Plug-in Estimators for Conditional Expectations and Probabilities

Steffen Grünewälder
Department of Mathematics and Statistics, Lancaster University.

Abstract

We study plug-in estimators of conditional expectations and probabilities, and we provide a systematic analysis of their rates of convergence. The plug-in approach is particularly useful in this setting since it introduces a natural link to VC- and empirical process theory. We make use of this link to derive rates of convergence that hold uniformly over large classes of functions and sets, and under various conditions. For instance, we demonstrate that elementary conditional probabilities are estimated by these plug-in estimators with a rate of $n^{\alpha-1/2}$ if one conditions with a VC-class of sets and where $\alpha \in [0, 1/2)$ controls a lower bound on the size of sets we can estimate given n samples. We gain similar results for Kolmogorov's conditional expectation and probability which generalize the elementary forms of conditioning. Due to their simplicity, plug-in estimators can be evaluated in linear time and there is no up-front cost for inference.

1 Introduction

Conditional distributions and conditional expectations are of importance in all areas of machine learning and statistics. We consider a simple and natural approach to estimate conditional distributions based on the plug-in principle. This allows us to leverage results from VC- and empirical process theory in order to control estimation errors uniformly over large families of sets and functions. Before giving a more detailed account of the estimators and their properties we go through some concrete examples in which conditional distributions play a major role. Many applications of conditional distributions make use of the Markov property, and essentially any method that uses this property relies on estimates of conditional distributions. For instance, hidden Markov models use the Baum-Welch algorithm to infer conditional distributions and, more generally, graphical models rely heavily on conditioning (Baum and Petrie, 1966; Rabiner, 1989; Bishop, 2006). Similarly, many reinforcement learning algorithms are developed for Markov decision processes (MDPs) and balance between optimization and the estimation of conditional pay-offs (Sutton and Barto, 1998; Szepesvári, 2010). Recently, causal inference has become increasingly popular, which is yet another branch of machine learning that relies heavily on conditioning (Schölkopf, Janzing, Peters, Sgouritsa, Zhang, and Mooij, 2012).

In the case that conditioning happens only with respect to finitely many events, elementary tools suffice to prove consistency of estimators and to derive rates of convergence. However, in practice, the number of events we condition on is typically very large or even infinite. For example, the number of states in an MDP is usually exceptionally large and approximation techniques are used in practice. Neural networks have proven useful in this context (Silver et al., 2016). Providing guarantees on estimates of conditional expectations is, in this case, a non-trivial challenge and falls within the area of non-parametric statistics. Classical tools in non-parametric statistics to address such challenges are VC- and empirical process theory. Before showing how these tools can be applied in our context we first provide an overview of the forms of conditioning we can deal with.

In machine learning we encounter conditional distributions in various forms. For instance, graphical models make use of conditional probabilities $P(A|B)$ of an event A given a second event B. Closely related to this is the conditional probability of a random variable Y attaining a value y and given that a second random variable X equals x when both X and Y can attain only finitely many values. Formally, we are working with $P(Y = y|X = x)$ and y and x are elements of the finite range spaces of Y and X respectively, which we will denote in the following by $\mathbb{Y}$ and $\mathbb{X}$. Similarly, $E(Y|B)$, the average value of the random variable Y over the set B, plays a role in various applications, often in the form of $E(Y|X = x)$ or $E(f(Y)|X = x)$. For example, in classical reinforcement learning we use the conditional expectation operator $E(f(Y)|X = x) = \sum_{i=1}^{n} f(y_i)P(Y = y_i|X = x)$, where $\mathbb{Y} = \{y_1, \ldots, y_n\}$, to infer, for instance, the value

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84.

of a policy. One can define conditional expectations also on continuous spaces if there exist densities $p(x, y)$ on the joint space $\mathbb{X} \times \mathbb{Y}$. In modern probability theory such conditional probabilities and expectations are treated as special cases of Kolmogorov's conditional expectation. Coming from these more elementary forms of conditioning Kolmogorov's approach can appear counter-intuitive at the outset. Kolmogorov's conditional expectation takes a random variable Y and a set of events grouped in $\mathcal{G}$ and returns a random variable $E(Y|\mathcal{G})$ that is measurable with respect to the events in $\mathcal{G}$ and which equals Y when compared across any event contained in $\mathcal{G}$. Intuitively, these properties say that information of the events $\mathcal{G}$ allows us to determine the value $E(Y|\mathcal{G})$ and $E(Y|\mathcal{G})$ corresponds to a suitable average value of Y. The approach appears sometimes unfamiliar since one often thinks of a conditional expectation $E(Y|X = x)$ as a fixed value and not as a random variable. This difference occurs since for $E(Y|X = x)$ we already assume that a particular event occured, i.e. the event that the random variable X attains value x. If we do not make this assumption then $E(Y|X)$ is a random variable that depends on the values that X attains. Furthermore, the move to a family of events $\mathcal{G}$ allows one to consider the average value of Y across a variety of events and not just for the event $X = x$. In Kolmogorov's approach $\mathcal{G}$ is a σ-algebra which guarantees, in particular, that if we can calculate the average value for $X = x_1, X = x_2, \ldots$ then we can also calculate the average value for any union of these events, i.e. we could ask what is the average probability of Y if $X \geq 1$ etc. Kolmogorov's approach is also used to define conditional probabilities. The way these are defined is based on the simple observation that $P(A) = E(\chi A)$, where χA denotes the characteristic function of the event A, i.e. $\chi A(x)$ attains value 1 if $x \in A$ and 0 otherwise. In particular, conditional probabilities $P(A|\mathcal{G})$ are defined as $E(\chi A|\mathcal{G})$ and one can regain the more elementary forms of conditional probabilities by suitable choices of $\mathcal{G}$. Here, our aim is not to derive estimators for the most general form of conditioning but to understand how one can control estimation errors not only uniformly over a set of elementary events, but also over combinations of these events. That is, we control estimation errors for at most countable unions and intersections, as well as complements, of these elementary events. Figure 1 provides an overview of the different forms of conditioning and we fill in missing details about these in the preliminaries.

