Towards Empirical Process Theory for Vector-Valued Functions: Metric Entropy of Smooth Function Classes

Junhyung Park JUNHYUNG.PARK@TUEBINGEN.MPG.DE
Max Planck Institute for Intelligent Systems, Max-Planck-Ring 4, 72076 Tübingen, Germany

Krikamol Muandet MUANDET@CISPA.DE
CISPA – Helmholtz Center for Information Security, Stuhlsatzenhaus 5, 66123 Saarbrücken, Germany

Editors: Shipra Agrawal and Francesco Orabona

Abstract

This paper provides some first steps in developing empirical process theory for functions taking values in a vector space. Our main results provide bounds on the entropy of classes of smooth functions taking values in a Hilbert space, by leveraging theory from differential calculus of vector-valued functions and fractal dimension theory of metric spaces. We demonstrate how these entropy bounds can be used to show the uniform law of large numbers and asymptotic equicontinuity of the function classes, and also apply it to statistical learning theory in which the output space is a Hilbert space. We conclude with a discussion on the extension of Rademacher complexities to vector-valued function classes.

Keywords: Empirical Processes, Vector-Valued Functions, Metric Entropy, Upper Box-Counting Dimension, Assouad Dimension, Empirical Risk Minimisation, Rademacher Complexity

1. Introduction

Empirical process theory is an important branch of probability theory that deals with the empirical measure $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ based on random independent and identically distributed (i.i.d.) copies $X_1, ..., X_n$ of a random variable X on a domain $\mathcal{X}$, and stochastic processes of the form $\{P_n f - Pf : f \in \mathcal{F}\}$, where $\mathcal{F}$ is a class of functions $\mathcal{X} \to \mathbb{R}$. Due to its very nature, the theory has found a wealth of applications in statistics (van der Vaart and Wellner, 1996; van de Geer, 2000; Kosorok, 2008; Shorack and Wellner, 2009; Dudley, 2014). In particular, it has been the major tool in analysing properties of estimators in supervised learning, both in regression and classification (Györfi et al., 2006; Steinwart and Christmann, 2008; Shalev-Shwartz and Ben-David, 2014).

In the traditional (and still dominant) supervised learning setting, the output space is (a subset of) $\mathbb{R}$, but there is a rapidly growing literature in machine learning and statistics on learning vector-valued functions (Micchelli and Pontil, 2005; Álvarez et al., 2012), and efforts are already under way to explore ways to make them faster and more robust (Laforgue et al., 2020; Lambert et al., 2022; Ahmad et al., 2022). This occurs, for example, in multi-task or multi-output learning (Evgeniou et al., 2005; Yousefi et al., 2018; Xu et al., 2019; Reeve and Kaban, 2020), functional response models (Morris, 2015; Kadri et al., 2016; Brault, 2017; Saha and Palaniappan, 2020), kernel conditional mean embeddings (Grünewälder et al., 2012; Park and Muandet, 2020a) or structured prediction (Ciliberto et al., 2020; Laforgue et al., 2020), among others. Very recently, there is even an interest in the more general setting of learning mappings between two metric spaces (Hanneke et al., 2020; Cohen and Kontorovich, 2022).

There are valuable works analysing the properties of vector-valued regressors with specific algorithms, notably integral operator techniques in vector-valued reproducing kernel Hilbert space regression (Caponnetto and De Vito, 2006; Kadri et al., 2016; Singh et al., 2019; Park and Muandet, 2020b; Cabannes et al., 2021), and in the form of (local) Rademacher complexities, empirical process theoretic techniques have been applied to cases where the output space is finite dimensional (Yousefi et al., 2018; Li et al., 2019; Reeve and Kaban, 2020; Wu et al., 2021). However, as general empirical process theory is developed, to the best of our knowledge, exclusively for classes of real-valued functions, the powerful armoury of empirical process theory has not been utilised fully to analyse vector-valued learning problems. The aim of this paper is to provide some first steps towards developing a theory of empirical processes with vector-valued functions.

An indispensable object in empirical process theory is metric entropy of function classes[1], and one of the most frequently used function classes is that of smooth functions. In our main results in Section 3, we investigate how we can bound the entropy of classes of smooth vector-valued functions. When the output space is infinite-dimensional, bounding the entropy becomes far less trivial, compared to the case of real-valued function classes. For example, seemingly benign function classes such as the classes of constant functions onto the unit ball clearly has infinite entropy with respect to any reasonable metric, since the unit ball in an infinite-dimensional Hilbert space is not totally bounded (Bollobás, 1999, p.62, Corollary 6).

This requires us to look for other ways to restrict the functions than in the norm sense. Our contributions are as follows.

- In the main results of this paper in Section 3, we propose considering subsets of the output space with specific geometric features. We leverage notions from dimension theory of metric spaces (Heinonen et al., 2001; Robinson, 2010; Fraser, 2020). We investigate how restricting our function classes to subsets of the output space in three different ways, to have (i) finite Assouad dimension, (ii) finite upper box-counting dimension and (iii) at most exponentially growing entropy, can help us bound the entropies of the function classes (Theorems 4, 5 and 6 respectively).
- We use these entropy bounds to show uniform law of large numbers and asymptotic equicontinuity of the corresponding function classes (Corollaries 7 and 8).
- In Section 4, we demonstrate applications in statistical learning theory, and discuss the generalisation of the popular Rademacher complexity to the vector-valued setting.

1.1. Mathematical Preliminaries & Notations

Let $\mathcal{Y}$ be a separable Hilbert space over $\mathbb{R}$, with its inner product and norm denoted by $\langle\cdot,\cdot\rangle_{\mathcal{Y}}$ and $\|\cdot\|_{\mathcal{Y}}$ respectively. We denote by $\mathscr{Y}$ the Borel σ-algebra of $\mathcal{Y}$, i.e. the σ-algebra generated by the open subsets of $\mathcal{Y}$. Let $(\mathcal{X}, \mathscr{X})$ be a measurable set, and Q a probability measure on it.

[1] In the usual theory of empirical processes with real-valued functions, there are two major tools. The first is to consider the entropy with respect to the empirical measure P_n. One usually requires this entropy to be uniformly bounded over all realisations of the samples $X_1, ..., X_n$, and the most widely-used example of function classes that satisfy this property are the celebrated *Vapnik-Chervonenkis (VC) subgraph classes*. The second tool is what is known as *entropy with bracketing* with respect to the underlying measure P (see, for example, van de Geer (2000, p.122, Theorem 2.4.1 and p.129, Section 2.5.2), van de Geer (2000, Sections 3.1 and 5.5) and Dudley (2014, Chapter 7)). However, both VC subgraph classes and entropy with bracketing make explicit use of the fact that the output space $\mathbb{R}$ is *totally-ordered*, and makes use of objects such as $\{x \in \mathcal{X} : x \leq g(x_0)\}$ and $\{x \in \mathcal{X} : g_1(x_0) \leq x \leq g_2(x_0)\}$, where $g, g_1, g_2 \in \mathcal{G}$ and $x_0 \in \mathcal{X}$. A direct extension is clearly not possible when our output space $\mathcal{Y}$ has any dimension greater than 1, and an attempt at an extension is even more difficult when $\mathcal{Y}$ is infinite-dimensional. In this paper, we do not investigate whether it is possible to obtain meaningful results by extending these ideas, and leave it for future work.

Bochner Integration A function $g : \mathcal{X} \to \mathcal{Y}$ is said to be *Bochner-integrable* with respect to Q if g is strongly measurable and if $\|g\|_{\mathcal{Y}}$ is Q-integrable (Dinculeanu, 2000, p.15, Definition 35), and denote its Bochner integral by $\int g dQ \in \mathcal{Y}$. We denote the space of Bochner Q-integrable functions by $L^1(\mathcal{X}, Q; \mathcal{Y})$. Further, for $1 \leq p < \infty$, we denote by $L^p(\mathcal{X}, Q; \mathcal{Y})$ the space of functions $g : \mathcal{X} \to \mathcal{Y}$ such that $\int \|g\|_{\mathcal{Y}}^p dQ < \infty$, and denote the corresponding seminorm by $\|g\|_{p,Q}^p = \int \|g\|_{\mathcal{Y}}^p dQ$. The case $p = 2$ is a special case, where $L^2(\mathcal{X}, Q; \mathcal{Y})$ can be equipped with a semi-inner product $\langle g_1, g_2 \rangle_{2,Q} = \int \langle g_1, g_2 \rangle_{\mathcal{Y}} dQ$. Finally, we denote by $L^\infty(\mathcal{X}; \mathcal{Y})$ the space of functions $g : \mathcal{X} \to \mathcal{Y}$ such that the uniform norm $\|g\|_\infty = \sup_{x \in \mathcal{X}} \left\|g(x)\right\|_{\mathcal{Y}}$ is bounded. Following van de Geer (2000, p.16), we do not consider the essential supremum (which depends on the measure Q), but the supremum over *all* $x \in \mathcal{X}$, so that the uniform norm does not depend on any measure.

Taylor's Theorem for Vector-Valued Functions The notions of (partial) differentiation and smoothness for $\mathcal{Y}$-valued functions are central in our bounds for entropy of smooth functions (see Appendix B for more details, and Cartan (1967); Coleman (2012) for full expositions). Suppose that U is an open subset of $\mathbb{R}^d$, and denote the Euclidean norm in $\mathbb{R}^d$ by $\|\cdot\|$. For $m \in \mathbb{N}$ and an m-times differentiable function $g : U \to \mathcal{Y}$, we write $g^{(m)}$ for the m^{th} derivative of g, an m-linear operator from $\mathbb{R}^d$ into $\mathcal{Y}$ (see Appendix B). We state the extension of Taylor's theorem to functions with values in $\mathcal{Y}$, with Lagrange's form of the remainder. To this end, for $a, b \in \mathbb{R}^d$, define the *segment* joining a and b as the set $[a, b] = \{x \in \mathbb{R}^d : x = va + (1 - v)b, v \in [0, 1]\}$ (Coleman, 2012, p.51).

Theorem 1 (Cartan (1967, p.77, Théorème 5.6.2)) *Suppose that $g : U \to \mathcal{Y}$ is $(m + 1)$-times differentiable, that the segment $[a, a + h]$ is contained in U and that, for some $K > 0$, we have $\|g^{(m+1)}(x)\|_{\text{op}} \leq K$ for all $x \in U$. Then*

$$\left\| g(a + h) - \sum_{k=0}^{m} \frac{1}{k!} g^{(k)}(a)((h)^k) \right\|_{\mathcal{Y}} \leq K \frac{\|h\|^{m+1}}{(m + 1)!},$$

where we wrote $(h)^k = (h, ..., h) \in (\mathbb{R}^d)^k$ for $k = 1, ..., m$.

Write $\mathbb{N}_0 = \{0, 1, 2, ...\}$, and for $p = (p_1, ..., p_d) \in \mathbb{N}_0^d$, write $[p] := p_1 + ... + p_d$. Then we denote the p^{th} partial derivative $\partial_1^{p_1} ... \partial_d^{p_d} g(a)$ of g at $a \in U$ as $D^p g(a) \in \mathcal{Y}$. For each $k = 1, ..., m + 1$, $g^{(k)}(a)((h)^k) = \sum_{l_1,...,l_k=1}^{d} h_{l_1} ... h_{l_k} \partial_{l_1} ... \partial_{l_k} g(a) = \sum_{[p]=k} \frac{k! h^p}{p!} D^p g(a)$, where we wrote h^p as a shorthand for $h_1^{p_1} ... h_d^{p_d}$ and $p!$ for $p_1! ... p_d!$. Hence, using partial derivatives, we can express Taylor's theorem above as

$$\left\| g(a + h) - \sum_{[p] \leq m} \frac{h^p}{p!} D^p g(a) \right\|_{\mathcal{Y}} \leq K \frac{\|h\|^{m+1}}{(m + 1)!}.$$

Metric Spaces, Covering Numbers and Dimensions Finally, we introduce some notions from the theory of metric spaces. In particular, covering numbers play a central role in entropy discussions, and different notions of dimensions based on covering numbers will be used to restrict the range of partial derivatives of functions, leading up to entropy bounds in our main results (Section 3).

Suppose $(\mathcal{Z}, \rho)$ is a metric space. For $r > 0$ and $z_0 \in \mathcal{Z}$, the *ball of radius r centred at z_0* is $\mathcal{B}(z_0, r) = \{z \in \mathcal{Z} : \rho(z, z_0) \leq r\}$. For any $\delta > 0$, the *δ-covering number* of $(\mathcal{Z}, \rho)$, denoted by $N(\delta, \mathcal{Z}, \rho)$, is the minimum number of balls of radius δ with centres in $\mathcal{Z}$ required to cover $\mathcal{Z}$, i.e. the minimal N such that there exists a set $\{z_1, ..., z_N\} \subset \mathcal{Z}$ such that for all $z \in \mathcal{Z}$, there exists a $j = j(z) \in \{1, ..., N\}$ with $\rho(z, z_j) \leq \delta$ (we take $N(\delta, \mathcal{Z}, \rho) = \infty$ if no finite covering by closed balls with radius δ exists). We say that $\mathcal{Z}$ is *totally bounded* if $N(\delta, \mathcal{Z}, \rho) < \infty$ for all $\delta > 0$. We define the δ-*entropy* as $H(\delta, \mathcal{Z}, \rho) = \log N(\delta, \mathcal{Z}, \rho)$.

Let E be a subset of $(\mathcal{Z}, \rho)$. The *upper box-counting dimension* of E is

$$\tau_{\text{box}}(E) := \limsup_{\delta \to 0} \frac{H(\delta, E, \rho)}{-\log \delta}$$

(Robinson, 2010, p.32, Definition 3.1). It is immediate from the definition (Robinson, 2010, p.32, (3.3)) that if $\tau > \tau_{\text{box}}(E)$, then there exists $\delta_0 > 0$ such that for all $\delta < \delta_0$,

$$N(\delta, E, \rho) < \delta^{-\tau}. \tag{box}$$

A subset E of $(\mathcal{Z}, \rho)$ is said to be (M, τ)-*homogeneous* (or simply *homogeneous*) if the intersection of E with any closed ball of radius R can be covered by at most $M \left(\frac{R}{r}\right)^{\tau}$ closed balls of smaller radius r, i.e. $N(r, \mathcal{B}(z, R) \cap E, \rho) \leq M \left(\frac{R}{r}\right)^{\tau}$ for all $z \in E$ and $R > r$ (Robinson, 2010, p.83, Definition 9.1). The *Assouad dimension* (Robinson, 2010, p.85, Definition 9.5), sometimes also known as the *doubling dimension*, of E is

$$\tau_{\text{asd}}(E) := \inf\{\tau : E \text{ is } (M, \tau)\text{-homogeneous for some } M \geq 1\}.$$

2. Empirical Process Theory for Functions Taking Values in a Hilbert Space

Take $(\Omega, \mathscr{F}, \mathbb{P})$ as the underlying probability space. Let $X : \Omega \to \mathcal{X}$ be a random variable, and let $X_1, X_2, \ldots$ be i.i.d. copies of X. Denote by P its distribution, i.e. for $A \in \mathscr{X}$, $P(A) = \mathbb{P}(X^{-1}(A))$, and by P_n the empirical measure on $\mathcal{X}$ based on $X_1, \ldots, X_n$, i.e.

$$P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}, \qquad \text{where, for } A \in \mathscr{X}, \delta_{X_i}(A) = \begin{cases} 0 & \text{if } X_i \notin A \\ 1 & \text{if } X_i \in A \end{cases}.$$

For a function $g \in L^1(\mathcal{X}, Q; \mathcal{Y})$, we adopt the notation $Qg = \int g dQ$. Hence,

$$Pg = \int g dP \qquad \text{and} \qquad P_n g = \frac{1}{n} \sum_{i=1}^{n} g(X_i).$$

Note that the integral Pg is a Bochner integral, and that we have $Pg, P_n g \in \mathcal{Y}$. Now, for fixed g, the law of large numbers in Hilbert (more generally, Banach) spaces (Mourier, 1953) tells us that $P_n g$ converges to Pg. One of the pillars of empirical process theory is to consider the convergence of $P_n g$ to Pg not for a fixed g, but uniformly over a class of functions. Let $\mathcal{G} \subset L^1(\mathcal{X}, P; \mathcal{Y})$. For a measure Q on $\mathcal{X}$, we denote $\|Q\|_{\mathcal{G}} := \sup_{g \in \mathcal{G}} \|Qg\|_{\mathcal{Y}}$.

Definition 2 *We say that the class $\mathcal{G}$ is a Glivenko Cantelli (GC) class, or that it satisfies the uniform law of large numbers (with respect to the measure P) if $\|P_n - P\|_{\mathcal{G}} = \sup_{g \in \mathcal{G}} \|P_n g - Pg\|_{\mathcal{Y}} \xrightarrow{P} 0$.*

Definition 2 could have been defined in terms of the weak convergence in Hilbert spaces, i.e. $y_n \to y_0$ if $\langle y, y_n \rangle_{\mathcal{Y}} \to \langle y, y_0 \rangle_{\mathcal{Y}}$ for every $y \in \mathcal{Y}$. In this paper, we only consider strong (norm) convergence. Next, we define the empirical process and the asymptotic equicontinuity.

Definition 3 *We regard $\left\{\nu_n(g) = \sqrt{n}\,(P_n - P)\,g : g \in \mathcal{G}\right\}$ as a stochastic process with values in $\mathcal{Y}$ indexed by $\mathcal{G}$, and call it the empirical process.*

We say that the empirical process $\left\{\nu_n(g) : g \in \mathcal{G}\right\}$ is asymptotically equicontinuous at $g_0 \in \mathcal{G}$ if, for every sequence $\{\hat{g}_n\} \subset \mathcal{G}$ with $\|\hat{g}_n - g_0\|_{2,P} \xrightarrow{P} 0$, we have $\left\|\nu_n\left(\hat{g}_n\right) - \nu_n\left(g_0\right)\right\|_{\mathcal{Y}} \xrightarrow{P} 0$.

Some of the first steps in empirical process theory are the symmetrisation and chaining techniques, and using them to prove uniform law of large numbers and asymptotic equicontinuity for classes of functions that satisfy certain entropy conditions. We provide the adaptation of some of these results for vector-valued function classes but defer them to Appendix C, because, while strictly speaking novel, the statements and proofs of these results carry over from the case of real-valued function classes with only minor adjustments, in particular with concentration inequalities for vector-valued random variables (Pinelis, 1992). A more challenging task, as mentioned in the Introduction, is to bound entropies of vector-valued function classes, and the main results of this paper will focus on this problem (Section 3).

We mention that in this work, we overlook the problem of measurability, which arise as we take suprema over possibly uncountable sets. This is commonly done in works treating statistical applications of empirical processes (see, e.g. van de Geer (2000, p.21, Section 2.5), Bartlett et al. (2005, p.7, first paragraph of Section 2) and Yousefi et al. (2018, p.5, last paragraph of Section 1)). We either assume that function classes and underlying distributions satisfy conditions that ensure measurability, or that notions of outer probabilities and expectations are used instead, as in van der Vaart and Wellner (1996) and Kosorok (2008).

3. Entropy of Classes of Smooth Vector-Valued Functions

In the usual empirical process theory with real-valued functions, classes of smooth functions on compact domains are some of the most frequently used examples that satisfy good entropy conditions (van de Geer, 2000, p.154, Example 9.3.2), (van der Vaart and Wellner, 1996, Section 2.7.1), (Dudley, 2014, Section 8.2). In this section, we give analogues of these results when the output space is the (not necessarily finite-dimensional) Hilbert space $\mathcal{Y}$.

Let $m \in \mathbb{N}$; this will determine the smoothness of our function class. Let $d \geq 1$, and take as our input space the unit cube in $\mathbb{R}^d$, $\mathcal{X} = \{x \in \mathbb{R}^d : 0 \leq x_j \leq 1, j = 1, ..., d\}$; this is only to simplify the exposition, and the subsequent results will clearly hold for any bounded convex subsets of $\mathbb{R}^d$.

In order to bound the entropy of classes of smooth real-valued functions, one bounds the absolute values of the range of the functions and their partial derivatives. When the output space is $\mathcal{Y}$, in particular, if $\mathcal{Y}$ has infinite dimensions, bounding the norm of the range is useless, because balls in infinite-dimensional spaces are not totally bounded. Therefore, to have any hope, the very least we need to do is to find a totally bounded subset $B \subset \mathcal{Y}$, and restrict our range and partial derivatives therein. As B is totally bounded, for some $K_B > 0$, $\|y\|_{\mathcal{Y}} \leq K_B$ for all $y \in B$.

Denote by $\mathcal{G}_B^m$ the set of m-times differentiable functions $g : \mathcal{X} \to \mathcal{Y}$ whose partial derivatives $D^p g : \mathcal{X} \to \mathcal{Y}$ of orders $[p] \leq m$ exist everywhere on the interior of $\mathcal{X}$, and such that $D^p g(x) \in B$ for all $x \in \mathcal{X}$ and $[p] \leq m$, where $D^0 g = g$. We present three results bounding $H(\delta, \mathcal{G}_B^m, \|\cdot\|_\infty)$ for $\delta > 0$ sufficiently small, each with different assumptions on B. Theorem 4 assumes that B is homogeneous, i.e. we impose local entropy conditions. In Theorems 5 and 6, we impose global entropy conditions on B, the former with finite upper box-counting dimension, and the latter with $N(\delta, B, \|\cdot\|_{\mathcal{Y}})$ allowed to grow exponentially as δ decreases. Proofs are deferred to Section 3.2.

Theorem 4 *Let $B \subset \mathcal{Y}$ be totally bounded and (M, τ_{asd})-homogeneous. Then for sufficiently small $\delta > 0$, there exists some constant K depending on K_B, m, d, M and τ_{asd} such that*

$$H\left(\delta, \mathcal{G}_B^m, \|\cdot\|_\infty\right) \leq K\delta^{-\frac{d}{m}}.$$

Theorem 4 gives the same rate for $\mathcal{G}_B^m$ as for smooth real-valued function classes (Dudley, 2014, p.288, Theorem 8.4(a)), which is a special case of the set-up in Theorem 4, since any bounded subset of $\mathbb{R}$ is a homogeneous subset (with Assouad dimension at most 1). In fact, Dudley (2014, Theorem 8.4(a)) shows that this rate of $\delta^{-\frac{d}{m}}$ cannot be improved, so the rate given in Theorem 4 is also optimal. We will later see from the proof that the dependence on τ_{asd} is linear.

