arXiv:2002.02901v1 [stat.ML] 7 Feb 2020

# Oblivious Data for Fairness with Kernels

**Steffen Grünewälder** [1] **Azadeh Khaleghi** [1]

## 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. Our theoretical results are accompanied by experimental evaluations.

## 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 post-graduate 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; R. Zemel and Dwork, 2013; Louizos et al., 2015; Hardt et al., 2016; M. Joseph and Roth, 2016; N. Kilbertus and Schölkopf, 2017; M. J. Kusner and Silva, 2017; F. Calmon and Varshney, 2017; Zafar et al., 2017; Kleinberg et al., 2017; Donini et al., 2018; Madras et al., 2018), 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.

**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. They are also easy to generate from samples

[1]Mathematics & Statistics, Lancaster University, UK.

$(X_1, S_1), \ldots, (X_n, S_n)$. 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. Having obtained the empirical estimates of the conditional expectations, we generate oblivious features and 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. Our main contributions are as follows.

- 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 in Section 5.2 that our solution satisfies the constraint of the optimization problem at hand, and show via Proposition 5.3 that it is indeed optimal.
- Through Proposition 1 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 $\phi$. 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 conditional expectations to estimate those taking values in a Hilbert space, and show via Proposition 3 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 $O(n^2)$.

While the key contributions of this work are theoretical, we also provide an evaluation of the proposed approach through examples and some experiments.

**Related Work.** Among the vast literature on algorithmic fairness, Donini et al. (2018); Madras et al. (2018), 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. The approach taken 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 to 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 authors do 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.

**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 examples and empirical evaluations in Section 7. We conclude in Section 8 with a discussion of the results and future directions.

## 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. We denote by $\sigma(X)$ the $\sigma$-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.** Let $S : \Omega \to \mathbb{S}$ be a random variable taking values in a measurable space $\mathbb{S}$. For the random variable $X$ 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 $\Omega$ 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 bilinear form acting on $\mathcal{L}^2(\mathcal{H})$ and we also denote these by $\|\cdot\|_2$ and $\langle \cdot, \cdot \rangle_2$.

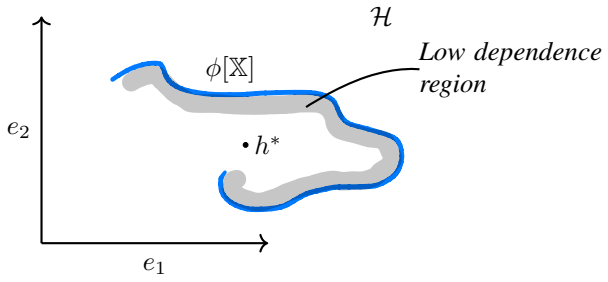


*Figure 1. The image of $\mathbb{X}$ under $\phi$ is sketched (blue curve). This is a subset of $\mathcal{H}$ whose projection onto the subspace spanned by two orthonormal basis elements $e_1$ and $e_2$ is shown here. The set $\phi[\mathbb{X}]$ is a low-dimensional manifold if $\phi$ is continuous. The 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 $\phi$ as introduced in Section 2, and assume that $\mathcal{H}$ is large enough to allow for the approximation of arbitrary indicator functions $\chi\{Z \in A'\}$ 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, indepen-

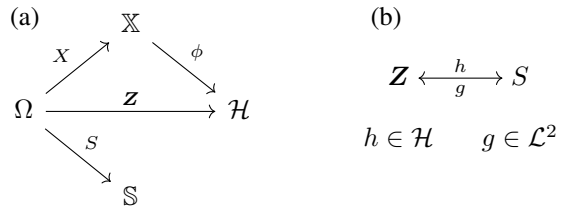


*Figure 2. (a) The three main random variables in Problem 1 are shown. The non-sensitive features $X$ attains values in $\mathbb{X}$ and is mapped onto the RKHS $\mathcal{H}$ through the feature map $\phi$; the sensitive features $S$ attains values in $\mathbb{S}$, and $\boldsymbol{Z}$ attains values in $\mathcal{H}$. All three random variables are defined on the same probability space $(\Omega, \mathcal{A}, P)$. (b) The random variable $\boldsymbol{Z}$ is compared to $S$ by means of RKHS functions $h$ and $\mathcal{L}^2$ functions $g$. In particular, the functions in $\mathcal{H}$ and $\mathcal{L}^2$ are used to guarantee that any non-linear estimator that uses $\boldsymbol{Z}$ is uncorrelated with the sensitive features $S$.*

dent. This is because $h$ and $g$ can be used to approximate arbitrary indicator functions, which together with (2) gives,

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

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

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

Observe that $\phi(Z)$ is a random variable that attains values in an arbitrary low-dimensional manifold; the image $\phi[\mathbb{X}]$ of $\mathbb{X}$ under $\phi$ is visualized as the blue curve in Figure 1. 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 that $\boldsymbol{Z} \in \mathcal{L}^2(\Omega, \mathcal{A}, P; \mathcal{H})$ 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 2.