The problem of estimating a probability measure P uniformly over large classes of sets is well understood when the empirical measure P_n is used as the estimator. P_n converges uniformly over a family of sets $\mathcal{C}$ to the measure P if $\mathcal{C}$ is a VC-class (recall that $\mathcal{C}$ is a VC-class if there exists a set $x_1, \ldots, x_n$ and a labeling $b_1, \ldots, b_n \in \{0, 1\}$ such that for every $A \in \mathcal{C}$ there is an i for which $b_i \neq \chi A(x_i)$). In fact, in this case the rate of convergence is known; indeed, $\sup_{A \in \mathcal{C}} |P_n(A) - P(A)|$ converges to zero at a rate of $n^{-1/2}$, and the error decreases uniformly over $\mathcal{C}$. We make use of this approach to derive plug-in estimators of conditional probabilities $P(A|B)$ where $A \in \mathcal{C}_1, B \in \mathcal{C}_2$. The plug-in principle suggests to replace the unknown probability measure P with the empirical measure P_n to gain the estimator $P_n(A|B)$. Suitable restrictions on $\mathcal{C}_1$ and $\mathcal{C}_2$ analogous to the VC-class approach above allow us to derive $n^{-1/2}$ rates for this estimator. A difficulty that arises here is that the sets B we condition on can have small probabilities and $\inf_{B \in \mathcal{C}_2} P(B)$ can be zero. We develop a simple technique which decreases the rate of convergence but circumvents this problem and allows us to work with infinite sets $\mathcal{C}_2$ for which $\inf_{B \in \mathcal{C}_2} P(B) = 0$. A similar approach allows us to derive estimators for conditional expectations $E(f|B)$ and to control their estimation error. We need restrictions on the function class $\mathcal{F}$ over which the guarantees should hold. We extend standard results which say that a rate of $n^{-1/2}$ uniformly over $\mathcal{F}$ can be achieved when estimating expectations $E(f)$ if $\mathcal{F}$ is a VC-subgraph class (the family of subgraphs $\{\{(x, t) : x \in \mathbb{X}, t \leq f(x)\} : f \in \mathcal{F}\}$ is a VC-class), or more generally, a Donsker class (Dudley, 2014). Examples of such classes are sufficiently constrained neural networks (Shalev-Shwartz and Ben-David, 2014)[Sec. 20.4], and unit balls in a reproducing kernel Hilbert space (RKHS, (Aronszajn, 1950)) if the kernel function is continuous and the input space is compact. We extend these results to Kolmogorov's conditional expectation and probability.

This extension is straightforward when the involved families are finite, but is non-trivial when they are infinite. This is because, for instance, the fact that $\mathcal{C}$ is a VC-class does not necessarily imply that the smallest σ-algebra that contains $\mathcal{C}$ is a VC-class. In fact, $\sigma(\mathcal{C})$ is a VC-class if, and only if, it is generated by finitely many sets. The implication of this is that conditional expectations and probabilities given a σ-algebra $\mathcal{G}$ consisting of infinitely many sets need to be approximated by conditional expectations and probabilities over finite σ-algebras $\mathcal{G}_n$ where either $\bigcup_{n \in \mathbb{N}} \mathcal{G}_n = \mathcal{G}$ or one can increasingly well represent elements in $\mathcal{G}$ by elements in $\mathcal{G}_n$. A further difficulty one faces is that it is often easy to construct families $\mathcal{G}_1 \subset \mathcal{G}_2 \subset \ldots$ such that for each $A \in \mathcal{G}$ there exists an n and a $B \in \mathcal{G}_n$ which approximates A well in the sense that $P(A \Delta B) < \epsilon$. However, this convergence is typically not uniform. In natural settings one has, for instance, $\sup_{A \in \mathcal{G}} \inf_{B \in \mathcal{G}_n} P(A \Delta B) = 1/2$ despite this approximation property of the families $\mathcal{G}_n$; see Section 2.3 for details. The problematic sets get increasingly irregular as n increases and one way to resolve this problem is to work with functions f that posses certain smoothness properties, such as Lipschitz-continuity. This allows for efficient estimation of conditional expectations in the sense that one attains inequalities of the type

$$\|E(f|\mathcal{G}) - E(f|\mathcal{G}_n)\|_{\mathcal{L}^1(P)} \leq L d_n,$$

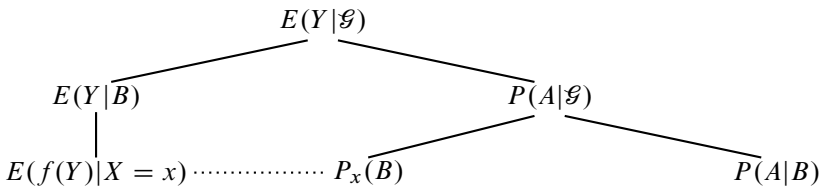

Figure 1: An overview of various forms of conditioning and their interrelation.

where f is an L-Lipschitz function and d_n is some measure of size of the elements in $\mathcal{G}_n$ which decreases to 0 in n. Such inequalities can then be combined with guarantees for estimators of $E(f|\mathcal{G}_n)$ to gain bounds on $\|E(f|\mathcal{G}) - E_n(f|\mathcal{G}_n)\|_{\mathcal{L}^1(P)}$ where we use $E_n(\cdot|\mathcal{G}_n)$ to denote the estimator of the conditional expectation. We explore this line of thought in Section 3.2. In particular, we use this approach to derive plug-in estimators for $E(f|\mathcal{G})$ and $P(A|\mathcal{G})$, and we control their estimation error over infinite families of events $\mathcal{G}$ and simultaneously over large function classes $\mathcal{F}$. In general, we focus on rates of convergence and we do not study finite sample guarantees. That being said, finite sample guarantees are amenable to the approach, e.g. by using data-driven Rademacher complexities as can be found in Giné and Nickl (2016)[Sec 3.5].

Comparison. There exists a variety of approaches to estimate conditional distributions and expectations. Of particular importance are the following two. 1. In the elementary setting where X and Y attain only finitely many values and one has observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ it is common to estimate $E(Y|X = x)$ by $\sum_{i=1}^n Y_i \times \chi\{X_i = x\}/|\{i : X_i = x\}|$, where $X_1, \ldots, X_n$ are i.i.d. according to P. From the central limit theorem it follows that these estimators converge with a rate of $n^{-1/2}$. Our estimators are the natural generalization of such estimators to not necessarily finite valued random variables. 2. Estimators of conditional expectations of the form $E(f|X = x)$ are popular in the kernel literature and have been pioneered by Song et al. (2009). f lies here in an RKHS and $E(f|X = x)$ is approximated through a function g in a second RKHS. It is known that under favorable conditions the kernel estimator converges with a rate of $\log(n)/n$ to an approximation of $E(f|X = x)$ – this corresponds to a rate of $\log^{1/2}(n)n^{-1/2}$ in our setting. An important condition is here that f is an element of the unit ball of a finite dimensional RKHS (Grünewälder et al., 2012). This approach is somewhat different from ours in that it controls the complexity of the conditional expectation estimator by representing it through an RKHS function that is constrained in its norm. Our techniques do not extend straightforwardly to this approach and we cannot use them to derive rates of convergence for the approximation. On the other hand, our approach does not rely on a particular function space like an RKHS or on assumptions on the dimension of the involved function spaces.

Contribution. Our main contributions are given below:

1. To the best of our knowledge plug-in estimators for conditional expectations and probabilities have not been systematically studied before. This comment applies, in particular, to the estimators we introduce in Equations (2,3) and (5-7).

2. Similarly, works that exploit VC-theory to study such conditional expectations and probabilities are lacking. The plug-in principle allows for a very efficient use of VC-theory and demonstrating this relation is, in our opinion, a significant contribution of our paper.

