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

Zeyu Jia zyjia@mit.edu
Yury Polyanskiy yp@mit.edu
Alexander Rakhlin rakhlin@mit.edu

October 13, 2025

**Abstract**

We study gapped scale-sensitive dimensions of a function class in both sequential and non-sequential settings. We demonstrate that covering numbers for any uniformly bounded class are controlled above by these gapped dimensions, generalizing the results of [AB00, ABDCBH97]. Moreover, we show that the gapped dimensions lead to lower bounds on offset Rademacher averages, thereby strengthening existing approaches for proving lower bounds on rates of convergence in statistical and online learning.

## 1 Introduction

The celebrated Vapnik-Chervonenkis dimension $\mathsf{vc}(\mathcal{F})$ of a binary-valued function class $\mathcal{F}$ and the scale-sensitive dimension $\mathsf{vc}(\mathcal{F},\alpha)$ of a real-valued function class $\mathcal{F}$ are central notions in the study of empirical processes and convergence of statistical learning methods [VC71, BLW94, KS94]. Sequential analogues of these notions—the Littlestone dimension $\mathsf{ldim}(\mathcal{F})$ and the sequential scale-sensitive dimension $\mathrm{sfat}(\mathcal{F},\alpha)$—have been shown to play an analogously central role in the study of uniform martingale laws and online prediction [Lit88, BDPSS09, RST10].

In this paper, we study "gapped" versions of $\mathsf{vc}(\mathcal{F},\alpha)$ and $\mathrm{sfat}(\mathcal{F},\alpha)$. The modification yields a dimension that is no larger than the original one, yet can still be shown to control covering numbers in both sequential and non-sequential cases. More importantly, the new notion gives us a more precise control on the functions involved in "shattering" and thus yields non-vacuous lower bounds for offset Rademacher complexities for *any* uniformly bounded class—both in the classical and sequential cases—and, as a consequence, tighter lower bounds for online prediction problems, such as online regression or transductive learning. Our definition in the non-sequential case can also be seen as a modification of the Natarajan dimension [NT88, Nat89], and was, in fact, introduced in [AB00].

We first motivate the development in this paper on the simpler case of non-sequential data. We start by recalling the definition of the Vapnik-Chervonenkis dimension and its scale-sensitive version. Given a class $\mathcal{F}\subseteq\{f:\mathcal{X}\to\{-1,1\}\}$ of binary-valued functions on some set $\mathcal{X}$, consider the projection $\mathcal{F}|_{x_1,\dots,x_d}=\{(f(x_1),\dots,f(x_d)):f\in\mathcal{F}\}$ onto $d$ elements $x_1,\dots,x_d\in\mathcal{X}$. The VC-dimension $\mathsf{vc}(\mathcal{F})$ is the largest $d$ such that there exist $\{x_1,\dots,x_d\}$ with $\mathcal{F}|_{x_1,\dots,x_d}=\{-1,1\}^d$. For a real-valued class $\mathcal{F}\subseteq\{f:\mathcal{X}\to[-1,1]\}$ and a scale $\alpha\geq 0$, the scale-sensitive dimension $\mathsf{vc}(\mathcal{F},\alpha)$ is defined to be the largest $d$ such that there exist $x_1,\dots,x_d\in\mathcal{X}$ and $\mathbf{s}\in[-1,1]^d$ with the following property: for any $\boldsymbol{\varepsilon}\in\{-1,1\}^d$ there exists $f^{\boldsymbol{\varepsilon}}\in\mathcal{F}$ with $\varepsilon_i(f^{\boldsymbol{\varepsilon}}(x_i)-s_i)\geq\alpha/2$ for all $i\in\{1,\dots,d\}$. We say that $\mathcal{F}$ *shatters* $x_1,\dots,x_n$ at scale $\alpha$ if the aforementioned property holds. Shattering can also be visualized as a property that $\mathcal{F}|_{x_1,\dots,x_d}$ "contains" a cube $\mathbf{s}+(\alpha/2)\{-1,+1\}^d$ at scale $\alpha$ with a center at $\mathbf{s}$, in the sense that for each direction $\boldsymbol{\varepsilon}$, there is a function $f^{\boldsymbol{\varepsilon}}\in\mathcal{F}$ whose projection onto the data lies in the quadrant outside the vertex $\mathbf{s}+(\alpha/2)\boldsymbol{\varepsilon}$.

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

$$f^{\varepsilon}(x_i)\approx s_i+\varepsilon_i\alpha/2 \tag{1}$$

for every $i \in \{1, \ldots, d\}$. We can see that such a requirement (in the non-sequential case described here) can be satisfied under the assumption of convexity of $\mathcal{F}$. Unfortunately, for the sequential case described in Section 3, the nature of the restriction that (1) imposes on $\mathcal{F}$ is less clear.

We finish this introductory section by motivating the utility of the additional requirement (1). Once again, we only discuss the non-sequential case here. Consider the following lower bound on Rademacher averages of a class $\mathcal{F}$ in terms of $\mathsf{vc}(\mathcal{F}, \alpha)$:

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

where $d = \mathsf{vc}(\mathcal{F}, \alpha)$ and $\{x_1, \ldots, x_d\}$ is a set whose existence is guaranteed by the definition. Here the expectation is with respect to independent Rademacher random variables $\varepsilon_i$ (taking values $\pm 1$ with equal probability). The lower bound argument can be extended to $d > \mathsf{vc}(\mathcal{F}, \alpha)$ by considering repeated blocks of points and appealing to Khintchine's inequality. Such an argument leads to lower bounds of order $\Omega(\alpha\sqrt{\mathsf{vc}(\mathcal{F}, \alpha)n})$, implying that the scale-sensitive dimension is an inherent barrier for Rademacher averages to be small, and, as a consequence, a barrier for certain learning problems. Indeed, the notion of Radmacher averages in (2) is known to be a key object in the study of prediction with i.i.d. data and "non-curved" losses. On the other hand, for loss functions such as square loss, it is the *offset Rademacher averages*—or closely related local Rademacher averages [BBM05]—that in many situations correctly quantify the rates of convergence. The (non-sequential) offset Rademacher averages are defined as:[1]

$$\mathbb{E}_\varepsilon \sup_{f \in \mathcal{F}-\mathcal{F}} \sum_{i=1}^{n} \varepsilon_i f(x_i) - c \cdot f(x_i)^2. \tag{3}$$

Unfortunately, the lack of control on the magnitudes of departures of $f^{\boldsymbol{\epsilon}}(x_i)$ from $s_i \pm \alpha/2$ prevents us from obtaining sufficiently strong lower bounds when considering a shattered set, as the negative term in (3) may render the lower bound vacuous. This question motivates the study of gapped scale-sensitive dimensions, as presented in the next sections for both the non-sequential and sequential cases.

We briefly mention that a number of other versions of combinatorial dimensions have been proposed over the last few decades (see [DSS14, BCD+22, HMZ23] and references therein). To the best of our knowledge, these notions are different from those proposed in the present paper, and do not immediately imply the lower bounds we seek.

**Organization** We start the technical part of the paper with the non-sequential version of the gapped dimension in Section 2. We first introduce the gapped dimension for *integer-valued* classes in Section 2.1 and state a version of the Sauer-Shelah-Vapnik-Chervonenkis lemma for this combinatorial definition due to [AB00],[2] yielding control of covering numbers. We then extend the definition to the real-valued case in Section 2.2 and prove that this scale-sensitive dimension controls covering numbers, similarly to the development in [ABDCBH97] for the standard definition. We then prove that offset Rademacher averages for any uniformly bounded class are lower bounded according to the behavior of the gapped scale-sensitive dimension of this class (Section 2.4), and present the ensuing lower bound for the problem of online transductive regression. Section 3 mirrors the development in Section 2 for the sequential case, with an application to online regression.

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

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

[2] After this paper was completed, we were informed by Peter Bartlett that the definition of the "gapped" dimension and Lemma 1 are contained in [AB00].

# 2 Non-Sequential Gapped Dimensions

## 2.1 Integer-Valued Functions

Suppose $M$ is a positive integer. Let $\mathfrak{c} : [M] \times [M] \to \mathbb{R}_+ \cup \{0\}$ be a distance metric. Consider the following combinatorial parameter [AB00]:

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

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

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

The gapped dimension in Definition 1 was introduced in [AB00]. The definition is similar to that of the Natarajan dimension [NT88, Nat89] for multi-class learning, with the important difference that the two "labels" (denoted by the choice $s_t[1]$ and $s_t[-1]$) are $\alpha$-separated (hence the name *gapped* dimension); unlike multi-class problems where the the labels are treated as a categorical variable, in our case they are ordinal.

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

**Definition 2** ((Non-Sequential) Covering Number for Integer-Valued Functions)**.** *Fix $x_1, \dots, x_n \in \mathcal{X}$. We say that the set $\mathcal{V} = \{\mathbf{v} = (v_1, \dots, v_n) \in [M]^n\}$, is a (non-sequential) cover of $\mathcal{F}$ on $x_{1:n}$ at scale $\alpha$ if for any function $f \in \mathcal{F}$, there exists $\mathbf{v} \in \mathcal{V}$ such that*

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

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

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

**Lemma 1** ([AB00])**.** *Let $\mathcal{F} \subseteq \{f : \mathcal{X} \to [M]\}$ and $\{x_1, \dots, x_n\} \subseteq \mathcal{X}$. Then we have*

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

The proof of Lemma 1 is deferred to Section A.1. This result is similar to that of [ABDCBH97], which provides an upper bound on the covering number in terms of the (original) scale-sensitive dimension. Since the gapped dimension can be smaller than the original definition, Lemma 1 is an improvement over the corresponding result in [ABDCBH97].

## 2.2 Real-Valued Functions

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

**Definition 3** ((Non-Sequential) Gapped Scale-Sensitive Dimension)**.** *Let $\mathcal{F} \subseteq \{f : \mathcal{X} \to [-1, 1]\}$ and fix $\alpha, \beta > 0$. We say that $\mathcal{F}$ shatters $\{x_1, \dots, x_d\} \subseteq \mathcal{X}$ at scale $(\alpha, \beta)$ if there exist $s_1 = (s_1[-1], s_1[1]), s_2 = (s_2[-1], s_2[1]), \dots, s_d = (s_d[-1], s_d[1]) \in [-1, 1] \times [-1, 1]$ with the following properties:*

1. *For any* $t \in [d]$, $\mathfrak{c}(s_t[1], s_t[-1]) \geq \alpha$;

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

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

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

Figure 1: Grey area depicts a set $\mathcal{F}|_{x_1,\ldots,x_d}$. The hypercube on the left is "inscribed" in the classical sense, while the hyperrectangle on the right is "inscribed" according to the proposed definition.

**Definition 4** ((Non-Sequential) Covering Number for Real-Valued Functions)**.** *We say that a set* $\mathcal{V} \subseteq [-1,1]^n$ *is a cover of* $\mathcal{F}$ *on* $\{x_1,\ldots,x_n\} \subseteq \mathcal{X}$ *at scale* $\alpha$ *if for any function* $f \in \mathcal{F}$, *there exists* $v = (v_1,\ldots,v_n) \in \mathcal{V}$ *such that*

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

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

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

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

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

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

A version of this result already appears as Lemma 4.3 in [AB00], and we present the proof in Section A.2 for completeness. We remark that the logarithmic dependence on $n$ is unavoidable. The question of reducing the power from 2 to 1 (with the classical definition of scale-sensitive dimension) has been studied in [RV06]; we did not attempt to answer this question for the gapped dimension.

## 2.3 Comparison to Scale-Sensitive Dimension

In this section, we compare the classic scale-sensitive dimension for real-valued function class and the non-sequential scale-sensitive dimension defined in Definition 3. We first recall the definition of the scale-sensitive dimension [KS94, BLW94].

**Definition 5.** *Given a function class $\mathcal{F} \subseteq \{f : \mathcal{X} \to [-1,1]\}$, we say that $\mathcal{F}$ shatters $\{x_1, \dots, x_d\} \subseteq \mathcal{X}$ at scale $\alpha > 0$ if there exists witnesses $s_1, \dots, s_d \in [-1,1]$ such that for any $\boldsymbol{\varepsilon} = (\varepsilon_{1:d}) \in \{-1,1\}^d$, there exists $f^{\boldsymbol{\varepsilon}} \in \mathcal{F}$ such that for all $t \in [n]$, $\varepsilon_t \cdot (f^{\boldsymbol{\varepsilon}}(x_t) - s_t) \geq \alpha/2$. The scale-sensitive dimension $\mathsf{vc}(\mathcal{F}, \alpha)$ is defined to be the largest $d$ such that there exists $\{x_1, \dots, x_d\} \subseteq \mathcal{X}$ shattered by $\mathcal{F}$ at scale $\alpha$.*

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

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

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

**Proposition 3.** *Given a function class $\mathcal{F} \subseteq \{\mathcal{X} \to [-1,1]\}$, for any $\alpha, \beta > 0$, we have*

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

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

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

The proofs of Proposition 6 and Proposition 7 are almost identical to the proof of [AB00, Theorem 4.2] and [AB00, Theorem 4.3]. We defer both proofs to Section A.3. We remark that Proposition 3 is proved *using* Proposition 1. We are not aware of a direct proof of Proposition 3, which would, of course, allow us to re-use existing estimates of covering numbers via non-gapped dimensions.

There is an extra squared logarithmic factor in Proposition 3 compared to Proposition 2. The following proposition indicates that at least one logarithmic factor in Proposition 3 is necessary.

**Proposition 4.** *There exists a class of contexts $\mathcal{X}$, and a class of functions $\mathcal{F}$, such that for any $0 < 2\beta < \alpha < 1$, we have*

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

The proof of Proposition 4 is deferred to Section A.3.

## 2.4 Lower Bounds for Offset Rademacher Complexity and Transductive Regression

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

$$\mathcal{R}_n(\mathcal{F}, \mu_{1:n}, x_{1:n}, \boldsymbol{\varepsilon}) = \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} C \cdot \varepsilon_t (f(x_t) - \mu_t) - (f(x_t) - \mu_t)^2. \tag{4}$$

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

**Theorem 1.** *Suppose $C \geq 2$. Let $\mathfrak{c}(a,b) = |a-b|$. If there exists $p \geq 0$ such that $d(\mathcal{F}, \alpha, \alpha/(20nC)) = \tilde{\Omega}\left(\alpha^{-p}\right)$ holds for any $\alpha > 0$ and positive integer $n$, then for any positive integer $n$,*

$$\sup_{\mu_{1:n}, x_{1:n}} \mathbb{E}[\mathcal{R}_n(\mathcal{F}, \mu_{1:n}, x_{1:n}, \boldsymbol{\varepsilon})] = \tilde{\Omega}\left(n^{\frac{p}{p+2}}\right), \tag{5}$$

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

The proof of Theorem 1 is deferred to Section A.5.

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

$$\mathcal{V}_n(\mathcal{F}) \coloneqq \sup_{x_{1:n} \in \mathcal{X}^n} \left\{ \inf_{\widehat{y}_t} \sup_{y_t} \right\}_{t=1}^{n} \left[ \sum_{t=1}^{n} (\widehat{y}_t - y_t)^2 - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} (f(x_t) - y_t)^2 \right], \tag{6}$$

where $\{\cdot\}_{t=1}^{n}$ denotes repeated application of the operators. We remark that a typical assumption in transductive regression is that the set $\{x_1, \ldots, x_n\}$ is known, but not the order of appearance of its elements [QRZ24]. Such a setting is more difficult than the minimax game in Eq. (6), as the forecaster has less information about contexts throughout the game. Hence, the lower bound for Lemma 2 also applies to this more widely studied setting.

The following theorem, a transductive analogue of [RS14], lower bounds the regret in Eq. (6) by the (non-sequential) offset Rademacher process (4).

**Lemma 2.** *For any function class $\mathcal{F} \subseteq \{\mathcal{X} \to [-1,1]\}$, we have the following lower bound on the transductive regression objective:*

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

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

We defer the proof of Lemma 2 to Section A.5. Lemma 2 together with Theorem 1 and Proposition 1 implies the following lower bound for transductive online regression in terms of the non-sequential covering number defined in Section 4. Notably, the lower bound holds for any $\mathcal{F} \subseteq \{f : \mathcal{X} \to [-1,1]\}$. On the downside, the $[-2,2]$ range of predictions and outcomes does not correspond to the $[-1,1]$ range of the functions in $\mathcal{F}$, making the problem misspecified in an atypical manner; this issue also appears in [RS14, Lemma 4], and we are not aware of other general proof techniques that circumvent this.

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

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

In a manner similar to [RS14, Lemma 9], we can also show that $\mathcal{V}_n(\mathcal{F})$ is lower bounded by $n^{\frac{p-1}{p}}$ up to constants, under the same setting of Corollary 1. Hence, we obtain a complete picture of lower bounds for transductive regression in terms of the non-sequential covering numbers, modulo the misspecification issue discussed above.

# 3 Sequential Gapped Dimensions

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

We recall that constant-level trees (those with $x_t(\boldsymbol{\varepsilon}) = x_t \in \mathcal{X}$ for all $t \in [n]$ and $\boldsymbol{\varepsilon}$) correspond to a "tuple" of points $(x_1, \dots, x_n)$. Likewise, sequential generalizations—such as sequential cover—reduce to the classical notions when the trees are constant-level [RS15b]. However, we remark that the relations between the various sequential quantities do not imply the analogous relations in the non-sequential case. For this reason, in this paper we developed both sequential and non-sequential results separately.

