TU Delft Delft University of Technology

Error bounds for learning in reproducing kernel Banach spaces

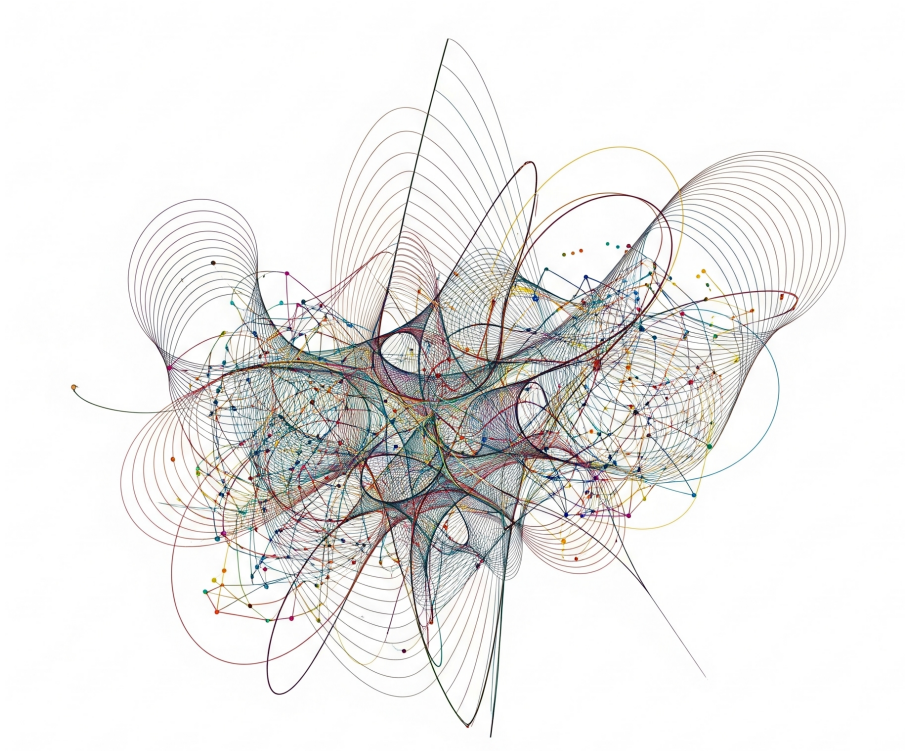
MASTER'S THESIS IN APPLIED MATHEMATICS

Supervised by Francesca Bartolucci and Nicholas Nelsen

Author:

Jiahong Liu

August 20, 2025



Contents

1 Introduction

Understanding errors in neural networks is a challenging problem in theoretical machine learning. There are two types of error bounds in theoretical machine learning analysis depending on whether the true function $f^\star$ or the learned model $\hat{f}_N$ enters into the bounds [33]. The former is referred to as an "a priori" error estimate and the latter as an "a posteriori" error estimate [33]. It was observed in [1, 10, 21] that the numerical values of the error bounds in the "a posteriori" error estimates are very large for real neural network models.

In this paper, we pursue a different line of inquiry by conducting an "a priori" error estimate on the regularized $f^\star$-population risk for one hidden layer neural networks of possibly infinite width in reproducing kernel Banach spaces (RKBSs). We work on a supervised learning setting with a squared loss function, where the goal of the learning is to estimate a function $\hat{f}_N : \mathcal{X} \to \mathcal{Y}$, given a set of input-output pairs $D = \{(x_1, y_1), \cdots, (x_N, y_N)\} \subset \mathcal{X} \times \mathcal{Y}$, so that the function $\hat{f}_N$ predicts well the paired output of a new input data, and the measurement of the "estimation error" is in L^2 sense.

Throughout this paper, we denote $\mathcal{X}$ as the input space and $\mathcal{Y}$ the output space. We view $\mathcal{X}$ and $\mathcal{Y}$ as measurable spaces equipped with their respective Borel σ-algebras. We denote $\nu \in \mathscr{P}(\mathcal{X})$ as the input distribution, and Θ as the parameter space. Let $\mathcal{M}(\Theta)$ be the Banach space of bounded measures defined on the Borel $\sigma-$algebra of Θ, and endow $\mathcal{M}(\Theta)$ with the total variation norm $\|\cdot\|_{\mathrm{TV}}$. The hypotheses space of the neural network models that we will be working on is the RKBS $(\mathcal{B}, \|\cdot\|_{\mathcal{B}})$ of real-valued functions on $\mathcal{X}$ defined by the feature map $\phi : \mathcal{X} \to \mathcal{M}(\Theta)'$ given by

$$\phi(x)(\mu) = {}_{\mathcal{M}(\Theta)}\langle \mu, \phi(x) \rangle_{\mathcal{M}(\Theta)'} = \int_\Theta \varphi(x, \theta) d\mu(\theta)$$

for every $x \in \mathcal{X}$ and $\mu \in \mathcal{M}(\Theta)$. It was shown in [6] that

$$\mathcal{B} = \{f_\mu : \mathcal{X} \to \mathbb{R} \mid \|\mu\|_{\mathrm{TV}} < \infty\}, \quad \text{where } f_\mu = \int_\Theta \varphi(\cdot, \theta) d\mu(\theta), \tag{1}$$

is a RKBS and the norm on $\mathcal{B}$ is the quotient norm

$$\|f\|_{\mathcal{B}} = \inf_{f = f_\mu} \|\mu\|_{\mathrm{TV}}.$$

If we choose $\mu = \sum_{k=1}^K \delta_{\theta_k} a_k$, where δ_θ is the Dirac measure at point θ, then

$$\int_\Theta \varphi(x, \theta) d\mu(\theta) = \sum_{k=1}^K a_k \varphi(x, \theta_k)$$

is the one hidden layer neural networks with finite width. In some literature, it is also referred to as two-layer neural networks.

1.1 Related work

The theoretical analysis of two-layer neural networks traces back to the work by Andrew Barron [4, 5]. Follow-up studies such as [2, 11, 13] have extended Barron's results in various aspects. Especially, [33] has conducted an "a priori" error estimate for the population risk for two-layer neural networks in Barron spaces.

Another line of research for feature-based models studies random feature models (RFMs), proposed in [23, 24]. The RFM is based on a linear combination of random features (RFs), which serves as a randomized basis. The RFMs provide computationally efficient approximation of underlying kernel methods [23, 24]. Existing theoretical analysis of RFMs such as [3, 8, 17] mainly focus on target functions in the corresponding reproducing kernel Hilbert space (RKHS) [1]. Especially, [14] has conducted an "a priori" error estimate for the population risk for RFMs in a RKHS.

There exists a theoretical equivalence between one hidden layer neural networks and RFMs. Specifically, a one hidden layer neural network can be conceptualized as an RFM in which the weight distribution is not random but is instead learned through an optimization process [11].

This paper is a subsequent study of Bartolucci et al. [6] and Lanthaler and Nelsen [14]. [6] shows that a wide class of reproducing kernel Banach spaces in the form of (1) can be well represented by neural networks. In fact, these spaces admit an integral representation and include one hidden layer neural networks of possibly infinite width. In [14], a squared error bound for the population risk of RFMs is established by decomposing the error into two components: the regularized empirical risk and the approximation error. The goal of this paper is to establish an "a priori" error estimate for the population risk for one hidden layer neural networks in RKBSs. We adopt the same error decomposition proof framework as in [14]. The research gap addressed by this work

stems from two key departures from existing literature: our focus on RKBSs instead of the more commonly studied RKHSs, and our use of integral representations as a generalization of linear combinations.

Drawing from the rich body of tools available in statistical learning theory (see, e.g., [22, 29, 32]), we establish our estimation bounds. Our approach employs a covering-number framework, which relies on fundamental concepts such as Rademacher complexity, symmetrization, Dudley's theorem, and the VC-dimension.

Let N denote the number of samples and d denote the dimension of the parameter space. For a one-hidden-layer neural network with N training samples, [33] derived a risk bound of $\mathcal{O}\big(1/M + \ln(N)\sqrt{\ln(d)/N}\big)$, where M denotes the number of parameters, under the assumption that the activation function is 1-Lipschitz continuous and positively homogeneous. This result is established in a Barron space setting for finite-length neural networks. Our result gives a risk bound of $\mathcal{O}(N^{-1/\max\{2,d\}}\sqrt{\ln N})$, working on the underlying RKBS of possibly infinite-length neural networks.

1.2 Outline

The paper is organized as follows. In Section 2 we recall the main ideas and results in measure theory, probability theory, statistical learning, and kernel methods. This section includes the methods and theoretical tools used to derive our main result. In Section 3 we review the paper by Lanthaler and Nelsen [14] which provides an upper bound of the population risk of RFMs in RKHSs. The proof framework of our main result follows the proof framework outlined in Section 3.2. In Section 4 we present the main result of the paper. In particular, in Section 4.1 we discuss the setting and assumptions of our theoretical analysis. Then, we state the main theorem of the paper. In Section 4.2, we present the full proof of the main theorem. In Section 4.3, we discuss some consequences of the main theorem and outline some limitations and future work.

2 Preliminaries

Section 2 reviews essential background knowledge used in the main result of the thesis. In section 2.1, we summarize essential definitions, propositions, and theorems from measure theory and probability theory. In section 2.2, we introduce the main yield of the thesis, that is, the scheme of supervised learning and the formulation of empirical risk minimization problems. In section 2.3, we introduce the objects that we aim to study in the thesis, that is, the neural networks. In section 2.4, we study the concepts of reproducing kernel Hilbert spaces and reproducing kernel Banach spaces. These concepts are important to the thesis because we consider them as the hypothesis spaces of random features and neural networks. In section 2.5, we study some existing results in terms of the reproducing kernel Banach spaces of neural networks.

2.1 Measure theory and probability theory

Empirical processes theory often relies on pseudometric spaces, because the empirical L^p distances of N data points are pseudometric. Pseudometric spaces generalize the definition of metric spaces.

Definition 2.1 (Pseudometric space)**.** *A pseudometric space (S, ρ) is a set S together with a non-negative real-valued function $\rho : S \times S \to \mathbb{R}_{\geq 0}$, called a pseudometric, such that for every $x, y, z \in S$,*

(i) $\rho(x, x) = 0$*,*

(ii) $\rho(x, y) = \rho(y, x)$ *(symmetry),*

(iii) $\rho(x, z) \leq \rho(x, y) + \rho(y, z)$ *(triangle inequality).*

Remark 2.1. *Unlike a metric space, points in a pseudometric space need not be distinguishable; that is, one may have $\rho(x, y) = 0$ for distinct values $x \neq y$.*

Although points in a pseudometric space may not be distinguishable, we may still define covers on pseudometric spaces. The ϵ-covering number of a compact subset of a pseudometric space describes the structure and the level of compactness of the set.

Definition 2.2 (ϵ-cover and ϵ-covering number)**.** *Let (S, ρ) be a pseudometric space and $\epsilon > 0$. A subset $\widehat{T} \subseteq S$ is called an ϵ-cover of $T \subseteq S$ if for every $m \in T$, there exists an $m' \in \widehat{T}$ such that $\rho(m, m') \leq \epsilon$. The set $\widehat{T} \subseteq S$ is a minimal ϵ-cover of $T \subseteq S$ if there is no other ϵ-cover with lower cardinality. The ϵ-covering number of T is the cardinality of any minimal ϵ-cover of T, that is*

$$\mathfrak{C}(\epsilon; T, \rho) = \min\{|\widehat{T}| : \widehat{T} \text{ is an } \epsilon\text{-cover of } T\}. \tag{2}$$

There exist various known results that bound the covering numbers in finite dimensional spaces.

Example 2.1 ([27], Example 27.1)**.** *Suppose that $T \subset \mathbb{R}^t$, let $c = \max_{\mathbf{m}\in T} \|\mathbf{m}\|_2$, and assume that T lies in a d-dimensional subspace of $\mathbb{R}^t$. Then,*

$$\mathfrak{C}(\epsilon; T, \|\cdot\|_2) \leq \left(\frac{2c\sqrt{d}}{\epsilon}\right)^d.$$

To see this, let $\mathbf{v}_1, \ldots, \mathbf{v}_d$ be an orthonormal basis of the subspace. Then, any $\mathbf{m} \in T$ can be written as $\mathbf{m} = \sum_{i=1}^d \alpha_i \mathbf{v}_i$ with $\|\boldsymbol{\alpha}\|_\infty \leq \|\boldsymbol{\alpha}\|_2 = \|\mathbf{m}\|_2 \leq c$. Let $\delta \in \mathbb{R}$ and consider the set

$$\widehat{T} = \left\{ \sum_{i=1}^d \alpha_i' \mathbf{v}_i : \forall i, \alpha_i' \in \{-c, -c+\delta, -c+2\delta, \ldots, c\} \right\}.$$

Given $\mathbf{m} \in T$ s.t. $\mathbf{m} = \sum_{i=1}^d \alpha_i \mathbf{v}_i$ with $\|\boldsymbol{\alpha}\|_\infty \leq c$. Then, there exists $\mathbf{m}' \in \widehat{T}$ such that

$$\|\mathbf{m} - \mathbf{m}'\|^2 = \left\| \sum_{i=1}^d (\alpha_i' - \alpha_i) \mathbf{v}_i \right\|^2 \leq \delta^2 \sum_{i=1}^d \|\mathbf{v}_i\|^2 \leq \delta^2 d.$$

Choose $\delta = \epsilon/\sqrt{d}$, then $\|\mathbf{m} - \mathbf{m}'\| \leq \epsilon$. Therefore, $\widehat{T}$ is an ϵ-cover of T. Hence,

$$\mathfrak{C}(\epsilon; T, \|\cdot\|) \leq \left|\widehat{T}\right| = \left(\frac{2c}{\delta}\right)^d = \left(\frac{2c\sqrt{d}}{\epsilon}\right)^d.$$

In [28], a more precise result is provided that gives both upper and lower bounds for the covering number of a compact subset of a Euclidean space. This is often referred to as the volumetric bound of the covering number in Euclidean space.

Example 2.2 ([28], Exercise 2.5.9, pp. 36-37)**.** *Suppose that $T \subset \mathbb{R}^t$, let $c = \max_{\mathbf{m}\in T} \|\mathbf{m}\|_2$, and assume that T lies in a d-dimensional subspace of $\mathbb{R}^t$. Then,*

$$\epsilon^{-d} \lesssim \mathfrak{C}(\epsilon; T, \|\cdot\|_2) \leq \left(\frac{2c}{\epsilon} + 1\right)^d.$$

Sometimes, directly bounding the covering number of a target set can be challenging. However, the following lemma shows that the covering number of the image of a set under a Lipschitz map can be controlled by the covering number of its pre-image, scaled by the Lipschitz constant. Consequently, if we can express the target set as the image of a Lipschitz map applied to another set with a known covering number bound, we can obtain a corresponding bound for the target set.

Lemma 2.1 ([22], Lemma 14.12, pp. 206)**.** *Let $(X_1, \rho_1), (X_2, \rho_2)$ be two metric spaces and let $f : X_1 \to X_2$ be Lipschitz continuous with Lipschitz constant C_{Lip}. For every relatively compact $T \subseteq X_1$, it holds that for all $\varepsilon > 0$,*

$$\mathfrak{C}\left(C_{\mathrm{Lip}}\varepsilon; f(T), \rho\right) \leq \mathfrak{C}\left(\varepsilon; T, \rho\right).$$

Proof. Fix $\varepsilon > 0$ and let $N := \mathcal{C}(\varepsilon; T, \rho_1)$. Thus, there exist points $x_1, \ldots, x_N \in X_1$ such that

$$T \subseteq \bigcup_{i=1}^N B_{\rho_1}(x_i, \varepsilon).$$

Now, for any $t \in T$, there exists some i such that $\rho_1(t, x_i) \leq \varepsilon$. By the Lipschitz continuity of f, we have

$$\rho_2\big(f(t), f(x_i)\big) \leq C_{\mathrm{Lip}} \rho_1(t, x_i) \leq C_{\mathrm{Lip}}\varepsilon.$$

Thus, $f(t) \in B_{\rho_2}(f(x_i), C_{\mathrm{Lip}}\varepsilon)$, and hence

$$f(T) \subseteq \bigcup_{i=1}^N B_{\rho_2}(f(x_i), C_{\mathrm{Lip}}\varepsilon).$$

This shows that $f(T)$ can be covered by N balls of radius $C_{\mathrm{Lip}}\varepsilon$ in X_2. □

In machine learning theory, we often define functions taking parameters in the space of measures. It is a known result that the space of measures is a Banach space when equipped with the TV-norm. The concept of the TV-norm relies on signed measures.

Definition 2.3 (Signed measure). *Given a measurable space $(\mathcal{X}, \Sigma)$ (that is, a set $\mathcal{X}$ with a σ-algebra Σ on it), a signed measure is a function*

$$\mu : \Sigma \to \mathbb{R}$$

such that $\mu(\emptyset) = 0$ and μ is σ-additive – that is, it satisfies the equality

$$\mu\big(\cup_{n=1}^{\infty} A_n\big) = \sum_{n=1}^{\infty} \mu(A_n)$$

for any sequence $A_1, A_2, \cdots, A_n, \cdots$ of disjoint sets in Σ.

Definition 2.4 (Total variation). *Given a signed measure μ on a measurable space $(\mathcal{X}, \Sigma)$, define*

$$\mu^+(E) = \sup\{\mu(A) : A \in \Sigma, A \subset E\}, \quad \forall E \in \Sigma,$$
$$\mu^-(E) = -\inf\{\mu(A) : A \in \Sigma, A \subset E\}, \quad \forall E \in \Sigma.$$

Then, the total variation of μ is defined as

$$\|\mu\|_{\mathrm{TV}} = |\mu|(\mathcal{X}) = \mu^+(\mathcal{X}) + \mu^-(\mathcal{X}).$$

Bochner spaces are a generalization of the concept of L^p spaces to functions whose values lie in a Banach space which is not necessarily the space $\mathbb{R}$ or $\mathbb{C}$.

Definition 2.5 (Bochner spaces). *Let $(\mathcal{X}, \mathcal{M}, \mu)$ be a measure space, where $\mathcal{M}$ is a σ-algebra on the set $\mathcal{X}$ and μ is a measure. Let $(\mathcal{Y}, \|\cdot\|_{\mathcal{Y}})$ be a Banach space. For $1 \leq p < \infty$, the space $L^p_\mu(\mathcal{X}; \mathcal{Y})$ consists of all functions $f : \mathcal{X} \to \mathcal{Y}$ that are Bochner measurable and have a finite norm, defined as:*

$$\|f\|_{L^p_\mu(\mathcal{X};\mathcal{Y})} = \left(\int_{\mathcal{X}} \|f(x)\|^p_{\mathcal{Y}}\, d\mu(x)\right)^{1/p} < \infty\,.$$

For the case of $p = \infty$, the space is defined by the essential supremum norm:

$$\|f\|_{L^\infty_\mu(\mathcal{X};\mathcal{Y})} = \mu\text{-ess sup}\, f := \inf\left\{a \in \mathcal{Y} \,\middle|\, \mu\big(\{x \in \mathcal{X} : f(x) > a\}\big) = 0\right\} < \infty\,.$$

A classical result establishes that if a class has a finite VC index (or VC dimension), then its covering number admits a polynomial upper bound, with the exponent determined by the index. We first give a definition of the VC index.

Definition 2.6 ([29], Section 2.6.1, pp. 134-135). *Let $\mathcal{C}$ be a collection of subsets of a set $\mathcal{X}$. Let $x_1, \cdots, x_n$ be elements in $\mathcal{X}$. Denote*

$$\Delta_n\left(\mathcal{C}, x_1, \ldots, x_n\right) = \#\left\{C \cap \{x_1, \ldots, x_n\} : C \in \mathcal{C}\right\}.$$

The VC index of $\mathcal{C}$, denoted by $V(\mathcal{C})$, is defined as

$$V(\mathcal{C}) = \inf\left\{n : \max_{x_1,\ldots,x_n} \Delta_n\left(\mathcal{C}, x_1, \ldots, x_n\right) < 2^n\right\}.$$

Here, the infimum over the empty set is taken to be infinity.

Definition 2.7 (VC-class). *A collection of measurable sets $\mathcal{C}$ is called a VC-class if its VC index is finite.*

In the context of functional analysis, we can define the VC-classes of functions by means of the notion of the subgraph of a function. Recall that the subgraph of a real-valued function is the subset of $\mathcal{X} \times \mathbb{R}$ given by

$$\{(x, t) : t < f(x)\}.$$

A collection $\mathcal{F}$ of measurable functions on a sample space $\mathcal{X}$ is called a VC-subgraph class (or just a VC-class), if the collection of all subgraphs of the functions in $\mathcal{F}$ forms a VC-class of sets (in $\mathcal{X} \times \mathbb{R}$). The VC-index of the set of subgraphs of functions in $\mathcal{F}$ is denoted by $V(\mathcal{F})$. Now, we are ready to state the classical result.

Theorem 2.2 ([29], Theorem 2.6.7, pp. 141). *For a VC-class of functions with measurable envelope function F and $p \geq 1$, one has that for any probability measure μ with $\|F\|_{L^p_\mu(\mathcal{X};\mathbb{R})} > 0$,*

$$\mathfrak{C}\left(\varepsilon\|F\|_{L^p_\mu(\mathcal{X};\mathbb{R})}; \mathcal{F}, \|\cdot\|_{L^p_\mu(\mathcal{X};\mathbb{R})}\right) \leq KV(\mathcal{F})(16e)^{V(\mathcal{F})}\left(\frac{1}{\varepsilon}\right)^{p(V(\mathcal{F})-1)}$$

for a universal constant K and $0 < \varepsilon < 1$.

Proof. See van der Vaart and Wellner [29, Theorem 2.6.7, pp. 141]. □

Now, we move on to the preliminary results in probability theory. We first list some fundamental definitions.

Definition 2.8 (Probability space)**.** *A probability space is a triple $(\Omega, \Sigma, \mathbb{P})$ consisting of:*

(i) the sample space Ω – an arbitrary non-empty set;

(ii) the σ-algebra $\Sigma \subseteq 2^\Omega$ – a set of subsets of Ω, called events, such that:

(a) Σ contains the sample space: $\Omega \in \Sigma$;

(b) Σ is closed under complements: if $A \in \Sigma$, then also $\Omega \setminus A \in \Sigma$;

(c) Σ is closed under countable unions: if $A_i \in \Sigma$ for $i = 1, 2 \cdots$, then also $(\cup_{i=1}^\infty A_i) \in \Sigma$;

(iii) the probability measure $\mathbb{P} : \Sigma \to [0, 1]$ is a function on Σ such that:

(a) $\mathbb{P}$ is σ-additive: if $\{A_i\}_{i=1}^\infty \subseteq \Sigma$ is a countable collection of pairwise disjoint sets, then $\mathbb{P}(\cup_{i=1}^\infty A_i) = \sum_{i=1}^\infty \mathbb{P}(A_i)$;

(b) the measure of the entire sample space is equal to one: $\mathbb{P}(\Omega) = 1$.

Definition 2.9 (Random variable)**.** *Let $(\Omega, \Sigma, \mathbb{P})$ be a probability space and $(E, \mathcal{E})$ a measurable space. Then, an $(E, \mathcal{E})$-valued random variable X is a measurable function $X : \Omega \to E$ from the sample space Ω to a measurable space E, that is, for every subset $B \in \mathcal{E}$, the preimage of X is Σ-measurable ($X^{-1}(B) \in \Sigma$, where $X^{-1}(B) = \{\omega : X(\omega) \in B\}$).*

Definition 2.10 (Expected value of real-valued random variables)**.** *Let X be a real-valued random variable defined on a probability space $(\Omega, \Sigma, \mathbb{P})$. Then, the expected value of X, denoted by $\mathbb{E}[X]$, is defined as the Lebesgue integral*

$$\mathbb{E}[X] = \int_\Omega X d\mathbb{P}.$$

Bernstein inequality is a powerful concentration inequality that is widely used in high-dimensional probability.

Theorem 2.3 (Vector-valued Bernstein inequality in Hilbert space)**.** *Let Z be an H-valued random variable, where $(H, \langle\cdot,\cdot\rangle, \|\cdot\|)$ is a separable Hilbert space. Suppose there exist positive numbers $b' > 0$ and $\sigma > 0$ such that*

$$\mathbb{E}\|Z - \mathbb{E}Z\|^p \le \frac{1}{2} p! \sigma^2 (b')^{p-2} \quad \text{for all} \quad p \ge 2. \tag{3}$$

For any $\delta \in (0, 1)$ and $N \in \mathbb{N}$, denoting by $\{Z_n\}_{n=1}^N$ a sequence of N iid copies of Z, it holds that

$$\mathbb{P}\left\{ \left\| \frac{1}{N} \sum_{n=1}^N Z_n - \mathbb{E}Z \right\| \le \frac{2b' \log(2/\delta)}{N} + \sqrt{\frac{2\sigma^2 \log(2/\delta)}{N}} \right\} \ge 1 - \delta, \tag{4}$$

where we use $\Pr$ to denote probability with respect to the underlying probability space.

Proof. See Lanthaler and Nelsen [14, Supplementary Material A, Theorem A.1]. □

The most common use case of Bernstein's inequality is in the following bounded setting.

Lemma 2.4. *([14], Supplementary Material A, Lemma A.2)**.** Let Z be a (potentially) uncentered random variable such that*

$$\|Z\| \le c \text{ almost surely} \quad \text{and} \quad \mathbb{E}\|Z - \mathbb{E}Z\|^2 \le v^2$$

for some $c > 0$ and $v > 0$. Then Z satisfies Bernstein's moment condition (3) with $b' = 2c$ and $\sigma = v$. If $\mathbb{E}Z = 0$, then taking $b' = c$ suffices.

Proof. The proof follows the original proof in the paper. By assumption we have that $\mathbb{E}\|Z\| \le c$. We establish a bound on the centered random variable. By the triangle inequality, we have $\|Z - \mathbb{E}Z\| \le \|Z\| + \|\mathbb{E}Z\|$. Using the inequality $\|\mathbb{E}Z\| \le \mathbb{E}\|Z\|$ and our initial assumption, we get

$$\|Z - \mathbb{E}Z\| \le \|Z\| + \mathbb{E}\|Z\| \le c + c = 2c$$

almost surely. Now, we use this bound to compute the p-th moment of the centered variable. For any integer $p \geq 2$, we can write

$$\begin{aligned}\mathbb{E}\|Z - \mathbb{E}Z\|^p &= \mathbb{E}\left[\|Z - \mathbb{E}Z\|^2 \cdot \|Z - \mathbb{E}Z\|^{p-2}\right] \\ &\leq \mathbb{E}\|Z - \mathbb{E}Z\|^2 \cdot (2c)^{p-2} \\ &\leq v^2(2c)^{p-2}\,.\end{aligned}$$

Finally, for any $p \geq 2$, we know that $1 \leq \frac{1}{2}p!$. Therefore, we can write the final bound as:

$$\mathbb{E}\|Z - \mathbb{E}Z\|^p \leq \frac{1}{2}p!v^2(2c)^{p-2}\,.$$

□

We may extend the use of the Bernstein inequality to Banach spaces.

Theorem 2.5 (Vector-valued Bernstein inequality in Banach space)**.** *Let Z be a $\mathcal{Z}$-valued random variable, where $(\mathcal{Z}, \|\cdot\|)$ is a separable Banach space. Suppose there exist positive numbers $b' > 0$ and $\sigma > 0$ such that*

$$\mathbb{E}\|Z - \mathbb{E}Z\|^p \leq \frac{1}{2}p!\sigma^2(b')^{p-2} \quad \textit{for all} \quad p \geq 2. \tag{5}$$

For any $\delta \in (0,1)$ and $N \in \mathbb{N}$, denoting by $\{Z_n\}_{n=1}^N$ a sequence of N i.i.d. copies of Z, it holds that

$$\Pr\left\{\left\|\frac{1}{N}\sum_{n=1}^N Z_n\right\| - \mathbb{E}\left\|\frac{1}{N}\sum_{n=1}^N Z_n\right\| \leq \frac{2b'\log(1/\delta)}{N} + \sqrt{\frac{2\sigma^2\log(1/\delta)}{N}}\right\} \geq 1 - \delta. \tag{6}$$

Proof. See Lanthaler and Nelsen [14, Supplementary Material A, Theorem A.3]. □

Subexponential variables generalize the concept of sub-Gaussian variables.

Definition 2.11 (Subexponential random variables)**.** *For a random variable Z with values in a Banach space $(\mathcal{Z}, \|\cdot\|_{\mathcal{Z}})$, its subexponential norm is defined by*

$$\|Z\|_{\psi_1(\mathcal{Z})} := \sup_{p\in[1,\infty)} \frac{(\mathbb{E}\|Z\|_{\mathcal{Z}}^p)^{1/p}}{p}. \tag{7}$$

Z is subexponential if its subexponential norm is finite.

Remark 2.2. *In theorem 2.5, a random variable Z satisfying the Bernstein moment condition (5) is subexponential in the sense that $\|Z - \mathbb{E}Z\|$ is subexponential on $\mathbb{R}$, i.e., exhibits exponential tail decay.*

It is a known result in probability theory that a subexponential variable concentrates around its expectation in Bernstein sense.

Proposition 2.6 (Subexponential implies Bernstein moment condition)**.** *Let Z be a $(\mathcal{Z}, \|\cdot\|)$-valued subexponential random variable, that is, $\|Z\|_{\psi_1(\mathcal{Z})} < \infty$. Then Z satisfies*

$$\mathbb{E}\|Z - \mathbb{E}Z\|^p \leq \frac{1}{2}p!\sigma^2(b')^{p-2} \quad \textit{for all} \quad p \geq 2,$$

where

$$\sigma^2 := 4e\sqrt{\mathbb{E}\|Z - \mathbb{E}Z\|^2}\|Z\|_{\psi_1(\mathcal{Z})} \quad \textit{and} \quad b' := 4e\|Z\|_{\psi_1(\mathcal{Z})}.$$

Proof. See Lanthaler and Nelsen [14, Supplementary Material A, Proposition A.4]. □

The sum of subexponential variables is subexponential. The following corollary is important in the thesis.

Corollary 2.7 (Subexponential tail bound in Banach space)**.** *Fix $N \in \mathbb{N}$. Let $\{Z_n\}_{n=1}^N$ be iid random variables with values in a separable Banach space $(\mathcal{Z}, \|\cdot\|)$. Suppose that $\|Z_1\|_{\psi_1(\mathcal{Z})} < \infty$. Let $S_N := \frac{1}{N}\sum_{n=1}^N Z_n$. Fix $\delta \in (0,1)$. With probability at least $1 - \delta$, it holds that*

$$\|S_N\| - \mathbb{E}\|S_N\| \leq \frac{8e\,\|Z_1\|_{\psi_1(\mathcal{Z})}\log(1/\delta)}{N} + \sqrt{\frac{8e\,\|Z_1 - \mathbb{E}Z_1\|_{L^2_{\mathrm{p}}(\Omega;\mathcal{Z})}\,\|Z_1\|_{\psi_1(\mathcal{Z})}\log(1/\delta)}{N}}. \tag{8}$$

In particular, if $N \geq \log(1/\delta)$, then with probability at least $1 - \delta$ it holds that

$$\|S_N\| - \mathbb{E}\|S_N\| \leq \sqrt{\frac{64e^3\,\|Z_1\|^2_{\psi_1(\mathcal{Z})}\log(1/\delta)}{N}}. \tag{9}$$

Proof. The proof of this corollary is similar to the proof of Lemma 4.2. This result is taken from Lanthaler and Nelsen [14, Supplementary Material A, Corollary A.5], which also contains the proof. □

More generally, we can extend the definition 2.11 to the Orlicz norms.

Definition 2.12 ([32], Definition 5.34)**.** *The ψ_q-Orlicz norm of a zero-mean random variable X is given by*

$$\|X\|_{\psi_q} := \inf\left\{\lambda > 0 \mid \mathbb{E}\left[\psi_q(|X|/\lambda)\right] \leq 1\right\}.$$

The Orlicz norm is infinite if there is no $\lambda \in \mathbb{R}$ for which the given expectation is finite.

Remark 2.3. *The subexponential norm defined in definition 2.11 is equivalent to the ψ_1-Orlicz norm where $\psi_q(x) = \exp(x^q) - 1$ and $q \in [1, 2]$.*

There are multiple equivalent definitions for subexponential variables.

Proposition 2.8 (Subexponential properties)**.** *$\|X\|_{\psi_1} \leq K$ can be equivalently reformulated as*

$$\begin{aligned}\mathbb{P}\{|X| > t\} &\leq 2\exp(-t/C_1K) \quad \forall t \geq 0,\\ \|X\|_{L^p} = (\mathbb{E}|X|^p)^{1/p} &\leq C_2Kp \quad \forall p \geq 1.\end{aligned}$$

Moreover, if $\mathbb{E}[X] = 0$, the following equivalent condition holds on the moment-generating function of X:

$$\mathbb{E}[\exp(\lambda X)] \leq \exp\left(C_3^2K^2\lambda^2\right) \quad \forall\lambda : |\lambda| \leq \frac{1}{C_3K}. \tag{10}$$

Proof. See Vershynin [31, Proposition 2.7.1]. □

Based on the notion of the Orlicz norm, we can define an interesting generalization of a sub-exponential process.

Definition 2.13 ([32], Definition 5.35)**.** *A zero-mean stochastic process $\{X_\theta, \theta \in \mathbb{T}\}$ is a ψ_q-process with respect to a metric ρ if*

$$\left\|X_\theta - X_{\widetilde{\theta}}\right\|_{\psi_q} \leq \rho(\theta, \widetilde{\theta}) \quad \text{for all } \theta, \widetilde{\theta} \in \mathbb{T}.$$

For the exponential-type functions considered here, note that we have

$$\psi_q^{-1}(u) = [\log(1+u)]^{1/q}.$$

With this set-up, we have the following result:

Theorem 2.9 (Tail bound of Orlicz processes)**.** *For a given parameter $q \in [1, 2]$, let $\{X_\theta, \theta \in \mathbb{T}\}$ be a ψ_q-process with respect to ρ. Then there is a universal constant c_1 such that*

$$\mathbb{P}\left[\sup_{\theta,\widetilde{\theta}\in\mathbb{T}} \left|X_\theta - X_{\widetilde{\theta}}\right| \geq c_1\left(\int_0^D \psi_q^{-1}(\mathfrak{C}(u;\mathbb{T},\rho))du + t\right)\right] \leq 2e^{-\frac{t^q}{D^q}} \quad \text{for all } t > 0,$$

where $D = \sup_{\theta,\widetilde{\theta}\in\mathbb{T}} \rho(\theta, \widetilde{\theta})$ is the diameter of the set $\mathbb{T}$ under ρ.

Proof. See Wainwright [32, Theorem 5.36, pp. 151-153]. □

The Rademacher complexity measures the richness of a class of sets with respect to a probability distribution.

