Sequential Complexities and Uniform Martingale Laws of Large Numbers

Alexander Rakhlin Karthik Sridharan Ambuj Tewari

September 19, 2013

Abstract

We establish necessary and sufficient conditions for a uniform martingale Law of Large Numbers. We extend the technique of symmetrization to the case of dependent random variables and provide "sequential" (non-i.i.d.) analogues of various classical measures of complexity, such as covering numbers and combinatorial dimensions from empirical process theory. We establish relationships between these various sequential complexity measures and show that they provide a tight control on the uniform convergence rates for empirical processes with dependent data. As a direct application of our results, we provide exponential inequalities for sums of martingale differences in Banach spaces.

1 Introduction

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be an arbitrary complete probability space. Let $\mathcal{Z}$ be a separable metric space and $\mathcal{F} = \{f : \mathcal{Z} \mapsto \mathbb{R}\}$ be a set of bounded real valued functions on $\mathcal{Z}$. Consider independent and identically distributed random variables $Z_1, \ldots, Z_n, \ldots$ in $\mathcal{Z}$ with the common distribution $\mathbf{P}$. The empirical process indexed by $f \in \mathcal{F}$ is defined as

$$f \mapsto \mathbb{G}_n(f) := \frac{1}{n} \sum_{t=1}^{n} \Big(\mathbb{E} f(Z) - f(Z_t) \Big) .$$

The study of the behavior of the *supremum* of this process is a central topic in empirical process theory, and it is well known that this behavior depends on the "richness" of $\mathcal{F}$. Statements about convergence of the supremum to zero are known as uniform Laws of Large Numbers (LLN). More precisely, a class $\mathcal{F}$ is said to be (strong) Glivenko-Cantelli for the distribution $\mathbf{P}$ if the supremum of $\mathbb{G}_n(f)$ converges to zero almost surely as $n \to \infty$. Of particular interest are classes for which this convergence happens uniformly for all distributions. A class $\mathcal{F}$ is said to be *uniform Glivenko-Cantelli* if

$$\forall \delta > 0, \lim_{n' \to \infty} \sup_{\mathbf{P}} \mathbb{P}\left(\sup_{n \geq n'} \sup_{f \in \mathcal{F}} |\mathbb{G}_n(f)| > \delta \right) = 0 \tag{1}$$

where $\mathbb{P}$ is the product measure $\mathbf{P}^\infty$. As a classical example, consider i.i.d. random variables $Z_1, \ldots, Z_n$ in $\mathbb{R}$ and a class $\mathcal{F} = \{\mathbf{1}\{z \leq \theta\} : \theta \in \mathbb{R}\}$, where $\mathbf{1}\{\}$ is the indicator function. For this class, (1) holds by the well known results of Glivenko and Cantelli: almost surely, the supremum of the difference between the cumulative distribution function and the empirical distribution function converges to zero. A number of necessary and sufficient conditions for the Glivenko-Cantelli and the uniform Glivenko-Cantelli properties have been derived over the past several decades [11].

In this paper, we are interested in the martingale analogues of the uniform LLN, as well as in the analogues to the various notions of complexity that appear in empirical process theory. Specifically, consider a sequence of random variables $(Z_t)_{t\geq 1}$ adapted to a filtration $(\mathcal{A}_t)_{t\geq 1}$. We are interested in the following process indexed by $f \in \mathcal{F}$:

$$f \mapsto \mathbb{M}_n(f) := \frac{1}{n}\sum_{t=1}^{n}\Big(\mathbb{E}[f(Z_t)|\mathcal{A}_{t-1}] - f(Z_t)\Big) .$$

The central object of study in this paper is the supremum of the process $\mathbb{M}_n(f)$, and in particular we address the question of whether a uniform convergence similar to (1) holds. Evidently, $\mathbb{M}_n(f)$ coincides with $\mathbb{G}_n(f)$ in the case when $Z_1, Z_2, \ldots$ are i.i.d. random variables. More generally, for any fixed $f \in \mathcal{F}$, the sequence $\big(\mathbb{E}\left[f(Z_t) \mid \mathcal{A}_{t-1}\right] - f(Z_t)\big)_{t\geq 1}$ is a martingale difference sequence. Similar to the notion of uniform Glivenko-Cantelli class $\mathcal{F}$, we can define the notion of uniform convergence for dependent random variables over function class $\mathcal{F}$ as follows.

Definition 1. A function class $\mathcal{F}$ satisfies *Sequential Uniform Convergence* if,

$$\forall \delta > 0, \; \lim_{n'\to\infty} \sup_{\mathbb{P}} \; \mathbb{P}\left(\sup_{n\geq n'} \sup_{f\in\mathcal{F}} |\mathbb{M}_n(f)| > \delta\right) = 0 \; , \tag{2}$$

where the supremum is over all distributions $\mathbb{P}$ on the space $(\Omega, \mathcal{A})$.

The gap between properties (1) and (2) is already witnessed by the example of the class $\mathcal{F} = \{\mathbf{1}\{z \leq \theta\} : \theta \in \mathbb{R}\}$ of functions on $\mathbb{R}$, discussed earlier. In contrast to the uniform Glivenko-Cantelli property, the martingale analogue (2) does not hold for this class. On the positive side, the necessary and sufficient conditions for a class $\mathcal{F}$ to satisfy sequential uniform convergence, as derived in this paper, can be verified for a wide range of interesting classes.

2 Summary of the Results

One of the main results in this paper is the following equivalence.

Theorem 1. *Let $\mathcal{F}$ be a class of $[-1,1]$-valued functions. Then the following statements are equivalent.*

1. *$\mathcal{F}$ satisfies Sequential Uniform Convergence.*
2. *For any $\alpha > 0$, the sequential fat-shattering dimension $\mathrm{fat}_\alpha(\mathcal{F})$ is finite.*
3. *Sequential Rademacher complexity $\mathfrak{R}_n(\mathcal{F})$ satisfies $\lim_{n\to\infty}\mathfrak{R}_n(\mathcal{F}) = 0$.*

Theorem 1 yields a characterization of the uniform convergence property in terms of two quantities. The first one is a combinatorial "dimension" of the class at scale α (Definition 7). The second is a measure of complexity of the class through random averages (Definition 3). In addition to these quantities, we define sequential versions of covering numbers and the associated Dudley-type entropy integral. En route to proving Theorem 1, we obtain key relationships between the introduced covering numbers, the combinatorial dimensions, and random averages. These relationships constitute the bulk of the paper, and can be considered as martingale extensions of the results in empirical process theory. Specifically, we show

- A relationship between the empirical process with dependent random variables and the sequential Rademacher complexity (Theorem 2), obtained through sequential symmetrization.
- An upper bound of sequential Rademacher complexity by a Dudley-type entropy integral through the chaining technique (Theorem 4).
- An upper bound on sequential covering numbers in terms of the combinatorial dimensions (Theorems 5 and 7), as well as Corollary 6. In particular, Theorem 7 is a sequential analogue of the celebrated Vapnik-Chervonenkis-Sauer-Shelah lemma.
- A relationship between the combinatorial dimension and sequential Rademacher complexity (Lemma 8) and, as a consequence, equivalence of many of the introduced complexity notions up to a poly-logarithmic factor.
- Properties of sequential Rademacher complexity and, in particular, the contraction inequality (Lemma 13).
- An extension of the above results to high-probability statements (Lemmas 10, 11, and 12) and an application to concentration of martingales in Banach spaces (Corollary 17).

This paper is organized as follows. In the next section we place the present paper in the context of previous work. In Sections 4-6 we introduce sequential complexities. A characterization of sequential uniform convergence appears in Section 7. We conclude the paper with some structural results in Section 8 and an application to exponential inequalities for sums of martingale difference sequences in Banach spaces in Section 9. Most proofs are deferred to the appendix.

3 Related Literature

The seminal work of Vapnik and Chervonenkis [36] provided the first necessary and sufficient conditions – via a notion of random VC entropy – for a class $\mathcal{F}$ of binary valued functions to be a Glivenko-Cantelli (GC) class. These results were strengthened by Steele [28], who showed almost sure convergence. A similar characterization of the GC property via a notion of a *covering number* in the case of uniformly bounded real-valued functions appears in [37]. For the binary-valued case, a distribution-independent version of the VC entropy (termed the growth function) was shown by Vapnik and Chervonenkis [36] to yield a sufficient condition for the *uniform* GC property. The "necessary" direction was first shown (according to [11, p. 229]) in an unpublished manuscript of Assouad, 1982. For real-valued classes of functions, the necessary and sufficient conditions for the uniform GC property were established in [12] through a notion of a covering number similar to the Koltchinskii-Pollard entropy. A characterization of GC classes for a fixed distribution were also given by Talagrand [29, 30] through a notion of a "witness of irregularity". Similar in spirit, the pseudo-dimension introduced in [23] was shown by Pollard to be sufficient, though not necessary, for the uniform GC property. A scale-sensitive version of pseudo-dimension (termed the *fat-shattering dimension* by [5]) was introduced by Kearns and Schapire [15]. Finiteness of this dimension at all scales was shown in [3] to characterize the uniform GC classes. We refer the reader to [11, Chapter 6] and [32, 33] for a much more detailed account of the results.

The GC-type theorems have also been extended to the case of weakly dependent random variables. For instance, Yukich [39] relies on a ϕ-mixing assumption, while Nobel and Dembo [20] and Yu [38] consider β-mixing sequences. For a countable class with a finite VC dimension, a GC theorem has been recently shown by Adams and Nobel [2] for ergodic sequences. We refer the reader to [2, 10] for a more comprehensive survey of results for non-i.i.d. data. Notably, the aforementioned papers prove a GC-type property under much the same type of complexity measures as in the i.i.d. case. This is in contrast to the present paper, where the classical notions do not provide answers to the questions of convergence.

In this paper, we do not make mixing or ergodicity assumptions on the sequence of random variables. However, the definition of $\mathbb{M}_n(f)$ imposes a certain structure which is not present when an average is compared with a single expected value. Thus, our results yield an extension of the GC property to non-i.i.d. data in a direction that is different from the papers mentioned above. Such an extension has already been considered in the literature: the quantity $\sup_{f\in\mathcal{F}}\mathbb{M}_n(f)$ has been studied by S. van de Geer [33] (see Chapter 8.2). Dudley integral type upper bounds for a given distribution $\mathbb{P}$ were provided in terms of the so called generalized entropy with bracketing, corresponding to the particular distribution $\mathbb{P}$. This is a *sufficient* condition for convergence of the supremum of $\mathbb{M}_n(f)$ for the given distribution. In this work, however, we are interested in providing necessary and sufficient conditions for the uniform analogue of the GC property, as well as in extending the ideas of symmetrization, covering numbers, and scale-sensitive dimensions to the non-i.i.d. case. Towards the end of Section 7, we discuss the relationship between the generalized entropy with bracketing of [33] and the tools provided in this work. We also stress that this paper studies martingale uniform laws of large numbers rather than a convergence of $n\mathbb{M}_n(f)$, which only holds under stringent conditions; such a convergence for reverse martingales has been studied in [35]. The question of the limiting behavior of $\sqrt{n}\mathbb{M}_n(f)$ (that is, the analogue of the Donsker property [11]) is also outside of the scope of this paper.

The study of the supremum of the process $\mathbb{M}_n(f)$ has many potential applications. For instance, in [34], the quantity $\sup_{f\in\mathcal{F}}\mathbb{M}_n(f)$ is used to provide bounds on estimation rates for autoregressive models. In [1, 24] connections between minimax rates of sequential prediction problems and the supremum of the process $\mathbb{M}_n(f)$ over the associated class of predictors $\mathcal{F}$ are established. In Section 9 of this work, we show how the supremum of $\mathbb{M}_n(f)$ over class of linear functionals can be used to derive exponential inequalities for sums of martingale differences in general Banach spaces.

4 Symmetrization and the Tree Process

A key tool in deriving classical uniform convergence theorems (for i.i.d. random variables) is symmetrization. The main idea behind symmetrization is to compare the empirical process $\mathbb{G}_n(f)$ over a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ to a symmetrized empirical process, called the Rademacher process, over the probability space $(\Omega^\epsilon, \mathcal{B}, \mathbb{P}_\epsilon)$ where $\Omega^\epsilon = \{-1, 1\}^{\mathbb{N}}$, $\mathcal{B}$ the Borel σ-algebra and $\mathbb{P}_\epsilon$ the uniform probability measure. We use the notation $\mathbb{E}_\epsilon$ to represent expectation under the measure $\mathbb{P}_\epsilon$, and $(\mathcal{B}_t)_{t\geq 0}$ to denote the dyadic filtration on Ω^ϵ given by $\mathcal{B}_t = \sigma(\epsilon_1, \ldots, \epsilon_t)$, where ϵ_t's are independent symmetric $\{\pm 1\}$-valued *Rademacher random variables* and $\mathcal{B}_0 = \{\{\}, \Omega^\epsilon\}$.

Given $z_1, \ldots, z_n \in \mathcal{Z}$, the Rademacher process $\mathbb{S}_n^{(z_{1:n})}(f)$ is defined[1] as

$$f \mapsto \mathbb{S}_n^{(z_{1:n})}(f) := \frac{1}{n} \sum_{t=1}^{n} \epsilon_t f(z_t) \ . \tag{3}$$

It is well-known (e.g. [32]) that the behavior of the supremum of the symmetrized process $\mathbb{S}_n^{(z_{1:n})}(f)$ is closely related to the behavior of the supremum of the empirical process as

$$\mathbb{E} \sup_{f\in\mathcal{F}} \mathbb{G}_n(f) \leq 2 \sup_{z_1,\ldots,z_n\in\mathcal{Z}} \mathbb{E} \sup_{f\in\mathcal{F}} \mathbb{S}_n^{(z_{1:n})}(f) \tag{4}$$

and a similar high-probability statement can also be proved. Note that the Rademacher process is defined on the probability space $(\Omega^\epsilon, \mathcal{B}, \mathbb{P}_\epsilon)$, which is potentially easier to handle than the original probability space for the empirical process.

In the non-i.i.d. case, however, a similar symmetrization argument requires significantly more care and relies on the notion of *decoupled tangent sequences* [9, Def. 6.1.4]. Fix a sequence of random variables $(Z_t)_{t\geq 1}$ adapted to the filtration $(\mathcal{A}_t)_{t\geq 1}$. A sequence of random variables $(Z'_t)_{t\geq 1}$ is said to be a decoupled sequence tangent to $(Z_t)_{t\geq 1}$ if for each t, conditioned on $Z_1, \ldots, Z_{t-1}$, the random variables Z_t and Z'_t are independent and identically distributed. Thus, the random variables $(Z'_t)_{t\geq 1}$ are conditionally independent given $(Z_t)_{t\geq 1}$. In Theorem 2 below, a sequential symmetrization argument is applied to the decoupled sequences, leading to a tree process – an analogue of the Rademacher process for the non-i.i.d. case. First, let us define the notion of a tree.

A *$\mathcal{Z}$-valued tree* $\mathbf{z}$ *of depth* n is a rooted complete binary tree with nodes labeled by elements of $\mathcal{Z}$. We identify the tree $\mathbf{z}$ with the sequence $(\mathbf{z}_1, \ldots, \mathbf{z}_n)$ of labeling functions $\mathbf{z}_i : \{\pm 1\}^{i-1} \mapsto \mathcal{Z}$ which provide the labels for each node. Here, $\mathbf{z}_1 \in \mathcal{Z}$ is the label for the *root* of the tree, while $\mathbf{z}_i$ for $i > 1$ is the label of the node obtained by following the path of length $i-1$ from the root, with $+1$ indicating 'right' and -1 indicating 'left'. A *path* of length n is given by the sequence $\epsilon = (\epsilon_1, \ldots, \epsilon_n) \in \{\pm 1\}^n$. For brevity, we shall often write $\mathbf{z}_t(\epsilon)$, but it is understood that $\mathbf{z}_t$ only depends only on the prefix $(\epsilon_1, \ldots, \epsilon_{t-1})$ of ϵ. Given a tree $\mathbf{z}$ and a function $f : \mathcal{Z} \mapsto \mathbb{R}$, we define the composition $f \circ \mathbf{z}$ as a real-valued tree given by the labeling functions $(f \circ \mathbf{z}_1, \ldots, f \circ \mathbf{z}_n)$.

Observe that if $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. Rademacher random variables, then

$$(\epsilon_t f(\mathbf{z}_t(\epsilon_1, \ldots, \epsilon_{t-1})))_{t=1}^{n}$$

is a martingale-difference sequence for any given f.

[1] For integers $a \leq b$, we denote a sequence of the form $(y_a, \ldots, y_b)$ by $y_{a:b}$. For any $n \in \mathbb{N}$, we use $[n]$ to denote the set $\{1, \ldots, n\}$.

Definition 2. Let $\epsilon_1, \dots, \epsilon_n$ be independent Rademacher random variables. Given a $\mathcal{Z}$-valued tree $\mathbf{z}$ of depth n, the stochastic process

$$f \mapsto \mathbb{T}_n^{(\mathbf{z})}(f) := \frac{1}{n} \sum_{t=1}^{n} \epsilon_t f(\mathbf{z}_t(\epsilon_1, \dots, \epsilon_{t-1}))$$

will be called the *tree process* indexed by $\mathcal{F}$.

We may view the tree process $\mathbb{T}_n^{(\mathbf{z})}(f)$ as a generalization of the Rademacher process $\mathbb{S}_n^{(z_{1:n})}(f)$. Indeed, suppose $(\mathbf{z}_1, \dots, \mathbf{z}_n)$ is a sequence of constant labeling functions such that for any $t \in [n]$, $\mathbf{z}_t(\epsilon_1, \dots, \epsilon_{t-1}) = z_t$ for any $(\epsilon_1, \dots, \epsilon_{t-1})$. In this case, $\mathbb{T}_n^{(\mathbf{z})}(f)$ and $\mathbb{S}_n^{(z_{1:n})}(f)$ coincide. In general, however, the tree process can behave differently (in a certain sense) from the Rademacher process.

Given $z_1, \dots, z_n$, the expected supremum of the Rademacher process in (4) is known as (empirical) Rademacher averages or Rademacher complexity of the function class. We propose the following definition for the tree process:

Definition 3. The *sequential Rademacher complexity* of a function class $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{Z}}$ on a $\mathcal{Z}$-valued tree $\mathbf{z}$ is defined as

$$\mathfrak{R}_n(\mathcal{F}, \mathbf{z}) = \mathbb{E} \sup_{f \in \mathcal{F}} \mathbb{T}_n^{(\mathbf{z})}(f) = \mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \epsilon_t f(\mathbf{z}_t(\epsilon))\right]$$

and

$$\mathfrak{R}_n(\mathcal{F}) = \sup_{\mathbf{z}} \mathfrak{R}_n(\mathcal{F}, \mathbf{z})$$

where the outer supremum is taken over all $\mathcal{Z}$-valued trees of depth n, and $\epsilon = (\epsilon_1, \dots, \epsilon_n)$ is a sequence of i.i.d. Rademacher random variables.

Theorem 2. *The following relation holds between the empirical process with dependent random variables and the sequential Rademacher complexity:*

$$\mathbb{E} \sup_{f \in \mathcal{F}} \mathbb{M}_n(f) \le 2 \mathfrak{R}_n(\mathcal{F}) \ . \tag{5}$$

Furthermore, this bound is tight, as we have

$$\frac{1}{2}\left(\mathfrak{R}_n(\mathcal{F}) - \frac{B}{2\sqrt{n}}\right) \le \sup_{\mathbb{P}} \mathbb{E} \sup_{f \in \mathcal{F}} \mathbb{M}_n(f) \tag{6}$$

where $B = \inf_{z \in \mathcal{Z}} \sup_{f, f' \in \mathcal{F}} (f(z) - f'(z)) \ge 0$.

We would like to point out that in general $\mathfrak{R}_n(\mathcal{F}) = \Omega\left(\frac{B}{\sqrt{n}}\right)$, and so in the worst case the behavior of the expected supremum of $\mathbb{M}_n$ is precisely given by the sequential Rademacher complexity. Further, we remark that for a class $\mathcal{F}$ of linear functions on some subset $\mathcal{Z}$ of a vector space such that $0 \in \mathcal{Z}$, we have $B \le 0$ and the lower bound becomes $\frac{1}{2}\mathfrak{R}_n(\mathcal{F})$.

The proof of Theorem 2 requires more work than the classical symmetrization proof [11, 18] due to the non-i.i.d. nature of the sequences. To readers familiar with the notion of *martingale type* in the theory of Banach spaces we would like to point out that the tree process can be viewed as an analogue of Walsh-Paley martingales. The upper bound of Theorem 2 is a generalization of the fact that the expected norm of a sum of martingale difference sequences can be upper bounded by the expected norm of sum of Walsh-Paley martingale difference sequences, as shown in [22].