If $\boldsymbol{Z}$ lies in the image of $\phi$ 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, $\boldsymbol{Z}$ can not be represented as some $\phi(W)$ and there can be dependencies between $\langle \hat{h}, \boldsymbol{Z}\rangle$ and $S$. In Section 4 below we generalize the above argument to bound the dependence between $\boldsymbol{Z}$ and $S$ depending on how well $\boldsymbol{Z}$ can be approximated by $\phi(W)$, for some $W$ appropriately chosen to minimize the distance between $\boldsymbol{Z}$ and $\phi(W)$.

## 4. Bounds on the dependence between $\boldsymbol{Z}$ & $S$

A common approach to quantifying a measure of 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 $\sigma$-algebras $\sigma(\boldsymbol{Z})$ and $\sigma(S)$, and the difference between $P(A \cap B)$ and $P(A)P(B)$, $A \in \sigma(\boldsymbol{Z}), B \in \sigma(S)$, quantifies the dependency between the random variables $\boldsymbol{Z}$ and $S$. Upper bounds on the absolute difference of these two quantities, which are independent of $P(A)$ and $P(B)$, correspond to the notion of $\alpha$-dependence which underlies $\alpha$-mixing. In times-series analysis mixing conditions like $\alpha$-mixing play a significant role since they provide means to control temporal dependencies (see, e.g., (Bradley, 2007; Doukhan, 1994)). We present Proposition 1, which gives a bound on the dependence between $\boldsymbol{Z}$ and $S$.

To improve readability, we summarize the notation used in the proposition statement below and give an intuitive exposition before stating the result.

**Notation 1.** *For $\boldsymbol{Z} \in \mathcal{L}^2(\Omega, \mathcal{A}, P; \mathcal{H})$ fix constants $c_1, c_2 > 0$, a function $h^* \in \mathcal{H}$, and a random variable $W : \Omega \to \mathbb{X}$ with $\phi(W) \in \mathcal{L}^2(\Omega, \mathcal{A}, P; \mathcal{H})$, such that*

*(i.) for all* $A_1 \in \sigma(\boldsymbol{Z})$ *there exists an* $A_2 \in \sigma(W)$ *with* $P(A_1 \triangle A_2) \leq c_1$

*(ii.)* $\|\boldsymbol{Z} - (\phi(W) - h^*)\|_2 \leq c_2$.

In words, for a given $\boldsymbol{Z}$, we first specify some $W$ whose dependence on $\boldsymbol{Z}$ is controlled by some $c_1 > 0$ in that any event $A_1 \in \sigma(\boldsymbol{Z})$ can be coupled with some event $A_2 \in \sigma(S)$ such that $P(A_1 \triangle A_2) \leq c_2$. Note that such a $W$ always exists, e.g. it could be a function of $\boldsymbol{Z}$ in which case the condition would be trivially satisfied. Next, we let $c_2 > 0$ denote an upper-bound on the error (as measured by the Hilbert space-valued $L^2$-norm) of approximating $\boldsymbol{Z}$ by some appropriate (translation of) $\phi(W)$. Observe that the error could be arbitrary large, as we do not require $c_2$ to be particularly small. The result stated below bounds the dependence between $\boldsymbol{Z}$ and $S$ as a function of $c_1$, $c_2$, and the size and approximation capacity of $\mathcal{H}$.

**Proposition 1.** *Suppose that* $\mathcal{H}$ *is separable and its feature map* $\phi$ *satisfies* $\sup_{x \in \mathbb{X}} \|\phi(x)\| \leq 1$. *Consider some* $\boldsymbol{Z} \in \mathcal{L}^2(\Omega, \mathcal{A}, P; \mathcal{H})$ *that is* $\mathcal{H}$*-independent from* $S$. *Let*

$$\psi(A) = \inf_{D \in \mathcal{B}_A} \inf_{f \in \mathcal{H}} 2\|\chi D(W) - f(W)\|_2 + c_2 \|f\| \quad (4)$$