Definition 2.14 (Rademacher complexity)**.** *Let $\mathscr{F}$ be a class of integrable real-valued functions with domain $\mathcal{X}$. For any fixed collection $x_1^n := (x_1, \ldots, x_n)$ of points, consider the subset of $\mathbb{R}^n$ given by*

$$\mathscr{F}\left(x_1^n\right) := \left\{\left(f\left(x_1\right), \ldots, f\left(x_n\right)\right) : f \in \mathscr{F}\right\}.$$

*The set $\mathscr{F}\left(x_1^n\right)$ corresponds to all those vectors in $\mathbb{R}^n$ that can be realized by applying a function $f \in \mathscr{F}$ to the collection $(x_1, \ldots, x_n)$, and the **empirical Rademacher complexity** is given by*

$$\mathcal{R}\left(\mathscr{F}\left(x_1^n\right)/n\right) := \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i f\left(x_i\right)\right|\right], \quad \sigma_i \overset{\text{i.i.d.}}{\sim} \text{Rademacher}.$$

*Given a collection $X_1^n := \{X_i\}_{i=1}^n$ of random samples, then the empirical Rademacher complexity $\mathcal{R}\left(\mathscr{F}\left(X_1^n\right)/n\right)$ is a random variable. Taking its expectation yields the **Rademacher complexity of the function class** $\mathscr{F}$*

$$\mathcal{R}_n(\mathscr{F}) := \mathbb{E}_X\left[\mathcal{R}\left(\mathscr{F}\left(X_1^n\right)/n\right)\right] = \mathbb{E}_{X,\boldsymbol{\sigma}}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i f\left(X_i\right)\right|\right], \quad \sigma_i \overset{\text{i.i.d.}}{\sim} \text{Rademacher}. \tag{11}$$

Symmetrization is an important technique that links the supremum of an empirical process to its Rademacher complexity. Notably, it can convert a class of possibly not centered functions to a class of conditionally centered functions.

Proposition 2.10. *[Symmetrization: sandwich bounds] Let $\mathscr{F}$ be a class of integrable real-valued functions with domain $\mathcal{X}$, and let $\{X_i\}_{i=1}^n$ be a collection of i.i.d. samples from some distribution $\mathbb{P}$ over $\mathcal{X}$. Consider the random variable*

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}} := \sup_{f\in\mathscr{F}} \left| \frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}[f(X)] \right|. \tag{12}$$

The symmetrized version of the random variable $\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}$ is

$$\|\mathbb{S}_n\|_{\mathscr{F}} := \sup_{f\in\mathscr{F}} \left| \frac{1}{n}\sum_{i=1}^n \sigma_i f(X_i) \right|, \quad \sigma_i \overset{\text{i.i.d.}}{\sim} \text{Rademacher}. \tag{13}$$

For any convex non-decreasing function $\Phi : \mathbb{R} \to \mathbb{R}$, we have

$$\mathbb{E}_{X,\boldsymbol{\sigma}}\left[\Phi\left(\frac{1}{2}\|\mathbb{S}_n\|_{\overline{\mathscr{F}}}\right)\right] \le \mathbb{E}_X\left[\Phi\left(\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}\right)\right] \le \mathbb{E}_{X,\boldsymbol{\sigma}}\left[\Phi\left(2\|\mathbb{S}_n\|_{\mathscr{F}}\right)\right],$$

where $\overline{\mathscr{F}} = \{f - \mathbb{E}[f] : f \in \mathscr{F}\}$ is the recentered function class.

Proof. See Wainwright [32, Section 4.2, Proposition 4.11, pp. 107–108]. □

Remark 2.4. *When applied with the convex non-decreasing function $\Phi(t) = t$, proposition 2.10 yields the inequalities*

$$\frac{1}{2}\mathbb{E}_{X,\boldsymbol{\sigma}}\|\mathbb{S}_n\|_{\overline{\mathscr{F}}} \le \mathbb{E}_X\left[\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}\right] \le 2\mathbb{E}_{X,\boldsymbol{\sigma}}\|\mathbb{S}_n\|_{\mathscr{F}}.$$

Massart's finite class lemma gives an upper bound of the Rademacher complexity of a finite class of functions. This lemma is vital in the technique of one-step discretization and Dudley's theorem.

Lemma 2.11 (Massart's lemma)**.** *Assume $|\mathscr{F}|$ is finite. Let $x_1^n = (x_1, \cdots, x_n)$ be a random i.i.d. sample, and let*

$$B = \max_{f\in\mathscr{F}} \left(\frac{1}{n}\sum_{i=1}^n f^2(x_i)\right)^{\frac{1}{2}},$$

then

$$\mathcal{R}(\mathscr{F}(x_1^n)/n) \le B\sqrt{\frac{2\log|\mathscr{F}|}{n}}.$$

Proof. For any $s > 0$, we have

$$\begin{aligned}
e^{sn\mathcal{R}(\mathscr{F}(x_1^n)/n)} &= e^{s\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathscr{F}}\sum_{i=1}^n \sigma_i f(x_i)\right]} \\
&\le \mathbb{E}_{\boldsymbol{\sigma}}\left[e^{s\sup_{f\in\mathscr{F}}\sum_{i=1}^n \sigma_i f(x_i)}\right] \qquad \text{(Jensen's inequality)} \\
&= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathscr{F}} e^{s\sum_{i=1}^n \sigma_i f(x_i)}\right] \\
&\le \sum_{f\in\mathscr{F}} \mathbb{E}_{\boldsymbol{\sigma}}\left[e^{s\sum_{i=1}^n \sigma_i f(x_i)}\right] \qquad \text{(inner part is positive)} \\
&= \sum_{f\in\mathscr{F}}\prod_{i=1}^n \mathbb{E}_{\boldsymbol{\sigma}}\left[e^{s\sigma_i f(x_i)}\right], \qquad \text{(independence of } \sigma)
\end{aligned}$$

where $\boldsymbol{\sigma} = \{\sigma_1, \sigma_2, \cdots, \sigma_n\}$ is the set of Rademacher random variables. Since for fixed i, $\mathbb{E}_{\boldsymbol{\sigma}}[\sigma_i f(x_i)] = 0$ and $-f(x_i) \le \sigma_i f(x_i) \le f(x_i)$, we can apply Hoeffding's Lemma ([12], which says that for X being a random variable with $X \in [a,b]$ and $\mathbb{E}[X] = 0$, then for every $\lambda > 0$, we have that $\mathbb{E}\left[e^{\lambda X}\right] \le e^{\lambda^2(b-a)^2/8}$) to obtain,

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[e^{s\sigma_i f(x_i)}\right] \le e^{s^2 f^2(x_i)/2}.$$

Plugging this into the previous inequality, we have

$$\begin{aligned} e^{sn\mathcal{R}(\mathscr{F}(x_1^n)/n)} &\le \sum_{f\in\mathscr{F}}\prod_{i=1}^{n} e^{s^2 f^2(x_i)/2} \\ &= \sum_{f\in\mathscr{F}} e^{\left(s^2/2\right)\sum_{i=1}^{n} f^2(x_i)} \\ &\le \sum_{f\in\mathscr{F}} e^{(sB)^2/2} \\ &= |\mathscr{F}| e^{(sB)^2/2}. \end{aligned}$$

Hence, for any $s > 0$,

$$\mathcal{R}(\mathscr{F}(x_1^n)/n) \le \frac{1}{n}\left(\frac{\log|\mathscr{F}|}{s} + \frac{sB^2}{2}\right).$$

By optimizing over s, we can find that setting $s = \frac{\sqrt{2\log|\mathscr{F}|}}{B}$ yields

$$\mathcal{R}(\mathscr{F}(x_1^n)/n) \le \frac{B\sqrt{2\log|\mathscr{F}|}}{n}.$$

□

Theorem 2.12 (Dudley's theorem)**.** *Let $\mathcal{F}$ be a class of real-valued functions. Write $\nu = \mathrm{Law}(x)$, $x_i \overset{\text{i.i.d.}}{\sim} \nu$, and $\nu_n \coloneqq \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$. Then*

$$\mathcal{R}(\mathcal{F}(x_1^n)/n) \le 12\int_0^\infty \sqrt{\frac{\log \mathfrak{C}\left(\varepsilon; \mathcal{F}, \mathsf{d}_{L^2_{\nu_n}(\mathcal{X};\mathbb{R})}\right)}{n}}\, d\varepsilon\,.$$

Proof. The proof is analogous to the one given in Section 4.2.1, particularly the "Chaining" part. Since this result is classical, it appears in many sources; see, for example, Ma [18, Theorem 4.26, pp. 40-43] for a dedicated proof. □

We notice that directly using the original Dudley's integral bound in our setting encounters difficulty as the Dudley's integral may have a singular point at zero. Therefore, we will apply the following result to deal with this issue.

Theorem 2.13 (Localized Dudley's theorem)**.** *Let $\mathcal{G}$ be a class of real-valued functions. Let $\alpha \ge 0$. Write $\nu = \mathrm{Law}(x)$, $x_i \overset{\text{i.i.d.}}{\sim} \nu$, and $\nu_n \coloneqq \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$. Then*

$$\mathcal{R}(\mathcal{G}(x_1^n)/n) \le 4\alpha + \frac{12}{\sqrt{n}}\int_\alpha^\infty \sqrt{\log \mathfrak{C}\left(\varepsilon; \mathcal{G}, \mathsf{d}_{L^2_{\nu_n}(\mathcal{X};\mathbb{R})}\right)}\, d\varepsilon\,.$$

The ∞ in the upper limit of the integral can be replaced with the $L^2_{\nu_N}(\mathcal{X};\mathbb{R})$-diameter of $\mathcal{G}$, and accordingly, the theorem holds for every fixed cutoff α less than the diameter.

Proof. The proof of this theorem is similar to the proof of the original Dudley's theorem, except that the iterative covering procedure is stopped at the threshold $\epsilon = \alpha$ at the cost of the extra 4α term above. □

In fact, Dudley's theorem applies to all zero-mean sub-Gaussian processes with respect to the induced pseudometric ρ_X. Recall that a collection of zero-mean random variables $\{X_\theta, \theta\in\mathbb{T}\}$ is a sub-Gaussian process with respect to a metric ρ_X on $\mathbb{T}$ if

$$\mathbb{E}\left[e^{\lambda(X_\theta - X_{\theta'})}\right] \le e^{\frac{\lambda^2\rho_X^2(\theta,\theta')}{2}} \quad \text{for all } \theta,\theta'\in\mathbb{T}, \text{ and } \lambda\in\mathbb{R}.$$

It is easy to verify that the canonical Gaussian processes and Rademacher processes are both sub-Gaussian processes with respect to the Euclidean metric $\|\theta-\theta'\|_2$. A result of Wainwright [32, Theorem 5.22, pp. 140-142] gives a form of truncated Dudley's entropy integral bound for zero-mean sub-Gaussian processes.

Theorem 2.14. *Let $\{X_\theta, \theta\in\mathbb{T}\}$ be a zero-mean sub-Gaussian process with respect to the induced pseudo-metric ρ_X. Then for any $\delta\in[0,D]$, we have*

$$\mathbb{E}\left[\sup_{\theta,\widetilde{\theta}\in\mathbb{T}}\left(X_\theta - X_{\widetilde{\theta}}\right)\right] \le 2\mathbb{E}\left[\sup_{\substack{\gamma,\gamma'\in\mathbb{T}\\ \rho_X(\gamma,\gamma')\le\delta}}\left(X_\gamma - X_{\gamma'}\right)\right] + 32\int_{\delta/4}^{D}\sqrt{\log\mathfrak{C}(\varepsilon;\mathbb{T},\rho_X)}\, d\varepsilon\,,$$

where $D = \sup_{\theta,\widetilde{\theta}\in\mathbb{T}}\rho_X(\theta,\widetilde{\theta})$.

Proof. See Wainwright [32, Theorem 5.22, pp. 140-142]. □

Motivated by this theorem, we extend Theorem 2.9 to a truncated form that incorporates a discretization error. This leads to the following result.

Theorem 2.15. *For a given parameter $q \in [1,2]$, let $\{X_\theta, \theta \in \mathbb{T}\}$ be a ψ_q-process with respect to ρ. Let $D = \sup_{\theta,\widetilde{\theta}\in\mathbb{T}} \rho(\theta,\widetilde{\theta})$ be the diameter of the set $\mathbb{T}$ under ρ. Let $D > \alpha \geq 0$. Then there is a universal constant c_2 such that*

$$\mathbb{P}\left[\sup_{\theta,\widetilde{\theta}\in\mathbb{T}} \left|X_\theta - X_{\widetilde{\theta}}\right| \geq c_2 \left(\int_{\alpha/4}^{D} \psi_q^{-1}(\mathfrak{C}(u;\mathbb{T},\rho))du + t\right)\right] \leq 2e^{-\frac{(t-E/c_2)^q}{D^q}} \quad \text{for all } t > 0\,,$$

where

$$E = 2\mathbb{E}\left[\sup_{\substack{\gamma,\gamma'\in\mathbb{T}\\ \rho(\gamma,\gamma')\leq\alpha}} (X_\gamma - X_{\gamma'})\right].$$

2.2 Supervised learning and empirical risk minimization

The task of supervised learning is to estimate a function $f : \mathcal{X} \to \mathbb{R}$, given a set of input-output pairs $D = \{(x_1,y_1),\cdots,(x_N,y_N)\} \subset \mathcal{X}\times\mathbb{R}$, so that the function f predicts well the paired output of a new input data. To formalize the problem, we assume that $\mathcal{X}\times\mathbb{R}$ is a probability space with distribution P and that the data set is sampled i.i.d. of P, that is, $D \sim P^N$. Then, the function of interest can be obtained by solving the minimization problem

$$\min_{f\in\mathcal{T}} \mathcal{L}(f), \quad \mathcal{L}(f) = \int L(y,f(x))dP(x,y),$$

where $\mathcal{L}(f)$ is called the expected risk, $\mathcal{T}$ is the largest space in which the expected risk is defined, and $L : \mathbb{R}\times\mathbb{R} \to [0,\infty)$ is a given loss function. A loss function measures the difference between the true label y and the predicted label $f(x)$. An example of a loss function is the squared loss function $L(y,f(x)) = |y-f(x)|^2$. Since we only have access to the training data set D, we estimate the expected risk using Monte Carlo integration. In practice, we restrict ourselves to a hypothesis space $\mathcal{H} \subset \mathcal{T}$ for searching of a solution of the minimization problem. Thus, we consider the empirical risk minimization (ERM)

$$\min_{f\in\mathcal{H}} \widehat{\mathcal{L}}(f), \quad \widehat{\mathcal{L}}(f) = \frac{1}{n}\sum_{i=1}^{N} L(y_i,f(x_i)).$$

As shown in [6], the usual hypothesis spaces used in both kernel methods and neural networks have the property

$$\min_{f\in\mathcal{H}} \mathcal{L}(f) = \min_{f\in\mathcal{T}} \mathcal{L}(f).$$

Hypothesis spaces that have the above property are sometimes called universal classes of functions [7]. In the case of kernel methods and neural networks, we often add a regularizing functional $J : \mathcal{H} \to \mathbb{R}$ to the ERM, that is, we consider

$$\min_{f\in\mathcal{H}} \widehat{\mathcal{L}}(f) + J(f).$$

The regularizer J measures the norm associated to the hypothesis space $\mathcal{H}$ of f. Thus, ERM is set to find a solution with small norm.

2.3 Neural networks

This subsection summarizes the results from [6, 15]. Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a continuous nonlinear function. A simplified model of a neuron is $\sigma(w\cdot x)$, where $w,x \in \mathbb{R}^d$. Here, x is a vector of inputs of the neuron, w is a vector of synaptic weights, $w\cdot x$ is the weighted sum of all the inputs to the neuron, $\sigma(\cdot)$ is an activation function. Given d inputs (i.e. x), a neuron is activated if $w\cdot x$ is greater than some threshold determined by $\sigma(\cdot)$. Examples of activation functions are the rectified linear unit (ReLU) $\sigma(x) = \max(0,x)$, and the sigmoid logistic function $\sigma(x) = 1/(1+e^{-x})$. A one hidden layer neural network is a function obtained as linear combination of several neurons

$$f(x) = \sum_{k=1}^{K} \alpha_k \sigma(w_k\cdot x + b_k), \tag{14}$$

where $w_k \in \mathbb{R}^d$ and $b_k \in \mathbb{R}$. Here, the number of neurons is K, which is called the width of the neural network.

2.4 Reproducing kernel Hilbert spaces and reproducing kernel Banach spaces

This section follows Bartolucci et al. [6, Section 2, Section 3]. The concept of RKHSs is important and widely-used in the literature of machine-learning theory.

Definition 2.15. *Let $\mathcal{X}$ be a set. A reproducing kernel Hilbert spaces (RKHS) $\mathcal{H}$ over $\mathcal{X}$ is a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$ such that:*

(i) as a vector space, $\mathcal{H}$ is endowed with the pointwise operations of sum and multiplication by a scalar;

(ii) for all $x \in \mathcal{X}$, there is a constant $C_x > 0$ such that

$$|f(x)| \leq C_x \|f\|_{\mathcal{H}}, \quad \forall f \in \mathcal{H}. \tag{15}$$

Remark 2.5. *Fix $x \in \mathcal{X}$. Define the point evaluation functional $ev_x : \mathcal{H} \to \mathbb{R}, ev_x f = f(x)$. Then, for $(f_n)_{n=0}^{\infty}$ being a sequence in $\mathcal{H}$ such that $f_n \to f$ in $\mathcal{H}$ as $n \to \infty$, we have that $|ev_x f_n - ev_x f| = |f_n(x) - f(x)| = |(f_n - f)(x)| \leq C_x(\|f_n - f\|_{\mathcal{H}})$. The right-hand side goes to zero as $n \to \infty$. Thus, ev_x is continuous on $\mathcal{H}$. By the Riesz representation theorem, property (15) is thus equivalent to the existence, for all $x \in \mathcal{X}$, of an element $K_x \in \mathcal{H}$ such that $f(x) = \langle f, K_x \rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$.*

Remark 2.5 leads to the following more practical characterization of RKHSs [1].

Proposition 2.16. *A Hilbert space $\mathcal{H}$ of functions on $\mathcal{X}$ is a RKHS if and only if there exists a function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that for all $x \in \mathcal{X}$*

1. $K(x, \cdot) \in \mathcal{H}$,
2. $f(x) = \langle f, K(x, \cdot) \rangle_{\mathcal{H}}, \quad \forall f \in \mathcal{H}$.

Proof. The "only if" direction follows from remark 2.5 by letting the function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ to satisfiy $K(x, \cdot) = K_x(\cdot)$ for all $x \in \mathcal{X}$. Now, we turn to the "if" direction. Since $\mathcal{H}$ is a Hilbert space, condition (i) of definition 2.15 is satisfied. It remains to show condition (ii) is satisfied. Let $x \in \mathcal{X}$ be arbitary. Fix $f \in \mathcal{H}$. Then, we have that $K(x, \cdot) \in \mathcal{H}$ and $|f(x)| = |\langle f, K(x, \cdot) \rangle_{\mathcal{H}}| \leq \|K(x, \cdot)\|_{\mathcal{H}} \|f\|_{\mathcal{H}}$ by Cauchy-Schwarz inequality. Thus, condition (ii) is satisfied with $C_x = \|K(x, \cdot)\|_{\mathcal{H}}$. □

Remark 2.6. *The function K is called the **reproducing kernel** and item (2) is called the **reproducing property**. Every reproducing kernel is symmetric and positive definite. It can be shown that each symmetric positive definite kernel can be used to define a unique RKHS [1].*

Another characterization of RKHSs that is popular in machine learning is the following proposition.

Proposition 2.17. *A Hilbert space $\mathcal{H}$ of functions on $\mathcal{X}$ is a RKHS if and only if there exists a Hilbert space $\mathcal{F}$ and a map $\phi : \mathcal{X} \to \mathcal{F}$ such that*

(i) $\mathcal{H} = \{f_w : w \in \mathcal{F}\}$ where $f_w(x) = \langle \phi(x), w \rangle_{\mathcal{F}}$;

(ii) $\|f\|_{\mathcal{H}} = \inf\{\|w\|_{\mathcal{F}} : w \in \mathcal{F}, f = f_w\}$.

Proof. Since Hilbert spaces are special cases of Banach spaces, the proof mimics the proof of proposition 2.18 with dual pairings being replaced by inner products. □

Remark 2.7. *The map ϕ is called **feature map** and $\mathcal{F}$ **feature space**. Each function in a RKHS can be seen as a hyperplane in the feature space.*

A reproducing kernel Banach space (RKBS) follows definition 2.15 with "Hilbert" being replaced by "Banach".

Definition 2.16. *Let $\mathcal{X}$ be a set. A reproducing kernel Banach spaces (RKBS) $\mathcal{B}$ over $\mathcal{X}$ is a Banach space of functions $f : \mathcal{X} \to \mathbb{R}$ such that:*

(i) as a vector space, $\mathcal{B}$ is endowed with the pointwise operations of sum and multiplication by a scalar;

(ii) for all $x \in \mathcal{X}$, there is a constant $C_x > 0$ such that

$$|f(x)| \leq C_x \|f\|_{\mathcal{B}}, \quad \forall f \in \mathcal{B}. \tag{16}$$

Similar to remark 2.5, property (16) is equivalent to the fact that for every $x \in \mathcal{X}$ there exists an element $ev_x \in \mathcal{B}'$ such that

$$f(x) = {}_{\mathcal{B}'}\langle ev_x, f\rangle_{\mathcal{B}}, \quad \forall f \in \mathcal{B}.$$

However, proposition 2.16 does not generalize to the Banach case, since in general the dual of $\mathcal{B}$ is not isomorphic to itself. Interestingly, the characterization of proposition 2.17 in terms of feature maps generalizes naturally [9, 16].

Proposition 2.18. *A Banach space $\mathcal{B}$ of functions on $\mathcal{X}$ is a RKBS if and only if there exists a Banach space $\mathcal{F}$ and a map $\phi : \mathcal{X} \to \mathcal{F}'$ such that*

(i) $\mathcal{B} = \{f_\mu : \mu \in \mathcal{F}\}$ *where* $f_\mu(x) = {}_{\mathcal{F}'}\langle \phi(x), \mu\rangle_{\mathcal{F}}$;

(ii) $\|f\|_{\mathcal{B}} = \inf\{\|\mu\|_{\mathcal{F}} : \mu \in \mathcal{F}, f = f_\mu\}$.

Proof. We begin with the "only if" direction. Let $\mathcal{B}$ be a RKBS of functions on $\mathcal{X}$. Define $\mathcal{F} = \mathcal{B}$ and the canonical feature map

$$\phi : \mathcal{X} \to \mathcal{B}', \quad \phi(x)(\cdot) = ev_x(\cdot) = {}_{\mathcal{B}'}\langle ev_x, \cdot\rangle_{\mathcal{B}}.$$

Then, for $\mu \in \mathcal{B}$, $f_\mu(x) = {}_{\mathcal{F}'}\langle \phi(x), \mu\rangle_{\mathcal{F}} = {}_{\mathcal{B}'}\langle ev_x, \mu\rangle_{\mathcal{B}} = \mu(x)$ for every $x \in \mathcal{X}$. Thus, $f_\mu = \mu$ for all $\mu \in \mathcal{B}$. Then, both item (i) and (ii) are clear.

Now, we turn to the "if" direction. Suppose we have a Banach space $\mathcal{F}$ and a map $\phi : \mathcal{X} \to \mathcal{F}'$, and define a vector space $\mathcal{B}$ of functions on $\mathcal{X}$ as in (i). Then, the norm in (ii) makes $\mathcal{B}$ a normed space. To see that $\mathcal{B}$ is complete, consider the linear map $\phi_* : \mathcal{F} \to \mathcal{B}$ given by

$$\phi_*(\mu) = f_\mu.$$

Let $\phi_*(\mu) = 0$. Then, $f_\mu(x) = 0$ for all $x \in \mathcal{X}$, that is, ${}_{\mathcal{F}'}\langle \phi(x), \mu\rangle_{\mathcal{F}} = 0$ for all $x \in \mathcal{X}$. Denote the kernel of ϕ_* as $\mathcal{N}$. Then,

$$\mathcal{N} = \bigcap_{x \in \mathcal{X}} Ker\phi(x).$$

Thus, $\mathcal{N}$ is a closed subspace of $\mathcal{F}$ as it is the intersection of closed subspaces in $\mathcal{F}$. Then, $\mathcal{F} \setminus \mathcal{N}$ is a Banach space with respect to the norm

$$\|\pi(\mu)\|_{\mathcal{F}\setminus\mathcal{N}} = \inf\{\|v\|_{\mathcal{F}} : v \in \mathcal{F}, \pi(v) = \pi(\mu)\},$$

where $\pi : \mathcal{F} \to \mathcal{F} \setminus \mathcal{N}$ is the canonical projection ([26], Chapter 1, Theorem 1.41). By definition, $\mathcal{B} = \{f_\mu : \mu \in \mathcal{F}\}$, so $\mathcal{B}$ is isomorphic to $\mathcal{F} \setminus \mathcal{N}$, and

$$\|f_\mu\|_{\mathcal{B}} = \inf\{\|v\|_{\mathcal{F}} : v \in \mathcal{F}, f_v = f_\mu\} = \|\pi(\mu)\|_{\mathcal{F}\setminus\mathcal{N}}.$$

Therefore, $\mathcal{B}$ is a normed space isometrically isomorphic to the Banach space $\mathcal{F} \setminus \mathcal{N}$, and consequently it is complete. Moreover, by definition, for every $f \in \mathcal{B}$ there exists $\mu \in \mathcal{F}$ such that $f = f_\mu$, and $|f(x)| = |f_\mu(x)| \leq \|\mu\|_{\mathcal{F}}\|\phi(x)\|_{\mathcal{F}'}$. Thus, for every $x \in \mathcal{X}$,

$$|f(x)| \leq \inf_{\mu \in \mathcal{F}, f_\mu = f} \|\mu\|_{\mathcal{F}}\|\phi(x)\|_{\mathcal{F}'} = \|f\|_{\mathcal{B}}\|\phi(x)\|_{\mathcal{F}'}.$$

Therefore, condition (16) is satisfied with $C_x = \|\phi(x)\|_{\mathcal{F}'}$. □

Remark 2.8. *The map ϕ is called* ***feature map*** *and $\mathcal{F}'$* ***feature space****. Feature maps are in general not unique. We can construct a RKBS starting from a Banach space $\mathcal{F}$ and a map $\phi : \mathcal{X} \to \mathcal{F}'$.*

2.5 Reproducing kernel Banach spaces of neural networks

This section follows Bartolucci et al. [6, Section 3]. We restrict our attention to one hidden layer networks in this work. Recall that a one hidden layer network is a function defined as in (14). Let $\rho(x, \theta) := \sigma(w \cdot x + b)$, where $\theta = (w, b)$. The key idea is to consider the limit for large K in (14), that is

$$\sum_{k=1}^{K} \rho(x, \theta_k)c_k \mapsto \int_\Theta \rho(x, \theta)d\mu(\theta).$$

The right-hand-side expression is the limit where the hidden layer has an infinite number of neurons. If we choose $\mu = \sum_{k=1}^{K} \delta_{\theta_k} c_k$, then

$$\int_\Theta \rho(x, \theta)d\mu(\theta) = \sum_{k=1}^{K} \rho(x, \theta_k)c_k.$$

We fix a (Hausdorff) locally compact second countable topological space Θ as the parameter space. Then, we denote by $\mathcal{M}(\Theta)$ the Banach space of bounded measures defined on the Borel $\sigma-$algebra of Θ, and endow $\mathcal{M}(\Theta)$ with the total variation norm $\|\cdot\|_{\mathrm{TV}}$. By Bartolucci et al. [6, Section 3.3], $\mathcal{M}(\Theta)$ can be identified with the dual of $C_0(\Theta)$, the Banach space of continuous functions going to zero at infinity endowed with the sup norm $\|\cdot\|_\infty$. Then, the TV norm is written as

$$\|\mu\|_{\mathrm{TV}} = \sup\{\langle \mu, \psi\rangle : \psi \in C_0(\Theta), \|\psi\|_\infty \leq 1\}.$$

The keys to our construction are a function $\rho : \mathcal{X} \times \Theta \to \mathbb{R}$ and a measurable function $\beta : \Theta \to \mathbb{R}$ satisfying the following conditions:

(i) for all $x \in \mathcal{X}$

$$\sup_{\theta\in\Theta} |\rho(x,\theta)\beta(\theta)| = D_x < \infty, \tag{17}$$

for some $D_x > 0$;

(ii) for all $x \in \mathcal{X}$, the function $\rho(x,\cdot)$ is measurable.

Next, we define the feature map

$$\phi : \mathcal{X} \to \mathcal{M}(\Theta)', \quad {}_{\mathcal{M}(\Theta)}\langle \mu, \phi(x)\rangle_{\mathcal{M}(\Theta)'} = \int_\Theta \rho(x,\theta)\beta(\theta)d\mu(\theta).$$

This is well-defined because of (17). Then, by proposition 2.18 the feature map ϕ defines a RKBS $\mathcal{B}$ explicitly given by

$$\mathcal{B} = \{f_\mu : \mu \in \mathcal{M}(\Theta)\}, \tag{18}$$

$$f_\mu(x) = \int_\Theta \rho(x,\theta)\beta(\theta)d\mu(\theta), \tag{19}$$

$$\|f\|_{\mathcal{B}} = \inf\{\|\mu\|_{\mathrm{TV}} : f_\mu = f\}. \tag{20}$$

The RKBS can be seen to be parameterized in terms of measures on the parameter space. Here, we regard $\{\rho(\cdot,\theta)\}_\theta$ as a family of basis functions (e.g. as identified by the choice of an activation function in neural networks), and β as a smoothing function needed to ensure that the integral in (19) converges for all μ. Equation (19) provides a pointwise integral representation of the elements of $\mathcal{B}$.

Note that all one hidden layer neural networks belong to $\mathcal{B}$. To see this, we choose the measure μ having finite support, i.e.

$$\mu = \sum_{k=1}^K a_k \delta_{\theta_k}, \quad a_k \in \mathbb{R}, \quad \theta_k \in \Theta,$$

where δ_θ is the Dirac measure at point θ, it follows that the elements of the form

$$f_\mu = \sum_{k=1}^K \alpha_k \rho(\cdot,\theta_k), \quad \alpha_k = a_k\beta(\theta_k) \in \mathbb{R}, \quad \theta_k \in \Theta, \tag{21}$$

belong to $\mathcal{B}$. Thus, the constructed RKBS $\mathcal{B}$ yields a direct connection with one hidden layer neural networks with possibly infinite width.

Next, we link the constructed RKBS given by (18) with ERM. We consider the problem

$$\inf_{f\in\mathcal{B}} \big(\frac{1}{N}\sum_{i=1}^N L(y_i, f(x_i)) + \|f\|_{\mathcal{B}}\big). \tag{22}$$

The following proposition shows that the minimization over the RKBS $\mathcal{B}$ is equivalent to minimizing over the space of measures $\mathcal{M}(\Theta)$.

Proposition 2.19 (Minimization over the RKBS is minimization over the space of measures)**.** *Take $\rho : \mathcal{X}\times\Theta \to \mathbb{R}$, $\beta : \Theta \to \mathbb{R}$ satisfying (17), and set $\mathcal{B}$ as the corresponding RKBS defined in (18). Then,*

$$(22) = \inf_{\mu\in\mathcal{M}(\Theta)} \big(\frac{1}{N}\sum_{i=1}^N L(y_i, f_\mu(x_i)) + \|\mu\|_{\mathrm{TV}}\big).$$

Furthermore, if $\mu^\star$ is any minimizer of

$$\inf_{\mu\in\mathcal{M}(\Theta)} \big(\frac{1}{N}\sum_{i=1}^N L(y_i, f_\mu(x_i)) + \|\mu\|_{\mathrm{TV}}\big), \tag{23}$$

then $f^\star = f_{\mu^\star}$ is a minimizer of problem (22).

Proof. See Bartolucci et al. [6, Proposition 3.7]. □

Thus, problem (22) is reformulated as a minimization over the space of measures $\mathcal{M}(\Theta)$, and we consider the regularizer to be the TV-norm of the measure μ.

Now we strengthen assumption (17) by

$$\rho(x,\cdot)\beta \in C_0(\Theta), \quad \forall x \in \mathcal{X}, \tag{24}$$

which clearly implies (17). Then, we provide a representer theorem (Bartolucci et al. [6, Theorem 3.9]) for the RKBS defined by (18), showing that ERM minimizers always exist, and are of the form (21).

Theorem 2.20 (Representer theorem)**.** *Assume that (24) holds true, and for every $y \in \mathbb{R}$, the function $L(y,\cdot)$ is convex and coercive in the second entry. Then, problem (22) admits solutions $f^\star$ such that, for all $x \in \mathcal{X}$,*

$$f^\star(x) = \sum_{k=1}^{K} \alpha_k \rho(x,\theta_k), \quad \alpha_k \in \mathbb{R}\setminus\{0\}, \quad \theta_k \in \Theta,$$
$$\|f^\star\|_{\mathcal{B}} \leq \sum_{k=1}^{K} |\alpha_k \beta(\theta_k)^{-1}|,$$

with $K \leq N$ and $\beta(\theta_k) \neq 0$ for all $k = 1, \cdots, K$.

Proof. See Bartolucci et al. [6, Theorem 3.9]. □

3 Error bounds for learning with vector-valued random features

Section 3 summarizes the paper of Lanthaler and Nelsen [14], in which an upper bound of the $\mathcal{G}$-population squared error of the random feature models is derived. Section 3.1 presents the key definitions, the assumptions, and the main theorem of Lanthaler and Nelsen [14]. Section 3.2 study the proof of the main theorem in the original paper. The thesis follows the same proof framework as shown in section 3.2. However, the hypothesis space of the thesis regarding neural networks is a reproducing kernel Banach space; whereas the hypothesis space of Lanthaler and Nelsen [14] regarding random features is a reproducing kernel Hilbert space. Therefore, the use of Cauchy–Schwarz inequality on the underlying model as shown in section 3.2 is forbidden in the proof of the main result of the thesis. Nevertheless, except for the proof framework, many results used in section 3.2 are useful in proving the main result of the thesis. For these reasons, we include this section for reference.

3.1 Setting and main results

This section presents the setting and main results of Lanthaler and Nelsen [14]. All definitions, assumptions, settings, and results are taken from the paper.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a sufficiently rich probability space on which all random variables in this section are defined. Let $\mathcal{X}$ be the input space, $\mathcal{Y}$ the output space, and Θ the parameter space. We denote elements of $\mathcal{X}$ as u and random features parameters in Θ as θ . The set of probability measures supported on a set $\mathcal{Q}$ is denoted by $\mathscr{P}(\mathcal{Q})$. We write expectation with respect to $u \sim v \in \mathscr{P}(\mathcal{X})$ as $\mathbb{E}_u[\cdot]$ and similarly for $\theta \sim \mu \in \mathscr{P}(\Theta)$. Independent and identically distributed (iid) samples $u_1, \cdots, u_N$ from v will be denoted by $\{u_n\} \sim v^{\otimes N}$ and similarly for $\{\theta_m\} \sim \mu^{\otimes M}$.