As mentioned earlier, the sequential Rademacher complexity is an object that is easier to study than the original empirical process $\mathbb{M}_n$. The following sections introduce additional notions of complexity of a function class that provide control of the sequential Rademacher complexity. Specific relations between these complexity notions will be shown, leading to the proof of Theorem 1.

5 Finite Classes, Covering Numbers, and Chaining

The first step in upper bounding sequential Rademacher complexity is the following result for a finite collection of trees.

Lemma 3. *For any finite set V of $\mathbb{R}$-valued trees of depth n we have that*

$$\mathbb{E}\left[\max_{\mathbf{v}\in V}\sum_{t=1}^{n}\epsilon_t\mathbf{v}_t(\epsilon)\right]\le\sqrt{2\log(|V|)\max_{\mathbf{v}\in V}\max_{\epsilon\in\{\pm1\}^n}\sum_{t=1}^{n}\mathbf{v}_t(\epsilon)^2}$$

where $|V|$ denotes the cardinality of the set V.

A simple consequence of the above lemma is that if $\mathcal{F}\subseteq[-1,1]^{\mathcal{Z}}$ is a finite class, then for any tree $\mathbf{z}$, we have that

$$\mathbb{E}\left[\max_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\epsilon_t f(\mathbf{z}_t(\epsilon))\right]\le\mathbb{E}\left[\max_{\mathbf{v}\in\mathcal{F}(\mathbf{z})}\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\mathbf{v}_t(\epsilon)\right]\le\sqrt{\frac{2\log(|\mathcal{F}|)}{n}}\ , \tag{7}$$

where $\mathcal{F}(\mathbf{z})=\{f\circ\mathbf{z}:f\in\mathcal{F}\}$ is the *projection* of $\mathcal{F}$ onto $\mathbf{z}$. It is clear that $|\mathcal{F}(\mathbf{z})|\le|\mathcal{F}|$ which explains the second inequality above.

To illustrate the next idea, consider a binary-valued function class $\mathcal{F}\subseteq\{\pm1\}^{\mathcal{Z}}$. In the i.i.d. case, the cardinality of the coordinate projection

$$\left\{(f(z_1),\ldots,f(z_n)):f\in\mathcal{F}\right\}$$

immediately yields a control of the supremum of the empirical process. For the tree-based definition, however, it is easy to see that the cardinality of $\mathcal{F}(\mathbf{z})$ is exponential in n for any interesting class $\mathcal{F}$, leading to a vacuous upper bound.

A key observation is that the first inequality in (7) holds with $\mathcal{F}(\mathbf{z})$ replaced by a potentially smaller set V of $\mathbb{R}$-valued trees with the property that

$$\forall f\in\mathcal{F},\ \forall\epsilon\in\{\pm1\}^n,\ \exists\mathbf{v}\in V\ \text{ s.t. }\quad\mathbf{v}_t(\epsilon)=f(\mathbf{z}_t(\epsilon)) \tag{8}$$

for all $t\in[n]$. Crucially, the choice of $\mathbf{v}$ may depend on ϵ. A set V of $\mathbb{R}$-valued trees satisfying (8) is termed a 0-*cover* of $\mathcal{F}\subseteq\mathbb{R}^{\mathcal{Z}}$ on a tree $\mathbf{z}$ of depth n. We denote by $\mathcal{N}(0,\mathcal{F},\mathbf{z})$ the size of a smallest 0-cover on $\mathbf{z}$ and define $\mathcal{N}(0,\mathcal{F},n)=\sup_{\mathbf{z}}\mathcal{N}(0,\mathcal{F},\mathbf{z})$.

To illustrate the gap between the size of a 0-cover and the cardinality of $\mathcal{F}(\mathbf{z})$, consider a tree $\mathbf{z}$ of depth n and suppose for simplicity that $|\mathrm{Img}(\mathbf{z})|=2^n-1$ where $\mathrm{Img}(\mathbf{z})=\cup_{t\in[n]}\mathrm{Img}(\mathbf{z}_t)$ and

$\mathrm{Img}(\mathbf{z}_t) = \{\mathbf{z}_t(\epsilon) : \epsilon \in \{\pm 1\}^n\}$. Suppose $\mathcal{F}$ consists of 2^{n-1} binary-valued functions defined as zero on all of $\mathrm{Img}(\mathbf{z})$ except for a single value of $\mathrm{Img}(\mathbf{z}_n)$. In plain words, each function is zero everywhere on the tree except for a single leaf. While the projection $\mathcal{F}(\mathbf{z})$ contains 2^{n-1} distinct trees, the size of a 0-cover is only 2: it is enough to take an all-zero function g_0 along with a function g_1 which is zero on all of $\mathrm{Img}(\mathbf{z})$ except $\mathrm{Img}(\mathbf{z}_n)$ (i.e. on the leaves). It is easy to verify that $g_0 \circ \mathbf{z}$ and $g_1 \circ \mathbf{z}$ provide a 0-cover for $\mathcal{F}$ on $\mathbf{z}$. Unlike $|\mathcal{F}(\mathbf{z})|$, the size of the cover does not grow with n, capturing the fact that the function class is "simple" on any given path.

For real-valued function classes, the notion of a 0-cover needs to be relaxed to incorporate scale. We propose the following definitions.

Definition 4. A set V of $\mathbb{R}$-valued trees of depth n is *a (sequential) α-cover* (with respect to ℓ_p-norm) of $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{Z}}$ on a tree $\mathbf{z}$ of depth n if

$$\forall f \in \mathcal{F},\ \forall \epsilon \in \{\pm 1\}^n,\ \exists \mathbf{v} \in V \quad \text{s.t.} \quad \left(\frac{1}{n} \sum_{t=1}^{n} |\mathbf{v}_t(\epsilon) - f(\mathbf{z}_t(\epsilon))|^p \right)^{1/p} \le \alpha \ .$$

The *(sequential) covering number* of a function class $\mathcal{F}$ on a given tree $\mathbf{z}$ is defined as

$$\mathcal{N}_p(\alpha, \mathcal{F}, \mathbf{z}) = \min \left\{ |V| : V \text{ is an } \alpha\text{-cover w.r.t. } \ell_p\text{-norm of } \mathcal{F} \text{ on } \mathbf{z} \right\}.$$

Further define $\mathcal{N}_p(\alpha, \mathcal{F}, n) = \sup_{\mathbf{z}} \mathcal{N}_p(\alpha, \mathcal{F}, \mathbf{z})$, the maximal ℓ_p covering number of $\mathcal{F}$ over depth-n trees.

In the study of the supremum of a stochastic process indexed by a set S, it is natural to endow the set with a pseudo-metric d. The structure and "richness" of the index set S (as given by covering numbers or, more generally, via the chaining technique [31, 33]) yield precise control on the supremum of the stochastic process. It is natural to ask whether we can endow the projection $\mathcal{F}(\mathbf{z})$ with a metric and appeal to known results. This turns out to be not quite the case, as the pseudo-metric needs to be random. Indeed, observe that the tree $\mathbf{v}$ providing the cover may depend on the path ϵ itself. We may define the random pseudo-metric between the $\mathbb{R}$-valued trees $\mathbf{v}', \mathbf{v}$ as

$$d^p_\epsilon(\mathbf{v}', \mathbf{v}) = \left(\frac{1}{n} \sum_{t=1}^{n} \left| \mathbf{v}'_t(\epsilon) - \mathbf{v}_t(\epsilon) \right|^p \right)^{1/p} \ .$$

An α-cover V then guarantees that, for any $\epsilon \in \{\pm 1\}^n$,

$$\sup_{\mathbf{v}' \in \mathcal{F}(\mathbf{z})} \inf_{\mathbf{v} \in V} d^p_\epsilon(\mathbf{v}', \mathbf{v}) \le \alpha \ .$$

Therefore, our results below can be seen as extending the chaining technique to the case of a random pseudo-metric d^p_ϵ.

With the definition of an α-cover with respect to ℓ_1 norm, it is immediate (using Lemma 3) that for any $\mathcal{F} \subset [-1, 1]^{\mathcal{Z}}$, for any $\alpha > 0$,

$$\mathfrak{R}_n(\mathcal{F}, \mathbf{z}) \le \alpha + \sqrt{\frac{2 \log \mathcal{N}_1(\alpha, \mathcal{F}, \mathbf{z})}{n}} \ . \tag{9}$$

It is recognized, however, that a tighter control is obtained by integrating the covering numbers at different scales. To this end, consider the following analogue of the Dudley entropy integral bound.

Definition 5. For $p \geq 2$, the *integrated complexity* of a function class $\mathcal{F} \subseteq [-1,1]^{\mathcal{Z}}$ on a $\mathcal{Z}$-valued tree of depth n is defined as

$$\mathfrak{D}_n^p(\mathcal{F}, \mathbf{z}) = \inf_{\alpha} \left\{ 4\alpha + \frac{12}{\sqrt{n}} \int_{\alpha}^{1} \sqrt{\log \ \mathcal{N}_p(\delta, \mathcal{F}, \mathbf{z})} \, d\delta \right\}$$

and

$$\mathfrak{D}_n^p(\mathcal{F}) = \sup_{\mathbf{z}} \mathfrak{D}_n^p(\mathcal{F}, \mathbf{z}).$$

We denote $\mathfrak{D}_n^2(\mathcal{F}, \mathbf{z})$ as $\mathfrak{D}_n(\mathcal{F}, \mathbf{z})$. Clearly, $\mathfrak{D}_n^p(\mathcal{F}, \mathbf{z}) \leq \mathfrak{D}_n^q(\mathcal{F}, \mathbf{z})$ for $p \leq q$.

Theorem 4. *For any function class $\mathcal{F} \subseteq [-1,1]^{\mathcal{Z}}$, we have that*

$$\mathfrak{R}_n(\mathcal{F}, \mathbf{z}) \leq \mathfrak{D}_n(\mathcal{F}, \mathbf{z})$$

for any $\mathcal{Z}$-valued tree $\mathbf{z}$ of depth n.

We conclude this section by mentioning that two distinct notions of a *packing* (or, α-separated set) exist for trees, according to the order of quantifiers in the definition. In one definition, it must be that every member of the packing set is α-separated from every other member on *some* path. For the other, there must be a path on which every member of the packing is α-separated from every other member. In the classical case the distinction does not arise, and the packing number is known to be closely related to the covering number. For the tree case, however, the two notions are distinct, one providing an upper bound and one a lower bound on the covering number. Due to this discrepancy, difficulties arise in attempting to replicate proofs that pass through the packing number, such as the Dudley's extraction technique [18] for obtaining estimates on the ℓ_2 covering numbers.

6 Combinatorial Parameters

For i.i.d. data, the uniform Glivenko-Cantelli property for classes of binary-valued functions is characterized by the Vapnik-Chervonenkis combinatorial dimension [36]. For real-valued function classes, the corresponding notions are the scale-sensitive dimensions, such as the fat-shattering dimension [3, 6]. In this section, we recall the definition of Littlestone dimension [17, 8] and propose its scale-sensitive versions for real-valued function classes. Subsequently, these combinatorial parameters are shown to control the growth of sequential covering numbers.

Definition 6. A $\mathcal{Z}$-valued tree $\mathbf{z}$ of depth d is *shattered* by a function class $\mathcal{F} \subseteq \{\pm 1\}^{\mathcal{Z}}$ if

$$\forall \epsilon \in \{\pm 1\}^d, \quad \exists f \in \mathcal{F} \quad \text{s.t.} \quad \forall t \in [d], \quad f(\mathbf{z}_t(\epsilon)) = \epsilon_t \ .$$

The *Littlestone dimension* $\text{Ldim}(\mathcal{F}, \mathcal{Z})$ is the largest d such that $\mathcal{F}$ shatters a $\mathcal{Z}$-valued tree of depth d.

We propose the following scale-sensitive version of Littlestone dimension.

Definition 7. A $\mathcal{Z}$-valued tree $\mathbf{z}$ of depth d is *α-shattered* by a function class $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{Z}}$ if there exists an $\mathbb{R}$-valued tree $\mathbf{s}$ of depth d such that

$$\forall \epsilon \in \{\pm 1\}^d, \quad \exists f \in \mathcal{F} \quad \text{s.t.} \quad \forall t \in [d], \quad \epsilon_t (f(\mathbf{z}_t(\epsilon)) - \mathbf{s}_t(\epsilon)) \geq \alpha/2 \ .$$

The tree $\mathbf{s}$ will be called a *witness to shattering*. The *(sequential) fat-shattering dimension* $\mathrm{fat}_\alpha(\mathcal{F}, \mathcal{Z})$ at scale α is the largest d such that $\mathcal{F}$ α-shatters a $\mathcal{Z}$-valued tree of depth d.

With these definitions it is easy to see that $\mathrm{fat}_\alpha(\mathcal{F}, \mathcal{Z}) = \mathrm{Ldim}(\mathcal{F}, \mathcal{Z})$ for a binary-valued function class $\mathcal{F} \subseteq \{\pm 1\}^{\mathcal{Z}}$ for any $0 < \alpha \leq 2$.

When $\mathcal{Z}$ and/or $\mathcal{F}$ are understood from the context, we will simply write fat_α or $\mathrm{fat}_\alpha(\mathcal{F})$ instead of $\mathrm{fat}_\alpha(\mathcal{F}, \mathcal{Z})$. Furthermore, we will write $\mathrm{fat}_\alpha(\mathcal{F}, \mathbf{z})$ for $\mathrm{fat}_\alpha(\mathcal{F}, \mathrm{Img}(\mathbf{z}))$. Hence, $\mathrm{fat}_\alpha(\mathcal{F}, \mathbf{z})$ is the largest d such that $\mathcal{F}$ α-shatters a tree $\mathbf{z}'$ of depth d with $\mathrm{Img}(\mathbf{z}') \subseteq \mathrm{Img}(\mathbf{z})$.

If trees $\mathbf{z}$ are defined by constant mappings $\mathbf{z}_t(\epsilon) = z_t$, the combinatorial parameters introduced in the definitions above coincide with the Vapnik-Chervonenkis dimension and its scale-sensitive version, the fat-shattering dimension. Therefore, the notions we are studying lead to a theory that can be viewed as a sequential generalization of the Vapnik-Chervonenkis theory.

We now relate the combinatorial parameters to the size of a sequential cover. In the binary case ($k = 1$ below), a reader might notice a similarity of Theorems 5 and 7 to the classical results due to Sauer [26], Shelah [27], and Vapnik and Chervonenkis [36]. There are several approaches to proving what is often called the Vapnik-Chervonenkis-Sauer-Shelah lemma. We opt for the inductive-style proof (e.g., see the book by Alon and Spencer [4]), which becomes more natural for the case of trees because of their recursive structure.

Theorem 5. *Let $\mathcal{F} \subseteq \{0, \ldots, k\}^{\mathcal{Z}}$ be a class of functions with $\mathrm{fat}_2(\mathcal{F}, \mathcal{Z}) = d$. Then for any $n > d$,*

$$\mathcal{N}_\infty(1/2, \mathcal{F}, n) \ \leq \ \sum_{i=0}^{d} \binom{n}{i} k^i \ \leq \ \left(\frac{ekn}{d}\right)^d .$$

Furthermore, for $n \leq d$,

$$\mathcal{N}_\infty(1/2, \mathcal{F}, n) \leq k^n .$$

Consequently, the upper bound $\mathcal{N}_\infty(1/2, \mathcal{F}, n) \leq (ekn)^d$ holds for any $n \geq 1$.

Armed with Theorem 5, we can approach the problem of bounding the size of a sequential cover at scale α through discretization. For the classical case of a cover based on a set points, the discretization idea appears in [3, 19]. When passing from the combinatorial result to the cover at scale α in Corollary 6, it is crucial that the statement of Theorem 5 is in terms of $\mathrm{fat}_2(\mathcal{F})$ rather than $\mathrm{fat}_1(\mathcal{F})$. This point can be seen in the proof of Corollary 6: unavoidably, the discretization process can map almost identical function values to distinct discrete values which differ by 1, forcing us to demand shattering at scale 2.

We now show that the sequential covering numbers are bounded in terms of the sequential fat-shattering dimension.

Corollary 6. *Let $\mathcal{F} \subseteq [-1, 1]^{\mathcal{Z}}$. For any $\alpha > 0$ and any $n \geq 1$, we have that*

$$\mathcal{N}_\infty(\alpha, \mathcal{F}, n) \leq \left(\frac{2en}{\alpha}\right)^{\mathrm{fat}_\alpha(\mathcal{F})} .$$

In the classical (non-sequential) case, it is known that the ℓ_∞ covering numbers cannot be dimension-independent, suggesting that the dependence on n in the above estimate cannot be removed altogether. It is interesting to note that the clean upper bound of Corollary 6 is not yet known for the classical case of ℓ_∞ covering numbers (see [25]). Finally, we remark that the question of proving a dimension-free bound for the sequential ℓ_2 covering number in the spirit of [19] is still open.

We state one more result, with a proof similar to that of Theorem 5. It provides a bound on the 0-cover in terms of the $\mathrm{fat}_1(\mathcal{F})$ combinatorial parameter. Of particular interest is the case $k = 1$, when $\mathrm{fat}_1(\mathcal{F}) = \mathrm{Ldim}(\mathcal{F})$.

Theorem 7. *Let $\mathcal{F} \subseteq \{0, \dots, k\}^{\mathcal{Z}}$ be a class of functions with $\mathrm{fat}_1(\mathcal{F}, \mathcal{Z}) = d$. Then for any $n > d$,*

$$\mathcal{N}(0, \mathcal{F}, n) \le \sum_{i=0}^{d} \binom{n}{i} k^i \le \left(\frac{ekn}{d}\right)^d .$$

Furthermore, for $n \le d$,

$$\mathcal{N}(0, \mathcal{F}, n) \le (k+1)^n .$$

Consequently, the upper bound $\mathcal{N}(0, \mathcal{F}, n) \le (ekn)^d$ holds for any $n \ge 1$.

In addition to the above connections between the combinatorial dimensions and covering numbers, we can also relate the scale-sensitive dimension to sequential Rademacher averages. This will allow us to "close the loop" and show equivalence of the introduced complexity measures. The following lemma asserts that the fat-shattering dimensions at "large enough" scales cannot be too large and provides a lower bound for the Rademacher complexity.

Lemma 8. *Let $\mathcal{F} \subseteq [-1,1]^{\mathcal{Z}}$. For any $\beta > 2\mathfrak{R}_n(\mathcal{F})$, we have that $\mathrm{fat}_\beta(\mathcal{F}) < n$. Furthermore, for any $\beta > 0$, it holds that*

$$\min\left\{\mathrm{fat}_\beta(\mathcal{F}), n\right\} \le \frac{32\, n\, \mathfrak{R}_n(\mathcal{F})^2}{\beta^2} .$$

The following lemma complements Theorem 4.

Lemma 9. *For any function class $\mathcal{F} \subseteq [-1,1]^{\mathcal{Z}}$, we have that*

$$\mathfrak{D}_n^\infty(\mathcal{F}) \le 8\ \mathfrak{R}_n(\mathcal{F}) \left(1 + 4\sqrt{2}\ \log^{3/2}\left(en^2\right)\right)$$

as long as $\mathfrak{R}_n(\mathcal{F}) \ge 1/n$.

Theorems 2 and 4, together with Lemma 9, imply that the quantities $\mathfrak{D}_n^\infty(\mathcal{F})$, $\mathfrak{D}_n^2(\mathcal{F})$, $\mathfrak{R}_n(\mathcal{F})$, and $\sup_{\mathbb{P}} \mathbb{E} \sup_{f \in \mathcal{F}} \mathbb{M}_n(f)$ are equivalent up to poly-logarithmic in n factors:

$$\begin{aligned} \frac{1}{2}\left(\mathfrak{R}_n(\mathcal{F}) - \frac{B}{2\sqrt{n}}\right) \le \sup_{\mathbb{P}} \mathbb{E} \sup_{f \in \mathcal{F}} \mathbb{M}_n(f) \le 2\mathfrak{R}_n(\mathcal{F}) \le 2\mathfrak{D}_n(\mathcal{F}) \\ \le 16\ \mathfrak{R}_n(\mathcal{F}) \left(1 + 4\sqrt{2}\ \log^{3/2}\left(en^2\right)\right) \end{aligned} \tag{10}$$

as long as $\mathfrak{R}_n(\mathcal{F}) \ge 1/n$, with B defined as in Theorem 2. Additionally, the upper and lower bounds in terms of the fat-shattering dimension follow, respectively, from the integrated complexity bound and Corollary 6, and from Lemma 8.

At this point, we have introduced all the key notions of sequential complexity and showed fundamental connections between them. In the next section, we turn to the proof of Theorem 1.

7 Sequential Uniform Convergence

In order to prove Theorem 1, we will need to show in-probability (rather than in-expectation) versions of some of the earlier results. Luckily, the proof techniques are not significantly different. First, we prove Lemma 10, an in-probability version of the sequential symmetrization technique of Theorem 2. Let us use the shorthand $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid Z_1, \ldots, Z_t]$.

