
\log 2 \cdot \mathsf{vc}(\mathcal{F}, 3(\alpha + \beta)) &\leq \sup_{x_{1:\mathsf{vc}(\mathcal{F},\alpha+\beta)}} \log \mathcal{N}_\infty(\mathcal{F}, x_{1:\mathsf{vc}(\mathcal{F},3(\alpha+\beta))}, \alpha + \beta) \\
&\leq 16 d(\mathcal{F}, \alpha, \beta) \log^2\left(\frac{2e \cdot \mathsf{vc}(\mathcal{F}, 3(\alpha + \beta))}{\beta}\right),
\end{aligned}$$

which implies

$$\mathsf{vc}(\mathcal{F}, 3(\alpha + \beta)) \leq 32 d(\mathcal{F}, \alpha, \beta) \log^2\left(\frac{6\mathsf{vc}(\mathcal{F}, 3(\alpha + \beta))}{\beta}\right). \tag{E.10}$$

Additionally, we notice that $\log(x) \le \sqrt{2} \cdot \sqrt[3]{x}$, hence

$$\mathsf{vc}(\mathcal{F}, 3(\alpha+\beta)) \le 64 d(\mathcal{F}, \alpha, \beta) \left(\frac{6\mathsf{vc}(\mathcal{F}, 3(\alpha+\beta))}{\beta} \right)^{2/3},$$

which implies that

$$\mathsf{vc}(\mathcal{F}, 3(\alpha+\beta)) \le 64^3 \cdot 6^2 \cdot \frac{d(\mathcal{F}, \alpha, \beta)^3}{\beta^2}.$$

Bringing this back to eq. (E.10), we obtain that

$$\mathsf{vc}(\mathcal{F}, 3(\alpha+\beta)) \le 288 d(\mathcal{F}, \alpha, \beta) \cdot \log^2 \left(\frac{384 d(\mathcal{F}, \alpha, \beta)}{\beta} \right).$$

The second part of Theorem 6.9 is implied by Theorem E.2 below. Indeed, if $x_{1:d}$ is shattered by $\mathcal{F}$ at scale α according to Theorem E.1, then $x_{1:d}$ is also shattered by $\mathcal{F}$ at scale (α, β) according to Theorem 6.4. □

Finally, we provide the proof of Theorem 6.10.

Proof of Theorem 6.10. Let $d = \lfloor \log(1/\alpha) \rfloor$ and $\mathcal{X} = \{x_1, \dots, x_d\}$. For any $\boldsymbol{\varepsilon} = (\varepsilon_{1:d}) \in \{-1, 1\}^d$, we define

$$a^{\boldsymbol{\varepsilon}} = \alpha + \sum_{i=1}^{d} \frac{\varepsilon_i + 1}{2} \cdot 2^{i-1} \cdot \alpha.$$

Then for any $\boldsymbol{\varepsilon}$, we have

$$\alpha \le a^{\boldsymbol{\varepsilon}} \le \alpha + \alpha \cdot (2^d - 1) \le 1.$$

We define function class $\mathcal{F} = \{f^{\boldsymbol{\varepsilon}} : \boldsymbol{\varepsilon} \in \{-1, 1\}^d\}$, where $f^{\boldsymbol{\varepsilon}} : \mathcal{X} \to [-1, 1]$ is defined as

$$f^{\boldsymbol{\varepsilon}}(x_i) = \varepsilon_i \cdot a^{\boldsymbol{\varepsilon}}, \quad \forall 1 \le i \le d.$$

Then it is easy to see that $\{x_1, \dots, x_d\}$ is shattered by $\mathcal{F}$ in terms of the classical shattering (defined in Theorem 6.7).

Next, we will verify that the non-sequential gapped dimension $d(\mathcal{F}, \alpha, \beta) = 1$ for any $\beta < \alpha/2$. First of all, it is easy to see that $\{x_1\}$ is shattered by $\mathcal{F}$, hence $d(\mathcal{F}, \alpha, \beta) \ge 1$. If there exists a size-two subset $\{x_i, x_j\}$ of $\mathcal{X}$ shattered by $\mathcal{F}$, in terms of Theorem 6.1, then there exists $s_i = (s_i[-1], s_i[1]) \in [-1, 1] \times [-1, 1]$ and $s_j = (s_j[-1], s_j[1]) \in [-1, 1] \times [-1, 1]$ such that for any $\mathbf{e} = (e_i, e_j) \in \{-1, 1\} \times \{-1, 1\}$,

$$\exists\, \boldsymbol{\varepsilon}[\mathbf{e}] \in \{-1, 1\}^d \quad \text{such that} \quad |f^{\boldsymbol{\varepsilon}[\mathbf{e}]}(x_i) - s_i[e_i]| \le \beta \quad \text{and} \quad |f^{\boldsymbol{\varepsilon}[\mathbf{e}]}(x_j) - s_j[e_j]| \le \beta.$$

Hence, we obtain

$$|f^{\boldsymbol{\varepsilon}[(-1,-1)]}(x_i) - f^{\boldsymbol{\varepsilon}[(-1,1)]}(x_i)| \le 2\beta < \alpha.$$

According to the construction of function $f^{\boldsymbol{\varepsilon}}$, we must have $\boldsymbol{\varepsilon}[(-1,-1)] = \boldsymbol{\varepsilon}[(-1,1)]$, which implies that

$$|s_j[1] - s_j[-1]| \le |s_j[1] - f^{\boldsymbol{\varepsilon}[(-1,1)]}(x_j)| + |s_j[-1] - f^{\boldsymbol{\varepsilon}[(-1,-1)]}(x_j)| \le \beta + \beta < \alpha.$$

This violates the definition of shattering in Theorem 6.1. Therefore, we have verified that $d(\mathcal{F}, \alpha, \beta) \le 1$ for any $\beta < \alpha/2$. □

E.1.4 Properties of Fixed-Scale Scale-Sensitive Dimension

We consider the following fixed-scale scale-sensitive dimension; the only modification with respect to Theorem 6.7 is that the inequality is turned into an equality. In terms of fig. 6.1, the requirement of shattering states that the vertices of the hypercube with side-length α are in the set.

Definition E.1 (Fixed-Scale Dimension)**.** *Given a function class $\mathcal{F} \subseteq \{\mathcal{X} \to [-1,1]\}$, we say that a set $\{x_1, x_2, \ldots, x_d\} \subseteq \mathcal{X}$ is shattered by $\mathcal{F}$ at scale $\alpha > 0$, if there exists $s_1, \ldots, s_d \in [-1,1]$ such that for any $\boldsymbol{\varepsilon} = (\varepsilon_{1:d}) \in \{-1,1\}^d$, there exists some $f^{\boldsymbol{\varepsilon}} \in \mathcal{F}$ such that*

$$\varepsilon_t \cdot (f^{\boldsymbol{\varepsilon}}(x_t) - s_t) = \frac{\alpha}{2}, \qquad \forall t \in [d].$$

The fixed-scale dimension $\mathsf{vc}_{\mathsf{fix}}(\mathcal{F}, \alpha)$ of $\mathcal{F}$ at scale α is the largest d such that there exists a size-d subset of $\mathcal{X}$ shattered by $\mathcal{F}$.

We have the following proposition showing that if the function class $\mathcal{F}$ is convex, then the scale-sensitive dimensions defined in Theorem 6.7 and in Theorem E.1 coincide.

Proposition E.2. *If function class $\mathcal{F}$ is convex, then for any $\alpha > 0$, $\mathsf{vc}(\mathcal{F}, \alpha) = \mathsf{vc}_{\mathsf{fix}}(\mathcal{F}, \alpha)$.*

Proof of Theorem E.2. According to Theorem 6.7 and Theorem E.1, we have $\mathsf{vc}(\mathcal{F}, \alpha) \geq \mathsf{vc}_{\mathsf{fix}}(\mathcal{F}, \alpha)$. Hence we only need to prove that $\mathsf{vc}_{\mathsf{fix}}(\mathcal{F}, \alpha) \geq \mathsf{vc}(\mathcal{F}, \alpha)$. In the following, we will show that if $\{x_1, \ldots, x_d\}$ is shattered by $\mathcal{F}$ at scale α under witness $s_1, \ldots, s_d$ under Theorem 6.7, then $\{x_1, \ldots, x_d\}$ is shattered by $\mathcal{F}$ at scale α under witness $s_1, \ldots, s_d$ under Theorem E.1.

Since $\{x_1, \ldots, x_d\}$ is shattered by $\mathcal{F}$ at scale α under witness $s_1, \ldots, s_d$ under Theorem 6.7, for any $\boldsymbol{\varepsilon} \in \{-1,1\}^d$ there exists $f^{\boldsymbol{\varepsilon}}$ such that

$$\varepsilon_t \cdot (f^{\boldsymbol{\varepsilon}}(x_t) - s_t) \geq \frac{\alpha}{2} \qquad \forall t \in [d].$$

We next iteratively construct $f_i^{\boldsymbol{\varepsilon}} \in \mathcal{F}$ for $i = 0, 1, \ldots, d$ such that $f_i^{\boldsymbol{\varepsilon}}$ satisfies

$$\varepsilon_t \cdot (f^{\boldsymbol{\varepsilon}_i}(x_t) - s_t) = \frac{\alpha}{2} \quad \text{if } t \leq i \qquad \text{and} \qquad \varepsilon_t \cdot (f^{\boldsymbol{\varepsilon}_i}(x_t) - s_t) \geq \frac{\alpha}{2} \quad \text{if } t > i. \tag{E.11}$$

For $i = 0$, we let $f_0^{\boldsymbol{\varepsilon}} = f^{\boldsymbol{\varepsilon}}$ and they satisfies conditions in eq. (E.11). Suppose we have constructed $f_i^{\boldsymbol{\varepsilon}}$, and we will now construct $f_{i+1}^{\boldsymbol{\varepsilon}}$. For any $\boldsymbol{\varepsilon} = (\varepsilon_{1:d}) \in \{-1,1\}^d$, we let

$$\tilde{\boldsymbol{\varepsilon}} = (\varepsilon_1, \ldots, \varepsilon_i, -\varepsilon_{i+1}, \varepsilon_{i+2}, \ldots, \varepsilon_d). \tag{E.12}$$

Let

$$f_{i+1}^{\boldsymbol{\varepsilon}} = \alpha^{\boldsymbol{\varepsilon}} \cdot f_i^{\boldsymbol{\varepsilon}} + (1 - \alpha^{\boldsymbol{\varepsilon}}) \cdot f_i^{\tilde{\boldsymbol{\varepsilon}}}, \tag{E.13}$$

where

$$\alpha^{\boldsymbol{\varepsilon}} = \begin{cases} \frac{\alpha/2 + s_{i+1} - f^{\tilde{\boldsymbol{\varepsilon}}}(x_{i+1})}{f_i^{\boldsymbol{\varepsilon}}(x_{i+1}) - f_i^{\tilde{\boldsymbol{\varepsilon}}}(x_{i+1})} & \text{if } \varepsilon_{i+1} = 1, \\ \frac{\alpha/2 + f^{\tilde{\boldsymbol{\varepsilon}}}(x_{i+1}) - s_{i+1}}{f_i^{\boldsymbol{\varepsilon}}(x_{i+1}) - f_i^{\tilde{\boldsymbol{\varepsilon}}}(x_{i+1})} & \text{if } \varepsilon_{i+1} = -1. \end{cases}$$

According to eq. (E.11), we can verify that $\alpha^{\tilde{\boldsymbol{\varepsilon}}} \in [0,1]$ for any $\boldsymbol{\varepsilon}$. Therefore $f_{i+1}^{\boldsymbol{\varepsilon}}$ constructed in eq. (E.13) belongs to $\mathcal{F}$ due to the convexity of $\mathcal{F}$. Next, we will verify that $f_{i+1}^{\boldsymbol{\varepsilon}}$ also satisfies eq. (E.11). For $t \le i$, since

$$f_i^{\boldsymbol{\varepsilon}}(x_t) = s_t + \varepsilon_t \cdot \frac{\alpha}{2}$$

according to eq. (E.11), according to our choice of $\tilde{\boldsymbol{\varepsilon}}$ in eq. (E.12) we have

$$\varepsilon_t \cdot (f_i^{\boldsymbol{\varepsilon}}(x_t) - s_t) = \frac{\alpha}{2}.$$

For $t = i+1$, based on our construction in eq. (E.13) and our choice of $\tilde{\boldsymbol{\varepsilon}}$ in eq. (E.12), we can calculate that

$$\varepsilon_t \cdot (f_i^{\boldsymbol{\varepsilon}}(x_t) - s_t) = \frac{\alpha}{2}.$$

For $t > i+1$, since

$$\varepsilon_t \cdot (f_i^{\boldsymbol{\varepsilon}}(x_t) - s_t) \ge \frac{\alpha}{2}$$

according to eq. (E.11), hence according to our choice of $\tilde{\boldsymbol{\varepsilon}}$ in eq. (E.12) we have

$$\varepsilon_t \cdot (f_i^{\boldsymbol{\varepsilon}}(x_t) - s_t) \ge \frac{\alpha}{2}.$$

Therefore, $f_{i+1}^{\boldsymbol{\varepsilon}}$ satisfies eq. (E.11) as well. By choosing $i = d$, we obtain that there exist $f_d^{\boldsymbol{\varepsilon}} \in \mathcal{F}$ such that

$$\varepsilon_t \cdot (f_d^{\boldsymbol{\varepsilon}}(x_t) - s_t) = \frac{\alpha}{2},$$

which implies that $\{x_1, \dots, x_d\}$ is shattered by $\mathcal{F}$ at scale α under witness $s_1, \dots, s_d$ under Theorem E.1. □

E.1.5 Missing Proofs in section 6.2.4

Proof of Theorem 6.11.

$$\alpha_n = \frac{1}{2} \cdot \arg\min_{\alpha > 0} \left\{ d(\mathcal{F}, \alpha, \alpha/(20nC)) \cdot \frac{16}{\alpha^2} \le n \right\}.$$

Since $d(\mathcal{F}, \alpha, \alpha/(20nC)) = \tilde{\Omega}\left(\alpha^{-p}\right)$ for every $\alpha \ge 0$, we have

$$d(\mathcal{F}, \alpha_n, \alpha_n/(20nC)) \wedge \frac{n\alpha_n^2}{4} = \tilde{\Omega}\left(n^{\frac{p}{p+2}}\right), \qquad \forall n \in \mathbb{Z}_+.$$

In the following, we will prove that for any positive integer n, we have

$$\sup_{\boldsymbol{\mu}, \mathbf{x}} \mathbb{E}[\mathcal{R}_n(\mathcal{F}, \boldsymbol{\mu}, \mathbf{x}, \boldsymbol{\varepsilon})] = \Omega\left(d(\mathcal{F}, \alpha_n, \alpha_n/(20nC))\right).$$

We fix n, and, for brevity, write

$$\alpha = \alpha_n \quad \text{and} \quad d = d(\mathcal{F}, \alpha_n, \alpha_n/(20nC)) \wedge n\alpha^2/4 \le d(\mathcal{F}, \alpha_n, \alpha_n/(20nC)).$$

Let $1 : d \in \mathcal{X}^d$ be a sequence shattered by $\mathcal{F}$ in the sense of Theorem 6.4. That is, there exist $s_1 = (s_1[-1], s_1[1]), s_2 = (s_2[-1], s_2[1]), \ldots, s_d = (s_d[-1], s_d[1]) \in [-1, 1] \times [-1, 1]$ such that for any $\tilde{\boldsymbol{\varepsilon}} = (\tilde{\varepsilon}_{1:d}) \in \{-1, 1\}^d$, there exists $f^{\tilde{\boldsymbol{\varepsilon}}} \in \mathcal{F}$ such that

$$\left| f^{\tilde{\boldsymbol{\varepsilon}}}(t) - s_t[\tilde{\varepsilon}_t] \right| \le \frac{\alpha}{20(nC)} \quad \text{and} \quad |s_t[1] - s_t[-1]| \ge \alpha \qquad \forall t \in [d]. \tag{E.14}$$

Without loss of generality, we assume $s_t[1] \ge s_t[-1]$. We now define $\tilde{\mu}_{1:d} \in [-1, 1]^d$ as $\tilde{\mu}_t = (s_t[-1] + s_t[1])/2$ and

$$k_t = \left\lfloor \frac{1}{(s_t[1] - s_t[-1])^2} \right\rfloor \vee 1 \le 4\alpha^{-2}$$

for any $t \in [d]$. Then we have

$$\sum_{t=1}^{d} k_t \le n.$$

Next, we construct $x_{1:n}$ and $\mu_{1:n}$ with the following block structure:

$$\begin{aligned}
(x_{1:n}) &= (\underbrace{1, 1, \ldots, 1}_{k_1 \text{ terms}}, \underbrace{2, 2, \ldots, 2}_{k_2 \text{ terms}}, \ldots, \underbrace{d-1, d-1, \ldots, d-1}_{k_{d-1} \text{ terms}}, \underbrace{d, d, \ldots, d}_{n-k_1-\ldots-k_{d-1} \text{ terms}}), \\
(\mu_{1:n}) &= (\underbrace{\tilde{\mu}_1, \tilde{\mu}_1, \ldots, \tilde{\mu}_1}_{k_1 \text{ terms}}, \underbrace{\tilde{\mu}_2, \tilde{\mu}_2, \ldots, \tilde{\mu}_2}_{k_2 \text{ terms}}, \ldots, \underbrace{\tilde{\mu}_{d-1}, \tilde{\mu}_{d-1}, \ldots, \tilde{\mu}_{d-1}}_{k_{d-1} \text{ terms}}, \underbrace{s_d[1], s_d[1], \ldots, s_d[1]}_{n-k_1-\ldots-k_d \text{ terms}}).
\end{aligned}$$

Next we write $\mathcal{R}_n(\mathcal{F}, \mu_{1:n}, x_{1:n}, \boldsymbol{\varepsilon})$ in terms of segments 1 to d. For any $\boldsymbol{\varepsilon} \in \{0, 1\}^n$, we construct $\tilde{\boldsymbol{\varepsilon}} \in \{-1, 1\}^d$:

$$\tilde{\varepsilon}_t = 2\mathbb{I}\left\{ \sum_{j=1}^{k_t} \varepsilon_{k_1+\ldots+k_{t-1}+j} \ge 0 \right\} - 1, \qquad \forall t \in [d-1] \quad \text{and} \quad \tilde{\varepsilon}_d = 1.$$

Recall that $f^{\tilde{\boldsymbol{\varepsilon}}} \in \mathcal{F}$ for any sequence $\tilde{\boldsymbol{\varepsilon}} \in \{-1, 1\}^d$. Then, using $[t, j]$ to denote $k_1 + \ldots + k_t + j$, and defining the random variable $\widehat{\varepsilon}_t = \frac{1}{k_t} \sum_{j=1}^{k_t} \varepsilon_{[t-1,j]} \in [-1, 1]$ for fixed $\boldsymbol{\varepsilon}$, we obtain

$$\begin{aligned}
&\mathcal{R}_n(\mathcal{F}, \mu_{1:n}, x_{1:n}, \boldsymbol{\varepsilon}) \\
&= \sup_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{d-1} \sum_{j=1}^{k_t} C \cdot \varepsilon_{[t-1,j]} (f(x_{[t-1,j]}) - \mu_{[t-1,j]}) - (f(x_{[t-1,j]}) - \mu_{[t-1,j]})^2 \right. \\
&\qquad \left. + \sum_{j=1}^{n-k_1-\ldots-k_{d-1}} C \cdot \varepsilon_{[d-1,j]} (f(x_{[d-1,j]}) - \mu_{[d-1,j]}) - (f(x_{[d-1,j]}) - \mu_{[d-1,j]})^2 \right\} \\
&\overset{(i)}{\ge} \sum_{t=1}^{d-1} \sum_{j=1}^{k_t} C \cdot \varepsilon_{[t-1,j]} (f^{\tilde{\boldsymbol{\varepsilon}}}(t) - \tilde{\mu}_t) - (f^{\tilde{\boldsymbol{\varepsilon}}}(t) - \tilde{\mu}_t)^2 - n \cdot 2C \cdot \frac{\alpha}{nC} \\
&\ge \sum_{t=1}^{d-1} k_t \cdot \left(C \cdot \widehat{\varepsilon}_t (f^{\tilde{\boldsymbol{\varepsilon}}}(t) - \tilde{\mu}_t) - (f^{\tilde{\boldsymbol{\varepsilon}}}(t) - \tilde{\mu}_t)^2 \right) - 4,
\end{aligned} \tag{E.15}$$

where (i) uses the choice $f = f^{\tilde{\boldsymbol{\varepsilon}}} \in \mathcal{F}$, and also $|f^{\tilde{\boldsymbol{\varepsilon}}}(d) - s_d[1]| \le \alpha/(20nC) < \alpha/(nC)$ since $\tilde{\varepsilon}_d = 1$, and also the fact that $\alpha \le 2$ due to eq. (E.14). According to eq. (E.14), we have for any $\tilde{\boldsymbol{\varepsilon}}$,

$$|f^{\tilde{\boldsymbol{\varepsilon}}}(t) - s_t[\tilde{\varepsilon}_t]| \le \frac{\alpha}{20} \quad \text{and} \quad s_t[\tilde{\varepsilon}_t] - \mu_t = \operatorname{sign}(\widehat{\varepsilon}_t) \cdot \frac{s_t[1] - s_t[-1]}{2}.$$

eq. (E.14) also gives that $s_t[1] - s_t[-1] \ge \alpha$, which implies

$$\frac{9}{10} \cdot \frac{s_t[1] - s_t[-1]}{2} \le \operatorname{sign}(\widehat{\varepsilon}_t)(f^{\tilde{\boldsymbol{\varepsilon}}}(t) - \tilde{\mu}_t) \le \frac{11}{10} \cdot \frac{s_t[1] - s_t[-1]}{2}.$$

Therefore we can lower bound eq. (E.15) by

$$\mathcal{R}_n(\mathcal{F}, \mu_{1:n}, x_{1:n}, \boldsymbol{\varepsilon}) \ge \sum_{t=1}^{d-1} k_t \cdot \left(C|\widehat{\varepsilon}_t| \cdot \frac{9}{10} \cdot \frac{s_t[1] - s_t[-1]}{2} - \frac{121}{100} \cdot \left(\frac{s_t[1] - s_t[-1]}{2} \right)^2 \right) - 4.$$

Next, we can further lower bound

$$\begin{aligned}
&\mathcal{R}_n(\mathcal{F}, \mu_{1:n}, x_{1:n}, \boldsymbol{\varepsilon}) \\
&\quad \ge \sum_{t=1}^{d-1} k_t \cdot \left(C|\widehat{\varepsilon}_t| \cdot \frac{9}{10} \cdot \frac{s_t[1] - s_t[-1]}{2} - \frac{121}{100} \cdot \left(\frac{s_t[1] - s_t[-1]}{2} \right)^2 \right) - 4 \\
&\quad = \sum_{t=1}^{d-1} k_t \cdot \left(C\mathbb{E}\left[|\widehat{\varepsilon}_t|\right] \cdot \frac{9}{10} \cdot \frac{s_t[1] - s_t[-1]}{2} - \frac{121}{100} \cdot \left(\frac{s_t[1] - s_t[-1]}{2} \right)^2 \right) - 4 \\
&\quad \ge \sum_{t=1}^{d-1} k_t \cdot \left(C\sqrt{\frac{1}{2k_t}} \cdot \frac{9}{10} \cdot \frac{s_t[1] - s_t[-1]}{2} - \frac{121}{100} \cdot \left(\frac{s_t[1] - s_t[-1]}{2} \right)^2 \right) - 4 \qquad \text{(E.16)}
\end{aligned}$$

where the last inequality uses the Khintchine inequality [222]. According to our choice of k_t,

$$k_t = \left\lfloor \frac{1}{(s_t[1] - s_t[-1])^2} \right\rfloor \vee 1 \in \left[\frac{1}{2(s_t[1] - s_t[-1])^2}, \frac{4}{(s_t[1] - s_t[-1])^2} \right],$$

which implies that $1/\sqrt{2} \le \sqrt{k_t}(s_t[1] - s_t[-1]) \le 2$. Therefore, since $C \ge 2$, we have

$$\text{RHS of } eq.\ (\text{E.16}) \ge \sum_{t=1}^{d-1} \left(\frac{9C}{20\sqrt{2}} - \frac{121}{200} \right) \cdot \sqrt{k_t} \cdot |s_t[1] - s_t[-1]| - 4 \ge \frac{d-1}{50} - 4.$$

Therefore, we obtain that

$$\sup_{\mu_{1:n}, x_{1:n}} \mathbb{E}[\mathcal{R}_n(\mathcal{F}, \mu_{1:n}, x_{1:n}, \boldsymbol{\varepsilon})] \ge \frac{d-1}{50} - 4 = \tilde{\Omega}\left(n^{\frac{p}{p+2}} \right).$$

□

Proof of Theorem 6.12. We first transform $\mathcal{V}_n(\mathcal{F})$ into the following dual form

$$\mathcal{V}_n(\mathcal{F}) = \sup_{x_{1:n} \in \mathcal{X}^n} \left\{ \inf_{\widehat{y}_t} \sup_{p_t \in \Delta([-2,2])} \mathbb{E}_{y_t \sim p_t} \right\}_{t=1}^{n} \left[\sum_{t=1}^{n} (\widehat{y}_t - y_t)^2 - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} (f(x_t) - y_t)^2 \right]$$

$$= \sup_{x_{1:n}\in\mathcal{X}^n} \left\{ \sup_{p_t\in\Delta([-2,2])} \mathbb{E}_{y_t\sim p_t} \right\}_{t=1}^{n} \left[\sum_{t=1}^{n} \left(\mathbb{E}[y_t \mid y_{1:t-1}] - y_t\right)^2 - \inf_{f\in\mathcal{F}} \sum_{t=1}^{n} \left(f(x_t) - y_t\right)^2 \right]$$
$$= \sup_{x_{1:n}} \sup_{\mathbf{p}\in\Delta([-2,2]^n)} \mathbb{E}_{\mathbf{y}\sim\mathbf{p}} \left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{n} 2(y_t - \mu_t(\mathbf{y}))(f(x_t) - \mu_t(\mathbf{y})) - (f(x_t) - \mu_t(\mathbf{y}))^2 \right]$$

where we use $\mu_t(\mathbf{y})$ to denote the expectation of y_t conditioned on $y_{1:t-1}$, i.e. $\mu_t(\mathbf{y}) = \mathbb{E}[y_t \mid y_{1:t-1}]$. Taking $\mathbf{p} = p_1 \otimes p_2 \otimes \ldots \otimes p_n$ where $p_t \in \Delta([-2,2])$ for each $t \in [n]$ in the above equation, we obtain that

$$\mathcal{V}_n(\mathcal{F}) \geq \sup_{x_{1:n}} \sup_{p_{1:n}\in\Delta([-2,2]} \mathbb{E}_{y_t\sim p_t} \left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{n} 2(y_t - \mu_t)(f(x_t) - \mu_t) - (f(x_t) - \mu_t)^2 \right],$$

where $\mu_t \coloneqq \mathbb{E}[y_t]$ denotes the expectation of distribution p_t.

We fix $\mu_1, \ldots, \mu_n \in [-1,1]$, and choose distributions $p_t = \mathrm{Unif}(\{\mu_t - 1, \mu_t + 1\}) \subseteq \Delta([-2,2])$ for all $t \in [n]$. Then we obtain that

$$\mathcal{V}_n(\mathcal{F}) \geq \sup_{x_{1:n}} \sup_{\mu_{1:n}\in[-1,1]} \mathbb{E}_{\varepsilon_t} \left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{n} 2\varepsilon_t(f(x_t) - \mu_t) - (f(x_t) - \mu_t)^2 \right],$$

where $\varepsilon_{1:n} \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(\{-1,1\})$. □

Proof of Theorem 6.13. Since $\sup_{x_{1:n}} \log \mathcal{N}_\infty(\mathcal{F}, x_{1:n}, \alpha) = \tilde{\Omega}\left(\alpha^{-p}\right)$, according to Theorem 6.6 we have

$$d(\mathcal{F}, \alpha, \alpha/(40n)) = \tilde{\Omega}\left(\alpha^{-p}\right).$$

Next, we call Theorem 6.12, and Theorem 6.11 with $C = 2$. Then we obtain

$$\mathcal{V}_n(\mathcal{F}) = \tilde{\Omega}\left(n^{\frac{p}{p+2}}\right).$$

□

E.2 Missing Proofs in section 6.3

E.2.1 Proof of Theorem 6.16

Proof of Lemma Theorem 6.16. We first define function

$$g_M(n,d) = \sum_{i=0}^{d} \binom{n}{i} \cdot (M-1)^i,$$

which satisfies (see [20])

$$g_M(n,d) = g_M(n-1,d) + (M-1) \cdot g_M(n-1,d-1). \tag{E.17}$$

We will prove the present Lemma by induction with the following induction statement:

$\mathfrak{G}(d,n)$: For any function class $\mathcal{F} \subseteq \{f : \mathcal{X} \to [M]\}$ with $d^{\mathsf{seq}}(\mathcal{F}, \alpha) \leq d$, and any depth-$n$ $\mathcal{X}$-valued tree $\mathbf{x}$, $\mathcal{N}_\infty^{\mathsf{seq}}(\mathcal{F}, \mathbf{x}, \alpha) \leq g_M(d,n)$.

Base: There are two base cases to consider: $n \le d$ and $d = 0$.

When $n \le d$, we let

$$\mathcal{V} = \{\mathbf{v}[i_1, i_2, \ldots, i_n] : i_1, \ldots, i_n \in [M]\},$$

where $\mathbf{v}[i_1, \ldots, i_n]$ denotes the tree with values i_t at depth t along any path. Then it is easy to see that for any $f \in \mathcal{F}$, depth-n $\mathcal{X}$-valued tree $\mathbf{x}$, and any path $\boldsymbol{\varepsilon} \in \{-1, 1\}^n$, there exists some $\mathbf{v} \in \mathcal{V}$ such that $f(x_t(\boldsymbol{\varepsilon})) = v_t(\boldsymbol{\varepsilon})$ for all $t \in [n]$. Hence $\mathcal{V}$ is an exact (that is, $\alpha = 0$) cover of $\mathcal{F}$ on $\mathbf{x}$; hence, $\mathcal{V}$ is also an α-sequential covering as well. Thus, we have

$$\mathcal{N}_\infty^{\mathsf{seq}}(\mathcal{F}, \mathbf{x}, \alpha) \le |\mathcal{V}| = M^n = \sum_{i=0}^{d} \binom{n}{i} \cdot (M-1)^i = g_M(n, d).$$

When $d = 0$, there is no depth-1 $\mathcal{X}$-valued tree which is shattered by $\mathcal{F}$ at scale α under the metric $\mathfrak{c}$. This implies for any $x \in \mathcal{X}$ and functions $f, f' \in \mathcal{F}$, we always have $\mathfrak{c}(f(x), f'(x)) \le \alpha$. Indeed, otherwise we can construct depth-1 tree $\mathbf{x}$ with $x_1 = x$ and depth-1 tree $\mathbf{s}$ with $s_1 = (f(x), f'(x))$, and $\mathbf{x}$ is shattered by $\mathcal{F}$. We let $f_0 \in \mathcal{F}$ to be an arbitrary function in $\mathcal{F}$. For any depth-n $\mathcal{X}$-valued $\mathbf{x}$, we construct depth-n $[-1, 1]$-valued tree $\mathbf{v}$ which takes the value $f_0(x_t(\boldsymbol{\varepsilon}))$ at depth t along any path $\boldsymbol{\varepsilon}$. Then for any $f \in \mathcal{F}$ and path $\boldsymbol{\varepsilon} \in \{-1, 1\}$, we always have $\mathfrak{c}(f(x_t(\boldsymbol{\varepsilon})), v_t(\boldsymbol{\varepsilon})) = \mathfrak{c}(f(x_t(\boldsymbol{\varepsilon})), f_0(x_t(\boldsymbol{\varepsilon}))) \le \alpha$. Hence $\mathcal{V}$ is an α-sequential covering of $\mathcal{F}$ on $\mathbf{x}$, and it satisfies $|\mathcal{V}| = 1 = g_M(n, 0)$.

Induction: Suppose the induction hypotheses $\mathfrak{G}(n-1, d-1)$ and $\mathfrak{G}(n-1, d)$ both hold. We will prove induction statement $\mathfrak{G}(n, d)$. For fixed function class $\mathcal{F}$ with $d^{\mathsf{seq}}(\mathcal{F}, \alpha) = d$ and depth-n $\mathcal{X}$-valued tree $\mathbf{x}$, we will construct a α-sequential covering to of $\mathcal{F}$ on $\mathbf{x}$ whose size is no more than $g_M(n, d)$. Suppose the root of tree $\mathbf{x}$ is $x_1 \in \mathcal{X}$, the left subtree of x_1 is denoted as $\mathbf{x}^l$, and the right subtree of x_1 is denoted as $\mathbf{x}^r$. We partition the function class $\mathcal{F}$ as:

$$\mathcal{F} = \mathcal{F}_1 \cup \mathcal{F}_2 \cup \ldots \cup \mathcal{F}_M \quad \text{where} \quad \mathcal{F}_k = \{f \in \mathcal{F} : f(x_1) = k\}, \quad \forall 1 \le k \le M.$$

Then we have $d^{\mathsf{seq}}(\mathcal{F}_k, \alpha) \le d^{\mathsf{seq}}(\mathcal{F}, \alpha) = d$ for all $k \in [M]$. We let $\mathcal{K} = \{k \in [M] : d^{\mathsf{seq}}(\mathcal{F}_k, \alpha) = d\}$. Then for any $a, b \in \mathcal{K}$ and $a < b$, there exist depth-d $\mathcal{X}$-valued trees $\mathbf{x}^a$ and $\mathbf{x}^b$, and also depth-d $[M] \times [M]$-valued trees $\mathbf{s}^a$ and $\mathbf{s}^b$ such that for any $\boldsymbol{\varepsilon} \in \{-1, 1\}^d$, there exists $f_a^{\boldsymbol{\varepsilon}} \in \mathcal{F}_a$ and $f_b^{\boldsymbol{\varepsilon}} \in \mathcal{F}_b$ such that for any $t \in [d]$,

$$\begin{aligned} f_a^{\boldsymbol{\varepsilon}}(x_t^a(\boldsymbol{\varepsilon})) = s_t^a(\boldsymbol{\varepsilon})[\varepsilon_t] \quad &\text{and} \quad f_b^{\boldsymbol{\varepsilon}}(x_t^b(\boldsymbol{\varepsilon})) = s_t^b(\boldsymbol{\varepsilon})[\varepsilon_t], \\ \mathfrak{c}(s_t^a(\boldsymbol{\varepsilon})[-1], s_t^a(\boldsymbol{\varepsilon})[1]) \ge \alpha \quad &\text{and} \quad \mathfrak{c}(s_t^b(\boldsymbol{\varepsilon})[-1], s_t^b(\boldsymbol{\varepsilon})[1]) \ge \alpha. \end{aligned}$$

For the sake of a contradiction, suppose it holds that $\mathfrak{c}(a, b) \ge \alpha$. Then we can construct a depth-$(d+1)$ $\mathcal{X}$-valued tree $\mathbf{x}$ with root x_1, left subtree of the root being $\mathbf{x}^a$, and right subtree of the root being $\mathbf{x}^b$, and also a depth-$(d+1)$ $[M] \times [M]$-valued tree $\mathbf{s}$ with root $(a, b) \in [M] \times [M]$, left subtree of the root being $\mathbf{s}^a$, and right subtree of the root being $\mathbf{s}^b$. Then we can verify that for any $\boldsymbol{\varepsilon} \in \{-1, 1\}^{d+1}$, and any $t \in [d+1]$, we have $s_t(\boldsymbol{\varepsilon})[-1] < s_t(\boldsymbol{\varepsilon})[1]$, and $\mathfrak{c}(s_t(\boldsymbol{\varepsilon})[-1], s_t(\boldsymbol{\varepsilon})[1]) \ge \alpha$. Further, we let $\boldsymbol{\varepsilon}' = (\varepsilon_2, \varepsilon_3, \ldots, \varepsilon_{d+1}) \in \{-1, 1\}^d$, and if $\varepsilon_1 = -1$, then letting $f^{\boldsymbol{\varepsilon}} = f_a^{\boldsymbol{\varepsilon}'}$ we can verify that $f^{\boldsymbol{\varepsilon}}(x_t(\boldsymbol{\varepsilon})) = s_t(\boldsymbol{\varepsilon})[\varepsilon_t]$ for any $t \in [d+1]$, and if $\varepsilon_1 = 1$,

then letting $f^{\boldsymbol{\varepsilon}} = f_b^{\boldsymbol{\varepsilon}'}$ we can verify that $f^{\boldsymbol{\varepsilon}}(x_t(\boldsymbol{\varepsilon})) = s_t(\boldsymbol{\varepsilon})[\varepsilon_t]$ for any $t \in [d+1]$. Hence, depth-$(k+1)$ tree $\mathbf{x}$ is shattered by $\mathcal{F}$, which leads to contradiction. Therefore, we have

$$\mathfrak{c}(a,b) < \alpha, \qquad \forall a, b \in \mathcal{K}. \tag{E.18}$$

Next, for any $k \in [M]$ with $d^{\mathsf{seq}}(\mathcal{F}_k, \alpha) \le d-1$, according to the induction hypothesis $\mathfrak{G}(n-1, d-1)$, there exists a sequential cover $\mathcal{V}_k^l$ of size $g_M(n-1,d-1)$ for $\mathcal{F}$ on the depth-$(n-1)$ $\mathcal{X}$-valued tree $\mathbf{x}^l$, and also a sequential cover $\mathcal{V}_k^r$ of size $g_M(n-1,d-1)$ for $\mathcal{F}$ on the depth-$(n-1)$ $\mathcal{X}$-valued tree $\mathbf{x}^r$. We then combine the elements in $\mathcal{V}_k^l$ and $\mathcal{V}_k^r$ into a set $\mathcal{V}_k$ of depth-n $[M]$-valued trees by a joining process as follows. We let $v_1 = k \in [M]$, and according to the construction of $\mathcal{F}_k$ we have for any $f \in \mathcal{F}$ that $f(x_1) = v_1$, and thus $\mathfrak{c}(f(x_1), v_1) \le \alpha$. For $\mathbf{v}^l \in \mathcal{V}_k^l$ and $\mathbf{v}^r \in \mathcal{V}_k^r$, we define depth-$n$ $[M]$-valued tree $\mathbf{v}[\mathbf{v}^l, \mathbf{v}^r]$ as: for any path $\boldsymbol{\varepsilon} \in \{-1,1\}^n$, we let $\boldsymbol{\varepsilon}' = (\varepsilon_{2:n}) \in \{-1,1\}^{n-1}$, and let $v_1[\mathbf{v}^l, \mathbf{v}^r](\boldsymbol{\varepsilon}) = v_1$. If $\varepsilon_1 = -1$, then let $v_t[\mathbf{v}^l, \mathbf{v}^r](\boldsymbol{\varepsilon}) = v_{t-1}^l(\boldsymbol{\varepsilon}')$, and if $\varepsilon_1 = 1$, then let $v_t[\mathbf{v}^l, \mathbf{v}^r](\boldsymbol{\varepsilon}) = v_{t-1}^r(\boldsymbol{\varepsilon}')$. We construct $\mathcal{V}_k = \{\mathbf{v}[\mathbf{v}^l, \mathbf{v}^r]\}$ with $|\mathcal{V}_k| \le \max\{|\mathcal{V}_k^l|, |\mathcal{V}_k^r|\}$ to make sure that every element in $\mathcal{V}_k^l$ and $\mathcal{V}_k^r$ appears at least once in the construction of $\mathcal{V}_k$. Next, we will argue that $\mathcal{V}_k$ is an α-sequential cover of $\mathcal{F}_k$ on $\mathbf{x}$. This is easy to see by construction: for any $f \in \mathcal{F}_k$ and $\boldsymbol{\varepsilon} \in \{-1,1\}^n$, if $\varepsilon_1 = -1$, then since $\mathcal{V}_k^l$ is a α-sequential cover of $\mathcal{F}_k$, there exists $\mathbf{v}^l \in \mathcal{V}_k^l$ such that for any $2 \le t \le n$, $\mathfrak{c}(f(x_t(\boldsymbol{\varepsilon})), v_t^l(\boldsymbol{\varepsilon})) \le \alpha$. Suppose $\mathbf{v} = \mathbf{v}[\mathbf{v}^l, \mathbf{v}^r] \in \mathcal{V}_k$ for some $\mathbf{v}^r \in \mathcal{V}_k^r$, and we also have $\mathfrak{c}(f(x_1(\boldsymbol{\varepsilon})), v_1(\boldsymbol{\varepsilon})) \le \alpha$ according to the construction of $\mathcal{F}_k$. Hence, for any $t \in [n]$, we always $\mathfrak{c}(f(x_t(\boldsymbol{\varepsilon})), v_t(\boldsymbol{\varepsilon})) \le \alpha$. Therefore, $\mathcal{V}_k$ is a cover of $\mathcal{F}_k$ on $\mathbf{x}$. Further by induction hypothesis we have $\max\{|\mathcal{V}_k^l|, |\mathcal{V}_k^r|\} \le g_M(n-1,d-1)$. Hence $|\mathcal{V}_k| \le g_M(n-1,d-1)$.

