On the Consistency of Kernel Methods with Dependent Observations

Pierre-François Massiani [1] **Sebastian Trimpe** [1] **Friedrich Solowjow** [1]

Abstract

The consistency of a learning method is usually established under the assumption that the observations are a realization of an independent and identically distributed (i.i.d.) or mixing process. Yet, kernel methods such as support vector machines (SVMs), Gaussian processes, or conditional kernel mean embeddings (CKMEs) all give excellent performance under sampling schemes that are obviously non-i.i.d., such as when data comes from a dynamical system. We propose the new notion of *empirical weak convergence (EWC)* as a general assumption explaining such phenomena for kernel methods. It assumes the existence of a random asymptotic data distribution and is a strict weakening of previous assumptions in the field. Our main results then establish consistency of SVMs, kernel mean embeddings, and general Hilbert-space valued empirical expectations with EWC data. Our analysis holds for both finite- and infinite-dimensional outputs, as we extend classical results of statistical learning to the latter case. In particular, it is also applicable to CKMEs. Overall, our results open new classes of processes to statistical learning and can serve as a foundation for a theory of learning beyond i.i.d. and mixing.

1. Introduction

A learning method is *consistent* when the learned function is optimal in a certain sense in the infinite-sample limit, and the most common assumption to prove consistency is that training data comes from an independent and identically distributed (i.i.d.) process. That assumption is often blatantly violated, however, leading to drops in performance for many learning algorithms. For instance, replay buffers were introduced in deep reinforcement learning to mitigate this issue; see (Mnih et al., 2013). In contrast, kernel methods such as support vector machines (SVMs), Gaussian processes, or conditional kernel mean embeddings (CKMEs) seem unaffected by dependencies in the training data and often perform well despite the absence (or limited amount) of theoretical justification; see (Buisson-Fenet et al., 2020; von Rohr et al., 2021) for examples.

A common such case is that of learning on a Markov chain, where data is neither independent (previous states influence those that follow) nor identically distributed (transition probabilities may differ in different states). The standard assumption to get consistency under such sampling is *mixing* (Steinwart & Christmann, 2009). It replaces independence with decaying correlations between samples as their temporal distance increases. There is a vast literature showing consistency under mixing; examples are (Steinwart et al., 2009; Steinwart & Christmann, 2009; Irle, 1997).

A possible explanation for the popularity of the mixing assumption is that it allows dependencies while retaining a central concept of statistical learning: there exists an asymptotic distribution describing future samples. Learning is then formalized as minimizing the risk of those future samples. There are currently two main approaches to generalizing statistical learning to non-mixing data: assuming the existence of an asymptotic distribution to define the risk (Steinwart et al., 2009), or defining a non-asymptotic notion of risk (Simchowitz et al., 2018; Ziemann & Tu, 2022). We focus on the first approach, as it leverages more standard tools whereas the alternatives impose additional conditions on the data-generating process. Central to this approach is examining whether the empirical measure has a limit instead of imposing stronger conditions enforcing the existence of this limit and convergence thereto with certain speed — e.g., i.i.d. This pertains to whether a form of the *law of large numbers for events (LLNE)* holds. While Steinwart et al. (2009) provide first results on learning in this context, their analysis still excludes simple, non-pathological processes such as Example 3.8 below. This shows the necessity for a theory of consistency under assumptions weaker than i.i.d. or mixing.

We propose empirically weakly converging (EWC)[1] pro-

[1]Institute for Data Science in Mechanical Engineering, RWTH Aachen University, Aachen, Germany. Correspondence to: Pierre-François Massiani <massiani@dsme.rwth-aachen.de>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

[1]In what follows, we use the abbreviation EWC indistinctly for the noun "empirical weak convergence" and the adjective "empiri-

cesses for the basis of such a theory. They are those that possess an asymptotic data distribution, which may be random. We focus on the weak limit of the empirical measure for that distribution since it is one of the weakest (and thus more general) and simplest notion of convergence of random measures. Our main results then establish consistency of SVMs and of some Hilbert-space valued empirical expectations under EWC. Interestingly, the weak convergence we assume imposes a *continuously differentiable* loss function for consistency, which is more restrictive than the continuity required in (Steinwart et al., 2009) under convergence in total variation. In other words, the gain of generality allowed by EWC processes comes at the price of weaker guarantees.

Our results formalize the general intuitive idea that *consistency is w.r.t. the data distribution effectively generated by the process*. Albeit simple, it offers the flexibility of a *path-dependent* asymptotic distribution, which is completely new to the best of our knowledge. Previous standard assumptions such as i.i.d., mixing, and the LLNE are then sufficient conditions that specify this asymptotic distribution a priori. With this shift of perspective, EWC pushes the reasoning of statistical learning to minimize an asymptotic risk beyond those assumptions.

All of our results hold in a framework more general than that of standard statistical learning theory of Steinwart & Christmann (2008). Specifically, we allow for a general separable Hilbert output space, whereas the above reference only considers scalar outputs. As an immediate benefit, our analysis also applies to cases that were so far excluded such as CKMEs, which are an instance of SVMs with infinite-dimensional outputs in general.

The rest of the paper is organized as follows. We begin with preliminaries in Section 2. Section 3 defines EWC processes and their properties. Specifically, Theorems 3.4 and 3.5 provide necessary and sufficient conditions for EWC, and Theorem 3.6 shows that empirical expectations of Hilbert-space valued continuous and bounded functions converge under EWC. Further, Theorems 3.17 and 3.19 characterize EWC of multivariate processes, which is particularly relevant as it constitutes one of our main assumptions. Finally, Section 4 states our main results on consistency of kernel methods, Theorems 4.1 and 4.3. With a few exceptions, proofs are in Appendix A. Appendix B contains technical results on random compact sets necessary for the proof of Theorem 3.6, Appendix C is the generalization of the general representer theorem of (Steinwart & Christmann, 2008) to separable Hilbert output spaces, and Appendix D justifies the measurability of risks with random measures and of SVMs with such output spaces, upon which our results rely. We finish with a remark on terminology: as in (Steinwart & Christmann, 2008), we mean by "SVM" a generalized SVM, that is, regularized empirical risk minimization over a reproducing kernel Hilbert space (RKHS) with an arbitrary loss. This differs from the historic definition (Schölkopf & Smola, 1998), which only considers the Hinge loss.

cally weakly converging".

1.1. Related work

Consistency under dependent sampling Consistency is a fundamental property a learning method should have. It has been thoroughly studied for SVMs in the case of i.i.d. data with finite-dimensional outputs. One method to obtain consistency are the so-called *oracle inequalities*, of which a complete exposition is in Chapter 6 of (Steinwart & Christmann, 2008). The other main method for consistency is the *integral operator* technique, which is limited to the square loss as it leverages the specific form of the solution to the SVM (Caponnetto & De Vito, 2007). While it allows infinite-dimensional outputs, the required assumptions on the operators historically limited the allowed class of kernels in that case. Recent results show consistency for a broad class of kernels; see (Park & Muandet, 2022; Li et al., 2022). Next, many bodies of work examine consistency of SVMs under *mixing* assumptions rather than i.i.d. (Steinwart et al., 2009; Steinwart & Christmann, 2009; Irle, 1997). Finally, all of these works also provide learning rates quantifying the speed of the convergence. Such rates also require restricting the class of data-generating distributions; see (Steinwart et al., 2009; Von Luxburg & Schölkopf, 2011; Caponnetto & De Vito, 2007). In the present work, we focus on *whether convergence occurs* and neglect the question of its speed; in fact, we allow arbitrarily slow convergence and leave this question for future work. Therefore, we rely on a third, non-quantitative method leveraging only laws of large numbers (Steinwart & Christmann, 2009). Our results generalize those of Section 2 in that reference to EWC data and infinite-dimensional outputs. Consistently with the no-free-lunch theorem (Steinwart et al., 2009), we find that the sequence of should be annealed at a rate that depends on the data-generating process in general. This is in accordance with results on the inconsistency of kernel ridgeless regression (Rakhlin & Zhai, 2019; Buchholz, 2022; Beaglehole et al., 2023), where the regularization is fixed to 0.

There are no further relaxations of independence that are specific to SVMs, to the best of our knowledge. Instead, *learnability* focuses on whether learning is possible at all without considering a specific learning method. Results on learnability are typically on the *existence* of a learning method, and are thus of independent interest. An overview of recent results in that field is available in (Hanneke, 2021).

Learning theory with infinite-dimensional outputs The classical framework of statistical learning theory from Steinwart & Christmann (2008) assumes a scalar output space. It immediately generalizes to finite-dimensional outputs by

reasoning component-wise, but the extension to infinite-dimensional outputs is less clear. In fact, there is no systematic exposition of the generalization of the definitions or the results of the above reference to that case, with the notable exception of the square loss as discussed above.

A contribution of this work is thus the generalization of the setup of statistical learning theory to separable Hilbert output spaces. As an immediate consequence, we obtain consistency of SVMs in that case under appropriate assumptions such as i.i.d. or EWC. In particular, our analysis applies to CKMEs, which are, when regularized, the solutions of infinite-dimensional SVMs where the output space is itself an RKHS and output data consists of kernel partial evaluations; cf. (Grünewälder et al., 2012). This is of particular interest since, historically, unregularized CKMEs have suffered from debates on meaningful assumptions for a well-posed definition; see (Klebanov et al., 2020; Mollenhauer & Koltai, 2020). A way to understand this debate is to notice that a CKME is the solution of an *unregularized* SVM, which becomes problematic when the solution does not exist in the RKHS where regression is performed. We avoid this concern by only considering the regularized problem; consistency is meaningful even when the sequence of regressors does not converge in the RKHS. There is a rich literature on the asymptotic properties of CKME estimation. For instance, Park & Muandet (2022) and Li et al. (2022) show consistency and (optimal) learning rates. While our theory is not quantitative, it applies in a more general case, as it does not require i.i.d. data, and we are not limited to the square loss.

2. Preliminaries and notations

We introduce in this section necessary definitions to state our results. We focus on the different spaces we consider, Markov kernels, statistical learning theory, and RKHSs.

2.1. Sets and topology

A Polish space is a topological space $(\mathcal{X}, \mathcal{T})$ that is separable and completely metrizable. In what follows, we abuse notation and simply say that $\mathcal{X}$ is a Polish space, omitting specifying the topology since it is always clear from context. Specifically, we equip product spaces with the natural product topology, and any normed vector space with the topology induced by its norm. In particular, $\mathbb{R}$ is equipped with its usual topology (for which it is Polish), a complete subset of a separable Hilbert space is Polish, and a product of Polish spaces is Polish. Then, a Polish space is *locally compact* if every point has a compact neighborhood; this is for instance the case of finite-dimensional vector spaces. Finally, if $\mathcal{X}$ and $\mathcal{Y}$ are Polish, then $C(\mathcal{X};\mathcal{Y})$ is the set of continuous functions from $\mathcal{X}$ to $\mathcal{Y}$. If, additionally, $\mathcal{Y}$ is metric, $C_{\mathrm{b}}(\mathcal{X};\mathcal{Y})$ is the set of continuous bounded functions from $\mathcal{X}$ to $\mathcal{Y}$. If $\mathcal{X}$ is compact, then $C_{\mathrm{b}}(\mathcal{X};\mathcal{Y})$ equipped with the topology of uniform convergence is separable.

We equip any Polish space $\mathcal{X}$ with its σ-algebra of Borel sets, that is, the σ-algebra generated by its topology. We denote it by $\mathcal{B}(\mathcal{X})$. In this context, we define $\mathcal{P}(\mathcal{X})$ as the set of probability measures on the measurable space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. For any probability measure $P \in \mathcal{P}(\mathcal{X})$ and $f \in C_{\mathrm{b}}(\mathcal{X};\mathbb{R})$, we define the standard notation

$$Pf := \int_{\mathcal{X}} f(x)\mathrm{d}P(x)$$

We also use this notation if P is not a probability measure but only a finite sum or difference thereof, making it a signed measure. We endow $\mathcal{P}(\mathcal{X})$ with the topology of weak convergence: recall that a sequence of probability measures $(P_n)_{n\in\mathbb{N}^\star} \subset \mathcal{P}(\mathcal{X})$ converges weakly to a probability measure $Q \in \mathcal{P}(\mathcal{X})$ if, for every $f \in C_{\mathrm{b}}(\mathcal{X};\mathbb{R})$,

$$P_n f \xrightarrow[n\to\infty]{} Q f.$$

We denote this as $P_n \rightsquigarrow Q$. Further, it is known that if $\mathcal{X}$ is Polish, then $\mathcal{P}(\mathcal{X})$ is metrizable; see for instance (Sriperumbudur et al., 2010).

Finally, if $\mathcal{X}$ is a topological space, $P \in \mathcal{P}(\mathcal{X})$, and $\mathcal{G}$ is a Banach space, we define $\mathcal{L}_0(\mathcal{X};\mathcal{G})$ the set of (Bochner) measurable functions from $\mathcal{X}$ to $\mathcal{G}$. Then, $\mathcal{L}_\infty(\mathcal{X};\mathcal{G})$ and $\mathcal{L}_2(\mathcal{X},P;\mathcal{G})$ are its subsets consisting of bounded functions and of functions with finite 2 norm w.r.t. P, $L_2(\mathcal{X},P,\mathcal{G})$ is the Bochner space of equivalence classes of functions in $\mathcal{L}_2(\mathcal{X},P;\mathcal{G})$, and $L_\infty(\mathcal{X},P;\mathcal{G})$ is the set of P-essentially bounded functions from $\mathcal{X}$ to $\mathcal{G}$.

2.2. Markov kernels and random elements

All the random variables that we consider in this work are defined on a complete, standard Borel probability space $(\Omega, \mathcal{A}, \mathbb{P})$. We consider discrete stochastic processes, which we index by the set of positive integers $\mathbb{N}^\star$ without loss of generality unless mentioned otherwise. If X is a process taking values in $\mathcal{X}$, we consistently use the notation X_n for its value at time $n \in \mathbb{N}^\star$ and do not introduce this notation again later. We also use the notation $X_{1:n}$ to denote the random $\mathcal{X}^n$-valued vector $(X_1, \dots, X_n)$, $n \in \mathbb{N}^\star$. Further, if Z is an $\mathcal{X}\times\mathcal{Y}$-valued process, the notation $Z = (X, Y)$ means that X and Y are $\mathcal{X}$- and $\mathcal{Y}$-valued processes, respectively, such that $Z_n = (X_n, Y_n)$ for all $n \in \mathbb{N}^\star$.

For an $\mathcal{X}$-valued random variable X and a sub-σ-algebra $F \subset \mathcal{B}(\mathcal{X})$, we denote by $\mathbb{E}[X \mid F]$ and $\mathbb{P}[X \in \cdot \mid F]$ versions of the conditional expectation and law of X given F. The specific choices of conditional expectation and law do not play a role in this article. If F is the σ-algebra generated by one or many random variables, we conveniently replace F with the appropriate variables in the above notations. Finally, if Y is a $\mathcal{Y}$-valued process, we allow the notation $Y_{1:0}$

when conditioning to denote conditioning w.r.t. the trivial σ-algebra $\{\emptyset, \Omega\}$.

Next, we introduce Markov kernels (Kallenberg, 2017), which we consider in two contexts: stochastic input-output maps, and random measures.

Definition 2.1 (Markov kernel)**.** Let $(\mathcal{X}, \mathcal{A}_\mathcal{X})$ and $(\mathcal{Y}, \mathcal{A}_\mathcal{Y})$ be measurable spaces. A *Markov kernel* from $\mathcal{X}$ to $\mathcal{Y}$ is a map $p : \mathcal{A}_\mathcal{Y} \times \mathcal{X} \to [0, \infty)$ such that

1. for all $x \in \mathcal{X}$, $p(\cdot, x)$ is a probability measure on $\mathcal{Y}$;
2. for all $A \in \mathcal{A}_\mathcal{Y}$, $p(A, \cdot)$ is $\mathcal{A}_\mathcal{X}$-measurable.

The following notion generalizes continuity to such kernels. It is classically defined for continuous processes, but readily applies to discrete ones as well (Li & Li, 2009).

Definition 2.2. Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces. We say that a Markov kernel p from $\mathcal{X}$ to $\mathcal{Y}$ is Feller-continuous if the map $x \in \mathcal{X} \mapsto p(\cdot, x)f$ is in $C_\mathrm{b}(\mathcal{X}; \mathbb{R})$ for any $f \in C_\mathrm{b}(\mathcal{Y}; \mathbb{R})$.

Definition 2.3. Let $\mathcal{X}$ be Polish space. We call a measurable map $P : \Omega \to \mathcal{P}(\mathcal{X})$ a *random measure*.

An alternative definition is through Markov kernels: a map $P : \Omega \to \mathcal{P}(\mathcal{X})$ is a random measure if, and only if, the map $p : \mathcal{B}(\mathcal{X}) \times \Omega \to \mathbb{R}$ defined by $p(B, \omega) = P(\omega)(B)$ for all $B \in \mathcal{B}(\mathcal{X})$ and $\omega \in \Omega$ is a Markov kernel; see Lemma 1.14 in (Kallenberg, 2017). If P is a random measure on $\mathcal{X}$, we abuse notation and often omit its dependency on $\omega \in \Omega$. The *empirical (probability) measure* of an $\mathcal{X}$-valued process X at time $n \in \mathbb{N}^\star$ is

$$\eta_n^X = n^{-1} \sum_{i=1}^{n} \delta_{X_i},$$

where $\delta_x \in \mathcal{P}(\mathcal{X})$ is the Dirac measure with mass at $x \in \mathcal{X}$. The empirical measure is a random probability measure. We also use the notation η_n^X when X is a deterministic sequence; then, η_n^X is a probability measure.

2.3. Elements of statistical learning theory

We generalize here the definitions of (Steinwart & Christmann, 2008) to the case of an infinite-dimensional output space, which is necessary for our main result (Theorem 4.3) to also encompass CKMEs. The definitions here are straightforward extensions of those of the above reference. More interesting is the resulting general representer theorem, which generalizes Theorem 5.9 in the reference to separable infinite-dimensional output spaces. Although it is new to the best of our knowledge, its proof is a direct generalization of its counterpart in (Steinwart & Christmann, 2008). For this reason, we defer the statement and proof of this general representer theorem to Appendix C.