Lemma 10. *Let $\mathcal{F} \subseteq [-1,1]^{\mathcal{Z}}$. For any $\alpha > 0$, it holds that*

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{t=1}^{n}\left(f(Z_t) - \mathbb{E}_{t-1}\left[f(Z_t)\right]\right)\right| > \alpha\right) \le 4\sup_{\mathbf{z}} \ \mathbb{P}_\epsilon\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{t=1}^{n}\epsilon_t f(\mathbf{z}_t(\epsilon))\right| > \alpha/4\right).$$

The next result is an analogue of Eq. (9).

Lemma 11. *Let $\mathcal{F} \subseteq [-1,1]^{\mathcal{Z}}$. For any $\alpha > 0$, we have that*

$$\mathbb{P}_\epsilon\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{t=1}^{n}\epsilon_t f(\mathbf{z}_t(\epsilon))\right| > \alpha/4\right) \le 2\mathcal{N}_1(\alpha/8, \mathcal{F}, \mathbf{z})e^{-n\alpha^2/128} \le 2\left(\frac{16en}{\alpha}\right)^{\mathrm{fat}_{\alpha/8}} e^{-n\alpha^2/128}$$

for any $\mathcal{Z}$-valued tree $\mathbf{z}$ of depth n.

Proof of Theorem 1. Let E_n^α denote the event

$$\frac{1}{n}\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{n}\left(f(Z_t) - \mathbb{E}_{t-1}[f(Z_t)]\right)\right| > \alpha.$$

Combining Lemma 10 and Lemma 11, for any distribution,

$$\mathbb{P}\left(E_n^\alpha\right) \le 8\left(\frac{16en}{\alpha}\right)^{\mathrm{fat}_{\alpha/8}} e^{-n\alpha^2/128}.$$

We have for a fixed n',

$$\mathbb{P}\left(\sup_{n\ge n'}\sup_{f\in\mathcal{F}}|\mathbb{M}_n(f)| > \alpha\right) \le \sum_{n\ge n'}\mathbb{P}\left(E_n^\alpha\right) \le \sum_{n\ge n'} 8\left(\frac{16en}{\alpha}\right)^{\mathrm{fat}_{\alpha/8}} e^{-n\alpha^2/128}.$$

Since the last sum does not depend on $\mathbb{P}$, we may take the supremum over $\mathbb{P}$ and then let $n' \to \infty$ to conclude that, if $\mathrm{fat}_{\alpha/8}$ is finite then

$$\limsup_{n'\to\infty}\sup_{\mathbb{P}} \ \mathbb{P}\left(\sup_{n\ge n'}\sup_{f\in\mathcal{F}}|\mathbb{M}_n(f)| > \alpha\right) \le \limsup_{n'\to\infty}\sum_{n\ge n'} 8\left(\frac{16en}{\alpha}\right)^{\mathrm{fat}_{\alpha/8}} e^{-n\alpha^2/128} = 0.$$

Therefore, if fat_α is finite for all $\alpha > 0$ then $\mathcal{F}$ satisfies sequential uniform convergence. This proves $2 \Rightarrow 1$.

Next, notice that if $\mathfrak{R}_n(\mathcal{F}) \to 0$ then for any $\alpha > 0$ there exists $n_\alpha < \infty$ such that $\alpha > 2\mathfrak{R}_{n_\alpha}(\mathcal{F})$. Therefore, by Lemma 8 we have that $\mathrm{fat}_\alpha(\mathcal{F}) < n_\alpha < \infty$. We can, therefore, conclude that $3 \Rightarrow 2$. Now, to see that $1 \Rightarrow 3$ notice that by Theorem 2,

$$\sup_{\mathbb{P}} \mathbb{E} \sup_{f\in\mathcal{F}} |\mathbb{M}_n(f)| \geq \frac{1}{2}\left(\mathfrak{R}_n(\mathcal{F}) - \frac{B}{2\sqrt{n}}\right)$$

and so $\lim_{n\to\infty} \sup_{\mathbb{P}} \mathbb{E} \sup_{f\in\mathcal{F}} |\mathbb{M}_n(f)| = 0$ implies that $\mathfrak{R}_n(\mathcal{F}) \to 0$. Since almost sure convergence implies convergence in expectation, we conclude $1 \Rightarrow 3$.

□

The final result of this section is a stronger version of Lemma 11, showing that sequential Rademacher complexity is, in some sense, the "right" complexity measure even when one considers high probability statements. This lemma will be used in Section 9.

Lemma 12. *Let $\mathcal{F} \subseteq [-1,1]^{\mathcal{Z}}$. Suppose $\mathrm{fat}_\alpha(\mathcal{F})$ is finite for all $\alpha > 0$ and that the following mild assumptions hold: $\mathfrak{R}_n(\mathcal{F}) \geq 1/n$, $\mathcal{N}_\infty(2^{-1}, \mathcal{F}, n) \geq 4$, and there exists a constant L such that $L > \sum_{j=1}^{\infty} \mathcal{N}_\infty(2^{-j}, \mathcal{F}, n)^{-1}$. Then for any $\theta > \sqrt{12/n}$, for any $\mathcal{Z}$-valued tree $\mathbf{z}$ of depth n,*

$$\begin{aligned}
&\mathbb{P}_\epsilon\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{t=1}^{n}\epsilon_t f(\mathbf{z}_t(\epsilon))\right| > 8\left(1+\theta\sqrt{8n\log^3(en^2)}\right)\cdot\mathfrak{R}_n(\mathcal{F})\right)\\
&\leq \mathbb{P}_\epsilon\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{t=1}^{n}\epsilon_t f(\mathbf{z}_t(\epsilon))\right| > \inf_{\alpha>0}\left\{4\alpha + 6\theta\int_\alpha^1\sqrt{\log\mathcal{N}_\infty(\delta,\mathcal{F},n)}d\delta\right\}\right)\\
&\leq 2Le^{-\frac{n\theta^2}{4}}\ .
\end{aligned}$$

We established that sequential Rademacher complexity, as well as the sequential versions of covering numbers and fat-shattering dimensions, provide necessary and sufficient conditions for sequential uniform convergence. Let us now make a connection to the results of [33] who studied the notion of *generalized entropy with bracketing.* In [33], this complexity measure was shown to provide upper bounds on uniform deviations of martingale difference sequences for a given distribution $\mathbb{P}$. We do not know whether having a "small" generalized entropy with bracketing (more precisely, decay of the Dudley integral with generalized bracketing entropy) with respect to *all* distributions is also a necessary condition for sequential uniform convergence. Nevertheless, using the results of this work we can show that the generalized entropy with bracketing can be used as a tool to establish sequential uniform convergence for *tree processes.* Specifically, by Theorem 1, to establish this convergence it is enough to show uniform convergence for the tree process $\sup_{f\in\mathcal{F}} \mathbb{T}_n^{(\mathbf{z})}(f)$ for any $\mathcal{Z}$-valued tree $\mathbf{z}$. It is therefore sufficient to only consider the generalized entropy with bracketing for tree processes. For a tree process on any $\mathbf{z}$, however, the notion of the generalized entropy coincides with the notion of a sequential cover in the ℓ_∞ sense. Indeed, the brackets for the tree process are pairs of real valued trees. By taking the center of each bracket, one obtains a sequential cover; conversely, by using a covering tree as a center one obtains a bracket. We conclude that convergence of the Dudley-type integral with the generalized entropy with bracketing for all tree processes is a necessary and sufficient condition for sequential uniform convergence.

We end this section by mentioning that measurability of $\sup_{f\in\mathcal{F}} \mathbb{M}_n(f)$ can be ensured with some regularity conditions on $\mathcal{Z}$ and $\mathcal{F}$. For instance, we may assume that $\mathcal{F}$ is a class of uniformly bounded measurable functions that is image admissible Suslin (there is a map Γ from a Polish space $\mathcal{Y}$ to $\mathcal{F}$ such that the composition of Γ with the evaluation map $(y,z) \mapsto \Gamma(y)(z)$ is jointly measurable). In this case, it is easy to check that $\sup_{f\in\mathcal{F}} \mathbb{M}_n(f)$ is indeed measurable (see, for instance, Corollary 5.3.5 in [11]).

8 Structural Results

Being able to bound the complexity of a function class by a complexity of a simpler class is of great utility for proving bounds. In empirical process theory, such structural results are obtained through properties of Rademacher averages [18, 7]. In particular, the contraction inequality due to Ledoux and Talagrand [16, Corollary 3.17], allows one to pass from a composition of a Lipschitz function with a class to the function class itself. This wonderful property permits easy convergence proofs for a vast array of problems.

We show that the notion of sequential Rademacher complexity also enjoys many of the same properties. In particular, the following is a sequential analogue of the Ledoux-Talagrand contraction inequality [16, 7].

Lemma 13. *Fix a class $\mathcal{F} \subseteq [-1,1]^{\mathcal{Z}}$ with $\mathfrak{R}_n(\mathcal{F}) \geq 1/n$. Let $\phi : \mathbb{R} \mapsto \mathbb{R}$ be a Lipschitz function with a constant L. Then*

$$\mathfrak{R}_n(\phi \circ \mathcal{F}) \leq 8\,L\left(1 + 4\sqrt{2}\log^{3/2}(en^2)\right) \cdot \mathfrak{R}_n(\mathcal{F}) \ .$$

In comparison to the classical result, here we get an extra logarithmic factor in n. Whether the result without this logarithmic factor can be proved for the tree process remains an open question.

In the next proposition, we summarize some other useful properties of sequential Rademacher complexity (see [18, 7] for the results in the i.i.d. setting).

Proposition 14. *Sequential Rademacher complexity satisfies the following properties. For any $\mathcal{Z}$-valued tree $\mathbf{z}$ of depth n:*

1. *If $\mathcal{F} \subseteq \mathcal{G}$, then $\mathfrak{R}_n(\mathcal{F},\mathbf{z}) \leq \mathfrak{R}_n(\mathcal{G},\mathbf{z})$.*
2. *$\mathfrak{R}_n(\mathcal{F},\mathbf{z}) = \mathfrak{R}_n(\mathrm{conv}(\mathcal{F}),\mathbf{z})$.*
3. *$\mathfrak{R}_n(c\mathcal{F},\mathbf{z}) = |c|\mathfrak{R}_n(\mathcal{F},\mathbf{z})$ for all $c \in \mathbb{R}$.*
4. *For any h, $\mathfrak{R}_n(\mathcal{F}+h,\mathbf{z}) = \mathfrak{R}_n(\mathcal{F},\mathbf{z})$ where $\mathcal{F}+h = \{f+h : f \in \mathcal{F}\}$.*

The structural results developed in this section are crucial for analyzing sequential prediction problems, a topic we explore in a separate paper.

9 Application: Concentration of Martingales in Banach Spaces

As a consequence of the uniform convergence results, one can obtain concentration inequalities for martingale difference sequences in Banach spaces. Before we provide the concentration inequality, we first state a rather straightforward lemma that follows from Lemmas 10 and 12.

Lemma 15. *For $\mathcal{F} \subseteq [-1,1]^{\mathcal{Z}}$, for $n \geq 2$ and any $\alpha > 0$, we have that*

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{t=1}^{n} f(Z_t) - \mathbb{E}_{t-1}\left[f(Z_t)\right]\right| > \alpha\right) \leq 8L\exp\left(-\frac{\alpha^2}{c\log^3(n)\mathfrak{R}_n^2(\mathcal{F})}\right)$$

under the mild assumptions $\mathfrak{R}_n(\mathcal{F}) \geq 1/n$ and $\mathcal{N}_\infty(2^{-1},\mathcal{F},n) \geq 4$. Here c is an absolute constant and $L > e^4$ is such that $L > \sum_{j=1}^{\infty}\mathcal{N}_\infty(2^{-j},\mathcal{F},n)^{-1}$.

Let us now consider the case of a unit ball in a Banach space and discuss the conditions under which the main result of this section (Corollary 17 below) is stated. Let $\mathcal{Z}$ be the unit ball of a Banach space with norm $\|\cdot\|$ and consider the class $\mathcal{F}$ of continuous linear mappings $z \mapsto \langle f, z\rangle$ with $\|f\|_* \leq 1$, where $\|\cdot\|_*$ is the dual to the norm $\|\cdot\|$. By definition of the norm,

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{t=1}^{n} f(Z_t) - \mathbb{E}_{t-1}\left[f(Z_t)\right]\right| = \left\|\frac{1}{n}\sum_{t=1}^{n} Z_t - \mathbb{E}_{t-1}\left[Z_t\right]\right\|$$

and

$$\mathfrak{R}_n(\mathcal{F}) = \sup_{\mathbf{z}}\mathbb{E}\left\|\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\mathbf{z}_t(\epsilon)\right\| \ .$$

Further note that for any linear class and any $\gamma > 0$, $\mathcal{N}_\infty(\gamma,\mathcal{F},n) \geq 1/\gamma$ and so $\sum_{j=1}^{\infty}\mathcal{N}_\infty(2^{-j},\mathcal{F},n)^{-1} \leq 2$. In view of Lemma 15, under the mild condition that $\mathcal{N}_\infty(2^{-1},\mathcal{F},n) \geq 4$,

$$\forall\alpha > 0, \qquad \mathbb{P}\left(\left\|\frac{1}{n}\sum_{t=1}^{n} Z_t - \mathbb{E}_{t-1}\left[Z_t\right]\right\| > \alpha\right) \leq C\ \exp\left(-\frac{\alpha^2}{c\log^3(n)\mathfrak{R}_n^2(\mathcal{F})}\right)$$

for absolute constants $c, C > 0$. It remains to provide an upper bound on the sequential Rademacher complexity $\mathfrak{R}_n(\mathcal{F})$. To this end, recall that a function $\Psi : \mathcal{F} \to \mathbb{R}$ is (σ,q)-uniformly convex (for $q \in [2,\infty)$) with respect to a norm $\|\cdot\|_*$ if, for all $\theta \in [0,1]$ and $f_1, f_2 \in \mathcal{F}$,

$$\Psi(\theta f_1 + (1-\theta)f_2) \leq \theta\Psi(f_1) + (1-\theta)\Psi(f_2) - \frac{\sigma\,\theta\,(1-\theta)}{q}\|f_1 - f_2\|_*^q \ .$$

Proposition 16. *Suppose that Ψ is (σ,q)-uniformly convex with respect to a given norm $\|\cdot\|_*$ on $\mathcal{F}$ and $0 \leq \Psi(f) \leq \Psi_{\max}$ for all $f \in \mathcal{F}$. Then we have*

$$\mathfrak{R}_n(\mathcal{F}) \leq C_p\left(\frac{\Psi_{\max}^{p-1}}{\sigma\, n^{p-1}}\right)^{1/p},$$

where $p > 1$ is such that $1/p + 1/q = 1$, and $C_p = (p/(p-1))^{\frac{p-1}{p}}$.

The following is the main result of this section:

Corollary 17. *Given any function Ψ that is (σ,q)-uniformly convex with respect to $\|\cdot\|_*$ such that $0 \leq \Psi(f) \leq \Psi_{\max}$ for all $f \in \mathcal{F}$, and given and any $\alpha > 0$ we have*

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{t=1}^{n} Z_t - \mathbb{E}_{t-1}\left[Z_t\right]\right\| > \alpha\right) \leq C\ \exp\left(-\frac{n^{2/q}\sigma^{2/p}\alpha^2}{c_p\ \Psi_{\max}^{2/q}\ \log^3(n)}\right)$$

where $\frac{1}{q} + \frac{1}{p} = 1$, $q \in [2,\infty)$, and $n \geq 2$. Here, C is an absolute constant, and c_p only depends on p.

We suspect that the $\log^3(n)$ term in the above bound is an artifact of the proof technique and can probably be avoided. For instance, when the norm is equivalent to a 2-smooth norm (that is, $p = 2$), Pinelis [21] shows concentration of martingale difference sequences in the Banach space without the extra $\log^3(n)$ term and with better constants. However, his argument is specific to the 2-smooth case. As a rather direct consequence of the uniform convergence results for dependent random variables, we are able to provide concentration of martingales in Banach spaces for general norms.

Let $(W_t)_{t\geq 1}$ be a martingale difference sequence in a Banach space such that for any t, $\|W_t\| \leq 1$. The celebrated result of Pisier [22] states that

$$\mathbb{E}\left\|\frac{1}{n}\sum_{t=1}^{n} W_t\right\| \to 0 \tag{11}$$

if and only if the Banach space can be given an equivalent p-smooth norm for some $p > 1$. Using duality this can equivalently be restated as: (11) holds for any martingale difference sequence $(W_t)_{t\geq 1}$ with $\|W_t\| \leq 1$ if and only if we can find a function $\Psi : \mathcal{F} \mapsto \mathbb{R}$ which is $(1, q)$ uniformly convex for some $q < \infty$ with respect to norm $\|\cdot\|_*$ (the dual norm) and is such that $\Psi_{\max} \leq C < \infty$. Furthermore, it can be shown that the rate of convergence of expected norm of the martingale is tightly governed by the smallest such q one can find. Hence, combined with Corollary 17 we conclude that whenever expected norm of martingales in Banach spaces converge, exponential bounds for martingales in the Banach space, like the one in Corollary 17, also have to hold.

Acknowledgements

We would like to thank J. Michael Steele, Dean Foster, and Ramon van Handel for helpful discussions. We gratefully acknowledge the support of NSF under grants CAREER DMS-0954737 and CCF-1116928.

A Proofs

Throughout, $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ denotes a sequence of n independent Rademacher random variables. Let $\mathbb{E}_{\epsilon_t}$ stand for the expectation $\frac{1}{2}\sum_{\epsilon_t \in \{\pm 1\}}$ over the random variable ϵ_t, and let $\mathbb{E}_\epsilon$ stand for the corresponding expectation over $\epsilon_1, \ldots, \epsilon_n$.

We first prove the following technical lemma.

Lemma 18. *Let $(Z_1, \ldots, Z_n) \in \mathcal{Z}^n$ be a sequence of random variables and let $(Z'_1, \ldots, Z'_n)$ be a decoupled tangent sequence. Let $\phi : \mathbb{R} \mapsto \mathbb{R}$ be a measurable function. We then have*

$$\mathbb{E}\left[\phi\left(\sup_{f\in\mathcal{F}}\sum_{t=1}^{n}(f(Z'_t) - f(Z_t))\right)\right] \quad \leq \quad \sup_{z_1, z'_1 \in \mathcal{Z}} \mathbb{E}_{\epsilon_1} \ldots \sup_{z_n, z'_n \in \mathcal{Z}} \mathbb{E}_{\epsilon_n}\left[\phi\left(\sup_{f\in\mathcal{F}}\sum_{t=1}^{n}\epsilon_t(f(z'_t) - f(z_t))\right)\right]$$

where $\epsilon_1, \ldots, \epsilon_n$ are independent Rademacher random variables. The inequality also holds when an absolute value of the sum is introduced on both sides.

Proof of Lemma 18. Let P be the joint distribution of the two sequences. We start by noting that since Z_n, Z'_n are conditionally independent and identically distributed,

$$\begin{aligned}
&\mathbb{E}\left[\phi\left(\sup_{f\in\mathcal{F}}\sum_{t=1}^{n} f(Z'_t) - f(Z_t)\right) \;\middle|\; Z_{1:n-1} = z_{1:n-1}, Z'_{1:n-1} = z'_{1:n-1}\right] \\
&= \int \phi\left(\sup_{f\in\mathcal{F}}\sum_{t=1}^{n-1}(f(z'_t) - f(z_t)) + (f(z'_n) - f(z_n))\right) dP(z_n, z'_n | z_{1:n-1}, z'_{1:n-1}) \\
&= \int \phi\left(\sup_{f\in\mathcal{F}}\sum_{t=1}^{n-1}(f(z'_t) - f(z_t)) - (f(z'_n) - f(z_n))\right) dP(z_n, z'_n | z_{1:n-1}, z'_{1:n-1})
\end{aligned}$$

for any $z_1, \dots, z_{n-1}, z'_1, \dots, z_{n-1} \in \mathcal{Z}$. The notation $Z_{1:n-1} = z_{1:n-1}$ is a shorthand for the event $\{Z_1 = z_1, \dots, Z_{n-1} = z_{n-1}\}$.