*and let* $\mathcal{B}_A = \{W[C] : C \in \sigma(W), P(C \triangle A) \leq c_1\}$ *where* $W$, $c_1$ *and* $c_2$ *are specified by Notation 1. For any* $A \in \sigma(\boldsymbol{Z})$ *and* $B \in \sigma(S)$ *it holds that*

$$|P(A \cap B) - P(A)P(B)| \leq 2c_1 + \psi(A)P(B)^{1/2}.$$

*Proof.* Proof is provided in Appendix B.3. □

A key factor in the bound is $\psi(A)$ given by (16) which measures how well indicator functions $\chi A$, $A \in \sigma(\boldsymbol{Z})$, can be approximated using RKHS functions acting on the random variable $W$ when we penalize with the norm of the RKHS function. The penalization is scaled by the bound on the $L^2$-norm between the random variables $\boldsymbol{Z}$ and $\phi(W)$. The 'size' of the RKHS also factors into the bound. When the RKHS is 'small' then not many indicator functions $\chi A$, $A \in \sigma(\boldsymbol{Z})$, can be approximated well and $\psi(A)$ can be large. On the other hand, if the RKHS 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 $\|\chi D(W) - f(W)\|_2$ in the proposition, can be made arbitrary small. See also Remark 1 in the Appendix.

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

## 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 $\boldsymbol{Z}$ in Section 5.1 using Hilbert-space-valued conditional expectations. In Sections 5.2 and 5.3 we respectively that these features are $\mathcal{H}$-independent of $S$ and that $\boldsymbol{Z}$ is the best $\mathcal{H}$-independent approximation of $\phi(X)$.

### 5.1. Specification of the oblivious features $\boldsymbol{Z}$

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 $\boldsymbol{X} \in \mathcal{L}^2(\Omega, \mathcal{A}, P; \mathcal{H})$, and any $\sigma$-subalgebra $\mathcal{B}$ of $\mathcal{A}$, conditional expectation $E^{\mathcal{B}}\boldsymbol{X}$ 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 $\sigma$-subalgebra which makes $S$ measurable, and we denote this conditional expectation by $E^S\boldsymbol{X}$. In the following, we use the notation $\boldsymbol{X} = \phi(X)$. A natural choice for the new features is

$$\boldsymbol{Z} = \boldsymbol{X} - E^S\boldsymbol{X} + E(\boldsymbol{X}). \quad (5)$$

The expectation $E(\boldsymbol{X})$ is to be interpreted as the Bochner-integral of $\boldsymbol{X}$ given measure $P$. Importantly, if $S$ and $\boldsymbol{X}$ are independent, we have with this choice that $\boldsymbol{Z} = \boldsymbol{X} = \phi(X)$ 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 $\boldsymbol{Z}$.

### 5.2. $\boldsymbol{Z}$ and $S$ are $\mathcal{H}$-independent

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 $g \in \mathcal{L}^2$,

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

Since $E(\boldsymbol{X})$ 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 3 the effect of the move from $\boldsymbol{X}$ 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

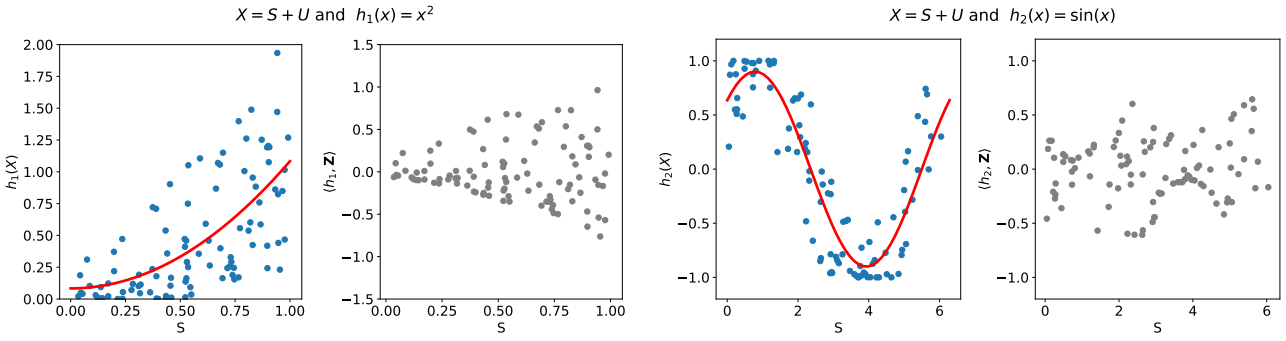


*Figure 3. The figure shows data from two different settings. In the left two plots $X = S + U$, where $S$ and $U$ are independent, $S$ is uniformly distributed on $[0, 1]$ and $U$ is uniformly distributed on $[-1/2, 1/2]$. The function $h_1$ is the quadratic function. The leftmost plot shows $h_1(X)$ 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 two 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 $S$.*

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 $h$.