Theorem 5 *Let B be a subset of $\mathcal{Y}$ with finite upper box-counting dimension τ_{box}. Then for sufficiently small $\delta > 0$, there exists some constant K depending on K_B, m, d and τ_{box} such that*

$$H\left(\delta, \mathcal{G}_B^m, \|\cdot\|_\infty\right) \leq K\delta^{-\frac{d}{m}} \log\left(\frac{1}{\delta}\right).$$

Theorem 6 *Let B be a subset of $\mathcal{Y}$ with $N(\epsilon, B, \|\cdot\|_\mathcal{Y}) \leq \exp\{M\epsilon^{-\tau_{\text{exp}}}\}$ for some $M, \tau_{\text{exp}} > 0$. Then for sufficiently small $\delta > 0$, there is some constant K depending on K_B, m, d, M and τ_{exp} such that*

$$H\left(\delta, \mathcal{G}_B^m, \|\cdot\|_\infty\right) \leq K\delta^{-\left(\frac{d}{m}+\tau_{\text{exp}}\right)}.$$

We can use results stated and proved in Appendix C to show that we have uniform law of large numbers over $\mathcal{G}_B^m$, where B satisfies the conditions in any one of Theorems 4, 5 or 6.

Corollary 7 *The function class $\mathcal{G}_B^m$, where B is either homogeneous, has finite upper box-counting dimension or satisfies $N(\epsilon, B, \|\cdot\|_\mathcal{Y}) \leq \exp\{M\epsilon^{-\tau_{\text{exp}}}\}$ for some $\tau_{\text{exp}} > 0$, is Glivenko-Cantelli.*

Further, the empirical process defined by $\mathcal{G}_B^m$ (c.f. Definition 3) is asymptotically equicontinuous.

Corollary 8 *Suppose that B is either homogeneous, has finite upper box-counting dimension or satisfies $N(\epsilon, B, \|\cdot\|_\mathcal{Y}) \leq \exp\{M\epsilon^{-\tau_{\text{exp}}}\}$ for some $\tau_{\text{exp}} > 0$. Then the empirical process $\{\nu_n(g) = \sqrt{n}(P_n - P)g : g \in \mathcal{G}_B^m\}$ defined by $\mathcal{G}_B^m$ is asymptotically equicontinuous.*

3.1. Examples

With these results in hand, it is now of interest to investigate which interesting examples of output space $\mathcal{Y}$ and subsets B satisfy the conditions of Theorems 4, 5 and 6.

Example 1 *Suppose that $\mathcal{Y}$ is a finite-dimensional Hilbert space, say with dimension $d_\mathcal{Y}$. Then balls are totally bounded, so we can let B be of the form $B = \{y \in \mathcal{Y} : \|y\|_\mathcal{Y} \leq K\}$ for any $K > 0$. Moreover, subsets of finite-dimensional spaces are homogeneous with Assouad dimension at most $d_\mathcal{Y}$ (Robinson, 2010, p.85, Lemma 9.6(iii)), and so we can apply Theorem 4. The case $\mathcal{Y} = \mathbb{R}$ corresponds to the usual regression with real-valued output. If $\mathcal{Y} = \mathbb{R}^{d_\mathcal{Y}}$, it corresponds to the multi-task learning setting (Evgeniou et al., 2005; Yousefi et al., 2018; Xu et al., 2019).*

A prominent application of vector-valued output spaces will be when we have functional responses; example data sets include speech, diffusion tensor imaging, mass spectrometry and glaucoma (see Morris (2015); Kadri et al. (2016) and references therein). Let $\mathcal{X}'$ be a domain, and $\mathcal{Y} = L^2(\mathcal{X}', P'; \mathbb{R})$ the space of real-valued functions that are square-integrable with respect to some distribution P' on $\mathcal{X}'$. By considering interesting subsets of $\mathcal{Y}$, we can derive bounds on the entropy $H(\delta, \mathcal{G}_B^m, \|\cdot\|_\infty)$ using Theorems 4, 5 and 6. The next 4 examples are considered in this set-up.

Example 2 *Suppose that $\psi_1, ..., \psi_r \in \mathcal{Y}$, and let $B = \{f = \theta_1\psi_1 + ... + \theta_r\psi_r : \theta = (\theta_1, ..., \theta_r)^T \in \mathbb{R}^r, \|f\|_{2,P'} \leq R\}$ Then van de Geer (2000, p.20, Lemma 2.5) tells us that B is homogeneous, and so Theorem 4 applies. This corresponds to the case where the responses are finite-dimensional functions, or adopting the nomenclature of van de Geer (2000, p.152, Example 9.3.1), "linear regressors".*

Example 3 *More generally, function classes with finite Assouad dimensions have been considered in classification problems, and their generalisation properties analysed (Li and Long, 2007; Bshouty et al., 2009). If these functions form the responses of a regression problem, then Theorem 4 can again be applied. Examples of such function classes include halfspaces with respect to the uniform distribution (i.e. where P' is the uniform distribution) (Bshouty et al., 2009, Proposition 6).*

Example 4 *Let $\mathcal{X}'$ be compact in $\mathbb{R}^{d'}$ (in general, $d \neq d'$), and suppose that $B \subset \mathcal{Y}$ consists of smooth functions. More specifically, for some $m' \in \mathbb{N}$ and $M > 0$, let B be the set of all m'-times differentiable functions $f : \mathcal{X}' \to \mathbb{R}$ whose partial derivatives $D^q f : \mathcal{X}' \to \mathbb{R}$ of orders $[q] \leq m'$ exist everywhere on the interior of $\mathcal{X}'$, and such that $|D^q f(x')| \leq M$ for all $x' \in \mathcal{X}'$ and $[q] \leq m'$. Then applying the result for real-valued function classes (Dudley, 2014, p.288, Theorem 8.4) (or Theorem 4 with $\mathcal{Y} = \mathbb{R}$ and B being the ball of radius M), we have $N(\delta, B, \|\cdot\|_\infty) \leq \exp\{K'\delta^{-\frac{d'}{m'}}\}$ for some constant $K' > 0$. This in turn allows us to apply Theorem 6 to bound the entropy of $\mathcal{G}^m_B$ as*

$$H(\delta, \mathcal{G}^m_B, \|\cdot\|_\infty) \leq K\delta^{-\left(\frac{d}{m}+\frac{d'}{m'}\right)}$$

for some constant $K > 0$. So when the output space is itself a class of smooth (real-valued) functions, the smoothness of the two function classes simply add in the negative exponent of δ in the entropy.

Example 5 *Let B be a ball in a reproducing kernel Hilbert space (RKHS) with a $\mathcal{C}^\infty$ Mercer kernel (see Cucker and Smale (2002) for details), then Cucker and Smale (2002, Theorem D) tells us that for some constant K', we have $N(\delta, B, \|\cdot\|_\infty) \leq \exp\{K'\delta^{-\frac{2d}{h}}\}$ for any $h > d$. Then we can again apply Theorem 6 to bound $H(\delta, \mathcal{G}^m_B, \|\cdot\|_\infty)$ by $K\delta^{-(\frac{d}{m}+\frac{2d}{h})}$ for some constant K and any $h > d$.*

3.2. Proofs of the Main Results

We now prove Theorems 4, 5 and 6. The idea is to approximate smooth functions by piecewise polynomials (Kolmogorov, 1955). We start with some development shared by the three Theorems.

Let $g \in \mathcal{G}^m_B$, $x, x+h \in \mathcal{X}$ and $p \in \mathbb{N}^d_0$ with $[p] \leq m-1$. Then $D^p g$ is $(m-[p])$-times differentiable, and $\|(D^p g)^{(m-[p])}(x)\|_{\text{op}} = \|\sum_{[q]\leq m-[p]} \frac{(m-[p])!}{q!} D^{p+q}g(x)\|_{\mathcal{Y}} \leq K_B \sum_{[q]\leq m-[p]} \frac{(m-[p])!}{q!}$. Hence,

$$D^p g(x+h) = \sum_{[q]\leq m-1-[p]} \frac{h^q}{q!} D^{p+q} g(x) + R_p(g, x, h) \qquad (*)$$

by Taylor's Theorem (Theorem 1), where $\|R_p(g,x,h)\|_{\mathcal{Y}} \leq K_B \frac{\|h\|^{m-[p]}}{(m-[p])!} \sum_{[q]\leq m-[p]} \frac{(m-[p])!}{q!}$. So there is a constant $K_1 = K_1(K_B, m, d) \geq 1$ such that, for all $g \in \mathcal{G}^m_B$, $x \in \mathcal{X}$, $x+h \in \mathcal{X}$ and $p \in \mathbb{N}^d_0$ with $[p] \leq m-1$,

$$\left\|R_p(g,x,h)\right\|_{\mathcal{Y}} \leq K_1 \|h\|^{m-[p]}. \qquad (**)$$

Let $\Delta := (\frac{\delta}{4K_1})^{\frac{1}{m}}$, and $x_{(1)}, ..., x_{(L)}$ a $\frac{\Delta}{2}$-net in $\mathcal{X}$, i.e. $\sup_{x\in\mathcal{X}}\{\inf_{1\leq l\leq L}\|x - x_{(l)}\|\} \leq \frac{\Delta}{2}$. By decomposing $\mathcal{X}$ into cubes of side $\left\lceil \frac{d^{1/2}}{\Delta} \right\rceil^{-1}$ and taking the $x_{(l)}$ as the centres thereof, we can take

$$L \leq K_2 \delta^{-\frac{d}{m}} \tag{†}$$

for some constant $K_2 = K_2(d, K_1)$. Now, for each $k = 0, 1, ..., m-1$, define $\delta_k = \frac{\delta}{2\Delta^k e^d}$. We construct a cover of B as follows. First, to ease the notation, write $N_k = N(\frac{1}{2}\delta_k, B, \|\cdot\|_{\mathcal{Y}})$, and find a set $\{a_j^k, j = 1, ..., N_k\} \subset B$ such that $\mathcal{B}(a_j^k, \frac{1}{2}\delta_k)$ cover B. Then define

$$A_1^k = \mathcal{B}(a_1^k, \frac{1}{2}\delta_k), A_2^k = \mathcal{B}(a_2^k, \frac{1}{2}\delta_k)\backslash\mathcal{B}(a_1^k, \frac{1}{2}\delta_k), ..., A_{N_k}^k = \mathcal{B}(a_{N_k}^k, \frac{1}{2}\delta_k)\backslash \cup_{j=1}^{N_k-1} \mathcal{B}(a_j^k, \frac{1}{2}\delta_k).$$

Then $\mathscr{A}_k := \{A_j^k, j = 1, ..., N_k\}$ is a cover of B of cardinality N_k, whose sets A_j^k have diameter at most δ_k and are disjoint. For each $l = 1, ..., L$, $g \in \mathcal{G}_B^m$ and $p \in \mathbb{N}_0^d$ with $[p] \leq m-1$, define $A_{l,p}(g)$ as the unique set in $\mathscr{A}_{[p]}$ such that $D^p g(x_{(l)}) \in A_{l,p}(g)$, and $a_{l,p}(g)$ as the centre of the ball from which $A_{l,p}(g)$ was created, so that $\|a_{l,p}(g) - D^p g(x_{(l)})\|_{\mathcal{Y}} \leq \frac{1}{2}\delta_{[p]}$. Then if $g_1, g_2 \in \mathcal{G}_B^m$ are such that $A_{l,p}(g_1) = A_{l,p}(g_2)$ for all $l = 1, ..., L$ and all $p \in \mathbb{N}_0^d$ with $[p] \leq m-1$, then

$$\|D^p(g_1 - g_2)(x_{(l)})\|_{\mathcal{Y}} \leq \delta_{[p]}, \tag{***}$$

since the diameter of $A_{l,p}(g_1) = A_{l,p}(g_2)$ is at most $\delta_{[p]}$. For each $x \in \mathcal{X}$, take $x_{(l)}$ such that $\|x - x_{(l)}\| \leq \frac{\Delta}{2}$. Then we have, by putting $p = 0$ into (*),

$$\begin{aligned}
&\left\|(g_1 - g_2)(x)\right\|_{\mathcal{Y}} \\
&= \left\| R_0(g_1, x_{(l)}, x - x_{(l)}) - R_0(g_2, x_{(l)}, x - x_{(l)}) + \sum_{[p]\leq m-1} \frac{(x - x_{(l)})^p}{p!} D^p(g_1 - g_2)(x_{(l)}) \right\|_{\mathcal{Y}} \\
&\leq 2K_1\|x - x_{(l)}\|^m + \sum_{[p]\leq m-1} \delta_{[p]} \frac{\|x - x_{(l)}\|^{[p]}}{p!} \qquad \text{by (**) with } p = 0 \text{ and (***)} \\
&\leq 2K_1\Delta^m + \sum_{k=0}^{m-1} \delta_k \Delta^k \left(\sum_{[p]=k} \frac{1}{p!} \right) \leq \frac{\delta}{2} + \left(\max_{k\leq m-1} \delta_k \Delta^k \right) \sum_{k=0}^{m-1} \frac{d^k}{k!} \leq \frac{\delta}{2} + \frac{\delta}{2e^d} e^d = \delta.
\end{aligned}$$

It follows that the δ-covering number $N(\delta, \mathcal{G}_B^m, \|\cdot\|_\infty)$ with respect to the supremum norm is bounded by the number of distinct possibilities for $\{A_{l,p}(g) : l = 1, ..., L, g \in \mathcal{G}_B^m, p \in \mathbb{N}_0^d, [p] \leq m-1\}$.

Proof of Theorem 4 Let $x_{(l)}$ be ordered so that for $1 < l \leq L$, $\|x_{(l')} - x_{(l)}\| \leq \Delta$ for some $l' < l$. For each $l = 1, ..., L$ and $p \in \mathbb{N}_0^d$ with $[p] \leq m-1$, we write $\mathcal{A}_{l,p}$ for the number of possibilities of $A_{l,p}(g)$ for $g \in \mathcal{G}_B^m$, and for each $l = 1, ..., L$, we write $\mathcal{A}_l$ for the number of possibilities of $A_{l,p}(g)$ as $p \in \mathbb{N}_0^d$ varies with $[p] \leq m-1$. For $l = 1$, we have $D^p g(x_{(1)}) \in B$ for each $p \in \mathbb{N}_0^d$ with $[p] \leq m-1$. So

$$\mathcal{A}_{1,p} \leq N_{[p]} = N\left(\frac{1}{4e^d} \delta^{\frac{m-[p]}{m}} (4K_1)^{\frac{[p]}{m}}, B, \|\cdot\|_{\mathcal{Y}} \right) \leq N\left(\frac{\delta}{4e^d}, B, \|\cdot\|_{\mathcal{Y}} \right),$$

where the last upper bound follows since $N(\cdot, B, \|\cdot\|_{\mathcal{Y}})$ is a decreasing function, and we have $K_1 \geq 1$ and $0 < \delta < 1$. This upper bound has no dependence on p. The number of different $p \in \mathbb{N}_0^d$ with

$[p] \le m-1$ is equal to $\binom{m+d-1}{d}$, which is bounded above by m^d, and so $\mathcal{A}_1 \le N(\frac{\delta}{4e^d}, B, \|\cdot\|_{\mathcal{Y}})^{m^d}$. Since $B = B \cap \mathcal{B}(0, K_B)$ is (M, τ_{asd})-homogeneous, $N(\frac{\delta}{4e^d}, B, \|\cdot\|_{\mathcal{Y}}) \le M(\frac{4e^d K_B}{\delta})^{\tau_{\text{asd}}}$, and so

$$\mathcal{A}_1 \le M^{m^d} \left(\frac{4e^d K_B}{\delta} \right)^{\tau_{\text{asd}} m^d} . \tag{$\dagger\dagger$}$$

Now, for $1 < l \le L$, suppose that $A_{l',q}(g)$ is given for all $l' < l$ and all $q \in \mathbb{N}_0^d$ with $[q] \le m-1$. Choose $l' < l$ such that $\|x_{(l')} - x_{(l)}\| \le \Delta$, and write $y_{l,p}(g) := \sum_{[q] \le m-1-[p]} \frac{(x_{(l')} - x_{(l)})^q}{q!} a_{l',p+q}(g)$. Then for any $p \in \mathbb{N}_0^d$ with $[p] \le m-1$, (*) tells us that

$$\begin{aligned}
&\left\| D^p g(x_{(l)}) - y_{l,p}(g) \right\|_{\mathcal{Y}} \\
&= \left\| R_p(g, x_{(l')}, x_{(l)} - x_{(l')}) \right\|_{\mathcal{Y}} + \sum_{[q] \le m-1-[p]} \frac{\|x_{(l')} - x_{(l)}\|^{[q]}}{q!} \left\| D^{p+q} g(x_{(l')}) - a_{l',p+q}(g) \right\|_{\mathcal{Y}} \\
&\le K_1 \Delta^{m-[p]} + \sum_{[q] \le m-1-[p]} \delta_{[p+q]} \frac{\Delta^q}{q!} = K_1 \frac{\Delta^m}{\Delta^{[p]}} + \delta_{[p]} \sum_{k=0}^{m-1-[p]} \delta_k \Delta^k \left(\sum_{[q]=k} \frac{1}{q!} \right) \le \frac{e^d+1}{2} \delta_{[p]}.
\end{aligned}$$

As $a_{l',p+q}(g)$ is given for all $[q] \le m-1-[p]$, $y_{l,p}(g)$ is a fixed point in $\mathcal{Y}$. So $\mathcal{A}_{l,p}$ is bounded by the number of sets in $\mathscr{A}_{[p]}$ that intersect with $B_{l,p}(g) := B \cap \mathcal{B}\left(y_{l,p}(g), \frac{e^d+1}{2}\delta_{[p]}\right)$. Define $\mathscr{A}_{l,p}(g) := \{A \in \mathscr{A}_{[p]} : A \cap B_{l,p}(g) = \emptyset\}$ and $\mathscr{A}'_{l,p}(g) := \{A \in \mathscr{A}_{[p]} : A \cap B_{l,p}(g) \neq \emptyset\}$, so that $\mathscr{A}_{[p]} = \mathscr{A}_{l,p}(g) \cup \mathscr{A}'_{l,p}(g)$, $N_{[p]} = |\mathscr{A}_{[p]}| = |\mathscr{A}_{l,p}(g)| + |\mathscr{A}'_{l,p}(g)|$ and $\mathcal{A}_{l,p} \le |\mathscr{A}'_{l,p}(g)|$. Now, write $B^+_{l,p}(g) := B \cap \mathcal{B}(y_{l,p}(g), \frac{e^d+3}{2}\delta_{[p]})$. Then we have $A \subset B^+_{l,p}(g)$ for all $A \in \mathscr{A}'_{l,p}(g)$. Let $\mathscr{A}^+_{l,p}(g)$ be a $\frac{1}{2}\delta_{[p]}$-cover of $B^+_{l,p}(g)$ with minimal cardinality $N(\frac{1}{2}\delta_{[p]}, B^+_{l,p}(g), \|\cdot\|_{\mathcal{Y}})$. Since B is (M, τ_{asd})-homogeneous, $N(\frac{1}{2}\delta_{[p]}, B^+_{l,p}(g), \|\cdot\|_{\mathcal{Y}}) \le M(e^d+3)^{\tau_{\text{asd}}}$. By taking the union $\mathscr{A}^+_{l,p}(g)$ with $\mathscr{A}_{l,p}(g)$, we have a $\frac{1}{2}\delta_{[p]}$-cover of B with cardinality at most $|\mathscr{A}_{l,p}(g)| + M(e^d+3)^{\tau_{\text{asd}}}$. So if $|\mathscr{A}'_{l,p}(g)| > M(e^d+3)^{\tau_{\text{asd}}}$, then we have found a $\frac{1}{2}\delta_{[p]}$-cover of B with cardinality strictly less than $N_{[p]}$, contradicting its minimality. Hence, we must have $\mathcal{A}_{l,p} \le |\mathscr{A}'_{l,p}(g)| \le M(e^d+3)^{\tau_{\text{asd}}}$. But the latter quantity is a constant that does not depend on δ or p. Thus

$$\mathcal{A}_l \le \prod_{[p] \le m-1} \mathcal{A}_{l,p} \le M^{m^d} \left(e^d + 3 \right)^{\tau_{\text{asd}} m^d} . \tag{$\dagger\dagger\dagger$}$$

Putting together (†), (††) and († † †), we arrive at

$$N\left(\delta, \mathcal{G}^m_B, \|\cdot\|_\infty\right) \le \prod_{l=1}^{L} \mathcal{A}_l \le M^{m^d} \left(\frac{4e^d K_B}{\delta} \right)^{\tau_{\text{asd}} m^d} M^{m^d K_2 \delta^{-\frac{d}{m}}} \left(e^d + 3 \right)^{\tau_{\text{asd}} m^d K_2 \delta^{-\frac{d}{m}}},$$

and so

$$H\left(\delta, \mathcal{G}^m_B, \|\cdot\|_\infty\right) \le \delta^{-\frac{d}{m}} \log\left(M^{m^d K_2} \left(e^d + 3 \right)^{\tau_{\text{asd}} m^d K_2} \right) + m^d \log\left(M \left(\frac{4e^d K_B}{\delta} \right)^{\tau_{\text{asd}}} \right)$$

$$\leq K\delta^{-\frac{d}{m}},$$

where K is a constant depending on $M, m, d, K_2, \tau_{\text{asd}}$ and K_B. With the second term, we bounded $\log\left(\frac{1}{\delta}\right)$ by a constant times $\delta^{-\frac{d}{m}}$. ■

Proof of Theorem 5 Suppose $g \in \mathcal{G}_B^m$. With notation as in the proof of Theorem 4, for each $l = 1, ..., L$ and each $p \in \mathbb{N}_0^d$ with $[p] \leq m-1$, we have

$$\mathcal{A}_{l,p} \leq N_{[p]} = N\left(\frac{\delta}{2\Delta^{[p]}e^d}, B, \|\cdot\|_{\mathcal{Y}}\right) \leq N\left(\frac{\delta}{2e^d}, B, \|\cdot\|_{\mathcal{Y}}\right) \leq \left(\frac{\delta}{2e^d}\right)^{-(\tau_{\text{box}}+1)},$$

where the second last upper bound follows since $N(\cdot, B, \|\cdot\|_{\mathcal{Y}})$ is a decreasing function, and we have $K_1 \geq 1$ and $0 < \delta < 1$, and the last upper bound follows from Equation (box) in Section 1.1. This upper bound has no dependence on l or p. So for each $l = 1, ..., L$, $\mathcal{A}_l \leq \left(\frac{2e^d}{\delta}\right)^{(\tau_{\text{box}}+1)m^d}$. Putting this together with (†), we arrive at

$$N\left(\delta, \mathcal{G}_B^m, \|\cdot\|_\infty\right) \leq \prod_{l=1}^L \mathcal{A}_l \leq \left(\frac{2e^d}{\delta}\right)^{(\tau_{\text{box}}+1)m^d K_2 \delta^{-\frac{d}{m}}},$$

and so

$$H\left(\delta, \mathcal{G}_B^m, \|\cdot\|_\infty\right) \leq (\tau_{\text{box}}+1)m^d K_2 \delta^{-\frac{d}{m}} \log\left(\frac{2e^d}{\delta}\right) \leq K\delta^{-\frac{d}{m}} \log\left(\frac{1}{\delta}\right),$$

where K is a constant depending on m, d, K_2 and τ_{box}. ■

Proof of Theorem 6 Suppose $g \in \mathcal{G}_B^m$. With notation as in the proof of Theorem 4, for each $l = 1, ..., L$ and each $p \in \mathbb{N}_0^d$ with $[p] \leq m-1$, we have

$$\mathcal{A}_{l,p} \leq N_{[p]} = N\left(\frac{\delta}{2\Delta^{[p]}e^d}, B, \|\cdot\|_{\mathcal{Y}}\right) \leq N\left(\frac{\delta}{2e^d}, B, \|\cdot\|_{\mathcal{Y}}\right) \leq \exp\left\{M\left(\frac{\delta}{2e^d}\right)^{-\tau_{\text{exp}}}\right\},$$

where the second last upper bound follows since $N(\cdot, B, \|\cdot\|_{\mathcal{Y}})$ is a decreasing function, and we have $K_1 \geq 1$ and $0 < \delta < 1$. This upper bound has no dependence on l or p. So for each $l = 1, ..., L$,

$$\mathcal{A}_l \leq \exp\left\{M\left(\frac{\delta}{2e^d}\right)^{-\tau_{\text{exp}}} m^d\right\}.$$

Putting this together with (†), we arrive at

$$N\left(\delta, \mathcal{G}_B^m, \|\cdot\|_\infty\right) \leq \prod_{l=1}^L \mathcal{A}_l \leq \exp\left\{M\left(\frac{\delta}{2e^d}\right)^{-\tau_{\text{exp}}} m^d K_2 \delta^{-\frac{d}{m}}\right\},$$

and so

$$H\left(\delta, \mathcal{G}_B^m, \|\cdot\|_\infty\right) \leq M\left(\frac{1}{2e^d}\right)^{-\tau_{\text{exp}}} m^d K_2 \delta^{-\frac{d}{m}-\tau_{\text{exp}}} \leq K\delta^{-\left(\frac{d}{m}+\tau_{\text{exp}}\right)},$$

where K is a constant depending on m, d, M, K_2 and τ_{exp}. ■

4. Applications to Statistical Learning Theory

In this Section, we discuss the application of the above main results to statistical learning theory.

