Journal of Machine Learning Research 22 (2021) 1-36 Submitted 11/20; Revised 5/21; Published 8/21

Oblivious Data for Fairness with Kernels

Steffen Grünewälder S.GRUNEWALDER@LANCASTER.AC.UK

Azadeh Khaleghi A.KHALEGHI@LANCASTER.AC.UK

Department of Mathematics and Statistics
Lancaster University
Lancaster, UK

Editor: Massimiliano Pontil

Abstract

We investigate the problem of algorithmic fairness in the case where sensitive and non-sensitive features are available and one aims to generate new, 'oblivious', features that closely approximate the non-sensitive features, and are only minimally dependent on the sensitive ones. We study this question in the context of kernel methods. We analyze a relaxed version of the Maximum Mean Discrepancy criterion which does not guarantee full independence but makes the optimization problem tractable. We derive a closed-form solution for this relaxed optimization problem and complement the result with a study of the dependencies between the newly generated features and the sensitive ones. Our key ingredient for generating such oblivious features is a Hilbert-space-valued conditional expectation, which needs to be estimated from data. We propose a plug-in approach and demonstrate how the estimation errors can be controlled. While our techniques help reduce the bias, we would like to point out that no post-processing of any dataset could possibly serve as an alternative to well-designed experiments.

Keywords: Algorithmic Fairness, Kernel Methods

1. Introduction

Machine learning algorithms trained on historical data may inherit implicit biases which can in turn lead to potentially unfair outcomes for some individuals or minority groups. For instance, gender-bias may be present in a historical dataset on which a model is trained to automate the postgraduate admission process at a university. This may in turn render the algorithm biased, leading it to inadvertently generate unfair decisions. In recent years, a large body of work has been dedicated to systematically addressing this problem, whereby various notions of fairness have been considered, see, e.g. (Calders et al., 2009; Zemel et al., 2013; Louizos et al., 2015; Hardt et al., 2016; Joseph et al., 2016; Kilbertus et al., 2017; Kusner et al., 2017; Calmon et al., 2017; Zafar et al., 2017; Kleinberg et al., 2017; Donini et al., 2018; Madras et al., 2018; Oneto et al., 2020), and references therein.

Among the several *algorithmic fairness* criteria, one important objective is to ensure that a model's prediction is not influenced by the presence of sensitive information in the data. In this paper, we address this objective from the perspective of (fair) representation learning. Thus, a central question which forms the basis of our work is as follows.

Can the observed features be replaced by close approximations
that are independent of the sensitive ones?

More formally, assume that we have a dataset such that each data-point is a realization of a random variable (X, S) where S and X are in turn vector-valued random variables corresponding to the sensitive and non-sensitive features respectively. We further allow X and S to be arbitrarily dependent, and ask whether it is possible to generate a new random variable Z which is ideally independent of S and close to X in some *meaningful probabilistic sense*. As an initial step, we may assume that X is zero-mean, and aim for decorrelation between Z and X. This can be achieved by letting $Z = X - E^S X$ where $E^S X$ is the conditional expectation of X given S. The random variable Z so-defined is not correlated with S and is close to X. In particular, it recovers X if X and S are independent. In fact, under mild assumptions, Z gives the best approximation (in the mean-squared sense) of X, while being uncorrelated with S. Observe that while the distribution of Z differs from that of X, this new random variable seems to serve the purpose well. For instance, if S corresponds to a subject's *gender* and X to a subject's *height*, then Z corresponds to height of the subject centered around the average height of the class corresponding to the subject's gender. The key contributions of this work, briefly summarized below, are theoretical; we also provide an evaluation of the proposed approach through experiments in the context of classification and regression[1]. Before giving an overview of our results, we would also like to point out that while our techniques help reduce the bias, it is important to note that no post-processing of any dataset could possibly serve as an alternative to well-designed experiments.

Contributions. Building upon this intuition, and using results inspired by testing for independence using the Maximum Mean Discrepancy (MMD) criterion (see e.g. (Gretton et al., 2008)), we obtain a related optimization problem in which X and $E^S X$ are replaced with Hilbert-space-valued random variables and Hilbert-space-valued conditional expectations. While the move to Hilbert spaces does not enforce complete independence between the new features and the sensitive features, it helps to significantly reduce the dependencies between the features. The new features $\boldsymbol{Z}$ have various useful properties which we explore in this paper (the notation $\boldsymbol{Z}$ is used to highlight that the features attain values in a Hilbert space and not in the space in which our samples lie). They are also easy to generate from samples $(X_1, S_1), \ldots, (X_n, S_n)$. The main challenge in generating the oblivious features $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ is that we do not have access to the Hilbert-space-valued conditional expectation and need to estimate it from data. Since we are concerned with Reproducing Kernel Hilbert Spaces (RKHSs) here, we use the reproducing property to extend the plugin approach of Grünewälder (2018) to the RKHS setting and tackle the estimation problem. We further show how estimation errors can be controlled. After obtaining the empirical estimates of the conditional expectations, we generate oblivious features as well as an oblivious kernel matrix to be used as input to any kernel method. This guarantees a significant reduction in the dependence between the predictions and the sensitive features. We cast the objective of finding oblivious features $\boldsymbol{Z}$ which approximate the original features X well while maintaining minimal dependence on the sensitive features S, as a constrained optimization problem. Making use of Hilbert-space-valued conditional expectations, we provide a closed form solution to the optimization problem proposed. Specifically, we first prove that our solution satisfies the constraint of the optimization problem at hand, and show via Proposition 4 that it is indeed optimal. Through Proposition 2 we relate the strength of the dependencies between $\boldsymbol{Z}$ and S to how close $\boldsymbol{Z}$ lies to the low-dimensional manifold corresponding to the image under the feature map ϕ. This result is key in providing some insight into the interplay between probabilistic independence and approximations in the Hilbert space. We extend known estimators for real-valued

1. Our implementations are available at `https://github.com/azalk/Oblivious.git`.

conditional expectations to estimate those taking values in a Hilbert space, and show via Proposition 5 how to control their estimation errors. This result in itself may be of independent interest in future research concerning Hilbert-space-valued conditional expectations. We provide a method to generate oblivious features and the oblivious kernel matrix which can be used instead of the kernel matrix to reduce the dependence of the prediction on the sensitive features; the computational complexity of the approach is $O(n^2)$.

Related Work. Among the vast literature on algorithmic fairness, (Donini et al., 2018; Madras et al., 2018; Oneto et al., 2020), which fit into the larger body of work on fair representation learning, are closest to our approach. Madras et al. (2018) describe a general framework for fair representation learning. Their approach is inspired by generative adversarial networks and is based on a game played between generative models and adversarial evaluations. Depending on which function classes one considers for the generative models and for the adversarial evaluations one can describe a vast array of approaches. Interestingly, it is possible to interpret our approach in this general context: the encoder f corresponds to a map from $\mathbb{X}$ and $\mathbb{S}$ to $\mathcal{H}$, where our new features $\boldsymbol{Z}$ live. We do not have a decoder but compare features directly (one could also take our decoder to be the identity map). Our adversary is different from that used by Madras et al. (2018). In their approach a regressor is inferred which maps the features to the sensitive features, while we compare sensitive features and new features by applying test functions to them. The regression approach performs well in their context because they only consider finitely many sensitive features. In the more general framework considered in the present paper where the sensitive features are allowed to take on continuous values, this approach would be sub-optimal since it cannot capture all dependencies. Finally, we ignore labels when inferring new features. It is also worth pointing out that our approach is not based on a game played between generative models and an adversary but we provide closed form solutions. On other hand, while the focus of Donini et al. (2018) is mostly on empirical risk minimization under fairness constraints, the authors briefly discuss representation learning for fairness as well. In particular, Equation (13) in the reference paper effectively describes a conditional expectation in Hilbert space, though it is not denoted or motivated as such. The conditional expectation is based on the binary features S only and the construction is applied in the linear kernel context to derive new features. The paper does not go beyond the linear case for representation learning but there is a clear link to the more general notions of conditional expectation on which we base our work. We discuss the relation to (Donini et al., 2018) in detail in Section 6.5 and we show how their approach can be extended beyond binary sensitive features by making use of our conditional expectation estimates. In (Oneto et al., 2020) the demographic parity constraint is considered in a multi-task setting. The sensitive features have to be discrete in this setting, attaining only finitely many values. The tasks share an unknown "representation" and it is assumed that this representation fulfills the demographic parity constraint. The authors propose an optimization problem that minimizes a sum of a multi-task loss function (for classification and regression) and a metric that quantifies how well demographic parity is achieved. The parameters of this optimization problem are the potential representations and task specific functions. Both the Sinkhorn divergence and MMD are considered as metrics to quantify demographic parity. The focus is on a case where the potential representations are one-layer networks and the task specific functions are linear.

Organization. The rest of the paper is organized as follows. In Section 2 we introduce our notation and provide preliminary definitions used in the paper. Our problem formulation and optimization objective are stated in Section 3. As part of the formulation we also define the notion of $\mathcal{H}$-

independence between Hilbert-space-valued features and the sensitive features. In Section 4 we study the relation between $\mathcal{H}$-independence and bounds on the dependencies between oblivious and sensitive features. In Section 5 we provide a solution to the optimization objective. In Section 6 we derive an estimator for the conditional expectation and use it to generate oblivious features and the oblivious kernel matrix. We provide some empirical evaluations in Section 7, and end with some concluding remarks and open directions in Section 8.

2. Preliminaries

In this section we introduce some notation and basic definitions. Consider a probability space $(\Omega, \mathcal{A}, P)$. For any $A \in \mathcal{A}$ we let $\chi A : \Omega \to \{0, 1\}$ be the indicator function such that $\chi A(\omega) = 1$ if, and only if, $\omega \in A$. Let $\mathbb{X}$ be a measurable space in which a random variable $X : \Omega \to \mathbb{X}$ takes values and $\mathbb{S}$ be a measurable space in which the random variables $S : \Omega \to \mathbb{S}$ takes values. We denote by $\sigma(X)$ the σ-algebra generated by X. Let $\mathcal{H}$ be an RKHS composed of functions $h : \mathbb{X} \to \mathbb{R}$ and denote its feature map by $\phi(x) : \mathbb{X} \to \mathcal{H}$ where, $\phi(x) = k(x, \cdot)$ for some positive definite kernel $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$. As follows from the reproducing kernel property of $\mathcal{H}$ we have $\langle \phi(x), h \rangle = h(x)$ for all $h \in \mathcal{H}$. Moreover, observe that $\phi(X)$ is in turn a random variable attaining values in $\mathcal{H}$. In Appendix A we provide some technical details concerning Hilbert-space-valued random variables such as $\phi(X)$.

Conditional Expectation. For the random variable X and S defined above, we denote by $E^S X$ the random variable corresponding to Kolmogorov's conditional expectation of X given S, i.e. $E^S X = E(X|\sigma(S))$, see, e.g. (Shiryaev, 1989). Recall that in a special case where $\mathbb{S} = \{0, 1\}$ we simply have

$$E(X|S = 0)\chi\{S = 0\} + E(X|S = 1)\chi\{S = 1\}$$

where, $E(X|S = i)$ is the familiar conditional expectation of X given the event $\{S = i\}$ for $i = 0, 1$. Thus, in this case, the random variable $E^S X$ is equal to $E(X|S = 0)$ if S attains value 0 and is equal to $E(X|S = 1)$ otherwise. Note that the above example is for illustration only, and that X and S may be arbitrary random variables: they are not required to be binary or discrete-valued. Unless otherwise stated, in this paper we use Kolmogorov's notion of conditional expectation. We will also be concerned with conditional expectations that attain values in a Hilbert space $\mathcal{H}$, which mostly behave like real-valued conditional expectations (see (Pisier, 2016) and Appendix B for details). Next, we introduce Hilbert-space-valued $\mathcal{L}^2$-spaces which play a prominent role in our results.

Hilbert-space-valued $\mathcal{L}^2$-spaces. For a Hilbert space $\mathcal{H}$, we denote by $\mathcal{L}^2(\mathcal{H}) = \mathcal{L}^2(\Omega, \mathcal{A}, P; \mathcal{H})$ the $\mathcal{H}$-valued $\mathcal{L}^2$ space. If $\mathcal{H}$ is an RKHS with a bounded and measurable kernel function then $\phi(X)$ is an element of $\mathcal{L}^2(\Omega, \mathcal{A}, P; \mathcal{H})$. The space $\mathcal{L}^2(\Omega, \mathcal{A}, P; \mathcal{H})$ consists of all (Bochner)-measurable functions $\boldsymbol{X}$ from Ω to $\mathcal{H}$ such that $E(\|\boldsymbol{X}\|^2) < \infty$ (see Appendix A for more details). We call these functions random variables or Hilbert-space-valued random variables and denote them with bold capital letters. As in the scalar case we have a corresponding space of equivalence classes which we denote by $L^2(\Omega, \mathcal{A}, P; \mathcal{H})$. For $\boldsymbol{X}, \boldsymbol{Y} \in \mathcal{L}^2(\Omega, \mathcal{A}, P; \mathcal{H})$ we use $\boldsymbol{X}^\bullet, \boldsymbol{Y}^\bullet$ for the corresponding equivalence classes in $L^2(\Omega, \mathcal{A}, P; \mathcal{H})$. The space $L^2(\Omega, \mathcal{A}, P; \mathcal{H})$ is itself a Hilbert space with norm and inner product given by $\|\boldsymbol{X}^\bullet\|_2^2 = E(\|\boldsymbol{X}\|^2)$ and $\langle \boldsymbol{X}^\bullet, \boldsymbol{Y}^\bullet \rangle_2 = E(\langle \boldsymbol{X}, \boldsymbol{Y} \rangle)$, where we use a subscript to distinguish this norm and inner product from the ones from $\mathcal{H}$. The norm and inner product have a corresponding pseudo-norm and bi-linear form acting on $\mathcal{L}^2(\mathcal{H})$ and we also denote these by $\|\cdot\|_2$ and $\langle \cdot, \cdot \rangle_2$.

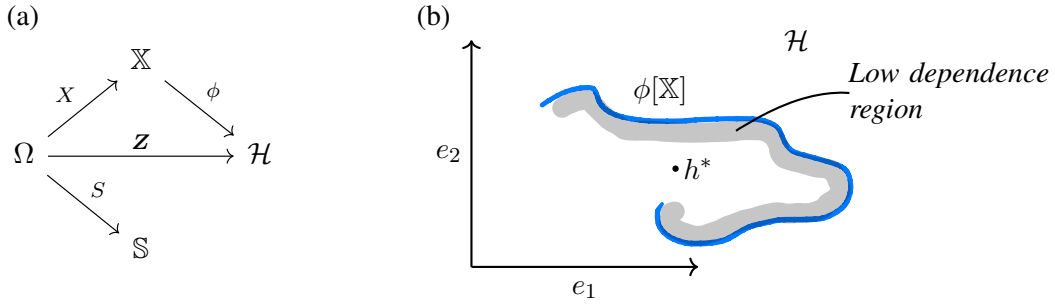


Figure 1: **(a)** The three main random variables in Problem 1 are shown. The non-sensitive features X attains values in $\mathbb{X}$ and is mapped onto the RKHS $\mathcal{H}$ through the feature map ϕ; the sensitive features S attains values in $\mathbb{S}$, and $\boldsymbol{Z}$ attains values in $\mathcal{H}$. All three random variables are defined on the same probability space $(\Omega, \mathcal{A}, P)$. **(b)** The image of $\mathbb{X}$ under ϕ is sketched (blue curve). This is a subset of $\mathcal{H}$ whose projection onto the subspace spanned by two orthonormal basis elements e_1 and e_2 is shown here. The set $\phi[\mathbb{X}]$ is a low-dimensional manifold if ϕ is continuous. The element $h^* = E(\phi(X))$ lies in the convex hull of $\phi[\mathbb{X}]$. Intuitively, if $\boldsymbol{Z}$ attains values mainly in the gray shaded area then $\boldsymbol{Z}$ is only weakly dependent on S.

3. Problem Formulation

We formulate the problem as follows. Given two random variables $X : \Omega \to \mathbb{X}$ and $S : \Omega \to \mathbb{S}$ corresponding to non-sensitive and sensitive features in a dataset, we wish to devise a random variable $Z : \Omega \to \mathbb{X}$ which is independent of S and closely approximates X in the sense that for all $Z' : \Omega \to \mathbb{X}$ we have,

$$\|Z - X\|_2 \leq \|Z' - X\|_2. \tag{1}$$

Dependencies between random variables can be very subtle and difficult to detect. Similarly, completely removing the dependence of X on S without changing X drastically is an intricate task that is rife with difficulties. Thus, we aim for a more tractable objective, described below, which still gives us control over the dependencies.