If $\mathcal{K} = \emptyset$, then by letting $\mathcal{V} = \cup_{k\in[M]}\mathcal{V}_k$, $\mathcal{V}$ is an α-sequential cover of $\mathcal{F}$ on $\mathbf{x}$, and also

$$|\mathcal{V}| \le M \cdot g_M(n-1,d-1) \le g_M(n-1,d) + (M-1)g_M(n-1,d-1) = g_M(n,d),$$

where the inequality follows from the fact that $g_M(n-1,d-1) \le g_M(n-1,d)$ for any n, d, and the last equation follows from eq. (E.17).

Next, we consider cases where $|\mathcal{K}| \ge 1$. Consider the function class $\mathcal{F}' = \cup_{k\in\mathcal{K}}\mathcal{F}_k \subseteq \mathcal{F}$. We have $d^{\mathsf{seq}}(\mathcal{F}', \alpha) \le d^{\mathsf{seq}}(\mathcal{F}, \alpha) = d$. According to the induction hypothesis $\mathfrak{G}(n-1,d)$, there exists a sequential cover $\mathcal{V}^l$ of size $g_M(n-1,d)$ for the depth-$(n-1)$ $\mathcal{X}$-valued tree $\mathbf{x}^l$, and also a sequential cover $\mathcal{V}^l$ of size $g_M(n-1,d)$ for the depth-$(n-1)$ $\mathcal{X}$-valued tree $\mathbf{x}^l$. As before, we combine the elements in $\mathcal{V}^l$ and $\mathcal{V}^r$ into a set $\mathcal{V}'$ of depth-n $[M]$-valued trees. We let $v_1 = f(x_1) \in [M]$ for some $f \in \mathcal{F}'$, chosen arbitrarily. Then, according to the construction of $\mathcal{F}'$, we have for any other $g \in \mathcal{F}'$, $\mathfrak{c}(g(x_1), v_1) \le \alpha$. For $\mathbf{v}^l \in \mathcal{V}^l$ and $\mathbf{v}^r \in \mathcal{V}^r$, we define depth-$n$ $[M]$-valued tree $\mathbf{v}[\mathbf{v}^l, \mathbf{v}^r]$ by joining them with v_1 at the root as before: for any path $\boldsymbol{\varepsilon} \in \{-1,1\}^n$, we let $\boldsymbol{\varepsilon}' = (\varepsilon_{2:n}) \in \{-1,1\}^{n-1}$, and let $v_1[\mathbf{v}^l, \mathbf{v}^r](\boldsymbol{\varepsilon}) = v_1$. If $\varepsilon_1 = -1$, then let $v_t[\mathbf{v}^l, \mathbf{v}^r](\boldsymbol{\varepsilon}) = v_{t-1}^l(\boldsymbol{\varepsilon}')$, and if $\varepsilon_1 = 1$, then let $v_t[\mathbf{v}^l, \mathbf{v}^r](\boldsymbol{\varepsilon}) = v_{t-1}^r(\boldsymbol{\varepsilon}')$. We construct $\mathcal{V}' = \{\mathbf{v}[\mathbf{v}^l, \mathbf{v}^r]\}$ with $|\mathcal{V}'| \le \max\{|\mathcal{V}^l|, |\mathcal{V}^r|\}$ to make sure that every element in $\mathcal{V}^l$ and $\mathcal{V}^r$ appears at least once in the construction of $\mathcal{V}'$. Next, we will argue that $\mathcal{V}'$ is a α-sequential cover of $\mathcal{F}'$ on $\mathbf{x}$. For any $f \in \mathcal{F}'$ and $\boldsymbol{\varepsilon} \in \{-1,1\}^n$, if $\varepsilon_1 = -1$, then since $\mathcal{V}^l$ is a α-sequential cover of $\mathcal{F}'$ on the tree $\mathbf{x}^l$, there exists $\mathbf{v}^l \in \mathcal{V}^l$ such that for any $2 \le t \le n$, $\mathfrak{c}(f(x_t(\boldsymbol{\varepsilon})), v_t^l(\boldsymbol{\varepsilon})) \le \alpha$. Suppose $\mathbf{v} = \mathbf{v}[\mathbf{v}^l, \mathbf{v}^r] \in \mathcal{V}'$ for some $\mathbf{v}^r \in \mathcal{V}^r$ (according to the construction of $\mathcal{V}'$ such $\mathbf{v}$ exists). Then since $f \in \mathcal{F}'$, according to eq. (E.18) we have $\mathfrak{c}(f(x_1(\boldsymbol{\varepsilon})), v_1(\boldsymbol{\varepsilon})) \le \alpha$. Hence for any $t \in [n]$, we always $\mathfrak{c}(f(x_t(\boldsymbol{\varepsilon})), v_t(\boldsymbol{\varepsilon})) \le \alpha$. Similarly, if $\varepsilon_1 = 1$, there also exists some $\mathbf{v} \in \mathcal{V}'$ such that $\mathfrak{c}(f(x_t(\boldsymbol{\varepsilon})), v_t(\boldsymbol{\varepsilon})) \le \alpha$ for any $t \in [n]$.

Therefore, $\mathcal{V}'$ is an α-sequential cover of $\mathcal{F}'$. Further by induction hypothesis we have $\max\{|\mathcal{V}^l|, |\mathcal{V}^r|\} \le g_M(n-1,d)$. Hence $|\mathcal{V}'| \le g_M(n-1,d)$. We now let $\mathcal{V} = \mathcal{V}' \cup (\cup_{k\notin\mathcal{K}}\mathcal{V}_k)$, and after noticing that $|[M]\backslash\mathcal{K}| \le (M-1)$, we obtain

$$|\mathcal{V}| \le (M-1)\cdot g_M(n-1,d-1) + g_M(n-1,d) = g_M(n,d),$$

where the last equation follows from eq. (E.17). This finishes the proof of the induction statement $\mathfrak{G}(n,d)$.

We conclude that, by induction, we have for any depth-n $\mathcal{X}$-valued tree $\mathbf{x}$,

$$\mathcal{N}_\infty^{\mathsf{seq}}(\mathcal{F},\mathbf{x},\alpha) \le g_M(n, d^{\mathsf{seq}}(\mathcal{F},\alpha)) = \sum_{i=0}^{d^{\mathsf{seq}}(\mathcal{F},\alpha)} \binom{n}{i}\cdot(M-1)^i \overset{(i)}{\le} \left(\frac{enM}{d^{\mathsf{seq}}(\mathcal{F},\alpha)}\right)^{d^{\mathsf{seq}}(\mathcal{F},\alpha)} \le (enM)^{d^{\mathsf{seq}}(\mathcal{F},\alpha)},$$

where (i) follows e.g. from [118, Theorem 7]. □

E.2.2 Proof of Theorem 6.19

Proof of Theorem 6.19. We define distance $\mathfrak{c}' : [M]\times[M] \to \mathbb{R}_+ \cup \{0\}$:

$$\mathfrak{c}'(a,b) = \mathfrak{c}(u_a, u_b). \tag{E.19}$$

For any $\mathcal{F} \subseteq \{f : \mathcal{X} \to [-1,1]\}$, since f maps $\mathcal{X}$ into $[-1,1]$, there exists $\bar{f} : \mathcal{X} \to [M]$ such that for any $x \in \mathcal{X}$, $\mathfrak{c}(f(x), u_{\bar{f}(x)}) \le \beta$. We define function class $\bar{\mathcal{F}} = \{\bar{f} : f \in \mathcal{F}\} \subseteq \{\mathcal{X} \to [M]\}$. We use $d_{\mathfrak{c}'}^{\mathsf{seq}}(\bar{\mathcal{F}},\alpha)$ to denote the sequential gapped dimension of integer-valued function class $\bar{\mathcal{F}}$ at scale α under distance $\mathfrak{c}'$, where the dimension is defined in Theorem 6.14, and for simplicity we let $d = d_{\mathfrak{c}'}^{\mathsf{seq}}(\bar{\mathcal{F}},\alpha)$. Suppose $\mathbf{x}$ is a depth-d tree shattered by $\mathcal{F}$. Then there exists a depth-d $([M]\times[M])$-valued tree $\bar{\mathbf{s}}$ such that for any $\boldsymbol{\varepsilon} \in \{-1,1\}^d$ and $t \in [d]$, $\mathfrak{c}'(\bar{s}_t(\boldsymbol{\varepsilon})[-1], \bar{s}_t(\boldsymbol{\varepsilon})[1]) \ge \alpha$, and also for any $\boldsymbol{\varepsilon} \in \{-1,1\}^d$, there exists $\bar{f}^{\boldsymbol{\varepsilon}} \in \bar{\mathcal{F}}$ such that $\bar{f}^{\boldsymbol{\varepsilon}}(x_t(\boldsymbol{\varepsilon})) = \bar{s}_t(\boldsymbol{\varepsilon})[\varepsilon_t]$ for any $t \in [d]$. Notice from the definition of $\mathcal{F}$ there exists some $f^{\boldsymbol{\varepsilon}} \in \mathcal{F}$ such that $\bar{f}^{\boldsymbol{\varepsilon}} = \overline{(f^{\boldsymbol{\varepsilon}})}$.

We next verify that tree $\mathbf{x}$ is also shattered by $\mathcal{F}$ at scale (α,β), according to the definition Theorem 6.17. We construct depth-d $([-1,1]\times[-1,1])$-tree $\mathbf{s}$ according to $\bar{\mathbf{s}}$ as follows:

$$s_t(\boldsymbol{\varepsilon})[-1] = u_{\bar{s}_t(\boldsymbol{\varepsilon})[-1]} \quad \text{and} \quad s_t(\boldsymbol{\varepsilon})[1] = u_{\bar{s}_t(\boldsymbol{\varepsilon})[1]}.$$

Then according to the definition of $\mathfrak{c}'$ in eq. (E.19), we have for any $\boldsymbol{\varepsilon} \in \{-1,1\}^d$ and $t \in [n]$,

$$\mathfrak{c}(s_t(\boldsymbol{\varepsilon})[-1], s_t(\boldsymbol{\varepsilon})[1]) = \mathfrak{c}'(\bar{s}_t(\boldsymbol{\varepsilon})[-1], \bar{s}_t(\boldsymbol{\varepsilon})[1]) \ge \alpha.$$

Additionally, for any $\boldsymbol{\varepsilon} \in \{-1,1\}^d$, there exists some $f^{\boldsymbol{\varepsilon}} \in \mathcal{F}$ such that $\overline{(f^{\boldsymbol{\varepsilon}})}(x_t(\boldsymbol{\varepsilon})) = \bar{s}_t(\boldsymbol{\varepsilon})[\varepsilon_t]$, which implies,

$$\mathfrak{c}(f(x_t(\boldsymbol{\varepsilon})), s_t(\boldsymbol{\varepsilon})[\varepsilon_t]) = \mathfrak{c}(f(x_t(\boldsymbol{\varepsilon})), u_{\bar{s}_t(\boldsymbol{\varepsilon})[\varepsilon_t]}) = \mathfrak{c}(f(x_t(\boldsymbol{\varepsilon})), u_{\overline{(f^{\boldsymbol{\varepsilon}})}(x_t(\boldsymbol{\varepsilon}))}) \le \beta, \quad \forall t \in [d],$$

where the last inequality follows from the definiton of $\bar{f}$. Therefore, $\mathbf{x}$ is shattered by $\mathcal{F}$ at scale (α,β), and according to Theorem 6.17 we have $d^{\mathsf{seq}}(\mathcal{F},\alpha,\beta) \ge d$.

Next we will upper bound the sequential covering number of $\mathcal{F}$ in terms of d. For a fixed depth-n $\mathcal{X}$-valued tree $\mathbf{x}$, according to Theorem 6.16, there exists a sequential covering $\bar{\mathcal{V}}$ of $\bar{\mathcal{F}}$ with size no more than $(neM)^d$. Hence for any $f \in \mathcal{F}$ and $\boldsymbol{\varepsilon} \in \{-1,1\}^d$, since $\bar{f} \in \bar{\mathcal{F}}$, there exists some $\bar{\mathbf{v}} \in \bar{\mathcal{V}}$ which satisfies

$$\mathfrak{c}'(\bar{f}(x_t(\boldsymbol{\varepsilon})), \bar{v}_t(\boldsymbol{\varepsilon})) \leq \alpha \qquad \forall t \in [n],$$

which implies that

$$\mathfrak{c}(f(x_t(\boldsymbol{\varepsilon})), u_{\bar{v}_t(\boldsymbol{\varepsilon})}) \leq \mathfrak{c}(f(x_t(\boldsymbol{\varepsilon})), u_{\bar{f}(x_t(\boldsymbol{\varepsilon}))}) + \mathfrak{c}(u_{\bar{f}(x_t(\boldsymbol{\varepsilon}))}, u_{\bar{v}_t(\boldsymbol{\varepsilon})}) \leq \beta + \alpha \qquad \forall t \in [n],$$

where the first inequality uses the triangle inequality of $\mathfrak{c}$, and the second inequality uses the definition of function $\bar{f}$. For every $\bar{\mathbf{v}} \in \bar{\mathcal{V}}$, we construct depth-$d$ $\mathbb{R}$-valued tree $\mathbf{v}_{\bar{\mathbf{v}}}$ where for any $\boldsymbol{\varepsilon} \in \{-1,1\}^d$ and $t \in [n]$, $(v_{\bar{\mathbf{v}}})_t(\boldsymbol{\varepsilon}) = u_{\bar{v}_t(\boldsymbol{\varepsilon})}$, for every $\bar{\mathbf{v}} \in \bar{\mathcal{V}}$. And we further let $\mathcal{V} = \{\mathbf{v}_{\bar{\mathbf{v}}} : \bar{\mathbf{v}} \in \bar{\mathcal{V}}\}$. Then $\mathcal{V}$ is an $(\alpha+\beta)$-cover of $\mathcal{F}$ on tree $\mathbf{x}$, which implies

$$\mathcal{N}_\infty^{\mathsf{seq}}(\mathcal{F}, \mathbf{x}, \alpha+\beta) \leq |\mathcal{V}| = |\bar{\mathcal{V}}| \leq (neM)^d.$$

□

E.2.3 Missing Proofs in section 6.3.3

Proof of Theorem 6.21. We only need to prove that if a depth-d $\mathcal{X}$-valued tree $\mathbf{x}$ is shattered by $\mathcal{F}$ at scale (α, β) according to Theorem 6.17, then $\mathbf{x}$ is shattered by $\mathcal{F}$ at scale $\alpha - 2\beta$ according to Theorem 6.20.

If $\mathbf{x}$ is shattered by $\mathcal{F}$ at scale (α, β) according to Theorem 6.17, then there exists a depth-d $([0,1] \times [0,1])$-valued tree $\mathbf{s}$ such that for any $\boldsymbol{\varepsilon} \in \{-1,1\}^n$, there exists a function $f^{\boldsymbol{\varepsilon}} \in \mathcal{F}$ such that for any $t \in [d]$, we have

$$|f^{\boldsymbol{\varepsilon}}(x_t(\boldsymbol{\varepsilon})) - s_t(\boldsymbol{\varepsilon})[\varepsilon_t]| \leq \beta \quad \text{and} \quad |s_t(\boldsymbol{\varepsilon})[1] - s_t(\boldsymbol{\varepsilon})[-1]| \geq \alpha. \tag{E.20}$$

Without loss of generality we assume $s_t(\boldsymbol{\varepsilon})[1] > s_t(\boldsymbol{\varepsilon})[-1]$ for any $\boldsymbol{\varepsilon}$ and $t \in [d]$. We define depth-d $[0,1]$-valued tree $\mathbf{v}$ as follows: for any $\boldsymbol{\varepsilon} \in \{-1,1\}^d$ and $t \in [d]$, let

$$v_t(\boldsymbol{\varepsilon}) = \frac{s_t(\boldsymbol{\varepsilon})[-1] + s_t(\boldsymbol{\varepsilon})[1]}{2}.$$

Since $\alpha > 2\beta$, according to eq. (E.20) we have for any $\boldsymbol{\varepsilon} \in \{-1,1\}^n$,

$$\varepsilon_t \cdot (f^{\boldsymbol{\varepsilon}}(x_t(\boldsymbol{\varepsilon})) - v_t(\boldsymbol{\varepsilon})) \geq \frac{\alpha}{2} - \beta.$$

Therefore, $\mathbf{x}$ is shattered by $\mathcal{F}$ with $\mathbf{v}$ being its witness at scale $\alpha - 2\beta$, according to Theorem 6.20. □

Proof of Theorem 6.22. We first show that if depth-d $\mathcal{X}$-valued tree $\mathbf{x}$ is shattered by function class $\mathcal{F}$ at scale α, according to Theorem 6.20, then we have

$$\mathcal{N}_\infty^{\mathsf{seq}}(\mathcal{F}, \mathbf{x}, \alpha/3) \geq 2^d. \tag{E.21}$$

We use depth-d $[-1,1]$-valued tree $\mathbf{s}$ to denote the witness of shattering of $\mathbf{x}$ via $\mathcal{F}$ at scale α. According to Theorem 6.20, for any $\boldsymbol{\varepsilon} \in \{-1,1\}^d$, there exists some function $f^{\boldsymbol{\varepsilon}} \in \mathcal{F}$ such that

$$\varepsilon_t \cdot (f^{\boldsymbol{\varepsilon}}(x_t(\boldsymbol{\varepsilon})) - s_t(\boldsymbol{\varepsilon})) \geq \frac{\alpha}{2} \tag{E.22}$$

Next we will prove eq. (E.21) via contradiction. Suppose there exists an ℓ_∞-sequential covering $\mathcal{V}$ at scale $\alpha/2$ of size less than 2^d. Then for any $\boldsymbol{\varepsilon} \in \{-1,1\}^d$ and function $f \in \mathcal{F}$, there exists some tree $\mathbf{v}[f,\boldsymbol{\varepsilon}] \in \mathcal{V}$ such that

$$|f(x_t(\boldsymbol{\varepsilon})) - v_t(\boldsymbol{\varepsilon})| \leq \frac{\alpha}{3}, \qquad \forall t \in [d]. \tag{E.23}$$

Since $|\mathcal{V}| \leq 2^d - 1$, according to the pigeonhole principle there exists different $\boldsymbol{\varepsilon} = (\varepsilon_{1:d}), \boldsymbol{\varepsilon}' = (\varepsilon'_{1:d}) \in \{-1,1\}^d$ such that $\mathbf{v}[f^{\boldsymbol{\varepsilon}},\boldsymbol{\varepsilon}] = \mathbf{v}[f^{\boldsymbol{\varepsilon}'},\boldsymbol{\varepsilon}']$. We let

$$\mathbf{v}[f^{\boldsymbol{\varepsilon}},\boldsymbol{\varepsilon}] = \mathbf{v}[f^{\boldsymbol{\varepsilon}'},\boldsymbol{\varepsilon}'] = \mathbf{v}$$

We then choose r to be smallest nonnegative integer such that $\varepsilon_r \neq \varepsilon'_r$. Since $\boldsymbol{\varepsilon} \neq \boldsymbol{\varepsilon}'$ we have $1 \leq r \leq d$. Then we have $\varepsilon_{1:r-1} = \varepsilon'_{1:r-1}$, which implies that

$$x_r(\boldsymbol{\varepsilon}) = x_r(\boldsymbol{\varepsilon}'), \quad v_r(\boldsymbol{\varepsilon}) = v_r(\boldsymbol{\varepsilon}') \quad \text{and} \quad s_r(\boldsymbol{\varepsilon}) = s_r(\boldsymbol{\varepsilon}'). \tag{E.24}$$

Therefore, we obtain that

$$\begin{aligned} |f^{\boldsymbol{\varepsilon}}(x_r(\boldsymbol{\varepsilon})) - f^{\boldsymbol{\varepsilon}'}(x_r(\boldsymbol{\varepsilon}))| &= |f^{\boldsymbol{\varepsilon}}(x_r(\boldsymbol{\varepsilon})) - f^{\boldsymbol{\varepsilon}'}(x_r(\boldsymbol{\varepsilon}'))| \\ &\leq |f^{\boldsymbol{\varepsilon}}(x_r(\boldsymbol{\varepsilon})) - v_r(\boldsymbol{\varepsilon})| + |v_r(\boldsymbol{\varepsilon}) - v_r(\boldsymbol{\varepsilon}')| + |v_r(\boldsymbol{\varepsilon}') - f^{\boldsymbol{\varepsilon}'}(x_r(\boldsymbol{\varepsilon}'))| \leq \frac{\alpha}{3} + \frac{\alpha}{3} < \alpha, \end{aligned} \tag{E.25}$$

where the second inequality uses eq. (E.23) and the fact that $\mathbf{v}[f^{\boldsymbol{\varepsilon}},\boldsymbol{\varepsilon}] = \mathbf{v}[f^{\boldsymbol{\varepsilon}'},\boldsymbol{\varepsilon}'] = \mathbf{v}$. Next according to eq. (E.22), we have

$$\varepsilon_r \cdot (f^{\boldsymbol{\varepsilon}}(x_r(\boldsymbol{\varepsilon})) - s_r(\boldsymbol{\varepsilon})) \geq \frac{\alpha}{2} \quad \text{and} \quad \varepsilon'_r \cdot \left(f^{\boldsymbol{\varepsilon}'}(x_r(\boldsymbol{\varepsilon}')) - s_r(\boldsymbol{\varepsilon}')\right) \geq \frac{\alpha}{2}.$$

According to the definition of r we have $\varepsilon_r \neq \varepsilon'_r$. Again using eq. (E.24), we obtain that

$$|f^{\boldsymbol{\varepsilon}}(x_t(\boldsymbol{\varepsilon})) - f^{\boldsymbol{\varepsilon}'}(x_t(\boldsymbol{\varepsilon}))| \geq \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha,$$

which contradicts eq. (E.25). Therefore, we proved eq. (E.21).

Next, according to Theorem 6.19 with $\mathfrak{c}(a,b) = |a-b|$, and tree $\mathbf{x}$ being the depth-$\mathrm{fat}(\mathcal{F}, 3(\alpha+\beta))$ $\mathcal{X}$-valued tree shattered by $\mathcal{F}$, we obtain that

$$\mathcal{N}_\infty^{\mathsf{seq}}(\mathcal{F},\mathbf{x},\alpha+\beta) \leq \left(\frac{2e \cdot \mathrm{fat}(\mathcal{F}, 3(\alpha+\beta))}{\beta}\right)^{d^{\mathsf{seq}}(\mathcal{F},\alpha,\beta)}.$$

According to eq. (E.21) we further have

$$\mathcal{N}_\infty^{\mathsf{seq}}(\mathcal{F},\mathbf{x},\alpha+\beta) \geq 2^{\mathrm{fat}(\mathcal{F},3\alpha+3\beta)}.$$

Therefore, we conclude that

$$\log 2 \cdot \mathrm{fat}(\mathcal{F}, 3(\alpha+\beta)) \le d^{\mathsf{seq}}(\mathcal{F}, \alpha, \beta) \cdot \log\left(\frac{2e \cdot \mathrm{fat}(\mathcal{F}, 3(\alpha+\beta))}{\beta}\right),$$

which implies that

$$\mathrm{fat}(\mathcal{F}, 3(\alpha+\beta)) \le 2d^{\mathsf{seq}}(\mathcal{F}, \alpha, \beta) \cdot \log\left(\frac{6\mathrm{fat}(\mathcal{F}, 3(\alpha+\beta))}{\beta}\right). \tag{E.26}$$

Additionally, since $\log x \le \sqrt{x}$ holds for any $x > 0$,

$$\mathrm{fat}(\mathcal{F}, 3(\alpha+\beta)) \le 2d^{\mathsf{seq}}(\mathcal{F}, \alpha, \beta) \cdot \sqrt{\frac{6\mathrm{fat}(\mathcal{F}, 3(\alpha+\beta))}{\beta}},$$

which implies

$$\mathrm{fat}(\mathcal{F}, 3(\alpha+\beta)) \le \frac{24 d^{\mathsf{seq}}(\mathcal{F}, \alpha, \beta)^2}{\beta}$$

Bringing this back to eq. (E.26), we obtain that

$$\mathrm{fat}(\mathcal{F}, 3(\alpha+\beta)) \le 4d^{\mathsf{seq}}(\mathcal{F}, \alpha, \beta) \cdot \log\left(\frac{12 d^{\mathsf{seq}}(\mathcal{F}, \alpha, \beta)}{\beta}\right).$$

□

Proof of Theorem 6.23. We first show that for any $0 < 2\beta < \alpha < 1$, $d^{\mathsf{seq}}(\mathcal{F}, \alpha, \beta) = 1$. It is easy to see that the depth-1 $\mathcal{X}$-valued tree $\mathbf{x}$ with $x_1 = x$ is shattered by $\mathcal{F}$ at scale (α, β), according to Theorem 6.17, hence $d^{\mathsf{seq}}(\mathcal{F}, \alpha, \beta) \ge 1$. Next we show that $d^{\mathsf{seq}}(\mathcal{F}, \alpha, \beta) \le 1$. Suppose there is a depth-2 $\mathcal{X}$-valued tree $\mathbf{x}$ shattered by $\mathcal{F}$, then all nodes equal to x whatever depth and path. We let $\mathbf{s}$ to be the depth-2 $([-1,1] \times [-1,1])$-tree which is the witness of the shattering. Then for any $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2) \in \{-1,1\}^2$, there exists functions $f^{\boldsymbol{\varepsilon}} \in \mathcal{F}$ such that $|f^{\boldsymbol{\varepsilon}}(x) - s_1(\boldsymbol{\varepsilon})[\varepsilon_1]| \le \beta$ and $|f^{\boldsymbol{\varepsilon}}(x) - s_2(\boldsymbol{\varepsilon})[\varepsilon_2]| \le \beta$. Therefore, we have $|s_1(\boldsymbol{\varepsilon})[\varepsilon_1] - s_2(\boldsymbol{\varepsilon})[\varepsilon_2]| \le 2\beta$ for any $\boldsymbol{\varepsilon} \in \{-1,1\}^2$. We choose $\boldsymbol{\varepsilon} = (1,1)$ and $\boldsymbol{\varepsilon}' = (1,-1)$, then we have

$$s_1(\boldsymbol{\varepsilon})[\varepsilon_1] = s_1(\boldsymbol{\varepsilon}')[\varepsilon_1'] \quad \text{and} \quad |s_2(\boldsymbol{\varepsilon})[\varepsilon_2] - s_2(\boldsymbol{\varepsilon}')[\varepsilon_2']| \ge \alpha.$$

When $\alpha > 2\beta$, the above inequality cannot hold. Hence we have $d^{\mathsf{seq}}(\mathcal{F}, \alpha, \beta) = 1$.

Next, we show that $\mathrm{fat}(\mathcal{F}, \alpha) \ge \log(1/\alpha)$. We let $d = \lfloor \log_2(1/\alpha) \rfloor$, and for depth-$d$ path $\boldsymbol{\varepsilon} \in \{-1,1\}^d$, we define the shattered tree $\mathbf{x}$ such that $x_t(\boldsymbol{\varepsilon}) = x$ for any $\boldsymbol{\varepsilon} \in \{-1,1\}^d$ and $t \in [d]$, and the witness tree $\mathbf{v}$ as:

$$v_t(\boldsymbol{\varepsilon}) = \sum_{i=1}^{t-1} \varepsilon_i \cdot 2^{d-i} \cdot \alpha,$$

and we choose function $f^{\boldsymbol{\varepsilon}}$ as

$$f^{\boldsymbol{\varepsilon}}(x) = \sum_{i=1}^{d} \varepsilon_i \cdot 2^{d-i} \cdot \alpha.$$

Since $d = \lfloor \log_2(1/\alpha) \rfloor$, $\mathbf{v}$ is a depth-d $[-1,1]$-valued tree, and also $f \in \mathcal{F}$. For any $\boldsymbol{\varepsilon} \in \{-1,1\}^d$, we have

$$\varepsilon_t \cdot (f^{\boldsymbol{\varepsilon}}(x_t(\boldsymbol{\varepsilon})) - v_t(\boldsymbol{\varepsilon})) = \varepsilon_t \cdot \left(\sum_{i=t}^{d} \varepsilon_i \cdot 2^{d-i} \right) \cdot \alpha = \left(2^{d-t} - \sum_{i=t+1}^{d} \varepsilon_i \varepsilon_t \cdot 2^{d-i} \right) \cdot \alpha \geq \alpha > \frac{\alpha}{2}.$$

Hence tree $\mathbf{x}$ is shattered by $\mathcal{F}$, which implies that $\mathrm{fat}(\mathcal{F}, \alpha) \geq \lfloor \log_2(1/\alpha) \rfloor$. □

E.2.4 Missing Proofs in section 6.3.4

Proof of Theorem 6.24. For fixed positive integer n, we let

$$\alpha_n = \frac{1}{2} \cdot \arg\min_{\alpha > 0} \left\{ d^{\mathsf{seq}}(\mathcal{F}, \alpha, \alpha/20) \cdot \frac{16}{\alpha^2} \leq n \right\}.$$

Since $d^{\mathsf{seq}}(\mathcal{F}, \alpha, \alpha/20) = \tilde{\Omega}\left(\alpha^{-p}\right)$ for every $\alpha \geq 0$, we have

$$d^{\mathsf{seq}}(\mathcal{F}, \alpha_n, \alpha_n/20) \wedge \frac{n\alpha_n^2}{4} = \tilde{\Omega}\left(n^{\frac{p}{p+2}}\right), \qquad \forall n \in \mathbb{Z}_+.$$

In the following, we will prove that for any positive integer n, we have

$$\sup_{\boldsymbol{\mu}, \mathbf{x}} \mathbb{E}[\mathcal{R}_n(\mathcal{F}, \boldsymbol{\mu}, \mathbf{x}, \boldsymbol{\varepsilon})] = \Omega\left(d^{\mathsf{seq}}(\mathcal{F}, \alpha_n, \alpha_n/20)\right).$$

We fix n, and let $\alpha = \alpha_n$, $d = d^{\mathsf{seq}}(\mathcal{F}, \alpha_n, \alpha_n/20) \wedge (n\alpha_n^2/4)$. We let $\tilde{\mathbf{x}}$ be the depth-d $\mathcal{X}$-valued tree shattered by $\mathcal{F}$. Then according to Theorem 6.17, there exists a depth-d $[-1,1] \times [-1,1]$-valued tree $\mathbf{s}$ such that for any path $\tilde{\boldsymbol{\varepsilon}} = (\tilde{\varepsilon}_{1:d}) \in \{-1,1\}^d$, there exists $f^{\tilde{\boldsymbol{\varepsilon}}} \in \mathcal{F}$ such that

$$\left| f^{\tilde{\boldsymbol{\varepsilon}}}(t(\tilde{\boldsymbol{\varepsilon}})) - s_t(\tilde{\boldsymbol{\varepsilon}})[\tilde{\varepsilon}_t] \right| \leq \frac{\alpha}{20} \quad \text{and} \quad |s_t(\tilde{\boldsymbol{\varepsilon}})[1] - s_t(\tilde{\boldsymbol{\varepsilon}})[-1]| \geq \alpha \qquad \forall t \in [d], \tag{E.27}$$

Without loss of generality, we assume $s_t(\tilde{\boldsymbol{\varepsilon}})[1] \geq s_t(\tilde{\boldsymbol{\varepsilon}})[-1]$. We define depth-$d$ $[-1,1]$-valued $\tilde{\boldsymbol{\mu}}$ as

$$\tilde{\mu}_t(\tilde{\boldsymbol{\varepsilon}}) = \frac{s_t(\tilde{\boldsymbol{\varepsilon}})[-1] + s_t(\tilde{\boldsymbol{\varepsilon}})[1]}{2}, \quad \forall \boldsymbol{\varepsilon} \in \{-1,1\}^d.$$

In the following, we will construct trees $\mathbf{x}$ and $\boldsymbol{\mu}$ such that $\mathbb{E}[\mathcal{R}_n(\mathcal{F}, \boldsymbol{\mu}, \mathbf{x}, \boldsymbol{\varepsilon})] \geq d/50$. For a fixed path $\boldsymbol{\varepsilon} \in \{-1,1\}^n$, we first define an auxiliary path $\tilde{\boldsymbol{\varepsilon}} = (\tilde{\varepsilon}_{1:d}) \in \{-1,1\}^d$ of length d, as well as d integers $k_1, k_2, \ldots, k_d$ in the following way: calculate $\tilde{\varepsilon}_{1:d}$ and $k_{1:d}$ iteratively as

$$k_t = \left\lfloor \frac{1}{(s_t(\tilde{\boldsymbol{\varepsilon}})[1] - s_t(\tilde{\boldsymbol{\varepsilon}})[-1])^2} \right\rfloor \vee 1, \qquad \forall t \in [d],$$

and

$$\tilde{\varepsilon}_t = 2\mathbb{I}\left\{ \sum_{j=1}^{k_t} \varepsilon_{k_1 + \ldots + k_{t-1} + j} \geq 0 \right\} - 1, \qquad \forall t \in [d].$$

Notice that according to the above definition, k_t only depends on $\varepsilon_{1:k_1+\ldots+k_{t-1}}$, and $\tilde{\varepsilon}_t$ depends on $\varepsilon_{1:k_1+\ldots+k_t}$. Additionally, since $|s_t(\tilde{\boldsymbol{\varepsilon}})[1] - s_t(\tilde{\boldsymbol{\varepsilon}})[-1]| \geq \alpha$, we have

$$k_t \leq \frac{1}{\alpha^2} \vee 1 \leq \frac{4}{\alpha^2},$$

which implies $k_1 + \ldots + k_d \leq n$ always holds due to the definition of $\alpha = \alpha_n$. Hence $k_{1:d}$ and $\tilde{\varepsilon}_{1:d}$ are all well-defined. We pad the tree $\mathbf{x}$ with an arbitrary value x_0, resulting in the following block structure of $(x_1(\boldsymbol{\varepsilon}), x_2(\boldsymbol{\varepsilon}), \ldots, x_n(\boldsymbol{\varepsilon}))$:

$$(\underbrace{1(\tilde{\boldsymbol{\varepsilon}}), 1(\tilde{\boldsymbol{\varepsilon}}), \ldots, 1(\tilde{\boldsymbol{\varepsilon}})}_{k_1 \text{ terms}}, \underbrace{2(\tilde{\boldsymbol{\varepsilon}}), 2(\tilde{\boldsymbol{\varepsilon}}), \ldots, 2(\tilde{\boldsymbol{\varepsilon}})}_{k_2 \text{ terms}}, \ldots, \underbrace{d(\tilde{\boldsymbol{\varepsilon}}), d(\tilde{\boldsymbol{\varepsilon}}), \ldots, d(\tilde{\boldsymbol{\varepsilon}})}_{k_d \text{ terms}}, \underbrace{x_0, x_0, \ldots, x_0}_{(n-k_1-k_2-\ldots-k_d) \text{ terms}}),$$

Similarly, the values $(\mu_1(\boldsymbol{\varepsilon}), \mu_2(\boldsymbol{\varepsilon}), \ldots, \mu_n(\boldsymbol{\varepsilon}))$ are of the form

$$(\underbrace{\tilde{\mu}_1(\tilde{\boldsymbol{\varepsilon}}), \tilde{\mu}_1(\tilde{\boldsymbol{\varepsilon}}), \ldots, \tilde{\mu}_1(\tilde{\boldsymbol{\varepsilon}})}_{k_1 \text{ terms}}, \underbrace{\tilde{\mu}_2(\tilde{\boldsymbol{\varepsilon}}), \tilde{\mu}_2(\tilde{\boldsymbol{\varepsilon}}), \ldots, \tilde{\mu}_2(\tilde{\boldsymbol{\varepsilon}})}_{k_2 \text{ terms}}, \ldots, \underbrace{\tilde{\mu}_d(\tilde{\boldsymbol{\varepsilon}}), \tilde{\mu}_d(\tilde{\boldsymbol{\varepsilon}}), \ldots, \tilde{\mu}_d(\tilde{\boldsymbol{\varepsilon}})}_{k_d \text{ terms}}, \underbrace{f^{\tilde{\boldsymbol{\varepsilon}}}(x_0), f^{\tilde{\boldsymbol{\varepsilon}}}(x_0), \ldots, f^{\tilde{\boldsymbol{\varepsilon}}}(x_0)}_{(n-k_1-k_2-\ldots-k_d) \text{ terms}})$$

We note that the construction of trees $x_t(\boldsymbol{\varepsilon})$ and $\mu_t(\boldsymbol{\varepsilon})$ here differs slightly from the construction used for the non-sequential setting in the proof of Theorem 6.11. In the sequential case, μ is structured as a tree and can adapt based on the history of ε, allowing us to choose μ as $f^{\boldsymbol{\varepsilon}}(x_0)$. In contrast, in the non-sequential case, the choice of μ must be independent of the history, so we are required to select global (i.e., fixed) values for μ. For this reason, Theorem 6.24 only needs $d(\mathcal{F}, \alpha, \alpha/20) = \tilde{\Omega}(\alpha^{-p})$, while Theorem 6.11, with the current proof, requires that $d(\mathcal{F}, \alpha, \alpha/(20nC)) = \tilde{\Omega}(\alpha^{-p})$.