4.1. Least-Squares Regression with Fixed Design

We first consider problem of least squares regression with fixed design, whereby the covariates $x_1, ..., x_n \in \mathcal{X}$ are considered fixed. Let $Y_1, ..., Y_n$ be random variables taking values in $\mathcal{Y}$ satisfying

$$Y_i = g_0(x_i) + \varepsilon_i, \qquad i = 1, ..., n,$$

where ε_i are independent (Hilbert space) Gaussian noise terms with zero mean and covariance with trace 1 (see Section A.2 for details), and g_0 is the unknown regression function in a given class $\mathcal{G}$ of functions $\mathcal{X} \to \mathcal{Y}$. We assume that the following least squares estimator exists:

$$\hat{g}_n := \arg\min_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n \left\|Y_i - g(x_i)\right\|_{\mathcal{Y}}^2,$$

and we are interested in the convergence of $\|\hat{g}_n - g_0\|_{2,P_n}$ to 0. Theorem 27 in Appendix C.4, whose proof is based on the "peeling device" (van de Geer, 2000) and concentration of Gaussian measures in Hilbert spaces as discussed in Section A, tells us that $\|\hat{g}_n - g_0\|_{2,P_n} = \mathcal{O}_P(\delta_n)$, with δ_n satisfying

$$\sqrt{n}\delta_n^2 \geq 8\left(J(\delta_n) + 4\delta_n\sqrt{1+t} + \delta_n\sqrt{8t/3}\right),$$

where $J(\delta) := 4\int_0^\delta \sqrt{2H(u, \mathcal{B}_{2,P_n}(\delta), \|\cdot\|_{2,P_n})}du$ and $\mathcal{B}_{2,P_n}(\delta) := \{g \in \mathcal{G} : \|g\|_{2,P_n} \leq \delta\}$.

As an example, let us return to the setting of Example 4, where $\mathcal{Y} = L^2(\mathcal{X}', P'; \mathbb{R})$, and $B \subset \mathcal{Y}$ is a class of m'-times differentiable functions. We saw that $H(\delta, \mathcal{G}_B^m, \|\cdot\|_\infty) \leq K\delta^{-(\frac{d}{m}+\frac{d'}{m'})}$ by Theorem 6. Thus, for another constant $K' > 0$, $J(\delta) \leq K'\delta^{1-\frac{1}{2}(\frac{d}{m}+\frac{d'}{m'})}$, and it can be shown that

$$\|\hat{g}_n - g_0\|_{2,P_n} = \mathcal{O}_P(n^{-1/(2+\frac{d}{m}+\frac{d'}{m'})}).$$

For smooth real-valued function classes, the rate is $n^{-1/(2+\frac{d}{m})}$ (Tsybakov, 2008, p.40, Theorem 1.6), so we can see that the terms $\frac{d}{m}$ and $\frac{d'}{m'}$ that correspond to the smoothness of $\mathcal{G}_B^m$ and B simply add up in the exponent. Note that as $m \to \infty$ and $m' \to \infty$, we have $\|\hat{g}_n - g_0\|_{2,P_n} = \mathcal{O}_P(n^{-\frac{1}{2}})$.

4.2. Empirical Risk Minimisation with Bounded Lipschitz Loss with Random Design

We discuss the random design setting with L-bounded c-Lipschitz loss $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. The population and empirical risks for $g \in \mathcal{G}$ are given by $\mathcal{R}(g) = \mathbb{E}[\mathcal{L}(Y, g(X))]$ and $\hat{\mathcal{R}}_n(g) = \frac{1}{n}\sum_{i=1}^n \mathcal{L}(Y_i, g(X_i))$ respectively. We assume that the minimiser $\hat{g}_n = \arg\min_{g\in\mathcal{G}} \hat{\mathcal{R}}_n(g)$ exists, and are interested in the convergence of $\mathcal{R}(\hat{g}_n)$ to $\mathcal{R}(g^*)$. Writing $\mathcal{J}(1) := 4\int_0^1 \sqrt{2H(u)}du$, where $H(u)$ is a function such that, for all $u > 0$ and any probability distribution Q with finite support, $H(uL, \mathcal{L} \circ \mathcal{G}, \|\cdot\|_{2,Q}) \leq H(u)$, Theorem 28 and Hoeffding's inequality (Prop. 10) give:

$$\mathbb{P}\left(\mathcal{R}(\hat{g}_n) - \mathcal{R}(g^*) > \frac{4\sqrt{2}L\mathcal{J}(1)}{\sqrt{n}} + 24L\sqrt{\frac{1+t}{n}} + \frac{L}{\sqrt{n}} + L\sqrt{\frac{2t}{n}}\right) \leq 3e^{-t}.$$

Returning to the setting of Example 4, the entropy contraction property (Lemma 33) gives

$$H(\delta,\mathcal{L}\circ\mathcal{G}_B^m,\|\cdot\|_{2,P_n})\leq H(\delta,\mathcal{L}\circ\mathcal{G}_B^m,\|\cdot\|_\infty)\leq H(\frac{1}{c}\delta,\mathcal{G}_B^m,\|\cdot\|_\infty)\leq K\delta^{-\left(\frac{d}{m}+\frac{d'}{m'}\right)}$$

for some constant K, and so as long as $\frac{d}{m}+\frac{d'}{m'}<2$, we do indeed have $\mathcal{J}(1)<\infty$.

4.3. Discussion on Rademacher Complexity for Vector-Valued Function Classes

As well as metric entropy, another common measure of complexity of function classes is the Rademacher complexity, the empirical version of which, for real-valued function classes $\mathcal{F}$, is $\hat{\mathfrak{R}}_n(\mathcal{F})=\mathbb{E}[\sup_{f\in\mathcal{F}}|\frac{1}{n}\sum_{i=1}^n\sigma_i f(X_i)|\mid X_1,...,X_n]$, where σ_i are independent Rademacher variables (Bartlett and Mendelson, 2002, Definition 2). In this Section, we briefly discuss Rademacher complexities for vector-valued function classes, and due to space constraints, defer a fuller discussion to Appendix C.6. Define the "Rademacher complexity" of a class $\mathcal{G}$ of vector-valued functions as

$$\hat{\mathfrak{R}}_n(\mathcal{G})=\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\|\frac{1}{n}\sum_{i=1}^n\sigma_i g(X_i)\right\|_{\mathcal{Y}}\mid X_1,...,X_n\right].$$

Indeed, in Section C.1, we use the symmetrised empirical measure $\frac{1}{n}\sum_{i=1}^n\sigma_i\delta_{X_i}$, which suggests the use of the above definition. However, there is a critical issue with this definition. Rademacher complexities are almost always used in conjunction with a loss function, i.e. what we end up using is the Rademacher complexity of the class $\mathcal{L}\circ\mathcal{G}$ (c.f. Section 4.2). With real-valued function classes $\mathcal{F}$, Ledoux and Talagrand (1991, p.112, Theorem 4.12) shows that for bounded Lipschitz losses $\mathcal{L}$, we have $\hat{\mathfrak{R}}_n(\mathcal{L}\circ\mathcal{F})\leq K\hat{\mathfrak{R}}_n(\mathcal{F})$ for a constant K, so it is meaningful to work with $\hat{\mathfrak{R}}_n(\mathcal{F})$. However, the proof makes use of the fact that the output space is $\mathbb{R}$, and Maurer (2016, Section 6) shows via a counterexample that contraction no longer holds for the above definition of $\hat{\mathfrak{R}}_n(\mathcal{G})$. Maurer (2016) in fact shows a contraction result for what we call the *coordinate-wise Rademacher complexity*:

$$\hat{\mathfrak{R}}_n^{\text{coord}}(\mathcal{G})=\mathbb{E}\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^n\sum_k\sigma_i^k g_k(X_i)\mid X_1,...,X_n\right],$$

where a particular basis of $\mathcal{Y}$ is fixed, k is the index on the coordinates of $\mathcal{Y}$ with respect to this basis and g_k are real-valued functions that map to each coordinate of g. Notice that in this case, we need a separate Rademacher variable for each coordinate, as well as for each sample. This has been used for finite-dimensional multi-task learning (Yousefi et al., 2018; Li et al., 2019). While we recognise its usefulness, the coordinate-wise Rademacher complexity, by definition, relies on a choice of basis of $\mathcal{Y}$, and we show in Appendix C.6 that $\hat{\mathfrak{R}}_n^{\text{coord}}$ is actually not independent of the choice of basis. We regard this as a critical issue in using $\hat{\mathfrak{R}}_n^{\text{coord}}$ as a "complexity measure of a function class", since it is intuitively clear that the complexity should not depend on the choice of basis of the output space.

A common way to bound the Rademacher complexity is to use Dudley's chaining and uniform entropy condition, in precisely the same manner as in Section C.3. In this case, we propose a workaround that avoids using either $\hat{\mathfrak{R}}_n(\mathcal{G})$ or $\hat{\mathfrak{R}}_n^{\text{coord}}(\mathcal{G})$. For a bounded Lipschitz loss function $\mathcal{L}$, $\hat{\mathfrak{R}}_n(\mathcal{L}\circ\mathcal{G})$ can be bounded by an expression involving the integral (with respect to δ) of the entropy $H(\delta,\mathcal{L}\circ\mathcal{G},\|\cdot\|_{2,P_n})$ (this is a standard result; see, for example, Shalev-Shwartz and Ben-David (2014, p.338, Lemma 27.4); we show a vector-valued analogue for $\hat{\mathfrak{R}}_n(\mathcal{G})$ in Theorem 31, using vector-valued Hoeffding-type inequality). But as discussed in Section 4.2, the entropies satisfy a simple contraction property given in Lemma 33. So applying the same argument, we can bound $\hat{\mathfrak{R}}_n(\mathcal{L} \circ \mathcal{G})$ by an expression involving $H(\delta, \mathcal{G}, \|\cdot\|_\infty)$, which has been the main topic of this paper. This does not contradict the counterexample of Maurer (2016, Section 6), since the latter is the space of linear operators between infinite-dimensional Hilbert spaces, and hence has infinite entropy.

5. Discussion & Future Directions

To summarize, we took some first steps towards establishing a theory of empirical processes for vector-valued functions. In particular, we investigated the metric entropy of smooth functions, by restricting the partial derivatives to take values in totally bounded subsets with specific properties, leveraging theory from fractal geometry, and demonstrated its application in empirical risk minimisation.

There is a plethora of possible future research directions. Considering other classes of functions than those of smooth functions is a natural next step. Also, we let $\mathcal{Y}$ be a Hilbert space, primarily because some simplifications occur for Hoeffding's inequality and Gaussian measures (Appendix A), but extensions to Banach spaces should be possible. Moreover, we used compact subsets of $\mathbb{R}^d$ as our input space due to the ease in considering partial derivatives, but interesting applications exist for which the input space $\mathcal{X}$ is a subset of an infinite-dimensional space (Li et al., 2020; Nelsen and Stuart, 2021; Lu et al., 2021). On the more theoretical side, measurability questions for empirical processes and uniform central limit theorems involving Gaussian elements in vector spaces are interesting questions. Also, obtaining complementary lower bounds, so that our upper bounds are minimax optimal, is an interesting problem. With empirical risk minimisation, extensions to more general noise with vector-valued Bernstein's inequality or misspecified models are important.

Acknowledgments

We thank Shashank Singh and Simon Buchholz at the Max Planck Institute for Intelligent Systems, Tübingen for their helpful comments to the initial drafts.

This work was inspired by the lecture course “Empirical Process Theory and Applications” given by Professor Sara van de Geer at Seminar für Statistik, ETH Zürich. JP is extremely grateful to her for the lectures, as well as readily engaging in post-lecture discussions.

References

Tamim El Ahmad, Pierre Laforgue, and Florence d'Alché Buc. p-Sparsified Sketches for Fast Multiple Output Kernel Methods. *arXiv preprint arXiv:2206.03827*, 2022.

Mauricio A Álvarez, Lorenzo Rosasco, Neil D Lawrence, et al. Kernels for Vector-Valued Functions: A Review. *Foundations and Trends® in Machine Learning*, 4(3):195–266, 2012.

Peter L Bartlett and Shahar Mendelson. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

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

Albert Turner Bharucha-Reid. *Random Integral Equations*. Academic Press, 1972.

Béla Bollobás. *Linear Analysis: An Introductory Course*. Cambridge University Press, 1999.

Romain Brault. *Large-Scale Operator-Valued Kernel Regression*. PhD thesis, Université Paris Saclay, 2017.

Nader H Bshouty, Yi Li, and Philip M Long. Using the Doubling Dimension to Analyze the Generalization of Learning Algorithms. *Journal of Computer and System Sciences*, 75(6):323–335, 2009.

Vivien A. Cabannes, Francis R. Bach, and Alessandro Rudi. Fast Rates for Structured Prediction. In *Conference on Learning Theory, COLT 2021*, pages 823–865, 2021.

Andrea Caponnetto and Ernesto De Vito. Optimal Rates for the Regularized Least-Squares Algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2006.

Henri Cartan. *Calcul Différentiel*. Hermann, 1967.

Carlo Ciliberto, Lorenzo Rosasco, and Alessandro Rudi. A General Framework for Consistent Structured Prediction with Implicit Loss Embeddings. *J. Mach. Learn. Res.*, 21(98):1–67, 2020.

Erhan Çınlar. *Probability and Stochastics*, volume 261. Springer Science & Business Media, 2011.

Dan Tsir Cohen and Aryeh Kontorovich. Metric-Valued Regression. *arXiv preprint arXiv:2202.03045*, 2022.

Rodney Coleman. *Calculus on Normed Vector Spaces*. Springer Science & Business Media, 2012.

Corinna Cortes, Vitaly Kuznetsov, Mehryar Mohri, and Scott Yang. Structured Prediction Theory based on Factor Graph Complexity. *Advances in Neural Information Processing Systems*, 29: 2514–2522, 2016.

Felipe Cucker and Steve Smale. On the Mathematical Foundations of Learning. *Bulletin of the American mathematical society*, 39(1):1–49, 2002.

Giuseppe Da Prato and Jerzy Zabczyk. *Stochastic Equations in Infinite Dimensions*. Cambridge University Press, 2014.

Nicolae Dinculeanu. *Vector Integration and Stochastic Integration in Banach Spaces*, volume 48. John Wiley & Sons, 2000.

Richard M Dudley. *Uniform Central Limit Theorems*, volume 142. Cambridge university press, 2014.

Theodoros Evgeniou, Charles A Micchelli, Massimiliano Pontil, and John Shawe-Taylor. Learning Multiple Tasks with Kernel Methods. *Journal of machine learning research*, 6(4), 2005.

Dylan J Foster and Alexander Rakhlin. l_∞ Vector Contraction for Rademacher Complexity. *arXiv preprint arXiv:1911.06468*, 6, 2019.

Jonathan M Fraser. *Assouad Dimension and Fractal Geometry*, volume 222. Cambridge University Press, 2020.

Steffen Grünewälder, Guy Lever, Luca Baldassarre, Sam Patterson, Arthur Gretton, and Massimilano Pontil. Conditional Mean Embeddings as Regressors. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1803–1810, 2012.

László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer Science & Business Media, 2006.

Steve Hanneke, Aryeh Kontorovich, Sivan Sabato, and Roi Weiss. Universal Bayes Consistency in Metric Spaces. In *2020 Information Theory and Applications Workshop (ITA)*, pages 1–33. IEEE, 2020.

Juha Heinonen et al. *Lectures on Analysis on Metric Spaces*. Springer Science & Business Media, 2001.

Wassily Hoeffding. Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Hachem Kadri, Emmanuel Duflos, Philippe Preux, Stéphane Canu, Alain Rakotomamonjy, and Julien Audiffren. Operator-Valued Kernels for Learning from Functional Response Data. *The Journal of Machine Learning Research*, 17(1):613–666, 2016.

AN Kolmogorov. Bounds for the Minimal Number of Elements of an ε-net in Various Classes of Functions and Their Applications to the Question of Representability of Functions of Several Variables by Superpositions of Functions of Fewer Variables. *Uspekhi Mat. Nauk (NS)*, 10: 192–194, 1955.

Michael R Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Springer, 2008.

Pierre Laforgue, Alex Lambert, Luc Brogat-Motte, and Florence d'Alché Buc. Duality in RKHSs with Infinite Dimensional Outputs: Application to Robust Losses. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Alex Lambert, Dimitri Bouche, Zoltan Szabo, and Florence d'Alché Buc. Functional Output Regression with Infimal Convolution: Exploring the Huber and ε-Insensitive Losses. In *International Conference on Machine Learning*, pages 11844–11867. PMLR, 2022.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*, volume 23. Springer Science & Business Media, 1991.

Jian Li, Yong Liu, and Weiping Wang. Learning Vector-Valued Functions with Local Rademacher Complexity. *arXiv preprint arXiv:1909.04883*, 2019.

Yi Li and Philip M Long. Learnability and the Doubling Dimension. *Advances in neural information processing systems*, 19:889, 2007.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural Operator: Graph Kernel Network for Partial Differential Equations. *arXiv preprint arXiv:2003.03485*, 2020.

Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning Nonlinear Operators via DeepONet based on the Universal Approximation Theorem of Operators. *Nature Machine Intelligence*, 3(3):218–229, 2021.

Andreas Maurer. A Vector-Contraction Inequality for Rademacher Complexities. In *International Conference on Algorithmic Learning Theory*, pages 3–17. Springer, 2016.

Ron Meir and Tong Zhang. Generalization Error Bounds for Bayesian Mixture Algorithms. *Journal of Machine Learning Research*, 4(Oct):839–860, 2003.

Charles A Micchelli and Massimiliano Pontil. On Learning Vector-Valued Functions. *Neural computation*, 17(1):177–204, 2005.

Jeffrey S Morris. Functional Regression. *Annual Review of Statistics and Its Application*, 2:321–359, 2015.

Edith Mourier. Eléments Aléatoires dans un Espace de Banach. In *Annales de l'institut Henri Poincaré*, volume 13, pages 161–244, 1953.

Nicholas H Nelsen and Andrew M Stuart. The Random Feature Model for Input-Output Maps between Banach Spaces. *SIAM Journal on Scientific Computing*, 43(5):A3212–A3243, 2021.

Junhyung Park and Krikamol Muandet. A Measure-Theoretic Approach to Kernel Conditional Mean Embeddings. *Advances in Neural Information Processing Systems*, 33:21247–21259, 2020a.

Junhyunng Park and Krikamol Muandet. Regularised Least-Squares Regression with Infinite-Dimensional Output Space. *arXiv preprint arXiv:2010.10973*, 2020b.

Iosif Pinelis. An Approach to Inequalities for the Distributions of Infinite-Dimensional Martingales. In *Probability in Banach Spaces, 8: Proceedings of the Eighth International Conference*, pages 128–134. Springer, 1992.

Henry Reeve and Ata Kaban. Optimistic Bounds for Multi-Output Learning. In *International Conference on Machine Learning*, pages 8030–8040. PMLR, 2020.

James C Robinson. *Dimensions, Embeddings, and Attractors*, volume 186. Cambridge University Press, 2010.

Lorenzo Rosasco, Mikhail Belkin, and Ernesto De Vito. On Learning with Integral Operators. *Journal of Machine Learning Research*, 11(2), 2010.

Akash Saha and Balamurugan Palaniappan. Learning with Operator-Valued Kernels in Reproducing Kernel Krein Spaces. *Advances in Neural Information Processing Systems*, 33, 2020.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge university press, 2014.

Galen R Shorack and Jon A Wellner. *Empirical Processes with Applications to Statistics*. SIAM, 2009.

Rahul Singh, Maneesh Sahani, and Arthur Gretton. Kernel instrumental variable regression. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer Science & Business Media, 2008.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 2008. ISBN 0387790519.

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

Aad W van der Vaart and Jon Wellner. *Weak Convergence and Empirical Processes: with Applications to Statistics*. Springer Science & Business Media, 1996.