We start by a *strategic shift* from probabilistic concepts to interactions between functions and random variables. Consider the RKHS $\mathcal{H}$ of functions $h : \mathbb{X} \to \mathbb{R}$ with feature map ϕ as introduced in Section 2, and assume that $\mathcal{H}$ is large enough to allow for the approximation of arbitrary indicator functions $\chi\{Z \in A'\}$, A' a measurable subset of $\mathbb{X}$, in the $\mathcal{L}^2$-pseudo-norm for any $\mathbb{X}$-valued random variable Z. Observe that if

$$E(h(Z) \times g(S)) = E(h(Z)) \cdot E(g(S)) \tag{2}$$

for all $h \in \mathcal{H}, g \in \mathcal{L}^2$ then Z and S are, indeed, independent. This is because h and g can be used to approximate arbitrary indicator functions. In particular, for any measurable subset A' of $\mathbb{X}$ and B' of $\mathbb{S}$ there exist functions h and g such that

$$P(\{Z \in A'\} \cap \{S \in B'\}) \approx E(h(Z) \times g(S)) = E(h(Z)) \cdot E(g(S)) \approx P(Z \in A') \cdot P(S \in B').$$

This means that the independence constraint of the optimization problem of (1) translates to (2). Note that using RKHS elements as test functions is a common approach for detecting dependencies and is used in the MMD-criterion (e.g. (Gretton et al., 2008)).

On the other hand, due to the reproducing property of the kernel of $\mathcal{H}$, we can also rewrite the constraint (2) as

$$E(\langle h, \phi(Z)\rangle \times g(S)) = E\langle h, \phi(Z)\rangle \cdot E(g(S)). \tag{3}$$

Observe that $\phi(Z)$ is a random variable that attains values in a low-dimensional manifold; if the kernel function is continuous and $\mathbb{X} = \mathbb{R}^d$ then the image $\phi[\mathbb{X}]$ of $\mathbb{X}$ under ϕ is a d-dimensional manifold which we denote in the following by $\mathcal{M}$. In Figure 1 this manifold is visualized as the blue curve. Therefore, while Equation (3) is linear in $\phi(Z)$, depending on the shape of the manifold, it can lead to an arbitrarily complex optimization problem.

We propose to relax (3) by moving away from the manifold, replacing $\phi(Z)$ with a random variable $\boldsymbol{Z} : \Omega \to \mathcal{H}$ which potentially has all of $\mathcal{H}$ as its range. This simplifies the original optimization problem to one over a vector space under a linear constraint. To formalize the problem, we rely on a notion of $\mathcal{H}$-*independence* introduced below.

Definition 1 ($\mathcal{H}$-Independence) *We say $\boldsymbol{Z} \in \mathcal{L}^2(\Omega, \mathcal{A}, P; \mathcal{H})$ and $S : \Omega \to \mathbb{S}$ are $\mathcal{H}$-independent if and only if for all $h \in \mathcal{H}$ and all bounded measurable $g : \mathbb{S} \to \mathbb{R}$ it holds that,*

$$E(\langle h, \boldsymbol{Z}\rangle \times g(S)) = E\langle h, \boldsymbol{Z}\rangle \times E(g(S)).$$

Thus, instead of solving for $Z : \Omega \to \mathbb{X}$ in (1), we seek a solution to the following optimization problem.

Problem 1 *Find $\boldsymbol{Z} \in \mathcal{L}^2(\Omega, \mathcal{A}, P; \mathcal{H})$ that is $\mathcal{H}$-independent from S (in the sense of Definition 1) and is close to X in the sense that*

$$\|\boldsymbol{Z} - \phi(X)\|_2 \leq \|\boldsymbol{Z}' - \phi(X)\|_2$$

for all $\boldsymbol{Z}'$ which are also $\mathcal{H}$-independent of S.

Observe that the $\mathcal{H}$-independence constraint imposed by Problem 1, ensures that all non-linear predictions based on $\boldsymbol{Z}$ are uncorrelated with the sensitive features S. The setting is summarized in Figure 1(a).

Projection onto $\mathcal{M}$. If $\boldsymbol{Z}$ lies in the image of ϕ and $\mathcal{H}$ is a 'large' RKHS then $\mathcal{H}$-independence also implies complete independence between the estimator $\langle \hat{h}, \boldsymbol{Z}\rangle$ and S. To see this, assume that there exists a random variable $W : \Omega \to \mathbb{X}$ such that $\boldsymbol{Z} = \phi(W)$ and that the RKHS is *characteristic*. Since for any $f \in \mathcal{H}$ and bounded measurable $g : \mathbb{S} \to \mathbb{R}$

$$E(f(W) \times g(S)) = E(\langle f, \boldsymbol{Z}\rangle \times g(S)) = E\langle f, \boldsymbol{Z}\rangle \cdot E(g(S)) = E(f(W)) \cdot E(g(S))$$

we can deduce that W and S is independent. Moreover, since $\boldsymbol{Z}$ is a function of W it is also independent of S. In general, $\boldsymbol{Z}$ will not be representable as some $\phi(W)$ and there can be dependencies between $\langle \hat{h}, \boldsymbol{Z}\rangle$ and S. However, if $\boldsymbol{Z}$ attains values close to the manifold $\mathcal{M}$ then we can find a random variable W such that $\phi(W)$ is close to $\boldsymbol{Z}$ and the dependence between $\phi(W)$ and S is controlled by how close $\boldsymbol{Z}$ is to the manifold. More exactly, we allow $\boldsymbol{Z}$ to be translated by $h^* \in \mathcal{H}$ before measuring the distance. This is important because the manifold itself can lie away from the origin while the $\boldsymbol{Z}$ we construct in Section 5 lies around the origin. The distance we consider is

$$d^2(\boldsymbol{Z} + h^*, \mathcal{M}) := E(\inf_{h \in \mathcal{M}} \|\boldsymbol{Z} + h^* - h\|^2),$$

the average Hilbert space distance between $\boldsymbol{Z} + h^*$ and the manifold. Observe that the expectation on the right side is well defined when $\mathcal{M}$ is compact since we can then replace $\mathcal{M}$ with a countable dense subset of $\mathcal{M}$.

Showing that a suitable W exists is not trivial; the difficulty is that for values that $\boldsymbol{Z}$ might attain in $\mathcal{H}$ there can be many points on the manifold closest to that value and selecting points on the manifold in a way that makes the random variable W well defined needs a result on *measurable selections*. The following proposition makes use of such a selection and guarantees the existence of a suitable W, i.e. it states that there exists a random variable W such that $\phi(W)$ achieves the minimal distance to $\boldsymbol{Z} + h^*$.

Proposition 1 *Consider $\boldsymbol{Z} \in \mathcal{L}^2(\Omega, \mathcal{A}, P; \mathcal{H})$, assume that the kernel function is continuous and (strictly) positive-definite, and $\mathbb{X}$ is compact. For any $h^* \in \mathcal{H}$ there exists a $\sigma(\boldsymbol{Z})$-measurable random variable W which attains values in $\mathbb{X}$ such that $\phi(W) \in \mathcal{L}^2(\Omega, \mathcal{A}, P; \mathcal{H})$ and*

$$\|\boldsymbol{Z} + h^* - \phi(W)\|_2 = d(\boldsymbol{Z} + h^*, \mathcal{M}).$$

The proof is provided in Appendix C.1.

We will call such a variable W provided by the proposition *a projection of $\boldsymbol{Z}$ on $\mathcal{M}$*. The variable W can be approximated algorithmically for a given $\boldsymbol{Z}$ and h^* (see Appendix E.3). Furthermore, $\phi(W)$ is a good approximation of $\phi(X)$ whenever $\boldsymbol{Z}$ is, as

$$\|\phi(W(\omega)) - \phi(X(\omega))\| \leq \|\phi(W(\omega)) - \boldsymbol{Z}(\omega)\| + \|\boldsymbol{Z}(\omega) - \phi(X(\omega))\| \leq 2\|\boldsymbol{Z}(\omega) - \phi(X(\omega))\|,$$

where we used that $\phi(W(\omega))$ is closest to $\boldsymbol{Z}(\omega)$ on $\mathcal{M} = \phi[\mathbb{X}]$. Therefore,

$$\|\phi(W) - \phi(X)\|_2 \leq 2\|\boldsymbol{Z} - \phi(X)\|_2.$$

4. Bounding the dependencies

A common approach to quantifying the dependence between random variables is to consider

$$|P(A \cap B) - P(A)P(B)|$$

where A and B run over suitable families of events. In our setting, these families are the σ-algebras $\sigma(\boldsymbol{Z})$ (or, alternatively, $\sigma(W)$) and $\sigma(S)$, and the difference between $P(A \cap B)$ and $P(A)P(B)$, $A \in \sigma(\boldsymbol{Z})$ or $A \in \sigma(W), B \in \sigma(S)$, quantifies the dependence between the random variables $\boldsymbol{Z}$ and S, and W and S, respectively. Upper bounds on the absolute difference of these two quantities are related to the notion of α-dependence which underlies α-mixing. In times-series analysis mixing conditions like α-mixing play a significant role since they provide means to control temporal dependencies (see, e.g., (Bradley, 2007; Doukhan, 1994)). The aim of this section is to show how the notion of $\mathcal{H}$-independence is related to the dependence between the random variables. In particular, Proposition 2 below states a bound on the dependence between W and S in terms of the distance of $\boldsymbol{Z}$ to the manifold $\mathcal{M}$. Furthermore, if $\boldsymbol{Z}$ and $\phi(W)$ are closely coupled in the sense that there exists a constant c such that for any event $A_1 \in \sigma(\boldsymbol{Z})$ there exist an event $A_2 \in \sigma(W)$ fulfilling $P(A_1 \triangle A_2) \leq c$ then the dependence between $\boldsymbol{Z}$ and S can also be bounded. For the bound to be useful we want a small value of c for which the above holds, e.g. if we let $c = 1$ then the above holds trivially but the bound we provide below becomes vacuous. In this context, observe that W,

as constructed above, is a function of $\boldsymbol{Z}$ and we know that $\sigma(W) \subset \sigma(\boldsymbol{Z})$. However, the opposite inclusion is not guaranteed to hold.

Coming back to bounding the dependence between W and S: the high level idea is that $\mathcal{H}$-independence would correspond to normal independence if we had function evaluations '$h(Z)$' instead of inner products $\langle h, \boldsymbol{Z}\rangle$ (given that $\mathcal{H}$ is sufficient to approximate indicator functions). While generally there is no such expression for the inner product we know that for $\phi(W)$ we actually have the equivalence $\langle h, \phi(W)\rangle = h(W)$ due to the reproducing property of the kernel function. In contrast to $\boldsymbol{Z}$ the random variable $\phi(W)$ does not need to be $\mathcal{H}$-independent of S, however, if $\boldsymbol{Z} + h^*$ and $\phi(W)$ are not too far from each other in $\|\cdot\|_2$-norm then $\phi(W)$ will be approximately $\mathcal{H}$-independent of S and we can say something about the dependence between W and S. Therefore, the bound below is stated in terms of $\|\boldsymbol{Z} + h^* - \phi(W)\|_2$, which is equal to the distance between $\boldsymbol{Z} + h^*$ and $\mathcal{M}$, and a measure of how well indicator functions can be approximated. More specifically, the bound is controlled by the functional

$$\psi(A) = \inf_{f \in \mathcal{H}} 2\|\chi A(W) - f(W)\|_2 + \|f\|\, d(\boldsymbol{Z} + h^*, \mathcal{M}), \tag{4}$$

where $A \in \{W[C] : C \in \sigma(W)\}$ and f has to balance between approximating the indicator function while keeping $\|f\|\, d(\boldsymbol{Z} + h^*, \mathcal{M})$ small. The function ψ has a natural interpretation as the minimal error that can be achieved in a regularized interpolation problem. If $\mathcal{H}$ lies dense in a certain space, then any relevant indicator can in principle be approximated arbitrary well. This is not saying that $\psi(A)$ will be small since the norm of the element that approximates the indicator might be large. But the approximation error, which is $\|\chi A(W) - f(W)\|_2$, can be made arbitrary small. With this notation in place the proposition is as follows.

Proposition 2 *Consider a $\boldsymbol{Z} \in \mathcal{L}^2(\Omega, \mathcal{A}, P; \mathcal{H})$ which is $\mathcal{H}$-independent from S, suppose that the kernel function is continuous and (strictly) positive-definite, and $\mathbb{X}$ is compact. Let W be a projection of $\boldsymbol{Z}$ on $\mathcal{M}$. For any $A \in \sigma(W)$ and $B \in \sigma(S)$, with $A' = W[A]$ being the image of A under W, the following holds,*

$$|P(A \cap B) - P(A)P(B)| \leq \psi(A').$$

Furthermore, for $A \in \sigma(\boldsymbol{Z})$, if $c > 0$ is such that $\mathcal{B}_A = \{W[C] : C \in \sigma(W), P(C \triangle A) \leq c\}$ is non-empty then for any $B \in \sigma(S)$,

$$|P(A \cap B) - P(A)P(B)| \leq 2c + \inf_{D \in \mathcal{B}_A} \psi(D).$$

The proof is provided in Appendix C.2.

Intuitively, as visualized in Figure 1, the proposition states that if $\boldsymbol{Z}$ mostly attains values in the gray area then the dependence between W and S is low and, if W and $\boldsymbol{Z}$ are strongly coupled, then the dependence between $\boldsymbol{Z}$ and S is also low.

4.1 Estimating $\psi(A)$

The key quantity in Proposition 2 is $\psi(A)$. To control $\psi(A)$ it is necessary to control how well the RKHS can approximate indicators and to estimate the distance $d(\boldsymbol{Z} + h^*, \mathcal{M})$. The former problem is more difficult and might be approached using the theory of interpolation spaces; we do not try to develop the necessary theory here but only mention a simple result on denseness at the end of this section. On the other hand, the latter problem is easy to deal with: the distance $d(\boldsymbol{Z} + h^*, \mathcal{M})$

between $\boldsymbol{Z} + h^*$ and $\mathcal{M}$ can be estimated efficiently. In the case where the space $\mathbb{X}$ is compact and ϕ is a continuous function, we propose an empirical estimate of $d(\boldsymbol{Z} + h^*, \mathcal{M})$ given by

$$d_n(\boldsymbol{Z} + h^*, \mathcal{M}) := \frac{1}{n} \sum_{i=1}^{n} \min_{h \in \mathcal{M}} \|\boldsymbol{Z}_i + h^* - h\| \tag{5}$$

where $\boldsymbol{Z}_i,\ i \leq n,\ n \in \mathbb{N}$, are n independent copies of $\boldsymbol{Z}$. Note that the compactness of $\mathbb{X}$ together with the continuity of ϕ make the $\min$ operator in (5) well-defined.

Proposition 3 *Consider a $\boldsymbol{Z} \in \mathcal{L}^2(\Omega, \mathcal{A}, P; \mathcal{H})$ which is $\mathcal{H}$-independent from S, suppose that the kernel function is continuous and (strictly) positive-definite, and $\mathbb{X}$ is compact. Let $\rho = \max_{x \in \mathbb{X}} \|\phi(x)\| < \infty$. For any $h^* \in \mathcal{H}$ with $\|h^*\| \leq \rho$ and every $\epsilon > 0$ we have,*

$$\Pr(|d_n(\boldsymbol{Z} + h^*, \mathcal{M}) - d(\boldsymbol{Z} + h^*, \mathcal{M})| \geq \epsilon) \leq 2 \exp\Big(-\frac{2n\epsilon^2}{25\rho^2}\Big).$$

The proof is provided in Appendix C.3.

Going back to the approximation error $\|\chi A(W) - f(W)\|_2$, where $A \subset \mathbb{X}$ is the image under W of some set $C \in \sigma(W)$ and $f \in \mathcal{H}$ we like to mention the following: let $\nu = PW^{-1}$ be the push-forward measure of P under W. If $\mathcal{H}$ lies dense in $\mathcal{L}^2(\mathbb{X}, \mathcal{B}, \nu)$ then for any such A and any $\epsilon > 0$ there exists a function f such that $\|\chi A(W) - f(W)\|_2 < \epsilon$, i.e. for the measurable set A there exists a function $f \in \mathcal{H}$ such that

$$\int (\chi A(W) - f(W))^2 \, dP = \int (\chi A(x) - f(x))^2 \, d\nu(x) < \epsilon^2,$$

using (Fremlin, 2001, Theorem 235Gb). In many cases the continuous functions $C(\mathbb{X})$ lie dense in $\mathcal{L}^2(\mathbb{X}, \mathcal{B}, \nu)$ and a universal RKHS $\mathcal{H}$ is sufficient to approximate the indicators χW (see (Sriperumbudur et al., 2011)).

5. Best $\mathcal{H}$-independent features

In this section we discuss how to obtain $\boldsymbol{Z}$ as a closed-form solution to Problem 1. To this end, inspired by the sub-problem in the linear case, we obtain $\boldsymbol{Z}$ using Hilbert-space-valued conditional expectations. We further show that these features are $\mathcal{H}$-independent of S and that $\boldsymbol{Z}$ is the best $\mathcal{H}$-independent approximation of $\phi(X)$.