Next we write $\mathcal{R}_n(\mathcal{F}, \boldsymbol{\mu}, \mathbf{x}, \boldsymbol{\varepsilon})$ in terms of segments 1 to d. Noticing that $f^{\tilde{\boldsymbol{\varepsilon}}} \in \mathcal{F}$ for any depth-d path $\tilde{\boldsymbol{\varepsilon}} \in \{-1, 1\}^d$, if we use $[t, j]$ to denote $k_1 + \ldots + k_t + j$, we obtain

$$\begin{aligned}
&\mathcal{R}_n(\mathcal{F}, \boldsymbol{\mu}, \mathbf{x}, \varepsilon) \\
&\quad = \sup_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{d} \sum_{j=1}^{k_t} C \cdot \varepsilon_{[t-1,j]} (f(x_{[t-1,j]}(\boldsymbol{\varepsilon})) - \mu_{[t-1,j]}(\boldsymbol{\varepsilon})) - (f(x_{[t-1,j]}(\boldsymbol{\varepsilon})) - \mu_{[t-1,j]}(\boldsymbol{\varepsilon}))^2 \right. \\
&\qquad \left. + \sum_{j=1}^{n-k_1-\ldots-k_d} C \cdot \varepsilon_{[d,j]} (f(x_{[t-1,j]}(\boldsymbol{\varepsilon})) - \mu_{[d,j]}(\boldsymbol{\varepsilon})) - (f(x_{[d,j]}(\boldsymbol{\varepsilon})) - \mu_{[d,j]}(\boldsymbol{\varepsilon}))^2 \right\} \\
&\quad \geq \sum_{t=1}^{d} \sum_{j=1}^{k_t} C \cdot \varepsilon_{[t-1,j]} (f^{\tilde{\boldsymbol{\varepsilon}}}(t(\boldsymbol{\varepsilon})) - \tilde{\mu}_t(\tilde{\boldsymbol{\varepsilon}})) - (f^{\tilde{\boldsymbol{\varepsilon}}}(t(\boldsymbol{\varepsilon})) - \tilde{\mu}_t(\tilde{\boldsymbol{\varepsilon}}))^2,
\end{aligned}$$

where in the last step we chose $f = f^{\tilde{\boldsymbol{\varepsilon}}} \in \mathcal{F}$. Next, for fixed $\boldsymbol{\varepsilon}$, we define the random variable

$$\widehat{\varepsilon}_t = \frac{1}{k_t} \sum_{j=1}^{k_t} \varepsilon_{[t-1,j]}.$$

Since $t(\tilde{\boldsymbol{\varepsilon}})$ and $\mu_t(\tilde{\boldsymbol{\varepsilon}})$ are independent of $\varepsilon_{[t-1,1]}, \ldots, \varepsilon_{[t-1,k_t]}$, we can write

$$\mathcal{R}_n(\mathcal{F}, \boldsymbol{\mu}, \mathbf{x}, \boldsymbol{\varepsilon}) \geq \sum_{t=1}^{d} k_t \cdot \left(C \cdot \widehat{\varepsilon}_t (f^{\tilde{\boldsymbol{\varepsilon}}}(t(\boldsymbol{\varepsilon})) - \tilde{\mu}_t(\tilde{\boldsymbol{\varepsilon}})) - (f^{\tilde{\boldsymbol{\varepsilon}}}(t(\boldsymbol{\varepsilon})) - \tilde{\mu}_t(\tilde{\boldsymbol{\varepsilon}}))^2 \right). \tag{E.28}$$

According to eq. (E.27), we have for any $\tilde{\boldsymbol{\varepsilon}}$,

$$|f^{\tilde{\boldsymbol{\varepsilon}}}(t(\tilde{\boldsymbol{\varepsilon}})) - s_t(\tilde{\boldsymbol{\varepsilon}})[\tilde{\varepsilon}_t]| \le \frac{\alpha}{20} \quad \text{and} \quad s_t(\tilde{\boldsymbol{\varepsilon}})[\tilde{\varepsilon}_t] - \mu_t(\tilde{\boldsymbol{\varepsilon}}) = \operatorname{sign}(\widehat{\varepsilon}_t) \cdot \frac{s_t(\tilde{\boldsymbol{\varepsilon}})[1] - s_t(\tilde{\boldsymbol{\varepsilon}})[-1]}{2}.$$

eq. (E.27) also gives that $s_t(\tilde{\boldsymbol{\varepsilon}})[1] - s_t(\tilde{\boldsymbol{\varepsilon}})[-1] \ge \alpha$, which implies

$$\frac{9}{10} \cdot \frac{s_t(\tilde{\boldsymbol{\varepsilon}})[1] - s_t(\tilde{\boldsymbol{\varepsilon}})[-1]}{2} \le \operatorname{sign}(\widehat{\varepsilon}_t)(f^{\tilde{\boldsymbol{\varepsilon}}}(t(\tilde{\boldsymbol{\varepsilon}})) - \tilde{\mu}_t(\tilde{\boldsymbol{\varepsilon}})) \le \frac{11}{10} \cdot \frac{s_t(\tilde{\boldsymbol{\varepsilon}})[1] - s_t(\tilde{\boldsymbol{\varepsilon}})[-1]}{2}.$$

Therefore we can lower bound eq. (E.28) by

$$\mathcal{R}_n(\mathcal{F}, \boldsymbol{\mu}, \mathbf{x}, \boldsymbol{\varepsilon}) \ge \sum_{t=1}^{d} k_t \cdot \left(C|\widehat{\varepsilon}_t| \cdot \frac{9}{10} \cdot \frac{s_t(\tilde{\boldsymbol{\varepsilon}})[1] - s_t(\tilde{\boldsymbol{\varepsilon}})[-1]}{2} - \frac{121}{100} \cdot \left(\frac{s_t(\tilde{\boldsymbol{\varepsilon}})[1] - s_t(\tilde{\boldsymbol{\varepsilon}})[-1]}{2} \right)^2 \right).$$

We notice that in the t-th summand in the right hand side, the only term which depend on $\varepsilon_{[t-1,1]:[t-1,k_t]}$ is $|\widehat{\varepsilon}_t|$, and all other terms only depends on $\varepsilon_{1:[t-1:0]}$. Hence we can lower bound $\mathbb{E}\left[\mathcal{R}_n(\mathcal{F}, \boldsymbol{\mu}, \mathbf{x}, \boldsymbol{\varepsilon})\right]$ by

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{R}_n(\mathcal{F}, \boldsymbol{\mu}, \mathbf{x}, \boldsymbol{\varepsilon})\right] \\
&\ge \sum_{t=1}^{d} \mathbb{E}\left[k_t \cdot \left(C|\widehat{\varepsilon}_t| \cdot \frac{9}{10} \cdot \frac{s_t(\tilde{\boldsymbol{\varepsilon}})[1] - s_t(\tilde{\boldsymbol{\varepsilon}})[-1]}{2} - \frac{121}{100} \cdot \left(\frac{s_t(\tilde{\boldsymbol{\varepsilon}})[1] - s_t(\tilde{\boldsymbol{\varepsilon}})[-1]}{2} \right)^2 \right) \right] \\
&= \sum_{t=1}^{d} \mathbb{E}\left[k_t \cdot \left(C\mathbb{E}\left[|\widehat{\varepsilon}_t| \mid t_{1:[t-1,0]}\right] \cdot \frac{9}{10} \cdot \frac{s_t(\tilde{\boldsymbol{\varepsilon}})[1] - s_t(\tilde{\boldsymbol{\varepsilon}})[-1]}{2} - \frac{121}{100} \cdot \left(\frac{s_t(\tilde{\boldsymbol{\varepsilon}})[1] - s_t(\tilde{\boldsymbol{\varepsilon}})[-1]}{2} \right)^2 \right) \right] \\
&\ge \sum_{t=1}^{d} \mathbb{E}\left[k_t \cdot \left(C\sqrt{\frac{1}{2k_t}} \cdot \frac{9}{10} \cdot \frac{s_t(\tilde{\boldsymbol{\varepsilon}})[1] - s_t(\tilde{\boldsymbol{\varepsilon}})[-1]}{2} - \frac{121}{100} \cdot \left(\frac{s_t(\tilde{\boldsymbol{\varepsilon}})[1] - s_t(\tilde{\boldsymbol{\varepsilon}})[-1]}{2} \right)^2 \right) \right]
\end{aligned} \tag{E.29}$$

where the last inequality uses the Khintchine inequality [222]. Next notice that according to our choice of k_t,

$$k_t = \left\lfloor \frac{1}{(s_t(\tilde{\boldsymbol{\varepsilon}})[1] - s_t(\tilde{\boldsymbol{\varepsilon}})[-1])^2} \right\rfloor \vee 1 \in \left[\frac{1}{2(s_t(\tilde{\boldsymbol{\varepsilon}})[1] - s_t(\tilde{\boldsymbol{\varepsilon}})[-1])^2}, \frac{4}{(s_t(\tilde{\boldsymbol{\varepsilon}})[1] - s_t(\tilde{\boldsymbol{\varepsilon}})[-1])^2} \right],$$

which implies that $1/\sqrt{2} \le \sqrt{k_t}(s_t(\tilde{\boldsymbol{\varepsilon}})[1] - s_t(\tilde{\boldsymbol{\varepsilon}})[-1]) \le 2$. Therefore, since $C \ge 2$, we have

$$\begin{aligned}
\text{RHS of } eq.\ (\text{E.29}) &\ge \sum_{t=1}^{d} \mathbb{E}\left[k_t \cdot \left(C\sqrt{\frac{1}{2k_t}} \cdot \frac{9}{10} \cdot \frac{s_t(\tilde{\boldsymbol{\varepsilon}})[1] - s_t(\tilde{\boldsymbol{\varepsilon}})[-1]}{2} - \frac{121}{100} \cdot \left(\frac{s_t(\tilde{\boldsymbol{\varepsilon}})[1] - s_t(\tilde{\boldsymbol{\varepsilon}})[-1]}{2} \right)^2 \right) \right] \\
&\ge \sum_{t=1}^{d} \mathbb{E}\left[\left(\frac{9C}{20\sqrt{2}} - \frac{121}{200} \right) \cdot \sqrt{k_t} \cdot |s_t(\tilde{\boldsymbol{\varepsilon}})[1] - s_t(\tilde{\boldsymbol{\varepsilon}})[-1]| \right] \ge \frac{d}{50}.
\end{aligned}$$

Therefore, we obtain that

$$\sup_{\boldsymbol{\mu}, \mathbf{x}} \mathbb{E}[\mathcal{R}_n(\mathcal{F}, \boldsymbol{\mu}, \mathbf{x}, \boldsymbol{\varepsilon})] \ge \frac{d}{50} = \tilde{\Omega}\left(n^{\frac{p}{p+2}} \right).$$

□

Proof of Theorem 6.25. Since $\sup_{\mathbf{x}} \log \mathcal{N}_{\infty}^{\mathsf{seq}}(\mathcal{F}, \mathbf{x}, \alpha) = \tilde{\Omega}\left(\alpha^{-p}\right)$, according to Theorem 6.19 we have

$$d^{\mathsf{seq}}(\mathcal{F}, \alpha, \alpha/20) = \tilde{\Omega}\left(\alpha^{-p}\right).$$

Next, we call Theorem 6.24 with $C = 2$. According to [109, Lemma 4], we obtain

$$\mathcal{V}_n^{\mathsf{seq}}(\mathcal{F}) = \tilde{\Omega}\left(n^{\frac{p}{p+2}}\right).$$

□

Appendix F

Proofs for Chapter 7

F.1 Missing Proofs in section 7.3

F.1.1 Proof of Change of Trajectory Measure Lemma

Proof of Theorem 7.3. In the following proof, for any trajectory $\tau = (s_1, a_1, s_2, a_2, \cdots, s_H, a_H)$ and $1 \le u \le v \le H$, we use $\tau_{u:v}$ to denote the partial trajectory $(s_u, a_u, s_{u+1}, a_{u+1}, \cdots, s_v, a_v)$. We use $f(\tau_{u:v}) \coloneqq \sum_{h=u}^{v} f(s_h, a_h)$ to denote the cumulative reward in this segment. Without loss of generality we assume for any trajectory τ, $f(\tau) \in [-1, 1]$. Since the state space of the MDP is layered, we write $\{s_h \in \tau\}$ for $s_h \in \mathcal{S}_h$ to denote the event that s_h is the time-h element of the trajectory τ. We let

$$L(\eta) = \mathbb{P}_{\tau \sim \pi_{\text{off}}} \left[|f(\tau)| \ge \eta \right]$$

for every $\eta > 0$. For every real number $r \in \mathbb{R}$ and $\eta, p > 0$, we define the following sets:

$$\mathcal{S}_h^{\uparrow}(r, \eta, p) \coloneqq \left\{ s_h \in \mathcal{S}_h : \mathbb{P}_{\tau \sim \pi_{\text{off}}} \left(|r - f(\tau_{h:H})| \le \eta \;\middle|\; s_h \in \tau \right) \ge 1 - p \right\},$$

$$\text{and} \quad \mathcal{S}_h^{\downarrow}(r, \eta, p) \coloneqq \left\{ s_h \in \mathcal{S}_h : \mathbb{P}_{\tau \sim \pi_{\text{off}}} \left(|r - f(\tau_{1:h-1})| \le \eta \;\middle|\; s_h \in \tau \right) \ge 1 - p \right\}.$$

Intuitively, $\mathcal{S}_h^{\downarrow}(r, \eta, p)$ denotes subset of states in $\mathcal{S}_h$ where, conditionally on arriving at that state under policy π_{off}, with probability at least $1 - p$, the total reward collected in the first $(h-1)$ steps is η-close to r. Similarly, $\mathcal{S}_h^{\uparrow}(r, \eta, p)$ denotes subset of states in $\mathcal{S}_h$ where, conditionally on arriving at that state under policy π_{off}, the probability that the total reward collected in the last $(H - h + 1)$ steps is η-close to r is at least $1 - p$.

We further define sets

$$\mathcal{S}_h^{\uparrow}(\eta, p) = \cup_{r \in \mathbb{R}} \mathcal{S}_h^{\uparrow}(r, \eta, p), \quad \mathcal{S}_h^{\downarrow}(\eta, p) = \cup_{r \in \mathbb{R}} \mathcal{S}_h^{\downarrow}(r, \eta, p).$$

We can now upper bound the occupancy measure of those states outside the set $\mathcal{S}_h^{\uparrow}(\eta, p)$ or $\mathcal{S}_h^{\downarrow}(\eta, p)$. This property is summarized in the following claim:

Claim 1 We have the following upper bounds on the occupancy measure outside $\mathcal{S}_h^{\uparrow}(\eta, p)$ and $\mathcal{S}_h^{\downarrow}(\eta, p)$:

$$\mathbb{P}_{\tau \sim \pi_{\text{off}}}(\tau \cap \mathcal{S}_h \not\subset \mathcal{S}_h^{\uparrow}(\eta, p)) \le \frac{L(\eta)}{p} \quad \text{and} \quad \mathbb{P}_{\tau \sim \pi_{\text{off}}}(\tau \cap \mathcal{S}_h \not\subset \mathcal{S}_h^{\downarrow}(\eta, p)) \le \frac{L(\eta)}{p}. \tag{F.1}$$

Proof We only prove the first inequality, as the proof of the second inequality is similar. Notice that for any state $s_h \in \mathcal{S}_h \backslash \mathcal{S}_h^{\uparrow}(\eta, p)$, according to the definition of $\mathcal{S}_h^{\uparrow}(\eta, p)$ we have for any $r \in \mathbb{R}$,

$$\mathbb{P}_{\tau \sim \pi_{\text{off}}}\left(|r - f(\tau_{h:H})| \le \eta \;\middle|\; s_h \in \tau\right) < 1 - p.$$

According to the Markov property, when sampling $\tau \sim \pi_{\text{off}}$, $\tau_{1:h-1} \perp\!\!\!\perp \tau_{h:H}$ conditioned on $s_h \in \tau$, which implies

$$\begin{aligned}
&\mathbb{P}_{\tau \sim \pi_{\text{off}}}\left(|f(\tau)| \le \eta \;\middle|\; s_h \in \tau\right) \\
&\quad = \mathbb{E}_{\tau \sim \pi_{\text{off}}}\left[\mathbb{P}_{\tau \sim \pi_{\text{off}}}\left[|f(\tau_{h:H}) - (-f(\tau_{1:h-1}))| \le \eta \;\middle|\; \tau_{1:h-1}, s_h \in \tau\right] \;\middle|\; s_h \in \tau\right] \le 1 - p.
\end{aligned}$$

Hence we have

$$L(\eta) = \mathbb{P}_{\tau \sim \pi_{\text{off}}}\left[|f(\tau)| \ge \eta\right] \ge \mathbb{P}_{\tau \sim \pi_{\text{off}}}(\tau \cap \mathcal{S}_h \not\subset \mathcal{S}_h^{\uparrow}(\eta, p)) \cdot p,$$

which implies the first inequality of eq. (F.1). The second inequality of eq. (F.1) follows similarly. □

Next, for every real number $r \in \mathbb{R}$ and $\eta, p > 0$, we define the set

$$\mathcal{S}_h(r, \eta, p) \coloneqq \left\{s_h \in \mathcal{S}_h : \mathbb{P}_{\tau \sim \pi_{\text{off}}}\left(|r - f(\tau_{1:h-1})| \le \eta \text{ and } |r + f(\tau_{h:H})| \le \eta \;\middle|\; s_h \in \tau\right) \ge 1 - p\right\}.$$

This set denotes subset of states in $\mathcal{S}_h$ where conditioned on arriving at that state under policy π_{off}, the probability that the first $(h-1)$ steps' total reward is η-close to r and also the last $(H - h + 1)$ steps' total reward is η-close to $-r$ is less than p. We further define the set

$$\mathcal{S}_h(\eta, p) = \cup_{r \in \mathbb{R}} \mathcal{S}_h(r, \eta, p),$$

Then we have the following claim, which shows that the occupancy measure of states outside $\mathcal{S}_h(\eta, p)$ is also upper bounded:

Claim 2 We have the following upper bounds on the occupancy measure outside $\mathcal{S}_h(\eta, p)$:

$$\mathbb{P}_{\tau \sim \pi_{\text{off}}}(\tau \cap \mathcal{S}_h \notin \mathcal{S}_h(\eta, p)) \le \frac{L(2\eta/3)}{1 - p} + \frac{4L(\eta/3)}{p}. \tag{F.2}$$

Proof For state $s_h \in \mathcal{S}_h \backslash \mathcal{S}_h(\eta, p)$ but $s_h \in \mathcal{S}_h^{\uparrow}(\eta/3, p/2) \cap \mathcal{S}_h^{\downarrow}(\eta/3, p/2)$, there exists some $r^{\uparrow}(s_h) \in \mathbb{R}$ and $r^{\downarrow}(s_h) \in \mathbb{R}$ such that

$$\mathbb{P}_{\tau \sim \pi_{\text{off}}}\left(\left|f(s_h)^{\downarrow} - f(\tau_{1:h-1})\right| \le \frac{\eta}{3} \;\middle|\; s_h \in \tau\right) \ge 1 - \frac{p}{2} \text{ and } \mathbb{P}_{\tau \sim \pi_{\text{off}}}\left(\left|f(s_h)^{\uparrow} - f(\tau_{h:H})\right| \le \frac{\eta}{3} \;\middle|\; s_h \in \tau\right) \ge 1 -$$

By union bound we have that for any such s_h,

$$\mathbb{P}_{\tau \sim \pi_{\text{off}}}\left(\left|r^{\downarrow}(s_h) - f(\tau_{1:h-1})\right| \le \frac{\eta}{3} \quad \text{and} \quad \left|r^{\uparrow}(s_h) - f(\tau_{h:H})\right| \le \frac{\eta}{3} \;\middle|\; s_h \in \tau\right) \ge 1 - p.$$

If for some $s_h \in \mathcal{S}_h \backslash \mathcal{S}_h(\eta, p)$, we have $|r^{\downarrow}(s_h) + r^{\uparrow}(s_h)| \leq \frac{4\eta}{3}$, then by letting $r = \frac{r^{\downarrow}(s_h) - r^{\uparrow}(s_h)}{2}$, we obtain

$$\begin{aligned}
&\mathbb{P}_{\tau \sim \pi_{\text{off}}} \left(|r - f(\tau_{1:h-1})| \leq \eta \text{ and } |r + f(\tau_{h:H})| \leq \eta \;\middle|\; s_h \in \tau \right) \\
&\quad \geq \mathbb{P}_{\tau \sim \pi_{\text{off}}} \left(\left|r^{\downarrow}(s_h) - f(\tau_{1:h-1})\right| \leq \frac{\eta}{3} \text{ and } \left|r^{\uparrow}(s_h) - f(\tau_{h:H})\right| \leq \frac{\eta}{3} \;\middle|\; s_h \in \tau \right) \\
&\quad \geq 1 - p.
\end{aligned}$$

This contradicts the definition of $\mathcal{S}_h \backslash \mathcal{S}_h(\eta, p)$. Hence for any $s_h \in (\mathcal{S}_h \backslash \mathcal{S}_h(\eta, p)) \cap (\mathcal{S}_h^{\uparrow}(\eta/3, p/2) \cap \mathcal{S}_h^{\downarrow}(\eta/3, p/2))$, we always have $|r^{\downarrow}(s_h) + r^{\uparrow}(s_h)| > \frac{4\eta}{3}$, which implies that

$$\begin{aligned}
\mathbb{P}_{\tau \sim \pi_{\text{off}}} \left(|f(\tau)| \geq \frac{2\eta}{3} \;\middle|\; s_h \in \tau \right) &\geq \mathbb{P}_{\tau \sim \pi_{\text{off}}} \left(\left|r^{\downarrow}(s_h) - f(\tau_{1:h-1})\right| \leq \frac{\eta}{3} \text{ and } \left|r^{\uparrow}(s_h) - f(\tau_{h:H})\right| \leq \frac{\eta}{3} \;\middle|\; s_h \in \tau \right) \\
&\geq 1 - p.
\end{aligned}$$

Further notice that

$$\begin{aligned}
L(2\eta/3) &= \mathbb{P}_{\tau \sim \pi_{\text{off}}} (|f(\tau)| \geq 2\eta/3) \\
&\geq \mathbb{P}_{\tau \sim \pi_{\text{off}}} (\tau \cap \mathcal{S}_h \subset (\mathcal{S}_h \backslash \mathcal{S}_h(\eta, p)) \cap (\mathcal{S}_h^{\uparrow}(\eta/3, p/2) \cap \mathcal{S}_h^{\downarrow}(\eta/3, p/2))) \cdot (1 - p).
\end{aligned}$$

Hence we obtain

$$\mathbb{P}_{\tau \sim \pi_{\text{off}}} (\tau \cap \mathcal{S}_h \subset (\mathcal{S}_h \backslash \mathcal{S}_h(\eta, p)) \cap (\mathcal{S}_h^{\uparrow}(\eta/3, p/2) \cap \mathcal{S}_h^{\downarrow}(\eta/3, p/2))) \leq \frac{L(2\eta/3)}{1 - p}.$$

Additionally, according to our previous claim, we have

$$\mathbb{P}_{\tau \sim \pi_{\text{off}}} (\tau \cap \mathcal{S}_h \subset \mathcal{S}_h^{\uparrow}(\eta/3, p/2)) \leq \frac{2L(\eta/3)}{p} \quad \text{and} \quad \mathbb{P}_{\tau \sim \pi_{\text{off}}} (\tau \cap \mathcal{S}_h \subset \mathcal{S}_h^{\downarrow}(\eta/3, p/2)) \leq \frac{2L(\eta/3)}{p}.$$

Combining the above three inequalities, we obtain that

$$\mathbb{P}_{\tau \sim \pi_{\text{off}}} (\tau \cap \mathcal{S}_h \not\subset \mathcal{S}_h(\eta, p)) \leq \frac{L(2\eta/3)}{1 - p} + \frac{4L(\eta/3)}{p},$$

and eq. (F.2) is verified. □

We next define the following sets of "good" state-action-state tuples:

$$\begin{aligned}
\mathcal{U}_h \coloneqq \{&(s_h, a_h, s_{h+1}) : s_h \in \mathcal{S}_h, a_h \in \mathcal{A}_h, s_{h+1} \in \mathcal{S}_{h+1}, \\
&\exists r, r' \in \mathbb{R} \text{ such that } s_h \in \mathcal{S}_h(r, \eta, p), s_{h+1} \in \mathcal{S}_{h+1}(r', \eta, p) \text{ and } |f(s_h, a_h) + r - r'| \leq 3\eta\}
\end{aligned}$$

and also 'bad' state-action-state tuples:

$$\begin{aligned}
\mathcal{U}_h' \coloneqq \{&(s_h, a_h, s_{h+1}) : s_h \in \mathcal{S}_h, a_h \in \mathcal{A}_h, s_{h+1} \in \mathcal{S}_{h+1}, \\
&\forall r, r' \in \mathbb{R} \text{ such that } s_h \in \mathcal{S}_h(r, \eta, p), s_{h+1} \in \mathcal{S}_{h+1}(r', \eta, p) \text{ and } |f(s_h, a_h) + r - r'| \geq 3\eta\}.
\end{aligned}$$

We will show that under policy π_{off}, the encountered state-action-state pairs are 'good', i.e., belong to $\mathcal{U}_h$, with high probability. Notice that the complement of set $\mathcal{U}_h$ satisfies

$$(\mathcal{U}_h)^c \subset \mathcal{U}_h' \cup \{(s_h, a_h, s_{h+1}) : s_h \in \mathcal{S}_h \backslash \mathcal{S}_h(\eta, p)\} \cup \{(s_h, a_h, s_{h+1}) : s_{h+1} \in \mathcal{S}_{h+1} \backslash \mathcal{S}_{h+1}(\eta, p)\}, \tag{F.3}$$

and according to the last claim we have

$$\mathbb{P}_{\tau\sim\pi_{\text{off}}}(\tau\cap\mathcal{S}_h\not\subset\mathcal{S}_h(\eta,p))\le\frac{L(2\eta/3)}{1-p}+\frac{4L(\eta/3)}{p}$$
$$\text{and}\quad\mathbb{P}_{\tau\sim\pi_{\text{off}}}(\tau\cap\mathcal{S}_{h+1}\not\subset\mathcal{S}_{h+1}(\eta,p))\le\frac{L(2\eta/3)}{1-p}+\frac{4L(\eta/3)}{p}. \tag{F.4}$$

We next upper bound $\mathbb{P}_{\tau\sim\pi_{\text{off}}}(\tau\cap\mathcal{U}_h'\neq\emptyset)$, which is summarized into the following claim.

Claim 3 We have the following upper bounds on the occupancy measure of $\mathcal{U}_h'$:

$$\mathbb{P}_{\tau\sim\pi_{\text{off}}}(\tau\cap\mathcal{U}_h'\neq\emptyset)\le\frac{L(\eta)}{1-2p} \tag{F.5}$$

Proof According to the Markov property, we have

$$(s_1,a_1,\cdots,s_{h-1},a_{h-1})\perp\!\!\!\perp(a_h,s_{h+1})\quad\text{conditioned on } s_h$$
$$\text{and}\quad(s_{h+1},a_{h+1},\cdots,s_H,a_H)\perp\!\!\!\perp(s_h,a_h)\quad\text{conditioned on } s_{h+1}.$$

Hence for any $(s_h,a_h,s_{h+1})\in\mathcal{U}_h'$, there exists r and r' such that $s_h\in\mathcal{S}_h(r,\eta,p)$ and $s_{h+1}\in\mathcal{S}_{h+1}(r',\eta,p)$, which implies

$$\mathbb{P}_{\tau\sim\pi_{\text{off}}}\left(|f(\tau_{1:h-1})-r|\le\eta\;\middle|\;(s_h,a_h,s_{h+1})\in\tau\right)=\mathbb{P}_{\tau\sim\pi_{\text{off}}}\left(|f(\tau_{1:h-1})-r|\le\eta\;\middle|\;s_h\in\tau\right)\ge1-p$$

and

$$\mathbb{P}_{\tau\sim\pi_{\text{off}}}\left(|f(\tau_{h:H})+r'|\le\eta\;\middle|\;(s_h,a_h,s_{h+1})\in\tau\right)=\mathbb{P}_{\tau\sim\pi_{\text{off}}}\left(|f(\tau_{h:H})+r'|\le\eta\;\middle|\;s_{h+1}\in\tau\right)\ge1-p.$$

By union bound, we obtain

$$\mathbb{P}_{\tau\sim\pi_{\text{off}}}\left(|f(\tau_{1:h-1})-r|\le\eta\text{ and }|f(\tau_{h:H})+r'|\le\eta\;\middle|\;(s_h,a_h,s_{h+1})\in\tau\right)\ge1-2p.$$

Additionally $(s_h,a_h,s_{h+1})\in\mathcal{U}_h'$ implies $|f(s_h,a_h)+r-r'|\ge3\eta$. Therefore, for any $(s_h,a_h,s_{h+1})\in\mathcal{U}_h'$,

$$\mathbb{P}_{\tau\sim\pi_{\text{off}}}\left(|f(\tau)|\ge\eta\;\middle|\;(s_h,a_h,s_{h+1})\in\tau\right)\ge1-2p.$$

This leads to

$$L(\eta)=\mathbb{P}_{\tau\sim\pi_{\text{off}}}\left(|f(\tau)|\ge\eta\right)\ge\mathbb{P}_{\tau\sim\pi_{\text{off}}}\left(\tau\cap\mathcal{U}_h'\neq\emptyset\right)\cdot(1-2p),$$

and eq. (F.5) follows. □

Combining eq. (F.5) and the two inequalities in eq. (F.4), and in view of eq. (F.3), we obtain

$$\mathbb{P}_{\tau\sim\pi_{\text{off}}}(\tau\cap\mathcal{U}_h=\emptyset)\le\frac{2L(2\eta/3)}{1-p}+\frac{8L(\eta/3)}{p}+\frac{L(\eta)}{1-2p}.$$

Next notice from the definition of state-action concentrability, we have for any policy π and layer $h \in [H]$,

$$\begin{aligned}
&\sup_{s_h \in \mathcal{S}_h, a_h \in \mathcal{A}, s_{h+1} \in \mathcal{S}_{h+1}} \frac{d^{\pi}(s_h, a_h, s_{h+1})}{d^{\pi_{\text{off}}}(s_h, a_h, s_{h+1})} \\
&= \sup_{s_h \in \mathcal{S}_h, a_h \in \mathcal{A}, s_{h+1} \in \mathcal{S}_{h+1}} \frac{d^{\pi}(s_h, a_h) \cdot T(s_{h+1} \mid s_h, a_h)}{d^{\pi_{\text{off}}}(s_h, a_h) \cdot T(s_{h+1} \mid s_h, a_h)} \\
&= \sup_{s_h \in \mathcal{S}_h, a_h \in \mathcal{A}} \frac{d^{\pi}(s, a)}{d^{\pi_{\text{off}}}(s, a)} \leq C_{\mathsf{sa}}(\pi, \pi_{\text{off}}),
\end{aligned}$$

which implies that for any policy π and layer $h \in [H]$,

$$\begin{aligned}
\mathbb{P}_{\tau \sim \pi}(\tau \cap \mathcal{U}_h = \emptyset) &= \sum_{(s_h, a_h, s_{h+1}) \in (\mathcal{S}_h \times \mathcal{A} \times \mathcal{S}_{h+1}) \setminus \mathcal{U}_h} d^{\pi}(s_h, a_h, s_{h+1}) \\
&\leq C_{\mathsf{sa}}(\pi, \pi_{\text{off}}) \cdot \sum_{(s_h, a_h, s_{h+1}) \in (\mathcal{S}_h \times \mathcal{A} \times \mathcal{S}_{h+1}) \setminus \mathcal{U}_h} d^{\pi_{\text{off}}}(s_h, a_h, s_{h+1}) \\
&= C_{\mathsf{sa}}(\pi, \pi_{\text{off}}) \cdot \mathbb{P}_{\tau \sim \pi_{\text{off}}}(\tau \cap \mathcal{U}_h = \emptyset) \\
&\leq C_{\mathsf{sa}}(\pi, \pi_{\text{off}}) \cdot \left(\frac{2L(2\eta/3)}{1-p} + \frac{8L(\eta/3)}{p} + \frac{L(\eta)}{1-2p} \right).
\end{aligned}$$

Hence by union bound, we have for any policy π,

$$\mathbb{P}_{\tau \sim \pi}(\tau \cap \mathcal{U}_h \neq \emptyset, \ \forall h \in [H]) \geq 1 - H C_{\mathsf{sa}}(\pi, \pi_{\text{off}}) \cdot \left(\frac{2L(2\eta/3)}{1-p} + \frac{8L(\eta/3)}{p} + \frac{L(\eta)}{1-2p} \right).$$

Finally, we have the last claim showing that if for all $h \in [H]$, $(s_h, a_h, s_{h+1}) \in \mathcal{U}_h$, then the total reward of the trajectory can be upper bounded.