In what follows, $\mathcal{X}$ is a Polish space, $\mathcal{G}$ is a Hilbert space, and $\mathcal{Y}$ is a complete subset of $\mathcal{G}$.

Definition 2.4. A function $L : \mathcal{X} \times \mathcal{Y} \times \mathcal{G} \to [0, \infty)$ is called a *loss function* (on $\mathcal{X} \times \mathcal{Y} \times \mathcal{G}$) if it is measurable. Then, it is convex (resp. continuous) if, for all $x \in \mathcal{X}, y \in \mathcal{Y}$, the function $L(x, y, \cdot)$ is convex (resp. continuous). Further, L is *locally bounded* if, for all bounded $A \subset \mathcal{G}$, the restriction $L_{|\mathcal{X} \times \mathcal{Y} \times A}$ is bounded, and *locally Lipschitz continuous* if, for all $a > 0$, the following quantity is finite:

$$|L|_{a,1} := \sup_{\substack{(t,t') \in \mathcal{G}^2 \\ 0 < \|t - t'\|_\mathcal{G} \leq 2a}} \sup_{\substack{x \in \mathcal{X} \\ y \in \mathcal{Y}}} \frac{|L(x, y, t) - L(x, y, t')|}{\|t - t'\|_\mathcal{G}}.$$

It is *Lipschitz continuous* if $|L|_1 = \sup_{a>0} |L|_{a,1} < \infty$. Finally, the loss is (Fréchet) differentiable if the map $L(x, y, \cdot)$ is Fréchet differentiable on $\mathcal{G}$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, that is, for all $t \in \mathcal{G}$ there exists a vector $A_{x,y,t} \in \mathcal{G}$ such that

$$\lim_{\substack{h \to 0 \\ h \neq 0}} \frac{|L(x, y, t + h) - L(x, y, t) - \langle A_{x,y,t}, h \rangle|}{\|h\|_\mathcal{G}} = 0.$$

The vector $A_{x,y,t}$ is then written $\nabla L(x, y, t)$. We say that L is *continuously (Fréchet) differentiable* if the map $(x, y, t) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{G} \mapsto \nabla L(x, y, t)$ is continuous.

Definition 2.5. A loss function $L : \mathcal{X} \times \mathcal{Y} \times \mathcal{G} \to \mathbb{R}$ is called a Nemitski loss function if there exists a measurable function $b : \mathcal{X} \times \mathcal{Y} \to [0, \infty)$ and an increasing function $h : [0, \infty) \to [0, \infty)$ such that, for all $(x, y, t) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{G}$,

$$L(x, y, t) \leq b(x, y) + h(\|t\|_\mathcal{G}).$$

Further, it is a Nemitski loss of order $p \in (0, \infty)$ if there exists a constant $c > 0$ such that the above holds with $h(s) = c \cdot s^p$, for $s > 0$. Finally, we say that L is a J-integrable Nemitski loss if $b \in \mathcal{L}_1(\mathcal{X} \times \mathcal{Y}, J; \mathbb{R})$ for $J \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$.

Definition 2.6. Let L be a loss function and $J \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$. For any measurable function $f : \mathcal{X} \to \mathcal{G}$, we define its *$L$-risk* as

$$\mathcal{R}_{L,J}(f) = \int_{\mathcal{X} \times \mathcal{Y}} L(x, y, f(x)) \mathrm{d}J(x, y).$$

The *Bayes L-risk* is then $\mathcal{R}^\star_{L,J} = \inf\{\mathcal{R}_{L,J}(f) \mid f \in \mathcal{L}_0(\mathcal{X}; \mathcal{G})\}$.

Definition 2.7. Let L be a loss function and $\mathcal{F}$ a set of measurable functions from $\mathcal{X}$ to $\mathcal{G}$. For $J \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, we say that $\mathcal{F}$ is (L, J)-rich if

$$\mathcal{R}_{\mathcal{F},L,J} := \inf_{f \in \mathcal{F}} \mathcal{R}_{L,J}(f) = \mathcal{R}^\star_{L,J}.$$

A *learning method* $\mathfrak{L}$ is then a mapping that maps any training set $Z = ((x_i, y_i))_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$, $n \in \mathbb{N}^\star$, to a unique measurable function $f_Z : \mathcal{X} \to \mathcal{G}$. We postpone introducing the notion of consistency that we consider to Section 4, since it requires first defining EWC processes.

2.4. Vector-valued RKHSs and SVMs

In this section, $\mathcal{X}$ is a measurable space and $\mathcal{G}$ a separable Hilbert space.

Definition 2.8. A $\mathcal{G}$-valued RKHS $\mathcal{H}$ on $\mathcal{X}$ is a Hilbert space $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ of functions such that for all $x \in \mathcal{X}$, the evaluation operator $S_x : f \in \mathcal{H} \mapsto f(x) \in \mathcal{G}$ is continuous. Then, we define $K(\cdot, x) = S_x^\star$ and $K(x, x') = S_x S_{x'}^\star$, for all $x, x' \in \mathcal{X}$. The map $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{G})$ is called the (operator-valued) reproducing kernel of $\mathcal{H}$, where $\mathcal{L}(\mathcal{G})$ is the Banach space of continuous linear operators on $\mathcal{G}$.

Theorem 2.9. *Let $\mathcal{H}$ be a $\mathcal{G}$-valued RKHS with kernel K. Then, K is Hermitian, positive semi-definite*[2]*, and the reproducing property holds for all $x \in \mathcal{X}$, $f \in \mathcal{H}$, and $g \in \mathcal{G}$:*

$$\langle f(x), g \rangle_{\mathcal{G}} = \langle f, K(\cdot, x) g \rangle_{\mathcal{H}} .$$

It is well known that, for every positive semi-definite function $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{G})$, there exists a unique $\mathcal{G}$-valued RKHS of which K is the unique reproducing kernel; see (Carmeli et al., 2006). A special case of vector-valued RKHSs is that of scalar-valued RKHSs, that is, when $\mathcal{G} = \mathbb{R}$. Then, we introduce a symmetric, positive definite function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that $K(x, x') = k(x, x')\mathrm{Id}_{\mathbb{R}}$ (which always exists), and refer to k as the reproducing kernel of $\mathcal{H}$.

Definition 2.10. We say that a kernel K is *bounded* if

$$\|K\|_\infty := \sup_{x \in \mathcal{X}} \sqrt{\|K(x, x)\|_{\mathcal{L}(\mathcal{G})}} < \infty.$$

Given a convex loss L, a $\mathcal{G}$-valued RKHS $\mathcal{H}$, and a training set $Z = ((x_i, y_i))_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$, $n \in \mathbb{N}^\star$, there exists a unique minimizer to the following minimization problem, where $\lambda > 0$ is a regularization parameter (cf. Lemma C.5):

$$f_{Z,\lambda} = \arg\min_{f \in \mathcal{H}} \mathcal{R}_{L, \eta_n^Z}(f) + \lambda \|f\|_{\mathcal{H}}^2 . \tag{1}$$

We call the learning method that maps, given a sequence of regularization parameters $(\lambda_n)_{n \in \mathbb{N}^\star}$, a sequence of training sets $(Z_n)_{n \in \mathbb{N}^\star}$ to the sequence (f_{Z_n, λ_n}) the (λ_n)-SVM.

3. Empirical weak convergence

In this section, we introduce the notion of EWC processes. We show some first properties and discuss the connections with other standard notions. Finally, we examine EWC of joint processes.

[2]Recall that a bivariate function $\phi : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{G})$ is Hermitian if $\phi(x, x') = \phi(x', x)^\star$. Further, it is positive semi-definite if for all $n \in \mathbb{N}^\star$, $(x_i)_{i=1}^n \in \mathcal{X}^n$, and $(g_i)_{i=1}^n \in \mathcal{G}^n$, $\sum_{i=1}^n \sum_{j=1}^n \langle g_i, \phi(x_i, x_j) g_j \rangle_{\mathcal{G}} \geq 0$.

3.1. Definition and first properties

We are interested in processes that possess an asymptotic data distribution. A reasonable definition is to take for such an asymptotic distribution the limit of $(\eta_n^X)_{n \in \mathbb{N}^\star}$ considered as an element of a topological space; here, $\mathcal{P}(\mathcal{X})$ endowed with a suitable topology. We choose that of weak convergence, as it is allows for a broad class of processes while still providing strong guarantees, as we will see it. Further, since $(\eta_n^X)_{n \in \mathbb{N}^\star}$ is a *random* measure, the convergence should be in a probabilistic sense.

Definition 3.1. Let $\mathcal{X}$ be a Polish space and d be any metric that metrizes the weak convergence on $\mathcal{P}(\mathcal{X})$. We say that an $\mathcal{X}$-valued process X is *empirically weakly converging (EWC) in probability* if there exists a random measure P on $\mathcal{X}$ such that and $\eta_n^X \rightsquigarrow P$ in probability, that is,

$$\lim_{n \to \infty} d(\eta_n^X, P) = 0, \tag{2}$$

where convergence is in probability. Further, we say that X is *EWC almost surely (a.s.)* if the weak convergence $\eta_n^X \rightsquigarrow P$ occurs a.s., that is, if the convergence in (2) holds a.s. In either case, the random measure P is called a limit measure of the process.

Proposition 3.2. *Let $\mathcal{X}$ be a Polish space. The notions of EWC in probability and EWC a.s. are independent of the metric d that metrizes weak convergence. Furthermore, the random limit measure P is unique $\mathbb{P}$-a.s..*

It is clear that a.s. EWC implies its counterpart in probability with the same limit measure, and that the converse is not true in general.

Remark 3.3. The above definition raises the question of the measurability of the map $\omega \mapsto d(\eta_n^X, P)(\omega)$. It is indeed measurable, since (η_n^X, P) is measurable (cf. Theorem 14.8 in (Klenke, 2013)) and d is trivially continuous.

A meaningful question is then on the asymptotic behaviour of the empirical averages $\eta_n^X f$, where f is a test function. Usual results on weak convergence hint that they should converge to Pf as long as $f \in C_{\mathrm{b}}(\mathcal{X}; \mathbb{R})$. This is what the following result guarantees, with the additional technicality that the convergence is here again probabilistic. In fact, this is an equivalent characterization of EWC under an additional compactness assumption.

Theorem 3.4. *Let $\mathcal{X}$ be a Polish space, X an $\mathcal{X}$-valued process, and P a random measure on $\mathcal{X}$. Consider the following statements:*

(i) X is EWC a.s. with limit measure P;

(ii) $\mathbb{P}\left[\forall f \in C_{\mathrm{b}}(\mathcal{X}; \mathbb{R}), \limsup_n |\eta_n^X f - Pf| = 0\right] = 1$;

(iii) for all $f \in C_{\mathrm{b}}(\mathcal{X}; \mathbb{R})$, $\lim_{n \to \infty} \eta_n^X f = Pf$, where the convergence is a.s.;

Then, (i) $\iff$ (ii) $\implies$ (iii). If, additionally, $\mathcal{X}$ is compact, the implication (iii) $\implies$ (ii) also holds and all of the statements are equivalent.

Theorem 3.5. *Let $\mathcal{X}$ be a Polish space, X an $\mathcal{X}$-valued process, and P a random measure. Consider the following statements:*

(i) X is EWC in probability with limit measure P;

(ii) for any strictly increasing sequence $(k_n)_{n\in\mathbb{N}^\star} \subset \mathbb{N}^\star$, there exists a subsequence $(k_{m_n})_{n\in\mathbb{N}^\star}$ such that

$$\mathbb{P}[\forall f \in C_{\mathrm{b}}(\mathcal{X};\mathbb{R}), \limsup_n |\eta^X_{k_{m_n}} f - Pf| = 0] = 1.$$

(iii) for all $f \in C_{\mathrm{b}}(\mathcal{X};\mathbb{R})$, $\lim_{n\to\infty} \eta^X_n f = Pf$, where the convergence is in probability.

Then, (i) $\iff$ (ii) $\implies$ (iii). If, additionally, $\mathcal{X}$ is compact, the implication (iii) $\implies$ (ii) also holds and all of the statements are equivalent.

We point that, in the second statement of the above two theorems, the set of which the probability is evaluated is indeed measurable. The next results are very useful properties of EWC processes: empirical averages of Hilbert-space-valued continuous and bounded maps also converge. They are key in the proof of Theorem 4.3, but their generality makes them of independent interest.

Theorem 3.6. *Let $\mathcal{X}$ be a locally compact Polish space and $\mathcal{H}$ be a separable Hilbert space. Let X be an $\mathcal{X}$-valued process, and assume that X is EWC in probability (resp. a.s.) with limit measure P. Then, for all $\phi \in C_{\mathrm{b}}(\mathcal{X};\mathcal{H})$, we have*

$$\lim_{n\to\infty} \eta^X_n \phi = P\phi,$$

where the convergence is in probability (resp. a.s.).

Corollary 3.7. *Let $\mathcal{X}$ be a locally compact Polish space, $\mathcal{H}$ a separable Hilbert space, and $\mathcal{F}$ a separable subset of $C_{\mathrm{b}}(\mathcal{X};\mathcal{H})$. Let X be an $\mathcal{X}$-valued process, and assume that X is EWC in probability (resp. a.s.) with limit measure P. Then, for any strictly increasing sequence $(k_n)_{n\in\mathbb{N}^\star} \subset \mathbb{N}^\star$, there exists a subsequence $(k_{m_n})_{n\in\mathbb{N}^\star}$ such that*

$$\mathbb{P}[\forall f \in \mathcal{F}, \limsup_n \left\|\eta_{k_{m_n}} f - Pf\right\|_{\mathcal{H}} = 0] = 1.$$

If, additionally, X is EWC a.s., then

$$\mathbb{P}[\forall f \in \mathcal{F}, \limsup_n \|\eta_n f - Pf\|_{\mathcal{H}} = 0] = 1.$$

3.2. Examples and connections

We begin with examples of EWC processes to build intuition and establish its relevance as a generalization of existing notions before exhibiting some general connections. If unspecified, $\mathcal{X}$ is a Polish space in this section.

The following example is a compelling and simple instance where EWC strictly generalizes mixing or the LLNE.

Example 3.8. *Let Y be a Rademacher variable and $X_n = (1-\frac{1}{n})Y$ for all $n \in \mathbb{N}^\star$. The process X is EWC a.s. with limit measure δ_Y.*

This example abstracts a dynamical system that takes an irreversible step. Notice indeed that an equivalent definition is $X_1 = Y$ and $X_{n+1} = \frac{n^2}{(n-1)(n+1)} X_n$; the irreversible step is taken at time $n = 1$. It can for instance model the position of a ball on a hill at the origin and subject to gravity: the ball falls on a random side of the hill and goes to the corresponding asymptotic position ± 1. The following example generalizes the idea so that every step is irreversible.

Example 3.9. *Let $Y = (Y_n)_{n\in\mathbb{N}^\star}$ be i.i.d. Rademacher random variables. Let $X_1 = \frac{1}{2} + \frac{1}{3}Y_1$, and $X_{n+1} = X_n + \frac{1}{3^{n+1}} Y_{n+1}$. The process X converges a.s. to a variable Z taking values in the Cantor set. In particular, X is EWC with limit measure δ_Z.*

Neither of these simple examples is captured by existing assumptions, and yet they are sufficiently well-behaved — for instance, every trajectory converges — so that one can expect meaningful learning. Consider, for instance, the case where data is of the form $(X_n, f(X_n) + \epsilon_n)$, with X_n defined in Example 3.9, f a function one wants to learn, and ϵ_n i.i.d. noise. Results that guarantee successful learning with such data (under technical assumptions) are in Section 4.2. Before that, we connect EWC to other notions, confirming the generality of the assumption.

Independent, ergodic, and mixing processes Consistently with our initial motivation, i.i.d. and mixing processes are EWC when $\mathcal{X}$ is compact. More generally, any ergodic process is. Indeed, recall that an $\mathcal{X}$-valued process X is ergodic if it satisfies the condition of Birkhoff's pointwise ergodic theorem, that is, if there exists a measure $P \in \mathcal{P}(\mathcal{X})$ such that for every $f \in \mathcal{L}_1(\mathcal{X}, P; \mathbb{R})$, $\lim_{n\to\infty} \eta^X_n f \to Pf$, where convergence is a.s. Importantly, ergodicity is the weakest of all of the above notions: α–mixing is the weakest notion of mixing among β– and ϕ–mixing, and is stronger than ergodicity (see (Bradley, 2005) for details). But then, it is immediately clear from (iii) in Theorem 3.4 that such a process is EWC a.s. with constant limit measure P. The question is more delicate if $\mathcal{X}$ is not compact, as the null set of non-convergence of $\eta^X_n f$ to Pf in the definition of an ergodic process depends on the function f in general. Albeit interesting, this question is out of scope.

Measure-preserving dynamical systems Birkhoff's pointwise ergodic theorem applied to a general non-ergodic,

measure-preserving dynamical systems yields convergence of $\eta_n^X f$ to a conditional expectation, i.e., an expectation w.r.t. a random measure (Adams & Quas, 2020). Hence, such trajectories are EWC a.s. if $\mathcal{X}$ is compact.

Weak Convergence The notions of EWC and weakly converging processes differ, despite similar naming. The latter condition involves weak convergence of the sequence of *marginals* $M_n \in \mathcal{P}(\mathcal{X})$ of the process X to a non-random probability measure $M \in \mathcal{P}(\mathcal{X})$. Intuitively, it thus involves ensemble averages (and thus disregards correlations), whereas EWC involves time averages (and thus considers correlations). Both concepts are independent; there exist EWC processes that do not converge weakly, and vice-versa:

Lemma 3.10. *The process X defined on $\mathcal{X} := \{-1, 1\}$ by $X_n = (-1)^n$ a.s. is not weakly converging, but it is EWC a.s. with limit measure $\frac{1}{2}(\delta_{-1} + \delta_1)$.*

Proof. For $f \in C_\mathrm{b}(\mathcal{X}; \mathbb{R})$ and $n \in \mathbb{N}^\star$, $M_n f = f((-1)^n)$, which does not not converge unless f is constant. In contrast, $\eta_n^X f$ is the Cesàro average of that sequence, which converges to $\frac{1}{2}(f(-1) + f(1)) = \left[\frac{1}{2}(\delta_{-1} + \delta_1)\right] f$. □

Lemma 3.11. *The process X defined on $\mathcal{X} = \{0, 1\}$ by $X_1 \sim \mathcal{B}(0.5)$ and*

$$X_n = \begin{cases} X_1, & \text{if } \lfloor \log_{10}(n) \rfloor \equiv 0 \pmod 2, \\ 1 - X_1, & \text{otherwise}, \end{cases}$$

is weakly converging with limit measure $\frac{1}{2}(\delta_0 + \delta_1)$, but it is not EWC in probability.