Definition 3.1 (Random feature model)**.** *Let (φ, μ) be a pair where $\varphi : \mathcal{X} \times \Theta \to \mathcal{Y}$ and $\mu \in \mathscr{P}(\Theta)$. Fixing $\theta \sim \mu$ defines a map $\varphi(\cdot,\theta) : \mathcal{X} \to \mathcal{Y}$. The map $\Phi(\cdot;\alpha) = \Phi(\cdot;\alpha,\{\theta_m\}) : \mathcal{X} \to \mathcal{Y}$ given by*

$$u \mapsto \Phi(u;\alpha) := \frac{1}{M}\sum_{m=1}^{M} \alpha_m \varphi(u;\theta_m) \tag{25}$$

is a random feature model (RFM) with coefficients $\alpha \in \mathbb{R}^M$ and fixed realizations $\{\theta_m\} \sim \mu^{\otimes M}$.

Associated to the pair (φ, μ) is a RKHS $\mathcal{H}$ of maps from $\mathcal{X}$ to $\mathcal{Y}$ ([20], section 2.3). Under mild assumptions assumed in the main results, it holds that

$$\mathcal{H} = \{\mathcal{G} \in L^2_v(\mathcal{X};\mathcal{Y}) : \mathcal{G} = \mathbb{E}[\alpha(\theta)\varphi(\cdot;\theta)], \alpha \in L^2_\mu(\Theta;\mathbb{R})\} \tag{26}$$

with RKHS norm $\|\mathcal{G}\|_{\mathcal{H}} = \min_\alpha \|\alpha\|_{L^2_\mu}$, where α ranges over all decompositions of $\mathcal{G}$ of the form in (26). A minimizer $\alpha_{\mathcal{H}}$ of this problem always exists ([3], section 2.2). We use this fact to identify any $\mathcal{G} \in \mathcal{H}$ with its minimizer $\alpha_{\mathcal{H}} \in L^2_\mu(\Theta;\mathbb{R})$.

Let $\mathcal{P} \in \mathscr{P}(\mathcal{X} \times \mathcal{Y})$ denote the *joint data distribution*. The goal of **random feature ridge regression** (RF-RR) is to estimate an underlying operator $\mathcal{G} : \mathcal{X} \to \mathcal{Y}$ from a finite sample of i.i.d. input-output pairs $\{u_n, y_n\}_{n=1}^N \sim \mathcal{P}^{\otimes N}$, where each y_n is typically a noisy observation of $\mathcal{G}(u_n)$. Writing $Y = \{y_n\}$ for the collection of observed output data and fixing a regularization parameter $\lambda > 0$, the *regularized Y-empirical risk* of $\alpha \in \mathbb{R}^M$ is given by

$$\mathscr{R}_N^\lambda(\alpha, Y) := \frac{1}{N}\sum_{n=1}^N \|y_n - \Phi(u_n;\alpha)\|_{\mathcal{Y}}^2 + \lambda\|\alpha\|_M^2, \quad \text{where } \|\alpha\|_M^2 := \frac{1}{M}\sum_{m=1}^M |\alpha_m|^2. \tag{27}$$

The *regularized $\mathcal{G}$-empirical risk* is defined by

$$\mathscr{R}_N^\lambda(\alpha, \mathcal{G}) := \frac{1}{N}\sum_{n=1}^N \|\mathcal{G}(u_n) - \Phi(u_n;\alpha)\|_{\mathcal{Y}}^2 + \lambda\|\alpha\|_M^2.$$

In the absence of regularization, i.e., $\lambda = 0$, these expressions define the *Y-empirical risk* and *$\mathcal{G}$-empirical risk*, denoted by $\mathscr{R}_N(\alpha; Y)$ and $\mathscr{R}_N(\alpha; \mathcal{G})$, respectively. RF-RR is formulated as the minimization problem $\min_{\alpha\in\mathbb{R}^M} \mathscr{R}_N^\lambda(\alpha; Y)$. The minimizer, denoted by $\widehat{\alpha}$, is called the *trained coefficients* and the corresponding random feature model $\Phi(\cdot;\widehat{\alpha})$ is referred to as the *trained RFM*.

Throughout this section, we assume that the input space $\mathcal{X}$ is a Polish space and the output space $\mathcal{Y}$ is a real seperable Hilbert space. We view $\mathcal{X}$ and $\mathcal{Y}$ as measurable spaces equipped with their respective Borel σ-algebras. Key assumptions of the setting are:

Assumption 3.1 (Random feature regularity)**.** *Let $v \in \mathscr{P}(\mathcal{X})$ be the input distribution and (Θ, Σ, μ) be a probability space. The random feature map $\varphi : \mathcal{X} \times \Theta \to \mathcal{Y}$ and the probability measure $\mu \in \mathscr{P}(\Theta)$ are such that (i) φ is measurable; (ii) φ is uniformly bounded; in fact, $\|\varphi\|_{L^\infty} := \operatorname{ess\,sup}_{(u,\theta)\sim v\otimes\mu}\|\varphi(u;\theta\|_{\mathcal{Y}} \le 1$; (iii) the RKHS $\mathcal{H}$ corresponding to (φ, μ) is separable.*

Assumption 3.2 (Misspecification)**.** *There exist $\rho \in L_v^\infty(\mathcal{X}; \mathcal{Y})$ and $\mathcal{G}_\mathcal{H} \in \mathcal{H}$ such that the operator $\mathcal{G} : \mathcal{X} \to \mathcal{Y}$ satisfies the decomposition $\mathcal{G} = \rho + \mathcal{G}_\mathcal{H}$.*

Assumption 3.3 (Joint data distribution)**.** *The joint distribution $\mathcal{P} \in \mathscr{P}(\mathcal{X} \times \mathcal{Y})$ of random variable $(u, y) \sim \mathcal{P}$ is given by $u \sim v$ with $v \in \mathscr{P}(\mathcal{X})$ and $y = \mathcal{G}(u) + \eta$. Here, $\mathcal{G}$ satisfies assumption 3.2. The addictive noise η is a random variable in $\mathcal{Y}$ that is conditionally centered, $\mathbb{E}[\eta|u] = 0$, and is subexponential: $\|\eta\|_{\psi_1(\mathcal{Y})} < \infty$.*

By (26), assumption 3.1 implies that $\mathcal{H} \subset L_v^\infty(\mathcal{X}; \mathcal{Y})$. So, any $\mathcal{G} = \rho + \mathcal{G}_\mathcal{H}$ as in assumption 3.2 is automatically bounded in the sense that $\mathcal{G} \in L_v^\infty(\mathcal{X}; \mathcal{Y})$. Assumption 3.3 implies that $\mathcal{G}(u) = \mathbb{E}[y|u]$.

For any $\mathcal{G}$, define the *$\mathcal{G}-$population risk functional* or *$\mathcal{G}$-population squared error* by

$$\mathscr{R}(\alpha; \mathcal{G}) := \mathbb{E}_{u\sim v}\|\mathcal{G}(u) - \Phi(u;\alpha, \{\theta_m\})\|_{\mathcal{Y}}^2 \quad \text{for } \alpha \in \mathbb{R}^M.$$

The main result (Lanthaler and Nelsen [14, Theorem 3.4]) establishes an upper bound for this quantity that holds with high probability, provided that the number of random features and number of data pairs are large enough.

Theorem 3.1 ($\mathcal{G}$-population squared error bound)**.** *Suppose that $\mathcal{G} = \rho + \mathcal{G}_\mathcal{H}$ satisfies assumption 3.2. Fix a failure probability $\delta \in (0, 1)$, regularization strength $\lambda \in (0, 1)$, and sample size N. Let $\{\theta_m\} \sim \mu^{\otimes M}$ be the M random feature parameters and $\{(u_n, y_n)\} \sim \mathcal{P}^{\otimes N}$ be the data according to assumption 3.3. For Φ the RFM (25) satisfying assumption 3.1, let $\widehat{\alpha} \in \mathbb{R}^M$ be the minimizer of the regularized Y-empirical risk $\mathscr{R}_N^\lambda(\cdot; Y)$ given by (27). If $M \ge \lambda^{-1}\log(32/\delta)$ and $N \ge \lambda^{-2}\log(16/\delta)$, then*

$$\mathbb{E}_{u\sim v}\|\mathcal{G}(u) - \Phi(u;\widehat{\alpha}, \{\theta_m\})\|_{\mathcal{Y}}^2 \le 79e^{3/2}\big(\|\mathcal{G}\|_{L_v^\infty}^2 + 2\beta(\rho, \lambda, \mathcal{G}_\mathcal{H}, \eta)\big)\lambda$$

with probability at least $1 - \delta$, where

$$\beta(\rho, \lambda, \mathcal{G}_\mathcal{H}, \eta) := 328\|\mathcal{G}_\mathcal{H}\|_\mathcal{H}^2 + 2023e^3\|\eta\|_{\psi_1(\mathcal{Y})}^2 + 8\lambda^{-1}\mathbb{E}_{u\sim v}\|\rho(u)\|_{\mathcal{Y}}^2 + 18\lambda\|\rho\|_{L_v^\infty}^2 \tag{28}$$

is a function of $\rho, \lambda, \mathcal{G}_\mathcal{H}$, and the law of the noise variable η.

3.2 Proof of Theorem 3.1

For the sake of completeness, this section displays the proof as in Lanthaler and Nelsen [14] in linear order. We begin with the error decomposition

$$\mathscr{R}(\widehat{\alpha}; \mathcal{G}) = \mathscr{R}_N(\widehat{\alpha}; \mathcal{G}) + [\mathscr{R}(\widehat{\alpha}; \mathcal{G}) - \mathscr{R}_N(\widehat{\alpha}; \mathcal{G})], \tag{29}$$

where the first term is called the approximation error, and the second term is called the generalization gap.

Step 1: We bound the approximation error $\mathscr{R}_N(\widehat{\alpha};\mathcal{G})$ with $\mathscr{R}_N(\widehat{\alpha};\mathcal{G}) \leq \mathscr{R}_N^{\lambda}(\widehat{\alpha};\mathcal{G})$.

Step 1.1: We further bound $\mathscr{R}_N^{\lambda}(\widehat{\alpha};\mathcal{G})$ as follows. Fix any $\alpha \in \mathbb{R}^M$. It holds that

$$\mathscr{R}_N^{\lambda}(\alpha;Y) = \mathscr{R}_N^{\lambda}(\alpha;\mathcal{G}) + \frac{2}{N}\sum_{n=1}^{N}\langle \eta_n, \mathcal{G}(u_n) - \Phi(u_n;\alpha)\rangle_{\mathcal{Y}} + \frac{1}{N}\sum_{n=1}^{N}\|\eta_n\|_{\mathcal{Y}}^2$$

because $\mathcal{Y}$ is a Hilbert space and $y_n = \mathcal{G}(u_n) + \eta_n$. Using this, performing $\mathscr{R}_N^{\lambda}(\widehat{\alpha};Y) - \mathscr{R}_N^{\lambda}(\alpha;Y)$ will cancel the noise term, and we have that

$$\begin{aligned}\mathscr{R}_N^{\lambda}(\widehat{\alpha};\mathcal{G}) &= \left[\mathscr{R}_N^{\lambda}(\widehat{\alpha};Y) - \mathscr{R}_N^{\lambda}(\alpha;Y)\right] + \mathscr{R}_N^{\lambda}(\alpha;\mathcal{G}) + \frac{2}{N}\sum_{n=1}^{N}\langle \eta_n, \Phi(u_n;\widehat{\alpha}) - \Phi(u_n;\alpha)\rangle_{\mathcal{Y}} \\ &\leq \mathscr{R}_N^{\lambda}(\alpha;\mathcal{G}) + \frac{2}{N}\sum_{n=1}^{N}\langle -\eta_n, \Phi(u_n;\alpha)\rangle_{\mathcal{Y}} + \frac{2}{N}\sum_{n=1}^{N}\langle \eta_n, \Phi(u_n;\widehat{\alpha})\rangle_{\mathcal{Y}}. \end{aligned} \tag{30}$$

We used the fact the $\widehat{\alpha}$ is a minimizer of the RF-RR so the first term is always less than or equal to zero. If $\|\widehat{\alpha}\|_M > 0$, then by linearity we have that

$$\begin{aligned}\frac{2}{N}\sum_{n=1}^{N}\langle \eta_n, \Phi(u_n;\widehat{\alpha})\rangle_{\mathcal{Y}} &= \|\widehat{\alpha}\|_M\left(\frac{2}{N}\sum_{n=1}^{N}\langle \eta_n, \Phi(u_n;\widehat{\alpha}/\|\widehat{\alpha}\|_M)\rangle_{\mathcal{Y}}\right) \\ &\leq \|\widehat{\alpha}\|_M\left(2\sup_{\alpha'\in\mathcal{A}_1}\left|\frac{1}{N}\sum_{n=1}^{N}\langle \eta_n, \Phi(u_n;\alpha')\rangle_{\mathcal{Y}}\right|\right),\end{aligned}$$

where $\mathcal{A}_1 = \left\{\alpha' \in \mathbb{R}^M \mid \|\alpha'\|_M^2 \leq 1\right\}$. Note that the inequality in the above display holds trivially if $\|\widehat{\alpha}\|_M = 0$. Next, we define

$$t := \frac{1}{M}\sum_{m=1}^{M}|\widehat{\alpha}_m|^2 = \|\widehat{\alpha}\|_M^2 \tag{31}$$

$$A_{N,M}^{\lambda} := \mathscr{R}_N^{\lambda}(\alpha;\mathcal{G}) + \frac{2}{N}\sum_{n=1}^{N}\langle -\eta_n, \Phi(u_n;\alpha)\rangle_{\mathcal{Y}}, \quad \text{and} \tag{32}$$

$$c_{N,M} := \left(2\sup_{\alpha'\in\mathcal{A}_1}\left|\frac{1}{N}\sum_{n=1}^{N}\langle \eta_n, \Phi(u_n;\alpha')\rangle_{\mathcal{Y}}\right|\right)^2. \tag{33}$$

The definition of the regularized $\mathcal{G}$-empirical risk gives that $\lambda t \leq \mathscr{R}_N^{\lambda}(\widehat{\alpha};\mathcal{G})$. Inequality (30) and the arithmetic-mean-geometric-mean inequality together imply that

$$\lambda t \leq \mathscr{R}_N^{\lambda}(\widehat{\alpha};\mathcal{G}) \leq A_{N,M}^{\lambda} + \sqrt{c_{N,M}t} \leq A_{N,M}^{\lambda} + \frac{1}{2}\lambda^{-1}c_{N,M} + \frac{1}{2}\lambda t. \tag{34}$$

Subtracting $\lambda t/2$ from both sides and multiplying through by $2\lambda^{-1}$ yields

$$t \leq 2\lambda^{-1}A_{N,M}^{\lambda} + \lambda^{-2}c_{N,M}. \tag{35}$$

Substituting (35) back into the right-most side of (34) gives

$$\mathscr{R}_N^{\lambda}(\widehat{\alpha};\mathcal{G}) \leq 2A_{N,M}^{\lambda} + \lambda^{-1}c_{N,M} \tag{36}$$

Now, it remains for us to bound the $A_{N,M}^{\lambda}$ (32) and $c_{N,M}$ (33) terms.

Step 1.2: We begin with bounding $c_{N,M}$ (33). We precede with using Bernstein-type inequality to show that $\sqrt{c_{N,M}}$ concentrates around its expectation. Then, we upper bound the expectation of $\sqrt{c_{N,M}}$.

Recall that

$$\sqrt{c_{N,M}} = 2\sup_{\alpha'\in\mathcal{A}_1}\left|\frac{1}{N}\sum_{n=1}^{N}\langle \eta_n, \Phi(u_n;\alpha')\rangle_{\mathcal{Y}}\right|.$$

The main idea is to apply (9) in corollary 2.7 to $\langle \eta_n, \Phi(u_n;\cdot)\rangle_{\mathcal{Y}}$, which takes value in the separable Banach space $C(\mathcal{A}_1;\mathbb{R})$ of continuous functions from the compact set $\mathcal{A}_1 \subset \mathbb{R}^M$ into $\mathbb{R}$, equipped with the supremum norm. We first need to check that $\langle \eta_n, \Phi(u_n;\cdot)\rangle_{\mathcal{Y}}$ satisfies the conditions of corollary 2.7. The i.i.d. random variables $\langle \eta_n, \Phi(u_n;\cdot)\rangle_{\mathcal{Y}} : \alpha \mapsto \langle \eta_n, \Phi(u_n;\alpha)\rangle_{\mathcal{Y}}$ (conditional on $\{\theta_m\}$) are linear and hence continuous. Then, it remains to check that $\langle \eta_n, \Phi(u_n;\cdot)\rangle_{\mathcal{Y}}$ is bounded in the $\psi_1(C(\mathcal{A}_t;\mathbb{R}))$ norm. For any $\alpha \in \mathcal{A}_1$, we compute

$$\begin{aligned}
\left|\langle \eta_1, \Phi(u_1;\alpha)\rangle_{\mathcal{Y}}\right| &= \left|\frac{1}{M}\sum_{m=1}^{M} \alpha_m \langle \eta_1, \varphi(u_1;\theta_m)\rangle_{\mathcal{Y}}\right| \\
&\leq \left(\frac{1}{M}\sum_{m=1}^{M} |\alpha_m|^2\right)^{1/2} \left(\frac{1}{M}\sum_{m=1}^{M} \langle \eta_1, \varphi(u_1;\theta_m)\rangle_{\mathcal{Y}}^2\right)^{1/2} \\
&\leq \left(\|\eta_1\|_{\mathcal{Y}}^2 \frac{1}{M}\sum_{m=1}^{M} \|\varphi(u_1;\theta_m)\|_{\mathcal{Y}}^2\right)^{1/2}.
\end{aligned}$$

We used the Cauchy-Schwarz inequality twice. By the boundedness of φ, the above display gives

$$\|\langle \eta_1, \Phi(u_1;\cdot)\rangle\|_{\psi_1(C(\mathcal{A}_1;\mathbb{R}))} = \left\| \sup_{\alpha\in\mathcal{A}_1} \left|\langle \eta_1, \Phi(u_1;\alpha)\rangle_{\mathcal{Y}}\right| \right\|_{\psi_1} \leq \|\eta_1\|_{\psi_1(\mathcal{Y})} \|\varphi\|_{L^\infty}.$$

Thus, (9) in corollary 2.7 gives that if $N \geq \log(1/\delta)$, then conditioned on the realizations $\{\theta_m\}$ in the family Φ, it holds that

$$\sqrt{c_{N,M}} \leq \mathbb{E}_{\{(u_n,\eta_n)\}} \sup_{\alpha\in\mathcal{A}_1} \left|\frac{1}{N}\sum_{n=1}^{N} \langle \eta_n, \Phi(u_n;\alpha)\rangle_{\mathcal{Y}}\right| + 16e^{3/2}\|\eta\|_{\psi_1(\mathcal{Y})}\|\varphi\|_{L^\infty}\sqrt{\frac{\log(1/\delta)}{N}}$$

with probability at least $1-\delta$. Next, we bound the expectation $\Xi := \mathbb{E}_{\{(u_n,\eta_n)\}} \sup_{\alpha\in\mathcal{A}_1} \left|\frac{1}{N}\sum_{n=1}^{N} \langle \eta_n, \Phi(u_n;\alpha)\rangle_{\mathcal{Y}}\right|$. For any $\alpha \in \mathcal{A}_1$,

$$\begin{aligned}
\left|\frac{1}{N}\sum_{n=1}^{N} \langle \eta_n, \Phi(u_n;\alpha)\rangle_{\mathcal{Y}}\right| &= \left|\frac{1}{M}\sum_{m=1}^{M} \alpha_m \frac{1}{N}\sum_{n=1}^{N} \langle \eta_n, \varphi(u_n;\theta_m)\rangle_{\mathcal{Y}}\right| \\
&\leq \left(\frac{1}{M}\sum_{m=1}^{M} |\alpha_m|^2\right)^{1/2} \left(\frac{1}{M}\sum_{m=1}^{M} \left[\frac{1}{N}\sum_{n=1}^{N} \langle \eta_n, \varphi(u_n;\theta_m)\rangle_{\mathcal{Y}}\right]^2\right)^{1/2}.
\end{aligned}$$

We used the Cauchy-Schwarz inequality in $\mathbb{R}^M$. We next note that

$$\begin{aligned}
\mathbb{E}_{\{(u_n,\eta_n)\}}\left[\langle \eta_n, \varphi(u_n;\theta_m)\rangle_{\mathcal{Y}}\right] &= \mathbb{E}_{u_n\sim\nu}\left[\mathbb{E}\left[\langle \eta_n, \varphi(u_n;\theta_m)\rangle_{\mathcal{Y}} \mid u_n\right]\right] \\
&= \mathbb{E}_{u_n\sim\nu}\left[\langle \mathbb{E}\left[\eta_n \mid u_n\right], \varphi(u_n;\theta_m)\rangle_{\mathcal{Y}}\right] \\
&= 0.
\end{aligned}$$

By independence of (u_n, η_n) and $(u_{n'}, \eta_{n'})$ for any two indices $n \neq n'$, together with the above observation, we obtain that

$$\begin{aligned}
&\mathbb{E}_{\{(u_n,\eta_n)\}}\left[\langle \eta_n, \varphi(u_n;\theta_m)\rangle_{\mathcal{Y}} \langle \eta_{n'}, \varphi(u_{n'};\theta_m)\rangle_{\mathcal{Y}}\right] \\
&\quad = \mathbb{E}_{(u_n,\eta_n)}\left[\langle \eta_n, \varphi(u_n;\theta_m)\rangle_{\mathcal{Y}}\right] \mathbb{E}_{(u_{n'},\eta_{n'})}\left[\langle \eta_{n'}, \varphi(u_{n'};\theta_m)\rangle_{\mathcal{Y}}\right] \\
&\quad = 0.
\end{aligned}$$

So, expanding the square $\left[\sum_{n=1}^{N} \langle \eta_n, \varphi(u_n;\theta_m)\rangle_{\mathcal{Y}}\right]^2$ and taking the expectation with respect to $\{(u_n, \eta_n)\}$, it remains the terms with same indices. This implies that

$$\begin{aligned}
\Xi &\leq \frac{1}{\sqrt{N}} \sqrt{\frac{1}{M}\sum_{m=1}^{M} \mathbb{E}_{(u_1,\eta_1)} \langle \eta_1, \varphi(u_1;\theta_m)\rangle_{\mathcal{Y}}^2} \\
&\leq \frac{1}{\sqrt{N}} \|\eta_1\|_{L^2_{\mathbb{P}}(\Omega;\mathcal{Y})} \|\varphi\|_{L^\infty},
\end{aligned}$$

where we used Jensen's inequality to bring the expectation inside the square root in the first line, and the Cauchy-Schwarz inequality in $\mathcal{Y}$ in the second line.

Combining all results above,

$$\begin{aligned}\sqrt{c_{N,M}} &\leq \frac{4\|\eta\|_{\psi_1(\mathcal{Y})}\|\varphi\|_{L^\infty}}{\sqrt{N}} + 16e^{3/2}\|\eta\|_{\psi_1(\mathcal{Y})}\|\varphi\|_{L^\infty}\sqrt{\frac{\log(1/\delta)}{N}} \\ &\leq 16e^{3/2}\|\eta\|_{\psi_1(\mathcal{Y})}\|\varphi\|_{L^\infty}\sqrt{\frac{6\log(2/\delta)}{N}}\end{aligned} \tag{37}$$

with probability at least $1-\delta$ if $N \geq \log(2/\delta) \geq \log(1/\delta)$. We used the inequalities $4 \leq 16e^{3/2}$, $\sqrt{a}+\sqrt{b} \leq \sqrt{2(a+b)}$, and $1 \leq 2\log(2/\delta)$ to get to the last line.

Step 1.3: Continuing, we bound $A^{\lambda}_{N,M}$ (32). It is made up of two terms. We bound each term separately.

Step 1.3.1: We begin with the first term in (32), that is $\mathscr{R}^{\lambda}_N(\alpha;\mathcal{G})$, where $\alpha \in \mathbb{R}^M$ is arbitrary. By assumption 3.2, $\mathcal{G} = \rho + \mathcal{G}_{\mathcal{H}}$ so that

$$\mathscr{R}^{\lambda}_N(\alpha;\mathcal{G}) \leq \lambda\|\alpha\|^2_M + 2\mathscr{R}_N\left(\alpha;\mathcal{G}_{\mathcal{H}}\right) + \frac{2}{N}\sum_{n=1}^N \|\rho\left(u_n\right)\|^2_{\mathcal{Y}} \leq 2\mathscr{R}^{\lambda}_N\left(\alpha;\mathcal{G}_{\mathcal{H}}\right) + \frac{2}{N}\sum_{n=1}^N \|\rho\left(u_n\right)\|^2_{\mathcal{Y}}. \tag{38}$$

Step 1.3.1.1: First, we bound the second term on the right hand side in (38). Let $Z_1 = \|\rho\left(u_1\right)\|^2_{\mathcal{Y}}$. Note that Z_1 is uncentered. The idea is to use vector-valued Bernstein inequality in Hilbert space (Theorem 2.3). We first need to show that Z_1 satisfies the Bernstein's moment condition. To achieve this, we can apply Lemma 2.4. We have that $Z_1 \leq \|\rho\|^2_{L^\infty_\nu}$ and

$$\mathbb{E}\left|Z_1 - \mathbb{E}Z_1\right|^2 \leq \mathbb{E}Z_1^2 = \mathbb{E}_{u\sim\nu}\|\rho(u)\|^4_{\mathcal{Y}} \leq \|\rho\|^2_{L^\infty_\nu}\mathbb{E}_{u\sim\nu}\|\rho(u)\|^2_{\mathcal{Y}}$$

almost surely. Thus with probability at least $1-\delta$, lemma 2.4 and theorem 2.3 provide the bound

$$\frac{1}{N}\sum_{n=1}^N \|\rho\left(u_n\right)\|^2_{\mathcal{Y}} \leq \mathbb{E}_u\|\rho(u)\|^2_{\mathcal{Y}} + \frac{4\|\rho\|^2_{L^\infty_\nu}\log(2/\delta)}{N} + \sqrt{\frac{2\mathbb{E}_u\|\rho(u)\|^2_{\mathcal{Y}}\|\rho\|^2_{L^\infty_\nu}\log(2/\delta)}{N}}$$

The arithmetic-mean-geometric-mean inequality $\sqrt{ab} \leq (a+b)/2$ gives that

$$\sqrt{\left(2\mathbb{E}_u\|\rho(u)\|^2_{\mathcal{Y}}\right)\left(\|\rho\|^2_{L^\infty_\nu}\log(2/\delta)/N\right)} \leq \mathbb{E}_u\|\rho(u)\|^2_{\mathcal{Y}} + \frac{\frac{1}{2}\|\rho\|^2_{L^\infty_\nu}\log(2/\delta)}{N}.$$

Thus, it holds with probability at least $1-\delta$ that

$$\frac{2}{N}\sum_{n=1}^N \|\rho\left(u_n\right)\|^2_{\mathcal{Y}} \leq 4\mathbb{E}_u\|\rho(u)\|^2_{\mathcal{Y}} + \frac{9\|\rho\|^2_{L^\infty_\nu}\log(2/\delta)}{N}. \tag{39}$$

Step 1.3.1.2: Next, we bound the first term on the right hand side in (38).

Let $T := T(\lambda) = \sqrt{\lambda^{-1}\mathbb{E}_{\theta\sim\mu}\left|\alpha_{\mathcal{H}}(\theta)\right|^2}$. Define $\alpha^\star \in \mathbb{R}^M$ componentwise by

$$\alpha^\star_m := \alpha_{\mathcal{H}}\left(\theta_m\right)\mathbb{1}_{\{|\alpha_{\mathcal{H}}(\theta_m)|\leq T\}}, \tag{40}$$

where $\mathcal{G}_{\mathcal{H}} = \mathbb{E}_{\theta\sim\mu}\left[\alpha_{\mathcal{H}}(\theta)\varphi(\cdot;\theta)\right]$ with $\|\mathcal{G}_{\mathcal{H}}\|^2_{\mathcal{H}} = \mathbb{E}_{\theta\sim\mu}\left|\alpha_{\mathcal{H}}(\theta)\right|^2$. Write $\alpha := \alpha_{\mathcal{H}}$. Next, define $\alpha_{\leq T} \in L^2_\mu(\Theta;\mathbb{R})$ by

$$\theta \mapsto \alpha_{\leq T}(\theta) := \alpha(\theta)\mathbb{1}_{\{|\alpha(\theta)|\leq T\}} = \begin{cases}\alpha(\theta), & \text{if } |\alpha(\theta)| \leq T, \\ 0, & \text{otherwise.}\end{cases}$$

We define $\alpha_{>T} := \alpha\mathbb{1}_{\{|\alpha|>T\}}$ similarly, so that $\alpha \equiv \alpha_{\leq T} + \alpha_{>T}$. Then for $\theta_1,\ldots,\theta_M$, we have $\alpha^\star \in \mathbb{R}^M$ given by $\alpha^\star_m = \alpha_{\leq T}\left(\theta_m\right)$ for each $m \in \{1,\ldots,M\}$. We make the error decomposition

$$\begin{aligned}\mathscr{R}^{\lambda}_N\left(\alpha^\star;\mathcal{G}\right) =& \frac{1}{N}\sum_{n=1}^N \left\|\mathcal{G}\left(u_n\right) - \frac{1}{M}\sum_{m=1}^M \alpha_{\leq T}\left(\theta_m\right)\varphi\left(u_n;\theta_m\right)\right\|^2_{\mathcal{Y}} + \frac{\lambda}{M}\sum_{m=1}^M \left|\alpha_{\leq T}\left(\theta_m\right)\right|^2 \\ \leq& \frac{2}{N}\sum_{n=1}^N \left\|\mathcal{G}\left(u_n\right) - \mathbb{E}_\theta\left[\alpha_{\leq T}(\theta)\varphi\left(u_n;\theta\right)\right]\right\|^2_{\mathcal{Y}} && \text{(I)} \\ &+ \frac{2}{N}\sum_{n=1}^N \left\|\frac{1}{M}\sum_{m=1}^M \alpha_{\leq T}\left(\theta_m\right)\varphi\left(u_n;\theta_m\right) - \mathbb{E}_\theta\left[\alpha_{\leq T}(\theta)\varphi\left(u_n;\theta\right)\right]\right\|^2_{\mathcal{Y}} && \text{(II)} \\ &+ \frac{\lambda}{M}\sum_{m=1}^M \left|\alpha_{\leq T}\left(\theta_m\right)\right|^2. && \text{(III)}\end{aligned}$$

Each of the three terms (I)-(III) is estimated as follows. For (I), it holds that

$$\|\mathcal{G}(u) - \mathbb{E}_\theta\left[\alpha_{\leq T}(\theta)\varphi(u;\theta)\right]\|_{\mathcal{Y}}^2 = \|\mathbb{E}_\theta\left[\alpha_{>T}(\theta)\varphi(u;\theta)\right]\|_{\mathcal{Y}}^2 \leq \|\varphi\|_{L^\infty}^2 \left(\mathbb{E}_\theta\left|\alpha_{>T}(\theta)\right|\right)^2 .$$

Since $\mathbb{E}_\theta\left|\alpha_{>T}(\theta)\right| \leq \mathbb{E}_\theta|\alpha(\theta)|^2/T$, we have that

$$\text{(I)} \leq \frac{2\|\varphi\|_{L^\infty}^2 \left(\mathbb{E}_\theta|\alpha(\theta)|^2\right)^2}{T^2} = 2\lambda\|\varphi\|_{L^\infty}^2 \mathbb{E}_\theta|\alpha(\theta)|^2 .$$

For (II), denote $\nu_N = \frac{1}{N}\sum_{n=1}^N \delta_{u_n}$ as the empirical measure, let $Z = Z(\theta)$ be the $L^2_{\nu_N}(\mathcal{X};\mathcal{Y})$-valued random variable defined for $\theta \sim \mu$ by

$$Z = \alpha_{\leq T}(\theta)\varphi(\cdot;\theta).$$

The idea here is agin to use the vector-valued Bernstein inequality in Hilbert space (Theorem 2.3). We next show that Z satisfies the Bernstein's moment condition by Lemma 2.4. By boundedness of $|\alpha_{\leq T}| \leq T$, we have $\|Z_m\|_{L^2_{\nu_N}} \leq T\|\varphi\|_{L^\infty}$ for each m. The variance is bounded above as

$$\sigma^2 := \mathbb{E}\|Z - \mathbb{E}Z\|_{L^2_{\nu_N}}^2 \leq \mathbb{E}\|Z\|_{L^2_{\nu_N}}^2 \leq \|\varphi\|_{L^\infty}^2 \mathbb{E}_\theta|\alpha(\theta)|^2.$$

By lemma 2.4 and theorem 2.3, it holds with probability at least $1-\delta$ that

$$\left\| \frac{1}{M}\sum_{m=1}^M Z_m - \mathbb{E}Z \right\|_{L^2_{\nu_N}} \leq \frac{4T\|\varphi\|_{L^\infty}\log(2/\delta)}{M} + \sqrt{\frac{2\sigma^2\log(2/\delta)}{M}},$$

where $Z_1, \ldots, Z_M$ are M iid copies of Z. Squaring both sides and substitution of the above bound on σ^2 yields that

$$\begin{aligned}\text{(II)} &\leq \frac{64T^2\|\varphi\|_{L^\infty}^2\log(2/\delta)^2}{M^2} + \frac{8\|\varphi\|_{L^\infty}^2\log(2/\delta)\mathbb{E}_\theta|\alpha(\theta)|^2}{M} \\ &= \lambda\mathbb{E}_\theta|\alpha(\theta)|^2\left(\frac{64\|\varphi\|_{L^\infty}^2\log(2/\delta)^2}{\lambda^2M^2} + \frac{8\|\varphi\|_{L^\infty}^2\log(2/\delta)}{\lambda M}\right)\end{aligned}$$

with probability at least $1-\delta$.
(III) is estimated in a similar way. We aim to apply Bernstein's inequality (4) to the random variable $Z(\theta) := |\alpha_{\leq T}(\theta)|^2$ with $\theta \sim \mu$ and $M \in \mathbb{N}$ iid copies $Z_1, \ldots, Z_M$ of Z defined by $Z_m = Z(\theta_m)$ for each m. Note that $|Z| \leq T^2$ by definition and the variance of Z satisfies the upper bound

$$\sigma^2 := \mathbb{E}|Z - \mathbb{E}Z|^2 \leq \mathbb{E}|Z|^2 = \mathbb{E}_\theta\left|\alpha_{\leq T}(\theta)\right|^4 \leq T^2\mathbb{E}_\theta|\alpha(\theta)|^2.$$

It follows from lemma 2.4 and theorem 2.3 that

$$\frac{1}{m}\sum_{m=1}^M Z_m \leq \mathbb{E}Z + \frac{4T^2\log(2/\delta)}{M} + \sqrt{\frac{2\sigma^2\log(2/\delta)}{M}}$$

with probability at least $1-\delta$. This implies that

$$\begin{aligned}\text{(III)} &\leq \lambda\mathbb{E}|\alpha(\theta)|^2 + \frac{4\lambda T^2\log(2/\delta)}{M} + \lambda\sqrt{\frac{2T^2\mathbb{E}|\alpha(\theta)|^2\log(2/\delta)}{M}} \\ &= \lambda\mathbb{E}_\theta|\alpha(\theta)|^2\left(1 + \frac{4\log(2/\delta)}{\lambda M} + \sqrt{\frac{2\log(2/\delta)}{\lambda M}}\right)\end{aligned}$$

with probability at least $1-\delta$.
Combining the three estimates, it follows that if

$$\frac{\log(2/\delta)}{\lambda M} \leq 1,$$

then

$$\mathscr{R}_N^\lambda\left(\alpha^\star;\mathcal{G}_\mathcal{H}\right) = \mathscr{R}_N^\lambda\left(\alpha^\star;\mathcal{G}\right) \leq \left(74\|\varphi\|_{L^\infty}^2 + 7\right)\lambda\mathbb{E}_\theta\left|\alpha_\mathcal{H}(\theta)\right|^2 \leq 81\lambda\left\|\mathcal{G}_\mathcal{H}\right\|_\mathcal{H}^2 .$$

with probability at least $1-2\delta$.