## 3.1 Integer-Valued Functions

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

**Definition 6** (Sequential Gapped Dimension)**.** *Let $\mathcal{F} \subseteq \{f : \mathcal{X} \to [M]\}$ be a function class and fix $\alpha \geq 0$. We say that $\mathcal{F}$ shatters an $\mathcal{X}$-valued tree $\mathbf{x}$ of depth $d$ at scale $\alpha$ if there exists an $[M] \times [M]$-valued tree $\mathbf{s}$ of depth $d$ with the following properties:*

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

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

**Definition 7** (Sequential Covering Number for Integer-Valued Functions)**.** *Given an $\mathcal{X}$-valued tree $\mathbf{x}$ of depth $n$, we say that a set $\mathcal{V}$ of $[M]$-valued trees of depth $n$ is a sequential cover of $\mathcal{F}$ on $\mathbf{x}$ at scale $\alpha$, if for any function $f \in \mathcal{F}$ and $\boldsymbol{\varepsilon} \in \{-1,1\}^n$, there exists $\mathbf{v} \in \mathcal{V}$ such that for any $t \in [n]$, $\mathfrak{c}(f(x_t(\boldsymbol{\varepsilon})), v_t(\boldsymbol{\varepsilon})) \leq \alpha$.*

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

The following lemma upper bounds the sequential covering number for an integer-valued function class in terms of the sequential gapped dimension of the class, an analogue of Lemma 1.

**Lemma 3.** *Let $\mathcal{F} \subseteq \{f : \mathcal{X} \to [M]\}$ and let $\mathbf{x}$ be an $\mathcal{X}$-valued tree of depth $n$. We have*

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

The proof of Lemma 3 is deferred to Section B.

## 3.2 Real-Valued Functions

Let $\mathfrak{c} : [-1,1] \times [-1,1] \to \mathbb{R}_+ \cup \{0\}$ be a distance metric, as in Section 2.2. We define the following notion of complexity of $\mathcal{F}$, with respect to this metric:

**Definition 8** (Sequential Gapped Scale-sensitive Dimension)**.** *Let $\mathcal{F} \subseteq \{f : \mathcal{X} \to [-1,1]\}$ be a function class and fix $\alpha, \beta > 0$. We say that $\mathcal{F}$ shatters an $\mathcal{X}$-valued tree $\mathbf{x}$ of depth $d$ at scale $(\alpha, \beta)$ if there exists a $([-1,1] \times [-1,1])$-valued tree $\mathbf{s}$ of depth $d$ with the following properties:*

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

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

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

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

**Definition 9** (Sequential Covering Number for Real-Valued Functions)**.** *Given an* $\mathcal{X}$*-valued tree* $\mathbf{x}$ *of depth* $n$, *we say that a set* $\mathcal{V}$ *of* $\mathbb{R}$*-valued trees of depth* $n$ *is a sequential cover of* $\mathcal{F}$ *on* $\mathbf{x}$ *at scale* $\alpha$, *if for any function* $f \in \mathcal{F}$ *and* $\boldsymbol{\varepsilon} \in \{-1,1\}^n$, *there exists* $\mathbf{v} \in \mathcal{V}$ *such that for any* $t \in [n]$, $\mathfrak{c}(f(x_t(\boldsymbol{\varepsilon})), v_t(\boldsymbol{\varepsilon})) \le \alpha$.

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

**Proposition 5.** *For given real numbers* $\alpha, \beta > 0$, *suppose there exists a positive integer* $M$ *and* $M$ *real numbers* $-1 \le u_1 < u_2 < \ldots < u_M \le 1$ *such that for any* $u \in [-1,1]$, *there exists some* $i \in [M]$ *such that* $\mathfrak{c}(u, u_i) \le \beta$. *Then for any function class* $\mathcal{F} \subseteq \{f : \mathcal{X} \to [-1,1]\}$, *positive integer* $n$ *and* $\alpha, \beta > 0$, *for any depth-*$n$ $\mathcal{X}$*-valued tree* $\mathbf{x}$ *we have*

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

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

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

The proof of Proposition 5 is deferred to Section B. The structure of the proof follows that in [RST10]; see also [RS15a] for a different sequential generalization of Natarajan and Steele dimensions.

As in the previous works, Proposition 5 relies (via discretization) on covering numbers and combinatorial dimensions for integer-valued function classes, as developed in Section 3.1.

### 3.3 Comparison to Sequential Scale-Sensitive Dimension

We recall the definition of sequential scale-sensitive dimension in [RST15].

**Definition 10** (Sequential Scale-sensitive Dimension [RST15])**.** *Given a set* $\mathcal{X}$ *and a function class* $\mathcal{F} \subseteq \{\mathcal{X} \to [-1,1]\}$, *a depth-*$d$ $\mathcal{X}$*-valued tree* $\mathbf{x}$ *is shattered by* $\mathcal{F}$ *at scale* $\alpha > 0$ *if and only if there exists a depth-*$d$ $[-1,1]$*-valued tree* $\mathbf{v}$ *such that for any* $\boldsymbol{\varepsilon} \in \{-1,1\}^n$, *there exists* $f^{\boldsymbol{\varepsilon}} \in \mathcal{F}$ *which satisfies*

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

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

We have the following relations between the sequential scale-sensitive dimension $\mathrm{sfat}(\mathcal{F}, \alpha)$ in Definition 10 and the gapped sequential scale-sensitive dimension $d^{\mathsf{seq}}(\mathcal{F}, \alpha)$ with the distance function $\mathfrak{c}(a,b) = |a-b|$ in Definition 8. First, observe that a tree that is $(\alpha,\beta)$-shattered in the above sense for $\beta \le \alpha/4$ is also $\alpha/2$-shattered in the sense of the sequential scale-sensitive dimension sfat. Indeed, we may choose $\bar{\mathbf{s}}$ as a $[-1,1]$-valued tree defined by $\bar{s}_t(\boldsymbol{\varepsilon}) = (s_t(\boldsymbol{\varepsilon})[1] + s_t(\boldsymbol{\varepsilon})[-1])/2$. We then have that for any $\boldsymbol{\varepsilon} \in \{-1,1\}^d$, there exists $f^{\boldsymbol{\varepsilon}} \in \mathcal{F}$ such that $\varepsilon_t(f^{\boldsymbol{\varepsilon}}(x_t(\boldsymbol{\varepsilon})) - \bar{s}_t(\boldsymbol{\varepsilon})) \ge \alpha/4$. More precisely, we have the following:

**Proposition 6.** *Given a set* $\mathcal{X}$ *and a function class* $\mathcal{F} \subseteq \{\mathcal{X} \to [-1,1]\}$, *for any* $0 < 2\beta < \alpha$, *we have*

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

For the reverse direction, the following holds:

**Proposition 7.** *Given set $\mathcal{X}$ and function class $\mathcal{F} \subseteq \{\mathcal{X} \to [-1,1]\}$, for any $\alpha, \beta > 0$, we have*

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

Just as in the non-sequential case, we remark that Proposition 7 is proved (in Section B.2) *using* Proposition 5, not the other way around. Furthermore, the following result says that the $\log(1/\beta)$ factor in Proposition 7 is indeed necessary.

**Proposition 8.** *Consider the case where $\mathcal{X} = \{x\}$, and function class $\mathcal{F} = \{f : \mathcal{X} \to [-1,1] : f(x) \in [-1,1]\}$. Then for any $0 < 2\beta < \alpha < 1$, we have*

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

The proof of Proposition 8 is deferred to Section B.2.

### 3.4 Lower Bounds for Sequential Offset Rademacher Complexity and Online Regression

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

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

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

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

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

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

The proof of Theorem 2 is deferred to Section B.

**Application: Online Regression** In parallel to Section 2.4, Theorem 2 implies the following lower bound (Corollary 2) for online regression for any $\mathcal{F}$, significantly improving upon the lower bound in [RS14]; the latter result only guaranteed *existence* of $\mathcal{F}$ with such lower bound properties. This improvement was the main motivation for this paper.

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

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

Compared to the transductive setting in Eq. (6), here the forecaster does not have access to the context $x_t$ until the beginning of round $t$.

**Corollary 2.** *Fix a function class $\mathcal{F} \subseteq \{\mathcal{X} \to [-1,1]\}$ over context space $\mathcal{X}$. Suppose there exists a real number $p \geq 0$ such that the sequential covering number (defined in Definition 9) under distance $\mathfrak{c}(a,b) = |a-b|$ satisfies $\sup_{\mathbf{x}} \log \mathcal{N}_\infty^{\mathsf{seq}}(\mathcal{F}, \mathbf{x}, \alpha) = \tilde{\Omega}\left(\alpha^{-p}\right)$, where the supremum is over all depth-$n$ $\mathcal{X}$-valued trees. Then we have the following lower bound for the minimax regret:*

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

## Acknowledgements

We acknowledge support from the Simons Foundation and the NSF through awards DMS-2031883 and PHY-2019786, as well as support from the DARPA AIQ program.

## References

[AB00] Martin Anthony and Peter L Bartlett. Function learning from interpolation. *Combinatorics, Probability and Computing*, 9(3):213–225, 2000.

[ABDCBH97] Noga Alon, Shai Ben-David, Nicolo Cesa-Bianchi, and David Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM (JACM)*, 44(4):615–631, 1997.

[BBM05] Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497 – 1537, 2005.

[BCD+22] Nataly Brukhim, Daniel Carmon, Irit Dinur, Shay Moran, and Amir Yehudayoff. A characterization of multiclass learnability. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 943–955. IEEE, 2022.

[BDPSS09] Shai Ben-David, Dávid Pál, and Shai Shalev-Shwartz. Agnostic online learning. In *COLT*, volume 3, page 1, 2009.

[BLW94] Peter L Bartlett, Philip M Long, and Robert C Williamson. Fat-shattering and the learnability of real-valued functions. In *Proceedings of the seventh annual conference on Computational learning theory*, pages 299–310, 1994.

[DSS14] Amit Daniely and Shai Shalev-Shwartz. Optimal learners for multiclass problems. In Maria Florina Balcan, Vitaly Feldman, and Csaba Szepesvári, editors, *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 287–316, Barcelona, Spain, 13–15 Jun 2014. PMLR.

[Haa81] Uffe Haagerup. The best constants in the khintchine inequality. *Studia Mathematica*, 70(3):231–283, 1981.

[HMZ23] Steve Hanneke, Shay Moran, and Qian Zhang. Universal rates for multiclass learning. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 5615–5681. PMLR, 2023.

[KS94] Michael J Kearns and Robert E Schapire. Efficient distribution-free learning of probabilistic concepts. *Journal of Computer and System Sciences*, 48(3):464–497, 1994.

[Lit88] Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine learning*, 2:285–318, 1988.

[Nat89] Balas K Natarajan. On learning sets and functions. *Machine Learning*, 4:67–97, 1989.

[NT88] Balas K Natarajan and Prasad Tadepalli. Two new frameworks for learning. In *Machine Learning Proceedings 1988*, pages 402–415. Elsevier, 1988.

[QRZ24] Jian Qian, Alexander Rakhlin, and Nikita Zhivotovskiy. Refined risk bounds for unbounded losses via transductive priors. *arXiv preprint arXiv:2410.21621*, 2024.

[RS14] Alexander Rakhlin and Karthik Sridharan. Online non-parametric regression. In *Conference on Learning Theory*, pages 1232–1264. PMLR, 2014.

[RS15a] Alexander Rakhlin and Karthik Sridharan. Combinatorial dimensions, uniform convergence, and prediction, 2015. Conference for J. Michael Steele's 65th birthday.

[RS15b] Alexander Rakhlin and Karthik Sridharan. *On Martingale Extensions of Vapnik–Chervonenkis Theory with Applications to Online Learning*, pages 197–215. Springer International Publishing, 2015.

[RST10] Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning: Random averages, combinatorial parameters, and learnability. *Advances in Neural Information Processing Systems*, 23, 2010.

[RST15] Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Sequential complexities and uniform martingale laws of large numbers. *Probability theory and related fields*, 161:111–153, 2015.

[RV06] Mark Rudelson and Roman Vershynin. Combinatorics of random processes and sections of convex bodies. *Annals of Mathematics*, pages 603–648, 2006.

[VC71] V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.

# A Missing Proofs in Section 2

## A.1 Proofs of Lemma 1

*Proof of Lemma 1.* Without loss of generality we assume $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$, since the largest subset of $\{x_1, \ldots, x_n\}$ shattered by $\mathcal{F}$ is also a subset of $\mathcal{X}$ shattered by $\mathcal{F}$. In the following, we only need to consider cases where

$$4d(\mathcal{F}, \alpha) \log(eMn) \le n. \tag{7}$$

In cases where Eq. (7) fails, by taking the covering of all functions which map $\{x_{1:n}\}$ to $[M]$, we have the covering number estimate

$$\mathcal{N}_\infty(\mathcal{F}, x_{1:n}, \alpha) \le M^n = \exp(n \cdot \log M) \le \exp(4d(\mathcal{F}, \alpha) \log(eMn) \cdot \log M) \le \exp(16 d(\mathcal{F}, \alpha) \log^2(eMn)).$$

In the following, we let $d = d(\mathcal{F}, \alpha)$. We will use an approach similar to that in [ABDCBH97] to prove Lemma 1. We define the packing number $\mathcal{M}_\infty(\mathcal{F}, x_{1:n}, \alpha)$ of class $\mathcal{F}$ under design $x_{1:n}$ at scale $\alpha$: we say $\widetilde{\mathcal{F}} \subseteq \mathcal{F}$ is a packing if for any $f, f' \in \widetilde{\mathcal{F}}$,

$$\max_{t \in [n]} \mathfrak{c}(f(x_t), f'(x_t)) \ge \alpha,$$

and we let $\mathcal{M}_\infty(\mathcal{F}, x_{1:n}, \alpha)$ be the size of the largest packing of class $\mathcal{F}$. Then according to covering-packing inequality we have

$$\mathcal{N}_\infty(\mathcal{F}, x_{1:n}, \alpha) \le \mathcal{M}_\infty(\mathcal{F}, x_{1:n}, \alpha).$$

Hence, it suffices to prove

$$\log \mathcal{M}_\infty(\mathcal{F}, x_{1:n}, \alpha) \le 16 d \log^2(eMn). \tag{8}$$

For set $X = (x_1, \ldots, x_l) \subseteq \{x_{1:l}\}$, and tuple $\mathbf{s} = (s_1, s_2, \ldots, s_l)$ where $s_t = (s_t[-1], s_t[1]) \in [M] \times [M]$ for any $t \in [l]$, we say the pair $(X, \mathbf{s})$ is shattered by $\mathcal{F}$ if the following two properties both hold:

(a) for any $t \in [l]$, $\mathfrak{c}(s_t[-1], s_t[-1]) \ge \alpha$;

(b) for any $\boldsymbol{\varepsilon} = (\varepsilon_{1:l}) \in \{-1, 1\}^l$, there exists $f^{\boldsymbol{\varepsilon}} \in \mathcal{F}$ such that $f^{\boldsymbol{\varepsilon}}(x_t) = s_t[\varepsilon_t]$ for any $t \in [l]$.

We define function

$$t(h, l) \coloneqq \min_{\tilde{\mathcal{X}} \subseteq \mathcal{X}, |\tilde{\mathcal{X}}| = l} \max\{k : \forall\ F \subseteq \mathcal{F} \text{ such that } |F| = h \text{ and } \forall f, f' \in F,\ \max_{x \in \tilde{\mathcal{X}}} \mathfrak{c}(f(x), f'(x)) \ge \alpha,$$
$$F \text{ shatters at least } k \text{ pairs } (X, \mathbf{s}) \text{ where } X \subseteq \tilde{\mathcal{X}}\}. \tag{9}$$

(if no packing of size $h$ exists, $t(h, l)$ is defined to be infinity). According to the definition of $d = d(\mathcal{F}, \alpha)$, if $(X, \mathbf{s})$ is shattered by $\mathcal{F}$, then we have $|X| \le d$. Additionally, for fixed $X$, there can be at most $(M^2)^{|X|}$ choices of $\mathbf{s}$ such that $(X, \mathbf{s})$ is shattered by $\mathcal{F}$. Therefore, by choosing $h$ to be the size of largest packing, i.e. $\mathcal{M}_\infty(\mathcal{F}, x_{1:n}, \alpha)$, the number of pairs $(X, \mathbf{s})$ shattered by the largest packing is no more than the number of pairs $(X, \mathbf{s})$ shattered by $\mathcal{F}$, which implies that

$$t(\mathcal{M}_\infty(\mathcal{F}, x_{1:n}, \alpha), n) \le \sum_{i=0}^{d} \binom{n}{i} \cdot M^{2i}. \tag{10}$$

It is sufficient to argue that for any $r + 1 \le l \le n$,

$$t(2(2lM^2)^r, l) > 2^r. \tag{11}$$

Indeed, with Eq. (11) holding, it is easy to see from the definition of function $t$ in Eq. (9) that for $h_1 \le h_2$ and any $l$, $t(h_1, l) \le t(h_2, l)$. Notice that according to Eq. (7),

$$n \ge 4d \log(eMn) \ge \log_2 \left\lfloor \sum_{i=1}^{d} \binom{n}{i} M^{2i} \right\rfloor + 2.$$

Hence, if

$$\mathcal{M}_\infty(\mathcal{F}, x_{1:n}, \alpha) > 2(2nM^2)^{\log_2 \lfloor \sum_{i=1}^d \binom{n}{i} M^{2i} \rfloor + 1},$$

we have

$$\begin{aligned} t(\mathcal{M}_\infty(\mathcal{F}, x_{1:n}, \alpha), n) &\geq t\left(2(2nM^2)^{\log_2 \lfloor \sum_{i=1}^d \binom{n}{i} M^{2i} \rfloor + 1}, n\right) \\ &\geq 2^{\log_2 \lfloor \sum_{i=1}^d \binom{n}{i} M^{2i} \rfloor + 1} > \sum_{i=1}^d \binom{n}{i} M^{2i}, \end{aligned}$$

which contradicts Eq. (10). Therefore,

$$\log \mathcal{N}_\infty(\mathcal{F}, x_{1:n}, \alpha) \leq \log \mathcal{M}_\infty(\mathcal{F}, x_{1:n}, \alpha) \leq \log\left(2(2nM^2)^{\log_2 \lfloor \sum_{i=1}^d \binom{n}{i} M^{2i} \rfloor + 1}\right) \leq 16d \log^2(eMn).$$

In the remainder of the proof, we argue that Eq. (11) holds by induction. We will verify the following two properties:

$$t(2, l) \geq 1 \qquad \forall l \geq 1, \tag{12}$$

$$t(2lM^2 \cdot 2m, l) \geq 2 \cdot t(2m, l-1) \qquad \forall m \geq 1, l \geq 2. \tag{13}$$

We first verify Eq. (12). For $\tilde{\mathcal{X}} \subseteq \mathcal{X}$ and any packing $F \subseteq \mathcal{F}$ with $|F| = 2$, there exists $f_1, f_2 \in F$ such that

$$\max_{x \in \tilde{\mathcal{X}}} \mathfrak{c}(f_1(x), f_2(x)) \geq \alpha.$$

We let $\tilde{x} \in \tilde{\mathcal{X}}$ to be the one which takes the maximum in the above inequality, and let $X = (\tilde{x}) \subseteq \mathcal{X}$ and $\mathbf{s} = \{s_1\}$ with $s_1 = (f_1(x), f_2(x)) \in [M] \times [M]$, then $(X, \mathbf{s})$ is shattered by $\mathcal{F}$. Therefore, $t(2, l) \geq 1$.