Claim 4 For any trajectory $\tau = (s_1, a_1, \cdots, s_H, a_H)$ where $(s_h, a_h, s_{h+1}) \in \mathcal{U}_h$ for every h, we have

$$|f(\tau)| \leq 5H\eta.$$

Proof We define tuple $u_h = (s_h, a_h, s_{h+1})$ for $h \in [H]$. According to the definition of $\mathcal{U}_h$, there exist real numbers $r(u_h) \in \mathbb{R}$ and $r'(u_{h+1}) \in \mathbb{R}$ such that for any $h \in [H]$,

$$s_h \in \mathcal{S}_h(r(u_h), \eta, p), \quad s_{h+1} \in \mathcal{S}_h(r'(u_h), \eta, p), \quad \text{and} \quad |f(s_h, a_h) + r(u_h) - r'(u_h)| \leq 3\eta.$$

Compare the condition on $s_h \in \mathcal{S}_h(r(u_h), \eta, p)$ with $s_h \in \mathcal{S}_h(r'(u_{h-1}), \eta, p)$, we have

$$\mathbb{P}_{\tau \sim \pi_{\text{off}}}\left(|r(u_h) - f(\tau_{1:h-1})| \leq \eta \;\middle|\; s_h \in \tau \right) \geq 1-p \ \text{ and } \ \mathbb{P}_{\tau \sim \pi_{\text{off}}}\left(|r'(u_{h-1}) - f(\tau_{1:h-1})| \leq \eta \;\middle|\; s_h \in \tau \right) \geq 1-p.$$

When $p < 1/2$, by union bound we obtain

$$\mathbb{P}_{\tau \sim \pi_{\text{off}}}\left(|r(u_h) - f(\tau_{1:h-1})| \leq \eta \text{ and } |r'(u_{h-1}) - f(\tau_{1:h-1})| \leq \eta \;\middle|\; s_h \in \tau \right) \geq 1 - 2p > 0,$$

which implies that there exists a trajectory τ such that $|r(u_h) - f(\tau_{1:h-1})| \le \eta$ and $|r'(u_{h-1}) - f(\tau_{1:h-1})| \le \eta$ both hold. Hence we obtain

$$|r(u_h) - r'(u_{h-1})| \le |r(u_h) - f(\tau_{1:h-1})| + |f(\tau_{1:h-1}) - r'(u_{h-1})| \le \eta + \eta = 2\eta.$$

Therefore, we obtain that

$$|f(\tau)| = \left|\sum_{h=1}^{H} f(s_h, a_h)\right| \le 3H\eta + \left|\sum_{h=1}^{H}\{r'(u_h) - r(u_h)\}\right| \le 5H\eta + |r'(u_H) - r(u_1)| = 5H\eta,$$

where the last equality uses the fact that s_{H+1} is the notational terminal state and s_1 is in the first layer. □

According to the previous proofs, we obtain that

$$\mathbb{P}_{\tau\sim\pi}\left(|f(\tau)| \ge 5H\eta\right) \le HC_{\mathsf{sa}}(\pi, \pi_{\text{off}}) \cdot \left(\frac{2L(2\eta/3)}{1-p} + \frac{8L(\eta/3)}{p} + \frac{L(\eta)}{1-2p}\right).$$

By choosing $p = 1/3$ and replacing η by $\eta/(5H)$, we have

$$\begin{aligned}
\mathbb{P}_{\tau\sim\pi}\left(|f(\tau)| \ge \eta\right) &\le HC_{\mathsf{sa}}(\pi, \pi_{\text{off}}) \cdot (6L(2\eta/(15H)) + 24L(\eta/(15H)) + 3L(\eta/(5H))) \\
&\overset{(i)}{\le} 33HC_{\mathsf{sa}}(\pi, \pi_{\text{off}}) \cdot L(\eta/(15H)) = 33HC_{\mathsf{sa}}(\pi, \pi_{\text{off}}) \cdot \mathbb{P}_{\tau\sim\pi_{\text{off}}}\left(|f(\tau)| \ge \eta/(15H)\right),
\end{aligned}$$

where (i) uses the fact that $L(\eta)$ is monotonically decreasing with η. Hence we obtain

$$\begin{aligned}
\mathbb{E}_{\tau\sim\pi}\left[f(\tau)^2\right] &\overset{(i)}{=} \int_0^1 2\eta \cdot \mathbb{P}_{\tau\sim\pi}\left(|f(\tau)| \ge \eta\right) d\eta \\
&\le \int_0^1 2\eta \cdot 33HC_{\mathsf{sa}}(\pi, \pi_{\text{off}}) \cdot \mathbb{P}_{\tau\sim\pi_{\text{off}}}\left(|f(\tau)| \ge \eta/(15H)\right) d\eta \\
&\le \int_0^{15H} 2\eta \cdot 33HC_{\mathsf{sa}}(\pi, \pi_{\text{off}}) \cdot \mathbb{P}_{\tau\sim\pi_{\text{off}}}\left(|f(\tau)| \ge \eta/(15H)\right) d\eta \\
&= 7425H^3C_{\mathsf{sa}}(\pi, \pi_{\text{off}}) \cdot \int_0^1 2(\eta/(15H)) \cdot \mathbb{P}_{\tau\sim\pi_{\text{off}}}\left(|f(\tau)| \ge \eta/(15H)\right) d(\eta/(15H)) \\
&= 7425H^3C_{\mathsf{sa}}(\pi, \pi_{\text{off}}) \cdot \mathbb{E}_{\tau\sim\pi_{\text{off}}}[f(\tau)^2],
\end{aligned}$$

where (i) uses integration by parts and also the assumption that for any trajectory τ, $f(\tau) \in [-1, 1]$. Additionally, we upper bound the expected total reward under policy π with Cauchy-Schwarz inequality:

$$\mathbb{E}_{\tau\sim\pi}[|f(\tau)|] \le \sqrt{\mathbb{E}_{\tau\sim\pi}[f(\tau)^2]} \lesssim \sqrt{H^3C_{\mathsf{sa}}(\pi, \pi_{\text{off}}) \cdot \mathbb{E}_{\tau\sim\pi_{\text{off}}}\left[f(\tau)^2\right]}.$$

□

F.1.2 Missing Proofs in section 7.3.1

Proof of Theorem 7.1. For any reward model r, we use $r^\star[r] = r - r^\star$ to denote the difference between reward model r and $r^\star$. Then we have

$$J_{r^\star[r]}(\pi) = J_r(\pi) - J(\pi),$$

For any reward $r \in \mathcal{R}$, we also have

$$\sum_{\tau \in \mathcal{D}_{\text{orm}}} \left(r(\tau) - r^\star(\tau)\right)^2 = \sum_{\tau \in \mathcal{D}_{\text{orm}}} \left(r^\star[r](\tau)\right)^2.$$

We notice that according to our assumption on the MDPs, for any trajectory τ and reward model r, $r(\tau) \in [0,1]$, hence $r^\star[r](\tau) \in [-1,1]$. According to Foster et al. [203, Lemma A.3] and union bound over $\mathcal{R}$, with probability $1-p$ we have for any $r \in \mathcal{R}$,

$$\left| \frac{1}{|\mathcal{D}_{\text{orm}}|} \sum_{\tau \in \mathcal{D}_{\text{orm}}} \left(r^\star[r](\tau)\right)^2 - \mathbb{E}_{\tau \sim \pi_{\text{off}}} \left[\left(r^\star[r](\tau)\right)^2\right] \right| \le \frac{1}{2} \cdot \mathbb{E}_{\tau \sim \pi_{\text{off}}} \left[\left(r^\star[r](\tau)\right)^2\right] + \frac{4 \log(2|\mathcal{R}|/p)}{|\mathcal{D}_{\text{orm}}|}. \tag{F.6}$$

When eq. (F.6) holds for all $r \in \mathcal{R}$, since $\widehat{r}$ is the solution of optimization problem eq. (7.3), we have

$$\begin{aligned}
\mathbb{E}_{\tau \sim \pi_{\text{off}}} \left[\left(r^\star[\widehat{r}](\tau)\right)^2\right] &\le \frac{2}{|\mathcal{D}_{\text{orm}}|} \sum_{\tau \in \mathcal{D}_{\text{orm}}} \left(r^\star[\widehat{r}](\tau)\right)^2 + \frac{8 \log(2|\mathcal{R}|/p)}{|\mathcal{D}_{\text{orm}}|} \\
&\le \frac{2}{|\mathcal{D}_{\text{orm}}|} \sum_{\tau \in \mathcal{D}_{\text{orm}}} \left(r^\star[r^\star](\tau)\right)^2 + \frac{8 \log(2|\mathcal{R}|/p)}{|\mathcal{D}_{\text{orm}}|} \\
&\le 3\mathbb{E}_{\tau \sim \pi_{\text{off}}} \left[\left(r^\star[r^\star](\tau)\right)^2\right] + \frac{16 \log(2|\mathcal{R}|/p)}{|\mathcal{D}_{\text{orm}}|} \\
&= \frac{16 \log(2|\mathcal{R}|/p)}{|\mathcal{D}_{\text{orm}}|},
\end{aligned}$$

where the last equation uses $r^\star[r^\star] = r^\star - r^\star = 0$. Therefore, according to Theorem 7.3, with probability at least $1-p$ we have for any policy π,

$$|J_{\widehat{r}}(\pi) - J(\pi)| = |J_{r^\star[\widehat{r}]}(\pi)| \lesssim \sqrt{H^3 \cdot C_{\mathsf{sa}}(\pi, \pi_{\text{off}}) \cdot \mathbb{E}_{\tau \sim \pi_{\text{off}}} \left[\left(r^\star[\widehat{r}](\tau)\right)^2\right]} \lesssim H^{3/2} \cdot \sqrt{\frac{C_{\mathsf{sa}}(\pi, \pi_{\text{off}}) \cdot \log(|\mathcal{R}|/p)}{|\mathcal{D}_{\text{orm}}|}}.$$

□

Proof of Theorem 7.2. Suppose $\pi^\star$ to be the best policy of the true MDP $M^\star$. By Theorem 7.1, with probability at least $1-\delta$, for any policy π we have

$$|J(\pi) - J_{\widehat{r}}(\pi)| \lesssim \sqrt{\frac{H^3 C_{\mathsf{sa}}(\pi, \pi_{\text{off}}) \cdot \log(|\mathcal{R}|/\delta)}{|\mathcal{D}_1|}}.$$

Especially since $\pi^\star, \widehat{\pi} \in \Pi$, we have

$$|J(\pi^\star) - J_{\widehat{r}}(\pi^\star)| \lesssim \sqrt{\frac{H^3 C_{\mathsf{sa}}(\pi^\star, \pi_{\mathrm{off}}) \cdot \log(|\mathcal{R}|/\delta)}{|\mathcal{D}_1|}} \leq \sqrt{\frac{H^3 \sup_{\pi \in \Pi} C_{\mathsf{sa}}(\pi^\star, \pi_{\mathrm{off}}) \cdot \log(|\mathcal{R}|/\delta)}{|\mathcal{D}_1|}}$$

and also

$$|J(\widehat{\pi}) - J_{\widehat{r}}(\widehat{\pi})| \lesssim \sqrt{\frac{H^3 C_{\mathsf{sa}}(\widehat{\pi}, \pi_{\mathrm{off}}) \cdot \log(|\mathcal{R}|/\delta)}{|\mathcal{D}_1|}} \leq \sqrt{\frac{H^3 \sup_{\pi \in \Pi} C_{\mathsf{sa}}(\pi^\star, \pi_{\mathrm{off}}) \cdot \log(|\mathcal{R}|/\delta)}{|\mathcal{D}_1|}}.$$

Next, by calling algorithm $\mathfrak{A}$, with probability at least $1 - \delta$ we have

$$\max_{\pi} J_{\widehat{r}}(\pi) - J_{\widehat{r}}(\widehat{\pi}) \leq \varepsilon_{\mathsf{alg}}(|\mathcal{D}|_2, \delta),$$

Hence with probability at least $1 - 2\delta$,

$$\begin{aligned}
\max_{\pi} J(\pi) - J(\widehat{\pi}) &= J(\pi^\star) - J(\widehat{\pi}) \\
&\leq J_{\widehat{r}}(\pi^\star) - J_{\widehat{r}}(\widehat{\pi}) + \mathcal{O}\left(\sqrt{\frac{H^3 C_{\mathsf{sa}}(\Pi, \pi_{\mathrm{off}}) \cdot \log(|\mathcal{M}|/\delta)}{|\mathcal{D}_1|}}\right) \\
&\leq \max_{\pi} J_{\widehat{r}}(\pi) - J_{\widehat{r}}(\widehat{\pi}) + \mathcal{O}\left(\sqrt{\frac{H^3 C_{\mathsf{sa}}(\Pi, \pi_{\mathrm{off}}) \cdot \log(|\mathcal{M}|/\delta)}{|\mathcal{D}_1|}}\right) \\
&\lesssim \varepsilon_{\mathsf{alg}}(|\mathcal{D}|_2, \delta) + \sqrt{\frac{H^3 C_{\mathsf{sa}}(\Pi, \pi_{\mathrm{off}}) \cdot \log(|\mathcal{M}|/\delta)}{|\mathcal{D}_1|}}.
\end{aligned}$$

□

F.1.3 Missing Proofs in section 7.3.4

In the following, we will prove Theorem 7.4 and Theorem 7.7. Before the proofs, we first present several useful lemmas.

Lemma F.1 (Lemma A.1 in [175]). *Suppose Theorem 7.5 holds, and $\widehat{\pi}$ satisfies*

$$\widehat{\pi} = \arg\max_{\pi \in \Pi} \mathbb{E}_{\tau \sim \pi}\left[\widehat{r}(\tau) - \log \frac{\pi(\tau)}{\pi_{\mathsf{ref}}(\tau)}\right] \tag{F.7}$$

then we have

$$\mathcal{J}_\beta(\pi^\star_\beta) - \mathcal{J}_\beta(\widehat{\pi}) \leq \mathbb{E}_{\tau \sim \pi^\star_\beta, \tilde{\tau} \sim \widehat{\pi}}\left[r^\star(\tau) - r^\star(\tilde{\tau}) - \widehat{r}(\tau) + \widehat{r}(\tilde{\tau})\right].$$

Proof of Theorem F.1. We calculate

$$\mathcal{J}_\beta(\pi^\star_\beta) - \mathcal{J}_\beta(\widehat{\pi}) = \mathbb{E}_{\tau \sim \pi^\star_\beta}\left[r^\star(\tau) - \log \frac{\pi^\star_\beta(\tau)}{\pi_{\mathsf{ref}}(\tau)}\right] - \mathbb{E}_{\tau \sim \widehat{\pi}}\left[r^\star(\tau) - \log \frac{\widehat{\pi}(\tau)}{\pi_{\mathsf{ref}}(\tau)}\right]$$

$$
\begin{aligned}
&= \mathbb{E}_{\tau\sim\pi^\star_\beta}\left[\widehat{r}(\tau) - \log\frac{\pi^\star_\beta(\tau)}{\pi_{\mathsf{ref}}(\tau)}\right] - \mathbb{E}_{\tau\sim\widehat{\pi}}\left[\widehat{r}(\tau) - \log\frac{\widehat{\pi}(\tau)}{\pi_{\mathsf{ref}}(\tau)}\right] \\
&\qquad + \mathbb{E}_{\tau\sim\pi^\star_\beta}[r^\star(\tau) - \widehat{r}(\tau)] - \mathbb{E}_{\tau\sim\widehat{\pi}}[r^\star(\tau) - \widehat{r}(\tau)] \\
&\overset{(i)}{\leq} \mathbb{E}_{\tau\sim\pi^\star_\beta}[r^\star(\tau) - \widehat{r}(\tau)] - \mathbb{E}_{\tau\sim\widehat{\pi}}[r^\star(\tau) - \widehat{r}(\tau)] \\
&= \mathbb{E}_{\tau\sim\pi^\star_\beta,\tilde{\tau}\sim\widehat{\pi}}\left[r^\star(\tau) - r^\star(\tilde{\tau}) - \widehat{r}(\tau) + \widehat{r}(\tilde{\tau})\right],
\end{aligned}
$$

where in (i) we uses eq. (F.7) and the realizability assumption Theorem 7.5. □

Lemma F.2 (Lemma C.5 in [177]). *We define reward model $\widehat{r}$:*

$$
\widehat{r}(\tau) = \log\frac{\widehat{\pi}(\tau)}{\pi_{\mathsf{ref}}(\tau)},
$$

where $\widehat{\pi}$ is given in eq. (7.9). Then with probability at least $1-\delta$, we have

$$
\mathbb{E}_{\tau,\tilde{\tau}\overset{\text{i.i.d.}}{\sim}\pi_{\mathsf{ref}}}\left[\left(r^\star(\tau) - r^\star(\tilde{\tau}) - \widehat{r}(\tau) + \widehat{r}(\tilde{\tau})\right)^2\right] \leq \kappa^2\cdot\frac{2\log(|\Pi|/\delta)}{|\mathcal{D}|},
$$

where $\kappa = 16V_{\max}e^{2V_{\max}}$.

Lemma F.3. *For MDP $M = (\mathcal{S},\mathcal{A},T,r,H)$ with reward function $f:\mathcal{S}\times\mathcal{A}\to[-1,1]$, then for any policy π and $\tilde{\pi}$, we have*

$$
\mathbb{E}_{\tau\sim\pi,\tilde{\tau}\sim\tilde{\pi}}[|f(\tau) - f(\tilde{\tau})|] \lesssim \sqrt{H^3(C_{\mathsf{sa}}(\pi,\pi_{\mathrm{off}})\vee C_{\mathsf{sa}}(\tilde{\pi},\pi_{\mathrm{off}}))\cdot\mathbb{E}_{\tau,\tilde{\tau}\overset{\text{i.i.d.}}{\sim}\pi_{\mathrm{off}}}\left[\left(f(\tau)-f(\tilde{\tau})\right)^2\right]},
$$

where $C_{\mathsf{sa}}(\pi,\pi_{\mathrm{off}})$ is defined in eq. (7.5).

Proof of Theorem F.3. Suppose the layered state space representation is $\mathcal{S} = \cup_{h=1}^H\mathcal{S}_h$. We construct a new MDP M' with horizon $2H$. The state spaces of M' among layer 1 to H, and among layer $H+1$ to $2H$, are both $\mathcal{S}_1,\cdots,\mathcal{S}_H$. The action space of $\mathcal{M}'$ is $\mathcal{A}$. The transition model of M' follows transition model T in the first H layers, and then transits to the initial state s_1 of M and follows transition model T again in the last H layers as well. The reward model g in the first H layers are set to be f and in the last H layers are set to be $-f$.

We define the policy π' of MDP M', which follows policy π in the first H layers, and follows policy $\tilde{\pi}$ in the last H layers. We further define the policy π'_{off} of MDP M', which follows policy π_{off} in the first H layers, and in the last H layers as well. Then it is easy to see that

$$
\mathbb{E}_{\tau\sim\pi,\tilde{\tau}\sim\tilde{\pi}}[|f(\tau) - f(\tilde{\tau})|] = \mathbb{E}_{\tau'\sim\pi'}[|g(\tau')|],
$$

where the right hand side is the expected rewards in the new MDP M'. We also have

$$
\mathbb{E}_{\tau,\tilde{\tau}\overset{\text{i.i.d.}}{\sim}\pi_{\mathrm{off}}}\left[\left(f(\tau)-f(\tilde{\tau})\right)^2\right)^2\right] = \mathbb{E}_{\tau'\sim\pi_{\mathrm{off}}}\left[g(\tau')^2\right].
$$

Next, according to the construction of M', we have

$$
C_{\mathsf{sa}}(\pi,\pi_{\mathrm{off}})\vee C_{\mathsf{sa}}(\tilde{\pi},\pi_{\mathrm{off}}) = \sup_{h\in[2H]}\sup_{s\in\mathcal{S}_h,a\in\mathcal{A}}\frac{d^{\pi'}(s,a)}{d^{\pi'_{\mathrm{off}}}(s,a)}.
$$

Hence according to Theorem 7.3, we have

$$\begin{aligned}\mathbb{E}_{\tau'\sim\pi'}[|g(\tau')|] &\lesssim \sqrt{H^3 \cdot \sup_{h\in[2H]} \sup_{s\in\mathcal{S}_h, a\in\mathcal{A}} \frac{d^{\pi'}(s,a)}{d^{\pi'}_{\mathsf{off}}(s,a)} \cdot \mathbb{E}_{\tau'\sim\pi_{\mathsf{off}}}[g(\tau')^2]} \\ &= \sqrt{H^3 (C_{\mathsf{sa}}(\pi,\pi_{\mathsf{off}}) \vee C_{\mathsf{sa}}(\tilde{\pi},\pi_{\mathsf{off}})) \cdot \mathbb{E}_{\tau'\sim\pi_{\mathsf{off}}}[g(\tau')^2]},\end{aligned}$$

which implies that

$$\mathbb{E}_{\tau\sim\pi,\tilde{\tau}\sim\tilde{\pi}}[|f(\tau)-f(\tilde{\tau})|] \lesssim \sqrt{H^3 (C_{\mathsf{sa}}(\pi,\pi_{\mathsf{off}}) \vee C_{\mathsf{sa}}(\tilde{\pi},\pi_{\mathsf{off}})) \cdot \mathbb{E}_{\tau,\tilde{\tau}\overset{\text{i.i.d.}}{\sim}\pi_{\mathsf{off}}}\left[(f(\tau)-f(\tilde{\tau}))^2\right]}.$$

□

Now we are ready to prove Theorem 7.4 and Theorem 7.7.

Proof of Theorem 7.4. According to Theorems F.2 and F.3, for any policy π, with probability at least $1-\delta$,

$$\begin{aligned}&\mathbb{E}_{\tau\sim\pi,\tau'\sim\pi'}[|r^\star(\tau)-r^\star(\tau')-\widehat{r}(\tau)+\widehat{r}(\tau')|] \\ &\lesssim \sqrt{H^3 \cdot (C_{\mathsf{sa}}(\pi,\pi_{\mathsf{ref}}) \vee C_{\mathsf{sa}}(\pi',\pi_{\mathsf{ref}})) \cdot \mathbb{E}_{\tau,\tilde{\tau}\overset{\text{i.i.d.}}{\sim}\pi_{\mathsf{ref}}}\left[(r^\star(\tau)-r^\star(\tilde{\tau})-\widehat{r}(\tau)+\widehat{r}(\tilde{\tau}))^2\right]} \\ &\lesssim H^{3/2} V_{\max} e^{2V_{\max}} \cdot \sqrt{\frac{(C_{\mathsf{sa}}(\pi,\pi_{\mathsf{ref}}) \vee C_{\mathsf{sa}}(\pi',\pi_{\mathsf{ref}}))\log(|\mathcal{R}|/\delta)}{|\mathcal{D}|}},\end{aligned}$$

where the second line uses Theorem F.2 with the class of policies $\{\widehat{\pi} : \widehat{\pi}(\tau) \propto \pi_{\mathsf{ref}}(\tau)\exp(r(\tau))\ \forall r \in \mathcal{R}\}$. □

Proof of Theorem 7.7. Let $\widehat{\pi}$ to be the policy given in eq. (7.9). We define the reward model $\widehat{r}$ as

$$\widehat{r}(\tau) = \log \frac{\widehat{\pi}(\tau)}{\pi_{\mathsf{ref}}(\tau)}.$$

Then it is easy to see that $\widehat{\pi}$ is the solution of eq. (F.7) according to the above reward model. Since $\widehat{\pi} \in \Pi$ and $\pi^\star_\beta \in \Pi$ according to Theorem 7.5, with probability at least $1-\delta$ we have

$$\begin{aligned}\mathcal{J}_\beta(\pi^\star_\beta) - \mathcal{J}_\beta(\widehat{\pi}) &\overset{(i)}{\leq} \mathbb{E}_{\tau\sim\pi^\star_\beta,\tilde{\tau}\sim\widehat{\pi}}[r^\star(\tau)-r^\star(\tilde{\tau})-\widehat{r}(\tau)+\widehat{r}(\tilde{\tau})] \\ &\overset{(ii)}{\lesssim} \sqrt{H^3 \cdot C_{\mathsf{sa}}(\Pi,\pi_{\mathsf{ref}}) \cdot \mathbb{E}_{\tau,\tilde{\tau}\overset{\text{i.i.d.}}{\sim}\pi_{\mathsf{ref}}}\left[(r^\star(\tau)-r^\star(\tilde{\tau})-\widehat{r}(\tau)+\widehat{r}(\tilde{\tau}))^2\right]} \\ &\overset{(iii)}{\lesssim} H^{3/2} V_{\max} e^{2V_{\max}} \cdot \sqrt{\frac{C_{\mathsf{sa}}(\Pi,\pi_{\mathsf{ref}})\log(|\Pi|/\delta)}{|\mathcal{D}|}},\end{aligned}$$

where (i) uses Theorem F.1, (ii) uses Theorem F.3 with the reward model to be $r-\widehat{r}$, and (iii) uses Theorem F.2. □

F.2 Analysis of ARMOR with Total Reward

ARMOR is an offline RL algorithm introduced by Xie et al. [161]. Given n trajectories of data in the form of $(s_1, a_1, r_1, \cdots, s_H, a_H, r_H) \sim \pi_{\text{off}}$, they proved that for the optimal policy $\pi^\star$, the output policy $\widehat{\pi}$ of their algorithm satisfies

$$J(\pi^\star) - J(\widehat{\pi}) \lesssim \sqrt{\frac{H^2 \cdot C(\pi^\star, \pi_{\text{off}}) \log(|\mathcal{M}|/\delta)}{n}}. \tag{F.8}$$

Notably, this error bound scales with the best-policy concentrability $C(\pi^\star, \pi_{\text{off}})$ rather than the all-policy concentrability $\sup_\pi C(\pi, \pi_{\text{off}})$ shown in Theorem 7.2. A natural question arises: can we develop a variant of ARMOR that takes data in the form of trajectory plus total reward, i.e., $(s_1, a_1, \cdots, s_H, a_H, R)$, while maintaining the sample complexity that scales with best-policy concentrability instead of all-policy concentrability? In this section, we answer this question affirmatively by presenting and analyzing such a variant of the ARMOR algorithm.

Algorithm 4 ARMOR with Total Rewards

1: **Input:** Batch data $\mathcal{D}$, model class $\mathcal{M}$, parameter α.

2: Construct version space

$$\mathcal{M}_\alpha = \left\{ M \in \mathcal{M} : \max_{M' \in \mathcal{M}} \mathcal{L}_{\mathcal{D}}(\mathcal{M}') - \mathcal{L}_{\mathcal{D}}(\mathcal{M}) \le \alpha \right\},$$

where for MDP model M with transition model P_M and reward function r_M,

$$\mathcal{L}_{\mathcal{D}}(M) := \sum_{(s_1, a_1, \cdots, s_H, a_H, R) \in \mathcal{D}} \left[\sum_{h=1}^{H-1} \log P_M(s_{h+1} \mid s_h, a_h) - \left(\sum_{h=1}^{H} r_M(s_h, a_h) - R \right)^2 \right]$$

3: Output the best policy with pessimism:

$$\widehat{\pi} = \arg\max_{\pi} \min_{M \in \mathcal{M}_\alpha} J_M(\pi),$$

where $J_M(\pi)$ denotes the value function of policy π under MDP M

Suppose the learner is given a model class $\mathcal{M}$, which realizes the ground truth model $M^\star$. The learner also have access to some offline batched data $\mathcal{D}$, which is collected from policy π_{off}. Specifically, $\mathcal{D}$ consists of i.i.d. sampled trajectories $\tau = (s_1, a_1, \cdots, s_H, a_H)$ together with its total reward $R = r(\tau)$. Each of the trajectories is collected by executing π_{off} under the ground truth MDP $M^\star$.

We consider Algorithm 4, which is a variant of the ARMOR algorithm introduced in Xie et al. [161] after specifically taylored for data of trajectories with total reward. We have the following guarantee on its sample complexity, which only relies on the single policy concentrability $C_{\mathsf{sa}}(\pi^\star, \pi_{\text{off}})$.

Theorem F.4. *Suppose the model class $\mathcal{M}$ realizes the ground truth model $M^\star$. Then there exists some positive constant c such that for any $\delta > 0$, by letting $\alpha = c \cdot \log(|\mathcal{M}|/\delta)$, with*

probability at least $1-\delta$*, the output of Algorithm 4 satisfies*

$$J_{M^\star}(\pi^\star) - J_{M^\star}(\widehat{\pi}) \lesssim \sqrt{\frac{H^3 \cdot C_{\mathsf{sa}}(\pi^\star, \pi_{\mathrm{off}}) \log(|\mathcal{M}|/\delta)}{n}}$$

Proof of Theorem F.4. Adopting the similar way as Xie et al. [161], we let

$$\ell_{\mathcal{D}}(M) = \prod_{(s_1,a_1,\cdots,s_H,a_H,R)\in\mathcal{D}} \prod_{h=1}^{H-1} P_M(s_{h+1}, s_h, a_h).$$

Then according to Xie et al. [161, Lemma 8], with probability at least $1-\delta$, we have

$$\max_{M\in\mathcal{M}} \log \ell_{\mathcal{D}}(M) - \ell_{\mathcal{D}}(M^\star) \le \log\left(\frac{|\mathcal{M}|}{\delta}\right),$$

where $M^\star$ denotes the ground truth model. Additionally, according to Xie et al. [142, Theorem A.1] (by letting $\gamma = 0$ and merge states or actions across the entire horizon into one state or action), we have

$$\sum_{(\tau,R)\in\mathcal{D}} (r_{M^\star}(\tau) - R)^2 - \min_{M\in\mathcal{M}} \sum_{(\tau,R)\in\mathcal{D}} (r_M(\tau) - R)^2 \lesssim \log\left(\frac{|\mathcal{M}|}{\delta}\right).$$

Combining the above inequalities together, we obtain that

$$\max_{M\in\mathcal{M}} \mathcal{L}_{\mathcal{D}}(M) - \mathcal{L}_{\mathcal{D}}(M^\star) \lesssim \log\left(\frac{|\mathcal{M}|}{\delta}\right).$$

Hence with our choice of α, with probability at least $1-\delta$ we have $M^\star \in \mathcal{M}$. Next, by Xie et al. [161, Lemma 7], with probability at least $1-\delta$, for any $M \in \mathcal{M}$,

$$\mathbb{E}_{(s_1,a_1,\cdots,s_H,a_H)\sim\pi_{\mathrm{off}}} \left[\sum_{h=1}^{H-1} D_{\mathrm{TV}}(P_M(s_{h+1} \mid s_h, a_h), P_{M^\star}(s_{h+1} \mid s_h, a_h)) + \left(\sum_{h=1}^{H} r_M(s_h, a_h) - r_{M^\star}(s_h, a_h)\right)^2\right]$$
$$\lesssim \frac{\max_{M'\in\mathcal{M}} \mathcal{L}_{\mathcal{D}}(M') - \mathcal{L}_{\mathcal{D}}(M^\star) + \log(|\mathcal{M}|/\delta)}{n} \lesssim \frac{\log(|\mathcal{M}|/\delta)}{n}. \tag{F.9}$$

Finally, according to the optimality of $\widehat{\pi}$, if letting $\widehat{M} = \arg\min_{M\in\mathcal{M}} J_M(\widehat{\pi})$, we have

$$\begin{aligned} J_{M^\star}(\pi^\star) - J_{M^\star}(\widehat{\pi}) &\le J_{M^\star}(\pi^\star) - \min_{M\in\mathcal{M}} J_M(\widehat{\pi}) \\ &\overset{(i)}{=} J_{M^\star}(\pi^\star) - \max_{\pi} \min_{M\in\mathcal{M}} J_M(\widehat{\pi}) \\ &\le \max_{M\in\mathcal{M}} \left\{J_{M^\star}(\pi^\star) - J_M(\pi^\star)\right\}, \end{aligned}$$

where (i) uses the definition of $\widehat{\pi}$ in Algorithm 4. According to simulation lemma [e.g., 163, Lemma 7], we have

$$|J_{M^\star}(\pi^\star) - J_M(\pi^\star)| \le \sum_{h=1}^{H-1} \mathbb{E}_{(s_h,a_h)\sim d^{\pi^\star}} [D_{\mathrm{TV}}(P_M(s_{h+1} \mid s_h, a_h), P_{M^\star}(s_{h+1} \mid s_h, a_h))]$$

$$
\begin{aligned}
&+ \mathbb{E}_{\tau\sim\pi^\star}[|r_{M^\star}(\tau) - r_M(\tau)|] \\
&\le H^{1/2} \cdot \sqrt{C_{\mathsf{sa}}(\pi^\star, \pi_{\mathrm{off}}) \sum_{h=1}^{H-1} \mathbb{E}_{(s_h,a_h)\sim d^{\pi_{\mathrm{off}}}}[D_{\mathrm{TV}}(P_M(s_{h+1} \mid s_h, a_h), P_{M^\star}(s_{h+1} \mid s_h, a_h))^2]} \\
&+ \mathbb{E}_{\tau\sim\pi^\star}[|r_{M^\star}(\tau) - r_M(\tau)|],
\end{aligned}
$$

where the last inequality uses the Cauchy-Schwarz inequality and the definition of state-action concentrability. Additionally, according to Theorem 7.3, we have

$$
\mathbb{E}_{\tau\sim\pi^\star}[|r_{M^\star}(\tau) - r_M(\tau)|] \lesssim H^{3/2}\sqrt{C_{\mathsf{sa}}(\pi^\star, \pi_{\mathrm{off}}) \cdot \mathbb{E}_{\tau\sim\pi_{\mathrm{off}}}[(r_{M^\star}(\tau) - r_M(\tau))^2]}.
$$

Therefore, with probability at least $1-\delta$,

$$
\begin{aligned}
&J_{M^\star}(\pi^\star) - J_{M^\star}(\widehat{\pi}) \le \max_{M\in\mathcal{M}}\{|J_{M^\star}(\pi^\star) - J_M(\pi^\star)|\} \lesssim \sqrt{H^3 C_{\mathsf{sa}}(\pi^\star, \pi_{\mathrm{off}})} \\
&\cdot \sqrt{\mathbb{E}_{(s_1,a_1,\cdots,s_H,a_H)\sim\pi_{\mathrm{off}}}\left[\sum_{h=1}^{H-1} D_{\mathrm{TV}}(P_M(s_{h+1} \mid s_h, a_h), P_{M^\star}(s_{h+1} \mid s_h, a_h))\right] + \mathbb{E}_{\tau\sim\pi_{\mathrm{off}}}\left(r_M(\tau) - r_{M^\star}(\tau)\right)^2} \\
&\lesssim \sqrt{\frac{H^3 C_{\mathsf{sa}}(\pi^\star, \pi_{\mathrm{off}}) \log(|\mathcal{M}|/\delta)}{n}}
\end{aligned}
$$

where the last inequality is due to eq. (F.9). □

We observe that similar to the ARMOR algorithm, the sample complexity of Algorithm 4 scales with the single-policy concentrability $C_{\mathsf{sa}}(\pi^\star, \pi_{\mathrm{off}})$ rather than the all-policy concentrability $\sup_\pi C_{\mathsf{sa}}(\pi, \pi_{\mathrm{off}})$ as shown in Theorem F.4. However, unlike ARMOR, Theorem F.4 requires the assumption that the total reward is bounded by $[0, 1]$. Without this assumption, if we include the scale of total reward in the proof of Theorem F.4, we would directly obtain a sample complexity of $\sqrt{\frac{H^5 \cdot C_{\mathsf{sa}}(\pi^\star, \pi_{\mathrm{off}}) \log(|\mathcal{M}|/\delta)}{n}}$. Comparing this with ARMOR's sample complexity (eq. (F.8)), we observe a gap of H^3 (when obtaining an ε-sub-optimality bound) in terms of the horizon dependency.

We suspect this gap arises fundamentally from the extra horizon dependency in the Change of Trajectory Measure Lemma (Theorem 7.3). To illustrate this, consider how changing measures affects complexity (ignoring log terms):

- Change of state-action measure: For any function $g : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$,

$$
\frac{\sum_{h=1}^H \mathbb{E}_{(s_h,a_h)\sim d^\pi}[g(s_h, a_h)^2]}{\sum_{h=1}^H \mathbb{E}_{(s_h,a_h)\sim d^{\pi_{\mathrm{off}}}}[g(s_h, a_h)^2]} \le \max_{h\in[H]} \frac{\mathbb{E}_{(s_h,a_h)\sim d^\pi}[g(s_h, a_h)^2]}{\mathbb{E}_{(s_h,a_h)\sim d^{\pi_{\mathrm{off}}}}[g(s_h, a_h)^2]} \le \max_{h\in[H]} \sup_{s_h\in\mathcal{S}_h, a_h\in\mathcal{A}} \frac{d^\pi(s_h, a_h)}{d^{\pi_{\mathrm{off}}}(s_h, a_h)}.
$$

- Change of trajectory measure (Theorem 7.3): For any function $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$,

$$
\frac{\mathbb{E}_{\tau\sim\pi}[f(\tau)^2]}{\mathbb{E}_{\tau\sim\pi_{\mathrm{off}}}[f(\tau)^2]} \lesssim H^3 \max_{h\in[H]} \sup_{s_h\in\mathcal{S}_h, a\in\mathcal{A}} \frac{d^\pi(s_h, a_h)}{d^{\pi_{\mathrm{off}}}(s_h, a_h)}.
$$

While this gap in horizon dependency raises an interesting theoretical question, the precise polynomial dependence on horizon is not the main focus of our paper. We leave a more thorough investigation of this horizon dependency gap between outcome and process supervision as an interesting direction for future work.