### 5.3. $\boldsymbol{Z}$ is the best $\mathcal{H}$-independent approximation

Besides being $\mathcal{H}$-independent of $S$ these new features $\boldsymbol{Z}$ also closely approximates our original features $\boldsymbol{X}$ if the influence from $S$ is not too strong, i.e. the mean squared distance is $E(\|\boldsymbol{X} - \boldsymbol{Z}\|^2) = E(\|E^S\boldsymbol{X} - E(\boldsymbol{X})\|^2)$ which is equal to zero if $X$ is independent of $S$. In fact, $\boldsymbol{Z}$ is the best approximation of $\boldsymbol{X}$ 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 2.** *Given $\boldsymbol{X}, \boldsymbol{Z}' \in \mathcal{L}^2(\Omega, \mathcal{A}, P; \mathcal{H})$ such that $\boldsymbol{Z}'$ is $\mathcal{H}$-independent of $S$, then*

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

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

*Proof.* Proof provided in Appendix B.4. □

When replacing $\boldsymbol{X}$ 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

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

where $\|E^S\boldsymbol{X} - E(\boldsymbol{X})\|$ 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 $\boldsymbol{X}$.

## 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 based on an extension of the method in (Grünewälder, 2018). After introducing this approach in Section 6.1 we show in Section 6.2 how the oblivious features can be generated and we introduce the oblivious kernel matrix. In Section 6.3 we discuss how the estimation errors of the plugin-estimator can be controlled. Finally, in Section 6.4, we demonstrate how the approach can be used for statistical problems.

### 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

$$E^S\boldsymbol{X} = g(S) \quad \textit{almost surely}, \tag{6}$$

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 $\delta_{S_i}$ the Dirac measure with mass one at location $S_i$. We define an estimate of the conditional expectation of $\boldsymbol{X}$ given that the sensitive variable falls into a set $B$ by

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

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

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

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

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

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 $E_n^S \boldsymbol{X}$ so that $\langle h, E_n^S \boldsymbol{X}\rangle = E_n^S h(X)$.

### 6.2. Generating an oblivious random variable $\boldsymbol{Z}$

We consider a simple approach where we split our data into two equal parts of size $n$. We use the second $n$ observations to infer the conditional expectation $E_n^S \boldsymbol{X}$ and and $E_n \boldsymbol{X}$. We use the remaining $n$ observations to generate oblivious features through $\boldsymbol{Z}_i = \boldsymbol{X}_i - E_n^{S_i} \boldsymbol{X}_i + E(\boldsymbol{X}), \quad i \leq n$. Most kernel methods work with the kernel matrix and do not need access to the observations themselves. The same holds in the oblivious case. Instead of the original kernel matrix algorithms use the *oblivious kernel matrix*, i.e.

$$\mathcal{O} = \begin{pmatrix} \|\boldsymbol{Z}_1\|^2 & \cdots & \langle \boldsymbol{Z}_1, \boldsymbol{Z}_n\rangle \\ \vdots & \ddots & \vdots \\ \langle \boldsymbol{Z}_n, \boldsymbol{Z}_1\rangle & \cdots & \|\boldsymbol{Z}_n\|^2 \end{pmatrix}. \tag{7}$$

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 $\mathbf{a} \in \mathbb{R}^n$. Importantly, the oblivious kernel matrix can be calculated by using kernel evaluations and we never need to represent $\boldsymbol{Z}$ explicitly in the Hilbert space. The complexity to compute the matrix is $O(n^2)$. See Appendix D for details on the algorithm.

### 6.3. 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,

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

and bounds on the latter term are known. Similarly,

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

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. This $\mathcal{L}^2$-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 $(X_1, S_1), \ldots, (X_n, S_n)$ let $\mathcal{F}_n = \sigma(X_1, S_1, \ldots, X_n, S_n)$ and define the 'conditional' $\mathcal{L}^2$-pseudo-norm by

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

Together with Equation (8) we obtain,

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

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 $\|\cdot\|_{2,n}$. The following proposition demonstrates this by showing that the rate of convergence of the estimator is $n^{-1/2}$, which is optimal.

**Proposition 3.** *Given a continuous kernel function acting on a compact set $\mathbb{X}$, sensitive features $S$ which attain only finitely many values, independent observations $(X_1, S_1), (X_2, S_2), \ldots$, it holds that*

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

*Proof.* The proof is given in Appendix B.5. □

### 6.4. Oblivious ridge regression

In this section we discuss how this approach can be combined with kernel methods. We showcase this 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

$$\hat{h} = \operatorname*{arg\,min}_{h\in\mathcal{H}} \sum_{i=1}^{n} (\langle h, \mathbf{Z}_i\rangle - Y_i)^2 + \lambda \|h\|^2, \qquad (9)$$

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 C 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 $\mathbf{Z}$ and then by evaluating $\langle \mathbf{Z}, \hat{h}\rangle = \sum_{i=1}^{n} \alpha_i \langle \mathbf{Z}, \mathbf{Z}_i\rangle$.