Since the last two lines are equal, they are both equal to their average and hence

$$\begin{aligned}
&\mathbb{E}\left[\phi\left(\sup_{f\in\mathcal{F}}\sum_{t=1}^{n} f(Z'_t) - f(Z_t)\right) \;\middle|\; Z_{1:n-1} = z_{1:n-1}, Z'_{1:n-1} = z'_{1:n-1}\right] \\
&= \mathbb{E}_{\epsilon_n}\int \phi\left(\sup_{f\in\mathcal{F}}\sum_{t=1}^{n-1}(f(z'_t) - f(z_t)) + \epsilon_n(f(z'_n) - f(z_n))\right) dP(z_n, z'_n | z_{1:n-1}, z'_{1:n-1}) \\
&= \int \mathbb{E}_{\epsilon_n}\phi\left(\sup_{f\in\mathcal{F}}\sum_{t=1}^{n-1}(f(z'_t) - f(z_t)) + \epsilon_n(f(z'_n) - f(z_n))\right) dP(z_n, z'_n | z_{1:n-1}, z'_{1:n-1}) \\
&\le \sup_{z_n, z'_n \in \mathcal{Z}} \mathbb{E}_{\epsilon_n}\phi\left(\sup_{f\in\mathcal{F}}\sum_{t=1}^{n-1}(f(z'_t) - f(z_t)) + \epsilon_n(f(z'_n) - f(z_n))\right) .
\end{aligned}$$

Repeating this calculation for step $n-1$ and using the inequality above,

$$\begin{aligned}
&\mathbb{E}\left[\phi\left(\sup_{f\in\mathcal{F}}\sum_{t=1}^{n} f(Z'_t) - f(Z_t)\right) \;\middle|\; Z_{1:n-2} = z_{1:n-2}, Z'_{1:n-2} = z'_{1:n-2}\right] \\
&= \int \mathbb{E}\left[\phi\left(\sup_{f\in\mathcal{F}}\sum_{t=1}^{n-1}(f(z'_t) - f(z_t)) + (f(Z'_n) - f(Z_n))\right) \;\middle|\; Z_{1:n-1} = z_{1:n-1}, Z'_{1:n-1} = z'_{1:n-1}\right] \\
&\qquad\qquad\qquad \times dP(z_{n-1}, z'_{n-1} | z_{1:n-2}, z'_{1:n-2}) \\
&\le \int \left[\sup_{z_n, z'_n} \mathbb{E}_{\epsilon_n}\left[\phi\left(\sup_{f\in\mathcal{F}}\sum_{t=1}^{n-1}(f(z'_t) - f(z_t)) + \epsilon_n(f(z'_n) - f(z_n))\right)\right]\right] \\
&\qquad\qquad\qquad \times dP(z_{n-1}, z'_{n-1} | z_{1:n-2}, z'_{1:n-2}) \\
&\le \sup_{z_{n-1}, z'_{n-1}} \mathbb{E}_{\epsilon_{n-1}} \sup_{z_n, z'_n} \mathbb{E}_{\epsilon_n}\left[\phi\left(\sup_{f\in\mathcal{F}}\sum_{t=1}^{n-2}(f(z'_t) - f(z_t)) + \sum_{s=n-1}^{n} \epsilon_s(f(z'_s) - f(z_s))\right)\right].
\end{aligned}$$

Proceeding in this fashion yields the main statement. The exact same argument shows that the statement holds with absolute value around the sum. □

Proof of Theorem 2. By the definition of $\mathbb{M}_n(f)$ and convexity of the supremum,

$$n \cdot \mathbb{E} \sup_{f\in\mathcal{F}} \mathbb{M}_n(f) = \mathbb{E} \sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \Big(\mathbb{E}[f(Z'_t)|\mathcal{A}_{t-1}] - f(Z_t) \Big) \tag{12}$$

$$\leq \mathbb{E} \sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \Big(f(Z'_t) - f(Z_t) \Big) \tag{13}$$

where $\{Z'_t\}$ is a decoupled sequence tangent to $\{Z_t\}$. That is, Z'_t and Z_t are (conditionally) independent and distributed identically given $Z_1, \ldots, Z_{t-1}$. Appealing to Lemma 18 with ϕ being the identity function,

$$\mathbb{E} \sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \Big(f(Z'_t) - f(Z_t) \Big) \leq \sup_{z_1, z'_1} \mathbb{E}_{\epsilon_1} \sup_{z_2, z'_2} \mathbb{E}_{\epsilon_2} \ldots \sup_{z_n, z'_n} \mathbb{E}_{\epsilon_n} \sup_{f\in\mathcal{F}} \left\{ \sum_{t=1}^{n} \epsilon_t \Big(f(z'_t) - f(z_t) \Big) \right\}$$

$$\leq 2 \sup_{z_1} \mathbb{E}_{\epsilon_1} \sup_{z_2} \mathbb{E}_{\epsilon_2} \ldots \sup_{z_n} \mathbb{E}_{\epsilon_n} \sup_{f\in\mathcal{F}} \left\{ \sum_{t=1}^{n} \epsilon_t f(z_t) \right\}. \tag{14}$$

We claim that the right-hand side of (14) is a tree process for the worst-case tree. Indeed, the first supremum is achieved at some $z^*_1 \in \mathcal{Z}$. The second supremum is achieved at $z^*_2(+1)$ if $\epsilon_1 = +1$ and at some potentially different value $z^*_2(-1)$ if $\epsilon_1 = -1$. In the case the suprema are not achieved, a simple limiting argument can be employed. Proceeding in this manner we get that

$$\sup_{z_1} \mathbb{E}_{\epsilon_1} \sup_{z_2} \mathbb{E}_{\epsilon_2} \ldots \sup_{z_n} \mathbb{E}_{\epsilon_n} \sup_{f\in\mathcal{F}} \left\{ \sum_{t=1}^{n} \epsilon_t f(z_t) \right\} = \mathbb{E}_{\epsilon_1, \ldots, \epsilon_n} \sup_{f\in\mathcal{F}} \left\{ \sum_{t=1}^{n} \epsilon_t f(\mathbf{z}^*_t(\epsilon_1, \ldots, \epsilon_{t-1})) \right\} \tag{15}$$

$$\leq n \sup_{\mathbf{z}} \mathbb{E} \sup_{f\in\mathcal{F}} \mathbb{T}^{(\mathbf{z})}_n(f).$$

The other direction also trivially holds: for any $\mathbf{z}$,

$$\sup_{z_1} \mathbb{E}_{\epsilon_1} \sup_{z_2} \mathbb{E}_{\epsilon_2} \ldots \sup_{z_n} \mathbb{E}_{\epsilon_n} \sup_{f\in\mathcal{F}} \left\{ \sum_{t=1}^{n} \epsilon_t f(z_t) \right\} \geq \mathbb{E}_{\epsilon_1, \ldots, \epsilon_n} \sup_{f\in\mathcal{F}} \left\{ \sum_{t=1}^{n} \epsilon_t f(\mathbf{z}_t(\epsilon_1, \ldots, \epsilon_{t-1})) \right\}.$$

This proves the upper bound (5) in the theorem statement. We now turn to the lower bound, (6). Fix some $z \in \mathcal{Z}$. Let ϵ_1 be a Rademacher random variable and define the marginal distribution of Z_1 to be $P(Z_1 = z_1|\epsilon_1 = -1) = 1$ and $P(Z_1 = z|\epsilon_1 = 1) = 1$ for some $z_1 \in \mathcal{Z}$. Next, conditionally on ϵ_1, let the distribution of Z_2 be given by $P(Z_2 = z_2|\epsilon_1, \epsilon_2 = -1) = 1$ and $P(Z_2 = z|\epsilon_1, \epsilon_2 = 1) = 1$ where $z_2 \in \mathcal{Z}$ is measurable w.r.t. $\mathcal{A}_1 = \sigma(\epsilon_1)$ and ϵ_2 is an independent Rademacher random variable. Proceeding in similar fashion we provide conditional distribution of Z_t as $P(Z_t = z_t|\epsilon_{1:t-1}, \epsilon_t = -1) = 1$ and $P(Z_t = z|\epsilon_{1:t-1}, \epsilon_t = 1) = 1$ where each $z_t \in \mathcal{Z}$ is measurable w.r.t. $\mathcal{A}_{t-1} = \sigma(\epsilon_{1:t-1})$. Further, by construction, $P(Z_t = z_t|\epsilon_{1:t-1}) = P(Z_t = z|\epsilon_{1:t-1}) = \frac{1}{2}$. Since our choice of $z_t \in \mathcal{Z}$ that is measurable w.r.t. $\mathcal{A}_{t-1}$ can be arbitrary, we conclude that

$$\begin{aligned}
\sup_{\mathbb{P}} \mathbb{E} \sup_{f\in\mathcal{F}} \mathbb{M}_n(f) &= \sup_{\mathbb{P}} \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^{n}\Big(\mathbb{E}[f(Z_t)|\mathcal{A}_{t-1}] - f(Z_t)\Big)\right] \\
&\geq \sup_{z\in\mathcal{Z}} \sup_{z_1\in\mathcal{Z}} \mathbb{E}_{\epsilon_1} \ldots \sup_{z_n\in\mathcal{Z}} \mathbb{E}_{\epsilon_n}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^{n}\Big(\frac{f(z_t)+f(z)}{2} - \tfrac{(1-\epsilon_t)}{2} f(z_t) - \tfrac{(1+\epsilon_t)}{2} f(z)\Big)\right] \\
&= \sup_{z\in\mathcal{Z}} \sup_{z_1\in\mathcal{Z}} \mathbb{E}_{\epsilon_1} \ldots \sup_{z_n\in\mathcal{Z}} \mathbb{E}_{\epsilon_n}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^{n}\epsilon_t\Big(\frac{f(z_t)-f(z)}{2}\Big)\right] \\
&\geq \sup_{z\in\mathcal{Z}} \sup_{z_1\in\mathcal{Z}} \mathbb{E}_{\epsilon_1} \ldots \sup_{z_n\in\mathcal{Z}} \mathbb{E}_{\epsilon_n}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^{n}\frac{\epsilon_t f(z_t)}{2} - \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^{n}\frac{\epsilon_t f(z)}{2}\right] \\
&= \frac{1}{2}\left(\sup_{z_1}\mathbb{E}_{\epsilon_1}\ldots\sup_{z_n}\mathbb{E}_{\epsilon_n}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\epsilon_t f(z_t)\right] + \sup_{z}\mathbb{E}_{\epsilon}\left[-\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\epsilon_t f(z)\right]\right) \\
&= \frac{1}{2}\left(\sup_{z_1}\mathbb{E}_{\epsilon_1}\ldots\sup_{z_n}\mathbb{E}_{\epsilon_n}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\epsilon_t f(z_t)\right] - \inf_{z}\mathbb{E}_{\epsilon}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\epsilon_t f(z)\right]\right) \\
&= \frac{1}{2}\left(\sup_{z_1}\mathbb{E}_{\epsilon_1}\ldots\sup_{z_n}\mathbb{E}_{\epsilon_n}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\epsilon_t f(z_t)\right] - \inf_{z}\left(\sup_{f\in\mathcal{F}} f(z) - \inf_{f\in\mathcal{F}} f(z)\right)\frac{1}{2}\mathbb{E}_{\epsilon}\left|\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\right|\right) \\
&\geq \frac{1}{2}\left(\mathfrak{R}_n(\mathcal{F}) - \left(\inf_{z}\sup_{f,f'\in\mathcal{F}}\big(f(z)-f'(z)\big)\right)\frac{1}{2\sqrt{n}}\right)
\end{aligned}$$

where the last step holds because $\mathbb{E}_\epsilon\left|\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\right| \leq \sqrt{\frac{1}{n}}$.

Before we conclude the proof we would like to point out that our assumption that $\mathcal{F}$ is Suslin admissible and that $\mathcal{Z}$ is a separable metric space ensures measurability of $\sup_{f\in\mathcal{F}}\mathbb{M}_n(f)$ and other terms in the proof. We briefly sketch the reasoning now. By Suslin admissibility of $\mathcal{F}$ and by [13, Corollary 10.2.3] (with the conditional probability measure and since $\mathcal{Z}$ is assumed to be separable), for each $t \in [n]$, $\mathbb{E}_{t-1}f - f(Z_t)$ is measurable and hence is $\sum_{t=1}^{n}\Big(\mathbb{E}_{t-1}f - f(Z_t)\Big)$. By Corollary 5.16 of [11] we can conclude that $\sup_{f\in\mathcal{F}}\mathbb{M}_n(f)$ are also uniformly measurable. On similar lines as the reasoning sketched above, measurability of other quantities appearing in the proof can be concluded. The sketch above is on similar lines as proof of [13, Theorem 10.3.2] and for more details on measurability issues we encourage readers to refer to [13] and [11]. □

Proof of Lemma 3. For any $\lambda > 0$, we invoke Jensen's inequality to get

$$\begin{aligned}
M(\lambda) := \exp\left\{\lambda\mathbb{E}_\epsilon\left[\max_{\mathbf{v}\in V}\sum_{t=1}^{n}\epsilon_t\mathbf{v}_t(\epsilon)\right]\right\} &\leq \mathbb{E}_\epsilon\left[\exp\left\{\lambda\max_{\mathbf{v}\in V}\sum_{t=1}^{n}\epsilon_t\mathbf{v}_t(\epsilon)\right\}\right] \\
&= \mathbb{E}_\epsilon\left[\max_{\mathbf{v}\in V}\exp\left\{\lambda\sum_{t=1}^{n}\epsilon_t\mathbf{v}_t(\epsilon)\right\}\right] \leq \sum_{\mathbf{v}\in V}\mathbb{E}_\epsilon\left[\exp\left\{\lambda\sum_{t=1}^{n}\epsilon_t\mathbf{v}_t(\epsilon)\right\}\right].
\end{aligned}$$

Fix a $\mathbf{v}\in V$. For $t \in \{0,\ldots,n-1\}$ define a function $A^t : \{\pm1\}^t \mapsto \mathbb{R}$ by

$$A^t(\epsilon_1,\ldots,\epsilon_t) = \max_{\epsilon_{t+1},\ldots,\epsilon_n}\exp\left\{\frac{\lambda^2}{2}\sum_{s=t+1}^{n}\mathbf{v}_s(\epsilon_{1:s-1})^2\right\}$$

and $A^n(\epsilon_1, \ldots, \epsilon_n) = 1$. We have that for any $t \in \{1, \ldots, n\}$, for any $(\epsilon_1, \ldots, \epsilon_{t-1}) \in \{\pm1\}^{t-1}$,

$$\begin{aligned}
&\mathbb{E}_{\epsilon_t}\left[\exp\left(\lambda \sum_{s=1}^{t} \epsilon_s \mathbf{v}_s(\epsilon_{1:s-1})\right) \times A^t(\epsilon_1, \ldots, \epsilon_t)\right] \\
&= \exp\left(\lambda \sum_{s=1}^{t-1} \epsilon_s \mathbf{v}_s(\epsilon_{1:s-1})\right) \times \left(\frac{1}{2} e^{\lambda \mathbf{v}_t(\epsilon_{1:t-1})} A^t(\epsilon_1, \ldots, \epsilon_{t-1}, +1) + \frac{1}{2} e^{-\lambda \mathbf{v}_t(\epsilon_{1:t-1})} A^t(\epsilon_1, \ldots, \epsilon_{t-1}, -1)\right) \\
&\leq \exp\left(\lambda \sum_{s=1}^{t-1} \epsilon_s \mathbf{v}_s(\epsilon_{1:s-1})\right) \times \max_{\epsilon_t \in \{\pm1\}} A^t(\epsilon_1, \ldots, \epsilon_t) \left(\frac{1}{2} e^{\lambda \mathbf{v}_t(\epsilon_{1:t-1})} + \frac{1}{2} e^{-\lambda \mathbf{v}_t(\epsilon_{1:t-1})}\right) \\
&\leq \exp\left(\lambda \sum_{s=1}^{t-1} \epsilon_s \mathbf{v}_s(\epsilon_{1:s-1})\right) \times A^{t-1}(\epsilon_1, \ldots, \epsilon_{t-1})
\end{aligned}$$

where in the last step we used the inequality $(e^a + e^{-a})/2 \leq e^{a^2/2}$. Hence,

$$\mathbb{E}_{\epsilon_1, \ldots, \epsilon_n}\left\{\exp\left(\lambda \sum_{s=1}^{n} \epsilon_s \mathbf{v}_s(\epsilon_{1:s-1})\right)\right\} \leq A^0 = \max_{\epsilon_1, \ldots, \epsilon_n} \exp\left\{\frac{\lambda^2}{2} \sum_{s=1}^{n} \mathbf{v}_s(\epsilon_{1:s-1})^2\right\} .$$

We arrive at

$$M(\lambda) \leq \sum_{\mathbf{v} \in V} \exp\left\{\frac{\lambda^2}{2} \max_{\epsilon_1 \ldots \epsilon_{n-1} \in \{\pm1\}} \sum_{t=1}^{n} \mathbf{v}_t(\epsilon_{1:t-1})^2\right\} \leq |V| \exp\left\{\frac{\lambda^2}{2} \max_{\mathbf{v} \in V} \max_{\epsilon \in \{\pm1\}^n} \sum_{t=1}^{n} \mathbf{v}_t(\epsilon)^2\right\} .$$

Taking logarithms on both sides, dividing by λ and setting

$$\lambda = \sqrt{\frac{2 \log(|V|)}{\max_{\mathbf{v} \in V} \max_{\epsilon \in \{\pm1\}^n} \sum_{t=1}^{n} \mathbf{v}_t(\epsilon)^2}}$$

we conclude the proof. □

Proof of Theorem 4. Define $\beta_0 = 1$ and $\beta_j = 2^{-j}$. For a fixed tree $\mathbf{z}$ of depth n, let V_j be an β_j-cover with respect to ℓ_2. For any path $\epsilon \in \{\pm1\}^n$ and any $f \in \mathcal{F}$, let $\mathbf{v}[f, \epsilon]^j \in V_j$ be the element of the cover such that

$$\sqrt{\frac{1}{n} \sum_{t=1}^{n} \left(\mathbf{v}[f, \epsilon]_t^j(\epsilon) - f(\mathbf{z}_t(\epsilon))\right)^2} \leq \beta_j.$$

By the definition such a $\mathbf{v}[f, \epsilon]^j \in V_j$ exists, and we assume for simplicity this element is unique (ties can be broken in an arbitrary manner). Thus, $f \mapsto \mathbf{v}[f, \epsilon]^j$ is a well-defined mapping for any fixed ϵ and j. As before, $\mathbf{v}[f, \epsilon]_t^j$ denotes the t-th mapping of $\mathbf{v}[f, \epsilon]^j$. For any $t \in [n]$ and N to be chosen later, we have

$$f(\mathbf{z}_t(\epsilon)) = f(\mathbf{z}_t(\epsilon)) - \mathbf{v}[f, \epsilon]_t^N(\epsilon) + \sum_{j=1}^{N} (\mathbf{v}[f, \epsilon]_t^j(\epsilon) - \mathbf{v}[f, \epsilon]_t^{j-1}(\epsilon))$$

where $\mathbf{v}[f,\epsilon]^0_t(\epsilon) = 0$. For any $\epsilon = (\epsilon_1, \dots, \epsilon_n) \in \{\pm 1\}^n$,

$$\begin{aligned}
&\sup_{f\in\mathcal{F}} \left\{ \sum_{t=1}^n \epsilon_t f(\mathbf{z}_t(\epsilon)) \right\} \\
&= \sup_{f\in\mathcal{F}} \left\{ \sum_{t=1}^n \epsilon_t \left(f(\mathbf{z}_t(\epsilon)) - \mathbf{v}[f,\epsilon]^N_t(\epsilon) + \sum_{j=1}^N (\mathbf{v}[f,\epsilon]^j_t(\epsilon) - \mathbf{v}[f,\epsilon]^{j-1}_t(\epsilon)) \right) \right\} \\
&= \sup_{f\in\mathcal{F}} \left\{ \sum_{t=1}^n \epsilon_t \left(f(\mathbf{z}_t(\epsilon)) - \mathbf{v}[f,\epsilon]^N_t(\epsilon) \right) + \sum_{t=1}^n \epsilon_t \left(\sum_{j=1}^N (\mathbf{v}[f,\epsilon]^j_t(\epsilon) - \mathbf{v}[f,\epsilon]^{j-1}_t(\epsilon)) \right) \right\} \\
&\le \sup_{f\in\mathcal{F}} \left\{ \sum_{t=1}^n \epsilon_t \left(f(\mathbf{z}_t(\epsilon)) - \mathbf{v}[f,\epsilon]^N_t(\epsilon) \right) \right\} + \sup_{f\in\mathcal{F}} \left\{ \sum_{t=1}^n \epsilon_t \left(\sum_{j=1}^N (\mathbf{v}[f,\epsilon]^j_t(\epsilon) - \mathbf{v}[f,\epsilon]^{j-1}_t(\epsilon)) \right) \right\} . \qquad (16)
\end{aligned}$$

The first term above can be bounded via the Cauchy-Schwarz inequality as

$$\frac{1}{n} \mathbb{E} \sup_{f\in\mathcal{F}} \sum_{t=1}^n \epsilon_t \left(f(\mathbf{z}_t(\epsilon)) - \mathbf{v}[f,\epsilon]^N_t(\epsilon) \right) \le \mathbb{E} \sup_{f\in\mathcal{F}} \sum_{t=1}^n \frac{\epsilon_t}{\sqrt{n}} \frac{\left(f(\mathbf{z}_t(\epsilon)) - \mathbf{v}[f,\epsilon]^N_t(\epsilon) \right)}{\sqrt{n}} \le \beta_N \ .$$