Asymptotic mean stationarity As announced, processes that satisfy the LLNE are EWC with limit measure the stationary mean.

Lemma 3.12. *Let $\mathcal{X}$ be a Polish space and X be an $\mathcal{X}$-valued process that satisfies the weak law of large numbers for events (WLLNE) with stationary mean $P \in \mathcal{P}(\mathcal{X})$. Then, X is EWC in probability with limit measure constant equal to P. If, additionally, X satisfies the strong law of large numbers for events (SLLNE), then X is EWC a.s.*

This is no longer true if X is only asymptotically mean stationary (AMS)[3], however. For instance, the process of Lemma 3.11 is AMS but not EWC in probability as the empirical averages only converge *on expectation*. In fact, there is no relationship between AMS and EWC, as the requirements for AMS are stronger on some aspects (AMS requires convergence in total variation instead of weak convergence) but weaker on others (EWC requires the whole random limit measure to exist). This leads us to introduce the following two notions, which strengthen EWC (resp. weaken AMS) to relate to AMS (resp. to EWC), as summarized on Figure 1.

[3] See Definition 2.2 in (Steinwart et al., 2009) for the definition.

$$\begin{tikzcd}
 & \mathrm{ESC} \arrow[r, Rightarrow] \arrow[d, Rightarrow] & \mathrm{AMS} \arrow[d, Rightarrow] \\
\mathrm{LLNE} \arrow[ur, Rightarrow] \arrow[r, Rightarrow] & \mathrm{EWC} \arrow[r, Rightarrow] & \text{weakly AMS}
\end{tikzcd}$$

Figure 1. Relation between the the different notions. All implications hold with the same limit measure or with the intensity measure thereof, and both in probability and a.s. when applicable. The implication LLNE $\implies$ AMS is shown in Steinwart et al. (2009).

Definition 3.13. Let $\mathcal{X}$ be a Polish space and d_TV a metric on $\mathcal{P}(\mathcal{X})$ that metrizes convergence in total variation. An $\mathcal{X}$-valued process X is *empirically strongly converging (ESC)* in probability (resp. a.s.) if there exists a random measure P on $\mathcal{X}$ such that $d_\mathrm{TV}(\eta_n^X, P)$ is measurable, $n \in \mathbb{N}^\star$, and (2) holds with d_TV instead of d, where the convergence is in probability (resp. a.s.). Then, P is unique $\mathbb{P}$-a.s. and is called the limit measure of X.

Definition 3.14. Let $\mathcal{X}$ be a Polish space. We say that an $\mathcal{X}$-valued process X is *weakly AMS* if there exists a probability measure $P \in \mathcal{P}(\mathcal{X})$ such that

$$\forall f \in C_\mathrm{b}(\mathcal{X}; \mathbb{R}), \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[f(X_i)] = Pf.$$

It is called the weak asymptotic mean of X and is unique.

Remark 3.15. Assuming that $d_\mathrm{TV}(\eta_n^X, P)$ is measurable in Definition 3.13 certainly lacks elegance, but it is also solves a nontrivial technical difficulty. Indeed, the topology of convergence in total variation on $\mathcal{P}(\mathcal{X})$ is strictly finer than that of weak convergence in general, and the definition of random measures only gives measurability w.r.t. the Borel σ-algebra generated by the latter. Further exploration of this assumption is left for future work; we refer the interested reader to Chapter 4 in (Kallenberg, 2017).

Proposition 3.16. *Let $\mathcal{X}$ be a Polish space, X an $\mathcal{X}$-valued process, and P a random measure on $\mathcal{X}$. If X is ESC in probability (resp. a.s.) with limit measure P, then it is both EWC in probability (resp. a.s.) with the same limit measure and AMS. If X is EWC in probability with limit measure P, then it is weakly AMS. In either case, the corresponding asymptotic mean $\bar{P}$ satisfies $\mathbb{E}P = \bar{P}$, where $\mathbb{E}P$ is the usual intensity measure of a random measure.*

In other words, ESC and EWC require the whole limit measures to exist whereas AMS and weak AMS only require convergence on expectation. It is still unclear whether the implication LLNE $\implies$ ESC holds, as it would require a result similar to Theorems 3.4 and 3.5 for ESC processes, for which the implication (iii) $\implies$ (ii) is more challenging because of the non-separability of $\mathcal{L}_\infty(\mathcal{X}; \mathbb{R})$.

3.3. Joint processes

We now focus on characterizing EWC for joint processes. Indeed, it is a core assumption for learning, as the main assumption of Theorem 4.3 is EWC of the input-output process. We begin with continuous transformations of a process, as they include among others marginalization.

Theorem 3.17. *Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces, X be an $\mathcal{X}$-valued process, and $g : \mathcal{X} \to \mathcal{Y}$ be continuous. Assume that X is EWC in probability (resp. a.s.) with limit measure P. Then, the $\mathcal{Y}$-valued process $g(X)$ is EWC in probability (resp. a.s.) with limit measure $Q = P \circ g^{-1}$.*

Proof. This follows immediately by showing that (ii) in Theorem 3.4 (resp. Theorem 3.5) holds for Y and Q. □

Next, we formalize the mapping of a process through a Markov kernel.

Definition 3.18. Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces and p be a Markov kernel from $\mathcal{X}$ to $\mathcal{Y}$. A *data set of transition pairs of p* is an $\mathcal{X} \times \mathcal{Y}$-valued process $Z = (X, Y)$ such that the following holds $\mathbb{P}$-a.s.:

$$\forall n \in \mathbb{N}^\star,\ \mathbb{P}\left[Y_n \in \cdot \mid X_{1:n}, Y_{1:n-1}\right] = p(\cdot, X_n). \quad (3)$$

An example of a data set of transition pairs is the process $((X_n, X_{n+1}))_{n\in\mathbb{N}^\star}$, where X is a Markov chain.

Theorem 3.19. *Let $\mathcal{X}$ and $\mathcal{Y}$ be compact Polish spaces and $Z = (X, Y)$ be a data set of transition pairs of a Feller-continuous Markov kernel p from $\mathcal{X}$ to $\mathcal{Y}$. Then Z is EWC in probability (resp a.s.) if, and only if, X is. In this case, with P the limit measure of X and J that of Z, we have $\mathbb{P}$-a.s. for all $A \in \mathcal{B}(\mathcal{X})$ and $B \in \mathcal{B}(\mathcal{Y})$*

$$J(A \times B) = \int_A p(B, x)\mathrm{d}P(x). \quad (4)$$

4. Consistency of kernel methods

We now state our results on the consistency of kernel methods. We focus first on kernel mean embeddings (KMEs) and show *statistical consistency* of the standard estimator, i.e., that it recovers the KME of the random limit measure in the infinite-sample limit. We then attend to SVMs and show *L-consistency*, which focuses on the achieved risk instead of on the estimator itself and is the standard notion for a learning method (see (Steinwart & Christmann, 2008)).

4.1. Statistical consistency of kernel mean embeddings

Theorem 4.1. *Let $\mathcal{X}$ be a Polish space, X an $\mathcal{X}$-valued process, and $\mathcal{H}$ a scalar-valued RKHS of measurable functions on $\mathcal{X}$ with bounded kernel k. Finally, let P be a random measure on $\mathcal{X}$. If X is EWC with limit measure P, then*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} k(\cdot, X_i) = \int_{\mathcal{X}} k(\cdot, x)\mathrm{d}P(x) =: \mu_P, \quad (5)$$

where the convergence is in probability (resp. a.s.). Conversely, if the maximum mean discrepancy (MMD)

$$\mathrm{MMD} : (Q, R) \in \mathcal{P}(\mathcal{X})^2 \mapsto \|\mu_Q - \mu_R\|_{\mathcal{H}}$$

metrizes weak convergence in $\mathcal{P}(\mathcal{X})$, then (5) implies that X is EWC in probability (resp. a.s.).

While this result seems to follow from Theorem 3.6, we emphasize that it has weaker assumptions. In particular, Theorem 4.1 assumes neither local compactness nor a continuous kernel. This remarkable generalization relies on the fact that the topology generated by the MMD is coarser than that of weak convergence under the assumptions of the theorem (by Theorem 21 in (Sriperumbudur et al., 2010)). The metrization of weak convergence by the MMD is well-studied; see (Simon-Gabriel et al., 2023) for a comprehensive overview. For instance, a sufficient condition is that $\mathcal{X}$ is compact and k is bounded, continuous, and characteristic (Theorem 7 in the reference; recall that any Polish space is Hausdorff).

4.2. L-consistency of SVMs

We generalize consistency of a learning method to EWC data.

Definition 4.2. Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces, L be a loss function, and $Z = (X, Y)$ be an $\mathcal{X} \times \mathcal{Y}$-valued process. Assume that Z is EWC with limit measure J. We say that a learning method $\mathfrak{L}$ is L-consistent if,

$$\lim_{n\to\infty} \mathcal{R}_{L,J}(f_{Z_{1:n}}) = \mathcal{R}^\star_{L,J},$$

where convergence is in probability. Further, if the convergence is a.s., we say that $\mathfrak{L}$ is strongly L-consistent.

This definition raises the immediate concern of the measurability of the sets $\{\mathcal{R}_{L,J}(f_{Z_{1:n}}) \leq \mathcal{R}^\star_{L,J} + \epsilon\}$, where $\epsilon > 0$. For an SVM on a Polish space with outputs in a separable Hilbert space and in a separable RKHS of measurable functions, such sets are measurable. This result was already known in the case of a non-random measure J and scalar outputs (Lemmas 6.3 and 6.23 in (Steinwart & Christmann, 2008)); we extend it to our setting in Appendix D.

The next theorem is our main result. It guarantees that SVMs are consistent with EWC data, up to technical assumptions on the loss, RKHS, and $C_{\mathrm{b}}(\mathcal{X} \times \mathcal{Y}; \mathcal{G})$.

Theorem 4.3. *Let $\mathcal{X}$ be a locally compact Polish space, $\mathcal{G}$ a separable Hilbert space, $\mathcal{Y}$ a complete subset of $\mathcal{G}$, and $\mathcal{H}$ a separable $\mathcal{G}$-valued RKHS of continuous functions on $\mathcal{X}$ with bounded kernel K and continuous feature map Φ :*

$x \mapsto K(\cdot, x)$. Assume that $C_{\mathrm{b}}(\mathcal{X} \times \mathcal{Y}; \mathcal{H})$ is separable. Let L be a convex, continuously differentiable, locally Lipschitz continuous, and locally bounded loss function. Let $Z = (X, Y)$ be an $\mathcal{X} \times \mathcal{Y}$-valued process, and assume that Z is EWC in probability (resp. a.s.) with limit measure J. Assume that $\mathcal{H}$ is (L, J)-rich, $\mathbb{P}$-a.s. Then, there exists a sequence of strictly positive real numbers $(\lambda_n)_{n \in \mathbb{N}^\star}$ such that $\lim_{n \to \infty} \lambda_n = 0$ and the (λ_n)-SVM is L-consistent (resp. strongly L-consistent) for Z.

We emphasize again that usual consistency results do *not* assume that the loss function is continuously differentiable. The necessity for this requirement comes directly from the weak convergence that EWC guarantees: this assumption is more general than what is usually assumed, but comes at the price of weaker guarantees. Furthermore, Theorem 4.3 does not indicate at what speed the sequence $(\lambda_n)_{n \in \mathbb{N}^\star}$ should decrease. Unfortunately, no such speed exists without further assumptions, for two reasons. First, similarly to works with assumptions stronger than EWC, there is no such choice of $(\lambda_n)_{n \in \mathbb{N}^\star}$ uniform in the limit measure J; see for instance Steinwart et al. (2009) or Section 6.1 in Steinwart & Christmann (2008). Second, EWC allows the data to approximate the limit distribution arbitrarily slowly, adding another layer of necessary assumptions. In other words, a finite-sample analysis requires handling both the classical question of bounding the approximation error function (Steinwart & Christmann, 2008) and imposing a minimum speed on the EWC convergence.

5. Conclusion and outlook

We introduce the new notion of empirical weak convergence to address the consistency of kernel methods in the presence of dependent data. We show consistency for SVMs for EWC processes and allow an infinite-dimensional output, which is crucial to include related methods such as CKMEs. We discuss in details how EWC relates to and generalizes the existing usual assumptions for consistency, establishing it as a suitable basis paving the way for a more general theory of learning with dependent data.

An important open question is on the generalization of Theorem 4.3 to non-separable sets of continuous bounded functions. Indeed, that assumption is typically achieved through compactness of the input and output sets, which is relatively restrictive. Another relevant topic is to extend the theorem to stronger forms of convergence to a random measure, such as ESC. Finally, the topic of learning rates is a promising area for future work. While it is clear that the convergence involved in EWC may be arbitrarily slow, translating bounds on this speed into oracle inequalities for SVMs is an interesting extension of our results.

Acknowledgements

We thank Christian Fiedler and the anonymous reviewers for detailed and helpful comments.

Impact statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Adams, T. and Quas, A. *Ergodicity and Mixing Properties*, pp. 1–26. Springer, 2020.

Aliprantis, C. D. and Border, K. C. *Infinite dimensional analysis*. Springer, 3 edition, 2006.

Bauer, H. *Measure and integration theory*, volume 26. Walter de Gruyter, 2011.

Beaglehole, D., Belkin, M., and Pandit, P. On the inconsistency of kernel ridgeless regression in fixed dimensions. *SIAM Journal on Mathematics of Data Science*, 5(4): 854–872, 2023.

Beer, G. *Topologies on closed and closed convex sets*, volume 268. Springer Science & Business Media, 1993.

Bertsekas, D. and Shreve, S. E. *Stochastic optimal control: the discrete-time case*, volume 5. Athena Scientific, 1996.

Bogachev, V. I. *Measure theory*, volume 2. Springer, 2007.

Bradley, R. C. Basic Properties of Strong Mixing Conditions. A Survey and Some Open Questions. *Probability Surveys*, 2:107 – 144, 2005.

Buchholz, S. Kernel interpolation in sobolev spaces is not consistent in low dimensions. In *Conference on Learning Theory*, pp. 3410–3440. PMLR, 2022.

Buisson-Fenet, M., Solowjow, F., and Trimpe, S. Actively learning gaussian process dynamics. In *Learning for dynamics and control*, pp. 5–15. PMLR, 2020.

Caponnetto, A. and De Vito, E. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7:331–368, 2007.

Carmeli, C., De Vito, E., and Toigo, A. Vector valued reproducing kernel hilbert spaces of integrable functions and mercer theorem. *Analysis and Applications*, 4(04): 377–408, 2006.

Castaing, C. and Valadier, M. *Convex Analysis and Measurable Multifunctions*. Springer, 1977.

Csörgö, M. On the strong law of large numbers and the central limit theorem for martingales. *Transactions of the American Mathematical Society*, 131(1):259–275, 1968.

Grünewälder, S., Lever, G., Baldassarre, L., Patterson, S., Gretton, A., and Pontil, M. Conditional mean embeddings as regressors. In *International Conference on Machine Learning*, pp. 1803–1810, 2012.

Hanneke, S. Learning whenever learning is possible: Universal learning under general stochastic processes. *Journal of Machine Learning Research*, 22:5751–5866, 2021.

Henrikson, J. Completeness and total boundedness of the Hausdorff metric. *MIT Undergraduate Journal of Mathematics*, 1(69-80):10, 1999.

Hytönen, T., Van Neerven, J., Veraar, M., and Weis, L. *Analysis in Banach spaces*, volume 1. Springer, 2016.

Irle, A. On consistency in nonparametric estimation under mixing conditions. *Journal of multivariate analysis*, 60 (1):123–147, 1997.

Kallenberg, O. *Random measures, theory and applications*. Springer, 2017.

Klebanov, I., Schuster, I., and Sullivan, T. J. A rigorous theory of conditional mean embeddings. *SIAM Journal on Mathematics of Data Science*, pp. 583–606, 2020.

Klenke, A. *Probability theory: a comprehensive course*. Springer Science & Business Media, 2013.

Kruger, A. Y. *Nonsmooth analysis: Fréchet subdifferentialsNonsmooth Analysis: Fréchet Subdifferentials*, pp. 2651–2658. Springer, 2009.

Li, Y. and Li, J. Criteria for Feller transition functions. *Journal of mathematical analysis and applications*, 359 (2):653–665, 2009.

Li, Z., Meunier, D., Mollenhauer, M., and Gretton, A. Optimal Rates for Regularized Conditional Mean Embedding Learning. *Advances in Neural Information Processing Systems*, 2022.

Megginson, R. E. *An introduction to Banach space theory*. Springer, 2012.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Mollenhauer, M. and Koltai, P. Nonparametric approximation of conditional expectation operators. *arXiv preprint arXiv:2012.12917*, 2020.

Park, J. and Muandet, K. Regularised Least-Squares Regression with Infinite-Dimensional Output Space. *arXiv preprint arXiv:2010.10973*, 2022.

Rakhlin, A. and Zhai, X. Consistency of interpolation with laplace kernels is a high-dimensional phenomenon. In *Conference on Learning Theory*, pp. 2595–2623. PMLR, 2019.

Schölkopf, B. and Smola, A. J. *Learning with kernels*, volume 4. Citeseer, 1998.

Simchowitz, M., Mania, H., Tu, S., Jordan, M. I., and Recht, B. Learning without mixing: Towards a sharp analysis of linear system identification. In *Conference On Learning Theory*, pp. 439–473. PMLR, 2018.

Simon-Gabriel, C.-J., Barp, A., Schölkopf, B., and Mackey, L. Metrizing weak convergence with maximum mean discrepancies. *Journal of Machine Learning Research*, 24(184):1–20, 2023.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. R. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.

Steinwart, I. and Christmann, A. *Support vector machines*. Springer, 2008.

Steinwart, I. and Christmann, A. Fast learning from non-iid observations. *Advances in neural information processing systems*, 22, 2009.

Steinwart, I., Hush, D., and Scovel, C. Learning from dependent observations. *Journal of Multivariate Analysis*, 100:175–194, 2009.