F.3 Missing Proofs in section 7.4

F.3.1 Proof of Theorem 7.8

First we present a lemma.

Lemma F.5. *For any MDP $M = (\mathcal{S}, \mathcal{A}, P, r, H)$ and policy μ, if we let $A^\mu : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ to be the advantage function of policy μ (defined in eq. (7.10)), then for any policy π, we have*

$$J_{A^\mu}(\pi) = J_r(\pi) - J_r(\mu).$$

Proof of Theorem F.5. In view of the performance difference lemma [190], we have

$$\begin{aligned} J_r(\pi) - J_r(\mu) &= H \cdot \mathbb{E}_{(s,a)\sim d^\pi}\left[Q^\mu(s,a) - Q^\mu(s,\mu)\right] \\ &= H \cdot \mathbb{E}_{(s,a)\sim d^\pi}\left[Q^\mu(s,a) - V^\mu(s)\right] \\ &= H \cdot \mathbb{E}_{(s,a)\sim d^\pi}\left[A^\mu(s,a)\right] \\ &= J_{A^\mu}(\pi). \end{aligned}$$

□

Next we present the proof of Theorem 7.8.

Proof of Theorem 7.8. For any policy π, we decompose

$$\begin{aligned} J_r(\pi) - J_r(\widehat{\pi}) &= J_r(\pi) - J_{\widehat{r}}(\pi) + J_{\widehat{r}}(\pi) - J_{\widehat{r}}(\widehat{\pi}) + J_{\widehat{r}}(\widehat{\pi}) - J_r(\widehat{\pi}) \\ &\overset{(i)}{=} J_{A^\mu}(\pi) - J_{\widehat{r}}(\pi) + J_{\widehat{r}}(\pi) - J_{\widehat{r}}(\widehat{\pi}) + J_{\widehat{r}}(\widehat{\pi}) - J_{A^\mu}(\widehat{\pi}), \end{aligned} \tag{F.10}$$

where in (i) we use Theorem F.5. Next, we notice that for any policy π, we can write down the value of policy π as the inner product between the occupancy measures of policy π and the reward function, which implies that

$$\begin{aligned} J_{A^\mu}(\pi) - J_{\widehat{r}}(\pi) &= H \cdot \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d^\pi(s,a) \cdot \left(A^\mu(s,a) - \widehat{r}(s,a)\right) \\ &\overset{(i)}{\le} H \cdot \sqrt{\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d^\pi(s,a) \cdot \left(A^\mu(s,a) - \widehat{r}(s,a)\right)^2} \\ &\overset{(ii)}{\le} H \cdot \sqrt{C_{\mathsf{sa}}(\nu) \cdot \mathbb{E}_{(s,a)\sim\nu}\left[\left(A^\mu(s,a) - \widehat{r}(s,a)\right)^2\right]} \\ &\overset{(iii)}{\le} H\sqrt{C_{\mathsf{sa}}(\nu)} \cdot \varepsilon_{\mathsf{stat}}, \end{aligned}$$

where (i) uses Cauchy-Schwarz inequality, (ii) adopts the definition of $C_{\mathsf{sa}}(\nu)$ in eq. (7.11), and finally in (iii) we use eq. (7.12). Additionally, according to eq. (7.13) we have

$$J_{\widehat{r}}(\pi) - J_{\widehat{r}}(\widehat{\pi}) \le \max_\pi J_{\widehat{r}}(\pi) - J_{\widehat{r}}(\widehat{\pi}) \le \varepsilon_{\mathsf{alg}}.$$

Bringing these inequalities back to eq. (F.10), we obtain that

$$J_r(\pi) - J_r(\widehat{\pi}) \le 2H\sqrt{C_{\mathsf{sa}}(\nu)} \cdot \varepsilon_{\mathsf{stat}} + \varepsilon_{\mathsf{alg}}.$$

□

F.3.2 Proof of Theorem 7.9

Proof of Theorem 7.9. We construct M as follows: We let $H = 2$, $\mathcal{A} = \{0, 1\}$, and $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2$ where $\mathcal{S}_1 = \{a\}$ and $\mathcal{S}_2 = \{b, c\}$. We further define the transition model P as

$$P(b \mid a, 0) = 1 \quad \text{and} \quad P(c \mid a, 1) = 1,$$

and the reward function R as

$$R(a, 0) = R(a, 1) = 0, \quad \text{and} \quad R(b, 0) = 1, R(b, 1) = 0, \quad \text{and} \quad R(c, 0) = \frac{2}{3}, R(c, 1) = \frac{1}{2}.$$

The best policy $\pi^\star$ of this MDP M satisfies

$$\pi^\star(a) = \pi^\star(b) = \pi^\star(c) = 0,$$

hence $J_r(\pi^\star) = 1$. We next choose policy μ as $\mu(a) = 0$, $\mu(b) = \mu(c) = 1$. Then we can calculate that

$$\begin{aligned}
Q^\mu(b, 0) = R(b, 0) = 1, \quad Q^\mu(b, 1) = R(b, 1) = 0, \\
Q^\mu(c, 0) = R(c, 0) = \frac{2}{3}, \quad Q^\mu(c, 1) = R(c, 1) = \frac{1}{2}, \\
Q^\mu(a, 0) = R(a, 0) + Q^\mu(b, \mu(b)) = 0, \quad Q^\mu(a, 1) = R(a, 1) + Q^\mu(c, \mu(c)) = \frac{1}{2}.
\end{aligned}$$

Therefore, the greedy policy $\widehat{\pi}$ with respect to the MDP with reward function to be Q^μ is

$$\widehat{\pi}(a) = 1, \quad \widehat{\pi}(b) = 0, \quad \widehat{\pi}(c) = 0,$$

which satisfies

$$\max_\pi J_r(\pi) - J_r(\widehat{\pi}) = 1 - \frac{2}{3} = \frac{1}{3}.$$

□

Remark F.6. *With the same choice of MDP M and policy μ in the above proof, we can calculate the advantage function A^μ as*

$$A^\mu(b, 0) = 1, \quad A^\mu(b, 1) = 0, \quad \textit{and} \quad A^\mu(c, 0) = \frac{1}{6}, \quad A^\mu(c, 1) = 0, \quad \textit{and} \quad A^\mu(a, 0) = 0, \quad A^\mu(a, 1) = \frac{1}{2}.$$

And it is easy to verify that the greedy policy with respect to the MDP with reward function to be A^μ coincides with $\pi^\star$.

Appendix G

Proofs for Chapter 8

G.1 More Related Works

We review more related works in this section.

The *coverability coefficient* has recently gained attention in the theory of online reinforcement learning [153–156]. This condition is in the same spirit as the widely used concentrability coefficient [142, 146–152], a concept frequently employed in the theory of offline (or batch) reinforcement learning. A well-known duality suggests that the coverability coefficient can be interpreted as the optimal concentrability coefficient attainable by any offline data distribution. For further discussion, see [153].

A related body of theoretical work explores *reinforcement learning with trajectory feedback* [27, 29–31, 193], where the learner receives only episode-level feedback at the end of each trajectory. This category also encompasses preference-based reinforcement learning [28, 32, 34, 173, 188], which relies on pairwise comparisons between trajectories. While most prior work focuses on tabular or linear MDP settings, we take a step further by studying learning with function approximation, and bound the complexity by the coverability coefficient.

In the context of *Online Reinforcement Learning*, numerous prior works have investigated the complexity of exploration and policy optimization, introducing various complexity measures such as Bellman rank [208], Eluder dimension [205, 225], witness rank [226], Bellman-Eluder dimension [200], the bilinear class [201], and decision-estimation coefficients [203]. These complexity notions characterize properties of the function or model class but are generally not instance-dependent. In contrast, Xie et al. [153] introduces the instance-dependent notion of *coverability coefficients* to provide complexity bounds in online reinforcement learning. For further discussion on instance-dependent complexity measures, we refer the reader to the discussions therein.

We further review some literatures on *online preference-based learning* or *online RLHF*. [28, 34, 173, 189, 227, 228] provides theoretical guarantees for tabular MDPs and linear MDPs. [182] studies RLHF with general function approximation for contextual bandits, which is equivalent to the case where $H = 1$. [32, 35] use the Eluder dimension type complexity measures to characterize the sample complexity of online RLHF, which sometimes can be too pessimistic. [176, 177, 229] proposed algorithms for online RLHF with function approximation, but their complexity depends on the trajectory coverability instead of the state coverability

G.2 Technical tools

G.2.1 Uniform convergence with square loss

To prove the uniform convergence results with square loss, we frequently use the following version Freedman's inequality [see e.g., 230].

Lemma G.1 (Freedman's inequality)**.** *Suppose that $Z_{(1)}, \cdots, Z_{(T)}$ is a martingale difference sequence that is adapted to the filtration $(\mathfrak{F}_{(t)})_{t=1}^T$, and $Z_{(t)} \le C$ almost surely for all $t \in [T]$. Then for any $\lambda \in [0, \frac{1}{C}]$, with probability at least $1-\delta$, for all $n \le T$,*

$$\sum_{t=1}^{n} Z_{(t)} \le \lambda \sum_{t=1}^{n} \mathbb{E}\left[(Z_{(t)})^2 \middle| \mathfrak{F}_{(t-1)} \right] + \frac{\log(1/\delta)}{\lambda}.$$

Lemma G.2. *Suppose that $(x_{(1)}, y_{(1)}), \cdots, (x_{(T)}, y_{(T)})$ is a sequence of random variable in $\mathcal{X} \times [0, C]$ that is adapted to the filtration $(\mathfrak{F}_{(t)})_{t=1}^T$, such that there exists a function $F^\star : \mathcal{X} \to [0,1]$ with $F^\star(x_{(t)}) = \mathbb{E}[y_{(t)} | \mathfrak{F}_{(t-1)}, x_{(t)}]$ almost surely. Then for any function $F : \mathcal{X} \to [0, C]$, it holds that with probability at least $1-\delta$, for all $n \in [T]$,*

$$\sum_{t=1}^{n} (F(x_{(t)}) - y_{(t)})^2 - \sum_{t=1}^{n} (F^\star(x_{(t)}) - y_{(t)})^2 \ge \frac{1}{2} \sum_{t=1}^{n} \mathbb{E}\left[(F(x_{(t)}) - F^\star(x_{(t)}))^2 \middle| \mathfrak{F}_{(t-1)} \right] - 10C^2 \log(1/\delta).$$

Conversely, it holds that with probability at least $1-\delta$, for all $n \in [T]$,

$$\sum_{t=1}^{n} (F(x_{(t)}) - y_{(t)})^2 - \sum_{t=1}^{n} (F^\star(x_{(t)}) - y_{(t)})^2 \le 2 \sum_{t=1}^{n} \mathbb{E}\left[(F(x_{(t)}) - F^\star(x_{(t)}))^2 \middle| \mathfrak{F}_{(t-1)} \right] + 5C^2 \log(1/\delta).$$

Proof of Theorem G.2 Denote

$$\begin{aligned} W_{(t)} :=& \ (F(x_{(t)}) - y_{(t)})^2 - (F^\star(x_{(t)}) - y_{(t)})^2 \\ =& \ (F(x_{(t)}) - F^\star(x_{(t)}))^2 + 2 (F(x_{(t)}) - F^\star(x_{(t)})) (F^\star(x_{(t)}) - y_{(t)}) . \end{aligned}$$

Note that

$$\mathbb{E}\left[W_{(t)} \middle| \mathfrak{F}_{(t-1)} \right] = \mathbb{E}\left[(F(x_{(t)}) - F^\star(x_{(t)}))^2 \middle| \mathfrak{F}_{(t-1)} \right],$$

and

$$Z_{(t)} := W_{(t)} - \mathbb{E}\left[W_{(t)} \middle| \mathfrak{F}_{(t-1)} \right] \le W_{(t)} \le C^2.$$

Therefore, using Freedman's inequality (Theorem G.1), for any fixed $\lambda \in [0, \frac{1}{C^2}]$, we have with probability at least $1-\delta$,

$$\sum_{t=1}^{n} Z_{(t)} \le \lambda \sum_{t=1}^{n} \mathbb{E}\left[(Z_{(t)})^2 \middle| \mathfrak{F}_{(t-1)} \right] + \frac{\log(1/\delta)}{\lambda}, \qquad \forall n \in [T].$$

Note that

$$\begin{aligned}\mathbb{E}\left[(Z_{(t)})^2 \middle| \mathfrak{F}_{(t-1)}\right] &\le \mathbb{E}\left[(W_{(t)})^2 \middle| \mathfrak{F}_{(t-1)}\right] \\ &= \mathbb{E}\left[(F(x_{(t)}) - F^\star(x_{(t)}))^4 + 4\,(F(x_{(t)}) - F^\star(x_{(t)}))^2\,(F^\star(x_{(t)}) - y_{(t)})^2 \middle| \mathfrak{F}_{(t-1)}\right] \\ &\le 5C^2\mathbb{E}\left[(F(x_{(t)}) - F^\star(x_{(t)}))^2 \middle| \mathfrak{F}_{(t-1)}\right].\end{aligned}$$

Therefore, by setting $\lambda = \frac{1}{5C^2}$, we get the desired upper bound.

Similarly, for the lower bound, we can apply Freedman's inequality with $(-Z_{(t)})$ to show that for $\lambda = \frac{1}{10C^2}$, with probability at least $1-\delta$,

$$\begin{aligned}-\sum_{t=1}^{n} Z_{(t)} &\le \lambda \sum_{t=1}^{n} \mathbb{E}\left[(Z_{(t)})^2 \middle| \mathfrak{F}_{(t-1)}\right] + \frac{\log(1/\delta)}{\lambda} \\ &\le \frac{1}{2}\sum_{t=1}^{n} \mathbb{E}\left[(F(x_{(t)}) - F^\star(x_{(t)}))^2 \middle| \mathfrak{F}_{(t-1)}\right] + 10C^2\log(1/\delta), \qquad \forall n \in [T].\end{aligned}$$

□

Proposition G.3. *Fix a parameter $\alpha \ge 0$. Under the assumption of Theorem G.2, suppose that $\mathcal{H} \subseteq (\mathcal{X} \to [0, C])$ is a fixed function class, and $F^\sharp \in \mathcal{H}$ satisfies $\left\|F^\star - F^\sharp\right\|_\infty \le \varepsilon_{\mathrm{app}}$. Define*

$$\mathcal{L}_n(F) := \sum_{t=1}^{n} (F(x_{(t)}) - y_{(t)})^2, \qquad \mathcal{E}_n(F) := \sum_{t=1}^{n} \mathbb{E}\left[(F(x_{(t)}) - F^\star(x_{(t)}))^2 \middle| \mathfrak{F}_{(t-1)}\right].$$

Let $\kappa := 15C^2\log(2N(\mathcal{H}, \alpha)/\delta) + 3Cn\alpha + 4n\varepsilon_{\mathrm{app}}^2$. Then the following holds simultaneously with probability at least $1-\delta$:

(1) For each $n \in [T]$,

$$\mathcal{L}_n(F^\sharp) - \inf_{F' \in \mathcal{H}} \mathcal{L}_n(F') \le \kappa.$$

(2) For each $n \in [T]$, for all $F \in \mathcal{H}$,

$$\frac{1}{2}\mathcal{E}_n(F) \le \mathcal{L}_n(F) - \inf_{F' \in \mathcal{H}} \mathcal{L}_n(F') + \kappa.$$

Proof of Theorem G.3 Denote $N := N(\mathcal{H}, \alpha)$. Let $\mathcal{H}_\alpha$ be a minimal α-covering of $\mathcal{H}$. Then applying Theorem G.2 and the union bound, we have with probability at least $1-\delta$, the following holds simultaneously for $n \in [T]$:

(1) For all $F' \in \mathcal{H}_\alpha$, it holds that

$$\frac{1}{2}\mathcal{E}_n(F') \le \mathcal{L}_n(F') - \mathcal{L}_n(F^\star) + 10C^2\log(2N/\delta).$$

(2) It holds that

$$\mathcal{L}_n(F^\sharp) - \mathcal{L}_n(F^\star) \le 2\mathcal{E}_n(F^\sharp) + 5C^2\log(2/\delta).$$

In the following, we condition on the above success event.

By definition, $\mathcal{E}_n(F^\sharp) \le n\varepsilon_{\rm app}^2$, and hence

$$\mathcal{L}_n(F^\star) \ge \mathcal{L}_n(F^\sharp) - 4n\varepsilon_{\rm app}^2 - 5C^2\log(2/\delta). \tag{G.1}$$

Furthermore, for any $F \in \mathcal{H}$, there exists $F' \in \mathcal{H}_\alpha$ with $\|F - F'\|_\infty \le \alpha$, which implies

$$|\mathcal{L}_n(F) - \mathcal{L}_n(F')| \le 2Cn\alpha, \qquad |\mathcal{E}_n(F) - \mathcal{E}_n(F')| \le 2Cn\alpha.$$

Therefore, under the success event, we have

$$\frac{1}{2}\mathcal{E}_n(F) \le \mathcal{L}_n(F) - \mathcal{L}_n(F^\star) + 10C^2\log(2N/\delta) + 3Cn\alpha.$$

holds for arbitrary $F \in \mathcal{F}$. Hence, by (G.1), we have

$$\frac{1}{2}\mathcal{E}_n(F) \le \mathcal{L}_n(F) - \mathcal{L}_n(F^\sharp) + 15C^2\log(2N/\delta) + 3Cn\alpha + 4n\varepsilon_{\rm app}^2.$$

Noting that $\mathcal{L}_n(F^\sharp) \ge \inf_{F'\in\mathcal{H}} \mathcal{L}_n(F')$ completes the proof of (2). Furthermore, using $\mathcal{E}_n(F) \ge 0$, we also have

$$\mathcal{L}_n(F^\sharp) \le \inf_{F'\in\mathcal{H}} \mathcal{L}_n(F') + 15C^2\log(2N/\delta) + 3Cn\alpha + 4n\varepsilon_{\rm app}^2.$$

This completes the proof of (1). □

G.2.2 Uniform convergence with log-loss

We prove the following result, which is a direct extension of the standard MLE guarantee [231].

Proposition G.4. *Suppose that $\{P_\theta(y|x)\}_{\theta\in\Theta} \subseteq (\mathcal{X} \to \Delta(\mathcal{Y}))$ is a class of condition densities parametrized by an abstract parameter class Θ. Without loss of generality, we assume $\mathcal{Y}$ is discrete.*

A α-covering of Θ is a subset $\Theta' \subseteq \Theta$ such that for any $\theta \in \Theta$, there exists $\theta' \in \Theta'$ such that $|\log P_\theta(y|x) - \log P_{\theta'}(y|x)| \le \alpha$ for all $x \in \mathcal{X}, y \in \mathcal{Y}$. We define the covering number of Θ under log-loss as

$$N_{\log}(\Theta, \alpha) := \min\{|\Theta'| : \Theta' \text{ is a } \alpha\text{-covering of } \Theta\}.$$

Suppose that $(x_{(1)}, y_{(1)}), \cdots, (x_{(T)}, y_{(T)})$ is a sequence of random variables adapted to the filtration $(\mathfrak{F}_{(t)})_{t=1}^T$, such that there exists $\theta^\star \in \Theta$ so that $\mathbb{P}(y_{(t)} = \cdot|x_{(t)}, \mathfrak{F}_{(t-1)}) = P_{\theta^\star}(y_{(t)} = \cdot|x_{(t)})$ almost surely for $t \in [T]$. Then it holds that for all $n \in [T]$, for all $\theta \in \Theta$,

$$\sum_{t=1}^n \mathbb{E}\left[D_{\rm H}^2\left(P_\theta(\cdot|x_{(t)}), P_{\theta^\star}(\cdot|x_{(t)})\right)\middle|\mathfrak{F}_{(t-1)}\right] \le -\frac{1}{2}\sum_{t=1}^n \left[\log P_\theta(y_{(t)}|x_{(t)}) - \log P_{\theta^\star}(y_{(t)}|x_{(t)})\right] + \log N_{\log}(\Theta, \alpha) + 2n\alpha.$$

Proof of Theorem G.4 Let $\Theta' \subseteq \Theta$ be a minimal α-covering, and let $N := |\Theta'| = N_{\log}(\Theta, \alpha)$. For each $\theta \in \Theta$, we consider

$$L_{(t)}(\theta) := \log P_\theta(y_{(t)}|x_{(t)}) - \log P_{\theta^\star}(y_{(t)}|x_{(t)}).$$

Then it holds that

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\frac{1}{2}L_{(t)}(\theta)\right)\middle| x_{(t)}, \mathfrak{F}_{(t-1)}\right] &= \mathbb{E}_{y\sim P_{\theta^\star}(\cdot|x_{(t)})}\sqrt{\frac{P_\theta(y|x_{(t)})}{P_{\theta^\star}(y|x_{(t)})}} \\
&= \sum_{y\in\mathcal{Y}} \sqrt{P_\theta(y|x_{(t)})P_{\theta^\star}(y|x_{(t)})} \\
&= 1 - D_{\mathrm{H}}^2\left(P_\theta(\cdot|x_{(t)}), P_{\theta^\star}(\cdot|x_{(t)})\right).
\end{aligned}$$

Therefore, applying Theorem G.5 and using union bound over $\theta \in \Theta'$, we have the following bound: with probability at least $1-\delta$, for any $\theta' \in \Theta'$, $n \in [T]$,

$$\sum_{t=1}^{n} -\log\left[\exp\left(\frac{1}{2}L_{(t)}(\theta')\right)\middle| \mathcal{F}_{(t-1)}\right] \le -\frac{1}{2}\sum_{t=1}^{n} L_{(t)}(\theta') + \log(N/\delta).$$

In the following, we condition on the above event. Fix any $\theta \in \Theta$. Then, there exists $\theta' \in \Theta'$ such that $|\log P_\theta(y|x) - \log P_{\theta'}(y|x)| \le \alpha$ for all $x \in \mathcal{X}, y \in \mathcal{Y}$, and hence $|L_{(t)}(\theta) - L_{(t)}(\theta')| \le \alpha$ almost surely. Therefore, combining the results above and using $\log w \le w - 1$ for $w > 0$, we have

$$\begin{aligned}
\sum_{t=1}^{n} \mathbb{E}\left[D_{\mathrm{H}}^2\left(P_\theta(\cdot|x_{(t)}), P_{\theta^\star}(\cdot|x_{(t)})\right)\middle| \mathfrak{F}_{(t-1)}\right] &\le \sum_{t=1}^{n} -\log\left[\exp\left(\frac{1}{2}L_{(t)}(\theta)\right)\middle| \mathcal{F}_{(t-1)}\right] \\
&\le n\alpha + \sum_{t=1}^{n} -\log\left[\exp\left(\frac{1}{2}L_{(t)}(\theta')\right)\middle| \mathcal{F}_{(t-1)}\right] \\
&\le n\alpha - \frac{1}{2}\sum_{t=1}^{n} L_{(t)}(\theta') + \log(N/\delta) \\
&\le 2n\alpha - \frac{1}{2}\sum_{t=1}^{n} L_{(t)}(\theta) + \log(N/\delta).
\end{aligned}$$

By the arbitrariness of $\theta \in \Theta$, the proof is completed. □

Lemma G.5 (Foster et al. [203, Lemma A.4])**.** *For any sequence of real-valued random variables $X_{(1)}, \cdots, X_{(T)}$ adapted to a filtration $(\mathfrak{F}_{(t)})_{t=1}^{T}$, it holds that with probability at least $1-\delta$, for all $n \in [T]$,*

$$\sum_{t=1}^{n} -\log\left[\exp(-X_{(t)})\middle| \mathcal{F}_{(t-1)}\right] \le \sum_{t=1}^{n} X_{(t)} + \log\left(1/\delta\right).$$

G.3 Missing Proofs in section 8.3.1

G.3.1 Proof of Theorem 8.5

We first present a more detailed statement of the upper bound of Theorem 8.5, as follows.

Theorem G.6. *Let $\delta \in (0,1)$, $\rho \in [0,1)$, and we denote $C_{\mathrm{cov}} = C_{\mathrm{cov}}(\Pi_{\mathcal{F}})$, where $\Pi_{\mathcal{F}} = \{\pi_f : f \in \mathcal{F}\}$ is the policy class induced by $\mathcal{F}$. Suppose that Theorem 8.2 and Theorem 8.4 hold. Then with probability at least $1-\delta$, the output policy $\widehat{\pi}$ of Algorithm 2 satisfies*

$$\begin{aligned} V^{\star}(s_1) - V^{\widehat{\pi}}(s_1) &= \frac{1}{T}\sum_{t=1}^{T}\left(V^{\star}(s_1) - V^{\pi^{(t)}}(s_1)\right) \\ &\le O(H)\cdot\left[\frac{\log(N(\rho)/\delta) + TH^2(\rho+\varepsilon_{\mathrm{app}}^2)}{\lambda} + \frac{\lambda C_{\mathrm{cov}}\log(T)}{T}\right]. \end{aligned}$$

Therefore, for any $\varepsilon \in (0,1)$, with the optimally-tuned parameter λ, it holds that $V^{\star}(s_1) - V^{\widehat{\pi}}(s_1) \le \varepsilon + \widetilde{O}\left(\sqrt{C_{\mathrm{cov}}}H^2\varepsilon_{\mathrm{app}}\right)$, as long as

$$T \ge \widetilde{O}\left(\frac{C_{\mathrm{cov}}H^2}{\varepsilon^2}\cdot\log N(\varepsilon^2/(C_{\mathrm{cov}}H^4))\right).$$

Recall that we let $Q^{\sharp} \in \mathcal{Q}, R^{\sharp} \in \mathcal{R}$ be such that

$$\max_{h\in[H]}\left\|Q_h^{\sharp} - Q_h^{\star}\right\|_{\infty} \le \varepsilon_{\mathrm{app}}, \qquad \max_{h\in[H]}\left\|R_h^{\sharp} - R_h^{\star}\right\|_{\infty} \le \varepsilon_{\mathrm{app}}.$$

For each $t \in [T]$, we write $\mathcal{D}_{(t)}$ to be the dataset maintained by Algorithm 2 at the end of the tth iteration, i.e.,

$$\mathcal{D}_{(t)} = \{(\tau^{(k,h)}, r^{(k,h)})\}_{k\le t, h\in[H]}.$$

We summarize the uniform concentration results for the loss $\mathcal{L}^{\mathsf{BE}}_{\mathcal{D}_{(t)}}$ and $\mathcal{L}^{\mathsf{RM}}_{\mathcal{D}_{(t)}}$ as follows. We note that these concentration bounds are fairly standard (see e.g. Jin, Liu, and Miryoosefi [200]), and for completeness, we present the proof in appendix G.3.3.

Proposition G.7. *Let $\delta \in (0,1), \rho \ge 0$. Suppose that Theorem 8.2 and Theorem 8.4 holds. Then with probability at least $1-\delta$, for all $t \in [T], f \in \mathcal{F}, R \in \mathcal{R}$, it holds that*

$$\begin{aligned} &\frac{1}{2}\sum_{k\le t}\sum_{h=1}^{H}\mathbb{E}^{\pi^{(k)}\circ_h\pi_{\mathrm{ref}}}\left(R(\tau) - R^{\star}(\tau)\right)^2 \le \mathcal{L}^{\mathsf{RM}}_{\mathcal{D}^{(t)}}(R) - \mathcal{L}^{\mathsf{RM}}_{\mathcal{D}^{(t)}}(R^{\sharp}) + H\kappa, \\ &\frac{1}{2}\sum_{k\le t}\sum_{h=1}^{H}\mathbb{E}^{\pi^{(k)}}\left(f_h(s_h,a_h) - [\mathcal{T}_{R,h}f_{h+1}](s_h,a_h)\right)^2 \le \mathcal{L}^{\mathsf{BE}}_{\mathcal{D}^{(t)}}(f;R) - \mathcal{L}^{\mathsf{BE}}_{\mathcal{D}^{(t)}}(Q^{\sharp};R^{\sharp}) + H\kappa, \end{aligned}$$

where

$$\kappa = C\left(\log N(\rho) + \log(H/\delta) + TH^2(\varepsilon_{\mathrm{app}}^2 + \rho)\right),$$

and $C > 0$ is an absolute constant.

Performance difference decomposition Denote $V^\sharp(s_1) := \max_{a\in\mathcal{A}} Q^\sharp(s_1,a)$. Then it is clear that $\left|V^\sharp(s_1) - V^\star(s_1)\right| \le \varepsilon_{\mathrm{app}}$. Therefore, for any $t \in [T]$, by optimism (the definition of $(f^{(t)}, R^{(t)})$), it holds that

$$\begin{aligned}
&V^\star(s_1) - \varepsilon_{\mathrm{app}} \le V^\sharp(s_1) \\
&= V^\sharp(s_1) - \frac{\mathcal{L}^{\mathsf{BE}}_{\mathcal{D}^{(t-1)}}(Q^\sharp, R^\sharp) + \mathcal{L}^{\mathsf{RM}}_{\mathcal{D}^{(t-1)}}(R^\sharp)}{\lambda} + \frac{\mathcal{L}^{\mathsf{BE}}_{\mathcal{D}^{(t-1)}}(Q^\sharp, R^\sharp) + \mathcal{L}^{\mathsf{RM}}_{\mathcal{D}^{(t-1)}}(R^\sharp)}{\lambda} \\
&\le f_1^{(t)}(s_1, \pi^{(t)}) - \frac{\mathcal{L}^{\mathsf{BE}}_{\mathcal{D}^{(t-1)}}(f^{(t)}, R^{(t)}) + \mathcal{L}^{\mathsf{RM}}_{\mathcal{D}^{(t-1)}}(R^{(t)})}{\lambda} + \frac{\mathcal{L}^{\mathsf{BE}}_{\mathcal{D}^{(t-1)}}(Q^\sharp, R^\sharp) + \mathcal{L}^{\mathsf{RM}}_{\mathcal{D}^{(t-1)}}(R^\sharp)}{\lambda}.
\end{aligned}$$

Furthermore, by the standard performance difference lemma [190], it holds that

$$f_1^{(t)}(s_1, \pi^{(t)}) - V^{\pi^{(t)}}(s_1) = \sum_{h=1}^{H} \mathbb{E}^{\pi^{(t)}} \left[f_h^{(t)}(s_h, a_h) - [\mathcal{T}_h^\star f_{h+1}^{(t)}](s_h, a_h) \right]. \tag{G.2}$$

Based on (G.2), the existing approaches (with per-step reward feedback) bound the expectation of the Bellman error $e_h^{(t)}(s_h,a_h) := f_h^{(t)}(s_h,a_h) - [\mathcal{T}_h^\star f_{h+1}](s_h,a_h)$ through various arguments (e.g., eluder argument [200] and coverability argument [153]). However, in outcome reward model, it is possible that $\mathbb{E}^{\pi^{(t)}} \left[f_h^{(t)}(s_h, a_h) - [\mathcal{T}_h^\star f_{h+1}](s_h, a_h) \right]$ is large even when the sub-optimality of $\pi^{(t)}$ is small, because the outcome reward is invariant under shifting of the ground-truth reward function $R^\star$.

Therefore, we consider the following refined decomposition:

$$\begin{aligned}
f_1^{(t)}(s_1) - V^{\pi^{(t)}}(s_1) = &\sum_{h=1}^{H} \mathbb{E}^{\pi^{(t)}} \left[f_h^{(t)}(s_h, a_h) - [\mathcal{T}_{R^{(t)}} f_{h+1}^{(t)}](s_h, a_h) \right] \\
&+ \mathbb{E}^{\pi^{(t)}} \left[\sum_{h=1}^{H} R_h^{(t)}(s_h, a_h) - \sum_{h=1}^{H} R_h^\star(s_h, a_h) \right].
\end{aligned} \tag{G.3}$$

Coverability argument Following the coverability argument of Xie et al. [153], we have the following upper bound on the expected Bellman errors.

Proposition G.8. *Denote $e_h^{(t)} := f_h^{(t)} - \mathcal{T}_h^\star f_{h+1}$. Then, for each $h \in [H]$, it holds that*

$$\sum_{t=1}^{T} \mathbb{E}^{\pi^{(t)}} \left| e_h^{(t)}(s_h, a_h) \right| \le \sqrt{2C_{\mathrm{cov}} \log\left(1 + \frac{C_{\mathrm{cov}} T}{\kappa}\right) \cdot \left[2T\kappa + \sum_{1 \le k < t \le T} \mathbb{E}^{\pi^{(k)}} e_h^{(t)}(s_h, a_h)^2 \right]}.$$

Following the ideas of Jia, Rakhlin, and Xie [199], we prove the following upper bound on the reward errors.

Proposition G.9. *It holds that*

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}^{\pi^{(t)}} \left| R^{(t)}(\tau) - R^\star(\tau) \right| \\
&\le \sqrt{8HC_{\mathrm{cov}} \log\left(1 + \frac{C_{\mathrm{cov}} T}{\kappa}\right) \cdot \left[HT\kappa + \sum_{1 \le k < t \le T} \sum_{h=1}^{H} \mathbb{E}^{\pi^{(k)} \circ_h \pi_{\mathrm{ref}}} \left(R^{(t)}(\tau) - R^\star(\tau) \right)^2 \right]}.
\end{aligned}$$

Based on the results above, we finalize the proof of Theorem 8.5.

Proof of Theorem 8.5 By optimism and the decomposition (G.3), it holds that for $t \in [T]$,

$$V^\star(s_1) - V^{\pi^{(t)}}(s_1) - \varepsilon_{\mathrm{app}} \le \sum_{h=1}^{H} \mathbb{E}^{\pi^{(t)}}\left[e_h^{(t)}(s_h, a_h)\right] - \frac{\mathcal{L}^{\mathsf{BE}}_{\mathcal{D}^{(t-1)}}(f^{(t)}, R^{(t)}) - \mathcal{L}^{\mathsf{BE}}_{\mathcal{D}^{(t-1)}}(Q^\sharp, R^\sharp)}{\lambda} + \mathbb{E}^{\pi^{(t)}}\left[R^{(t)}(\tau) - R^\star(\tau)\right] - \frac{\mathcal{L}^{\mathsf{RM}}_{\mathcal{D}^{(t-1)}}(R^{(t)}) - \mathcal{L}^{\mathsf{RM}}_{\mathcal{D}^{(t-1)}}(R^\sharp)}{\lambda}.$$

Then, under the success event of Theorem G.7, we have

$$\begin{aligned} & V^\star(s_1) - V^{\pi^{(t)}}(s_1) - \varepsilon_{\mathrm{app}} - \frac{2H\kappa}{\lambda} \\ \le & \sum_{h=1}^{H}\left(\mathbb{E}^{\pi^{(t)}}\left[e_h^{(t)}(s_h, a_h)\right] - \frac{1}{2\lambda}\sum_{k<t}\mathbb{E}^{\pi^{(k)}} e_h^{(t)}(s_h, a_h)^2\right) \\ & + \mathbb{E}^{\pi^{(t)}}\left[R^{(t)}(\tau) - R^\star(\tau)\right] - \frac{1}{2\lambda}\sum_{k<t}\sum_{h=1}^{H}\mathbb{E}^{\pi^{(k)}\circ_h\pi_{\mathrm{ref}}}\left(R^{(t)}(\tau) - R^\star(\tau)\right)^2. \end{aligned} \tag{G.4}$$

Applying Theorem G.8 and Cauchy inequality gives for all $h \in [H]$,

$$\sum_{t=1}^{T}\mathbb{E}^{\pi^{(t)}}\left|e_h^{(t)}(s_h, a_h)\right| \le \lambda C_{\mathrm{cov}}\log\left(1 + \frac{C_{\mathrm{cov}}T}{\kappa}\right) + \frac{1}{2\lambda}\left[2T\kappa + \sum_{1\le k<t\le T}\mathbb{E}^{\pi^{(k)}} e_h^{(t)}(s_h, a_h)^2\right],$$

and similarly, applying Theorem G.9 and Cauchy inequality gives

$$\begin{aligned} & \sum_{t=1}^{T}\mathbb{E}^{\pi^{(t)}}\left|R^{(t)}(\tau) - R^\star(\tau)\right| \\ \le & 4\lambda H C_{\mathrm{cov}}\log\left(1 + \frac{C_{\mathrm{cov}}T}{\kappa}\right) + \frac{1}{2\lambda}\left[HT\kappa + \sum_{1\le k<t\le T}\sum_{h=1}^{H}\mathbb{E}^{\pi^{(k)}\circ_h\pi_{\mathrm{ref}}}\left(R^{(t)}(\tau) - R^\star(\tau)\right)^2\right]. \end{aligned}$$

Therefore, we take summation of (G.4) over $t \in [T]$, and combining the inequalities above gives

$$\sum_{t=1}^{T}\left(V^\star(s_1) - V^{\pi^{(t)}}(s_1)\right) \le T\varepsilon_{\mathrm{app}} + \frac{4TH\kappa}{\lambda} + 5\lambda H C_{\mathrm{cov}}\log\left(1 + \frac{C_{\mathrm{cov}}T}{\kappa}\right).$$

This is the desired upper bound. □

G.3.2 Proof of Theorem G.8 and Theorem G.9

The following proposition is an generalized version of the results in Xie et al. [153, Appendix D]. For proof, see e.g. Chen et al. [232].