## 7. Examples and experiments

We start with a fundamental example. Let $X$ and $S$ be standard normal random variables with covariance $c \in [-1, 1]$. First, let us 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, $\boldsymbol{Z} = X - E^S X$ 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, we 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 $X^2 - E^S X^2$ and is not independent of $S$. Hence, the kernel function affects the dependence between $\boldsymbol{Z}$ and $S$. Also, within the same RKHS we again have linear functions and $\langle \boldsymbol{Z}, h_{lin}\rangle$ is independent of $S$ for any linear function $h_{lin}$. Therefore, within the same RKHS we can have directions in which $\boldsymbol{Z}$ is independent of $S$ and directions where there are dependencies left.

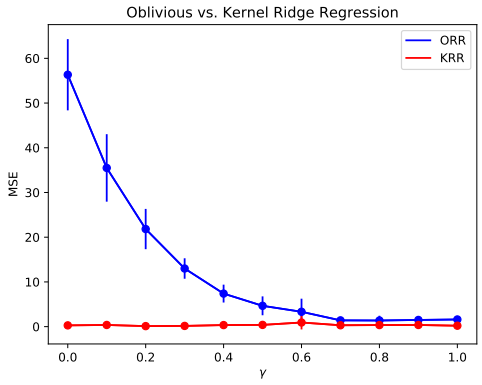


*Figure 4. In this figure we compare KRR with ORR. The $x$-axis corresponds to the coefficient $\gamma$ and the $y$-axis to the MSE over 25 runs; $n = 500$ training samples and $m = 100$ test samples were used. When $\gamma = 0$ there are no non-sensitive features and ORR has a high MSE. As $\gamma$ approaches 1 the effect of the sensitive features vanishes and ORR performs essentially as well as KRR.*

Finally, we compare ORR and KRR in a simple experiment, see Figure 4. We have samples sensitive features $S$ and non-sensitive features $U$ which are both uniformly distributed between $[-5, 5]$ and are independent. The features $X$ are a convex combination of these two, i.e. $X = \gamma U + (1-\gamma)S$, $\gamma \in [0, 1]$. The response variable is $Y = X^2 + \epsilon$, where $\epsilon$ is normally distributed with variance 0.1 and is independent of $U$ and $S$. In particular, for $\gamma = 0$ $X = S$ and ORR behaves poorly in terms of the Mean Squared Error (MSE). On the other hand when $\gamma = 1$ we have $X = U$ and ORR has as much information about $Y$ as the standard KRR. In the plot we can see that the MSE for ORR is slightly higher than the MSE for KRR for high $\gamma$ values. This is due to the empirical estimation errors of the conditional expectations.

## 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. 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 used to further reduce dependencies, and hopefully achieve full independence. Another question concerns the interplay between the errors induced by the empirical estimation of the conditional expectations and those of the kernel methods applied to $\boldsymbol{Z}$.

## References


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*, pages 13–18. IEEE, 2009.

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, NeurIPS*, 2018.

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

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

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

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, NeurIPS*, 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, NeurIPS*, 2016.

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 (ITCS 2017)*, volume 67. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2017.

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

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

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

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

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

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

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

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

M. B. Zafar, I. Valera, Gomez Rodriguez M., 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*, pages 1171–1180, 2017.

# Appendix

## A. Probability in Hilbert spaces: elementary results

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

### A.1. Measurable functions

There are three natural definitions of what it means for a function $\boldsymbol{X} : \Omega \to \mathcal{H}$ 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 $\sigma$-algebra.

1. $\boldsymbol{X}$ is *Bochner-measurable* iff $\boldsymbol{X}$ 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
$$\boldsymbol{S}(\omega) = \sum_{i=1}^{n} h_i \chi A_i(\omega)$$
for some $n \in \mathbb{N}$, $A_1, \ldots, A_n \in \mathcal{A}$ and $h_1, \ldots, h_n \in \mathcal{H}$.
2. $\boldsymbol{X}$ is *strongly-measurable* iff $\boldsymbol{X}^{-1}[B] \in \mathcal{A}$ for every Borel-measurable subset $B$ of $\mathcal{H}$. The topology that is used here is the norm-topology.
3. $\boldsymbol{X}$ is *weakly-measurable* iff for every element $h \in \mathcal{H}$ the function $\langle h, \boldsymbol{X}\rangle : \Omega \to \mathbb{R}$ 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 $\boldsymbol{X} : \Omega \to \mathcal{H}$ a *random variable* if it is measurable in this sense.