Recall that the goal of Step 1.3.1 is to bound the first term on the right hand side in (38). Since $\mathcal{G}_\mathcal{H} \in \mathcal{H}$, there exists $\alpha_\mathcal{H} \in L^2_\mu(\Theta;\mathbb{R})$ such that

$$\mathcal{G}_\mathcal{H} = \mathbb{E}_{\theta\sim\mu}\left[\alpha_\mathcal{H}(\theta)\varphi(\cdot;\theta)\right] \quad \text{and} \quad \left\|\mathcal{G}_\mathcal{H}\right\|_\mathcal{H}^2 = \mathbb{E}_{\theta\sim\mu}\left|\alpha_\mathcal{H}(\theta)\right|^2 .$$

With $\alpha_{\mathcal{H}}$ as in the above display, choose once and for all $\alpha \equiv \alpha^\star \in \mathbb{R}^M$ as in (40). Then, we have that

$$2\mathscr{R}_N^\lambda\left(\alpha^\star; \mathcal{G}_{\mathcal{H}}\right) \leq 162\lambda \left\|\mathcal{G}_{\mathcal{H}}\right\|_{\mathcal{H}}^2 \tag{41}$$

with probability at least $1-\delta$ if $M \geq \lambda^{-1}\log(4/\delta)$.

Step 1.3.2: It remains to bound the second term in $A_{N,M}^\lambda$ (32), that is $\frac{2}{N}\sum_{n=1}^N \langle -\eta_n, \Phi\left(u_n;\alpha\right)\rangle_{\mathcal{Y}}$. Define $Z_n(\alpha) := \langle -\eta_n, \Phi\left(u_n;\alpha\right)\rangle_{\mathcal{Y}}$ for each n. Conditioned on $\{\theta_m\}$, it holds that Z_n is an iid copy of Z_1. We will show below that Z_n is subexponential. Then, the idea to bound the sum is to apply Corollary 2.7, which gives an upper bound for the sum of subexponentials. By the assumption that $\mathbb{E}\left[\eta_1 \mid u_1\right] = 0$, we have

$$\begin{aligned}
\mathbb{E}Z_1(\alpha) &= \mathbb{E}_{(u_1,\eta_1)}\left[\langle -\eta_1, \Phi\left(u_1;\alpha\right)\rangle_{\mathcal{Y}}\right] \\
&= \mathbb{E}_{u_1\sim\nu}\left[\mathbb{E}\left[\langle -\eta_1, \Phi\left(u_1;\alpha\right)\rangle_{\mathcal{Y}} \mid u_1\right]\right] \\
&= \mathbb{E}_{u_1\sim\nu}\left[\langle -\mathbb{E}\left[\eta_1 \mid u_1\right], \Phi\left(u_1;\alpha\right)\rangle_{\mathcal{Y}}\right] \\
&= 0.
\end{aligned}$$

Next,

$$|Z_1(\alpha)| \leq \|\eta_1\|_{\mathcal{Y}} \left\|\Phi\left(u_1;\alpha\right)\right\|_{\mathcal{Y}} \leq \|\eta_1\|_{\mathcal{Y}} \|\alpha\|_M \|\varphi\|_{L^\infty}$$

by two applications of the Cauchy-Schwarz inequality, one in $\mathcal{Y}$ and the other in $\mathbb{R}^M$. We deduce that conditioned on $\{\theta_m\}$, it holds that $\|Z_1(\alpha)\|_{\psi_1} \leq \|\eta_1\|_{\psi_1(\mathcal{Y})}\|\alpha\|_M\|\varphi\|_{L^\infty}$. Then, Proposition 2.6, Thmorem 2.3 (Bernstein's inequality), and a similar argument to that in the proof of Corollary 2.7 gives that with probability at least $1-\delta$, it holds that

$$\frac{2}{N}\sum_{n=1}^N \langle -\eta_n, \Phi(u_n;\alpha)\rangle_{\mathcal{Y}} \leq 16e^{3/2}\|\eta_1\|_{\psi_1(\mathcal{Y})}\|\alpha\|_M\sqrt{\frac{\log(2/\delta)}{N}}.$$

On the same event as in (41), it holds that

$$\|\alpha^\star\|_M^2 \leq \mathbb{E}_\theta\left|\alpha_{\mathcal{H}}(\theta)\right|^2\left(1 + \frac{2\log(2/\delta)}{\lambda M} + \sqrt{\frac{2\log(2/\delta)}{\lambda M}}\right) \leq 5\mathbb{E}_\theta\left|\alpha_{\mathcal{H}}(\theta)\right|^2 = 5\left\|\mathcal{G}_{\mathcal{H}}\right\|_{\mathcal{H}}^2$$

due to the result we obtained when bounding the (III) term in Step 1.3.1. Using this fact and choosing $\alpha \equiv \alpha^\star$ as in (40), we have that the second and final term in $A_{N,M}^\lambda$ (32) satisfies with probability at least $1-\delta$ the upper bound

$$\frac{2}{N}\sum_{n=1}^N \langle -\eta_n, \Phi\left(u_n;\alpha^\star\right)\rangle_{\mathcal{Y}} \leq 40e^{3/2}\left\|\mathcal{G}_{\mathcal{H}}\right\|_{\mathcal{H}}\|\eta\|_{\psi_1(\mathcal{Y})}\|\varphi\|_{L^\infty}\sqrt{\frac{\log(2/\delta)}{N}}. \tag{42}$$

Step 1.4: Combining the estimates (37), (38), (39), (41), and (42), recalling (36), and invoking the union bound, we deduce that if $N \geq \lambda^{-2}\log(2/\delta)$, then

$$\mathscr{R}_N^\lambda(\widehat{\alpha};\mathcal{G}) \leq C_0\lambda + 8\mathbb{E}_{u\sim\nu}\|\rho(u)\|_{\mathcal{Y}}^2 + 18\|\rho\|_{L_\nu^\infty}^2\lambda^2$$

with probability at least $1-4\delta$, where

$$\begin{aligned}
C_0 &:= 324\left\|\mathcal{G}_{\mathcal{H}}\right\|_{\mathcal{H}}^2 + 80e^{3/2}\|\eta\|_{\psi_1(\mathcal{Y})}\|\varphi\|_{L^\infty}\left\|\mathcal{G}_{\mathcal{H}}\right\|_{\mathcal{H}} + 1536e^3\|\eta\|_{\psi_1(\mathcal{Y})}^2\|\varphi\|_{L^\infty}^2 \\
&\leq (324+4)\left\|\mathcal{G}_{\mathcal{H}}\right\|_{\mathcal{H}}^2 + 1936e^3\|\eta\|_{\psi_1(\mathcal{Y})}^2\|\varphi\|_{L^\infty}^2.
\end{aligned}$$

In the last line, we used Young's inequality with $\varepsilon = 1/8$, that is, $ab \leq \varepsilon a^2/2 + b^2/(2\varepsilon)$ with $a = 80e^{3/2}\|\eta\|_{\psi_1(\mathcal{Y})}\|\varphi\|_{L^\infty}$ and $b = \left\|\mathcal{G}_{\mathcal{H}}\right\|_{\mathcal{H}}$. Thus, we conclude that if $M \geq \lambda^{-1}\log(16/\delta)$ and $N \geq \lambda^{-2}\log(8/\delta)$, then

$$\mathscr{R}_N^\lambda(\widehat{\alpha};\mathcal{G}) \leq \lambda\beta\left(\rho,\lambda,\mathcal{G}_{\mathcal{H}},\eta\right) \tag{43}$$

with probability at least $1-\delta$, where the multiplicative factor $\beta\left(\rho,\lambda,\mathcal{G}_{\mathcal{H}},\eta\right)$ is given by (28).

Step 2: We bound the generalization gap $\mathscr{R}(\widehat{\alpha};\mathcal{G}) - \mathscr{R}_N(\widehat{\alpha};\mathcal{G})$ with suprema techniques. Define the random variable

$$\mathcal{E}_\beta\left(\{u_n\},\{\theta_m\}\right) := \sup_{\alpha\in\mathcal{A}_\beta}\left|\frac{1}{N}\sum_{n=1}^N \left\|\mathcal{G}\left(u_n\right) - \Phi\left(u_n;\alpha\right)\right\|_{\mathcal{Y}}^2 - \mathbb{E}_u\|\mathcal{G}(u) - \Phi(u;\alpha)\|_{\mathcal{Y}}^2\right| \tag{44}$$

where $\mathcal{A}_\beta := \left\{\alpha' \in \mathbb{R}^M \mid \|\alpha'\|_M^2 \leq \beta\right\}$ and the deterministic radius $\beta = \beta(\rho, \lambda, \mathcal{G}_\mathcal{H}, \eta)$ is given in (28). For any $\alpha \in \mathcal{A}_\beta$ and $n \in \{1, \ldots, N\}$, let

$$X_n(\beta, \alpha) := \|\mathcal{G}(u_n) - \Phi(u_n; \alpha)\|_\mathcal{Y}^2 - \mathbb{E}_{u\sim\nu}\|\mathcal{G}(u) - \Phi(u; \alpha)\|_\mathcal{Y}^2.$$

We will show below that $X_n(\beta, \cdot)$ is subexponential. Then, the idea to bound the sum in (44) is again using Corollary 2.7. We compute

$$\begin{aligned}|X_1(\beta, \alpha)| &\leq 2\|\mathcal{G}(u_1)\|_\mathcal{Y}^2 + 2\mathbb{E}_{u\sim\nu}\|\mathcal{G}(u)\|_\mathcal{Y}^2 + 2\|\Phi(u_1; \alpha)\|_\mathcal{Y}^2 + 2\mathbb{E}_{u\sim\nu}\|\Phi(u; \alpha)\|_\mathcal{Y}^2 \\ &\leq 4\|\mathcal{G}\|_{L^\infty_\nu}^2 + 4\|\varphi\|_{L^\infty}^2\beta.\end{aligned}$$

We used the fact that for any $u \in \mathcal{X}$ ν-almost surely, $\|\Phi(u; \alpha)\|_\mathcal{Y}^2 \leq \beta\|\varphi\|_{L^\infty}^2$ on the set $\mathcal{A}_\beta$ (by the Cauchy-Schwarz inequality). This implies that

$$\|X_1(\beta, \cdot)\|_{\psi_1(C(\mathcal{A}_\beta;\mathbb{R}))} = \left\|\sup_{\alpha\in\mathcal{A}_\beta} |X_1(\beta, \alpha)|\right\|_{\psi_1} \leq 4\|\mathcal{G}\|_{L^\infty_\nu}^2 + 4\|\varphi\|_{L^\infty}^2\beta.$$

The $X_n(\beta, \cdot)$ do indeed belong to $C(\mathcal{A}_\beta; \mathbb{R})$ almost surely, as they can be written as a sum of affine and quadratic forms on $\mathbb{R}^M$ in the α variable. Application of (9) in corollary 2.7 (taking the separable Banach space to be $C(\mathcal{A}_\beta; \mathbb{R})$ equipped with the supremum norm) shows that if $N \geq \log(1/\delta)$, then conditioned on the realizations $\{\theta_m\}$ in the family Φ, it holds that

$$\mathcal{E}_\beta \leq \mathbb{E}_{\{u_n\}}[\mathcal{E}_\beta] + 32e^{3/2}\left(\|\mathcal{G}\|_{L^\infty_\nu}^2 + \|\varphi\|_{L^\infty}^2 t\right)\sqrt{\frac{\log(1/\delta)}{N}} \tag{45}$$

with probability at least $1 - \delta$.

Since the supremum concentrates around its mean, it remains to show that its mean is small as a function of the sample size. We do this by means of the technique of symmetrization.

By Giné-Zinn symmetrization [see [32], Section 4.2, Proposition 4.11, pp. 107-108],

$$\mathbb{E}_{\{u_n\}}[\mathcal{E}_\beta] \leq 2\mathbb{E}\sup_{\alpha\in\mathcal{A}_\beta}\left|\frac{1}{N}\sum_{n=1}^N \varepsilon_n\|\mathcal{G}(u_n) - \Phi(u_n; \alpha)\|_\mathcal{Y}^2\right|, \quad \text{where} \quad \varepsilon_n \overset{\text{iid}}{\sim} \text{Unif}(\{+1, -1\}),$$

since the original summands (conditioned on $\{\theta_m\}$) are independent. The expectation on the right-hand side is to be understood as the conditional expectation given $\{\theta_m\}$, i.e., $\mathbb{E}_{\{u_n\},\{\varepsilon_n\}}$ taken only with respect to the data and the Rademacher variables. The right-hand side is the Rademacher complexity of the RFM class composed with the square loss. Expanding the square, it is bounded above by

$$2\mathbb{E}_{\{u_n\},\{\varepsilon_n\}}\left|\frac{1}{N}\sum_{n=1}^N \varepsilon_n\|\mathcal{G}(u_n)\|_\mathcal{Y}^2\right| \tag{I}$$

$$+ 4\mathbb{E}_{\{u_n\},\{\varepsilon_n\}}\sup_{\alpha\in\mathcal{A}_\beta}\left|\frac{1}{N}\sum_{n=1}^N \varepsilon_n\langle\mathcal{G}(u_n), \Phi(u_n; \alpha)\rangle_\mathcal{Y}\right| \tag{II}$$

$$+ 2\mathbb{E}_{\{u_n\},\{\varepsilon_n\}}\sup_{\alpha\in\mathcal{A}_\beta}\left|\frac{1}{N}\sum_{n=1}^N \varepsilon_n\|\Phi(u_n; \alpha)\|_\mathcal{Y}^2\right| \tag{III}$$

We now estimate each term. The first term (I) satisfies the standard Monte Carlo bound

$$\text{(I)} \leq 2\left(\mathbb{E}\left|\frac{1}{N}\sum_{n=1}^N \varepsilon_n\|\mathcal{G}(u_n)\|_\mathcal{Y}^2\right|^2\right)^{1/2} = \frac{2}{\sqrt{N}}\left(\frac{1}{N}\sum_{n=1}^N \mathbb{E}\|\mathcal{G}(u_n)\|_\mathcal{Y}^4\right)^{1/2} \leq \frac{2\|\mathcal{G}\|_{L^\infty_\nu}^2}{\sqrt{N}}.$$

For the second term (II), we begin by estimating the empirical average on the set $\mathcal{A}_\beta$ as

$$\begin{aligned}\left|\frac{1}{N}\sum_{n=1}^N \varepsilon_n\langle\mathcal{G}(u_n), \Phi(u_n; \alpha)\rangle_\mathcal{Y}\right| &= \left|\frac{1}{M}\sum_{m=1}^M \alpha_m\left(\frac{1}{N}\sum_{n=1}^N \varepsilon_n\langle\mathcal{G}(u_n), \varphi(u_n; \theta_m)\rangle_\mathcal{Y}\right)\right| \\ &\leq \sqrt{\beta}\left(\frac{1}{M}\sum_{m=1}^M\left|\frac{1}{N}\sum_{n=1}^N \varepsilon_n\langle\mathcal{G}(u_n), \varphi(u_n; \theta_m)\rangle_\mathcal{Y}\right|^2\right)^{1/2}\end{aligned}$$

by the Cauchy-Schwarz inequality in $\mathbb{R}^M$. We deduce by Jensen's inequality and independence that

$$\begin{aligned}(\text{II}) &\leq \frac{4\sqrt{\beta}}{N}\left(\frac{1}{M}\sum_{m=1}^{M}\sum_{n,n'=1}^{N}\mathbb{E}\left[\varepsilon_n\varepsilon_{n'}\right]\mathbb{E}_{\{u_n\}}\left[\left\langle\mathcal{G}\left(u_n\right),\varphi\left(u_n;\theta_m\right)\right\rangle_{\mathcal{Y}}\left\langle\mathcal{G}\left(u_{n'}\right),\varphi\left(u_{n'};\theta_m\right)\right\rangle_{\mathcal{Y}}\right]\right)^{1/2}\\ &= \frac{4\sqrt{\beta}}{N}\left(\frac{1}{M}\sum_{m=1}^{M}\sum_{n=1}^{N}\mathbb{E}_{u\sim\nu}\left\langle\mathcal{G}(u),\varphi\left(u;\theta_m\right)\right\rangle_{\mathcal{Y}}^2\right)^{1/2}.\end{aligned}$$

A final application of the Cauchy-Schwarz inequality in $\mathcal{Y}$ in the last line shows that the second term (II) is bounded above by $4\sqrt{\beta}\|\mathcal{G}\|_{L^\infty_\nu}\|\varphi\|_{L^\infty}/\sqrt{N}$. By Young's inequality $ab \leq a^2/2 + b^2/2$, we further bound

$$\frac{4\|\mathcal{G}\|_{L^\infty_\nu}\|\varphi\|_{L^\infty}\sqrt{\beta}}{\sqrt{N}} = \left(\frac{2\|\mathcal{G}\|_{L^\infty_\nu}}{N^{1/4}}\right)\left(\frac{2\|\varphi\|_{L^\infty}\sqrt{\beta}}{N^{1/4}}\right) \leq \frac{2\|\mathcal{G}\|^2_{L^\infty_\nu}}{\sqrt{N}} + \frac{2\|\varphi\|^2_{L^\infty}\beta}{\sqrt{N}}.$$

The third term (III) is estimated in a similar way. Expanding the empirical average on $\mathcal{A}_\beta$ yields

$$\left|\frac{1}{N}\sum_{n=1}^{N}\varepsilon_n\left\|\Phi\left(u_n;\alpha\right)\right\|_{\mathcal{Y}}^2\right| = \left|\frac{1}{M}\sum_{m=1}^{M}\alpha_m\left(\frac{1}{M}\sum_{m'=1}^{M}\alpha_{m'}\gamma^{(N)}_{m,m'}\right)\right|, \quad \text{where}$$

$$\gamma^{(N)}_{m,m'} = \frac{1}{N}\sum_{n=1}^{N}\varepsilon_n\left\langle\varphi\left(u_n;\theta_m\right),\varphi\left(u_n;\theta_{m'}\right)\right\rangle_{\mathcal{Y}}$$

The first equality in the above display satisfies the upper bound

$$\begin{aligned}\sqrt{\beta}\left(\frac{1}{M}\sum_{m=1}^{M}\left|\frac{1}{M}\sum_{m'=1}^{M}\alpha_{m'}\gamma^{(N)}_{m,m'}\right|^2\right)^{1/2} &\leq \sqrt{\beta}\left(\frac{1}{M}\sum_{m=1}^{M}\beta\left[\frac{1}{M}\sum_{m'=1}^{M}\left|\gamma^{(N)}_{m,m'}\right|^2\right]\right)^{1/2}\\ &= \frac{\beta}{N}\sqrt{\frac{1}{M^2}\sum_{m,m'=1}^{M}\left|\sum_{n=1}^{N}\varepsilon_n\left\langle\varphi\left(u_n;\theta_m\right),\varphi\left(u_n;\theta_{m'}\right)\right\rangle_{\mathcal{Y}}\right|^2}\end{aligned}$$

by two applications of the Cauchy-Schwarz inequality in $\mathbb{R}^M$. Finally, we deduce that

$$\begin{aligned}(\text{III}) &\leq \frac{2\beta}{N}\sqrt{\frac{1}{M^2}\sum_{m,m'=1}^{M}\sum_{n=1}^{N}\mathbb{E}_{\{u_n\}}\left\langle\varphi\left(u_n;\theta_m\right),\varphi\left(u_n;\theta_{m'}\right)\right\rangle_{\mathcal{Y}}^2}\\ &\leq \frac{2\beta}{\sqrt{N}}\sqrt{\frac{1}{M^2}\sum_{m,m'=1}^{M}\mathbb{E}_u\left[\left\|\varphi\left(u;\theta_m\right)\right\|_{\mathcal{Y}}^2\left\|\varphi\left(u;\theta_{m'}\right)\right\|_{\mathcal{Y}}^2\right]}\\ &\leq \frac{2\beta\|\varphi\|^2_{L^\infty}}{\sqrt{N}}\end{aligned}$$

by Jensen's inequality, the fact that $\mathbb{E}\left[\varepsilon_n\varepsilon_{n'}\right] = \delta_{n,n'}$, and the Cauchy-Schwarz inequality in $\mathcal{Y}$. Combining the three estimates shows that

$$\mathbb{E}_{\{u_n\}}\left[\mathcal{E}_\beta\right] \leq \frac{4\|\mathcal{G}\|^2_{L^\infty_\nu}}{\sqrt{N}} + \frac{4\|\varphi\|^2_{L^\infty}}{\sqrt{N}}\beta. \tag{46}$$

Combining (45) and (46) yields that if $N \geq \log(1/\delta)$, then

$$\mathcal{E}_\beta\left(\{u_n\},\{\theta_m\}\right) \leq \frac{4\left(\|\mathcal{G}\|^2_{L^\infty_\nu} + \|\varphi\|^2_{L^\infty}\beta\right)}{\sqrt{N}} + 32e^{3/2}\left(\|\mathcal{G}\|^2_{L^\infty_\nu} + \|\varphi\|^2_{L^\infty}\beta\right)\sqrt{\frac{\log(1/\delta)}{N}} \tag{47}$$

with conditional probability (over $\{\theta_m\}$) at least $1-\delta$. Since δ does not depend on $\{\theta_m\}$, by the tower law of conditional expectation, we have that the event implied by (47) has $\mathbb{P}$-probability at least $1-\delta$ as well. The expression in (47) is bounded above by

$$32e^{3/2}\left(\|\mathcal{G}\|^2_{L^\infty_\nu} + \|\varphi\|^2_{L^\infty}\beta\right)\sqrt{\frac{2(1+\log(1/\delta))}{N}},$$

where we used the inequalities $4 \leq 32e^{3/2}$ and $\sqrt{a}+\sqrt{b} \leq \sqrt{2(a+b)}$. For $\delta \in (0,1)$, the inequality $1 \leq 2\log(2/\delta)$ holds true. We conclude that, if $N \geq \log(1/\delta)$, then with probability at least $1-\delta$ it holds that

$$\mathcal{E}_\beta\left(\{u_n\},\{\theta_m\}\right) \leq 32e^{3/2}\left(\|\mathcal{G}\|_{L^\infty_\nu}^2 + \beta\left(\rho,\lambda,\mathcal{G}_{\mathcal{H}},\eta\right)\right)\sqrt{\frac{6\log(2/\delta)}{N}}. \tag{48}$$

Finally, recall (43) as we have proved in **Step 1**, combining the fact that $\lambda\|\widehat{\alpha}\|_M^2 \leq \mathscr{R}_N^\lambda(\widehat{\alpha};\mathcal{G})$, we have that, if $M \geq \lambda^{-1}\log(16/\delta)$ and $N \geq \lambda^{-2}\log(8/\delta)$, then $\widehat{\alpha} \in \mathcal{A}_\beta$ with probability at least $1-\delta$, where the set $\mathcal{A}_\beta$ is defined as in (44). Thus, on the same event that (43) holds,

$$\mathscr{R}(\widehat{\alpha};\mathcal{G}) - \mathscr{R}_N(\widehat{\alpha};\mathcal{G}) \leq \sup_{\alpha\in\mathcal{A}_\beta}\left|\mathscr{R}_N(\alpha;\mathcal{G}) - \mathscr{R}(\alpha;\mathcal{G})\right|.$$

According to (48), we have that the right-hand side of the above display is bounded above by

$$32\sqrt{6}e^{3/2}\left(\|\mathcal{G}\|_{L^\infty_\nu}^2 + \|\varphi\|_{L^\infty}^2\beta\left(\rho,\lambda,\mathcal{G}_{\mathcal{H}},\eta\right)\right)\lambda \leq 79e^{3/2}\left(\|\mathcal{G}\|_{L^\infty_\nu}^2 + \|\varphi\|_{L^\infty}^2\beta\left(\rho,\lambda,\mathcal{G}_{\mathcal{H}},\eta\right)\right)\lambda \tag{49}$$

with probability at least $1-\delta$ because $N \geq \lambda^{-2}\log(8/\delta) \geq \log(2/\delta)$.

Step 3: Recall the error decomposition (29). We bound the first term in (29) with (43), and bound the second term with (49). Using $1 \leq 79e^{3/2}$, $\|\varphi\|_{L^\infty}^2 \leq 1$, and applying a union bound completes the proof of theorem 3.1.

4 Error bounds for learning in reproducing kernel Banach spaces

Section 4 presents the main results of the thesis. In Section 4.1, we introduce the assumptions of our setting, the key definitions in terms of empirical risk and population risk, and the main theorem of the thesis. In Section 4.2, we present a detailed proof of the main theorem of the thesis. The key contributions of the thesis include the tools we developed for the reproducing Banach spaces setting and a general proof framework of an upper bound of the test error performance for learning with neural networks. In Section 4.3, we discuss some consequences of the main theorem and outline some limitations and future work.

4.1 Setting and the goal of the thesis

The goal of this thesis is to establish generalization bounds for "convex neural networks" in reproducing kernel Banach spaces. One general proof framework is given in Lanthaler and Nelsen [14] as summarized in section 3.2. However, the regularizer in that setting is a reproducing kernel Hilbert space norm. It remains a challenge to adapt that framework to an RKBS setting. This thesis aims to bridge this gap by developing Banach space tools to deal with the RKBS regularizer.

Throughout this section, we assume that the input space $\mathcal{X}$ is a Polish space and the output space $\mathcal{Y}$ is a real separable Hilbert space. We view $\mathcal{X}$ and $\mathcal{Y}$ as measurable spaces equipped with their respective Borel σ-algebras. Let $\nu \in \mathscr{P}(\mathcal{X})$ be the input distribution. Key assumptions of our setting are:

Assumption 4.1 (Parameter space)**.** *Let Θ be a Hausdorff second-countable topological space, that can be seen as the parameter space. In addition, Θ is compact.*

Since Θ is compact Hausdorff, it is normal and thus regular. By Urysohn's metrization theorem (see, e.g. Munkres [19, Theorem 34.1, pp. 215-217]), Θ is metrizable. Then, Θ is separable because every compact metric space is separable. If Θ has finite topological dimension, then it is homeomorphic to a subspace of $\mathbb{R}^d$ for some d.

We denote by $\mathcal{M}(\Theta)$ the Banach space of bounded measures defined on the Borel $\sigma-$algebra of Θ, and endow $\mathcal{M}(\Theta)$ with the total variation norm $\|\cdot\|_{\mathrm{TV}}$. Following Bartolucci et al. [6, Section 3.3], the elements of $\mathcal{M}(\Theta)$ are finite Radon measures, and the Markov-Riesz representation theorem ensures that $\mathcal{M}(\Theta)$ can be identified with the dual of $C(\Theta)$, the Banach space of continuous functions on Θ endowed with the sup norm $\|\cdot\|_\infty$.

Assumption 4.2 (Feature map regularity)**.** *Let $\varphi : \mathcal{X}\times\Theta \to \mathbb{R}$ be a function such that $\varphi(x,\cdot)$ is measurable for every $x \in \mathcal{X}$. The function φ satisfies that, for every $x \in \mathcal{X}$, $\varphi(x,\cdot) \in C(\Theta)$. In addition, the map $\theta \mapsto \varphi(x,\theta)$ is K-Lipschitz on θ, uniformly in $x \in \mathcal{X}$. Futhurmore, φ is uniformly bounded over $\mathcal{X}\times\Theta$, that is, for all $\theta \in \Theta$, $\operatorname{ess\,sup}_{x\sim\nu}|\varphi(x,\theta)| < \infty$.*

Consider the RKBS $(\mathcal{B}, \|\cdot\|)$ of real-valued functions on $\mathcal{X}$ defined by the feature map $\phi : \mathcal{X} \to \mathcal{M}(\Theta)'$ given by

$$\phi(x)(\mu) = {}_{\mathcal{M}(\Theta)}\langle \mu, \phi(x) \rangle_{\mathcal{M}(\Theta)'} = \int_\Theta \varphi(x,\theta) d\mu(\theta) \tag{50}$$

for every $x \in \mathcal{X}$ and $\mu \in \mathcal{M}(\Theta)$. By (18), and (19), it holds that

$$\mathcal{B} = \{f_\mu : \mathcal{X} \to \mathbb{R} \mid \|\mu\|_{\mathrm{TV}} < \infty\}, \quad \text{where } f_\mu = \int_\Theta \varphi(\cdot,\theta) d\mu(\theta). \tag{51}$$

By (20), the norm on $\mathcal{B}$ is the quotient norm

$$\|f\|_{\mathcal{B}} = \inf_{f=f_\mu} \|\mu\|_{\mathrm{TV}}.$$

Under Assumption 4.2, we can assume that $\|\varphi\|_{L^\infty} \leq b$ for some $b > 0$, and we have $\mathcal{B} \subset L^\infty(\mathcal{X}; \mathbb{R})$.

Assumption 4.3 (Joint data distribution)**.** *The joint distribution $\mathbb{Q} \in \mathscr{P}(\mathcal{X} \times \mathcal{Y})$ of random variable $(x, y) \sim \mathbb{Q}$ is given by $x \sim \nu$ with $\nu \in \mathscr{P}(\mathcal{X})$ and $y = f^\star(x) + \eta$. The addictive noise η is a random variable in $\mathcal{Y}$ that is conditionally centered: $\mathbb{E}[\eta|x] = 0$, and is subexponential: $\|\eta\|_{\psi_1(\mathcal{Y})} < \infty$.*

Here, we think of $f^\star$ as the "true function" of the data such that the output y is the transform of the input x through $f^\star$ with some additional noise. The properties that the noise η is conditionally centered and subexponential are crucial in our arguments.

Assumption 4.4 (Well-specification)**.** *The operator $f^\star : \mathcal{X} \to \mathcal{Y}$ in Assumption 4.3 satisfies $f^\star \in \mathcal{B}$.*

Assumption 4.4 states that the true function $f^\star$ proposed in Assumption 4.3 belongs to the RKBS generated by the feature map ϕ, where ϕ is defined following Assumption 4.2. Thus, there exists $\mu^\star \in \mathcal{M}(\Theta)$ such that $f^\star = f_{\mu^\star}$. Future work may consider scenarios in which $f^\star \notin \mathcal{B}$, such as misspecified models, and address them by employing existing approximation bounds.

Assumption 4.5 (Output space)**.** *The output space $\mathcal{Y} = \mathbb{R}$.*

Assumption 4.5 is introduced solely to enable the application of a result showing that the covering numbers of VC-classes of functions has a polynomial decay, which is essential for bounding the estimation error in our approach. The classical theory of VC-classes of functions, as developed in [29], is formulated for real-valued function classes and therefore does not directly extend to vector-valued functions. With the exception of this technical restriction, the entire analysis remains valid in the vector-valued setting. In our analysis, we consistently denote the norm in $\mathcal{Y}$ by $\|\cdot\|_{\mathcal{Y}}$ rather than by the absolute value, to ease future extensions to the vector-valued settings. Future work may extend the VC-classes framework to vector-valued function classes by employing generalizations of VC-dimension, for example, the Natarajan dimension (see, e.g. Shalev-Shwartz and Ben-David [27, Section 29, pp. 402-409]).

Now, suppose we have access to a dataset $\{(x_n, y_n)\}_{n=1}^N \sim \mathbb{Q}^{\otimes N}$ according to Assumption 4.3 and Assumption 4.4. We are interested in the regularized empirical risk minimization (ERM)

$$\inf_{\mu \in \mathcal{M}(\Theta)} \frac{1}{N} \sum_{n=1}^N \|y_n - f_\mu(x_n)\|_{\mathcal{Y}}^2 + \lambda \|\mu\|_{\mathrm{TV}}, \quad \text{where } f_\mu(x) = \int_\Theta \varphi(x,\theta) d\mu(\theta). \tag{52}$$

Here we used Proposition 2.19 for the squared loss function. The goal of the theoretical analysis is to develop error bounds for the test error performance

$$\mathbb{E}_{x \sim v} \|f^\star(x) - f_{\widehat{\mu}}(x)\|_{\mathcal{Y}}^2$$

of a minimizer $\widehat{\mu}$ of ERM, especially as $N \to \infty$ and $\lambda = \lambda_N \to 0$.

Next, we outline the notation used in this section.

Definition 4.1 (Empirical risk)**.** *Writing $Y = \{y_n\}$ for the collection of observed output data and fixing a regularization parameter $\lambda > 0$, the regularized Y-empirical risk of $\mu \in \mathcal{M}(\Theta)$ is given by*

$$\mathscr{R}_N^\lambda(\mu; Y) := \frac{1}{N} \sum_{n=1}^N \|y_n - f_\mu(x_n)\|_{\mathcal{Y}}^2 + \lambda \|\mu\|_{\mathrm{TV}}.$$

The regularized $f^\star$-empirical risk, $\mathscr{R}_N^\lambda(\mu, f^\star)$, is defined analogously by

$$\mathscr{R}_N^\lambda(\mu; f^\star) := \frac{1}{N} \sum_{n=1}^N \|f^\star(x_n) - f_\mu(x_n)\|_{\mathcal{Y}}^2 + \lambda \|\mu\|_{\mathrm{TV}}. \tag{53}$$

In the absence of regularization, i.e., $\lambda = 0$, these expressions define the Y-empirical risk and $f^\star$-empirical risk, denoted by $\mathscr{R}_N(\mu; Y)$ and $\mathscr{R}_N(\mu; f^\star)$, respectively.

Definition 4.2 (Population risk)**.** *The regularized joint population risk of $\mu \in \mathcal{M}(\Theta)$ is defined by*

$$\mathscr{R}^{\lambda}(\mu; \mathbb{Q}) := \mathbb{E}_{(x,y)\sim\mathbb{Q}} \|y - f_{\mu}(x)\|_{\mathcal{Y}}^2 + \lambda \|\mu\|_{\mathrm{TV}}.$$

The regularized $f^{\star}$-population risk, $\mathscr{R}^{\lambda}(\mu, f^{\star})$, is defined analogously by

$$\mathscr{R}^{\lambda}(\mu; f^{\star}) := \mathbb{E}_{x\sim v} \|f^{\star}(x) - f_{\mu}(x)\|_{\mathcal{Y}}^2 + \lambda \|\mu\|_{\mathrm{TV}}.$$

In the absence of regularization, i.e., $\lambda = 0$, these expressions define the Y-population risk and $f^{\star}$-population risk, denoted by $\mathscr{R}(\mu; \mathbb{Q})$ and $\mathscr{R}(\mu; f^{\star})$, respectively.

The main result of this thesis is the following theorem.