The second term in (16) is bounded by considering successive refinements of the cover. The argument, however, is more delicate than in the classical case, as the trees $\mathbf{v}[f,\epsilon]^j$, $\mathbf{v}[f,\epsilon]^{j-1}$ depend on the particular path. Consider all possible pairs of $\mathbf{v}^s \in V_j$ and $\mathbf{v}^r \in V_{j-1}$, for $1 \le s \le |V_j|$, $1 \le r \le |V_{j-1}|$, where we assumed an arbitrary enumeration of elements. For each pair $(\mathbf{v}^s, \mathbf{v}^r)$, define a real-valued tree $\mathbf{w}^{(s,r)}$ by

$$\mathbf{w}^{(s,r)}_t(\epsilon) = \begin{cases} \mathbf{v}^s_t(\epsilon) - \mathbf{v}^r_t(\epsilon) & \text{if there exists } f \in \mathcal{F} \text{ s.t. } \mathbf{v}^s = \mathbf{v}[f,\epsilon]^j, \mathbf{v}^r = \mathbf{v}[f,\epsilon]^{j-1} \\ 0 & \text{otherwise.} \end{cases}$$

for all $t \in [n]$ and $\epsilon \in \{\pm 1\}^n$. It is crucial that $\mathbf{w}^{(s,r)}$ can be non-zero only on those paths ϵ for which $\mathbf{v}^s$ and $\mathbf{v}^r$ are indeed the members of the covers (at successive resolutions) close to $f(\mathbf{z}(\epsilon))$ (in the ℓ_2 sense) *for some* $f \in \mathcal{F}$. It is easy to see that $\mathbf{w}^{(s,r)}$ is well-defined. Let the set of trees W_j be defined as

$$W_j = \left\{ \mathbf{w}^{(s,r)} : 1 \le s \le |V_j|, 1 \le r \le |V_{j-1}| \right\} \ .$$

Now, the second term in (16) can be written as

$$\begin{aligned}
\sup_{f\in\mathcal{F}} \sum_{t=1}^n \epsilon_t \sum_{j=1}^N (\mathbf{v}[f,\epsilon]^j_t(\epsilon) - \mathbf{v}[f,\epsilon]^{j-1}_t(\epsilon)) &\le \sum_{j=1}^N \sup_{f\in\mathcal{F}} \sum_{t=1}^n \epsilon_t (\mathbf{v}[f,\epsilon]^j_t(\epsilon) - \mathbf{v}[f,\epsilon]^{j-1}_t(\epsilon)) \\
&\le \sum_{j=1}^N \max_{\mathbf{w}\in W_j} \sum_{t=1}^n \epsilon_t \mathbf{w}_t(\epsilon) \ .
\end{aligned}$$

The last inequality holds because for any $j \in [N]$, $\epsilon \in \{\pm 1\}^n$ and $f \in \mathcal{F}$ there is some $\mathbf{w}^{(s,r)} \in W_j$ with $\mathbf{v}[f,\epsilon]^j = \mathbf{v}^s$, $\mathbf{v}[f,\epsilon]^{j-1} = \mathbf{v}^r$ and

$$\mathbf{v}^s_t(\epsilon) - \mathbf{v}^r_t(\epsilon) = \mathbf{w}^{(s,r)}_t(\epsilon) \quad \forall t \le n \ .$$

Clearly, $|W_j| \le |V_j| \cdot |V_{j-1}|$. To invoke Lemma 3, it remains to bound the magnitude of all $\mathbf{w}^{(s,r)} \in W_j$ along all paths. For this purpose, fix $\mathbf{w}^{(s,r)}$ and a path ϵ. If there exists $f \in \mathcal{F}$ for which $\mathbf{v}^s = \mathbf{v}[f,\epsilon]^j$ and $\mathbf{v}^r = \mathbf{v}[f,\epsilon]^{j-1}$, then $\mathbf{w}_t^{(s,r)}(\epsilon) = \mathbf{v}[f,\epsilon]_t^j - \mathbf{v}[f,\epsilon]_t^{j-1}$ for any $t \in [n]$. By triangle inequality

$$\begin{aligned}\sqrt{\sum_{t=1}^n \mathbf{w}_t^{(s,r)}(\epsilon)^2} &\le \sqrt{\sum_{t=1}^n (\mathbf{v}[f,\epsilon]_t^j(\epsilon) - f(\mathbf{z}_t(\epsilon)))^2} + \sqrt{\sum_{t=1}^n (\mathbf{v}[f,\epsilon]_t^{j-1}(\epsilon) - f(\mathbf{z}_t(\epsilon)))^2} \\ &\le \sqrt{n}(\beta_j + \beta_{j-1}) = 3\sqrt{n}\beta_j \ .\end{aligned}$$

If there exists no such $f \in \mathcal{F}$ for the given ϵ and (s,r), then $\mathbf{w}_t^{(s,r)}(\epsilon)$ is zero for all $t \ge t_o$, for some $1 \le t_o < n$, and thus

$$\sqrt{\sum_{t=1}^n \mathbf{w}_t^{(s,r)}(\epsilon)^2} \le \sqrt{\sum_{t=1}^n \mathbf{w}_t^{(s,r)}(\epsilon')^2}$$

for any other path ϵ' which agrees with ϵ up to t_o. If $t_o = 1$, the desired bound holds. If $t_o > 1$, there must be an ϵ' such that there exists $f \in \mathcal{F}$ for which $\mathbf{v}^s = \mathbf{v}[f,\epsilon']^j$ and $\mathbf{v}^r = \mathbf{v}[f,\epsilon']^{j-1}$, leading us to the case that we analyzed before. Hence, the bound

$$\sqrt{\sum_{t=1}^n \mathbf{w}_t^{(s,r)}(\epsilon)^2} \le 3\sqrt{n}\beta_j$$

holds for all $\epsilon \in \{\pm 1\}^n$ and all $\mathbf{w}^{(s,r)} \in W_j$.

Now, back to (16), we take the expectation with respect to the path ϵ and apply Lemma 3:

$$\begin{aligned}\mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{t=1}^n \epsilon_t f(\mathbf{z}_t(\epsilon))\right] &\le \beta_N + \frac{1}{\sqrt{n}} \sum_{j=1}^N 3\beta_j \sqrt{2\log(|V_j|\ |V_{j-1}|)} \\ &\le \beta_N + \frac{1}{\sqrt{n}} \sum_{j=1}^N 6\beta_j \sqrt{\log(|V_j|)} \ .\end{aligned}$$

Since $\beta_j = 2(\beta_j - \beta_{j+1})$ and $\log \mathcal{N}_2(\beta_0, \mathcal{F}, \mathbf{z}) = 0$, the above sum can be upper bounded by

$$\beta_N + \frac{12}{\sqrt{n}} \sum_{j=1}^N (\beta_j - \beta_{j+1}) \sqrt{\log \mathcal{N}_2(\beta_j, \mathcal{F}, \mathbf{z})} \le \beta_N + \frac{12}{\sqrt{n}} \int_{\beta_{N+1}}^{\beta_0} \sqrt{\log\ \mathcal{N}_2(\delta, \mathcal{F}, \mathbf{z})}\, d\delta \ .$$

Now for any $\alpha > 0$, pick $N = \max\{j : \beta_j > 2\alpha\}$. In this case we see that by our choice of N, $\beta_{N+1} \le 2\alpha$ and so $\beta_N = 2\beta_{N+1} \le 4\alpha$. Also note that since $\beta_N > 2\alpha$, $\beta_{N+1} = \frac{\beta_N}{2} > \alpha$. Hence, we conclude that

$$\mathfrak{R}_n(\mathcal{F}, \mathbf{z}) \le \inf_{\alpha} \left\{4\alpha + \frac{12}{\sqrt{n}} \int_{\alpha}^1 \sqrt{\log\ \mathcal{N}_2(\delta, \mathcal{F}, \mathbf{z})}\, d\delta\right\} \ .$$

□

Proof of Theorem 5. First, a few definitions. The *left subtree* $\mathbf{z}^\ell$ of $\mathbf{z}$ at the root is defined as $n-1$ mappings $(\mathbf{z}_1^\ell, \ldots, \mathbf{z}_{n-1}^\ell)$ with $\mathbf{z}_i^\ell(\epsilon) = \mathbf{z}_{i+1}((-1,\epsilon))$ for $\epsilon \in \{\pm 1\}^{n-1}$. The right subtree $\mathbf{z}^r$ is defined analogously by setting the first coordinate to be $+1$. Given two subtrees $\mathbf{z}$, $\mathbf{v}$ of the same depth $n-1$ and a constant mapping $\mathbf{w}_1$, we can *join* the two subtrees to obtain a new set of mappings $(\mathbf{w}_1, \dots, \mathbf{w}_n)$ as follows. The root is the constant $\mathbf{w}_1$. For $i \in \{2, \dots, n\}$ and $\epsilon \in \{\pm 1\}^n$, $\mathbf{w}_i(\epsilon) = \mathbf{z}_{i-1}(\epsilon_{2:n})$ if $\epsilon_1 = -1$ and $\mathbf{w}_i(\epsilon) = \mathbf{v}_{i-1}(\epsilon_{2:n})$ if $\epsilon_1 = +1$.

For integers $d \geq 0$ and $n \geq 1$, define the function

$$g_k(d,n) = \sum_{i=0}^{d} \binom{n}{i} k^i \tag{17}$$

with the convention $\binom{n}{0} = 1$. It is not difficult to verify that this function satisfies the recurrence

$$g_k(d, n-1) + k g_k(d-1, n-1) = g_k(d,n)$$

for all $1 \leq d \leq n-1$. To visualize this recursion, consider a $k \times n$ matrix and ask for ways to choose at most d columns followed by a choice among the k rows for each chosen column. The task can be decomposed into (a) making the d column choices out of the first $n-1$ columns, followed by picking rows (there are $g_k(d, n-1)$ ways to do it) or (b) choosing $d-1$ columns (followed by choices of rows) out of the first $n-1$ columns and choosing a row for the nth column (there are $k g_k(d-1, n-1)$ ways to do it). This gives the recursive formula.

First consider the case $n \leq d$. Let V be a set of trees $\mathbf{v}$ with constant mappings $\mathbf{v}_t(\epsilon) = v_t \in \{1/2, 3/2, \dots, k-1/2\}$. There are k^n such trees, and thus $|V| = k^n$. Clearly, V is an 1/2-cover of $\mathcal{F}$ on any tree $\mathbf{z}$ of depth n and $\mathcal{N}_\infty(1/2, \mathcal{F}, n) \leq k^n < (ekn)^d$.

We now turn to the case $n > d$. The proof proceeds by induction on (d,n), $0 \leq d < n$, for the following statement:

$\mathfrak{S}(d,n)$: For any set $\mathcal{Z}$ and a function class $\mathcal{F}$ of functions from $\mathcal{Z}$ to $\{0, \dots, k\}$ with $\text{fat}_2(\mathcal{F}, \mathcal{Z}) \leq d$, for any $\mathcal{Z}$-valued tree $\mathbf{z}$ of depth n, $\mathcal{N}_\infty(1/2, \mathcal{F}, \mathbf{z}) \leq g_k(d,n)$.

In what follows, we shall refer to an ℓ_∞ cover at scale 1/2 simply as a 1/2-cover. Recall the notation $\text{fat}_2(\mathcal{F}, \mathbf{z}) = \text{fat}_2(\mathcal{F}, \text{Img}(\mathbf{z}))$.

Base: We prove two base cases. For the first one, consider $n = d \geq 1$. Above, we argued that the size of the 1/2-cover is at most k^n, which is smaller than $g(n,n) = \sum_{i=0}^{n} \binom{n}{i} k^i$. For the second base case, consider $d = 0$ and any $n \geq 1$. Observe that $\text{fat}_2(\mathcal{F}, \mathcal{Z}) = 0$ means that there is no element of $\mathcal{Z}$ which is 2-shattered by $\mathcal{F}$. In other words, functions in $\mathcal{F}$ differ by at most 1 over $\mathcal{Z}$ and, therefore, there exists a 1/2-cover of size $1 = g_k(0,n)$ on any tree $\mathbf{z}$ of depth n. Given these two base cases, it is enough to prove that $\mathfrak{S}(d, n-1)$ and $\mathfrak{S}(d-1, n-1)$ imply $\mathfrak{S}(d,n)$ for $1 \leq d \leq n-1$.

Induction step: Suppose by the way of induction that the statement holds for $(d, n-1)$ and $(d-1, n-1)$. Consider any set $\mathcal{Z}$ and any $\mathcal{F} \subseteq \{0, \dots, k\}^{\mathcal{Z}}$ with $\text{fat}_2(\mathcal{F}, \mathcal{Z}) = d$. Consider any $\mathcal{Z}$-valued tree $\mathbf{z}$ of depth n. If $\text{fat}_2(\mathcal{F}, \mathbf{z}) < d$, the desired size of the 1/2-cover on $\mathbf{z}$ follows from the induction hypothesis $\mathfrak{S}(d-1, n)$ with $\mathcal{Z} = \text{Img}(\mathbf{z})$; therefore, consider the case $\text{fat}_2(\mathcal{F}, \mathbf{z}) = \text{fat}_2(\mathcal{F}, \mathcal{Z}) = d$. Define the partition $\mathcal{F} = \mathcal{F}_0 \cup \ldots \cup \mathcal{F}_k$ with $\mathcal{F}_i = \{f \in \mathcal{F} : f(\mathbf{z}_1) = i\}$ for $i \in \{0, \dots, k\}$, where $\mathbf{z}_1$ is the root of $\mathbf{z}$. Let $m = |\{i : \text{fat}_2(\mathcal{F}_i, \mathbf{z}) = d\}|$.

Suppose first, for the sake of contradiction, that $\text{fat}_2(\mathcal{F}_i, \mathbf{z}) = \text{fat}_2(\mathcal{F}_j, \mathbf{z}) = d$ for some $0 \leq i, j \leq k$ with $j - i \geq 2$. Then there exist two trees $\mathbf{w}$ and $\mathbf{v}$ of depth d which are 2-shattered by $\mathcal{F}_i$ and $\mathcal{F}_j$, respectively, and with $\text{Img}(\mathbf{w}), \text{Img}(\mathbf{v}) \subseteq \text{Img}(\mathbf{z})$. Since functions within each subset $\mathcal{F}_i$ take on the same values on $\mathbf{z}_1$, we conclude that $\mathbf{z}_1 \notin \text{Img}(\mathbf{w}), \mathbf{z}_1 \notin \text{Img}(\mathbf{v})$. This follows immediately from the definition of shattering. We now *join* the two shattered trees $\mathbf{w}$ and $\mathbf{v}$ with $\mathbf{z}_1$ at the root and observe that $\mathcal{F}_i \cup \mathcal{F}_j$ 2-shatters this resulting tree of depth $d+1$, which is a contradiction. Indeed, the witness $\mathbb{R}$-valued tree $\mathbf{s}$ is constructed by joining the two witnesses for the 2-shattered trees $\mathbf{w}$ and $\mathbf{v}$ and by defining the root as $\mathbf{s}_1 = (i+j)/2$. It is easy to see that $\mathbf{s}$ is a witness to the shattering. Given any $\epsilon \in \{\pm 1\}^{d+1}$, there is a function $f^i \in \mathcal{F}_i$ which realizes the desired separation under the signs $(\epsilon_2, \ldots, \epsilon_{d+1})$ for the tree $\mathbf{w}$ and there is a function $f^j \in \mathcal{F}_j$ which does the same for $\mathbf{v}$. Depending on $\epsilon_1 = -1$ or $\epsilon_1 = +1$, either f^i or f^j realize the separation over ϵ, a contradiction.

We conclude that the number of subsets of $\mathcal{F}$ with fat-shattering dimension equal to d cannot be more than two (for otherwise at least two indices will be separated by 2 or more). We have three cases: $m = 0$, $m = 1$, or $m = 2$, and in the last case it must be that the indices of the two subsets differ by 1.

First, consider any $\mathcal{F}_i$ with $\mathrm{fat}_2(\mathcal{F}_i, \mathbf{z}) \le d-1$, $i \in \{0, \ldots, k\}$. By the induction hypothesis $\mathfrak{S}(d-1, n-1)$ with the function class $\mathcal{F}_i$, there are 1/2-covers V^ℓ and V^r of $\mathcal{F}_i$ on the subtrees $\mathbf{z}^\ell$ and $\mathbf{z}^r$, respectively, both of size at most $g_k(d-1, n-1)$. Informally, out of these 1/2-covers we can create a 1/2-cover V for $\mathcal{F}_i$ on $\mathbf{z}$ by pairing the 1/2-covers in V^ℓ and V^r. The resulting cover of $\mathcal{F}_i$ will be of size at most $g_k(d-1, n-1)$. Formally, consider a set of pairs $(\mathbf{v}^\ell, \mathbf{v}^r)$ of trees, with $\mathbf{v}^\ell \in V^\ell$, $\mathbf{v}^r \in V^r$ and such that each tree in V^ℓ and V^r appears in at least one of the pairs. Clearly, this can be done using at most $g_k(d-1, n-1)$ pairs, and such a construction is not unique. We join the subtrees in every pair $(\mathbf{v}^\ell, \mathbf{v}^r)$ with a constant i as the root, thus creating a set V of trees, $|V| \le g_k(d-1, n-1)$. We claim that V is a 1/2-cover for $\mathcal{F}_i$ on $\mathbf{z}$. Note that all the functions in $\mathcal{F}_i$ take on the same value i on $\mathbf{z}_1$ and by construction $\mathbf{v}_1 = i$ for any $\mathbf{v} \in V$. Now, consider any $f \in \mathcal{F}_i$ and $\epsilon \in \{\pm 1\}^n$. Assume $\epsilon_1 = -1$. By assumption, there is a $\mathbf{v}^\ell \in V^\ell$ such that $|\mathbf{v}^\ell_t(\epsilon_{2:n}) - f(\mathbf{z}_{t+1}(\epsilon_{1:n}))| \le 1/2$ for any $t \in [n-1]$. By construction $\mathbf{v}^\ell$ appears as a left subtree of at least one tree in V. The same argument holds for $\epsilon_1 = +1$ by finding an appropriate subtree in V^r. We conclude that V is a 1/2-cover of $\mathcal{F}_i$ on $\mathbf{z}$, and such a V can be found for any $i \in \{0, \ldots, k\}$ with $\mathrm{fat}_2(\mathcal{F}_i, \mathbf{z}) \le d-1$. Therefore, the total size of a 1/2-cover for the union $\cup_{i:\mathrm{fat}_2(\mathcal{F}_i, \mathbf{z}) \le d-1} \mathcal{F}_i$ is at most $(k+1-m) g_k(d-1, n-1)$.

If $m = 0$, the induction step is proven because $g_k(d-1, n-1) \le g_k(d, n-1)$ and so the total size of the constructed cover is at most

$$(k+1) g_k(d-1, n-1) \le g_k(d, n-1) + k g_k(d-1, n-1) = g_k(d, n).$$

Now, consider the case $m = 1$ and let $\mathrm{fat}_2(\mathcal{F}_i, \mathbf{z}) = d$. An argument exactly as above yields a 1/2-cover for $\mathcal{F}_i$, and this cover is of size at most $g_k(d, n-1)$ by induction. The total 1/2-cover is therefore of size at most

$$g_k(d, n-1) + k g_k(d-1, n-1) = g_k(d, n).$$

Lastly, for $m = 2$, suppose $\mathrm{fat}_2(\mathcal{F}_i, \mathbf{z}) = \mathrm{fat}_2(\mathcal{F}_j, \mathbf{z}) = d$ for $j = i+1$. Let $\mathcal{F}' = \mathcal{F}_i \cup \mathcal{F}_j$. Note that $\mathrm{fat}_2(\mathcal{F}', \mathbf{z}) = d$. Just as before, the 1/2-covering for $\mathbf{z}$ can be constructed by considering the 1/2-covers for the two subtrees. However, when joining any $(\mathbf{v}^\ell, \mathbf{v}^r)$, we take $(i+j)/2$ as the root. It is straightforward to check that the resulting cover is indeed an 1/2-cover of $\mathcal{F}'$ on $\mathbf{z}$, thanks to the relation $|i-j| = 1$. The size of the constructed cover is at most $g_k(d, n-1)$ by the induction hypothesis $\mathfrak{S}(d, n-1)$ with the class $\mathcal{F}'$, and the induction step follows. This concludes the induction proof, yielding the main statement of the theorem.