Liang Wu, Antoine Ledent, Yunwen Lei, and Marius Kloft. Fine-Grained Generalization Analysis of Vector-Valued Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10338–10346, 2021.

Donna Xu, Yaxin Shi, Ivor W Tsang, Yew-Soon Ong, Chen Gong, and Xiaobo Shen. Survey on Multi-Output Learning. *IEEE transactions on neural networks and learning systems*, 31(7): 2409–2429, 2019.

Niloofar Yousefi, Yunwen Lei, Marius Kloft, Mansooreh Mollaghasemi, and Georgios C Anagnostopoulos. Local Rademacher Complexity-based Learning Guarantees for Multi-Task Learning. *The Journal of Machine Learning Research*, 19(1):1385–1431, 2018.

Oscar Zatarain-Vera. A Vector-Contraction Inequality for Rademacher Complexities Using p-Stable Variables. *arXiv preprint arXiv:1912.10136*, 2019.

The appendix is structured as follows. In empirical process theory, concentration inequalities are an essential tool, and in Appendix A, we state and prove concentration inequalities in Hilbert spaces, in particular, the extensions of Hoeffding's inequality and Gaussian measures to Hilbert spaces. In Appendix B, we develop the theory of differential calculus between Banach spaces, which plays a vital role in our paper in considering smooth functions. In Appendix C, we develop the theory of empirical process theory for vector-valued functions, briefly introduced in Section 2, in more detail. In particular, we develop the symmetrisation technique (Appendix C.1); we establish the uniform law of large numbers for vector-valued function classes with bounded entropy (Appendix C.2); we develop the chaining technique for vector-valued functions and use it to establish asymptotic equicontinuity of empirical processes satisfying uniform entropy condition (Appendix C.3); we use the chaining technique in tandem with the peeling device for least-squares regression, as briefly discussed in Section 4.1 (Appendix C.4); and finally, we discuss in full connections with the popular Rademacher complexity, as briefly touched upon in Section 4.3 (Appendix C.6).

Appendix A. Concentration Inequalities in Hilbert Spaces

First, we state Markov's inequality, on which all subsequent results are based.

Proposition 9 (Markov's inequality) *For any non-negative real random variable Z and $a > 0$,* $\mathbb{P}\left(Z \geq a\right) \leq \frac{\mathbb{E}[Z]}{a}$.

A.1. Hoeffding's inequality

Hoeffding's inequality is a concentration result for sums of bounded random variables. We first state the real version, due to Hoeffding (1963).

Proposition 10 (Hoeffding's inequality) *Let $Z_1, ..., Z_n$ be independent real random variables such that for all i, $\mathbb{E}[Z_i] = 0$ and $|Z_i| \leq c_i$ almost surely for some constants $c_i > 0$. Then writing $S_n = \sum_{i=1}^n Z_i$ and $b^2 = \sum_{i=1}^n c_i^2$,*

$$\mathbb{P}\left(S_n \geq a\right) \leq e^{-\frac{a^2}{2b^2}}$$

for any $a > 0$, or reformulated, for any $t > 0$,

$$\mathbb{P}\left(S_n \geq b\sqrt{2t}\right) \leq e^{-t}.$$

In Pinelis (1992), Hoeffding's inequality was extended to martingales in Banach spaces with certain smoothness properties (see also Rosasco et al. (2010, Eqn. (3)) and Steinwart and Christmann (2008, p.217, Corollary 6.15)). As we only require the result for sums of independent random variables taking values in a separable Hilbert space $\mathcal{Y}$, we give the corresponding simplified statement First, we state its expectation form, then the probability inequality is stated separately.

Proposition 11 *Let $Y_1, ..., Y_n$ be independent random variables in $\mathcal{Y}$, such that for all i, $\mathbb{E}[Y_i] = 0$ and $\|Y_i\|_{\mathcal{Y}} \leq c_i$ almost surely for some constants $c_i > 0$. Then writing $S_n = \sum_{i=1}^n Y_i$, for any $\lambda > 0$, we have*

$$\mathbb{E}\left[\cosh\left(\lambda \left\|S_n\right\|_{\mathcal{Y}}\right)\right] \leq \prod_{i=1}^n e^{\lambda^2 c_i^2}.$$

Proposition 12 (Hoeffding's inequality in Hilbert spaces) *Let $Y_1, ..., Y_n$ be independent random variables in $\mathcal{Y}$, such that for all i, $\mathbb{E}[Y_i] = 0$ and $\|Y_i\|_{\mathcal{Y}} \leq c_i$ almost surely for some constants $c_i > 0$. Then writing $S_n = \sum_{i=1}^n Y_i$ and $b^2 = \sum_{i=1}^n c_i^2$,*

$$\mathbb{P}\left(\|S_n\|_{\mathcal{Y}} \geq a\right) \leq 2e^{-\frac{a^2}{4b^2}}$$

for any $a > 0$, or reformulated, for any $t > 0$,

$$\mathbb{P}\left(\|S_n\|_{\mathcal{Y}} \geq 2b\sqrt{t}\right) \leq 2e^{-t}.$$

A.2. Gaussian Measures in Hilbert Spaces

Next, we consider concentration of the Gaussian measure. In the real case, the Gaussian measure with mean μ and variance q is defined as the measure that is absolutely continuous with respect to the Lebesgue measure and has density $\frac{1}{\sqrt{2\pi q}}e^{-\frac{1}{2q}(x-\mu)^2}$. The Gaussian measure with mean 0 and variance 1 is called the standard Gaussian measure. For a real variable with the standard Gaussian distribution, the following concentration inequality can easily be derived.

Lemma 13 *Let Z have the standard Gaussian distribution. Then for any $a > 0$, $\mathbb{P}(Z \geq a) \leq e^{-\frac{1}{2}a^2}$.*

The definition of Gaussian measures can be extended to the separable Hilbert space $\mathcal{Y}$.

Definition 14 *Da Prato and Zabczyk (2014, pp.46-47)] A random variable Y in $\mathcal{Y}$ is Gaussian if, for any $y \in \mathcal{Y}$, $\langle Y, y\rangle_{\mathcal{Y}}$ is a real Gaussian random variable (with some mean and variance).*

The next two lemmas are concerned with the *mean* and *covariance operator* of a $\mathcal{Y}$-valued Gaussian random variable. The proofs are given in Da Prato and Zabczyk (2014, Section 2.3).

Lemma 15 *If Y is a Gaussian random variable in $\mathcal{Y}$, then $\mathbb{E}[\|Y\|_{\mathcal{Y}}^2] < \infty$. As a consequence, Y is Bochner integrable, and we call $\mu = \mathbb{E}[Y] \in \mathcal{Y}$ the* mean *of Y.*

Denote by $\mathscr{L}(\mathcal{Y})$ the Banach space of continuous linear operators from $\mathcal{Y}$ into itself, with the operator norm.

Lemma 16 *For a $\mathcal{Y}$-valued Gaussian variable Y with mean μ, the random operator $(Y-\mu)\otimes(Y-\mu) : \mathcal{Y} \to \mathcal{Y}$ defined by $(Y-\mu)\otimes(Y-\mu)(y) = \langle Y-\mu, y\rangle_{\mathcal{Y}}(Y-\mu)$ is continuous and linear, and as a random variable taking values in $\mathscr{L}(\mathcal{Y})$, is Bochner integrable. We call $\Phi = \mathbb{E}\left[(Y-\mu)\otimes(Y-\mu)\right] \in \mathscr{L}(\mathcal{Y})$ the* covariance operator *of Y. The covariance operator Φ is self-adjoint and trace-class.*

Some authors (e.g. Bharucha-Reid (1972, p.24)) refer to the quantity

$$\text{Tr}\Phi = \mathbb{E}\left[\left\|(Y-\mu)\otimes(Y-\mu)\right\|_{\text{op}}\right] = \mathbb{E}\left[\|Y-\mu\|_{\mathcal{Y}}^2\right]$$

as the "variance" of Y .

For a random variable Y on $\mathcal{Y}$, its *characteristic function* is defined as the functional $\varphi_Y : \mathcal{Y} \to \mathbb{C}$ defined by $\varphi_Y(y) = \mathbb{E}\left[e^{i\langle Y,y\rangle_{\mathcal{Y}}}\right]$ (Da Prato and Zabczyk, 2014, pp.34-35). As for real variables, the characteristic function uniquely determines the distribution of the random variable (Da Prato and

Zabczyk, 2014, p.35, Proposition 2.5(i)). Clearly, the characteristic function of a Gaussian variable Y with mean μ and covariance operator Φ is given by

$$\varphi_Y(y) = e^{i\langle \mu, y\rangle_{\mathcal{Y}} - \frac{1}{2}\langle \Phi y, y\rangle_{\mathcal{Y}}}, \qquad y \in \mathcal{Y},$$

so a Gaussian distribution is uniquely determined by its mean and covariance operator.

The next result gives a concentration result for Gaussian random variables in separable Hilbert spaces. The proof can again be found in Da Prato and Zabczyk (2014, Section 2.3).

Proposition 17 *Suppose that Y is a Gaussian random variable in $\mathcal{Y}$ with mean 0 and covariance operator Φ. Then for any $0 < \lambda < \frac{1}{2\mathrm{Tr}\Phi}$,*

$$\mathbb{E}\left[e^{\lambda\|Y\|_{\mathcal{Y}}^2}\right] \leq \frac{1}{\sqrt{1 - 2\lambda\mathrm{Tr}\Phi}}$$

and consequently,

$$\mathbb{P}\left(\|Y\|_{\mathcal{Y}} \geq a\right) \leq 2e^{-\frac{3a^2}{8\mathrm{Tr}\Phi}}.$$

Appendix B. Differential Calculus

Recall that $\mathcal{Y}$ is a Hilbert space. Suppose that U is an open subset of $\mathbb{R}^d$, and denote the Euclidean norm in $\mathbb{R}^d$ by $\|\cdot\|$. We say that $f_1, f_2 : U \to \mathcal{Y}$ are *tangent* at a point $a \in U$ (Cartan, 1967, p.28) if the quantity

$$m(r) = \sup_{\|x-a\|\leq r} \left\|f_1(x) - f_2(x)\right\|_{\mathcal{Y}},$$

which is defined for $r > 0$ small enough (since U is open), satisfies the condition

$$\lim_{r\to 0}\frac{m(r)}{r} = 0, \qquad \text{which we also write as} \qquad m(r) = o(r).$$

We say that the map $g : U \to \mathcal{Y}$ is *differentiable* at $a \in U$ if g is continuous at a and there exists a linear map $g'(a) : \mathbb{R}^d \to \mathcal{Y}$ such that the maps $x \mapsto g(x) - g(a)$ and $x \mapsto g'(a)(x-a)$ are tangent at a (Cartan, 1967, p.29). This condition is also written as

$$\left\|g(x) - g(a) - g'(a)(x-a)\right\|_{\mathcal{Y}} = o(\|x-a\|).$$

This immediately implies that $g'(a)$ is continuous, so $g'(a)$ belongs to $\mathscr{L}(\mathbb{R}^d, \mathcal{Y})$, the space of continuous linear operators from $\mathbb{R}^d$ into $\mathcal{Y}$. We call $g'(a) \in \mathscr{L}(\mathbb{R}^d, \mathcal{Y})$ the *derivative* of g at a. We say that g is differentiable on U if g is differentiable at every point in U, and the map $g' : U \to \mathscr{L}(\mathbb{R}^d, \mathcal{Y})$ is called the *derivative map* of g. We say that g is *continuously differentiable*, or *of class C^1*, if g is differentiable at every point of U and the map $g' : U \to \mathscr{L}(\mathbb{R}^d, \mathcal{Y})$ is continuous (Cartan, 1967, p.30).

Let $g : U \to \mathcal{Y}$ be a continuous map. For each $a = (a_1, ..., a_d) \in U$ and each $l = 1, ..., d$, consider the inclusion $\lambda_l : \mathbb{R} \to \mathbb{R}^d$ defined by

$$\lambda_l(x_l) = (a_1, ..., a_{l-1}, x_l, a_{l+1}, ..., a_d).$$

The composition $g \circ \lambda_l$ is defined on an open subset $\lambda_l^{-1}(\mathcal{X}) \subset \mathbb{R}$, which contains a_l. If g is differentiable at a, then for each $l = 1, ..., d$, the map $g \circ \lambda_l$ is differentiable at a_l (Cartan, 1967, p.38, Proposition 2.6.1). The derivative of $g \circ \lambda_l$ at a is called the *partial derivative* of g, denoted by $\partial_l g(a)$, and lives in $\mathscr{L}(\mathbb{R}, \mathcal{Y})$. But $\mathscr{L}(\mathbb{R}, \mathcal{Y})$ is isometrically isomorphic to $\mathcal{Y}$ (Cartan, 1967, p.20, Exemple 1), so we can view $\partial_l g(a)$ as an element of $\mathcal{Y}$. Moreover,

$$g'(a)(h) = g'(a)(h_1, ..., h_d) = \sum_{l=1}^{d} h_l \partial_l g(a), \qquad \text{for } h = (h_1, ..., h_d) \in \mathbb{R}^d.$$

Cartan (1967, p.40, Proposition 2.6.2) tells us that g is of class C^1 if and only if $\partial_l g : U \to \mathcal{Y}$ is continuous for each $l = 1, ..., d$.

Next, we consider higher-order derivatives. For an integer m, a map $F : (\mathbb{R}^d)^m \to \mathcal{Y}$ is *m-linear* if, for each $k = 1, ..., m$ and any $a^{(1)}, ..., a^{(k-1)}, a^{(k+1)}, ..., a^{(m)} \in \mathbb{R}^d$, the map $x \mapsto F(a^{(1)}, ..., a^{(k-1)}, x, a^{(k+1)}, ..., a^{(m)})$ is linear from $\mathbb{R}^d$ into $\mathcal{Y}$ (Cartan, 1967, p.24). We say that F is an m-linear map from $\mathbb{R}^d$ into $\mathcal{Y}$, and denote by $\mathscr{L}_m(\mathbb{R}^d; \mathcal{Y})$ the space of all continuous m-linear maps from $\mathbb{R}^d$ into $\mathcal{Y}$[2]. The space $\mathscr{L}_m(\mathbb{R}^d; \mathcal{Y})$ can then be equipped with a natural operator norm defined by

$$\|F\|_{\text{op}} = \sup_{\|x^{(1)}\| \le 1, ..., \|x^{(m)}\| \le 1} \left\| F(x^{(1)}, ..., x^{(m)}) \right\|_{\mathcal{Y}}.$$

For any integer m, Coleman (2012, p.88, Theorem 4.4) tells us that $\Psi_m : \mathscr{L}(\mathbb{R}^d, \mathscr{L}_{m-1}(\mathbb{R}^d; \mathcal{Y})) \to \mathscr{L}_m(\mathbb{R}^d; \mathcal{Y})$ defined by $\Psi_m(F)(x^{(1)}, x^{(2)}, ..., x^{(m)}) = F(x^{(1)})(x^{(2)}, ..., x^{(m)})$ is an isometric isomorphism.

We say that $g : U \to \mathcal{Y}$ is *twice differentiable at $a \in U$* if the derivative map $g' : U \to \mathscr{L}(\mathbb{R}^d, \mathcal{Y})$ is differentiable at a. We denote by $g''(a) = g^{(2)}(a) \in \mathscr{L}(\mathbb{R}^d, \mathscr{L}(\mathbb{R}^d, \mathcal{Y})) \simeq \mathscr{L}_2(\mathbb{R}^d; Y)$ the *second derivative of g at a*. We say that g is *twice differentiable on U* if it is twice differentiable at all points in U. Then we have a map $g^{(2)} : U \to \mathscr{L}_2(\mathbb{R}^d, \mathcal{Y})$. We say that g is *twice continuously differentiable on U*, or *of class C^2 on U*, if g is twice differentiable and if the map $g^{(2)}$ is continuous (Cartan, 1967, p.64). By continuing in this way, we say that g is *m-times differentiable at $a \in U$* if $g^{(m-1)} : U \to \mathscr{L}_{m-1}(\mathbb{R}^d; \mathcal{Y})$ is differentiable at a, define the m^{th} *derivative* $g^{(m)}(a) \in \mathscr{L}_m(\mathbb{R}^d; \mathcal{Y})$ of g at a as the derivative of $g^{(m-1)}$ at a, and say that g is m-times differentiable on U if it is m-times differentiable at all points in U. We say that g is *of class C^m on U* if g is m-times differentiable at all points in U and the map $g^{(m)} : U \to \mathscr{L}_m(\mathbb{R}^d; \mathcal{Y})$ is continuous; we say that g is *of class C^∞* if it is of class C^m for all $m \in \mathbb{N}$ (Cartan, 1967, pp.69–70).

Similarly, for $l_1 \in \{1, ..., d\}$, if the partial derivative $\partial_{l_1} g : U \to \mathcal{Y}$ is defined in some neighbourhood of $x \in U$ and is differentiable, then for $l_2 \in \{1, ..., d\}$ (which may or may not be distinct from l_1), we may define the second partial derivative $\partial_{l_1}\partial_{l_2} g(a) \in \mathcal{Y}$. If $l_1 = l_2 = l$, then we write $\partial_l \partial_l g = \partial_l^2 g$. Analogously to the first partial derivative, we have a formula that expresses the second derivative as a sum of second partial derivatives:

$$g''(a)((h_1^{(1)}, ..., h_d^{(1)}), (h_1^{(2)}, ..., h_d^{(2)})) = \sum_{l_1, l_2 = 1}^{d} h_{l_1}^{(1)} h_{l_2}^{(2)} \partial_{l_1} \partial_{l_2} g(a),$$

where $h^{(1)} = (h_1^{(1)}, ..., h_d^{(1)}), h^{(2)} = (h_1^{(2)}, ..., h_d^{(2)}) \in \mathbb{R}^d$ (Cartan, 1967, p.68, (5.2.5)). Continuing in the same way, we can define the m^{th} partial derivative $\partial_{l_1} ... \partial_{l_m} g(a) \in \mathcal{Y}$. Then writing $\mathbf{h} =$

[2] Beware that $\mathscr{L}_m(\mathbb{R}^d; \mathcal{Y})$, the space of continuous m-linear maps from $\mathbb{R}^d$ into $\mathcal{Y}$, is different to $\mathscr{L}((\mathbb{R}^d)^m, \mathcal{Y})$, the space of continuous linear maps from $(\mathbb{R}^d)^m$ into $\mathcal{Y}$.

$(h^{(1)},...,h^{(m)}) \in (\mathbb{R}^d)^m$, we have

$$g^{(m)}(a)(\mathbf{h}) = \sum_{l_1,...,l_m=1}^{d} h^{(1)}_{l_1}...h^{(m)}_{l_m}\partial_{l_1}...\partial_{l_m}g(a).$$

Finally, we state the extension of Taylor's theorem to functions with values in $\mathcal{Y}$, with Lagrange's form of the remainder. To this end, for $a, b \in \mathbb{R}^d$, define the *segment* joining a and b as the set (Coleman, 2012, p.51).

$$[a, b] = \{x \in \mathbb{R}^d : x = va + (1 - v)b, v \in [0, 1]\}.$$

Theorem 18 (Cartan (1967, p.77, Théorème 5.6.2)) *Suppose that $g : U \to \mathcal{Y}$ is $(m + 1)$-times differentiable, that the segment $[a, a + h]$ is contained in U and that, for some $K > 0$, we have*

$$\left\|g^{(m+1)}(x)\right\|_{\text{op}} \leq K \qquad \textit{for all } x \in U.$$

Then

$$\left\|g(a + h) - \sum_{k=0}^{m} \frac{1}{k!} g^{(k)}(a)((h)^k)\right\|_{\mathcal{Y}} \leq K\frac{\|h\|^{m+1}}{(m + 1)!},$$

where we wrote $(h)^k = (h, ..., h) \in (\mathbb{R}^d)^k$ for $k = 1, ..., m$.

Write $\mathbb{N}_0 = \{0, 1, 2, ...\}$, and for $p = (p_1, ..., p_d) \in \mathbb{N}_0^d$, write $[p] := p_1 + ... + p_d$. Then we denote the p^{th} partial derivative $\partial_1^{p_1}...\partial_d^{p_d}g(a)$ of g at $a \in U$ as $D^p g(a) \in \mathcal{Y}$. This is possible since the order of partial differentiation is immaterial by repeated application of Cartan (1967, p.69, Proposition 5.2.2). Hence, for each $k = 1, ..., m + 1$, we have

$$g^{(k)}(a)((h)^k) = \sum_{l_1,...,l_k=1}^{d} h_{l_1}...h_{l_k}\partial_{l_1}...\partial_{l_k}g(a) = \sum_{[p]=k} \frac{k! h^p}{p!} D^p g(a),$$

where we wrote h^p as a shorthand for $h_1^{p_1}...h_d^{p_d}$ and $p!$ for $p_1!...p_d!$. Hence, using partial derivatives, we can express Taylor's theorem above as

$$\left\|g(a + h) - \sum_{[p]\leq m} \frac{h^p}{p!} D^p g(a)\right\|_{\mathcal{Y}} \leq K\frac{\|h\|^{m+1}}{(m + 1)!}.$$

Appendix C. Empirical Process Theory with Vector-Valued Functions

In this Section, we state and prove some basic empirical process-theoretic results, adapted to our setting of vector-valued functions. Although technically new, the ideas and proofs carry over from the real case with ease, by applying vector-valued concentration inequalities from Section A.

C.1. Symmetrisation

Symmetrisation is an indispensable technique in empirical process theory. Let $X_1', ..., X_n'$ be another set of independent copies of X, independent of $X_1, ..., X_n$. Denote by P_n' the empirical measure on $X_1', ..., X_n'$, i.e. $P_n' = \frac{1}{n}\sum_{i=1}^n \delta_{X_i'}$.

