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

Zeyu Jia ZYJIA@MIT.EDU
Yury Polyanskiy YP@MIT.EDU
Alexander Rakhlin RAKHLIN@MIT.EDU
Massachusetts Institute of Technology

Editors: Nika Haghtalab and Ankur Moitra

Abstract

We study the problem of sequential probability assignment under logarithmic loss, both with and without side information. Our objective is to analyze the *minimax regret*—a notion extensively studied in the literature—in terms of geometric quantities, such as covering numbers and scale-sensitive dimensions. We show that the minimax regret for the case of no side information (equivalently, the Shtarkov sum) can be upper bounded in terms of *sequential square-root entropy*, a notion closely related to Hellinger distance. For the problem of sequential probability assignment with side information, we develop both upper and lower bounds based on the aforementioned entropy. The lower bound matches the upper bound, up to log factors, for classes in the Donsker regime (according to our definition of entropy).

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 Cesa-Bianchi and Lugosi (2006). 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

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

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

$$\sum_{t=1}^{n} -\log \widehat{p}_t(y_t) - \inf_{\mathbf{q}\in\mathcal{Q}} \sum_{t=1}^{n} -\log q_t(y_t|x_t) \tag{2}$$

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

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

$$\mathcal{R}_n(\mathcal{Q}) \coloneqq \inf_{\widehat{\mathbf{p}}\in\Delta(\mathcal{Y}^n)} \sup_{\mathbf{y}\in\mathcal{Y}^n} \sup_{\mathbf{q}\in\mathcal{Q}} \log\left(\frac{\mathbf{q}(\mathbf{y})}{\widehat{\mathbf{p}}(\mathbf{y})}\right), \tag{3}$$

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 $(x_1, \dots, x_n)$ also form an individual (i.e. arbitrary) sequence; however, for brevity, we defer this definition to Section 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 Cover (1974); Rissanen (1983); Shtar'kov (1987); Cover (1991); Merhav and Feder (1993, 1998); Cesa-Bianchi and Lugosi (1999), and more recently in Bilodeau et al. (2020); Rakhlin and Sridharan (2015b), among many others.

In particular, the classical result of Shtar'kov (1987) states that in the non-contextual case, $\mathcal{R}_n(\mathcal{Q})$ has the following closed form:

$$\mathcal{R}_n(\mathcal{Q}) = \log \sum_{\mathbf{y}\in\mathcal{Y}^n} \sup_{\mathbf{q}\in\mathcal{Q}} \mathbf{q}(\mathbf{y}), \tag{4}$$

and the optimal strategy in Eq. (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 (4) is more succinct than (3), it is still not amenable to analysis with standard tools, except for special cases Cesa-Bianchi and Lugosi (2006).

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 Opper and Haussler (1999), who presented an upper bound involving a Dudley integral. This work was extended to a general $\mathcal{Q}$ in Cesa-Bianchi and Lugosi (1999), who observed that, owing to the equalizing property of the optimal strategy $\mathbf{p}^*$, the Shtarkov sum (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 van de Geer (2000). Notably, the process was shown to be subgaussian with respect to a pseudometric

$$d(f,g) = \left(\frac{1}{n}\sum_{t=1}^{n} \max_{y_{1:n}} (\log f(y_t|y_{1:t-1}) - \log g(y_t|y_{1:t-1}))^2\right)^{1/2}, \tag{5}$$

where $y_{1:t} := (y_1, \ldots, y_t)$. Cesa-Bianchi and Lugosi (1999) 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 (Cesa-Bianchi and Lugosi, 1999, Theorem 3) assumes a form typical of bounds obtained via localized or *offset Rademacher* complexities, as encountered in square loss regression van de Geer (2000); Liang et al. (2015); Mourtada (2023).

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, Rakhlin and Sridharan (2015b) 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 (Rakhlin and Sridharan, 2015b, Lemma 2).

The important work of Bilodeau et al. (2020) 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. Further, to analyze the contextual version of the problem, Liu et al. (2024) introduced the notion of a contextual Shtarkov sum, which is equivalent to the minimax regret; however, this equivalent reformulation does not offer guidance on how geometric concepts—such as covering numbers—can be employed.

The Hellinger distance has long been recognized as a convenient metric on the space of distributions LeCam (1973); Haussler et al. (1997); Yang and Barron (1999); van de Geer (2000); Bilodeau et al. (2023). 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 Cesa-Bianchi and Lugosi (1999), 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 (Rakhlin and Sridharan, 2014, 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 Jia et al. (2025).

We now summarize our contributions.

1.2. Summary of Main Results

We study minimax regret in both non-contextual (Section 2) and contextual (Section 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 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 (3) (and, hence, the Shtarkov sum (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 (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

$$\mathcal{R}_n(\mathcal{Q}) = \begin{cases} \tilde{\mathcal{O}}\left(n^{\frac{p}{p+2}}\right) & \text{if } 0 < p \leq 2, \\ \tilde{\mathcal{O}}\left(n^{\frac{p-1}{p}}\right) & \text{if } p > 2. \end{cases} \tag{6}$$

In addition, we establish a lower bound demonstrating the tightness of (6) 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 Rakhlin and Sridharan (2015b).

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 Rakhlin and Sridharan (2015b) 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).

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.

1.4. Organization

In Section 2, we present our results in upper bounding the Shtarkov sum using sequential square-root entropy. In Section 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 4 we provide a proof sketch of our main result, Theorem 2. All the technical proofs are deferred to the appendix.

2. Upper Bound for Shtarkov Sum through Sequential Square-Root Covering

In this section, we upper bound the minimax regret Eq. (3) or Shtarkov sum Eq. (4) in terms of the ℓ_∞ sequential square-root covering defined as follows.

Definition 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*

$$\sup_{\mathbf{q}\in\mathcal{Q}} \max_{\mathbf{w}\in\mathcal{Y}^n} \min_{\mathbf{v}\in\mathcal{V}} \max_{t\in[n]} \max_{y\in\mathcal{Y}} \left| \sqrt{q_t(y \mid \mathbf{w})} - \sqrt{v_t(y \mid \mathbf{w})} \right| \le \alpha. \tag{7}$$

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 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 2 is deferred to Section B, and we provide a sketch of the proof in Section 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 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*

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

2.1. Comparison with Previous Results

We compare our results with Cesa-Bianchi and Lugosi (1999, 2006), which also provide an upper bound on the minimax regret using entropy. Taking (5) 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. Cesa-Bianchi and Lugosi (1999) establish that

$$\mathcal{R}_n(\mathcal{Q}) \lesssim \inf_{\gamma > 0}\left\{\sqrt{n}\int_0^\gamma \sqrt{\mathcal{H}_{\mathsf{log}}(\mathcal{F},\varepsilon,n)}d\varepsilon + \mathcal{H}_{\mathsf{log}}(\mathcal{F},\gamma,n)\right\}. \tag{8}$$

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 (7), we have

$$\sup_{\mathbf{q}\in\mathcal{Q}} \min_{\mathbf{v}\in\mathcal{V}} \max_{\mathbf{w}\in\mathcal{Y}^n} \max_{t\in[n]} \max_{y\in\mathcal{Y}} |\log q_t(y \mid \mathbf{w}) - \log v_t(y \mid \mathbf{w})| \leq \alpha. \tag{9}$$

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 (9) is clearly more stringent than (7) 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 $\sup_{y\in\mathcal{Y}} |\sqrt{p(y)} - \sqrt{q(y)}| \leq \sup_{y\in\mathcal{Y}} |\log p(y) - \log q(y)|$ 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 Cesa-Bianchi and Lugosi (1999), 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 Section F). It is known from Csiszar and Shields (1996) that minimax regret is $\Theta(\sqrt{n})$. In Section F, however, we show that the sequential square-root entropy is always lower bounded by $\Omega(n)$.

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 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

$$\ell(\widehat{p}_t, y_t) := -y_t \log \widehat{p}_t - (1 - y_t) \log(1 - \widehat{p}_t). \tag{10}$$

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 (2)) as

$$\mathcal{R}_n(\mathcal{F}, \widehat{p}_{1:n}, x_{1:n}, y_{1:n}) := \sum_{t=1}^{n} \ell(\widehat{p}_t, y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t). \tag{11}$$

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

$$\mathcal{R}_n(\mathcal{F}) = \sup_{x_1 \in \mathcal{X}} \inf_{\widehat{p}_1 \in [0,1]} \sup_{y_1 \in \{0,1\}} \cdots \sup_{x_n \in \mathcal{X}} \inf_{\widehat{p}_n \in [0,1]} \sup_{y_n \in \{0,1\}} \mathcal{R}(\mathcal{F}, \widehat{p}_{1:n}, x_{1:n}, y_{1:n}), \tag{12}$$

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 $\mathbf{x} = (x_1, \ldots, x_n)$ 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 4 *For any $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ which is convex in its first argument, and for any $\phi : \mathcal{Y}^n \times \mathcal{X}^n \to \mathbb{R}$,*

$$\left\{ \sup_{x_t} \inf_{\widehat{p}_t} \sup_{y_t} \right\}_{t=1}^{n} \left[\sum_{t=1}^{n} \ell(\widehat{p}_t, y_t) - \phi(y_{1:n}, x_{1:n}) \right]$$

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

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 (Liu et al., 2024, Theorem 3.2). The proof of Lemma 4 is deferred to Section C.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 2 with $\mathcal{Q}_\mathbf{x} = \mathcal{F} \circ \mathbf{x}$.

3.1. Upper Bound with Sequential Square-Root Entropy

We now repeat the definition Definition 1, adapting it to the case of binary alphabet and the real-valued re-parametrization of probabilities:

Definition 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 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 $\mathbf{x}$.

1. Note that the optimal learning algorithm for this $\mathbf{x}$ tree is not guaranteed to be optimal for the actual problem (12).

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 $p \geq 0$.

Corollary 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 $p \geq 0$. The minimax regret $\mathcal{R}_n(\mathcal{F})$ is upper bounded by*

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

The proofs of Theorem 6 and Corollary 7 are deferred to Section C.1.

3.2. Comparison with Previous Upper Bounds

We compare our results to those of Rakhlin and Sridharan (2015b) and Bilodeau et al. (2020). These two works use the sequential entropy $\mathcal{H}_\infty(\mathcal{F}, \alpha, n, \mathbf{x})$, defined as

$$\mathcal{H}_\infty(\mathcal{F}, \alpha, n, \mathbf{x}) = \log \mathcal{N}_\infty(\mathcal{F} \circ \mathbf{x}, \alpha, n), \tag{13}$$

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

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

With proper choice of parameters, our results can recover the upper bounds in (Rakhlin and Sridharan, 2015b, Theorem 1 and Theorem 4). This follows from the next result relating sequential square-root entropy and the sequential entropy defined above.

Proposition 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 $\mathbf{x}$,*

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

For nonparametric class $\mathcal{F}$ which satisfies $\sup_{\mathbf{x}} \mathcal{H}_\infty(\mathcal{F}, \alpha, n, \mathbf{x}) \asymp \alpha^{-q}$ for some $q > 0$, (Bilodeau et al., 2020, Theorem 2) proves the following upper bound for the minimax regret:

$$\mathcal{R}_n(\mathcal{F}) = \mathcal{O}\left(n^{\frac{q}{q+1}}\right). \tag{14}$$

Corollary 7 recovers this result for $0 < q \leq 1$, up to logarithmic factors, via the following proposition.

Proposition 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*

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

The proofs of Proposition 8 and Proposition 9 are deferred to Section C.2.

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. (12) via sequential square-root entropy.

Theorem 10 *Suppose function class $\mathcal{F} \subseteq [0,1]^{\mathcal{X}}$ satisfies*

$$\sup_{\mathbf{x}} \mathcal{H}_{\mathsf{sq}}(\mathcal{F}, \alpha, n, \mathbf{x}) = \tilde{\Omega}\left(\alpha^{-p}\right), \tag{15}$$

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 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 Section D. Notice that according to Corollary 7 and Theorem 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 10 and the upper bound in Corollary 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 11 *Suppose the function class $\mathcal{F} \subseteq \{f : \mathcal{X} \to [7/16, 9/16]\}$ satisfies*

$$\sup_{\mathbf{x}} \mathcal{H}_{\mathsf{sq}}(\mathcal{F}, \alpha, n, \mathbf{x}) = \tilde{\Omega}\left(\alpha^{-p}\right). \tag{16}$$

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. (16) holds.

The proof of Theorem 11 is deferred to Section D. Comparing Theorem 11 and Corollary 7, we see a dichotomy between the regime of $0 \le p \le 2$ 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 Rakhlin and Sridharan (2014) and to misspecified regression with i.i.d. data Rakhlin et al. (2017).

3.4. Examples

In this section, we provide several examples to illustrate Theorem 6 and Corollary 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

$$\mathcal{F} = \left\{ f : f(x) = \frac{1 + \langle w, x \rangle}{2} \text{ for some } w \in B_2(1) \right\}. \tag{17}$$

This class is generally viewed as 'hard case' in existing literature. Rakhlin and Sridharan (2015b) 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 Bilodeau et al. (2020); Wu et al. (2022) 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:

$$\mathcal{F}_{1/n} = \left\{ f : f(x) = \frac{1 + \langle w, x \rangle}{2} \text{ for some } w \in B_2(1 - 1/n) \right\}. \tag{18}$$

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

Lemma 12 *For $\mathcal{F}$ and $\mathcal{F}_{1/n}$ defined in Eq. (17) and Eq. (18), the minimax regret satisfies*

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

The proof of Lemma 12 is deferred to Section E. Equipped with this lemma, we only need to bound the sequential square-root entropy of function class $\mathcal{F}_{1/n}$.

Proposition 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 Proposition 13 is deferred to Section E. As a consequence, in view of Corollary 7, we conclude:

Corollary 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 Bilodeau et al. (2020); Wu et al. (2022); Foster and Krishnamurthy (2021), among others. In this case, the context set is $\mathcal{X} = [0,1]$, and the function class $\mathcal{F}$ is defined to be

$$\mathcal{F} = \{f : [0,1] \to [0,1], f \text{ is 1-Lipschitz}\}. \tag{19}$$

In Bilodeau et al. (2020), 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 Proposition 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 (Bilodeau et al., 2020, 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 15 *For $\mathcal{X} = [0,1]$ and Lipschitz function class $\mathcal{F}$ defined in Eq. (19),*

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

4. Proof Sketch of Theorem 2

In this section, we sketch the proof of Theorem 2. The detailed proof is deferred to Section B. 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 Bilodeau et al. (2020); Rakhlin and Sridharan (2015b):

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

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 Bilodeau et al. (2020).

(ii) **Nonnegativity of divergence:** $\mathbb{E}_{y\sim p}\left\{-\zeta(f(y)/p(y)) - c\cdot\zeta(f(y)/p(y))^2\right\} \ge 0$ for any $f, p \in \Delta(\mathcal{Y})$. This property is inspired by the proof of Cesa-Bianchi and Lugosi (1999) 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 Section B.1.3. As a consequence, we obtain the following inequality after a sequential symmetrization argument:

$$\begin{aligned}
\mathbb{E}_{\mathbf{y}\sim\mathbf{p}} \mathcal{R}_n(\mathcal{Q}, \mathbf{p}, \mathbf{y}) &= \mathbb{E}_{\mathbf{w}\sim\mathbf{p}}\left[\sup_{\mathbf{q}\in\mathcal{Q}} \sum_{t=1}^n \left(\log q_t(w_t \mid \mathbf{w}) - \log p_t(w_t \mid \mathbf{w})\right)\right] \\
&\le \mathbb{E}_{\mathbf{w}\sim\mathbf{p}}\left[\sup_{\mathbf{q}\in\mathcal{Q}} \sum_{t=1}^n \left\{\zeta\left(\frac{q_t(w_t \mid \mathbf{w})}{p_t(w_t \mid \mathbf{w})}\right) - c\cdot\zeta\left(\frac{q_t(w_t \mid \mathbf{w})}{p_t(w_t \mid \mathbf{w})}\right)^2\right\}\right] \\
&\le \mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}} \sum_{t=1}^n \varepsilon_t \zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - c\cdot\zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)^2\right] \\
&\quad + \mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}} \sum_{t=1}^n (-\varepsilon_t)\zeta\left(\frac{q_t(z_t \mid \mathbf{w})}{p_t(z_t \mid \mathbf{w})}\right) - c\cdot\zeta\left(\frac{q_t(z_t \mid \mathbf{w})}{p_t(z_t \mid \mathbf{w})}\right)^2\right], \qquad (20)
\end{aligned}$$

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 Rakhlin et al. (2011).