Finally, the upper bound on $g_k(d,n)$ is

$$\sum_{i=1}^{d} \binom{n}{i} k^i \le \left(\frac{kn}{d}\right)^d \sum_{i=1}^{d} \binom{n}{i} \left(\frac{d}{n}\right)^i \le \left(\frac{kn}{d}\right)^d \left(1 + \frac{d}{n}\right)^n \le \left(\frac{ekn}{d}\right)^d$$

whenever $n > d$.

□

Proof of Corollary 6. For any $\alpha > 0$ define an α-discretization of the $[-1,1]$ interval as $B_\alpha = \{-1 + \alpha/2, -1 + 3\alpha/2, \ldots, -1 + (2k+1)\alpha/2, \ldots\}$ for $0 \le k$ and $(2k+1)\alpha \le 4$. Also for any $a \in [-1,1]$ define $\lfloor a \rfloor_\alpha = \underset{r \in B_\alpha}{\operatorname{argmin}} |r - a|$ with ties being broken by choosing the smaller discretization point. For a function $f : \mathcal{Z} \mapsto [-1,1]$ let the function $\lfloor f \rfloor_\alpha$ be defined pointwise as $\lfloor f(z) \rfloor_\alpha$, and let $\lfloor \mathcal{F} \rfloor_\alpha = \{\lfloor f \rfloor_\alpha : f \in \mathcal{F}\}$. Fix an arbitrary $\mathcal{Z}$-valued tree $\mathbf{z}$ of depth n. First, we prove that $\mathcal{N}_\infty(\alpha, \mathcal{F}, \mathbf{z}) \le \mathcal{N}_\infty(\alpha/2, \lfloor \mathcal{F} \rfloor_\alpha, \mathbf{z})$. Indeed, suppose the set of trees V is a minimal $\alpha/2$-cover of $\lfloor \mathcal{F} \rfloor_\alpha$ on $\mathbf{z}$. That is,

$$\forall f_\alpha \in \lfloor \mathcal{F} \rfloor_\alpha,\ \forall \epsilon \in \{\pm 1\}^n\ \exists \mathbf{v} \in V \text{ s.t. } \quad \forall t \in [n],\ \ |\mathbf{v}_t(\epsilon) - f_\alpha(\mathbf{z}_t(\epsilon))| \le \alpha/2.$$

Pick any $f \in \mathcal{F}$ and let $f_\alpha = \lfloor f \rfloor_\alpha$. Then $\|f - f_\alpha\|_\infty \le \alpha/2$. Then for all $\epsilon \in \{\pm 1\}^n$ and $t \in [n]$

$$|f(\mathbf{z}_t(\epsilon)) - \mathbf{v}_t(\epsilon)| \le |f(\mathbf{z}_t(\epsilon)) - f_\alpha(\mathbf{z}_t(\epsilon))| + |f_\alpha(\mathbf{z}_t(\epsilon)) - \mathbf{v}_t(\epsilon)| \le \alpha,$$

and so V also provides an ℓ_∞ cover at scale α.

We conclude that $\mathcal{N}_\infty(\alpha, \mathcal{F}, \mathbf{z}) \le \mathcal{N}_\infty(\alpha/2, \lfloor \mathcal{F} \rfloor_\alpha, \mathbf{z}) = \mathcal{N}_\infty(1/2, \mathcal{G}, \mathbf{z})$ where $\mathcal{G} = \frac{1}{\alpha}\lfloor \mathcal{F} \rfloor_\alpha$. The functions of $\mathcal{G}$ take on a discrete set of at most $\lfloor 2/\alpha \rfloor + 1$ values. Obviously, by adding a constant to all the functions in $\mathcal{G}$, we can make the set of values to be $\{0, \ldots, \lfloor 2/\alpha \rfloor\}$. We now apply Theorem 5 with the upper bound $(ekn)^d$ which holds for any $n \ge 1$. This yields $\mathcal{N}_\infty(1/2, \mathcal{G}, \mathbf{z}) \le (2en/\alpha)^{\mathrm{fat}_2(\mathcal{G})}$.

It remains to prove $\mathrm{fat}_2(\mathcal{G}) \le \mathrm{fat}_\alpha(\mathcal{F})$, or, equivalently (by scaling) $\mathrm{fat}_{2\alpha}(\lfloor \mathcal{F} \rfloor_\alpha) \le \mathrm{fat}_\alpha(\mathcal{F})$. To this end, suppose there exists an $\mathbb{R}$-valued tree $\mathbf{w}$ of depth $d = \mathrm{fat}_{2\alpha}(\lfloor \mathcal{F} \rfloor_\alpha)$ such that there is an witness tree $\mathbf{s}$ with

$$\forall \epsilon \in \{\pm 1\}^d,\ \exists f_\alpha \in \lfloor \mathcal{F} \rfloor_\alpha \quad \text{s.t. } \forall t \in [d],\ \epsilon_t(f_\alpha(\mathbf{w}_t(\epsilon)) - \mathbf{s}_t(\epsilon)) \ge \alpha\ .$$

Using the fact that for any $f \in \mathcal{F}$ and $f_\alpha = \lfloor f \rfloor_\alpha$ we have $\|f - f_\alpha\|_\infty \le \alpha/2$, it follows that

$$\forall \epsilon \in \{\pm 1\}^d,\ \exists f \in \mathcal{F} \quad \text{s.t. } \forall t \in [d],\ \epsilon_t(f(\mathbf{w}_t(\epsilon)) - \mathbf{s}_t(\epsilon)) \ge \alpha/2\ .$$

That is, $\mathbf{s}$ is a witness to α-shattering by $\mathcal{F}$. We conclude that for any $\mathbf{z}$,

$$\mathcal{N}_\infty(\alpha, \mathcal{F}, \mathbf{z}) \le \mathcal{N}_\infty(\alpha/2, \lfloor \mathcal{F} \rfloor_\alpha, \mathbf{z}) \le \left(\frac{2en}{\alpha}\right)^{\mathrm{fat}_{2\alpha}(\lfloor \mathcal{F} \rfloor_\alpha)} \le \left(\frac{2en}{\alpha}\right)^{\mathrm{fat}_\alpha(\mathcal{F})}\ .$$

□

Proof of Theorem 7. The proof is very close to the proof of Theorem 5, with a few key differences. Recall the definition of $g_k(d,n)$ in (17). First, consider the case $d \ge n$. As in the proof of Theorem 5, we can construct a set of $(k+1)^n$ trees $\mathbf{v}$ consisting of constant mappings $\mathbf{v}_t(\epsilon) = v_t \in \{0, \ldots, k\}$. This set trivially provides a 0-cover for $\mathcal{F}$ on any $\mathcal{Z}$-valued tree of depth n.

We now turn to the case $n > d$. The proof proceeds by induction on (d,n), $0 \le d < n$, for the following statement:

$\mathfrak{S}(d,n)$: For any set $\mathcal{Z}$ and a function class $\mathcal{F}$ of functions from $\mathcal{Z}$ to $\{0,\ldots,k\}$ with $\mathrm{fat}_1(\mathcal{F},\mathcal{Z}) \le d$, for any $\mathcal{Z}$-valued tree $\mathbf{z}$ of depth n, $\mathcal{N}(0,\mathcal{F},\mathbf{z}) \le g_k(d,n)$.

Base: Let $n = d \ge 1$. We already argued that the size of the 0-cover is at most $(k+1)^n$, which is exactly the value $g_k(n,n) = \sum_{i=0}^{n} \binom{n}{i} k^i$. The second base case is $d = 0$ and $n \ge 1$, which happens if all functions in $\mathcal{F}$ coincide over $\mathcal{Z}$ and, hence, $\mathcal{N}(0,\mathcal{F},\mathbf{z}) = 1 = g_k(0,n)$ for any $\mathcal{Z}$-valued tree $\mathbf{z}$ of depth n.

Induction step: Suppose by the way of induction that the statements $\mathfrak{S}(d-1,n-1)$ and $\mathfrak{S}(d,n-1)$ hold. Consider any set $\mathcal{Z}$ and any $\mathcal{F} \subseteq \{0,\ldots,k\}^{\mathcal{Z}}$ with $\mathrm{fat}_1(\mathcal{F},\mathcal{Z}) = d$. Consider any $\mathcal{Z}$-valued tree $\mathbf{z}$ of depth n. If $\mathrm{fat}_1(\mathcal{F},\mathbf{z}) < d$, the desired size of the 0-cover on $\mathbf{z}$ follows from the induction hypothesis $\mathfrak{S}(d-1,n)$ with $\mathcal{Z} = \mathrm{Img}(\mathbf{z})$; therefore, consider the case $\mathrm{fat}_1(\mathcal{F},\mathbf{z}) = \mathrm{fat}_1(\mathcal{F},\mathcal{Z}) = d$. Define the partition $\mathcal{F} = \mathcal{F}_0 \cup \ldots \cup \mathcal{F}_k$ with $\mathcal{F}_i = \{f \in \mathcal{F} : f(\mathbf{z}_1) = i\}$ for $i \in \{0,\ldots,k\}$.

We first argue that $\mathrm{fat}_1(\mathcal{F}_i,\mathbf{z}) = d$ for at most one value $i \in \{0,\ldots,k\}$. By the way of contradiction, suppose we do have $\mathrm{fat}_1(\mathcal{F}_i,\mathbf{z}) = \mathrm{fat}_1(\mathcal{F}_j,\mathbf{z}) = d$ for $i \ne j$. Then there exist two trees $\mathbf{w}$ and $\mathbf{v}$ of depth d 1-shattered by $\mathcal{F}_i$ and $\mathcal{F}_j$, respectively, and with $\mathrm{Img}(\mathbf{w}), \mathrm{Img}(\mathbf{v}) \subseteq \mathrm{Img}(\mathbf{z})$. Since functions within each subset $\mathcal{F}_i$ take on the same values on $\mathbf{z}_1$, we conclude that $\mathbf{z}_1 \notin \mathrm{Img}(\mathbf{w}), \mathbf{z}_1 \notin \mathrm{Img}(\mathbf{v})$. We join the two shattered $\mathbf{w}$ and $\mathbf{v}$ trees with $\mathbf{z}_1$ at the root and observe that $\mathcal{F}_i \cup \mathcal{F}_j$ 1-shatters this resulting tree of depth $d+1$, which is a contradiction. Indeed, the witness $\mathbb{R}$-valued tree $\mathbf{s}$ is constructed by joining the two witnesses for the 1-shattered trees $\mathbf{w}$ and $\mathbf{v}$ and by defining the root as $\mathbf{s}_1 = (i+j)/2$.

Without loss of generality, assume $\mathrm{fat}_1(\mathcal{F}_0,\mathbf{z}) \le d$ and $\mathrm{fat}_1(\mathcal{F}_i,\mathbf{z}) \le d-1$ for $i \in \{1,\ldots,k\}$. By the induction hypothesis $\mathfrak{S}(d-1,n-1)$, for any $\mathcal{F}_i$, $i \in \{1,\ldots,k\}$, there are 0-covers V^ℓ and V^r of $\mathcal{F}_i$ on the subtrees $\mathbf{z}^\ell$ and $\mathbf{z}^r$, respectively, both of size at most $g_k(d-1,n-1)$. Out of these 0-covers we can create a 0-cover V for $\mathcal{F}_i$ on $\mathbf{z}$ by pairing the 0-covers in V^ℓ and V^r. Formally, consider a set of pairs $(\mathbf{v}^\ell,\mathbf{v}^r)$ of trees, with $\mathbf{v}^\ell \in V^\ell$, $\mathbf{v}^r \in V^r$ and such that each tree in V^ℓ and V^r appears in at least one of the pairs. Clearly, this can be done using at most $g_k(d-1,n-1)$ pairs, and such a pairing is not unique. We join the subtrees in every pair $(\mathbf{v}^\ell,\mathbf{v}^r)$ with a constant i as the root, thus creating a set V of trees, $|V| \le g_k(d-1,n-1)$. We claim that V is a 0-cover for $\mathcal{F}_i$ on $\mathbf{z}$. Note that all the functions in $\mathcal{F}_i$ take on the same value i on $\mathbf{z}_1$ and by construction $\mathbf{v}_1 = i$ for any $\mathbf{v} \in V$. Now, consider any $f \in \mathcal{F}_i$ and $\epsilon \in \{\pm 1\}^n$. Assume $\epsilon_1 = -1$. By assumption, there is a $\mathbf{v}^\ell \in V^\ell$ such that $\mathbf{v}^\ell_t(\epsilon_{2:n}) = f(\mathbf{z}_{t+1}(\epsilon_{1:n}))$ for any $t \in [n-1]$. By construction $\mathbf{v}^\ell$ appears as a left subtree of at least one tree in V, which, therefore, matches the values of f for $\epsilon_{1:n}$. The same argument holds for $\epsilon_1 = +1$ by finding an appropriate subtree in V^r. We conclude that V is a 0-cover of $\mathcal{F}_i$ on $\mathbf{z}$, and this holds for any $i \in \{1,\ldots,k\}$.

Therefore, the total size of a 0-cover for $\mathcal{F}_1 \cup \ldots \cup \mathcal{F}_k$ on $\mathbf{z}$ is at most $k g_k(d-1,n-1)$. A similar argument yields a 0-cover for $\mathcal{F}_0$ on $\mathbf{z}$ of size at most $g_k(d,n-1)$ by induction. Thus, the size of the resulting 0-cover of $\mathcal{F}$ on $\mathbf{z}$ is at most $g_k(d,n-1) + k g_k(d-1,n-1) = g_k(d,n)$, completing the induction step and yielding the main statement of the theorem.

□

Proof of Lemma 8. Consider some $\beta > 0$ for which $\mathrm{fat}_\beta = \mathrm{fat}_\beta(\mathcal{F}) \ge n$. By definition, there exists a tree $\mathbf{z}^*$ of depth n that is β-shattered (with some witness $\mathbf{s}$) by the function class $\mathcal{F}$. Hence we have that

$$n\,\mathfrak{R}_n(\mathcal{F}) \geq \mathbb{E}_\epsilon \sup_{f\in\mathcal{F}} \sum_{t=1}^n \epsilon_t f(\mathbf{z}_t^*(\epsilon)) = \mathbb{E}_\epsilon \sup_{f\in\mathcal{F}} \sum_{t=1}^n \epsilon_t \left(f(\mathbf{z}_t^*(\epsilon)) - \mathbf{s}_t(\epsilon)\right) \geq \frac{n\beta}{2}\ .$$

We conclude that $\mathrm{fat}_\beta \geq n$ implies that $\beta \leq 2\mathfrak{R}_n(\mathcal{F})$. The converse is the first statement of the lemma.

We now prove the second statement. For a given $\beta > 0$, first consider any n such that $\mathrm{fat}_\beta(\mathcal{F}) \leq n$. Let $\mathbf{z}$ be a $\mathcal{Z}$-valued tree of depth $\mathrm{fat}_\beta = \mathrm{fat}_\beta(\mathcal{F})$, β-shattered by $\mathcal{F}$ w.r.t. the witness tree $\mathbf{s}$. To show a lower bound on Rademacher complexity we construct a tree of depth $n' = \lceil \frac{n}{\mathrm{fat}_\beta} \rceil \mathrm{fat}_\beta$ using the shattered tree $\mathbf{z}$. For convenience, define $k = \lceil \frac{n}{\mathrm{fat}_\beta} \rceil = \frac{n'}{\mathrm{fat}_\beta}$ and consider the $\mathcal{Z}$-valued tree $\tilde{\mathbf{z}}$ of depth n' constructed as follows. For any path $\epsilon \in \{\pm 1\}^{n'}$ and any $t \in [n']$, set

$$\tilde{\mathbf{z}}_t(\epsilon) = \mathbf{z}_{\lceil \frac{t}{k} \rceil}(\tilde{\epsilon})$$

where $\tilde{\epsilon} \in \{\pm 1\}^{\mathrm{fat}_\beta}$ is the sequence of signs specified as

$$\tilde{\epsilon} = \left(\mathrm{sign}\left(\sum_{i=1}^k \epsilon_i\right), \mathrm{sign}\left(\sum_{i=k+1}^{2k} \epsilon_i\right), \ldots, \mathrm{sign}\left(\sum_{i=k\left(\mathrm{fat}_\beta - 1\right)}^{k\,\mathrm{fat}_\beta} \epsilon_i\right)\right).$$

That is for each ϵ the corresponding $\tilde{\epsilon}$ is the sequence of signs of length $\mathrm{fat}_\beta(\mathcal{F})$ obtained by dividing ϵ into fat_β blocks each of length k and taking the majority vote of signs in the corresponding block.

Now we proceed to lower bound the sequential Rademacher complexity as follows :

$$n'\ \mathfrak{R}_{n'}(\mathcal{F}) \geq \mathbb{E}_\epsilon \left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{n'} \epsilon_t f(\tilde{\mathbf{z}}_t(\epsilon))\right] = \mathbb{E}_\epsilon \left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{n'} \epsilon_t f(\mathbf{z}_{\lceil \frac{t}{k} \rceil}(\tilde{\epsilon}))\right].$$

Since $\mathbb{E}_\epsilon \left[\sum_{t=1}^{n'} \epsilon_t \mathbf{s}_{\lceil \frac{t}{k} \rceil}(\tilde{\epsilon})\right] = 0$, the above quantity is equal to

$$\begin{aligned}
\mathbb{E}_\epsilon \left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{n'} \epsilon_t \left(f(\mathbf{z}_{\lceil \frac{t}{k} \rceil}(\tilde{\epsilon})) - \mathbf{s}_{\lceil \frac{t}{k} \rceil}(\tilde{\epsilon})\right)\right] &= \mathbb{E}_\epsilon \left[\sup_{f\in\mathcal{F}} \sum_{i=1}^{\mathrm{fat}_\beta} \sum_{j=(i-1)k+1}^{i\cdot k} \epsilon_j \left(f(\mathbf{z}_i(\tilde{\epsilon})) - \mathbf{s}_i(\tilde{\epsilon})\right)\right] \\
&= \mathbb{E}_\epsilon \left[\sup_{f\in\mathcal{F}} \sum_{i=1}^{\mathrm{fat}_\beta} \left(\sum_{j=(i-1)k+1}^{i\cdot k} \epsilon_j\right) \left(f(\mathbf{z}_i(\tilde{\epsilon})) - \mathbf{s}_i(\tilde{\epsilon})\right)\right] \\
&= \mathbb{E}_\epsilon \left[\sup_{f\in\mathcal{F}} \sum_{i=1}^{\mathrm{fat}_\beta} \left|\sum_{j=(i-1)k+1}^{i\cdot k} \epsilon_j\right| \tilde{\epsilon}_i \left(f(\mathbf{z}_i(\tilde{\epsilon})) - \mathbf{s}_i(\tilde{\epsilon})\right)\right].
\end{aligned}$$

Since $\mathbf{z}$ is shattered by the function class $\mathcal{F}$, for any choice of signs $\tilde{\epsilon}$ there exists $f_{\tilde{\epsilon}} \in \mathcal{F}$ such that for any $i \in [\mathrm{fat}_\beta(\mathcal{F})]$, $\tilde{\epsilon}_i(f_{\tilde{\epsilon}}(\mathbf{z}_i(\tilde{\epsilon})) - \mathbf{s}_i(\tilde{\epsilon})) \geq \frac{\beta}{2}$. Hence, we pass to the lower bound

$$\mathbb{E}_\epsilon \left[\sum_{i=1}^{\mathrm{fat}_\beta} \left|\sum_{j=(i-1)k+1}^{i\cdot k} \epsilon_j\right| \tilde{\epsilon}_i \left(f_{\tilde{\epsilon}}(\mathbf{z}_i(\tilde{\epsilon})) - \mathbf{s}_i(\tilde{\epsilon})\right)\right] \geq \mathbb{E}_\epsilon \left[\sum_{i=1}^{\mathrm{fat}_\beta} \left|\sum_{j=(i-1)k+1}^{i\cdot k} \epsilon_j\right| \frac{\beta}{2}\right] = \frac{\beta}{2} \mathrm{fat}_\beta\ \mathbb{E}_\epsilon \left[\left|\sum_{i=1}^k \epsilon_j\right|\right].$$

The last expression is lower bounded by Khinchine's inequality by

$$\frac{\beta}{2}\mathrm{fat}_\beta(\mathcal{F})\sqrt{\frac{k}{2}} = \frac{\beta}{2}\mathrm{fat}_\beta(\mathcal{F})\sqrt{\frac{n'}{2\,\mathrm{fat}_\beta(\mathcal{F})}} = \sqrt{\frac{\beta^2\,n'\,\mathrm{fat}_\beta(\mathcal{F})}{8}}\ .$$

Thus we have shown that

$$\mathrm{fat}_\beta(\mathcal{F}) \le \frac{8\,n'\,\mathfrak{R}_{n'}(\mathcal{F})^2}{\beta^2}\ .$$