Von Luxburg, U. and Schölkopf, B. Statistical learning theory: Models, concepts, and results. In *Handbook of the History of Logic*, volume 10, pp. 651–706. Elsevier, 2011.

von Rohr, A., Neumann-Brosig, M., and Trimpe, S. Probabilistic robust linear quadratic regulators with gaussian processes. In *Learning for Dynamics and Control*, pp. 324–335. PMLR, 2021.

Ziemann, I. and Tu, S. Learning with little mixing. *Advances in Neural Information Processing Systems*, 35: 4626–4637, 2022.

A. Proofs

A.1. Proofs for Section 3.1

Proof of Proposition 3.2. Let d_1 and d_2 be two metrics that metrize weak convergence on $\mathcal{P}(\mathcal{X})$. By definition, sequences of $\mathcal{P}(\mathcal{X})$ that converge for d_1 also converge for d_2, and vice versa. It immediately follows that the notion of EWC a.s. is unaffected by the choice of the metric. Furthermore, we can apply Corollary 20.8 in (Bauer, 2011) to the sequence $(d(\eta_n^X, P))_{n\in\mathbb{N}^\star}$ to show that a process X is EWC in probability if, and only if, for any strictly increasing sequence $k \subset \mathbb{N}^\star$, there exists a subsequence $(k_{m_n}) \subset \mathbb{N}^\star$ such that $\lim_{n\to\infty} \eta^X_{k_{m_n}} = P$, where the convergence is a.s. Since a.s. convergence of random measures is independent of the chosen metric that metrizes weak convergence, we deduce that so is the notion of a process EWC in probability.

We are left to show uniqueness of the limit measure. We only consider the case where X is EWC in probability as the other case follows immediately since any process EWC a.s. is also EWC in probability with the same limit measure. The result follows immediately from uniqueness of the limit in probability in a metric space; we repeat the proof for completeness. Let d be a metric that metrizes weak convergence on $\mathcal{P}(\mathcal{X})$ and P and Q be two limit (random) measures of X. For any $\epsilon > 0$ and $n \in \mathbb{N}^\star$, we have

$$\begin{aligned}\mathbb{P}[d(P,Q) > \epsilon] &\leq \mathbb{P}[d(\eta_n^X, P) > \epsilon \text{ or } d(\eta_n^X, Q) > \epsilon] \\ &\leq \mathbb{P}[d(\eta_n^X, P) > \epsilon] + \mathbb{P}[d(\eta_n^X, Q) > \epsilon],\end{aligned}$$

by the triangle inequality for d and the union bound. Since the right-hand side (RHS) goes to 0 as $n \to \infty$, we deduce that $d(P,Q) = 0$ a.s., which is the desired result. □

Proof of Theorem 3.4. The equivalence (i) $\iff$ (ii) immediately follows from the definition of weak convergence. Indeed, we have

$$\begin{aligned}\{\eta_n^X \rightsquigarrow P\} &= \{\forall f \in C_\mathrm{b}(\mathcal{X};\mathbb{R}), \lim_{n\to\infty} \eta_n^X f = Pf\} \\ &= \{\forall f \in C_\mathrm{b}(\mathcal{X};\mathbb{R}), \limsup_n |\eta_n^X f - Pf| = 0\}.\end{aligned}$$

Since d is continuous, it follows that both sets are measurable. Taking probabilities on both sides shows the equivalence.

Next, it is also clear that (ii) $\implies$ (iii), since for any $g \in C_\mathrm{b}(\mathcal{X};\mathbb{R})$,

$$\{\forall f \in C_\mathrm{b}(\mathcal{X};\mathbb{R}), \lim_{n\to\infty} \eta_n^X f = Pf\} \subset \{\lim_{n\to\infty} \eta_n^X g = Pg\},$$

and thus the RHS has probability 1 if (ii) holds.

Finally, we further assume that $\mathcal{X}$ is compact and show (iii) $\implies$ (ii). Assume that (iii) holds. By compactness and separability of $\mathcal{X}$, $(C_\mathrm{b}(\mathcal{X};\mathbb{R}), \|\cdot\|_\infty)$ is separable; let $\Phi = (\phi_m)_{m\in\mathbb{N}^\star}$ be a dense family thereof. Define for all $n \in \mathbb{N}^\star$ the event

$$N_m = \left\{\limsup_n |\eta_n^X \phi_m - P\phi_m| > 0\right\}.$$

It is the set of non-convergence of $\eta_n^X \phi_m$ to $P\phi_m$. Let

$$N = \bigcup_{m=1}^{\infty} N_m.$$

By assumption, $\mathbb{P}[N_m] = 0$ for all $m \in \mathbb{N}^\star$ and thus $\mathbb{P}[N] = 0$. Now, let $f \in C_\mathrm{b}(\mathcal{X};\mathbb{R})$ be fixed, and take $(\psi_m)_{m\in\mathbb{N}^\star} \subset \Phi$ be such that $\|\psi_m - f\|_\infty \to 0$ as $m \to \infty$. Let $\omega \in \Omega \setminus N$; from now on, we consider that all variables are evaluated in ω but drop the explicit dependency for conciseness. Let $\epsilon > 0$ be arbitrary, and take $M \in \mathbb{N}^\star$ such that for all $m \geq M$, $\|\psi_m - f\|_\infty \leq \frac{\epsilon}{2}$. Then, for all $m \geq M$ and $n \in \mathbb{N}^\star$,

$$\begin{aligned}\left|\eta_n^X f - Pf\right| &\leq \left|(\eta_n^X - P)(f - \psi_m)\right| + \left|\eta_n^X \psi_m - P\psi_m\right| \\ &\leq |\eta_n^X - P| \cdot |f - \psi_m| + \left|\eta_n^X \psi_m - P\psi_m\right| \\ &\leq 2 \cdot \frac{\epsilon}{2} + \left|\eta_n^X \psi_m - P\psi_m\right|,\end{aligned} \tag{6}$$

where the second inequality comes from the triangle inequality for the signed measure $\eta_n^X - P$ of which $|\eta_n^X - P|$ is the absolute variation, and the third uses the fact that $|\eta_n^X - P|\mathbf{1}_{\mathcal{X}} \leq \eta_n^X \mathbf{1}_{\mathcal{X}} + P\mathbf{1}_{\mathcal{X}} = 2$. Taking the $\limsup$ over n thus yields

$$\limsup_n \left|\eta_n^X f - Pf\right| \leq \epsilon,$$

for all $\epsilon > 0$. Indeed, the second term vanishes by assumption on ω. We deduce that $\limsup_n \left|\eta_n^X f - Pf\right| = 0$ for all $f \in C_{\mathrm{b}}(\mathcal{X};\mathbb{R})$ and $\omega \in \Omega \setminus N$, which concludes the proof. □

Proof of Theorem 3.5. Let d be a metric that metrizes weak convergence on $\mathcal{P}(\mathcal{X})$. The equivalence (i) $\iff$ (ii) immediately follows from Corollary (20.8) in (Bauer, 2011). Indeed, by the cited corollary, the process X is EWC with limit measure P if, and only if, for any strictly increasing sequence $k \subset \mathbb{N}^\star$, there exists another strictly increasing sequence $m \subset \mathbb{N}^\star$ such that $\lim_{n\to\infty} d(\eta_{k_{m_n}}^X, P) = 0$ with probability 1. Then, for any fixed such sequence k and m, the definition of weak convergence shows that $\lim_{n\to\infty} d(\eta_{k_{m_n}}^X, P) = 0$ holds with probability 1 if, and only if, the event $\{\forall f \in C_{\mathrm{b}}(\mathcal{X};\mathbb{R}), \lim_{n\to\infty} \eta_{k_{m_n}} f = Pf\}$ has probability 1. The equivalence follows.

It is also clear that (ii) implies (iii) by using the converse implication of Corollary (20.8) in (Bauer, 2011). Indeed, (ii) immediately implies that for all $f \in C_{\mathrm{b}}(\mathcal{X};\mathbb{R})$ and all strictly increasing sequence $k \subset \mathbb{N}^\star$, there exists another strictly increasing sequence $m \subset \mathbb{N}^\star$ such that $\lim_{n\to\infty} \eta_{k_{m_n}} f = Pf$ a.s. The cited result then implies (iii).

Finally, we further assume that $\mathcal{X}$ is compact and show (iii) $\implies$ (ii). Let $k \subset \mathbb{N}^\star$ be a strictly increasing sequence. We proceed by constructing a strictly increasing sequence $m \subset \mathbb{N}^\star$ so that the event $\{\forall f \in C_{\mathrm{b}}(\mathcal{X};\mathbb{R}), \lim_{n\to\infty} \eta_{k_{m_n}}^X f = Pf\}$ has probability 1; i.e., neither the subsequence m nor the null set of non-convergence depend on the chosen function f, contrary to the ones given by (iii). Since $\mathcal{X}$ is compact, $(C_{\mathrm{b}}(\mathcal{X};\mathbb{R}), \|\cdot\|_\infty)$ is separable; let $\Phi = (\phi_p)_{p\in\mathbb{N}^\star}$ be a dense family thereof. The construction of the sequence m follows from a diagonal argument. First, we construct inductively a double-indexed sequence $(m_{n,p})_{(n,p)\in(\mathbb{N}^\star)^2}$ that satisfies the following properties for all $p \in \mathbb{N}^\star$:

- the sequence $(m_{n,p})_{n\in\mathbb{N}^\star}$ is strictly increasing;
- $\lim_{n\to\infty} \eta_{k_{m_{n,p}}} \phi_p = P\phi_p$ on all of $\Omega \setminus N_p$ for some measurable $N_p \in \mathcal{A}$ with $\mathbb{P}[N_p] = 0$;
- $(k_{m_{n,p+1}})_{n\in\mathbb{N}^\star}$ is a subsequence of $(k_{m_{n,p}})_{n\in\mathbb{N}^\star}$.

Such a construction is indeed possible by applying (iii) and Corollary 20.8 in (Bauer, 2011) to the sequence $(\eta_{k_{m_{n,p}}}^X \phi_{p+1})_{n\in\mathbb{N}^\star}$ at every induction step; we forego detailing it for conciseness. Then, we define inductively the sequence m as follows. We first take $m_1 := m_{1,1}$. Then, assuming that m_i is constructed for all $i \leq p$ for some $p \in \mathbb{N}^\star$, define

$$m_{p+1} := \min\{m_{i,p+1} \mid i \in \mathbb{N}^\star \wedge m_{i,p+1} > m_p\}.$$

The set over which we minimize is indeed nonempty since $m_{i,p+1} \to \infty$ as $i \to \infty$. We claim that the hereby constructed sequence m has the announced property. First, m is clearly strictly increasing, by construction. Next, let $N = \bigcup_{p\in\mathbb{N}^\star} N_p$ and take $\omega \in \Omega \setminus N$; from now on, we consider that all variables are evaluated in ω but drop the explicit dependency for conciseness. For any $p \in \mathbb{N}^\star$, the sequence m satisfies $\lim_{n\to\infty} \eta_{k_{m_n}} \phi_p = P\phi_p$. Indeed, by construction, the sequence $(k_{m_n})_{n\geq p}$ is a subsequence of the sequence $(k_{m_{n,p}})_{n\geq p}$, and $\lim_{n\to\infty} \eta_{k_{m_{n,p}}}^X \phi_p = P\phi_p$. Now, let $f \in C_{\mathrm{b}}(\mathcal{X};\mathbb{R})$ be arbitrary, and take $(\psi_q)_{q\in\mathbb{N}^\star} \subset \Phi$ be such that $\|\psi_q - f\|_\infty \to 0$ as $n \to \infty$. Let $\epsilon > 0$, and take $Q \in \mathbb{N}^\star$ such that $\|f - \psi_q\|_\infty \leq \frac{\epsilon}{2}$ for all $q \geq Q$. The exact same calculations as the ones in (6) with the index k_{m_n} instead of n directly show that for all $n \in \mathbb{N}^\star$ and $q \geq Q$,

$$|\eta_{k_{m_n}} f - Pf| \leq \epsilon + |\eta_{k_{m_n}} \psi_q - P\psi_q|.$$

The second term vanishes when taking the $\limsup$ over n by the property of the sequence m. This shows that for all $f \in C_{\mathrm{b}}(\mathcal{X};\mathbb{R})$ and $\epsilon > 0, \limsup_n |\eta_{k_{m_n}} f - Pf| \leq \epsilon$, and thus the left-hand side (LHS) is equal to 0. This is true for all $\omega \in \Omega \setminus N$ and N has probability 0; hence, $\lim_{n\to\infty} \eta_{k_{m_n}} f = Pf$, where the convergence is almost sure. Since the sequence m is independent of f, this shows the result and concludes the proof. □

Proof of Theorem 3.6. We begin by showing the claim in the case where X is EWC a.s., and assume without loss of generality that for all $x \in \mathcal{X}$, $\|\phi(x)\|_{\mathcal{H}} \leq 1$.

Let $\epsilon > 0$. We apply theorem B.6 to the countable family of random measures $\mathcal{Q} = \{\eta_n^X \mid n \in \mathbb{N}^\star\} \cup \{P\}$, which is sequentially compact a.s. by assumption (here, we rely on completeness of $(\Omega, \mathcal{A})$ to guarantee the measurability of the event). Therefore, there exists a random compact set K_ϵ such that the following holds a.s.:

$$\begin{aligned} P(K_\epsilon) &\geq 1 - \epsilon, \\ \forall n \in \mathbb{N}^\star,\ \eta_n^X(K_\epsilon) &\geq 1 - \epsilon. \end{aligned} \tag{7}$$

Next, since $\mathcal{H}$ is a Hilbert space, it has the b-approximation property for some $b > 0$ (see the discussion after Definition 4.1.34 in (Megginson, 2012)). Applying this property to the compact set $\phi(K_\epsilon)$ shows that for all $\omega \in \Omega$, there exists a bounded operator $S(\omega) : \mathcal{H} \to \mathcal{H}$ with finite rank $m(\omega) \in \mathbb{N}^\star$ such that

$$\forall x \in K_\epsilon(\omega), \|S(\omega)\phi(\omega)(x) - \phi(\omega)(x)\|_{\mathcal{H}} \leq \epsilon, \text{ and } \|S(\omega)\|_{\mathcal{L}(\mathcal{H})} \leq b. \tag{8}$$

It is clear from the proof of the bounded approximation property (Theorem 4.1.33 in (Megginson, 2012)) that the map $\omega \mapsto S(\omega)$ can be taken to be measurable; it defines a random operator $S : \mathcal{H} \to \mathcal{H}$ with finite rank m that satisfies (8) a.s. We now take an $\mathcal{H}$-valued process[4] e such that $(e_1, \dots, e_m)$ is a.s. an orthonormal basis (ONB) of $S\mathcal{H}$, and let

$$\forall j \in \mathbb{N}^\star, f_j : \omega \in \Omega \mapsto \langle e_j(\omega),\ S(\omega)\phi(\omega)(\cdot)\rangle_{\mathcal{H}} \in \mathbb{R}^{\mathcal{X}}. \tag{9}$$

By continuity of ϕ, of S, and of the scalar product, the process f takes values in $C_{\mathrm{b}}(\mathcal{X}; \mathbb{R})$.

We are now equipped to show the convergence announced. We have the following bound:

$$\left\|\eta_n^X \phi - P\phi\right\|_{\mathcal{H}} \leq \left\|\eta_n^X \phi - \eta_n^X S\phi\right\|_{\mathcal{H}} + \left\|\eta_n^X S\phi - PS\phi\right\|_{\mathcal{H}} + \|PS\phi - P\phi\|_{\mathcal{H}}. \tag{10}$$

We further bound these three terms separately to show a.s. convergence of the LHS to 0. For this, we leverage the equivalent characterization of a.s. convergence of Lemma 20.6 in (Bauer, 2011) and show that for all $\alpha > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left[\sup_{\ell \geq n} \left\|\eta_\ell^X \phi - P\phi\right\|_{\mathcal{H}} > \alpha\right] = 0.$$

To this end, we first apply that same characterization to the variable $\sqrt{m}\sup_{j\in\{1,\dots,m\}}|\eta_n^X f_j - Pf_j|$, which converges to 0 a.s., yielding the existence of $n_\epsilon > 0$ such that the following holds with probability not less than $1 - \epsilon$:

$$\sup_{n \geq n_\epsilon} \sup_{j \in \{1,\dots,m\}} |\eta_n^X f_j - Pf_j| \leq \epsilon \cdot m^{-1/2}. \tag{11}$$

Let now $A \in \mathcal{A}$ be an event with probability 1 such that 1. (7) and (8) hold; and 2. $(e_1, \dots, e_m)$ is an ONB of $S\mathcal{H}$ for all $\omega \in A$, $n \geq n_\epsilon$, and $\omega \in A$ that satisfies (11). We assume that all variables are evaluated in $\omega \in A$, but drop the explicit dependency for conciseness until mentioned otherwise. The last term in (10) is bounded as follows:

$$\begin{aligned} \|PS\phi - P\phi\|_{\mathcal{H}} &\leq P\,\|S\phi - \phi\|_{\mathcal{H}} \\ &= P\,\|S\phi - \phi\|_{\mathcal{H}} \cdot \mathbf{1}_{K_\epsilon} + P\,\|S\phi - \phi\|_{\mathcal{H}} \cdot \mathbf{1}_{\mathcal{X}\setminus K_\epsilon} \\ &\leq \epsilon P\mathbf{1}_{K_\epsilon} + \sup_{x \in \mathcal{X}\setminus K_\epsilon} \|S\phi(x) - \phi(x)\|_{\mathcal{H}}\, P\mathbf{1}_{\mathcal{X}\setminus K_\epsilon} \\ &\leq \epsilon + \sup_{x \in \mathcal{X}\setminus K_\epsilon} \|S\phi(x) - \phi(x)\|_{\mathcal{H}} \cdot \epsilon \\ &\leq \epsilon + (b+1)\cdot\epsilon = (b+2)\epsilon, \end{aligned}$$

where we use the facts that $\|S\phi - \phi\|_{\mathcal{H}} \cdot \mathbf{1}_{K_\epsilon} \leq \epsilon \cdot \mathbf{1}_{K_\epsilon}$, $P(K_\epsilon) \geq 1 - \epsilon$, $\sup_x \|\phi(x)\|_{\mathcal{H}} \leq 1$, and $\|S\|_{\mathcal{L}(\mathcal{H})} \leq b$. The first term in (10) is bounded similarly, as the same computations with η_n^X instead of P show

$$\left\|\eta_n^X \phi - \eta_n^X S\phi\right\|_{\mathcal{H}} \leq \eta_n^X\,\|S\phi - \phi\|_{\mathcal{H}} \cdot \mathbf{1}_{K_\epsilon} + \eta_n^X\,\|S\phi - \phi\|_{\mathcal{H}} \cdot \mathbf{1}_{\mathcal{X}\setminus K_\epsilon} \leq (b+2)\epsilon.$$

[4]There is indeed a choice of the map $\omega \mapsto e(\omega)$ that is measurable from separability of $\mathcal{H}$; we skip the proof for conciseness.

where we use $\eta_n^X(K_\epsilon) \geq 1-\epsilon$. Finally, we bound the second term by leveraging the fact that S has finite rank:

$$\begin{aligned}\left\|\eta_n^X S\phi - PS\phi\right\|_{\mathcal{H}} &= \left(\sum_{j=1}^m \left\langle e_j,\, \eta_n^X S\phi - PS\phi\right\rangle^2\right)^{1/2} \\ &= \left(\sum_{j=1}^m (\eta_n^X f_j - Pf_j)^2\right)^{1/2} \\ &\leq \sqrt{m} \sup_{j\in\{1,\dots,m\}} |\eta_n^X f_j - Pf_j| \\ &\leq \epsilon,\end{aligned}$$

where the last inequality comes from the fact that ω satisfies (11). Since $n \geq n_\epsilon$ was arbitrary, we have shown that

$$\sup_{n\geq n_\epsilon} \left\|\eta_n^X \phi - P\phi\right\| \leq (b+2)\epsilon + \epsilon + (b+2)\epsilon = (2b+5)\epsilon. \tag{12}$$

This bound holds for any $\omega \in A$ that satisfies (11); therefore, we deduce that it holds with probability at least $1-\epsilon$. Since $\epsilon > 0$ was arbitrary, the conclusion follows as announced by leveraging (20.6) in (Bauer, 2011).