3. We use VC-theory to establish rates of convergence for the plug-in estimators, i.e. Prop. 3.1 to Prop. 3.4 provide rates of convergence for the estimators under various conditions and these propositions are novel.

4. Another contribution worth pointing out is the study of the interplay between σ-algebras and VC-classes. This is crucial for an understanding of rates of convergence of the plug-in estimators.

1.1 Preliminaries

We start by adding details about how conditional expectations and probabilities are defined (see (Dudley, 2002) for more details). To define Kolmogorov's conditional expectation formally consider a probability space $(\Omega, \mathcal{A}, P)$ and a σ-subalgebra $\mathcal{G} \subseteq \mathcal{A}$. A conditional expectation $E(Y|\mathcal{G})$ is a random variable that is $\mathcal{G}$-measurable and which agrees with Y over any element $B \in \mathcal{G}$, i.e. $\int_B E(Y|\mathcal{G})\, dP = \int_B Y\, dP$, for all $B \in \mathcal{G}$. $E(Y|\mathcal{G})$ is a random variable and is guaranteed to exist if $E|Y|$ is finite. Recall that conditional probabilities are defined by $P(A|\mathcal{G}) := E(\chi A|\mathcal{G})$. Specific choices of $\mathcal{G}$ yield the elementary conditional probabilities. For instance, with $\mathcal{G} = \{\emptyset, \Omega, B, \Omega\backslash B\}$, $P(A|\mathcal{G})(\omega) = P(A|B)$ for any $\omega \in B$, if $P(B) > 0$. Conditioning with respect to a random variable X is achieved by letting $\mathcal{G} = \{X^{-1}[B] : B \in \mathcal{B}\} =: \sigma(X)$ where $\mathcal{B}$ is the Borel-algebra on $\mathbb{R}$. In this case we also use $P(A|X) := P(A|\mathcal{G})$ and $E(Y|X) := E(Y|\mathcal{G})$. A caveat is that $E(Y|X)$ is a random variable that acts on the probability space Ω which might be $\mathbb{X} \times \mathbb{Y}$ or some abstract probability space. Ideally, one wants a conditional

expectation or a distribution that acts on $\mathbb{Y}$, e.g. if Y attains values in $\mathbb{R}$ then the conditional distribution should be a distribution on $\mathbb{R}$. For simplicity let $\mathbb{Y} = \mathbb{R}$ and let $\mathcal{B}$ be the Borel-algebra. If there exists $P_{Y|\mathcal{G}}(B, \omega)$ such that $P_{Y|\mathcal{G}}(B, \omega) := P(Y^{-1}[B]|\mathcal{G})$ almost surely, $P_{Y|\mathcal{G}}(B, \omega)$ is a probability measure on the space $(\mathbb{R}, \mathcal{B})$ for almost all ω and $P_{Y|\mathcal{G}}(B, \cdot)$ is $\mathcal{G}$-measurable then we call $P_{Y|\mathcal{G}}$ a conditional distribution. We also use the notation P_x in the product space case if for each $x \in \mathbb{X}$, P_x is a probability measure on the Borel sets $\mathcal{B}_{\mathbb{Y}}$ of $\mathbb{Y}$, $x \mapsto P_x(B)$ is $\mathcal{B}_{\mathbb{X}}$-measurable and $P(A \times B) = \int_B P_x(A) d\mu(x)$ for all $A \in \mathcal{B}_{\mathbb{Y}}, B \in \mathcal{B}_{\mathbb{X}}$. $\mu = PX^{-1}$ is the marginal measure on $\mathbb{X}$. The conditional expectation wrt. $f : \mathbb{Y} \to \mathbb{R}$ can then be written as $E(f(Y)|X \in B) = \int_B \int f(y) P_x(y)\, d\mu(x)$. We will use $E(f|X \in B)$ to denote such estimators if it is obvious that f is a function of Y only.

Empirical Processes. In this paper we assume that we have been given a sequence $S_1, S_2, \ldots$ of i.i.d. random variables attaining values in some sample space $\mathbb{S} \subseteq \mathbb{R}^d$, where we equip $\mathbb{S}$ with the Borel-algebra $\mathcal{B}_{\mathbb{S}}$. In the simplest case we are interested in estimating the probability law P of these random variables. One way to estimate P is to use the empirical measure $P_n = n^{-1} \sum_{i=1}^n \delta_{S_i}$, where δ_{S_i} denotes the measure that has point mass at S_i. The empirical measure P_n is a random probability measure on $\mathcal{B}_{\mathbb{S}}$. One can also view P_n as a stochastic process indexed by the sets in $\mathcal{B}_{\mathbb{S}}$, i.e. a stochastic process that maps $A \mapsto P_n A, A \in \mathcal{B}_{\mathbb{S}}$. The empirical process $\nu_n := n^{1/2}(P_n - P)$ is a centered and normalized version of this stochastic process. The empirical process can be indexed by subsets of $\mathcal{B}_{\mathbb{S}}$ or by sets of functions, where we use then the notation $Pf := \int f\, dP$. To avoid some technical difficulties it is useful to assume that the underlying probability space on which the S_i live is the product $(\mathbb{S}, \mathcal{B}_{\mathbb{S}}, P)^{\mathbb{N}+}$ and to identify the S_i with the projections onto the i'th coordinate (Dudley, 2014)[Ex. on p. 234]. Furthermore, the process ν_n is in general not measurable and one needs to use outer probabilities when studying its convergence behavior (Dudley, 2014)[Ch. 3]. For us these technicalities play only a minor role, i.e. we will have to consider rates of convergence in O_P^* instead of the more familiar O_P-notation. A sequence of random variables $Y_1, Y_2, \ldots$ lies in $O_P^*(a_n)$ for a sequence of positive real numbers $a_1, a_2, \ldots$ if, and only if, for every $\epsilon > 0$ there exists an $M > 0$ and $N \in \mathbb{N}$ such that $\Pr^*(|Y_n/a_n| > M) \leq \epsilon$ for all $n \geq N$. $\Pr^*$ denotes an outer measure. The most important result for us is that if a class of functions is a Donsker class then $\|P_n - P\|_{\mathcal{F}} \in O_P^*(n^{-1/2})$, where we use here the supremum norm $\|\nu_n\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |\nu_n(f)|$. This follows from the definition of Donsker classes (Dudley, 2014)[p. 137], Remark 3.7.27 in (Giné and Nickl, 2016) and the Portmanteau theorem (Dudley, 2014)[Thm. 3.27]. The result extends to families of sets $\mathcal{C}$ by considering $\mathcal{C} := \{\chi A : A \in \mathcal{C}\}$, If $\mathcal{C}$ is a Donsker class then $\|P_n - P\|_{\mathcal{C}} \in O_P^*(n^{-1/2})$, where $\|\nu_n\|_{\mathcal{C}} = \sup_{A \in \mathcal{C}} |\nu_n(C)|$.