Analysis through chaining technique Finally, to upper bound the right hand side of Eq. (20), we adopt the chaining technique Dudley (1978); Rakhlin and Sridharan (2014); Rakhlin et al. (2015b); Rakhlin and Sridharan (2015b). We sketch the beginning of the argument. The first term (and, analogously, the second term) in (20) can be decomposed through a chain of N approximating representatives as

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\sum_{t=1}^{n}\varepsilon_t\left\{\zeta\left(\frac{q_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)-c\cdot\zeta\left(\frac{q_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)^2\right\}\right] \qquad (21)\\
&\le \mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\sum_{t=1}^{n}\varepsilon_t\left\{\zeta\left(\frac{q_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)-\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_1](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)\right\}\right]\\
&\quad+\sum_{i=1}^{N-1}\mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\sum_{t=1}^{n}\varepsilon_t\left\{\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_i](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)-\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_{i+1}](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)\right\}\right]\\
&\quad+\mathbb{E}\left[\sup_{\mathbf{v}\in\mathcal{V}(\alpha_N)\cup\{\mathbf{p}\}}\sum_{t=1}^{n}\left\{\varepsilon_t\zeta\left(\frac{v_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)-\frac{c}{4}\cdot\zeta\left(\frac{v_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)^2\right\}\right].
\end{aligned}$$

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 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 (Lemma 16, Lemma 17), 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.

Acknowledgements

ZJ and AR acknowledge support of the Simons Foundation and the NSF through awards DMS-2031883 and PHY-2019786, as well as from the ARO through award W911NF-21-1-0328.

References

Jacob Abernethy, Alekh Agarwal, Peter L Bartlett, and Alexander Rakhlin. A stochastic view of optimal regret through minimax duality. *arXiv preprint arXiv:0903.5328*, 2009.

Daniel Berend and Aryeh Kontorovich. A sharp estimate of the binomial mean absolute deviation with applications. *Statistics & Probability Letters*, 83(4):1254–1259, 2013.

Blair Bilodeau, Dylan Foster, and Daniel Roy. Tight bounds on minimax regret under logarithmic loss via self-concordance. In *International Conference on Machine Learning*, pages 919–929. PMLR, 2020.

Blair Bilodeau, Dylan J Foster, and Daniel M Roy. Minimax rates for conditional density estimation via empirical entropy. *The Annals of Statistics*, 51(2):762–790, 2023.

Nicolo Cesa-Bianchi and Gabor Lugosi. Minimax regret under log loss for general classes of experts. In *Proceedings of the Twelfth annual conference on computational learning theory*, pages 12–18, 1999.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Thomas M Cover. Universal gambling schemes and the complexity measures of kolmogorov and chaitin. *Technical Report, no. 12*, 1974.

Thomas M Cover. Universal portfolios. *Mathematical finance*, 1(1):1–29, 1991.

Imre Csiszar and Paul C Shields. Redundancy rates for renewal and other processes. *IEEE Transactions on Information theory*, 42(6):2065–2072, 1996.

Richard M Dudley. The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967.

Richard M Dudley. Central limit theorems for empirical measures. *The Annals of Probability*, pages 899–929, 1978.

Ky Fan. Minimax theorems. *Proceedings of the National Academy of Sciences*, 39(1):42–47, 1953.

Dylan J Foster and Akshay Krishnamurthy. Efficient first-order contextual bandits: Prediction, allocation, and triangular discrimination. *Advances in Neural Information Processing Systems*, 34:18907–18919, 2021.

Dylan J Foster, Satyen Kale, Haipeng Luo, Mehryar Mohri, and Karthik Sridharan. Logistic regression: The importance of being improper. In *Conference on learning theory*, pages 167–208. PMLR, 2018.

Evarist Giné and Joel Zinn. Some limit theorems for empirical processes. *The Annals of Probability*, pages 929–989, 1984.

Uffe Haagerup. The best constants in the khintchine inequality. *Studia Mathematica*, 70(3):231–283, 1981.

David Haussler, Manfred Opper, et al. Mutual information, metric entropy and cumulative relative entropy risk. *The Annals of Statistics*, 25(6):2451–2492, 1997.

Zeyu Jia, Yury Polyanskiy, and Alexander Rakhlin. A modified scale-sensitive dimension and lower bounds for offset sequential rademacher processes, 2025.

L. LeCam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, 1 (1):38–53, 1973.

Tengyuan Liang, Alexander Rakhlin, and Karthik Sridharan. Learning with square loss: Localization through offset rademacher complexity. In *Conference on Learning Theory*, pages 1260–1285. PMLR, 2015.

Ziyi Liu, Idan Attias, and Daniel M Roy. Sequential probability assignment with contexts: Minimax regret, contextual shtarkov sums, and contextual normalized maximum likelihood. *arXiv preprint arXiv:2410.03849*, 2024.

Neri Merhav and Meir Feder. Universal schemes for sequential decision from individual data sequences. *IEEE Transactions on Information Theory*, 39(4):1280–1292, 1993.

Neri Merhav and Meir Feder. Universal prediction. *IEEE Transactions on Information Theory*, 44 (6):2124–2147, 1998.

Jaouad Mourtada. Universal coding, intrinsic volumes, and metric complexity. *arXiv preprint arXiv:2303.07279*, 2023.

Manfred Opper and David Haussler. Worst case prediction over sequences under log loss. In *The Mathematics of Information Coding, Extraction and Distribution*, pages 81–90. Springer, 1999.

Yury Polyanskiy and Yihong Wu. Lecture notes on information theory. *Lecture Notes for ECE563 (UIUC) and*, 6(2012-2016):7, 2014.

Yury Polyanskiy and Yihong Wu. *Information theory: From coding to learning*. Cambridge university press, 2024.

Alexander Rakhlin. Mathematical statistics: A non-asymptotic approach, May 2024. URL https://www.mit.edu/~rakhlin/course_mathstat.pdf.

Alexander Rakhlin and Karthik Sridharan. Online non-parametric regression. In *Conference on Learning Theory*, pages 1232–1264. PMLR, 2014.

Alexander Rakhlin and Karthik Sridharan. Combinatorial dimensions, uniform convergence, and prediction, 2015a. URL http://www.mit.edu/~rakhlin/papers/talk_duke_steele.pdf. Conference for J. Michael Steele's 65th birthday.

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

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning: Stochastic, constrained, and smoothed adversaries. *Advances in neural information processing systems*, 24, 2011.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning via sequential complexities. *J. Mach. Learn. Res.*, 16(1):155–186, 2015a.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Sequential complexities and uniform martingale laws of large numbers. *Probability theory and related fields*, 161:111–153, 2015b.

Alexander Rakhlin, Karthik Sridharan, and Alexandre B Tsybakov. Empirical entropy, minimax regret and minimax risk. *Bernoulli*, 23(2):789–824, 2017.

Jorma Rissanen. A universal data compression system. *IEEE Transactions on information theory*, 29(5):656–664, 1983.

Yurii Mikhailovich Shtar'kov. Universal sequential coding of single messages. *Problemy Peredachi Informatsii*, 23(3):3–17, 1987.

J v. Neumann. Zur theorie der gesellschaftsspiele. *Mathematische annalen*, 100(1):295–320, 1928.

Sara van de Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

Changlong Wu, Mohsen Heidari, Ananth Grama, and Wojciech Szpankowski. Precise regret bounds for log-loss via a truncated bayesian algorithm. *Advances in Neural Information Processing Systems*, 35:26903–26914, 2022.

Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, pages 1564–1599, 1999.

Appendix A. Finite Class Lemmas

We first provide a version of (Rakhlin and Sridharan, 2014, Lemma 10).

Lemma 16 *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 $\mathcal{S}_1, \dots, \mathcal{S}_n$, we suppose $s_1, s_2, \dots, s_n$ are $\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_n$-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 $\mathbf{a} = (a_1, a_2, \dots, a_n)$ 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 $\mathbf{a} = (a_1, a_2, \dots, a_n)$.

Proof We observe that

$$\begin{aligned}
&\{\mathbb{E}_{s_t}\mathbb{E}_{\varepsilon_t}\}_{t=1}^n \left[\sup_{\mathbf{a}\in\mathcal{A}} \sum_{t=1}^n a_t(s_t)\varepsilon_t - \lambda a_t(s_t)^2\right] \\
&\overset{(i)}{\le} \frac{1}{2\lambda} \log \{\mathbb{E}_{s_t}\mathbb{E}_{\varepsilon_t}\}_{t=1}^n \sup_{\mathbf{a}\in\mathcal{A}} \left[\exp\left(2\lambda \sum_{t=1}^n a_t(s_t)\varepsilon_t - 2\lambda^2 a_t(s_t)^2\right)\right] \\
&\overset{(ii)}{\le} \frac{1}{2\lambda} \log \sum_{\mathbf{a}\in\mathcal{A}} \{\mathbb{E}_{s_t}\mathbb{E}_{\varepsilon_t}\}_{t=1}^n \exp\left(2\lambda \sum_{t=1}^n a_t(s_t)\varepsilon_t - 2\lambda^2 a_t(s_t)^2\right)
\end{aligned}$$

$$= \frac{1}{2\lambda} \log \sum_{\mathbf{a}\in\mathcal{A}} \{\mathbb{E}_{s_t}\mathbb{E}_{\varepsilon_t}\}_{t=1}^{n-1} \left[\exp\left(2\lambda \sum_{t=1}^{n-1} a_t(s_t)\varepsilon_t - 2\lambda^2 a_t(s_t)^2 \right) \right.$$
$$\left. \cdot\, \mathbb{E}_{s_n}\left[\exp\left(-2\lambda^2 a_n(s_n)^2\right) \left(\frac{\exp(2\lambda a_n(s_n))}{2} + \frac{\exp(-2\lambda a_n(s_n))}{2} \right) \mid \mathcal{G}_n \right] \right]$$
$$\overset{(iii)}{\leq} \frac{1}{2\lambda} \log \sum_{\mathbf{a}\in\mathcal{A}} \{\mathbb{E}_{s_t}\mathbb{E}_{\varepsilon_t}\}_{t=1}^{n-1} \exp\left(2\lambda \sum_{t=1}^{n-1} a_t(s_t)\varepsilon_t - 2\lambda^2 a_t(s_t)^2 \right),$$

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

$$\{\mathbb{E}_{s_t}\mathbb{E}_{\varepsilon_t}\}_{t=1}^{n} \left[\sup_{\mathbf{a}\in\mathcal{A}} \sum_{t=1}^{n} a_t(s_t)\varepsilon_t - \lambda a_t(s_t)^2 \right] \leq \frac{1}{2\lambda} \log \sum_{\mathbf{a}\in\mathcal{A}} 1 = \frac{\log|\mathcal{A}|}{2\lambda}.$$

■

Lemma 16 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 17 *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 $s_t = (\varepsilon_{1:t-1}) \in \mathcal{S}_t$,

$$\mathbb{E}_{\varepsilon_{1:n}} \left[\sup_{\mathbf{a}\in\mathcal{A}} \sum_{t=1}^{n} a_t(\varepsilon_{1:t-1})\varepsilon_t \right] \leq \sqrt{2\log|\mathcal{A}|} \cdot \sqrt{\mathbb{E}\left[\sup_{\mathbf{a}\in\mathcal{A}} \sum_{t=1}^{n} a_t(\varepsilon_{1:t-1})^2 \right]}. \tag{22}$$

Proof According to Lemma 16, 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

$$\begin{aligned}
&\{\mathbb{E}_{s_t}\mathbb{E}_{\varepsilon_t}\}_{t=1}^n \left[\sup_{\mathbf{a}\in\mathcal{A}} \sum_{t=1}^n a_t(s_t)\varepsilon_t\right] \\
&\quad \le \{\mathbb{E}_{s_t}\mathbb{E}_{\varepsilon_t}\}_{t=1}^n \left[\sup_{\mathbf{a}\in\mathcal{A}} \sum_{t=1}^n a_t(s_t)\varepsilon_t - \lambda a_t(s_t)^2\right] + \lambda \cdot \{\mathbb{E}_{s_t}\mathbb{E}_{\varepsilon_t}\}_{t=1}^n \left[\sup_{\mathbf{a}\in\mathcal{A}} \sum_{t=1}^n a_t(s_t)^2\right] \\
&\quad \le \frac{\log|\mathcal{A}|}{2\lambda} + \lambda\beta = \sqrt{2\beta\log|\mathcal{A}|} = \sqrt{2\log|\mathcal{A}|} \cdot \sqrt{\mathbb{E}\left[\sup_{\mathbf{a}\in\mathcal{A}} \sum_{t=1}^n a_t(s_t)^2\right]}.
\end{aligned}$$

■

Lemma 17 is an improvement on the finite class lemma in (Rakhlin et al., 2015b, Lemma 1); the latter result was proved with the supremum (rather than the expected value) over $\varepsilon_{1:n}$ under the square root in (22).

Appendix B. Proof of Theorem 2

B.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 Abernethy et al. (2009); Rakhlin and Sridharan (2014); Rakhlin et al. (2015a); Rakhlin and Sridharan (2015b); Foster et al. (2018); Bilodeau et al. (2020). The next step is to truncate the functions and forecaster's strategies away from 0. This analysis technique is also used in Cesa-Bianchi and Lugosi (1999); Rakhlin et al. (2015a). Our main steps in the proof include constructing an offset Rademacher process using a symmetrization argument Giné and Zinn (1984) and using chaining techniques Dudley (1967); van de Geer (2000) 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.

B.1.1. CONVERSION TO DUAL FORM GAME

We have the following standard result (see e.g. (Bilodeau et al., 2020, Lemma 6) or (Rakhlin and Sridharan, 2015b, Eq. 27))):

Lemma 18 *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

$$\mathcal{R}_n(\mathcal{Q}, \mathbf{p}, \mathbf{y}) := \sup_{\mathbf{q}\in\mathcal{Q}} \log\left(\frac{\mathbf{q}(\mathbf{y})}{\mathbf{p}(\mathbf{y})}\right). \tag{23}$$

Proof [Proof of Lemma 18] 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 v. Neumann (1928) 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 $\mathbf{p} \in \Delta(\mathcal{Y}^n)$.

B.1.2. 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

$$p_t^\delta(y_t \mid \mathbf{y}) = \begin{cases} \delta & \text{if } p_t(y_t \mid \mathbf{y}) < \delta \\ p_t(y_t \mid \mathbf{y}) & \text{if } \delta \le p_t(y_t \mid \mathbf{y}) \le 2\delta, \\ p_t(y_t \mid \mathbf{y}) \cdot \frac{1-\sum_{y\in\mathcal{Y}} p_t(y|\mathbf{y})\mathbb{I}[\delta \le p_t(y|\mathbf{y}) < 2\delta] - \delta \sum_{y\in\mathcal{Y}} \mathbb{I}[p_t(y|\mathbf{y}) < \delta]}{1-\sum_{y\in\mathcal{Y}} p_t(y|\mathbf{y})\mathbb{I}[p_t(y|\mathbf{y}) < 2\delta]} & \text{if } p_t(y \mid \mathbf{y}) \ge 2\delta. \end{cases} \tag{24}$$

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 19 *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 20 *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 Lemma 19 and Lemma 20 are deferred to Section B.2. In the following, we use $\Delta_n(\mathcal{Y}^n)$ to denote the following joint distribution set:

$$\Delta_n(\mathcal{Y}^n) := \left\{\mathbf{q} \in \Delta(\mathcal{Y}^n) : q_t(y_t \mid \mathbf{y}) \ge 1/(n^2|\mathcal{Y}|), \forall \mathbf{y} \in \mathcal{Y}^n\right\}. \tag{25}$$

B.1.3. 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 $t \geq 0$,

$$\zeta(t) = \begin{cases} 2\left(\sqrt{t} - 1\right), & t \leq 1, \\ 2\log\left(\frac{t+1}{2}\right), & t > 1. \end{cases} \tag{26}$$

For the justification of this choice of ζ see Section 4. We next introduce the following three properties of the function ζ, whose proofs are deferred to Section B.3.