In the linear case discussed in the Introduction it turned out that $Z = X - E^S X + EX$ is a good candidate for the new features Z. In the Hilbert-space-valued case a similar result holds. The main difference here is that we do have to work with Hilbert-space-valued conditional expectations. For any random variable $\phi(X) \in \mathcal{L}^2(\Omega, \mathcal{A}, P; \mathcal{H})$, and any σ-subalgebra $\mathcal{B}$ of $\mathcal{A}$, conditional expectation $E^{\mathcal{B}}\phi(X)$ is defined and is again an element of $\mathcal{L}^2(\Omega, \mathcal{A}, P; \mathcal{H})$. We are particularly interested in conditioning with respect to the sensitive random variable S. In this case, $\mathcal{B}$ is chosen as $\sigma(S)$, the smallest σ-subalgebra which makes S measurable, and we denote this conditional expectation by $E^S\phi(X)$. A natural choice for the new features is

$$\boldsymbol{Z} = \phi(X) - E^S\phi(X) + E(\phi(X)). \tag{6}$$

The expectation $E(\phi(X))$ is to be interpreted as the Bochner-integral of $\phi(X)$ given measure P. Importantly, if S and $\phi(X)$ are independent, we have with this choice that $\boldsymbol{Z} = \phi(X) = \phi(X)$ and we are back to the standard kernel setting. Also, if $\phi(X) \in \mathcal{L}^2(\Omega, \mathcal{A}, P; \mathcal{H})$ then so is $\boldsymbol{Z}$.

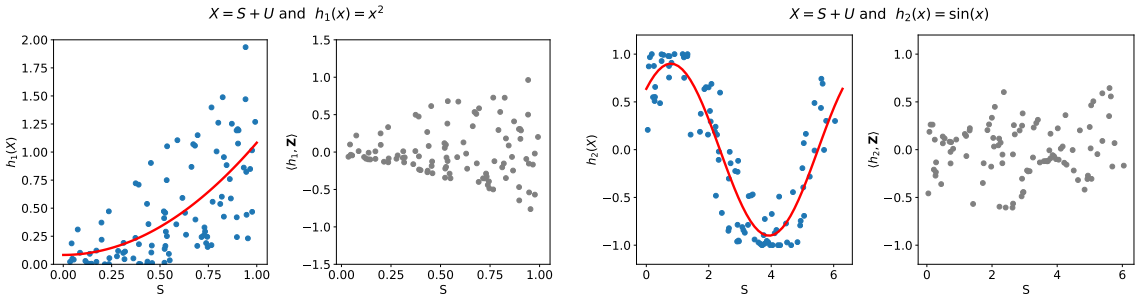


Figure 2: The figure shows data from two different settings. In the left two plots $X = S + U$, where S and U are independent, S is uniformly distributed on $[0, 1]$ and U is uniformly distributed on $[-1/2, 1/2]$. The function h_1 is the quadratic function. The leftmost plot shows $h_1(X)$ against S and the plot to its right shows a centered version of $\langle h_1, \boldsymbol{Z} \rangle$ plotted against S. Similarly for the right plots with the difference that S is uniformly distributed on $[0, 2\pi]$ and U is uniformly distributed on $[0, \pi/2]$. The function $h_2(x)$ is $\sin(x)$. The red curves show the best regression curve, predicting $h_1(X)$ and $h_2(X)$ using S.

We can verify that the features $\boldsymbol{Z}$ are, in fact, $\mathcal{H}$-independent of S. In particular, for any $h \in \mathcal{H}$ and $g \in \mathcal{L}^2$,

$$\begin{aligned} &E(\langle \phi(X) - E^S \phi(X), h \rangle \times g(S)) \\ &= \langle E(\phi(X) \times g(S)) - E((E^S \phi(X)) \times g(S)), h \rangle \\ &= \langle E(\phi(X) \times g(S)) - E(E^S(\phi(X) \times g(S))), h \rangle = 0. \end{aligned}$$

Since $E(\phi(X))$ is a constant this implies that $E(\langle \boldsymbol{Z}, h \rangle \times g(S)) = E(h(X)) \cdot E(g(S))$ A similar argument shows that $E\langle \boldsymbol{Z}, h \rangle = E(h(X))$. Thus, $\boldsymbol{Z}$ is $\mathcal{H}$-independent of S.

In Figure 2 the effect of the move from $\phi(X)$ to $\boldsymbol{Z}$ is visualized. In the figure S is plotted against $h_1(X)$ and $h_2(X)$ (blue dots), where h_1 corresponds to the quadratic function and h_2 to the sinus function. The dependencies between $h_1(X)$ and S, as well as $h_2(X)$ and S, are high and there is clear trend in the data. The two red curves correspond to the best regression functions, using S to predict $h_1(X)$ and $h_2(X)$. The relation between the new features and S is shown in the other two plots (gray dots). In the case of h_1 one can observe that the dependence between $\langle h_1, \boldsymbol{Z} \rangle$ and S is much smaller and, by the design of $\boldsymbol{Z}$, $\langle h_1, \boldsymbol{Z} \rangle$ and S are uncorrelated. Similarly, for $\langle h_2, \boldsymbol{Z} \rangle$, whereas here the dependence to S seems to be even lower and it is difficult to visually verify any remaining dependence between S and $\langle h_2, \boldsymbol{Z} \rangle$.

An interesting aspect of this transformation from X to $\boldsymbol{Z}$ is that $\boldsymbol{Z}$ is automatically uncorrelated with S for all functions h in the corresponding RKHS, without the need to ever explicitly consider a particular h. Besides being $\mathcal{H}$-independent of S these new features $\boldsymbol{Z}$ also closely approximate our original features $\phi(X)$ if the influence from S is not too strong, i.e. the mean squared distance is

$$E(\|\phi(X) - \boldsymbol{Z}\|^2) = E(\|E^S \phi(X) - E(\phi(X))\|^2)$$

which is equal to zero if X is independent of S. In fact, $\boldsymbol{Z}$ is the best approximation of $\phi(X)$ in the mean squared sense under the $\mathcal{H}$-independent constraint. This is essentially a property of

the conditional expectation which corresponds to an orthogonal projection in $L^2(\Omega, \mathcal{A}, P; \mathcal{H})$. We summarize this property in the following result.

Proposition 4 *Given $\phi(X), \boldsymbol{Z}' \in \mathcal{L}^2(\Omega, \mathcal{A}, P; \mathcal{H})$ such that $\boldsymbol{Z}'$ is $\mathcal{H}$-independent of S, then*

$$E(\|\phi(X) - \boldsymbol{Z}'\|^2) \geq E(\|\phi(X) - \boldsymbol{Z}\|^2),$$

where $\boldsymbol{Z} = \phi(X) - E^S\phi(X) + E(\phi(X))$. Furthermore, $\boldsymbol{Z}$ is the unique minimizer (up to almost sure equivalence).

The proof provided is in Appendix C.4.

Change in predictions. When replacing $\phi(X)$ by $\boldsymbol{Z}$ we lose information (we reduce the influence of the sensitive features). An interesting question to ask is, 'how much does the reduction in information change our predictions?' A simple way to bound the difference in predictions is as follows. Consider any $h \in \mathcal{H}$, for instance corresponding to a regression function, then

$$|h(X) - \langle h, \boldsymbol{Z}\rangle| \leq \|h\|\|\phi(X) - \boldsymbol{Z}\| \leq \|h\|\|E^S\phi(X) - E(\phi(X))\|$$

where $\|E^S\phi(X) - E(\phi(X))\|$ effectively measures the influence of S. Hence, the difference in prediction is upper bound by the norm of the predictor (here h) and a quantity that measures the dependence between S and $\phi(X)$.

Example. To demonstrate that the effect of the move from X to $\boldsymbol{Z}$ can be profound we consider the following fundamental example: suppose that X and S are standard normal random variables with covariance $c \in [-1, 1]$ and consider the *linear kernel* $k(x, y) = xy$, $x, y \in \mathbb{R}$. In this case $\phi(X) = X$ and $E^S X = cS$ is also normally distributed (see (Bertsekas and Tsitsiklis, 2002)[Sec4.7]). Hence, $\boldsymbol{Z} = X - E^S X + E(X)$ is normally distributed and $E(\boldsymbol{Z} \times S) = c - cE(S^2) = 0$. This implies that $\boldsymbol{Z}$ and S are, in fact, *fully independent*, regardless of how large the dependence between the original features X and the sensitive features S may be. In the case where X and S are fully dependent, i.e. $X = aS$ for some $a \in \mathbb{R}$, the features $\boldsymbol{Z}$ are equal to zero and do not approximate X.

Next, consider a *polynomial kernel* of second order such that the quadratic function $h(x) = x^2$ lies within the corresponding RKHS. The inner product between this h and $\boldsymbol{Z}$ is equal to $X^2 - E^S X^2 + E(X^2)$ and *is not independent* of S. Hence, the *kernel function affects the dependence* between $\boldsymbol{Z}$ and S. Also, within the same RKHS there lie linear functions and for any linear function h' it holds that $\langle \boldsymbol{Z}, h' \rangle$ is independent of S. Therefore, within the same RKHS we can have directions in which $\boldsymbol{Z}$ is independent of S and directions where both variables are dependent.

6. Generating oblivious features from data

To be able to generate the features $\boldsymbol{Z}$ we need to first estimate the conditional expectation $E^S\phi(X)$ from data. To this end, we devise a plugin-approach. After introducing this approach in Section 6.1 we discuss how the estimation errors of the plugin-estimator can be controlled in Section 6.2. In Section 6.3 we show how the oblivious features can be generated. Finally, in Section 6.4, we demonstrate how the approach can be applied to statistical problems and we discuss relations to the approach of Donini et al. (2018) in Section 6.5.

6.1 Plug-in estimator

A common method for estimation is the plug-in approach whereby an unknown probability measure is replaced by the empirical measure. This approach is used in (Grünewälder, 2018) for deriving estimators of conditional expectations. To see how the approach can be generalized to our setting, first observe that we can write

$$E^S\phi(X) = g \circ S \quad \textit{almost surely}, \tag{7}$$

where $g : \mathbb{S} \to \mathcal{H}$ is a Bochner-measurable function (see Appendix A and Lemma 2 for details). Our aim is to estimate this function g from i.i.d. observations $\{(X_i, S_i)\}_{i\leq n}$. For any subset B of the range space $\mathbb{S}$ of the sensitive features define the empirical measure $P_n(S \in B) = (1/n)\sum_{i=1}^n \delta_{S_i}(B)$, where δ_{S_i} the Dirac measure with mass one at location S_i. We define an estimate of the conditional expectation of $\phi(X)$ given that the sensitive variable falls into a set B by

$$E_n(\phi(X)|S \in B) = \frac{1}{nP_n(S \in B)} \sum_{i=1}^n \phi(X_i) \times \delta_{S_i}(B),$$

when $P_n(S \in B) > 0$ and through $E_n(\phi(X)|S \in B) = 0$ otherwise. Observe that for $h \in \mathcal{H}$ we have,

$$\Big\langle h, \frac{1}{nP_n(S \in B)} \sum_{i=1}^n \phi(X_i) \times \delta_{S_i}(B)\Big\rangle = \frac{1}{nP_n(S \in B)} \sum_{i=1}^n h(X_i) \times \delta_{S_i}(B).$$

We can also write this as $\langle h, E_n(\phi(X)|S \in B)\rangle = E_n(h(X)|S \in B)$. An estimate of the conditional expectation given S is provided by

$$E_n^S\phi(X) = \sum_{B\in\wp_S} E_n(\phi(X)|S \in B) \times \chi\{S \in B\},$$

where $\wp_S$ is a finite partition of the range space $\mathbb{S}$ of S. A common choice for $\wp_S$ if $\mathbb{S}$ is the hypercube $[0, 1]^d$, $d \geq 1$, are the dyadic sets. Observe, that we can move inner products inside the conditional expectation $E_n^S\phi(X)$ so that $\langle h, E_n^S\phi(X)\rangle = E_n^S h(X)$, where $E_n^S h(X)$ is the empirical conditional expectation introduced in (Grünewälder, 2018).

6.2 Controlling the estimation error

The estimation error when estimating $E^S\phi(X)$ using $E_n^S\phi(X)$ is relatively easy to control thanks to the plug-in approach. Essentially, standard results concerning the empirical measure carry over to conditional expectation estimates in the real-valued case (Grünewälder, 2018). But through scalarization we can transfer some of these results straight away to the Hilbert-space-valued case. For instance, using $\phi(X)$ in place of $\phi(X)$,

$$\begin{aligned}
&\|E_n(\phi(X)|S \in B) - E(\phi(X)|S \in B)\| \\
&= \sup_{\|h\|\leq 1} |\langle E_n(\phi(X)|S \in B) - E(\phi(X)|S \in B), h\rangle| \\
&= \sup_{\|h\|\leq 1} |E_n(h(X)|S \in B) - E(h(X)|S \in B)|
\end{aligned}$$

and bounds on the latter term are known. Similarly,

$$\|E_n^S\phi(X) - E^S\phi(X)\| = \sup_{\|h\|\leq 1} |E_n^S h(X) - E^S(h(X))|. \tag{8}$$

However, both $E_n^S\phi(X)$ and $E^S\phi(X)$ are random variables and a useful measure of their difference is the $\mathcal{L}^2$-pseudo-norm. The $\mathcal{L}^2$-pseudo-norm should in this case not be taken with respect to P itself but conditional on the training sample. Hence, for i.i.d. pairs $(X, S), (X_1, S_1), \ldots, (X_n, S_n)$ let $\mathcal{F}_n = \sigma(X_1, S_1, \ldots, X_n, S_n)$ and define the 'conditional' $\mathcal{L}^2$-pseudo-norm by

$$\|E_n^S\phi(X) - E^S\phi(X)\|_{2,n}^2 = E^{\mathcal{F}_n}\|E_n^S\phi(X) - E^S\phi(X)\|^2.$$

Substituting Equation (8) in shows that this expression is equal to

$$E^{\mathcal{F}_n}\Big(\sup_{\|h\|\leq 1} |E_n^S h(X) - E^S h(X)|^2\Big).$$

The supremum cannot be taken out of the conditional expectation, however, by writing $E_n^S h(X)$ and $E^S h(X)$ as simple functions (see Appendix A.1) we can get around this difficulty and control the error in $\|\cdot\|_{2,n}$. We demonstrate this in the following by deriving rates of convergence for two cases: for the case where $\mathbb{S}$ is finite, and for the case where $\mathbb{S}$ is the unit cube in $\mathbb{R}^d$ for some $d \geq 1$ and S has a density that is bounded away from zero.

To derive these rates we rely, among other things, on the convergence of the empirical process uniformly over families of functions related to the unit ball of $\mathcal{H}$ and partitions of $\mathbb{S}$. For instance, in the case where $\mathbb{S}$ is finite we need to assume that

$$\mathcal{H}_{\mathbb{S}} := \{(h \circ \pi_1) \times \chi(\mathbb{X} \times \{s\}) : h \in \mathcal{H}, \|h\| \leq 1, s \in \mathbb{S}\},$$

as a family of real-valued functions on $\mathbb{X} \times \mathbb{S}$, is a P-Donsker class. The function $\pi_1 : \mathbb{X} \times \mathbb{S} \to \mathbb{X}$ is here the projection onto the first argument, i.e. $\pi_1(x, s) = x$. For the definition of *P-Donsker classes* see (Dudley, 2014; Giné and Nickl, 2016).

There are various ways to verify this condition in concrete settings. For example, if $\mathcal{H}$ is a finite dimensional RKHS then $\mathcal{H}_{\mathbb{S}}$ is a P-Donsker class under a mild measurability assumption. This follows from a few simple arguments: any finite dimensional space of functions is a VC-subgraph class (Giné and Nickl, 2016, Ex.3.6.11); this implies directly that $\{(h \circ \pi_1) \times \chi(\mathbb{X} \times \{s\}) : h \in \mathcal{H}, \|h\| \leq 1\}$ is a VC-subgraph class for every $s \in \mathbb{S}$. Furthermore, finite unions of VC-subgraph classes are again a VC-subgraph class; under a mild measurbility assumption it follows now from (Dudley, 2014, Cor.6.19) that $\mathcal{H}_{\mathbb{S}}$ is a P-Donsker class.

There are obviously other ways to prove this statement. In particular, one might use that the unit ball of $\mathcal{H}$ is a *universal Donsker class* (see (Dudley, 2014; Giné and Nickl, 2016) for details) when the kernel function is continuous and $\mathbb{X}$ is compact (this also holds when $\mathcal{H}$ is infinite dimensional): due to Marcus (1985) the unit ball of a Hilbert space is a universal Donsker class if $\sup_{x\in\mathbb{X}} |h(x)| \leq c\|h\|$ for some constant c that does not depend on h. If the kernel function is bounded $c = \sqrt{k(x, x)}$ witnesses that this property holds.

Case 1: finitely many sensitive features. Our first proposition states that the estimator converges with the optimal rate $n^{-1/2}$ when $\mathbb{S}$ is finite and $\mathcal{H}_{\mathbb{S}}$ is a P-Donsker class.