We next verify Eq. (13). Without loss of generality we assume the minimum over $\tilde{\mathcal{X}} \subseteq \mathcal{X}$ with $|\tilde{\mathcal{X}}| = l$ in Eq. (9) is achieved by $\tilde{\mathcal{X}} = \{x_{1:l}\}$. Suppose packing $F \subseteq \mathcal{F}$ satisfies $|F| \geq 4lM^2 \cdot m$. We arbitrarily pair up functions in $F$ to form:

$$F = \bigcup_{t=1}^{2lM^2 m} \{f_t, g_t\}.$$

For any $t \in [2lM^2 m]$, since $F$ is a packing, there exists $x[t] \in \{x_{1:l}\}$ such that

$$\mathfrak{c}(f_t(x[t]), g_t(x[t])) \geq \alpha.$$

Next, for any $x \in \{x_{1:l}\}$ and $i, j \in [M]$ with $\mathfrak{c}(i, j) \geq \alpha$, we define the set

$$\mathcal{T}(x, i, j) = \{t \in [2lM^2 m] : x[t] = x \text{ and } f_t(x[t]) = i,\ g_t(x[t]) = j\}.$$

Then we have

$$\bigcup_{\substack{x \in \{x_{1:l}\} \\ i,j \in M, \mathfrak{c}(i,j) \geq \alpha}} \mathcal{T}(x, i, j) = [2lM^2 m].$$

According to the pigeonhole principle, there exists some $x_{t^\star} \in \{x_{1:l}\}$ and $i^\star, j^\star \in [M]$ with $\mathfrak{c}(i^\star, j^\star) \geq \alpha$ such that $|\mathcal{T}(x_{t^\star}, i^\star, j^\star)| \geq 2m$. We define two function classes $F_1, F_2 \subseteq F$ as

$$F_1 = \{f_t : t \in \mathcal{T}(x_t^\star, i^\star, j^\star)\} \quad \text{and} \quad F_2 = \{g_t : t \in \mathcal{T}(x_t^\star, i^\star, j^\star)\}.$$

Then we have $|F_1| = |F_2| \geq 2m$. Since functions in $F_1$ all take value $i^\star$ at $x_{t^\star}$, $F_1$ is a packing under design $\{x_{1:l}\} \backslash \{x_{t^\star}\}$. Hence, there exists a set $V$ consisting of pairs $(X, \mathbf{s})$ shattered by class $F_1$, and $|V| \geq t(2m, l-1)$. Since functions in $F_1$ takes the same value at $x_{t^\star}$, for any $(X, \mathbf{s}) \in V$, $x_{t^\star} \notin X$. Similarly, there exists a set $U$ shattered by $F_2$ with $|U| \geq t(2m, l-1)$, and for any $(X, \mathbf{s}) \in U$, $x_{t^\star} \notin X$. Since $F_1 \subseteq F$ and $F_2 \subseteq F$, any

pairs $(X, \mathbf{s})$ in $U \cup V$ are also shattered by $F$. Additionally, for any pair $(X, \mathbf{s}) \in U \cap V$, we construct a new pair $(X \cup (x_{t^\star}), \mathbf{s} \cup (i^\star, j^\star)) \notin U \cup V$. Since $\mathfrak{c}(i^\star, j^\star) \geq \alpha$, this pair is also shattered by $F$. Hence $F$ shatters at least

$$|U \cap V| + |U \cup V| = |U| + |V| = 2t(2m, l-1)$$

pairs, which implies that $t(4lM^2 m, l) \geq 2t(2m, l-1)$. This completes the proof of Eq. (13).

With Eq. (12) and Eq. (13), when $r \leq l-1$ we have

$$t(2(2lM^2)^r, l) \geq 2^r t(2, l-r) \geq 2^r.$$

which verifies Eq. (11). □

## A.2 Proof of Proposition 1

*Proof of Proposition 1.* We define distance $\mathfrak{c}' : [M] \times [M] \to \mathbb{R}_+ \cup \{0\}$:

$$\mathfrak{c}'(a, b) = \mathfrak{c}(u_a, u_b). \tag{14}$$

For any $f \in \mathcal{F}$, since $f$ maps $\mathcal{X}$ into $[-1, 1]$, we define $\bar{f} : \mathcal{X} \to [M]$ to be an arbitrary function mapping $\mathcal{X}$ into $[M]$, which satisfies that for any $x \in \mathcal{X}$, $\mathfrak{c}(f(x), u_{\bar{f}(x)}) \leq \beta$. We construct function class $\bar{\mathcal{F}} = \{\bar{f} : f \in \mathcal{F}\} \subseteq \{\mathcal{X} \to [M]\}$. We use $d_{\mathfrak{c}'}(\bar{\mathcal{F}}, \alpha)$ to denote the non-sequential gapped dimension (defined in Definition 1) of integer-valued function class $\bar{\mathcal{F}}$ at scale $\alpha$ under distance $\mathfrak{c}'$, and for simplicity we let $d = d_{\mathfrak{c}'}(\bar{\mathcal{F}}, \alpha)$. Suppose $x_{1:d} \in \mathcal{X}^d$ is shattered by $\mathcal{F}$. Then there exists $\bar{s}_1 = (\bar{s}_1[-1], \bar{s}_1[1]), \ldots, \bar{s}_d = (\bar{s}_d[-1], \bar{s}_d[1])$ such that for any $\boldsymbol{\varepsilon} \in \{-1, 1\}^d$ and $t \in [d]$, $\mathfrak{c}'(\bar{s}_t[-1], \bar{s}_t[1]) \geq \alpha$, and also for any $\boldsymbol{\varepsilon} \in \{-1, 1\}^d$, there exists $\bar{f}^{\boldsymbol{\varepsilon}} \in \bar{\mathcal{F}}$ such that $\bar{f}^{\boldsymbol{\varepsilon}}(x_t) = \bar{s}_t[\varepsilon_t]$ for any $t \in [d]$. Notice from the definition of $\mathcal{F}$ that there exists some $f^{\boldsymbol{\varepsilon}} \in \mathcal{F}$ such that $\bar{f}^{\boldsymbol{\varepsilon}} = \overline{(f^{\boldsymbol{\varepsilon}})}$.

We next verify that $x_{1:n}$ is also shattered by $\mathcal{F}$ at scale $(\alpha, \beta)$, according to the definition Definition 3. We construct $s_1 = (s_1[-1], s_1[1]), \ldots, s_d = (s_d[-1], s_d[1])$ according to $\bar{s}_{1:d}$ as follows:

$$s_t[-1] = u_{\bar{s}_t[-1]} \quad \text{and} \quad s_t[1] = u_{\bar{s}_t[1]}.$$

Then according to the definition of $\mathfrak{c}'$ in Eq. (14), we have for any $t \in [d]$,

$$\mathfrak{c}(s_t[-1], s_t[1]) = \mathfrak{c}'(\bar{s}_t[-1], \bar{s}_t[1]) \geq \alpha.$$

Additionally, for any $\boldsymbol{\varepsilon} \in \{-1, 1\}^d$, there exists some $f^{\boldsymbol{\varepsilon}} \in \mathcal{F}$ such that $\overline{(f^{\boldsymbol{\varepsilon}})}(x_t) = \bar{s}_t[\varepsilon_t]$ for any $t \in [d]$, which implies that

$$\mathfrak{c}(f(x_t), s_t[\varepsilon_t]) = \mathfrak{c}(f(x_t), u_{\bar{s}_t[\varepsilon_t]}) = \mathfrak{c}(f(x_t), u_{\overline{(f^{\boldsymbol{\varepsilon}})}(x_t)}) \leq \beta, \quad \forall t \in [d],$$

where the last inequality follows from the definiton of $\bar{f}$. Therefore, $x_{1:d}$ is shattered by $\mathcal{F}$ at scale $(\alpha, \beta)$, and, according to Definition 3, we have $d(\mathcal{F}, \alpha, \beta) \geq d$.

Next, we will upper bound the non-sequential covering number of $\mathcal{F}$ in terms of $d$. According to Lemma 1, for any $x_{1:n} \in \mathcal{X}^n$ there exists a non-sequential covering $\bar{\mathcal{V}}$ of $\bar{\mathcal{F}}$ with size no more than $\exp\left(16d \log^2(enM)\right)$. Hence, for any $f \in \mathcal{F}$, there exists some $\bar{\mathbf{v}} = (\bar{v}_{1:n}) \in \bar{\mathcal{V}}$ which satisfies

$$\mathfrak{c}'(\bar{f}(x_t), \bar{v}_t) \leq \alpha \qquad \forall t \in [n],$$

which implies that

$$\mathfrak{c}(f(x_t), u_{\bar{v}_t}) \leq \mathfrak{c}(f(x_t), u_{\bar{f}(x_t)}) + \mathfrak{c}(u_{\bar{f}(x_t)}, u_{\bar{v}_t}) \leq \beta + \alpha \qquad \forall t \in [n],$$

where the first inequality uses the triangle inequality of $\mathfrak{c}$, and the second inequality uses the definition of function $\bar{f}$. For every $\bar{\mathbf{v}} \in \bar{\mathcal{V}}$, we construct $\mathbf{v}^{\bar{\mathbf{v}}} = (v^{\bar{\mathbf{v}}}_{1:n}) \in [-1, 1]^n$: for any $t \in [n]$, $v^{\bar{\mathbf{v}}}_t = u_{\bar{v}_t}$. We further let $\mathcal{V} = \{\mathbf{v}^{\bar{\mathbf{v}}} : \bar{\mathbf{v}} \in \bar{\mathcal{V}}\}$. Then $\mathcal{V}$ is an $(\alpha + \beta)$-cover of $\mathcal{F}$ on $x_{1:n}$, which implies

$$\log \mathcal{N}_\infty(\mathcal{F}, \mathbf{x}, \alpha + \beta) \leq \log |\mathcal{V}| = \log |\bar{\mathcal{V}}| \leq 16d \log^2(enM) \leq 16d(\mathcal{F}, \alpha, \beta) \log^2(enM).$$

□

## A.3 Missing Proofs in Section 2.3

*Proof of Proposition 2.* We only need to prove that if $x_{1:d} \in \mathcal{X}^d$ is shattered by $\mathcal{F}$ at scale $(\alpha, \beta)$ according to Definition 3, then $x_{1:d}$ is shattered by $\mathcal{F}$ at scale $\alpha - 2\beta$ according to the traditional notion of shattering as in Definition 5.

If $x_{1:d}$ is shattered by $\mathcal{F}$ at scale $(\alpha, \beta)$ according to Definition 3, then there exists $s_{1:d}$ where $s_t = (s_t[-1], s_t[1]) \in [-1, 1] \times [-1, 1]$ for any $t \in [d]$, such that $s_t[1] - s_t[-1] \geq \alpha$ for any $t \in [d]$, and also for any $\boldsymbol{\varepsilon} \in \{-1, 1\}^d$, there exists a function $f^{\boldsymbol{\varepsilon}} \in \mathcal{F}$ which satisfies $|f^{\boldsymbol{\varepsilon}}(x_t) - s_t[\varepsilon_t]| \leq \beta$ for any $t \in [d]$. We let

$$v_t = \frac{s_t[-1] + s_t[1]}{2}, \quad \forall t \in [n].$$

Since $\alpha > 2\beta$, we have for any $\boldsymbol{\varepsilon} \in \{-1, 1\}^d$,

$$\varepsilon_t \cdot (f^{\boldsymbol{\varepsilon}}(x_t) - v_t) \geq \frac{\alpha}{2} - \beta,$$

which implies that $x_{1:d}$ is shattered by $\mathcal{F}$ at scale $\alpha - 2\beta$, according to Definition 5. □

*Proof of Proposition 3.* Let $d = \mathsf{vc}(\mathcal{F}, \alpha)$, then there exists $x_{1:d}$ shattered by $\mathcal{F}$ according to Definition 5. Hence there exists $v_{1:d} \in [-1, 1]^d$ and also function $f^{\boldsymbol{\varepsilon}} \in \mathcal{F}$ for each $\boldsymbol{\varepsilon} \in \{-1, 1\}^d$, such that for any $\boldsymbol{\varepsilon} \in \{-1, 1\}^d$,

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

Further notice for every $\boldsymbol{\varepsilon} \neq \boldsymbol{\varepsilon}'$, there exists $t \in [d]$ such that $\varepsilon_t \neq \varepsilon'_t$, which implies that for this specific $t$,

$$|f^{\boldsymbol{\varepsilon}}(x_t) - f^{\boldsymbol{\varepsilon}'}(x_t)| \geq \alpha.$$

Hence $\{f^{\boldsymbol{\varepsilon}} : \boldsymbol{\varepsilon} \in \{-1, 1\}^d\}$ forms a $\alpha$-packing of $\mathcal{F}$ under $\ell_\infty$-norm under design $x_{1:d}$. Therefore, the covering-packing duality indicates that

$$\mathcal{N}_\infty(\mathcal{F}, x_{1:\mathsf{vc}(\mathcal{F},\alpha)}, \alpha/3) \geq 2^d = 2^{\mathsf{vc}(\mathcal{F},\alpha)}. \tag{15}$$

Next according to Proposition 1, we have for any $n$ and $x_{1:n} \in \mathcal{X}^n$,

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

Hence by replacing $\alpha$ in Eq. (15) with $3(\alpha + \beta)$, and choosing $n$ to be $\mathsf{vc}(\mathcal{F}, 3(\alpha + \beta))$, we obtain that

$$\begin{aligned}
\log 2 \cdot \mathsf{vc}(\mathcal{F}, 3(\alpha + \beta)) &\leq \sup_{x_{1:\mathsf{vc}(\mathcal{F},\alpha+\beta)}} \log \mathcal{N}_\infty(\mathcal{F}, x_{1:\mathsf{vc}(\mathcal{F},3(\alpha+\beta))}, \alpha + \beta) \\
&\leq 16 d(\mathcal{F}, \alpha, \beta) \log^2\left(\frac{2e \cdot \mathsf{vc}(\mathcal{F}, 3(\alpha + \beta))}{\beta}\right),
\end{aligned}$$

which implies

$$\mathsf{vc}(\mathcal{F}, 3(\alpha + \beta)) \leq 32 d(\mathcal{F}, \alpha, \beta) \log^2\left(\frac{6\mathsf{vc}(\mathcal{F}, 3(\alpha + \beta))}{\beta}\right). \tag{16}$$

Additionally, we notice that $\log(x) \leq \sqrt{2} \cdot \sqrt[3]{x}$, hence

$$\mathsf{vc}(\mathcal{F}, 3(\alpha + \beta)) \leq 64 d(\mathcal{F}, \alpha, \beta) \left(\frac{6\mathsf{vc}(\mathcal{F}, 3(\alpha + \beta))}{\beta}\right)^{2/3},$$

which implies that

$$\mathsf{vc}(\mathcal{F}, 3(\alpha + \beta)) \leq 64^3 \cdot 6^2 \cdot \frac{d(\mathcal{F}, \alpha, \beta)^3}{\beta^2}.$$

Bringing this back to Eq. (16), we obtain that

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

The second part of Proposition 3 is implied by Proposition 9 below. Indeed, if $x_{1:d}$ is shattered by $\mathcal{F}$ at scale $\alpha$ according to Definition 11, then $x_{1:d}$ is also shattered by $\mathcal{F}$ at scale $(\alpha, \beta)$ according to Definition 3. □

Finally, we provide the proof of Proposition 4.