Proposition 21 *For every $0 < x \leq n^2|\mathcal{Y}|$,*

$$\log x \leq \zeta(x) - \frac{1}{4\log(n|\mathcal{Y}|)} \cdot \zeta(x)^2. \tag{27}$$

Proposition 22 *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 (Polyanskiy and Wu, 2014, Theorem 7.5).

Proposition 23 *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 24 (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 $\boldsymbol{\varepsilon} = (\varepsilon_{1:n}) \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}\{-1, 1\}$, 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 25 (Symmetrization) *For any joint distribution $\mathbf{p} \in \Delta_n(\mathcal{Y}^n)$ where $\Delta_n(\mathcal{Y}^n)$ is defined in Eq. (25), 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:*

$$\begin{aligned} &\mathbb{E}_{\mathbf{y}\sim\mathbf{p}}\mathcal{R}_n(\mathcal{Q}, \mathbf{p}, \mathbf{y}) \\ &\quad\leq \mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}} \sum_{t=1}^n \varepsilon_t \zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \frac{1}{4\log(n|\mathcal{Y}|)}\zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)^2\right] \\ &\qquad+ \mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}} \sum_{t=1}^n (-\varepsilon_t) \zeta\left(\frac{q_t(z_t \mid \mathbf{w})}{p_t(z_t \mid \mathbf{w})}\right) - \frac{1}{4\log(n|\mathcal{Y}|)}\zeta\left(\frac{q_t(z_t \mid \mathbf{w})}{p_t(z_t \mid \mathbf{w})}\right)^2\right], \end{aligned} \tag{28}$$

where the expectation is with respect to $(\boldsymbol{\varepsilon}, \mathbf{w}, \mathbf{y}, \mathbf{z}) \sim \odot\mathbf{p}$.

The proof of Lemma 25 is deferred to Section B.3. It is based on the aforementioned properties of function ζ, and the symmetrization technique in Rakhlin et al. (2011).

B.1.4. CHAINING

We next analyze the right hand side of Eq. (28) 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. (28) 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 Section B.4:

Lemma 26 *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*

$$\sum_{t=1}^{n} \left(\zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \zeta\left(\frac{v_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)\right)^2 \le \sum_{t=1}^{n} \left(\zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \zeta\left(\frac{v'_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)\right)^2, \tag{29}$$

and

$$\sum_{t=1}^{n} \zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)^2 \ge \frac{1}{4} \sum_{t=1}^{n} \zeta\left(\frac{v_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)^2. \tag{30}$$

Remark 27 *The above lemma is similar to (Rakhlin and Sridharan, 2015b, Eq. (40)), (Rakhlin and Sridharan, 2014, Eq. (46)). The additional atom $\mathbf{p}$ serves as the 'zero' element in (Rakhlin and Sridharan, 2015b, 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 Definition 1. Then we have the following lemma.

Lemma 28 *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. (30) with $\mathbf{v} = \mathbf{v}[\mathbf{q}, \mathbf{p}, \mathbf{w}, \mathbf{y}, \alpha_N]$. We then hwave*

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}} \sum_{t=1}^{n} \left\{ \varepsilon_t \zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \frac{1}{4\log(n|\mathcal{Y}|)} \zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)^2 \right\}\right] \\
&\le \mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}} \sum_{t=1}^{n} \varepsilon_t \left\{ \zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \zeta\left(\frac{v_t[\mathbf{q}, \mathbf{p}, \mathbf{w}, \mathbf{y}, \alpha_1](y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)\right\}\right] \\
&\quad + \sum_{i=1}^{N-1} \mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}} \sum_{t=1}^{n} \varepsilon_t \left\{ \zeta\left(\frac{v_t[\mathbf{q}, \mathbf{p}, \mathbf{w}, \mathbf{y}, \alpha_i](y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \zeta\left(\frac{v_t[\mathbf{q}, \mathbf{p}, \mathbf{w}, \mathbf{y}, \alpha_{i+1}](y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)\right\}\right] \\
&\quad + \mathbb{E}\left[\sup_{\mathbf{v}\in\mathcal{V}(\alpha_N)\cup\{\mathbf{p}\}} \sum_{t=1}^{n} \left\{ \varepsilon_t \zeta\left(\frac{v_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \frac{1}{16\log(n|\mathcal{Y}|)} \cdot \zeta\left(\frac{v_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)^2 \right\}\right],
\end{aligned} \tag{31}$$

where the expectation is with respect to $(\boldsymbol{\varepsilon}, \mathbf{w}, \mathbf{y}, \mathbf{z}) \sim \odot\mathbf{p}$.

The proof of Lemma 28 is deferred to Section B.4. Next, we further upper bound the three terms in Eq. (31). 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 Lemma 17 (see Remark 29 in the proof of Theorem 2 for a discussion). The formal proof of these arguments, together with the full proof of Theorem 2, is deferred to Section B.5.

B.2. Missing Proofs in Section B.1.2

Proof [Proof of Lemma 19] Given the formula of $\mathcal{R}_n(\mathcal{Q}, \mathbf{p}, \mathbf{y})$ in Eq. (23), we only need to verify

$$\sup_{\mathbf{q}\in\mathcal{Q}} \sum_{t=1}^{n} \frac{1}{q_t(y_t \mid \mathbf{y})} \geq \sup_{\mathbf{q}\in\mathcal{Q}^\delta} \sum_{t=1}^{n} \frac{1}{q_t(y_t \mid \mathbf{y})} - 4n|\mathcal{Y}|\delta. \tag{32}$$

Notice that according to our construction of truncation in Eq. (24), we have for any $y \in \mathcal{Y}$ and $p \in \Delta(\mathcal{Y})$,

$$\log p(y) - \log p^\delta(y) = \log \frac{p(y)}{p^\delta(y)} \leq -\log\left(1 - 2\delta \cdot |\mathcal{Y}|\right) \leq 4\delta \cdot |\mathcal{Y}|, \tag{33}$$

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. (32). ■

Proof [Proof of Lemma 20] 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

$$\begin{aligned} &\mathbb{E}_{\mathbf{y}\sim\mathbf{p}} \mathcal{R}_n(\mathcal{Q}^\delta, \mathbf{p}, \mathbf{y}) - \mathbb{E}_{\mathbf{y}\sim\mathbf{p}^\delta} \mathcal{R}_n(\mathcal{Q}^\delta, \mathbf{p}^\delta, \mathbf{y}) \\ &= \sum_{s=1}^{n} \left[\mathbb{E}_{\mathbf{y}\sim\mathbf{p}^{s+1,\delta}} \mathcal{R}_n(\mathcal{Q}^\delta, \mathbf{p}^{s+1,\delta}, \mathbf{y}) - \mathbb{E}_{\mathbf{y}\sim\mathbf{p}^{s,\delta}} \mathcal{R}_n(\mathcal{Q}^\delta, \mathbf{p}^{s,\delta}, \mathbf{y}) \right]. \end{aligned} \tag{34}$$

We expand the right hand side of Eq. (34) for each $s \in [n]$:

$$\begin{aligned} &\mathbb{E}_{\mathbf{y}\sim\mathbf{p}^{s+1,\delta}} \mathcal{R}_n(\mathcal{Q}^\delta, \mathbf{p}^{s+1,\delta}, \mathbf{y}) - \mathbb{E}_{\mathbf{y}\sim\mathbf{p}^{s,\delta}} \mathcal{R}_n(\mathcal{Q}^\delta, \mathbf{p}^{s,\delta}, \mathbf{y}) \\ &= \left\{\mathbb{E}_{y_t\sim p_t(\cdot|\mathbf{y})}\right\}_{t=1}^{s-1} \left[\mathbb{E}_{y_s\sim p_s(\cdot|\mathbf{y})} \left\{\mathbb{E}_{y_t\sim p_t^\delta(\cdot|\mathbf{y})}\right\}_{t=s+1}^{n} \mathcal{R}_n(\mathcal{Q}^\delta, \mathbf{p}^{s+1,\delta}, \mathbf{y}) \right. \\ &\quad \left. - \mathbb{E}_{y_s\sim p_s^\delta(\cdot|\mathbf{y})} \left\{\mathbb{E}_{y_t\sim p_t^\delta(\cdot|\mathbf{y})}\right\}_{t=s+1}^{n} \mathcal{R}_n(\mathcal{Q}^\delta, \mathbf{p}^{s,\delta}, \mathbf{y}) \right]. \end{aligned} \tag{35}$$

Next, we fix $y_{1:s-1}$ and upper bound the expression inside the expectation:

$$\mathbb{E}_{y_s\sim p_s(\cdot|\mathbf{y})} \left\{\mathbb{E}_{y_t\sim p_t^\delta(\cdot|\mathbf{y})}\right\}_{t=s+1}^{n} \mathcal{R}_n(\mathcal{Q}^\delta, \mathbf{p}^{s+1,\delta}, \mathbf{y}) - \mathbb{E}_{y_s\sim p_s^\delta(\cdot|\mathbf{y})} \left\{\mathbb{E}_{y_t\sim p_t^\delta(\cdot|\mathbf{y})}\right\}_{t=s+1}^{n} \mathcal{R}_n(\mathcal{Q}^\delta, \mathbf{p}^{s,\delta}, \mathbf{y})$$

$$= \mathbb{E}_{y_s\sim p_s(\cdot|\mathbf{y})}\{\mathbb{E}_{y_t\sim p_t^\delta(\cdot|\mathbf{y})}\}_{t=s+1}^n \left[\mathcal{R}_n(\mathcal{Q}^\delta, \mathbf{p}^{s+1,\delta}, \mathbf{y}) - \mathcal{R}_n(\mathcal{Q}^\delta, \mathbf{p}^{s,\delta}, \mathbf{y})\right]$$
$$+ \left[\mathbb{E}_{y_s\sim p_s(\cdot|\mathbf{y})}\{\mathbb{E}_{y_t\sim p_t^\delta(\cdot|\mathbf{y})}\}_{t=s+1}^n \mathcal{R}_n(\mathcal{Q}^\delta, \mathbf{p}^{s,\delta}, \mathbf{y}) - \mathbb{E}_{y_s\sim p_s^\delta(\cdot|\mathbf{y})}\{\mathbb{E}_{y_t\sim p_t^\delta(\cdot|\mathbf{y})}\}_{t=s+1}^n \mathcal{R}_n(\mathcal{Q}^\delta, \mathbf{p}^{s,\delta}, \mathbf{y})\right] \tag{36}$$

For the first term in the right hand side of Eq. (36), 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. (36), when fixing $y_{1:s-1}$, we have

$$\mathbb{E}_{y_s\sim p_s(\cdot|\mathbf{y})}\{\mathbb{E}_{y_t\sim p_t^\delta(\cdot|\mathbf{y})}\}_{t=s+1}^n \mathcal{R}_n(\mathcal{Q}^\delta, \mathbf{p}^{s,\delta}, \mathbf{y}) - \mathbb{E}_{y_s\sim p_s^\delta(\cdot|\mathbf{y})}\{\mathbb{E}_{y_t\sim p_t^\delta(\cdot|\mathbf{y})}\}_{t=s+1}^n \mathcal{R}_n(\mathcal{Q}^\delta, \mathbf{p}^{s,\delta}, \mathbf{y})$$
$$\overset{(i)}{=} \mathbb{E}_{y_s\sim p_s(\cdot|\mathbf{y})}\{\mathbb{E}_{y_t\sim p_t^\delta(\cdot|\mathbf{y})}\}_{t=s+1}^n \left[\sup_{\mathbf{q}\in\mathcal{Q}^\delta}\left\{\sum_{t=1}^{s-1}\log\frac{q_t(y_t \mid \mathbf{y})}{p_t(y_t \mid \mathbf{y})} + \sum_{t=s}^n \log\frac{q_t(y_t \mid \mathbf{y})}{p_t^\delta(y_t \mid \mathbf{y})}\right\}\right]$$
$$- \mathbb{E}_{y_s\sim p_s^\delta(\cdot|\mathbf{y})}\{\mathbb{E}_{y_t\sim p_t^\delta(\cdot|\mathbf{y})}\}_{t=s+1}^n \left[\sup_{\mathbf{q}\in\mathcal{Q}^\delta}\left\{\sum_{t=1}^{s-1}\log\frac{q_t(y_t \mid \mathbf{y})}{p_t(y_t \mid \mathbf{y})} + \sum_{t=s}^n \log\frac{q_t(y_t \mid \mathbf{y})}{p_t^\delta(y_t \mid \mathbf{y})}\right\}\right]$$
$$\overset{(ii)}{=} \mathbb{E}_{y_s\sim p_s(\cdot|\mathbf{y})}\{\mathbb{E}_{y_t\sim p_t^\delta(\cdot|\mathbf{y})}\}_{t=s+1}^n \left[\sup_{\mathbf{q}\in\mathcal{Q}^\delta}\sum_{t=1}^n \log\frac{q_t(y_t \mid \mathbf{y})}{p_t^\delta(y_t \mid \mathbf{y})}\right]$$
$$- \mathbb{E}_{y_s\sim p_s^\delta(\cdot|\mathbf{y})}\{\mathbb{E}_{y_t\sim p_t^\delta(\cdot|\mathbf{y})}\}_{t=s+1}^n \left[\sup_{\mathbf{q}\in\mathcal{Q}^\delta}\sum_{t=1}^n \log\frac{q_t(y_t \mid \mathbf{y})}{p_t^\delta(y_t \mid \mathbf{y})}\right], \tag{37}$$

where (i) uses the formula of $\mathcal{R}_n(\mathcal{Q}, \mathbf{p}, \mathbf{y})$ in Eq. (23) 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 $y_{1:s-1}$,

$$\text{RHS of Eq. (37)} \le 2\mathrm{TV}\left(p_s(\cdot \mid \mathbf{y}), p_s^\delta(\cdot \mid \mathbf{y})\right) \cdot n\log\frac{1}{\delta}.$$

Based on Eq. (24), 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. (36) and then further back to Eq. (35), 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. (34), we have

$$\mathbb{E}_{\mathbf{y}\sim\mathbf{p}}\mathcal{R}_n(\mathcal{Q}^\delta,\mathbf{p},\mathbf{y})\le\mathbb{E}_{\mathbf{y}\sim\mathbf{p}^\delta}\mathcal{R}_n(\mathcal{Q}^\delta,\mathbf{p}^\delta,\mathbf{y})+2n^2|\mathcal{Y}|\delta\log\frac{1}{\delta}.$$

■

B.3. Missing Proofs in Section B.1.3

Proof [Proof of Proposition 21] We first verify the upper bounds part in Eq. (27). 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 [Proof of Proposition 22] 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

$$\mathbb{E}_{y\sim p}\left[-2\cdot\left(\sqrt{\frac{f(y)}{p(y)}}-1\right)-\frac{1}{4\log(n|\mathcal{Y}|)}\cdot\left(2\cdot\left(\sqrt{\frac{f(y)}{p(y)}}-1\right)\right)^2\right]\geq 0.$$

This can be verified by

$$\begin{aligned}
&\mathbb{E}_{y\sim p}\left[-\sqrt{\frac{f(y)}{p(y)}}+1-\frac{1}{2\log(n|\mathcal{Y}|)}\cdot\left(\sqrt{\frac{f(y)}{p(y)}}-1\right)^2\right]\\
&\geq \mathbb{E}_{y\sim p}\left[-\sqrt{\frac{f(y)}{p(y)}}+1-\frac{1}{2}\cdot\left(\sqrt{\frac{f(y)}{p(y)}}-1\right)^2\right]\\
&=\sum_{y\in\mathcal{Y}}\left[-\sqrt{f(y)p(y)}+p(y)-\frac{1}{2}f(y)+\sqrt{f(y)p(y)}-\frac{1}{2}p(y)\right]\\
&=0.
\end{aligned}$$

