Returning The Favour: When Regression Benefits From Probabilistic Causal Knowledge

Shahine Bouabid [*1] **Jake Fawkes** [*1] **Dino Sejdinovic** [2]

Abstract

A directed acyclic graph (DAG) provides valuable prior knowledge that is often discarded in regression tasks in machine learning. We show that the independences arising from the presence of collider structures in DAGs provide meaningful inductive biases, which constrain the regression hypothesis space and improve predictive performance. We introduce *collider regression*, a framework to incorporate probabilistic causal knowledge from a collider in a regression problem. When the hypothesis space is a reproducing kernel Hilbert space, we prove a strictly positive generalisation benefit under mild assumptions and provide closed-form estimators of the empirical risk minimiser. Experiments on synthetic and climate model data demonstrate performance gains of the proposed methodology.

1. Introduction

Causality has recently become a main pillar of research in the machine learning community. Historically, machine learning has been used to help solve problems in the field of causal inference (Shalit et al., 2017; Zhang et al., 2012). But recently a different focus has emerged, asking what causality can do to return the favour to machine learning (Schölkopf et al., 2021). In this work we continue in this vein, and aim to answer whether the knowledge of a causal directed acyclic graph (DAG) underpinning the data generating process can assist and improve performance in regression tasks.

When a causal DAG is available, it constitutes a source of prior knowledge that is typically discarded when addressing a regression problem. It can however guide the setup of the regression problem. Classically, the structure of a DAG informs on which predictors should be selected to regress a given response variable Y. This process, known as feature selection, is solved by selecting the predictors that are either adjacent to Y, or that influence children of Y. The resulting set of predictors is called the Markov boundary of Y (Pearl, 1987).

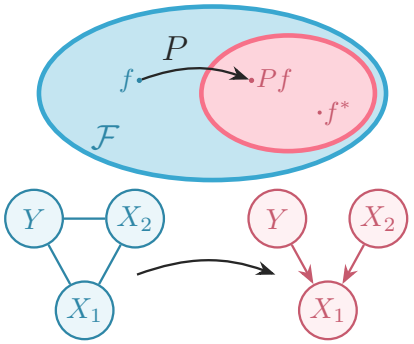

Figure 1. When performing regression in a hypothesis space $\mathcal{F}$ (blue), we implicitly assume that the data generating process could follow any DAG structure. The optimal regressor f^* lies in the subspace of function that satisfy the independence structure arising from the collider (pink), onto which the projection P maps.

As we will see, the presence of a particular structure in a Markov boundary is typically overlooked in regression problems: colliders of the form $Y \to X_1 \leftarrow X_2$. In this work, we investigate how the conditional independence constraints arising due to colliders in the Markov boundary can be used to construct useful inductive biases in a regression problem and to guide the choice of the hypothesis space. We will see that the colliders are also unique in that regard: beyond colliders, the Markov boundary cannot contain any graphical structure implying a conditional independence with Y.

To understand the intuition behind colliders, consider this classic example: imagine we have a randomly timed sprinkler (X_2) and we want to infer whether it has rained (Y), having observed whether the sidewalk is wet (X_1). Although the sprinkler and the rain are marginally independent, knowing whether the sprinkler has been active is important for determining whether it has rained. Colliders arise naturally in many application domains. For example, in climate science, the objective may be to regress an environmental driver Y that, independently from human activity X_2, influences observed global temperatures X_1.

[*]Equal contribution [1]Department of Statistics, University of Oxford, UK [2]School of CMS & AIML, University of Adelaide, Australia. Correspondence to: Shahine Bouabid <shahine.bouabid@stats.ox.ac.uk>, Jake Fawkes <jake.fawkes@stats.ox.ac.uk>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

As illustrated in Figure 1, when performing least-square regression over a hypothesis space $\mathcal{F}$, only a subset of $\mathcal{F}$ will comply with the independences arising from the collider. By considering the projection operator P that maps onto this subspace, we propose a framework called *collider regression* to incorporate inductive biases arising from colliders into any regressor. We show that when the data generating process follows a collider, projecting any given regressor onto this subspace provides a positive generalisation benefit.

We then consider the specific case where the hypothesis space is a reproducing kernel Hilbert space (RKHS). Because RKHSs are rich functional spaces that also enjoy closed analytical solutions to the least-squares regression problem, they allow us to build intuition for the general case. We prove a strictly positive generalisation benefit from projecting the least-squares empirical risk minimiser in a RKHS, where the size of the generalisation gap increases with the complexity of the problem. We also show that for a RKHS, it is possible to solve the least-squares regression problem directly inside the projected hypothesis subspace and provide closed-form estimators.