Lemma 19 *We have*

$$\mathbb{E}\left[\|P_n - P\|_{\mathcal{G}}\right] \leq \mathbb{E}\left[\|P_n - P_n'\|_{\mathcal{G}}\right].$$

Proof Denote by $\mathcal{F}_n$ the σ-algebra generated by $X_1, ..., X_n$. Then for each $g \in \mathcal{G}$, we have

$$\mathbb{E}\left[P_n g \mid \mathcal{F}_n\right] = P_n g \qquad \text{and} \qquad \mathbb{E}\left[P_n' g \mid \mathcal{F}_n\right] = Pg,$$

and so

$$(P_n - P)g = \mathbb{E}\left[\left(P_n - P_n'\right) g \mid \mathcal{F}_n\right].$$

Now see that

$$\begin{aligned}
\|P_n - P\|_{\mathcal{G}} &= \sup_{g\in\mathcal{G}} \left\|\mathbb{E}\left[\left(P_n - P_n'\right) g \mid \mathcal{F}_n\right]\right\|_{\mathcal{Y}} \\
&\leq \sup_{g\in\mathcal{G}} \mathbb{E}\left[\left\|\left(P_n - P_n'\right) g\right\|_{\mathcal{Y}} \mid \mathcal{F}_n\right] \quad \text{by Jensen's inequality} \\
&\leq \mathbb{E}\left[\sup_{g\in\mathcal{G}} \left\|\left(P_n - P_n'\right) g\right\|_{\mathcal{Y}} \mid \mathcal{F}_n\right].
\end{aligned}$$

Now take expectations on both sides and apply the law of iterated expectations arrive at the result. ■

We let $\{\sigma_i\}_{i=1}^n$ be a *Rademacher sequence*, i.e. a sequence of independent random variables σ_i with

$$\mathbb{P}\left(\sigma_i = 1\right) = \mathbb{P}\left(\sigma_i = -1\right) = \frac{1}{2}, \qquad \text{for all } i = 1, ..., n.$$

We define the symmetrised empirical measures $P_n^\sigma = \frac{1}{n}\sum_{i=1}^n \sigma_i \delta_{X_i}$ and $P_n'^\sigma = \frac{1}{n}\sum_{i=1}^n \sigma_i \delta_{X_i'}$, and denote

$$P_n^\sigma g = \frac{1}{n}\sum_{i=1}^n \sigma_i g(X_i) \qquad \text{and} \qquad P_n'^\sigma g = \frac{1}{n}\sum_{i=1}^n \sigma_i g(X_i').$$

Lemma 20 (Symmetrisation with means) *We have*

$$\mathbb{E}\left[\|P_n - P\|_{\mathcal{G}}\right] \leq 2\mathbb{E}\left[\|P_n^\sigma\|_{\mathcal{G}}\right]$$

Proof Note that $\left\|P_n - P_n'\right\|_{\mathcal{G}}$ has the same distribution as $\left\|P_n^\sigma - P_n'^\sigma\right\|_{\mathcal{G}}$, since, for each $i = 1, ..., n$ and $g \in \mathcal{G}$. $g(X_i) - g(X_i')$ and $\sigma_i\left(g(X_i) - g(X_i')\right)$ have the same distribution. Hence, the triangle inequality gives us

$$\mathbb{E}\left[\left\|P_n - P_n'\right\|_{\mathcal{G}}\right] = \mathbb{E}\left[\left\|P_n^\sigma - P_n'^\sigma\right\|_{\mathcal{G}}\right] \leq \mathbb{E}\left[\|P_n^\sigma\|_{\mathcal{G}} + \left\|P_n'^\sigma\right\|_{\mathcal{G}}\right] = 2\mathbb{E}\left[\|P_n^\sigma\|_{\mathcal{G}}\right].$$

Now apply Lemma 19. ■

Lemma 21 (Symmetrisation with probabilities) *Let $a > 0$. Suppose that for all $g \in \mathcal{G}$,*

$$\mathbb{P}\left(\left\|(P_n - P)\,g\right\|_{\mathcal{Y}} > \frac{a}{2}\right) \leq \frac{1}{2}.$$

Then

$$\mathbb{P}\left(\|P_n - P\|_{\mathcal{G}} > a\right) \leq 4\mathbb{P}\left(\|P_n^\sigma\|_{\mathcal{G}} > \frac{a}{4}\right).$$

Proof Denote again by $\mathcal{F}_n$ the σ-algebra generated by $X_1, ..., X_n$. If $\|P_n - P\|_{\mathcal{G}} > a$, then we know that for some random function g_* depending on $X_1, ..., X_n$, $\left\|(P_n - P)\,g_*\right\|_{\mathcal{Y}} > a$. Because $X_1', ..., X_n'$ are independent of $\mathcal{F}_n$,

$$\mathbb{P}\left(\left\|\left(P_n' - P\right)g_*\right\|_{\mathcal{Y}} > \frac{a}{2} \mid \mathcal{F}_n\right) = \mathbb{P}\left(\left\|(P_n - P)\,g_*\right\|_{\mathcal{Y}} > \frac{a}{2}\right) \leq \frac{1}{2}. \tag{*}$$

Then see that,

$$\begin{aligned}
\mathbb{P}\left(\|P_n - P\|_{\mathcal{G}} > a\right) &\leq \mathbb{P}\left(\left\|(P_n - P)\,g_*\right\|_{\mathcal{Y}} > a\right) \\
&= \mathbb{E}\left[\mathbf{1}\left\{\left\|(P_n - P)\,g_*\right\|_{\mathcal{Y}} > a\right\}\right] \\
&\leq 2\mathbb{E}\left[\mathbb{P}\left(\left\|\left(P_n' - P\right)g_*\right\|_{\mathcal{Y}} \leq \frac{a}{2} \mid \mathcal{F}_n\right)\mathbf{1}\left\{\left\|(P_n - P)\,g_*\right\|_{\mathcal{Y}} > a\right\}\right] \text{by (*)} \\
&= 2\mathbb{E}\left[\mathbb{P}\left(\left\|\left(P_n' - P\right)g_*\right\|_{\mathcal{Y}} \leq \frac{a}{2} \text{ and } \left\|(P_n - P)\,g_*\right\|_{\mathcal{Y}} > a \mid \mathcal{F}_n\right)\right] \\
&= 2\mathbb{P}\left(\left\|\left(P_n' - P\right)g_*\right\|_{\mathcal{Y}} \leq \frac{a}{2} \text{ and } \left\|(P_n - P)\,g_*\right\|_{\mathcal{Y}} > a\right).
\end{aligned}$$

But if the two inequalities in the probability on the last line hold, then the reverse triangle inequality gives us

$$\frac{a}{2} < \left\|(P_n - P)\,g_*\right\|_{\mathcal{Y}} - \left\|\left(P_n' - P\right)g_*\right\|_{\mathcal{Y}} \leq \left\|\left(P_n - P_n'\right)g_*\right\|_{\mathcal{Y}},$$

so

$$\begin{aligned}
\mathbb{P}\left(\|P_n - P\|_{\mathcal{G}} > a\right) &\leq 2\mathbb{P}\left(\left\|\left(P_n - P_n'\right)g_*\right\|_{\mathcal{Y}} > \frac{a}{2}\right) \\
&\leq 2\mathbb{P}\left(\|P_n - P_n'\|_{\mathcal{G}} > \frac{a}{2}\right) \\
&= 2\mathbb{P}\left(\|P_n^\sigma - P_n'^\sigma\|_{\mathcal{G}} > \frac{a}{2}\right) \\
&\leq 2\mathbb{P}\left(\|P_n^\sigma\|_{\mathcal{G}} > \frac{a}{4} \text{ or } \|P_n'^\sigma\|_{\mathcal{G}} > \frac{a}{4}\right) \\
&\leq 4\mathbb{P}\left(\|P_n^\sigma\|_{\mathcal{G}} > \frac{a}{4}\right).
\end{aligned}$$

■

A simple application of the above symmetrisation argument and Hoeffding's inequality in Hilbert spaces (Proposition 12) shows that finite function classes are Glivenko-Cantelli.

Lemma 22 *Let $\mathcal{G} = \{g_1, ..., g_N\} \in L^1(\mathcal{X}, P; \mathcal{Y})$ be a finite class of functions with cardinality $N > 1$. Then we have*

$$\|P_n - P\|_{\mathcal{G}} \to 0.$$

Proof Take any $K > 0$. Define the function $G : \mathcal{X} \to \mathbb{R}$ by $G(x) = \max_{1\leq j\leq N} \left\|g_j(x)\right\|_{\mathcal{Y}}$. Since each $\left\|g_j\right\|_{\mathcal{Y}}$ is integrable, and we have a finite collection, G is also integrable. Then, for each $j = 1, ..., N$, define the function $\tilde{g}_j : \mathcal{X} \to \mathcal{Y}$ by $\tilde{g}_j = g_j \mathbf{1}\{G \leq K\}$. Then for all $i = 1, ..., n$, letting σ_i be independent Rademacher variables again, we have

$$\mathbb{E}\left[\sigma_i \tilde{g}_j(X_i)\right] = 0 \qquad \text{and} \qquad \left\|\sigma_i \tilde{g}_j(X_i)\right\|_{\mathcal{Y}} \leq K \text{ almost surely.}$$

Hence, for each $j = 1, ..., N$, by Hoeffding's inequality (Proposition 12), for any $t > 0$, we have

$$\mathbb{P}\left(\left\|P_n^\sigma \tilde{g}_j\right\|_{\mathcal{Y}} \geq 2K\sqrt{\frac{t}{n}}\right) = \mathbb{P}\left(\left\|\sum_{i=1}^n \sigma_i \tilde{g}_j(X_i)\right\|_{\mathcal{Y}} \geq 2K\sqrt{nt}\right) \leq 2e^{-t}.$$

By the union bound, for any $t > 0$, we have

$$\begin{aligned}\mathbb{P}\left(\max_{1\leq j\leq N} \left\|P_n^\sigma \tilde{g}_j\right\|_{\mathcal{Y}} \geq 2K\sqrt{\frac{t+\log N}{n}}\right) &\leq N \max_{1\leq j\leq N} \mathbb{P}\left(\left\|P_n^\sigma \tilde{g}_j\right\|_{\mathcal{Y}} \geq 2K\sqrt{\frac{t+\log N}{n}}\right)\\ &\leq 2e^{-t}.\end{aligned}$$

Now see that, for each $j = 1, ..., N$, Chebyshev's inequality gives

$$\mathbb{P}\left(\left\|(P_n - P)\tilde{g}_j\right\|_{\mathcal{Y}} > 4K\sqrt{\frac{t+\log N}{n}}\right) \leq \frac{n\mathbb{E}\left[\left\|(P_n - P)\tilde{g}_j\right\|_{\mathcal{Y}}^2\right]}{16K^2(t+\log N)} \leq \frac{1}{16(t+\log N)} \leq \frac{1}{2},$$

where the last inequality follows since $8t + 8\log N \geq 8\log 2 \geq 1$. Now apply Lemma 21 to see that

$$\begin{aligned}\mathbb{P}\left(\max_{1\leq j\leq N} \left\|(P_n - P)\tilde{g}_j\right\|_{\mathcal{Y}} > 8K\sqrt{\frac{t+\log N}{n}}\right) &\leq 4\mathbb{P}\left(\max_{1\leq j\leq N} \left\|P_n^\sigma \tilde{g}_j\right\|_{\mathcal{Y}} > 2K\sqrt{\frac{t+\log N}{n}}\right)\\ &\leq 8e^{-t}.\end{aligned}$$

This tells us that

$$\max_{1\leq j\leq N} \left\|(P_n - P)\tilde{g}_j\right\|_{\mathcal{Y}} \xrightarrow{P} 0.$$

Finally, see that

$$\|P_n - P\|_{\mathcal{G}} \leq \max_{1\leq j\leq N} \left\|(P_n - P)\tilde{g}_j\right\|_{\mathcal{Y}} + \max_{1\leq j\leq N} \left\|(P_n - P) g_j \mathbf{1}\{G > K\}\right\|_{\mathcal{Y}}.$$

Here, the first term converges to 0 in probability for any $K > 0$, as shown above, and the second term decomposes as

$$\begin{aligned}\max_{1\leq j\leq N} \left\|(P_n - P) g_j \mathbf{1}\{G > K\}\right\|_{\mathcal{Y}} &\leq (P_n + P) G\mathbf{1}\{G > K\}\\ &= (P_n - P) G\mathbf{1}\{G > K\} + 2PG\mathbf{1}\{G > K\}\\ &\leq (P_n - P) G + 2PG\mathbf{1}\{G > K\}.\end{aligned}$$

Here, the first term converges to 0 in probability by the weak law of large numbers, and the second term converges to 0 as $K \to \infty$, by Çınlar (2011, p.71, Lemma 3.10). ■

C.2. Uniform law of large numbers

We start with a definition.

Definition 23 (Adapted from van de Geer (2000, p.26, Definition 3.1)) *The function $G : \mathcal{X} \to \mathbb{R}$ defined by $G(\cdot) = \sup_{g \in \mathcal{G}} \left\| g(\cdot) \right\|_{\mathcal{Y}}$ is called the envelope of $\mathcal{G}$.*

The following is a uniform law of large numbers based on conditions on the entropy $H(\delta, \mathcal{G}, \|\cdot\|_{1,P_n})$ and the envelope G.

Theorem 24 *Suppose that*

$$G \in L^1(\mathcal{X}, P; \mathbb{R}) \qquad \textit{and} \qquad \frac{1}{n} H(\delta, \mathcal{G}, \|\cdot\|_{1,P_n}) \xrightarrow{P} 0 \textit{ for each } \delta > 0.$$

Then $\mathcal{G}$ is a Glivenko Cantelli class, i.e. $\|P_n - P\|_{\mathcal{G}} \xrightarrow{P} 0$.

Proof Take any $K > 0$ and $\delta > 0$. Denote again by $\mathcal{F}_n$ the σ-algebra generated by $X_1, ..., X_n$, and define $\mathcal{G}_K = \{g\mathbf{1}\{G \leq K\} : g \in \mathcal{G}\}$. Let $g_1, ..., g_N$, with $N = N(\delta, \mathcal{G}, \|\cdot\|_{1,P_n})$, be a minimal δ-covering of $\mathcal{G}$. Then N is a random variable, that is measurable with respect to $\mathcal{F}_n$. Moreover, writing $\tilde{g}_j = g_j \mathbf{1}\{G \leq K\}$ for each $j = 1, ..., N$, $\tilde{g}_1, ..., \tilde{g}_N$ form a δ-covering of $\mathcal{G}_K$, since, for any $\tilde{g} = g\mathbf{1}\{G \leq K\} \in \mathcal{G}_K$ for $g \in \mathcal{G}$, there exists $j \in \{1, ..., N\}$ with $\|g - g_j\|_{1,P_n} \leq \delta$, so $\|\tilde{g} - \tilde{g}_j\|_{1,P_n} \leq \|g - g_j\|_{1,P_n} \leq \delta$.

Note that, when $\left\|\tilde{g} - \tilde{g}_j\right\|_{1,P_n} = P_n \left\|\tilde{g} - \tilde{g}_j\right\|_{\mathcal{Y}} \leq \delta$, we have

$$\|P_n^\sigma \tilde{g}\|_{\mathcal{Y}} \leq \left\|P_n^\sigma \tilde{g}_j\right\|_{\mathcal{Y}} + \left\|P_n^\sigma \tilde{g} - P_n^\sigma \tilde{g}_j\right\|_{\mathcal{Y}} \leq \left\|P_n^\sigma \tilde{g}_j\right\|_{\mathcal{Y}} + P_n \left\|\tilde{g} - \tilde{g}_j\right\|_{\mathcal{Y}} \leq \left\|P_n^\sigma \tilde{g}_j\right\|_{\mathcal{Y}} + \delta.$$

So for any $\tilde{g} \in \mathcal{G}_K$,

$$\|P_n^\sigma \tilde{g}\|_{\mathcal{Y}} \leq \max_{1 \leq j \leq N} \left\|P_n^\sigma \tilde{g}_j\right\|_{\mathcal{Y}} + \delta. \qquad (*)$$

By Hoeffding's inequality and union bound (as in the proof of Lemma 22, since N is measurable with respect to $\mathcal{F}_n$), for any $t > 0$, we have

$$\mathbb{P}\left(\max_{1 \leq j \leq N} \left\|P_n^\sigma \tilde{g}_j\right\|_{\mathcal{Y}} \geq 2K\sqrt{\frac{t + \log N}{n}} \mid \mathcal{F}_n\right) \leq 2e^{-t}.$$

We then apply (*) and integrate both sides (to remove the conditioning on $\mathcal{F}_n$) to see that, for any $t > 0$,

$$\mathbb{P}\left(\|P_n^\sigma\|_{\mathcal{G}_K} \geq \delta + 2K\sqrt{\frac{t + \log N}{n}}\right) \leq 2e^{-t}.$$

Then see that, using the elementary inequality $\sqrt{a} + \sqrt{b} \geq \sqrt{a + b}$,

$$\begin{aligned}
&\mathbb{P}\left(\|P_n^\sigma\|_{\mathcal{G}_K} \geq 2\delta + 2K\sqrt{\frac{t}{n}}\right) \\
&\leq \mathbb{P}\left(\|P_n^\sigma\|_{\mathcal{G}_K} \geq \delta + 2K\sqrt{\frac{t}{n}} + 2K\sqrt{\frac{\log N}{n}}\right) + \mathbb{P}\left(2K\sqrt{\frac{\log N}{n}} \geq \delta\right)
\end{aligned}$$

$$\begin{aligned}&\leq \mathbb{P}\left(\|P_n^\sigma\|_{\mathcal{G}_K} \geq \delta + 2K\sqrt{\frac{t+\log N}{n}}\right) + \mathbb{P}\left(2K\sqrt{\frac{1}{n}H(\delta,\mathcal{G},\|\cdot\|_{1,P_n})} \geq \delta\right)\\ &\leq 2e^{-t} + \mathbb{P}\left(2K\sqrt{\frac{1}{n}H(\delta,\mathcal{G},\|\cdot\|_{1,P_n})} \geq \delta\right).\end{aligned}$$

Also, by Chebyshev's inequality, for each $\tilde{g} \in \mathcal{G}_K$, we have, for any $t \geq \frac{1}{8}$

$$\begin{aligned}\mathbb{P}\left(\left\|(P_n - P)\,\tilde{g}\right\|_{\mathcal{Y}} > 4\delta + 4K\sqrt{\frac{t}{n}}\right) &\leq \mathbb{P}\left(\left\|(P_n - P)\,\tilde{g}\right\|_{\mathcal{Y}} > 4K\sqrt{\frac{t}{n}}\right)\\ &\leq \frac{n\mathbb{E}\left[\left\|(P_n - P)\,\tilde{g}\right\|_{\mathcal{Y}}^2\right]}{16K^2t}\\ &\leq \frac{1}{16t}\\ &\leq \frac{1}{2}.\end{aligned}$$

Hence, we can apply symmetrisation with probabilities again (Lemma 21) to see that, for any $t \geq \frac{1}{8}$,

$$\begin{aligned}\mathbb{P}\left(\|P_n - P\|_{\mathcal{G}_K} \geq 8\delta + 8K\sqrt{\frac{t}{n}}\right) &\leq 4\mathbb{P}\left(\|P_n^\sigma\|_{\mathcal{G}_K} \geq 2\delta + 2K\sqrt{\frac{t}{n}}\right)\\ &\leq 2^{-t} + \mathbb{P}\left(2K\sqrt{\frac{1}{n}H(\delta,\mathcal{G},\|\cdot\|_{1,P_n})} \geq \delta\right).\end{aligned}$$

Here, since $\delta > 0$ was arbitrary and $\frac{1}{n}H(\delta,\mathcal{G},\|\cdot\|_{1,P_n}) \xrightarrow{P} 0$ by hypothesis, we have that $\mathcal{G}_K$ is Glivenko Cantelli.

Finally, see that

$$\|P_n - P\|_{\mathcal{G}} \leq \sup_{\tilde{g}\in\mathcal{G}_K} \left\|(P_n - P)\,\tilde{g}\right\|_{\mathcal{Y}} + \sup_{g\in\mathcal{G}} \left\|(P_n - P)\,g\mathbf{1}\left\{G > K\right\}\right\|_{\mathcal{Y}}.$$

Here, the first term converges to 0 in probability for any $K > 0$, as shown above, and the second term decomposes as

$$\begin{aligned}\sup_{g\in\mathcal{G}} \left\|(P_n - P)\,g\mathbf{1}\left\{G > K\right\}\right\|_{\mathcal{Y}} &\leq (P_n + P)\,G\mathbf{1}\left\{G > K\right\}\\ &= (P_n - P)\,G\mathbf{1}\left\{G > K\right\} + 2PG\mathbf{1}\left\{G > K\right\}\\ &\leq (P_n - P)\,G + 2PG\mathbf{1}\left\{G > K\right\}.\end{aligned}$$

Here, the first term converges to 0 in probability by the weak law of large numbers, and the second term converges to 0 as $K \to \infty$, by Çınlar (2011, p.71, Lemma 3.10), since G is integrable by hypothesis. ■

C.3. Chaining and asymptotic equicontinuity with empirical entropy

In this subsection we show that, with additional conditions on the entropy of $\mathcal{G}$ (which we assume to be totally bounded with respect to the appropriate metric) and a technique called "chaining", we can derive explicit finite-sample bounds, and show the asymptotic continuity of the empirical process indexed by $\mathcal{G}$ (see Definition 3). As before, we work conditionally on the samples, and denote the σ-algebra generated by $X_1, ..., X_n$ as $\mathcal{F}_n$.

Suppose that $\mathcal{G}$ has an envelope $G \in L^2(\mathcal{X}, P; \mathbb{R})$ (see Definition 23). Then the quantity $R = \sup_{g\in\mathcal{G}} \|g\|_{2,P}$ is finite, since

$$R^2 = \sup_{g\in\mathcal{G}} \mathbb{E}\left[\|g(X)\|_{\mathcal{Y}}^2\right] \leq \mathbb{E}\left[\sup_{g\in\mathcal{G}} \|g(X)\|_{\mathcal{Y}}^2\right] = \mathbb{E}\left[G^2\right] < \infty.$$

Similarly, the quantity $R_n = \sup_{g\in\mathcal{G}} \|g\|_{2,P_n}$ is almost surely finite. We call R and R_n the *theoretical radius* and *empirical radius* of $\mathcal{G}$, respectively. Note that R_n is a random quantity, measurable with respect to $\mathcal{F}_n$.

Let us fix $S \in \mathbb{N}$. To ease the notation, for $s = 0, 1, ..., S$, write $N_s = N(2^{-s}R_n, \mathcal{G}, \|\cdot\|_{2,P_n})$ for the $2^{-s}R_n$-covering number of $\mathcal{G}$ with respect to the $\|\cdot\|_{2,P_n}$-metric, which we assume to be finite. Let $\{g_j^s\}_{j=1}^{N_s} \subset \mathcal{G}$ be a $2^{-s}R_n$-covering set of $\mathcal{G}$ with respect to the $\|\cdot\|_{2,P_n}$-metric. Note that $\{g^0\} = \{0\}$ is an R_n-covering set of $\mathcal{G}$, since, for any $g \in \mathcal{G}$, $\|g\|_{2,P_n} \leq R_n$. Similarly, write $H_s = \log N_s$ for each $s = 0, 1, ..., S$, for the corresponding entropy. Note that the quantities N_s and H_s, as well as the covering set $\{g_j^s\}_{j=1}^{N_s}$, are random quantities that are measurable with respect to $\mathcal{F}_n$.