■

Proof [Proof of Proposition 23] We only need to verify that the function

$$h(t)=\begin{cases}2(t-1) & \text{if } 0<t\leq 1\\ 2\log\left(\frac{1+t^2}{2}\right) & \text{if } t>1\end{cases}$$

is a Lipschitz function with Lipschitz constant 2. This can be seen from

$$\frac{dh(t)}{dt}=\begin{cases}2 & \text{if } 0<t\leq 1,\\ \frac{4t}{1+t^2} & \text{if } t>1,\end{cases}$$

which satisfies $\left|\frac{dh(t)}{dt}\right|\leq 2$ for any $t>0$. ■

Proof [Proof of Lemma 25] 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

$$\begin{aligned}
\mathbb{E}_{\mathbf{y}\sim\mathbf{p}}\mathcal{R}_n(\mathcal{Q},\mathbf{p},\mathbf{y})&=\mathbb{E}_{\mathbf{w}\sim\mathbf{p}}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\sum_{t=1}^n\left(\log q_t(w_t\mid\mathbf{w})-\log p_t(w_t\mid\mathbf{w})\right)\right]\\
&\leq\mathbb{E}_{\mathbf{w}\sim\mathbf{p}}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\sum_{t=1}^n\left\{\zeta\left(\frac{q_t(w_t\mid\mathbf{w})}{p_t(w_t\mid\mathbf{w})}\right)-\frac{1}{4\log(n|\mathcal{Y}|)}\zeta\left(\frac{q_t(w_t\mid\mathbf{w})}{p_t(w_t\mid\mathbf{w})}\right)^2\right\}\right],
\end{aligned}\tag{38}$$

where the last steps follows from Proposition 21, 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 Proposition 22 gives that for any $\mathbf{q} \in \mathcal{Q}$ and $t \in [n]$,

$$\mathbb{E}\left[-\zeta\left(\frac{q_t(v_t \mid \mathbf{w})}{p_t(v_t \mid \mathbf{w})}\right) - \frac{1}{4\log(n|\mathcal{Y}|)}\zeta\left(\frac{q_t(v_t \mid \mathbf{w})}{p_t(v_t \mid \mathbf{w})}\right)^2 \,\middle|\, w_{1:t-1}\right] \geq 0. \tag{39}$$

Hence we can further upper bound

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\sum_{t=1}^{n}\left\{\zeta\left(\frac{q_t(w_t \mid \mathbf{w})}{p_t(w_t \mid \mathbf{w})}\right) - \frac{1}{4\log(n|\mathcal{Y}|)}\zeta\left(\frac{q_t(w_t \mid \mathbf{w})}{p_t(w_t \mid \mathbf{w})}\right)^2\right\}\right] \\
&\overset{(i)}{\leq} \mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\sum_{t=1}^{n}\left\{\zeta\left(\frac{q_t(w_t \mid \mathbf{w})}{p_t(w_t \mid \mathbf{w})}\right) - \frac{1}{4\log(n|\mathcal{Y}|)}\zeta\left(\frac{q_t(w_t \mid \mathbf{w})}{p_t(w_t \mid \mathbf{w})}\right)^2\right\}\right] \\
&\quad + \sum_{t=1}^{n}\mathbb{E}\left[-\zeta\left(\frac{q_t(v_t \mid \mathbf{w})}{p_t(v_t \mid \mathbf{w})}\right) - \frac{1}{4\log(n|\mathcal{Y}|)}\zeta\left(\frac{q_t(v_t \mid \mathbf{w})}{p_t(v_t \mid \mathbf{w})}\right)^2 \,\middle|\, w_{1:t-1}\right]\Bigg] \\
&\overset{(ii)}{\leq} \mathbb{E}\Bigg[\sup_{\mathbf{q}\in\mathcal{Q}}\sum_{t=1}^{n}\Bigg\{\zeta\left(\frac{q_t(w_t \mid \mathbf{w})}{p_t(w_t \mid \mathbf{w})}\right) - \zeta\left(\frac{q_t(v_t \mid \mathbf{w})}{p_t(v_t \mid \mathbf{w})}\right) \\
&\qquad - \frac{1}{4\log(n|\mathcal{Y}|)}\zeta\left(\frac{q_t(w_t \mid \mathbf{w})}{p_t(w_t \mid \mathbf{w})}\right)^2 - \frac{1}{4\log(n|\mathcal{Y}|)}\zeta\left(\frac{q_t(v_t \mid \mathbf{w})}{p_t(v_t \mid \mathbf{w})}\right)^2\Bigg\}\Bigg],
\end{aligned} \tag{40}$$

where in (i) we use Eq. (39), 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. (40) and further back to Eq. (38), we obtain that

$$\begin{aligned}
&\mathbb{E}_{\mathbf{y}\sim\mathbf{p}}\mathcal{R}_n(\mathcal{Q}, \mathbf{p}, \mathbf{y}) \\
&\leq \mathbb{E}\Bigg[\sup_{\mathbf{q}\in\mathcal{Q}}\sum_{t=1}^{n}\Bigg\{\varepsilon_t\zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \varepsilon_t\zeta\left(\frac{q_t(z_t \mid \mathbf{w})}{p_t(z_t \mid \mathbf{w})}\right) \\
&\qquad - \frac{1}{4\log(n|\mathcal{Y}|)}\zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)^2 - \frac{1}{4\log(n|\mathcal{Y}|)}\zeta\left(\frac{q_t(z_t \mid \mathbf{w})}{p_t(z_t \mid \mathbf{w})}\right)^2\Bigg\}\Bigg] \\
&\leq \mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\sum_{t=1}^{n}\left\{\varepsilon_t\zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \frac{1}{4\log(n|\mathcal{Y}|)}\zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)^2\right\}\right] \\
&\quad + \mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\sum_{t=1}^{n}\left\{(-\varepsilon_t)\zeta\left(\frac{q_t(z_t \mid \mathbf{w})}{p_t(z_t \mid \mathbf{w})}\right) - \frac{1}{4\log(n|\mathcal{Y}|)}\zeta\left(\frac{q_t(z_t \mid \mathbf{w})}{p_t(z_t \mid \mathbf{w})}\right)^2\right\}\right],
\end{aligned}$$

where the last inequality uses Jensen's inequality. ■

B.4. Missing Proofs in Section B.1.4

Proof [Proof of Lemma 26] 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

$$\sum_{t=1}^{n} \zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)^2 \geq \frac{1}{4}\sum_{t=1}^{n} \zeta\left(\frac{v'_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)^2$$

then we let $\mathbf{v} = \mathbf{v}'$ and it is easy to see that Eq. (29) and Eq. (30) both hold. Next we assume

$$\sum_{t=1}^{n} \zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)^2 < \frac{1}{4}\sum_{t=1}^{n} \zeta\left(\frac{v'_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)^2. \tag{41}$$

With $\mathbf{v} = \mathbf{p} \in \mathcal{V}(\alpha) \cup \{\mathbf{p}\}$ we will verify Eq. (29) and Eq. (30). First, since $\zeta(1) = 0$, we have

$$\sum_{t=1}^{n} \zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)^2 \geq 0 = \frac{1}{4}\sum_{t=1}^{n} \zeta\left(\frac{v_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)^2,$$

hence Eq. (30) holds. Next, according to Eq. (41) 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. (29) holds. ∎

Proof [Proof of Lemma 28] 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. (30) 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. (31) as follows:

$$\begin{aligned}&\sup_{\mathbf{q}\in\mathcal{Q}} \sum_{t=1}^{n}\left\{\varepsilon_t \zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \frac{1}{4\log(n|\mathcal{Y}|)}\zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)^2\right\} \\ &\quad= \sup_{\mathbf{q}\in\mathcal{Q}}\left\{\sum_{t=1}^{n}\varepsilon_t\left\{\zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \zeta\left(\frac{v_t[\mathbf{q}, \mathbf{p}, \mathbf{w}, \mathbf{y}, \alpha_N](y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)\right\}\right. \\ &\qquad\left. + \sum_{t=1}^{n}\left\{\varepsilon_t \zeta\left(\frac{v_t[\mathbf{q}, \mathbf{p}, \mathbf{w}, \mathbf{y}, \alpha_N](y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \frac{1}{16\log(n|\mathcal{Y}|)}\zeta\left(\frac{v_t[\mathbf{q}, \mathbf{p}, \mathbf{w}, \mathbf{y}, \alpha_N](y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)^2\right\}\right.\end{aligned}$$

$$
\begin{aligned}
&+ \frac{1}{16\log(n|\mathcal{Y}|)}\sum_{t=1}^{n}\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_N](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)^2 - \frac{1}{4\log(n|\mathcal{Y}|)}\sum_{t=1}^{n}\zeta\left(\frac{q_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)^2\Bigg\} \\
&\overset{(i)}{\leq} \sup_{\mathbf{q}\in\mathcal{Q}}\Bigg\{\sum_{t=1}^{n}\varepsilon_t\left\{\zeta\left(\frac{q_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)-\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_N](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)\right\} \\
&\quad + \sum_{t=1}^{n}\left\{\varepsilon_t\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_N](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right) - \frac{1}{16\log(n|\mathcal{Y}|)}\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_N](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)^2\right\}\Bigg\} \\
&\overset{(ii)}{\leq} \sup_{\mathbf{q}\in\mathcal{Q}}\left\{\sum_{t=1}^{n}\varepsilon_t\left\{\zeta\left(\frac{q_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)-\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_N](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)\right\}\right\} \\
&\quad + \sup_{\mathbf{v}\in\mathcal{V}(\alpha_N)\cup\{\mathbf{p}\}}\left\{\sum_{t=1}^{n}\left\{\varepsilon_t\zeta\left(\frac{v_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right) - \frac{1}{16\log(n|\mathcal{Y}|)}\zeta\left(\frac{v_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)^2\right\}\right\}, \qquad (42)
\end{aligned}
$$

where (i) uses the condition Eq. (30), 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:

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

where the last inequality is due to Jensen's inequality. Bringing this back to Eq. (42), we obtain Eq. (31). ■

B.5. Proof of Theorem 2

Proof [Proof of Theorem 2] First of all, according to Lemma 18, 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 Lemma 19 and Lemma 20, we have

$$\mathbb{E}_{\mathbf{y}\sim\mathbf{p}}\left[\mathcal{R}_n(\mathcal{Q},\mathbf{p},\mathbf{y})\right] \leq \mathbb{E}_{\mathbf{y}\sim\mathbf{p}^\delta}\left[\mathcal{R}_n(\mathcal{Q}^\delta,\mathbf{p}^\delta,\mathbf{y})\right] + 6n^2|\mathcal{Y}|\delta\log\frac{1}{\delta}.$$

Choosing $\delta = 1/(n^2|\mathcal{Y}|)$ for $n \geq 2$, we conclude that

$$\mathbb{E}\left[\mathcal{R}_n(\mathcal{Q}, \mathbf{p}, \mathbf{y})\right] \leq \mathbb{E}\left[\mathcal{R}_n(\mathcal{Q}^\delta, \mathbf{p}^\delta, \mathbf{y})\right] + 12\log(n|\mathcal{Y}|).$$

Hence in order to prove Theorem 2, we only need to prove that

$$\mathbb{E}_{\mathbf{y}\sim\mathbf{p}}\left[\mathcal{R}_n(\mathcal{Q}, \mathbf{p}, \mathbf{y})\right] = \tilde{\mathcal{O}}\left(\inf_{\gamma>\delta>0}\left\{n\delta\sqrt{|\mathcal{Y}|} + \sqrt{n|\mathcal{Y}|}\int_\delta^\gamma \sqrt{\mathcal{H}_{\mathsf{sq}}(\mathcal{Q}, \alpha, n)}d\alpha + \mathcal{H}_{\mathsf{sq}}(\mathcal{Q}, \gamma, n)\right\}\right) \tag{43}$$

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. (25). To prove this, we first notice that according to Lemma 25, for $\mathbf{p} \in \Delta_n(\mathcal{Y}^n)$ and $\mathcal{Q} \subseteq \Delta_n(\mathcal{Y}^n)$, we have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{y}\sim\mathbf{p}}\mathcal{R}_n(\mathcal{Q}, \mathbf{p}, \mathbf{y}) \\
&\quad\leq \mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\sum_{t=1}^n \varepsilon_t\zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \frac{1}{4\log(n|\mathcal{Y}|)}\zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)^2\right] \\
&\qquad+ \mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\sum_{t=1}^n (-\varepsilon_t)\zeta\left(\frac{q_t(z_t \mid \mathbf{w})}{p_t(z_t \mid \mathbf{w})}\right) - \frac{1}{4\log(n|\mathcal{Y}|)}\zeta\left(\frac{q_t(z_t \mid \mathbf{w})}{p_t(z_t \mid \mathbf{w})}\right)^2\right],
\end{aligned}$$

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 Definition 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

$$\mathbf{v}'[\mathbf{q}, \mathbf{p}, \mathbf{w}, \mathbf{y}, \alpha_N] = \operatorname*{arg\,min}_{\mathbf{v}\in\mathcal{V}(\alpha_N)}\left\{\sum_{t=1}^n\left(\zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \zeta\left(\frac{v_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)\right)^2\right\}. \tag{44}$$

According to Lemma 26, 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

$$\begin{aligned}
&\sum_{t=1}^n\left(\zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \zeta\left(\frac{v_t[\mathbf{q}, \mathbf{p}, \mathbf{w}, \mathbf{y}, \alpha_N](y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)\right)^2 \\
&\qquad\leq \sum_{t=1}^n\left(\zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \zeta\left(\frac{v'_t[\mathbf{q}, \mathbf{p}, \mathbf{w}, \mathbf{y}, \alpha_N](y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)\right)^2,
\end{aligned} \tag{45}$$

and

$$\sum_{t=1}^n \zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)^2 \geq \frac{1}{4}\sum_{t=1}^n \zeta\left(\frac{v_t[\mathbf{q}, \mathbf{p}, \mathbf{w}, \mathbf{y}, \alpha_N](y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)^2. \tag{46}$$

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

$$\mathbf{v}[\mathbf{q}, \mathbf{p}, \mathbf{w}, \mathbf{y}, \alpha_i] = \operatorname*{arg\,min}_{\mathbf{v}\in\mathcal{V}(\alpha_i)}\left\{\sum_{t=1}^n\left(\zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \zeta\left(\frac{v_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)\right)^2\right\}. \tag{47}$$

Then according to Lemma 28, we have

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\sum_{t=1}^{n}\varepsilon_t\left\{\zeta\left(\frac{q_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)-\frac{1}{4\log(n|\mathcal{Y}|)}\zeta\left(\frac{q_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)^2\right\}\right]\\
&\quad\le\mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\left\{\sum_{t=1}^{n}\varepsilon_t\left\{\zeta\left(\frac{q_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)-\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_1](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)\right\}\right\}\right]\\
&\qquad+\mathbb{E}\left[\sum_{i=1}^{N-1}\sup_{\mathbf{q}\in\mathcal{Q}}\left\{\sum_{t=1}^{n}\varepsilon_t\left\{\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_i](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)-\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_{i+1}](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)\right\}\right\}\right]\\
&\qquad+\mathbb{E}\left[\sup_{\mathbf{v}\in\mathcal{V}(\alpha_N)\cup\{\mathbf{p}\}}\left\{\sum_{t=1}^{n}\left\{\varepsilon_t\zeta\left(\frac{v_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)-\frac{1}{16\log(n|\mathcal{Y}|)}\zeta\left(\frac{v_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)^2\right\}\right\}\right]. \qquad (48)
\end{aligned}$$

In the following, we upper bound the three terms in Eq. (48) respectively.