Now, we handle the case when X is only EWC in probability. Corollary 20.8 in (Bauer, 2011) shows that this is equivalent to the fact that any subsequence $(d(\eta_{k_n}^X, P))_{n\in\mathbb{N}^\star}$ of $(d(\eta_n^X, P))_{n\in\mathbb{N}^\star}$ has itself a subsequence that converges to 0 a.s., say, $(d(\eta_{k_{m_n}}^X, P))_{n\in\mathbb{N}^\star}$. Here $k : \mathbb{N}^\star \to \mathbb{N}^\star$ and $m : \mathbb{N}^\star \to \mathbb{N}^\star$ are strictly increasing sequences. But then, $(\eta_{k_{m_n}}^X)_{n\in\mathbb{N}^\star}$ converges weakly to P with probability 1. We can thus repeat the above proof *as is* by replacing $(\eta_n^X)_{n\in\mathbb{N}^\star}$ with $(\eta_{k_{m_n}}^X)_{n\in\mathbb{N}^\star}$. Indeed, the assumption that X is EWC a.s. only appears in two places: when applying Theorem B.6, and to establish the existence of n_ϵ such that (11) hold. Both of these results only rely on the fact that $(\eta_n^X)_{n\in\mathbb{N}^\star}$ converges weakly to P a.s., which holds for $(\eta_{k_{m_n}}^X)_{n\in\mathbb{N}^\star}$ as well by construction of ℓ. Overall, this shows that (12) holds for this specific sub-subsequence $(\eta_{k_{m_n}}^X)_{n\in\mathbb{N}^\star}$, and thus $\lim_{n\to\infty} \eta_{k_{m_n}}^X \phi = P\phi$ a.s. by (20.6) in (Bauer, 2011). In other words, every subsequence of $(\eta_n^X \phi)_{n\in\mathbb{N}^\star}$ has itself a subsequence that converges to $P\phi$ a.s., which is equivalent to convergence in probability of $(\eta_n^X \phi)_{n\in\mathbb{N}^\star}$ to $P\phi$ and concludes the proof. □

Proof of Corollary 3.7. The proof is an identical repetition of the proofs of the implications (iii) $\Longrightarrow$ (ii) of Theorems 3.4 and 3.5, up to the replacement of the dense family of $C_{\mathrm{b}}(\mathcal{X};\mathbb{R})$ by a dense family of $\mathcal{F}$ and of absolute values by $\|\cdot\|_{\mathcal{H}}$. □

A.2. Proofs for Section 3.2

Proof of Lemma 3.11. It is clear that the process X converges weakly. We show that it is not EWC in probability. To this end, we show that (iii) in Theorem 3.5 does not hold. Let $f \in C_{\mathrm{b}}(\{0,1\})$, and assume that $u_n = \eta_n^X f$ converges in probability to some real-valued random variable ℓ. We show that f is constant. Let $n_p = 10^p - 1$ for all $p \in \mathbb{N}$. Since X is constant on index intervals of the form $[n_q+1, n_{q+1}] \cap \mathbb{N}$, grouping the terms in the sum on these intervals yields

$$\begin{aligned}u_{n_p} &= \frac{1}{n_p}\sum_{q=0}^{p-1}(n_{q+1}-n_q) f(X_{n_q+1}) \\ &= \frac{1}{n_p}\sum_{q=0}^{p-1}(n_{q+1}-n_q)\cdot\big(\mathbf{1}_{2\mathbb{N}}(q) f(X_1) + \mathbf{1}_{2\mathbb{N}+1}(q) f(1-X_1)\big),\end{aligned}$$

where $\mathbf{1}_{\mathcal{S}}$ is the indicator function of a set $\mathcal{S}$. Now, let $v_p = u_{n_{p+1}} - u_{n_p}$. For an even value of p, we have

$$\begin{aligned}v_p &= \left(\frac{n_p}{n_{p+1}} - 1\right) u_{n_p} + \frac{n_{p+1}-n_p}{n_{p+1}} f(X_1) \\ &= \frac{n_{p+1}-n_p}{n_{p+1}}\left(f(X_1) - u_{n_p}\right).\end{aligned}$$

For an odd value of p, similar calculations reveal

$$v_p = \frac{n_{p+1} - n_p}{n_{p+1}} \left(f(1 - X_1) - u_{n_p} \right).$$

We now use the facts that $v_p \to 0$ in probability as $n \to \infty$ (by convergence in probability of u and the definition of v_p) and $\frac{n_{p+1}-n_p}{n_{p+1}} \to \frac{9}{10}$ as $p \to \infty$ to conclude that both of the following hold $\mathbb{P}$-a.s., by $\mathbb{P}$-a.s.-uniqueness of the limit of convergence in probability:

$$\frac{9}{10} \left(f(X_1) - \ell \right) = 0, \mathbb{P}\text{-a.s.}$$
$$\frac{9}{10} \left(f(1 - X_1) - \ell \right) = 0, \mathbb{P}\text{-a.s.}$$

Therefore, $f(X_1) = f(1 - X_1)$, $\mathbb{P}$-a.s., which shows that f is constant. We can now conclude the proof: if f is not constant, then the contraposition of the above shows that $\eta_n^X f$ does not converge in probability to any random variable. Since, by Theorem 3.5, this is a necessary condition for EWC in probability, X is not EWC in probability. □

Proof of Lemma 3.12. The result follows immediately from Lemma 2.5 in (Steinwart et al., 2009) together with Theorems 3.4 and 3.5. Indeed, assume that X satisfies the WLLNE, and let $P \in \mathcal{P}(\mathcal{X})$ be its asymptotic mean (which exists by Theorem 2.4 in (Steinwart et al., 2009)). Then, for all f in $\mathcal{L}_\infty(\mathcal{X}; \mathbb{R})$, the following convergence in probability holds:

$$Pf = \lim_{n \to \infty} \eta_n^X f, \tag{13}$$

from Lemma 2.5 in the above reference. In particular, $C_\mathrm{b}(\mathcal{X}; \mathbb{R}) \subset \mathcal{L}_\infty(\mathcal{X}; \mathbb{R})$, showing that (iii) in Theorem 3.5 holds. Since $\mathcal{X}$ is separable, we deduce that X is EWC with limit measure P. Further, if X satisfies the SLLNE, then Lemma 2.5 in the same reference states that (13) holds where the convergence is now almost sure. The argument $C_\mathrm{b}(\mathcal{X}; \mathbb{R}) \subset \mathcal{L}_\infty(\mathcal{X}; \mathbb{R})$ now enables to conclude that (iii) in Theorem 3.4 holds, and thus X is EWC a.s. □

Proof of Proposition 3.16. The implication that ESC implies EWC with the same limit measure follows trivially from the fact that the topology of convergence in total variation is finer than that of weak convergence. Finally, we only show that EWC in probability implies weak AMS, since the case for EWC a.s. follows, and the proof that ESC implies AMS is similar by leveraging a result similar to (i) $\implies$ (iii) in Theorem 3.5 for ESC processes involving $\mathcal{L}_\infty(\mathcal{X}; \mathbb{R})$ instead of $C_\mathrm{b}(\mathcal{X}; \mathbb{R})$, and whose proof is left to the reader.

Let $f \in C_\mathrm{b}(\mathcal{X}; \mathbb{R})$. For any $n \in \mathbb{N}^\star$, we have

$$\left| \frac{1}{n} \sum_{i=1}^{N} \mathbb{E}[f(X_i)] - (\mathbb{E}P)f \right| = \left| \mathbb{E}[\eta_n^X f - Pf] \right| \leq \mathbb{E}[|\eta_n^X f - Pf|],$$

where the identity $(\mathbb{E}P)f = \mathbb{E}[Pf]$ follows from usual properties of the intensity measure. Now, from Theorem 3.5, we have

$$\lim_{n \to \infty} \eta_n^X f = Pf, \tag{14}$$

where convergence is in probability. Furthermore, the sequence $(\eta_n^X f)_{n \in \mathbb{N}^\star}$ is uniformly integrable by boundedness of f; indeed, we have $\mathbb{E}[|\eta_n^X f|] \leq \|f\|_\infty$ for all $n \in \mathbb{N}^\star$. It thus follows from known results on convergence of random variables that (14) is equivalent to convergence of $(\eta_n^X f)_{n \in \mathbb{N}^\star}$ to Pf in $L_1(\Omega, \mathbb{P}; \mathbb{R})$. Hence, $\mathbb{E}[|\eta_n^X f - Pf|] \to 0$ as $n \to \infty$ and the result follows. □

A.3. Proofs for Section 3.3

Proof of Theorem 3.19. We begin with the direct implication and focus on the proof for X, as that for Y is identical. Let $\pi_\mathcal{X} : (x, y) \in \mathcal{X} \times \mathcal{Y} \mapsto x \in \mathcal{X}$. Clearly, $\pi_\mathcal{X}$ is continuous. Now, if Z is EWC a.s., it follows that (ii) in Theorem 3.4 holds for Z, and in particular with J the limit measure of Z

$$\mathbb{P}[\forall f \in C_\mathrm{b}(\mathcal{X}; \mathbb{R}),\ \limsup_n |\eta_n^Z (f \circ \pi_\mathcal{X}) - J(f \circ \pi_\mathcal{X})| = 0] = 1,$$

because $f \circ \pi_{\mathcal{X}} \in C_{\mathrm{b}}(\mathcal{X} \times \mathcal{Y}; \mathbb{R})$ for any $f \in C_{\mathrm{b}}(\mathcal{X}; \mathbb{R})$. Furthermore, $\eta_n^Z(f \circ \pi_{\mathcal{X}}) = \eta_n^X f$ and $J(f \circ \pi_{\mathcal{X}}) = Pf$, where P is the random measure defined as $P[A] = J[A \times \mathcal{Y}]$ for all $A \in \mathcal{B}(\mathcal{X})$. This shows that

$$\mathbb{P}[\forall f \in C_{\mathrm{b}}(\mathcal{X}; \mathbb{R}),\ \limsup_n |\eta_n^X f - Pf| = 0] = 1,$$

and shows that X is EWC a.s. by Theorem 3.4. If Z is only EWC in probability, then the exact same reasoning as above but using Theorem 3.5 instead of Theorem 3.4 shows that X is EWC in probability.

We now assume that $\mathcal{X}$ and $\mathcal{Y}$ are compact and show the converse implication. We begin with the case where X is EWC a.s. Let $f \in C_{\mathrm{b}}(\mathcal{X} \times \mathcal{Y}; \mathbb{R})$ be fixed, and define for all $n \in \mathbb{N}^\star$ $\Delta_n = f(X_n, Y_n) - \mathbb{E}[f(X_n, Y_n) \mid X_{1:n}, Y_{1:n-1}]$, where we recall the convention that conditioning on $Y_{1:0}$ means conditioning on the trivial σ-algebra $\{\emptyset, \Omega\}$. It satisfies $\mathbb{E}[\Delta_1] = 0$ and, for all $n \geq 2$, $\mathbb{E}[\Delta_n \mid X_{1:n}, Y_{1:n-1}] = 0$ a.s. Further, the sequence $(\Delta_n)_{n \in \mathbb{N}^\star}$ is uniformly bounded by $2 \cdot \|f\|_\infty$. Hence, it satisfies the assumptions of Theorem 1 in (Csörgö, 1968) with $b_n = n$ for all $n \in \mathbb{N}^\star$. We deduce from that same theorem that

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^n \Delta_i = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^n f(X_i, Y_i) - \mathbb{E}[f(X_i, Y_i) \mid X_{1:i}, Y_{1:i-1}] = 0, \tag{15}$$

where convergence is a.s. Next, we show that

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^n \mathbb{E}[f(X_i, Y_i) \mid X_{1:i}, Y_{1:i-1}] = Jf, \tag{16}$$

with probability 1, where J is defined in (4). Define $g : x \in \mathcal{X} \mapsto \int_{\mathcal{Y}} f(x, y) p(x, \mathrm{d}y)$; we have $g \in C_{\mathrm{b}}(\mathcal{X}; \mathbb{R})$ since p is Feller continuous. It holds with probability 1 that

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \mathbb{E}[f(X_i, Y_i) \mid X_{1:i}, Y_{1:i-1}] &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}[f(X_i, Y_i) \mid X_i] \\ &= \eta_n^X g, \end{aligned}$$

and thus by (iii) in Theorem 3.4 applied to X, which is EWC,

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^n \mathbb{E}[f(X_i, Y_i) \mid X_{1:i}, Y_{1:i-1}] = Pg, \quad \mathbb{P}\text{-a.s.}$$

Furthermore,

$$Pg = \int_{\mathcal{X}} g(x) \mathrm{d}P(x) = \int_{\mathcal{X}} \int_{\mathcal{Y}} f(x, y) p(x, \mathrm{d}y) \mathrm{d}P(x) = Jf,$$

showing that (16) holds. Combining this with (15) shows

$$\lim_{n \to \infty} \eta_n^Z f = Jf, \tag{17}$$

where the convergence is a.s. The conclusion follows by leveraging Theorem 3.4, since $\mathcal{X} \times \mathcal{Y}$ is compact.

Finally, the case where X is only EWC in probability follows by repeating the above steps by leveraging Theorem 3.5 instead of Theorem 3.4. The only difference is that the convergence in (16) is now in probability, and thus so is the one in (17), concluding the proof. □

A.4. Proofs for Section 4.1

Proof of Theorem 4.1. We begin with the following observation: from (ii) in Theorem 21 in (Sriperumbudur et al., 2010), there exists a constant $C > 0$ such that for all Q_1 and Q_2 in $\mathcal{P}(\mathcal{X})$,

$$\mathrm{MMD}(Q_1, Q_2) \leq C \cdot \beta(Q_1, Q_2),$$

where β is the Dudley metric and is known to metrize weak convergence on $\mathcal{P}(\mathcal{X})$ (see the discussion above (28) in the same reference). We emphasize that, although (ii) in Theorem 21 requires an additional separability assumption, that assumption

is only used in the proof to show the upper bound on the Dudley metric; therefore, the lower bound still holds without the assumption. Consequently, if $(Q_n)_{n\in\mathbb{N}^\star} \subset \mathcal{P}(\mathcal{X})$ converges weakly to $Q \in \mathcal{P}(\mathcal{X})$, then $\mathrm{MMD}(Q_n, Q) \to 0$ as $n \to \infty$.

Let us now assume that X is EWC in probability with limit measure P. We use the same argument as in the end of the proof of Theorem 3.6 to deduce from Corollary 20.8 in (Bauer, 2011) that this is equivalent to the fact that, for any strictly increasing $\ell \subset \mathbb{N}^\star$, there exists another strictly increasing $m \subset \mathbb{N}^\star$ such that $\eta^X_{\ell_{m_n}} \rightsquigarrow P$ a.s.. But then, $\lim_{n\to\infty} \mathrm{MMD}(\eta^X_{\ell_{m_n}}, P) = 0$ by the above upper bound, where convergence is a.s. By definition of the MMD, this is equivalent to

$$\lim_{n\to\infty} \frac{1}{\ell_{m_n}} \sum_{i=1}^{\ell_{m_n}} k(\cdot, X_i) = \mu_P,$$

where convergence is in $\|\cdot\|_{\mathcal{H}}$ and is a.s. It then follows from the converse implication of Corollary 20.8 in (Bauer, 2011) applied to $\left\| \frac{1}{\ell_{m_n}} \sum_{i=1}^{\ell_{m_n}} k(\cdot, X_i) - \mu_P \right\|_{\mathcal{H}}$ that (5) holds, where convergence is in probability. Finally, if X is EWC a.s., we can repeat the above proof without going to a subsequence, showing the result in that case as well.

The converse implication follows immediately from Proposition 3.2. □

A.5. Proofs for Section 4.2

The general structure of the proof of Theorem 4.3 is identical to that of Theorem 2.17 in (Steinwart et al., 2009) thanks to Theorem C.8 and Theorem 3.6 which extend to our case Lemmas 4.2 and 4.4 in the same reference. Only the last step differs, since the randomness of the limit measure J makes the convergence of $\mathcal{R}_{L,J}(f_{J,\lambda})$ to $\mathcal{R}_{L,J,\mathcal{H}}$ as $\lambda \to 0$ also a random event, and that convergence may not occur uniformly on (a full probability subset of) Ω.