## 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 $\sigma$-subalgebra of $\mathcal{A}$. Due to Pisier (2016)[Eq. (1.7)], for any $f \in \mathcal{H}$

$$\langle f, E^{\mathcal{B}}\boldsymbol{X}\rangle = E^{\mathcal{B}}\langle f, \boldsymbol{X}\rangle \qquad \text{(a.s.)} \tag{10}$$

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

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

because $\chi A_i$ 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 $\omega$ 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 $\eta$.

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.

Now, 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 $\eta_n$ 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

$$\lim_{n\to\infty} g_n(x) = \lim_{n\to\infty} g_n(S(\omega)) \tag{11}$$

$$= \lim_{n\to\infty} \eta_n(\omega) \tag{12}$$

$$= \eta(\omega). \tag{13}$$

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

$$g(x) := \begin{cases} \lim_{n\to\infty} g_n(x) & x \in B \\ 0 & x \notin B \end{cases} \tag{14}$$

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

$$\eta(\omega) = \lim_{n\to\infty} \eta_n(\omega) = \lim_{n\to\infty} g_n(S(\omega)) = g(S(\omega)), \tag{15}$$

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 $\boldsymbol{X} : \Omega \to \mathcal{H}$ 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*

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

*Proof.* Observing that by definition of conditional expectation, $E^S \boldsymbol{X}$ is a $\sigma(S)$-measurable function from $\Omega$ to $\mathcal{H}$, the result readily follows from Lemma 1. □

### B.3. Proof of Proposition 1

For $\boldsymbol{Z} \in \mathcal{L}^2(\Omega, \mathcal{A}, P; \mathcal{H})$ fix constants $c_1, c_2 > 0$, a function $h^* \in \mathcal{H}$, and a random variable $W : \Omega \to \mathbb{X}$ with $\phi(W) \in \mathcal{L}^2(\Omega, \mathcal{A}, P; \mathcal{H})$, such that

(i.) for all $A_1 \in \sigma(\boldsymbol{Z})$ there exists an $A_2 \in \sigma(W)$ with $P(A_1 \triangle A_2) \leq c_1$

(ii.) $\|\boldsymbol{Z} - (\phi(W) - h^*)\|_2 \leq c_2$.

**Proposition.** *Suppose that $\mathcal{H}$ is separable and its feature map $\phi$ satisfies $\sup_{x\in\mathbb{X}} \|\phi(x)\| \leq 1$. Consider some $\boldsymbol{Z} \in \mathcal{L}^2(\Omega, \mathcal{A}, P; \mathcal{H})$ that is $\mathcal{H}$-independent from $S$. Let*

$$\psi(A) = \inf_{D\in\mathcal{B}_A} \inf_{f\in\mathcal{H}} 2\|\chi_D(W) - f(W)\|_2 + c_2\|f\| \tag{16}$$

*and let $\mathcal{B}_A = \{W[C] : C \in \sigma(W), P(C \triangle A) \leq c_1\}$ where $W$, $c_1$ and $c_2$ are specified by Notation 1. For any $A \in \sigma(\boldsymbol{Z})$ and $B \in \sigma(S)$ it holds that*

$$|P(A \cap B) - P(A)P(B)| \leq 2c_1 + \psi(A)P(B)^{1/2}.$$

*Proof.* **(a)** Let $\boldsymbol{W} = \phi(W) - h^*$. Observe that two applications of the Cauchy-Schwarz inequality yield

$$E(|\langle f, \boldsymbol{Z}\rangle - f(W) - \langle f, h^*\rangle| \times \chi_B) \leq E|\langle f, (\boldsymbol{Z} - \phi(W) - h^*) \times \chi_B\rangle| \leq \|f\| E(\chi_B \times \|\boldsymbol{Z} - \mathbf{W}\|) \leq P(B)^{1/2}\|f\|\|\boldsymbol{Z} - \mathbf{W}\|_2$$

for all $f \in \mathcal{H}$. Similarly, for any $f \in \mathcal{H}$ it holds that $E|\langle f, \boldsymbol{Z} - \phi(W) - h^*\rangle| \leq \|f\| E\|\boldsymbol{Z} - \mathbf{W}\| \leq \|f\|\|\boldsymbol{Z} - \mathbf{W}\|_2$. Using that $\boldsymbol{Z}$ is $\mathcal{H}$-independent we find that for any $f \in \mathcal{H}$ and $B \in \sigma(S)$

$$\begin{aligned} |E(f(W) \times \chi_B) - Ef(W)P(B)| &= |E((f(W) - \langle f, h^*\rangle) \times \chi_B) - E(f(W) - \langle f, h^*\rangle)P(B)| \\ &\leq 2P(B)^{1/2}\|f\|\|\boldsymbol{Z} - \mathbf{W}\|_2 \leq c_2 P(B)^{1/2}\|f\|. \end{aligned}$$