Proposition 5 *Given a finite space $\mathbb{S}$ and a P-Donsker class $\mathcal{H}_{\mathbb{S}}$, it holds that*

$$\|E_n^S\phi(X) - E^S\phi(X)\|_{2,n} \in O_P^*(n^{-1/2}).$$

The proof is given in Appendix C.5.

Case 2: $[0,1]^d$-valued sensitive features. We extend Proposition 5 to the case where S is not confined to taking finitely many values. In order to state the result, we introduce the following notation. Set $\mathbb{S} := [0,1]^d$ for some $d \in \mathbb{N}$ and let $g : \mathbb{S} \to \mathcal{H}$ be such that with probability one $E^S\phi(X) = g \circ S$ (which is possible by Lemma 2). Consider a discretization of $\mathbb{S}$ into dyadic cubes $\Delta_1, \Delta_2, \ldots, \Delta_{\ell^d}$ of side-length $1/\ell$ for some $\ell \in \mathbb{N}$. Define $\mathfrak{C}_\ell := \{\mathbb{X} \times \Delta : \Delta \in \mathfrak{D}_\ell\}$ and let $\mathcal{H}_\mathfrak{C} := \{h \times \chi D : h \in \mathcal{H},\ \|h\| \leq 1,\ D \in \bigcup_{\ell \in \mathbb{N}} \mathfrak{C}_\ell\}$.

Proposition 6 *Suppose that the push forward measure $\mu := PS^{-1}$ has density u with respect to the Lebesgue measure λ on $\mathbb{S}$ with the property that $\inf_{s\in\mathbb{S}} u(s) \geq b$ for some $b > 0$. Assume that $g \circ S$ is L-lipschitz continuous and that $\mathcal{H}_\mathfrak{C}$ is a P-Donsker class. We have*

$$\|E_n^S\phi(X) - E^S\phi(X)\|_{2,n} \in O_P^*(n^{-\frac{1}{2(d+1)}}).$$

The proof is given in Appendix C.6.

6.3 Generating an oblivious random variable

Given a data-point (X, S) composed of non-sensitive and sensitive features X and S respectively, we can generate an *oblivious* random variable $\boldsymbol{Z}$ as

$$\boldsymbol{Z} := \phi(X) - E_n^S\phi(X) + E_n(\phi(X)). \tag{9}$$

Most kernel methods work with the kernel matrix and do not need access to the features themselves. The same holds in our setting. More specifically, we never need to represent $\boldsymbol{Z}$ explicitly in the Hilbert space but only require inner-product calculations. In order to calculate the empirical estimates of the conditional expectation $E_n^S\phi(X)$ and of $E_n(\phi(X))$ in (9) we consider a simple approach whereby we split the training set into two subsets of size n, and use half the observations to obtain the empirical estimates of the expectations. The remaining n observations are used to obtain an *oblivious* predictor; we have two cases as follows.

Case 1 (M-Oblivious). The standard kernel matrix K is calculated with the remaining n observations and a kernel-method is applied to K to obtain a predictor g. When applying the predictor to a new unseen data-point (X, S) we first transform X into $\boldsymbol{Z}$ via (9) and calculate the prediction as $\langle g, \boldsymbol{Z}\rangle$. As discussed in the Introduction, we conjecture that this approach is suitable in the case where the labels Y are conditionally independent of the sensitive features S given the non-sensitive features X, i.e. when S, X, Y form a Markov chain $S \to X \to Y$. As such we call this approach M-Oblivious.

Case 2 (Oblivious). Instead of calculating the kernel matrix K an *oblivious kernel matrix*, i.e.

$$\mathcal{O} = \begin{pmatrix} \|\boldsymbol{Z}_1\|^2 & \cdots & \langle \boldsymbol{Z}_1, \boldsymbol{Z}_n\rangle \\ \vdots & \ddots & \vdots \\ \langle \boldsymbol{Z}_n, \boldsymbol{Z}_1\rangle & \cdots & \|\boldsymbol{Z}_n\|^2 \end{pmatrix}, \tag{10}$$

is calculated by applying Equation (9) to the remaining training samples (X_i, S_i) before taking inner products. The oblivious matrix is then passed to the kernel-method to gain a predictor g. The matrix is positive semi-definite since $\mathbf{a}^\top\mathcal{O}\mathbf{a} = \|\sum_{i=1}^n a_i\boldsymbol{Z}_i\|^2 \geq 0$, for any $\mathbf{a} \in \mathbb{R}^n$. The complexity to compute the matrix is $O(n^2)$ (see Appendix E for details on the algorithm). Prediction for a new unseen data-point (X, S) is now done in the same way as in Case 1.

6.4 Oblivious ridge regression

In this section we showcase our approach in the context of kernel ridge regression. We have three relevant random variables, namely, the non-sensitive features X, the sensitive features S and labels Y which are real valued. We assume that we have $2n$ i.i.d. observations $\{(X_i, S_i, Y_i)\}_{i\leq 2n}$. We use the observations $n+1, \ldots, 2n$ to generate the oblivious random variables $\mathbf{Z}_i$ and then use oblivious data $\{(\mathbf{Z}_i, Y_i)\}_{i\leq n}$ for oblivious ridge regression (ORR).

The ORR problem has the following form. Given a positive definite kernel function $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$, a corresponding RKHS $\mathcal{H}$ and oblivious features $\mathbf{Z}_i$. Our aim is to find a regression function $h \in \mathcal{H}$ such that the mean squared error between $\langle h, \boldsymbol{Z}\rangle$ and Y is small. Replacing the mean squared error by the empirical least-squares error and adding a regularization term for h gives us the optimization problem

$$\hat{h} = \operatorname*{arg\,min}_{h\in\mathcal{H}} \sum_{i=1}^{n} (\langle h, \mathbf{Z}_i\rangle - Y_i)^2 + \lambda \|h\|^2, \tag{11}$$

where $\lambda > 0$ is the regularization parameter.

It is easy to see that the setting is not substantially different from standard kernel ridge regression and derive a closed form solution for $\hat{h}$. More specifically, we have a representer theorem in this setting which tells us that the minimizer lies in the span of $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$. One can then solve the optimization problem in the same way as for standard kernel ridge regression, see Appendix D for details. The solution to the optimization problem is $\hat{h} = \sum_{i=1}^{n} \alpha_i \boldsymbol{Z}_i$, where $\boldsymbol{\alpha} = (\mathcal{O} + \lambda I)^{-1}\boldsymbol{y}$. The vector $\boldsymbol{y}$ is given by $(Y_1, \ldots, Y_n)^\top$. Predicting Y for a new observation (X, S) is achieved by first generating the oblivious features $\mathbf{Z}$ (see Appendix E.2) and then by evaluating $\langle \mathbf{Z}, \hat{h}\rangle = \sum_{i=1}^{n} \alpha_i \langle \mathbf{Z}, \mathbf{Z}_i\rangle$.

6.5 Comparison to (Donini, Oneto, Ben-David, Shawe-Taylor, and Pontil, 2018)

Our focus in this paper is on generating features that are less dependent on the sensitive features than the original non-sensitive features. However, the conditional expectation $E^S\phi(X)$, which is at the heart of our approach, also features prominently in methods that add constraints to SVM classifiers. In particular, in (Donini et al., 2018) a constraint is used to achieve approximately equal opportunity in classification where the sensitive feature is binary. While their approach does not make explicit use of conditional expectations one can recognize that the key object in their approach (Eq. (13) in (Donini et al., 2018)) is, in fact, closely related to our conditional expectation when used in the case where S can attain only two values (say $\mathbb{S} = \{0, 1\}$). In detail, the optimization problem (14) is constraint by enforcing for a given $\epsilon > 0$ that the solution $h^* \in \mathcal{H}$ fulfills

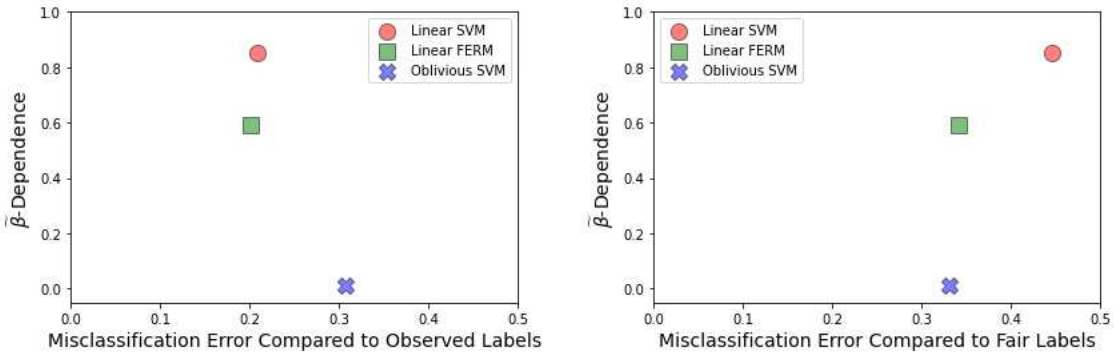
$$|E_n(h^*(X)|S = 0) - E_n(h^*(X)|S = 1)| \leq \epsilon. \tag{12}$$

Considering $\boldsymbol{Z} = \phi(X) - E^S\phi(X) + E(\phi(X))$ we can observe right away that in this setting for all $h \in \mathcal{H}$,

$$E(\langle h, \boldsymbol{Z}\rangle|S = 0) = E(\langle h, \boldsymbol{Z}\rangle|S = 1).$$

To see this observe that $E^S\langle h, \boldsymbol{Z}\rangle$ is almost surely equal to the $E(h(X))$. In other words

$$E(\langle h, \boldsymbol{Z}\rangle|S = 0) \times \chi\{S = 0\} + E(\langle h, \boldsymbol{Z}\rangle|S = 1) \times \chi\{S = 1\} = E^S\langle h, \boldsymbol{Z}\rangle$$



(a) (b)

Figure 3: Binary classification error vs. $\widetilde{\beta}$-dependence between prediction and sensitive features is shown for three different methods: classical Linear SVM, Linear FERM, and Oblivious SVM. In Figure 3a the error is calculated with respect to the observed labels which are intrinsically biased and in Figure 3a the error is calculated with respect to the true *fair* classification rule.

is almost surely constant. Unless $P(S = 0) = 0$ or $P(S = 1) = 0$ this implies that $E(\langle h, \boldsymbol{Z}\rangle | S = 0) = E(\langle h, \boldsymbol{Z}\rangle | S = 1)$. Hence, for the max-margin classifier h^* and $\boldsymbol{Z}$ it holds that $E(\langle h^*, \boldsymbol{Z}\rangle | S = 0) = E(\langle h^*, \boldsymbol{Z}\rangle | S = 1)$ and on the population level our new features $\boldsymbol{Z}$ guarantee that constraint (12) is automatically fulfilled.

7. Empirical evaluation

In this section we report our experimental results for classification and regression. Our objective in the classification experiment is to point out an important property of supervised learning problems where sensitive features affect both the non-sensitive features and the labels: the estimation error of the observed labels can be misleading as a quality measure. The aim is much rather to predict values in an unbiased fashion. The first experiment highlights this difference by considering a synthetic data set for which we know the unbiased labels (though the true unbiased labels are not available to the methods). We measure the dependencies between the predicted values and the sensitive features, and compare against a standard SVM and to FERM. The second set of experiments aims to investigate how dependencies between sensitive and non-sensitive features affect ORR and M-ORR. We are investigating this relationship by considering a family of synthetic problems for which we can adjust the dependency between the features using a parameter γ. In this set of experiments we are also concerned with clarifying the relationship between ORR and M-ORR, where the latter is the M-Oblivious version of KRR, see Section 6.3. Our implementation can be found at the following repository: `https://github.com/azalk/Oblivious.git`.

7.1 Binary Classification

We carried out an experiment to mimic a scenario where a class of students should normally receive grades between 0 and 5, and anyone with a grade above a fixed threshold $\theta = 2$ should pass. Half of the class, representing a "minority group", are disadvantaged in that their grades are almost systematically reduced, while the other half receive a boost on average. More specifically, let the

sensitive feature S be a $\{0,1\}$-valued Bernoulli random variable with parameter 0.5, and let X_0 be distributed according to a truncated normal distribution with support $[1,4]$. Let the non-sensitive feature X, representing a student's grade, be given by

$$X := (X_0 - B)\chi\{S=0\} + (X_0 + B)\chi\{S=1\}$$

where B is a Bernoulli random variable with parameter 0.9 independent of X_0 and of S. The label Y is defined as a noisy decision influenced by the student's "original grade" X_0 prior to the S-based modification. More formally, let U be a random variable independent of X_0 and of S, and uniformly distributed on $[0,1]$. Let $Y_0 := \chi\{U \geq X_0\}$ and define

$$Y := Y_0\chi\{X + S \geq \theta\}.$$

Classification Error. In a typical classification problem, the labels Y depend on both X and S so when we remove the bias it is not clear what we should compare against when calculating the classification performance. Observe that our experimental construction here allows access to the true ground-truth labels

$$Y^* := \chi\{X_0 \geq \theta\}. \tag{13}$$

Therefore, we are able to calculate the true (unbiased) errors as well. However, this is not always the case in practice. In fact, we argue that the question of how to evaluate fair classification performance is an important open problem which has yet to be addressed.

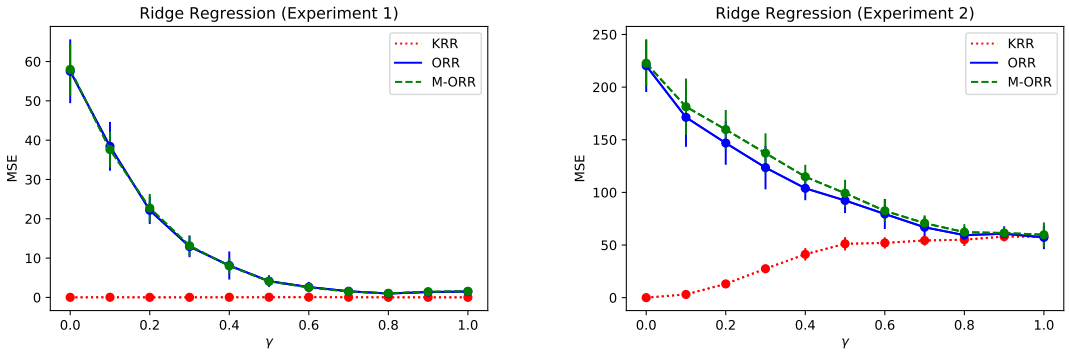
Measure of Dependence. Let $\mathcal{F}_n := \sigma(X_1, \ldots, X_n, S_1, \ldots, S_n),\ n \in \mathbb{N}$ be the σ-algebra generated by the training samples. In this experiment, we measure the dependence between the predicted labels $\widehat{Y}$ produced by any algorithm and the sensitive features S as

$$\widetilde{\beta}(\widehat{Y}, S) := \frac{1}{2} \sum_{s\in\{0,1\}} \sum_{y\in\{0,1\}} E\left| P(\widehat{Y} = y, S = s|\mathcal{F}_n) - P(\widehat{Y} = y|\mathcal{F}_n)P(S = s)\right| \tag{14}$$

which is closely related to the β-dependence (see, e.g. (Bradley, 2007, vol. I, p. 67)) between their respective σ-algebras. We obtain an empirical estimate of $\widetilde{\beta}(\sigma(\widehat{Y}), \sigma(S))$ by simply replacing the probabilities in (14) with corresponding empirical frequencies.

Synthetic Experiment. We generated $n = 1000$ training and test samples as described above and the errors reported for each experiment are averaged over 10 repetitions. Figure 3 shows binary classification error vs. dependence between prediction and sensitive features for three different methods: classical Linear SVM, Linear FERM, and Oblivious SVM. In Figure 3a the error is calculated with respect to the observed labels which are intrinsically biased and in Figure 3a the error is calculated with respect to the true *fair* classification rule Y^* given by (13). As can be seen in the plots, the *true* classification error of Oblivious SVM is smaller than that of the other two methods. Moreover, in both plots the β-dependence between the predicted labels produced by Oblivious SVM and the sensitive feature is close to 0 and is much smaller than that of the other two methods.