*Proof of Proposition 4.* Let $d = \lfloor \log(1/\alpha) \rfloor$ and $\mathcal{X} = \{x_1, \dots, x_d\}$. For any $\boldsymbol{\varepsilon} = (\varepsilon_{1:d}) \in \{-1, 1\}^d$, we define

$$a^{\boldsymbol{\varepsilon}} = \alpha + \sum_{i=1}^{d} \frac{\varepsilon_i + 1}{2} \cdot 2^{i-1} \cdot \alpha.$$

Then for any $\boldsymbol{\varepsilon}$, we have

$$\alpha \le a^{\boldsymbol{\varepsilon}} \le \alpha + \alpha \cdot (2^d - 1) \le 1.$$

We define function class $\mathcal{F} = \{f^{\boldsymbol{\varepsilon}} : \boldsymbol{\varepsilon} \in \{-1, 1\}^d\}$, where $f^{\boldsymbol{\varepsilon}} : \mathcal{X} \to [-1, 1]$ is defined as

$$f^{\boldsymbol{\varepsilon}}(x_i) = \varepsilon_i \cdot a^{\boldsymbol{\varepsilon}}, \quad \forall 1 \le i \le d.$$

Then it is easy to see that $\{x_1, \dots, x_d\}$ is shattered by $\mathcal{F}$ in terms of the classical shattering (defined in Definition 5).

Next, we will verify that the non-sequential gapped dimension $d(\mathcal{F}, \alpha, \beta) = 1$ for any $\beta < \alpha/2$. First of all, it is easy to see that $\{x_1\}$ is shattered by $\mathcal{F}$, hence $d(\mathcal{F}, \alpha, \beta) \ge 1$. If there exists a size-two subset $\{x_i, x_j\}$ of $\mathcal{X}$ shattered by $\mathcal{F}$, in terms of Definition 1, then there exists $s_i = (s_i[-1], s_i[1]) \in [-1, 1] \times [-1, 1]$ and $s_j = (s_j[-1], s_j[1]) \in [-1, 1] \times [-1, 1]$ such that for any $\mathbf{e} = (e_i, e_j) \in \{-1, 1\} \times \{-1, 1\}$,

$$\exists\, \boldsymbol{\varepsilon}[\mathbf{e}] \in \{-1, 1\}^d \quad \text{such that} \quad |f^{\boldsymbol{\varepsilon}[\mathbf{e}]}(x_i) - s_i[e_i]| \le \beta \quad \text{and} \quad |f^{\boldsymbol{\varepsilon}[\mathbf{e}]}(x_j) - s_j[e_j]| \le \beta.$$

Hence, we obtain

$$|f^{\boldsymbol{\varepsilon}[(-1,-1)]}(x_i) - f^{\boldsymbol{\varepsilon}[(-1,1)]}(x_i)| \le 2\beta < \alpha.$$

According to the construction of function $f^{\boldsymbol{\varepsilon}}$, we must have $\boldsymbol{\varepsilon}[(-1, -1)] = \boldsymbol{\varepsilon}[(-1, 1)]$, which implies that

$$|s_j[1] - s_j[-1]| \le |s_j[1] - f^{\boldsymbol{\varepsilon}[(-1,1)]}(x_j)| + |s_j[-1] - f^{\boldsymbol{\varepsilon}[(-1,-1)]}(x_j)| \le \beta + \beta < \alpha.$$

This violates the definition of shattering in Definition 1. Therefore, we have verified that $d(\mathcal{F}, \alpha, \beta) \le 1$ for any $\beta < \alpha/2$. □

## A.4 Properties of Fixed-Scale Scale-Sensitive Dimension

We consider the following fixed-scale scale-sensitive dimension; the only modification with respect to Definition 5 is that the inequality is turned into an equality. In terms of Figure 1, the requirement of shattering states that the vertices of the hypercube with side-length $\alpha$ are in the set.

**Definition 11** (Fixed-Scale Dimension)**.** *Given a function class $\mathcal{F} \subseteq \{\mathcal{X} \to [-1, 1]\}$, we say that a set $\{x_1, x_2, \dots, x_d\} \subseteq \mathcal{X}$ is shattered by $\mathcal{F}$ at scale $\alpha > 0$, if there exists $s_1, \dots, s_d \in [-1, 1]$ such that for any $\boldsymbol{\varepsilon} = (\varepsilon_{1:d}) \in \{-1, 1\}^d$, there exists some $f^{\boldsymbol{\varepsilon}} \in \mathcal{F}$ such that*

$$\varepsilon_t \cdot (f^{\boldsymbol{\varepsilon}}(x_t) - s_t) = \frac{\alpha}{2}, \qquad \forall t \in [d].$$

*The fixed-scale dimension $\mathsf{vc}_{\mathsf{fix}}(\mathcal{F}, \alpha)$ of $\mathcal{F}$ at scale $\alpha$ is the largest $d$ such that there exists a size-$d$ subset of $\mathcal{X}$ shattered by $\mathcal{F}$.*

We have the following proposition showing that if the function class $\mathcal{F}$ is convex, then the scale-sensitive dimensions defined in Definition 5 and in Definition 11 coincide.

**Proposition 9.** *If function class $\mathcal{F}$ is convex, then for any $\alpha > 0$, $\mathsf{vc}(\mathcal{F}, \alpha) = \mathsf{vc}_{\mathsf{fix}}(\mathcal{F}, \alpha)$.*

*Proof of Proposition 9.* According to Definition 5 and Definition 11, we have $\mathsf{vc}(\mathcal{F}, \alpha) \geq \mathsf{vc}_{\mathsf{fix}}(\mathcal{F}, \alpha)$. Hence we only need to prove that $\mathsf{vc}_{\mathsf{fix}}(\mathcal{F}, \alpha) \geq \mathsf{vc}(\mathcal{F}, \alpha)$. In the following, we will show that if $\{x_1, \ldots, x_d\}$ is shattered by $\mathcal{F}$ at scale $\alpha$ under witness $s_1, \ldots, s_d$ under Definition 5, then $\{x_1, \ldots, x_d\}$ is shattered by $\mathcal{F}$ at scale $\alpha$ under witness $s_1, \ldots, s_d$ under Definition 11.

Since $\{x_1, \ldots, x_d\}$ is shattered by $\mathcal{F}$ at scale $\alpha$ under witness $s_1, \ldots, s_d$ under Definition 5, for any $\boldsymbol{\varepsilon} \in \{-1, 1\}^d$ there exists $f^{\boldsymbol{\varepsilon}}$ such that

$$\varepsilon_t \cdot (f^{\boldsymbol{\varepsilon}}(x_t) - s_t) \geq \frac{\alpha}{2} \qquad \forall t \in [d].$$

We next iteratively construct $f_i^{\boldsymbol{\varepsilon}} \in \mathcal{F}$ for $i = 0, 1, \ldots, d$ such that $f_i^{\boldsymbol{\varepsilon}}$ satisfies

$$\varepsilon_t \cdot (f^{\boldsymbol{\varepsilon}_i}(x_t) - s_t) = \frac{\alpha}{2} \quad \text{if } t \leq i \qquad \text{and} \qquad \varepsilon_t \cdot (f^{\boldsymbol{\varepsilon}_i}(x_t) - s_t) \geq \frac{\alpha}{2} \quad \text{if } t > i. \tag{17}$$

For $i = 0$, we let $f_0^{\boldsymbol{\varepsilon}} = f^{\boldsymbol{\varepsilon}}$ and they satisfies conditions in Eq. (17). Suppose we have constructed $f_i^{\boldsymbol{\varepsilon}}$, and we will now construct $f_{i+1}^{\boldsymbol{\varepsilon}}$. For any $\boldsymbol{\varepsilon} = (\varepsilon_{1:d}) \in \{-1, 1\}^d$, we let

$$\tilde{\boldsymbol{\varepsilon}} = (\varepsilon_1, \ldots, \varepsilon_i, -\varepsilon_{i+1}, \varepsilon_{i+2}, \ldots, \varepsilon_d). \tag{18}$$

Let

$$f_{i+1}^{\boldsymbol{\varepsilon}} = \alpha^{\boldsymbol{\varepsilon}} \cdot f_i^{\boldsymbol{\varepsilon}} + (1 - \alpha^{\boldsymbol{\varepsilon}}) \cdot f_i^{\tilde{\boldsymbol{\varepsilon}}}, \tag{19}$$

where

$$\alpha^{\boldsymbol{\varepsilon}} = \begin{cases} \frac{\alpha/2 + s_{i+1} - f^{\tilde{\boldsymbol{\varepsilon}}}(x_{i+1})}{f_i^{\tilde{\boldsymbol{\varepsilon}}}(x_{i+1}) - f_i^{\boldsymbol{\varepsilon}}(x_{i+1})} & \text{if } \varepsilon_{i+1} = 1, \\ \frac{\alpha/2 + f^{\tilde{\boldsymbol{\varepsilon}}}(x_{i+1}) - s_{i+1}}{f_i^{\tilde{\boldsymbol{\varepsilon}}}(x_{i+1}) - f_i^{\boldsymbol{\varepsilon}}(x_{i+1})} & \text{if } \varepsilon_{i+1} = -1. \end{cases}$$

According to Eq. (17), we can verify that $\alpha^{\tilde{\boldsymbol{\varepsilon}}} \in [0, 1]$ for any $\boldsymbol{\varepsilon}$. Therefore $f_{i+1}^{\boldsymbol{\varepsilon}}$ constructed in Eq. (19) belongs to $\mathcal{F}$ due to the convexity of $\mathcal{F}$. Next, we will verify that $f_{i+1}^{\boldsymbol{\varepsilon}}$ also satisfies Eq. (17). For $t \leq i$, since

$$f_i^{\boldsymbol{\varepsilon}}(x_t) = s_t + \varepsilon_t \cdot \frac{\alpha}{2}$$

according to Eq. (17), according to our choice of $\tilde{\boldsymbol{\varepsilon}}$ in Eq. (18) we have

$$\varepsilon_t \cdot (f_i^{\boldsymbol{\varepsilon}}(x_t) - s_t) = \frac{\alpha}{2}.$$

For $t = i + 1$, based on our construction in Eq. (19) and our choice of $\tilde{\boldsymbol{\varepsilon}}$ in Eq. (18), we can calculate that

$$\varepsilon_t \cdot (f_i^{\boldsymbol{\varepsilon}}(x_t) - s_t) = \frac{\alpha}{2}.$$

For $t > i + 1$, since

$$\varepsilon_t \cdot (f_i^{\boldsymbol{\varepsilon}}(x_t) - s_t) \geq \frac{\alpha}{2}$$

according to Eq. (17), hence according to our choice of $\tilde{\boldsymbol{\varepsilon}}$ in Eq. (18) we have

$$\varepsilon_t \cdot (f_i^{\boldsymbol{\varepsilon}}(x_t) - s_t) \geq \frac{\alpha}{2}.$$

Therefore, $f_{i+1}^{\boldsymbol{\varepsilon}}$ satisfies Eq. (17) as well. By choosing $i = d$, we obtain that there exist $f_d^{\boldsymbol{\varepsilon}} \in \mathcal{F}$ such that

$$\varepsilon_t \cdot (f_d^{\boldsymbol{\varepsilon}}(x_t) - s_t) = \frac{\alpha}{2},$$

which implies that $\{x_1, \ldots, x_d\}$ is shattered by $\mathcal{F}$ at scale $\alpha$ under witness $s_1, \ldots, s_d$ under Definition 11. □

## A.5 Missing Proofs in Section 2.4

*Proof of Theorem 1.*

$$\alpha_n = \frac{1}{2} \cdot \arg\min_{\alpha>0} \left\{ d(\mathcal{F}, \alpha, \alpha/(20nC)) \cdot \frac{16}{\alpha^2} \le n \right\}.$$

Since $d(\mathcal{F}, \alpha, \alpha/(20nC)) = \tilde{\Omega}\left(\alpha^{-p}\right)$ for every $\alpha \ge 0$, we have

$$d(\mathcal{F}, \alpha_n, \alpha_n/(20nC)) \wedge \frac{n\alpha_n^2}{4} = \tilde{\Omega}\left(n^{\frac{p}{p+2}}\right), \qquad \forall n \in \mathbb{Z}_+.$$

In the following, we will prove that for any positive integer $n$, we have

$$\sup_{\boldsymbol{\mu}, \mathbf{x}} \mathbb{E}[\mathcal{R}_n(\mathcal{F}, \boldsymbol{\mu}, \mathbf{x}, \boldsymbol{\varepsilon})] = \Omega\left(d(\mathcal{F}, \alpha_n, \alpha_n/(20nC))\right).$$

We fix $n$, and, for brevity, write

$$\alpha = \alpha_n \quad \text{and} \quad d = d(\mathcal{F}, \alpha_n, \alpha_n/(20nC)) \wedge n\alpha^2/4 \le d(\mathcal{F}, \alpha_n, \alpha_n/(20nC)).$$

Let $\tilde{x}_{1:d} \in \mathcal{X}^d$ be a sequence shattered by $\mathcal{F}$ in the sense of Definition 3. That is, there exist $s_1 = (s_1[-1], s_1[1]), s_2 = (s_2[-1], s_2[1]), \ldots, s_d = (s_d[-1], s_d[1]) \in [-1,1] \times [-1,1]$ such that for any $\tilde{\boldsymbol{\varepsilon}} = (\tilde{\varepsilon}_{1:d}) \in \{-1,1\}^d$, there exists $f^{\tilde{\boldsymbol{\varepsilon}}} \in \mathcal{F}$ such that

$$\left| f^{\tilde{\boldsymbol{\varepsilon}}}(\tilde{x}_t) - s_t[\tilde{\varepsilon}_t] \right| \le \frac{\alpha}{20(nC)} \quad \text{and} \quad |s_t[1] - s_t[-1]| \ge \alpha \qquad \forall t \in [d]. \tag{20}$$

Without loss of generality, we assume $s_t[1] \ge s_t[-1]$. We now define $\tilde{\mu}_{1:d} \in [-1,1]^d$ as $\tilde{\mu}_t = (s_t[-1] + s_t[1])/2$ and

$$k_t = \left\lfloor \frac{1}{(s_t[1] - s_t[-1])^2} \right\rfloor \vee 1 \le 4\alpha^{-2}$$

for any $t \in [d]$. Then we have

$$\sum_{t=1}^{d} k_t \le n.$$

Next, we construct $x_{1:n}$ and $\mu_{1:n}$ with the following block structure:

$$(x_{1:n}) = (\underbrace{\tilde{x}_1, \tilde{x}_1, \ldots, \tilde{x}_1}_{k_1 \text{ terms}}, \underbrace{\tilde{x}_2, \tilde{x}_2, \ldots, \tilde{x}_2}_{k_2 \text{ terms}}, \ldots, \underbrace{\tilde{x}_{d-1}, \tilde{x}_{d-1}, \ldots, \tilde{x}_{d-1}}_{k_{d-1} \text{ terms}}, \underbrace{\tilde{x}_d, \tilde{x}_d, \ldots, \tilde{x}_d}_{n-k_1-\ldots-k_{d-1} \text{ terms}}),$$

$$(\mu_{1:n}) = (\underbrace{\tilde{\mu}_1, \tilde{\mu}_1, \ldots, \tilde{\mu}_1}_{k_1 \text{ terms}}, \underbrace{\tilde{\mu}_2, \tilde{\mu}_2, \ldots, \tilde{\mu}_2}_{k_2 \text{ terms}}, \ldots, \underbrace{\tilde{\mu}_{d-1}, \tilde{\mu}_{d-1}, \ldots, \tilde{\mu}_{d-1}}_{k_{d-1} \text{ terms}}, \underbrace{s_d[1], s_d[1], \ldots, s_d[1]}_{n-k_1-\ldots-k_d \text{ terms}}).$$

Next we write $\mathcal{R}_n(\mathcal{F}, \mu_{1:n}, x_{1:n}, \boldsymbol{\varepsilon})$ in terms of segments 1 to $d$. For any $\boldsymbol{\varepsilon} \in \{0,1\}^n$, we construct $\tilde{\boldsymbol{\varepsilon}} \in \{-1,1\}^d$:

$$\tilde{\varepsilon}_t = 2\mathbb{I}\left\{ \sum_{j=1}^{k_t} \varepsilon_{k_1+\ldots+k_{t-1}+j} \ge 0 \right\} - 1, \qquad \forall t \in [d-1] \quad \text{and} \quad \tilde{\varepsilon}_d = 1.$$