2 Families of Sets and Functions

We collect in this section results about families of sets and functions that allow us to control estimation errors uniformly. VC-theory is one of the main tools to derive uniform rates of convergence. In particular, $\mathcal{C} = \{\chi C : C \in \mathcal{C}\}$ with $\mathcal{C}$ being a VC-class, and VC-subclasses of functions are Donsker classes if a measurability assumption is fulfilled (Dudley, 2014)[p. 258] and one attains a rate of convergence of $n^{-1/2}$ uniformly over any Donsker-class. Note that we will use the notation $\mathcal{C}$ throughout for the characteristic functions of sets in $\mathcal{C}$. Our running example in this paper is the following set

$$\{A : A = [\mathbf{x}, \mathbf{x} + h\mathbf{1}] \subseteq \mathbb{X}, \mathbf{x} \in \mathbb{X}, h \in [0, \infty[\} \cup \{\emptyset\}, \quad (1)$$

where we consider both $\mathbb{X} = \mathbb{R}^d$ and $\mathbb{X} = [0, 1]^d$. The family of sets $\mathcal{C}$ defined in eq. 1 is the family of hypercubes, h is the length of the sides of the hypercube and $\mathbf{x}$ specifies its location. Such families of sets are natural when trying to estimate a probability measure P and resembles the density estimation approach where a kernel is used to average the values of the density of P. This family of sets has finite VC-dimension and one can estimate P uniformly over $\mathcal{C}$ with a rate of $n^{-1/2}$.

Lemma 2.1 (Proof on p.10)**.** *$\mathcal{C}$ is a universal Donsker class if $\mathbb{X}$ is $\mathbb{R}^d$ or $[0, 1]^d$.*

The term universal refers here to the property that $\mathcal{C}$ is a Donsker class for any probability measure P, i.e. P does not need to posses any particular property like having a density function. Variations of this family can be considered, however, the family cannot be chosen much large since, for example, the family of all convex closed subsets of a bounded open set in $\mathbb{R}^d, d \geq 3$, is not a universal Donsker class (Dudley, 2014)[Thm 11.3].

2.1 Products and Intersections

Intersections between sets and products between functions and characteristic functions of sets play a major role for conditioning. In particular, for elementary conditional probabilities $P(A|B) = P(A \cap B)/P(B), A \in \mathcal{C}_1, B \in \mathcal{C}_2$, it is beneficial if the chracteristic functions of the sets in $\mathcal{C}_1 \sqcap \mathcal{C}_2 := \{A \cap B : A \in \mathcal{C}_1, B \in \mathcal{C}_2\}$ form a Donsker class. Similarly, for the general conditional probability $P(A|\mathcal{G}), A \in \mathcal{C}$, we gain fast rates of convergence if the characteristic functions of sets in $\mathcal{C} \sqcap \mathcal{G}$ are a Donsker class. Closely related is the importance of the product for conditional expectations $E(f|\mathcal{G}), f \in \mathcal{F}$. We gain fast rates if $\mathcal{F} \times \{\chi A : A \in \mathcal{G}\}$ is a Donsker class. VC theory is here useful: if $\mathcal{C}_1$ and $\mathcal{C}_2$ are VC-classes then so is $\mathcal{C}_1 \sqcap \mathcal{C}_2$ Dudley (2014)[Thm 4.34]. Similarly, the product $\mathcal{C}_1 \times \mathcal{C}_2 = \{A \times B : A \in \mathcal{C}_1, B \in \mathcal{C}_2\}$ is a VC-class if both $\mathcal{C}_1$ and $\mathcal{C}_2$ are. This implies that $\mathcal{F} \times \mathcal{C}$ is a VC-subgraph class if $\mathcal{F}$ is a VC-subgraph class and $\mathcal{C}$ is a VC-class.

2.2 Growing Families

Donsker classes are restricted in complexity and size. It is often useful to increase the size of the involved function class or family of sets with the number of samples at the cost of a slower rate of convergence. One way to achieve this is to use nested function classes $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \ldots$. In particular, if $\mathcal{F}_1$ is a Donsker class such that $c\mathcal{F}_1 \subseteq \mathcal{F}_1$ for all $c \in [0,1]$ and $\lambda_1 \leq \lambda_2 \leq \ldots$ is a non-decreasing sequence in $[0,\infty)$ then by letting $\mathcal{F}_n := \{cf : 0 \leq c \leq \lambda_n, f \in \mathcal{F}_1\}$ we gain a sequence of nested function classes. The rate of convergence is then slowed down to $\lambda_n n^{-1/2}$, but the guarantees are uniform over $\mathcal{F}_n$. A typical case is where $\mathcal{F}_1$ is a unit ball in some function space and we increase the ball size with n to exhaust the function space successively. This approach is closely related to Vapnik's *structural risk minimization (SRM)*.

Nested families of sets $\mathcal{C}_1 \subseteq \mathcal{C}_2 \ldots$. can also be used. Standard bounds for VC-classes (Devroye et al., 1996)[chp. 12] show that for VC-classes $\{\mathcal{C}_n\}_{n\geq 1}$ with VC-dimension v_n one has for any $n \geq 1$ and all $\delta \in (0,1)$ that

$$\Pr^*\Big(\sup_{A\in\mathcal{C}_n} \frac{\sqrt{n}|P_n(A) - P(A)|}{\sqrt{8\log(16/\delta) + 8v_n\log(n)}} \geq 1\Big) \leq \delta.$$

The argument allows us to derive a rate of up to $\sqrt{\log(n)}n^{-1/2}$ which is a $\sqrt{\log(n)}$ factor slower than the correct rate if one uses, for instance, a single VC-class $\mathcal{C} = \mathcal{C}_1 = \mathcal{C}_2 = \ldots$. Beside this minor reduction in rate the observation allows for a useful control of nested VC-classes. Given some $\delta > 0, q \in (0,1)$, we can choose the classes $\mathcal{C}_n$, for example, with VC-dimension $v_n = n^q$ and with outer probability of at least $1-\delta$ we have for any $n \geq 1$ that

$$\sup_{A\in\mathcal{C}_n} \frac{\sqrt{n}|P_n(A) - P(A)|}{\sqrt{8\log(16/\delta) + 8n^q\log(n)}} \leq 1.$$

Particularly, this implies that $\|P_n - P\|_{\mathcal{C}_n}$ lies in the order class $O_P^*(\sqrt{\log(n)}n^{(q-1)/2})$.