It is easy to show that $n'\mathfrak{R}_{n'}(\mathcal{F})$ is non-decreasing in n' and that $\mathfrak{R}_{2n}(\mathcal{F}) \le \mathfrak{R}_n(\mathcal{F})$. Thus,

$$n'\mathfrak{R}_{n'}(\mathcal{F}) \le 2n\mathfrak{R}_{2n}(\mathcal{F}) \le 2n\mathfrak{R}_n(\mathcal{F})\ .$$

Since $n \le n'$ we can conclude that

$$\mathrm{fat}_\beta(\mathcal{F}) \le \frac{32\,n\ \mathfrak{R}_n(\mathcal{F})^2}{\beta^2}\ .$$

This proves the required statement for any n such that $\mathrm{fat}_\beta(\mathcal{F}) \le n$. For the case $\mathrm{fat}_\beta(\mathcal{F}) > n$, we use the first statement of the lemma which implies $\beta^2 \le 4\mathfrak{R}_n(\mathcal{F})^2$, concluding the proof. □

Proof of Lemma 9. Let $\mathrm{fat}_\beta = \mathrm{fat}_\beta(\mathcal{F},\mathcal{Z})$. By Corollary 6, we have that

$$\mathfrak{D}_n^\infty(\mathcal{F}) \le \inf_\alpha\left\{4\alpha + \frac{12}{\sqrt{n}}\int_\alpha^1 \sqrt{\mathrm{fat}_\beta \log(2en/\beta)}\ d\beta\right\}\ .$$

Choosing $\alpha = 2\mathfrak{R}_n(\mathcal{F})$,

$$\mathfrak{D}_n^\infty(\mathcal{F}) \le 8\ \mathfrak{R}_n(\mathcal{F}) + \frac{12}{\sqrt{n}}\int_{2\mathfrak{R}_n(\mathcal{F})}^1 \sqrt{\mathrm{fat}_\beta \log(2en/\beta)}d\beta\ .$$

Lemma 8 implies that for any $\beta > 2\mathfrak{R}_n(\mathcal{F})$,

$$\mathrm{fat}_\beta \le \frac{32n\ \mathfrak{R}_n(\mathcal{F})^2}{\beta^2}\ .$$

The following inequality is useful for bounding the integral: For any $b > 1$ and $\alpha \in (0,1)$

$$\int_\alpha^1 \frac{1}{\beta}\sqrt{\log(b/\beta)}d\beta = \int_b^{b/\alpha}\frac{1}{x}\sqrt{\log x}dx = \frac{2}{3}\log^{3/2}(x)\Big|_b^{b/\alpha} \le \frac{2}{3}\log^{3/2}(b/\alpha) \tag{18}$$

where we performed a change of variables with $x = b/\beta$. Using Eq. (18),

$$\begin{aligned}\mathfrak{D}_n^\infty(\mathcal{F}) &\le 8\ \mathfrak{R}_n(\mathcal{F}) + 48\sqrt{2}\ \mathfrak{R}_n(\mathcal{F})\int_{2\mathfrak{R}_n(\mathcal{F})}^1 \frac{1}{\beta}\sqrt{\log(2en/\beta)}d\beta\\ &\le 8\ \mathfrak{R}_n(\mathcal{F}) + 32\sqrt{2}\ \mathfrak{R}_n(\mathcal{F})\ \log^{3/2}\left(\frac{en}{\mathfrak{R}_n(\mathcal{F})}\right)\ .\end{aligned}$$

Using the assumption $\mathfrak{R}_n(\mathcal{F}) \ge 1/n$ concludes the proof. We remark that the assumption is very mild and satisfied for any non-trivial class $\mathcal{F}$. □

Proof of Lemma 10. Let $(Z'_1, \ldots, Z'_n)$ be a decoupled sequence tangent to $(Z_1, \ldots, Z_n)$. Let us use the notation $\mathbb{E}_{t-1} f = \mathbb{E}\{f(Z'_t) \mid Z_{1:t-1}\}$. By Chebychev's inequality, for any $f \in \mathcal{F}$,

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{n}\left|\sum_{t=1}^{n}\left(f(Z'_t) - \mathbb{E}_{t-1}f\right)\right| > \alpha/2 \;\middle|\; Z_{1:n}\right) &\le \frac{\mathbb{E}\left[\left(\sum_{t=1}^n \left(f(Z'_t) - \mathbb{E}_{t-1}f\right)\right)^2 \;\middle|\; Z_{1:n}\right]}{n^2\alpha^2/4} \\
&= \frac{\sum_{t=1}^n \mathbb{E}\left[\left(f(Z'_t) - \mathbb{E}_{t-1}f\right)^2 \;\middle|\; Z_{1:n}\right]}{n^2\alpha^2/4} \\
&\le \frac{4n}{n^2\alpha^2/4} = \frac{16}{n\alpha^2}.
\end{aligned}$$

The second step is due to the fact that the cross terms are zero:

$$\mathbb{E}\left\{\left(f(Z'_t) - \mathbb{E}_{t-1}f\right)\left(f(Z'_s) - \mathbb{E}_{s-1}f\right) \;\middle|\; Z_{1:n}\right\} = 0 \; .$$

Hence

$$\inf_{f \in \mathcal{F}} \mathbb{P}\left(\frac{1}{n}\left|\sum_{t=1}^{n}\left(f(Z'_t) - \mathbb{E}_{t-1}f\right)\right| \le \alpha/2 \;\middle|\; Z_{1:n}\right) \ge 1 - \frac{16}{n\alpha^2} \ge \frac{1}{2}$$

whenever $\alpha^2 \ge \frac{32}{n}$. Given $Z_1, \ldots, Z_n$, let f^* be the function that maximizes $\frac{1}{n}\left|\sum_{t=1}^n \left(f(Z_t) - \mathbb{E}_{t-1}f\right)\right|$. Then

$$\begin{aligned}
\frac{1}{2} &\le \inf_{f \in \mathcal{F}} \mathbb{P}\left(\frac{1}{n}\left|\sum_{t=1}^{n}\left(f(Z'_t) - \mathbb{E}_{t-1}f\right)\right| \le \alpha/2 \;\middle|\; Z_{1:n}\right) \\
&\le \mathbb{P}\left(\frac{1}{n}\left|\sum_{t=1}^{n}\left(f^*(Z'_t) - \mathbb{E}_{t-1}f^*]\right)\right| \le \alpha/2 \;\middle|\; Z_{1:n}\right) \; .
\end{aligned}$$

Define the event $A = \left\{\sup_{f \in \mathcal{F}} \frac{1}{n}\left|\sum_{t=1}^n (f(Z_t) - \mathbb{E}_{t-1}f)\right| > \alpha\right\}$. We can thus assert that

$$\frac{1}{2} \le \mathbb{P}\left(\frac{1}{n}\left|\sum_{t=1}^{n}\left(f^*(Z'_t) - \mathbb{E}_{t-1}f^*\right)\right| \le \alpha/2 \;\middle|\; A\right) \; .$$

This, in turn, implies that

$$\begin{aligned}
\frac{1}{2}\mathbb{P}\left(\sup_{f \in \mathcal{F}} \frac{1}{n}\left|\sum_{t=1}^{n}\left(f(Z_t) - \mathbb{E}_{t-1}f\right)\right| > \alpha\right) &\le \mathbb{P}\left(\frac{1}{n}\left|\sum_{t=1}^{n}\left(f^*(Z'_t) - \mathbb{E}_{t-1}f^*\right)\right| \le \alpha/2 \;\middle|\; A\right) \\
&\quad \times \mathbb{P}\left(\sup_{f \in \mathcal{F}} \frac{1}{n}\left|\sum_{t=1}^{n}\left(f(Z_t) - \mathbb{E}_{t-1}f\right)\right| > \alpha\right) \; .
\end{aligned}$$

The latter product is a joint probability that can be further upper bounded by

$$\mathbb{P}\left(\frac{1}{n}\left|\sum_{t=1}^{n}\left(f^*(Z_t) - f^*(Z'_t)\right)\right| > \alpha/2\right) \le \mathbb{P}\left(\frac{1}{n}\sup_{f \in \mathcal{F}}\left|\sum_{t=1}^{n}\left(f(Z_t) - f(Z'_t)\right)\right| > \alpha/2\right) \; .$$

Now we apply Lemma 18 with $\phi(u) = \mathbf{1}\{u > n\alpha/2\}$,

$$\begin{aligned}
&\mathbb{E}\mathbf{1}\left\{\sup_{f \in \mathcal{F}}\left|\sum_{t=1}^{n} f(Z_t) - f(Z'_t)\right| > n\alpha/2\right\} \\
&\qquad \le \sup_{z_1, z'_1} \mathbb{E}_{\epsilon_1} \ldots \sup_{z_n, z'_n} \mathbb{E}_{\epsilon_n} \mathbf{1}\left\{\sup_{f \in \mathcal{F}}\left|\sum_{t=1}^{n} \epsilon_t\left(f(z_t) - f(z'_t)\right)\right| > n\alpha/2\right\} . \qquad (19)
\end{aligned}$$

Since

$$\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{n}\epsilon_t\left(f(z_t)-f(z_t')\right)\right| \le \sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{n}\epsilon_t f(z_t)\right| + \sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{n}\epsilon_t f(z_t')\right|$$

it is true that

$$\begin{aligned}\mathbf{1}\left\{\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{n}\epsilon_t\left(f(z_t)-f(z_t')\right)\right| > n\alpha/2\right\} &\le \mathbf{1}\left\{\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{n}\epsilon_t f(z_t)\right| > n\alpha/4\right\} \\ &\quad + \mathbf{1}\left\{\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{n}\epsilon_t f(z_t')\right| > n\alpha/4\right\} .\end{aligned}$$

An upper bound on the right-hand side of Eq. (19) is obtained by splitting into two parts:

$$\begin{aligned}&\sup_{z_1}\mathbb{E}_{\epsilon_1}\ldots\sup_{z_n}\mathbb{E}_{\epsilon_n}\mathbf{1}\left\{\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{n}\epsilon_t f(z_t)\right| > n\alpha/4\right\} + \sup_{z_1'}\mathbb{E}_{\epsilon_1}\ldots\sup_{z_n'}\mathbb{E}_{\epsilon_n}\mathbf{1}\left\{\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{n}\epsilon_t f(z_t')\right| > n\alpha/4\right\} \\ &= 2\sup_{z_1}\mathbb{E}_{\epsilon_1}\ldots\sup_{z_n}\mathbb{E}_{\epsilon_n}\mathbf{1}\left\{\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{n}\epsilon_t f(z_t)\right| > n\alpha/4\right\} .\end{aligned}$$

Moving to the tree representation (see proof of Theorem 2),

$$\begin{aligned}\mathbb{P}\left(\frac{1}{n}\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{n}\left(f(Z_t)-f(Z_t')\right)\right| > \alpha/2\right) &\le 2\sup_{\mathbf{z}}\mathbb{E}\left[\mathbf{1}\left\{\frac{1}{n}\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{n}\epsilon_t f(\mathbf{z}_t(\epsilon))\right| > \alpha/4\right\}\right] \\ &= 2\sup_{\mathbf{z}}\mathbb{P}\left(\frac{1}{n}\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{n}\epsilon_t f(\mathbf{z}_t(\epsilon))\right| > \alpha/4\right) .\end{aligned}$$

We can now conclude that

$$\mathbb{P}\left(\frac{1}{n}\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{n}\left(f(Z_t)-\mathbb{E}_{t-1}f\right)\right| > \alpha\right) \le 4\sup_{\mathbf{z}}\ \mathbb{P}_{\epsilon}\left(\frac{1}{n}\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{n}\epsilon_t f(\mathbf{z}_t(\epsilon))\right| > \alpha/4\right) .$$

□

Proof of Lemma 11. Fix a $\mathcal{Z}$-valued tree $\mathbf{z}$ of depth n. Let V be a minimum $\alpha/8$-cover of $\mathcal{F}$ over $\mathbf{z}$ with respect to ℓ_1. Corollary 6 ensures that

$$|V| = \mathcal{N}_1(\alpha/8,\mathcal{F},\mathbf{z}) \le \left(\frac{16en}{\alpha}\right)^{\mathrm{fat}_{\frac{\alpha}{8}}} .$$

By definition, for any $f\in\mathcal{F}$ and $\epsilon\in\{\pm1\}^n$, there exists $\mathbf{v}[f,\epsilon]\in V$ such that

$$\frac{1}{n}\sum_{t=1}^{n}|f(\mathbf{z}_t(\epsilon)) - \mathbf{v}[f,\epsilon]_t(\epsilon)| \le \alpha/8,$$

implying that

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{n}\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{n}\epsilon_t f(\mathbf{z}_t(\epsilon))\right| > \alpha/4\right)\\
&= \mathbb{P}\left(\frac{1}{n}\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{n}\epsilon_t\left(f(\mathbf{z}_t(\epsilon)) - \mathbf{v}[f,\epsilon]_t(\epsilon) + \mathbf{v}[f,\epsilon]_t(\epsilon)\right)\right| > \alpha/4\right)\\
&\le \mathbb{P}\left(\frac{1}{n}\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{n}\epsilon_t\left(f(\mathbf{z}_t(\epsilon)) - \mathbf{v}[f,\epsilon]_t(\epsilon)\right)\right| + \frac{1}{n}\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{n}\epsilon_t\mathbf{v}[f,\epsilon]_t(\epsilon)\right| > \alpha/4\right)\\
&\le \mathbb{P}\left(\frac{1}{n}\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{n}\epsilon_t\mathbf{v}[f,\epsilon]_t(\epsilon)\right| > \alpha/8\right) .
\end{aligned}$$

For fixed ϵ,

$$\frac{1}{n}\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{n}\epsilon_t\mathbf{v}[f,\epsilon]_t(\epsilon)\right| > \alpha/8 \qquad \Longrightarrow \qquad \frac{1}{n}\max_{\mathbf{v}\in V}\left|\sum_{t=1}^{n}\epsilon_t\mathbf{v}_t(\epsilon)\right| > \alpha/8$$

and, therefore, for any $\mathbf{z}$,

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{n}\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{n}\epsilon_t f(\mathbf{z}_t(\epsilon))\right| > \alpha/4\right) \;\le\; \mathbb{P}\left(\frac{1}{n}\max_{\mathbf{v}\in V}\left|\sum_{t=1}^{n}\epsilon_t\mathbf{v}_t(\epsilon)\right| > \alpha/8\right)\\
&\le \sum_{\mathbf{v}\in V}\mathbb{P}\left(\frac{1}{n}\left|\sum_{t=1}^{n}\epsilon_t\mathbf{v}_t(\epsilon)\right| > \alpha/8\right) \;\le\; 2|V|e^{-n\alpha^2/128} \;\le\; 2\left(\frac{16en}{\alpha}\right)^{\mathrm{fat}_{\frac{\alpha}{8}}} e^{-n\alpha^2/128} .
\end{aligned}$$

□

Proof of Lemma 12. We first prove the second inequality in Lemma 12. The proof closely follows that of Theorem 4. Define $\beta_0 = 1$ and $\beta_j = 2^{-j}$. For a fixed tree $\mathbf{z}$ of depth n, let V_j be an β_j-cover with respect to ℓ_∞. For any path $\epsilon \in \{\pm1\}^n$ and any $f \in \mathcal{F}$, let $\mathbf{v}[f,\epsilon]^j \in V_j$ a β_j-close element of the cover in the ℓ_∞ sense. Now, for any $f \in \mathcal{F}$,

$$\begin{aligned}
\left|\frac{1}{n}\sum_{t=1}^{n}\epsilon_t f(\mathbf{z}_t(\epsilon))\right| &\le \left|\frac{1}{n}\sum_{t=1}^{n}\epsilon_t(f(\mathbf{z}_t(\epsilon)) - \mathbf{v}[f,\epsilon]_t^N)\right| + \sum_{j=1}^{N}\left|\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\left(\mathbf{v}[f,\epsilon]_t^j - \mathbf{v}[f,\epsilon]_t^{j-1}\right)\right|\\
&\le \max_{t=1,\dots,n}\left|f(\mathbf{z}_t(\epsilon)) - \mathbf{v}[f,\epsilon]_t^N\right| + \sum_{j=1}^{N}\left|\frac{1}{n}\sum_{t=1}^{n}\epsilon_t(\mathbf{v}[f,\epsilon]_t^j - \mathbf{v}[f,\epsilon]_t^{j-1})\right| .
\end{aligned}$$

Thus,

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{t=1}^{n}\epsilon_t f(\mathbf{z}_t(\epsilon))\right| \le \beta_N + \sup_{f\in\mathcal{F}}\left\{\sum_{j=1}^{N}\left|\frac{1}{n}\sum_{t=1}^{n}\epsilon_t(\mathbf{v}[f,\epsilon]_t^j - \mathbf{v}[f,\epsilon]_t^{j-1})\right|\right\} .$$

We now proceed to upper bound the second term. Consider all possible pairs of $\mathbf{v}^s \in V_j$ and $\mathbf{v}^r \in V_{j-1}$, for $1 \le s \le |V_j|$, $1 \le r \le |V_{j-1}|$, where we assumed an arbitrary enumeration of elements. For each pair $(\mathbf{v}^s, \mathbf{v}^r)$, define a real-valued tree $\mathbf{w}^{(s,r)}$ by

$$\mathbf{w}_t^{(s,r)}(\epsilon) = \begin{cases} \mathbf{v}_t^s(\epsilon) - \mathbf{v}_t^r(\epsilon) & \text{if there exists } f \in \mathcal{F} \text{ s.t. } \mathbf{v}^s = \mathbf{v}[f,\epsilon]^j, \mathbf{v}^r = \mathbf{v}[f,\epsilon]^{j-1}\\ 0 & \text{otherwise.}\end{cases}$$

for all $t \in [n]$ and $\epsilon \in \{\pm 1\}^n$. It is crucial that $\mathbf{w}^{(s,r)}$ can be non-zero only on those paths ϵ for which $\mathbf{v}^s$ and $\mathbf{v}^r$ are indeed the members of the covers (at successive resolutions) close in the ℓ_∞ sense *to some* $f \in \mathcal{F}$. It is easy to see that $\mathbf{w}^{(s,r)}$ is well-defined. Let the set of trees W_j be defined as

$$W_j = \left\{ \mathbf{w}^{(s,r)} : 1 \le s \le |V_j|, 1 \le r \le |V_{j-1}| \right\} \ .$$

Using the above notations we see that

$$\begin{aligned}
\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{t=1}^{n} \epsilon_t f(\mathbf{x}_t(\epsilon)) \right| &\le \beta_N + \sup_{f \in \mathcal{F}} \left\{ \sum_{j=1}^{N} \left| \frac{1}{n} \sum_{t=1}^{n} \epsilon_t (\mathbf{v}[f,\epsilon]_t^j - \mathbf{v}[f,\epsilon]_t^{j-1}) \right| \right\} \\
&\le \beta_N + \sum_{j=1}^{N} \sup_{\mathbf{w}^j \in W_j} \left| \frac{1}{n} \sum_{t=1}^{n} \epsilon_t \mathbf{w}_t^j(\epsilon) \right| \ . \qquad (20)
\end{aligned}$$

Similarly to the proof of Theorem 4, we can show that $\max_{t=1}^n |\mathbf{w}_t^j(\epsilon)| \le 3\beta_j$ for any $\mathbf{w}^j \in W_j$ and any path ϵ. In the remainder of the proof we will use the shorthand $\mathcal{N}_\infty(\beta) = \mathcal{N}_\infty(\beta, \mathcal{F}, n)$. By Azuma-Hoeffding inequality for real-valued martingales,

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{t=1}^{n} \epsilon_t \mathbf{w}_t^j(\epsilon) \right| > 3\theta\beta_j \sqrt{\log \mathcal{N}_\infty(\beta_j)} \right) \le 2 \exp\left\{ -\frac{n\theta^2 \log \mathcal{N}_\infty(\beta_j)}{2} \right\} \ .$$

Hence, by union bound,

$$\mathbb{P}\left(\sup_{\mathbf{w}^j \in W_j} \left| \frac{1}{n} \sum_{t=1}^{n} \epsilon_t \mathbf{w}_t^j(\epsilon) \right| > 3\theta\beta_j \sqrt{\log \mathcal{N}_\infty(\beta_j)} \right) \le 2\mathcal{N}_\infty(\beta_j)^2 \ \exp\left\{ -\frac{n\theta^2 \log \mathcal{N}_\infty(\beta_j)}{2} \right\}$$

and so

$$\begin{aligned}
&\mathbb{P}\left(\exists j \in [N], \quad \sup_{\mathbf{w}^j \in W_j} \left| \frac{1}{n} \sum_{t=1}^{n} \epsilon_t \mathbf{w}_t^j(\epsilon) \right| > 3\theta\beta_j \sqrt{\log \mathcal{N}_\infty(\beta_j)} \right) \\
&\qquad \le 2 \sum_{j=1}^{N} \mathcal{N}_\infty(\beta_j)^2 \ \exp\left\{ -\frac{n\theta^2 \log \mathcal{N}_\infty(\beta_j)}{2} \right\} \ .
\end{aligned}$$