Proof of Theorem 4.3. We begin with the case where Z is EWC a.s. First, since L is locally bounded, the function $L(\cdot,\cdot,0)$ is bounded, and thus we can assume without loss of generality that $\mathcal{R}_{L,I}(0) \le 1$ for all $I \in \mathcal{P}(\mathcal{X}\times\mathcal{Y})$. By Lemma C.5, we thus have $\|f_{I,\lambda}\|_{\mathcal{H}} \le \lambda^{-1/2}$ for all $\lambda > 0$. Furthermore, since K is bounded, we can also assume without loss of generality that $\|K\|_\infty \le 1$ so that $\|f\|_\infty \le \|f\|_{\mathcal{H}}$ for all $f \in \mathcal{H}$. Finally, by completeness of $(\Omega, \mathcal{A})$ and measurability of $\mathcal{R}_{L,J,\mathcal{H}}$ (cf. Corollary D.4), the assumption that $\mathcal{H}$ is (L, J)-rich $\mathbb{P}$-a.s. implies that $\mathcal{R}^\star_{L,J}$ is measurable and is a.s. equal to $\mathcal{R}_{L,J,\mathcal{H}}$. We denote by Ω_1 a full probability set such that $\mathcal{R}_{L,J,\mathcal{H}} = \mathcal{R}^\star_{L,J}$.

We have for all $n \in \mathbb{N}^\star$ and $\lambda > 0$ the upper bound

$$\begin{aligned}
\left|\mathcal{R}_{L,J}(f_{\eta^Z_n,\lambda}) - \mathcal{R}^\star_{L,J}\right| &\le \left|\mathcal{R}_{L,J}(f_{\eta^Z_n,\lambda}) - \mathcal{R}_{L,J}(f_{J,\lambda})\right| + \left|\mathcal{R}_{L,J}(f_{J,\lambda}) - \mathcal{R}^\star_{L,J}\right| \\
&= \left|\mathcal{R}_{L,J}(f_{\eta^Z_n,\lambda}) - \mathcal{R}_{L,J}(f_{J,\lambda})\right| + \left|\mathcal{R}_{L,J}(f_{J,\lambda}) - \mathcal{R}_{L,J,\mathcal{H}}\right| \\
&\le |L|_{\lambda^{-1/2},1} \cdot \left\|f_{\eta^Z_n,\lambda} - f_{J,\lambda}\right\|_\infty + \left|\mathcal{R}_{L,J}(f_{J,\lambda}) - \mathcal{R}_{L,J,\mathcal{H}}\right| \\
&\le \frac{|L|_{\lambda^{-1/2},1}}{\lambda} \cdot \left\|\eta^Z_n(\phi h_{J,\lambda}) - J(\phi h_{J,\lambda})\right\|_{\mathcal{H}} + \left|\mathcal{R}_{L,J}(f_{J,\lambda}) - \mathcal{R}_{L,J,\mathcal{H}}\right|,
\end{aligned} \tag{18}$$

where the second step only holds on Ω_1, the third one comes from local Lipschitz continuity and the integral definition of $\mathcal{R}_{L,J}$, the last step comes from Theorem C.8, and $h_{J,\lambda}$ is as given in Corollary C.9. We can now pick the regularization sequence $(\lambda_n)_{n\in\mathbb{N}^\star}$ as follows. Let $\epsilon > 0$. First, note that by the expression of $h_{J,\lambda}$ in Corollary C.9 and by Lemma D.6, the map $\omega \mapsto \Phi h_{J(\omega),\lambda}$ is measurable, and thus so are its integrals w.r.t. random measures (by Lemma 1.15.(i) in (Kallenberg, 2017)). Next, for all $n \in \mathbb{N}^\star$, and $\lambda > 0$, let

$$F(\lambda, n) := \mathbb{P}\left[\sup_{m\ge n} \left\|\eta^Z_m(\phi h_{J,\lambda}) - J(\phi h_{J,\lambda})\right\|_{\mathcal{H}} \ge \frac{\lambda}{|L|_{\lambda^{-1/2},1}}\epsilon\right].$$

Now, $h_{J,\lambda}$ takes values in $C_{\mathrm{b}}(\mathcal{X}\times\mathcal{Y};\mathcal{G})$ by Corollary C.9, and thus $\phi h_{J,\lambda}$ takes values in $C_{\mathrm{b}}(\mathcal{X}\times\mathcal{Y};\mathcal{H})$, which is separable by assumption. Therefore, by Corollary 3.7, $\lim_{n\to\infty} \eta^Z_n(\phi h_{J,\lambda}) = J(\phi h_{J,\lambda})$, where convergence is a.s. Consequently, it follows from Lemma 20.6 in (Bauer, 2011) that $\lim_{n\to\infty} F(\lambda, n) = 0$ for all $\lambda > 0$. We now use Lemma 4.4 in (Steinwart et al., 2009) to obtain a sequence $(\lambda_n)_{n\in\mathbb{N}^\star}$ such that $\lim_{n\to\infty} \lambda_n = 0$ and $\lim_{n\to\infty} F(\lambda_n, n) = 0$. Let now $\delta > 0$ be arbitrary, and take $n_0 \in \mathbb{N}^\star$ such that for all $n \ge n_0$, both $F(n, \lambda_n) \le \frac{\delta}{2}$ and $G_n \le \frac{\delta}{2}$, where

$$G_n := \mathbb{P}\left[\sup_{m\ge n} \left|\mathcal{R}_{L,J}(f_{J,\lambda_m}) - \mathcal{R}_{L,J,\mathcal{H}}\right| \ge \epsilon\right].$$

Such an n_0 exists since $\lim_{n\to\infty} G_n = 0$ by applying again Lemma 20.6 in (Bauer, 2011) to $|\mathcal{R}_{L,J}(f_{J,\lambda_n}) - \mathcal{R}_{L,J,\mathcal{H}}|$, which converges to 0 a.s. by a simple argument using Lemma C.5. It follows from (18) that for all $n \geq n_0$,

$$\begin{aligned}\mathbb{P}\left[\sup_{m\geq n}\left|\mathcal{R}_{L,J}(f_{\eta_m^Z,\lambda_m}) - \mathcal{R}^\star_{L,J}\right| \geq \epsilon\right] &\leq \mathbb{P}\left[\sup_{m\geq n}\frac{|L|_{\lambda_m^{-1/2},1}}{\lambda_m}\cdot\left\|\eta_n^Z(\phi h_{J,\lambda_m}) - J(\phi h_{J,\lambda_m})\right\|_{\mathcal{H}} \geq \epsilon \text{ or}\right.\\ &\qquad\left.\sup_{m\geq n}|\mathcal{R}_{L,J}(f_{J,\lambda_m}) - \mathcal{R}_{L,J,\mathcal{H}}| \geq \epsilon\right]\\ &\leq F(\lambda_n, n) + G_n\\ &\leq \delta.\end{aligned}$$

The converse implication of Lemma 20.6 in (Bauer, 2011) completes the proof of this case. The case when Z is only EWC in probability follows immediately by repeating the above steps without introducing the supremum $\sup_{m\geq n}$ in the definition of $F(\lambda, n)$ and of G_n, which concludes the proof. □

B. Random compact sets

This section contains technical results necessary for the proof of Theorem 3.6. The main result is Lemma B.6, which is a generalization of Prokhorov's theorem to finite or countable families of random measures.

Definition B.1. Let $(\mathcal{X}, d)$ be a metric space. We say that $\mathcal{X}$ has *nice closed balls* if for any $x \in \mathcal{X}$ and $\epsilon > 0$, the closed ball $B(x, \epsilon) = \{y \in \mathcal{X} \mid d(x, y) \leq \epsilon\}$ is either compact of equal to $\mathcal{X}$.

Lemma B.2. *Let $\mathcal{X}$ be a metrizable space. Then, $\mathcal{X}$ has a compatible metric with nice closed balls if, and only if, $\mathcal{X}$ is locally compact.*

Proof. This follows immediately from Theorem 5.1.12 in (Beer, 1993). □

Lemma B.3. *Let $\mathcal{X}$ be a Polish space and let d be a metric that metrizes its topology. The set $\mathcal{K}(\mathcal{X})$ of nonempty compact subsets of $\mathcal{X}$ can be equipped with the Hausdorff metric*

$$h:\ (K_1, K_2) \in \mathcal{K}(\mathcal{X})^2 \mapsto \max\left\{\sup_{x\in K_1}\inf_{y\in K_2} d(x,y), \sup_{x\in K_2}\inf_{y\in K_1} d(x,y)\right\}.$$

Then, $(\mathcal{K}(\mathcal{X}), h)$ is complete and separable. In particular, $\mathcal{K}(\mathcal{X})$ equipped with the resulting topology is Polish.

Proof. Completeness is proven in Theorem 3-3 in (Henrikson, 1999), and separability follows from Corollary 3.90 and Theorem 3.91 in (Aliprantis & Border, 2006). □

In what follows, we assume without loss of generality that, if $\mathcal{X}$ is locally compact, then the Hausdorff distance h on $\mathcal{K}(\mathcal{X})$ is defined w.r.t. a compatible metric d that has nice closed balls.

Before moving on to the main result, we show two useful preliminary results.

Lemma B.4. *Let $(\mathcal{X}, d)$ be a metric space with nice closed balls. Then, for any $K \in \mathcal{K}(\mathcal{X})$ and $\epsilon > 0$, the set*

$$\mathrm{dil}(K, \epsilon) = \{x \in \mathcal{X} \mid \inf_{y\in K} d(x, y) \leq \epsilon\} = \{x \in \mathcal{X} \mid d(x, K) \leq \epsilon\}$$

is either compact or equal to $\mathcal{X}$.

Proof. Let $K \in \mathcal{K}(\mathcal{X})$ and $\epsilon > 0$. Let $(x_n)_{n\in\mathbb{N}^\star} \subset \mathrm{dil}(K, \epsilon)$. Take, for all $n \in \mathbb{N}^\star$, $z_n \in K$ such that $d(x_n, z_n) \leq \epsilon$; such a sequence exists by definition of $\mathrm{dil}(K, \epsilon)$. Since $(z_n)_{n\in\mathbb{N}^\star}$ is a sequence of the compact set K, it has a converging subsequence; say, $(z_{m_n})_{n\in\mathbb{N}^\star}$, and denote its limit by z. But then, for every $\delta > 0$, there exists $N \in \mathbb{N}^\star$ such that $x_{m_n} \in B(z, \epsilon + \delta)$ for all $n \geq N$. We conclude by discussing two cases. If $B(z, \epsilon + \delta) = \mathcal{X}$ for all $\delta > 0$, then in particular

$$B(z, \epsilon) = \bigcap_{n\in\mathbb{N}^\star} B(z, \epsilon + n^{-1}) = \bigcap_{n\in\mathbb{N}^\star} \mathcal{X} = \mathcal{X}.$$

Since $B(z,\epsilon) \subset \mathrm{dil}(K,\epsilon)$, we deduce that $\mathrm{dil}(K,\epsilon) = \mathcal{X}$. Otherwise, there exists $\delta > 0$ such that $B(z,\epsilon+\delta)$ is a proper subset of $\mathcal{X}$. Then, $(x_{m_n})_{n\in\mathbb{N}^\star}$ takes values in the compact set $B(z,\epsilon+\delta)$ from some finite rank $N \in \mathbb{N}^\star$ on. The sequence $(x_n)_{n\in\mathbb{N}^\star}$ thus has a subsequence converging in $B(z,\epsilon+\delta)$. Furthermore, since $\mathrm{dil}(K,\epsilon)$ is closed, we deduce that it must contain the limit of $(x_n)_{n\in\mathbb{N}^\star}$. In other words, either $\mathrm{dil}(K,\epsilon) = \mathcal{X}$, or each sequence thereof has a converging subsequence, showing compactness and concluding the proof. □

Lemma B.5. *Let $\mathcal{X}$ be a Polish space and $P \in \mathcal{P}(\mathcal{X})$. The map $e_P : K \in \mathcal{K}(\mathcal{X}) \mapsto P(K) \in [0,1]$ is upper semi-continuous, that is, for every $K \in \mathcal{K}(\mathcal{X})$ and sequence $(K_n)_{n\in\mathbb{N}^\star}$ such that $h(K_n,K) \to 0$ as $n \to \infty$, we have $\limsup_n e_P(K_n) \le e_P(K)$.*

Proof. First of all, our definition of upper semi-continuity is indeed equivalent to the general notion since $\mathcal{X}$ is Polish; see for instance Lemma 7.13.b in (Bertsekas & Shreve, 1996) for details. Let K and $(K_n)_{n\in\mathbb{N}^\star}$ be as described in the lemma. The result follows from the fact that $\limsup_n K_n \subset K$. Indeed, if $x \in \limsup_n K_n$, then there exists a strictly increasing sequence $j : \mathbb{N}^\star \to \mathbb{N}^\star$ such that $x \in K_{j_n}$ for all $n \in \mathbb{N}^\star$. But then, for all $n \in \mathbb{N}^\star$,

$$d(x,K) = \inf_{y\in K} d(x,y) \le \sup_{z\in K_{j_n}} \inf_{y\in K} d(z,y) \le h(K_{j_n},K).$$

Since the RHS goes to 0 as $n \to \infty$, we deduce that $d(x,K) = 0$, and thus $x \in K$ since K is closed, showing the inclusion. By the reverse Fatou's lemma, it follows that

$$\limsup_n e_P(K_n) \le e_P(\limsup_n K_n) \le e_P(K),$$

concluding the proof. □

Theorem B.6. *Let $\mathcal{X}$ be a locally compact Polish space and $\mathcal{Q} = (P_n)_{n\in\mathbb{N}^\star}$ be a finite or countable collection of random measures on $\mathcal{X}$. Assume that the set $E = \{\mathcal{Q}$ is sequentially compact$\}$ is measurable and has probability 1. Then, for all $\epsilon > 0$, there exists a random compact subset K_ϵ of $\mathcal{X}$ such that for all $P \in \mathcal{Q}$, $P(K_\epsilon) \ge 1-\epsilon$, a.s.*

Proof. Let L be an arbitrary element of $\mathcal{K}(\mathcal{X})$ and define

$$A : \omega \in \Omega \mapsto \begin{cases} \{K \in \mathcal{K}(\mathcal{X}) \mid \forall P \in \mathcal{Q},\, P_\omega(K) \ge 1-\epsilon\}, & \text{if } \omega \in E, \\ \{L\}, & \text{otherwise.} \end{cases}$$

The map A takes values in the set of subsets of $\mathcal{K}(\mathcal{X})$. We show that there exists a measurable selection of A by using the Kuratowksi-Ryll-Nardzewski theorem; see for instance Theorem 6.9.3 in (Bogachev, 2007). We need to show that A is weakly measurable and takes values in the set of nonempty closed subsets of $\mathcal{K}(\mathcal{X})$. First, by Prokhorov's theorem, $A(\omega) \neq \emptyset$. Furthermore, it is also closed. Indeed, let $(K_n)_{n\in\mathbb{N}^\star} \subset A(\omega)$, and assume that it converges to some $K \in \mathcal{K}(\mathcal{X})$. For any $P \in \mathcal{Q}$, we have $P_\omega(K) \ge \limsup_n P_\omega(K_n) \ge 1-\epsilon$, where the first inequality is from Lemma B.5 and the second one from the assumption on the sequence $(K_n)_{n\in\mathbb{N}^\star}$. Since $P \in \mathcal{Q}$ is arbitrary, this shows closedness of $A(\omega)$. Summarizing, A takes values in the set of nonempty closed subsets of $\mathcal{K}(\mathcal{X})$.

We are thus left to show that A is weakly measurable, that is, that for all U open in $\mathcal{K}(\mathcal{X})$, the set

$$\begin{aligned} \hat{A}(U) &= \{\omega \in \Omega \mid A(\omega) \cap U \neq \emptyset\} \\ &= \{\omega \in E \mid \exists K \in U, \forall P \in \mathcal{Q},\, P_\omega(K) \ge 1-\epsilon\} \cup \{\omega \in \Omega \setminus E \mid L \in U\}, \end{aligned}$$

belongs to $\mathcal{A}$. Since the second element of the RHS is either $\emptyset$ or $\Omega \setminus E$, which are both measurable, we focus on the first element, which we denote by $\bar{A}_P(U)$. We have

$$\bar{A}_P(U) = \{\omega \in E \mid \exists K \in U, \inf_{P\in\mathcal{Q}} P_\omega(K) \ge 1-\epsilon\}.$$

It is clear that it is sufficient to consider the case where U is an open ball of $\mathcal{K}(\mathcal{X})$, i.e.,

$$U = \{K \in \mathcal{K}(\mathcal{X}) \mid h(K_0,K) < \eta\},$$

where $K_0 \in \mathcal{K}(\mathcal{X})$ and $\eta > 0$ are arbitrary. Define for all $n \in \mathbb{N}^\star$

$$F_n = \operatorname{dil}\left(K_0, \eta(1-n^{-1})\right) = \{x \in \mathcal{X} \mid d(x, K_0) \leq \eta(1-n^{-1})\}.$$

where d is the metric on $\mathcal{X}$ used to define h. It is clear that $K_0 \subset F_n$ and that the sequence F is nondecreasing. We discuss two cases. First, if $F_n = \mathcal{X}$ for some $n \in \mathbb{N}^\star$, then $U \supset \{K \in \mathcal{K}(\mathcal{X}) \mid K_0 \subset K\}$; indeed, then any compact set K in that RHS satisfies $K \subset F_n$, and thus

$$\begin{aligned} h(K, K_0) &= \max\left\{\sup_{x\in K}\inf_{y\in K_0} d(x,y), \sup_{x\in K_0}\inf_{y\in K} d(x,y)\right\} \\ &\leq \max\left\{\sup_{x\in F_n}\inf_{y\in K_0} d(x,y), 0\right\} \\ &= h(F_n, K_0) \leq \eta(1-n^{-1}) < \eta, \end{aligned}$$

which shows $K \in U$. In that case, $\bar{A}_P(U) = E$ by Prokhorov's theorem and is thus measurable. Second, and more interestingly, assume that $F_n \neq \mathcal{X}$ for all $n \in \mathbb{N}^\star$. Then, F_n is compact by Lemma B.4, and thus $F_n \in U$ since

$$h(F_n, K_0) = \max\left\{\sup_{x\in F_n}\inf_{y\in K_0} d(x,y), 0\right\} = \sup_{x\in F_n} d(x, K_0) < \eta.$$

Furthermore, for any $K \in U$, it holds that $K \subset F_n$ for some $n \in \mathbb{N}^\star$; take for instance any $n \geq \left(1 - \frac{h(K,K_0)}{\eta}\right)^{-1}$. Consequently, we have the equality

$$\bar{A}_P(U) = \{\omega \in \Omega \mid \exists n \in \mathbb{N}^\star, \inf_{P\in\mathcal{Q}} P_\omega(F_n) \geq 1-\epsilon\}.$$

Indeed, the inclusion of the LHS in the RHS follows from the fact that any $K \in U$ such that $\inf_{P\in\mathcal{Q}} P_\omega(K) \geq 1-\epsilon$ must be in F_n for some $n \in \mathbb{N}^\star$ by what precedes, and the converse inclusion simply from the fact that $F_n \in U$. This shows that

$$\bar{A}_P(U) = \bigcup_{n\in\mathbb{N}^\star} \{\omega \in \Omega \mid \inf_{P\in\mathcal{Q}} P(F_n) \geq 1-\epsilon\}.$$

Since $\mathcal{Q}$ is finite or countable and the function $\omega \mapsto P_\omega(F_n)$ is measurable for any $n \in \mathbb{N}^\star$ and $P \in \mathcal{Q}$, so is $\omega \mapsto \inf_{P\in\mathcal{Q}} P_\omega(F_n)$. Therefore, $\bar{A}_P(U) \in \mathcal{A}$ as the preimage of $[1-\epsilon, 1]$ by a measurable map. Finally, Theorem 6.9.3 in (Bogachev, 2007) guarantees the existence of a measurable selection of A and concludes the proof. □

C. A General representer theorem with separable Hilbert output space

C.1. Preliminaries on Fréchet subdifferentials and convex optimization

Definition C.1 ((Kruger, 2009))**.** Let $\mathcal{B}$ be a Banach space, $F : \mathcal{B} \to \mathbb{R}$ a functional, and $x \in \mathcal{B}$. Let $\mathcal{B}'$ be the dual space of $\mathcal{B}$, and $\langle u', u\rangle$ be the evaluation of u' in u, for all $u' \in \mathcal{B}'$ and $u \in \mathcal{B}$. The Fréchet subdifferential of F in x is the set

$$\partial F(x) = \left\{x' \in \mathcal{B}' \middle| \liminf_{y\to x} \frac{F(y) - F(x) - \langle x', y-x\rangle}{\|y-x\|_{\mathcal{B}}} \geq 0\right\}.$$

The Fréchet subdifferential may be empty in general, but the following well-known result characterizes it for cases relevant for us.