Recall that $f^{\tilde{\boldsymbol{\varepsilon}}} \in \mathcal{F}$ for any sequence $\tilde{\boldsymbol{\varepsilon}} \in \{-1,1\}^d$. Then, using $[t,j]$ to denote $k_1 + \ldots + k_t + j$, and defining the random variable $\widehat{\varepsilon}_t = \frac{1}{k_t} \sum_{j=1}^{k_t} \varepsilon_{[t-1,j]} \in [-1,1]$ for fixed $\boldsymbol{\varepsilon}$, we obtain

$$\begin{aligned}
&\mathcal{R}_n(\mathcal{F}, \mu_{1:n}, x_{1:n}, \boldsymbol{\varepsilon}) \\
&\quad = \sup_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{d-1} \sum_{j=1}^{k_t} C \cdot \varepsilon_{[t-1,j]}(f(x_{[t-1,j]}) - \mu_{[t-1,j]}) - (f(x_{[t-1,j]}) - \mu_{[t-1,j]})^2 \right.
\end{aligned}$$

$$+ \sum_{j=1}^{n-k_1-\ldots-k_{d-1}} C \cdot \varepsilon_{[d-1,j]}(f(x_{[d-1,j]}) - \mu_{[d-1,j]}) - (f(x_{[d-1,j]}) - \mu_{[d-1,j]})^2 \Bigg\}$$

$$\overset{(i)}{\geq} \sum_{t=1}^{d-1} \sum_{j=1}^{k_t} C \cdot \varepsilon_{[t-1,j]}(f^{\tilde{\boldsymbol{\varepsilon}}}(\tilde{x}_t) - \tilde{\mu}_t) - (f^{\tilde{\boldsymbol{\varepsilon}}}(\tilde{x}_t) - \tilde{\mu}_t)^2 - n \cdot 2C \cdot \frac{\alpha}{nC}$$

$$\geq \sum_{t=1}^{d-1} k_t \cdot \left(C \cdot \widehat{\varepsilon}_t (f^{\tilde{\boldsymbol{\varepsilon}}}(\tilde{x}_t) - \tilde{\mu}_t) - (f^{\tilde{\boldsymbol{\varepsilon}}}(\tilde{x}_t) - \tilde{\mu}_t)^2\right) - 4, \tag{21}$$

where $(i)$ uses the choice $f = f^{\tilde{\boldsymbol{\varepsilon}}} \in \mathcal{F}$, and also $|f^{\tilde{\boldsymbol{\varepsilon}}}(\tilde{x}_d) - s_d[1]| \leq \alpha/(20nC) < \alpha/(nC)$ since $\tilde{\varepsilon}_d = 1$, and also the fact that $\alpha \leq 2$ due to Eq. (20). According to Eq. (20), we have for any $\tilde{\boldsymbol{\varepsilon}}$,

$$|f^{\tilde{\boldsymbol{\varepsilon}}}(\tilde{x}_t) - s_t[\tilde{\varepsilon}_t]| \leq \frac{\alpha}{20} \quad \text{and} \quad s_t[\tilde{\varepsilon}_t] - \mu_t = \operatorname{sign}(\widehat{\varepsilon}_t) \cdot \frac{s_t[1] - s_t[-1]}{2}.$$

Eq. (20) also gives that $s_t[1] - s_t[-1] \geq \alpha$, which implies

$$\frac{9}{10} \cdot \frac{s_t[1] - s_t[-1]}{2} \leq \operatorname{sign}(\widehat{\varepsilon}_t)(f^{\tilde{\boldsymbol{\varepsilon}}}(\tilde{x}_t) - \tilde{\mu}_t) \leq \frac{11}{10} \cdot \frac{s_t[1] - s_t[-1]}{2}.$$

Therefore we can lower bound Eq. (21) by

$$\mathcal{R}_n(\mathcal{F}, \mu_{1:n}, x_{1:n}, \boldsymbol{\varepsilon}) \geq \sum_{t=1}^{d-1} k_t \cdot \left(C|\widehat{\varepsilon}_t| \cdot \frac{9}{10} \cdot \frac{s_t[1] - s_t[-1]}{2} - \frac{121}{100} \cdot \left(\frac{s_t[1] - s_t[-1]}{2}\right)^2\right) - 4.$$

Next, we can further lower bound

$$\begin{aligned}
&\mathcal{R}_n(\mathcal{F}, \mu_{1:n}, x_{1:n}, \boldsymbol{\varepsilon}) \\
&\geq \sum_{t=1}^{d-1} k_t \cdot \left(C|\widehat{\varepsilon}_t| \cdot \frac{9}{10} \cdot \frac{s_t[1] - s_t[-1]}{2} - \frac{121}{100} \cdot \left(\frac{s_t[1] - s_t[-1]}{2}\right)^2\right) - 4 \\
&= \sum_{t=1}^{d-1} k_t \cdot \left(C\mathbb{E}\left[|\widehat{\varepsilon}_t|\right] \cdot \frac{9}{10} \cdot \frac{s_t[1] - s_t[-1]}{2} - \frac{121}{100} \cdot \left(\frac{s_t[1] - s_t[-1]}{2}\right)^2\right) - 4 \\
&\geq \sum_{t=1}^{d-1} k_t \cdot \left(C\sqrt{\frac{1}{2k_t}} \cdot \frac{9}{10} \cdot \frac{s_t[1] - s_t[-1]}{2} - \frac{121}{100} \cdot \left(\frac{s_t[1] - s_t[-1]}{2}\right)^2\right) - 4
\end{aligned} \tag{22}$$

where the last inequality uses the Khintchine inequality [Haa81]. According to our choice of $k_t$,

$$k_t = \left\lfloor \frac{1}{(s_t[1] - s_t[-1])^2} \right\rfloor \vee 1 \in \left[\frac{1}{2(s_t[1] - s_t[-1])^2}, \frac{4}{(s_t[1] - s_t[-1])^2}\right],$$

which implies that $1/\sqrt{2} \leq \sqrt{k_t}(s_t[1] - s_t[-1]) \leq 2$. Therefore, since $C \geq 2$, we have

$$\text{RHS of } \textit{Eq. (22)} \geq \sum_{t=1}^{d-1} \left(\frac{9C}{20\sqrt{2}} - \frac{121}{200}\right) \cdot \sqrt{k_t} \cdot |s_t[1] - s_t[-1]| - 4 \geq \frac{d-1}{50} - 4.$$

Therefore, we obtain that

$$\sup_{\mu_{1:n}, x_{1:n}} \mathbb{E}[\mathcal{R}_n(\mathcal{F}, \mu_{1:n}, x_{1:n}, \boldsymbol{\varepsilon})] \geq \frac{d-1}{50} - 4 = \tilde{\Omega}\left(n^{\frac{p}{p+2}}\right).$$

□

*Proof of Lemma 2.* We first transform $\mathcal{V}_n(\mathcal{F})$ into the following dual form

$$\begin{aligned}\mathcal{V}_n(\mathcal{F}) &= \sup_{x_{1:n}\in\mathcal{X}^n}\left\{\inf_{\widehat{y}_t}\sup_{p_t\in\Delta([-2,2])}\mathbb{E}_{y_t\sim p_t}\right\}_{t=1}^{n}\left[\sum_{t=1}^{n}(\widehat{y}_t-y_t)^2-\inf_{f\in\mathcal{F}}\sum_{t=1}^{n}(f(x_t)-y_t)^2\right]\\ &= \sup_{x_{1:n}\in\mathcal{X}^n}\left\{\sup_{p_t\in\Delta([-2,2])}\mathbb{E}_{y_t\sim p_t}\right\}_{t=1}^{n}\left[\sum_{t=1}^{n}(\mathbb{E}[y_t\mid y_{1:t-1}]-y_t)^2-\inf_{f\in\mathcal{F}}\sum_{t=1}^{n}(f(x_t)-y_t)^2\right]\\ &= \sup_{x_{1:n}}\sup_{\mathbf{p}\in\Delta([-2,2]^n)}\mathbb{E}_{\mathbf{y}\sim\mathbf{p}}\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^{n}2(y_t-\mu_t(\mathbf{y}))(f(x_t)-\mu_t(\mathbf{y}))-(f(x_t)-\mu_t(\mathbf{y}))^2\right]\end{aligned}$$

where we use $\mu_t(\mathbf{y})$ to denote the expectation of $y_t$ conditioned on $y_{1:t-1}$, i.e. $\mu_t(\mathbf{y})=\mathbb{E}[y_t\mid y_{1:t-1}]$. Taking $\mathbf{p}=p_1\otimes p_2\otimes\ldots\otimes p_n$ where $p_t\in\Delta([-2,2])$ for each $t\in[n]$ in the above equation, we obtain that

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

where $\mu_t\coloneqq\mathbb{E}[y_t]$ denotes the expectation of distribution $p_t$.

We fix $\mu_1,\ldots,\mu_n\in[-1,1]$, and choose distributions $p_t=\mathrm{Unif}(\{\mu_t-1,\mu_t+1\})\subseteq\Delta([-2,2])$ for all $t\in[n]$. Then we obtain that

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

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

*Proof of Corollary 1.* Since $\sup_{x_{1:n}}\log\mathcal{N}_\infty(\mathcal{F},x_{1:n},\alpha)=\tilde{\Omega}\left(\alpha^{-p}\right)$, according to Proposition 1 we have

$$d(\mathcal{F},\alpha,\alpha/(40n))=\tilde{\Omega}\left(\alpha^{-p}\right).$$

Next, we call Lemma 2, and Theorem 1 with $C=2$. Then we obtain

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

□

# B Missing Proofs in Section 3

## B.1 Proof of Lemma 3

*Proof of Lemma 3.* We first define function

$$g_M(n,d)=\sum_{i=0}^{d}\binom{n}{i}\cdot(M-1)^i,$$

which satisfies (see [RST10])

$$g_M(n,d)=g_M(n-1,d)+(M-1)\cdot g_M(n-1,d-1). \tag{23}$$

We will prove the present Lemma by induction with the following induction statement:

$\mathfrak{G}(d,n)$: For any function class $\mathcal{F}\subseteq\{f:\mathcal{X}\to[M]\}$ with $d^{\mathsf{seq}}(\mathcal{F},\alpha)\leq d$, and any depth-$n$ $\mathcal{X}$-valued tree $\mathbf{x}$, $\mathcal{N}_\infty^{\mathsf{seq}}(\mathcal{F},\mathbf{x},\alpha)\leq g_M(d,n)$.

**Base:** There are two base cases to consider: $n \le d$ and $d = 0$.

When $n \le d$, we let

$$\mathcal{V} = \{\mathbf{v}[i_1, i_2, \ldots, i_n] : i_1, \ldots, i_n \in [M]\},$$

where $\mathbf{v}[i_1, \ldots, i_n]$ denotes the tree with values $i_t$ at depth $t$ along any path. Then it is easy to see that for any $f \in \mathcal{F}$, depth-$n$ $\mathcal{X}$-valued tree $\mathbf{x}$, and any path $\boldsymbol{\varepsilon} \in \{-1, 1\}^n$, there exists some $\mathbf{v} \in \mathcal{V}$ such that $f(x_t(\boldsymbol{\varepsilon})) = v_t(\boldsymbol{\varepsilon})$ for all $t \in [n]$. Hence $\mathcal{V}$ is an exact (that is, $\alpha = 0$) cover of $\mathcal{F}$ on $\mathbf{x}$; hence, $\mathcal{V}$ is also an $\alpha$-sequential covering as well. Thus, we have

$$\mathcal{N}_\infty^{\mathsf{seq}}(\mathcal{F}, \mathbf{x}, \alpha) \le |\mathcal{V}| = M^n = \sum_{i=0}^{d} \binom{n}{i} \cdot (M-1)^i = g_M(n, d).$$

When $d = 0$, there is no depth-1 $\mathcal{X}$-valued tree which is shattered by $\mathcal{F}$ at scale $\alpha$ under the metric $\mathfrak{c}$. This implies for any $x \in \mathcal{X}$ and functions $f, f' \in \mathcal{F}$, we always have $\mathfrak{c}(f(x), f'(x)) \le \alpha$. Indeed, otherwise we can construct depth-1 tree $\mathbf{x}$ with $x_1 = x$ and depth-1 tree $\mathbf{s}$ with $s_1 = (f(x), f'(x))$, and $\mathbf{x}$ is shattered by $\mathcal{F}$. We let $f_0 \in \mathcal{F}$ to be an arbitrary function in $\mathcal{F}$. For any depth-$n$ $\mathcal{X}$-valued $\mathbf{x}$, we construct depth-$n$ $[-1, 1]$-valued tree $\mathbf{v}$ which takes the value $f_0(x_t(\boldsymbol{\varepsilon}))$ at depth $t$ along any path $\boldsymbol{\varepsilon}$. Then for any $f \in \mathcal{F}$ and path $\boldsymbol{\varepsilon} \in \{-1, 1\}$, we always have $\mathfrak{c}(f(x_t(\boldsymbol{\varepsilon})), v_t(\boldsymbol{\varepsilon})) = \mathfrak{c}(f(x_t(\boldsymbol{\varepsilon})), f_0(x_t(\boldsymbol{\varepsilon}))) \le \alpha$. Hence $\mathcal{V}$ is an $\alpha$-sequential covering of $\mathcal{F}$ on $\mathbf{x}$, and it satisfies $|\mathcal{V}| = 1 = g_M(n, 0)$.

**Induction:** Suppose the induction hypotheses $\mathfrak{G}(n-1, d-1)$ and $\mathfrak{G}(n-1, d)$ both hold. We will prove induction statement $\mathfrak{G}(n, d)$. For fixed function class $\mathcal{F}$ with $d^{\mathsf{seq}}(\mathcal{F}, \alpha) = d$ and depth-$n$ $\mathcal{X}$-valued tree $\mathbf{x}$, we will construct a $\alpha$-sequential covering to of $\mathcal{F}$ on $\mathbf{x}$ whose size is no more than $g_M(n, d)$. Suppose the root of tree $\mathbf{x}$ is $x_1 \in \mathcal{X}$, the left subtree of $x_1$ is denoted as $\mathbf{x}^l$, and the right subtree of $x_1$ is denoted as $\mathbf{x}^r$. We partition the function class $\mathcal{F}$ as:

$$\mathcal{F} = \mathcal{F}_1 \cup \mathcal{F}_2 \cup \ldots \cup \mathcal{F}_M \quad \text{where} \quad \mathcal{F}_k = \{f \in \mathcal{F} : f(x_1) = k\}, \quad \forall 1 \le k \le M.$$

Then we have $d^{\mathsf{seq}}(\mathcal{F}_k, \alpha) \le d^{\mathsf{seq}}(\mathcal{F}, \alpha) = d$ for all $k \in [M]$. We let $\mathcal{K} = \{k \in [M] : d^{\mathsf{seq}}(\mathcal{F}_k, \alpha) = d\}$. Then for any $a, b \in \mathcal{K}$ and $a < b$, there exist depth-$d$ $\mathcal{X}$-valued trees $\mathbf{x}^a$ and $\mathbf{x}^b$, and also depth-$d$ $[M] \times [M]$-valued trees $\mathbf{s}^a$ and $\mathbf{s}^b$ such that for any $\boldsymbol{\varepsilon} \in \{-1, 1\}^d$, there exists $f_a^{\boldsymbol{\varepsilon}} \in \mathcal{F}_a$ and $f_b^{\boldsymbol{\varepsilon}} \in \mathcal{F}_b$ such that for any $t \in [d]$,

$$\begin{aligned} f_a^{\boldsymbol{\varepsilon}}(x_t^a(\boldsymbol{\varepsilon})) = s_t^a(\boldsymbol{\varepsilon})[\varepsilon_t] \quad \text{and} \quad f_b^{\boldsymbol{\varepsilon}}(x_t^b(\boldsymbol{\varepsilon})) = s_t^b(\boldsymbol{\varepsilon})[\varepsilon_t], \\ \mathfrak{c}(s_t^a(\boldsymbol{\varepsilon})[-1], s_t^a(\boldsymbol{\varepsilon})[1]) \ge \alpha \quad \text{and} \quad \mathfrak{c}(s_t^b(\boldsymbol{\varepsilon})[-1], s_t^b(\boldsymbol{\varepsilon})[1]) \ge \alpha. \end{aligned}$$

For the sake of a contradiction, suppose it holds that $\mathfrak{c}(a, b) \ge \alpha$. Then we can construct a depth-$(d+1)$ $\mathcal{X}$-valued tree $\mathbf{x}$ with root $x_1$, left subtree of the root being $\mathbf{x}^a$, and right subtree of the root being $\mathbf{x}^b$, and also a depth-$(d+1)$ $[M] \times [M]$-valued tree $\mathbf{s}$ with root $(a, b) \in [M] \times [M]$, left subtree of the root being $\mathbf{s}^a$, and right subtree of the root being $\mathbf{s}^b$. Then we can verify that for any $\boldsymbol{\varepsilon} \in \{-1, 1\}^{d+1}$, and any $t \in [d+1]$, we have $s_t(\boldsymbol{\varepsilon})[-1] < s_t(\boldsymbol{\varepsilon})[1]$, and $\mathfrak{c}(s_t(\boldsymbol{\varepsilon})[-1], s_t(\boldsymbol{\varepsilon})[1]) \ge \alpha$. Further, we let $\boldsymbol{\varepsilon}' = (\varepsilon_2, \varepsilon_3, \ldots, \varepsilon_{d+1}) \in \{-1, 1\}^d$, and if $\varepsilon_1 = -1$, then letting $f^{\boldsymbol{\varepsilon}} = f_a^{\boldsymbol{\varepsilon}'}$ we can verify that $f^{\boldsymbol{\varepsilon}}(x_t(\boldsymbol{\varepsilon})) = s_t(\boldsymbol{\varepsilon})[\varepsilon_t]$ for any $t \in [d+1]$, and if $\varepsilon_1 = 1$, then letting $f^{\boldsymbol{\varepsilon}} = f_b^{\boldsymbol{\varepsilon}'}$ we can verify that $f^{\boldsymbol{\varepsilon}}(x_t(\boldsymbol{\varepsilon})) = s_t(\boldsymbol{\varepsilon})[\varepsilon_t]$ for any $t \in [d+1]$. Hence, depth-$(k+1)$ tree $\mathbf{x}$ is shattered by $\mathcal{F}$, which leads to contradiction. Therefore, we have