Theorem 4.1. *Suppose that the parameter space Θ satisfies assumption 4.1. Let $(\mathcal{B}, \|\cdot\|_{\mathcal{B}})$ be the RKBS of the real-valued functions on $\mathcal{X}$ defined by the feature map $\phi : \mathcal{X} \to \mathcal{M}(\Theta)'$ given by (50), where φ satisfies assumption 4.2. Denote K_{φ} as the Lipschitz constant, uniformly in $x \in \mathcal{X}$, of the map $\theta \mapsto \varphi(x, \theta)$. Fix a failure probability $\delta \in (0, 1)$ and sample size N. Let $\{(x_n, y_n)\} \sim \mathbb{Q}^{\otimes N}$ be the data according to assumption 4.3. For $f^{\star}$ the true function in assumption 4.3 satisfying assumption 4.4, let $\mu^{\star} \in \mathcal{M}(\Theta)$ be such that $f^{\star} = f_{\mu^{\star}}$, and let $\widehat{\mu} \in \mathcal{M}(\Theta)$ be the minimizer of the regularized $f^{\star}$-empirical risk $\mathscr{R}_N^{\lambda}(\cdot, f^{\star})$ given by (53), where $\lambda \in \mathbb{R}_+$ is the regularization strength. If $N \geq \log(4/\delta)$, $\lambda \gtrsim b \|\eta\|_{\psi_1(\mathcal{Y})} \log^{1/2}(2/\delta) N^{-1/2}$, and Θ has a finite topological dimension $d > 2$, then*

$$\mathbb{E}_{x\sim v} \|f^{\star}(x) - f_{\widehat{\mu}}(x)\|_{\mathcal{Y}}^2 \lesssim \|\mu^{\star}\|_{\mathrm{TV}} \lambda + b \|\mu^{\star}\|_{\mathrm{TV}} \mathcal{I} \lambda^{2/d} + b \|\mu^{\star}\|_{\mathrm{TV}}^2 \lambda \tag{54}$$

with probability at least $1 - \delta$, where

$$\mathcal{I} \lesssim C_d \log^{1/2}\left(b \|\mu^{\star}\|_{\mathrm{TV}} \lambda^{-1}\right) \|\mu^{\star}\|_{\mathrm{TV}}^{d/2}. \tag{55}$$

Remark 4.1. *The order of convergence that Theorem 4.1 gives is $\mathcal{O}(N^{-1/d}\sqrt{\log N})$. This rate is mainly due to the second term of (54).*

Remark 4.2. *The cases where $d = 1$ and $d = 2$ can also be handled. Indeed, we obtain that if $N \geq \log(4/\delta)$ and $\lambda \gtrsim b \|\eta\|_{\psi_1(\mathcal{Y})} \log^{1/2}(2/\delta) N^{-1/2}$, then*

$$\mathbb{E}_{x\sim v} \|f^{\star}(x) - f_{\widehat{\mu}}(x)\|_{\mathcal{Y}}^2 \lesssim \begin{cases} \|\mu^{\star}\|_{\mathrm{TV}} \lambda + \lambda + b \|\mu^{\star}\|_{\mathrm{TV}}^2 \lambda, & \text{if } d = 1, \\ \|\mu^{\star}\|_{\mathrm{TV}} \lambda + \lambda \log^{3/2}\left(b \|\mu^{\star}\|_{\mathrm{TV}} \lambda^{-1}\right) + b \|\mu^{\star}\|_{\mathrm{TV}}^2 \lambda, & \text{if } d = 2, \end{cases}$$

with probability at least $1 - \delta$ over $\{x_n\} \sim \nu^{\otimes N}$. The order of convergence of the above is $\mathcal{O}\left(N^{-1/2}\right)$ if $d = 1$ and $\mathcal{O}\left(N^{-1/2} \log^{3/2}(N)\right)$ if $d = 2$.

4.2 Proof of theorem 4.1

We begin with the standard error decomposition

$$\mathscr{R}\left(\widehat{\mu}; f^{\star}\right) = \mathscr{R}_N\left(\widehat{\mu}; f^{\star}\right) + \left[\mathscr{R}\left(\widehat{\mu}; f^{\star}\right) - \mathscr{R}_N\left(\widehat{\mu}; f^{\star}\right)\right].$$

The first term is the approximation error and the second term is the generalization gap, i.e., the estimation error.

4.2.1 Step 1: Approximation error

We aim to bound the approximation error $\mathscr{R}_N(\widehat{\mu}; f^{\star})$. We proceed by bounding $\mathscr{R}_N(\widehat{\mu}; f^{\star}) \leq \mathscr{R}_N^{\lambda}(\widehat{\mu}; f^{\star})$.

Fix any $\mu \in \mathcal{M}(\Theta)$. Using $y_n = f^{\star}(x_n) + \eta_n$ and expanding the square, we get

$$\begin{aligned} \mathscr{R}_N^{\lambda}(\mu; Y) &= \frac{1}{N} \sum_{n=1}^{N} \|y_n - f_{\mu}(x_n)\|_{\mathcal{Y}}^2 + \lambda \|\mu\|_{\mathrm{TV}} \\ &= \frac{1}{N} \sum_{n=1}^{N} \|f^{\star}(x_n) + \eta_n - f_{\mu}(x_n)\|_{\mathcal{Y}}^2 + \lambda \|\mu\|_{\mathrm{TV}} \\ &= \mathscr{R}_N^{\lambda}\left(\mu; f^{\star}\right) + \frac{2}{N} \sum_{n=1}^{N} \left\langle \eta_n, f^{\star}(x_n) - f_{\mu}(x_n) \right\rangle_{\mathcal{Y}} + \frac{1}{N} \sum_{n=1}^{N} \|\eta_n\|_{\mathcal{Y}}^2. \end{aligned}$$

It follows that

$$\begin{aligned}\mathscr{R}_N^\lambda(\widehat{\mu};f^\star)-\mathscr{R}_N^\lambda(\mu;f^\star)&=\mathscr{R}_N^\lambda(\widehat{\mu};Y)-\frac{2}{N}\sum_{n=1}^N\langle\eta_n,f^\star(x_n)-f_{\widehat{\mu}}(x_n)\rangle_{\mathcal{Y}}\\&\quad-\mathscr{R}_N^\lambda(\mu;Y)+\frac{2}{N}\sum_{n=1}^N\langle\eta_n,f^\star(x_n)-f_{\mu}(x_n)\rangle_{\mathcal{Y}}\\&=\mathscr{R}_N^\lambda(\widehat{\mu};Y)-\mathscr{R}_N^\lambda(\mu;Y)+\frac{2}{N}\sum_{n=1}^N\langle\eta_n,f_{\widehat{\mu}}(x_n)-f_{\mu}(x_n)\rangle_{\mathcal{Y}}.\end{aligned}$$

Using this and the fact that $\widehat{\mu}$ minimizes $\mathscr{R}_N^\lambda(\cdot;Y)$, we obtain

$$\begin{aligned}\mathscr{R}_N^\lambda(\widehat{\mu};f^\star)&=\left[\mathscr{R}_N^\lambda(\widehat{\mu};f^\star)-\mathscr{R}_N^\lambda(\mu;f^\star)\right]+\mathscr{R}_N^\lambda(\mu;f^\star)\\&=\left[\mathscr{R}_N^\lambda(\widehat{\mu};Y)-\mathscr{R}_N^\lambda(\mu;Y)\right]+\mathscr{R}_N^\lambda(\mu;f^\star)+\frac{2}{N}\sum_{n=1}^N\langle\eta_n,f_{\widehat{\mu}}(x_n)-f_{\mu}(x_n)\rangle_{\mathcal{Y}}\\&\le\mathscr{R}_N^\lambda(\mu;f^\star)+\frac{2}{N}\sum_{n=1}^N\langle-\eta_n,f_{\mu}(x_n)\rangle_{\mathcal{Y}}+\frac{2}{N}\sum_{n=1}^N\langle\eta_n,f_{\widehat{\mu}}(x_n)\rangle_{\mathcal{Y}}.\end{aligned}\tag{56}$$

Since μ is arbitrary, we can choose $\mu=\mu^\star$ such that the first term in (56) is small. In this case,

$$\mathscr{R}_N^\lambda(\mu;f^\star)=\lambda\|\mu^\star\|_{\mathrm{TV}}.\tag{57}$$

Next, define

$$t:=\|\widehat{\mu}\|_{\mathrm{TV}},\tag{58}$$

$$A_N^\lambda:=\mathscr{R}_N^\lambda(\mu;f^\star)+\frac{2}{N}\sum_{n=1}^N\langle-\eta_n,f_{\mu}(x_n)\rangle_{\mathcal{Y}},\quad\text{and}\tag{59}$$

$$B_N:=2\sup_{\|\mu\|_{\mathrm{TV}}\le1}\left|\frac{1}{N}\sum_{n=1}^N\langle\eta_n,f_{\mu}(x_n)\rangle_{\mathcal{Y}}\right|.\tag{60}$$

By the linearity of the map $\mu\mapsto f_\mu$ (following from (51)), it holds that

$$\begin{aligned}\frac{2}{N}\sum_{n=1}^N\langle\eta_n,f_{\widehat{\mu}}(x_n)\rangle_{\mathcal{Y}}&=\frac{2}{N}\sum_{n=1}^N\left\langle\eta_n,\int_\Theta\varphi(x_n;\theta)d\widehat{\mu}(\theta)\right\rangle_{\mathcal{Y}}\\&\le\left|\frac{2}{N}\sum_{n=1}^N\|\widehat{\mu}\|_{\mathrm{TV}}\left\langle\eta_n,\sup_{\|\mu\|_{\mathrm{TV}}\le1}\int_\Theta\varphi(x_n;\theta)d\mu(\theta)\right\rangle_{\mathcal{Y}}\right|\\&\le tB_N.\end{aligned}\tag{61}$$

Thus,

$$t\lambda\le\mathscr{R}_N^\lambda(\widehat{\mu};f^\star)\le A_N^\lambda+tB_N,\tag{62}$$

where the first inequality follows from definition (53) of $\mathscr{R}_N^\lambda(\mu;f^\star)$, and the second inequality follows from (56) and (61).

We now depart from the RKHS setting of Lanthaler and Nelsen [14]. The following argument is required for the loss of Cauchy-Schwarz inequality in RKBSs. Define the event

$$\mathrm{E}:=\left\{B_N\le\frac{\lambda}{2}\right\}.\tag{63}$$

This event couples λ with N, and in particular, we expect $\lambda\gtrsim N^{-1/2}$ on this event. We must later show that $\mathbb{P}(\mathrm{E})$ is close to one. On the event E, it holds that

$$tB_N\le\frac{t\lambda}{2}.\tag{64}$$

Applying (64) on tB_N in (62), subtracting $t\lambda/2$ and multiplying through by two on both sides in (62) shows that

$$t\lambda\le2A_N^\lambda\quad\text{on}\quad\mathrm{E},$$

and hence

$$t \leq 2\lambda^{-1} A_N^\lambda \quad \text{on} \quad \mathrm{E}. \tag{65}$$

Plugging (64) and (65) back in to the right hand side of (62) leads to

$$\mathscr{R}_N^\lambda\left(\widehat{\mu}; f^\star\right) \leq 2A_N^\lambda \quad \text{on} \quad \mathrm{E}. \tag{66}$$

It remains to show that A_N^λ is small and $\mathbb{P}(\mathrm{E})$ is large.

We first show the former. With μ fixed, the second term in (56) averages to zero by our assumptions on the noise, and hence, we expect it to be small with high probability. Indeed, it holds that

$$\begin{aligned}
\mathbb{E}\left[\langle \eta_1, f_\mu(x_1)\rangle_{\mathcal{Y}}\right] &= \mathbb{E}_{(x_1,\eta_1)}\left[\langle \eta_1, f_\mu(x_1)\rangle_{\mathcal{Y}}\right] \\
&= \mathbb{E}_{x_1\sim\nu}\left[\mathbb{E}\left[\langle \eta_1, f_\mu(x_1)\rangle_{\mathcal{Y}} \mid x_1\right]\right] \\
&= \mathbb{E}_{x_1\sim\nu}\left[\langle \mathbb{E}\left[\eta_1 \mid x_1\right], f_\mu(x_1)\rangle_{\mathcal{Y}}\right] \\
&= 0,
\end{aligned} \tag{67}$$

and

$$|\langle \eta_1, f_\mu(x_1)\rangle_{\mathcal{Y}}| \leq \|\eta_1\|_{\mathcal{Y}} \|f_\mu(x_1)\|_{\mathcal{Y}} \leq \|\eta_1\|_{\mathcal{Y}}\, b\, \|\mu\|_{\mathrm{TV}}, \tag{68}$$

where the first inequality used Cauchy-Schwarz inequality in Hilbert space $\mathcal{Y}$, and the second inequality follows from (51) and our assumption of uniform boundedness on φ. Then, a similar argument to that of Lanthaler and Nelsen [14, Lemma 4.6] gives the next result.

Lemma 4.2 (single noise cross term)**.** *Given any $\delta > 0$, for $N \geq \log(1/\delta)$, it holds that*

$$\frac{2}{N}\sum_{n=1}^{N} \langle -\eta_n, f_\mu(x_n)\rangle_{\mathcal{Y}} \leq 16e^{3/2} b\, \|\eta_1\|_{\psi_1(\mathcal{Y})}\, \|\mu\|_{\mathrm{TV}} \sqrt{\frac{\log(1/\delta)}{N}}$$

with probability at least $1-\delta$ over $\{x_n\} \sim \nu^{\otimes N}$.

Proof. Define $Z_n := \langle -\eta_n, f_\mu(x_n)\rangle_{\mathcal{Y}}$ for each n. It holds that Z_n is an i.i.d. copy of Z_1. By definition 7 of ψ_1 norm and (68), we have that

$$\begin{aligned}
\|Z_1\|_{\psi_1} &= \sup_{p\in[1,\infty)} \frac{\left(\mathbb{E}_{\boldsymbol{\eta},\boldsymbol{x}} \left|\langle -\eta_1, f_\mu(x_1)\rangle_{\mathcal{Y}}\right|^p\right)^{1/p}}{p} \\
&\leq b\, \|\mu\|_{\mathrm{TV}} \sup_{p\in[1,\infty)} \frac{\left(\mathbb{E}\, \|\eta_1\|_{\mathcal{Y}}^p\right)^{1/p}}{p} \\
&= \|\eta_1\|_{\psi_1(\mathcal{Y})}\, b \|\mu\|_{\mathrm{TV}}.
\end{aligned} \tag{69}$$

Since η is assumed to be subexponential, we have that Z_1 is also subexponential. Proposition 2.6 states that a subexponential random variable satisfies Bernstein moment condition (5) with $\sigma^2 := 4e\sqrt{\mathbb{E}\|Z - \mathbb{E}Z\|^2}\|Z\|_{\psi_1}$ and $b' := 4e\|Z\|_{\psi_1}$. Then, theorem 2.5 (Bernstein inequality) combining with (67) yields that

$$\frac{1}{N}\sum_{n=1}^{N} \langle -\eta_n, f_\mu(x_n)\rangle_{\mathcal{Y}} \leq \frac{8e\, \|Z_1\|_{\psi_1} \log(1/\delta)}{N} + \sqrt{\frac{8e\sqrt{\mathbb{E}\, \|Z_1 - \mathbb{E}Z_1\|^2}\, \|Z_1\|_{\psi_1} \log(1/\delta)}{N}} \tag{70}$$

with probability at least $1-\delta$ over $\{x_n\} \sim \nu^{\otimes N}$. Note that $\mathbb{E}\, \|Z_1 - \mathbb{E}Z_1\|^2 \leq 4\mathbb{E}\, \|Z_1\|^2$ (by triangle inequality and using $(a+b)^2 \leq 2\left(a^2+b^2\right)$) and $4\, \|Z_1\|_{\psi_1} \geq 2\sqrt{\mathbb{E}\, \|Z_1\|^2}$ (by definition of ψ_1 norm (7)). Since $N \geq \log(1/\delta)$, we have $\log(1/\delta)/N \leq \sqrt{\log(1/\delta)/N}$. Combining these facts, it follows that the right hand side of (70) is bounded above by

$$\sqrt{\frac{64e^2\, \|Z_1\|_{\psi_1}^2 \log(1/\delta)}{N}} + \sqrt{\frac{32e\, \|Z_1\|_{\psi_1}^2 \log(1/\delta)}{N}} \leq \sqrt{\frac{64\left(2e^2+e\right) \|Z_1\|_{\psi_1}^2 \log(1/\delta)}{N}},$$

where we used $\sqrt{a} + \sqrt{b} \leq \sqrt{2(a+b)}$ on the right. Noting that $2e^2 + e \leq e^3$ and combining (69) completes the proof. □

By (57) and lemma 4.2, we have that for any $\delta > 0$ and $N \geq \log(2/\delta)$,

$$\begin{aligned} A_N^\lambda &\leq \lambda \|\mu^\star\|_{\mathrm{TV}} + 16e^{3/2} b \, \|\eta_1\|_{\psi_1(\mathcal{Y})} \, \|\mu^\star\|_{\mathrm{TV}} \sqrt{\frac{\log(2/\delta)}{N}} \\ &= \left(16e^{3/2} b \, \|\eta_1\|_{\psi_1(\mathcal{Y})} \sqrt{\frac{\log(2/\delta)}{N}} + \lambda \right) \|\mu^\star\|_{\mathrm{TV}} \end{aligned} \tag{71}$$

on some event E_δ with probability at least $1 - \frac{\delta}{2}$ over $\{x_n\} \sim \nu^{\otimes N}$.

We next show that $\mathbb{P}(\mathrm{E})$ is large. The key step is to control B_N. Our initial strategy follows the argument from Step 1 of Section 3.2: applying a Bernstein-type concentration inequality to bound B_N, and then estimating its expectation via a covering-number method. This involves a one-step discretization followed by chaining, using Dudley's entropy integral bound, together with remarks on the growth rates of the covering numbers. These results are presented in Section 4.2.1 (**Alternative approaches**). A closer examination shows that B_N is bounded by the supremum of a 1-Orlicz process. As the process is already conditionally centered, symmetrization is unnecessary, and Theorem 2.15 can be applied directly. This refined approach is presented in Section 4.2.1 (**Refined approach**).

Alternative approaches We proceed by first deriving a bound on the expectation of B_N, and then show that B_N concentrates around its expectation.

To bound the expectation of B_N, we adopt a technique known as symmetrization, which relates the absolute deviation between the sample average and the population average (uniformly over the class $\mathscr{F}$) with the Rademacher complexity. Then, we aim to apply Massart's finite class lemma to upper bound the Rademacher complexity.

$$\begin{aligned} \mathbb{E}\left[B_N/2\right] &= \mathbb{E}_{\boldsymbol{\eta},\boldsymbol{x}} \sup_{\|\mu\|_{\mathrm{TV}} \leq 1} \left| \frac{1}{N} \sum_{n=1}^N \langle \eta_n, f_\mu(x_n) \rangle_{\mathcal{Y}} \right| \\ &= \mathbb{E}_{\boldsymbol{\eta},\boldsymbol{x}} \sup_{\|\mu\|_{\mathrm{TV}} \leq 1} \left| \frac{1}{N} \sum_{n=1}^N \langle \eta_n, f_\mu(x_n) \rangle_{\mathcal{Y}} - \mathbb{E}\left[\langle \eta_1, f_\mu(x_1) \rangle_{\mathcal{Y}}\right] \right| \\ &= \mathbb{E}_{\boldsymbol{\eta},\boldsymbol{x}} \left[\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}} \right], \quad \text{by definition (12)} \\ &\leq 2\mathbb{E}_{\boldsymbol{\sigma},\boldsymbol{\eta},\boldsymbol{x}} \left[\|\mathbb{S}_n\|_{\mathscr{F}} \right] \\ &= 2\mathbb{E}_{\boldsymbol{\sigma},\boldsymbol{\eta},\boldsymbol{x}} \sup_{\|\mu\|_{\mathrm{TV}} \leq 1} \left| \frac{1}{N} \sum_{n=1}^N \sigma_n \langle \eta_n, f_\mu(x_n) \rangle_{\mathcal{Y}} \right|, \quad \sigma_n \overset{\text{i.i.d.}}{\sim} \text{Rademacher} \\ &= 2\mathbb{E}_{\boldsymbol{\sigma},\boldsymbol{\eta},\boldsymbol{x}} \sup_{\|\mu\|_{\mathrm{TV}} \leq 1} \left| \int_\Theta \frac{1}{N} \sum_{n=1}^N \sigma_n \langle \eta_n, \varphi(x_n;\theta) \rangle_{\mathcal{Y}} \, d\mu(\theta) \right| \\ &\leq 2\mathbb{E}_{\boldsymbol{\sigma},\boldsymbol{\eta},\boldsymbol{x}} \left\| \frac{1}{N} \sum_{n=1}^N \sigma_n \langle \eta_n, \varphi(x_n;\cdot) \rangle_{\mathcal{Y}} \right\|_{L^\infty(\Theta;\mathbb{R})}, \end{aligned}$$

where the second equality is due to (67), and the first inequality follows from applying Giné-Zinn symmetrization (proposition 2.10 and remark 2.4) with $\mathscr{F} := \{(\eta, x) \mapsto \langle \eta, f_\mu(x) \rangle_{\mathcal{Y}} : \|\mu\|_{\mathrm{TV}} \leq 1\}$. Since $\theta \mapsto \varphi(\cdot;\theta)$ is assumed to be continuous, the L^∞ norm is equal to the $C(\Theta)$ supremum norm. That is,

$$\mathbb{E}\left[B_N/2\right] \leq 2\mathbb{E}_{\boldsymbol{\sigma},\boldsymbol{\eta},\boldsymbol{x}} \sup_{\theta \in \Theta} \left| \frac{1}{N} \sum_{n=1}^N \sigma_n \langle \eta_n, \varphi(x_n;\theta) \rangle_{\mathcal{Y}} \right|. \tag{72}$$

We then upper bound the expectation of the supremum of the empirical process through a covering-number-type analysis. To do so, we first need to define a suitable pseudometric on a suitable set.

For $\theta, \theta' \in \Theta$, define

$$D(\theta, \theta') := \left(\frac{1}{N} \sum_{n=1}^N \|\varphi(x_n;\theta) - \varphi(x_n;\theta')\|_{\mathcal{Y}}^2 \right)^{1/2}. \tag{73}$$

It is easy to check that $D(\cdot,\cdot)$ is a pseudometric (see definition 2.1) on Θ. In the following, we apply the technique of one-step discretization and the technique of chaining, respectively, to deliver an upper bound of $\mathbb{E}[B_N]$. Although chaining helps to obtain a better upper bound, it does require a stronger assumption on the noise η.

(1) One-step discretization: Let $\epsilon > 0$ be arbitrary. Following definition 2.2 of an ϵ-cover, let $\widehat{\Theta} = \{\hat{\theta}_i\}_{i=1}^{\mathfrak{C}(\epsilon;\Theta,D)}$ be an ϵ-cover for Θ. Since Θ is assumed to be compact with $\Theta \subseteq \mathbb{R}^{d+1}$, the ϵ-covering number $\mathfrak{C}(\epsilon;\Theta,D)$ is finite. Now, fix $\theta \in \Theta$. Let $\hat{\theta}_i \in \widehat{\Theta}$ be such that $D(\theta,\hat{\theta}_i) \leq \epsilon$ for some $i \in \{1,\cdots,\mathfrak{C}(\epsilon;\Theta,D)\}$. Then,

$$\begin{aligned}
\frac{1}{N}\sum_{n=1}^{N}\sigma_n \langle \eta_n, \varphi(x_n;\theta)\rangle_{\mathcal{Y}} &= \frac{1}{N}\sum_{n=1}^{N}\left[\sigma_n \langle \eta_n, \varphi(x_n;\theta)\rangle_{\mathcal{Y}} - \sigma_n \left\langle \eta_n, \varphi(x_n;\hat{\theta}_i)\right\rangle_{\mathcal{Y}}\right] + \frac{1}{N}\sum_{n=1}^{N}\sigma_n \left\langle \eta_n, \varphi(x_n;\hat{\theta}_i)\right\rangle_{\mathcal{Y}} \\
&\leq \frac{1}{N}\sum_{n=1}^{N}|\sigma_n|\,\|\eta_n\|_{\mathcal{Y}}\left\|\varphi(x_n;\theta)-\varphi(x_n;\hat{\theta}_i)\right\|_{\mathcal{Y}} + \frac{1}{N}\sum_{n=1}^{N}\sigma_n \left\langle \eta_n, \varphi(x_n;\hat{\theta}_i)\right\rangle_{\mathcal{Y}} \\
&\leq \left(\frac{1}{N}\sum_{n=1}^{N}\|\eta_n\|_{\mathcal{Y}}^2\right)^{1/2} D(\theta,\hat{\theta}_i) + \frac{1}{N}\sum_{n=1}^{N}\sigma_n \left\langle \eta_n, \varphi(x_n;\hat{\theta}_i)\right\rangle_{\mathcal{Y}},
\end{aligned}$$

where we used Cauchy-Schwarz inequality twice, one on the Hilbert space $\mathcal{Y}$, and the other on $\mathbb{R}^N$.
Note that

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{1}{N}\sum_{n=1}^{N}\|\eta_n\|_{\mathcal{Y}}^2\right)^{1/2}\right] &\leq \left(\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\,\|\eta_n\|_{\mathcal{Y}}^2\right)^{1/2} \\
&= \left(\mathbb{E}\,\|\eta_1\|_{\mathcal{Y}}^2\right)^{1/2} < \infty,
\end{aligned} \tag{74}$$

where we used Jensen's inequality on the first line, and since η is assumed to be subexponential, it has finite moments (see, e.g. Vershynin [31, Proposition 2.7.1, pp. 32-34]). Then, (72) is bounded above by

$$2\left(\mathbb{E}\,\|\eta_1\|_{\mathcal{Y}}^2\right)^{1/2}\epsilon + 2\mathbb{E}_{\boldsymbol{\sigma},\boldsymbol{\eta},\boldsymbol{x}}\sup_{\theta\in\Theta}\left[\frac{1}{N}\sum_{n=1}^{N}\sigma_n\left\langle \eta_n, \varphi(x_n;\hat{\theta}_i)\right\rangle_{\mathcal{Y}}\right]. \tag{75}$$

Note that

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\sigma},\boldsymbol{\eta},\boldsymbol{x}}\sup_{\theta\in\Theta}\left[\frac{1}{N}\sum_{n=1}^{N}\sigma_n\left\langle \eta_n, \varphi(x_n;\hat{\theta}_i)\right\rangle_{\mathcal{Y}}\right] &\leq \mathbb{E}_{\boldsymbol{\sigma},\boldsymbol{\eta},\boldsymbol{x}}\sup_{\theta\in\Theta}\sup_{\hat{\theta}\in\widehat{\Theta}}\left[\frac{1}{N}\sum_{n=1}^{N}\sigma_n\left\langle \eta_n, \varphi(x_n;\hat{\theta})\right\rangle_{\mathcal{Y}}\right] \\
&= \mathbb{E}_{\boldsymbol{\sigma},\boldsymbol{\eta},\boldsymbol{x}}\sup_{\hat{\theta}\in\widehat{\Theta}}\left[\frac{1}{N}\sum_{n=1}^{N}\sigma_n\left\langle \eta_n, \varphi(x_n;\hat{\theta})\right\rangle_{\mathcal{Y}}\right] \\
&\leq \max_{\hat{\theta}\in\widehat{\Theta}}\left(\frac{1}{N}\sum_{n=1}^{N}\left\langle \eta_n, \varphi(x_n;\hat{\theta})\right\rangle_{\mathcal{Y}}^2\right)^{1/2}\sqrt{\frac{2\log\mathfrak{C}(\epsilon;\Theta,D)}{N}} \qquad (76) \\
&\leq b\mathbb{E}\left(\frac{1}{N}\sum_{n=1}^{N}\|\eta_n\|_{\mathcal{Y}}^2\right)^{1/2}\sqrt{\frac{2\log\mathfrak{C}(\epsilon;\Theta,D)}{N}} \qquad (77) \\
&\leq b\left(\mathbb{E}\,\|\eta_1\|_{\mathcal{Y}}^2\right)^{1/2}\sqrt{\frac{2\log\mathfrak{C}(\epsilon;\Theta,D)}{N}}, \qquad (78)
\end{aligned}$$

where we used Massart's lemma (lemma 2.11) for $\mathscr{F} = \{(\eta,x) \mapsto \left\langle \eta, \varphi(x;\hat{\theta})\right\rangle_{\mathcal{Y}} : \hat{\theta} \in \widehat{\Theta}, |\widehat{\Theta}| = \mathfrak{C}(\epsilon;\Theta,D) < \infty\}$ in (76), we used Cauchy-Schwarz inequity and the fact that $\|\varphi\|_{L^\infty} \leq b$ in (77), and we used Jensen's inequality in (78).
Since $\epsilon > 0$ is arbitrary, combining (72), (75), (78), we have that

$$\mathbb{E}[B_N] \leq 4\left(\mathbb{E}\,\|\eta_1\|_{\mathcal{Y}}^2\right)^{1/2}\inf_{\epsilon>0}\left\{\epsilon + b\sqrt{\frac{2\log\mathfrak{C}(\epsilon;\Theta,D)}{N}}\right\}. \tag{79}$$

Remark 4.3. *Equation* (79) *still holds true for* Θ *being a compact subspace of an infinite dimensional space.*

(2) Chaining: The idea is to apply Dudley integral covering number bound on our setting. Consider the class of real-valued functions $\mathfrak{F} := \{f_\theta : \theta \in \Theta\} \cup \{0\}$, where $f_\theta(\eta,x) = f_\theta(\eta)(x) = \langle \eta, \varphi(x;\theta)\rangle_{\mathcal{Y}}$. For $f_\theta, f_{\theta'} \in \mathfrak{F}$, define $\widetilde{D}$ as

$$\widetilde{D}(f_\theta, f_{\theta'}) = \left(\frac{1}{N}\sum_{n=1}^{N}\left|\left(f_\theta - f_{\theta'}\right)(\eta_n,x_n)\right|^2\right)^{1/2}.$$

It is easy to check that $\widetilde{D}(\cdot,\cdot)$ is a pseudometric (see definition 2.1) on $\mathfrak{F}$. Here, we aim to do a m-steps discritization on the function class $\mathfrak{F}$ with respect to the pseudometric $\widetilde{D}$ for some nicely chosen integer m. However, although Θ is compact and $\varphi(\cdot;\cdot)$ is bounded, the functions f_θ may not be bounded due to the inner product with the noise η. To establish a m-steps discritization, we need some property of boundedness of $\mathfrak{F}$ with respect to $\widetilde{D}$. By assumption we have that $\|\varphi\|_{L^\infty} \le b$ for some $b > 0$. Using this and Cauchy-Schwarz inequity,

$$\sup_{\theta\in\Theta}\left(\frac{1}{N}\sum_{n=1}^{N}|f_\theta(\eta_n,x_n)|^2\right)^{1/2} \le b\left(\frac{1}{N}\sum_{n=1}^{N}\|\eta_n\|_{\mathcal{Y}}^2\right)^{1/2}. \tag{80}$$

This leads to an additional assumption on the noise η in order to apply the technique of chaining in our setting.

Assumption 4.6 (Empirical L^2 boundedness of noise)**.** *Suppose that the right-hand side of* (80) *is bounded above by c, that is*

$$\left(\frac{1}{N}\sum_{n=1}^{N}\|\eta_n\|_{\mathcal{Y}}^2\right)^{1/2} \le c/b < \infty. \tag{81}$$

Assumption 4.6 says that the noise is empirically bounded in L^2 sense. This property ensures that every $f_\theta \in \mathfrak{F}$ is inside some ball with respect to $\widetilde{D}$, that is, $\widetilde{D}(f_\theta, 0) \le c$ for all $f_\theta \in \mathfrak{F}$.