Proposition G.10 ([153]). *Let $C \ge 1$ be a parameter. Suppose that $p_{(1)}, \cdots, p_{(T)}$ is a sequence of distributions over $\mathcal{X}$, and there exists $\mu \in \Delta(\mathcal{X})$ such that $p_{(t)}(x)/\mu(x) \le C$ for all $x \in \mathcal{X}$,*

$t \in [T]$. *Then for any sequence* $\psi_{(1)}, \cdots, \psi_{(T)}$ *of functions* $\mathcal{X} \to [0,1]$ *and constant* $B \geq 1$, *it holds that*

$$\sum_{t=1}^{T} \mathbb{E}_{x \sim p_{(t)}} \psi_{(t)}(x) \leq \sqrt{2C \log\left(1 + \frac{CT}{B}\right) \left[2TB + \sum_{t=1}^{T} \sum_{k<t} \mathbb{E}_{x \sim p_{(k)}} \psi_{(t)}(x)^2\right]}.$$

As a warm-up, we prove Theorem G.8 by directly invoking Theorem G.10.

Proof of Theorem G.8 Fix a $h \in [H]$. To apply Theorem G.10, we consider $\mathcal{X} = \mathcal{S} \times \mathcal{A}$, and define

$$p_{(t)} := \mathbb{P}^{\pi^{(t)}}\left((s_h, a_h) = \cdot\right) \in \Delta(\mathcal{S} \times \mathcal{A}), \qquad \psi_{(t)} := \left|e_h^{(t)}\right| \in (\mathcal{S} \times \mathcal{A} \to [0,1]).$$

By the definition of coverability (Theorem 8.1), for $C = C_{\mathrm{cov}}(\Pi)$, there exists $\mu \in \Delta(\mathcal{S} \times \mathcal{A})$ such that $p_{(t)}(s,a)/\mu(s,a) \leq C$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. Therefore, applying Theorem G.10 with $B = \kappa \geq 1$ gives the desired upper bound. □

Next, we proceed to prove Theorem G.9. Our key proof technique is summarized in the following proposition, which is inspired by the (rather sophisticated) analysis of Jia, Rakhlin, and Xie [199].

Proposition G.11. *Recall that for any* $D = (D_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R})$, *we denote*

$$D(\tau) = \sum_{h=1}^{H} D_h(s_h, a_h), \qquad \forall \tau = (s_1, a_1, \cdots, s_H, a_H) \in (\mathcal{S} \times \mathcal{A})^H.$$

Fix a Markov policy π_{ref}, *we denote* $\overline{D}_1(s) = \overline{D}_{H+1}(s) = 0$, *and*

$$\overline{D}_h(s) := \mathbb{E}^{\pi_{\mathrm{ref}}}\left[\sum_{\ell=h}^{H} D_\ell(s_\ell, a_\ell) \middle| s_h = s\right], \qquad \forall 1 < h \leq H, s \in \mathcal{S}.$$

Then for any policy π, *it holds that*

$$\sum_{h=1}^{H} \mathbb{E}^{\pi}\left(D_h(s_h, a_h) + \overline{D}_{h+1}(s_{h+1}) - \overline{D}_h(s_h)\right)^2 \leq 4 \sum_{h=1}^{H} \mathbb{E}^{\pi \circ_h \pi_{\mathrm{ref}}} D(\tau)^2.$$

The proof of Theorem G.11 is deferred to the end of this subsection. With Theorem G.11, we prove Theorem G.9 as follows.

Proof of Theorem G.9 To apply Theorem G.11, for each $t \in [T]$, we consider

$$\Delta_{h^{(t)}}(s) := \mathbb{E}^{\pi_{\mathrm{ref}}}\left[\sum_{\ell=h}^{H} R_\ell^{(t)}(s_\ell, a_\ell) - \sum_{\ell=h}^{H} R_\ell^{\star}(s_\ell, a_\ell) \middle| s_h = s\right], \quad \forall h = 2, \cdots, H, s \in \mathcal{S},$$

Then, by Theorem G.11, it holds that for any policy π,

$$\begin{aligned}&\sum_{h=1}^{H}\mathbb{E}^{\pi}\left(R_h^{(t)}(s_h,a_h)-R_h^{\star}(s_h,a_h)+\Delta_{h+1^{(t)}}(s_{h+1})-\Delta_{h^{(t)}}(s_h)\right)^2\\&\qquad\le 4\sum_{h=1}^{H}\mathbb{E}^{\pi\circ_h\pi_{\mathrm{ref}}}\left(R^{(t)}(\tau)-R^{\star}(\tau)\right)^2.\end{aligned}\tag{G.5}$$

Furthermore,

$$\begin{aligned}\mathbb{E}^{\pi}\left|R^{(t)}(\tau)-R^{\star}(\tau)\right|&=\mathbb{E}^{\pi}\left|\sum_{h=1}^{H}\left[R_h^{(t)}(s_h,a_h)-R_h^{\star}(s_h,a_h)+\Delta_{h+1^{(t)}}(s_{h+1})-\Delta_{h^{(t)}}(s_h)\right]\right|\\&\le\sum_{h=1}^{H}\mathbb{E}^{\pi}\left|R_h^{(t)}(s_h,a_h)-R_h^{\star}(s_h,a_h)+\Delta_{h+1^{(t)}}(s_{h+1})-\Delta_{h^{(t)}}(s_h)\right|.\end{aligned}\tag{G.6}$$

Therefore, to apply Theorem G.10, we consider the space $\mathcal{X}=\mathcal{S}\times\mathcal{A}\times\mathcal{S}$, and for each $h\in[H]$, we define

$$\begin{aligned}p_{(t)h}&:=\mathbb{P}^{\pi^{(t)}}\left((s_h,a_h,s_{h+1})=\cdot\right)\in\Delta(\mathcal{S}\times\mathcal{A}\times\mathcal{S}),\\\psi_{(t)h}(s,a,s')&:=\left|R_h^{(t)}(s,a)-R_h^{\star}(s,a)+\Delta_{h+1^{(t)}}(s')-\Delta_{h^{(t)}}(s)\right|.\end{aligned}$$

Note that for any h, we have $\psi_{(t)h}:\mathcal{S}\times\mathcal{A}\times\mathcal{S}\to[0,1]$. Further, let $\mu_h\in\Delta(\mathcal{S}\times\mathcal{A})$ be the distribution such that $\|d_h^{\pi}/\mu_h\|_{\infty}\le C_{\mathrm{cov}}$ for any policy π. Then we can consider the distribution $\overline{\mu}_h\in\Delta(\mathcal{S}\times\mathcal{A}\times\mathcal{S})$ given by $\overline{\mu}_h(s,a,s')=\mu_h(s,a)\mathbb{T}_h(s'|s,a)$. Then it holds that

$$\left\|\frac{p_{(t)h}}{\overline{\mu}_h}\right\|_{\infty}=\sup_{s,a,s'}\frac{p_{(t)h}(s,a,s')}{\overline{\mu}_h(s,a,s')}=\sup_{s,a,s'}\frac{d_h^{\pi^{(t)}}(s,a)\mathbb{T}_h(s'|s,a)}{\mu_h(s,a)\mathbb{T}_h(s'|s,a)}\le C_{\mathrm{cov}}.$$

Therefore, for $h\in[H]$, applying Theorem G.10 on the sequence $(p_{(1)h},\cdots,p_{(T)h})$ and $(\psi_{(1)h},\cdots,\psi_{(T)h})$ gives

$$\sum_{t=1}^{T}\mathbb{E}_{x\sim p_{h^{(t)}}}\psi_{h^{(t)}}(x)\le\sqrt{2C_{\mathrm{cov}}\log\left(1+\frac{C_{\mathrm{cov}}T}{\kappa}\right)\left[2T\kappa+\sum_{t=1}^{T}\sum_{k<t}\mathbb{E}_{x\sim p_{h^{(k)}}}\psi_{h^{(t)}}(x)^2\right]}.$$

To conclude, we combine the inequalities above and bound

$$\begin{aligned}&\sum_{t=1}^{T}\mathbb{E}^{\pi^{(t)}}\left|R^{(t)}(\tau)-R^{\star}(\tau)\right|\\\le&\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}^{\pi^{(t)}}\left|R_h^{(t)}(s_h,a_h)-R_h^{\star}(s_h,a_h)+\Delta_{h+1^{(t)}}(s_{h+1})-\Delta_{h^{(t)}}(s_h)\right|\\=&\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{x\sim p_{h^{(t)}}}\psi_{h^{(t)}}(x)=\sum_{h=1}^{H}\sum_{t=1}^{T}\mathbb{E}_{x\sim p_{h^{(t)}}}\psi_{h^{(t)}}(x)\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{h=1}^{H} \sqrt{2C_{\mathrm{cov}} \log\left(1+\frac{C_{\mathrm{cov}}T}{\kappa}\right)\left[2T\kappa + \sum_{t=1}^{T}\sum_{k<t} \mathbb{E}_{x\sim p_{h^{(k)}}} \psi_{h^{(t)}}(x)^2\right]} \\
&\leq \sqrt{2HC_{\mathrm{cov}} \log\left(1+\frac{C_{\mathrm{cov}}T}{\kappa}\right)\left[2TH\kappa + \sum_{h=1}^{H}\sum_{t=1}^{T}\sum_{k<t} \mathbb{E}_{x\sim p_{h^{(k)}}} \psi_{h^{(t)}}(x)^2\right]} \\
&\leq \sqrt{8HC_{\mathrm{cov}} \log\left(1+\frac{C_{\mathrm{cov}}T}{\kappa}\right)\cdot\left[HT\kappa + \sum_{1\leq k<t\leq T}\sum_{h=1}^{H} \mathbb{E}^{\pi^{(k)}\circ_h \pi_{\mathrm{ref}}}\left(R^{(t)}(\tau)-R^{\star}(\tau)\right)^2\right]},
\end{aligned}
$$

where the first inequality follows from (G.6), the second last line follows from Cauchy inequality, and last inequality follows from the definition of $(p_{h^{(t)}}, \psi_{h^{(t)}})$ and (G.5). □

Proof of Theorem G.11 For $\tau = (s_1, a_1, \cdots, s_H, a_H) \in (\mathcal{S}\times\mathcal{A})^H$, we denote $\tau_h = (s_1, a_1, \cdots, s_h, a_h, s_{h+1})$ to be the prefix sequence of τ for each $h \in [H]$. Then, we note that

$$
\begin{aligned}
\mathbb{E}^{\pi\circ_h\pi_{\mathrm{ref}}}\left[D(\tau)\middle|\,\tau_h\right] &= \sum_{\ell=1}^{h} D_\ell(s_\ell,a_\ell) + \mathbb{E}^{\pi\circ_h\pi_{\mathrm{ref}}}\left[\sum_{\ell=h+1}^{H} D_\ell(s_\ell,a_\ell)\middle|\,\tau_h\right] \\
&= \sum_{\ell=1}^{h} D_\ell(s_\ell,a_\ell) + \overline{D}_{h+1}(s_{h+1}),
\end{aligned}
$$

because the policy $\pi\circ_h\pi_{\mathrm{ref}}$ executes the Markov policy π_{ref} starting at the $(h+1)$-th step. Therefore, it holds that

$$
\begin{aligned}
\mathbb{E}_{\tau_h\sim\pi}\left(\sum_{\ell=1}^{h} D_\ell(s_\ell,a_\ell) + \overline{D}_{h+1}(s_{h+1})\right)^2 &= \mathbb{E}_{\tau_h\sim\pi\circ_h\pi_{\mathrm{ref}}}\left(\mathbb{E}^{\pi\circ_h\pi_{\mathrm{ref}}}\left[D(\tau)\middle|\,\tau_h\right]\right)^2 \\
&\leq \mathbb{E}_{\tau\sim\pi\circ_h\pi_{\mathrm{ref}}} D(\tau)^2 = \mathbb{E}^{\pi\circ_h\pi_{\mathrm{ref}}} D(\tau)^2,
\end{aligned}
$$

where the first equality follows from the fact that the policy $\pi\circ_h\pi_{\mathrm{ref}}$ executes π for the first h steps. Therefore, for $h>1$, it holds that

$$
\begin{aligned}
&\mathbb{E}^{\pi}\left(D_h(s_h,a_h) + \overline{D}_{h+1}(s_{h+1}) - \overline{D}_h(s_h)\right)^2 \\
&= \mathbb{E}_{\tau_h\sim\pi}\left(\sum_{\ell=1}^{h} D_\ell(s_\ell,a_\ell) + \overline{D}_{h+1}(s_{h+1}) - \sum_{\ell=1}^{h-1} D_\ell(s_\ell,a_\ell) - \overline{D}_h(s_h)\right)^2 \\
&\leq 2\mathbb{E}_{\tau_h\sim\pi}\left(\sum_{\ell=1}^{h} D_\ell(s_\ell,a_\ell) + \overline{D}_{h+1}(s_{h+1})\right)^2 + \mathbb{E}_{\tau_{h-1}\sim\pi}\left(\sum_{\ell=1}^{h-1} D_\ell(s_\ell,a_\ell) + \overline{D}_h(s_h)\right)^2 \\
&\leq 2\mathbb{E}^{\pi\circ_h\pi_{\mathrm{ref}}} D(\tau)^2 + \mathbb{E}^{\pi\circ_{h-1}\pi_{\mathrm{ref}}} D(\tau)^2.
\end{aligned}
$$

For $h=1$, because $\overline{D}_1(s)=0$, we already have

$$
\mathbb{E}^{\pi}\left(D_1(s_1,a_1) + \overline{D}_2(s_2) - \overline{D}_1(s_1)\right)^2 = \mathbb{E}^{\pi}\left(D_1(s_1,a_1) + \overline{D}_2(s_2)\right)^2 \leq \mathbb{E}^{\pi\circ_1\pi_{\mathrm{ref}}} D(\tau)^2.
$$

Taking summation over $h\in[H]$ completes the proof. □

G.3.3 Proof of Theorem G.7

We prove Theorem G.7 in Theorem G.12 and Theorem G.13 separately. Recall again that under Theorem 8.2, the function $Q^\sharp \in \mathcal{F}, R^\sharp \in \mathcal{R}$ satisfy

$$\max_{h\in[H]} \left\| Q_h^\sharp - Q_h^\star \right\|_\infty \le \varepsilon_{\mathrm{app}}, \qquad \max_{h\in[H]} \left\| R_h^\sharp - R_h^\star \right\|_\infty \le \varepsilon_{\mathrm{app}}.$$

Lemma G.12. *Under Theorem 8.2, with probability at least $1-\delta$, for any $t \in [T]$, for all $R \in \mathcal{R}$, it holds that*

$$\frac{1}{2}\sum_{k\le t}\sum_{h=1}^{H} \mathbb{E}^{\pi^{(k)}\circ_h \pi_{\mathrm{ref}}} \left(R(\tau) - R^\star(\tau)\right)^2 \le \mathcal{L}^{\mathsf{RM}}_{\mathcal{D}^{(t)}}(R) - \mathcal{L}^{\mathsf{RM}}_{\mathcal{D}^{(t)}}(R^\sharp) + 15H\log N(\rho) + 15(2/\delta) + 2TH^3\varepsilon_{\mathrm{app}}^2 + 4TH^2\rho.$$

Proof of Theorem G.12 To apply Theorem G.3, we consider the whole history

$$\left\{(\tau^{(t,h)}, r^{(t,h)})\right\}_{t\in[T],h\in[H]}.$$

generated by executing Algorithm 2, and recall that $\mathcal{D}_{(t-1)} = \left\{(\tau^{(k,h)}, r^{(k,h)})\right\}_{k<t,h\in[H]}$ is the history up to the t-th iteration. Note that $(\tau^{(t,1)}, r^{(t,1)}), \cdots, (\tau^{(t,H)}, r^{(t,H)})$ are pairwise independent given $\mathcal{D}_{(t-1)}$, with

$$\tau^{(t,h)} \sim \pi^{(t)} \circ_h \pi_{\mathrm{ref}}, \qquad \mathbb{E}\left[r^{(t,h)} \middle| \mathcal{D}_{(t-1)}, \tau^{(t,h)}\right] = R^\star(\tau^{(t,h)}).$$

Also note that $r \in [0,1]$ almost surely, and we regard $\mathcal{R} \subseteq ((\mathcal{S}\times\mathcal{A})^H \to [0,1])$, and $R^\sharp \in \mathcal{R}$ satisfies $\left|R^\sharp(\tau) - R^\star(\tau)\right| \le H\varepsilon_{\mathrm{app}}$ for all $\tau \in (\mathcal{S}\times\mathcal{A})^H$.

Therefore, applying Theorem G.3 on the function class $\mathcal{R}$ and the sequence $\left\{(\tau^{(t,h)}, r^{(t,h)})\right\}_{t\in[T],h\in[0,H]}$ gives that with probability at least $1-\delta$, for all $R \in \mathcal{R}$, $t \in [T]$, it holds that

$$\frac{1}{2}\sum_{k=1}^{t}\sum_{h=1}^{H} \mathbb{E}^{\pi^{(k)}\circ_h \pi_{\mathrm{ref}}} \left(R(\tau) - R^\star(\tau)\right)^2 = \frac{1}{2}\sum_{k=1}^{t}\sum_{h=1}^{H} \mathbb{E}\left[\left(R(\tau^{(k,h)}) - R^\star(\tau^{(k,h)})\right)^2 \middle| \mathcal{D}_{(k-1)}\right]$$
$$\le \mathcal{L}^{\mathsf{RM}}_{\mathcal{D}^{(t)}}(R) - \mathcal{L}^{\mathsf{RM}}_{\mathcal{D}^{(t)}}(R^\sharp) + 15\log(2N(\mathcal{R}, H\rho)/\delta) + 2TH^3\varepsilon_{\mathrm{app}}^2 + 4TH^2\rho.$$

Finally, we note that $\log N(\mathcal{R}, H\rho) \le \sum_{h=1}^{H} \log N(\mathcal{R}_h, \rho) \le H\log N(\rho)$. This gives the desired upper bound. □

Similarly, we prove Theorem G.7 (2) as follows, following Jin, Liu, and Miryoosefi [200].

Lemma G.13. *Fix $h \in [H]$ and $\delta \in (0,1)$, $\rho \ge 0$. Suppose that Theorem 8.2 and Theorem 8.4 holds. Then with probability at least $1-\delta$, the following holds:*

(1) For each $t \in [T]$,

$$\mathcal{E}_{\mathcal{D}^{(t)},h}(Q_h^\sharp, Q_{h+1}^\sharp; R^\sharp) - \inf_{g_h\in\mathcal{G}_h} \mathcal{E}_{\mathcal{D}^{(t)},h}(g_h, Q_{h+1}^\sharp; R^\sharp) \le O\left(TH\varepsilon_{\mathrm{app}}^2 + TH\rho + \log(N(\rho)/\delta)\right).$$

(2) For each $t \in [T]$, for all $f_h \in \mathcal{F}_h$, $f_{h+1} \in \mathcal{F}_{h+1}$, and $R_h \in \mathcal{R}_h$,

$$\frac{1}{2}\sum_{k=1}^{t} \mathbb{E}^{\pi^{(k)}} \left(f_h(s_h,a_h) - [\mathcal{T}_{R,h} f_{h+1}](s_h,a_h)\right)^2$$
$$\le \mathcal{E}_{\mathcal{D}^{(t)},h}(f_h, f_{h+1}; R_h) - \inf_{g_h\in\mathcal{G}_h} \mathcal{E}_{\mathcal{D}^{(t)},h}(g_h, f_{h+1}; R_h) + O\left(TH\varepsilon_{\mathrm{app}}^2 + TH\rho + \log(N(\rho)/\delta)\right),$$

where we use $O(\cdot)$ to hide absolute constant for simplicity.

Proof of Theorem G.13 Fix $h \in [H]$ and denote $N := N(\rho)$. We let $\mathcal{F}'_{h+1}$ be a minimal ρ-covering of $\mathcal{F}_{h+1}$, and let $\mathcal{R}'_h$ be a minimal ρ-covering of $\mathcal{R}_h$. By definition, $|\mathcal{F}'_{h+1}| \le N, |\mathcal{R}'_h| \le N$.

In the following, we adopt the notation of the proof of Theorem G.12. Recall that conditional on $\mathcal{D}_{(t-1)}$,

$$\tau^{(t,\ell)} = (s_1^{(t,\ell)}, a_1^{(t,\ell)}, \cdots, s_H^{(t,\ell)}, a_H^{(t,\ell)}) \sim \pi^{(t)} \circ_\ell \pi_{\mathrm{ref}},$$

and $\tau^{(t,1)}, \cdots, \tau^{(t,H)}$ are independent conditional on $\mathcal{D}_{(t-1)}$. For simplicity, we denote $x_{(t,\ell)} := (s_h^{(t,\ell)}, a_h^{(t,\ell)})$.

Fix $f_{h+1} \in \mathcal{F}'_{h+1} \cup \{Q^\sharp_{h+1}\}$ and $R_h \in \mathcal{R}'_h \cup \{R^\sharp_h\}$, we consider

$$y_{(t,\ell)} := f_{h+1}(s_{h+1}^{(t,\ell)}) + R_h(s_h^{(t,\ell)}, a_h^{(t,\ell)}).$$

and it holds that

$$\mathbb{E}\left[y_{(t,\ell)} \middle| \mathcal{D}_{(t-1)}, x_{(t,\ell)}\right] = [\mathcal{T}_{R,h} f_{h+1}](x_{(t,\ell)}).$$

Then, for any $g_h \in \mathcal{G}_h$, it holds that

$$\mathcal{E}_{\mathcal{D}^{(t)},h}(g_h, f_{h+1}; R_h) = \sum_{k=1}^{t}\sum_{\ell=1}^{H} \left(g_h(x_{(t,\ell)}) - y_{(t,\ell)}\right)^2,$$

and we also have

$$\sum_{k=1}^{t}\sum_{\ell=1}^{H} \mathbb{E}^{\pi^{(k)}\circ_\ell \pi_{\mathrm{ref}}} \left(g_h(s_h,a_h) - [\mathcal{T}_{R,h} f_{h+1}](s_h,a_h)\right)^2 = \sum_{k=1}^{t}\sum_{\ell=1}^{H} \mathbb{E}\left[\left(g_h(x_{(k,\ell)}) - [\mathcal{T}_{R,h} f_{h+1}](x_{(k,\ell)})\right)^2 \middle| \mathcal{D}_{(k-1)}\right]$$

Then, applying Theorem G.3 with the function class $\mathcal{H} = \mathcal{G}_h$ yields that with probability at least $1 - \frac{\delta}{2N}$, the following holds:

(a) For each $t \in [T]$, for any $g_h \in \mathcal{G}_h$,

$$\begin{aligned}
&\frac{1}{2}\sum_{k=1}^{t} \mathbb{E}^{\pi^{(k)}} \left(g_h(s_h,a_h) - [\mathcal{T}_{R,h} f_{h+1}](s_h,a_h)\right)^2 \\
\le &\mathcal{E}_{\mathcal{D}^{(t)},h}(g_h, f_{h+1}; R_h) - \inf_{g'_h \in \mathcal{G}_h} \mathcal{E}_{\mathcal{D}^{(t)},h}(g'_h, f_{h+1}; R_h) + O\left(\log(N/\delta) + TH\varepsilon_{\mathrm{app}}^2 + TH\rho\right).
\end{aligned}$$

(b) When $f_{h+1} = Q^\sharp_{h+1}$ and $R_h = R^\sharp_h$, the function $Q^\sharp_h \in \mathcal{F}_h \subseteq \mathcal{G}_h$ satisfies the inequality $\|Q^\sharp_h - \mathcal{T}_{R^\sharp,h} Q^\sharp_{h+1}\|_\infty \le 3\varepsilon_{\mathrm{app}}$, and thus

$$\mathcal{E}_{\mathcal{D}^{(t)},h}(Q^\sharp_h, Q^\sharp_{h+1}; R^\sharp_h) \le \inf_{g_h \in \mathcal{G}_h} \mathcal{E}_{\mathcal{D}^{(t)},h}(g_h, Q^\sharp_{h+1}; R^\sharp_h) + O\left(\log(N/\delta) + TH\varepsilon_{\mathrm{app}}^2 + TH\rho\right).$$

Therefore, taking the union bound, we know that the inequalities (a) and (b) above hold simultaneously with probability at least $1-\delta$ for all $f_{h+1} \in \mathcal{F}'_{h+1} \cup \{Q^\sharp_{h+1}\}$ and $R_h \in \mathcal{R}'_h \cup \{R^\sharp_h\}$. In particular, we have completed the proof of (1).

To prove (2), we only need to note that $\mathcal{G}_h \subseteq \mathcal{F}_h$, and for any $f_{h+1} \in \mathcal{F}_{h+1}, R_h \in \mathcal{R}_h$, there exists $f'_{h+1} \in \mathcal{F}'_{h+1}, R'_h \in \mathcal{R}'_h$ such that $\|f_{h+1} - f'_{h+1}\|_\infty \le \rho$, $\|R_h - R'_h\|_\infty \le \rho$. Therefore, by the standard covering argument and the fact that $\pi^{(k)} \circ_H \pi_{\mathrm{ref}} = \pi^{(k)}$, we have also shown (2). □

G.4 Proof of Theorem 8.8

In this section, we provide the proof of Theorem 8.8, which is a direct adaption of the proof of Theorem 8.5 in appendix G.3. We first present a more detailed statement of the upper bound (with any parameter $\lambda > 0$).

Theorem G.14. *Suppose that Theorem 8.2 holds. Then with probability at least $1-\delta$, Algorithm 3 achieves*

$$\frac{1}{T}\sum_{t=1}^{T}\left(V^{\star}(s_1^{(t)}) - V^{\pi^{(t)}}(s_1^{(t)})\right) \le \varepsilon_{\mathrm{app}} + O(1)\cdot\left[\frac{H^3\log(N_{\mathcal{F},T}/\delta) + T\varepsilon_{\mathrm{app}}^2}{\lambda} + \frac{\lambda H C'_{\mathrm{cov}}(\Pi)}{T}\right],$$

We also work with a slightly relaxed version of Theorem 8.7.

Assumption G.15. *Under any policy π, for each $h \in [H]$, it holds that almost surely*

$$Q_h^{\star}(s_h, a_h) = R_h^{\star}(s_h, a_h) + V_{h+1}^{\star}(s_{h+1}).$$

Further, to simply the notation, for each $f \in \mathcal{F}$, we recall that the induced reward model R^f is defined as $R_1^f(s,a) := f_1(s,a)$, $R_h^f(s,a) = f_h(s,a) - f_h(s)$, which implies

$$R^f(\tau) \ = \sum_{h=1}^{H} f_h(s_h,a_h) - f_{h+1}(s_{h+1}).$$

Uniform convergence For each $t \in [T]$, we define $\mathcal{D}_{(t-1)} := \{(\tau^{(k)}, r^{(k)})\}_{k<t}$ be the data collected before tth iteration. We also recall that by definition (8.8), we have

$$\mathcal{L}^{\mathsf{BR}}_{\mathcal{D}^{(t)}}(f) := \sum_{k=1}^{t}\left(\sum_{h=1}^{H}\left[f_h(s_h^{(k)}, a_h^{(k)}) - f_{h+1}(s_{h+1}^{(k)})\right] - r^{(k)}\right)^2 = \sum_{k=1}^{t}\left(R^f(\tau^{(k)}) - r^{(k)}\right)^2.$$

Therefore, a direct instantiation of Theorem G.3 on the class $\mathcal{R} := \{R^f : f \in \mathcal{F}\}$ yields the following proposition.

Proposition G.16. *Let $\delta \in (0,1), \rho \ge 0$. Suppose that Theorem 8.2 and Theorem G.15 holds. Then with probability at least $1-\delta$, for all $t \in [T], f \in \mathcal{F}$, it holds that*

$$\frac{1}{2}\sum_{k=1}^{t}\mathbb{E}^{\pi^{(k)}}\left(R^f(\tau) - R^{\star}(\tau)\right)^2 \le \mathcal{L}^{\mathsf{BR}}_{\mathcal{D}^{(t)}}(f) - \mathcal{L}^{\mathsf{BE}}_{\mathcal{D}^{(t)}}(Q^{\sharp}) + \kappa,$$

where

$$\kappa = C\left(H^3\log(N_{\mathcal{F}}(\alpha)/\delta) + TH\alpha + T\varepsilon_{\mathrm{app}}^2\right),$$

$C > 0$ is an absolute constant, and we denote $N_{\mathcal{F}}(\alpha) := \max_{h\in[H]} N(\mathcal{F}_h, \alpha)$ for any $\alpha \ge 0$.

Performance difference decomposition In this setting, we can rewrite the decomposition (G.2) as

$$\begin{aligned} f_1^{(t)}(s_1) - V^{\pi^{(t)}}(s_1) &= \sum_{h=1}^{H} \mathbb{E}^{\pi^{(t)}} \left[f_h^{(t)}(s_h, a_h) - R_h^\star(s_h, a_h) - f_{h+1}^{(t)}(s_{h+1}) \right] \\ &= \mathbb{E}^{\pi^{(t)}} \left[\sum_{h=1}^{H} \left[f_h^{(t)}(s_h, a_h) - f_{h+1}^{(t)}(s_{h+1}) \right] - \sum_{h=1}^{H} R_h^\star(s_h, a_h) \right] \\ &= \mathbb{E}^{\pi^{(t)}} \left[R^{(t)}(\tau) - R^\star(\tau) \right], \end{aligned} \tag{G.7}$$

where we denote $R^{(t)} := R^{f^{(t)}}$, which is a reward model given by

$$R_1^{(t)}(s, a) := f_1^{(t)}(s, a), \qquad R_h^{(t)}(s, a) = f_h^{(t)}(s, a) - f_h^{(t)}(s).$$

Optimism Similar to appendix G.3.1, we use the fact that from (8.9),

$$f^{(t)} = \max_{f \in \mathcal{F}} \lambda f_1(s_1^{(t)}) - \mathcal{L}^{\mathsf{BR}}_{\mathcal{D}^{(t-1)}}(f),$$

and hence

$$\lambda f_1^{(t)}(s_1^{(t)}) - \mathcal{L}^{\mathsf{BR}}_{\mathcal{D}^{(t-1)}}(f^{(t)}) \geq \lambda V_1^\sharp(s_1^{(t)}) - \mathcal{L}^{\mathsf{BR}}_{\mathcal{D}^{(t-1)}}(f^{(t)}).$$

Using $\left| V_1^\sharp(s_1^{(t)}) - V_1^\star(s_1^{(t)}) \right| \leq \varepsilon_{\text{app}}$, (G.7) and Theorem G.16, we now deduce that

$$\begin{aligned} V^\star(s_1^{(t)}) - V^{\pi^{(t)}}(s_1^{(t)}) &\leq \varepsilon_{\text{app}} + \mathbb{E}^{\pi^{(t)}} \left[R^{(t)}(\tau) - R^\star(\tau) \right] - \frac{\mathcal{L}^{\mathsf{BR}}_{\mathcal{D}^{(t-1)}}(f^{(t)}) - \mathcal{L}^{\mathsf{BE}}_{\mathcal{D}^{(t-1)}}(Q^\sharp)}{\lambda} \\ &\leq \varepsilon_{\text{app}} + \frac{\kappa}{\lambda} + \mathbb{E}^{\pi^{(t)}} \left[R^{(t)}(\tau) - R^\star(\tau) \right] - \frac{1}{2\lambda} \sum_{k=1}^{t-1} \mathbb{E}^{\pi^{(k)}} \left(R^{(t)}(\tau) - R^\star(\tau) \right)^2. \end{aligned} \tag{G.8}$$

Therefore, it remains to prove an analogue to Theorem G.9.

Coverability argument We strength Theorem G.9 using the deterministic nature of the underlying MDP. For each $s \in \mathcal{S}$ and $h \in [H]$, we define

$$\mathcal{S}_h(s; \Pi) := \{(s', a) : \exists \pi \in \Pi, \text{ under } \pi \text{ and } s_1 = s, \text{ it holds that } s_h = s', a_h = a\},$$

and $N_h(s; \Pi) := |\mathcal{S}_h(s; \Pi)|$.

Proposition G.17. *Let $B \geq 1$. For any initial state $s_1 \in \mathcal{S}$, any sequence of reward functions $R^{(1)}, \cdots, R^{(T)}$ and any sequence of policies $\pi^{(1)}, \cdots, \pi^{(T)}$, it holds that*

$$\sum_{t=1}^{T} \mathbb{E}^{\pi^{(t)}} \left[R^{(t)}(\tau) - R^\star(\tau) \middle| s_1 \right]$$

$$\leq \sqrt{2N_{\mathrm{cov}} \log\left(1+\frac{4TH}{B}\right) \cdot \left[2TB + \sum_{1\leq k<t\leq T} \mathbb{E}^{\pi^{(k)}}\left[\left(R^{(t)}(\tau)-R^{\star}(\tau)\right)^2 \middle| s_1\right]\right]},$$

where $N(s_1) \coloneqq \sum_{h=1}^{H} N_h(s_1;\Pi)$*, and the conditional distribution* $\mathbb{E}^{\pi^{(t)}}[\cdot|\, s_1]$ *is taken over the expectation of* τ *generated by executing policy* π *starting with the initial state* s_1*.*

The proof of Theorem G.17 is deferred to the end of this section.

Finalizing the proof With the above preparation, we now finalize the proof of Theorem 8.8. Taking summation of (G.8) over $t = 1, 2, \cdots, T$, we have

$$\begin{aligned}
&\sum_{t=1}^{T} V^{\star}(s_1^{(t)}) - V^{\pi^{(t)}}(s_1^{(t)}) \\
&\leq T\varepsilon_{\mathrm{app}} + \frac{T\kappa}{\lambda} + \sum_{t=1}^{T} \mathbb{E}^{\pi^{(t)}}\left[R^{(t)}(\tau)-R^{\star}(\tau)\right] - \frac{1}{2\lambda}\sum_{1\leq k<t\leq T} \mathbb{E}^{\pi^{(k)}}\left(R^{(t)}(\tau)-R^{\star}(\tau)\right)^2 \\
&= T\varepsilon_{\mathrm{app}} + \frac{T\kappa}{\lambda} + \mathbb{E}_{s_1\sim\rho}\left[\sum_{t=1}^{T} \mathbb{E}^{\pi^{(t)}}\left[R^{(t)}(\tau)-R^{\star}(\tau)\middle| s_1\right] - \frac{1}{2\lambda}\sum_{1\leq k<t\leq T} \mathbb{E}^{\pi^{(k)}}\left[\left(R^{(t)}(\tau)-R^{\star}(\tau)\right)^2 \middle| s_1\right]\right] \\
&\leq T\varepsilon_{\mathrm{app}} + \frac{2T\kappa}{\lambda} + \mathbb{E}_{s_1\sim\rho}\left[N(s_1)\lambda\log\left(1+\frac{TH}{\kappa}\right)\right],
\end{aligned}$$

where the last inequality follows from Theorem G.17 and Cauchy inequality. This is the desired upper bound. □

Proof of Theorem G.17 In the following proof, we assume $s_1 \in \mathcal{S}$ is fixed. Consider

$$\mathcal{I} \coloneqq \{(h,s,a) : h\in[H], (s,a)\in\mathcal{S}_h(s_1;\Pi)\} \subseteq [H]\times\mathcal{S}\times\mathcal{A}.$$

Note that $|\mathcal{I}| = \sum_{h=1}^{H} N_h(s_1;\Pi) = N_{\mathrm{cov}}$. By definition, for any policy π, there is a unique pair $(s_h^{\pi}, a_h^{\pi}) \in \mathcal{S}_h(s_1;\Pi)$, such that under π and starting from s_1, we have $s_h = s_h^{\pi}, a_h = a_h^{\pi}$ deterministically.