Real-World Experiment. We evaluated the performance of our method on the so-called "Adult" dataset, which is a benchmark dataset publicly available on the UCI Machine Learning Repository (Dua and Graff, 2017). The dataset consists of more than 40K data-points, each composed of 14 features including an individual's age, education, marital status, gender, occupation, and capital



(a) (b)

Figure 4: Plots 4a and 4b correspond to Ridge Regression Experiments 1 and 2 respectively. In both plots, the performance of three estimators (KRR,ORR and M-ORR) is plotted against γ, where γ controls the dependence of X on S. The case of $\gamma = 0$ corresponds to the highest dependence while $\gamma = 1$ corresponds to the case in which X is independent of S.

gain/loss. The objective is to predict whether an individual's income exceeds \$50K per year. Table 1 outlines the result of our comparison against standard SVM and the FERM algorithm of Donini et al. (2018), with gender marked as a sensitive feature. We used the repository provided by Donini et al. (2018) to extract the data and run the experiments corresponding to both standard SVM and FERM. The "smaller" option was used in data extraction, giving a total of 1628 training and 12661 testing data-points respectively. Both linear and Gaussian (RBF) kernels were considered. The Gaussian kernel was parametrized by $\gamma > 0$, i.e. $k(x, x') = e^{-\gamma\|x-x'\|^2}$. The hyperparameters were selected using a 5-fold cross-validation. The SVM regularization parameter C was varied between $2^{-4}, 2^{-3}, \ldots, 2^4$, and γ was selected from $\{0.001, 0.01, 0.1, 1\}$. As can be observed in the table, our Oblivious (linear and non-linear) SVM reduces $\widetilde{\beta}$-dependence between the label and the sensitive feature as desired, while incurring a minimal increase in Mean Prediction Error. On the other hand, FERM achieves the lowest DEO (see Donini et al. (2018)) for the linear kernel, while in the case of RBF kernel our model provides the highest reduction in DEO.

7.2 Ridge-Regression

In this section we compare ORR with KRR and the 'Markov' version of ORR, M-ORR, which applies the KRR solution to oblivious test features Z. We use an RBF kernel with $\sigma = 1$. We are particularly interested in how the dependence of S on X affects the performance and in a comparison of ORR to M-ORR. We use synthetic data to be able to control the dependence between S and X. The basic data generating process is as follows. Sensitive features S and non-sensitive features U are sampled independently from a uniform distribution with support $[-5, 5]$. The features X are a convex combination of these two of the form $X = \gamma U + (1-\gamma)S$, $\gamma \in [0, 1]$. We consider two ways to generate the response variable Y. In *Experiment 1*, the response variable is $Y = X^2 + \epsilon$, where ϵ is normally distributed with variance 0.1 and is independent of U and S. In this case $S \to X \to Y$ forms a Markov chain and we expect M-ORR to do well. In *Experiment 2*, the variable S influences

Algorithm	Mean Prediction Error	DEO	$\widetilde{\beta}$-Dependence	Relative $\widetilde{\beta}$-Dependence
SVM (Linear)	0.20	0.05	0.06	100%
FERM (Linear)	0.20	0.03	0.07	116%
Oblivious (Linear)	0.21	0.05	0.02	33%
SVM (RBF)	0.17	0.22	0.14	100%
FERM (RBF)	0.17	0.12	0.16	114%
Oblivious (RBF)	0.19	0.03	0.06	43%

Table 1: Comparison against standard SVM and FERM of Donini et al. (2018) on "Adult" benchmark dataset (Dua and Graff, 2017), with gender marked as sensitive feature. Our algorithm reduces the $\widetilde{\beta}$-dependence between the label and the sensitive feature by 67% in the case of linear SVM and by 57% for non-linear SVM, while only incurring a minimal increase in Mean Prediction Error in both cases. FERM achieves the lowest DEO for the linear kernel, while our model provides the highest DEO reduction in the case of RBF kernel.

Y also directly and not only through X, i.e. $Y = X^2 + S^2 + \epsilon$. We use here S^2 instead of S because S^2 is not a zero mean random variable and cannot simply be consumed into the noise term.

Figure 4 shows the results of these experiments. In these experiments, γ varies between $0, 0.1, 0.2 \ldots, 1$. For each value of γ we generate 500 data points for ORR and M-ORR to infer the conditional expectations and further 500 data points are used by all three methods to calculate the ridge regression solution. For simplicity, we fixed a partition for the conditional expectation: the set $\mathbb{S} = [-5, 5]$ is split into a dyadic partition consisting of 16 sets. Each method uses a validation set of 100 data points (which are different from the 500 training data points) to select the regularization parameter λ from $2^{-5}, 2^{-4}, \ldots, 2^5$. A test set of size 100 is used to calculate the mean squared error (MSE). For each γ the experiment is repeated 20 times. Figure 4 reports the average MSE and the standard deviation of the MSE over these 20 experiments.

We make the following observations from Figure 4a. KRR is the best estimator as it uses the features X directly and not the new features $\boldsymbol{Z}$. As $\gamma \to 1$ both the ORR and M-ORR estimators approach the KRR estimator since the effect of S on X vanishes. Both estimators do not quite reach the performance of the KRR estimator. This is due to the additional uncertainty introduced by estimating the conditional expectations. By definition, the ORR estimator will achieve the best fit of the training data given the new features $\boldsymbol{Z}$. We can observe that the M-ORR estimator is performing as well as the ORR estimator even though the M-ORR estimator uses the KRR solution and applies it to $\boldsymbol{Z}$. This is due to the fact that $S \to X \to Y$ forms a Markov chain. Finally, when $\gamma = 0$ both the M-ORR and ORR estimator achieve an MSE that is very close to the best MSE that can be achieved by a regressor that generates values which are independent of S: assume that some new features Z' are given which are a function of X and are independent of S. when $\gamma = 0$ this random variable Z' can only be independent of S if Z' is a constant. However, if Z' is a constant then the ridge-regressor using Z' is also a constant and the MSE $E(Y - c)^2 = E(S^4) - 2cE(S^2) + c^2 + 0.1$ is minimized for $c = E(S^2)$. The minimal value is approximately 56.2 which is very close to the values of the ORR and M-ORR estimator.

Figure 4b shares a few characteristics with Figure 4a as follows. For $\gamma = 0$ both M-ORR and ORR attain an MSE that is close to the best possible (in the above sense) which is approximately equal to 224.8. As before, KRR is the overall best estimator and ORR is the best estimator using features $\boldsymbol{Z}$. Furthermore, as $\gamma \to 1$ both estimators become close to the KRR solution. A crucial difference in this experiment is that S, X, Y does not form a Markov chain anymore and the performance of M-ORR is worse than that of ORR for values of γ between 0.2 and 0.8. The performance of M-ORR and ORR is essentially the same for $\gamma = 0$ and $\gamma = 1$. This is not surprising given that when $\gamma = 0$ then $Y = 2S^2 + \epsilon$ and we are back in the Markov chain setting, while when $\gamma = 1$ then X is already independent of S.

8. Discussion

We have introduced a novel approach to derive oblivious features which approximate non-sensitive features well while maintaining only minimal dependence on sensitive features. We make use of Hilbert-space-valued conditional expectations and estimates thereof; our plug-in estimators in this case can be of independent interest in future research, and in turn open grounds for interesting questions involving their guarantees. The application of our approach to kernel methods is facilitated by an oblivious kernel matrix which we have derived to be used in place of the original kernel matrix. We characterize the dependencies between the oblivious and the sensitive features in terms of how 'close' the sensitive features are to the low-dimensional manifold $\phi[\mathbb{X}]$. One may wonder if this relation can be exploited to further reduce dependencies, and potentially achieve complete independence. Another important question concerns the interplay between the estimation errors introduced by estimating conditional expectations and the estimation errors introduced by kernel methods which are applied to the oblivious data.

Appendix A. Probability in Hilbert spaces: elementary results

In this section we summarize a few elementary results concerning random variables that attain values in a separable Hilbert space which we use in the paper.

A.1 Measurable functions

There are three natural definitions of what it means for a function $f : \Omega \to \mathcal{H}$ to be measurable. Denote the measure space in the following by $(\Omega, \mathcal{A})$ with the understanding that these definitions apply, in particular, to $\Omega = \mathbb{R}^d$ and $\mathcal{A}$ being the corresponding Borel σ-algebra.

1. f is *Bochner-measurable* iff f is the point-wise limit of a sequence of simple functions, where $\boldsymbol{S} : \Omega \to \mathcal{H}$ is a simple function if it can be written as

$$\boldsymbol{S}(\omega) = \sum_{i=1}^{n} h_i \times \chi A_i(\omega)$$

   for some $n \in \mathbb{N}$, $A_1, \ldots, A_n \in \mathcal{A}$ and $h_1, \ldots, h_n \in \mathcal{H}$.

2. f is *strongly-measurable* iff $f^{-1}[B] \in \mathcal{A}$ for every Borel-measurable subset B of $\mathcal{H}$. The topology that is used here is the norm-topology.

3. f is *weakly-measurable* iff for every element $h \in \mathcal{H}$ the function $\langle h, f\rangle : \Omega \to \mathbb{R}$ is measurable in the usual sense (using the Borel-algebra on $\mathbb{R}$).

All three definitions of measurability are equivalent in our setting. We call a function $f : \Omega \to \mathcal{H}$ a *random variable* if it is measurable in this sense.

The main example in our paper is $f = \phi(X)$. This is a well defined random variable whenever $X : \Omega \to \mathbb{R}^d$ is a random variable and $\phi : \mathbb{R}^d \to \mathcal{H}$ is Bochner-measurable. In particular, when the kernel function is continuous ϕ is Bochner-measurable.

Appendix B. Hilbert space-valued conditional expectations

B.1 Basic properties

We recall a few important properties of Hilbert space valued conditional expectations. These often follow from properties of real-valued conditional expectations through 'scalarization' (Pisier, 2016). In the following, let $\boldsymbol{X}, \boldsymbol{Z} \in \mathcal{L}^2(\Omega, \mathcal{A}, P; \mathcal{H})$ and $\mathcal{B}$ some σ-subalgebra of $\mathcal{A}$. Due to Pisier (2016)[Eq. (1.7)], for any $f \in \mathcal{H}$

$$\langle f, E^{\mathcal{B}}\boldsymbol{X}\rangle = E^{\mathcal{B}}\langle f, \boldsymbol{X}\rangle \qquad \text{(a.s.)} \tag{15}$$

and the right hand side is just the usual real-valued conditional expectation. It is also worth highlighting that the same holds for the Bochner-integral $E(\boldsymbol{X})$, i.e. for any $f \in \mathcal{H}$, $\langle f, E(\boldsymbol{X})\rangle = E\langle f, \boldsymbol{X}\rangle$. This can be used to derive properties of $E^{\mathcal{B}}\boldsymbol{X}$. For instance, since $E(E^{\mathcal{B}}\langle f, \boldsymbol{X}\rangle) = E\langle f, \boldsymbol{X}\rangle$ is a property of real-valued conditional expectations we find right away that

$$\langle f, E(\boldsymbol{X})\rangle = E\langle f, \boldsymbol{X}\rangle = E(E^{\mathcal{B}}\langle f, \boldsymbol{X}\rangle) = E\langle f, E^{\mathcal{B}}\boldsymbol{X}\rangle = \langle f, E(E^{\mathcal{B}}\boldsymbol{X})\rangle.$$

Because $E(\boldsymbol{X})$ and $E(E^{\mathcal{B}}\boldsymbol{X})$ are elements of $\mathcal{H}$ and for all $f \in \mathcal{H}$

$$\langle f, E(\boldsymbol{X}) - E(E^{\mathcal{B}}\boldsymbol{X})\rangle = 0$$

it follows that $E(\boldsymbol{X}) = E(E^{\mathcal{B}}\boldsymbol{X})$.

Another result we need is that if $\boldsymbol{Z}$ is $\mathcal{B}$-measurable then

$$E^{\mathcal{B}}\langle \boldsymbol{X}, \boldsymbol{Z}\rangle = \langle E^{\mathcal{B}}\boldsymbol{X}, \boldsymbol{Z}\rangle \qquad \text{(a.s.)}.$$

Showing this needs a bit more work. Since $\boldsymbol{Z} \in \mathcal{L}^2(\Omega, \mathcal{B}, P; \mathcal{H})$ there exist $\mathcal{B}$-measurable simple functions U_n such that U_n converges point-wise to $\boldsymbol{Z}$, $\lim_{n\to\infty} \|U_n^\bullet - \boldsymbol{Z}^\bullet\|_2 = 0$ and the sequence fulfills $\|U_n\| \leq 3\|\boldsymbol{Z}\|$ for all $n \in \mathbb{N}$ (Pisier, 2016)[Prop.1.2]. Consider some n and write

$$U_n = \sum_{i=1}^{m} h_i \times \chi A_i,$$

for a suitable $m \in \mathbb{N}, h_i \in \mathcal{H}, A_i \in \mathcal{B}$, then

$$\begin{aligned} E^{\mathcal{B}}\langle \boldsymbol{X}, U_n\rangle &= \sum_{i=1}^{m} E^{\mathcal{B}}(\langle \boldsymbol{X}, h_i\rangle \times \chi A_i) \\ &= \sum_{i=1}^{m} (E^{\mathcal{B}}\langle \boldsymbol{X}, h_i\rangle) \times \chi A_i \\ &= \sum_{i=1}^{m} \langle E^{\mathcal{B}}\boldsymbol{X}, h_i\rangle \times \chi A_i \\ &= \langle E^{\mathcal{B}}\boldsymbol{X}, U_n\rangle \qquad \text{(a.s.)}, \end{aligned}$$

because χA_i is $\mathcal{B}$-measurable. For the right hand side point-wise convergence of U_n to $\boldsymbol{Z}$ tells us that for all $\omega \in \Omega$ we have $\lim_{n\to\infty} \|U_n(\omega) - \boldsymbol{Z}(\omega)\| = 0$. Because $E^{\mathcal{B}}\boldsymbol{X}^{\bullet} \in L^2(\Omega, \mathcal{A}, P; \mathcal{H})$ we also know that $E^{\mathcal{B}}\boldsymbol{X}$ is finite almost surely. Therefore, for ω in the corresponding co-negligible set,

$$\lim_{n\to\infty} |\langle (E^{\mathcal{B}}\boldsymbol{X})(\omega), U_n(\omega)\rangle - \langle (E^{\mathcal{B}}\boldsymbol{X})(\omega), \boldsymbol{Z}(\omega)\rangle| \leq \lim_{n\to\infty} \|(E^{\mathcal{B}}\boldsymbol{X})(\omega)\| \|U_n(\omega) - \boldsymbol{Z}(\omega)\| = 0$$

and $\lim_{n\to\infty}\langle E^{\mathcal{B}}\boldsymbol{X}, U_n\rangle = \langle E^{\mathcal{B}}\boldsymbol{X}, \boldsymbol{Z}\rangle$ almost surely.

By the same argument it follows that $\lim_{n\to\infty}\langle \boldsymbol{X}, U_n\rangle = \langle \boldsymbol{X}, \boldsymbol{Z}\rangle$ almost surely. Let $h_n = \langle \boldsymbol{X}, U_n\rangle$ and $h = \langle \boldsymbol{X}, \boldsymbol{Z}\rangle$ then $|h_n - h| \leq \|\boldsymbol{X}\| \|U_n - \boldsymbol{Z}\| \leq 3\|\boldsymbol{X}\| \|\boldsymbol{Z}\|$. Furthermore, $|h_n| \leq |h| + 3\|\boldsymbol{X}\| \|\boldsymbol{Z}\| \leq 4\|\boldsymbol{X}\| \|\boldsymbol{Z}\| \leq 4(\|\boldsymbol{X}\|^2 + \|\boldsymbol{Z}\|^2)$. The right hand side lies in $\mathcal{L}^1$ and dominates h_n. Using Shiryaev (1989)[II.§7.Thm.2(a)], we conclude that

$$\lim_{n\to\infty} E^{\mathcal{B}}\langle \boldsymbol{X}, U_n\rangle = E^{\mathcal{B}}\langle \boldsymbol{X}, \boldsymbol{Z}\rangle \quad \text{(a.s.)}$$

and the result follows.

The operator $E^{\mathcal{B}}$ is also idempotent and self-adjoint, i.e.

$$E^{\mathcal{B}}\boldsymbol{X} = E^{\mathcal{B}}(E^{\mathcal{B}}\boldsymbol{X}) \text{ (a.s)} \quad \text{and} \quad \langle \boldsymbol{X}^{\bullet}, E^{\mathcal{B}}\boldsymbol{Z}^{\bullet}\rangle_2 = \langle E^{\mathcal{B}}\boldsymbol{X}^{\bullet}, \boldsymbol{Z}^{\bullet}\rangle_2.$$

B.2 Representation of conditional expectations

A well known result in probability theory states that a conditional expectation $E^S X$ of a real-valued random variable X given another real-valued random variable S can be written as $g(S)$ with some suitable measurable function $g : \mathbb{R} \to \mathbb{R}$. This result generalizes to our setting. Here, we include the generalized result together with a short proof for reference.

Lemma 1 *Consider a probability space $(\Omega, \mathcal{A}, P)$, and let $\mathcal{H}$ be a separable Hilbert space. Let $S : \Omega \to \mathbb{R}^d$ be a random variable and suppose that $\eta : \Omega \to \mathcal{H}$ is a $\sigma(S)$-measurable function. There exists a Bochner-measurable function $g : \mathbb{R}^d \to \mathcal{H}$ such that*

$$\eta = g \circ S \quad \textit{almost surely.}$$

Proof We first show the statement for simple functions, and observing that any arbitrary Bochner-measurable function can be written as the point-wise limit of a sequence of simple functions, we extend the result to arbitrary η.