2.3 Extension to $\sigma(\mathcal{C})$

If $\mathcal{C}$ is a VC-class then we can estimate $P(A)$ well uniformly over $\mathcal{C}$. This does not guarantee us, however, that we estimate $P(\bigcup_{n\geq 1} A_n)$ or $P(\bigcap_{n\geq 1} A_n)$ well for arbitrary sequences $\{A_n\}_{n\geq 1}$ in $\mathcal{C}$ and it prevents us from approximating and estimating probabilities of more complicated sets. Furthermore, it introduces difficulties when trying to estimate general conditional probabilities $P(A|\mathcal{G})$ and expectations $E(A|\mathcal{G})$. The smallest family of sets which is closed under countable unions, intersections and complements and which contains $\mathcal{C}$ is $\sigma(\mathcal{C})$, the smallest σ-algebra that contains $\mathcal{C}$. There are a few ways by which one can extend guarantees to $\sigma(\mathcal{C})$ from $\mathcal{C}$. In interesting cases the rate of convergence of P_n to P is, however, significantly slower when measured over all of $\sigma(\mathcal{C})$. This difficulty is already indicated by the VC-dimension: the VC-dimension increases significantly when one moves from $\mathcal{C}$ to $\sigma(\mathcal{C})$. For example, a partition of $[0,1]$ consisting of n intervals of equal length has VC-dimension one whilst the smallest σ-algebra that contains this partition has VC-dimension n. A simple way to extend the guarantees that hold uniformly over $\mathcal{C}$ to $\sigma(\mathcal{C})$ makes use of some weak assumption about the probability measure P and the size of the sets in $\mathcal{C}$. This approach is based on the following lemma.

Lemma 2.2 (Proof on p. 11)**.** *Let $(\Omega, \mathcal{A}, P)$ be a probability space and $\mathcal{C} \subset \mathcal{A}$ a disjoint family of sets such that for each $A \in \mathcal{C}$ there exists $\{A_n\}_{n\in\mathbb{N}}$ in $\mathcal{C}$ with $\Omega\backslash A = \bigcup_{n\in\mathbb{N}} A_n$ and $\emptyset \in \mathcal{C}$. For any measure Q for which there exists a constant $c > 0$ such that for all $A \in \mathcal{C}$, $|Q(A) - P(A)| \leq cP(A)$, we have that $\sup_{A\in\sigma(\mathcal{C})} |Q(A) - P(A)|/P(A) \leq c$.*

We can make use of this lemma by letting Q be the empirical measure P_n. Now, if we impose a density assumption, say the density of P is lower bounded by $b > 0$, and if we assume the sets A to have at least a volume of d then $|P_n(A) - P(A)| \leq c$ implies that $|P_n(A) - P(A)| \leq P(A)c/(bd)$ and the lemma tells us that $\sup_{A\in\sigma(\mathcal{C})} |P_n(A) - P(A)| \leq c/(bd)$. So the guarantees that we have for elements in $\mathcal{C}$ transfer to guarantees for elements in $\sigma(\mathcal{C})$ by scaling the upper bound by $1/(bd)$.

When $\mathcal{C}$ is of infinite cardinality then we face a further difficulty since $\sigma(\mathcal{C})$ cannot be a VC-class in this case, even if $\mathcal{C}$ is a VC-class. This is a consequence of the following lemma.

Lemma 2.3 (Proof on p. 10)**.** *Let $\mathbb{X}$ be any set and $\mathcal{G}$ be a σ-algebra of subsets of $\mathbb{X}$. $\mathcal{G}$ is a VC-class if, and only if, $\mathcal{G}$ is a finite family of sets.*

This means that we need to approximate large σ-algebras $\mathcal{G}$ in a suitable way, for instance, through a sequence of finite σ-subalgebras $\mathcal{G}_1 \subset \mathcal{G}_2 \subset \ldots$ of $\mathcal{G}$ such that any element in $\mathcal{G}$ can eventually be approximated by elements in some $\mathcal{G}_n$. In particular, for each $A \in \mathcal{G}$ and $\epsilon > 0$ we want to have an element $B \in \bigcup_n \mathcal{G}_n$ such that $P(A\Delta B) < \epsilon$.

We demonstrate this approach on a simple example. We use in this example the Borel-algebra on the d-dimensional hypercube. We approximate this large σ-algebra with families $\mathcal{C}_n = \{\emptyset\} \cup \{[\mathbf{x}, \mathbf{x} + 2^{-n}\mathbf{1}) : \mathbf{x}_i \in \{0, 2^{-n}, \ldots, 1 - 2^{-n}\}, i \leq d\}$ with union $\mathcal{C} = \bigcup_{n\geq 1} \mathcal{C}_n$. $\mathcal{C}$ is a VC-class, however, $\sigma(\mathcal{C})$ is not. Also observe that the VC -dimension of $\mathcal{C}_n$ is 1 and of $\sigma(\mathcal{C}_n)$ is 2^{dn}. This family of sets is well studied, for instance in the context of classification (Devroye et al., 1996; Scott and Nowak, 2006). It is well known that $\mathcal{C}$ is large enough to approximate any Borel set in the above sense arbitrary well. We can now either use the SRM like approach discussed in Section 2.2

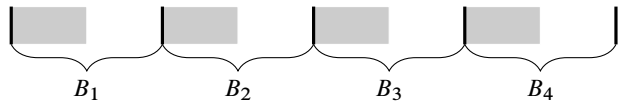

Figure 2: The figure shows a dyadic partition of the unit interval consisting of the set $B_1, \ldots, B_4$. The shaded area is a measurable set A which cannot be approximated well by this dyadic partition, i.e. if B is any union of the sets $B_1, \ldots, B_4$ then $P(A \Delta B) = 1/2$.

or use Lemma 2.2 together with the observation that $\mathcal{C}$ is a VC-class. The former approach tells us that if given n-samples we use the class $\sigma(\mathcal{C}_{\lambda_n})$, where $\{\lambda_n\}_{n\geq 1}$ is a non-decreasing sequence, then the rate of convergence is not slower than $\sqrt{\log(n)} 2^{d\lambda_n/2} n^{-1/2}$ without any assumption on P. Alternatively, imposing an assumption on the density, using Lemma 2.2 together with the fact that $\mathcal{C}$ is a VC-class, we gain a rate of convergence that is upper bounded by $2^{d\lambda_n} n^{-1/2}$, i.e. a rate of convergence that is inferior by a factor 2 in the exponent to the SRM like approach. We summarize this result in the following corollary. A simple proof of the approximation property of the families $\sigma(\mathcal{C}_n)$ is contained in the appendix for the reader's convenience.

Corollary 2.1 (Proof on p. 11)**.** *Let $([0,1]^d, \mathcal{A}, P)$ be a probability space such that P has a density p that is lower bounded by $b > 0$. Let $\{\lambda_n\}_{n\geq 1}$ be a non-decreasing sequence in $\mathbb{N}_+$ such that $\lim_{n\to\infty} \lambda_n = \infty$ then $\|\nu_n(A)\|_{\sigma(\mathcal{C}_{\lambda_n})} \in O_P^*(\sqrt{\log(n)} 2^{d\lambda_n/2})$. Furthermore, for any Borel set A and $\epsilon > 0$ there exists an $n \in \mathbb{N}$ and $B \in \sigma(\mathcal{C}_{\lambda_n})$ such that $P(A \Delta B) \leq \epsilon$.*

We can increase λ_n only logarithmically since we have an exponential increase in the number of intervals, e.g. $\lambda_n \approx 1/(2d) \log n$ results in a rate of convergence of $n^{-1/4}$.