For each $t\in[T]$, we consider the following vectors indexed by $\mathcal{I}$:

$$\begin{aligned}
\psi_{(t)} &\coloneqq \left[R_h^{(t)}(s,a) - R_h^{\star}(s,a)\right]_{(h,s,a)\in\mathcal{I}} \in \mathbb{R}^{\mathcal{I}}, \\
\phi_{(t)} &\coloneqq \left[\mathbb{P}^{\pi^{(t)}}(s_h=s, a_h=a|s_1)\right]_{(h,s,a)\in\mathcal{I}} = \sum_{h=1}^{H} e_{(h,s_h^{\pi^{(t)}},a_h^{\pi^{(t)}})} \in \mathbb{R}^{\mathcal{I}}.
\end{aligned}$$

With this definition, it holds that for any $k,t\in[T]$,

$$\mathbb{E}^{\pi^{(k)}}\left[R^{(t)}(\tau)-R^{\star}(\tau)\middle| s_1\right] = \sum_{h=1}^{H}\left[R^{(t)}(s_h^{\pi^{(k)}}, a_h^{\pi^{(k)}}) - R^{\star}(s_h^{\pi^{(k)}}, a_h^{\pi^{(k)}})\right] = \langle\phi_{(k)}, \psi_{(t)}\rangle.$$

Algorithm 5 Outcome-Based Exploration for Preference-based RL

input: Function class $\mathcal{F}$, parameter $\lambda > 0$, reference policy π_{ref}.
initialize: $\mathcal{D} \leftarrow \emptyset$.

for $t = 1, 2, \ldots, T$ **do**
  Compute the optimistic estimates through (8.12): $(f^{(t)}, R^{(t)}) = \max_{f \in \mathcal{F}, R \in \mathcal{R}} \lambda \left[f_1(s_1) - \widehat{V}^{\mathrm{ref}}_{\mathcal{D},R} \right] - \mathcal{L}^{\mathsf{BE}}_{\mathcal{D}}(f; R) - \mathcal{L}^{\mathsf{PbRM}}_{\mathcal{D}}(R)$,
  Select policy $\pi^{(t)} \leftarrow \pi_{f^{(t)}}$.
  for $h = 1, 2, \cdots, H$ **do**
    Execute $\pi^{(t)} \circ_h \pi_{\mathrm{ref}}$ for two episode and obtain two trajectories $(\tau^{(t,h,+)}, \tau^{(t,h,-)})$ and preference feedback $y^{(t,h)}$.

    Update dataset: $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\tau^{(t,h,+)}, \tau^{(t,h,-)}, y^{(t,h)})\}$.
  end for
end for
Output $\widehat{\pi} = \mathrm{Unif}(\pi^{(1:T)})$.

Therefore, we apply the elliptical potential argument [184]. Let $V_t := \sum_{k<t} \phi_{(k)} \left(\phi_{(k)}\right)^\top + B\mathbf{I}$. Then it holds that

$$\begin{aligned}
\sum_{t=1}^{T} |\langle \phi_{(t)}, \psi_{(t)} \rangle| &\leq \sum_{t=1}^{T} \min\{\|\phi_{(t)}\|_{V_t^{-1}}, 1\} \cdot \max\{\|\psi_{(t)}\|_{V_t}, 1\} \\
&\leq \sqrt{\sum_{t=1}^{T} \min\{\|\phi_{(t)}\|^2_{V_t^{-1}}, 1\}} \cdot \sqrt{\sum_{t=1}^{T} \max\{\|\psi_{(t)}\|^2_{V_t}, 1\}}.
\end{aligned}$$

Note that

$$\begin{aligned}
\sum_{t=1}^{T} \max\{\|\psi_{(t)}\|^2_{V_t}, 1\} &\leq \sum_{t=1}^{T} \left[1 + B \|\psi_{(t)}\|^2 + \sum_{k=1}^{t-1} \langle \phi_{(k)}, \psi_{(t)} \rangle^2 \right] \\
&\leq T(1 + 4B|\mathcal{I}|) + \sum_{1 \leq k < t \leq T} \mathbb{E}^{\pi^{(k)}} \left[\left(R^{(t)}(\tau) - R^\star(\tau) \right)^2 \middle| s_1 \right],
\end{aligned}$$

and by Lattimore and Szepesvári [184], we have

$$\sum_{t=1}^{T} \min\{\|\phi_{(t)}\|^2_{V_t^{-1}}, 1\} \leq 2|\mathcal{I}| \log\left(1 + \frac{TH}{|\mathcal{I}|B}\right).$$

Combining the inequalities above and rescale $B \leftarrow \frac{B}{4|\mathcal{I}|}$ completes the proof. □

G.5 Proofs from section 8.4

We present the full description of our algorithm or preference-based RL as follows.

G.5.1 Proof of Theorem 8.10

For each $t \in [T]$, we write $\mathcal{D}_{(t)}$ to be the dataset maintained by Algorithm 2 at the end of the tth iteration, i.e.,

$$\mathcal{D}_{(t)} = \{(\tau^{(k,h,+)}, \tau^{(k,h,-)}, y^{(k,h)})\}_{k \le t, h \in [H]}.$$

Note that for each $t \in [T]$, $h \in [H]$, we have $\pi^{(t,h,-)} = \pi_{\mathrm{ref}}$. Therefore, for each $R \in \mathcal{R}$, we define $V_R^{\mathrm{ref}} := \mathbb{E}^{\pi_{\mathrm{ref}}}[R(\tau)]$ and recall that

$$\widehat{V}_{\mathcal{D},R}^{\mathrm{ref}} := \frac{1}{|\mathcal{D}|} \sum_{(\tau^+,\tau^-,y) \in \mathcal{D}} R(\tau^-).$$

The following lemma follows from the standard uniform convergence rate with Hoeffding's inequality and the union bound.

Lemma G.18. *Let $\delta \in (0,1), \rho \ge 0$. Suppose that Theorem 8.2 and Theorem 8.4 holds. Then with probability at least $1-\delta$, for all $t \in [T], R \in \mathcal{R}$, it holds that*

$$\left|\widehat{V}_{\mathcal{D}^{(t)},R}^{\mathrm{ref}} - V_R^{\mathrm{ref}}\right| \le \sqrt{\frac{\log(2TN(\rho)/\delta)}{t}} + H\rho.$$

We summarize the uniform concentration results for the loss $\mathcal{L}^{\mathsf{BE}}_{\mathcal{D}_{(t)}}$ and $\mathcal{L}^{\mathsf{PbRM}}_{\mathcal{D}_{(t)}}$ as follows. The proof is analogous to Theorem G.7 and is provided in appendix G.5.2.

Proposition G.19. *Let $\delta \in (0,1), \rho \ge 0$. Suppose that Theorem 8.2 and Theorem 8.4 holds. Then with probability at least $1-\delta$, for all $t \in [T], f \in \mathcal{F}, R \in \mathcal{R}$, it holds that*

$$\left|\widehat{V}_{\mathcal{D}^{(t)},R}^{\mathrm{ref}} - V_R^{\mathrm{ref}}\right| \le \sqrt{\frac{\kappa}{tH}},$$

$$\sum_{k \le t}\sum_{h=1}^{H} \mathbb{E}^{\pi^{(k,h,+)},\pi^{(k,h,-)}} \left(\left[R(\tau^+) - R(\tau^-)\right] - \left[R^\star(\tau^+) - R^\star(\tau^-)\right]\right)^2 \le C_\beta \left[\mathcal{L}^{\mathsf{PbRM}}_{\mathcal{D}^{(t)}}(R) - \mathcal{L}^{\mathsf{PbRM}}_{\mathcal{D}^{(t)}}(R^\sharp)\right] + C_\beta H\kappa,$$

$$\sum_{k \le t}\sum_{h=1}^{H} \mathbb{E}^{\pi^{(k)}} \left(f_h(s_h,a_h) - [\mathcal{T}_{R,h} f_{h+1}](s_h,a_h)\right)^2 \le 2\left[\mathcal{L}^{\mathsf{BE}}_{\mathcal{D}^{(t)}}(f;R) - \mathcal{L}^{\mathsf{BE}}_{\mathcal{D}^{(t)}}(Q^\sharp;R^\sharp)\right] + H\kappa,$$

where $C_\beta = \frac{4e^{2\beta}}{\beta^2}$,

$$\kappa = C\left(\log N(\rho) + \log(TH/\delta) + TH^2(\beta+1)(\varepsilon_{\mathrm{app}}^2 + \rho)\right),$$

and $C > 0$ is an absolute constant.

In the following, we condition on the success event of Theorem G.19. Note that $\pi^{(t,h,-)} \equiv \pi_{\mathrm{ref}}$, and hence Theorem G.19 implies that for all $R \in \mathcal{R}$, $t \in [T]$,

$$\sum_{k \le t}\sum_{h=1}^{H} \mathbb{E}^{\pi^{(k)} \circ_h \pi_{\mathrm{ref}}} \left([R(\tau) - R^\star(\tau)] - \left[V_R^{\mathrm{ref}} - V_{R^\star}^{\mathrm{ref}}\right]\right)^2 \le C_\beta \left[\mathcal{L}^{\mathsf{PbRM}}_{\mathcal{D}^{(t)}}(R) - \mathcal{L}^{\mathsf{PbRM}}_{\mathcal{D}^{(t)}}(R^\sharp)\right] + C_\beta H\kappa.$$

Therefore, for any reward function R, we define $\widetilde{R}$ as $\widetilde{R}_1(s,a) = R_1(s,a) - V_R^{\mathrm{ref}}$ and $\widetilde{R}_h(s,a) = R_h(s,a)$ for $h > 1$. Then it is clear that $\widetilde{R}(\tau) = R(\tau) - V_R^{\mathrm{ref}}$, and for all $R \in \mathcal{R}$, $t \in [T]$, we have

$$\sum_{k \le t} \sum_{h=1}^{H} \mathbb{E}^{\pi^{(k)} \circ_h \pi_{\mathrm{ref}}} \left(\widetilde{R}(\tau) - \widetilde{R}^\star(\tau) \right)^2 \le C_\beta \left[\mathcal{L}^{\mathsf{PbRM}}_{\mathcal{D}^{(t)}}(R) - \mathcal{L}^{\mathsf{PbRM}}_{\mathcal{D}^{(t)}}(R^\sharp) \right] + C_\beta H \kappa. \tag{G.9}$$

Performance difference decomposition In this setting, we re-write (G.3) as follows:

$$\begin{aligned} f_1^{(t)}(s_1) - V^{\pi^{(t)}}(s_1) = & \sum_{h=1}^{H} \mathbb{E}^{\pi^{(t)}} \left[f_h^{(t)}(s_h, a_h) - [\mathcal{T}_{R^{(t)}} f_{h+1}^{(t)}](s_h, a_h) \right] \\ & + \mathbb{E}^{\pi^{(t)}} \left[\sum_{h=1}^{H} R_h^{(t)}(s_h, a_h) - \sum_{h=1}^{H} R_h^\star(s_h, a_h) \right] \\ = & \sum_{h=1}^{H} \mathbb{E}^{\pi^{(t)}} e_h^{(t)}(s_h, a_h) + \mathbb{E}^{\pi^{(t)}} \left[\widetilde{R}^{(t)}(\tau) - \widetilde{R}^\star(\tau) \right] + V_{R^{(t)}}^{\mathrm{ref}} - V_{R^\star}^{\mathrm{ref}}, \end{aligned}$$

where we recall that we denote $e_h^{(t)} := f_h^{(t)} - \mathcal{T}_{R^{(t)}} f_{h+1}^{(t)}$. Therefore, we re-organize the equality as

$$\left[f_1^{(t)}(s_1) - V_{R^{(t)}}^{\mathrm{ref}} \right] - \left[V^{\pi^{(t)}}(s_1) - V_{R^\star}^{\mathrm{ref}} \right] = \mathbb{E}^{\pi^{(t)}} \left[\widetilde{R}^{(t)}(\tau) - \widetilde{R}^\star(\tau) \right] + \sum_{h=1}^{H} \mathbb{E}^{\pi^{(t)}} e_h^{(t)}(s_h, a_h). \tag{G.10}$$

With the above preparation, we present the proof of Theorem 8.10, which closely follows the proof of Theorem 8.5 in appendix G.3.1.

Proof of Theorem 8.10 By definition, for each $t \in [T]$,

$$(f^{(t)}, R^{(t)}) = \max_{f \in \mathcal{F}, R \in \mathcal{R}} \lambda \left[f_1(s_1) - \widehat{V}^{\mathrm{ref}}_{\mathcal{D}^{(t-1)}, R} \right] - \mathcal{L}^{\mathsf{BE}}_{\mathcal{D}^{(t-1)}}(f; R) - \mathcal{L}^{\mathsf{PbRM}}_{\mathcal{D}^{(t-1)}}(R).$$

Therefore, using $Q^\sharp \in \mathcal{F}, R^\sharp \in \mathcal{R}$, we have

$$\begin{aligned} & \left[f_1^{(t)}(s_1) - \widehat{V}^{\mathrm{ref}}_{\mathcal{D}^{(t-1)}, R^{(t)}} \right] - \left[V_1^\sharp(s_1) - \widehat{V}^{\mathrm{ref}}_{\mathcal{D}^{(t-1)}, R^\sharp} \right] \\ \le & - \frac{\mathcal{L}^{\mathsf{BE}}_{\mathcal{D}^{(t-1)}}(f^{(t)}; R^{(t)}) - \mathcal{L}^{\mathsf{BE}}_{\mathcal{D}^{(t-1)}}(Q^\sharp; R^\sharp)}{\lambda} - \frac{\mathcal{L}^{\mathsf{PbRM}}_{\mathcal{D}^{(t-1)}}(R^{(t)}) - \mathcal{L}^{\mathsf{PbRM}}_{\mathcal{D}^{(t-1)}}(R^\sharp)}{\lambda}. \end{aligned}$$

Using the decomposition (G.10), Theorem G.19, and the fact that $\left|V_1^\sharp(s_1) - V_1^\star(s_1)\right| \le \varepsilon_{\text{app}}$, $\left|R^\sharp(\tau) - R^\star(\tau)\right| \le H\varepsilon_{\text{app}}$, we have

$$\begin{aligned} V_1^\star(s_1) - V^{\pi^{(t)}}(s_1) \le (H+1)\varepsilon_{\text{app}} + \left|V^{\text{ref}}_{R^{(t)}} - \widehat{V}^{\text{ref}}_{\mathcal{D}^{(t-1)},R^{(t)}}\right| + \left|V^{\text{ref}}_{R^\sharp} - \widehat{V}^{\text{ref}}_{\mathcal{D}^{(t-1)},R^\sharp}\right| + \frac{2H\kappa}{\lambda} \\ + \sum_{h=1}^H \left(\mathbb{E}^{\pi^{(t)}}\left[e_h^{(t)}(s_h,a_h)\right] - \frac{1}{C_\beta\lambda}\sum_{k<t}\mathbb{E}^{\pi^{(k)}} e_h^{(t)}(s_h,a_h)^2\right) \\ + \mathbb{E}^{\pi^{(t)}}\left[\widetilde{R}^{(t)}(\tau) - \widetilde{R}^\star(\tau)\right] - \frac{1}{2\lambda}\sum_{k<t}\sum_{h=1}^H \mathbb{E}^{\pi^{(k)}\circ_h\pi_{\text{ref}}}\left(\widetilde{R}^{(t)}(\tau) - \widetilde{R}^\star(\tau)\right)^2. \end{aligned} \tag{G.11}$$

Taking summation over $t = 1, 2, \cdots, T$ and apply Theorem G.8, Theorem G.9, and Theorem G.18 yields

$$\sum_{t=1}^T V_1^\star(s_1) - V^{\pi^{(t)}}(s_1) \le O(1)\cdot\left[H(\varepsilon_{\text{app}}+\rho) + \sqrt{T\kappa} + \frac{TH\kappa}{\lambda} + C_\beta\lambda H C_{\text{cov}}\log\left(1+\frac{C_{\text{cov}}T}{\kappa}\right)\right].$$

This is the desired upper bound. □

G.5.2 Proof of Theorem G.19

The inequality involving $\mathcal{L}^{\mathsf{BE}}_{\mathcal{D}^{(t)}}$ is implied by Theorem G.7 and proven in appendix G.3.3. In the following, we only need to prove the inequality involving $\mathcal{L}^{\mathsf{PbRM}}_{\mathcal{D}^{(t)}}$ by invoking Theorem G.4.

Consider the class $\Theta = \mathcal{R}\cup\{R^\star\}$, $\mathcal{X} = (\mathcal{S}\times\mathcal{A})^H\times(\mathcal{S}\times\mathcal{A})^H$, and $\mathcal{Y} = \{0,1\}$. For any $R\in\Theta$, we define

$$P_R(1|\tau^+,\tau^-) = \frac{\exp(\beta R(\tau^+))}{\exp(\beta R(\tau^+)) + \exp(\beta R(\tau^-))}, \quad P_R(0|\tau^+,\tau^-) = \frac{\exp(\beta R(\tau^-))}{\exp(\beta R(\tau^+)) + \exp(\beta R(\tau^-))},$$

following Theorem 8.9.

Recall that $\mathcal{D}_{(t-1)} = \{(\tau^{(k,h,+)}, \tau^{(k,h,-)}, y^{(k,h)})\}_{k<t,h\in[H]}$ is the history up to the t-th iteration. For simplicity, we denote $x^{(t,h)} := (\tau^{(t,h,+)}, \tau^{(t,h,-)})$. Note that $(x^{(t,1)}, y^{(t,1)}), \cdots, (x^{(t,H)}, y^{(t,H)})$. Then it is clear that for all $t\in[T], h\in[H]$,

$$\mathbb{P}(y^{(t,h)}|x^{(t,h)}, \mathcal{D}^{(t-1)}) = P_{R^\star}(y^{(t,h)}|x^{(t,h)}),$$

and it also holds that

$$L(R(\tau^+) - R(\tau^-), y) = -\log P_R(y|\tau^+,\tau^-), \qquad \forall y\in\{0,1\}.$$

Further, noting that $N_{\log}(\Theta, 2H\beta\rho) \le N(\mathcal{R},\rho) + 1$. Therefore, applying Theorem G.4 gives the following result: with probability at least $1-\frac{\delta}{2}$, for any $R\in\mathcal{R}$, $t\in[T]$,

$$\sum_{k=1}^t\sum_{h=1}^H \mathbb{E}^{\pi^{(k,h,+)},\pi^{(k,h,-)}} D^2_{\text{H}}\left(P_R(\cdot|\tau^+,\tau^-), P_{R^\star}(\cdot|\tau^+,\tau^-)\right)$$

$$
\begin{aligned}
&\leq \frac{1}{2} \sum_{(\tau^+,\tau^-,y)} \left[L(R(\tau^+) - R(\tau^-), y) - L(R^\star(\tau^+) - R^\star(\tau^-), y)\right] \\
&\quad + \log(N(\mathcal{R}, H\rho) + 1) + \log(2/\delta) + TH^3\beta\rho \\
&\leq \frac{1}{2}\left[\mathcal{L}^{\mathsf{PbRM}}_{\mathcal{D}^{(t)}}(R) - \mathcal{L}^{\mathsf{PbRM}}_{\mathcal{D}^{(t)}}(R^\sharp)\right] + \frac{1}{2}H\kappa,
\end{aligned}
$$

where the second inequality uses the fact that $\left|R^\sharp(\tau) - R^\star(\tau)\right| \leq H\varepsilon_{\mathrm{app}}$. Finally, note that $D^2_{\mathrm{H}}\left(\mathrm{Bern}\left(p\right), \mathrm{Bern}\left(q\right)\right) \geq \frac{1}{2}(p-q)^2$ and

$$
\left|\frac{1}{e^{\beta w}+1} - \frac{1}{e^{\beta w'}+1}\right| \geq \frac{\beta}{2e^\beta}\left|w - w'\right|, \qquad \forall w, w' \in [-1, 1].
$$

Therefore, using the definition of P_R completes the proof. □

G.6 Proofs of Lower Bounds

G.6.1 Hard Case of Learning with Fitted Reward Models

As mentioned in section 8.3.1, in Algorithm 2 the learner has to optimize over the reward class and value function class jointly. In the following, we argue that if the learner first learns a fitted reward model in the reward class, then optimizes the value function with the fitted rewards, the output policies at each iteration never converge to the optimal policy.

In detail, we consider algorithms in the form of Algorithm 6, where the learner fits the reward model $R_{(t)}$ at iteration t first, then the learner calls algorithm alg, which takes per-step rewards data as input and outputs a policy π_t at each iteration. To align with the structure of Algorithm 2, we take alg to be a single iteration of the GOLF algorithm in Jin, Liu, and Miryoosefi [200], i.e. $\pi_{(t)} = \pi_{f_{(t)}}$ where $f_{(t)} = \arg\max_{f \in \mathcal{F}_{(t)}} f(x_1, \pi_f(x_1))$. Here the confidence set $\mathcal{F}_{(t)}$ is defined as

$$
\mathcal{F}_{(t)} = \left\{f \in \mathcal{F} : \mathcal{L}^{\mathsf{BE}}_{\mathcal{D}_{(t-1)}}(f; \widehat{R}_{(t-1)}) \leq \beta\right\}
$$

with $\mathcal{L}^{\mathsf{BE}}_{\mathcal{D}}$ defined in eq. (8.5).

Algorithm 6 RL with fitted reward models

input: Algorithm alg, reward regression oracle O.

1: **Initialize** $\mathcal{D}_{h(0)} = \emptyset$ for every $h \in [H]$
2: **for** $t = 1, 2, \ldots, T$ **do**
3: Feed $\mathcal{D}_{(t-1)}$ to alg and receive $\pi^{(t)}$ from alg
4: Execute $\pi^{(t)}$ and receive $(\tau^{(t)}, r^{(t)})$, where $\tau_{(t)} = (s_{1(t)}, a_{1(t)}, \cdots, s_{H(t)}, a_{H(t)})$
5: Receive the fitted reward function from O: $\widehat{R}^{(t)} = \min_{R \in \mathcal{R}} \sum_{k=1}^{t} (R(\tau^{(k)}) - r^{(k)})^2$.
6: Let $\widehat{r}_{h(t)} = \widehat{R}_{(t)h}(s_{h(t)}, a_{h(t)})$ for each $h \in [H]$.
7: Let $\mathcal{D}_{(t)} = \mathcal{D}_{(t-1)} \cup \{(s_{1(t)}, a_{1(t)}, \widehat{r}_{1(t)}, \cdots, s_{H(t)}, a_{H(t)}, \widehat{r}_{H(t)})\}$.
8: **end for**

Then we have the following proposition, which shows that this approach outputs suboptimal policies at every iteration in some special hard cases.

Proposition G.20. *Consider Algorithm 6 with* alg *to a single iteration of the GOLF algorithm. After running T iterations, the learner averages over all policies to output a policy. There exists an MDP class that realizes the ground truth MDP, such that the above algorithm outputs a policy which is at least* 0.01*-suboptimal.*

Proof of Theorem G.20. We consider the following class of two-layer MDP, where $\mathcal{S}_1 = \{s_1\}$, $\mathcal{S}_2 = \{s_2\}$, and the action space to be $\mathcal{A} = \{a_1, a_2\}$. The transition models $\mathbb{T}$ are identical across the class, and have the following form:

$$\mathbb{T}(s_2 \mid s_1, a_i) = 1, \qquad \forall i \in \{1, 2\}.$$

The reward class is defined as $\mathcal{R} = \{R^1, R^2\}$, where

$$\begin{aligned} & R^1(s_1, a_1) = R^1(s_1, a_2) = 0.20, \quad R^1(s_2, a_1) = 0.20, \quad R^1(s_2, a_2) = 0.19, \\ \text{and} \quad & R^2(s_1, a_1) = R^2(s_1, a_2) = 0.00, \quad R^2(s_2, a_1) = 0.38, \quad R^2(s_2, a_2) = 0.39. \end{aligned}$$

The Q-function class $\mathcal{Q}$ is defined as $\mathcal{Q} = \{Q^1, Q^2, Q^3, Q^4\}$, which takes value in table G.1 respectively. Notice that in all possible reward models and Q-functions, the values at (s_1, a_1)

Table G.1: Value of Q^1, Q^2, Q^3, Q^4

	(s_1, a_1)	(s_1, a_2)	(s_2, a_1)	(s_2, a_2)
Q^1	0.40	0.40	0.20	0.19
Q^2	0.20	0.20	0.20	0.19
Q^3	0.59	0.59	0.38	0.39
Q^4	0.39	0.39	0.38	0.39

and at (s_1, a_2) are the same. In the following, when without ambiguity we simply use $R(s_1)$ to denote $R(s_1, a_1)$ and $R(s_1, a_2)$, and use $Q(s_1)$ to denote $Q(s_1, a_1)$ and $Q(s_2, a_2)$.

We further suppose the ground truth model reward satisfies $R = R^1$, then we can verify that the optimal Q-function is Q^1. It is easy to verify that sets $\mathcal{Q}$ and $\mathcal{R}$ satisfy the completeness assumption. Hence, sets $\mathcal{Q}$ and $\mathcal{R}$ satisfy the realizability assumption Theorem 8.2 and the completeness assumption Theorem 8.4 with $\mathcal{G} = \mathcal{Q}$.

To see why this is a hard-case for GOLF type algorithms, we first notice that for any trajectory $\tau = (s_1, \tilde{a}_1, s_2, \tilde{a}_2)$ with outcome reward $r = R^1(s_1, \tilde{a}_1) + R^1(s_2, \tilde{a}_2)$ collected by the algorithm, we always have

$$r = R^2(s_1) + R^2(s_2, \tilde{a}_2).$$

Hence as long as $\mathcal{D}$ does not contain state-action pair (s_2, a_1), when fitting the reward function using the following ERM oracle:

$$R = \arg\min_{R \in \mathcal{R}} \sum_{(\tau, r) \in \mathcal{D}} (r(\tau) - r)^2,$$

the reward model R^2 always achieves the minimum. In the worst case, we assume the fitted reward models encountered by the learner at such rounds are always R^2.

In the following, we verify that by running the GOLF algorithm, the learner will not encounter the state-action pair (s_2, a_1) at any round. We notice that the optimal policies of Q^3 and Q^4 all take a_2 at state s_2, and also that

$$Q^3(s_1) \geq Q^1(s_1) \quad \text{and} \quad Q^3(s_1) \geq Q^2(s_1).$$

Hence, to verify that the algorithm never chooses a_1 at state s_2, we only need to verify that if either Q^1 or Q^2 belongs to the confidence set, then Q^3 also belongs to the confidence set.

When the learner collects a new trajectory, two new pieces of data will be added to the dataset $\mathcal{D}$. If the trajectory does not pass through the state-action pair (s_2, a_1), these two pieces of data will be in the following form:

$$(s_1, a_1, R^2(s_1)), \quad (s_2, a_2, R^2(s_2, a_2)) \quad \text{or} \quad (s_1, a_2, R^2(s_1)), \quad (s_2, a_2, R^2(s_2, a_2)).$$

No matter which one of these two, we have the following inequality for the sum of squared Bellman error across these two pieces of data

$$\begin{aligned}
\mathcal{E}_1(Q^1)^2 + \mathcal{E}_2(Q^1)^2 &= 0.20^2 + 0.20^2 \geq 0.20^2 = \mathcal{E}_1(Q^3)^2 + \mathcal{E}_2(Q^3)^2, \\
\mathcal{E}_1(Q^2)^2 + \mathcal{E}_2(Q^2)^2 &= 0.01^2 + 0.28^2 \geq 0.20^2 = \mathcal{E}_1(Q^3)^2 + \mathcal{E}_2(Q^3)^2.
\end{aligned}$$

According to the construction of the confidence set, if either Q^1 or Q^2 belongs to the confidence set, then Q^3 belongs to the confidence set as well.

Therefore, no matter how many rounds the algorithm runs, the optimistic policy always takes action a_2 at state s_2. Hence the average policy $\widehat{\pi}$ also takes a_2 at s_2, which implies that

$$J(\pi^\star) - J(\widehat{\pi}) \geq 0.01.$$

□

G.6.2 Proof of Theorem 8.11

Fix a parameter $\varepsilon \in (0, 1)$ and $N \leq \left(\frac{1}{2\varepsilon}\right)^{d/2}$. Then, by the standard packing argument over sphere [see e.g., 207], there exists a set $\Theta = \{\theta_1, \cdots, \theta_N\} \subseteq \mathbb{S}^{d-1}$ such that

$$\|\theta_i - \theta_j\| \geq \sqrt{2\varepsilon}, \qquad \forall i \neq j.$$

This implies $\langle \theta_i, \theta_j \rangle \leq 1 - \varepsilon$ for any $i \neq j$.

Construction In the following, we set $b = 1 - \varepsilon$, and construct state space $\mathcal{S}$ as

$$\mathcal{S} = \mathcal{S}_1 \sqcup \mathcal{S}_2, \qquad \mathcal{S}_1 = \{s_1\}, \qquad \mathcal{S}_2 = \Theta,$$

and let action space $\mathcal{A} = \Theta$. The initial state is always s_1, and we define the transition $\mathbb{T}$ as

$$\mathbb{T}(s_2 = \theta \mid s_1, a = \theta) = 1, \qquad \forall \theta \in \Theta,$$

i.e., taking action $a = \theta$ at s_1 transits to $s_2 = \theta$ deterministically.

Reward functions For any $v \in \Theta$, we define the reward model R^v as follows:

$$R_1^v(s,a) = \frac{1}{3}\left[\mathrm{ReLU}(\langle a, v\rangle - b) + \langle a, v\rangle + 1\right] \in [0,1], \qquad \forall a \in \Theta,$$
$$R_2^v(s,a) = \frac{1}{3}\left[1 - \langle s, v\rangle\right] \in [0,1], \qquad \forall s \in \Theta.$$

Note that we can write $g_1(x) = \frac{1}{3}\left[\mathrm{ReLU}(x-b) + x + 1\right], g_2(x) = \frac{1-x}{3}$, and then g_2 is a linear function, and

$$\frac{1}{3}|x-y| \le |g_1(x) - g_2(y)| \le \frac{2}{3}|x-y|, \qquad \forall x, y \in \mathbb{R}.$$

Hence, g_1 and g_2 are (well-conditioned) generalized linear functions, and hence R_1^v and R_2^v are both (well-conditioned) d-dimensional generalized linear functions.

We let M^v be the MDP with transition $\mathbb{T}$ and mean reward function R^v, and $\mathcal{M} = \{M^v : v \in \Theta\}$ be the corresponding class of MDPs. We next show that $\mathcal{M}$ can be learned with polynomial process-based samples, but cannot be learned with polynomial outcome-based samples.

Exponential Lower Bound for Outcome-Based Setting When executing a policy π in MDP M^v, we have $a_1 = \pi_1(s_1)$, $s_2 = a_1$, and $a_2 = \pi_2(s_2)$, and the data $(\tau_{\theta,\theta'}, R)$ observed are in the following form of trajectory together outcome-based rewards:

$$\tau_\pi = (s_1, a_1, s_2, a_2), \qquad R|\tau_\pi \sim \mathrm{Bern}\left(\frac{1}{3}\mathrm{ReLU}(\langle a_1, v\rangle - b) + \frac{2}{3}\right),$$

where τ_π is a deterministic function of π. In the following, we denote $a_\pi = \pi_1(s_1)$, and then $\mathbb{E}[R|\pi] = \frac{1}{4}\mathrm{ReLU}(\langle a_\pi, v\rangle - b) + \frac{1}{2}$. Further, under M^v,

$$J(\pi) = \frac{2}{3} + \frac{\varepsilon}{3}\mathbb{I}\{a_\pi = v\},$$

and in particular, $J(\pi^\star) = \frac{2}{3} + \frac{\varepsilon}{3}$. Therefore, for any policy π, it is $(\varepsilon/3)$-optimal under M^v only when $a_\pi = v$. Hence, we can apply the standard lower bound argument for multi-arm bandits [see e.g., 184] to show that: If there any T-round algorithm that returns an $(\varepsilon/3)$-optimal policy with probability at least $\frac{3}{4}$ for any MDP $M^v \in \mathcal{M}$, then it must hold that $T \ge c\frac{N}{\varepsilon^2}$ (where $c > 0$ is an absolute constant). Setting $\varepsilon = \frac{1}{3}$ completes the proof of the lower bound. □

Polynomial Upper Bound with Process-Based Samples Notice that for fixed $v \in \Theta$, under $M^v \in \mathcal{M}$, we have

$$J(\pi^\star) - J(\pi) = \frac{\varepsilon}{3}\left[1 - \mathbb{I}\{a_\pi = v\}\right] = \frac{1}{3}\left[\langle v, v\rangle - \langle a_\pi, v\rangle\right].$$

Thus, for any $\theta \in \Theta$, we define π^θ as $\pi_1^\theta(s) = \pi_2^\theta(s) = \theta$ for $\forall s \in \mathcal{S}$. Then it holds that under M^v,

$$\mathbb{E}\left[\frac{1}{3} - R_2 \middle| \pi^\theta\right] = \frac{1}{3}\langle \theta, v\rangle.$$

Therefore, given access to process reward feedback, we can reduce learning $\mathcal{M}$ to learning a class of linear bandits. Hence, for any $\alpha > 0$, with process reward feedback, there are algorithms that returns an α-optimal policy with high probability, using $T \leq \widetilde{O}\left(d^2/\alpha^2\right)$ episodes with process rewards [see e.g., 233]. □

References

[1] L. LeCam. "Convergence of estimates under dimensionality restrictions". In: *The Annals of Statistics* (1973), pp. 38–53.

[2] I. A. Ibragimov and R. Z. Khas' minskii. "Estimation of distribution density". In: *Journal of Soviet Mathematics* 21 (1983), pp. 40–57.

[3] N. Littlestone. "Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm". In: *Machine learning* 2 (1988), pp. 285–318.

[4] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

[5] R. S. Sutton, A. G. Barto, et al. *Reinforcement learning: An introduction*. Vol. 1. MIT press Cambridge, 1998.

[6] C. Villani. *Topics in optimal transportation*. Vol. 58. American Mathematical Society, 2003.

[7] R. M. Dudley. "The speed of Mean Glivenko-Cantelli convergence". In: *The Annals of Mathematical Statistics* 40.1 (1969), pp. 40–50.

[8] Z. Goldfeld, K. Greenewald, J. Niles-Weed, and Y. Polyanskiy. "Convergence of smoothed empirical measures with applications to entropy estimation". In: *IEEE Transactions on Information Theory* 66.7 (2020), pp. 4368–4391.

[9] L. Birgé. "Approximation dans les espaces métriques et théorie de l'estimation". In: *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 65 (1983), pp. 181–237.

[10] L. Birgé. "On estimating a density using Hellinger distance and some other strange facts". In: *Probability theory and related fields* 71.2 (1986), pp. 271–291.

[11] Y. Yang and A. R. Barron. "An asymptotic property of model selection criteria". In: *IEEE Transactions on Information Theory* 44.1 (1998), pp. 95–116.

[12] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. 1st. Springer Publishing Company, Incorporated, 2008. ISBN: 0387790519.

[13] Y. I. Ingster. "On the minimax nonparametric detection of signals in white gaussian noise". In: *Problemy Peredachi Informatsii* 18.2 (1982), pp. 61–73.

[14] Y. I. Ingster. "An asymptotic minimax testing of nonparametric hypotheses on the density of the distribution of an independent sample". In: *Journal of Soviet Mathematics* 33.1 (1986), pp. 744–758.

[15] P. R. Gerber and Y. Polyanskiy. "Likelihood-free hypothesis testing". In: *IEEE Transactions on Information Theory* (2024).

[16] N. Cesa-Bianchi and G. Lugosi. "Minimax regret under log loss for general classes of experts". In: *Proceedings of the Twelfth annual conference on computational learning theory*. 1999, pp. 12–18.

[17] A. K. Rakhlin and K. Sridharan. "Sequential probability assignment with binary alphabets and large classes of experts". In: *arXiv preprint arXiv:1501.07340* (2015).

[18] B. Bilodeau, D. J. Foster, and D. M. Roy. "Tight bounds on minimax regret under logarithmic loss via self-concordance". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 919–929.

[19] P. L. Bartlett, P. M. Long, and R. C. Williamson. "Fat-shattering and the learnability of real-valued functions". In: *Proceedings of the seventh annual conference on Computational learning theory*. 1994, pp. 299–310.

[20] A. K. Rakhlin and K. Sridharan. "Online learning: Random averages, combinatorial parameters, and learnability". In: *Advances in Neural Information Processing Systems* 23 (2010).

[21] M. Anthony and P. L. Bartlett. "Function learning from interpolation". In: *Combinatorics, Probability and Computing* 9.3 (2000), pp. 213–225.

[22] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. "Training language models to follow instructions with human feedback". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 27730–27744.

[23] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, et al. "Training a helpful and harmless assistant with reinforcement learning from human feedback". In: *arXiv preprint arXiv:2204.05862* (2022).

[24] A. Jaech, A. Kalai, A. Lerer, A. Richardson, A. El-Kishky, A. Low, A. Helyar, A. Madry, A. Beutel, A. Carney, et al. "OpenAI o1 System Card". In: *arXiv preprint arXiv:2412.16720* (2024).

[25] J. Uesato, N. Kushman, R. Kumar, F. Song, N. Siegel, L. Wang, A. Creswell, G. Irving, and I. Higgins. "Solving math word problems with process-and outcome-based feedback". In: *arXiv preprint arXiv:2211.14275* (2022).

[26] H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, and K. Cobbe. "Let's verify step by step". In: *arXiv preprint arXiv:2305.20050* (2023).

[27] Y. Efroni, N. Merlis, and S. Mannor. "Reinforcement learning with trajectory feedback". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 35. 2021, pp. 7288–7295.

[28] A. Pacchiano, A. Saha, and J. Lee. "Dueling rl: reinforcement learning with trajectory preferences". In: *arXiv preprint arXiv:2111.04850* (2021).

[29] N. Chatterji, A. Pacchiano, P. Bartlett, and M. Jordan. "On the theory of reinforcement learning with once-per-episode feedback". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 3401–3412.

[30] A. Cassel, H. Luo, A. Rosenberg, and D. Sotnikov. "Near-optimal regret in linear MDPs with aggregate bandit feedback". In: *arXiv preprint arXiv:2405.07637* (2024).