First, assume that $\eta := h\chi A$ for some $h \in \mathcal{H}$ and $A \in \sigma(S)$. Since S is measurable with respect to $\mathcal{B}(\mathbb{R}^d)$ there exists some $B \in \mathcal{B}(\mathbb{R}^d)$ such that $\{\omega : S(\omega) \in B\} = A$. Define $g : \mathbb{R}^d \to \mathcal{H}$ as $g := h\tilde{\chi}B$, where $\tilde{\chi}$ denotes the indicator function on $\mathbb{R}^d$. We obtain, $\eta(\omega) = h\chi A(\omega) = h\tilde{\chi}B(S(\omega))$ so that $\eta = g \circ S$. Next, let $\eta := \sum_{i=1}^m h_i\chi A_i$ for some $m \in \mathbb{N},\ h_1, \ldots, h_m \in \mathcal{H}$ and $A_1, \ldots, A_m \in \sigma(S)$. As above, by measurability of S, there exists a sequence $B_1, \ldots, B_m \in \mathcal{B}(\mathbb{R}^d)$ such that $A_i = S^{-1}[B_i],\ i \in 1, \ldots, m$. It follows that $\eta(\omega) = \sum_{i=1}^m h_i\chi A_i(\omega) = \sum_{i=1}^m h_i\tilde{\chi}B_i(S(\omega)),\ \omega \in \Omega$; hence, $\eta = g \circ S$ for $g = \sum_{i=1}^m h_i\tilde{\chi}B_i$. Observe that in both cases g is trivially Bochner-measurable by construction, since it is a simple function.

Let $\eta : \Omega \to \mathcal{H}$ be an arbitrary Bochner-measurable function that is also measurable with respect to $\sigma(S)$. There exists a sequence of simple functions $\eta_n,\ n \in \mathbb{N}$ such that for every $\omega \in \Omega$ we have

$$\eta(\omega) = \lim_{n\to\infty} \eta_n(\omega).$$

Since each η_n is a simple function, by our argument above, there exists a sequence of Bochner-measurable functions $g_n : \mathbb{R}^d \to \mathcal{H}$ such that $\eta_n = g_n \circ S$ where for each $n \in \mathbb{N}$ the function g_n is simple of the form $g_n = \sum_{i=1}^{m_n} h_{i,n}\tilde{\chi}B_{i,n}$ for some $m_n \in \mathbb{N}$ and a sequence of functions $h_{1,n}, \ldots, h_{m_n,n} \in \mathcal{H}$ and a sequence of Borel sets $B_{1,n}, \ldots, B_{m_n,n} \in \mathcal{B}(\mathbb{R}^d)$.

Denote by $B := \{S(\omega) : \omega \in \Omega\} \subset \mathbb{R}^d$ the image of S, and observe that for each $x \in B$ $\lim_{n\to\infty} g_n(x)$ exists. To see this, note that by construction, for each $x \in B$ we have $x = S(\omega)$ for some $\omega \in \Omega$, thus, it holds that

$$\lim_{n\to\infty} g_n(x) = \lim_{n\to\infty} g_n(S(\omega)) = \lim_{n\to\infty} \eta_n(\omega) = \eta(\omega).$$

Moreover, we have $P(S^{-1}[B]) = P(\{\omega \in \Omega : S(\omega) \in B\}) = P(\Omega) = 1$. Define $g : \mathbb{R}^d \to \mathcal{H}$ as

$$g(x) := \begin{cases} \lim_{n\to\infty} g_n(x) & x \in B \\ 0 & x \notin B \end{cases} \tag{16}$$

Thus, for each $\omega \in \Omega$ with probability 1, we have

$$\begin{aligned} \eta(\omega) &= \lim_{n\to\infty} \eta_n(\omega) \\ &= \lim_{n\to\infty} g_n(S(\omega)) \\ &= g(S(\omega)), \end{aligned}$$

so that $\eta = g \circ S$ almost surely. On the other hand, since by definition, g is the pointwise limit of a sequence of simple functions g_n, it is Bochner-measurable, (see Property 1 in Section A.1) and the result follows. ■

Lemma 2 *Consider a separable Hilbert space $\mathcal{H}$, a probability space $(\Omega, \mathcal{A}, P)$, a Bochner-integrable random variable $\phi(X) : \Omega \to \mathcal{H}$ and a random variable $S : \Omega \to \mathbb{R}^d$. There exists a Bochner-measurable function $g : \mathbb{R}^d \to \mathcal{H}$ such that*

$$E^S\phi(X) = g \circ S \quad \textit{almost surely.}$$

Proof Observing that $E^S\phi(X)$ is a $\sigma(S)$-measurable function from Ω to $\mathcal{H}$, the result readily follows from Lemma 1. ■

Appendix C. Proofs

C.1 Proof of Proposition 1

Proof Let $\mathcal{M} := \phi[\mathbb{X}]$ denote the manifold corresponding to the image of $\mathbb{X}$ under ϕ, equipped with the subspace topology and corresponding Borel σ-algebra $\mathcal{B}(\mathcal{M})$. Define the *metric projection map* $\pi : \mathcal{H} \rightrightarrows \mathcal{M}$ as a multi-valued function such that

$$\pi(g) = \left\{ h \in \mathcal{M} : \|h - g\| = \min_{h' \in \mathcal{M}} \|h' - g\| \right\}. \tag{17}$$

Note that the $\min$ operator in Equation (17) is well-defined since by definition $h = \phi(x)$ for some $x \in \mathbb{X}$, the space $\mathbb{X}$ is compact and ϕ is a continuous function. Observe that π is not a function, but a multi-valued function which assigns to each element $g \in \mathcal{M}$ a subset of $\mathcal{M}$, see, e.g. (Beer, 1993, Section 6.1) for more on this notion.

π maps to non-empty compact subsets of $\mathcal{M}$. For each $g \in \mathcal{H}$, set $f_g(h) = \|h - g\|$ with $h \in \mathcal{M}$, and note that it is a continuous function from $\mathcal{M}$ to $\mathbb{R}$. Let $m(g) := \min_{h \in \mathcal{M}} \|h - g\|$, which, by the above argument, is well-defined, and observe that, since $\{m(g)\}$ is a closed subset of $\mathbb{R}$, then $\pi(g) = f_g^{-1}[\{m(g)\}]$ is a closed subset of $\mathcal{M}$. Since $\mathcal{M}$ is compact as the continuous image of the compact space $\mathbb{X}$ it follows that $\pi(g)$ is compact.

π is upper-semicontinuous. As follows from the standard definition, see, e.g. (Beer, 1993, Definition 6.2.4 and Theorem 6.2.5), the multi-valued function π is said to be upper-semicontinuous at a point $g_0 \in \mathcal{H}$ if for any open subset V of $\mathcal{M}$ such that $\pi(g_0) \subseteq V$ it holds that $\pi(g) \subseteq V$ for each g in some neighbourhood of g_0.[2] To show the upper-semicontinuity of π we proceed as follows. Take $g_0 \in \mathcal{H}$. Let V be an open subset of $\mathcal{M}$ such that $\pi(g_0) \subseteq V$. Denote by $\widetilde{\mathcal{M}} := \mathcal{M} \setminus V$. Note that $\widetilde{\mathcal{M}}$ is compact since it is a closed subset of $\mathcal{M}$ which is in turn compact. Therefore, in much the same way as for $\mathcal{M}$, the $\min$ operator is well-defined for $\widetilde{\mathcal{M}}$, i.e. the minimum $\widetilde{m}(g_0) := \min_{h \in \widetilde{\mathcal{M}}} \|g_0 - h\|$ exists. Moreover, since $\pi(g_0) \subseteq V$ and $V \cap \widetilde{\mathcal{M}} = \emptyset$, it holds that $\widetilde{m}(g_0) > m(g_0)$. Therefore, there exists some $\delta > 0$ such that $\widetilde{m}(g_0) \geq \delta + m(g_0)$. Consider an open ball $B_{g_0}(\delta/3)$ of radius $\delta/3$ around g_0. For every $g \in B_{g_0}(\delta/3)$ and all $h \in \pi(g_0)$ we have

$$\|g - h\| \leq \|g_0 - g\| + \|g_0 - h\| \leq \delta/3 + m(g_0). \tag{18}$$

On the other hand, we have

$$\min_{h' \in \widetilde{\mathcal{M}}} \|g - h'\| \geq \big| \|g - g_0\| - \|g_0 - h'\| \big| \geq m(g_0) + 2\delta/3. \tag{19}$$

This implies that $\pi(g) \cap \widetilde{\mathcal{M}} = \emptyset$ because there are already better candidates (closer to g) in $\pi(g_0)$ which is in turn contained in V and thus does not intersect $\widetilde{\mathcal{M}}$). Hence, it must hold that $\pi(g) \subseteq V$. Finally, since the choice of $g \in B_{g_0}(\delta/3)$ is arbitrary, it follows that for all $g \in B_{g_0}(\delta/3)$ we have $\pi(g) \subseteq V$ and π is upper-semicontinuous.

ϕ is a homeomorphism. To see this, note that ϕ is bijective and continuous since the kernel is positive definite and continuous: it is by definition surjective and it is injective since $\phi(x) = \phi(y)$ for $x \neq y$ would imply that $a_1^2 \|\phi(x)\|^2 - 2a_1 a_2 \langle \phi(x), \phi(y) \rangle + a_2^2 \|\phi(y)\|^2 = 0$ when $a_1 = a_2 = 1$. The statement follows now from (Engelking, 1989, Theorem 3.1.13) since $\mathbb{X}$ is compact and $\mathcal{M}$ is a Hausdorff space.

2. Upper-semicontinuity is also referred to as upper-hemicontinuity for multi-valued functions in the literature.

Measurable selection. Since π is upper-semicontinuous and maps to compact sets it is *usco-compact* (Fremlin, 2001, Definition 422A). This implies that π is measurable as a function from $\mathcal{H}$ to the compact subsets of $\mathcal{M}$ where the latter is equipped with the Vietoris topology and the corresponding Borel algebra (Fremlin, 2001, Proposition 5A4Db). Furthermore, there exists a Borel-measurable function f from the compact, non-empty, subsets of $\mathcal{M}$ to $\mathcal{M}$ such that $f(K) \in K$ for every compact, non-empty, subset K of $\mathcal{M}$. Define $W' = f(\pi(Z + h^*))$ then $W = \phi^{-1}(W')$ is the continuous image of the measurable function W' and W has the stated properties. ■

C.2 Proof of Proposition 2

Proof **(a)** Let W be the random variable provided by Proposition 1 and let $\boldsymbol{W} = \phi(W) - h^*$. Then $\|\boldsymbol{Z} - \boldsymbol{W}\|_2 = d(\boldsymbol{Z} + h^*, \mathcal{M})$. Observe that two applications of the Cauchy-Schwarz inequality yield

$$\begin{aligned} E(|\langle f, \boldsymbol{Z}\rangle - f(W) - \langle f, h^*\rangle| \times \chi B) &\leq E|\langle f, (\boldsymbol{Z} - \phi(W) - h^*) \times \chi B\rangle| \\ &\leq \|f\| E(\chi B \times \|\boldsymbol{Z} - \mathbf{W}\|) \\ &\leq \sqrt{P(B)}\|f\|\|\boldsymbol{Z} - \mathbf{W}\|_2 \end{aligned}$$

for all $f \in \mathcal{H}$. Similarly, for any $f \in \mathcal{H}$ it holds that

$$E|\langle f, \boldsymbol{Z} - \phi(W) - h^*\rangle| \leq \|f\| E\|\boldsymbol{Z} - \mathbf{W}\| \leq \|f\|\|\boldsymbol{Z} - \mathbf{W}\|_2.$$

Noting that $\boldsymbol{Z}$ is $\mathcal{H}$-independent of S we find that for any $f \in \mathcal{H}$ and $B \in \sigma(S)$

$$\begin{aligned} &|E(f(W) \times \chi B) - Ef(W)P(B)| \\ &= |E((f(W) - \langle f, h^*\rangle) \times \chi B) - E(f(W) - \langle f, h^*\rangle)P(B)| \\ &\leq |E(\langle f, \boldsymbol{Z}\rangle \times \chi B) - E\langle f, \boldsymbol{Z}\rangle P(B)| + (1 + \sqrt{P(B)})\|f\|\|\boldsymbol{Z} - \mathbf{W}\|_2 \\ &\leq 2\|f\|\, d(\boldsymbol{Z} + h^*, \mathcal{M}). \end{aligned}$$

(b) For $C \in \sigma(W)$ let D be the image of C under W, i.e. $D = W[C]$, $D \subset \mathbb{X}$. For $f \in \mathcal{H}$ let

$$\xi_C(f) = \|\chi D(W) - f(W)\|_2.$$

Now, for any $B \in \sigma(S)$,

$$|P(C \cap B) - E(f(W) \times \chi B)| \leq P(B)^{1/2}(E(\chi D(W) - f(W))^2)^{1/2} \leq \xi_C(f).$$

Moreover, we have $|P(C) - Ef(W)| \leq \xi_C(f)$. Hence, for any $f \in \mathcal{H}$ it holds that

$$\begin{aligned} |P(C \cap B) - P(C)P(B)| &\leq 2\xi_C(f) + |E(f(W) \times \chi B) - Ef(W)P(B)| \\ &\leq 2(\xi_C(f) + \|f\| d(\boldsymbol{Z} + h^*, \mathcal{M})). \end{aligned}$$

This proves the first part of the proposition.

(c) For the second part: by assumption for $A \in \sigma(\boldsymbol{Z})$ there exists a $C \in \sigma(W)$ such that $P(A \triangle C) \leq c$. For any such C we have that $|P(C) - P(A)| \leq P(C \triangle A) \leq c$ and

$$|P(C \cap B) - P(A \cap B)| \leq P((C \triangle A) \cap B) \leq c.$$

Hence, $|P(A \cap B) - P(A)P(B)| \leq 2c + 2(\xi_C(f) + \|f\| d(\boldsymbol{Z} + h^*, \mathcal{M}))$ for all $f \in \mathcal{H}$. Taking the infimum over f and C proves the second part of the proposition. ■

C.3 Proof of Proposition 3

First note that since ϕ is continuous and $\mathbb{X}$ is compact it follows that ρ is finite and

$$\begin{aligned}\|\boldsymbol{Z}\| &= \|\phi(X) - E^S\phi(X) + E\phi(X)\| \\ &\leq \|\phi(X)\| + \|E^S\phi(X)\| + \|E\phi(X)\| \\ &\leq \|\phi(X)\| + E^S\|\phi(X)\| + E\|\phi(X)\| \qquad (20) \\ &\leq 3\rho\end{aligned}$$

where (20) follows from (Diestel and Uhl, 1977, Theorem II.4) and (Pisier, 2016, Proposition 1.12). Let $\boldsymbol{Z}_i,\ i \leq n$ be n independent copies of $\boldsymbol{Z}$ and define $Y_i := \min_{h\in\mathcal{M}} \|\boldsymbol{Z}_i + h^* - h\|,\ i \leq n$, and $Y := \min_{h\in\mathcal{M}} \|\boldsymbol{Z} + h^* - h\|$. By Hoeffding's inequality we have,

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^n Y_i - EY\right| \geq \epsilon\right) \leq 2\exp\Big(-\frac{2n\epsilon^2}{25\rho^2}\Big)$$

and the result follows.

C.4 Proof of Proposition 4

Proof (a) We first show that

$$\langle E^S\phi(X)^\bullet, (\boldsymbol{Z}')^\bullet\rangle_2 = \langle E(\phi(X)^\bullet), (\boldsymbol{Z}')^\bullet\rangle_2. \qquad (21)$$

$E^S\phi(X)^\bullet$ is an element of $L^2(\Omega, \sigma(S), P; \mathcal{H})$ and there exists a sequence of $\sigma(S)$-measurable simple function $\{U_n\}_{n\in\mathbb{N}}$ such that $\lim_{n\to\infty} \|U_n^\bullet - E^S\phi(X)^\bullet\|_2 = 0$. In particular,

$$\lim_{n\to\infty} \langle U_n^\bullet, (\boldsymbol{Z}')^\bullet\rangle_2 = \langle E^S\phi(X)^\bullet, (\boldsymbol{Z}')^\bullet\rangle_2. \qquad (22)$$

Furthermore,

$$\|E(U_n^\bullet) - E(\phi(X)^\bullet)\|_2 \leq E\|U_n^\bullet - E^S\phi(X)^\bullet\|_2 = \|U_n^\bullet - E^S\phi(X)^\bullet\|_2 \qquad (23)$$

goes to zero in n.