$$\mathfrak{c}(a, b) < \alpha, \qquad \forall a, b \in \mathcal{K}. \tag{24}$$

Next, for any $k \in [M]$ with $d^{\mathsf{seq}}(\mathcal{F}_k, \alpha) \le d-1$, according to the induction hypothesis $\mathfrak{G}(n-1, d-1)$, there exists a sequential cover $\mathcal{V}_k^l$ of size $g_M(n-1, d-1)$ for $\mathcal{F}$ on the depth-$(n-1)$ $\mathcal{X}$-valued tree $\mathbf{x}^l$, and also a sequential cover $\mathcal{V}_k^r$ of size $g_M(n-1, d-1)$ for $\mathcal{F}$ on the depth-$(n-1)$ $\mathcal{X}$-valued tree $\mathbf{x}^r$. We then combine the elements in $\mathcal{V}_k^l$ and $\mathcal{V}_k^r$ into a set $\mathcal{V}_k$ of depth-$n$ $[M]$-valued trees by a joining process as follows. We let $v_1 = k \in [M]$, and according to the construction of $\mathcal{F}_k$ we have for any $f \in \mathcal{F}$ that $f(x_1) = v_1$, and thus $\mathfrak{c}(f(x_1), v_1) \le \alpha$. For $\mathbf{v}^l \in \mathcal{V}_k^l$ and $\mathbf{v}^r \in \mathcal{V}_k^r$, we define depth-$n$ $[M]$-valued tree $\mathbf{v}[\mathbf{v}^l, \mathbf{v}^r]$ as: for

any path $\boldsymbol{\varepsilon} \in \{-1,1\}^n$, we let $\boldsymbol{\varepsilon}' = (\varepsilon_{2:n}) \in \{-1,1\}^{n-1}$, and let $v_1[\mathbf{v}^l, \mathbf{v}^r](\boldsymbol{\varepsilon}) = v_1$. If $\varepsilon_1 = -1$, then let $v_t[\mathbf{v}^l, \mathbf{v}^r](\boldsymbol{\varepsilon}) = v^l_{t-1}(\boldsymbol{\varepsilon}')$, and if $\varepsilon_1 = 1$, then let $v_t[\mathbf{v}^l, \mathbf{v}^r](\boldsymbol{\varepsilon}) = v^r_{t-1}(\boldsymbol{\varepsilon}')$. We construct $\mathcal{V}_k = \{\mathbf{v}[\mathbf{v}^l, \mathbf{v}^r]\}$ with $|\mathcal{V}_k| \le \max\{|\mathcal{V}^l_k|, |\mathcal{V}^r_k|\}$ to make sure that every element in $\mathcal{V}^l_k$ and $\mathcal{V}^r_k$ appears at least once in the construction of $\mathcal{V}_k$. Next, we will argue that $\mathcal{V}_k$ is an $\alpha$-sequential cover of $\mathcal{F}_k$ on $\mathbf{x}$. This is easy to see by construction: for any $f \in \mathcal{F}_k$ and $\boldsymbol{\varepsilon} \in \{-1,1\}^n$, if $\varepsilon_1 = -1$, then since $\mathcal{V}^l_k$ is a $\alpha$-sequential cover of $\mathcal{F}_k$, there exists $\mathbf{v}^l \in \mathcal{V}^l_k$ such that for any $2 \le t \le n$, $\mathfrak{c}(f(x_t(\boldsymbol{\varepsilon})), v^l_t(\boldsymbol{\varepsilon})) \le \alpha$. Suppose $\mathbf{v} = \mathbf{v}[\mathbf{v}^l, \mathbf{v}^r] \in \mathcal{V}_k$ for some $\mathbf{v}^r \in \mathcal{V}^r_k$, and we also have $\mathfrak{c}(f(x_1(\boldsymbol{\varepsilon})), v_1(\boldsymbol{\varepsilon})) \le \alpha$ according to the construction of $\mathcal{F}_k$. Hence, for any $t \in [n]$, we always $\mathfrak{c}(f(x_t(\boldsymbol{\varepsilon})), v_t(\boldsymbol{\varepsilon})) \le \alpha$. Therefore, $\mathcal{V}_k$ is a cover of $\mathcal{F}_k$ on $\mathbf{x}$. Further by induction hypothesis we have $\max\{|\mathcal{V}^l_k|, |\mathcal{V}^r_k|\} \le g_M(n-1, d-1)$. Hence $|\mathcal{V}_k| \le g_M(n-1, d-1)$.

If $\mathcal{K} = \emptyset$, then by letting $\mathcal{V} = \cup_{k\in[M]}\mathcal{V}_k$, $\mathcal{V}$ is an $\alpha$-sequential cover of $\mathcal{F}$ on $\mathbf{x}$, and also

$$|\mathcal{V}| \le M \cdot g_M(n-1, d-1) \le g_M(n-1, d) + (M-1) g_M(n-1, d-1) = g_M(n, d),$$

where the inequality follows from the fact that $g_M(n-1, d-1) \le g_M(n-1, d)$ for any $n, d$, and the last equation follows from Eq. (23).

Next, we consider cases where $|\mathcal{K}| \ge 1$. Consider the function class $\mathcal{F}' = \cup_{k\in\mathcal{K}}\mathcal{F}_k \subseteq \mathcal{F}$. We have $d^{\mathsf{seq}}(\mathcal{F}', \alpha) \le d^{\mathsf{seq}}(\mathcal{F}, \alpha) = d$. According to the induction hypothesis $\mathfrak{G}(n-1, d)$, there exists a sequential cover $\mathcal{V}^l$ of size $g_M(n-1, d)$ for the depth-$(n-1)$ $\mathcal{X}$-valued tree $\mathbf{x}^l$, and also a sequential cover $\mathcal{V}^l$ of size $g_M(n-1, d)$ for the depth-$(n-1)$ $\mathcal{X}$-valued tree $\mathbf{x}^l$. As before, we combine the elements in $\mathcal{V}^l$ and $\mathcal{V}^r$ into a set $\mathcal{V}'$ of depth-$n$ $[M]$-valued trees. We let $v_1 = f(x_1) \in [M]$ for some $f \in \mathcal{F}'$, chosen arbitrarily. Then, according to the construction of $\mathcal{F}'$, we have for any other $g \in \mathcal{F}'$, $\mathfrak{c}(g(x_1), v_1) \le \alpha$. For $\mathbf{v}^l \in \mathcal{V}^l$ and $\mathbf{v}^r \in \mathcal{V}^r$, we define depth-$n$ $[M]$-valued tree $\mathbf{v}[\mathbf{v}^l, \mathbf{v}^r]$ by joining them with $v_1$ at the root as before: for any path $\boldsymbol{\varepsilon} \in \{-1,1\}^n$, we let $\boldsymbol{\varepsilon}' = (\varepsilon_{2:n}) \in \{-1,1\}^{n-1}$, and let $v_1[\mathbf{v}^l, \mathbf{v}^r](\boldsymbol{\varepsilon}) = v_1$. If $\varepsilon_1 = -1$, then let $v_t[\mathbf{v}^l, \mathbf{v}^r](\boldsymbol{\varepsilon}) = v^l_{t-1}(\boldsymbol{\varepsilon}')$, and if $\varepsilon_1 = 1$, then let $v_t[\mathbf{v}^l, \mathbf{v}^r](\boldsymbol{\varepsilon}) = v^r_{t-1}(\boldsymbol{\varepsilon}')$. We construct $\mathcal{V}' = \{\mathbf{v}[\mathbf{v}^l, \mathbf{v}^r]\}$ with $|\mathcal{V}'| \le \max\{|\mathcal{V}^l|, |\mathcal{V}^r|\}$ to make sure that every element in $\mathcal{V}^l$ and $\mathcal{V}^r$ appears at least once in the construction of $\mathcal{V}'$. Next, we will argue that $\mathcal{V}'$ is a $\alpha$-sequential cover of $\mathcal{F}'$ on $\mathbf{x}$. For any $f \in \mathcal{F}'$ and $\boldsymbol{\varepsilon} \in \{-1,1\}^n$, if $\varepsilon_1 = -1$, then since $\mathcal{V}^l$ is a $\alpha$-sequential cover of $\mathcal{F}'$ on the tree $\mathbf{x}^l$, there exists $\mathbf{v}^l \in \mathcal{V}^l$ such that for any $2 \le t \le n$, $\mathfrak{c}(f(x_t(\boldsymbol{\varepsilon})), v^l_t(\boldsymbol{\varepsilon})) \le \alpha$. Suppose $\mathbf{v} = \mathbf{v}[\mathbf{v}^l, \mathbf{v}^r] \in \mathcal{V}'$ for some $\mathbf{v}^r \in \mathcal{V}^r$ (according to the construction of $\mathcal{V}'$ such $\mathbf{v}$ exists). Then since $f \in \mathcal{F}'$, according to Eq. (24) we have $\mathfrak{c}(f(x_1(\boldsymbol{\varepsilon})), v_1(\boldsymbol{\varepsilon})) \le \alpha$. Hence for any $t \in [n]$, we always $\mathfrak{c}(f(x_t(\boldsymbol{\varepsilon})), v_t(\boldsymbol{\varepsilon})) \le \alpha$. Similarly, if $\varepsilon_1 = 1$, there also exists some $\mathbf{v} \in \mathcal{V}'$ such that $\mathfrak{c}(f(x_t(\boldsymbol{\varepsilon})), v_t(\boldsymbol{\varepsilon})) \le \alpha$ for any $t \in [n]$. Therefore, $\mathcal{V}'$ is an $\alpha$-sequential cover of $\mathcal{F}'$. Further by induction hypothesis we have $\max\{|\mathcal{V}^l|, |\mathcal{V}^r|\} \le g_M(n-1, d)$. Hence $|\mathcal{V}'| \le g_M(n-1, d)$. We now let $\mathcal{V} = \mathcal{V}' \cup (\cup_{k\notin\mathcal{K}}\mathcal{V}_k)$, and after noticing that $|[M]\backslash\mathcal{K}| \le (M-1)$, we obtain

$$|\mathcal{V}| \le (M-1) \cdot g_M(n-1, d-1) + g_M(n-1, d) = g_M(n, d),$$

where the last equation follows from Eq. (23). This finishes the proof of the induction statement $\mathfrak{G}(n, d)$.

We conclude that, by induction, we have for any depth-$n$ $\mathcal{X}$-valued tree $\mathbf{x}$,

$$\mathcal{N}^{\mathsf{seq}}_\infty(\mathcal{F}, \mathbf{x}, \alpha) \le g_M(n, d^{\mathsf{seq}}(\mathcal{F}, \alpha)) = \sum_{i=0}^{d^{\mathsf{seq}}(\mathcal{F},\alpha)} \binom{n}{i} \cdot (M-1)^i \overset{(i)}{\le} \left(\frac{enM}{d^{\mathsf{seq}}(\mathcal{F}, \alpha)}\right)^{d^{\mathsf{seq}}(\mathcal{F},\alpha)} \le (enM)^{d^{\mathsf{seq}}(\mathcal{F},\alpha)},$$

where $(i)$ follows e.g. from [RST15, Theorem 7]. □

*Proof of Proposition 5.* We define distance $\mathfrak{c}' : [M] \times [M] \to \mathbb{R}_+ \cup \{0\}$:

$$\mathfrak{c}'(a, b) = \mathfrak{c}(u_a, u_b). \tag{25}$$

For any $f \in \mathcal{F}$, since $f$ maps $\mathcal{X}$ into $[-1, 1]$, there exists $\bar{f} : \mathcal{X} \to [M]$ such that for any $x \in \mathcal{X}$, $\mathfrak{c}(f(x), u_{\bar{f}(x)}) \le \beta$. We define function class $\bar{\mathcal{F}} = \{\bar{f} : f \in \mathcal{F}\} \subseteq \{\mathcal{X} \to [M]\}$. We use $d^{\mathsf{seq}}_{\mathfrak{c}'}(\bar{\mathcal{F}}, \alpha)$ to denote the sequential gapped dimension of integer-valued function class $\bar{\mathcal{F}}$ at scale $\alpha$ under distance $\mathfrak{c}'$, where the dimension is defined in Definition 6, and for simplicity we let $d = d^{\mathsf{seq}}_{\mathfrak{c}'}(\bar{\mathcal{F}}, \alpha)$. Suppose $\mathbf{x}$ is a depth-$d$ tree shattered

by $\mathcal{F}$. Then there exists a depth-$d$ $([M]\times[M])$-valued tree $\bar{\mathbf{s}}$ such that for any $\boldsymbol{\varepsilon}\in\{-1,1\}^d$ and $t\in[d]$, $\mathfrak{c}'(\bar{s}_t(\boldsymbol{\varepsilon})[-1],\bar{s}_t(\boldsymbol{\varepsilon})[1])\geq\alpha$, and also for any $\boldsymbol{\varepsilon}\in\{-1,1\}^d$, there exists $\bar{f}^{\boldsymbol{\varepsilon}}\in\bar{\mathcal{F}}$ such that $\bar{f}^{\boldsymbol{\varepsilon}}(x_t(\boldsymbol{\varepsilon}))=\bar{s}_t(\boldsymbol{\varepsilon})[\varepsilon_t]$ for any $t\in[d]$. Notice from the definition of $\mathcal{F}$ there exists some $f^{\boldsymbol{\varepsilon}}\in\mathcal{F}$ such that $\bar{f}^{\boldsymbol{\varepsilon}}=\overline{(f^{\boldsymbol{\varepsilon}})}$.

We next verify that tree $\mathbf{x}$ is also shattered by $\mathcal{F}$ at scale $(\alpha,\beta)$, according to the definition Definition 8. We construct depth-$d$ $([-1,1]\times[-1,1])$-tree $\mathbf{s}$ according to $\bar{\mathbf{s}}$ as follows:

$$s_t(\boldsymbol{\varepsilon})[-1]=u_{\bar{s}_t(\boldsymbol{\varepsilon})[-1]}\quad\text{and}\quad s_t(\boldsymbol{\varepsilon})[1]=u_{\bar{s}_t(\boldsymbol{\varepsilon})[1]}.$$

Then according to the definition of $\mathfrak{c}'$ in Eq. (25), we have for any $\boldsymbol{\varepsilon}\in\{-1,1\}^d$ and $t\in[n]$,

$$\mathfrak{c}(s_t(\boldsymbol{\varepsilon})[-1],s_t(\boldsymbol{\varepsilon})[1])=\mathfrak{c}'(\bar{s}_t(\boldsymbol{\varepsilon})[-1],\bar{s}_t(\boldsymbol{\varepsilon})[1])\geq\alpha.$$

Additionally, for any $\boldsymbol{\varepsilon}\in\{-1,1\}^d$, there exists some $f^{\boldsymbol{\varepsilon}}\in\mathcal{F}$ such that $\overline{(f^{\boldsymbol{\varepsilon}})}(x_t(\boldsymbol{\varepsilon}))=\bar{s}_t(\boldsymbol{\varepsilon})[\varepsilon_t]$, which implies,

$$\mathfrak{c}(f(x_t(\boldsymbol{\varepsilon})),s_t(\boldsymbol{\varepsilon})[\varepsilon_t])=\mathfrak{c}(f(x_t(\boldsymbol{\varepsilon})),u_{\bar{s}_t(\boldsymbol{\varepsilon})[\varepsilon_t]})=\mathfrak{c}(f(x_t(\boldsymbol{\varepsilon})),u_{\overline{(f^{\boldsymbol{\varepsilon}})}(x_t(\boldsymbol{\varepsilon}))})\leq\beta,\quad\forall t\in[d],$$

where the last inequality follows from the definiton of $\bar{f}$. Therefore, $\mathbf{x}$ is shattered by $\mathcal{F}$ at scale $(\alpha,\beta)$, and according to Definition 8 we have $d^{\mathsf{seq}}(\mathcal{F},\alpha,\beta)\geq d$.

Next we will upper bound the sequential covering number of $\mathcal{F}$ in terms of $d$. For a fixed depth-$n$ $\mathcal{X}$-valued tree $\mathbf{x}$, according to Lemma 3, there exists a sequential covering $\bar{\mathcal{V}}$ of $\bar{\mathcal{F}}$ with size no more than $(neM)^d$. Hence for any $f\in\mathcal{F}$ and $\boldsymbol{\varepsilon}\in\{-1,1\}^d$, since $\bar{f}\in\bar{\mathcal{F}}$, there exists some $\bar{\mathbf{v}}\in\bar{\mathcal{V}}$ which satisfies

$$\mathfrak{c}'(\bar{f}(x_t(\boldsymbol{\varepsilon})),\bar{v}_t(\boldsymbol{\varepsilon}))\leq\alpha\qquad\forall t\in[n],$$

which implies that

$$\mathfrak{c}(f(x_t(\boldsymbol{\varepsilon})),u_{\bar{v}_t(\boldsymbol{\varepsilon})})\leq\mathfrak{c}(f(x_t(\boldsymbol{\varepsilon})),u_{\bar{f}(x_t(\boldsymbol{\varepsilon}))})+\mathfrak{c}(u_{\bar{f}(x_t(\boldsymbol{\varepsilon}))},u_{\bar{v}_t(\boldsymbol{\varepsilon})})\leq\beta+\alpha\qquad\forall t\in[n],$$

where the first inequality uses the triangle inequality of $\mathfrak{c}$, and the second inequality uses the definition of function $\bar{f}$. For every $\bar{\mathbf{v}}\in\bar{\mathcal{V}}$, we construct depth-$d$ $\mathbb{R}$-valued tree $\mathbf{v}_{\bar{\mathbf{v}}}$ where for any $\boldsymbol{\varepsilon}\in\{-1,1\}^d$ and $t\in[n]$, $(v_{\bar{\mathbf{v}}})_t(\boldsymbol{\varepsilon})=u_{\bar{v}_t(\boldsymbol{\varepsilon})}$, for every $\bar{\mathbf{v}}\in\bar{\mathcal{V}}$. And we further let $\mathcal{V}=\{\mathbf{v}_{\bar{\mathbf{v}}}:\bar{\mathbf{v}}\in\bar{\mathcal{V}}\}$. Then $\mathcal{V}$ is an $(\alpha+\beta)$-cover of $\mathcal{F}$ on tree $\mathbf{x}$, which implies

$$\mathcal{N}_\infty^{\mathsf{seq}}(\mathcal{F},\mathbf{x},\alpha+\beta)\leq|\mathcal{V}|=|\bar{\mathcal{V}}|\leq(neM)^d.$$

□

## B.2 Missing Proofs in Section 3.3

*Proof of Proposition 6.* We only need to prove that if a depth-$d$ $\mathcal{X}$-valued tree $\mathbf{x}$ is shattered by $\mathcal{F}$ at scale $(\alpha,\beta)$ according to Definition 8, then $\mathbf{x}$ is shattered by $\mathcal{F}$ at scale $\alpha-2\beta$ according to Definition 10.