**(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,

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

Also, $|P(C) - Ef(W)| \leq \xi_C(f)$. Hence, for any $f \in \mathcal{H}$,

$$\begin{aligned}|P(C \cap B) - P(C)P(B)| &\leq 2\xi_C(f)P(B)^{1/2} + |E(f(W) \times \chi B) - Ef(W)P(B)| \\ &\leq (2\xi_C(f) + c_1\|f\|)P(B)^{1/2}.\end{aligned}$$

**(c)** By assumption for $A \in \sigma(\boldsymbol{Z})$ there exists a $C \in \sigma(W)$ such that $P(A \triangle C) \leq c_2$ and we have that $|P(C) - P(A)| \leq P(C \triangle A) \leq c_1$ and

$$|P(C \cap B) - P(A \cap B)| \leq P((C \triangle A) \cap B) \leq c_1.$$

Hence, $|P(A \cap B) - P(A)P(B)| \leq 2c_1 + (2\xi_C(f) + c_2\|f\|)P(B)^{1/2}$ for all $C$ and $f$. Taking the infimum over $f$ and $C$ gives the stated result. □

**Remark 1.** Observe that when $\mathcal{H}$ lies dense in $\mathcal{L}^2(\mathbb{X}, \mathcal{B}, PW^{-1})$ then for any $D$ as in part **(b)** of the proof and any $\epsilon > 0$ there exists a function $f$ such that $\xi_C(f) < \epsilon$, i.e. for the measurable set $D$ there exists a function $f \in \mathcal{H}$ such that

$$\epsilon^2 > \int (\chi D(w) - f(w))^2 \, dPW^{-1}(w) = E(\chi D(W) - f(W))^2$$

using, for example, (Fremlin, 2001)[235Gb].

### B.4. Proof of Proposition 2

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

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

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

*Proof.* **(a)** We first show that

$$\langle E^S\boldsymbol{X}^\bullet, (\boldsymbol{Z}')^\bullet\rangle_2 = \langle E(\boldsymbol{X}^\bullet), (\boldsymbol{Z}')^\bullet\rangle_2. \tag{17}$$

$E^S\boldsymbol{X}^\bullet$ is an element of $L^2(\Omega, \sigma(S), P; \mathcal{H})$ and there exists a sequence of simple function $\{U_n\}_{n\in\mathbb{N}}$ such that $\lim_{n\to\infty}\|U_n^\bullet - \boldsymbol{X}^\bullet\| = 0$. In particular, $\lim_{n\to\infty}\langle U_n^\bullet, (\boldsymbol{Z}')^\bullet\rangle = \langle \boldsymbol{X}^\bullet, (\boldsymbol{Z}')^\bullet\rangle_2$ and $\|E(U_n^\bullet) - E(\boldsymbol{X}^\bullet)\| \leq E\|U_n^\bullet - \boldsymbol{X}^\bullet\| = \|U_n^\bullet - \boldsymbol{X}^\bullet\|$ goes to zero in $n$. Consider some $U_n = \sum_{i=1}^m h_i \times \chi A_i$, $h_i \in \mathcal{H}$, $A_i \in \sigma(S)$, $m \in \mathbb{N}$, 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 $\chi A_i$ is $\sigma(S)$-measurable, and, hence, can be written as a function of $S$ (Shiryaev, 1989)[II.§4.Thm.3]. Now,

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

and $\langle U_n^\bullet, (\boldsymbol{Z}')^\bullet\rangle_2 = \langle E(U_n^\bullet), (\boldsymbol{Z}')^\bullet\rangle_2$. Equation (17) follows since $U_n^\bullet$ converges to $\boldsymbol{X}^\bullet$ and $E(U_n^\bullet)$ converges to $E(\boldsymbol{X}^\bullet) = 0$ in $L^2(\Omega, \mathcal{A}, P; \mathcal{H})$.