We let

$$\bar{\mathbf{v}}[\mathbf{q},\mathbf{p},\mathbf{w},\alpha_N]=\arg\min_{\mathbf{v}\in\mathcal{V}(\alpha_N)}\left\{\max_{t\in[n]}\max_{y\in\mathcal{Y}}\left|\sqrt{q_t(y\mid\mathbf{w})}-\sqrt{v_t(y\mid\mathbf{w})}\right|\right\}. \qquad (49)$$

Then for $(\boldsymbol{\varepsilon},\mathbf{w},\mathbf{y},\mathbf{z})\sim\odot\mathbf{p}$, we have

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\sum_{t=1}^{n}\left(\zeta\left(\frac{q_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)-\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_N](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)\right)^2\right]\\
&\overset{(i)}{\le}\mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\sum_{t=1}^{n}\left(\zeta\left(\frac{q_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)-\zeta\left(\frac{v'_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_N](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)\right)^2\right]\\
&\overset{(ii)}{\le}\mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\sum_{t=1}^{n}\left(\zeta\left(\frac{q_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)-\zeta\left(\frac{\bar{v}_t[\mathbf{q},\mathbf{p},\mathbf{w},\alpha_N](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}\right)\right)^2\right]\\
&\overset{(iii)}{\le}4\mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\sum_{t=1}^{n}\left(\sqrt{\frac{q_t(y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}}-\sqrt{\frac{\bar{v}_t[\mathbf{q},\mathbf{p},\mathbf{w},\alpha_N](y_t\mid\mathbf{w})}{p_t(y_t\mid\mathbf{w})}}\right)^2\right]\\
&\overset{(iv)}{\le}4\mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\sum_{t=1}^{n}\frac{\alpha_N^2}{p_t(y_t\mid\mathbf{w})}\right]=4\alpha_N^2\cdot\mathbb{E}\left[\sum_{t=1}^{n}\frac{1}{p_t(y_t\mid\mathbf{w})}\right]\\
&\overset{(v)}{\le}4n\alpha_N^2|\mathcal{Y}|, \qquad (50)
\end{aligned}$$

where (i) uses Eq. (45), (ii) uses the definition of $\mathbf{v}'[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_N]$ in Eq. (44), (iii) uses the Lipschitz property of ζ function in Proposition 23, (iv) uses the definition of $\bar{\mathbf{v}}$ in Eq. (49) and Definition 1: for fixed $\mathbf{p},\mathbf{q}$ and $\mathbf{w}$, for any $t\in[n]$ and $y_t\in\mathcal{Y}$,

$$\begin{aligned}
&\left|\sqrt{q_t(y_t\mid\mathbf{w})}-\sqrt{\bar{v}_t[\mathbf{q},\mathbf{p},\mathbf{w},\alpha_N](y_t\mid\mathbf{w})}\right|\\
&\quad=\min_{\mathbf{v}\in\mathcal{V}(\alpha_N)}\max_{t\in[n]}\max_{y\in\mathcal{Y}}\left|\sqrt{q_t(y_t\mid\mathbf{w})}-\sqrt{v_t(y_t\mid\mathbf{w})}\right|\\
&\quad\le\sup_{\mathbf{q}\in\mathcal{Q}}\max_{\mathbf{w}\in\mathcal{Y}^n}\min_{\mathbf{v}\in\mathcal{V}(\alpha_N)}\max_{t\in[n]}\max_{y\in\mathcal{Y}}\left|\sqrt{q_t(y_t\mid\mathbf{w})}-\sqrt{v_t(y_t\mid\mathbf{w})}\right|\le\alpha_N,
\end{aligned}$$

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. (50) (the definition of $\mathbf{v}[\mathbf{q}, \mathbf{p}, \mathbf{w}, \mathbf{y}, \alpha_i]$ for $1 \le i \le N-1$ 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. (50)), we can show that with $(\boldsymbol{\varepsilon}, \mathbf{w}, \mathbf{y}, \mathbf{z}) \sim \odot\mathbf{p}$, for any $i \in [N-1]$,

$$\mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}} \sum_{t=1}^{n} \left(\zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \zeta\left(\frac{v_t[\mathbf{q}, \mathbf{p}, \mathbf{w}, \mathbf{y}, \alpha_i](y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)\right)^2\right] \le 4n\alpha_i^2|\mathcal{Y}|. \tag{51}$$

We are now ready to provide upper bounds for the three terms in Eq. (48). For the first term in Eq. (48), we have

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}} \left\{\sum_{t=1}^{n} \varepsilon_t \left\{\zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \zeta\left(\frac{v_t[\mathbf{q}, \mathbf{p}, \mathbf{w}, \mathbf{y}, \alpha_1](y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)\right\}\right\}\right] \\
&\overset{(i)}{\le} \sqrt{n} \cdot \mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}} \left\{\sqrt{\sum_{t=1}^{n} \left(\zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \zeta\left(\frac{v_t[\mathbf{q}, \mathbf{p}, \mathbf{w}, \mathbf{y}, \alpha_1](y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)\right)^2}\right\}\right] \\
&\overset{(ii)}{\le} \sqrt{n} \cdot \sqrt{\mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}} \sum_{t=1}^{n} \left(\zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \zeta\left(\frac{v_t[\mathbf{q}, \mathbf{p}, \mathbf{w}, \mathbf{y}, \alpha_1](y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)\right)^2\right]} \\
&\overset{(iii)}{\le} 2n\alpha_1\sqrt{|\mathcal{Y}|},
\end{aligned} \tag{52}$$

where (i) uses Cauchy-Schwarz inequality, (ii) uses Jensen's inequality and (iii) uses Eq. (51).

We next analyze the second term in Eq. (48). Notice that for any $i \in [N-1]$ and $\lambda > 0$, we can decompose the second term in Eq. (48) as

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}} \left\{\sum_{t=1}^{n} \varepsilon_t \left\{\zeta\left(\frac{v_t[\mathbf{q}, \mathbf{p}, \mathbf{w}, \mathbf{y}, \alpha_i](y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \zeta\left(\frac{v_t[\mathbf{q}, \mathbf{p}, \mathbf{w}, \mathbf{y}, \alpha_{i+1}](y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)\right\}\right\}\right] \\
&\le \mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}} \left\{\sum_{t=1}^{n} \varepsilon_t \left\{\zeta\left(\frac{v_t[\mathbf{q}, \mathbf{p}, \mathbf{w}, \mathbf{y}, \alpha_i](y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \zeta\left(\frac{v_t[\mathbf{q}, \mathbf{p}, \mathbf{w}, \mathbf{y}, \alpha_{i+1}](y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)\right\}\right.\right. \\
&\qquad \left.\left. - \lambda \cdot \left\{\zeta\left(\frac{v_t[\mathbf{q}, \mathbf{p}, \mathbf{w}, \mathbf{y}, \alpha_i](y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \zeta\left(\frac{v_t[\mathbf{q}, \mathbf{p}, \mathbf{w}, \mathbf{y}, \alpha_{i+1}](y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)\right\}^2\right\}\right] \\
&\quad + \lambda \cdot \mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}} \sum_{t=1}^{n} \left\{\zeta\left(\frac{v_t[\mathbf{q}, \mathbf{p}, \mathbf{w}, \mathbf{y}, \alpha_i](y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \zeta\left(\frac{v_t[\mathbf{q}, \mathbf{p}, \mathbf{w}, \mathbf{y}, \alpha_{i+1}](y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)\right\}^2\right]
\end{aligned} \tag{53}$$

For fixed $\mathbf{p}$ and $i \in [N-1]$, we define set $\mathcal{U}_i$ as

$$\mathcal{U}_i := \{\mathbf{u} : \mathbf{u} = \mathbf{u}[\mathbf{v}^i, \mathbf{v}^{i+1}] \text{ for some } \mathbf{v}^i \in \mathcal{V}(\alpha_i) \text{ and } \mathbf{v}^{i+1} \in \mathcal{V}(\alpha_{i+1}) \cup \{\mathbf{p}\}\},$$

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. (53) by

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\left\{\sum_{t=1}^n \varepsilon_t\left\{\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_i](y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_{i+1}](y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)\right\}\right\}\right] \\
&\quad \le \mathbb{E}\left[\sup_{\mathbf{u}\in\mathcal{U}}\sum_{t=1}^n \left\{\varepsilon_t u_t(y_t \mid \mathbf{w}) - \lambda \cdot u_t(y_t \mid \mathbf{w})^2\right\}\right] \\
&\qquad + \lambda \cdot \mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\sum_{t=1}^n \left\{\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_i](y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_{i+1}](y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)\right\}^2\right]
\end{aligned} \tag{54}$$

For the first term in Eq. (54), we adopt Lemma 16 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

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\mathbf{u}\in\mathcal{U}}\sum_{t=1}^n \left\{\varepsilon_t u_t(y_t \mid \mathbf{w}) - \lambda \cdot u_t(y_t \mid \mathbf{w})^2\right\}\right] \\
&\quad \le \frac{\log|\mathcal{U}|}{2\lambda} \le \frac{\log(|\mathcal{V}(\alpha_i)|(|\mathcal{V}(\alpha_{i+1})|+1))}{2\lambda} \le \frac{2\log|\mathcal{V}(\alpha_i)|}{\lambda},
\end{aligned} \tag{55}$$

where the last inequality uses the fact that $\alpha_i \le \alpha_{i+1}$ hence $|\mathcal{V}(\alpha_{i+1})| \le |\mathcal{V}(\alpha_i)|$. For the second term in Eq. (54), we have

$$\begin{aligned}
&\lambda \cdot \mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\sum_{t=1}^n \left\{\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_i](y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_{i+1}](y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)\right\}^2\right] \\
&\overset{(i)}{\le} 2\lambda \cdot \mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\sum_{t=1}^n \left(\zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_i](y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)\right)^2\right] \\
&\qquad + 2\lambda \cdot \mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\sum_{t=1}^n \left(\zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_{i+1}](y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)\right)^2\right] \\
&\overset{(ii)}{\le} 8\lambda n\alpha_i^2|\mathcal{Y}| + 8\lambda n\alpha_{i+1}^2|\mathcal{Y}| \overset{(iii)}{\le} 16\lambda n\alpha_{i+1}^2|\mathcal{Y}|,
\end{aligned} \tag{56}$$

where (i) we uses Cauchy-Schwarz inequality, (ii) we uses Eq. (50) and Eq. (51), and (iii) uses $\alpha_i \le \alpha_{i+1}$. Finally we choose $\lambda = \sqrt{\frac{\log|\mathcal{V}(\alpha_i)|}{8n\alpha_{i+1}^2|\mathcal{Y}|}}$. Then bringing Eq. (56) and Eq. (55) into Eq. (54), we obtain

$$\mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\left\{\sum_{t=1}^n \varepsilon_t\left\{\zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_i](y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \zeta\left(\frac{v_t[\mathbf{q},\mathbf{p},\mathbf{w},\mathbf{y},\alpha_{i+1}](y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)\right\}\right\}\right]$$

$$\leq \frac{2\log|\mathcal{V}(\alpha_i)|}{\lambda} + 16\lambda n\alpha_{i+1}^2|\mathcal{Y}| \leq 8\alpha_{i+1}\sqrt{2n|\mathcal{Y}|\log|\mathcal{V}(\alpha_i)|}. \tag{57}$$

Remark 29 *Let us briefly discuss the novel and key aspect of the approach used to upper bound the left-hand side of (57). 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 (55), which can be controlled for any predictable coefficients* irrespective of their magnitude*, as well as the expected squared distance between the coefficients in (56). We believe that this technique, which is summarized in Lemma 17, will be useful beyond this paper.*

Finally, we analyze the last term (offset term) in Eq. (48). 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 Lemma 16 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 Lemma 16 with $\lambda = \frac{1}{16\log(n|\mathcal{Y}|)}$, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\mathbf{v}\in\mathcal{V}(\alpha_N)\cup\{\mathbf{p}\}}\left\{\sum_{t=1}^n\left\{\varepsilon_t\zeta\left(\frac{v_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \frac{1}{16\log(n|\mathcal{Y}|)}\zeta\left(\frac{v_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)^2\right\}\right\}\right] \\
&\quad \leq 8\log(n|\mathcal{Y}|)\cdot\log(|\mathcal{V}(\alpha_N)|+1).
\end{aligned} \tag{58}$$

Finally, we specify the scales $\alpha_i = 2^{i-l}$ for some positive integer $l \geq N$. Bringing Eq. (52), Eq. (57) and Eq. (58) into Eq. (48), we obtain that

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\sum_{t=1}^n\varepsilon_t\left\{\zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right) - \frac{1}{4\log(n|\mathcal{Y}|)}\zeta\left(\frac{q_t(y_t \mid \mathbf{w})}{p_t(y_t \mid \mathbf{w})}\right)^2\right\}\right] \\
&\leq 2n\alpha_1\sqrt{|\mathcal{Y}|} + \sum_{i=1}^{N-1} 8\alpha_{i+1}\sqrt{2n|\mathcal{Y}|}\cdot\sqrt{\log|\mathcal{V}(\alpha_i)|} + 8\log(n|\mathcal{Y}|)\cdot(\log|\mathcal{V}(\alpha_N)|+1) \\
&\overset{(i)}{\lesssim} n\alpha_1\sqrt{|\mathcal{Y}|} + \sum_{i=1}^{N-1}\alpha_{i+1}\sqrt{n|\mathcal{Y}|}\cdot\sqrt{\mathcal{H}_{\mathsf{sq}}(\mathcal{Q},\alpha_i,n)} + \log(n|\mathcal{Y}|)\cdot\mathcal{H}_{\mathsf{sq}}(\mathcal{Q},\alpha_N,n) \\
&\overset{(ii)}{\lesssim} n\alpha_1\sqrt{|\mathcal{Y}|} + \sum_{i=1}^{N-1}(\alpha_i-\alpha_{i-1})\sqrt{n|\mathcal{Y}|}\cdot\sqrt{\mathcal{H}_{\mathsf{sq}}(\mathcal{Q},\alpha_i,n)} + \log(n|\mathcal{Y}|)\cdot\mathcal{H}_{\mathsf{sq}}(\mathcal{Q},\alpha_N,n) \\
&\overset{(iii)}{\leq} \frac{n\sqrt{|\mathcal{Y}|}}{2^{-l}} + \sqrt{n|\mathcal{Y}|}\int_{2^{-l}}^{2^{N-l}}\sqrt{\mathcal{H}_{\mathsf{sq}}(\mathcal{Q},\alpha,n)}d\alpha + \log(n|\mathcal{Y}|)\cdot\mathcal{H}_{\mathsf{sq}}(\mathcal{Q},2^{N-l},n),
\end{aligned}$$

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 $\alpha_{i-1} \leq \alpha \leq \alpha_i$,

$$\mathcal{H}_{\mathsf{sq}}(\mathcal{Q},\alpha_i,n) \leq \mathcal{H}_{\mathsf{sq}}(\mathcal{Q},\alpha,n).$$

For $0 < \delta < \gamma \leq 1$, letting $l = \log_2(1/\delta)$ and $N = l + \log_2(\gamma)$, and according to Lemma 25, 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. (43) is verified. ■

Appendix C. Missing Proofs in Section 3

C.1. Missing Proofs in Section 3.1

Proof [Proof of Lemma 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 (Fan, 1953, 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 [Proof of Theorem 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

$$f(\mathbf{x})_t(y_t \mid \mathbf{y}) = \begin{cases} f(x_t(\mathbf{y})) & \text{if } y_t = 1, \\ 1 - f(x_t(\mathbf{y})) & \text{if } y_t = 0. \end{cases} \tag{59}$$

According to Lemma 4 (see also (Bilodeau et al., 2020, Lemma 6), Liu et al. (2024) and (Rakhlin and Sridharan, 2015b, 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. (23), and $\mathcal{R}_n(\mathcal{F}(\mathbf{x}))$ is the Shtarkov sum Eq. (3) for joint distribution class $\mathcal{F}(\mathbf{x})$.