Proposition C.2 ((Kruger, 2009))**.** *Under the same setting as Definition C.1:*

1. *If F is Fréchet differentiable in x, then $\partial F(x) = \{\nabla F(x)\}$;*
2. *If F is convex, then*
$$\partial F(x) = \{x' \in \mathcal{B}' \mid \forall y \in \mathcal{B},\ f(y) \geq f(x) + \langle x', y-x\rangle\}.$$

In other words the Fréchet subdifferential coincides with the convex subdifferential for convex functionals. Finally, the following result extends usual ones on minimization of convex functions defined on finite-dimensional spaces.

Theorem C.3. *Let $\mathcal{B}$ be a reflexive Banach space and $F : \mathcal{B} \to \mathbb{R}$ be a functional. Assume that F is coercive*[5] *and α-convex*[6] *with $\alpha > 0$. Then, there exists a unique $x \in \mathcal{B}$ such that*

$$F(x) = \inf_{y \in \mathcal{B}} F(y).$$

Furthermore, x is the unique element of $\mathcal{B}$ that satisfies $0 \in \partial F(x)$.

Proof. This follows immediately from Theorem A.6.9 in (Steinwart & Christmann, 2008). □

Finally, we emphasize that Hilbert spaces are reflexive Banach spaces, as noted in (Park & Muandet, 2022).

C.2. The representer theorem

The results and proofs of this section follow closely the lines of the corresponding results of Chapter 5 in (Steinwart & Christmann, 2008). Indeed, these results generalize readily to the case where $\mathcal{G}$ is separable Hilbert; the novelty and contribution here reside in making sure that all of the arguments of the reference still hold.

Lemma C.4 (Properties of the risk)**.** *Let $\mathcal{X}$ be a measurable space, $\mathcal{G}$ a separable Hilbert space, and $\mathcal{Y}$ a complete subset of $\mathcal{G}$. Let $J \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ and L be a continuous, J-integrable Nemitski loss. Let P be the marginal of J on $\mathcal{X}$. Then, the following statements hold:*

1. *for all uniformly bounded sequence $(f_n)_{n \in \mathbb{N}^\star} \subset \mathcal{L}_\infty(\mathcal{X}; \mathcal{G})$ that converges to some $f \in \mathcal{L}_\infty(\mathcal{X}; \mathcal{G})$ P-almost-everywhere (a.e.), we have*
$$\lim_{n \to \infty} \mathcal{R}_{L,J}(f_n) = \mathcal{R}_{L,J}(f);$$
2. *the map $\mathcal{R}_{L,J} : L_\infty(\mathcal{X}, P; \mathcal{G}) \to [0, \infty)$ is well-defined and continuous. In particular, it only takes finite values;*
3. *if, additionally, L is a Nemitski loss of order 2, then $\mathcal{R}_{L,J} : L_2(\mathcal{X}, P; \mathcal{G}) \to [0, \infty)$ is well-defined and continuous.*

Proof. The proof is an identical repetition of the proof of Lemma 2.17 in (Steinwart & Christmann, 2008) and is omitted for conciseness. □

Lemma C.5 (Existence and uniqueness of SVM solutions)**.** *Let $\mathcal{X}$ be a measurable space, $\mathcal{G}$ be a separable Hilbert space, $\mathcal{Y}$ a complete subset of $\mathcal{G}$, and $\mathcal{H}$ a $\mathcal{G}$-valued RKHS of measurable functions on $\mathcal{X}$ with bounded kernel K. Let L be a convex loss function and $J \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ such that L is a J-integrable Nemitski loss of order 2. For all $\lambda > 0$, there exists a unique $f_{J,\lambda} \in \mathcal{H}$ such that*

$$\mathcal{R}_{L,J}(f_{J,\lambda}) + \lambda \cdot \|f_{J,\lambda}\|_{\mathcal{H}}^2 = \inf_{f \in \mathcal{H}} \mathcal{R}_{L,J}(f) + \lambda \cdot \|f\|_{\mathcal{H}}^2 .$$

Furthermore, $\|f_{J,\lambda}\|_{\mathcal{H}} \leq \sqrt{\frac{\mathcal{R}_{L,J}(0)}{\lambda}}$.

Proof. By Lemma C.4, $\mathcal{R}_{L,J} : L_2(\mathcal{X}, P; \mathcal{G}) \to [0, \infty)$ is a well-defined functional. Since K is bounded, the inclusion operator $\iota_P : \mathcal{H} \to L_2(\mathcal{X}, P; \mathcal{G})$ is bounded, and thus $R_{\mathcal{H}} = \mathcal{R}_{L,J} \circ \iota_P$ is also a well-defined functional defined on $\mathcal{H}$. Furthermore, convexity of L implies that $R_{\mathcal{H}}$ is also convex, and thus $R_{\mathcal{H}} + \lambda \cdot \|\cdot\|_{\mathcal{H}}$ is λ-convex and coercive. The existence and uniqueness of $f_{J,\lambda}$ then follows from Theorem C.3. The second claim follows immediately from the first one since

$$\mathcal{R}_{L,J}(f_{J,\lambda}) + \lambda \cdot \|f_{J,\lambda}\|_{\mathcal{H}}^2 \leq \mathcal{R}_{L,J}(0)$$

and $\mathcal{R}_{L,J}(f_{J,\lambda}) \geq 0$. □

Lemma C.6. *Under the same setting as Lemma C.5, let R be the risk functional $\mathcal{R}_{L,J}$ defined on $L_2 := L_2(\mathcal{X} \times \mathcal{Y}, J; \mathcal{G})$, that is, for $f \in L_2$,*

$$R(f) = \int_{\mathcal{X} \times \mathcal{Y}} L(x, y, f(x, y)) \mathrm{d}J(x, y).$$

[5] Recall that F is coercive if $F(x) \to \infty$ as $\|x\|_{\mathcal{B}} \to \infty$

[6] Recall that F is α-convex if the functional $F - \alpha \cdot \|\cdot\|_{\mathcal{B}}^2$ is convex.

Then, R is finite everywhere on L_2, is convex, and for all $f \in L_2$,

$$\partial R(f) = \{h \in L_2 \mid h(x,y) \in \partial L(x,y,f(x,y)) \textit{for } J\textit{-almost-every } (x,y) \in \mathcal{X} \times \mathcal{Y}\}. \tag{19}$$

Proof. Everywhere-finiteness and convexity of R follow immediately from the assumptions on L. We show the expression of the subdifferential, and denote by D the RHS of (19). First note that L_2 is equal to its dual, and thus $\partial R(f)$ is indeed a subset of it. Next, let $h \in D$, and let $g \in L_2$ be arbitrary. By definition of D, it holds J-a.e. that

$$L(x,y,g(x,y)) \geq L(x,y,f(x,y)) + \langle h(x,y),\, g(x,y) - f(x,y)\rangle_{\mathcal{G}}\,.$$

Integrating w.r.t. J yields

$$\begin{aligned} R(g) = \int_{\mathcal{X}\times\mathcal{Y}} L(x,y,g(x,y))\mathrm{d}J(x,y) &\geq \int_{\mathcal{X}\times\mathcal{Y}} L(x,y,f(x,y))\mathrm{d}J(x,y) + \int_{\mathcal{X}\times\mathcal{Y}} \langle h(x,y),\, (g-f)(x,y)\rangle_{\mathcal{G}}\,\mathrm{d}J(x,y) \\ &= R(f) + \langle h,\, g-f\rangle_{L_2}\,, \end{aligned}$$

where the equality between the second terms of the sums in the last step comes from Corollary 1.3.13 in (Hytönen et al., 2016) since $\mathcal{G}$ is Hilbert and thus has the Radon-Nikodym property. Since g is arbitrary, this shows that $h \in \partial R(f)$ and the inclusion $D \subset \partial R(f)$. Conversely, let $h \in \partial R(f)$. We prove $h \in D$ by constructing a set based on evaluations of h and showing that it has measure 0. This requires being precise and distinguishing functions from their equivalence classes for the relation of J-a.e. equality: in what follows, if $u \in L_2$, we denote by $\bar{u}$ one of its representations (and conversely). For all $(x,y) \in \mathcal{X} \times \mathcal{Y}$, define the sets $A(x,y)$ and B as follows:

$$\begin{aligned} A(x,y) &= \{t \in \mathcal{G} \mid L(x,y,t) < L(x,y,\bar{f}(x,y)) + \left\langle \bar{h}(x,y), t - \bar{f}(x,y)\right\rangle_{\mathcal{G}}\} \\ B &= \{(x,y) \in \mathcal{X}\times\mathcal{Y} \mid A(x,y) \neq \emptyset\}. \end{aligned}$$

It is clear from the definition of $A(x,y)$ that $B \in \mathcal{B}(X \times \mathcal{Y})$. Assume $J[B] > 0$, and take any measurable function $\bar{g}$ such that, for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$, $\bar{g}(x,y) \in A(x,y)$ if $(x,y) \in B$, and $\bar{g}(x,y) = \bar{f}(x,y)$ otherwise. We have

$$\begin{aligned} R(g) &= \int_{\mathcal{X}\times\mathcal{Y}} L(x,y,\bar{g}(x,y))\mathrm{d}J(x,y) \\ &= \int_{B^{\complement}} L(x,y,\bar{f}(x,y))\mathrm{d}J(x,y) + \int_B L(x,y,\bar{g}(x,y))\mathrm{d}J(x,y) \\ &< \int_{B^{\complement}} L(x,y,\bar{f}(x,y))\mathrm{d}J(x,y) + \int_B L(x,y,\bar{f}(x,y))\mathrm{d}J(x,y) + \int_B \left\langle \bar{h}(x,y), \bar{g}(x,y) - \bar{f}(x,y)\right\rangle \mathrm{d}J(x,y) \\ &= \int_{\mathcal{X}\times\mathcal{Y}} L(x,y,\bar{f}(x,y))\mathrm{d}J(x,y) + \int_{\mathcal{X}\times\mathcal{Y}} \left\langle \bar{h}(x,y), \bar{g}(x,y) - \bar{f}(x,y)\right\rangle \mathrm{d}J(x,y) \\ &= R(f) + \langle h, g-f\rangle_{L_2}\,, \end{aligned}$$

where we used successively the definition of R, of g, the property that the strict inequality in the definition of $A(x,y)$ is preserved by integrating over the set B since it has positive measure, the fact that $\bar{g}(x,y) - \bar{f}(x,y) = 0$ for $(x,y) \notin B$, and properties of the dual pairing $\langle \cdot,\cdot\rangle_{L_2}$. The above resulting inequality contradicts the assumption that $h \in \partial R(f)$. We deduce that, for any choice of representations, the corresponding set B has measure 0. This shows $h \in D$ by contradiction, concluding the proof. □

Theorem C.7 (General representer theorem with separable Hilbert output space)**.** *Let $\mathcal{X}$ be a measurable space, $\mathcal{G}$ a separable Hilbert space, $\mathcal{Y}$ a complete subset of $\mathcal{G}$, and $\mathcal{H}$ a separable $\mathcal{G}$-valued RKHS of measurable functions on $\mathcal{X}$ with bounded kernel K and canonical feature map $\Phi : x \in \mathcal{X} \to K(\cdot,x) \in \mathcal{L}(\mathcal{G};\mathcal{H})$. Let L be a convex loss function and $J \in \mathcal{P}(\mathcal{X}\times\mathcal{Y})$ such that L is a J-integrable Nemitski loss of order 2. Then, for all $\lambda > 0$, there exists $h_{J,\lambda} \in \mathcal{L}_2(\mathcal{X}\times\mathcal{Y}, J; \mathcal{G})$ such that, for all $(x,y) \in \mathcal{X}\times\mathcal{Y}$,*

$$\begin{aligned} h_{J,\lambda}(x,y) &\in \partial L(x,y,f_{J,\lambda}(x)), \\ f_{J,\lambda} &= -\frac{1}{2\lambda} J(\Phi h_{J,\lambda}), \end{aligned}$$

where the notation $\Phi h_{J,\lambda}$ stands for the function $(x,y) \in \mathcal{X} \times \mathcal{Y} \mapsto \Phi(x) h_{J,\lambda}(x,y) \in \mathcal{H}$. If, additionally, L is locally Lipschitz continuous, then any such $h_{J,\lambda}$ satisfies

$$\|h_{J,\lambda}\|_\infty \leq |L|_{B_\lambda,1},$$

where $B_\lambda = \sqrt{\frac{\mathcal{R}_{J,\lambda}(0)}{\lambda}}$.

Proof. We consider R_{L_2} the risk functional defined on the Bochner space $L_2 := L_2(\mathcal{X} \times \mathcal{Y}, J; \mathcal{G})$ as defined in Lemma C.6. By Lemma C.6, R_{L_2} is finite and convex with subdifferential in $f \in L_2$ given by

$$\partial R_{L_2}(f) = \{h \in L_2 \mid h(x,y) \in \partial L(x,y,f(x)) \text{ for } J\text{-almost-every } (x,y) \in \mathcal{X} \times \mathcal{Y}\}.$$

By boundedness of Φ, the inclusion operator $\iota : \mathcal{H} \to L_2$ defined as $(\iota f)(x,y) = f(x)$ for all $f \in \mathcal{H}$ and $(x,y) \in \mathcal{X} \times \mathcal{Y}$ is bounded and for all $h \in L_2$,

$$\langle h, \iota f \rangle_{L_2} = \int_{\mathcal{X}\times\mathcal{Y}} \langle h(x,y), f(x) \rangle_{\mathcal{G}} \, \mathrm{d}J(x,y) = \left\langle \int_{\mathcal{X}\times\mathcal{Y}} K(\cdot,x) h(x,y) \mathrm{d}J(x,y), \; f \right\rangle_{\mathcal{H}},$$

and thus the adjoint $\iota^\star$ of ι is $\iota^\star : h \in L_2 \mapsto J(\Phi h) \in \mathcal{H}$. Furthermore, the restriction $R_{\mathcal{H}}$ of $\mathcal{R}_{L,J}$ to $\mathcal{H}$ satisfies $R_{\mathcal{H}} = R_{L_2} \circ \iota$, and thus the chain rule for Fréchet subdifferentials yields $\partial R_{\mathcal{H}}(f) = \partial(R_{L_2} \circ \iota)(f) = \iota^\star \partial R_{L_2}(\iota f)$, for all $f \in \mathcal{H}$. In addition, $f \mapsto \|f\|_{\mathcal{H}}^2$ is Fréchet differentiable with derivative $2f$ for all $f \in \mathcal{H}$. Therefore, the subdifferential of the regularized risk $R_{\mathcal{H},\lambda} = R_{\mathcal{H}} + \lambda \cdot \|\cdot\|_{\mathcal{H}}$ at $f \in \mathcal{H}$ is

$$\begin{aligned}\partial R_{\mathcal{H},\lambda}(f) &= 2\lambda f + \partial R_{\mathcal{H}}(f) \\ &= 2\lambda f + \{J(\Phi h) \mid h \in \mathcal{L}_2(\mathcal{X} \times \mathcal{Y}, J; \mathcal{G}) \wedge \forall (x,y) \in \mathcal{X} \times \mathcal{Y},\, h(x,y) \in \partial L(x,y,f(x))\}\end{aligned}$$

Recall now that $R_{\mathcal{H},\lambda}$ has a minimum at $f_{J,\lambda}$ by Lemma C.5; thus, $0 \in \partial R_{\mathcal{H},\lambda}(f_{J,\lambda})$, establishing the existence of a function $h \in \mathcal{L}_2(\mathcal{X} \times \mathcal{Y}, J; \mathcal{G})$ satisfying $h(x,y) \in \partial L(x,y,f_{J,\lambda}(x))$ for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$ and $f_{J,\lambda} = -\frac{1}{2\lambda} J(\Phi h)$.