Now fix $g \in \mathcal{G}$. Then define

$$\begin{aligned}
g^{S+1} &:= \underset{\{g_j^{S+1}\}_{j=1}^{N_{S+1}}}{\arg\min} \left\{\left\|g - g_j^{S+1}\right\|_{2,P_n}\right\} \\
g^{S} &:= \underset{\{g_j^{S}\}_{j=1}^{N_{S}}}{\arg\min} \left\{\left\|g^{S+1} - g_j^{S}\right\|_{2,P_n}\right\} \\
&\vdots \quad \vdots \\
g^{s} &:= \underset{\{g_j^{s}\}_{j=1}^{N_{s}}}{\arg\min} \left\{\left\|g^{s+1} - g_j^{s}\right\|_{2,P_n}\right\} \\
&\vdots \quad \vdots \\
g^0 &:= 0.
\end{aligned}$$

Proposition 25 (Chaining) *We fix $S \in \mathbb{N}$. Define*

$$J_n := \sum_{s=0}^{S} 2^{-s} R_n \sqrt{2H_{s+1}}.$$

(i) For all $t > 0$,

$$\mathbb{P}\left(\sup_{g\in\mathcal{G}}\left\|\sum_{s=0}^{S} P_n^\sigma\left(g^{s+1}-g^s\right)\right\|_{\mathcal{Y}} \geq \frac{\sqrt{2}J_n}{\sqrt{n}} + 6R_n\sqrt{\frac{1+t}{n}} \mid \mathcal{F}_n\right) \leq 2e^{-t}.$$

(ii) Suppose that $\varepsilon_1, ..., \varepsilon_n$ are i.i.d. Gaussian random variables in $\mathcal{Y}$ with mean 0 and covariance operator Q. Without loss of generality (by rescaling if necessary), assume $\mathrm{Tr}Q = 1$*. For each $g \in \mathcal{G}$, we can consider the following inner product:*

$$\langle \varepsilon, g\rangle_{2,P_n} = \frac{1}{n}\sum_{i=1}^{n}\langle \varepsilon_i, g(X_i)\rangle_{\mathcal{Y}}.$$

Then for all $t > 0$,

$$\mathbb{P}\left(\sup_{g\in\mathcal{G}}\sum_{s=0}^{S}\left\langle \varepsilon, g^{s+1}-g^s\right\rangle_{2,P_n} \geq \frac{J_n}{\sqrt{n}} + 4R_n\sqrt{\frac{1+t}{n}} \mid \mathcal{F}_n\right) \leq e^{-t}.$$

Proof

(i) Fix $s \in \{0, 1, ..., S\}$ and $k \in \{1, ..., N_{s+1}\}$. Denote

$$g_k^{s+1,s} = \arg\min_{\{g_j^s\}_{j=1}^{N_s}}\left\{\left\|g_k^{s+1} - g_j^s\right\|_{2,P_n}\right\}.$$

Then

$$\left\|P_n^\sigma\left(g_k^{s+1} - g_k^{s+1,s}\right)\right\|_{\mathcal{Y}} \leq \frac{1}{n}\sum_{i=1}^{n}\left\|g_k^{s+1}(X_i) - g_k^{s+1,s}(X_i)\right\|_{\mathcal{Y}},$$

where

$$\sqrt{\sum_{i=1}^{n}\left\|g_k^{s+1}(X_i) - g_k^{s+1,s}(X_i)\right\|_{\mathcal{Y}}^2} = \sqrt{n}\left\|g_k^{s+1} - g_k^{s+1,s}\right\|_{2,P_n} \leq \sqrt{n}2^{-s}R_n,$$

since the $\{g_j^s\}_{j=1}^{N_s}$ form a $2^{-s}R_n$-covering of $(\mathcal{G}, \|\cdot\|_{2,P_n})$. Hence, noting that R_n is measurable with respect to $\mathcal{F}_n$, Hoeffding's inequality (Proposition 12) gives, for any $t > 0$,

$$\mathbb{P}\left(\left\|P_n^\sigma\left(g_k^{s+1} - g_k^{s+1,s}\right)\right\|_{\mathcal{Y}} \geq 2^{-(s-1)}R_n\sqrt{\frac{t}{n}} \mid \mathcal{F}_n\right) \leq 2e^{-t}.$$

Therefore (by the union bound), for each $s = 0, 1, ..., S$ and all $t > 0$,

$$\mathbb{P}\left(\max_{k\in\{1,...,N_{s+1}\}}\left\|P_n^\sigma\left(g_k^{s+1} - g_k^{s+1,s}\right)\right\|_{\mathcal{Y}} \geq 2^{-(s-1)}R_n\sqrt{\frac{H_{s+1}+t}{n}} \mid \mathcal{F}_n\right) \leq 2e^{-t}.$$

Fix t and for $s = 0, 1, ..., S$, let

$$\begin{aligned}\alpha_s :&= 2^{-(s-1)} R_n \left(\sqrt{H_{s+1}} + \sqrt{(1+s)(1+t)} \right) \\ &\geq 2^{-(s-1)} R_n \left(\sqrt{H_{s+1} + (1+s)(1+t)} \right),\end{aligned}$$

using $\sqrt{a} + \sqrt{b} \geq \sqrt{a+b}$. Then using $\sum_{s=0}^{S} 2^{-(s-1)} \sqrt{1+s} \leq 6$,

$$\begin{aligned}\sum_{s=0}^{S} \alpha_s &= \sqrt{2} J_n + \sum_{s=0}^{S} 2^{-(s-1)} R_n \sqrt{(1+s)(1+t)} \\ &\leq \sqrt{2} J_n + 6 R_n \sqrt{1+t}.\end{aligned}$$

Therefore

$$\begin{aligned}&\mathbb{P}\left(\sup_{g \in \mathcal{G}} \left\| \sum_{s=0}^{S} P_n^\sigma \left(g^{s+1} - g^s \right) \right\|_{\mathcal{Y}} \geq \frac{\sqrt{2} J_n}{\sqrt{n}} + 6 R_n \sqrt{\frac{1+t}{n}} \mid \mathcal{F}_n \right) \\ &\leq \mathbb{P}\left(\sup_{g \in \mathcal{G}} \left\| \sum_{s=0}^{S} P_n^\sigma \left(g^{s+1} - g^s \right) \right\|_{\mathcal{Y}} \geq \frac{1}{\sqrt{n}} \sum_{s=0}^{S} \alpha_s \mid \mathcal{F}_n \right) \\ &\leq \mathbb{P}\left(\sum_{s=0}^{S} \sup_{g \in \mathcal{G}} \left\| P_n^\sigma \left(g^{s+1} - g^s \right) \right\|_{\mathcal{Y}} \geq \frac{1}{\sqrt{n}} \sum_{s=0}^{S} \alpha_s \mid \mathcal{F}_n \right) \\ &\leq \sum_{s=0}^{S} \mathbb{P}\left(\sup_{g \in \mathcal{G}} \left\| P_n^\sigma \left(g^{s+1} - g^s \right) \right\|_{\mathcal{Y}} \geq \frac{1}{\sqrt{n}} \alpha_s \mid \mathcal{F}_n \right) \\ &= \sum_{s=0}^{S} \mathbb{P}\left(\max_{k=1,\ldots,N_{s+1}} \left\| P_n^\sigma \left(g_k^{s+1} - g_k^{s+1,s} \right) \right\|_{\mathcal{Y}} \geq \frac{1}{\sqrt{n}} \alpha_s \mid \mathcal{F}_n \right) \\ &\leq 2 \sum_{s=0}^{S} e^{-(1+s)(1+t)} \\ &\leq 2e^{-t}.\end{aligned}$$

(ii) Fix $s \in \{0, 1, ..., S\}$ and $k \in \{1, ..., N_{s+1}\}$. Denote

$$g_k^{s+1,s} = \underset{\{g_j^s\}_{j=1}^{N_s}}{\arg\min} \left\{ \left\| g_k^{s+1} - g_j^s \right\|_{2,P_n} \right\}.$$

Let $\lambda > 0$ be arbitrary. Then Markov's inequality gives us, for any $t > 0$,

$$\begin{aligned}&\mathbb{P}\left(\left\langle \varepsilon, g_k^{s+1} - g_k^{s+1,s} \right\rangle_{2,P_n} \geq 2^{-s} R_n \sqrt{\frac{2t}{n}} \mid \mathcal{F}_n \right) \\ &\leq e^{-\lambda 2^{-s} R_n \sqrt{\frac{2t}{n}}} \mathbb{E}\left[e^{\frac{\lambda}{n} \sum_{i=1}^{n} \left\langle \varepsilon_i, g_k^{s+1}(X_i) - g_k^{s+1,s}(X_i) \right\rangle_{\mathcal{Y}}} \mid \mathcal{F}_n \right]\end{aligned}$$

$$= e^{-\lambda 2^{-s} R_n \sqrt{\frac{2t}{n}}} \prod_{i=1}^{n} \mathbb{E}\left[e^{\frac{\lambda}{n}\left\langle \varepsilon_i, g_k^{s+1}(X_i) - g_k^{s+1,s}(X_i) \right\rangle_{\mathcal{Y}}} \mid \mathcal{F}_n \right].$$

Here, since ε_i is a $\mathcal{Y}$-valued Gaussian random variable with mean 0 and covariance operator Q for each $i = 1, ..., n$, the distribution of the real variable $\frac{\lambda}{n}\left\langle \varepsilon_i, g_k^{s+1}(X_i) - g_k^{s+1,s}(X_i) \right\rangle_{\mathcal{Y}}$ conditioned on $\mathcal{F}_n$ is real Gaussian with mean 0 and variance

$$\frac{\lambda^2}{n^2}\mathbb{E}\left[\left\langle g_k^{s+1}(X_i) - g_k^{s+1,s}(X_i), \varepsilon_i \right\rangle_{\mathcal{Y}}^2 \mid \mathcal{F}_n \right] \le \frac{\lambda^2}{n^2} \left\| g_k^{s+1}(X_i) - g_k^{s+1,s}(X_i) \right\|_{\mathcal{Y}}^2,$$

which follows from the Cauchy-Schwarz inequality and the fact that $\mathbb{E}\left[\|\varepsilon_i\|_{\mathcal{Y}}^2 \right] = \mathrm{Tr}Q = 1$. Hence,

$$\begin{aligned}
&\mathbb{P}\left(\left\langle \varepsilon, g_k^{s+1} - g_k^{s+1,s} \right\rangle_{2,P_n} \ge 2^{-s} R_n \sqrt{\frac{2t}{n}} \mid \mathcal{F}_n \right) \\
&\le e^{-\lambda 2^{-s} R_n \sqrt{\frac{2t}{n}}} \prod_{i=1}^{n} e^{\frac{\lambda^2}{2n^2}\left\| g_k^{s+1}(X_i) - g_k^{s+1,s}(X_i) \right\|_{\mathcal{Y}}^2} \\
&= e^{-\lambda 2^{-s} R_n \sqrt{\frac{2t}{n}}} e^{\frac{\lambda^2}{2n^2}\sum_{i=1}^{n}\left\| g_k^{s+1}(X_i) - g_k^{s+1,s}(X_i) \right\|_{\mathcal{Y}}^2} \\
&= e^{-\lambda 2^{-s} R_n \sqrt{\frac{2t}{n}}} e^{\frac{\lambda^2}{2n}\left\| g_k^{s+1} - g_k^{s+1,s} \right\|_{2,P_n}^2} \\
&\le e^{-\lambda 2^{-s} R_n \sqrt{\frac{2t}{n}}} e^{\frac{\lambda^2}{2n}\left(2^{-s} R_n\right)^2}.
\end{aligned}$$

Now let $\lambda = \frac{\sqrt{2nt}}{2^{-s} R_n}$ to see that

$$\mathbb{P}\left(\left\langle \varepsilon, g_k^{s+1} - g_k^{s+1,s} \right\rangle_{2,P_n} \ge 2^{-s} R_n \sqrt{\frac{2t}{n}} \mid \mathcal{F}_n \right) \le e^{-t}.$$

Therefore, by the union bound, for each $s = 0, 1, ..., S$ and all $t > 0$,

$$\mathbb{P}\left(\max_{k \in \{1, ..., N_{s+1}\}} \left\langle \varepsilon, g_k^{s+1} - g_k^{s+1,s} \right\rangle_{2,P_n} \ge 2^{-s} R_n \sqrt{\frac{2(t + H_{s+1})}{n}} \mid \mathcal{F}_n \right) \le e^{-t}.$$

Fix t and for $s = 0, 1, ..., S$, let

$$\alpha_s := 2^{-s} R_n \left(\sqrt{2H_{s+1}} + \sqrt{2(1+s)(1+t)} \right) \ge 2^{-s} R_n \sqrt{2\left(H_{s+1} + (1+s)(1+t) \right)}$$

using $\sqrt{a} + \sqrt{b} \ge \sqrt{a+b}$. Then using $\sum_{s=0}^{\infty} 2^{-s}\sqrt{2(1+s)} \le 4$,

$$\sum_{s=0}^{\infty} \alpha_s = J_n + \sum_{s=0}^{\infty} 2^{-s} R_n \sqrt{2(1+s)(1+t)} \le J_n + 4R_n\sqrt{1+t}.$$

Then

$$\mathbb{P}\left(\sup_{g \in \mathcal{G}} \sum_{s=0}^{S} \left\langle \varepsilon, g^{s+1} - g^{s} \right\rangle_{2,P_n} \ge \frac{J_n}{\sqrt{n}} + 4R_n \sqrt{\frac{1+t}{n}} \mid \mathcal{F}_n \right)$$

$$
\begin{aligned}
&\leq \mathbb{P}\left(\sum_{s=0}^{S} \sup_{g\in\mathcal{G}} \left\langle \varepsilon, g^{s+1} - g^{s} \right\rangle_{2,P_n} \geq \frac{1}{\sqrt{n}} \sum_{s=0}^{S} \alpha_s \mid \mathcal{F}_n \right) \\
&\leq \sum_{s=0}^{S} \mathbb{P}\left(\sup_{g\in\mathcal{G}} \left\langle \varepsilon, g^{s+1} - g^{s} \right\rangle_{2,P_n} \geq \frac{1}{\sqrt{n}} \alpha_s \mid \mathcal{F}_n \right) \\
&= \sum_{s=0}^{S} \mathbb{P}\left(\max_{k=1,\dots,N_{s+1}} \left\langle \varepsilon, g_k^{s+1} - g_k^{s+1,s} \right\rangle_{2,P_n} \geq \frac{1}{\sqrt{n}} \alpha_s \mid \mathcal{F}_n \right) \\
&\leq \sum_{s=0}^{S} e^{-(1+s)(1+t)} \\
&\leq e^{-t}.
\end{aligned}
$$

■

Recall from Definition 3 the empirical process, $\left\{\nu_n(g) = \sqrt{n}\,(P_n - P)\,g : g \in \mathcal{G}\right\}$. Under additional conditions, we can use the previous lemma to show its asymptotic equicontinuity. We continue to assume that the envelope $G = \sup_{g\in\mathcal{G}} \|g\|_{\mathcal{Y}}$ satisfies $G \in L^2(\mathcal{X}, P; \mathbb{R})$.

Theorem 26 *Suppose that $\mathcal{G}$ satisfies the "uniform entropy condition", i.e. there exists a decreasing function $H : \mathbb{R} \to \mathbb{R}$ satisfying*

$$\int_0^1 \sqrt{H(u)}du < \infty$$

such that, for all $u > 0$ and any probability distribution Q with finite support,

$$H(u\,\|G\|_{2,Q}\,, \mathcal{G}, \|\cdot\|_{2,Q}) \leq H(u).$$

Then the empirical process ν_n is asymptotically equicontinuous.

Proof Take any arbitrary $g_0 \in \mathcal{G}$. We will show that ν_n is asymptotically equicontinuous at g_0. Take arbitrary $\epsilon_1, \epsilon_2 > 0$, and fix $S \in \mathbb{N}$. Define, for $\delta > 0$, the closed δ-ball around the origin:

$$\mathcal{B}(\delta) := \left\{ g \in \mathcal{G} : \|g\|_{2,P} \leq \delta \right\}.$$

Then clearly, the theoretical radius of $\mathcal{B}(\delta)$ is $\sup_{g\in\mathcal{B}(\delta)} \|g\|_{2,P} = \delta$. Denote the empirical radius of $\mathcal{B}(\delta)$ by $R_{n,\delta} = \sup_{g\in\mathcal{B}(\delta)} \|g\|_{2,P_n}$, and analogously to the proof of Proposition 25, define

$$J_{n,\delta} := \sum_{s=0}^{S} 2^{-s} R_{n,\delta} \sqrt{2H\left(2^{-(s+1)} R_{n,\delta}, \mathcal{B}(\delta), \|\cdot\|_{2,P_n}\right)}.$$

Also define

$$\mathcal{J}(\rho) := 8 \int_0^{\rho} \sqrt{2H(u)}du, \qquad \rho > 0,$$

which is bounded for any finite $\rho > 0$, by the uniform entropy condition.

Define $A \in \mathcal{F}$ as the event on which $R_{n,\delta} \leq 2\delta$ and $\|G\|_{2,P_n} \leq 2\|G\|_{2,P}$. Then on this event, we have

$$\begin{aligned}
J_{n,\delta} &= \sum_{s=0}^{S} 2^{-s} R_{n,\delta} \sqrt{2H\left(2^{-(s+1)}R_{n,\delta}, \mathcal{B}(\delta), \|\cdot\|_{2,P_n}\right)} \\
&\leq 4 \int_0^{R_{n,\delta}} \sqrt{2H(u, \mathcal{B}(\delta), \|\cdot\|_{2,P_n})} du \\
&\leq 4 \int_0^{2\delta} \sqrt{2H\left(u, \mathcal{G}, \|\cdot\|_{2,P_n}\right)} du \qquad \text{since } R_{n,\delta} \leq 2\delta \text{ on } A \text{ and } \mathcal{B}(\delta) \subseteq \mathcal{G} \\
&\leq 4 \int_0^{2\delta} \sqrt{2H\left(\frac{u}{\|G\|_{2,P_n}}\right)} du \qquad \text{by the uniform entropy condition} \\
&\leq 4 \int_0^{2\delta} \sqrt{2H\left(\frac{u}{2\|G\|_{2,P}}\right)} du \qquad \text{since } \|G\|_{2,P_n} \leq 2\|G\|_{2,P} \text{ on } A \text{ and } H \text{ is decreasing.} \\
&= 8\|G\|_{2,P} \int_0^{\frac{\delta}{\|G\|_{2,P}}} \sqrt{2H(u)} du \qquad \text{by substitution} \\
&= \|G\|_{2,P} \mathcal{J}\left(\frac{\delta}{\|G\|_{2,P}}\right).
\end{aligned}$$

On A, we also have

$$\begin{aligned}
\sup_{g \in \mathcal{B}(\delta)} \left\| P_n^\sigma\left(g - g^{S+1}\right)\right\|_{\mathcal{Y}} &\leq \sup_{g \in \mathcal{B}(\delta)} \left\| g - g^{S+1}\right\|_{1,P_n} \\
&\leq \sup_{g \in \mathcal{B}(\delta)} \left\| g - g^{S+1}\right\|_{2,P_n} \\
&\leq 2^{-(S+1)} R_{n,\delta} \\
&\leq 2^{-S}\delta. \qquad (*)
\end{aligned}$$

So on A, noting that

$$\begin{aligned}
\|P_n^\sigma\|_{\mathcal{B}(\delta)} &= \sup_{g \in \mathcal{B}(\delta)} \left\| P_n^\sigma\left(g - g^{S+1}\right) + \sum_{s=0}^{S} P_n^\sigma\left(g^{s+1} - g^s\right)\right\|_{\mathcal{Y}} \\
&\leq \sup_{g \in \mathcal{B}(\delta)} \left\| P_n^\sigma\left(g - g^{S+1}\right)\right\|_{\mathcal{Y}} + \sup_{g \in \mathcal{B}(\delta)} \left\| \sum_{s=0}^{S} P_n^\sigma\left(g^{s+1} - g^s\right)\right\|_{\mathcal{Y}},
\end{aligned}$$

we have, for all $t > 0$,

$$\mathbb{P}\left(\|P_n^\sigma\|_{\mathcal{B}(\delta)} \geq \frac{\sqrt{2}\|G\|_{2,P}\mathcal{J}\left(\frac{\delta}{\|G\|_{2,P}}\right)}{\sqrt{n}} + 12\delta\sqrt{\frac{1+t}{n}} + 2^{-S}\delta \mid \mathcal{F}_n\right)$$

$$\begin{aligned}
&= \mathbb{P}\left(\sup_{g\in\mathcal{B}(\delta)} \left\| P_n^\sigma \left(g - g^{S+1}\right)\right\|_{\mathcal{Y}} + \sup_{g\in\mathcal{B}(\delta)} \left\| \sum_{s=0}^{S} P_n^\sigma \left(g^{s+1} - g^{s}\right)\right\|_{\mathcal{Y}} \right. \\
&\qquad\qquad \left. \geq \frac{\sqrt{2}J_{n,\delta}}{n} + 6R_{n,\delta}\sqrt{\frac{1+t}{n}} + 2^{-S}\delta \mid \mathcal{F}_n \right) \\
&\leq \mathbb{P}\left(\sup_{g\in\mathcal{B}(\delta)} \left\| \sum_{s=0}^{S} P_n^\sigma \left(g^{s+1} - g^{s}\right)\right\|_{\mathcal{Y}} \geq \frac{\sqrt{2}J_{n,\delta}}{\sqrt{n}} + 6R_{n,\delta}\sqrt{\frac{1+t}{n}} \mid \mathcal{F}_n \right) \\
&\leq 2e^{-t},
\end{aligned}$$

where the term $\mathbb{P}\left(\sup_{g\in\mathcal{B}(\delta)} \left\| P_n^\sigma \left(g - g^{S+1}\right)\right\|_{\mathcal{Y}} \geq 2^{-S}\delta \mid \mathcal{F}_n \right)$ vanishes by (*) and the last inequality follows Proposition 25(i). Then we can de-symmetrise using Lemma 21:

$$\begin{aligned}
&\mathbb{P}\left(\|P_n - P\|_{\mathcal{B}(\delta)} \geq \frac{4\sqrt{2}\,\|G\|_{2,P}\,\mathcal{J}\left(\frac{\delta}{\|G\|_{2,P}}\right)}{\sqrt{n}} + 48\delta\sqrt{\frac{1+t}{n}} + 2^{-(S-2)}\delta \right) \\
&\leq 4\mathbb{P}\left(\|P_n^\sigma\|_{\mathcal{B}(\delta)} \geq \frac{\sqrt{2}\,\|G\|_{2,P}\,\mathcal{J}\left(\frac{\delta}{\|G\|_{2,P}}\right)}{\sqrt{n}} + 12\delta\sqrt{\frac{1+t}{n}} + 2^{-S}\delta \right) \\
&\leq 8e^{-t} + 4\mathbb{P}\left(R_{n,\delta} > 2\delta \text{ or } \|G\|_{2,P_n} > 2\,\|G\|_{2,P} \right) \\
&= 8e^{-t} + 4\mathbb{P}\left(\sup_{g\in\mathcal{B}(\delta)\cup\{G\}} \|g\|_{2,P_n}^2 > 4 \sup_{g\in\mathcal{B}(\delta)\cup\{G\}} \|g\|_{2,P}^2 \right).
\end{aligned}$$