Hence,

$$\mathbb{P}\left(\sum_{j=1}^{N} \sup_{\mathbf{w}^j \in W_j} \left| \frac{1}{n} \sum_{t=1}^{n} \epsilon_t \mathbf{w}_t^j(\epsilon) \right| > 3\theta \sum_{j=1}^{N} \beta_j \sqrt{\log \mathcal{N}_\infty(\beta_j)} \right) \le 2 \sum_{j=1}^{N} \mathcal{N}_\infty(\beta_j)^2 \ \exp\left\{ -\frac{n\theta^2 \log \mathcal{N}_\infty(\beta_j)}{2} \right\}.$$

Using the above with Equation (20) yields

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{t=1}^{n} \epsilon_t f(\mathbf{z}_t(\epsilon)) \right| > \beta_N + 3\theta \sum_{j=1}^{N} \beta_j \sqrt{\log \mathcal{N}_\infty(\beta_j)} \right) \le 2 \sum_{j=1}^{N} \exp\left\{ \log \mathcal{N}_\infty(\beta_j) \left(2 - \frac{n\theta^2}{2} \right) \right\}.$$

Since we assume that $3 < \frac{n\theta^2}{4}$ and $\mathcal{N}_\infty(\beta_1) \ge e$, the right-hand side of the last inequality is bounded above by

$$2 \sum_{j=1}^{N} \exp\left\{ \log \mathcal{N}_\infty(\beta_j) \left(-1 - \frac{n\theta^2}{4} \right) \right\} \le 2e^{-\frac{n\theta^2}{4}} \sum_{j=1}^{N} \mathcal{N}_\infty(\beta_j)^{-1} \le 2Le^{-\frac{n\theta^2}{4}}.$$

Now picking N appropriately and bounding the sum by an integral,

$$\beta_N + 3\theta \sum_{j=1}^{N} \beta_j \sqrt{\log \mathcal{N}_\infty(\beta_j)} \le \inf_{\alpha>0} \left\{ 4\alpha + 6\theta \int_\alpha^1 \sqrt{\log \mathcal{N}_\infty(\delta)} d\delta \right\} .$$

Hence we conclude that

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} \left| \frac{1}{n} \sum_{t=1}^{n} \epsilon_t f(\mathbf{z}_t(\epsilon)) \right| > \inf_{\alpha>0} \left\{ 4\alpha + 6\theta \int_\alpha^1 \sqrt{\log \mathcal{N}_\infty(\delta, \mathcal{F}, n)} d\delta \right\} \right) \le 2Le^{-\frac{n\theta^2}{4}} .$$

This proves the second inequality of Lemma 12. To prove the first inequality, we note that by a proof identical to that of Lemma 9 (with the factor of $\theta/2$ replacing $1/\sqrt{n}$),

$$\inf_{\alpha>0} \left\{ 4\alpha + 6\theta \int_\alpha^1 \sqrt{\log \mathcal{N}_\infty(\delta, \mathcal{F}, n)} d\delta \right\} \le \mathfrak{R}_n(\mathcal{F}) \cdot 8 \left(1 + 2\sqrt{2}\theta \sqrt{n \log^3(en^2)} \right) .$$

□

Proof of Lemma 13. Without loss of generality we may assume that the Lipschitz constant $L = 1$ because the general case follows by scaling ϕ. Then $\phi \circ \mathcal{F} \subseteq [\phi(0) - 1, \phi(0) + 1]^{\mathcal{Z}}$. The centering $\phi(0)$ does not play a role in the proof of Theorem 4, and we conclude that

$$\mathfrak{R}_n(\phi \circ \mathcal{F}) \le \inf_\alpha \left\{ 4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^1 \sqrt{\log\ \mathcal{N}_2(\delta, \phi \circ \mathcal{F}, n)}\ d\delta \right\} . \tag{21}$$

Fix a $\mathcal{Z}$-valued tree $\mathbf{z}$ of depth n. Suppose V is a minimal δ-cover with respect to ℓ_∞ for $\mathcal{F}$ on the tree $\mathbf{z}$. Consider the set

$$V^\phi = \{ \mathbf{v}^\phi \ :\ \mathbf{v} \in V,\ \ \mathbf{v}^\phi_t(\epsilon) = \phi(\mathbf{v}_t(\epsilon)) \} .$$

For any $f \in \mathcal{F}$ and any $\epsilon \in \{\pm 1\}^n$, there is a representative $\mathbf{v} \in V$ such that

$$\max_{t\in[n]} \left| \phi(f(\mathbf{z}_t(\epsilon))) - \mathbf{v}^\phi_t(\epsilon) \right| = \max_{t\in[n]} |\phi(f(\mathbf{z}_t(\epsilon))) - \phi(\mathbf{v}_t(\epsilon))| \le \max_{t\in[n]} |f(\mathbf{z}_t(\epsilon))) - \mathbf{v}_t(\epsilon)| \le \delta .$$

Hence,

$$\log\ \mathcal{N}_2(\delta, \phi \circ \mathcal{F}, n) \le \log\ \mathcal{N}_\infty(\delta, \phi \circ \mathcal{F}, n) \le \log\ \mathcal{N}_\infty(\delta, \mathcal{F}, n),$$

which, together with Equation (21), implies

$$\mathfrak{R}_n(\phi \circ \mathcal{F}) \le \mathfrak{D}^\infty_n(\mathcal{F}) .$$

Invoking Lemma 9 concludes the proof. □

Proof of Proposition 14. The results follow similarly to Theorem 15 in [7] and we provide the proofs for completeness. Note that, unlike the Rademacher complexity defined in [7], sequential Rademacher complexity considered in this paper does not have the absolute value around the sum.

Part 1 is immediate because for any fixed tree $\mathbf{z}$ and fixed realization of $\{\epsilon_t\}$,

$$\sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(\mathbf{z}_t(\epsilon)) \le \sup_{f\in\mathcal{G}} \sum_{t=1}^{n} \epsilon_t f(\mathbf{z}_t(\epsilon)) ,$$

Taking expectation over ϵ and supremum over $\mathbf{z}$ completes the argument. To show Part 2, first observe that, according to Part 1,

$$\mathfrak{R}_n(\mathcal{F}) \le \mathfrak{R}_n(\mathrm{conv}(\mathcal{F})) \ .$$

Now, any $h \in \mathrm{conv}(\mathcal{F})$ there exists an $m \ge 1$ such that h can be written as $h = \sum_{j=1}^m \alpha_j f_j$ with $\sum_{j=1}^m \alpha_j = 1$, $\alpha_j \ge 0$. Then, for fixed tree $\mathbf{z}$ and sequence ϵ,

$$\sum_{t=1}^n \epsilon_t h(\mathbf{z}_t(\epsilon)) = \sum_{j=1}^m \alpha_j \sum_{t=1}^n \epsilon_t f_j(\mathbf{z}_t(\epsilon)) \le \sup_{f\in\mathcal{F}} \sum_{t=1}^n \epsilon_t f(\mathbf{z}_t(\epsilon))$$

and thus

$$\sup_{h\in\mathrm{conv}(\mathcal{F})} \sum_{t=1}^n \epsilon_t h(\mathbf{z}_t(\epsilon)) \le \sup_{f\in\mathcal{F}} \sum_{t=1}^n \epsilon_t f(\mathbf{z}_t(\epsilon)) \ .$$

Taking expectation over ϵ and supremum over $\mathbf{z}$ completes the proof.

To prove Part 3, first observe that the statement follows directly from the definition for $c \ge 0$. Hence, it remains to prove the statement for $c = -1$. Consider a tree $\mathbf{z}^R$ that is a reflection of $\mathbf{z}$. That is, $\mathbf{z}_t^R(\epsilon) = \mathbf{z}_t(-\epsilon)$ for all $t \in [n]$. It is then enough to observe that

$$\begin{aligned}\mathfrak{R}_n(-\mathcal{F}) &= \mathbb{E}_\epsilon\left[\sup_{f\in-\mathcal{F}} \sum_{t=1}^n \epsilon_t f(\mathbf{z}_t(\epsilon))\right] = \mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^n -\epsilon_t f(\mathbf{z}_t(\epsilon))\right] \\ &= \mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^n \epsilon_t f(\mathbf{z}_t(-\epsilon))\right] = \mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^n \epsilon_t f(\mathbf{z}_t^R(\epsilon))\right]\end{aligned}$$

where we used the fact that ϵ and $-\epsilon$ have the same distribution. As $\mathbf{z}$ varies over all trees, $\mathbf{z}^R$ also varies over all trees. Hence taking the supremum over $\mathbf{z}$ above finishes the argument.

Finally, for Part 4,

$$\sup_{f\in\mathcal{F}} \left\{ \sum_{t=1}^n \epsilon_t \left(f+h\right)\left(\mathbf{z}_t(\epsilon)\right) \right\} = \left\{ \sup_{f\in\mathcal{F}} \sum_{t=1}^n \epsilon_t f(\mathbf{z}_t(\epsilon)) \right\} + \left\{ \sum_{t=1}^n \epsilon_t h(\mathbf{z}_t(\epsilon)) \right\}.$$

Note that, since $h(\mathbf{z}_t(\epsilon))$ only depends on $\epsilon_{1:t-1}$, we have $\mathbb{E}_\epsilon\left[\epsilon_t h(\mathbf{z}_t(\epsilon))\right] = 0$. This concludes the proof. □

Proof of Lemma 15. We divide the proof into two cases. If $\alpha \le 512\sqrt{2\log^3(en^2)}\mathfrak{R}_n(\mathcal{F})$ then

$$L\exp\left(-\frac{\alpha^2}{256^2 \cdot 2 \cdot \log^3(en^2)\mathfrak{R}_n^2(\mathcal{F})}\right) \ge L/e^4 > 1$$

by our assumption. In this case the statement of the lemma is trivially satisfied with $c = 2^{17}$. Now consider the case $\alpha > 512\sqrt{2\log^3(en^2)}\mathfrak{R}_n(\mathcal{F})$. By Lemma 10,

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{t=1}^n \left(f(Z_t) - \mathbb{E}_{t-1}\left[f(Z_t)\right]\right)\right| > \alpha\right) \qquad \le \qquad 4\sup_{\mathbf{z}} \qquad \mathbb{P}_\epsilon\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{t=1}^n \epsilon_t f(\mathbf{z}_t(\epsilon))\right| > \alpha/4\right).$$

For an upper bound on the last quantity, we turn to Lemma 12. For any $\theta > 0$

$$\mathbb{P}_\epsilon\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{t=1}^{n}\epsilon_t f(\mathbf{z}_t(\epsilon))\right| > 8\left(1+\sqrt{96\log^3(en^2)}+\theta\sqrt{8n\log^3(en^2)}\right)\cdot\mathfrak{R}_n(\mathcal{F})\right) \le 2Le^{-\frac{n\theta^2}{4}} .$$

Since $1+\sqrt{96\log^3(en^2)} \le 4\sqrt{8\log^3(en^2)}$ we can conclude that for any $\theta > 0$,

$$\mathbb{P}_\epsilon\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{t=1}^{n}\epsilon_t f(\mathbf{z}_t(\epsilon))\right| > 16\sqrt{2\log^3(en^2)}\left(4+\theta\sqrt{n}\right)\cdot\mathfrak{R}_n(\mathcal{F})\right) \le 2Le^{-\frac{n\theta^2}{4}}.$$

The above inequality with $\theta = \frac{\alpha}{64\sqrt{2n\log^3(en^2)}\mathfrak{R}_n(\mathcal{F})} - \frac{4}{\sqrt{n}}$ yields

$$\mathbb{P}_\epsilon\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{t=1}^{n}\epsilon_t f(\mathbf{z}_t(\epsilon))\right| > \alpha/4\right) \le 2Le^{-\frac{n\theta^2}{4}}.$$

Now since we are considering the case when $\alpha > 512\sqrt{2\log^3(en^2)}\mathfrak{R}_n(\mathcal{F})$ we have that $\theta > \frac{\alpha}{128\sqrt{2n\log^3(en^2)}\mathfrak{R}_n(\mathcal{F})}$ and hence we can conclude that

$$\mathbb{P}_\epsilon\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{t=1}^{n}\epsilon_t f(\mathbf{z}_t(\epsilon))\right| > \alpha/4\right) \le 2L\exp\left(-\frac{\alpha^2}{256^2\cdot 2\cdot\log^3(en^2)\mathfrak{R}_n^2(\mathcal{F})}\right) .$$

Thus we have proved that in both the cases considered, we have that

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{t=1}^{n}\left(f(Z_t)-\mathbb{E}_{t-1}\left[f(Z_t)\right]\right)\right| > \alpha\right) \le 8L\exp\left(-\frac{\alpha^2}{256^2\cdot 2\cdot\log^3(en^2)\mathfrak{R}_n^2(\mathcal{F})}\right).$$

This concludes the proof, and we may choose $c = 2^{17}$ (we did not attempt to optimize constants). □

Proof of Proposition 16. We use linearity of the functions to write

$$n\cdot\mathfrak{R}_n(\mathcal{F}) = \sup_{\mathbf{z}}\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^{n}\epsilon_t\left\langle f,\mathbf{z}_t(\epsilon)\right\rangle\right] = \sup_{\mathbf{z}}\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\left\langle f,\sum_{t=1}^{n}\epsilon_t\mathbf{z}_t(\epsilon)\right\rangle\right] .$$

Let $\Psi^\star$ be the Fenchel conjugate of Ψ. By Fenchel-Young inequality, for any $\lambda > 0$,

$$\left\langle f,\sum_{t=1}^{n}\epsilon_t\mathbf{z}_t(\epsilon)\right\rangle \le \frac{\Psi(f)}{\lambda} + \frac{\Psi^\star\left(\sum_{t=1}^n\lambda\,\epsilon_t\,\mathbf{z}_t(\epsilon)\right)}{\lambda} .$$

Taking supremum over $f\in\mathcal{F}$, we get,

$$\sup_{f\in\mathcal{F}}\left\langle f,\sum_{t=1}^{n}\epsilon_t\mathbf{z}_t(\epsilon)\right\rangle \le \frac{\Psi_{\max}}{\lambda} + \frac{\Psi^\star\left(\sum_{t=1}^n\lambda\,\epsilon_t\,\mathbf{z}_t(\epsilon)\right)}{\lambda} . \qquad (22)$$

Since Ψ is uniformly convex, its conjugate $\Psi^\star$ is $(1/\sigma,p)$-uniformly smooth,

$$\Psi^\star(z_1) \le \Psi^\star(z_2) + \langle\nabla\Psi^\star(z_2), z_1-z_2\rangle + \frac{1}{p\sigma}\|z_1-z_2\|^p.$$

By taking expectation w.r.t. ϵ in (22) and repeatedly applying the above inequality (proof can be found e.g. in [14]),

$$\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\left\langle f, \sum_{t=1}^n \epsilon_t \mathbf{z}_t(\epsilon)\right\rangle\right] \le \frac{\Psi_{\max}}{\lambda} + \frac{\lambda^p n}{p\,\sigma\,\lambda}$$

since $\|\mathbf{z}_t(\epsilon)\| \le 1$. Simplifying and optimizing over $\lambda > 0$, gives

$$n \cdot \mathfrak{R}_n(\mathcal{F}) \le \left(\frac{p}{p-1}\right)^{\frac{p-1}{p}} \frac{\Psi_{\max}^{\frac{p-1}{p}}\, n^{\frac{1}{p}}}{\sigma^{\frac{1}{p}}} .$$

□

References

[1] J. Abernethy, A. Agarwal, P. Bartlett, and A. Rakhlin. A stochastic view of optimal regret through minimax duality. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.

[2] T. M. Adams and A. B. Nobel. Uniform convergence of Vapnik-Chervonenkis classes under ergodic sampling. *The Annals of Probability*, 38(4):1345–1367, 2010.

[3] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM*, 44:615–631, 1997.

[4] N. Alon and J. Spencer. *The Probabilistic Method*. John Wiley & Sons, 2nd edition, 2000.

[5] P. L. Bartlett, P. M. Long, and R. C. Williamson. Fat-shattering and the learnability of real-valued functions. In *Proceedings of the 7th Annual ACM Conference on Computational Learning Theory*, pages 299–310. ACM Press, 1994.

[6] P. L. Bartlett, P. M. Long, and R. C. Williamson. Fat-shattering and the learnability of real-valued functions. *Journal of Computer and System Sciences*, 52(3):434–452, 1996.

[7] P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2003.

[8] S. Ben-David, D. Pal, and S. Shalev-Shwartz. Agnostic online learning. In *Proceedings of the 22th Annual Conference on Learning Theory*, 2009.

[9] V. de la Peña and E. Giné. *Decoupling: From Dependence to Independence*. Springer, 1998.

[10] H. Dehling, T. Mikosch, and M. Sørensen. *Empirical Process Techniques for Dependent Data*. Birkhäuser, 2002.

[11] R. M. Dudley. *Uniform Central Limit Theorems*. Cambridge University Press, 1999.

[12] R. M. Dudley, E. Gine, and J. Zinn. Uniform and universal Glivenko-Cantelli classes. *Journal of Theoretical Probability*, 4:485–510, 1991.

[13] R. M. Dudley, H. Kunita, F. Ledrappier, and P. L. Hennequin. A course on empirical processes. In *École d'Été de Probabilités de Saint-Flour XII - 1982*, volume 1097 of *Lecture Notes in Mathematics*, pages 1–142. Springer, 1984.

[14] S. M. Kakade, K. Sridharan, and A. Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in Neural Information Processing Systems 21*, pages 793–800. MIT Press, 2009.

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

[16] M. Ledoux and M. Talagrand. *Probability in Banach Spaces.* Springer-Verlag, New York, 1991.

[17] N. Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2(4):285–318, 04 1988.

[18] S. Mendelson. A few notes on statistical learning theory. In S. Mendelson and A. J. Smola, editors, *Advanced Lectures in Machine Learning, LNCS 2600, Machine Learning Summer School 2002, Canberra, Australia, February 11-22*, pages 1–40. Springer, 2003.

[19] S. Mendelson and R. Vershynin. Entropy and the combinatorial dimension. *Inventiones mathematicae*, 152(1):37–55, 2003.

[20] A. Nobel and A. Dembo. A note on uniform laws of averages for dependent processes. *Statistics and Probability Letters*, 17:169–172, 1993.

[21] I. Pinelis. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, 22(4):1679–1706, 1994.

[22] G. Pisier. Martingales with values in uniformly convex spaces. *Israel Journal of Mathematics*, 20:326–350, 1975.

[23] D. Pollard. *Empirical Processes: Theory and Applications*, volume 2 of *NSF-CBMS Regional Conference Series in Probability and Statistics.* Institute of Mathematical Statistics, Hayward, CA, 1990.

[24] A. Rakhlin, K. Sridharan, and A. Tewari. Online learning: Random averages, combinatorial parameters, and learnability. *Advances in Neural Information Processing Systems 23*, pages 1984–1992, 2010.

[25] M. Rudelson and R. Vershynin. Combinatorics of random processes and sections of convex bodies. *The Annals of Mathematics*, 164(2):603–648, 2006.

[26] N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13:145–147, 1972.

[27] S. Shelah. A combinatorial problem: Stability and order for models and theories in infinitary languages. *Pacific Journal of Mathematics*, 4:247–261, 1972.

[28] J. M. Steele. Empirical discrepancies and subadditive processes. *The Annals of Probability*, 6(1):118–127, 1978.

[29] M. Talagrand. The Glivenko-Cantelli problem. *Annals of Probability*, 15:837–870, 1987.

[30] M. Talagrand. The Glivenko-Cantelli problem, ten years later. *Journal of Theoretical Probability*, 9(2):371–384, 1996.

[31] M. Talagrand. *The Generic Chaining: Upper and Lower Bounds for Stochastic Processes.* Springer, 2005.

[32] A. W. Van Der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics.* Springer Series, March 1996.

[33] S. A. van de Geer. *Empirical Processes in M-Estimation.* Cambridge University Press, 2000.

[34] S. A. van de Geer. On Hoeffding's inequality for dependent random variables. In *Empirical process techniques for dependent data*, pages 161–169. Springer, 2002.

[35] R. van Handel. The universal Glivenko–Cantelli property. *Probability Theory and Related Fields*, pages 1–24, 2012.

[36] 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.

[37] V.N. Vapnik and A. Ya. Chervonenkis. The necessary and sufficient conditions for the uniform convergence of averages to their expected values. *Teoriya Veroyatnostei i Ee Primeneniya*, 26(3):543–564, 1981.

[38] B. Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, 22(1):pp. 94–116, 1994.

[39] J. E. Yukich. Rates of convergence for classes of functions: The non-i.i.d. case. *Journal of Multivariate Analysis*, 20(2):175 – 189, 1986.