We now assume that L is locally Lipschitz continuous and show the claimed bound on $\|h\|_\infty$. For all $(x,y) \in \mathcal{X} \times \mathcal{Y}$,

$$\begin{aligned}\|h(x,y)\|_{\mathcal{G}} &\leq \sup\{\|v\|_{\mathcal{G}} \mid v \in \partial L(x,y,f_{J,\lambda}(x))\} \\ &\leq \sup\left\{ \|v\|_{\mathcal{G}} \,\middle|\, v \in \bigcup_{\substack{t \in \mathcal{G} \\ \|t\|_{\mathcal{G}} \leq B_\lambda}} \partial L(x,y,t) \right\},\end{aligned}$$

where the second inequality comes from $\|f_{J,\lambda}\|_\infty \leq \|f_{J,\lambda}\|_{\mathcal{H}} \leq B_\lambda$, by Lemma C.5. By Proposition A.6.11 in (Steinwart & Christmann, 2008), it holds that for every $r > 0$ and $(x,y,t) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{G}$ such that $\|t\|_{\mathcal{G}} \leq r$, every $v \in \partial L(x,y,t)$ satisfies $\|v\|_{\mathcal{G}} \leq |L|_{r,1}$. This shows $\|h(x,y)\|_{\mathcal{G}} \leq |L|_{B_\lambda,1}$ for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$, concluding the proof. □

C.3. Stability of solutions

Theorem C.8 (Stability)**.** *Let $\mathcal{X}$ be a measurable space, $\mathcal{G}$ a separable Hilbert space, $\mathcal{Y}$ a complete subset of $\mathcal{G}$, and $\mathcal{H}$ a separable $\mathcal{G}$-valued RKHS of measurable functions on $\mathcal{X}$ with bounded kernel K and canonical feature map $\Phi : x \in \mathcal{X} \to K(\cdot,x) \in \mathcal{L}(\mathcal{G};\mathcal{H})$. Let L be a convex loss function and $J \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ such that L is a J-integrable Nemitski loss of order 2. For all $\lambda > 0$, there exists $h_{J,\lambda} \in \mathcal{L}_2(\mathcal{X} \times \mathcal{Y}, J; \mathcal{G})$ such that for all $I \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ for which L is a J-integrable Nemitski loss of order 2 and all $(x,y) \in \mathcal{X} \times \mathcal{Y}$,*

$$h_{J,\lambda}(x,y) \in \partial L(x,y,f_{J,\lambda}(x)), \tag{20}$$

$$f_{J,\lambda} = -\frac{1}{2\lambda} J(h \cdot \Phi), \tag{21}$$

$$\|f_{J,\lambda} - f_{I,\lambda}\|_{\mathcal{H}} \leq \frac{1}{\lambda} \|J(\Phi h_{J,\lambda}) - I(\Phi h_{J,\lambda})\|_{\mathcal{H}}. \tag{22}$$

If, additionally, L is locally Lipschitz continuous, then any such $h_{J,\lambda}$ satisfies

$$\|h_{J,\lambda}\|_\infty \leq |L|_{B_\lambda,1}, \tag{23}$$

where $B_\lambda = \sqrt{\frac{\mathcal{R}_{J,\lambda}(0)}{\lambda}}$.

Proof. Let $\lambda > 0$. By Theorem C.7, there exists a function $h \in \mathcal{L}_2(\mathcal{X} \times \mathcal{Y}, J; \mathcal{G})$ that satisfies (20) and (21), and (23) as well if L is assumed to be locally Lipschitz continuous. We thus only need to show that (22) holds. For this, notice that for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$,

$$L(x, y, f_{I,\lambda}(x)) \geq L(x, y, f_{J,\lambda}(x)) + \langle h(x, y), f_{I,\lambda}(x) - f_{J,\lambda}(x)\rangle_{\mathcal{G}},$$

by (20). Integrating against I and using the reproducing property thus yields

$$\mathcal{R}_{L,I}(f_{I,\lambda}) \geq \mathcal{R}_{L,I}(f_{J,\lambda}) + \langle I(\Phi h), f_{I,\lambda} - f_{J,\lambda}\rangle_{\mathcal{H}}.$$

Furthermore,

$$2\lambda \langle f_{I,\lambda} - f_{J,\lambda},\, f_{J,\lambda}\rangle + \lambda \|f_{I,\lambda} - f_{J,\lambda}\|_{\mathcal{H}}^2 = \lambda \|f_{I,\lambda}\|_{\mathcal{H}}^2 - \lambda \|f_{J,\lambda}\|_{\mathcal{H}}^2,$$

by manipulation of properties of $\|\cdot\|_{\mathcal{H}}^2$. Combining the last two equations shows

$$\langle f_{I,\lambda} - f_{J,\lambda},\, I(\Phi h) + 2\lambda f_{J,\lambda}\rangle_{\mathcal{H}} + \lambda \|f_{I,\lambda} - f_{J,\lambda}\|_{\mathcal{H}}^2 \leq \mathcal{R}_{L,I}(f_{I,\lambda}) + \lambda \|f_{I,\lambda}\|_{\mathcal{H}}^2 - (\mathcal{R}_{L,I}(f_{J,\lambda}) + \lambda \|f_{J,\lambda}\|_{\mathcal{H}}^2) \quad \leq 0,$$

where the second inequality comes from Lemma C.5 applied to I. Consequently, the representation (21) and the Cauchy-Schwarz inequality yield

$$\begin{aligned}\lambda \|f_{J,\lambda} - f_{I,\lambda}\|_{\mathcal{H}}^2 &\leq \langle f_{J,\lambda} - f_{I,\lambda},\, I(\Phi h) - J(\Phi h)\rangle \\ &\leq \|f_{J,\lambda} - f_{I,\lambda}\|_{\mathcal{H}} \cdot \|I(\Phi h) - J(\Phi h)\|_{\mathcal{H}},\end{aligned}$$

which immediately implies (22) and concludes the proof. □

Corollary C.9. *With the same notations and assumptions as in Theorem C.8, assume additionally that $\mathcal{X}$ is a topological space, $\mathcal{H}$ consists of continuous functions, and that L is continuously differentiable. Then, the function $h_{J,\lambda}$ is unique and in $C_{\mathrm{b}}(\mathcal{X} \times \mathcal{Y}; \mathcal{G})$. Specifically, $h_{J,\lambda} : (x, y) \mapsto \nabla L(x, y, f_{J,\lambda}(x))$.*

Proof. Since $\mathcal{H}$ consists of continuous functions, we have in particular that $f_{J,\lambda} \in \mathcal{H}$ is continuous. The result follows immediately from (20). □

D. Measurability of risks with random measures and of SVMs with infinite-dimensional outputs

The results of this section generalize Lemmas 6.3 (measurability of risks) and 6.23 (measurability of SVMs) from (Steinwart & Christmann, 2008) to our setting. The generalization of the first result consists of Lemma D.3 and of Corollary D.4. The challenge compared to the reference is that the risk is now defined w.r.t. a random measure instead of a fixed one. The generalization of the second result is in Lemma D.6. It extends the reference to SVMs with separable Hilbert output spaces; our proof follows closely the lines of the original one. Finally, we conclude on measurability of the sets involved in Definition 4.2 with Corollary D.7.

We first recall the definition of a measurable learning method.

Definition D.1. Let $(\mathcal{X}, \mathcal{A}_{\mathcal{X}})$ be a nonempty measurable space, $\mathcal{G}$ be a separable Hilbert space, and $\mathcal{Y} \subset \mathcal{G}$ a nonempty, complete subset. We say that a learning method $\mathfrak{L}$ on $\mathcal{X} \times \mathcal{Y}$ is *measurable* if, for all $n \in \mathbb{N}^\star$, the map

$$\begin{aligned}(\mathcal{X} \times \mathcal{Y})^n \times \mathcal{X} &\to \mathcal{G} \\ (Z, x) &\mapsto f_Z(x)\end{aligned}$$

is measurable with respect to the universal completion of the product σ-algebra on $(\mathcal{X} \times \mathcal{Y}) \times \mathcal{X}$, where f_Z is the decision function produced by $\mathfrak{L}$ with data set Z.

We refer to Lemma A.3.3 in (Steinwart & Christmann, 2008) for a definition of the universal completion of a σ-algebra. Importantly, we chose in Section 2.1 to equip any topological space with its Borel σ-algebra. The following lemma ensures that this is not problematic for Polish spaces, as product σ-algebras coincide with Borel ones.

Lemma D.2. *Let $\mathcal{X}$ be a Polish space, $\mathcal{G}$ be a separable Hilbert space, and $\mathcal{Y} \subset \mathcal{G}$ a nonempty, complete subset. Then, for all $n \in \mathbb{N}^\star$, the product σ-algebra on $(\mathcal{X} \times \mathcal{Y})^n \times \mathcal{X}$ is equal to $\mathcal{B}\left((\mathcal{X} \times \mathcal{Y})^n \times \mathcal{X}\right)$. Further, their universal completions also coincide.*

Proof. From Theorem 14.8 in (Klenke, 2013), we have $\mathcal{B}\left((\mathcal{X} \times \mathcal{Y})^n \times \mathcal{X}\right) = \left(\mathcal{B}(\mathcal{X}) \otimes \mathcal{B}(\mathcal{Y})\right)^{\otimes n} \otimes \mathcal{B}(\mathcal{X})$ since $\mathcal{X}$ and $\mathcal{Y}$ are Polish, where $\otimes$ denotes the product σ-algebra. This shows the first part of the result. The second part follows immediately by definition of the universal completion. □

The next lemma ensures measurability of events involving the risk w.r.t. a random measure given a measurable learning method.

Lemma D.3. *Let $\mathcal{X}$ be a Polish space, $\mathcal{G}$ be a separable Hilbert space, and $\mathcal{Y} \subset \mathcal{G}$ a nonempty, complete subset. Let $\mathfrak{L}$ be a measurable learning method and $f_\cdot$ its decision function. Let Z be an $\mathcal{X} \times \mathcal{Y}$-valued process and J a random measure on $\mathcal{X} \times \mathcal{Y}$. Then, for any loss function $L : \mathcal{X} \times \mathcal{Y} \times \mathcal{G} \to [0, \infty)$ and any $n \in \mathbb{N}^\star$, the map $\omega \in \Omega \mapsto \mathcal{R}_{L,J(\omega)}(f_{Z_{1:n}(\omega)})$ is measurable.*

Proof. By measurability of $\mathfrak{L}$ and of L, we obtain measurability of the map $g : (\omega, x, y) \in \Omega \times \mathcal{X} \times \mathcal{Y} \mapsto L(x, y, f_{Z_{1:n}(\omega)}(x))$ by composition. It follows from item (i) in Lemma 1.15 in (Kallenberg, 2017) that $\omega \in \Omega \mapsto J(\omega)g(\omega, \cdot, \cdot) =: \mathcal{R}_{L,J(\omega)}(f_{Z_{1:n}(\omega)})$ is also measurable, concluding the proof. □

Corollary D.4. *Let $\mathcal{X}$ be a Polish space, $\mathcal{G}$ a separable Hilbert space, $\mathcal{Y} \subset \mathcal{G}$ a nonempty, complete subset, and $\mathcal{H}$ a separable $\mathcal{G}$-valued RKHS of measurable functions on $\mathcal{X}$. Let J be a random measure on $\mathcal{X} \times \mathcal{Y}$. Then, the map $\omega \mapsto \mathcal{R}_{\mathcal{H},L,J(\omega)}$ is measurable.*

Proof. Let $(f_n)_{n\in\mathbb{N}^\star}$ be a dense family of $\mathcal{H}$. We have $\mathcal{R}_{\mathcal{H},L,J(\omega)} = \inf_{n\in\mathbb{N}^\star} \mathcal{R}_{L,J}(f_n)$. Since $\mathcal{H}$ consists of measurable functions, the trivial learning method $\mathfrak{L}_n$ that returns f_n no matter the incoming data set is measurable for all $n \in \mathbb{N}^\star$. Consequently, the map $\omega \mapsto \mathcal{R}_{L,J(\omega)}(f_n)$ is measurable for all $n \in \mathbb{N}^\star$. The result follows since a countable infimum of measurable functions is measurable. □

We are thus left to show that SVMs are measurable. For this, we adapt the proof of Lemma 6.23 in (Steinwart & Christmann, 2008) to separable output spaces. We begin with the following technical lemma, whose proof follows the lines of that of Lemma 2.11 in the same reference.

Lemma D.5. *Let $\mathcal{X}$ be a Polish space, $\mathcal{G}$ be a separable Hilbert space, and $\mathcal{Y} \subset \mathcal{G}$ a nonempty, complete subset. Let L be a loss function on $\mathcal{X} \times \mathcal{Y} \times \mathcal{G}$ and $\mathcal{F} \subset \mathcal{L}_0(\mathcal{X}; \mathcal{G})$ a subset equipped with a complete and separable metric d. Assume that d dominates the pointwise convergence, that is,*

$$\lim_{n\to\infty} d(f_n, f) \implies \forall x \in \mathcal{X}, \ \lim_{n\to\infty} f_n(x) = f(x),$$

for all $f \in \mathcal{F}$ and $(f_n)_{n\in\mathbb{N}^\star} \subset \mathcal{F}$. Then, the evaluation map $(f, x) \in \mathcal{F} \times \mathcal{X} \mapsto f(x) \in \mathcal{G}$ is measurable. In particular, the map $(x, y, f) \mapsto L(x, y, f(x))$ is measurable.

Proof. Since d dominates pointwise convergence, for any fixed $x \in \mathcal{X}$ the $\mathcal{G}$-valued map $f \mapsto f(x)$ defined on $\mathcal{F}$ is continuous. Furthermore, since $\mathcal{F}$ consists of measurable functions, the map $x \mapsto f(x)$ is measurable for any fixed $f \in \mathcal{F}$. The first assertion then follows from Lemma III.14 in (Castaing & Valadier, 1977). The second assertion follows immediately by applying the first one to the Polish space $\mathcal{X} \times \mathcal{Y}$. □

Lemma D.6. *Let $\mathcal{X}$ be a Polish space, $\mathcal{G}$ a separable Hilbert space, $\mathcal{Y} \subset \mathcal{G}$ a nonempty, complete subset, and $\mathcal{H}$ a separable $\mathcal{G}$-valued RKHS of measurable functions on $\mathcal{X}$ with bounded kernel K. Let L be a convex loss function on $\mathcal{X} \times \mathcal{Y} \times \mathcal{G}$. For all $\lambda > 0$, the corresponding SVM that produces the decision function $f_{Z,\lambda}$ for $Z \in (\mathcal{X} \times \mathcal{Y})^n$ for some $n \in \mathbb{N}^\star$ is a measurable learning method.*

Proof. From the assumptions, $\mathcal{H}$ is a separable metric space of measurable functions from $\mathcal{X}$ to $\mathcal{G}$. Further, its metric dominates pointwise convergence. Indeed, for any sequences $(f_n)_{n\in\mathbb{N}^\star} \in \mathcal{H}$ and $(x_n)_{n\in\mathbb{N}^\star} \subset \mathcal{X}$ converging to $f \in \mathcal{H}$ and $x \in \mathcal{X}$, respectively, we have

$$\|f_n(x) - f(x)\|_{\mathcal{G}} = \langle f_n - f, K(\cdot, x)(f_n - f)\rangle_{\mathcal{H}} \leq \|K(\cdot, x)\|_{\mathcal{L}(\mathcal{H};\mathcal{G})} \|f_n - f\|_{\mathcal{H}},$$

and the RHS goes to 0 by assumption. By Lemma D.5, the map $(x, y, f) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{H} \mapsto L(x, y, f(x))$ is measurable. We deduce that the map

$$\begin{aligned}\varphi : (\mathcal{X} \times \mathcal{Y})^n \times \mathcal{H} &\to [0, \infty) \\ (Z, f) &\mapsto \mathcal{R}_{L,\eta_n^Z}(f) + \lambda \|f\|_{\mathcal{H}}^2\end{aligned}$$

is measurable. Next, we apply (iii) in Lemma A.3.18 in (Steinwart & Christmann, 2008) with $F(Z) := \mathcal{H}$ for $Z \in (\mathcal{X} \times \mathcal{Y})^n$ combined with Lemma C.5 to show that the map $Z \mapsto f_{Z,\lambda}$ is measurable with respect to the universal completion of the product σ-algebra of $(\mathcal{X} \times \mathcal{Y})^n$. Consequently, the map $(\mathcal{X} \times \mathcal{Y})^n \times \mathcal{X} \to \mathcal{H} \times \mathcal{X}$ defined by $(Z, x) \mapsto (f_{Z,\lambda}, x)$ is measurable. Finally, we deduce from the first conclusion of Lemma D.5 that the map $(\mathcal{X} \times \mathcal{Y})^n \times \mathcal{X} \to \mathcal{G}$ defined by $(Z, x) \mapsto f_{Z,\lambda}(x)$ is measurable, concluding the proof. □

Corollary D.7. *Let $\mathcal{X}$ be a Polish space, $\mathcal{G}$ a separable Hilbert space, $\mathcal{Y} \subset \mathcal{G}$ a nonempty, complete subset, and $\mathcal{H}$ a separable $\mathcal{G}$-valued RKHS of measurable functions on $\mathcal{X}$ with bounded kernel K. Let L be a convex loss function on $\mathcal{X} \times \mathcal{Y} \times \mathcal{G}$, Z be an $\mathcal{X} \times \mathcal{Y}$-valued process, and J be a random measure on $\mathcal{X} \times \mathcal{Y}$. For all $\lambda > 0$, $\epsilon > 0$, and $n \in \mathbb{N}^\star$ the set $\{\omega \in \Omega \mid \mathcal{R}_{L,J(\omega)}(f_{Z_{1:n}(\omega),\lambda}) \leq \mathcal{R}_{\mathcal{H},L,J(\omega)} + \epsilon\}$ is measurable.*

Proof. By Lemmas D.6 and D.3, the map $\omega \in \Omega \mapsto \mathcal{R}_{L,J(\omega)}(f_{Z_{1:n}(\omega),\lambda})$ is measurable. By Corollary D.4, the map $\omega \in \Omega \mapsto \mathcal{R}_{\mathcal{H},L,J(\omega)}$ is also measurable. The result follows immediately. □