We did not quantify the approximation error. This is, in fact, not straight forward: consider the case where $d = 1$ for simplicity and observe that for any n there exists a Borel set A_n such that $\inf_{B\in\sigma(\mathcal{C}_n)} P(A_n \Delta B) = 1/2$ (A_n can actually be chosen as an element of $\sigma(\mathcal{C}_{n+1})$). Figure 2 visualizes such a set A_n. Let us consider now one particular $n_0 \in \mathbb{N}$ and the corresponding set A_{n_0}. This particular set has the property that $\inf_{B\in\sigma(\mathcal{C}_n)} P(A_{n_0} \Delta B) = 1/2$ for all $n \leq n_0$ and $\inf_{B\in\sigma(\mathcal{C}_n)} P(A_{n_0} \Delta B) = 0$ for all $n > n_0$.

There exists another interesting approach to curtail the increase in model complexity when passing from $\mathcal{C}$ to $\sigma(\mathcal{C})$ which uses the symmetric convex hull of a VC-class. Let $\mathcal{C}$ be a VC-class then $\mathcal{C}$ is a VC-subgraph class and the symmetric convex hull of $\mathcal{C}$, which we will denote with $\operatorname{sco}\mathcal{C} = \{\sum_{i\leq n} \alpha_i \chi_{A_i} : A_i \in \mathcal{C}, \sum_{i\leq n} |\alpha_i| \leq 1, n \in \mathbb{N}\}$, is a Donsker class if certain measurability assumptions are fulfilled (Giné and Nickl, 2016)[Thm.3.7.34]. The symmetric convex hull of a VC class can be used in the following way: for finitely many disjoint elements $A_1, \ldots, A_n \in \mathcal{C}$ the function $m^{-1}\chi(\bigcup_{i\leq m} A_i) = m^{-1}\sum_{i\leq m} \chi_{A_i} \in \operatorname{sco}\mathcal{C}$ and $m^{-1}\chi(\bigcup_{i\leq m} A_i)$ can be estimated with a rate of $n^{-1/2}$ if $\mathcal{C}$ is a VC-class. In fact, the A_i do not even need to be disjoint for this to hold. Using the nested function class approach we can work with $\lambda_n \operatorname{sco}\mathcal{C}$, where λ_n is a non-decreasing sequence. This approach results in a rate of convergence of $\lambda_n n^{-1/2}$ and we can estimate unions of arbitrary m elements in $\mathcal{C}$ for any $m \leq \lambda_n$. Intuitively this approach is based on how difficult it is to represent the elements we want to estimate. Sets that can be represented by the union of few sets of $\mathcal{C}$ can be estimated well while sets that need large numbers of such sets to be represented have weak guarantees. Due to space constraints we will focus in the following only on the earlier discussed approach which uses growing families of σ-algebras.

3 Conditioning

We introduce estimators for conditional expectations and probabilities and we provide rates of convergence for these. We start with elementary forms of conditioning and define corresponding plug-in estimators. We then extend these to gain estimators of Kolmogorov's conditional expectation and probabilities, and we provide estimators of conditional expectations given a random variable X.

3.1 Elementary

The simplest form of a conditional expectation is the average of a random variable $Y : \mathbb{S} \to \mathbb{R}$ over a measurable set $B \in \mathcal{B}_{\mathbb{S}}$ which has positive measure $P(B) > 0$, that is $E(Y \mid B) = \int_B Y \, dP \big/ P(B)$, where $Y \in \mathcal{L}^1(P)$. Given i.i.d. samples $S_1, \ldots, S_n$ the plug-in estimate of this conditional expectation is

$$E_n(Y \mid B) := \left(\int_B Y \, dP_n\right) \Big/ P_n(B), \qquad (2)$$

$$\text{with } P_n(B) = \frac{1}{n}\sum_{i=1}^{n} \delta_{S_i}(B),$$

$$\text{and } \int_B Y \, dP_n = \frac{1}{n}\sum_{i=1}^{n} Y(S_i) \times \delta_{S_i}(B)$$

if $P_n(B) > 0$ and $E_n(Y \mid B) := 0$ otherwise. Similarly, $P_n(A|B) := P_n(A \cap B)/P_n(B)$ if $P_n(B) > 0$ and $P_n(B|A) = 0$ otherwise is the plug-in estimator for the elementary conditional probability. To emphasize that Y can be any measurable function acting on $\mathbb{S}$ we replace in the following Y by $f : \mathbb{S} \to \mathbb{R}$. The plug-in estimators converge uniformly over Donsker classes to $E(f|B)$ and $P(A|B)$ under various assumptions on the sets B we condition with. In particular, with the notation $\mathcal{F}_{\mathcal{C}} := \{f \times \chi_C : f \in \mathcal{F}, C \in \mathcal{C}\}$, we have the following first result.

Proposition 3.1 (Proof on p. 12)**.** *If $\mathcal{C} \subseteq \mathcal{B}_{\mathbb{S}}$ is a finite set with $\inf_{B\in\mathcal{C}} P(B) > 0$, $\mathcal{F}$ is a subset of $\mathcal{L}^1(P)$ uniformly bounded in supremum norm and $\mathcal{F}_{\mathcal{C}}$ is a $P-$Donsker class then*

$$\sup_{f\in\mathcal{F}} \sup_{B\in\mathcal{C}} |E_n(f \mid B) - E(f \mid B)| = O_P^*(n^{-1/2}).$$

Furthermore, if $\mathcal{C}' \subseteq \mathcal{B}_{\mathbb{S}}$, is such that $\mathcal{C}'_{\mathcal{C}}$ is a P-Donsker class then

$$\sup_{A\in\mathcal{C}'} \sup_{B\in\mathcal{C}} |P_n(A|B) - P(A|B)| = O_P^*(n^{-1/2}).$$

Here, $\mathcal{F}$ can be the unit ball of an RKHS with continuous kernel acting on a compact set or a neural-network, suitably restricted in its complexity and $\mathcal{C}'$ can be any VC-class (up to measurability assumptions). The restriction that $\mathcal{C}$ is finite can be weakened if we are willing to accept a slower rate of convergence. The following holds if the measure P has a density wrt. Lebesgue measure μ.