In order to prove Theorem 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 Definition 5 can form a sequential square-root covering of $\mathcal{F}(\mathbf{x})$ defined in Definition 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 Definition 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 Definition 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 Definition 1. Noticing that $|\mathcal{U}(\alpha)| = |\mathcal{V}(\alpha)|$, Theorem 6 directly follows from Theorem 2. ■

Proof [Proof of Corollary 7] Corollary 7 follows from Theorem 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:

$$(\gamma, \delta) = \begin{cases} \left(n^{-\frac{1}{p+2}}, n^{-1}\right) & \text{if } 0 \le p \le 2, \\ \left(1, n^{-\frac{1}{p}}\right) & \text{if } p > 2. \end{cases} \tag{60}$$

■

C.2. Missing Proofs in Section 3.2

Proof [Proof of Proposition 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

$$\begin{aligned}&\max\left\{\left|\sqrt{p}-\sqrt{q}\right|, \left|\sqrt{1-p}-\sqrt{1-q}\right|\right\}\\ &\quad= |p-q| \cdot \max\left\{\frac{1}{\sqrt{p}+\sqrt{q}}, \frac{1}{\sqrt{1-p}+\sqrt{1-q}}\right\} \le \frac{|p-q|}{\sqrt{\delta}}.\end{aligned}$$

■

Proof [Proof of Proposition 9] This proposition is a direct corollary of the standard inequality, e.g. (Polyanskiy and Wu, 2024, (7.22)), which shows that the squared Hellinger distance and TV distance satisfy the bound:

$$H(p,q)^2 \le 2D_{\mathrm{TV}}(p,q).$$

■

Appendix D. Missing Proofs in Section 3.3

In this section, we will present the formal proof to Theorem 10 and Theorem 11.

D.1. Proof of Theorem 10

To prove Theorem 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 $[0,1]$:

$$h(a,b) = \max\left\{\left|\sqrt{a}-\sqrt{b}\right|, \left|\sqrt{1-a}-\sqrt{1-b}\right|\right\}, \qquad \forall a, b \in [0,1]. \tag{61}$$

Definition 30 *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 31 *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 32 *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 10.

Proof [Proof of Theorem 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 Proposition 31, we have

$$\mathfrak{D}(F, \alpha, \alpha^4/16) \ge \sup_{\mathbf{x}} \mathcal{H}_{\mathsf{sq}}(\mathcal{F}, \alpha + \alpha^2/(2\sqrt{2}), n, \mathbf{x}) \cdot \log^{-1}\left(\frac{16en}{\alpha^4}\right) = \tilde{\Omega}\left(\alpha^{-p}\right).$$

Hence according to Proposition 32, 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 Proposition 31 and Proposition 32. A more general treatment of this approach, including the non-sequential analogue, will appear in the companion paper Jia et al. (2025).

D.1.1. PROOF OF PROPOSITION 31

Before proving Proposition 31, 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 33 *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 (61)), 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 34 *For function class $\mathcal{F} : \mathcal{X} \to U_\beta$, then for any depth-$n$ $\mathcal{X}$-valued tree $\mathbf{x}$, we have*

$$\mathcal{N}_{\mathsf{sq}}(\mathcal{F} \circ \mathbf{x}, \alpha, n) \leq \left(\frac{en}{\beta}\right)^{\mathfrak{D}(\mathcal{F},\alpha)}$$

Proof [Proof of Lemma 34] 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 Rakhlin and Sridharan (2015a)

$$g_\beta(n, d) = g_\beta(n-1, d) + (M-1) \cdot g_\beta(n-1, d-1). \tag{62}$$

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 $\mathfrak{D}(\mathcal{F}, \alpha) \leq d$, and any depth-$n$ $\mathcal{X}$-valued tree $\mathbf{x}$, $\mathcal{N}_{\mathsf{sq}}(\mathcal{F} \circ \mathbf{x}, \alpha, n) \leq g_\beta(d, n)$.

Base: There are two base case: $n \leq d$ and $d = 0$.

When $n \leq d$, 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

$$\mathcal{N}_{\mathsf{sq}}(\mathcal{F} \circ \mathbf{x}, \alpha, n) \leq |\mathcal{V}| = |U_\beta|^n = M^n = \sum_{i=0}^{d} \binom{n}{i} \cdot (M-1)^i = g_\beta(n, d).$$

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 $h(f(x), h(x')) \leq \alpha$ (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 $h(f(x_t(\mathbf{y})), v_t(\mathbf{y})) = h(f(x_t(\mathbf{y})), f(x_0)) \leq \alpha$. 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:

$$\mathcal{F} = \mathcal{F}_1 \cup \mathcal{F}_2 \cup \cdots \cup \mathcal{F}_{1/2\beta} \quad \text{where} \quad \mathcal{F}_k = \{f \in \mathcal{F} : f(x_1) = (2k-1)\beta\}, \forall 1 \leq k \leq M.$$

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

$$h((2a-1)\beta, (2b-1)\beta) \le \alpha, \qquad \forall a, b \in \mathcal{K} \tag{63}$$

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. (62).

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. (63) 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. (62). 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 Lemma 34, we are ready to prove Proposition 31 that works for real-valued function classes.

Proof [Proof of Proposition 31] 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 Definition 30 and Definition 33 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 Lemma 34 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

$$\sup_{\mathbf{x}} \mathcal{H}_{\mathsf{sq}}(\mathcal{F}, \alpha + \sqrt{2\beta}, n, \mathbf{x}) = \sup_{\mathbf{x}} \log \mathcal{N}(\mathcal{F} \circ \mathbf{x}, \alpha + \sqrt{2\beta}, n) \le \mathfrak{D}(\mathcal{F}, \alpha, \beta) \log\left(\frac{en}{\beta}\right).$$

■

D.1.2. PROOF OF PROPOSITION 32

Before proving Proposition 32, we present the following helper lemma.

Lemma 35 *Suppose real number $p, \alpha, \beta \in [0,1]$ and integer n satisfies $\beta < \alpha^2$ and*

$$\alpha + \beta < p < 1 - \alpha - \beta, \quad \text{and} \quad n \le \frac{p(1-p)}{324\alpha^2} \vee 1. \tag{64}$$

We collect n samples from $\mathrm{Ber}(p)$ to form an empirical estimation $\widehat{p} \in [0,1]$, and we define

$$\varepsilon = \begin{cases} 1 & \text{if } \widehat{p} \ge p \\ -1 & \text{if } \widehat{p} < p. \end{cases} \tag{65}$$

Then we have

$$\mathbb{E}\left[\widehat{p} \log \frac{p + \varepsilon\alpha - \beta}{p} + (1 - \widehat{p}) \log \frac{1 - p - \varepsilon\alpha - \beta}{1 - p}\right] \ge \frac{\alpha^2}{8p(1-p)}, \tag{66}$$

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

$$n \cdot \mathbb{E}\left[\widehat{p} \log \frac{p + \varepsilon\alpha - \beta}{p} + (1 - \widehat{p}) \log \frac{1 - p - \varepsilon\alpha - \beta}{1 - p}\right] \ge \frac{1}{5184}. \tag{67}$$

Proof [Proof of Lemma 35] 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

$$\mathbb{E}\left[\widehat{p}\log\frac{p+\varepsilon\alpha-\beta}{p} + (1-\widehat{p})\log\frac{1-p-\varepsilon\alpha-\beta}{1-p}\right]$$
$$= p\cdot\log\left(1+\frac{\alpha-\beta}{p}\right) + (1-p)\cdot\log\left(1+\frac{\alpha-\beta}{1-p}\right) \overset{(i)}{\ge} \frac{\alpha-\beta}{2} \overset{(ii)}{\ge} \frac{\alpha}{4} \overset{(iii)}{\ge} \frac{\alpha^2}{8p(1-p)},$$

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,

$$\mathbb{E}\left[\widehat{p}\log\frac{p+\varepsilon\alpha-\beta}{p} + (1-\widehat{p})\log\frac{1-p-\varepsilon\alpha-\beta}{1-p}\right]$$
$$= \Pr(\widehat{p}=0)\cdot\log\left(1+\frac{\alpha-\beta}{1-p}\right) + \mathbb{E}\left[\left(\widehat{p}\log\frac{p+\alpha-\beta}{p} + (1-\widehat{p})\log\frac{1-p-\alpha-\beta}{1-p}\right)\cdot\mathbb{I}[\widehat{p}>0]\right]. \tag{68}$$

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. (68) by

$$\Pr(\widehat{p}=0)\cdot\left(\left(\frac{\alpha-\beta}{1-p}\right) - \left(\frac{\alpha-\beta}{1-p}\right)^2\right)$$
$$+ \mathbb{E}\left[\left(\widehat{p}\cdot\left(\left(\frac{\alpha-\beta}{p}\right) - \left(\frac{\alpha-\beta}{p}\right)^2\right) + (1-\widehat{p})\cdot\left(\left(\frac{-\alpha-\beta}{1-p}\right) - \left(\frac{-\alpha-\beta}{1-p}\right)^2\right)\right)\cdot\mathbb{I}[\widehat{p}>0]\right]$$
$$= \mathbb{E}\left[\widehat{p}\cdot\left(\left(\frac{\varepsilon\alpha-\beta}{p}\right) - \left(\frac{\varepsilon\alpha-\beta}{p}\right)^2\right) + (1-\widehat{p})\cdot\left(\left(\frac{-\varepsilon\alpha-\beta}{1-p}\right) - \left(\frac{-\varepsilon\alpha-\beta}{1-p}\right)^2\right)\right]$$
$$\overset{(i)}{\ge} \mathbb{E}\left[\frac{\varepsilon\widehat{p}\alpha}{p} + \frac{-\varepsilon(1-\widehat{p})\alpha}{1-p}\right] - \beta\cdot\mathbb{E}\left[\frac{\widehat{p}}{p} + \frac{1-\widehat{p}}{1-p}\right] - (\alpha+\beta)^2\cdot\mathbb{E}\left[\frac{\widehat{p}}{p^2} + \frac{1-\widehat{p}}{(1-p)^2}\right]$$

$$\overset{(ii)}{=} \mathbb{E}\left[\frac{\varepsilon(\widehat{p}-p)\alpha}{p} + \frac{-\varepsilon(p-\widehat{p})\alpha}{1-p}\right] - 2\beta - (\alpha+\beta)^2\left(\frac{1}{p}+\frac{1}{1-p}\right)$$
$$= \alpha\cdot\mathbb{E}\left[\frac{|\widehat{p}-p|}{p(1-p)}\right] - 2\beta - \frac{(\alpha+\beta)^2}{p(1-p)}$$
$$\overset{(iii)}{\leq} \alpha\cdot\mathbb{E}\left[\frac{|\widehat{p}-p|}{p(1-p)}\right] - \frac{2\alpha^2}{p(1-p)}$$

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. (65) and (iii) uses the fact that $0<\beta\leq\alpha^2\leq 1/1296$. According to Khintchine inequality Haagerup (1981), we have

$$\mathbb{E}[|\widehat{p}-p|]\geq\sqrt{\frac{p(1-p)}{2n}},$$

which implies

$$\text{LHS of } Eq.\ (66) \geq \frac{\alpha}{\sqrt{2np(1-p)}} - \frac{2\alpha^2}{p(1-p)} \geq \frac{\alpha^2}{p(1-p)},$$

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

LHS of *Eq.* (66)
$$\geq \mathbb{E}\left[\widehat{p}\cdot\left(\left(\frac{\varepsilon\alpha-\beta}{p}\right)-\left(\frac{\varepsilon\alpha-\beta}{p}\right)^2\right)+(1-\widehat{p})\cdot\left(\left(\frac{-\varepsilon\alpha-\beta}{1-p}\right)-\left(\frac{-\varepsilon\alpha-\beta}{1-p}\right)^2\right)\right]$$
$$= \mathbb{E}\left[\widehat{p}\left(\frac{\varepsilon\alpha-\beta}{p}\right)+(1-\widehat{p})\left(\frac{-\varepsilon\alpha-\beta}{1-p}\right)-\widehat{p}\left(\frac{\varepsilon\alpha-\beta}{p}\right)^2-(1-\widehat{p})\left(\frac{-\varepsilon\alpha-\beta}{1-p}\right)^2\right]$$
$$\overset{(i)}{\geq} \mathbb{E}\left[\frac{\varepsilon\widehat{p}\alpha}{p}+\frac{-\varepsilon(1-\widehat{p})\alpha}{1-p}\right]-\beta\cdot\mathbb{E}\left[\frac{\widehat{p}}{p}+\frac{1-\widehat{p}}{1-p}\right]-(\alpha+\beta)^2\cdot\mathbb{E}\left[\frac{\widehat{p}}{p^2}+\frac{1-\widehat{p}}{(1-p)^2}\right]$$
$$\overset{(ii)}{=} \mathbb{E}\left[\frac{\varepsilon(\widehat{p}-p)\alpha}{p}+\frac{-\varepsilon(p-\widehat{p})\alpha}{1-p}\right]-2\beta-(\alpha+\beta)^2\left(\frac{1}{p}+\frac{1}{1-p}\right)$$
$$= \alpha\cdot\mathbb{E}\left[\frac{|\widehat{p}-p|}{p(1-p)}\right]-2\beta-\frac{(\alpha+\beta)^2}{p(1-p)}$$
$$\overset{(iii)}{\leq} \alpha\cdot\mathbb{E}\left[\frac{|\widehat{p}-p|}{p(1-p)}\right]-\frac{2\alpha^2}{p(1-p)}$$

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. (65) and (iii) uses the fact that $0 < \beta \le \alpha^2 \le 1$. According to (Berend and Kontorovich, 2013, Theorem 1), we have

$$\mathbb{E}[|\widehat{p} - p|] \ge \sqrt{\frac{p(1-p)}{2n}},$$

which implies

$$\text{LHS of } Eq.\ (66) \ge \frac{\alpha}{\sqrt{2np(1-p)}} - \frac{2\alpha^2}{p(1-p)} \ge \frac{\alpha^2}{p(1-p)},$$

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. (66), and Eq. (67) follows from Eq. (66) directly. ■

Now we are ready to prove Proposition 32.

Proof [Proof of Proposition 32] 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 Definition 30, 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

$$\left| f^{\tilde{\mathbf{y}}}(\tilde{x}_t(\tilde{\mathbf{y}})) - s_t(\tilde{\mathbf{y}})[\tilde{y}_t] \right| < \frac{\alpha^4}{16} \quad \text{and} \quad h(s_t(\tilde{\mathbf{y}})[0], s_t(\tilde{\mathbf{y}})[1]) \ge \alpha \qquad \forall t \in [d], \tag{69}$$

After noticing that $h(u,v)^2/2 \le |u - v|$ for any $u, v \in [0,1]$, we have

$$\left| f^{\tilde{\mathbf{y}}}(\tilde{x}_t(\tilde{\mathbf{y}})) - s_t(\tilde{\mathbf{y}})[\tilde{y}_t] \right| \le \frac{\left(s_t(\tilde{\mathbf{y}})[0] - s_t(\tilde{\mathbf{y}})[1]\right)^2}{4}$$

We define depth-d $[0,1]$-valued $\mathbf{v}$ as

$$v_t(\tilde{\mathbf{y}}) = \frac{s_t(\tilde{\mathbf{y}})[0] + s_t(\tilde{\mathbf{y}})[1]}{2}, \quad \forall \mathbf{y} \in \{0,1\}^n. \tag{70}$$

We can further verify that

$$h(s_t(\tilde{\mathbf{y}})[0], s_t(\tilde{\mathbf{y}})[1])^2$$

$$
\begin{aligned}
&= \left(s_t(\tilde{\mathbf{y}})[0] - s_t(\tilde{\mathbf{y}})[1]\right)^2 \\
&\quad \cdot \max\left\{ \frac{1}{\left(\sqrt{s_t(\tilde{\mathbf{y}})[0]} + \sqrt{s_t(\tilde{\mathbf{y}})[1]}\right)^2}, \frac{1}{\left(\sqrt{1-s_t(\tilde{\mathbf{y}})[0]} + \sqrt{1-s_t(\tilde{\mathbf{y}})[1]}\right)^2} \right\} \\
&\le \left(s_t(\tilde{\mathbf{y}})[0] - s_t(\tilde{\mathbf{y}})[1]\right)^2 \cdot \left(\frac{1}{s_t(\tilde{\mathbf{y}})[0] + s_t(\tilde{\mathbf{y}})[1]} + \frac{1}{1 - s_t(\tilde{\mathbf{y}})[0] + 1 - s_t(\tilde{\mathbf{y}})[1]} \right) \\
&= \frac{2\left(s_t(\tilde{\mathbf{y}})[0] - s_t(\tilde{\mathbf{y}})[1]\right)^2}{v_t(\tilde{y})(1 - v_t(\tilde{y}))}
\end{aligned} \tag{71}
$$