**(b)** Since $\langle E^S\boldsymbol{X}^\bullet, (\boldsymbol{Z}')^\bullet\rangle_2 = \langle E(\boldsymbol{X}^\bullet), (\boldsymbol{Z}')^\bullet\rangle_2$ and $\langle \boldsymbol{X}^\bullet, E^S\boldsymbol{X}^\bullet\rangle_2 = \|E^S\boldsymbol{X}^\bullet\|^2$ it follows right away that

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

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

### B.5. Proof of Proposition 3

**Proposition.** *Given a continuous kernel function acting on a compact set $\mathbb{X}$, sensitive features $S$ which attain only finite many values, independent observations $(X_1, S_1), (X_2, S_2), \ldots$, it holds that*

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

*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

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

using Grünewälder (2018). Since there are only $l$-many terms in the sum this result carries over to the whole sum. □

## C. 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

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

Let $\hat{f}$ be the minimizer of the regularized least-squares error as given by (9). By the representer theorem there exist scalars $\alpha_1, \ldots, \alpha_n$ 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,

$$\sum_{i=1}^{n} (\langle \hat{f}, \mathbf{Z}_i\rangle - Y_i)^2 + \lambda \|\hat{f}\|^2 = (\mathcal{O}\boldsymbol{\alpha} - \mathbf{y})(\mathcal{O}\boldsymbol{\alpha} - \mathbf{y})^\top + \lambda \boldsymbol{\alpha}^\top \mathcal{O} \boldsymbol{\alpha} \tag{18}$$

where $\boldsymbol{\alpha} := (\alpha_1, \ldots, \alpha_n)^\top$ and $\mathbf{y} := (Y_1, \ldots, Y_n)^\top$. Noting that $\hat{f}$ is the minimizer, and thus taking the gradient of (18) 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

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

## D. Algorithm for calculating the oblivious kernel matrix

Let $A_1, \ldots, A_l$ be a partition of $\mathbb{S}$ and assume that we have $2n$ samples $(X_i, S_i)$. Use samples $n+1, \ldots, 2n$ to estimate the conditional expectation and the remaining $n$ samples to generate the features $\boldsymbol{Z}_i$. The features $\boldsymbol{Z}_i$ will not be explicitly stored. The only thing that will be stored is the matrix $\mathcal{O}$. To calculate the matrix we only need kernel evaluations. To see this consider any $i \leq n$, then

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

Let for $u = 1, \ldots, l$,

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

be the number of samples that fall into set $A_u$. The basic conditional expectation estimate 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, to show how we can avoid explicitly representing $\boldsymbol{Z}_i$, consider $i, j \leq n$ and

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

This reduces to calculations involving only the kernel

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

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

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

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

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

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 calculated in the same way. The resulting algorithm is presented below.

For predicting values we will need to calculate terms of the form

$$\langle \boldsymbol{Z}, \boldsymbol{Z}_i \rangle$$

where $i \leq n$ and $\boldsymbol{Z}$ is corresponding to a new unseen pair $(X, S)$. The calculations are the same as for $\langle \boldsymbol{Z}_i, \boldsymbol{Z}_j \rangle$.

**Algorithm 1** Generating the oblivious kernel matrix

**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**
  **for** $j = i$ **to** $n$ **do**
    set $\mathcal{O}_{i,j} = k(x_i, x_j)$
    **if** $|\mathcal{I}_a| \cdot |\mathcal{I}_b| > 0$ **then**
      set $a$ such that $s_j \in A_a$
      set $b$ such that $s_i \in A_b$
      set $\mathcal{O}_{i,j} \leftarrow \mathcal{O}_{i,j} - \sum_{u \in \mathcal{I}_a} k(x_i, x_u)/|\mathcal{I}_a|$
      set $\mathcal{O}_{i,j} \leftarrow \mathcal{O}_{i,j} - \sum_{u \in \mathcal{I}_b} k(x_u, x_j)/|\mathcal{I}_b|$
      set $\mathcal{O}_{i,j} \leftarrow \mathcal{O}_{i,j} + \sum_{u \in \mathcal{I}_a, v \in \mathcal{I}_b} k(x_u, x_v)/(|\mathcal{I}_a||\mathcal{I}_b|)$
      set $\mathcal{O}_{i,j} \leftarrow \mathcal{O}_{i,j} + M - \sum_{u=n+1}^{2n} k(x_i, x_u)/n - \sum_{u=n+1}^{2n} k(x_u, x_j)/n$
      set $\mathcal{O}_{i,j} \leftarrow \mathcal{O}_{i,j} - \sum_{u \in \mathcal{I}_a} \sum_{v=n+1}^{2n} k(x_u, x_v)/(n|\mathcal{I}_a|)$
      set $\mathcal{O}_{i,j} \leftarrow \mathcal{O}_{i,j} - \sum_{u=n+1}^{2n} \sum_{v \in \mathcal{I}_b} k(x_u, x_v)/(n|\mathcal{I}_b|)$
    **end if**
    set $\mathcal{O}_{j,i} \leftarrow \mathcal{O}_{i,j}$
  **end for**
**end for**
**Return:** $\mathcal{O}$