Consider any $U_n = \sum_{i=1}^m h_i \times \chi A_i$ where $h_i \in \mathcal{H}$, $A_i \in \sigma(S)$, $m \in \mathbb{N}$, and observe that

$$\langle U_n^\bullet, (\boldsymbol{Z}')^\bullet\rangle_2 = \sum_{i=1}^m E\langle h_i \times \chi A_i, \boldsymbol{Z}'\rangle = \sum_{i=1}^m E(\langle h_i, \boldsymbol{Z}'\rangle \times \chi A_i) = \sum_{i=1}^m E\langle h_i, \boldsymbol{Z}'\rangle \times E(\chi A_i),$$

using the assumption on $\boldsymbol{Z}'$. The assumption can be applied because χA_i is $\sigma(S)$-measurable, and, hence, can be written as a function of S (Shiryaev, 1989)[II.§4.Thm.3]. Furthermore,

$$\sum_{i=1}^m E\langle h_i, \boldsymbol{Z}'\rangle \times E(\chi A_i) = E\langle \sum_{i=1}^m h_i \times E(\chi A_i), \boldsymbol{Z}'\rangle = E\langle E(U_n^\bullet), \boldsymbol{Z}'\rangle$$

and $\langle U_n^\bullet, (\boldsymbol{Z}')^\bullet\rangle_2 = \langle E(U_n^\bullet), (\boldsymbol{Z}')^\bullet\rangle_2$. Equation (21) follows now from Equation (22), (23) and $E(E^S\phi(X)^\bullet) = E(\phi(X)^\bullet)$.

(b) Since

$$\langle E^S\phi(X)^\bullet, (\boldsymbol{Z}')^\bullet\rangle_2 = \langle E(\phi(X)^\bullet), (\boldsymbol{Z}')^\bullet\rangle_2$$
$$\text{and } \langle \phi(X)^\bullet, E^S\phi(X)^\bullet\rangle_2 = \|E^S\phi(X)^\bullet\|_2^2,$$

it readily follows that

$$\begin{aligned}\|\phi(X)^\bullet - (\boldsymbol{Z}')^\bullet\|_2^2 &= \|\phi(X)^\bullet - \boldsymbol{Z}^\bullet\|_2^2 \\ &\quad + 2\langle E^S\phi(X)^\bullet - E(\phi(X)^\bullet), \phi(X)^\bullet - E^S\phi(X)^\bullet + E(\phi(X)^\bullet) - (\boldsymbol{Z}')^\bullet\rangle_2 \\ &\quad + \|\boldsymbol{Z}^\bullet - (\boldsymbol{Z}')^\bullet\|_2^2 \\ &= \|\phi(X)^\bullet - \boldsymbol{Z}^\bullet\|_2^2 + \|\boldsymbol{Z}^\bullet - (\boldsymbol{Z}')^\bullet\|_2^2.\end{aligned}$$

Hence, $\boldsymbol{Z}$ is a minimizer and it is almost surely unique because $\|\boldsymbol{Z}^\bullet - (\boldsymbol{Z}')^\bullet\|_2^2$ is only zero if $\boldsymbol{Z}^\bullet = (\boldsymbol{Z}')^\bullet$. ■

C.5 Proof of Proposition 5

Proof **(a)** In the following, let $s_1, \ldots, s_l$ be the values S can attain. Furthermore, let $f_i = E_n(\phi(X)|S = s_i) - E(\phi(X)|S = s_i)$, and let $\mathcal{F} = \sigma(X_1, S_1, \ldots, X_n, S_n)$. Each f_i is $\mathcal{F}$-measurable. Observe that for $i \neq j$,

$$\begin{aligned}E^{\mathcal{F}}(\langle f_i \times \chi\{S = s_i\}, f_j \times \chi\{S = s_j\}\rangle) &= E^{\mathcal{F}}(\langle f_i, f_j\rangle \times \chi\{S = s_i, S = s_j\}) \\ &= \langle f_i, f_j\rangle \cdot E^{\mathcal{F}}(\chi\{S = s_i, S = s_j\}) \\ &= \langle f_i, f_j\rangle \cdot P(S = s_i, S = s_j) \\ &= 0\end{aligned}$$

since f_i, f_j are $\mathcal{F}$-measurable and S is independent of $\mathcal{F}$. Hence,

$$\begin{aligned}E^{\mathcal{F}}(\|E_n^S\phi(X) - E^S\phi(X)\|^2) &= E^{\mathcal{F}}\Big(\Big\|\sum_{i=1}^{l} f_i \times \chi\{S = s_i\}\Big\|^2\Big) \\ &= \sum_{i=1}^{l} E^{\mathcal{F}}(\|f_i \times \chi\{S = s_i\}\|^2) \\ &= \sum_{i=1}^{l} E^{\mathcal{F}}(\|f_i\|^2 \times \chi\{S = s_i\}) \\ &= \sum_{i=1}^{l} \|f_i\|^2 P(S = s_i) \\ &= \sum_{i=1}^{l} P(S = s_i) \sup_{\|h\|\leq 1} |E_n(h(X)|S = s_i) - E(h(X)|S = s_i)|^2.\end{aligned}$$

(b) For each i either $P(S = s_i) = 0$ or

$$\sup_{\|h\|\leq 1} |E_n(h(X)|S = s_i) - E(h(X)|S = s_i)|^2 \in O_P^*(n^{-1})$$

by an argument similar to that given for (Grünewälder, 2018, Proposition 3.1). Since there are only l terms in the sum this result carries over to the whole sum. ■

C.6 Proof of Proposition 6

Proof Recall the notation $\mathfrak{D}_\ell := \{\Delta_i : i \in 1, \ldots, \ell^d\},\ \ell \in \mathbb{N}$ where $\Delta_1, \Delta_2, \ldots, \Delta_{\ell^d}$ are the dyadic cubes $\Delta_1, \Delta_2, \ldots, \Delta_{\ell^d}$ of side-length $1/\ell$ discretizing $\mathbb{S}$. Let $\mathcal{G} := \sigma(\{S^{-1}[\Delta] :\ \Delta \in \mathfrak{D}_\ell\})$ and choose a Bochner measurable $g : \mathbb{S} \to \mathcal{H}$ according to Lemma 2 such that $g(S) = E^S\phi(X)$ (a.s.). Since $\mathcal{G} \subseteq \sigma(S)$ we have,

$$E^{\mathcal{G}}\phi(X) = E^{\mathcal{G}}(E^S\phi(X))) = E^{\mathcal{G}}(g(S)) \quad \textit{almost surely}. \tag{24}$$

In the following, we use $g \circ S$ instead of $g(S)$ for readability. With probability one it holds that,

$$\begin{aligned}
&E^{\mathcal{F}}\|g \circ S - E^{\mathcal{G}}(g \circ S)\|^2 \\
&\quad = E^{\mathcal{F}}\Big(\sum_{\Delta \in \mathfrak{D}_\ell} \|g \circ S - E^{\mathcal{G}}(g \circ S)\|^2 \chi\{S \in \Delta\}\Big) \\
&\quad = \sum_{\Delta \in \mathfrak{D}_\ell} E^{\mathcal{F}}\|(g \circ S - E^{\mathcal{G}}(g \circ S))\chi\{S \in \Delta\}\|^2 \\
&\quad = \sum_{\Delta \in \mathfrak{D}_\ell} E^{\mathcal{F}}\|(g \circ S - \sum_{\Delta' \in \mathfrak{D}_\ell} E(g \circ S | S \in \Delta')\chi\{S \in \Delta'\})\chi\{S \in \Delta\}\|^2 \\
&\quad = \sum_{\Delta \in \mathfrak{D}_\ell} E^{\mathcal{F}}\left(\|g \circ S - E(g \circ S | S \in \Delta)\|^2 \chi\{S \in \Delta\}\right)
\end{aligned}$$

By (Diestel and Uhl, 1977, II.Corollary 8) for any $\Delta \in \mathfrak{D}_\ell$ it holds that the conditional expectation of g given Δ is in the closed convex hull of $g[\Delta] := \{g(s) : s \in \Delta\}$. That is,

$$\frac{1}{\mu(\Delta)} \int_\Delta g \, d\mu \in \mathrm{cch}(g[\Delta])$$

This means that for every $\epsilon > 0$ there exist $k \in \mathbb{N}$, and some $s_1, \ldots, s_k \in \Delta$ and $\alpha_1, \ldots, \alpha_k > 0$ with $\sum_{j=1}^k \alpha_i = 1$ such that

$$\Big\|\frac{1}{\mu(\Delta)} \int_\Delta g \, d\mu - \sum_{j=1}^k \alpha_j g(s_j)\Big\|^2 \leq \epsilon.$$

Let $D := S^{-1}[\Delta]$. We obtain

$$\frac{1}{\mu(\Delta)} \int_\Delta g \, d\mu = \frac{1}{P(D)} \int_D (g \circ S) \, dP = E(g \circ S | S \in \Delta).$$

Since $g \circ S$ is assumed to be L-Lipschitz-continuous, for all $\Delta \in \mathfrak{D}_\ell$ we have

$$\begin{aligned}
&\|g \circ S - \sum_{j=1}^{k} \alpha_j g(s_j)\|^2 \chi\{S \in \Delta\} \\
&\quad \leq \sup_{s \in \Delta} \|\sum_{j=1}^{k} \alpha_j (g(s) - g(s_j))\|^2 \\
&\quad \leq \sup_{s \in \Delta} (\sum_{j=1}^{k} \alpha_j \|g(s) - g(s_j)\|)^2 \\
&\quad \leq \Big(\sum_{j=1}^{k} \alpha_j \sup_{s \in \Delta} \|g(s) - g(s_j)\|\Big)^2 \\
&\quad \leq L^2 \Big(\sum_{j=1}^{k} \alpha_j \sup_{s \in \Delta} \|s - s_j\|\Big)^2 \\
&\quad \leq dL^2 \ell^{-2}.
\end{aligned}$$

Moreover, noting that $\chi\{\cdot\} = \chi^2\{\cdot\}$ we obtain, $\|g \circ S - \sum_{j=1}^{k} \alpha_j g(s_j)\| \chi\{S \in \Delta\} \leq L\sqrt{d}\ell^{-1}$. It follows that,

$$\begin{aligned}
&\|(g \circ S - E(g \circ S | S \in \Delta))\chi\{S \in \Delta\}\|^2 \\
&= \|(g \circ S - E(g \circ S | S \in \Delta))\|^2 \chi\{S \in \Delta\} \\
&\leq \Big(\big\|g \circ S - \sum_{j=1}^{k} \alpha_j g(s_j)\big\| + \big\|\sum_{j=1}^{k} \alpha_j g(s_j) - E(g \circ S | S \in \Delta)\big\|\Big)^2 \chi\{S \in \Delta\} \\
&\leq \Big(\big\|(g \circ S - \sum_{j=1}^{k} \alpha_j g(s_j))\big\| + \epsilon\Big)^2 \chi\{S \in \Delta\}.
\end{aligned}$$

Since this holds for every $\epsilon > 0$ we have,

$$\|(g \circ S - E(g \circ S | S \in \Delta))\chi\{S \in \Delta\}\|^2 \leq dL^2 \ell^{-2}.$$

Observe that for $\Delta \neq \Delta'$, $\Delta, \Delta' \in \mathfrak{D}_\ell$,

$$E^{\mathcal{F}}\Big(\|g \circ S - E(g \circ S | S \in \Delta)\|\chi\{S \in \Delta\} \times \|g \circ S - E(g \circ S | S \in \Delta')\|\chi\{S \in \Delta'\}\Big) = 0$$

and

$$\begin{aligned}
E^{\mathcal{F}}\|g \circ S - E^{\mathcal{G}}(g \circ S)\|^2 &= \sum_{\Delta \in \mathfrak{D}_\ell} E^{\mathcal{F}}\|g \circ S - E(g \circ S | S \in \Delta)\|^2 \chi\{S \in \Delta\} \\
&= \sum_{\Delta \in \mathfrak{D}_\ell} E^{\mathcal{F}}\|g \circ S - E(g \circ S | S \in \Delta)\|^2 \chi^2\{S \in \Delta\} \\
&= \sum_{\Delta \in \mathfrak{D}_\ell} E^{\mathcal{F}}\|(g \circ S - E(g \circ S | S \in \Delta))\chi\{S \in \Delta\}\|^2 \chi\{S \in \Delta\} \\
&\leq dL^2 \ell^{-2}.
\end{aligned}$$

In particular,

$$E^{\mathcal{F}}\|g \circ S - E^{\mathcal{G}}\phi(X)\|^2 \leq dL^2\ell^{-2}. \tag{25}$$

On the other hand, in much the same way as in the proof of Proposition 5, we have

$$E^{\mathcal{F}}(\|E_n^S\phi(X) - E^{\mathcal{G}}\phi(X)\|^2) = \sum_{\Delta\in\mathfrak{D}_\ell} P(S\in\Delta) \sup_{\|h\|\leq 1} |E_n(h(X)|S\in\Delta) - E(h(X)|S\in\Delta)|^2.$$

Let $U := (X, S)$ and define the push forward measure $\nu := P \circ U^{-1}$ of P onto $\mathbb{X}\times\mathbb{S}$ under U. Set $\nu_n := \frac{1}{n}\sum_{i=1}^n \delta_{(X_i,S_i)}$ where $\delta_{(X,S)}$ denotes the measure that has point mass at (X, S). Define the projection map $\pi : \mathbb{X}\times\mathbb{S} \to \mathbb{X}$ which maps a tuple $(x, s) \in \mathbb{X}\times\mathbb{S}$ to its first element so that $\pi((x, s)) = x$. For each $h \in \mathcal{H}$ such that $\|h\| \leq 1$ and every $\Delta \in \mathfrak{D}_\ell$ we obtain

$$\begin{aligned}
&|E_n(h(X)|S\in\Delta) - E(h(X)|S\in\Delta)|^2 \\
&\qquad = |E_n(h(\pi(U))|U\in\mathbb{X}\times\Delta) - E(h(\pi(U))|U\in\mathbb{X}\times\Delta)|^2 \\
&\qquad = \Big|\int_{\mathbb{X}\times\Delta} h\circ\pi\, d\nu_n - \int_{\mathbb{X}\times\Delta} h\circ\pi\, d\nu\Big|^2.
\end{aligned}$$

For each $\ell \in \mathbb{N}$ define $\mathfrak{C}_\ell := \{\mathbb{X}\times\Delta : \Delta\in\mathfrak{D}_\ell\}$. By assumption, $\mathcal{H}_{\mathfrak{C}} = \{h\times\chi D : h\in\mathcal{H},\ \|h\|\leq 1,\ D\in\bigcup_{\ell\in\mathbb{N}}\mathfrak{C}_\ell\}$ is P-Donsker and for $D\in\mathfrak{C}_\ell$, with $D = \mathbb{X}\times\Delta_i$ for $i\leq\ell$, $\nu(D) = PS^{-1}[\Delta_i] \geq b\ell^{-d}$. For a given $\alpha\in(0, 1/2)$ let ℓ be $\lfloor n^{\alpha/d}\rfloor$ so that $\ell^{-d}\geq n^{-\alpha}$. Similarly to (Grünewälder, 2018, Proposition 3.2) it follows that there exists a constant M such that for all $n\geq 1$ and corresponding ℓ,

$$\sup_{\|h\|\leq 1}\sup_{C\in\mathfrak{C}_\ell}\Big|\int_C h\circ\pi\, d\nu_n - \int_C h\circ\pi\, d\nu\Big| \leq 2M\ell^d n^{-1/2}/b.$$

Thus,

$$E^{\mathcal{F}}(\|E_n^S\phi(X) - E^S\phi(X)\|^2) \leq 4M^2\ell^{2d}/nb^2. \tag{26}$$

Using Equation (25) and (26) as well as the Cauchy-Schwarz inequality for conditional expectations we obtain,

$$\begin{aligned}
E^{\mathcal{F}}(\|E_n^S\phi(X) - E^S\phi(X)\|^2) &\leq E^{\mathcal{F}}(\|E_n^S\phi(X) - E^{\mathcal{G}}\phi(X)\| + \|E^{\mathcal{G}}\phi(X) - E^S\phi(X)\|)^2 \\
&\leq 4M^2\ell^{2d}/nb^2 + 4M\sqrt{d}L\ell^{d-1}n^{-1/2}/b + dL^2\ell^{-2}.
\end{aligned}$$

Because $\ell = \lfloor n^{\alpha/d}\rfloor$ the upper-bound becomes

$$4M^2n^{2\alpha-1}/b^2 + dL^2/(n^{-\alpha/d}-1)^2 + 4M\sqrt{d}Ln^{\alpha(1-\frac{1}{d})-\frac{1}{2}}/b.$$

We claim that the rate of convergence in n is optimized by $\alpha^* = d/2(d+1)$: For $\alpha\geq\alpha^*$ we have

$$2\alpha - 1 \geq \alpha(1-1/d) - 1/2 \geq -2\alpha/d$$

and the dominant term $2\alpha-1$ is minimized at α^*. On the other hand, for $\alpha\leq\alpha^*$,

$$2\alpha - 1 \leq -2\alpha/d \quad\text{and}\quad \alpha(1-1/d) - 1/2 \leq -2\alpha/d.$$