Now let $t = \log\left(\frac{8}{\epsilon_2}\right)$ and S large enough such that $2^{-(S-2)} \leq \frac{1}{\sqrt{n}}$, and δ small enough such that

$$4\sqrt{2}\,\|G\|_{2,P}\,\mathcal{J}\left(\frac{\delta}{\|G\|_{2,P}}\right) + 48\delta\sqrt{1 + \log\left(\frac{8}{\epsilon_2}\right)} + \delta \leq \epsilon_1.$$

Then

$$\mathbb{P}\left(\sqrt{n}\,\|P_n - P\|_{\mathcal{B}(\delta)} > \epsilon_1\right) \leq \epsilon_2 + 4\mathbb{P}\left(\sup_{g\in\mathcal{B}(\delta)\cup\{G\}} \|g\|_{2,P_n}^2 > 4 \sup_{g\in\mathcal{B}(\delta)\cup\{G\}} \|g\|_{2,P}^2 \right).$$

Hence, for any $g \in \mathcal{G}$ such that $\|g - g_0\|_{2,P} \leq \delta$,

$$\begin{aligned}
\mathbb{P}\left(\left\|\nu_n(g) - \nu_n(g_0)\right\|_{\mathcal{Y}} > \epsilon_1\right) &= \mathbb{P}\left(\sqrt{n}\,\left\|(P_n - P)\,(g - g_0)\right\|_{\mathcal{Y}} > \epsilon_1\right) \\
&\leq \mathbb{P}\left(\sqrt{n}\,\|P_n - P\|_{\mathcal{B}(\delta)} > \epsilon_1\right)
\end{aligned}$$

$$\leq \epsilon_2 + 4\mathbb{P}\left(\sup_{g \in \mathcal{B}(\delta)\cup\{G\}} \|g\|^2_{2,P_n} > 4 \sup_{g \in \mathcal{B}(\delta)\cup\{G\}} \|g\|^2_{2,P} \right).$$

Here, by the uniform law of large numbers on $\mathcal{B}(\delta) \cap \{G\}$ (Theorem 24), the second term converges to 0 as $n \to \infty$. Hence, as ϵ_1 and ϵ_2 were arbitrary, we have asymptotic equicontinuity. ■

C.4. Peeling and Least-Squares Regression with Fixed Design and Gaussian Noise

Theorem 27 *Suppose that $\varepsilon_1, ..., \varepsilon_n$ are i.i.d. with Gaussian distribution with mean 0 and covariance operator Q (c.f. Definition 14 and Lemmas 15 and 16), and that* $\operatorname{Tr} Q = 1$. *Further, suppose that*

$$J(\delta) := 4\int_0^\delta \sqrt{2H(u, \mathcal{B}_{2,P_n}(\delta), \|\cdot\|_{2,P_n})}du < \infty, \quad \textit{for each } \delta > 0, \textit{ and } \frac{J(\delta)}{\delta^2} \textit{ is decreasing in } \delta,$$

where $\mathcal{B}_{2,P_n}(\delta) := \{g \in \mathcal{G} : \|g\|_{2,P_n} \leq \delta\}$. Then for all $t \geq \frac{3}{8}$ and all δ_n satisfying

$$\sqrt{n}\delta_n^2 \geq 8\left(J(\delta_n) + 4\delta_n\sqrt{1+t} + \delta_n\sqrt{\frac{8}{3}t} \right),$$

we have

$$\mathbb{P}\left(\|\hat{g}_n - g_0\|_{2,P_n} > \delta_n \right) \leq \left(1 + \frac{2}{e-1} \right) e^{-t}.$$

Proof First, recall the notation

$$\langle \varepsilon, g \rangle_{2,P_n} = \frac{1}{n}\sum_{i=1}^n \langle \varepsilon_i, g(X_i) \rangle_{\mathcal{Y}}$$

from Proposition 25(ii), and note that we have the following basic inequality

$$\|\hat{g}_n - g_0\|^2_{2,P_n} \leq 2\langle \varepsilon, \hat{g}_n - g_0 \rangle_{2,P_n}, \tag{*}$$

which follows from the fact that $\hat{g}_n$ minimises $\left\|Y_i - g(X_i)\right\|^2_{2,P_n}$ over $g \in \mathcal{G}$, giving

$$\left\|\varepsilon_i - (g_0 - \hat{g}_n)\right\|^2_{2,P_n} = \left\|Y_i - \hat{g}_n(X_i)\right\|^2_{2,P_n} \leq \left\|Y_i - g_0(X_i)\right\|^2_{2,P_n} = \|\varepsilon_i\|^2_{2,P_n}.$$

We use a technique called the "peeling device", first introduced by van de Geer (2000). See that

$$\begin{aligned}
\mathbb{P}\left(\|\hat{g}_n - g_0\|_{2,P_n} > \delta_n \right) &= \mathbb{P}\left(\bigcup_{j=1}^\infty \left\{ 2^{j-1}\delta_n < \|\hat{g}_n - g_0\|_{2,P_n} \leq 2^j\delta_n \right\} \right) \\
&\leq \sum_{j=1}^\infty \mathbb{P}\left(2^{j-1}\delta_n < \|\hat{g}_n - g_0\|_{2,P_n} \leq 2^j\delta_n \right) \qquad \text{by the union bound} \\
&= \sum_{j=1}^\infty \mathbb{P}\left(\left\{ 2^{j-1}\delta_n < \|\hat{g}_n - g_0\|_{2,P_n} \right\} \bigcap \left\{ \hat{g}_n - g_0 \in \mathcal{B}_n(2^j\delta_n) \right\} \right)
\end{aligned}$$

$$
\begin{aligned}
&\le \sum_{j=1}^{\infty} \mathbb{P}\left(\left\{\left(2^{j-1}\delta_n\right)^2 < 2\left\langle \varepsilon, \hat{g}_n - g_0\right\rangle_{2,P_n}\right\} \bigcap \left\{\hat{g}_n - g_0 \in \mathcal{B}_n(2^j\delta_n)\right\}\right) \qquad \text{by (*)}\\
&\le \sum_{j=1}^{\infty} \mathbb{P}\left(\sup_{g\in\mathcal{B}_n(2^j\delta_n)} 2\left\langle \varepsilon, g\right\rangle_{2,P_n} > \left(2^{j-1}\delta_n\right)^2\right)\\
&= \sum_{j=1}^{\infty} \mathbb{P}\left(\sup_{g\in\mathcal{B}_n(2^j\delta_n)} \left\langle \varepsilon, g\right\rangle_{2,P_n} > \frac{1}{8}\left(2^{j}\delta_n\right)^2\right).
\end{aligned}
$$

Now, applying the hypothesis on δ_n, we see that, for each j,

$$
\begin{aligned}
\frac{1}{8}\left(2^j\delta_n\right)^2 &\ge \frac{(2^j)^2 J(\delta_n)}{\sqrt{n}} + 4(2^j)^2\delta_n\sqrt{\frac{1+t}{n}} + \frac{\sqrt{\frac{8}{3}}t(2^j)^2\delta_n}{\sqrt{n}}\\
&\ge \frac{J(2^j\delta_n)}{\sqrt{n}} + 4(2^j\delta_n)\sqrt{\frac{1+t+j}{n}} + \frac{\sqrt{\frac{8}{3}}t(2^j)^2\delta_n}{\sqrt{n}}\\
&\ge \frac{J_n}{\sqrt{n}} + 4(2^j\delta_n)\sqrt{\frac{1+t+j}{n}} + \frac{\sqrt{\frac{8}{3}}t(2^j)^2\delta_n}{\sqrt{n}}
\end{aligned}
$$

where we used the fact that $\frac{J(\delta)}{\delta^2}$ is decreasing in δ and $\sqrt{1+t+j} \le 2^j\sqrt{1+t}$, and J_n is defined as in Proposition 25 with $\mathcal{G} = \mathcal{B}_n(2^j\delta_n)$ and $R_n = 2^j\delta_n$. On the other hand, we can write, for any $S \in \mathbb{N}$,

$$
\left\langle \varepsilon, g\right\rangle_{2,P_n} = \left\langle \varepsilon, g - g^{S+1}\right\rangle_{2,P_n} + \sum_{s=0}^{S}\left\langle \varepsilon, g^{s+1} - g^{s}\right\rangle_{2,P_n},
$$

using the chaining notation in Section C.3. Hence,

$$
\begin{aligned}
&\mathbb{P}\left(\left\|\hat{g}_n - g_0\right\|_{2,P_n} > \delta_n\right)\\
&\le \sum_{j=1}^{\infty} \mathbb{P}\left(\sup_{g\in\mathcal{B}_n(2^j\delta_n)} \left\langle \varepsilon, g - g^{S+1}\right\rangle_{2,P_n} > \frac{\sqrt{\frac{8}{3}}t(2^j)^2\delta_n}{\sqrt{n}}\right)\\
&\quad + \sum_{j=1}^{\infty} \mathbb{P}\left(\sup_{g\in\mathcal{B}_n(2^j\delta_n)} \sum_{s=0}^{S}\left\langle \varepsilon, g^{s+1} - g^{s}\right\rangle_{2,P_n} > \frac{J_n}{\sqrt{n}} + 4(2^j\delta_n)\sqrt{\frac{1+t+j}{n}}\right)\\
&\le \sum_{j=1}^{\infty} \mathbb{P}\left(\frac{2^j}{2^{S+1}}\delta_n\left\|\varepsilon\right\|_{2,P_n} > \frac{\sqrt{\frac{8}{3}}t2^{2j}\delta_n}{\sqrt{n}}\right) + \sum_{j=1}^{\infty} e^{-(t+j)} \quad \text{by Proposition 25(ii)}\\
&= \sum_{j=1}^{\infty} \mathbb{P}\left(\left\|\varepsilon\right\|_{2,P_n} > 2^j\sqrt{\frac{8}{3}t}\right) + \frac{1}{e-1}e^{-t} \quad \text{letting } S \text{ such that } \sqrt{n} \le 2^{S+1}\\
&\le \sum_{j=1}^{\infty} \mathbb{P}\left(\left\|\varepsilon\right\|_{2,P_n} > 2^j + \sqrt{\frac{8}{3}t}\right) + \frac{1}{e-1}e^{-t} \quad \text{since } t \ge \frac{3}{8}
\end{aligned}
$$

$$
\begin{aligned}
&\leq \sum_{j=1}^{\infty} \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\|\varepsilon_i\|_{\mathcal{Y}}^2 > 2^{2j} + \frac{8}{3}t\right) + \frac{1}{e-1}e^{-t} \\
&\leq \sum_{j=1}^{\infty} e^{-\frac{3}{8}2^{2j}-t}\mathbb{E}\left[e^{\frac{3}{8}\frac{1}{n}\sum_{i=1}^{n}\|\varepsilon_i\|_{\mathcal{Y}}^2}\right] + \frac{1}{e-1}e^{-t} \qquad \text{by Markov's inequality} \\
&\leq \sum_{j=1}^{\infty} e^{-\frac{3}{8}2^{2j}-t}\prod_{i=1}^{n}\mathbb{E}\left[e^{\frac{3}{8}\frac{1}{n}\|\varepsilon_i\|_{\mathcal{Y}}^2}\right] + \frac{1}{e-1}e^{-t} \qquad \text{by independence} \\
&\leq \sum_{j=1}^{\infty} e^{-\frac{3}{8}2^{2j}-t}\mathbb{E}\left[e^{\frac{3}{8}\|\varepsilon_1\|_{\mathcal{Y}}^2}\right] + \frac{1}{e-1}e^{-t} \qquad \text{by Jensen's inequality} \\
&\leq \sum_{j=1}^{\infty} e^{-\frac{3}{8}2^{2j}-t} + \frac{1}{e-1}e^{-t} \qquad \text{by Proposition 17} \\
&\leq e^{-t}\left(e^{-\frac{3}{4}} - e^{-1} + \sum_{j=1}^{\infty} e^{-j} + \frac{1}{e-1}\right) \\
&\leq e^{-t}\left(1 + \frac{2}{e-1}\right).
\end{aligned}
$$

■

C.5. Empirical Risk Minimisation with Lipschitz Loss and Random Design

In this Section, we discuss the setting where we have an L-bounded, c-Lipschitz loss function $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. Suppose we have a given class $\mathcal{G}$ of functions $\mathcal{X} \to \mathcal{Y}$. Then given samples $(X_1, Y_1), ..., (X_n, Y_n)$, the empirical risk minimiser, which we assume exists, is given by

$$\hat{g}_n = \operatorname*{arg\,min}_{g\in\mathcal{G}} \hat{\mathcal{R}}_n(g), \qquad \hat{\mathcal{R}}_n(g) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(Y_i, g(X_i)).$$

We are interested in the convergence of $\hat{g}_n$ to the population risk minimiser,

$$g^* = \operatorname*{arg\,min}_{g\in\mathcal{G}} \mathcal{R}(g), \qquad \mathcal{R}(g) = \mathbb{E}[\mathcal{L}(Y, g(X))],$$

in terms of the population risk $\mathcal{R}$. First, see that

$$
\begin{aligned}
\mathcal{R}(\hat{g}_n) - \mathcal{R}(g^*) &= \mathcal{R}(\hat{g}_n) - \hat{\mathcal{R}}(\hat{g}_n) + \hat{\mathcal{R}}(\hat{g}_n) - \hat{\mathcal{R}}(g^*) + \hat{\mathcal{R}}(g^*) - \mathcal{R}(g^*) \\
&\leq \sup_{g\in\mathcal{G}}\left|\mathcal{R}(g) - \hat{\mathcal{R}}(g)\right| + \hat{\mathcal{R}}(g^*) - \mathcal{R}(g^*),
\end{aligned}
$$

where, going from the first line to the second, the first two terms on the right-hand side were bounded by the supremum over the whole function class $\mathcal{G}$ (since, although $\hat{g}_n$ varies as the samples and the size n of the dataset vary, it always lives in $\mathcal{G}$), the middle two terms were bounded above by 0 since the empirical risk minimiser $\hat{g}_n$ minimises $\hat{\mathcal{R}}$, and the last two terms remain unchanged.

Theorem 28 *Suppose the following uniform entropy condition is satisfied: there exists some function $H : \mathbb{R} \to \mathbb{R}$ satisfying*

$$\mathcal{J}(1) := 4 \int_0^1 \sqrt{2H(u)} du < \infty,$$

such that, for all $u > 0$ and any probability distribution Q with finite support,

$$H(uL, \mathcal{L} \circ \mathcal{G}, \|\cdot\|_{2,Q}) \leq H(u).$$

Then

$$\mathbb{P}\left(\sup_{g\in\mathcal{G}} \left|\mathcal{R}(g) - \hat{\mathcal{R}}(g)\right| > \frac{4\sqrt{2}L\mathcal{J}(1)}{\sqrt{n}} + 24L\sqrt{\frac{1+t}{n}} + \frac{L}{\sqrt{n}}\right) \leq 2e^{-t}.$$

Proof First, denote by $\mathcal{L} \circ \mathcal{G}$ the class of functions $\mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ given by $(x, y) \mapsto \mathcal{L}(y, g(x))$ for $g \in \mathcal{G}$. Also, by an abuse of notation, for each $g \in \mathcal{G}$, denote by $\mathcal{L} \circ g$ the function $(x, y) \mapsto \mathcal{L}(y, g(x))$. Then we have

$$P\mathcal{L} \circ g = \mathcal{R}(g), \qquad P_n\mathcal{L} \circ g = \hat{\mathcal{R}}(g).$$

Since the loss $\mathcal{L}$ is bounded above by L, the empirical radius and the theoretical radius of $\mathcal{L} \circ \mathcal{G}$ are both bounded above by L. In the chaining notation of Section C.3, define

$$J_n = \sum_{s=0}^{S} 2^{-s}L\sqrt{2H(2^{-(s+1)}L, \mathcal{L} \circ \mathcal{G}, \|\cdot\|_{2,P_n})}.$$

Then from the very definition of the chains, we have

$$\begin{aligned} \sup_{g\in\mathcal{G}} \left|P_n^\sigma\left(\mathcal{L} \circ g - \mathcal{L} \circ g^{S+1}\right)\right| &\leq \sup_{g\in\mathcal{G}} \left\|\mathcal{L} \circ g - \mathcal{L} \circ g^{S+1}\right\|_{1,P_n} \\ &\leq \sup_{g\in\mathcal{G}} \left\|\mathcal{L} \circ g - \mathcal{L} \circ g^{S+1}\right\|_{2,P_n} \\ &\leq 2^{-(S+1)}L. \end{aligned} \tag{*}$$

First, see that

$$\begin{aligned} J_n &= \sum_{s=0}^{S} 2^{-s}L\sqrt{2H(2^{-(s+1)}L, \mathcal{L} \circ \mathcal{G}, \|\cdot\|_{2,P_n})} \\ &\leq 4\int_0^L \sqrt{2H(u, \mathcal{L} \circ \mathcal{G}, \|\cdot\|_{2,P_n})}du \\ &\leq 4\int_0^L \sqrt{2H\left(\frac{u}{L}\right)}du && \text{by the uniform entropy condition} \\ &= 4L\int_0^1 \sqrt{2H(u)}du && \text{by substitution} \\ &= L\mathcal{J}(1). \end{aligned} \tag{**}$$

Then, by the symmetrisation lemma (Lemma 21) followed by chaining (Proposition 25), we have

$$\begin{aligned}
&\mathbb{P}\left(\sup_{g\in\mathcal{G}}\left|\mathcal{R}(g)-\hat{\mathcal{R}}(g)\right| > \frac{4\sqrt{2}L\mathcal{J}(1)}{\sqrt{n}} + 24L\sqrt{\frac{1+t}{n}} + 2^{-(S-1)}L\right)\\
&= \mathbb{P}\left(\|P-P_n\|_{\mathcal{L}\circ\mathcal{G}} > \frac{4\sqrt{2}J_n}{\sqrt{n}} + 24L\sqrt{\frac{1+t}{n}} + 2^{-(S-1)}L\right) \qquad \text{by (**)}\\
&\leq 4\mathbb{P}\left(\|P_n^\sigma\|_{\mathcal{L}\circ\mathcal{G}} > \frac{\sqrt{2}J_n}{\sqrt{n}} + 6L\sqrt{\frac{1+t}{n}} + 2^{-(S+1)}L\right) \qquad \text{by symmetrisation (Lemma 21)}\\
&\leq 4\mathbb{P}\left(\sup_{g\in\mathcal{G}}\left|P_n^\sigma\left(\mathcal{L}\circ g - \mathcal{L}\circ g^{S+1}\right)\right| + \sup_{g\in\mathcal{G}}\left|\sum_{s=0}^{S} P_n^\sigma\left(\mathcal{L}\circ g^{s+1} - \mathcal{L}\circ g^{s}\right)\right|\right.\\
&\qquad\qquad\qquad \left. > \frac{\sqrt{2}J_n}{\sqrt{n}} + 6L\sqrt{\frac{1+t}{n}} + 2^{-(S+1)}L\right)\\
&\leq 4\mathbb{P}\left(\sup_{g\in\mathcal{G}}\left|\sum_{s=0}^{S} P_n^\sigma\left(\mathcal{L}\circ g^{s+1} - \mathcal{L}\circ g^{s}\right)\right| > \frac{\sqrt{2}J_n}{\sqrt{n}} + 6L\sqrt{\frac{1+t}{n}}\right)\\
&\qquad + 4\mathbb{P}\left(\sup_{g\in\mathcal{G}}\left|P_n^\sigma\left(\mathcal{L}\circ g - \mathcal{L}\circ g^{S+1}\right)\right| > 2^{-(S+1)}L\right) \qquad \text{by the union bound}\\
&\leq 2e^{-t},
\end{aligned}$$

where the second term disappears by (*) and the first term is bounded by Proposition 25(i). Now letting S be large enough such that $\sqrt{n} \leq 2^{S+1}$,

$$\mathbb{P}\left(\sup_{g\in\mathcal{G}}\left|\mathcal{R}(g)-\hat{\mathcal{R}}(g)\right| > \frac{4\sqrt{2}L\mathcal{J}(1)}{\sqrt{n}} + 24L\sqrt{\frac{1+t}{n}} + \frac{L}{\sqrt{n}}\right) \leq 2e^{-t}.$$

■

C.6. Rademacher Complexities

In this Section, we discuss the extension of the concept of Rademacher complexities to classes of vector-valued functions in more depth (c.f. Section 4.3). We first give the definition of Rademacher complexities of classes of real-valued functions.

Definition 29 (Bartlett and Mendelson (2002, Definition 2)) *Suppose $\mathcal{G}$ is a class of real-valued functions $\mathcal{X} \to \mathbb{R}$. Then the empirical (or conditional) Rademacher complexity of $\mathcal{G}$ is defined as*

$$\hat{\mathfrak{R}}_n(\mathcal{G}) = \mathbb{E}\left[\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i g(X_i)\right| \mid X_1, ..., X_n\right],$$

where the expectation is taken with respect to the Rademacher variables $\{\sigma_i\}_{i=1}^n$. The Rademacher complexity of $\mathcal{G}$ is defined as

$$\mathfrak{R}_n(G) = \mathbb{E}\left[\hat{\mathfrak{R}}_n(G)\right].$$

Since this seminal definition, it was realised that the absolute value around $\frac{1}{n}\sum_{i=1}^{n}\sigma_i g(X_i)$ was unnecessary (see, for example, Meir and Zhang (2003, paragraph between Corollary 4 and Lemma 5) or Maurer (2016, last paragraph of Section 1)). However, in order to facilitate the following direct extension to classes of vector-valued functions, we retain the absolute value sign.