Proposition 3.2 (Proof on p. 13)**.** *If $\mathcal{C} \subseteq \mathcal{B}_{\mathbb{S}}$, $\mathcal{C}$ and $\mathcal{F}_{\mathcal{C}}$ are P-Donsker classes, $\mathcal{F}$ is a subset of $\mathcal{L}^1(P)$ uniformly bounded in supremum norm and P has a density which is lower bounded by a constant $b > 0$ then with $\mathcal{C}_n := \{C : C \in \mathcal{C}, \mu(C) \geq n^{-\alpha}\}$ and $\alpha \in [0, 1/2)$*

$$\sup_{f\in\mathcal{F}} \sup_{B\in\mathcal{C}_n} |E_n(f \mid B) - E(f \mid B)| \in O_P^*(n^{\alpha-1/2}).$$

Furthermore, if $\mathcal{C}' \subseteq \mathcal{B}_{\mathbb{S}}$ is such that $\mathcal{C}'_{\mathcal{C}}$ is a P-Donsker class then

$$\sup_{A\in\mathcal{C}'} \sup_{B\in\mathcal{C}_n} |P_n(A|B) - P(A|B)| \in O_P^*(n^{\alpha-1/2}).$$

If $\mathcal{C}$ is the set of cubes in $[0,1]^d$ then $\mu(C) = h^d$ for a cube C with edge length h. Hence, given n samples we can evaluate conditional expectations and probabilities over cubes with $h \geq n^{-\alpha/d}$.

3.2 Kolmogorov's Conditional Expectation and Probability

The previous results generalize straightforwardly to σ-algebras $\sigma(\mathcal{C})$ if $\mathcal{C}$ is finite, because $\sigma(\mathcal{C})$ has then finite VC-dimension. Such results are of interest if one wants to be able to combine events in $\mathcal{C}$ in various ways. For example, in a graphical model we have a set of events $A_1, \ldots, A_n$ and we aim to infer dependencies between the A_i but also between combinations like $A_1 \cap A_2 \cap A_3$. This approach can also be used to control estimation errors of Kolmogorov's conditional expectation and probability. Let $\mathcal{G}$ be a σ-algebra consisting of finitely many sets then there exists a unique finite partition $\mathcal{P}_{\mathcal{G}} \subseteq \mathcal{G}$ of $\mathbb{S}$ such that each $B \in \mathcal{G}$ can be written as a finite union of elements of the partition and the σ-algebra generated by the partition equals $\mathcal{G}$ (see Lemma B.1 in the Supplementary Material). Using this partition we define estimators $E_n(f|\mathcal{G})$ and $P_n(A|\mathcal{G})$ by

$$E_n(f|\mathcal{G}) := \sum_{B\in\mathcal{P}_{\mathcal{G}}} E_n(f|B) \times \chi B \quad \text{and} \qquad (3)$$
$$P_n(A|\mathcal{G}) := \sum_{B\in\mathcal{P}_{\mathcal{G}}} P_n(A|B) \times \chi B.$$

An important property of conditional expectations $E(f|\mathcal{G})$ is that they are $\mathcal{G}$-measurable. Observe that the estimators $E_n(f|\mathcal{G})$ and $P_n(A|\mathcal{G})$ have the same property, i.e. they are also $\mathcal{G}$-measurable. The following result shows that these estimators are sensible and estimate $E(f|\mathcal{G})$ and $P(A|\mathcal{G})$ uniformly with a rate of $n^{-1/2}$ if $\mathcal{G}$ consists of finitely many elements only.

Proposition 3.3 (Proof on p. 14)**.** *If $\mathcal{G} \subseteq \mathcal{B}_{\mathbb{S}}$ is a σ-algebra consisting of finitely many sets, $\mathcal{F}$ is a subset of $\mathcal{L}^1(P)$ uniformly bounded in supremum norm, $\mathcal{F}_{\mathcal{G}}$ is a $P-$Donsker class then*

$$\sup_{f\in\mathcal{F}} \|E_n(f \mid \mathcal{G}) - E(f \mid \mathcal{G})\|_{\mathcal{L}^1(P)} \in O_P^*(n^{-1/2}).$$

Furthermore, if $\mathcal{C} \subseteq \mathcal{B}_{\mathbb{S}}$, is such that $\mathcal{C}_{\mathcal{G}}$ is a P-Donsker class then

$$\sup_{A\in\mathcal{C}} \|P_n(A \mid \mathcal{G}) - P(A \mid \mathcal{G}))\|_{\mathcal{L}^1(P)} \in O_P^*(n^{-1/2}).$$

Using Proposition 3.2 these results can be extended to infinite collections of sets. We demonstrate this on the sequence $\{\mathcal{C}_n\}_{n\geq 1}$ introduced in Section 2.3. We used there a non-decreasing sequence $\{\lambda_n\}_{n\geq 1}$ to balance the rate of convergence against the size of the family of sets we condition with. We use in the following the notation $\mathcal{G}_{\lambda_n} := \sigma(\mathcal{C}_{\lambda_n})$ and we assume that the function class $\mathcal{F}$ fulfills

$$\sup_{f\in\mathcal{F}} \sup_{B\in\mathcal{G}_{\lambda_n}} |\nu_n(f \times \chi B)| \in O_P^*(\sqrt{\log n}\, 2^{d\lambda_n/2}). \qquad (4)$$

Recall that ν_n denotes the empirical process. This assumption is effectively saying that Corollary 2.1 also holds for $\mathcal{F} \times \{\chi C : C \in \mathcal{G}_{\lambda_n}\}$. This assumption can be verified for concrete function classes in a similar way as in Cor. 2.1.

Proposition 3.4 (Proof on p. 14)**.** *Let $([0,1]^d, \mathcal{B}, P)$ be a probability space such that P has a density p that is lower bounded by $b > 0$ and let $\{\lambda_n\}_{n\geq 1}$ be a non-decreasing sequence in $\mathbb{N}_+$ such that*

$$\lambda_n \in o\Big(\frac{1}{3d\log(2)} \log\Big(\frac{n}{\sqrt{\log(n)}}\Big)\Big).$$

If $\mathcal{F}$ is a subset of $\mathcal{L}^1(P)$ uniformly bounded in supremum norm which fulfills Equation 4, then

$$\sup_{f\in\mathcal{F}} \|E_n\left(f \,\middle|\, \mathcal{G}_{\lambda_n}\right) - E\left(f \,\middle|\, \mathcal{G}_{\lambda_n}\right)\|_\infty$$
$$\in O_P^*(\sqrt{\log(n)}\, 2^{(3/2)d\lambda_n} n^{-1/2}).$$

Furthermore, if $\mathcal{C} \subseteq \mathcal{B}_{\mathbb{S}}$, is such that $\mathcal{C}$ fulfills Equation 4

$$\sup_{A \in \mathcal{C}} \| P_n\left(A \,\middle|\, \mathcal{G}_{\lambda_n}\right) - P\left(A \,\middle|\, \mathcal{G}_{\lambda_n}\right) \|_\infty \in O_P^*(\sqrt{\log(n)} 2^{(3/2) d \lambda_n} n^{-1/2}).$$

Like in Corollary 2.1 λ_n can increase only logarithmically since the σ-algebras grow exponentially in size. Also observe that we kept $\mathcal{F}$ fixed. We could let this family grow with n at a further expense of the rate of convergence.