In this case the dominant term is also minimized for α^*. Therefore, we must set

$$\ell^* = \lfloor n^{\alpha^*/d}\rfloor = \lfloor n^{\frac{1}{2(d+1)}}\rfloor.$$

■

Appendix D. Solution to the oblivious kernel ridge regression optimization problem

Define $\mathbf{z}_i := (\langle \mathbf{Z}_1, \mathbf{Z}_i \rangle \ \cdots \ \langle \mathbf{Z}_n, \mathbf{Z}_i \rangle)^\top$, $i \in 1..n$, and observe that

$$\mathcal{O} = \begin{pmatrix} | & | & \dots & | \\ \mathbf{z}_1 & \mathbf{z}_2 & \dots & \mathbf{z}_n \\ | & | & \dots & | \end{pmatrix}.$$

Let $\hat{f}$ be the minimizer of the regularized least-squares error as given by (11). By the representer theorem there exist scalars $\alpha_1, \dots, \alpha_n$ such that $\hat{f} = \sum_{j=1}^n \alpha_j \mathbf{Z}_j$. It follows that $\langle \hat{f}, \mathbf{Z}_i \rangle = \sum_{j=1}^n \alpha_j \langle \mathbf{Z}_j, \mathbf{Z}_i \rangle$ so that,

$$\sum_{i=1}^n (\langle \hat{f}, \mathbf{Z}_i \rangle - Y_i)^2 + \lambda \|\hat{f}\|^2 = (\mathcal{O}\boldsymbol{\alpha} - \mathbf{y})(\mathcal{O}\boldsymbol{\alpha} - \mathbf{y})^\top + \lambda \boldsymbol{\alpha}^\top \mathcal{O} \boldsymbol{\alpha} \tag{27}$$

where $\boldsymbol{\alpha} := (\alpha_1, \dots, \alpha_n)^\top$ and $\mathbf{y} := (Y_1, \dots, Y_n)^\top$. Noting that $\hat{f}$ is the minimizer, and thus taking the gradient of (27) with respect to $\boldsymbol{\alpha}$ we obtain,

$$\nabla_{\boldsymbol{\alpha}} \Big((\mathcal{O}\boldsymbol{\alpha} - \mathbf{y})(\mathcal{O}\boldsymbol{\alpha} - \mathbf{y})^\top + \lambda \boldsymbol{\alpha}^\top \mathcal{O} \boldsymbol{\alpha} \Big) = 0.$$

Solving for $\boldsymbol{\alpha}$ and noting that $\mathcal{O}$ is symmetric, we obtain

$$\begin{aligned}
\boldsymbol{\alpha} &= \mathcal{O}^{-1} \Big(\mathcal{O}^\top + \lambda I \Big)^{-1} \mathcal{O}^\top \mathbf{y} \\
&= \mathcal{O}^{-1} \Big(\mathcal{O}^\top + \lambda I \Big)^{-1} \mathcal{O} \mathbf{y} && \text{since } \mathcal{O} \text{ is symmetric} \\
&= \mathcal{O}^{-1} \Big(\mathcal{O}^\top + \lambda I \Big)^{-1} (\mathcal{O}^{-1})^{-1} \mathbf{y} \\
&= \Big(\mathcal{O}^{-1} \Big(\mathcal{O}^\top + \lambda I \Big) \mathcal{O} \Big)^{-1} \mathbf{y} \\
&= \Big((\mathcal{O}^{-1}\mathcal{O} + \lambda \mathcal{O}^{-1}) \mathcal{O} \Big)^{-1} \mathbf{y} && \text{since } \mathcal{O} \text{ is symmetric} \\
&= (\mathcal{O} + \lambda I)^{-1} \mathbf{y}.
\end{aligned}$$

Appendix E. Algorithms

We discuss three algorithms in this section: an algorithm to calculate the oblivious kernel matrix (Section E.1), an algorithm to calculate $\langle \boldsymbol{Z}, \boldsymbol{Z}_i \rangle$ which is needed for prediction (Section E.2), and an algorithm to calculate W, the projection of $\boldsymbol{Z}_i$ onto $\mathcal{M}$, which also allows us to estimate the distance between $\boldsymbol{Z}$ and $\mathcal{M}$ (Section E.3).

E.1 Calculating the oblivious kernel matrix

We start by deriving the algorithm for calculating the oblivious matrix. The result algorithm is summarized in Algorithm 1 on page 32. Throughout we assume that $A_1, \dots, A_l$ is a partition of $\mathbb{S}$ and we assume that $2n$ samples (X_i, S_i) are available. The algorithm splits the data into two parts of size n and uses the samples $n+1, \dots, 2n$ to estimate the conditional expectation. The remaining n

Algorithm 1 Generating the oblivious kernel matrix; the sum over an empty index set is treated as 0

Input: data $(x_1, s_1), \ldots (x_{2n}, s_{2n})$, disjoint sets $A_1, \ldots, A_\ell$ which cover $\mathbb{S}$
set $M = \sum_{i=n+1}^{2n} \sum_{j=n+1}^{2n} k(x_i, x_j)/n^2$
set $\mathcal{I}_i = \emptyset,\ i \in 1, \ldots, \ell$
for $i = n+1$ **to** $2n$ **do**
  find index u such that $s_i \in A_u$
  update $\mathcal{I}_u \leftarrow \mathcal{I}_u \cup \{i\}$
end for
for $i = 1$ **to** n **do**
  set $\rho_i = \sum_{u=n+1}^{2n} k(x_i, x_u)/n$
  for $a = 1$ **to** l **do**
    set $\xi_{i,a} = \sum_{u \in \mathcal{I}_a} k(x_i, x_u)/|\mathcal{I}_a|$
  end for
end for
for $a = 1$ **to** l **do**
  set $\tau_a = \sum_{u \in \mathcal{I}_a} \sum_{v=n+1}^{2n} k(x_u, x_v)/(n|\mathcal{I}_a|)$
  for $b = 1$ **to** l **do**
    set $o_{a,b} = \sum_{u \in \mathcal{I}_a, v \in \mathcal{I}_b} k(x_u, x_v)/(|\mathcal{I}_a||\mathcal{I}_b|)$
  end for
end for
for $i = 1$ **to** n **do**
  for $j = i$ **to** n **do**
    set a such that $s_j \in A_a$
    set b such that $s_i \in A_b$
    set $\mathcal{O}_{i,j} = k(x_i, x_j) - \xi_{i,a} - \xi_{j,b} + o_{a,b} + M - \rho_i - \rho_j - \tau_a - \tau_b$
    set $\mathcal{O}_{j,i} = \mathcal{O}_{i,j}$
  end for
end for
Return: $\mathcal{O}$

samples are then used to generate the features $\boldsymbol{Z}_i$, $i = 1, \ldots, n$. The features $\boldsymbol{Z}_i$ will not be explicitly stored. The only thing that will be stored is the oblivious matrix $\mathcal{O}$. To calculate the oblivious matrix we only need kernel evaluations. To see this consider any $i \leq n$, then

$$\boldsymbol{Z}_i = \phi(X_i) - E_n^{S_i}\phi(X) = \phi(X_i) - \sum_{u=1}^{l} E_n(\phi(X)|S \in A_u) \times \chi\{S_i \in A_u\}.$$

For $u = 1, \ldots, l$ let

$$N_u = \sum_{v=n+1}^{2n} \chi\{S_v \in A_u\}$$

be the number of samples with indices within $n+1, \ldots, 2n$ that fall into set A_u. The estimate of the elementary conditional expectation is

$$E_n(\phi(X)|S \in A_u) = \frac{1}{N_u} \sum_{v=n+1}^{2n} \phi(X_v) \times \chi\{S_v \in A_u\},$$

which attains values in $\mathcal{H}$.

Now consider the inner product between $\boldsymbol{Z}_i$ and $\boldsymbol{Z}_j$, $i, j \leq n$:

$$\begin{aligned}\langle \boldsymbol{Z}_i, \boldsymbol{Z}_j \rangle =& \langle \phi(X_i), \phi(X_j) \rangle - \langle \phi(X_i), E_n^{S_j} \phi(X) \rangle - \langle E_n^{S_i} \phi(X), \phi(X_j) \rangle \\ &+ \langle E_n^{S_i} \phi(X), E_n^{S_j} \phi(X) \rangle + \langle \phi(X_i), E_n(\phi(X)) \rangle \\ &+ \langle E_n(\phi(X)), \phi(X_j) \rangle - \langle E_n^{S_i} \phi(X), E_n(\phi(X)) \rangle \\ &- \langle E_n(\phi(X)), E_n^{S_j} \phi(X) \rangle + \langle E_n(\phi(X)), E_n(\phi(X)) \rangle.\end{aligned}$$

This reduces to calculations involving only the kernel function and no other functions from $\mathcal{H}$. In detail,

$$\langle \phi(X_i), \phi(X_j) \rangle = k(X_i, X_j),$$

and

$$\langle \phi(X_i), E_n^{S_j} \phi(X) \rangle = \sum_{u=1}^{l} \langle \phi(X_i), E_n(\phi(X)|S \in A_u) \rangle \times \chi\{S_j \in A_u\},$$

where

$$\begin{aligned}\langle \phi(X_i), E_n(\phi(X)|S \in A_u) \rangle &= \frac{1}{N_u} \sum_{l=n+1}^{2n} \langle \phi(X_i), \phi(X_l) \rangle \times \chi\{S_l \in A_u\} \\ &= \frac{1}{N_u} \sum_{l=n+1}^{2n} k(X_i, X_l) \times \chi\{S_l \in A_u\}.\end{aligned}$$

The inner product $\langle E_n(\phi(X)|S \in A_u), \phi(X_j) \rangle$ can be calculated in the same way. Furthermore,

$$\langle E_n^{S_i} \phi(X), E_n^{S_j} \phi(X) \rangle = \sum_{u=1}^{l} \sum_{v=1}^{l} \langle E_n(\phi(X)|S \in A_u), E_n(\phi(X)|S \in A_v) \rangle \chi\{S_i \in A_u, S_j \in A_v\}$$

and

$$\begin{aligned}&\langle E_n(\phi(X)|S \in A_u), E_n(\phi(X)|S \in A_v) \rangle \\ &= \frac{1}{N_u N_v} \sum_{l=n+1}^{2n} \sum_{m=n+1}^{2n} \langle \phi(X_l), \phi(X_m) \rangle \times \chi\{S_l \in A_u, S_m \in A_v\} \\ &= \frac{1}{N_u N_v} \sum_{l=n+1}^{2n} \sum_{m=n+1}^{2n} k(X_l, X_m) \times \chi\{S_l \in A_u, S_m \in A_v\}.\end{aligned}$$

The terms involving $E_n(\phi(X)) = (1/n) \sum_{i=n+1}^{2n} \phi(X_i)$ are reduced in a similar way to kernel evaluations. Combining these calculations leads to Algorithm 1.

E.2 Prediction based on oblivious features

To be able to predict labels for new observations (X, S) in a regression or classification setting we need to transform (X, S) into an oblivious feature $\boldsymbol{Z}$. The approach to do is the same as for the training data. In particular, the conditional expectation estimates $E_n^{S_j}\phi(X)$ are needed to transform (X, S) into $\boldsymbol{Z}$. For kernel methods $\boldsymbol{Z}$ itself is never calculated explicitly but it appears in algorithms in the form of inner products $\langle \boldsymbol{Z}, \boldsymbol{Z}_i \rangle$, where $i \leq n$ and $\boldsymbol{Z}_i$ are the oblivious features corresponding to the training set. These inner product can be calculated in exactly the same way as the inner products $\langle \boldsymbol{Z}_i, \boldsymbol{Z}_j \rangle$ in Section E.1.

E.3 Projecting the oblivious features onto the manifold

The quadratic distance between $\boldsymbol{Z}$, or more precisely $\boldsymbol{Z}(\omega)$, and $\mathcal{M}$ in $\mathcal{H}$ is equal to

$$\inf_{x \in \mathbb{X}} \|\boldsymbol{Z} - \phi(x)\|^2 = \|\boldsymbol{Z}\|^2 + \inf_{x \in \mathbb{X}} (k(x,x) - 2\langle \boldsymbol{Z}, \phi(x) \rangle).$$

The constant $\|\boldsymbol{Z}\|^2$ is of no relevance and we are looking for a minimum (when this is well-defined) of the function

$$f(x) = k(x,x) - 2\langle \boldsymbol{Z}, \phi(x) \rangle$$

in $\mathbb{X}$. Using the conditional expectation $E_n^S\phi(X)$ and $\boldsymbol{Z} = \phi(X) - E_n^S\phi(X) + E(\phi(X))$ we can rewrite $f(x)$ as

$$f(x) = k(x,x) - 2(k(X,x) - E_n^S k(X,x) + E(k(X,x)).$$

The function f is (α, L)-Hölder continuous whenever $k(x, \cdot)$ is (α, L')-Hölder-continuous for all $x \in \mathbb{X}$ with $L = 8L'$, since then

$$\begin{aligned} |f(x) - f(y)| &\leq |k(x,x) - k(y,y)| + 2(|k(X,x) - k(X,y)| + E_n^S|k(X,y) - k(X,x)| \\ &\qquad + E|k(X,x) - k(X,y)|) \\ &\leq |k(x,x) - k(x,y)| + |k(x,y) - k(y,y)| + 6L'\|y - x\|^\alpha = 8L'\|y - x\|^\alpha. \end{aligned}$$

This property of f is useful because various kernel functions are Hölder-continuous and efficient algorithms are available to optimize Hölder-continuous functions. In particular, there exist classical global optimization algorithms (Vanderbei, 1997) and bandit algorithms (Munos, 2014) for this task.

The projection of $\boldsymbol{Z}$ onto $\mathcal{M}$ can also be used directly to approximate $d_n(\boldsymbol{Z} + h^*, \mathcal{M})$ and, by applying Proposition 3, to estimate $d(\boldsymbol{Z} + h^*, \mathcal{M})$.

References

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

D. P. Bertsekas and J. N. Tsitsiklis. *Introduction to Probability*. Athena Scientific, 1st edition, 2002.

R.V. Bradley. *Introduction to Strong Mixing Conditions, Vols. 1, 2 and 3*. Kendrick Press, 2007.

T. Calders, F. Kamiran, and M. Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, 2009.

F. Calmon, D. Wei, B. Vinzamuri, K. Natesan Ramamurthy, and K. R. Varshney. Optimized preprocessing for discrimination prevention. In *In Advances in Neural Information Processing Systems*, 2017.

J. Diestel and J.J. Uhl. *Vector measures*. American Mathematical Soc., 1977.

M. Donini, L. Oneto, S. Ben-David, J. Shawe-Taylor, and M. Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, 2018.

P. Doukhan. *Mixing: Properties and Examples*. Springer Lecture Notes, 1994.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

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

R. Engelking. *General Topology*. Heldermann Verlag Berlin, 1989.

D.H. Fremlin. *Measure Theory*. Torres Fremlin, 2001.

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

A. Gretton, K. Fukumizu, CH. Teo, L. Song, B. Schölkopf, and AJ. Smola. A kernel statistical test of independence. In *Advances in neural information processing systems*, 2008.

S. Grünewälder. Plug-in estimators for conditional expectations and probabilities. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, 2018.

M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, 2016.

M. Joseph, M. Kearns, J. H. Morgenstern, and A. Roth. Fairness in learning: Classic and contextual bandits. In *In Advances in Neural Information Processing Systems*, 2016.

N. Kilbertus, M. Rojas-Carulla G., Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf. Avoiding discrimination through causal reasoning. In *In Advances in Neural Information Processing Systems*, 2017.

J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent Trade-Offs in the Fair Determination of Risk Scores. In Christos H. Papadimitriou, editor, *8th Innovations in Theoretical Computer Science Conference*, volume 67. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2017.

M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. In *In Advances in Neural Information Processing Systems*, 2017.

C. Louizos, K. Swersky, Y. Li, M. Welling, and R. S. Zemel. The variational fair autoencoder. In *International Conference on Learning Representations*, 2015.

D. Madras, E. Creager, T. Pitassi, and R. Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, 2018.

D.J. Marcus. Relationships between donsker classes and sobolev spaces. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 1985.

R. Munos. From bandits to monte-carlo tree search: The optimistic principle applied to optimization and planning. *Foundations and Trends in Machine Learning*, 2014.

L. Oneto, M. Donini, G. Luise, C. Ciliberto, A. Maurer, and M. Pontil. Exploiting mmd and sinkhorn divergences for fair and transferable representation learning. In *Advances in Neural Information Processing Systems*, 2020.

G. Pisier. *Martingales in Banach Spaces*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2016.

A. Shiryaev. *Probability*. Springer: Graduate Texts in Mathematics, second edition, 1989.

B. K. Sriperumbudur, K. Fukumizu, and G.R.G. Lanckriet. Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 2011.

R.J. Vanderbei. Extension of piyavskii's algorithm to continuous global optimization. Technical report, Princeton University, 1997.

M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K.P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, 2017.

R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *In International Conference on Machine Learning*, 2013.