If $\mathbf{x}$ is shattered by $\mathcal{F}$ at scale $(\alpha,\beta)$ according to Definition 8, then there exists a depth-$d$ $([0,1]\times[0,1])$-valued tree $\mathbf{s}$ such that for any $\boldsymbol{\varepsilon}\in\{-1,1\}^n$, there exists a function $f^{\boldsymbol{\varepsilon}}\in\mathcal{F}$ such that for any $t\in[d]$, we have

$$|f^{\boldsymbol{\varepsilon}}(x_t(\boldsymbol{\varepsilon}))-s_t(\boldsymbol{\varepsilon})[\varepsilon_t]|\leq\beta\quad\text{and}\quad|s_t(\boldsymbol{\varepsilon})[1]-s_t(\boldsymbol{\varepsilon})[-1]|\geq\alpha.\tag{26}$$

Without loss of generality we assume $s_t(\boldsymbol{\varepsilon})[1]>s_t(\boldsymbol{\varepsilon})[-1]$ for any $\boldsymbol{\varepsilon}$ and $t\in[d]$. We define depth-$d$ $[0,1]$-valued tree $\mathbf{v}$ as follows: for any $\boldsymbol{\varepsilon}\in\{-1,1\}^d$ and $t\in[d]$, let

$$v_t(\boldsymbol{\varepsilon})=\frac{s_t(\boldsymbol{\varepsilon})[-1]+s_t(\boldsymbol{\varepsilon})[1]}{2}.$$

Since $\alpha>2\beta$, according to Eq. (26) we have for any $\boldsymbol{\varepsilon}\in\{-1,1\}^n$,

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

Therefore, $\mathbf{x}$ is shattered by $\mathcal{F}$ with $\mathbf{v}$ being its witness at scale $\alpha-2\beta$, according to Definition 10. □

*Proof of Proposition 7.* We first show that if depth-$d$ $\mathcal{X}$-valued tree $\mathbf{x}$ is shattered by function class $\mathcal{F}$ at scale $\alpha$, according to Definition 10, then we have

$$\mathcal{N}_\infty^{\mathsf{seq}}(\mathcal{F}, \mathbf{x}, \alpha/3) \geq 2^d. \tag{27}$$

We use depth-$d$ $[-1,1]$-valued tree $\mathbf{s}$ to denote the witness of shattering of $\mathbf{x}$ via $\mathcal{F}$ at scale $\alpha$. According to Definition 10, for any $\boldsymbol{\varepsilon} \in \{-1,1\}^d$, there exists some function $f^{\boldsymbol{\varepsilon}} \in \mathcal{F}$ such that

$$\varepsilon_t \cdot (f^{\boldsymbol{\varepsilon}}(x_t(\boldsymbol{\varepsilon})) - s_t(\boldsymbol{\varepsilon})) \geq \frac{\alpha}{2} \tag{28}$$

Next we will prove Eq. (27) via contradiction. Suppose there exists an $\ell_\infty$-sequential covering $\mathcal{V}$ at scale $\alpha/2$ of size less than $2^d$. Then for any $\boldsymbol{\varepsilon} \in \{-1,1\}^d$ and function $f \in \mathcal{F}$, there exists some tree $\mathbf{v}[f, \boldsymbol{\varepsilon}] \in \mathcal{V}$ such that

$$|f(x_t(\boldsymbol{\varepsilon})) - v_t(\boldsymbol{\varepsilon})| \leq \frac{\alpha}{3}, \qquad \forall t \in [d]. \tag{29}$$

Since $|\mathcal{V}| \leq 2^d - 1$, according to the pigeonhole principle there exists different $\boldsymbol{\varepsilon} = (\varepsilon_{1:d}), \boldsymbol{\varepsilon}' = (\varepsilon'_{1:d}) \in \{-1,1\}^d$ such that $\mathbf{v}[f^{\boldsymbol{\varepsilon}}, \boldsymbol{\varepsilon}] = \mathbf{v}[f^{\boldsymbol{\varepsilon}'}, \boldsymbol{\varepsilon}']$. We let

$$\mathbf{v}[f^{\boldsymbol{\varepsilon}}, \boldsymbol{\varepsilon}] = \mathbf{v}[f^{\boldsymbol{\varepsilon}'}, \boldsymbol{\varepsilon}'] = \mathbf{v}$$

We then choose $r$ to be smallest nonnegative integer such that $\varepsilon_r \neq \varepsilon'_r$. Since $\boldsymbol{\varepsilon} \neq \boldsymbol{\varepsilon}'$ we have $1 \leq r \leq d$. Then we have $\varepsilon_{1:r-1} = \varepsilon'_{1:r-1}$, which implies that

$$x_r(\boldsymbol{\varepsilon}) = x_r(\boldsymbol{\varepsilon}'), \quad v_r(\boldsymbol{\varepsilon}) = v_r(\boldsymbol{\varepsilon}') \quad \text{and} \quad s_r(\boldsymbol{\varepsilon}) = s_r(\boldsymbol{\varepsilon}'). \tag{30}$$

Therefore, we obtain that

$$\begin{aligned} |f^{\boldsymbol{\varepsilon}}(x_r(\boldsymbol{\varepsilon})) - f^{\boldsymbol{\varepsilon}'}(x_r(\boldsymbol{\varepsilon}))| &= |f^{\boldsymbol{\varepsilon}}(x_r(\boldsymbol{\varepsilon})) - f^{\boldsymbol{\varepsilon}'}(x_r(\boldsymbol{\varepsilon}'))| \\ &\leq |f^{\boldsymbol{\varepsilon}}(x_r(\boldsymbol{\varepsilon})) - v_r(\boldsymbol{\varepsilon})| + |v_r(\boldsymbol{\varepsilon}) - v_r(\boldsymbol{\varepsilon}')| + |v_r(\boldsymbol{\varepsilon}') - f^{\boldsymbol{\varepsilon}'}(x_r(\boldsymbol{\varepsilon}'))| \leq \frac{\alpha}{3} + \frac{\alpha}{3} < \alpha, \end{aligned} \tag{31}$$

where the second inequality uses Eq. (29) and the fact that $\mathbf{v}[f^{\boldsymbol{\varepsilon}}, \boldsymbol{\varepsilon}] = \mathbf{v}[f^{\boldsymbol{\varepsilon}'}, \boldsymbol{\varepsilon}'] = \mathbf{v}$. Next according to Eq. (28), we have

$$\varepsilon_r \cdot (f^{\boldsymbol{\varepsilon}}(x_r(\boldsymbol{\varepsilon})) - s_r(\boldsymbol{\varepsilon})) \geq \frac{\alpha}{2} \quad \text{and} \quad \varepsilon'_r \cdot \left(f^{\boldsymbol{\varepsilon}'}(x_r(\boldsymbol{\varepsilon}')) - s_r(\boldsymbol{\varepsilon}')\right) \geq \frac{\alpha}{2}.$$

According to the definition of $r$ we have $\varepsilon_r \neq \varepsilon'_r$. Again using Eq. (30), we obtain that

$$|f^{\boldsymbol{\varepsilon}}(x_t(\boldsymbol{\varepsilon})) - f^{\boldsymbol{\varepsilon}'}(x_t(\boldsymbol{\varepsilon}))| \geq \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha,$$

which contradicts Eq. (31). Therefore, we proved Eq. (27).

Next, according to Proposition 5 with $\mathfrak{c}(a,b) = |a-b|$, and tree $\mathbf{x}$ being the depth-$\mathrm{sfat}(\mathcal{F}, 3(\alpha+\beta))$ $\mathcal{X}$-valued tree shattered by $\mathcal{F}$, we obtain that

$$\mathcal{N}_\infty^{\mathsf{seq}}(\mathcal{F}, \mathbf{x}, \alpha+\beta) \leq \left(\frac{2e \cdot \mathrm{sfat}(\mathcal{F}, 3(\alpha+\beta))}{\beta}\right)^{d^{\mathsf{seq}}(\mathcal{F}, \alpha, \beta)}.$$

According to Eq. (27) we further have

$$\mathcal{N}_\infty^{\mathsf{seq}}(\mathcal{F}, \mathbf{x}, \alpha+\beta) \geq 2^{\mathrm{sfat}(\mathcal{F}, 3\alpha+3\beta)}.$$

Therefore, we conclude that

$$\log 2 \cdot \mathrm{sfat}(\mathcal{F}, 3(\alpha+\beta)) \leq d^{\mathsf{seq}}(\mathcal{F}, \alpha, \beta) \cdot \log\left(\frac{2e \cdot \mathrm{sfat}(\mathcal{F}, 3(\alpha+\beta))}{\beta}\right),$$

which implies that

$$\mathrm{sfat}(\mathcal{F}, 3(\alpha+\beta)) \le 2d^{\mathsf{seq}}(\mathcal{F}, \alpha, \beta) \cdot \log\left(\frac{6\mathrm{sfat}(\mathcal{F}, 3(\alpha+\beta))}{\beta}\right). \tag{32}$$

Additionally, since $\log x \le \sqrt{x}$ holds for any $x > 0$,

$$\mathrm{sfat}(\mathcal{F}, 3(\alpha+\beta)) \le 2d^{\mathsf{seq}}(\mathcal{F}, \alpha, \beta) \cdot \sqrt{\frac{6\mathrm{sfat}(\mathcal{F}, 3(\alpha+\beta))}{\beta}},$$

which implies

$$\mathrm{sfat}(\mathcal{F}, 3(\alpha+\beta)) \le \frac{24d^{\mathsf{seq}}(\mathcal{F}, \alpha, \beta)^2}{\beta}$$

Bringing this back to Eq. (32), we obtain that

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

□

*Proof of Proposition 8.* We first show that for any $0 < 2\beta < \alpha < 1$, $d^{\mathsf{seq}}(\mathcal{F}, \alpha, \beta) = 1$. It is easy to see that the depth-1 $\mathcal{X}$-valued tree $\mathbf{x}$ with $x_1 = x$ is shattered by $\mathcal{F}$ at scale $(\alpha, \beta)$, according to Definition 8, hence $d^{\mathsf{seq}}(\mathcal{F}, \alpha, \beta) \ge 1$. Next we show that $d^{\mathsf{seq}}(\mathcal{F}, \alpha, \beta) \le 1$. Suppose there is a depth-2 $\mathcal{X}$-valued tree $\mathbf{x}$ shattered by $\mathcal{F}$, then all nodes equal to $x$ whatever depth and path. We let $\mathbf{s}$ to be the depth-2 $([-1,1]\times[-1,1])$-tree which is the witness of the shattering. Then for any $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2) \in \{-1,1\}^2$, there exists functions $f^{\boldsymbol{\varepsilon}} \in \mathcal{F}$ such that $|f^{\boldsymbol{\varepsilon}}(x) - s_1(\boldsymbol{\varepsilon})[\varepsilon_1]| \le \beta$ and $|f^{\boldsymbol{\varepsilon}}(x) - s_2(\boldsymbol{\varepsilon})[\varepsilon_2]| \le \beta$. Therefore, we have $|s_1(\boldsymbol{\varepsilon})[\varepsilon_1] - s_2(\boldsymbol{\varepsilon})[\varepsilon_2]| \le 2\beta$ for any $\boldsymbol{\varepsilon} \in \{-1,1\}^2$. We choose $\boldsymbol{\varepsilon} = (1,1)$ and $\boldsymbol{\varepsilon}' = (1,-1)$, then we have

$$s_1(\boldsymbol{\varepsilon})[\varepsilon_1] = s_1(\boldsymbol{\varepsilon}')[\varepsilon_1'] \quad \text{and} \quad |s_2(\boldsymbol{\varepsilon})[\varepsilon_2] - s_2(\boldsymbol{\varepsilon}')[\varepsilon_2']| \ge \alpha.$$

When $\alpha > 2\beta$, the above inequality cannot hold. Hence we have $d^{\mathsf{seq}}(\mathcal{F}, \alpha, \beta) = 1$.

Next, we show that $\mathrm{sfat}(\mathcal{F}, \alpha) \ge \log(1/\alpha)$. We let $d = \lfloor \log_2(1/\alpha) \rfloor$, and for depth-$d$ path $\boldsymbol{\varepsilon} \in \{-1,1\}^d$, we define the shattered tree $\mathbf{x}$ such that $x_t(\boldsymbol{\varepsilon}) = x$ for any $\boldsymbol{\varepsilon} \in \{-1,1\}^d$ and $t \in [d]$, and the witness tree $\mathbf{v}$ as:

$$v_t(\boldsymbol{\varepsilon}) = \sum_{i=1}^{t-1} \varepsilon_i \cdot 2^{d-i} \cdot \alpha,$$

and we choose function $f^{\boldsymbol{\varepsilon}}$ as

$$f^{\boldsymbol{\varepsilon}}(x) = \sum_{i=1}^{d} \varepsilon_i \cdot 2^{d-i} \cdot \alpha.$$

Since $d = \lfloor \log_2(1/\alpha) \rfloor$, $\mathbf{v}$ is a depth-$d$ $[-1,1]$-valued tree, and also $f \in \mathcal{F}$. For any $\boldsymbol{\varepsilon} \in \{-1,1\}^d$, we have

$$\varepsilon_t \cdot (f^{\boldsymbol{\varepsilon}}(x_t(\boldsymbol{\varepsilon})) - v_t(\boldsymbol{\varepsilon})) = \varepsilon_t \cdot \left(\sum_{i=t}^{d} \varepsilon_i \cdot 2^{d-i}\right) \cdot \alpha = \left(2^{d-t} - \sum_{i=t+1}^{d} \varepsilon_i\varepsilon_t \cdot 2^{d-i}\right) \cdot \alpha \ge \alpha > \frac{\alpha}{2}.$$

Hence tree $\mathbf{x}$ is shattered by $\mathcal{F}$, which implies that $\mathrm{sfat}(\mathcal{F}, \alpha) \ge \lfloor \log_2(1/\alpha) \rfloor$. □

## B.3 Missing Proofs in Section 3.4

*Proof of Theorem 2.* For fixed positive integer $n$, we let

$$\alpha_n = \frac{1}{2} \cdot \arg\min_{\alpha > 0} \left\{ d^{\mathsf{seq}}(\mathcal{F}, \alpha, \alpha/20) \cdot \frac{16}{\alpha^2} \le n \right\}.$$

Since $d^{\mathsf{seq}}(\mathcal{F}, \alpha, \alpha/20) = \tilde{\Omega}\left(\alpha^{-p}\right)$ for every $\alpha \geq 0$, we have

$$d^{\mathsf{seq}}(\mathcal{F}, \alpha_n, \alpha_n/20) \wedge \frac{n\alpha_n^2}{4} = \tilde{\Omega}\left(n^{\frac{p}{p+2}}\right), \qquad \forall n \in \mathbb{Z}_+.$$

In the following, we will prove that for any positive integer $n$, we have

$$\sup_{\boldsymbol{\mu}, \mathbf{x}} \mathbb{E}[\mathcal{R}_n(\mathcal{F}, \boldsymbol{\mu}, \mathbf{x}, \boldsymbol{\varepsilon})] = \Omega\left(d^{\mathsf{seq}}(\mathcal{F}, \alpha_n, \alpha_n/20)\right).$$

We fix $n$, and let $\alpha = \alpha_n$, $d = d^{\mathsf{seq}}(\mathcal{F}, \alpha_n, \alpha_n/20) \wedge (n\alpha_n^2/4)$. We let $\tilde{\mathbf{x}}$ be the depth-$d$ $\mathcal{X}$-valued tree shattered by $\mathcal{F}$. Then according to Definition 8, there exists a depth-$d$ $[-1,1] \times [-1,1]$-valued tree $\mathbf{s}$ such that for any path $\tilde{\boldsymbol{\varepsilon}} = (\tilde{\varepsilon}_{1:d}) \in \{-1,1\}^d$, there exists $f^{\tilde{\boldsymbol{\varepsilon}}} \in \mathcal{F}$ such that

$$\left|f^{\tilde{\boldsymbol{\varepsilon}}}(\tilde{x}_t(\tilde{\boldsymbol{\varepsilon}})) - s_t(\tilde{\boldsymbol{\varepsilon}})[\tilde{\varepsilon}_t]\right| \leq \frac{\alpha}{20} \quad \text{and} \quad |s_t(\tilde{\boldsymbol{\varepsilon}})[1] - s_t(\tilde{\boldsymbol{\varepsilon}})[-1]| \geq \alpha \qquad \forall t \in [d], \tag{33}$$

Without loss of generality, we assume $s_t(\tilde{\boldsymbol{\varepsilon}})[1] \geq s_t(\tilde{\boldsymbol{\varepsilon}})[-1]$. We define depth-$d$ $[-1,1]$-valued $\tilde{\boldsymbol{\mu}}$ as

$$\tilde{\mu}_t(\tilde{\boldsymbol{\varepsilon}}) = \frac{s_t(\tilde{\boldsymbol{\varepsilon}})[-1] + s_t(\tilde{\boldsymbol{\varepsilon}})[1]}{2}, \quad \forall \boldsymbol{\varepsilon} \in \{-1,1\}^d.$$