[31] T. Lancewicki and Y. Mansour. "Near-optimal Regret Using Policy Optimization in Online MDPs with Aggregate Bandit Feedback". In: *arXiv preprint arXiv:2502.04004* (2025).

[32] X. Chen, H. Zhong, Z. Yang, Z. Wang, and L. Wang. "Human-in-the-loop: Provably Efficient Preference-based Reinforcement Learning with General Function Approximation". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 3773–3793.

[33] F. Chen, S. Mei, and Y. Bai. "Unified algorithms for rl with decision-estimation coefficients: pac, reward-free, preference-based learning, and beyond". In: *arXiv preprint arXiv:2209.11745* (2022).

[34] R. Wu and W. Sun. "Making rl with preference-based feedback efficient via randomization". In: *arXiv preprint arXiv:2310.14554* (2023).

[35] Y. Wang, Q. Liu, and C. Jin. "Is RLHF More Difficult than Standard RL?" In: *arXiv preprint arXiv:2306.14111* (2023).

[36] E. Boissard and T. Le Gouic. "On the mean speed of convergence of empirical and occupation measures in Wasserstein distance". In: *Annales de l'IHP Probabilités et statistiques*. Vol. 50. 2. 2014, pp. 539–563.

[37] S. Bobkov and M. Ledoux. *One-dimensional empirical measures, order statistics, and Kantorovich transport distances*. Vol. 261. American Mathematical Society, 2019.

[38] P. Berthet and J.-C. Fort. "Exact rate of convergence of the expected W_2 distance between the empirical and true Gaussian distribution". In: *arXiv preprint arXiv:2012.01955* (2020).

[39] M. Ledoux. "On optimal matching of Gaussian samples". In: *Journal of Mathematical Sciences* 238.4 (2019), pp. 495–522.

[40] M. Talagrand. "Scaling and non-standard matching theorems". In: *Comptes Rendus Mathematique* 356.6 (2018), pp. 692–695.

[41] S. Dereich, M. Scheutzow, and R. Schottstedt. "Constructive quantization: Approximation by empirical measures". In: *Annales de l'IHP Probabilités et statistiques*. Vol. 49. 4. 2013, pp. 1183–1203.

[42] N. Fournier and A. Guillin. "On the rate of convergence in Wasserstein distance of the empirical measure". In: *Probability Theory and Related Fields* 162.3 (2015), pp. 707–738.

[43] J. Weed and F. Bach. "Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance". In: *Bernoulli* 25.4A (2019), pp. 2620–2648.

[44] A. K. Kim. "Minimax bounds for estimation of normal mixtures". In: *Bernoulli* 20.4 (2014), pp. 1802–1818.

[45] A. K. Kim and A. Guntuboyina. "Minimax bounds for estimating multivariate Gaussian location mixtures". In: *Electronic Journal of Statistics* 16.1 (2022), pp. 1461–1484.

[46] S. van de Geer. "Hellinger-consistency of certain nonparametric maximum likelihood estimators". In: *The Annals of Statistics* 21.1 (1993), pp. 14–44.

[47] W. H. Wong and X. Shen. "Probability inequalities for likelihood ratios and convergence rates of sieve MLEs". In: *The Annals of Statistics* 23.2 (1995), pp. 339–362.

[48] S. van de Geer. "Rates of convergence for the maximum likelihood estimator in mixture models". In: *Journal of Nonparametric Statistics* 6.2-4 (1996), pp. 293–301.

[49] C. R. Genovese and L. Wasserman. "Rates of convergence for the Gaussian mixture sieve". In: *The Annals of Statistics* 28.4 (2000), pp. 1105–1127. DOI: 10.1214/aos/1015956709.

[50] S. Ghosal and A. W. van der Vaart. "Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities". In: *The Annals of Statistics* 29.5 (2001), pp. 1233–1263. DOI: 10.1214/aos/1013203452.

[51] S. Ghosal and A. van der Vaart. "Posterior convergence rates of Dirichlet mixtures at smooth densities". In: *The Annals of Statistics* 35.2 (2007), pp. 697–723.

[52] C.-H. Zhang. "Generalized maximum likelihood estimation of normal mixture densities". In: *Statistica Sinica* 19.3 (2009), pp. 1297–1318.

[53] S. Saha and A. Guntuboyina. "On the nonparametric maximum likelihood estimator for Gaussian location mixture densities with application to Gaussian denoising". In: *The Annals of Statistics* 48.2 (2020), pp. 738–762. DOI: 10.1214/19-AOS1817.

[54] Y. Polyanskiy and Y. Wu. "Sharp regret bounds for empirical Bayes and compound decision problems". In: *arXiv preprint arXiv:2109.03943* (2021).

[55] Y. G. Yatracos. "Rates of convergence of minimum distance estimators and Kolmogorov's entropy". In: *The Annals of Statistics* 13.2 (1985), pp. 768–774.

[56] Y. Yang and A. R. Barron. "Information-theoretic determination of minimax rates of convergence". In: *Annals of Statistics* (1999), pp. 1564–1599.

[57] D. Haussler, M. Opper, et al. "Mutual information, metric entropy and cumulative relative entropy risk". In: *The Annals of Statistics* 25.6 (1997), pp. 2451–2492.

[58] Y. Polyanskiy and Y. Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, 2022.

[59] N. Doss, Y. Wu, P. Yang, and H. H. Zhou. "Optimal estimation of high-dimensional Gaussian location mixtures". In: *The Annals of Statistics* 51.1 (2023), pp. 62–95. DOI: 10.1214/22-AOS2207.

[60] Y. I. Ingster. "Minimax testing of nonparametric hypotheses on a distribution density in the L_p metrics". In: *Theory of Probability & Its Applications* 31.2 (1987), pp. 333–337.

[61] L. Le Cam. *Asymptotic methods in statistical decision theory*. Springer Science & Business Media, 2012.

[62] Y. I. Ingster and I. A. Suslina. *Nonparametric goodness-of-fit testing under Gaussian models*. Vol. 169. Springer Science & Business Media, 2003.

[63] D. L. Donoho, R. C. Liu, and B. MacGibbon. "Minimax risk over hyperrectangles, and implications". In: *The Annals of Statistics* (1990), pp. 1416–1437.

[64] M. Neykov. "Signal Detection with Quadratically Convex Orthosymmetric Constraints". In: *arXiv preprint arXiv:2308.13036* (2023).

[65] Y. Baraud. "Non-asymptotic minimax rates of testing in signal detection". In: *Bernoulli* (2002), pp. 577–606.

[66] M. Nussbaum. "Asymptotic equivalence of density estimation and Gaussian white noise". In: *The Annals of Statistics* (1996), pp. 2399–2430.

[67] L. D. Brown and M. G. Low. "Asymptotic equivalence of nonparametric regression and white noise". In: *The Annals of Statistics* 24.6 (1996), pp. 2384–2398.

[68] A. van der Vaart. "The statistical work of lucien le cam". In: *The Annals of Statistics* 30.3 (2002), pp. 631–682.

[69] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. "Density estimation by wavelet thresholding". In: *The Annals of statistics* (1996), pp. 508–539.

[70] I. M. Johnstone. *Gaussian estimation: Sequence and wavelet models*. 2019.

[71] A. Juditsky and A. Nemirovski. "Statistical inference via convex optimization". In: (2020).

[72] M. S. Pinsker. "Optimal filtering of square-integrable signals in Gaussian noise". In: *Problemy Peredachi Informatsii* 16.2 (1980), pp. 52–68.

[73] C. Cheng, D. Levy, and J. C. Duchi. "Geometry, Computation, and Optimality in Stochastic Optimization". In: *arXiv preprint arXiv:1909.10455* (2019).

[74] Y. Ingster and I. A. Suslina. *Nonparametric goodness-of-fit testing under Gaussian models*. Vol. 169. Springer Science & Business Media, 2012.

[75] O. V. Lepski and V. G. Spokoiny. "Minimax nonparametric hypothesis testing: the case of an inhomogeneous alternative". In: (1999).

[76] M. S. Ermakov. "Minimax detection of a signal in a Gaussian white noise". In: *Theory of Probability & Its Applications* 35.4 (1991), pp. 667–679.

[77] Y. Wei and M. J. Wainwright. "The local geometry of testing in ellipses: Tight control via localized kolmogorov widths". In: *IEEE Transactions on Information Theory* 66.8 (2020), pp. 5110–5129.

[78] Y. Wei, M. J. Wainwright, and A. Guntuboyina. "The geometry of hypothesis testing over convex cones: Generalized likelihood ratio tests and minimax radii". In: (2019).

[79] O. Goldreich, S. Goldwasser, and D. Ron. "Property testing and its connection to learning and approximation". In: *Journal of the ACM (JACM)* 45.4 (1998), pp. 653–750.

[80] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. "Testing that distributions are close". In: *Proceedings 41st Annual Symposium on Foundations of Computer Science*. IEEE. 2000, pp. 259–269.

[81] L. Paninski. "A coincidence-based test for uniformity given very sparsely sampled discrete data". In: *IEEE Transactions on Information Theory* 54.10 (2008), pp. 4750–4755.

[82] G. Valiant and P. Valiant. "An automatic inequality prover and instance optimal identity testing". In: *SIAM Journal on Computing* 46.1 (2017), pp. 429–455.

[83] G. Valiant and P. Valiant. *Instance Optimal Distribution Testing and Learning*. 2020.

[84] C. L. Canonne and K. Wimmer. "Identity testing under label mismatch". In: *arXiv preprint arXiv:2105.01856* (2021).

[85] C. Canonne, S. B. Hopkins, J. Li, A. Liu, and S. Narayanan. "The full landscape of robust mean testing: Sharp separations between oblivious and adaptive contamination". In: *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2023, pp. 2159–2168.

[86] I. Diakonikolas, D. M. Kane, and A. Stewart. "Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures". In: *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2017, pp. 73–84.

[87] C. L. Canonne, X. Chen, G. Kamath, A. Levi, and E. Waingarten. "Random restrictions of high dimensional distributions and uniformity testing with subcube conditioning". In: *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM. 2021, pp. 321–336.

[88] I. Diakonikolas, D. M. Kane, and A. Pensia. "Gaussian mean testing made simple". In: *Symposium on Simplicity in Algorithms (SOSA)*. SIAM. 2023, pp. 348–352.

[89] C. L. Canonne. *Topics and techniques in distribution testing*. now Publishers, 2022.

[90] M. Gutman. "Asymptotically optimal classification for multiple tests with empirically observed statistics". In: *IEEE Transactions on Information Theory* 35.2 (1989), pp. 401–408.

[91] J. Ziv. "On classification with empirically observed statistics and universal data compression". In: *IEEE Transactions on Information Theory* 34.2 (1988), pp. 278–286.

[92] L. Zhou, V. Y. F. Tan, and M. Motani. "Second-Order Asymptotically Optimal Statistical Classification". In: *2019 IEEE International Symposium on Information Theory (ISIT)*. 2019, pp. 231–235. DOI: 10.1109/ISIT.2019.8849721.

[93] H.-W. Hsu and I.-H. Wang. "On binary statistical classification from mismatched empirically observed statistics". In: *2020 IEEE International Symposium on Information Theory (ISIT)*. IEEE. 2020, pp. 2533–2538.

[94] H. He, L. Zhou, and V. Y. Tan. "Distributed detection with empirically observed statistics". In: *IEEE Transactions on Information Theory* 66.7 (2020), pp. 4349–4367.

[95] M. Haghifam, V. Y. Tan, and A. Khisti. "Sequential classification with empirically observed statistics". In: *IEEE Transactions on Information Theory* 67.5 (2021), pp. 3095–3113.

[96] P. Boroumand and A. Guillén i Fàbregas. "Universal Neyman-Pearson Classification with a Known Hypothesis". In: *arXiv preprint arXiv:2206.11700* (2022).

[97] T. M. Cover. "Universal gambling schemes and the complexity measures of kolmogorov and chaitin". In: *Technical Report, no. 12* (1974).

[98] J. Rissanen. "A universal data compression system". In: *IEEE Transactions on information theory* 29.5 (1983), pp. 656–664.

[99] Y. M. Shtar'kov. "Universal sequential coding of single messages". In: *Problemy Peredachi Informatsii* 23.3 (1987), pp. 3–17.

[100] T. M. Cover. "Universal portfolios". In: *Mathematical finance* 1.1 (1991), pp. 1–29.

[101] N. Merhav and M. Feder. "Universal schemes for sequential decision from individual data sequences". In: *IEEE Transactions on Information Theory* 39.4 (1993), pp. 1280–1292.

[102] N. Merhav and M. Feder. "Universal prediction". In: *IEEE Transactions on Information Theory* 44.6 (1998), pp. 2124–2147.

[103] M. Opper and D. Haussler. "Worst case prediction over sequences under log loss". In: *The Mathematics of Information Coding, Extraction and Distribution*. Springer, 1999, pp. 81–90.

[104] S. A. Geer. *Empirical Processes in M-estimation*. Vol. 6. Cambridge university press, 2000.

[105] T. Liang, A. Rakhlin, and K. Sridharan. "Learning with Square Loss: Localization through Offset Rademacher Complexity". In: *Proceedings of The 28th Conference on Learning Theory*. 2015, pp. 1260–1285.

[106] J. Mourtada. "Universal coding, intrinsic volumes, and metric complexity". In: *arXiv preprint arXiv:2303.07279* (2023).

[107] L. LeCam. "Convergence of estimates under dimensionality restrictions". In: *The Annals of Statistics* 1.1 (1973), pp. 38–53.

[108] B. Bilodeau, D. J. Foster, and D. M. Roy. "Minimax rates for conditional density estimation via empirical entropy". In: *The Annals of Statistics* 51.2 (2023), pp. 762–790.

[109] A. K. Rakhlin and K. Sridharan. "Online non-parametric regression". In: *Conference on Learning Theory*. PMLR. 2014, pp. 1232–1264.

[110] Z. Jia, Y. Polyanskiy, and A. Rakhlin. "A modified scale-sensitive dimension and lower bounds for offset sequential Rademacher processes". In: *arXiv preprint arXiv:2503.11183* (2025).

[111] I. Csiszar and P. C. Shields. "Redundancy rates for renewal and other processes". In: *IEEE Transactions on Information theory* 42.6 (1996), pp. 2065–2072.

[112] Z. Liu, I. Attias, and D. M. Roy. "Sequential probability assignment with contexts: Minimax regret, contextual shtarkov sums, and contextual normalized maximum likelihood". In: *arXiv preprint arXiv:2410.03849* (2024).

[113] A. Rakhlin, K. Sridharan, and A. B. Tsybakov. "Empirical entropy, minimax regret and minimax risk". In: *Bernoulli* 23.2 (2017), pp. 789–824.

[114] C. Wu, M. Heidari, A. Grama, and W. Szpankowski. "Precise regret bounds for log-loss via a truncated bayesian algorithm". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 26903–26914.

[115] D. J. Foster and A. Krishnamurthy. "Efficient first-order contextual bandits: Prediction, allocation, and triangular discrimination". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 18907–18919.

[116] A. K. Rakhlin, K. Sridharan, and A. Tewari. "Online learning: Stochastic, constrained, and smoothed adversaries". In: *Advances in neural information processing systems* 24 (2011).

[117] R. M. Dudley. "Central limit theorems for empirical measures". In: *The Annals of Probability* (1978), pp. 899–929.

[118] A. K. Rakhlin, K. Sridharan, and A. Tewari. "Sequential complexities and uniform martingale laws of large numbers". In: *Probability theory and related fields* 161 (2015), pp. 111–153.

[119] V. N. Vapnik and A. Y. Chervonenkis. "On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities". In: *Theory of Probability and its Applications* 16.2 (1971), pp. 264–280.

[120] M. J. Kearns and R. E. Schapire. "Efficient distribution-free learning of probabilistic concepts". In: *Journal of Computer and System Sciences* 48.3 (1994), pp. 464–497.

[121] S. Ben-David, D. Pál, and S. Shalev-Shwartz. "Agnostic Online Learning." In: *COLT*. Vol. 3. 2009, p. 1.

[122] B. K. Natarajan and P. Tadepalli. "Two new frameworks for learning". In: *Machine Learning Proceedings 1988*. Elsevier, 1988, pp. 402–415.

[123] B. K. Natarajan. "On learning sets and functions". In: *Machine Learning* 4.1 (1989), pp. 67–97.

[124] P. L. Bartlett, O. Bousquet, and S. Mendelson. "Local rademacher complexities". In: *The Annals of Statistics* 33.4 (2005), pp. 1497–1537.

[125] A. Daniely and S. Shalev-Shwartz. "Optimal learners for multiclass problems". In: *Proceedings of The 27th Conference on Learning Theory*. Ed. by M. F. Balcan, V. Feldman, and C. Szepesvári. Vol. 35. Proceedings of Machine Learning Research. Barcelona, Spain: PMLR, 13–15 Jun 2014, pp. 287–316. URL: https://proceedings.mlr.press/v35/daniely14b.html.

[126] N. Brukhim, D. Carmon, I. Dinur, S. Moran, and A. Yehudayoff. "A characterization of multiclass learnability". In: *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2022, pp. 943–955.

[127] S. Hanneke, S. Moran, and Q. Zhang. "Universal Rates for Multiclass Learning". In: *The Thirty Sixth Annual Conference on Learning Theory*. PMLR. 2023, pp. 5615–5681.

[128] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. "Scale-sensitive dimensions, uniform convergence, and learnability". In: *Journal of the ACM* 44 (1997), pp. 615–631.

[129] M. Rudelson and R. Vershynin. "Combinatorics of Random Processes and Sections of Convex Bodies". In: *The Annals of Mathematics* 164.2 (2006), pp. 603–648.

[130] J. Qian, A. Rakhlin, and N. Zhivotovskiy. "Refined risk bounds for unbounded losses via transductive priors". In: *arXiv preprint arXiv:2410.21621* (2024).

[131] A. Rakhlin and K. Sridharan. "On Martingale Extensions of Vapnik–Chervonenkis Theory with Applications to Online Learning". In: Springer International Publishing, 2015, pp. 197–215.

[132] A. Rakhlin and K. Sridharan. "Combinatorial dimensions, uniform convergence, and prediction". In: *Conference for J. Michael Steele's 65th birthday*. 2015.

[133] D. Silver, S. Singh, D. Precup, and R. S. Sutton. "Reward is enough". In: *Artificial Intelligence* 299 (2021), p. 103535.

[134] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, et al. "Mastering the game of go without human knowledge". In: *nature* 550.7676 (2017), pp. 354–359.

[135] P. Wang, L. Li, Z. Shao, R. Xu, D. Dai, Y. Li, D. Chen, Y. Wu, and Z. Sui. "Math-shepherd: Verify and reinforce llms step-by-step without human annotations". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2024, pp. 9426–9439.

[136] L. Luo, Y. Liu, R. Liu, S. Phatale, H. Lara, Y. Li, L. Shu, Y. Zhu, L. Meng, J. Sun, et al. "Improve Mathematical Reasoning in Language Models by Automated Process Supervision". In: *arXiv preprint arXiv:2406.06592* (2024).

[137] H. Zhong, G. Feng, W. Xiong, X. Cheng, L. Zhao, D. He, J. Bian, and L. Wang. "Dpo meets ppo: Reinforced token optimization for rlhf". In: *arXiv preprint arXiv:2404.18922* (2024).

[138] L. Yuan, W. Li, H. Chen, G. Cui, N. Ding, K. Zhang, B. Zhou, Z. Liu, and H. Peng. "Free Process Rewards without Process Labels". In: *arXiv preprint arXiv:2412.01981* (2024).

[139] A. Setlur, C. Nagpal, A. Fisch, X. Geng, J. Eisenstein, R. Agarwal, A. Agarwal, J. Berant, and A. Kumar. "Rewarding Progress: Scaling Automated Process Verifiers for LLM Reasoning". In: *arXiv preprint arXiv:2410.08146* (2024).

[140] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, et al. "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning". In: *arXiv preprint arXiv:2501.12948* (2025).

[141] M. Uehara, J. Huang, and N. Jiang. "Minimax weight and q-function learning for off-policy evaluation". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 9659–9668.

[142] T. Xie, C.-A. Cheng, N. Jiang, P. Mineiro, and A. Agarwal. "Bellman-consistent pessimism for offline reinforcement learning". In: *Advances in neural information processing systems* 34 (2021), pp. 6683–6694.

[143] K. Dong and T. Ma. "First steps toward understanding the extrapolation of nonlinear models to unseen domains". In: *arXiv preprint arXiv:2211.11719* (2022).

[144] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn. "Direct preference optimization: Your language model is secretly a reward model". In: *Advances in Neural Information Processing Systems* 36 (2023).

[145] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. "Deep reinforcement learning from human preferences". In: *Advances in neural information processing systems* 30 (2017).

[146] R. Munos. "Error bounds for approximate policy iteration". In: *ICML*. Vol. 3. Citeseer. 2003, pp. 560–567.

[147] A. Antos, C. Szepesvári, and R. Munos. "Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path". In: *Machine Learning* 71 (2008), pp. 89–129.

[148] A.-m. Farahmand, C. Szepesvári, and R. Munos. "Error propagation for approximate policy and value iteration". In: *Advances in neural information processing systems* 23 (2010).

[149] J. Chen and N. Jiang. "Information-theoretic considerations in batch reinforcement learning". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 1042–1051.

[150] Y. Jin, Z. Yang, and Z. Wang. "Is pessimism provably efficient for offline rl?" In: *International Conference on Machine Learning*. PMLR. 2021, pp. 5084–5096.

[151] T. Xie and N. Jiang. "Batch value-function approximation with only realizability". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 11404–11413.

[152] M. Bhardwaj, T. Xie, B. Boots, N. Jiang, and C.-A. Cheng. "Adversarial model for offline reinforcement learning". In: *Advances in Neural Information Processing Systems* 36 (2023).

[153] T. Xie, D. J. Foster, Y. Bai, N. Jiang, and S. M. Kakade. "The role of coverage in online reinforcement learning". In: *arXiv preprint arXiv:2210.04157* (2022).

[154] F. Liu, L. Viano, and V. Cevher. "What can online reinforcement learning with function approximation benefit from general coverage conditions?" In: *International Conference on Machine Learning*. PMLR. 2023, pp. 22063–22091.

[155] P. Amortila, D. J. Foster, N. Jiang, A. Sekhari, and T. Xie. "Harnessing density ratios for online reinforcement learning". In: *arXiv preprint arXiv:2401.09681* (2024).

[156] P. Amortila, D. J. Foster, and A. Krishnamurthy. "Scalable Online Exploration via Coverability". In: *arXiv preprint arXiv:2403.06571* (2024).

[157] Q. Liu, L. Li, Z. Tang, and D. Zhou. "Breaking the curse of horizon: Infinite-horizon off-policy estimation". In: *Advances in neural information processing systems* 31 (2018).

[158] T. Xie, Y. Ma, and Y.-X. Wang. "Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling". In: *Advances in neural information processing systems* 32 (2019).

[159] O. Nachum, Y. Chow, B. Dai, and L. Li. "Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections". In: *Advances in neural information processing systems* 32 (2019).

[160] N. Jiang and T. Xie. "Offline reinforcement learning in large state spaces: Algorithms and guarantees". In: *Statistical Science* (2024).

[161] T. Xie, M. Bhardwaj, N. Jiang, and C.-A. Cheng. "Armor: A model-based framework for improving arbitrary baseline policies with offline data". In: *arXiv preprint arXiv:2211.04538* (2022).

[162] C.-A. Cheng, T. Xie, N. Jiang, and A. Agarwal. "Adversarially trained actor critic for offline reinforcement learning". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 3852–3878.

[163] M. Uehara and W. Sun. "Pessimistic model-based offline reinforcement learning under partial coverage". In: *arXiv preprint arXiv:2107.06226* (2021).

[164] S. Levine, A. Kumar, G. Tucker, and J. Fu. "Offline reinforcement learning: Tutorial, review, and perspectives on open problems". In: *arXiv preprint arXiv:2005.01643* (2020).

[165] R. Munos and C. Szepesvári. "Finite-Time Bounds for Fitted Value Iteration." In: *Journal of Machine Learning Research* 9.5 (2008).

[166] T. Xie and N. Jiang. "Q* approximation schemes for batch reinforcement learning: A theoretical comparison". In: *Conference on Uncertainty in Artificial Intelligence*. PMLR. 2020, pp. 550–559.

[167] T. Xie, N. Jiang, H. Wang, C. Xiong, and Y. Bai. "Policy finetuning: Bridging sample-efficient offline and online reinforcement learning". In: *Advances in neural information processing systems* 34 (2021), pp. 27395–27407.

[168] W. B. Knox and P. Stone. "Tamer: Training an agent manually via evaluative reinforcement". In: *2008 7th IEEE international conference on development and learning*. IEEE. 2008, pp. 292–297.

[169] R. Akrour, M. Schoenauer, and M. Sebag. "April: Active preference learning-based reinforcement learning". In: *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23*. Springer. 2012, pp. 116–131.

[170] C. Wirth, R. Akrour, G. Neumann, J. Fürnkranz, et al. "A survey of preference-based reinforcement learning methods". In: *Journal of Machine Learning Research* 18.136 (2017), pp. 1–46.

[171] S. Griffith, K. Subramanian, J. Scholz, C. L. Isbell, and A. L. Thomaz. "Policy shaping: Integrating human feedback with reinforcement learning". In: *Advances in neural information processing systems* 26 (2013).

[172] N. Stiennon, L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano. "Learning to summarize with human feedback". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 3008–3021.

[173] W. Zhan, M. Uehara, N. Kallus, J. D. Lee, and W. Sun. "Provable offline preference-based reinforcement learning". In: *arXiv preprint arXiv:2305.14816* (2023).

[174] Z. Liu, M. Lu, S. Zhang, B. Liu, H. Guo, Y. Yang, J. Blanchet, and Z. Wang. "Provably mitigating overoptimization in rlhf: Your sft loss is implicitly an adversarial regularizer". In: *arXiv preprint arXiv:2405.16436* (2024).

[175] Y. Song, G. Swamy, A. Singh, D. Bagnell, and W. Sun. "The importance of online data: Understanding preference fine-tuning via coverage". In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. 2024.

[176] S. Zhang, D. Yu, H. Sharma, H. Zhong, Z. Liu, Z. Yang, S. Wang, H. Hassan, and Z. Wang. "Self-exploring language models: Active preference elicitation for online alignment". In: *arXiv preprint arXiv:2405.19332* (2024).

[177] T. Xie, D. J. Foster, A. Krishnamurthy, C. Rosset, A. Awadallah, and A. Rakhlin. "Exploratory Preference Optimization: Harnessing Implicit Q*-Approximation for Sample-Efficient RLHF". In: *arXiv preprint arXiv:2405.21046* (2024).

[178] R. A. Bradley and M. E. Terry. "Rank analysis of incomplete block designs: I. The method of paired comparisons". In: *Biometrika* 39.3/4 (1952), pp. 324–345.

[179] R. Rafailov, J. Hejna, R. Park, and C. Finn. "From r to Q^*: Your Language Model is Secretly a Q-Function". In: *arXiv preprint arXiv:2404.12358* (2024).

[180] J. Hejna, R. Rafailov, H. Sikchi, C. Finn, S. Niekum, W. B. Knox, and D. Sadigh. "Contrastive prefence learning: Learning from human feedback without rl". In: *arXiv preprint arXiv:2310.13639* (2023).

[181] W. Xiong, H. Dong, C. Ye, H. Zhong, N. Jiang, and T. Zhang. "Gibbs sampling from human feedback: A provable kl-constrained framework for rlhf". In: *arXiv preprint arXiv:2312.11456* (2023).

[182] C. Ye, W. Xiong, Y. Zhang, N. Jiang, and T. Zhang. "A theoretical analysis of nash learning from human feedback under general kl-regularized preference". In: *arXiv preprint arXiv:2402.07314* (2024).

[183] A. Agarwal, N. Jiang, S. M. Kakade, and W. Sun. "Reinforcement learning: Theory and algorithms". In: *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep* 32 (2019), p. 96.

[184] T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[185] D. J. Foster and A. Rakhlin. "Foundations of reinforcement learning and interactive decision making". In: *arXiv preprint arXiv:2312.16730* (2023).

[186] C. Rosset, C.-A. Cheng, A. Mitra, M. Santacroce, A. Awadallah, and T. Xie. "Direct nash optimization: Teaching language models to self-improve with general preferences". In: *arXiv preprint arXiv:2404.03715* (2024).

[187] J. Watson, S. Huang, and N. Heess. "Coherent soft imitation learning". In: *Advances in Neural Information Processing Systems* 36 (2023).

[188] B. Zhu, M. Jordan, and J. Jiao. "Principled reinforcement learning with human feedback from pairwise or k-wise comparisons". In: *International Conference on Machine Learning*. PMLR. 2023, pp. 43037–43067.

[189] N. Das, S. Chakraborty, A. Pacchiano, and S. R. Chowdhury. "Provably sample efficient rlhf via active preference optimization". In: *arXiv preprint arXiv:2402.10500* (2024).

[190] S. Kakade and J. Langford. "Approximately optimal approximate reinforcement learning". In: *ICML*. Vol. 2. 2002, pp. 267–274.

[191] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, et al. "Training verifiers to solve math word problems". In: *arXiv preprint arXiv:2110.14168* (2021).

[192] N. Lambert, V. Pyatkin, J. Morrison, L. Miranda, B. Y. Lin, K. Chandu, N. Dziri, S. Kumar, T. Zick, Y. Choi, et al. "Rewardbench: Evaluating reward models for language modeling". In: *arXiv preprint arXiv:2403.13787* (2024).

[193] G. Neu and G. Bartók. "An efficient algorithm for learning with semi-bandit feedback". In: *International Conference on Algorithmic Learning Theory*. Springer. 2013, pp. 234–248.

[194] A. Y. Ng, D. Harada, and S. Russell. "Policy invariance under reward transformations: Theory and application to reward shaping". In: *International Conference on Machine Learning*. Vol. 99. 1999, pp. 278–287.

[195] A. Trott, S. Zheng, C. Xiong, and R. Socher. "Keeping your distance: Solving sparse reward tasks using self-balancing shaped rewards". In: *Advances in Neural Information Processing Systems* 32 (2019).

[196] A. Gupta, A. Pacchiano, Y. Zhai, S. Kakade, and S. Levine. "Unpacking reward shaping: Understanding the benefits of reward engineering on sample complexity". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 15281–15295.

[197] D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell. "Curiosity-driven exploration by self-supervised prediction". In: *International conference on machine learning*. PMLR. 2017, pp. 2778–2787.

[198] Y. Burda, H. Edwards, A. Storkey, and O. Klimov. "Exploration by random network distillation". In: *arXiv preprint arXiv:1810.12894* (2018).

[199] Z. Jia, A. Rakhlin, and T. Xie. "Do we need to verify step by step? rethinking process supervision from a theoretical perspective". In: *arXiv preprint arXiv:2502.10581* (2025).

[200] C. Jin, Q. Liu, and S. Miryoosefi. "Bellman eluder dimension: New rich classes of RL problems, and sample-efficient algorithms". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 13406–13418.

[201] S. Du, S. Kakade, J. Lee, S. Lovett, G. Mahajan, W. Sun, and R. Wang. "Bilinear classes: A structural framework for provable generalization in RL". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 2826–2836.

[202] A. Zanette, A. Lazaric, M. Kochenderfer, and E. Brunskill. "Learning near optimal policies with low inherent bellman error". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 10978–10989.

[203] D. J. Foster, S. M. Kakade, J. Qian, and A. Rakhlin. "The statistical complexity of interactive decision making". In: *arXiv preprint arXiv:2112.13487* (2021).

[204] D. J. Foster, A. Rakhlin, A. Sekhari, and K. Sridharan. "On the complexity of adversarial decision making". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 35404–35417.

[205] D. Russo and B. Van Roy. "Eluder dimension and the sample complexity of optimistic exploration". In: *Advances in Neural Information Processing Systems*. Vol. 26. 2013, pp. 2256–2264.

[206] K. Dong, J. Yang, and T. Ma. "Provable model-based nonlinear bandit and reinforcement learning: Shelve optimism, embrace virtual curvature". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 26168–26182.

[207] G. Li, P. Kamath, D. J. Foster, and N. Srebro. "Understanding the Eluder Dimension". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 23737–23750.

[208] N. Jiang, A. Krishnamurthy, A. Agarwal, J. Langford, and R. E. Schapire. "Contextual decision processes with low bellman rank are pac-learnable". In: *International Conference on Machine Learning*. PMLR. 2017, pp. 1704–1713.

[209] T. Van Erven and P. Harremos. "Rényi divergence and Kullback-Leibler divergence". In: *IEEE Transactions on Information Theory* 60.7 (2014), pp. 3797–3820.

[210] H. L. Van Trees. *Detection, estimation, and modulation theory, part I: detection, estimation, and linear modulation theory*. John Wiley & Sons, 2004.

[211] Y. Polyanskiy and Y. Wu. *Information theory: From coding to learning*. Cambridge university press, 2025.

[212] R. Vershynin. *High-dimensional probability: An introduction with applications in data science*. Vol. 47. Cambridge University Press, 2018.

[213] J. Abernethy, A. Agarwal, P. L. Bartlett, and A. Rakhlin. "A stochastic view of optimal regret through minimax duality". In: *arXiv preprint arXiv:0903.5328* (2009).

[214] A. K. Rakhlin, K. Sridharan, and A. Tewari. "Online learning via sequential complexities." In: *J. Mach. Learn. Res.* 16.1 (2015), pp. 155–186.

[215] D. J. Foster, S. Kale, H. Luo, M. Mohri, and K. Sridharan. "Logistic regression: The importance of being improper". In: *Conference on learning theory*. PMLR. 2018, pp. 167–208.

[216] E. Giné and J. Zinn. "Some limit theorems for empirical processes". In: *The Annals of Probability* (1984), pp. 929–989.

[217] R. M. Dudley. "The sizes of compact subsets of Hilbert space and continuity of Gaussian processes". In: *Journal of Functional Analysis* 1.3 (1967), pp. 290–330.

[218] J. v. Neumann. "Zur theorie der gesellschaftsspiele". In: *Mathematische annalen* 100.1 (1928), pp. 295–320.

[219] Y. Polyanskiy and Y. Wu. "Lecture notes on information theory". In: *Lecture Notes for ECE563 (UIUC) and* 6.2012-2016 (2014), p. 7.

[220] K. Fan. "Minimax theorems". In: *Proceedings of the National Academy of Sciences* 39.1 (1953), pp. 42–47.

[221] Y. Polyanskiy and Y. Wu. *Information theory: From coding to learning*. Cambridge university press, 2024.

[222] U. Haagerup. "The best constants in the khintchine inequality". In: *Studia Mathematica* 70.3 (1981), pp. 231–283.

[223] D. Berend and A. Kontorovich. "A sharp estimate of the binomial mean absolute deviation with applications". In: *Statistics & Probability Letters* 83.4 (2013), pp. 1254–1259.

[224] A. Rakhlin. *Mathematical Statistics: A Non-Asymptotic Approach*. May 2024. URL: https://www.mit.edu/~rakhlin/course_mathstat.pdf.

[225] I. Osband and B. Van Roy. "Model-based reinforcement learning and the eluder dimension". In: *Advances in Neural Information Processing Systems*. Vol. 27. 2014, pp. 1466–1474.

[226] W. Sun, N. Jiang, A. Krishnamurthy, A. Agarwal, and J. Langford. "Model-based RL in contextual decision processes: PAC bounds and exponential improvements over model-free approaches". In: *Conference on learning theory*. PMLR. 2019, pp. 2898–2933.

[227] Y. Xu, R. Wang, L. Yang, A. Singh, and A. Dubrawski. "Preference-based reinforcement learning with finite-time guarantees". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 18784–18794.

[228] E. Novoseller, Y. Wei, Y. Sui, Y. Yue, and J. Burdick. "Dueling posterior sampling for preference-based reinforcement learning". In: *Conference on Uncertainty in Artificial Intelligence*. PMLR. 2020, pp. 1029–1038.

[229] S. Cen, J. Mei, K. Goshvadi, H. Dai, T. Yang, S. Yang, D. Schuurmans, Y. Chi, and B. Dai. "Value-incentivized preference optimization: A unified approach to online and offline rlhf". In: *arXiv preprint arXiv:2405.19320* (2024).

[230] A. Beygelzimer, J. Langford, L. Li, L. Reyzin, and R. Schapire. "Contextual bandit algorithms with supervised learning guarantees". In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings. 2011, pp. 19–26.

[231] T. Zhang. "Covering number bounds of certain regularized linear function classes". In: *Journal of Machine Learning Research* 2.Mar (2002), pp. 527–550.

[232] F. Chen, C. Daskalakis, N. Golowich, and A. Rakhlin. "Near-optimal learning and planning in separated latent mdps". In: *The Thirty Seventh Annual Conference on Learning Theory*. PMLR. 2024, pp. 995–1067.

[233] V. Dani, T. P. Hayes, and S. M. Kakade. "Stochastic linear optimization under bandit feedback". In: *21st Annual Conference on Learning Theory*. 101. 2008, pp. 355–366.