For each $j \in \mathbb{N}_+$, let $\epsilon_j := c2^{-j}$ and $\widehat{\mathfrak{F}}_j$ be a minimal ϵ_j-cover of $\mathfrak{F}$ w.r.t. $\widetilde{D}$. For each $f_\theta \in \mathfrak{F}$ and $j \in \mathbb{N}$, let $\hat{f}_{\theta_j} \in \widehat{\mathfrak{F}}_j$ be such that $\widetilde{D}(f_\theta, \hat{f}_{\theta_j}) \le \epsilon_j$. Under assumption 4.6, the covering number $\mathfrak{C}(\epsilon;\mathfrak{F},\widetilde{D})$ is finite for all $\epsilon > 0$. Now, fix $f_\theta \in \mathfrak{F}$. For a given $m \in \mathbb{N}_+$ to be chosen later, we define the telescoping sum:

$$f_\theta = f_\theta - \hat{f}_{\theta_m} + \sum_{j=1}^{m}\left(\hat{f}_{\theta_j} - \hat{f}_{\theta_{j-1}}\right),$$

where $\hat{f}_{\theta_0} = 0$. Following (72),

$$\begin{aligned}
\mathbb{E}\left[B_N/4\right] &\le \mathbb{E}_{\boldsymbol{\sigma},\boldsymbol{\eta},\boldsymbol{x}}\sup_{\theta\in\Theta}\left|\frac{1}{N}\sum_{n=1}^{N}\sigma_n f_\theta(\eta_n,x_n)\right| \\
&= \mathbb{E}_{\boldsymbol{\sigma},\boldsymbol{\eta},\boldsymbol{x}}\sup_{\theta\in\Theta}\left|\frac{1}{N}\sum_{n=1}^{N}\sigma_n\left(f_\theta - \hat{f}_{\theta_m} + \sum_{j=1}^{m}\left(\hat{f}_{\theta_j} - \hat{f}_{\theta_{j-1}}\right)\right)(\eta_n,x_n)\right| \\
&\le \mathbb{E}_{\boldsymbol{\sigma},\boldsymbol{\eta},\boldsymbol{x}}\sup_{\theta\in\Theta}\left|\frac{1}{N}\sum_{n=1}^{N}\sigma_n\left(f_\theta - \hat{f}_{\theta_m}\right)(\eta_n,x_n)\right| \qquad \text{(I)} \\
&\quad + \sum_{j=1}^{m}\mathbb{E}_{\boldsymbol{\sigma},\boldsymbol{\eta},\boldsymbol{x}}\sup_{\theta\in\Theta}\left|\frac{1}{N}\sum_{n=1}^{N}\sigma_n\left(\hat{f}_{\theta_j} - \hat{f}_{\theta_{j-1}}\right)(\eta_n,x_n)\right|. \qquad \text{(II)}
\end{aligned}$$

(I) is bounded above by

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\sigma},\boldsymbol{\eta},\boldsymbol{x}}\sup_{\theta\in\Theta}\frac{1}{N}\sum_{n=1}^{N}|\sigma_n|\left|\left(f_\theta - \hat{f}_{\theta_m}\right)(\eta_n,x_n)\right| \\
&\le \sup_{\theta\in\Theta}\widetilde{D}(f_\theta,\hat{f}_{\theta_m}) \le \epsilon_m,
\end{aligned} \tag{82}$$

where we used Cauchy-Schwarz inequality on $\mathbb{R}^N$.
For (II), since the covering number $\mathfrak{C}(\epsilon;\mathfrak{F},\widetilde{D})$ is finite for all $\epsilon > 0$, we can apply Massart's lemma (lemma 2.11) on $\mathscr{F} = \{\hat{f}_{\theta_j} - \hat{f}_{\theta_{j-1}} : \hat{f}_{\theta_j} \in \widehat{\mathfrak{F}}_j, \hat{f}_{\theta_{j-1}} \in \widehat{\mathfrak{F}}_{j-1}\}$, where $|\mathscr{F}| \le |\widehat{\mathfrak{F}}_j\|\widehat{\mathfrak{F}}_{j-1}| < \infty$. Then, (II) is bounded above by

$$\sum_{j=1}^{m}\max_{\substack{\hat{\theta}_j\in\widehat{\Theta}_j \\ \hat{\theta}_{j-1}\in\widehat{\Theta}_{j-1}}}\mathbb{E}_{\boldsymbol{\eta},\boldsymbol{x}}\left(\sum_{n=1}^{N}\left(\hat{f}_{\theta_j} - \hat{f}_{\theta_{j-1}}\right)^2(\eta_n,x_n)\right)^{1/2}\frac{\sqrt{2\log\left(\mathfrak{C}(\epsilon_j;\mathfrak{F},\widetilde{D})\mathfrak{C}(\epsilon_{j-1};\mathfrak{F},\widetilde{D})\right)}}{N}. \tag{83}$$

Note that

$$\left(\sum_{n=1}^{N}\left(\hat{f}_{\theta_j}-\hat{f}_{\theta_{j-1}}\right)^2(\eta_n,x_n)\right)^{1/2} \le \left(\sum_{n=1}^{N}\left(\hat{f}_{\theta_j}-f_\theta\right)^2(\eta_n,x_n)\right)^{1/2}+\left(\sum_{n=1}^{N}\left(f_\theta-\hat{f}_{\theta_{j-1}}\right)^2(\eta_n,x_n)\right)^{1/2} \tag{84}$$

$$\le \sqrt{N}\left(\widetilde{D}(\hat{f}_{\theta_j},f_\theta)+\widetilde{D}(f_\theta,\hat{f}_{\theta_{j-1}})\right)$$

$$\le \sqrt{N}\left(\epsilon_j+\epsilon_{j-1}\right)$$

$$= 6\sqrt{N}\left(\epsilon_j-\epsilon_{j+1}\right), \tag{85}$$

where we used Minkowski inequality in (84), and (85) is due to our construction that $\epsilon_{j-1}=2\epsilon_j$ for all $j\in\mathbb{N}_+$. Thus, (83) is bounded above by

$$\sum_{j=1}^{m} 6\left(\epsilon_j-\epsilon_{j+1}\right)\sqrt{\frac{2\log\left(\mathfrak{C}(\epsilon_j;\mathfrak{F},\widetilde{D})\mathfrak{C}(\epsilon_{j-1};\mathfrak{F},\widetilde{D})\right)}{N}}. \tag{86}$$

By the definition of ϵ-covering number (definition 2.2), $\mathfrak{C}(\epsilon;\mathfrak{F},\widetilde{D})$ is non-increasing w.r.t. ϵ. Since $\epsilon_{j-1}=2\epsilon_j$ for all $j\in\mathbb{N}_+$, we have that $\mathfrak{C}(\epsilon_j;\mathfrak{F},\widetilde{D})\ge\mathfrak{C}(\epsilon_{j-1};\mathfrak{F},\widetilde{D})$ for all $j\in\mathbb{N}_+$.
Thus, combining (82), and (86), we have that

$$\mathbb{E}[B_N/4] \le \epsilon_m+\frac{12}{\sqrt{N}}\sum_{j=1}^{m}\left(\epsilon_j-\epsilon_{j+1}\right)\sqrt{\log\mathfrak{C}(\epsilon_j;\mathfrak{F},\widetilde{D})}$$

$$= \epsilon_m+\frac{12}{\sqrt{N}}\sum_{j=1}^{m}\int_{\epsilon_{j+1}}^{\epsilon_j}\sqrt{\log\mathfrak{C}(\epsilon_j;\mathfrak{F},\widetilde{D})}d\tau$$

$$\le \epsilon_m+\frac{12}{\sqrt{N}}\sum_{j=1}^{m}\int_{\epsilon_{j+1}}^{\epsilon_j}\sqrt{\log\mathfrak{C}(\tau;\mathfrak{F},\widetilde{D})}d\tau, \quad \text{since } \tau\in[\epsilon_{j+1},\epsilon_j]$$

$$\le \epsilon_m+\frac{12}{\sqrt{N}}\int_{\epsilon_{m+1}}^{c/2}\sqrt{\log\mathfrak{C}(\tau;\mathfrak{F},\widetilde{D})}d\tau. \tag{87}$$

For any $\epsilon\in[0,c/4]$, we can choose m such that $\epsilon\le\epsilon_{m+1}\le 2\epsilon$ or equivalently $m=\sup\{j\in\mathbb{N}_+:\epsilon_j\ge 2\epsilon\}$. Thus,

$$\mathbb{E}[B_N]\le\inf_{\epsilon\in[0,c/4]}\left\{16\epsilon+\frac{48}{\sqrt{N}}\int_{\epsilon}^{c/2}\sqrt{\log\mathfrak{C}(\tau;\mathfrak{F},\widetilde{D})}d\tau\right\}. \tag{88}$$

According to (79) and (88), we have that $\mathbb{E}[B_N]\to 0$ as $N\to\infty$. With suitable assumptions on the growth rates of the covering numbers, we can make this argument more precise.

Remark 4.4. *If Θ has finite topological dimension, then Θ can be identified with a closed and bounded subspace of a finite dimensional space, let say*

$$\Theta=\{\theta\in\mathbb{R}^d:\|\theta\|_2\le a\}.$$

In addition, suppose that φ is K-Lipschitz on θ. Then,

$$D(\theta,\theta')^2=\frac{1}{N}\sum_{n=1}^{N}\|\varphi(x_n;\theta)-\varphi(x_n;\theta')\|_{\mathcal{Y}}^2$$

$$\le K^2\|\theta-\theta'\|_2^2.$$

By a similar argument as in example 2.1, we have that

$$\mathfrak{C}(\epsilon;\Theta,D)\le\left(\frac{2aK\sqrt{d}}{\epsilon}\right)^d. \tag{89}$$

Now, assume that $\log(\frac{1}{\epsilon}) \lesssim \log(N)$. *Plugging* (89) *into* (79) *gives*

$$\begin{aligned}\mathbb{E}[B_N] &\lesssim \epsilon + \sqrt{\frac{d\log(\frac{2aK\sqrt{d}}{\epsilon})}{N}} \\ &\lesssim \sqrt{\frac{d\log N}{N}}.\end{aligned} \tag{90}$$

Furthermore, suppose that assumption 4.6 holds. Then,

$$\begin{aligned}\widetilde{D}(f_\theta, f_{\theta'})^2 &= \frac{1}{N}\sum_{n=1}^{N} \left|\langle \eta_n, \varphi(x_n;\theta)\rangle_{\mathcal{Y}} - \langle \eta_n, \varphi(x_n;\theta')\rangle_{\mathcal{Y}}\right|^2 \\ &\le \frac{1}{N}\sum_{n=1}^{N} \|\eta_n\|_{\mathcal{Y}}^2 \left\|\varphi(x_n;\theta) - \varphi(x_n;\theta')\right\|_{\mathcal{Y}}^2 \\ &\le \frac{c^2}{b^2}K^2 \left\|\theta - \theta'\right\|_2^2.\end{aligned}$$

Let $\widehat{\mathfrak{F}} = \{f_{\hat\theta} : \hat\theta \in \widehat{\Theta}\}$, *where* $f_{\hat\theta} : (\eta, x) \mapsto \langle \eta, \varphi(x;\hat\theta)\rangle_{\mathcal{Y}}$ *and*

$$\widehat{\Theta} = \left\{\sum_{i=1}^{d} \alpha_i' \mathbf{v}_i : \forall i, \alpha_i' \in \{-a, -a+\delta, -a+2\delta, \ldots, a\}\right\}.$$

Given $f_\theta \in \mathfrak{F}$ *such that* $\theta \in \Theta$ *with* $\theta = \sum_{i=1}^d \alpha_i \mathbf{v}_i$ *satisfying* $\|\boldsymbol{\alpha}\|_\infty \le a$. *Then, there exists* $f_{\theta'} \in \widehat{\mathfrak{F}}$ *with* $\theta' = \sum_{i=1}^d \alpha_i' \mathbf{v}_i$ *such that*

$$\widetilde{D}(f_\theta, f_{\theta'})^2 \le \frac{c^2}{b^2}K^2\left\|\theta - \theta'\right\|^2 = \frac{c^2}{b^2}K^2\left\|\sum_{i=1}^{d}(\alpha_i' - \alpha_i)\mathbf{v}_i\right\|^2 \le \frac{c^2}{b^2}K^2\delta^2\sum_{i=1}^{d}\|\mathbf{v}_i\|^2 \le \frac{c^2}{b^2}K^2\delta^2 d.$$

Choose $\delta = \epsilon/(\frac{c}{b}K\sqrt{d})$, *then* $\widetilde{D}(f_\theta, f_{\theta'}) \le \epsilon$. *Therefore,* $\widehat{\mathfrak{F}}$ *is an* ϵ*-cover of* $\mathfrak{F}$. *Hence,*

$$\mathfrak{C}(\epsilon; \mathfrak{F}, \widetilde{D}) \le \left|\widehat{\mathfrak{F}}\right| = \left(\frac{2a}{\delta}\right)^d = \left(\frac{2a\frac{c}{b}K\sqrt{d}}{\epsilon}\right)^d. \tag{91}$$

Plugging (91) *into* (88), *we realize that the function*

$$\sqrt{\log \mathfrak{C}(\epsilon; \mathfrak{F}, \widetilde{D})} \le \sqrt{d\left(\log(2a\frac{c}{b}\sqrt{d}) + \log\frac{1}{\epsilon}\right)} \lesssim \sqrt{d\log d} + \sqrt{d\log\frac{1}{\epsilon}}$$

is (improperly) integrable on $(0, c/2)$. *Thus, setting* $\epsilon = 0$ *(and disregarding the term depending on* c*), we obtain the bound*

$$\mathbb{E}[B_N] \lesssim \sqrt{\frac{d\log d}{N}}. \tag{92}$$

This shows that chaining does help us to obtain a better bound on $\mathbb{E}[B_N]$. *However, this relies on Assumption 4.6. If we abandon this assumption, then we have to control the noise term* η *by its maximum* $\max_{n\in\{1,\cdots,N\}}\|\eta_n\|_{\mathcal{Y}}$. *The expectation of the maximum of* N *sub-exponential random variables grows as* $\mathcal{O}(\log N)$. *The* $\log(N)$ *factor makes the rate worse than that the rate obtained using simple one-step discretization.*

Remark 4.5. *We can assume a more general scenario where the growth rates of the covering numbers* $\mathfrak{C}(\epsilon; \Theta, D)$ *and* $\mathfrak{C}(\epsilon; \mathfrak{F}, \widetilde{D})$ *only satisfy*

$$\log \mathfrak{C}(\epsilon; \Theta, D) \lesssim \epsilon^{-1/s}, \quad \log \mathfrak{C}(\epsilon; \mathfrak{F}, \widetilde{D}) \lesssim \epsilon^{-1/s} \tag{93}$$

for some $s > 0$ *(see Ratti [25, Section 7.1, p. 20] for an example of this scenario). Note that in this scenario, the growth rates of the covering numbers are dimensionally independent.*

Plugging (93) *into* (79) *gives*

$$\mathbb{E}[B_N] \lesssim \epsilon + \sqrt{\frac{\epsilon^{-1/s}}{N}}. \tag{94}$$

Choosing $\epsilon \sim N^{-\frac{s}{1+2s}}$ *in* (94) *yields*

$$\mathbb{E}[B_N] \lesssim N^{-\frac{s}{1+2s}}. \tag{95}$$

Furthermore, suppose that assumption 4.6 holds. Plugging (93) *into* (88) *gives*

$$\mathbb{E}[B_N] \lesssim \epsilon + \frac{1}{\sqrt{N}} \int_{\epsilon}^{c/2} \tau^{-\frac{1}{2s}} d\tau.$$

If $s > \frac{1}{2}$*, then* $\tau^{-\frac{1}{2s}}$ *is integrable on* $(0, c/2)$*. So, setting* $\epsilon = 0$*, we have that*

$$\mathbb{E}[B_N] \lesssim N^{-1/2}. \tag{96}$$

If $0 < s \leq \frac{1}{2}$*, we have that*

$$\mathbb{E}[B_N] \lesssim \epsilon + \frac{1}{\sqrt{N}} \epsilon^{1-\frac{1}{2s}}. \tag{97}$$

Choosing $\epsilon \sim N^{-s}$ *in* (97) *yields*

$$\mathbb{E}[B_N] \lesssim N^{-s}. \tag{98}$$

In this more general scenario, chaining still gives a better rate under Assumption 4.6. Without this assumption, the rate is worse than the rate given by the simple one-step discretization method due to the multiplication of the $\log(N)$ *factor induced by the expectation of the maximum of* N *sub-exponential random variables.*

Next, we show that B_N concentrates around its expectation with high probability using Bernstein-type inequalities. The key idea is to apply Corollary 2.7 (Lanthaler and Nelsen [14, Cor. A.5, p. 15, Eqn. A.11]) with i.i.d. random functionals $Z_n \colon \Theta \to \mathbb{R}$ given by $\theta \mapsto Z_n(\theta) := \langle \eta_n, \varphi(x_n; \theta) \rangle_{\mathcal{Y}}$. By assumption 4.2, the set Θ is a compact metrizable space. Then, by Stone–Weierstrass theorem, $C(\Theta)$ is a separable Banach space when equipped with the supremum norm. Moreover, for every $x \in \mathcal{X}$, the map $\theta \mapsto \varphi(x; \theta)$ is continuous on Θ. Furthermore, we have that $\|\varphi\|_{L^\infty(\mathcal{X} \times \Theta; \mathcal{Y})} \leq b$. To apply the corollary, it remains to control the subexponential norm of Z_1 and show that $Z_1 \in C(\Theta)$. We estimate

$$\sup_{\theta \in \Theta} |Z_1(\theta)| \leq \|\eta_1\|_{\mathcal{Y}} \sup_{\theta \in \Theta} \|\varphi(x_1; \theta)\|_{\mathcal{Y}} \leq b \|\eta_1\|_{\mathcal{Y}}.$$

Thus, Z_1 is bounded and

$$\|Z_1\|_{\psi_1(C(\Theta))} = \left\| \sup_{\theta \in \Theta} |Z_1(\theta)| \right\|_{\psi_1(\mathbb{R})} \leq b \|\eta_1\|_{\psi_1(\mathcal{Y})}.$$

Next, $Z_1 \in C(\Theta)$ is continuous a.s. because it is the composition of a.s. continuous maps. Thus, we apply Corollary 2.7 to obtain that, for any $\delta \in (0, 1)$ and $N \geq \log(1/\delta)$, it holds with probability at least $1 - \delta$ that

$$\sup_{\theta \in \Theta} \left| \frac{1}{N} \sum_{n=1}^{N} \langle \eta_n, \varphi(x_n; \theta) \rangle_{\mathcal{Y}} \right| \leq \mathbb{E} \sup_{\theta \in \Theta} \left| \frac{1}{N} \sum_{n=1}^{N} \langle \eta_n, \varphi(x_n; \theta)_{\mathcal{Y}} \rangle \right| + \sqrt{\frac{64 b^2 e^3 \|\eta_1\|^2_{\psi_1(\mathcal{Y})} \log(1/\delta)}{N}}.$$

Now, let

$$S_N := \sup_{\theta \in \Theta} \left| \frac{1}{N} \sum_{n=1}^{N} \langle \eta_n, \varphi(x_n; \theta) \rangle_{\mathcal{Y}} \right|, \tag{99}$$

$$r := \sqrt{\frac{256 b^2 e^3 \|\eta_1\|^2_{\psi_1(\mathcal{Y})} \log(2/\delta)}{N}}. \tag{100}$$

Recalling (107), we have that for any $\delta \in (0, 1)$ and $N \geq \log(2/\delta)$,

$$\mathbb{P}\{B_N \geq 2\mathbb{E}[S_N] + r\} \leq \mathbb{P}\{S_N \geq \mathbb{E}[S_N] + r/2\} \leq \delta/2. \tag{101}$$

With (79) in mind, we have

$$\mathbb{P}\left\{ B_N \geq 8 \left(\mathbb{E} \|\eta_1\|^2_{\mathcal{Y}} \right)^{1/2} \inf_{\epsilon > 0} \left\{ \epsilon + b \sqrt{\frac{2 \log \mathfrak{C}(\epsilon; \Theta, D)}{N}} \right\} + r \right\} \leq \mathbb{P}\{B_N \geq 2\mathbb{E}[S_N] + r\} \leq \delta/2. \tag{102}$$

Following remark 4.5, plugging (95) into (102), it holds

$$\mathbb{P}\left\{ B_N \geq C' N^{-\frac{1+3s}{2+4s}} + r \right\} \leq \delta/2,$$

where $s > 0$ (see (93)) and C' is an absolute constant. Thus, given an arbitrary $\delta > 0$, with S_N defined in (99) and r defined in (100), we can choose

$$\lambda \geq 2C' N^{-\frac{1+3s}{2+4s}} + 2r \tag{103}$$

such that $\mathbb{P}(\mathrm{E}) \geq 1 - \frac{\delta}{2}$.

Furthermore, suppose that assumption 4.6 holds. Then, by (88), we have

$$\mathbb{P}\left\{ B_N \geq \inf_{\epsilon \in [0, c/4]} \left\{ 32\epsilon + \frac{96}{\sqrt{N}} \int_{\epsilon}^{c/2} \sqrt{\log \mathfrak{C}(\tau; \mathfrak{F}, \widetilde{D})} d\tau \right\} + r \right\} \leq \mathbb{P}\{ B_N \geq 2\mathbb{E}[S_N] + r \} \leq \delta/2. \tag{104}$$

Following remark 4.5, if $s > \frac{1}{2}$, then we can plug (96) into (104), it holds

$$\mathbb{P}\left\{ B_N \geq C'' N^{-\frac{3}{4}} + r \right\} \leq \delta/2,$$

where C'' is an absolute constant. Thus, given an arbitrary $\delta > 0$, with S_N defined in (99) and r defined in (100), we can choose

$$\lambda \geq 2C'' N^{-\frac{3}{4}} + 2r \tag{105}$$

such that $\mathbb{P}(\mathrm{E}) \geq 1 - \frac{\delta}{2}$.

If $0 < s \leq \frac{1}{2}$, then we can plug (98) into (104), it holds

$$\mathbb{P}\left\{ B_N \geq C''' N^{-\frac{s+1}{2}} + r \right\} \leq \delta/2,$$

where C''' is an absolute constant. Thus, given an arbitrary $\delta > 0$, with S_N defined in (99) and r defined in (100), we can choose

$$\lambda \geq 2C''' N^{-\frac{s+1}{2}} + 2r \tag{106}$$

such that $\mathbb{P}(\mathrm{E}) \geq 1 - \frac{\delta}{2}$.

Refined approach In fact, since the summands of our linear empirical process $\langle \eta_n, \varphi(x_n; \theta) \rangle_{\mathcal{Y}}$ have mean zero, there is no need to use Gine-Zinn symmetrization. The property that the $\{\eta_n\}$ are conditionally centered enables them to play the role as the Rademacher random variables would. Instead, we can directly use the Dudley metric entropy integral inequalities for q-Orlicz random processes (e.g. theorem 2.9) to deliver a tail bound on B_N. This approach helps us get rid of the additional assumption 4.6 on the noise η.

Recalling (51), we have

$$\begin{aligned} B_N/2 &= \sup_{\|\mu\|_{\mathrm{TV}} \leq 1} \left| \int_{\Theta} \frac{1}{N} \sum_{n=1}^{N} \langle \eta_n, \varphi(x_n; \theta) \rangle_{\mathcal{Y}} \, d\mu(\theta) \right| \\ &\leq \left\| \frac{1}{N} \sum_{n=1}^{N} \langle \eta_n, \varphi(x_n; \ \cdot) \rangle_{\mathcal{Y}} \right\|_{L^\infty(\Theta; \mathbb{R})}. \end{aligned}$$

By Assumption 4.2, $\theta \mapsto \varphi(\cdot\,; \theta)$ is continuous, then the L^∞ norm is equal to the $C(\Theta)$ supremum norm. That is,

$$B_N/2 \leq \sup_{\theta \in \Theta} \left| \frac{1}{N} \sum_{n=1}^{N} \langle \eta_n, \varphi(x_n; \theta) \rangle_{\mathcal{Y}} \right|. \tag{107}$$

Condition on $\{x_n\}$, we can view the sum on the right-hand side of (107) as only a function of the $\{\eta_n\}$ with the $\{x_n\}$ as fixed. We claim that condition on $\{x_n\}$, the sum is a ψ_1-process (see definition 2.13) with respect to the empirical pseudo-metric

$$\mathrm{d}_N(\theta, \theta') := \left(\frac{1}{N} \sum_{n=1}^{N} \| \varphi(x_n; \theta) - \varphi(x_n; \theta') \|_{\mathcal{Y}}^2 \right)^{1/2}.$$

To see this, the random variables hold that

$$\left\| \frac{1}{N} \langle \eta_n, \varphi(x_n;\theta) - \varphi(x_n;\theta') \rangle_{\mathcal{Y}} \right\|_{\mathcal{Y}} \leq \frac{1}{N} \|\eta_n\|_{\mathcal{Y}} \|\varphi(x_n;\theta) - \varphi(x_n;\theta')\|_{\mathcal{Y}} ,$$

and thus condition on $\{x_n\}$,

$$\left\| \frac{1}{N} \langle \eta_n, \varphi(x_n;\theta) - \varphi(x_n;\theta') \rangle_{\mathcal{Y}} \right\|_{\psi_1(\mathcal{Y})} \leq \frac{1}{N} \|\eta_1\|_{\psi_1(\mathcal{Y})} \|\varphi(x_n;\theta) - \varphi(x_n;\theta')\|_{\mathcal{Y}} .$$

Fix arbitrary $\theta, \theta' \in \Theta$. Let

$$\begin{aligned} W_{x_n}(\eta_n) &:= \frac{1}{N} \langle \eta_n, \varphi(x_n;\theta) - \varphi(x_n;\theta') \rangle_{\mathcal{Y}} , \\ K_{x_n} &:= \frac{1}{N} \|\eta_1\|_{\psi_1(\mathcal{Y})} \|\varphi(x_n;\theta) - \varphi(x_n;\theta')\|_{\mathcal{Y}} , \\ K &:= \max_{n=1,\cdots,N} K_{x_n} . \end{aligned}$$

Later, we write $W_{x_n}(\eta_n)$ as W_{x_n}. Using (10), it holds that

$$\begin{aligned} \mathbb{E}\left[\exp\left(\lambda \sum_{n=1}^N W_{x_n} \right) \right] &= \prod_{n=1}^N \mathbb{E}\left[\exp\left(\lambda W_{x_n} \right) \right] \\ &\leq \prod_{n=1}^N \exp\left(C^2 K_{x_n}^2 \lambda^2 \right) \quad \forall \lambda \text{ s.t. } |\lambda| \leq \frac{1}{CK} \\ &\leq \exp\left(C_1^2 \frac{1}{N} \|\eta_1\|_{\psi_1(\mathcal{Y})}^2 \, \mathrm{d}_N(\theta,\theta')^2 \lambda^2 \right) \quad \forall \lambda \text{ s.t. } |\lambda| \leq \frac{1}{C_1 \|\eta_1\|_{\psi_1(\mathcal{Y})} \, \mathrm{d}_N(\theta,\theta')/\sqrt{N}} , \end{aligned}$$

where C, C_1 are absolute constants such that $C_1 \geq C$ and

$$C_1 \geq C \frac{\max_{n=1,\cdots,N} \|\varphi(x_n;\theta) - \varphi(x_n;\theta')\|_{\mathcal{Y}}}{\left(\sum_{n=1}^N \|\varphi(x_n;\theta) - \varphi(x_n;\theta')\|_{\mathcal{Y}}^2 \right)^{1/2}}, \tag{108}$$

where (108) ensures that

$$\frac{1}{C_1 \|\eta_1\|_{\psi_1(\mathcal{Y})} \, \mathrm{d}_N(\theta,\theta')/\sqrt{N}} \leq \frac{1}{CK}.$$

We thus deduce that $\|W_{x_1} + \ldots + W_{x_N}\|_{\psi_1} \lesssim \|\eta_1\|_{\psi_1(\mathcal{Y})} \, \mathrm{d}_N(\theta,\theta')/\sqrt{N}$. Thus, by the homogeneity of the norm, condition on $\{x_n\}$, the increments satisfy

$$\left\| \frac{1}{N} \sum_{n=1}^N \langle \eta_n, \varphi(x_n;\theta) \rangle_{\mathcal{Y}} - \frac{1}{N} \sum_{n=1}^N \langle \eta_n, \varphi(x_n;\theta') \rangle_{\mathcal{Y}} \right\|_{\psi_1(\mathbb{R})} \leq C_2 \frac{\|\eta_1\|_{\psi_1(\mathcal{Y})} \, \mathrm{d}_N(\theta,\theta')}{\sqrt{N}}.$$

More explicitly, condition on $\{x_n\}$,

$$\left\| \frac{\sqrt{N}}{C_2 \|\eta_1\|_{\psi_1(\mathcal{Y})}} \frac{1}{N} \sum_{n=1}^N \langle \eta_n, \varphi(x_n;\theta) \rangle_{\mathcal{Y}} - \frac{\sqrt{N}}{C_2 \|\eta_1\|_{\psi_1(\mathcal{Y})}} \frac{1}{N} \sum_{n=1}^N \langle \eta_n, \varphi(x_n;\theta') \rangle_{\mathcal{Y}} \right\|_{\psi_1(\mathbb{R})} \leq \mathrm{d}_N(\theta,\theta').$$

For any measurable set A and random variable Y, we denote $\mathbb{E}_A[Y] = \int_A Y d\mathbb{P}$. Note that we have $\mathbb{E}_A[Y] = \mathbb{E}[Y \mid Y \in A]\mathbb{P}[A]$ by construction. Since the random variable $\frac{1}{N} \sum_{n=1}^N \langle \eta_n, \varphi(x_n;\theta) \rangle_{\mathcal{Y}}$ is centered, we have

$$\begin{aligned} \mathbb{E}_A\left[\sup_{\theta \in \Theta} \left| \frac{1}{N} \sum_{n=1}^N \langle \eta_n, \varphi(x_n;\theta) \rangle_{\mathcal{Y}} \right| \right] &= \mathbb{E}_A\left[\sup_{\theta \in \Theta} \left| \frac{1}{N} \sum_{n=1}^N \langle \eta_n, \varphi(x_n;\theta) \rangle_{\mathcal{Y}} - \frac{1}{N} \sum_{n=1}^N \langle \eta_n, \varphi(x_n;\theta_0) \rangle_{\mathcal{Y}} \right| \right] \\ &\leq \mathbb{E}_A\left[\sup_{\theta,\theta' \in \Theta} \left| \frac{1}{N} \sum_{n=1}^N \langle \eta_n, \varphi(x_n;\theta) \rangle_{\mathcal{Y}} - \frac{1}{N} \sum_{n=1}^N \langle \eta_n, \varphi(x_n;\theta') \rangle_{\mathcal{Y}} \right| \right] . \end{aligned}$$

Following the proof of theorem 2.9 and theorem 2.15, there is a universal constant C_3 such that for all $q > 0$,

$$\begin{aligned} &\mathbb{P}\left[\sup_{\theta \in \Theta} \left| \frac{1}{N} \sum_{n=1}^N \langle \eta_n, \varphi(x_n;\theta) \rangle_{\mathcal{Y}} \right| \geq C_3 \frac{\|\eta_1\|_{\psi_1(\mathcal{Y})}}{\sqrt{N}} \left(\int_\alpha^{D(x_1,\cdots,x_N)} \log\left(1 + \mathfrak{C}(\tau;\Theta,\mathrm{d}_N)\right) d\tau + q \right) \,\middle|\, x_1, \cdots, x_N \right] \\ &\leq 2e^{-(q - E/C_3)/D(x_1,\cdots,x_N)}, \end{aligned} \tag{109}$$

where $D(x_1, \cdots, x_N) = \sup_{\theta,\theta' \in \Theta} \mathrm{d}_N(\theta, \theta')$ is the diameter of the set Θ under d_N for given $\{x_n\}$, $D > \alpha \geq 0$ is a fixed cutoff, and

$$\begin{aligned} E &= 2\mathbb{E} \sup_{\substack{\theta,\theta' \in \Theta \\ \mathrm{d}_N(\theta,\theta') \leq 4\alpha}} \left| \frac{\sqrt{N}}{C_2 \|\eta_1\|_{\psi_1(\mathcal{Y})}} \frac{1}{N} \sum_{n=1}^{N} \langle \eta_n, \varphi(x_n;\theta) - \varphi(x_n;\theta') \rangle_{\mathcal{Y}} \right| \\ &\leq \frac{2\sqrt{N}}{C_2 \|\eta_1\|_{\psi_1(\mathcal{Y})}} \mathbb{E} \sup_{\substack{\theta,\theta' \in \Theta \\ \mathrm{d}_N(\theta,\theta') \leq 4\alpha}} \left| \left(\frac{1}{N} \sum_{n=1}^{N} \|\eta_n\|_{\mathcal{Y}}^2 \right)^{1/2} \left(\frac{1}{N} \sum_{n=1}^{N} \|\varphi(x_n;\theta) - \varphi(x_n;\theta')\|_{\mathcal{Y}}^2 \right)^{1/2} \right| \\ &\leq \frac{8\alpha \sqrt{N \mathbb{E} \|\eta_1\|_{\mathcal{Y}}^2}}{C_2 \|\eta_1\|_{\psi_1(\mathcal{Y})}}, \end{aligned}$$

where we used Cauchy–Schwarz inequality for the first inequality, and Jensen's inequality for the second inequality. Since η_1 is sub-exponential, it has finite moments, so $\mathbb{E} \|\eta_1\|_{\mathcal{Y}}^2 < \infty$. Note that

$$D(x_1, \cdots, x_N) = \sup_{\theta,\theta' \in \Theta} \mathrm{d}_N(\theta, \theta') = \sup_{\theta,\theta' \in \Theta} \left(\frac{1}{N} \sum_{n=1}^{N} \|\varphi(x_n;\theta) - \varphi(x_n;\theta')\|_{\mathcal{Y}}^2 \right)^{1/2} \leq 2b. \tag{110}$$

Following Remark 4.5, we may suppose that the growth rate of the covering number $\mathfrak{C}(\tau; \Theta, \mathrm{d}_N)$ satisfies

$$\log \mathfrak{C}(\tau; \Theta, \mathrm{d}_N) \lesssim \tau^{-1/s_1}$$

for some $s_1 > 0$ Since $\mathfrak{C}(\tau; \Theta, \mathrm{d}_N)$ is a positive integer, we have that $\log(1 + \mathfrak{C}(\tau; \Theta, \mathrm{d}_N))$ is positive, and $\log(1 + \mathfrak{C}(\tau; \Theta, \mathrm{d}_N)) \leq \log(\mathfrak{C}(\tau; \Theta, \mathrm{d}_N)) + 1 \leq C_4 \tau^{-1/s_1} + 1$. Thus, using this fact and (110), we have

$$\int_{\alpha}^{D(x_1,\cdots,x_N)} \log(1 + \mathfrak{C}(\tau; \Theta, \mathrm{d}_N))\, d\tau \leq C_4 \int_{\alpha}^{2b} \tau^{-1/s_1} d\tau + 2b\,. \tag{111}$$

Now, let

$$\begin{aligned} S_N &:= \sup_{\theta \in \Theta} \left| \frac{1}{N} \sum_{n=1}^{N} \langle \eta_n, \varphi(x_n;\theta) \rangle_{\mathcal{Y}} \right|, \\ I_{\alpha,s_1,b} &:= C_4 \int_{\alpha}^{2b} \tau^{-1/s_1} d\tau + 2b\,, \\ E_{\alpha,\eta,N} &:= \frac{8\alpha \sqrt{N \mathbb{E} \|\eta_1\|_{\mathcal{Y}}^2}}{C_2 C_3 \|\eta_1\|_{\psi_1(\mathcal{Y})}}. \end{aligned}$$

We may avoid the singular point of the integral $I_{\alpha,s_1,b}$ at zero by choosing $0 < \alpha < 2b$. Then, the integral is finite for all $s_1 > 0$. To get rid of the $\sqrt{N}$ factor in $E_{\alpha,\eta,N}$, we choose $\alpha = N^{-1/2}$. So, we get

$$I_{N,s_1,b} = C_4 \int_{N^{-1/2}}^{2b} \tau^{-1/s_1} d\tau + 2b\,, \qquad E_{\eta} = \frac{8\sqrt{\mathbb{E} \|\eta_1\|_{\mathcal{Y}}^2}}{C_2 C_3 \|\eta_1\|_{\psi_1(\mathcal{Y})}}. \tag{112}$$

Recalling (109), we have that for all $q > 0$,

$$\begin{aligned} &\mathbb{P}\left[S_N \geq C_3 \frac{\|\eta_1\|_{\psi_1(\mathcal{Y})}}{\sqrt{N}} (I_{N,s_1,b} + q) \,\middle|\, x_1, \cdots, x_N \right] \\ &\leq \mathbb{P}\left[S_N \geq C_3 \frac{\|\eta_1\|_{\psi_1(\mathcal{Y})}}{\sqrt{N}} \left(\int_{\alpha}^{D(x_1,\cdots,x_N)} \log(1 + \mathfrak{C}(\tau; \Theta, \mathrm{d}_N))\, d\tau + q \right) \,\middle|\, x_1, \cdots, x_N \right] \\ &\leq 2e^{-(q - E/C_3)/D(x_1,\cdots,x_N)} \\ &\leq 2e^{-(q - E_\eta)/2b}. \end{aligned} \tag{113}$$

Taking expectation with respect to $\{x_1, \cdots, x_N\}$ on both sides of (113), we obtain that for all $q > 0$,

$$\mathbb{P}\left[S_N \geq C_3 \frac{\|\eta_1\|_{\psi_1(\mathcal{Y})}}{\sqrt{N}} \left(I_{N,s_1,b} + q \right) \right] \leq 2e^{-(q - E_\eta)/2b}. \tag{114}$$

Recalling (107), we have that for all $q > 0$,

$$\mathbb{P}\left\{ B_N \geq 2C_3 \frac{\|\eta_1\|_{\psi_1(\mathcal{Y})}}{\sqrt{N}} \left(I_{N,s_1,b} + q \right) \right\} \leq \mathbb{P}\left\{ S_N \geq C_3 \frac{\|\eta_1\|_{\psi_1(\mathcal{Y})}}{\sqrt{N}} \left(I_{N,s_1,b} + q \right) \right\} \leq 2e^{-(q-E_\eta)/2b}.$$

Then, for any $\delta \in (0,1)$ and $N \geq \log(2/\delta)$, by choosing $q = E_\eta - 2b\log(\delta/4)$, it holds that

$$\mathbb{P}\left\{ B_N \geq 2C_3 \frac{\|\eta_1\|_{\psi_1(\mathcal{Y})}}{\sqrt{N}} \left(I_{N,s_1,b} + E_\eta - 2b\log(\delta/4) \right) \right\} \leq \delta/2.$$

Thus, given an arbitrary $\delta > 0$, with $I_{N,s_1,b}$ and E_η defined in (112), we can choose

$$\lambda \geq 4C_3 \frac{\|\eta_1\|_{\psi_1(\mathcal{Y})}}{\sqrt{N}} \left(I_{N,s_1,b} + E_\eta - 2b\log(\delta/4) \right) \tag{115}$$

such that $\mathbb{P}(\mathrm{E}) \geq 1 - \frac{\delta}{2}$.