Definition 30 *Suppose $\mathcal{G}$ is a class of $\mathcal{X} \to \mathcal{Y}$ functions. Then the empirical (or conditional) Rademacher complexity of $\mathcal{G}$ is defined as*

$$\hat{\mathfrak{R}}_n(\mathcal{G}) = \mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\|\frac{1}{n}\sum_{i=1}^{n}\sigma_i g(X_i)\right\|_{\mathcal{Y}} \mid X_1, ..., X_n\right] = \mathbb{E}\left[\|P_n^\sigma g\|_{\mathcal{G}} \mid X_1, ..., X_n\right],$$

using the notation from Section 2. The Rademacher complexity of $\mathcal{G}$ is defined as

$$\mathfrak{R}_n(G) = \mathbb{E}\left[\hat{\mathfrak{R}}_n(G)\right].$$

Note that our definition is different to the “vector-valued Rademacher complexity” already in use in the literature, mostly for $\mathcal{Y}$ being a finite-dimensional Euclidean space (Yousefi et al., 2018, Definition 1; Li et al., 2019, Definition 3), but also for $\mathcal{Y} = l_2$, the space of square-summable sequences (Maurer, 2016). These papers define the “Rademacher complexity” of vector-valued function classes not as in Definition 30, where we have one Rademacher variable σ_i per sample X_i, but introduce a Rademacher variable for every coordinate of $\mathcal{Y}$. The resulting quantity looks something like

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sum_{k}\sigma_i^k g_k(X_i) \mid X_1, ..., X_n\right],$$

where g_k is the k^{th} coordinate of g with respect to a basis, and $\{\sigma_i^k\}_{i,k}$ are Rademacher random variables. For convenience, in what follows, we call this the “coordinate-wise Rademacher complexity”, and denote it by $\hat{\mathfrak{R}}_n^{\text{coord}}(\mathcal{G})$.

While we recognise the usefulness of this definition, especially thanks to the contraction result shown in Maurer (2016), Cortes et al. (2016), Zatarain-Vera (2019) and Foster and Rakhlin (2019), for several reasons, we insist on using Definition 30. Firstly, as it is clear from the definition, and as admitted by Maurer (2016, paragraph just above Conjecture 2), Definition 30 is a more natural definition in view of the real-valued Rademacher complexity. Moreover, our work in Section C.1 uses the empirical symmetrised measure $\frac{1}{n}\sum_{i=1}^{n}\sigma_i\delta_{X_i}$ to good effect and in a way that directly generalises from the real-valued case, which suggests that Definition 30 is natural. Finally, and perhaps most critically, the coordinate-wise Rademacher complexity is not independent of the choice of the basis of $\mathcal{Y}$. For a simple counterexample, let $\mathcal{X} = \mathcal{Y} = \mathbb{R}^2$, and $\mathcal{G} = \{g_1, g_2\}$, where g_1 is the orthogonal projection onto the line $y = x$, and g_2 is the orthogonal projection onto the line $y = -x$. This means that, letting $X_1 = \begin{pmatrix}1\\0\end{pmatrix}$ and $X_2 = \begin{pmatrix}0\\1\end{pmatrix}$, we have

$$g_1(X_1) = \begin{pmatrix}\frac{1}{2}\\\frac{1}{2}\end{pmatrix}, \quad g_1(X_2) = \begin{pmatrix}\frac{1}{2}\\\frac{1}{2}\end{pmatrix}, \quad g_2(X_1) = \begin{pmatrix}\frac{1}{2}\\-\frac{1}{2}\end{pmatrix}, \quad g_2(X_2) = \begin{pmatrix}-\frac{1}{2}\\\frac{1}{2}\end{pmatrix}.$$

Then the coordinate-wise Rademacher complexity of $\mathcal{G}$ with respect to the standard basis $\{X_1, X_2\}$ is

$$\begin{aligned}
\hat{\mathfrak{R}}_n^{\text{coord}}(\mathcal{G}) &= \mathbb{E}\left[\sup_{g\in\mathcal{G}} \sum_{i=1}^{2}\sum_{k=1}^{2} \sigma_i^k g_k(X_i)\right] \\
&= \mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\{\sigma_1^1 \big(g(X_1)\big)_1 + \sigma_1^2\big(g(X_1)\big)_2 + \sigma_2^1\big(g(X_2)\big)_1 + \sigma_2^2\big(g(X_2)\big)_2\right\}\right] \\
&= \mathbb{E}\left[\frac{\sigma_1^1}{2} + \frac{\sigma_2^2}{2} + \sup_{g\in\mathcal{G}}\left\{\sigma_1^2\big(g(X_1)\big)_2 + \sigma_2^1\big(g(X_2)\big)_1\right\}\right] \\
&= \sup_{g\in\mathcal{G}}\left\{\big(g(X_1)\big)_2 + \big(g(X_2)\big)_1\right\} + \sup_{g\in\mathcal{G}}\left\{-\big(g(X_1)\big)_2 + \big(g(X_2)\big)_1\right\} \\
&\qquad + \sup_{g\in\mathcal{G}}\left\{\big(g(X_1)\big)_2 - \big(g(X_2)\big)_1\right\} + \sup_{g\in\mathcal{G}}\left\{-\big(g(X_1)\big)_2 - \big(g(X_2)\big)_1\right\} \\
&= 1 + 0 + 0 + 1 \\
&= 2.
\end{aligned}$$

But if we use the orthonormal basis $\left\{\begin{pmatrix}\frac{1}{\sqrt{2}}\\ \frac{1}{\sqrt{2}}\end{pmatrix}, \begin{pmatrix}-\frac{1}{\sqrt{2}}\\ \frac{1}{\sqrt{2}}\end{pmatrix}\right\}$, then we have

$$\begin{aligned}
&(g_1(X_1))_1 = \frac{1}{\sqrt{2}}, \quad (g_1(X_1))_2 = 0, \quad (g_1(X_2))_1 = \frac{1}{\sqrt{2}} \quad (g_1(X_2))_2 = 0 \\
&(g_2(X_1))_1 = 0, \qquad (g_2(X_1))_2 = -\frac{1}{\sqrt{2}}, \qquad (g_2(X_2))_1 = 0, \quad (g_2(X_2))_2 = \frac{1}{\sqrt{2}}.
\end{aligned}$$

So the complexity with respect to the standard basis $\{X_1, X_2\}$ is

$$\begin{aligned}
&\mathbb{E}\left[\sup_{g\in\mathcal{G}} \sum_{i=1}^{2}\sum_{k=1}^{2} \sigma_i^k g_k(X_i)\right] \\
&= \mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\{\sigma_1^1 \big(g(X_1)\big)_1 + \sigma_1^2\big(g(X_1)\big)_2 + \sigma_2^1\big(g(X_2)\big)_1 + \sigma_2^2\big(g(X_2)\big)_2\right\}\right] \\
&= \sup_{g\in\mathcal{G}}\left\{\big(g(X_1)\big)_1 + \big(g(X_1)\big)_2 + \big(g(X_2)\big)_1 + \big(g(X_2)\big)_2\right\} \\
&\qquad + \sup_{g\in\mathcal{G}}\left\{\big(g(X_1)\big)_1 + \big(g(X_1)\big)_2 + \big(g(X_2)\big)_1 - \big(g(X_2)\big)_2\right\} \\
&\qquad + \sup_{g\in\mathcal{G}}\left\{\big(g(X_1)\big)_1 + \big(g(X_1)\big)_2 - \big(g(X_2)\big)_1 + \big(g(X_2)\big)_2\right\} \\
&\qquad + \sup_{g\in\mathcal{G}}\left\{\big(g(X_1)\big)_1 - \big(g(X_1)\big)_2 + \big(g(X_2)\big)_1 + \big(g(X_2)\big)_2\right\} \\
&\qquad + \sup_{g\in\mathcal{G}}\left\{-\big(g(X_1)\big)_1 + \big(g(X_1)\big)_2 + \big(g(X_2)\big)_1 + \big(g(X_2)\big)_2\right\}
\end{aligned}$$

$$
\begin{aligned}
&+\sup_{g\in\mathcal{G}}\left\{\big(g(X_1)\big)_1+\big(g(X_1)\big)_2-\big(g(X_2)\big)_1-\big(g(X_2)\big)_2\right\}\\
&+\sup_{g\in\mathcal{G}}\left\{\big(g(X_1)\big)_1-\big(g(X_1)\big)_2+\big(g(X_2)\big)_1-\big(g(X_2)\big)_2\right\}\\
&+\sup_{g\in\mathcal{G}}\left\{-\big(g(X_1)\big)_1+\big(g(X_1)\big)_2+\big(g(X_2)\big)_1-\big(g(X_2)\big)_2\right\}\\
&+\sup_{g\in\mathcal{G}}\left\{\big(g(X_1)\big)_1-\big(g(X_1)\big)_2-\big(g(X_2)\big)_1+\big(g(X_2)\big)_2\right\}\\
&+\sup_{g\in\mathcal{G}}\left\{-\big(g(X_1)\big)_1+\big(g(X_1)\big)_2-\big(g(X_2)\big)_1+\big(g(X_2)\big)_2\right\}\\
&+\sup_{g\in\mathcal{G}}\left\{-\big(g(X_1)\big)_1-\big(g(X_1)\big)_2+\big(g(X_2)\big)_1+\big(g(X_2)\big)_2\right\}\\
&+\sup_{g\in\mathcal{G}}\left\{\big(g(X_1)\big)_1-\big(g(X_1)\big)_2-\big(g(X_2)\big)_1-\big(g(X_2)\big)_2\right\}\\
&+\sup_{g\in\mathcal{G}}\left\{-\big(g(X_1)\big)_1+\big(g(X_1)\big)_2-\big(g(X_2)\big)_1-\big(g(X_2)\big)_2\right\}\\
&+\sup_{g\in\mathcal{G}}\left\{-\big(g(X_1)\big)_1-\big(g(X_1)\big)_2+\big(g(X_2)\big)_1-\big(g(X_2)\big)_2\right\}\\
&+\sup_{g\in\mathcal{G}}\left\{-\big(g(X_1)\big)_1-\big(g(X_1)\big)_2-\big(g(X_2)\big)_1+\big(g(X_2)\big)_2\right\}\\
&+\sup_{g\in\mathcal{G}}\left\{-\big(g(X_1)\big)_1-\big(g(X_1)\big)_2-\big(g(X_2)\big)_1-\big(g(X_2)\big)_2\right\}\\
={}&\sqrt{2}+\sqrt{2}+0+\sqrt{2}+0+0+\sqrt{2}+0+\sqrt{2}+0+\sqrt{2}+0-\sqrt{2}+0+0+0\\
={}&5\sqrt{2}.
\end{aligned}
$$

Hence, we see that the coordinate-wise Rademacher complexity is not independent of the chosen orthonormal basis. We deem this to be a critical issue with the coordinate-wise Rademacher complexity, because it is intuitively clear that the “complexity” of a function class should not depend on the choice of the basis of the output space. This is especially pertinent in our context, considering that our interest is primarily in the case when the output space $\mathcal{Y}$ is infinite-dimensional in which there may be no “standard basis”.

One of the main ways of bounding the Rademacher complexity of real-valued function classes is to use the entropy. We show that the Rademacher complexity of vector-valued function classes $\mathcal{G}$ can be bounded using the entropy, a vector-valued analogue of Shalev-Shwartz and Ben-David (2014, p.338, Lemma 27.4). We use the chaining notation in Section C.3, and also use Proposition 11, the expectation form of vector-valued Hoeffding’s inequality.

Theorem 31 *Let $S \in \mathbb{N}$ be any (large) integer. The empirical Rademacher complexity is bounded as*

$$\hat{\mathfrak{R}}_n(\mathcal{G}) \le 2^{-(S+1)} R_n + \frac{2}{\sqrt{n}} J_n,$$

where we recall that $R_n = \sup_{g\in\mathcal{G}} \|g\|_{2,P_n}$ is the empirical radius and $J_n = \sum_{s=0}^{S} 2^{-s} R_n \sqrt{2H_{s+1}}$ is the uniform entropy bound.

Proof See that

$$\begin{aligned}
\hat{\mathfrak{R}}_n(\mathcal{G}) &= \mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\|\frac{1}{n}\sum_{i=1}^{n}\sigma_i g(X_i)\right\|_{\mathcal{Y}} \mid \mathcal{F}_n\right] \\
&= \mathbb{E}\left[\sup_{g\in\mathcal{G}}\|P_n^\sigma g\|_{\mathcal{Y}} \mid \mathcal{F}_n\right] \\
&= \mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\|P_n^\sigma\left(g-g^{S+1}\right)+\sum_{s=0}^{S}P_n^\sigma\left(g^{s+1}-g^{s}\right)\right\|_{\mathcal{Y}} \mid \mathcal{F}_n\right] \\
&\le \mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\|P_n^\sigma\left(g-g^{S+1}\right)\right\|_{\mathcal{Y}} \mid \mathcal{F}_n\right]+\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\|\sum_{s=0}^{S}P_n^\sigma\left(g^{s+1}-g^{s}\right)\right\|_{\mathcal{Y}} \mid \mathcal{F}_n\right] \\
&\le \sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\left\|g(X_i)-g^{S+1}(X_i)\right\|_{\mathcal{Y}}+\sum_{s=0}^{S}\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\|P_n^\sigma\left(g^{s+1}-g^{s}\right)\right\|_{\mathcal{Y}} \mid \mathcal{F}_n\right] \\
&\le \sup_{g\in\mathcal{G}}\left\|g-g^{S+1}\right\|_{2,P_n}+\sum_{s=0}^{S}\mathbb{E}\left[\max_{k\in\{1,\ldots,N_{s+1}\}}\left\|P_n^\sigma\left(g_k^{s+1}-g_k^{s+1,s}\right)\right\|_{\mathcal{Y}} \mid \mathcal{F}_n\right] \\
&\le 2^{-(S+1)}R_n+\sum_{s=0}^{S}\frac{1}{\lambda_s}\log\left(\mathbb{E}\left[\sum_{k=1}^{N_{s+1}}e^{\lambda_s\left\|P_n^\sigma\left(g_k^{s+1}-g_k^{s+1,s}\right)\right\|_{\mathcal{Y}}} \mid \mathcal{F}_n\right]\right) \qquad (a) \\
&\le 2^{-(S+1)}R_n+\sum_{s=0}^{S}\frac{1}{\lambda_s}\log\left(\sum_{k=1}^{N_{s+1}}\mathbb{E}\left[2\cosh\left(\lambda_s\left\|P_n^\sigma\left(g_k^{s+1}-g_k^{s+1,s}\right)\right\|_{\mathcal{Y}}\right) \mid \mathcal{F}_n\right]\right) \quad (b) \\
&\le 2^{-(S+1)}R_n+\sum_{s=0}^{S}\frac{1}{\lambda_s}\log\left(2\sum_{k=1}^{N_{s+1}}e^{\frac{\lambda_s^2}{n}(2^{-s}R_n)^2}\right) \qquad (c) \\
&= 2^{-(S+1)}R_n+\sum_{s=0}^{S}\frac{1}{\lambda_s}\log\left(2N_{s+1}e^{\frac{\lambda_s^2}{n}(2^{-s}R_n)^2}\right) \\
&= 2^{-(S+1)}R_n+\sum_{s=0}^{S}\frac{1}{\lambda_s}\left(H_{s+1}+\log 2\right)+\frac{\lambda_s}{n}\sum_{s=0}^{S}(2^{-s}R_n)^2 \\
&\le 2^{-(S+1)}R_n+\sum_{s=0}^{S}\frac{1}{\lambda_s}2H_{s+1}+\frac{\lambda_s}{n}\sum_{s=0}^{S}(2^{-s}R_n)^2 \qquad (d) \\
&= 2^{-(S+1)}R_n+\frac{2}{\sqrt{n}}\sum_{s=0}^{S}2^{-s}R_n\sqrt{2H_{s+1}} \qquad (e) \\
&= 2^{-(S+1)}R_n+\frac{2}{\sqrt{n}}J_n
\end{aligned}$$

where, in (a), we used Jensen's inequality and the fact that the sum of positive numbers is greater than their maximum; in (b), we used the basic fact $e^x \leq 2\cosh x$; in (c), we used Proposition 11; in (d), we used the fact that $H_{s+1} \geq \log 2$; and in (e), we let

$$\lambda_s = \frac{\sqrt{2nH_{s+1}}}{2^{-s}R_n}.$$

■

When the Rademacher complexity is used in empirical risk minimisation for real-valued function classes $\mathcal{F}$, what we end up using is not the Rademacher complexity $\mathfrak{R}_n(\mathcal{F})$ of the function class itself, but that of the composition of the loss with the function class. The same is true for vector-valued empirical risk minimisation problems. More precisely, suppose we have a loss function $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, and we denote by $\hat{g}_n$ the solution of the following empirical risk minimisation problem:

$$\hat{g}_n = \arg\min_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(Y_i, g(X_i)) = \arg\min_{g\in\mathcal{G}} \hat{\mathcal{R}}_n(g).$$

Denote by g^* the minimiser of the population risk:

$$g^* := \arg\min_{g\in\mathcal{G}} \mathbb{E}\left[\mathcal{L}(Y, g(X))\right] = \arg\min_{g\in\mathcal{G}} \mathcal{R}(g).$$

We want to know how fast $\mathcal{R}(\hat{g}_n)$ converges to the minimal risk $\mathcal{R}(g^*)$ as the sample size n increases. Here, actually, the standard result concerning Rademacher complexities applies directly – we will quote the following result.

Theorem 32 (Shalev-Shwartz and Ben-David (2014, p.328, Theorem 26.5)) *Assume that for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and $g \in \mathcal{G}$, we have $|\mathcal{L}(y, g(x))| \leq c$ for some constant $c > 0$. Then with probability at least $1 - \delta$, we have*

$$\mathcal{R}(\hat{g}_n) - \mathcal{R}(g^*) \leq 2\mathfrak{R}_n(\mathcal{L} \circ \mathcal{G}) + 5c\sqrt{\frac{2\log\left(\frac{8}{\delta}\right)}{n}}$$

where we used the notation $\mathcal{L} \circ \mathcal{G}$ for the class of functions $\mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ defined as

$$\mathcal{L} \circ \mathcal{G} := \left\{(x, y) \mapsto \mathcal{L}(y, g(x)) : g \in \mathcal{G}\right\}.$$

Now, the question is how to obtain a meaningful bound on the Rademacher complexity $\mathfrak{R}_n(\mathcal{L} \circ \mathcal{G})$ as $n \to \infty$. When $\mathcal{G}$ is a class of real-valued functions, the Contraction Lemma (Shalev-Shwartz and Ben-David, 2014, p.331, Lemma 26.9) tells us that if, for each $Y_i \in \mathbb{R}$, the map $y \mapsto \mathcal{L}(Y_i, y)$ is c-Lipschitz, then $\mathfrak{R}_n(\mathcal{L} \circ \mathcal{G})$ is bounded by $c\mathfrak{R}_n(\mathcal{G})$, so it is meaningful to work with $\mathfrak{R}_n(\mathcal{G})$. However, an analogue of this result when $\mathcal{G}$ is a class of $\mathcal{Y}$-valued functions is shown to be impossible via a counterexample, in Maurer (2016, Section 6).

As mentioned above, one of the main ways of bounding the Rademacher complexity is to use entropy. As our end goal is to bound the Rademacher complexity of $\mathcal{L} \circ \mathcal{G}$, there are two ways of going about this task with entropy. For real-valued function classes $\mathcal{F}$, what is commonly done is to bound the Rademacher complexity of $\mathcal{L} \circ \mathcal{F}$ with the Rademacher complexity of $\mathcal{F}$ using contraction, then to bound the Rademacher complexity of $\mathcal{F}$ by an expression involving the entropy, using chaining. As discussed before, contraction becomes difficult with vector-valued function classes. But we propose a different way that avoids contraction of Rademacher complexities. We can first bound the Rademacher complexity of $\mathcal{L} \circ \mathcal{G}$ with an expression involving the entropy of $\mathcal{L} \circ \mathcal{G}$, and use the following contraction result of entropies.

Lemma 33 *Suppose that for each $Y \in \mathcal{Y}$, the $\mathcal{Y} \to \mathbb{R}$ map $y \mapsto \mathcal{L}(Y, y)$ is c-Lipschitz for some constant $c > 0$, i.e. for $y_1, y_2 \in \mathcal{Y}$, $|\mathcal{L}(Y, y_1) - \mathcal{L}(Y, y_2)| \leq c\|y_1 - y_2\|_{\mathcal{Y}}$. Then for any $\delta > 0$, we have*

$$H(c\delta, \mathcal{L} \circ \mathcal{G}, \|\cdot\|_{2,P_n}) \leq H(\delta, \mathcal{G}, \|\cdot\|_{2,P_n}).$$

Proof To ease the notation, write $N = N(\delta, \mathcal{G}, \|\cdot\|_{2,P_n})$, and let $g_1, ..., g_N$ be a minimal δ-covering of $\mathcal{G}$. Then for any $\mathcal{L} \circ g \in \mathcal{L} \circ \mathcal{G}$, there exists some g_j, $j \in \{1, ..., N\}$ with $\|g - g_j\|_{2,P_n} = (\frac{1}{n}\sum_{i=1}^n \|g(X_i) - g_j(X_i)\|_{\mathcal{Y}}^2)^{1/2} \leq \delta$. Then by the Lipschitz condition on $\mathcal{L}$,

$$\begin{aligned} \left\|\mathcal{L} \circ g - \mathcal{L} \circ g_j\right\|_{2,P_n} &= \left(\frac{1}{n}\sum_{i=1}^n \left|\mathcal{L}(Y_i, g(X_i)) - \mathcal{L}(Y_i, g_j(X_i))\right|^2\right)^{\frac{1}{2}} \\ &\leq \left(\frac{1}{n}\sum_{i=1}^n c^2 \left\|g(X_i) - g_j(X_i)\right\|_{\mathcal{Y}}^2\right)^{\frac{1}{2}} \\ &= c\left\|g - g_j\right\|_{2,P_n} \\ &\leq c\delta. \end{aligned}$$

Hence $\mathcal{L} \circ g_1, ..., \mathcal{L} \circ g_N$ is a $c\delta$-covering of $\mathcal{L} \circ \mathcal{G}$, i.e.

$$N(c\delta, \mathcal{L} \circ \mathcal{G}, \|\cdot\|_{2,P_n}) \leq N(\delta, \mathcal{G}, \|\cdot\|_{2,P_n}).$$

Now finish the proof by taking logarithms of both sides. ■

So for empirical risk minimisation problems with appropriate loss functions, it does make sense to consider the entropy of vector-valued function classes $\mathcal{G}$, while it remains as future work to investigate the use of the Rademacher complexity of $\mathcal{G}$.