In the following, we will construct trees $\mathbf{x}$ and $\boldsymbol{\mu}$ such that $\mathbb{E}[\mathcal{R}_n(\mathcal{F}, \boldsymbol{\mu}, \mathbf{x}, \boldsymbol{\varepsilon})] \geq d/50$. For a fixed path $\boldsymbol{\varepsilon} \in \{-1,1\}^n$, we first define an auxiliary path $\tilde{\boldsymbol{\varepsilon}} = (\tilde{\varepsilon}_{1:d}) \in \{-1,1\}^d$ of length $d$, as well as $d$ integers $k_1, k_2, \ldots, k_d$ in the following way: calculate $\tilde{\varepsilon}_{1:d}$ and $k_{1:d}$ iteratively as

$$k_t = \left\lfloor \frac{1}{(s_t(\tilde{\boldsymbol{\varepsilon}})[1] - s_t(\tilde{\boldsymbol{\varepsilon}})[-1])^2} \right\rfloor \vee 1, \qquad \forall t \in [d],$$

and

$$\tilde{\varepsilon}_t = 2\mathbb{I}\left\{ \sum_{j=1}^{k_t} \varepsilon_{k_1+\ldots+k_{t-1}+j} \geq 0 \right\} - 1, \qquad \forall t \in [d].$$

Notice that according to the above definition, $k_t$ only depends on $\varepsilon_{1:k_1+\ldots+k_{t-1}}$, and $\tilde{\varepsilon}_t$ depends on $\varepsilon_{1:k_1+\ldots+k_t}$. Additionally, since $|s_t(\tilde{\boldsymbol{\varepsilon}})[1] - s_t(\tilde{\boldsymbol{\varepsilon}})[-1]| \geq \alpha$, we have

$$k_t \leq \frac{1}{\alpha^2} \vee 1 \leq \frac{4}{\alpha^2},$$

which implies $k_1 + \ldots + k_d \leq n$ always holds due to the definition of $\alpha = \alpha_n$. Hence $k_{1:d}$ and $\tilde{\varepsilon}_{1:d}$ are all well-defined. We pad the tree $\mathbf{x}$ with an arbitrary value $x_0$, resulting in the following block structure of $(x_1(\boldsymbol{\varepsilon}), x_2(\boldsymbol{\varepsilon}), \ldots, x_n(\boldsymbol{\varepsilon}))$ :

$$(\underbrace{\tilde{x}_1(\tilde{\boldsymbol{\varepsilon}}), \tilde{x}_1(\tilde{\boldsymbol{\varepsilon}}), \ldots, \tilde{x}_1(\tilde{\boldsymbol{\varepsilon}})}_{k_1 \text{ terms}}, \underbrace{\tilde{x}_2(\tilde{\boldsymbol{\varepsilon}}), \tilde{x}_2(\tilde{\boldsymbol{\varepsilon}}), \ldots, \tilde{x}_2(\tilde{\boldsymbol{\varepsilon}})}_{k_2 \text{ terms}}, \ldots, \underbrace{\tilde{x}_d(\tilde{\boldsymbol{\varepsilon}}), \tilde{x}_d(\tilde{\boldsymbol{\varepsilon}}), \ldots, \tilde{x}_d(\tilde{\boldsymbol{\varepsilon}})}_{k_d \text{ terms}}, \underbrace{x_0, x_0, \ldots, x_0}_{(n-k_1-k_2-\ldots-k_d) \text{ terms}}),$$

Similarly, the values $(\mu_1(\boldsymbol{\varepsilon}), \mu_2(\boldsymbol{\varepsilon}), \ldots, \mu_n(\boldsymbol{\varepsilon}))$ are of the form

$$(\underbrace{\tilde{\mu}_1(\tilde{\boldsymbol{\varepsilon}}), \tilde{\mu}_1(\tilde{\boldsymbol{\varepsilon}}), \ldots, \tilde{\mu}_1(\tilde{\boldsymbol{\varepsilon}})}_{k_1 \text{ terms}}, \underbrace{\tilde{\mu}_2(\tilde{\boldsymbol{\varepsilon}}), \tilde{\mu}_2(\tilde{\boldsymbol{\varepsilon}}), \ldots, \tilde{\mu}_2(\tilde{\boldsymbol{\varepsilon}})}_{k_2 \text{ terms}}, \ldots, \underbrace{\tilde{\mu}_d(\tilde{\boldsymbol{\varepsilon}}), \tilde{\mu}_d(\tilde{\boldsymbol{\varepsilon}}), \ldots, \tilde{\mu}_d(\tilde{\boldsymbol{\varepsilon}})}_{k_d \text{ terms}}, \underbrace{f^{\tilde{\boldsymbol{\varepsilon}}}(x_0), f^{\tilde{\boldsymbol{\varepsilon}}}(x_0), \ldots, f^{\tilde{\boldsymbol{\varepsilon}}}(x_0)}_{(n-k_1-k_2-\ldots-k_d) \text{ terms}})$$

We note that the construction of trees $x_t(\boldsymbol{\varepsilon})$ and $\mu_t(\boldsymbol{\varepsilon})$ here differs slightly from the construction used for the non-sequential setting in the proof of Theorem 1. In the sequential case, $\mu$ is structured as a tree and can adapt based on the history of $\varepsilon$, allowing us to choose $\mu$ as $f^{\boldsymbol{\varepsilon}}(x_0)$. In contrast, in the non-sequential case, the choice of $\mu$ must be independent of the history, so we are required to select global (i.e., fixed) values

for $\mu$. For this reason, Theorem 2 only needs $d(\mathcal{F}, \alpha, \alpha/20) = \tilde{\Omega}(\alpha^{-p})$, while Theorem 1, with the current proof, requires that $d(\mathcal{F}, \alpha, \alpha/(20nC)) = \tilde{\Omega}(\alpha^{-p})$.

Next we write $\mathcal{R}_n(\mathcal{F}, \boldsymbol{\mu}, \mathbf{x}, \boldsymbol{\varepsilon})$ in terms of segments 1 to $d$. Noticing that $f^{\tilde{\boldsymbol{\varepsilon}}} \in \mathcal{F}$ for any depth-$d$ path $\tilde{\boldsymbol{\varepsilon}} \in \{-1, 1\}^d$, if we use $[t, j]$ to denote $k_1 + \ldots + k_t + j$, we obtain

$$
\begin{aligned}
&\mathcal{R}_n(\mathcal{F}, \boldsymbol{\mu}, \mathbf{x}, \boldsymbol{\varepsilon}) \\
&\quad = \sup_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{d} \sum_{j=1}^{k_t} C \cdot \varepsilon_{[t-1,j]} (f(x_{[t-1,j]}(\boldsymbol{\varepsilon})) - \mu_{[t-1,j]}(\boldsymbol{\varepsilon})) - (f(x_{[t-1,j]}(\boldsymbol{\varepsilon})) - \mu_{[t-1,j]}(\boldsymbol{\varepsilon}))^2 \right. \\
&\qquad \left. + \sum_{j=1}^{n-k_1-\ldots-k_d} C \cdot \varepsilon_{[d,j]} (f(x_{[t-1,j]}(\boldsymbol{\varepsilon})) - \mu_{[d,j]}(\boldsymbol{\varepsilon})) - (f(x_{[d,j]}(\boldsymbol{\varepsilon})) - \mu_{[d,j]}(\boldsymbol{\varepsilon}))^2 \right\} \\
&\quad \geq \sum_{t=1}^{d} \sum_{j=1}^{k_t} C \cdot \varepsilon_{[t-1,j]} (f^{\tilde{\boldsymbol{\varepsilon}}}(\tilde{x}_t(\boldsymbol{\varepsilon})) - \tilde{\mu}_t(\tilde{\boldsymbol{\varepsilon}})) - (f^{\tilde{\boldsymbol{\varepsilon}}}(\tilde{x}_t(\boldsymbol{\varepsilon})) - \tilde{\mu}_t(\tilde{\boldsymbol{\varepsilon}}))^2,
\end{aligned}
$$

where in the last step we chose $f = f^{\tilde{\boldsymbol{\varepsilon}}} \in \mathcal{F}$. Next, for fixed $\boldsymbol{\varepsilon}$, we define the random variable

$$
\widehat{\varepsilon}_t = \frac{1}{k_t} \sum_{j=1}^{k_t} \varepsilon_{[t-1,j]}.
$$

Since $\tilde{x}_t(\tilde{\boldsymbol{\varepsilon}})$ and $\mu_t(\tilde{\boldsymbol{\varepsilon}})$ are independent of $\varepsilon_{[t-1,1]}, \ldots, \varepsilon_{[t-1,k_t]}$, we can write

$$
\mathcal{R}_n(\mathcal{F}, \boldsymbol{\mu}, \mathbf{x}, \boldsymbol{\varepsilon}) \geq \sum_{t=1}^{d} k_t \cdot \left( C \cdot \widehat{\varepsilon}_t (f^{\tilde{\boldsymbol{\varepsilon}}}(\tilde{x}_t(\boldsymbol{\varepsilon})) - \tilde{\mu}_t(\tilde{\boldsymbol{\varepsilon}})) - (f^{\tilde{\boldsymbol{\varepsilon}}}(\tilde{x}_t(\boldsymbol{\varepsilon})) - \tilde{\mu}_t(\tilde{\boldsymbol{\varepsilon}}))^2 \right). \tag{34}
$$

According to Eq. (33), we have for any $\tilde{\boldsymbol{\varepsilon}}$,

$$
|f^{\tilde{\boldsymbol{\varepsilon}}}(\tilde{x}_t(\tilde{\boldsymbol{\varepsilon}})) - s_t(\tilde{\boldsymbol{\varepsilon}})[\tilde{\varepsilon}_t]| \leq \frac{\alpha}{20} \quad \text{and} \quad s_t(\tilde{\boldsymbol{\varepsilon}})[\tilde{\varepsilon}_t] - \mu_t(\tilde{\boldsymbol{\varepsilon}}) = \operatorname{sign}(\widehat{\varepsilon}_t) \cdot \frac{s_t(\tilde{\boldsymbol{\varepsilon}})[1] - s_t(\tilde{\boldsymbol{\varepsilon}})[-1]}{2}.
$$

Eq. (33) also gives that $s_t(\tilde{\boldsymbol{\varepsilon}})[1] - s_t(\tilde{\boldsymbol{\varepsilon}})[-1] \geq \alpha$, which implies

$$
\frac{9}{10} \cdot \frac{s_t(\tilde{\boldsymbol{\varepsilon}})[1] - s_t(\tilde{\boldsymbol{\varepsilon}})[-1]}{2} \leq \operatorname{sign}(\widehat{\varepsilon}_t)(f^{\tilde{\boldsymbol{\varepsilon}}}(\tilde{x}_t(\tilde{\boldsymbol{\varepsilon}})) - \tilde{\mu}_t(\tilde{\boldsymbol{\varepsilon}})) \leq \frac{11}{10} \cdot \frac{s_t(\tilde{\boldsymbol{\varepsilon}})[1] - s_t(\tilde{\boldsymbol{\varepsilon}})[-1]}{2}.
$$

Therefore we can lower bound Eq. (34) by

$$
\mathcal{R}_n(\mathcal{F}, \boldsymbol{\mu}, \mathbf{x}, \boldsymbol{\varepsilon}) \geq \sum_{t=1}^{d} k_t \cdot \left( C|\widehat{\varepsilon}_t| \cdot \frac{9}{10} \cdot \frac{s_t(\tilde{\boldsymbol{\varepsilon}})[1] - s_t(\tilde{\boldsymbol{\varepsilon}})[-1]}{2} - \frac{121}{100} \cdot \left( \frac{s_t(\tilde{\boldsymbol{\varepsilon}})[1] - s_t(\tilde{\boldsymbol{\varepsilon}})[-1]}{2} \right)^2 \right).
$$

We notice that in the $t$-th summand in the right hand side, the only term which depend on $\varepsilon_{[t-1,1]:[t-1,k_t]}$ is $|\widehat{\varepsilon}_t|$, and all other terms only depends on $\varepsilon_{1:[t-1:0]}$. Hence we can lower bound $\mathbb{E}\left[\mathcal{R}_n(\mathcal{F}, \boldsymbol{\mu}, \mathbf{x}, \boldsymbol{\varepsilon})\right]$ by

$$
\begin{aligned}
&\mathbb{E}\left[\mathcal{R}_n(\mathcal{F}, \boldsymbol{\mu}, \mathbf{x}, \varepsilon)\right] \\
&\quad \geq \sum_{t=1}^{d} \mathbb{E}\left[ k_t \cdot \left( C|\widehat{\varepsilon}_t| \cdot \frac{9}{10} \cdot \frac{s_t(\tilde{\boldsymbol{\varepsilon}})[1] - s_t(\tilde{\boldsymbol{\varepsilon}})[-1]}{2} - \frac{121}{100} \cdot \left( \frac{s_t(\tilde{\boldsymbol{\varepsilon}})[1] - s_t(\tilde{\boldsymbol{\varepsilon}})[-1]}{2} \right)^2 \right) \right] \\
&\quad = \sum_{t=1}^{d} \mathbb{E}\left[ k_t \cdot \left( C\mathbb{E}\left[|\widehat{\varepsilon}_t| \mid t_{1:[t-1,0]}\right] \cdot \frac{9}{10} \cdot \frac{s_t(\tilde{\boldsymbol{\varepsilon}})[1] - s_t(\tilde{\boldsymbol{\varepsilon}})[-1]}{2} - \frac{121}{100} \cdot \left( \frac{s_t(\tilde{\boldsymbol{\varepsilon}})[1] - s_t(\tilde{\boldsymbol{\varepsilon}})[-1]}{2} \right)^2 \right) \right] \\
&\quad \geq \sum_{t=1}^{d} \mathbb{E}\left[ k_t \cdot \left( C\sqrt{\frac{1}{2k_t}} \cdot \frac{9}{10} \cdot \frac{s_t(\tilde{\boldsymbol{\varepsilon}})[1] - s_t(\tilde{\boldsymbol{\varepsilon}})[-1]}{2} - \frac{121}{100} \cdot \left( \frac{s_t(\tilde{\boldsymbol{\varepsilon}})[1] - s_t(\tilde{\boldsymbol{\varepsilon}})[-1]}{2} \right)^2 \right) \right]
\end{aligned}
$$

(35)

where the last inequality uses the Khintchine inequality [Haa81]. Next notice that according to our choice of $k_t$,

$$k_t = \left\lfloor \frac{1}{(s_t(\tilde{\boldsymbol{\varepsilon}})[1] - s_t(\tilde{\boldsymbol{\varepsilon}})[-1])^2} \right\rfloor \vee 1 \in \left[ \frac{1}{2(s_t(\tilde{\boldsymbol{\varepsilon}})[1] - s_t(\tilde{\boldsymbol{\varepsilon}})[-1])^2}, \frac{4}{(s_t(\tilde{\boldsymbol{\varepsilon}})[1] - s_t(\tilde{\boldsymbol{\varepsilon}})[-1])^2} \right],$$

which implies that $1/\sqrt{2} \le \sqrt{k_t}(s_t(\tilde{\boldsymbol{\varepsilon}})[1] - s_t(\tilde{\boldsymbol{\varepsilon}})[-1]) \le 2$. Therefore, since $C \ge 2$, we have

$$\begin{aligned}
\text{RHS of } \textit{Eq. (35)} &\ge \sum_{t=1}^{d} \mathbb{E}\left[ k_t \cdot \left( C\sqrt{\frac{1}{2k_t}} \cdot \frac{9}{10} \cdot \frac{s_t(\tilde{\boldsymbol{\varepsilon}})[1] - s_t(\tilde{\boldsymbol{\varepsilon}})[-1]}{2} - \frac{121}{100} \cdot \left( \frac{s_t(\tilde{\boldsymbol{\varepsilon}})[1] - s_t(\tilde{\boldsymbol{\varepsilon}})[-1]}{2} \right)^2 \right) \right] \\
&\ge \sum_{t=1}^{d} \mathbb{E}\left[ \left( \frac{9C}{20\sqrt{2}} - \frac{121}{200} \right) \cdot \sqrt{k_t} \cdot |s_t(\tilde{\boldsymbol{\varepsilon}})[1] - s_t(\tilde{\boldsymbol{\varepsilon}})[-1]| \right] \ge \frac{d}{50}.
\end{aligned}$$

Therefore, we obtain that

$$\sup_{\boldsymbol{\mu},\mathbf{x}} \mathbb{E}[\mathcal{R}_n(\mathcal{F}, \boldsymbol{\mu}, \mathbf{x}, \boldsymbol{\varepsilon})] \ge \frac{d}{50} = \tilde{\Omega}\left( n^{\frac{p}{p+2}} \right).$$

□

*Proof of Corollary 2.* Since $\sup_{\mathbf{x}} \log \mathcal{N}_\infty^{\mathsf{seq}}(\mathcal{F}, \mathbf{x}, \alpha) = \tilde{\Omega}\left(\alpha^{-p}\right)$, according to Proposition 5 we have

$$d^{\mathsf{seq}}(\mathcal{F}, \alpha, \alpha/20) = \tilde{\Omega}\left(\alpha^{-p}\right).$$

Next, we call Theorem 2 with $C = 2$. According to [RS14, Lemma 4], we obtain

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

□