Remark 4.6 (Order of convergence)**.** *If $s_1 > 1$, then τ^{-1/s_1} is integrable on $(0, 2b)$. So, $I_{N,s_1,b}$ is bounded by some constant B for all N. Then, the order of convergence is $\mathcal{O}(N^{-1/2})$. If $0 < s_1 < 1$, directly integrating $I_{N,s_1,b}$ yields that*

$$I_{N,s_1,b} = C_4 \frac{(2b)^{1-\frac{1}{s_1}} - N^{-\frac{1}{2}(1-\frac{1}{s_1})}}{1 - 1/s_1} + 2b = \mathcal{O}\left(N^{\frac{1-s_1}{2s_1}} \right).$$

In this case, if $\frac{1}{2} < s_1 < 1$, then the order of convergence is $\mathcal{O}\left(N^{-\left(1-\frac{1}{2s_1}\right)} \right)$; if $0 < s_1 \leq \frac{1}{2}$, then this result does not give convergence of B_N to zero as N goes to infinity.

In the case where Θ is a compact metric space and has a finite topological dimension, then by the volumetric bound of the covering number in Euclidean space (see example 2.2), we have that $\log \mathfrak{C}(\tau; \Theta, \mathrm{d}_N) \lesssim \log \tau^{-1} \lesssim \tau^{-1}$. So, in this case, $s_1 = 1$. Integrating $I_{N,s_1,b}$ yields that

$$I_{N,s_1,b} = C_4 \left(\log(2b) + \frac{1}{2}\log N \right) + 2b = \mathcal{O}\left(\log N \right).$$

In this case, the convergence rate is $\mathcal{O}(N^{-1/2}\log N)$. This rate is, of course, not sharp due to the $\log N$ factor. Instead, if we have that

$$\Theta = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq a\}$$

for some diameter $a > 0$, then the volumetric bound gives

$$\mathfrak{C}(\tau; \Theta, \mathrm{d}_N) \leq \left(\frac{2a}{\tau} + 1 \right)^d.$$

Plugging this bound into (111), we get that the Dudley's integral is bounded by

$$d \int_\alpha^{2b} \log\left(\frac{2a}{\tau} + 1 \right) d\tau + 2b\,.$$

Since the integral converges to a constant (depending on a, b and d) as $\alpha \to 0^+$, by setting $\alpha = 0$, we get that the convergence rate of $B_N \to 0$ as $N \to \infty$ in this case is actually $\mathcal{O}(N^{-1/2})$ for

$$\lambda \gtrsim b\,\|\eta_1\|_{\psi_1(\mathcal{Y})} \log(4/\delta) N^{-1/2}\,. \tag{116}$$

In this scenario, the rate of the refined method is better than the previous two methods, where one-step discretization gives a convergence rate of $\mathcal{O}\left(\sqrt{\frac{\log N}{N}} \right)$ and chaining gives a convergence rate of $\mathcal{O}(N^{-1/2}\log N)$ without Assumption 4.6.

Finally, to complete Step 1, we intersect event E and event E_δ, and union bound the failure probabilities. Explicitly, $\mathbb{P}(\mathrm{E} \cap \mathrm{E}_\delta) = 1 - \mathbb{P}(\mathrm{E}^\mathrm{c} \cup \mathrm{E}_\delta^\mathrm{c}) \geq 1 - \mathbb{P}(\mathrm{E}^\mathrm{c}) - \mathbb{P}(\mathrm{E}_\delta^\mathrm{c})$. Thus, for any $\delta > 0$ and $N \geq \log(2/\delta)$, with a suitable choice of λ as in (115), it holds

$$\mathscr{R}_N\left(\widehat{\mu}; f^\star\right) \leq \left(32e^{3/2} b\,\|\eta_1\|_{\psi_1(\mathcal{Y})} \sqrt{\frac{\log(2/\delta)}{N}} + 2\lambda \right) \|\mu^\star\|_{\mathrm{TV}} \tag{117}$$

with probability at least $1 - \delta$ on the event $\mathrm{E} \cap \mathrm{E}_\delta$ over $\{x_n\} \sim \nu^{\otimes N}$.

4.2.2 Step 2: Estimation error

With (58), (65), and (71) in mind, we have that for any $\delta > 0$ and $N \geq \log(2/\delta)$,

$$\|\widehat{\mu}\|_{\mathrm{TV}} = t \leq 2\lambda^{-1} A_N^{\lambda} \leq \left(32\lambda^{-1} e^{3/2} b \, \|\eta_1\|_{\psi_1(\mathcal{Y})} \sqrt{\frac{\log(2/\delta)}{N}} + 2 \right) \|\mu^\star\|_{\mathrm{TV}} \tag{118}$$

with probability at least $1 - \frac{\delta}{2}$ on the event $\mathrm{E} \cap \mathrm{E}_\delta$ over $\{x_n\} \sim \nu^{\otimes N}$. If Θ is infinite-dimensional, then applying the technique of one-step discretization in Step 1 yields that λ satisfies (103). If Θ has a finite topological dimension d, then applying the refined approach in Step 1 yields that λ satisfies (116). They both show that

$$\lambda \gtrsim b \, \|\eta_1\|_{\psi_1(\mathcal{Y})} \sqrt{\frac{\log(2/\delta)}{N}} \tag{119}$$

So,

$$t \leq c \|\mu^\star\|_{\mathrm{TV}} =: t' \tag{120}$$

for some absolute constant $c > 0$. Then, we proceed by bounding

$$\begin{aligned} \mathscr{R}(\widehat{\mu}; f^\star) - \mathscr{R}_N(\widehat{\mu}; f^\star) &\leq \sup_{\|\mu\|_{\mathrm{TV}} \leq t'} |\mathscr{R}(\mu; f^\star) - \mathscr{R}_N(\mu; f^\star)| \\ &= \sup_{\|\mu\|_{\mathrm{TV}} \leq t'} \left| \frac{1}{N} \sum_{n=1}^N \|f^\star(x_n) - f_\mu(x_n)\|_{\mathcal{Y}}^2 - \mathbb{E}_{x\sim\nu} \|f^\star(x) - f_\mu(x)\|_{\mathcal{Y}}^2 \right|. \end{aligned} \tag{121}$$

Note that for any $\delta > 0$, the inequality above holds for all $N \geq \log(2/\delta)$ with probability at least $1 - \frac{\delta}{2}$ on the event $\mathrm{E} \cap \mathrm{E}_\delta$ over $\{x_n\} \sim \nu^{\otimes N}$. Define

$$Z_{t'} := \sup_{\mu \in \mathcal{A}_{t'}} \left| \frac{1}{N} \sum_{n=1}^N \|f^\star(x_n) - f_\mu(x_n)\|_{\mathcal{Y}}^2 - \mathbb{E}_{x\sim\nu} \|f^\star(x) - f_\mu(x)\|_{\mathcal{Y}}^2 \right|,$$

where

$$\mathcal{A}_{t'} := \{\mu \in \mathcal{M}(\Theta) \mid \|\mu\|_{\mathrm{TV}} \leq t'\}.$$

For any $\mu \in \mathcal{A}_{t'}$ and $n \in \{1, \ldots, N\}$, let

$$X_n(t', \mu) := \|f^\star(x_n) - f_\mu(x_n)\|_{\mathcal{Y}}^2 - \mathbb{E}_{x\sim\nu} \|f^\star(x) - f_\mu(x)\|_{\mathcal{Y}}^2 .$$

We compute

$$\begin{aligned} |X_1(t', \mu)| &\leq 2 \|f^\star(x_1)\|_{\mathcal{Y}}^2 + 2\mathbb{E}_{x\sim\nu} \|f^\star(x)\|_{\mathcal{Y}}^2 + 2 \|f_\mu(x_1)\|_{\mathcal{Y}}^2 + 2\mathbb{E}_{u\sim\nu} \|f_\mu(x)\|_{\mathcal{Y}}^2 \\ &\leq 4\|f^\star\|_{L_\nu^\infty}^2 + 4b^2 (t')^2, \end{aligned}$$

where we used $|a - b|^2 \leq 2|a|^2 + 2|b|^2$ in the first inequality, and for the second inequality, we used the fact that for any $x \in \mathcal{X}$ ν-almost surely, $\|f_\mu(x)\|_{\mathcal{Y}}^2 = \left| \int_\Theta \varphi(x, \theta) d\mu(\theta) \right|^2 \leq (t')^2 \|\varphi\|_{L^\infty}^2$ on the set $\mathcal{A}_{t'}$. Note also that by Assumption 4.4, $f^\star \in \mathcal{B} \subset L^\infty(\mathcal{X}, \mathbb{R})$. This implies that

$$\|X_1(t', \cdot)\|_{\psi_1(C(\mathcal{A}_{t'}; \mathbb{R}))} = \left\| \sup_{\mu \in \mathcal{A}_{t'}} |X_1(t', \mu)| \right\|_{\psi_1} \leq 4\|f^\star\|_{L_\nu^\infty}^2 + 4b^2 (t')^2.$$

Note that by Assumption 4.2, for any $x \in \mathcal{X}$, we have $\varphi(x, \cdot) \in C(\Theta)$, so $f_\mu(x) = \int_\Theta \varphi(x, \theta) d\mu(\theta)$ is continuous with respect to the weak* topology on $\mathcal{M}(\Theta)$ in the μ variable by the definition of weak* convergence of $\mu_k \to \mu$. In addition, by Fubini's theorem,

$$\mathbb{E}_{x\sim\nu} f_\mu(x) = \int_{\mathcal{X}} \int_\Theta \varphi(x, \theta) d\mu(\theta) d\nu(x) = \int_\Theta \int_{\mathcal{X}} \varphi(x, \theta) d\nu(x) d\mu(x).$$

Let $\{\theta_k\} \to \theta$ in Θ. Since $\|\varphi\|_{L_\nu^\infty} \leq b$, by the dominated convergence theorem, it holds

$$\lim_{k\to\infty} \int_{\mathcal{X}} \varphi(x, \theta_k) d\nu(x) = \int_{\mathcal{X}} \lim_{k\to\infty} \varphi(x, \theta_k) d\nu(x) = \int_{\mathcal{X}} \varphi(x, \theta) d\nu(x).$$

Also, since $\nu \in \mathscr{P}(\mathcal{X})$ is a probability measure, it holds that $\int_{\mathcal{X}} \varphi(x, \theta) d\nu(x) \leq b\nu(\mathcal{X}) = b$. So, the map $\theta \mapsto \int_{\mathcal{X}} \varphi(x, \theta) d\nu(x)$ is bounded and continuous. Thus, $\mathbb{E}_{x\sim\nu} f_\mu(x)$ is continuous with respect to the weak*

topology on $\mathcal{M}(\Theta)$ in the μ variable by the definition of weak* convergence of $\mu_k \to \mu$. Thus, the $X_n(t', \cdot)$ do indeed belong to $C\left(\mathcal{A}_{t'}; \mathbb{R}\right)$ almost surely, as they can be written as a sum of affine and quadratic forms on $\mathcal{M}(\Theta)$ in the μ variable.

Next, we claim that $C\left(\mathcal{A}_{t'}; \mathbb{R}\right)$ is a separable Banach space when equipped with the supremum norm. The reason is that, by the Stone–Weierstrass theorem, if S is a compact metric space, then $C(S)$ is separable. In our case, $S = \mathcal{A}_{t'}$ is a compact metric space with respect to the weak* topology: the compactness follows by the Banach–Alaoglu theorem (see e.g. Rudin [26, Theorem 3.15, p. 68]) and the metrizability follows by the fact that $C(\Theta)$ is separable. Indeed, a result of van Neerven [30, Proposition 4.51, pp. 154] states that if X is a separable Banach space, then the weak* topology of the closed unit ball of X^* is metrizable.

Now, applying (9) in corollary 2.7 (taking the separable Banach space to be $C\left(\mathcal{A}_{t'}; \mathbb{R}\right)$ equipped with the supremum norm) yields that if $N \geq \log(2/\delta)$, it holds that

$$Z_{t'} \leq \mathbb{E}_{\{x_n\}}\left[Z_{t'}\right] + 32e^{3/2}\left(\|f^\star\|_{L^\infty_\nu}^2 + b^2 (t')^2\right)\sqrt{\frac{\log(2/\delta)}{N}} \tag{122}$$

on some event $\widetilde{\mathrm{E}}_\delta$ with probability at least $1 - \frac{\delta}{2}$ over $\{x_n\} \sim \nu^{\otimes N}$.

Since the supremum concentrates around its mean, it remains to show that its mean is small as a function of the sample size. We do this with Rademacher symmetrization. By Giné-Zinn symmetrization [see e.g. [32], Section 4.2, Proposition 4.11, pp. 107-108],

$$\mathbb{E}_{\{x_n\}}\left[Z_{t'}\right] \leq 2\mathbb{E}_{\{x_n\},\{\sigma_n\}} \sup_{\mu \in \mathcal{A}_{t'}} \left| \frac{1}{N}\sum_{n=1}^N \sigma_n \left\|f^\star\left(x_n\right) - f_\mu\left(x_n\right)\right\|_{\mathcal{Y}}^2 \right|, \quad \text{where} \quad \sigma_n \overset{\text{iid}}{\sim} \mathrm{Unif}(\{+1, -1\}). \tag{123}$$

With theorem 2.13 in mind, in order to bound (123), we first use the law of total expectation to get

$$\mathbb{E}_{\{x_n\},\{\sigma_n\}} \sup_{\mu \in \mathcal{A}_{t'}} \left| \frac{1}{N}\sum_{n=1}^N \sigma_n \left\|f^\star\left(x_n\right) - f_\mu\left(x_n\right)\right\|_{\mathcal{Y}}^2 \right| = \mathbb{E}_{\{x_n\}}\left[\mathbb{E}_{\{\sigma_n\}}\left[\sup_{\mu \in \mathcal{A}_{t'}} \frac{1}{N}\sum_{n=1}^N \sigma_n \left\|f^\star(x_n) - f_\mu(x_n)\right\|_{\mathcal{Y}}^2 \,\middle|\, x_1, \ldots, x_N\right]\right]. \tag{124}$$

To apply theorem 2.13 in our setting, we define the sets of functions

$$\mathcal{G} := \{x \mapsto \|f^\star(x) - f_\mu(x)\|_{\mathcal{Y}}^2 \mid \mu \in \mathcal{A}_{t'}\}$$

and

$$\mathcal{F} := \{f_\mu \mid \mu \in \mathcal{A}_{t'}\}$$

Now, the conditional expectation in (124) is the empirical Rademacher complexity of $\mathcal{G}$. Let $\alpha \geq 0$. Write $\nu_N := \frac{1}{N}\sum_{n=1}^N \delta_{x_n}$. Then, theorem 2.13 gives that

$$\mathbb{E}\left[\sup_{\mu \in \mathcal{A}_{t'}} \frac{1}{N}\sum_{n=1}^N \sigma_n \left\|f^\star(x_n) - f_\mu(x_n)\right\|_{\mathcal{Y}}^2 \,\middle|\, x_1, \ldots, x_N\right] \leq 4\alpha + \frac{12}{\sqrt{N}}\int_\alpha^\infty \sqrt{\log \mathfrak{C}\left(\varepsilon; \mathcal{G}, \mathrm{d}_{L^2_{\nu_N}(\mathcal{X};\mathbb{R})}\right)}\, d\varepsilon\,,$$

where the ∞ in the upper limit of the integral can be replaced by the $L^2_{\nu_N}(\mathcal{X};\mathbb{R})$-diameter of $\mathcal{G}$. Notice that any $g \in \mathcal{G}$ can be written as $g(x) = (\psi_\star(f))(x)$ for some $f \in \mathcal{F}$, where

$$\begin{aligned} \psi_\star \colon (\mathcal{F}, \mathrm{d}_{L^2_{\nu_N}(\mathcal{X};\mathcal{Y})}) &\to (\mathcal{G}, \mathrm{d}_{L^2_{\nu_N}(\mathcal{X};\mathbb{R})}) \\ f &\mapsto \left(x \mapsto \|f^\star(x) - f(x)\|_{\mathcal{Y}}^2\right) \end{aligned}$$

is a Nemytskii operator between pseudo-metric spaces. We claim that

$$\mathfrak{C}\left(\epsilon; \mathcal{G}, \mathrm{d}_{L^2_{\nu_N}(\mathcal{X};\mathbb{R})}\right) \leq \mathfrak{C}\left(\epsilon/\mathrm{Lip}(\psi_\star); \mathcal{F}, \mathrm{d}_{L^2_{\nu_N}(\mathcal{X};\mathcal{Y})}\right).$$

This follows from a result of Petersen and Zech [22, Lem. 14.12, p. 206] using that $\mathcal{G} = \psi_\star(\mathcal{F})$. It remains to bound $\mathrm{Lip}(\psi_\star)$. To do so, let f_1 and f_2 belong to $\mathcal{F}$. Then for any $x \in \mathcal{X}$, it holds that

$$\begin{aligned} |\psi_\star(f_1)(x) - \psi_\star(f_2)(x)| &= \left| \|f^\star(x) - f_1(x)\|_{\mathcal{Y}}^2 - \|f^\star(x) - f_2(x)\|_{\mathcal{Y}}^2 \right| \\ &= \left(\|f^\star(x) - f_1(x)\|_{\mathcal{Y}} + \|f^\star(x) - f_2(x)\|_{\mathcal{Y}}\right) \left| \|f^\star(x) - f_1(x)\|_{\mathcal{Y}} - \|f^\star(x) - f_2(x)\|_{\mathcal{Y}} \right| \\ &\leq 2(\|f^\star\|_\infty + bt')\, \|f_1(x) - f_2(x)\|_{\mathcal{Y}} \end{aligned}$$

by the reverse triangle inequality and the definition of $\mathcal{F}$. Then

$$\begin{aligned}\mathsf{d}_{L^2_{\nu_N}(\mathcal{X};\mathbb{R})}\big(\psi_\star(f_1),\psi_\star(f_2)\big) &= \left(\frac{1}{N}\sum_{n=1}^{N}|\psi_\star(f_1)(x_n)-\psi_\star(f_2)(x_n)|^2\right)^{1/2}\\ &\le 2(\|f^\star\|_{L^\infty_\nu}+bt')\left(\frac{1}{N}\sum_{n=1}^{N}\|f_1(x_n)-f_2(x_n)\|_{\mathcal{Y}}^2\right)^{1/2}\\ &=: \mathrm{Lip}(\psi_\star)\,\mathsf{d}_{L^2_{\nu_N}(\mathcal{X};\mathcal{Y})}(f_1,f_2)\,,\end{aligned}$$

where $\mathrm{Lip}(\psi_\star)=2(\|f^\star\|_{L^\infty_\nu}+bt')$. Notice that the diameter of $\mathcal{F}$ satisfies

$$\sup_{(f,f')\in\mathcal{F}\times\mathcal{F}}\mathsf{d}_{L^2_{\nu_N}(\mathcal{X};\mathcal{Y})}(f,f')\le 2bt'\,.$$

We deduce after changing variables that

$$\begin{aligned}\mathbb{E}\left[\sup_{\mu\in\mathcal{A}_{t'}}\frac{1}{N}\sum_{n=1}^{N}\sigma_n\,\|f^\star(x_n)-f_\mu(x_n)\|_{\mathcal{Y}}^2\,\middle|\,x_1,\dots,x_N\right] &\le 8\alpha(\|f^\star\|_{L^\infty_\nu}+bt')\\ &\quad+\frac{24(\|f^\star\|_{L^\infty_\nu}+bt')}{\sqrt{N}}\int_\alpha^{2bt'}\sqrt{\log\mathfrak{C}\big(\varepsilon;\mathcal{F},\mathsf{d}_{L^2_{\nu_N}(\mathcal{X};\mathcal{Y})}\big)}\,d\varepsilon\,.\end{aligned}\tag{125}$$

Now, our goal is to bound the covering number $\mathfrak{C}(\epsilon;\mathcal{F},\mathrm{d}_{L^2_{\nu_N}(\mathcal{X};\mathcal{Y})})$. Recall that the Banach–Alaoglu theorem (see e.g. Rudin [26, Theorem 3.15, p. 68]) gives that the set $\mathcal{A}_{t'}$ is weakly* compact as a subspace of the Banach space $\mathcal{M}(\Theta)$. In addition, since $\mathcal{A}_{t'}$ is a ball, it is convex. Thus, by Krein-Milman theorem (see e.g. Rudin [26, Theorem 3.23, p. 75]) , $\mathcal{A}_{t'}$ is the closed convex hull of the set of its extreme points, that is

$$\mathcal{A}_{t'}=\overline{\mathrm{conv}}\left(\mathrm{Ext}(\mathcal{A}_{t'})\right).$$

Bartolucci et al. [6, Lemma B.2] shows that

$$\mathrm{Ext}(\mathcal{A}_{t'})=\{\pm t'\delta_\theta:\theta\in\Theta\}.$$

Consider the transformation $T:\mathcal{A}_{t'}\to\mathcal{F}$ given by

$$T(\mu)(x)=f_\mu(x)=\int_\Theta\varphi(x,\theta)d\mu(\theta).$$

It is obvious that T is weakly* continuous and linear on μ. Then, we have

$$\mathcal{F}=T(\mathcal{A}_{t'})=T(\overline{\mathrm{conv}}\left(\mathrm{Ext}(\mathcal{A}_{t'})\right))\subseteq\overline{\mathrm{conv}}\left(T\left((\mathrm{Ext}(\mathcal{A}_{t'}))\right)\right).$$

We compute

$$T(t'\delta_\theta)=t'\int_\Theta\varphi(\cdot,\theta)d(\delta_\theta)(\theta)=t'\varphi(\cdot,\theta).$$

Similarly, $T(-t'\delta_\theta)=-t'\varphi(\cdot,\theta)$. Define the generating function class as

$$\mathfrak{G}:=\{t'\varphi(\cdot,\theta):\theta\in\Theta\}.$$

Then, we have that

$$\mathcal{F}\subseteq\overline{\mathrm{conv}}\left(\mathfrak{G}\cup(-\mathfrak{G})\right).$$

Since φ is continuous on θ for every $x\in\mathcal{X}$ and $\theta\in\Theta$ is compact, $\mathfrak{G}\cup(-\mathfrak{G})$ is compact as a subspace of $C(\Theta)$, because the continuous image of a compact set is compact. Then, the set $\overline{\mathrm{conv}}\left(\mathfrak{G}\cup(-\mathfrak{G})\right.$ is totally bounded. So, the covering numbers $\mathfrak{C}(\tau;\mathfrak{G}\cup(-\mathfrak{G}),\mathrm{d}_{L^2_{\nu_N}(\mathcal{X};\mathcal{Y})})$ and $\mathfrak{C}(\tau;\overline{\mathrm{conv}}\left(\mathfrak{G}\cup(-\mathfrak{G})\right.,\mathrm{d}_{L^2_{\nu_N}(\mathcal{X};\mathcal{Y})})$ are finite for all $\tau>0$.

Let $g\in\mathrm{conv}\left(\mathfrak{G}\cup(-\mathfrak{G})\right)$. Then, there exists $\{\alpha_j\}_{j=1}^M\subset\mathbb{R}$ and $\{g_j\}_{j=1}^M\subset\mathfrak{G}\cup(-\mathfrak{G})$ such that

$$g(x)=\sum_{j=1}^{M}\alpha_j g_j(x)$$

holds pointwise for all $x\in\mathcal{X}$, where $\alpha_j\ge 0$ for all j and $\sum_{j=1}^M\alpha_j=1$. Now, let $\mathcal{C}_\mathfrak{G}$ be a minimal $\epsilon/2$-cover of the set $\mathfrak{G}\cup(-\mathfrak{G})$. For each $g_j\in\mathfrak{G}\cup(-\mathfrak{G})$, we can find an $h_j\in\mathcal{C}_\mathfrak{G}$ such that $\mathrm{d}_{L^2_{\nu_N}(\mathcal{X};\mathcal{Y})}(g_j,h_j)\le\epsilon/2$. Consider

$$\sum_{j=1}^{M}\alpha_j h_j\in\mathrm{conv}(\mathcal{C}_\mathfrak{G}).$$

It holds that

$$\begin{aligned}
\mathrm{d}^2_{L^2_{\nu_N}(\mathcal{X};\mathcal{Y})}\left(g, \sum_{j=1}^M \alpha_j h_j\right) &= \frac{1}{N}\sum_{n=1}^N \left\| g(x_n) - \sum_{j=1}^M \alpha_j h_j(x_n) \right\|^2_{\mathcal{Y}} \\
&= \frac{1}{N}\sum_{n=1}^N \left\| \sum_{j=1}^M \alpha_j \Big(g_j(x_n) - h_j(x_n) \Big) \right\|^2_{\mathcal{Y}} \\
&\leq \frac{1}{N}\sum_{n=1}^N \sum_{j=1}^M \alpha_j \left\| g_j(x_n) - h_j(x_n) \right\|^2_{\mathcal{Y}}, \quad \text{by Jensen's inequality} \\
&= \sum_{j=1}^M \alpha_j \mathrm{d}^2_{L^2_{\nu_N}(\mathcal{X};\mathcal{Y})}(g_j, h_j) \\
&\leq \sum_{j=1}^M \alpha_j \frac{\epsilon^2}{4}, \quad \text{because } \alpha_j \geq 0 \quad \forall i \\
&= \epsilon^2/4, \quad \text{because } \sum_{j=1}^M \alpha_j = 1.
\end{aligned}$$

Thus, it holds that

$$\mathrm{d}_{L^2_{\nu_N}(\mathcal{X};\mathcal{Y})}\left(g, \sum_{j=1}^M \alpha_j h_j\right) \leq \epsilon/2.$$

For $g \in \partial\, \overline{\mathrm{conv}}\,(\mathfrak{G} \cup (-\mathfrak{G}))$, given any $\delta > 0$, we can find an $g' \in \mathrm{conv}\,(\mathfrak{G} \cup (-\mathfrak{G}))$ such that $\|g - g'\|_\infty < \delta$. We compute

$$\mathrm{d}_{L^2_{\nu_N}(\mathcal{X};\mathcal{Y})}(g, g') = \left(\frac{1}{N}\sum_{n=1}^N \|g(x_n) - g'(x_n)\|^2_{\mathcal{Y}} \right)^{1/2} < \delta.$$

Then, we have

$$\mathrm{d}_{L^2_{\nu_N}(\mathcal{X};\mathcal{Y})}\left(g, \sum_{j=1}^M \alpha_j h_j\right) \leq \mathrm{d}_{L^2_{\nu_N}(\mathcal{X};\mathcal{Y})}(g, g') + \mathrm{d}_{L^2_{\nu_N}(\mathcal{X};\mathcal{Y})}\left(g', \sum_{j=1}^M \alpha_j h_j\right) < \delta + \epsilon/2$$

for all $\delta > 0$. It follows that $\mathrm{conv}(\mathcal{C}_{\mathfrak{G}})$ is an $\epsilon/2$-cover of $\overline{\mathrm{conv}}(\mathfrak{G} \cup (-\mathfrak{G}))$.

Now, let $\mathcal{C}_{\mathrm{conv}(\mathcal{C}_{\mathfrak{G}})}$ be a minimal $\epsilon/2$-cover of $\mathrm{conv}(\mathcal{C}_{\mathfrak{G}})$. For an arbitary $g \in \overline{\mathrm{conv}}(\mathfrak{G} \cup (-\mathfrak{G}))$, we can find an $h \in \mathrm{conv}(\mathcal{C}_{\mathfrak{G}})$ such that $\mathrm{d}_{L^2_{\nu_N}(\mathcal{X};\mathcal{Y})}(g, h) \leq \epsilon/2$. We can also find an $h' \in \mathcal{C}_{\mathrm{conv}(\mathcal{C}_{\mathfrak{G}})}$ such that $\mathrm{d}_{L^2_{\nu_N}(\mathcal{X};\mathcal{Y})}(h, h') \leq \epsilon/2$. So, we have that $\mathrm{d}_{L^2_{\nu_N}(\mathcal{X};\mathcal{Y})}(g, h') \leq \epsilon$. It follows that $\mathcal{C}_{\mathrm{conv}(\mathcal{C}_{\mathfrak{G}})}$ is an ϵ-cover of $\overline{\mathrm{conv}}(\mathfrak{G} \cup (-\mathfrak{G}))$. Thus,

$$\begin{aligned}
\mathfrak{C}\left(\epsilon; \mathcal{F}, \mathrm{d}_{L^2_{\nu_N}(\mathcal{X};\mathcal{Y})}\right) &\leq \mathfrak{C}(\epsilon; \overline{\mathrm{conv}}(\mathfrak{G} \cup (-\mathfrak{G})), \mathrm{d}_{L^2_{\nu_N}(\mathcal{X};\mathcal{Y})}) \\
&\leq |\mathcal{C}_{\mathrm{conv}(\mathcal{C}_{\mathfrak{G}})}| \\
&= \mathfrak{C}\left(\epsilon/2; \mathrm{conv}(\mathcal{C}_{\mathfrak{G}}), \mathrm{d}_{L^2_{\nu_N}(\mathcal{X};\mathcal{Y})}\right). \qquad (126)
\end{aligned}$$

Recall that $\mathcal{C}_{\mathfrak{G}}$ is a minimal $\epsilon/2$-cover of the set $\mathfrak{G} \cup (-\mathfrak{G})$. So, as the convex hull of a finite set, $\mathrm{conv}(\mathcal{C}_{\mathfrak{G}})$ is a finite-dimensional subspace of $C(\Theta)$, and we have

$$\dim \mathrm{conv}(\mathcal{C}_{\mathfrak{G}}) \leq \mathfrak{C}\left(\epsilon/2; \mathfrak{G} \cup (-\mathfrak{G}), \mathrm{d}_{L^2_{\nu_N}(\mathcal{X};\mathcal{Y})}\right) - 1 < \infty.$$

A result due to van der Vaart and Wellner [29, Lemma 2.6.15, p. 146] gives that the set $\mathrm{conv}(\mathcal{C}_{\mathfrak{G}})$ has finite VC-subgraph index, and its VC-subgraph index satisfies

$$\begin{aligned}
V(\mathrm{conv}(\mathcal{C}_{\mathfrak{G}})) &\leq \dim(\mathrm{conv}(\mathcal{C}_{\mathfrak{G}})) + 2 \\
&\leq \mathfrak{C}\left(\epsilon/2; \mathfrak{G} \cup (-\mathfrak{G}), \mathrm{d}_{L^2_{\nu_N}(\mathcal{X},\mathcal{Y})}\right) + 1 < \infty.
\end{aligned}$$

Since φ is uniformly bounded, $\mathfrak{G} \cup (-\mathfrak{G})$ is in the $t'b$-ball of $C(\Theta)$, and thus the set $\mathcal{C}_{\mathfrak{G}}$ is inside the same ball due to the minimality of the cover. Then $\mathrm{conv}(\mathcal{C}_{\mathfrak{G}})$ is inside the same ball. So, applying van der Vaart and

Wellner [29, Theorem 2.6.7, p. 141] yields

$$\begin{aligned}\mathfrak{C}\left(\epsilon/2;\operatorname{conv}(\mathcal{C}_{\mathfrak{G}}),\mathrm{d}_{L^2_{\nu_N}}\right) &\leq K_1 V(\operatorname{conv}(\mathcal{C}_{\mathfrak{G}}))(16e)^{V(\operatorname{conv}(\mathcal{C}_{\mathfrak{G}}))}\left(\frac{2t'b}{\epsilon}\right)^{2V(\operatorname{conv}(\mathcal{C}_{\mathfrak{G}}))-2}\\ &\leq K_1\left[\mathfrak{C}\left(\frac{\epsilon}{2};\mathfrak{G}\cup(-\mathfrak{G}),\mathrm{d}_{L^2_{\nu_N}}\right)+1\right](16e)^{\mathfrak{C}\left(\frac{\epsilon}{2};\mathfrak{G}\cup(-\mathfrak{G}),\mathrm{d}_{L^2_{\nu_N}}\right)+1}\left(\frac{2t'b}{\epsilon}\right)^{2\mathfrak{C}\left(\frac{\epsilon}{2};\mathfrak{G}\cup(-\mathfrak{G}),\mathrm{d}_{L^2_{\nu_N}}\right)}\\ &\leq K_1\left[2\mathfrak{C}\left(\frac{\epsilon}{2};\mathfrak{G},\mathrm{d}_{L^2_{\nu_N}}\right)+1\right](16e)^{2\mathfrak{C}\left(\frac{\epsilon}{2};\mathfrak{G},\mathrm{d}_{L^2_{\nu_N}}\right)+1}\left(\frac{2t'b}{\epsilon}\right)^{4\mathfrak{C}\left(\frac{\epsilon}{2};\mathfrak{G},\mathrm{d}_{L^2_{\nu_N}}\right)}, \end{aligned} \tag{127}$$

where K_1 is a universal constant. It remains to bound the covering number $\mathfrak{C}\left(\epsilon/2;\mathfrak{G},\mathrm{d}_{L^2_{\nu_N}(\mathcal{X},\mathcal{Y})}\right)$. Suppose that $\theta\mapsto\varphi(x,\theta)$ is K-Lipschitz on θ, uniformly in $x\in\mathcal{X}$. We compute

$$\begin{aligned}\mathrm{d}_{L^2_{\nu_N}}\left(t'\varphi(\cdot,\theta),t'\varphi(\cdot,\theta')\right)^2 &= \frac{1}{N}\sum_{n=1}^{N}\left\|t'\varphi(\cdot,\theta)-t'\varphi(\cdot,\theta')\right\|_{\mathcal{Y}}^2\\ &\leq \frac{(t')^2}{N}\sum_{n=1}^{N}K^2\left\|\theta-\theta'\right\|_{\Theta}^2\\ &= (Kt')^2\left\|\theta-\theta'\right\|_{\Theta}^2.\end{aligned}$$

Consider the operator

$$\begin{aligned}\Psi_\star\colon\ (\Theta,\|\cdot\|_\Theta)&\to\left(\mathfrak{G},\mathrm{d}_{L^2_{\nu_N}(\mathcal{X},\mathcal{Y})}\right)\\ \theta&\mapsto\left(x\mapsto t'\varphi(x,\theta)\right).\end{aligned}$$

We have that $\Psi_\star$ is Kt'-Lipschitz on θ because

$$\mathrm{d}_{L^2_{\nu_N}}\left(\Psi_\star(\theta)-\Psi_\star(\theta')\right)=\mathrm{d}_{L^2_{\nu_N}}\left(t'\varphi(\cdot,\theta),t'\varphi(\cdot,\theta')\right)\leq Kt'\left\|\theta-\theta'\right\|_\Theta.$$

Then, we apply Petersen and Zech [22, Lemmma 14.12, p. 206] using $\mathfrak{G}=\Psi_\star(\Theta)$, it yields that

$$\mathfrak{C}\big(\epsilon/2;\mathfrak{G},\mathrm{d}_{L^2_{\nu_N}(\mathcal{X};\mathcal{Y})}\big)\leq\mathfrak{C}\big(\epsilon/(2Kt');\Theta,\mathrm{d}_{L^2_{\nu_N}(\mathcal{X};\mathcal{Y})}\big). \tag{128}$$

Suppose that Θ has finite topological dimension and that Θ is homeomorphic to a subspace of $\mathbb{R}^d$ for some $d>0$. By the volumetric bound of the covering number in Euclidean space (see Example 2.2), we may assume that the growth rate of the covering number $\mathfrak{C}\big(\epsilon/(2Kt');\Theta,\mathrm{d}_{L^2_{\nu_N}(\mathcal{X};\mathcal{Y})}\big)$ satisfies

$$\mathfrak{C}\big(\epsilon/(2Kt');\Theta,\mathrm{d}_{L^2_{\nu_N}(\mathcal{X};\mathcal{Y})}\big)\lesssim\left(\frac{Kt'}{\epsilon}\right)^d. \tag{129}$$

Combining (126), (127), (128), and (129), it hodls that

$$\mathfrak{C}\left(\epsilon;\mathcal{F},\mathrm{d}_{L^2_{\nu_N}(\mathcal{X};\mathcal{Y})}\right)\lesssim\left(\frac{Kt'}{\epsilon}\right)^d(16e)^{2\left(\frac{Kt'}{\epsilon}\right)^d}\left(\frac{2t'b}{\epsilon}\right)^{4\left(\frac{Kt'}{\epsilon}\right)^d}.$$

It follows that

$$\begin{aligned}\log\mathfrak{C}\left(\epsilon;\mathcal{F},\mathrm{d}_{L^2_{\nu_N}(\mathcal{X};\mathcal{Y})}\right)&\lesssim d\log\left(\frac{Kt'}{\epsilon}\right)+2\left(\frac{Kt'}{\epsilon}\right)^d\log(16e)+4\left(\frac{Kt'}{\epsilon}\right)^d\log\left(\frac{2t'b}{\epsilon}\right)\\ &\lesssim\left(\frac{Kt'}{\epsilon}\right)^d\log\left(\frac{2t'b}{\epsilon}\right).\end{aligned} \tag{130}$$

Now, we plug (130) into (125) to get

$$\begin{aligned}\mathbb{E}\left[\sup_{\mu\in\mathcal{A}_{t'}}\frac{1}{N}\sum_{n=1}^{N}\sigma_n\left\|f^\star(x_n)-f_\mu(x_n)\right\|_{\mathcal{Y}}^2\,\middle|\,x_1,\ldots,x_N\right]&\leq 8\alpha(\|f^\star\|_{L^\infty_\nu}+bt')\\ &+C\frac{24(\|f^\star\|_{L^\infty_\nu}+bt')}{\sqrt{N}}\int_\alpha^{2bt'}\left(\frac{Kt'}{\epsilon}\right)^{d/2}\sqrt{\log\left(\frac{2t'b}{\epsilon}\right)}\,d\varepsilon,\end{aligned} \tag{131}$$

where C is an absolute constant. Let $\alpha = \alpha(N)$ be a function in N that is close to zero for large N be chosen later. Especially, α satisfies $0 < \alpha \ll 2bt'$ for large N. Then, the integral is finite for all $s > 0$ as the function inside the integral is continuous (with respect to ϵ) and the integral is over a closed and bounded interval.