Next, we will show $\mathcal{R}_n(\mathcal{F}) \ge d$. According to Lemma 18, for any $\mathbf{p} \in \Delta(\{0,1\}^n)$ and depth-n $\mathcal{X}$-valued tree $\mathbf{x}$, we have

$$\mathcal{R}_n(\mathcal{F}) \ge \mathbb{E}_{\mathbf{y}\sim\mathbf{p}} \mathcal{R}_n(\mathcal{F}(\mathbf{x}), \mathbf{p}, \mathbf{y}). \tag{72}$$

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:

$$k_t = \left\lfloor \frac{v_t(\tilde{y})(1 - v_t(\tilde{y}))}{324\left(s_t(\tilde{\mathbf{y}})[1] - s_t(\tilde{\mathbf{y}})[0]\right)^2} \right\rfloor \vee 1, \qquad \forall t \in [d]. \tag{73}$$

and

$$\tilde{y}_t = \mathbb{I}\left\{ \sum_{j=1}^{k_t} y_{k_1+\cdots+k_{t-1}+j} \ge k_t \cdot v_t(\tilde{\mathbf{y}}) \right\}, \qquad \forall t \in [d], \tag{74}$$

where $v_t(\tilde{\mathbf{y}})$ is defined in Eq. (70). 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. (71) and Eq. (69), 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:

$$\Big(\underbrace{\tilde{x}_1(\tilde{\mathbf{y}}), \tilde{x}_1(\tilde{\mathbf{y}}), \cdots, \tilde{x}_1(\tilde{\mathbf{y}})}_{k_1 \text{ pieces}}, \underbrace{\tilde{x}_2(\tilde{\mathbf{y}}), \tilde{x}_2(\tilde{\mathbf{y}}), \cdots, \tilde{x}_2(\tilde{\mathbf{y}})}_{k_2 \text{ pieces}}, \cdots, \underbrace{\tilde{x}_d(\tilde{\mathbf{y}}), \tilde{x}_d(\tilde{\mathbf{y}}), \cdots, \tilde{x}_d(\tilde{\mathbf{y}})}_{k_d \text{ pieces}}, \underbrace{x_0, x_0, \cdots, x_0}_{(n-k_1-k_2-\cdots-k_d) \text{ pieces}} \Big),$$

Similarly, the value of $(p_1(\mathbf{y}), p_t(\mathbf{y}), \cdots, p_n(\mathbf{y}))$ are in the following form:

$$\Big(\underbrace{v_1(\tilde{\mathbf{y}}), v_1(\tilde{\mathbf{y}}), \cdots, v_1(\tilde{\mathbf{y}})}_{k_1 \text{ pieces}}, \underbrace{v_2(\tilde{\mathbf{y}}), v_2(\tilde{\mathbf{y}}), \cdots, v_2(\tilde{\mathbf{y}})}_{k_2 \text{ pieces}}, \cdots, \underbrace{v_d(\tilde{\mathbf{y}}), v_d(\tilde{\mathbf{y}}), \cdots, v_d(\tilde{\mathbf{y}})}_{k_d \text{ pieces}}, \underbrace{f^{\tilde{\mathbf{y}}}(x_0), f^{\tilde{\mathbf{y}}}(x_0) \cdots f^{\tilde{\mathbf{y}}}(x_0)}_{(n-k_1-k_2-\cdots-k_d) \text{ pieces}} \Big),$$

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

$\mathcal{R}_n(\mathcal{F}(\mathbf{x}), \mathbf{p}, \mathbf{y})$

$$
\begin{aligned}
&= \sup_{f\in\mathcal{F}} \left\{ \sum_{t=1}^{d}\sum_{j=1}^{k_d} \log \frac{f(y_{k_1+\cdots+k_{t-1}+j} \mid x_{k_1+\cdots+k_{t-1}+j}(\mathbf{y}))}{p_{k_1+\cdots+k_{t-1}+j}(y_{k_1+\cdots+k_{t-1}+j} \mid \mathbf{y})} \right. \\
&\qquad \left. + \sum_{j=1}^{n-k_1-\cdots-k_d} \log \frac{f(y_{k_1+\cdots+k_d+j} \mid x_{k_1+\cdots+k_d+j}(\mathbf{y}))}{p_{k_1+\cdots+k_d+j}(y_{k_1+\cdots+k_d+j} \mid \mathbf{y})} \right\} \\
&= \sup_{f\in\mathcal{F}} \left\{ \sum_{t=1}^{d}\sum_{j=1}^{k_d} \log \frac{f(y_{k_1+\cdots+k_{t-1}+j} \mid \tilde{x}_t(\tilde{\mathbf{y}}))}{v_t(y_{k_1+\cdots+k_{t-1}+j} \mid \tilde{\mathbf{y}})} + \sum_{j=1}^{n-k_1-\cdots-k_d} \log \frac{f(y_{k_1+\cdots+k_d+j} \mid x)}{f^{\tilde{\mathbf{y}}}(y_{k_1+\cdots+k_d+j} \mid x_0)} \right\} \\
&\geq \sum_{t=1}^{d}\sum_{j=1}^{k_d} \log \frac{f^{\tilde{\mathbf{y}}}(y_{k_1+\cdots+k_{t-1}+j} \mid \tilde{x}_t(\tilde{\mathbf{y}}))}{v_t(y_{k_1+\cdots+k_{t-1}+j} \mid \tilde{\mathbf{y}})}, \qquad (75)
\end{aligned}
$$

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

$$\gamma_t(\tilde{\mathbf{y}}) = \frac{s_t(\tilde{\mathbf{y}})[1] - s_t(\tilde{\mathbf{y}})[0]}{2}. \qquad (76)$$

Then using inequality $h(u,v)^2/2 \leq |u-v|$ for any $u, v \in [0,1]$, we have

$$\gamma_t(\tilde{\mathbf{y}}) \geq \frac{h(s_t(\tilde{\mathbf{y}})[1], s_t(\tilde{\mathbf{y}})[0])^2}{4} \geq \frac{\alpha^2}{4}. \qquad (77)$$

Notice that $\tilde{x}_t(\tilde{\mathbf{y}})$ 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. (75) as

$$\mathcal{R}_n(\mathcal{F}(\mathbf{x}), \mathbf{p}, \mathbf{y}) \geq \sum_{t=1}^{d} k_t \cdot \left(\hat{v}_t \log \frac{f^{\tilde{\mathbf{y}}}(\tilde{x}_t(\tilde{\mathbf{y}}))}{v_t(\tilde{\mathbf{y}})} + (1-\hat{v}_t) \log \frac{1 - f^{\tilde{\mathbf{y}}}(\tilde{x}_t(\tilde{\mathbf{y}}))}{1 - v_t(\tilde{\mathbf{y}})} \right).$$

Next, we will calculate $\mathbb{E}_{\mathbf{y}\sim\mathbf{p}}\left[\mathcal{R}_n(\mathcal{F}(\mathbf{x}), \mathbf{p}, \mathbf{y})\right]$. According to the above equation, we can separate the expectation into the sum of d expectations as follows:

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{y}\sim\mathbf{p}}\left[\mathcal{R}_n(\mathcal{F}(\mathbf{x}), \mathbf{p}, \mathbf{y})\right] \\
&= \sum_{t=1}^{d} \mathbb{E}\left[k_t \cdot \left(\hat{v}_t \log \frac{f^{\tilde{\mathbf{y}}}(\tilde{x}_t(\tilde{\mathbf{y}}))}{v_t(\tilde{\mathbf{y}})} + (1-\hat{v}_t) \log \frac{1 - f^{\tilde{\mathbf{y}}}(\tilde{x}_t(\tilde{\mathbf{y}}))}{1 - v_t(\tilde{\mathbf{y}})} \right) \right] \\
&= \sum_{t=1}^{d} \mathbb{E}\left[\mathbb{E}\left[k_t \cdot \left(\hat{v}_t \log \frac{f^{\tilde{\mathbf{y}}}(\tilde{x}_t(\tilde{\mathbf{y}}))}{v_t(\tilde{\mathbf{y}})} + (1-\hat{v}_t) \log \frac{1 - f^{\tilde{\mathbf{y}}}(\tilde{x}_t(\tilde{\mathbf{y}}))}{1 - v_t(\tilde{\mathbf{y}})} \right) \,\middle|\, y_{1:(k_1+\cdots k_{t-1})} \right] \right] \\
&\overset{(i)}{\geq} \sum_{t=1}^{d} \mathbb{E}\left[\mathbb{E}\left[k_t \cdot \left(\hat{v}_t \log \frac{s_t(\tilde{\mathbf{y}})[\tilde{y}_t] - \alpha^4/16}{v_t(\tilde{\mathbf{y}})} + (1-\hat{v}_t) \log \frac{1 - s_t(\tilde{\mathbf{y}})[\tilde{y}_t] - \alpha^4/16}{1 - v_t(\tilde{\mathbf{y}})} \right) \,\middle|\, y_{1:(k_1+\cdots k_{t-1})} \right] \right] \\
&\overset{(ii)}{\geq} \sum_{t=1}^{d} \mathbb{E}\left[\mathbb{E}\left[k_t \cdot \left(\hat{v}_t \log \frac{v_t(\tilde{\mathbf{y}}) + \varepsilon(\hat{v}_t)\gamma_t(\tilde{\mathbf{y}}) - \gamma_t(\tilde{\mathbf{y}})^2}{v_t(\tilde{\mathbf{y}})} \right. \right. \right.
\end{aligned}
$$

$$+ (1-\hat{v}_t)\log\frac{1-v_t(\tilde{\mathbf{y}})-\varepsilon(\hat{v}_t)v_t(\tilde{\mathbf{y}})-\gamma_t(\tilde{\mathbf{y}})^2}{1-v_t(\tilde{\mathbf{y}})}\Bigg)\,\Bigg|\, y_{1:(k_1+\cdots k_{t-1})}\Bigg]\Bigg] \tag{78}$$

where (i) uses the choice of $f^{\tilde{\mathbf{y}}}$ in Eq. (69), 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. (74) and also $\alpha^2/4 \leq \gamma_t(\tilde{\mathbf{y}})$ from Eq. (77). In Eq. (78), conditioned on $y_{1:(k_1+\cdots+k_{t-1})}$, there is no randomness on k_t, $v_t(\tilde{\mathbf{y}})$ and also $\tilde{x}_t(\tilde{\mathbf{y}})$, 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. (76), 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. (73), we can verify that the conditions in Lemma 35 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. (67) in Lemma 35 implies that

$$\mathbb{E}\left[k_t \cdot \left(\hat{v}_t \log\frac{v_t(\tilde{\mathbf{y}})+\varepsilon(\hat{v}_t)\gamma_t(\tilde{\mathbf{y}})-\gamma_t(\tilde{\mathbf{y}})^2}{v_t(\tilde{\mathbf{y}})} + (1-\hat{v}_t)\log\frac{1-v_t(\tilde{\mathbf{y}})-\varepsilon(\hat{v}_t)v_t(\tilde{\mathbf{y}})-\gamma_t(\tilde{\mathbf{y}})^2}{1-v_t(\tilde{\mathbf{y}})}\right)\,\Bigg|\, y_{1:(k_1+\cdots k_{t-1})}\right] \geq \frac{1}{518}$$

Bringing this back to Eq. (78) and then further back to Eq. (72), 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.2. Proof of Theorem 11

We present the proof of Theorem 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 36 *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

$$\begin{aligned} h(f,p)^2 &\geq \frac{1}{2}\cdot\left(\left(\sqrt{f}-\sqrt{p}\right)^2 + \left(\sqrt{1-f}-\sqrt{1-p}\right)^2\right) \\ &= \frac{1}{2}\cdot(f-p)^2\cdot\left(\frac{1}{(\sqrt{f}+\sqrt{p})^2} + \frac{1}{(\sqrt{1-f}+\sqrt{1-p})^2}\right) \\ &\geq \frac{1}{2}\cdot(f-p)^2\cdot\left(\frac{1}{2(f+p)} + \frac{1}{2(2-f-p)}\right) \end{aligned}$$

$$\geq \frac{1}{2}(f-p)^2,$$

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 11.

Proof [Proof of Theorem 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 Definition 30, 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 Definition 30, 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

$$|f^{\mathbf{y}}(x_t(\mathbf{y})) - s_t(\mathbf{y})[y_t]| < \frac{\beta^2}{2} \quad \text{and} \quad h\left(s_t(\mathbf{y})[0], s_t(\mathbf{y})[1]\right) > \beta, \qquad \forall t \in [n]. \tag{79}$$

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 Lemma 36 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

$$\sum_{t=1}^{n} \log \frac{f(y_t \mid x_t(\mathbf{y}))}{u_t(y_t \mid \mathbf{y})} \geq \sum_{t=1}^{n} \left\{ \left(\frac{f(y_t \mid x_t(\mathbf{y}))}{u_t(y_t \mid \mathbf{y})} - 1 \right) - \frac{3}{2} \cdot \left(\frac{f(y_t \mid x_t(\mathbf{y}))}{u_t(y_t \mid \mathbf{y})} - 1 \right)^2 \right\}. \tag{80}$$

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. (80), we obtain that

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_{\mathbf{y} \sim \mathbf{p}} \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \log \frac{f^{\mathbf{y}}(y_t \mid x_t(\mathbf{y}))}{u_t(y_t \mid \mathbf{y})} \right] \geq \frac{n\beta}{2} = \tilde{\Omega}\left(n^{\frac{p-1}{p}} \right).$$

■

Appendix E. Missing Proofs in Section 3.4

We first prove Lemma 12.

Proof [Proof of Lemma 12] Recall from Lemma 18 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 $\mathbf{y} = (y_{1:n}) \in \{0,1\}^n$,

$$\sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \log f(y_t \mid x_t(\mathbf{y})) \leq \sup_{f \in \mathcal{F}_{1/n}} \sum_{t=1}^{n} \log f(y_t \mid x_t(\mathbf{y})) + 2. \tag{81}$$

According to the definition of Hilbert ball class Eq. (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. (81). ■

We next prove Proposition 13. Our proof requires the following definition of skipping binary tree.

Definition 37 *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 38 *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* $n \ge k(d-1)+1$*,* $\mathbf{x}$ *has a skipping tree of depth* d *whose nodes are of the same color.*

Proof

We prove a stronger result: For integers $d_1, d_2, \ldots, d_k \ge 0$, 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 $1 \le i \le k$, 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 ($1 \le j \le k$). 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 $d_j \ge 1$. 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, $\exists 1 \le i_1, i_2 \le k$ 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 $d_1, \ldots, d_n \ge 0$. ■

Now we are ready to prove Proposition 13.