The last proposition gives us only bounds for the difference between $E_n(f|\mathcal{G}_{\lambda_n})$ and $E(f|\mathcal{G}_{\lambda_n})$. We would obviously prefer to replace the latter term with $E(f|\mathcal{G})$, where $\mathcal{G} = \sigma(\bigcup_n \mathcal{G}_n)$ is in this example the Borel-algebra. This brings us back to the discussion centered around Figure 2 and the difficulties of approximating arbitrary sets in $\mathcal{G}$. This problem can be alleviated if the functions in $\mathcal{F}$ are Lipschitz continuous with common constant L. We demonstrate this for $d = 1$: observe that $\mathcal{G}_n \subset \mathcal{G}$ for all n, and that the conditional expectation can be written as

$$E(f|\mathcal{G}_n) = \sum_{i=1}^{2^n} \frac{1}{P(I_i)} \int_{I_i} f dP \times \chi I_i,$$

where $I_i = [a_i, b_i)$ denotes the i'th Dyadic interval of length 2^{-n}. Now, using these observations and Jensen's inequality for conditional expectations gives us

$$\begin{aligned} &\|E(f|\mathcal{G}) - E(f|\mathcal{G}_n)\|_{\mathcal{L}^1(P)} \\ &\leq \|f - E(f|\mathcal{G}_n)\|_{\mathcal{L}^1(P)} \\ &= \textstyle\sum_{i=1}^{2^n} \int_{a_i}^{b_i} |f - \int_{a_i}^{b_i} f dP / P(I_i)| dP \\ &\leq 2^{-n} L \textstyle\sum_{i=1}^{2^n} P(I_i) = 2^{-n} L. \end{aligned}$$

Therefore the approximation error between $E(f|\mathcal{G})$ and $E(f|\mathcal{G}_{\lambda_n})$ decreases as $1/n$ if we allow for a logarithmic increase of λ_n and the dominant term in the bound on $\sup_{f \in \mathcal{F}} \|E_n(f|\mathcal{G}_{\lambda_n}) - E(f|\mathcal{G})\|_{\mathcal{L}^1(P)}$ is the error term bounded in Proposition 3.4.

3.3 Conditioning with respect to a Random Variable

As discussed in the preliminaries there are multiple ways to condition a random variable Y by a second random variable X. Here, we are discussing plug-in estimators for the case where the sample space $\mathbb{S} := \mathbb{X} \times \mathbb{Y}$ and X and Y are the projections onto the components, i.e. $X(x, y) = x$ and $Y(x, y) = y$. First, we provide an estimator for the conditional probability measure $P_{Y|\mathcal{G}}(A, \omega) := P(Y^{-1}[A]|\mathcal{G})(\omega)$, $A \in \mathcal{B}_{\mathbb{Y}}$, $\omega \in \mathbb{S}$ and $\mathcal{G} = \sigma(X)$. The plug-in estimator based on i.i.d. random variables $S_1, \ldots S_n$ attaining values in $\mathbb{S}$, using the partition $\mathcal{P}_X \subseteq \sigma(X)$ implied by Lemma B.1 and assuming that X attains only finitely many values, is

$$P^n_{Y|\mathcal{G}}(A, \omega) := \sum_{B \in \mathcal{P}_X} P_n(\mathbb{X} \times A|B) \times \chi B(\omega) \quad (5)$$

where we moved n, which indicates that we are dealing with an estimator, into the superscript on the left side. The rate of convergence of this estimator can be studied in the same way as in the previous section.

Conditional expectations of the form $E(f|X \in B) := \int_B \int f(y) P_x(y) d\mu(x)$, $f : \mathbb{Y} \to \mathbb{R}$, can also be estimated if we use sets $B \in \mathcal{B}_{\mathbb{X}}$ that have non-zero measure. The plug-in estimator is

$$E_n(f|X \in B) = \sum_{i=1}^{n} f(Y_i) \times \delta_{X_i}(B) \Big/ P_n(B), \quad (6)$$

if $P_n(B) > 0$ and $E_n(f|X \in B) = 0$ otherwise. This resembles the estimator in eq. (2). Generalizing this to finite σ-subalgebras $\mathcal{S}$ of $\sigma(X)$ we gain

$$E_n(f|\mathcal{S}) = \sum_{B \in \mathcal{P}_{\mathcal{S}}} E_n(f|X \in B) \times \chi B. \quad (7)$$

Observe, that it is in general not possible to evaluate $E_n(f|X = x)$ directly since the set $B = \{X = x\}$ has probability 0 if X is not a discrete random variable. However, it is possible to shrink the sets B with increasing sample size. For instance, we can use the hypercubes defined in Equation 1 together with Proposition 3.2 to successively refine the estimate with respect to X and to approximate the sets $\{X = x\}$ well for large n. Similarly, we can use non-decreasing families of σ-subalgebras $\mathcal{S}_1 \subset \mathcal{S}_2 \subset \ldots$ to reproduce the result of Proposition 3.4 in the context of conditioning with respect to random variables.

4 Discussion

This paper presents a plug-in approach to the estimation of conditional expectations and probabilities and demonstrates how one can derive rates of convergence for these estimators in various settings. It is unclear if these rates are minimax optimal or if the approaches can be improved to gain faster rates of convergence. In particular, under stronger density assumptions we expect that Nadaraya-Watson style estimators can prove superior. While we focused here only on rates it would be useful to complement these with finite sample guarantees. Another open problem is the extension of our techniques to the case where the conditional expectation is approximated by a function (Song et al., 2009; Grünewälder et al., 2012). Furthermore, it would be interesting to see how these estimators perform in real-world problems. The application of conditional expectations and probabilities to real-world problems is not always straightforward since one often has no i.i.d. data available but samples from a dependent process (e.g. in reinforcement learning).

Acknowledgement

I would like to thank Łukasz Grabowski for a very helpful discussion concerning Lemma 2.3.

References

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

L.E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state markov chains. *Ann. Math. Statist.*, 37(6):1554–1563, 12 1966.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, Inc., 2006.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

R.M. Dudley. *Real Analysis and Probability*. Cambridge University Press, 2002.

R.M. Dudley. *Uniform Central Limit Theorems*. Cambridge University Press, 2nd edition, 2014.

D.H. Fremlin. *Measure Theory - Vol. 4: Topological Measure Spaces*. Torres Fremlin, 2003.

E. Giné and R. Nickl. *Mathematical Foundations of Infinite-dimensional Statistical Models*. Cambridge University Press, 2016.

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

L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, pages 257–286, 1989.

B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. On causal and anticausal learning. In *International Conference on Machine Learning*, 2012.

C. Scott and R.D. Nowak. Minimax-optimal classification with dyadic decision trees. *IEEE Transactions on Information Theory*, 2006.

S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 01 2016.

L. Song, J. Huang, A.J. Smola, and K. Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *International Conference on Machine Learning*, 2009.

R. S. Sutton and A.G. Barto. *Introduction to Reinforcement Learning*. MIT Press, Cambridge, MA, USA, 1st edition, 1998.

C. Szepesvári. *Algorithms for Reinforcement Learning*. Morgan & Claypool, 2010.