Now, we evaluate the asymptotic behavior as $\alpha \to 0^+$ of the integral:

$$I = \int_{\alpha}^{2bt'} \left(\frac{Kt'}{\epsilon}\right)^{d/2} \sqrt{\log\left(\frac{2bt'}{\epsilon}\right)}\, d\epsilon.$$

Note that if $0 < d < 2$, then the integrand converges on $(0, 2bt')$. So, the integral I is bounded by some constant B for all $\alpha(N)$. In this case, we set the α in the right hand side of (131) to be zero, then the expectation is bounded above by

$$24CB(\|f^\star\|_{L^\infty_\nu} + bt')N^{-1/2}\,.$$

Since d is the dimension of the parameter space Θ, we are interested in the situation where $d \geq 2$, especially when d is large. Let

$$u = \log\left(\frac{2bt'}{\epsilon}\right), \qquad \text{so} \quad \epsilon = 2bt'e^{-u}, \qquad d\epsilon = -\epsilon\, du.$$

Then, the integral becomes

$$\begin{aligned} I &= \int_0^{\log\left(\frac{2bt'}{\alpha}\right)} \left(\frac{K}{2b}\right)^{d/2} e^{\frac{d}{2}u}\sqrt{u}\cdot(2bt'e^{-u})\, du \\ &= 2bt'\left(\frac{K}{2b}\right)^{d/2} \int_0^{\log\left(\frac{2bt'}{\alpha}\right)} e^{\left(\frac{d}{2}-1\right)u}\sqrt{u}\, du. \end{aligned}$$

Define

$$\beta := \frac{d}{2} - 1, \quad C' := 2bt'\left(\frac{K}{2b}\right)^{d/2}, \quad L := \log\left(\frac{2bt'}{\alpha}\right).$$

Then,

$$I = C' \int_0^L e^{\beta u}\sqrt{u}\, du.$$

The asymptotic behavior of the integral depends on the value of β, i.e., on the dimension d.

- **Case 1:** $d = 2 \Rightarrow \beta = 0$**.**

  The integral simplifies to:

  $$\int_0^L \sqrt{u}\, du = \frac{2}{3}L^{3/2} = \frac{2}{3}\left(\log\left(\frac{2bt'}{\alpha}\right)\right)^{3/2}.$$

- **Case 2:** $d > 2 \Rightarrow \beta > 0$**.**

  Since the exponential function grows rapidly, the dominant contribution to the integral comes from values of u near the upper limit $u = L$. Note that for α being a small number which is close to zeor, it holds that $L \gg 1$. We use a change of variables to expand around $u = L$.

  Let

  $$v = \beta(L - u), \qquad \text{so} \quad u = L - \frac{v}{\beta}, \qquad du = -\frac{dv}{\beta}.$$

  Then as $u \to L$, we have $v \to 0$, and the integral becomes

  $$\begin{aligned} I/C' &= e^{\beta L}\int_0^L e^{-\beta(L-u)}\sqrt{u}\, du \\ &= e^{\beta L}\int_0^L e^{-\beta v}\sqrt{L - \frac{v}{\beta}}\cdot\frac{dv}{\beta}. \end{aligned}$$

  For large L, we approximate

  $$\sqrt{L - \frac{v}{\beta}} \approx \sqrt{L} \quad \text{(since } v \ll L \text{ when } u \to L\text{)}.$$

  Thus,

  $$I \lesssim C'\frac{\sqrt{L}}{\beta}e^{\beta L}\int_0^\infty e^{-\beta v}\, dv = C'\frac{\sqrt{L}}{\beta}e^{\beta L}\cdot\frac{1}{\beta}.$$

This gives the leading-order asymptotic behavior:

$$I \lesssim K^{d/2} b^{1-d/2} t' \frac{\sqrt{L}}{\beta^2} e^{\beta L} = K^{d/2} b^{1-d/2} t' \frac{\sqrt{\log\left(\frac{2bt'}{\alpha}\right)}}{\beta^2} \left(\frac{2bt'}{\alpha}\right)^{\beta}.$$

Choosing $\alpha = N^{-1}$ if $d = 2$ and $\alpha = N^{-1/d}$ if $d > 2$ yields that

$$I \lesssim \begin{cases} \log^{3/2}(bt'N)\,, & \text{if } d = 2\,, \\ K^{d/2} \frac{\sqrt{\log\left(bt' N^{1/d}\right)}}{(\frac{d}{2}-1)^2} (t')^{\frac{d}{2}} N^{\frac{1}{2}-\frac{1}{d}}\,, & \text{if } d > 2. \end{cases}$$

Since the right hand side of (131) is independent of the data $\{x_n\}$, we take expectation with respect to $\{x_n\}$ on both sides, combining (123) and (124), yields that for $d > 2$,

$$\mathbb{E}_{\{x_n\}}[Z_{t'}] \le 16(\|f^\star\|_{L^\infty_\nu} + bt')\mathcal{I} N^{-1/d}\,, \tag{132}$$

where

$$\mathcal{I} = \mathcal{I}(t', d, b, K, N) = 1 + C'' K^{d/2} \frac{\sqrt{\log\left(bt' N^{1/d}\right)}}{(\frac{d}{2}-1)^2} (t')^{\frac{d}{2}}\,, \tag{133}$$

and C'' is an absolute constant.

Plugging (132) into (122) yields that, for all $N \ge \log(2/\delta)$, with t' being defined in (120), it holds that for $d > 2$,

$$Z_{t'} \le 16(\|f^\star\|_{L^\infty_\nu} + bt')\mathcal{I} N^{-1/d} + 32 e^{3/2} \left(\|f^\star\|^2_{L^\infty_\nu} + b^2 (t')^2\right) \sqrt{\frac{\log(2/\delta)}{N}} \tag{134}$$

on the event $\widetilde{\mathrm{E}}_\delta$ with probability at least $1 - \frac{\delta}{2}$ over $\{x_n\} \sim \nu^{\otimes N}$.

Recall (121) that for any $\delta > 0$ and $N \ge \log(2/\delta)$,

$$\mathscr{R}(\widehat{\mu}; f^\star) - \mathscr{R}_N(\widehat{\mu}; f^\star) \le Z_{t'}$$

with probability at least $1 - \frac{\delta}{2}$ on the event $\mathrm{E} \cap \mathrm{E}_\delta$ over $\{x_n\} \sim \nu^{\otimes N}$.

To complete Step 2, we intersect event $\mathrm{E} \cap \mathrm{E}_\delta$ and event $\widetilde{\mathrm{E}}_\delta$, and union bound the failure probabilities. Explicitly, $\mathbb{P}((\mathrm{E} \cap \mathrm{E}_\delta) \cap \widetilde{\mathrm{E}}_\delta) = 1 - \mathbb{P}((\mathrm{E} \cap \mathrm{E}_\delta)^{\mathrm{c}} \cup \widetilde{\mathrm{E}}_\delta^{\mathrm{c}}) \ge 1 - \mathbb{P}((\mathrm{E} \cap \mathrm{E}_\delta)^{\mathrm{c}}) - \mathbb{P}(\widetilde{\mathrm{E}}_\delta^{\mathrm{c}})$. Thus, for any $\delta > 0$ and $N \ge \log(2/\delta)$, with t' being a function defined in (120) and $\mathcal{I}$ being a function defined in (133), it holds that for $d > 2$,

$$\mathscr{R}(\widehat{\mu}; f^\star) - \mathscr{R}_N(\widehat{\mu}; f^\star) \le 16(\|f^\star\|_{L^\infty_\nu} + bt')\mathcal{I} N^{-1/d} + 32 e^{3/2} \left(\|f^\star\|^2_{L^\infty_\nu} + b^2 (t')^2\right) \sqrt{\frac{\log(2/\delta)}{N}} \tag{135}$$

with probability at least $1 - \delta$ on the event $(\mathrm{E} \cap \mathrm{E}_\delta) \cap \widetilde{\mathrm{E}}_\delta$ over $\{x_n\} \sim \nu^{\otimes N}$.

The cases where $d = 1$ and $d = 2$ are treated similarly. We get that for any $\delta > 0$ and $N \ge \log(2/\delta)$, with t' being a function defined in (120), it holds that

$$\mathscr{R}(\widehat{\mu}; f^\star) - \mathscr{R}_N(\widehat{\mu}; f^\star) \lesssim \begin{cases} (\|f^\star\|_{L^\infty_\nu} + bt') N^{-1/2} + \left(\|f^\star\|^2_{L^\infty_\nu} + b^2 (t')^2\right) \sqrt{\log(2/\delta)} N^{-1/2}\,, & \text{if } d = 1\,, \\ (\|f^\star\|_{L^\infty_\nu} + bt') N^{-1/2} \log^{3/2}(bt'N) + \left(\|f^\star\|^2_{L^\infty_\nu} + b^2 (t')^2\right) \sqrt{\log(2/\delta)} N^{-1/2}, & \text{if } d = 2 \end{cases}$$

with probability at least $1 - \delta$ on the event $(\mathrm{E} \cap \mathrm{E}_\delta) \cap \widetilde{\mathrm{E}}_\delta$ over $\{x_n\} \sim \nu^{\otimes N}$.

Remark 4.7. *The approach presented above can handle an infinite-dimensional parameter space Θ if the covering numbers $\mathfrak{C}\left(\varepsilon; \Theta, \mathrm{d}_{L^2_{\nu_N}(\mathcal{X};\mathcal{Y})}\right)$ have a polynomial decay.*

4.2.3 Step 3: Conclusion

Recall the error decomposition

$$\mathscr{R}\left(\widehat{\mu}; f^\star\right) = \mathscr{R}_N\left(\widehat{\mu}; f^\star\right) + \left[\mathscr{R}\left(\widehat{\mu}; f^\star\right) - \mathscr{R}_N\left(\widehat{\mu}; f^\star\right)\right].$$

In Step 1, we obtain a high probability bound (117) for the approximation error $\mathscr{R}_N\left(\widehat{\mu}; f^\star\right)$. In Step 2, we obtain a high probability bound (135) for the estimation error $\mathscr{R}\left(\widehat{\mu}; f^\star\right) - \mathscr{R}_N\left(\widehat{\mu}; f^\star\right)$.

Before combining these two bounds, we note that they can be simplified by the results we have obtained. For λ satisfying (119), it holds that

$$\left(32e^{3/2}b\left\|\eta_1\right\|_{\psi_1(\mathcal{Y})}\sqrt{\frac{\log(2/\delta)}{N}}+2\lambda\right)\|\mu^\star\|_{\mathrm{TV}}\lesssim\lambda\|\mu^\star\|_{\mathrm{TV}}\,.$$

By the definition (120) of t', we have that $t'\lesssim\|\mu^\star\|_{\mathrm{TV}}$. In addition, $\|f^\star\|_{L^\infty_\nu}\leq b\|\mu^\star\|_{\mathrm{TV}}$. Furthermore,

$$\mathcal{I}\lesssim C_d\sqrt{\log(b\left\|\mu^\star\right\|_{\mathrm{TV}}\lambda^{-1})}\left\|\mu^\star\right\|_{\mathrm{TV}}^{d/2}=:\mathcal{I}'\,.$$

Finally, since for any $\delta>0$, $\log(4/\delta)\geq\log(2/\delta)$, we conclude that for any $\delta>0$ and $N\geq\log(4/\delta)$, with a suitable choice of λ satisfying (119), it holds that for $d>2$,

$$\mathscr{R}\left(\widehat{\mu};f^\star\right)\lesssim\|\mu^\star\|_{\mathrm{TV}}\lambda+b\left\|\mu^\star\right\|_{\mathrm{TV}}\mathcal{I}'\lambda^{2/d}+b\|\mu^\star\|_{\mathrm{TV}}^2\lambda$$

with probability at least $1-\delta$ over $\{x_n\}\sim\nu^{\otimes N}$.

The cases where $d=1$ and $d=2$ are treated similarly. We conclude that for any $\delta>0$ and $N\geq\log(4/\delta)$, with a suitable choice of λ satisfying (119), it holds that

$$\mathscr{R}\left(\widehat{\mu};f^\star\right)\lesssim\begin{cases}\|\mu^\star\|_{\mathrm{TV}}\lambda+\lambda+b\|\mu^\star\|_{\mathrm{TV}}^2\lambda\,, & \text{if } d=1\,,\\ \|\mu^\star\|_{\mathrm{TV}}\lambda+\lambda\log^{3/2}\left(b\|\mu^\star\|_{\mathrm{TV}}\lambda^{-1}\right)+b\|\mu^\star\|_{\mathrm{TV}}^2\lambda\,, & \text{if } d=2\,,\end{cases}$$

with probability at least $1-\delta$ over $\{x_n\}\sim\nu^{\otimes N}$.

4.3 Discussion and future work

By Remark 4.1 and Remark 4.2, Theorem 4.1 gives a convergence rate, which ensures statistical consistency in the well-specified setting. However, when the dimension of the parameter space d exceeds 2, the rate has an explicit dependence on d, deteriorating as d increases. This highlights the curse of dimensionality.

However, the convergence rate we obtain is not sharp, as it results from using the covering-number argument in bounding the estimation error. For future work, we may consider an alternative approach, analogous to Step 2 in Section 3.2. We outline this approach below.

Alternative approach for Step 2 We begin with Equation (123). The right-hand side is the Rademacher complexity of the neural network function class composed with the square loss. Expanding the square, it is bounded above by

$$2\mathbb{E}_{\{x_n\},\{\sigma_n\}}\left|\frac{1}{N}\sum_{n=1}^N\sigma_n\left\|f^\star\left(x_n\right)\right\|_{\mathcal{Y}}^2\right| \tag{I}$$

$$+4\mathbb{E}_{\{x_n\},\{\sigma_n\}}\sup_{\mu\in\mathcal{A}_{t'}}\left|\frac{1}{N}\sum_{n=1}^N\sigma_n\left\langle f^\star\left(x_n\right),f_\mu\left(x_n\right)\right\rangle_{\mathcal{Y}}\right| \tag{II}$$

$$+2\mathbb{E}_{\{x_n\},\{\sigma_n\}}\sup_{\mu\in\mathcal{A}_{t'}}\left|\frac{1}{N}\sum_{n=1}^N\sigma_n\left\|f_\mu\left(x_n\right)\right\|_{\mathcal{Y}}^2\right| \tag{III}$$

We now estimate each term. Apply Jensen's inequality on the first term (I) yields

$$\text{(I)}\leq 2\left(\mathbb{E}\left|\frac{1}{N}\sum_{n=1}^N\sigma_n\left\|f^\star\left(x_n\right)\right\|_{\mathcal{Y}}^2\right|^2\right)^{1/2}=\frac{2}{\sqrt{N}}\left(\frac{1}{N}\sum_{n=1}^N\mathbb{E}\left\|f^\star\left(x_n\right)\right\|_{\mathcal{Y}}^4\right)^{1/2}\leq\frac{2\|f^\star\|_{L^\infty_\nu}^2}{\sqrt{N}},$$

where we expand the square of the sum inside the expectation and use the fact that $\sigma_n^2=1$ and $\mathbb{E}[\sigma_n]=0$ to get the equality.

For the second term (II), we have

$$\begin{aligned}
&\mathbb{E}_{\{x_n\},\{\sigma_n\}} \sup_{\mu\in\mathcal{A}_{t'}} \left| \frac{1}{N}\sum_{n=1}^{N} \sigma_n \left\langle f^\star(x_n), f_\mu(x_n)\right\rangle_{\mathcal{Y}} \right| \\
&= \mathbb{E}_{\{x_n\},\{\sigma_n\}} \sup_{\mu\in\mathcal{A}_{t'}} \left| \int_\Theta \frac{1}{N}\sum_{n=1}^{N} \sigma_n \left\langle f^\star(x_n), \varphi(x_n,\theta)\right\rangle_{\mathcal{Y}} d\mu(\theta) \right| \\
&\le t' \mathbb{E}_{\{x_n\},\{\sigma_n\}} \left\| \frac{1}{N}\sum_{n=1}^{N} \sigma_n \left\langle f^\star(x_n), \varphi(x_n,\cdot)\right\rangle_{\mathcal{Y}} \right\|_{L^\infty(\Theta,\mathbb{R})} \\
&= t' \mathbb{E}_{\{x_n\},\{\sigma_n\}} \sup_{\theta\in\Theta} \left| \frac{1}{N}\sum_{n=1}^{N} \sigma_n \left\langle f^\star(x_n), \varphi(x_n,\theta)\right\rangle_{\mathcal{Y}} \right|. \qquad (136)
\end{aligned}$$

Next, we aim to bound the expectation of the supremum above via chaining using Dudley integral covering number bound. Consider the class of real-valued functions $\mathfrak{G} := \{g_\theta : \theta \in \Theta\} \cup \{0\}$, where $g_\theta : x \mapsto \langle f^\star(x), \varphi(x;\theta)\rangle_{\mathcal{Y}}$. For $g_\theta, g_{\theta'} \in \mathfrak{G}$, define $\overline{D}$ as

$$\overline{D}(g_\theta, g_{\theta'}) = \left(\frac{1}{N}\sum_{n=1}^{N} |g_\theta(x_n) - g_{\theta'}(x_n)|^2 \right)^{1/2}.$$

It is easy to check that $\overline{D}(\cdot,\cdot)$ is a pseudometric on $\mathfrak{G}$. Note that we have

$$\sup_{\theta\in\Theta} \left(\frac{1}{N}\sum_{n=1}^{N} |g_\theta(x_n)|^2 \right)^{1/2} \le b \left\| f^\star \right\|_{L^\infty_\nu}. \qquad (137)$$

For each $j \in \mathbb{N}_+$, let $\epsilon_j := b\left\|f^\star\right\|_{L^\infty_\nu} 2^{-j}$ and $\widehat{\mathfrak{G}}_j$ be a minimal ϵ_j-cover of $\mathfrak{G}$ w.r.t. $\overline{D}$. For each $g_\theta \in \mathfrak{G}$ and $j \in \mathbb{N}$, let $\hat{g}_{\theta_j} \in \widehat{\mathfrak{G}}_j$ be such that $\overline{D}(g_\theta, \hat{g}_{\theta_j}) \le \epsilon_j$. From (137), the covering number $\mathfrak{C}(\epsilon; \mathfrak{G}, \overline{D})$ is finite for all $\epsilon > 0$. Now, fix $g_\theta \in \mathfrak{G}$. For a given $m \in \mathbb{N}_+$ to be chosen later, we define the telescoping sum:

$$g_\theta = g_\theta - \hat{g}_{\theta_m} + \sum_{j=1}^{m} \left(\hat{g}_{\theta_j} - \hat{g}_{\theta_{j-1}} \right),$$

where $\hat{g}_{\theta_0} = 0$. We estimate

$$\begin{aligned}
&\mathbb{E}_{\{x_n\},\{\sigma_n\}} \sup_{\theta\in\Theta} \left| \frac{1}{N}\sum_{n=1}^{N} \sigma_n g_\theta(x_n) \right| \\
&= \mathbb{E}_{\{x_n\},\{\sigma_n\}} \sup_{\theta\in\Theta} \left| \frac{1}{N}\sum_{n=1}^{N} \sigma_n \left(g_\theta(x_n) - \hat{g}_{\theta_m}(x_n) + \sum_{j=1}^{m} \left[\hat{g}_{\theta_j}(x_n) - \hat{g}_{\theta_{j-1}}(x_n) \right] \right) \right| \\
&\le \mathbb{E}_{\{x_n\},\{\sigma_n\}} \sup_{\theta\in\Theta} \left| \frac{1}{N}\sum_{n=1}^{N} \sigma_n \left(g_\theta(x_n) - \hat{g}_{\theta_m}(x_n) \right) \right| \qquad \text{(i)} \\
&\quad + \sum_{j=1}^{m} \mathbb{E}_{\{x_n\},\{\sigma_n\}} \sup_{\theta\in\Theta} \left| \frac{1}{N}\sum_{n=1}^{N} \sigma_n \left(\hat{g}_{\theta_j}(x_n) - \hat{g}_{\theta_{j-1}}(x_n) \right) \right|. \qquad \text{(ii)}
\end{aligned}$$

(i) is bounded above by

$$\begin{aligned}
&\mathbb{E}_{\{x_n\},\{\sigma_n\}} \sup_{\theta\in\Theta} \frac{1}{N}\sum_{n=1}^{N} |\sigma_n| \left| (g_\theta(x_n) - \hat{g}_{\theta_m}(x_n)) \right| \\
&\le \sup_{\theta\in\Theta} \overline{D}(g_\theta, \hat{g}_{\theta_m}) \le \epsilon_m. \qquad (138)
\end{aligned}$$

For (ii), since the covering number $\mathfrak{C}(\epsilon; \mathfrak{G}, \overline{D})$ is finite for all $\epsilon > 0$, we can apply Massart's lemma on $\mathscr{F} = \{\hat{g}_{\theta_j} - \hat{g}_{\theta_{j-1}} : \hat{g}_{\theta_j} \in \widehat{\mathfrak{G}}_j, \hat{g}_{\theta_{j-1}} \in \widehat{\mathfrak{G}}_{j-1}\}$, where $|\mathscr{F}| \le |\widehat{\mathfrak{G}}_j \| \widehat{\mathfrak{G}}_{j-1}| < \infty$. Then, (ii) is bounded above by

$$\sum_{j=1}^{m} \max_{\substack{\hat{\theta}_j \in \widehat{\Theta}_j \\ \hat{\theta}_{j-1} \in \widehat{\Theta}_{j-1}}} \mathbb{E}_{\{x_n\},\{\sigma_n\}} \left(\frac{1}{N}\sum_{n=1}^{N} \left[\hat{g}_{\theta_j}(x_n) - \hat{g}_{\theta_{j-1}}(x_n) \right]^2 \right)^{1/2} \sqrt{\frac{2\log\left(\mathfrak{C}(\epsilon_j; \mathfrak{G}, \overline{D}) \mathfrak{C}(\epsilon_{j-1}; \mathfrak{G}, \overline{D}) \right)}{N}}. \qquad (139)$$

Note that

$$\left(\frac{1}{N}\sum_{n=1}^{N}\left[\hat{g}_{\theta_j}(x_n)-\hat{g}_{\theta_{j-1}}(x_n)\right]^2\right)^{1/2} \le \left(\frac{1}{N}\sum_{n=1}^{N}\left[\hat{g}_{\theta_j}(x_n)-g_\theta(x_n)\right]^2\right)^{1/2} + \left(\frac{1}{N}\sum_{n=1}^{N}\left[g_\theta(x_n)-\hat{g}_{\theta_{j-1}}(x_n)\right]^2\right)^{1/2} \tag{140}$$

$$\begin{aligned} &\le \overline{D}(\hat{g}_{\theta_j}, g_\theta) + \overline{D}(g_\theta, \hat{g}_{\theta_{j-1}}) \\ &\le \epsilon_j + \epsilon_{j-1} \\ &= 6\left(\epsilon_j - \epsilon_{j+1}\right), \end{aligned} \tag{141}$$

where we used Minkowski inequality in (140), and (141) is due to our construction that $\epsilon_{j-1} = 2\epsilon_j$ for all $j \in \mathbb{N}_+$.

Thus, (139) is bounded above by

$$\sum_{j=1}^{m} 6\left(\epsilon_j - \epsilon_{j+1}\right)\sqrt{\frac{2\log\left(\mathfrak{C}(\epsilon_j;\mathfrak{G},\overline{D})\mathfrak{C}(\epsilon_{j-1};\mathfrak{G},\overline{D})\right)}{N}}. \tag{142}$$

By the definition of ϵ-covering number, $\mathfrak{C}(\epsilon;\mathfrak{G},\overline{D})$ is non-increasing w.r.t. ϵ. Since $\epsilon_{j-1} = 2\epsilon_j$ for all $j \in \mathbb{N}_+$, we have that $\mathfrak{C}(\epsilon_j;\mathfrak{G},\overline{D}) \ge \mathfrak{C}(\epsilon_{j-1};\mathfrak{G},\overline{D})$ for all $j \in \mathbb{N}_+$.

Thus, combining (138), and (142), we have that

$$\begin{aligned} (136)/t' &\le \epsilon_m + \frac{12}{\sqrt{N}}\sum_{j=1}^{m}\left(\epsilon_j - \epsilon_{j+1}\right)\sqrt{\log\mathfrak{C}(\epsilon_j;\mathfrak{G},\overline{D})} \\ &= \epsilon_m + \frac{12}{\sqrt{N}}\sum_{j=1}^{m}\int_{\epsilon_{j+1}}^{\epsilon_j}\sqrt{\log\mathfrak{C}(\epsilon_j;\mathfrak{G},\overline{D})}d\tau \\ &\le \epsilon_m + \frac{12}{\sqrt{N}}\sum_{j=1}^{m}\int_{\epsilon_{j+1}}^{\epsilon_j}\sqrt{\log\mathfrak{C}(\tau;\mathfrak{G},\overline{D})}d\tau, \quad \text{since } \tau \in [\epsilon_{j+1},\epsilon_j] \\ &\le \epsilon_m + \frac{12}{\sqrt{N}}\int_{\epsilon_{m+1}}^{c/2}\sqrt{\log\mathfrak{C}(\tau;\mathfrak{G},\overline{D})}d\tau. \end{aligned}$$

For any $\epsilon \in [0, c/4]$, we can choose m such that $\epsilon \le \epsilon_{m+1} \le 2\epsilon$ or equivalently $m = \sup\{j \in \mathbb{N}_+ : \epsilon_j \ge 2\epsilon\}$. Thus,

$$(\mathrm{II}) \le 4t' \inf_{\epsilon \in \left[0, \frac{b}{4}\|f^\star\|_{L^\infty_\nu}\right]}\left\{16\epsilon + \frac{48}{\sqrt{N}}\int_{\epsilon}^{\frac{b}{2}\|f^\star\|_{L^\infty_\nu}}\sqrt{\log\mathfrak{C}(\tau;\mathfrak{G},\overline{D})}d\tau\right\}.$$

Note that $f^\star(x)$ is bounded by $b\|\mu^\star\|_{\mathrm{TV}}$ for all $x \in \mathcal{X}$. A similar reasoning as in Remark 4.4 implies that the covering numbers $\mathfrak{C}(\tau;\mathfrak{G},\overline{D})$ have a polynomial decay.

Estimation of the third term (III) is a bit tricky since it involves a quadratic empirical process. We have

$$\begin{aligned} &\mathbb{E}_{\{x_n\},\{\sigma_n\}}\sup_{\mu\in\mathcal{A}_{t'}}\left|\frac{1}{N}\sum_{n=1}^{N}\sigma_n\left\|f_\mu(x_n)\right\|_{\mathcal{Y}}^2\right| \\ &= \mathbb{E}_{\{x_n\},\{\sigma_n\}}\sup_{\mu\in\mathcal{A}_{t'}}\left|\frac{1}{N}\sum_{n=1}^{N}\sigma_n\left\|\int_\Theta\varphi(x_n,\theta)d\mu(\theta)\right\|_{\mathcal{Y}}^2\right|. \end{aligned}$$

One approach to handling the square is through (vector-valued) contraction mapping theorems. This allows us to move the square inside the integral, after which we can apply a standard Dudley metric entropy bound to complete the argument.

Other future directions A natural direction for future work is to extend our results to the vector-valued setting. Secondly, the results may be extended to the misspecified setting, where there exist $\rho \in L^\infty_v(\mathcal{X};\mathcal{Y})$ and $f_{\mathcal{B}} \in \mathcal{B}$ such that the operator $f^\star : \mathcal{X} \to \mathcal{Y}$ admits the decomposition $f^\star = \rho + f_{\mathcal{B}}$.

References

[1] Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404.

[2] Bach, F. (2017a). Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53.

[3] Bach, F. (2017b). On the equivalence between kernel quadrature rules and random feature expansions. *Journal of Machine Learning Research*, 18(21):1–38.

[4] Barron, A. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945.

[5] Barron, A. (1994). Approximation and estimation bounds for artificial neural networks. *Machine Learning*, 14:115–133.

[6] Bartolucci, F., De Vito, E., Rosasco, L., and Vigogna, S. (2023). Understanding neural networks with reproducing kernel Banach spaces. *Applied and Computational Harmonic Analysis*, 62:194–236.

[7] Carmeli, C., De Vito, E., Toigo, A., and Umanità, V. (2010). Vector valued reproducing kernel hilbert spaces and universality. *Analysis and Applications*, 08(01):19–61.

[8] Carratino, L., Rudi, A., and Rosasco, L. (2018). Learning with sgd and random features. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

[9] Combettes, P. L., Salzo, S., and Villa, S. (2018). Regularized learning schemes in feature banach spaces. *Analysis and Applications*, 16(01):1–54.

[10] Dziugaite, G. K. and Roy, D. M. (2017). Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence (UAI 2017)*, pages 173–182. AUAI Press.

[11] E, W., Ma, C., and Wu, L. (2022). The barron space and the flow-induced function spaces for neural network models. *Constructive Approximation*, 55(2):369–406.

[12] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30.

[13] Kurkova, V. and Sanguineti, M. (2001). Bounds on rates of variable-basis and neural-network approximation. *IEEE Transactions on Information Theory*, 47(6):2659–2665.

[14] Lanthaler, S. and Nelsen, N. H. (2023). Error bounds for learning with vector-valued random features. In Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 71834–71861. Curran Associates, Inc.

[15] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436.

[16] Lin, R. R., Zhang, H. Z., and Zhang, J. (2022). On reproducing kernel banach spaces: Generic definitions and unified framework of constructions. *Acta Mathematica Sinica, English Series*, 38(6):1459–1483.

[17] Liu, F., Huang, X., Chen, Y., and Suykens, J. A. K. (2022). Random features for kernel approximation: A survey on algorithms, theory, and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7128–7148.

[18] Ma, T. (2022). Lecture notes for machine learning theory. Technical report, Stanford University. Lecture notes, June 26, 2022.

[19] Munkres, J. R. (2000). *Topology*. Prentice Hall, Upper Saddle River, NJ, 2nd edition.

[20] Nelsen, N. H. and Stuart, A. M. (2021). The random feature model for input-output maps between banach spaces. *SIAM Journal on Scientific Computing*, 43(5):A3212–A3243.

[21] Neyshabur, B., Li, Z., Bhojanapalli, S., LeCun, Y., and Srebro, N. (2018). Towards understanding the role of over-parametrization in generalization of neural networks. *arXiv preprint arXiv:1805.12076*.

[22] Petersen, P. and Zech, J. (2024). Mathematical theory of deep learning. *preprint arXiv:2407.18384*.

[23] Rahimi, A. and Recht, B. (2007). Random features for large-scale kernel machines. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc.

[24] Rahimi, A. and Recht, B. (2008). Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc.

[25] Ratti, L. (2025). *Learned reconstruction methods for inverse problems: sample error estimates*, pages 163–200. De Gruyter, Berlin, Boston.

[26] Rudin, W. (1991). *Functional Analysis*, volume 8 of *International Series in Pure and Applied Mathematics*. McGraw-Hill, 2nd edition.

[27] Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 1st edition.

[28] Talagrand, M. (2014). *Upper and Lower Bounds for Stochastic Processes*. A Series of Modern Surveys in Mathematics. Springer Berlin, Heidelberg, 1 st edition.

[29] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer.

[30] van Neerven, J. (2022). *Functional Analysis*, volume 201 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge; New York.

[31] Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

[32] Wainwright, M. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

[33] Weinan, E., Ma, C., and Wu, L. (2019). A priori estimates of the population risk for two-layer neural networks. *Communications in Mathematical Sciences*, 17(5):1407–1425.