Proof [Proof of Proposition 13] First by choosing $\beta = \alpha^2/2$ in Proposition 31 and replacing α by 4α, we obtain

$$\sup_{\mathbf{x}} \mathcal{H}_{\mathsf{sq}}(\mathcal{F}_{1/n}, 5\alpha, n, \mathbf{x}) \le \mathfrak{D}(\mathcal{F}, 4\alpha, \alpha^2/2) \cdot \log\left(\frac{2en}{\alpha^2}\right).$$

Therefore, in order to prove Proposition 13, we only need to verify

$$\mathfrak{D}(\mathcal{F}_{1/n}, 4\alpha, \alpha^2/2) = \mathcal{O}\left(\frac{\log n}{\alpha^2}\right). \tag{82}$$

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

$$\left|\frac{1 + \langle w^{\mathbf{y}'}, x'_t(\mathbf{y}')\rangle}{2} - s'_t(\mathbf{y}')[y_t]\right| < \frac{\alpha^2}{2} \quad \text{and} \quad h\left(s'_t(\mathbf{y}')[0], s'_t(\mathbf{y}')[1]\right) > 4\alpha, \qquad \forall t \in [d']. \tag{83}$$

We further construct a depth-d' $[0,1]$-valued tree $\mathbf{u}'$ where for any $\mathbf{y}' \in \{0,1\}^{d'}$ and $t \in [d']$,

$$s'_t(\mathbf{y}')[0] < u'_t(\mathbf{y}') < s'_t(\mathbf{y}')[1], \qquad h(u'_t(\mathbf{y}'), s'_t(\mathbf{y}')[0]) \ge 2\alpha \quad \text{and} \quad h(u'_t(\mathbf{y}'), s'_t(\mathbf{y}')[1]) \ge 2\alpha.$$

According to Eq. (83), we have for any $\mathbf{y}' \in \{0,1\}^{d'}$ and $t \in [d']$,

$$(2y_t - 1) \cdot \left(\frac{1 + \langle w^{\mathbf{y}'}, x'_t(\mathbf{y}')\rangle}{2} - u'_t(\mathbf{y}')\right) \ge 0 \quad \text{and} \quad h\left(\frac{1 + \langle w^{\mathbf{y}'}, x'_t(\mathbf{y}')\rangle}{2}, u'_t(\mathbf{y}')\right) \ge \alpha.$$

We color the tree $\mathbf{x}'$ with two colors according to $u_t(\mathbf{y}')$: for each node $x'_t(\mathbf{y}')$, if $u'_t(\mathbf{y}') \le 1/2$ we color it with color 1, otherwise we color it with color 0. According to Lemma 38, 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 $t \in [d]$,

$$(2y_t - 1) \cdot \left(\frac{1 + \langle w, x_t(\mathbf{y}) \rangle}{2} - \bar{u}_t(\mathbf{y}) \right) \ge 0 \quad \text{and} \quad h\left(\frac{1 + \langle w, x_t(\mathbf{y}) \rangle}{2}, \bar{u}_t(\mathbf{y}) \right) \ge \alpha.$$

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

$$|\log p - \log f| \ge 2|\sqrt{p} - \sqrt{f}|.$$

Additionally, since for any $p \in [0,1/2]$ and $f \in [0,1]$, we always have

$$|\sqrt{p} - \sqrt{f}| = \frac{|p-f|}{\sqrt{p}+\sqrt{f}} \ge \frac{(\sqrt{2}+1)|p-f|}{\sqrt{1-p}+\sqrt{1-f}} = (\sqrt{2}+1) \cdot \left| \sqrt{1-p} - \sqrt{1-f} \right|,$$

which implies that $|\sqrt{p} - \sqrt{f}| \ge h(f,p)/4$. Hence we obtain

$$\log f - \log p \ge \frac{h(f,p)}{2}.$$

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 $t \in [d]$,

$$(2y_t - 1) \cdot \left(\log \frac{1 + \langle w, x_t(\mathbf{y}) \rangle}{2} - s_t(\mathbf{y}) \right) \ge \frac{\alpha}{2}$$

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 $w \in B_2(1-1/n)$,

$$\log \frac{1 + \langle w, x_t(\mathbf{y}) \rangle}{2} - s_t(\mathbf{y}) \ge \frac{\alpha}{2} > 0,$$

which implies that

$$s_t(\mathbf{y}) < \log \frac{1 + \langle w, x_t(\mathbf{y}) \rangle}{2} \le \log \frac{1 + \|w\|_2 \|x_t(\mathbf{y})\|_2}{2} \le \log \frac{1+1}{2} = 0.$$

Similarly by choosing $y_t = 0$, we obtain for some $w \in B_2(1-1/n)$,

$$\log \frac{1 + \langle w, x_t(\mathbf{y}) \rangle}{2} - s_t(\mathbf{y}) \le -\frac{\alpha}{2} < 0,$$

which implies that

$$s_t(\mathbf{y}) > \log \frac{1 + \langle w, x_t(\mathbf{y}) \rangle}{2} \ge \log \frac{1 - \|w\|_2 \|x_t(\mathbf{y})\|_2}{2} \ge \log \frac{1 - (1-1/n)}{2} = -\log(2n).$$

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 38, 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

$$v_i(\mathbf{y}) = x_{i,1} \to \cdots \to x_{i,l_i} = v_{i+1}(\mathbf{y}) \tag{84}$$

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 $\mathbf{x}$:

$$x_{1,1} \to x_{1,2} \cdots \to x_{1,l_1} \to x_{2,2} \to \cdots \to x_{2,l_2} \to x_{3,2} \to \cdots \to x_{3,l_3} \to \cdots x_{\bar{d},2} \to \cdots x_{\bar{d},l_{\bar{d}}}. \tag{85}$$

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

$$(2\tilde{y}_{i,j} - 1) \cdot \left(\log \frac{1 + \langle w, x_{i,j} \rangle}{2} - s_{i,j}(\tilde{\mathbf{y}}) \right) \geq \frac{\alpha}{2}, \qquad \forall 1 \leq i \leq \bar{d}, 1 \leq j \leq l_i. \tag{86}$$

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. (86) and notice that $v_i(\mathbf{y}) = x_{i,1}$ from Eq. (84) and $y_i = \tilde{y}_{i,1}$ from Eq. (85), 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 $i \in [\bar{d}]$,

$$(2y_i - 1) \cdot \left(\log \frac{1 + \langle w, v_i(\mathbf{y}) \rangle}{2} - u_i(\mathbf{y}) \right) \geq \frac{\alpha}{2} \quad \text{and} \quad u_i(\mathbf{y}) \in [-k-1, -k). \tag{87}$$

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 $i \in [\bar{d}]$,

$$\langle w, v_t(\mathbf{y})\rangle \geq 2e^{u_t(\mathbf{y})}e^{\alpha/2}-1 \quad \text{if } y_t = 1, \quad \text{and} \quad \langle w, v_t(\mathbf{y})\rangle \leq 2e^{u_t(\mathbf{y})}e^{-\alpha/2}-1 \quad \text{if } y_t = 0. \tag{88}$$

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. (88) 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. (87).

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

$$\|w[\mathbf{y}]^{\parallel} + v_0\|_2 \leq 1-(1-2e^{-k}) = 2e^{-k} \quad \text{and} \quad \|w[\mathbf{y}]^{\perp}\|_2 \leq \sqrt{1-(1-2e^{-k})^2} \leq 2e^{-k/2}. \tag{89}$$

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 (88) and the second inequality of Eq. (87), 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,

$$\|v_t(\mathbf{y})^{\perp}\|_2 \leq \|v_t(\mathbf{y}) - v_0\|_2 \leq 4e^{-k/2}. \tag{90}$$

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. (89) and Eq. (90), 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. (89) indicates that

$$\|v_2(\mathbf{y})^{\|} - v_0\|_2 \le 12e^{-k}.$$

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 Rakhlin et al. (2015a)), hence according to (Rakhlin, 2024, Page 67-68) we have an upper bound on $\bar{d}$.

$$\mathcal{G} = \{f | f(x) = \langle w, x\rangle, w, x \in B_2(1)\} \tag{91}$$

For any $\mathbf{y} \in \{0,1\}^{\bar{d}-1}$, we let

$$z_t(\mathbf{y}) = \frac{1}{5}e^{k/2}v_{t+1}((0,\mathbf{y}))^{\perp} + \frac{1}{5}v_{t+1}((0,\mathbf{y}))^{\|}, \quad \forall 1 \le t \le \bar{d}-1, \tag{92}$$

where we use $(0,\mathbf{y})$ to denote the path of length $\bar{d}$ whose t-th element equals to y_{t-1} for $t \ge 2$ and the first element equals to 0. Then according to Eq. (90), for any path $\mathbf{y}$ we have

$$\|z_t(\mathbf{y})\|_2 \le \frac{1}{5}e^{k/2}\|v_{t+1}((0,\mathbf{y}))^{\perp}\|_2 + \frac{1}{5}\|v_{t+1}((0,\mathbf{y}))^{\|}\|_2 \le \frac{4}{5} + \frac{1}{5} = 1,$$

which implies that $z_t(\mathbf{y}) \in B_2(1)$. We further notice from Eq. (89) that

$$\|w((0,\mathbf{y}))^{\|} + v_0\|_2 \le 2e^{-k} \quad \text{and} \quad \|w((0,\mathbf{y}))^{\perp}\|_2 \le 2e^{-k/2}.$$

Hence by choosing

$$\bar{w}(\mathbf{y}) = \frac{1}{4}e^{k}\left(w((0,\mathbf{y}))^{\|} + v_0\right) + \frac{1}{4}e^{k/2}w((0,\mathbf{y}))^{\perp}, \tag{93}$$

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. (93) and $z_t(\mathbf{y})$ in Eq. (92), 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. (88) and also $u_{t+1}((0,\mathbf{y})) \ge -k-1$ according to the second inequality of Eq. (87), we have

$$\langle \bar{w}(\mathbf{y}), z_t(\mathbf{y})\rangle - \bar{s}_t(\mathbf{y}) = \frac{1}{20}e^{k}\cdot\left(\langle w((0,\mathbf{y})), v_{t+1}((0,\mathbf{y}))\rangle - e^{u_{t+1}((0,\mathbf{y}))}\left(e^{\alpha/2} + e^{-\alpha/2}\right) + 1\right)$$

$$\begin{aligned}
&\geq \frac{1}{20}e^k \cdot \left(2e^{u_{t+1}((0,\mathbf{y}))}e^{\alpha/2} - 1 - e^{u_{t+1}((0,\mathbf{y}))}\left(e^{\alpha/2} + e^{-\alpha/2}\right) + 1\right) \\
&= \frac{1}{20}e^k e^{u_{t+1}((0,\mathbf{y}))}(e^{\alpha/2} - e^{-\alpha/2}) \\
&\geq \frac{1}{20}e^k e^{u_{t+1}((0,\mathbf{y}))}\alpha \geq \frac{1}{20}e^k e^{-k-1}\alpha = \frac{1}{20e}\alpha,
\end{aligned}$$

where in the second inequality we use the fact that $e^{\alpha/2} - e^{-\alpha/2} \geq \alpha$. And when $y_t = 0$, we have

$$\begin{aligned}
\langle \bar{w}(\mathbf{y}), z_t(\mathbf{y})\rangle - \bar{s}_t(\mathbf{y}) &= \frac{1}{20}e^k \cdot \left(\langle w((0,\mathbf{y})), v_{t+1}((0,\mathbf{y}))\rangle - e^{u_{t+1}((0,\mathbf{y}))}\left(e^{\alpha/2} + e^{-\alpha/2}\right) + 1\right) \\
&\leq \frac{1}{20}e^k \cdot \left(2e^{u_{t+1}((0,\mathbf{y}))}e^{-\alpha/2} - 1 - e^{u_{t+1}((0,\mathbf{y}))}\left(e^{\alpha/2} + e^{-\alpha/2}\right) + 1\right) \\
&= -\frac{1}{20}e^k e^{u_{t+1}((0,\mathbf{y}))}(e^{\alpha/2} - e^{-\alpha/2}) \\
&\leq -\frac{1}{20}e^k e^{u_{t+1}((0,\mathbf{y}))}\alpha \leq -\frac{1}{20}e^k e^{-k-1}\alpha = -\frac{1}{20e}\alpha.
\end{aligned}$$

Therefore, tree $\mathbf{z} \in B_2(1)$ is shattered by function class $\mathcal{G}$ (defined in Eq. (91)) at scale $1/(20e)\alpha$.

According to (Rakhlin, 2024, Page 67-68), the sequential fat shattering dimension of $\mathcal{G}$ at scale α is upper bounded by $16/\alpha^2$. Hence we have

$$\bar{d} - 1 \leq \frac{16}{(1/(20e)\alpha)^2} = \frac{6400e^2}{\alpha^2}.$$

This inequality, together with the fact that $\bar{d} \geq \frac{d-1}{\lceil \log(2n)\rceil}$, implies that

$$d \leq 1 + \lceil \log(2n)\rceil \cdot \left(1 + \frac{6400e^2}{\alpha^2}\right) = \mathcal{O}\left(\frac{\log n}{\alpha^2}\right).$$

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. (82). ■

Appendix F. Renewal Process and Hardness through Sequential Square-root Entropy

We consider the following class of renewal process, originally introduced in Csiszar and Shields (1996).

Definition 39 (Renewal Process Class Csiszar and Shields (1996)) *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 $i \geq 1$ and otherwise $y_t = 0$.*

For this class $\mathcal{Q}$ the work Csiszar and Shields (1996) 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 Definition 1), and also the sequential log entropy defined in Cesa-Bianchi and Lugosi (1999, 2006) are both $\Omega(n)$, no matter what scale we choose. Therefore, by simply applying Theorem 2 or the entropy bound in Cesa-Bianchi and Lugosi (1999, 2006) will only give a vacuous bound $O(n)$ on regret.

Proposition 40 *For any $0 < \alpha < 1/6$, we have $\mathcal{H}_{\mathsf{sq}}(\mathcal{Q}, \alpha, n) \geq n$. As for the log entropy (entropy with respect to distance Eq. (5)) defined in Cesa-Bianchi and Lugosi (1999, 2006), we have $\mathcal{H}_{\mathsf{log}}(\mathcal{Q}, \alpha, n) \geq (1 - \log 2)n - o(n)$.*

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 Definition 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

$$\max_{\mathbf{w}} \max_{y \in \{0,1\}} \max_{t \in [n]} \left| \sqrt{v_t^{\boldsymbol{\varepsilon}}(y_t \mid \mathbf{w})} - \sqrt{q_t^{\boldsymbol{\varepsilon}}(y_t \mid \mathbf{w})} \right| \leq \alpha,$$

which implies that

$$\max_{t \in [n]} \left| \sqrt{v_t^{\boldsymbol{\varepsilon}}(0 \mid \mathbf{y}^0)} - \sqrt{q_t^{\boldsymbol{\varepsilon}}(0 \mid \mathbf{y}^0)} \right| \leq \alpha.$$

If there exists $\boldsymbol{\varepsilon}$ and $\boldsymbol{\varepsilon}'$ such that $\mathbf{v}^{\boldsymbol{\varepsilon}} = \mathbf{v}^{\boldsymbol{\varepsilon}'}$, then

$$\max_{t \in [n]} \left| \sqrt{q_t^{\boldsymbol{\varepsilon}'}(0 \mid \mathbf{y}^0)} - \sqrt{q_t^{\boldsymbol{\varepsilon}}(0 \mid \mathbf{y}^0)} \right| \leq 2\alpha.$$

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 $|\mathcal{V}| \geq 2^n$, hence

$$\mathcal{H}(\mathcal{Q}, \alpha, n) \geq n.$$

Next we lower bound the log-entropy $\mathcal{H}_{\mathsf{log}}$, which is the entropy with respect to the distance defined in Eq. (5). 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

$$\sum_{t=1}^{n} \mathbb{I}[\varepsilon_t \neq \varepsilon_t'] \geq \frac{n}{4}. \tag{94}$$

According to the lower bound of packing number under Hamming distances (see (Polyanskiy and Wu, 2024, Theorem 27.5)), there exists such set E which satisfies

$$\log |E| \geq (1 - \log 2)n - o(n).$$

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

$$\sum_{t=1}^{n} \sup_{\mathbf{y}} (\log v_t^{\boldsymbol{\varepsilon}}(y_t \mid \mathbf{y}) - \log q_t^{\boldsymbol{\varepsilon}}(y_t \mid \mathbf{y}))^2 \leq n\alpha^2,$$

which implies that

$$\sum_{t=1}^{n} (\log v_t^{\boldsymbol{\varepsilon}}(0 \mid \mathbf{y}^0) - \log q_t^{\boldsymbol{\varepsilon}}(0 \mid \mathbf{y}^0))^2 \leq n\alpha^2.$$

If there exists $\boldsymbol{\varepsilon}$ and $\boldsymbol{\varepsilon}'$ such that $\mathbf{v}^{\boldsymbol{\varepsilon}} = \mathbf{v}^{\boldsymbol{\varepsilon}'}$, then

$$\sum_{t=1}^{n} (\log q_t^{\boldsymbol{\varepsilon}}(0 \mid \mathbf{y}^0) - \log q_t^{\boldsymbol{\varepsilon}'}(0 \mid \mathbf{y}^0))^2 \leq 4n\alpha^2. \tag{95}$$

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. (94). This contradicts to Eq. (95). 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 Definition 1 and (5) 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 (23) 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.