We experimentally validate the effectiveness of our methodology on a synthetic dataset and on a real world climate science dataset. Results demonstrate that collider regression consistently provides an improvement in generalisation at test time in comparison with standard least-squares regressors. Results also suggest that collider regression is particularly beneficial when few training samples are available, but samples from the covariates can easily be obtained, i.e. in a semi-supervised learning setting.

2. Background

Regression notation Let Y be our target variable over $\mathcal{Y} \subseteq \mathbb{R}$ and X be our covariates over $\mathcal{X}$. Our goal is a standard regression task where we have access to a dataset $\mathcal{D} = \{\mathbf{x}, \mathbf{y}\} \in (\mathcal{X} \times \mathcal{Y})^n$ of n samples $(x^{(i)}, y^{(i)})$ from (X, Y). We aim to minimise the regularised empirical risk

$$\hat{f} = \underset{f \in \mathcal{F}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \left(y^{(i)} - f(x^{(i)}) \right)^2 + \lambda \Omega(f) \qquad (1)$$

where $\mathcal{F}$ is a specified hypothesis space of functions $f : \mathcal{X} \to \mathcal{Y}$, $\lambda > 0$ and $\Omega(f) > 0$ is a regularisation term. This corresponds to finding a function $\hat{f}$ that best estimates the optimal regression function for the squared loss:

$$f^*(x) = \mathbb{E}[Y|X = x]. \qquad (2)$$

For any two functions $h, h' \in \mathcal{F}$, the squared-error generalisation gap between h and h' is defined as the difference in their true risk:

$$\Delta(h, h') = \mathbb{E}[(Y - h(X))^2] - \mathbb{E}[(Y - h'(X))^2]. \qquad (3)$$

Therefore if $\Delta(h, h') \geq 0$, it means that h' generalises better from the training data than h.

Reproducing kernel Hilbert spaces Let $\mathcal{X}$ be some non-empty space. A real-valued RKHS $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ is a complete inner product space of functions $f : \mathcal{X} \to \mathbb{R}$ that admits a bounded evaluation functional. For $x \in \mathcal{X}$, the Riesz representer of the evaluation functional is denoted $k_x \in \mathcal{H}$ and satisfies the *reproducing property* $f(x) = \langle f, k_x \rangle_{\mathcal{H}}$, $\forall f \in \mathcal{H}$. The bivariate symmetric positive definite function defined by $k(x, x') = \langle k_x, k_{x'} \rangle_{\mathcal{H}}$ is referred to as the *reproducing kernel* of $\mathcal{H}$. Conversely, the Moore-Aronszajn theorem (Aronszajn, 1950) shows that any symmetric positive definite function k is the unique reproducing kernel of an RKHS. For more details on RKHS theory, we refer the reader to Berlinet & Thomas-Agnan (2011).

Conditional Mean Embeddings Conditional mean embeddings (CMEs) provide a powerful framework to represent conditional distributions in a RKHS (Fukumizu et al., 2004; Song et al., 2013; Muandet et al., 2016). Given random variables X, Z on $\mathcal{X}, \mathcal{Z}$ and an RKHS $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}}$ with reproducing kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, the CME of $\mathbb{P}(X|Z = z)$ is defined as

$$\mu_{X|Z=z} = \mathbb{E}[k_X | Z = z] \in \mathcal{H}. \qquad (4)$$

It corresponds to the Riesz representer of the conditional expectation functional $f \mapsto \mathbb{E}[f(X)|Z = z]$ and can thus be used to evaluate conditional expectations by taking an inner product $\mathbb{E}[f(X)|Z = z] = \langle f, \mu_{X|Z=z} \rangle_{\mathcal{H}}$.

Introducing a second RKHS $\mathcal{G} \subseteq \mathbb{R}^{\mathcal{Z}}$ with reproducing kernel $\ell : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$, Grünewälder et al. (2012) propose an alternative view of CMEs as the solution to the least-square regression of canonical feature maps ℓ_Z onto k_X

$$\begin{cases} E^* = \underset{C \in \mathsf{B}_2(\mathcal{G}, \mathcal{H})}{\arg\min} \mathbb{E}[\|k_X - C\ell_Z\|^2_{\mathcal{H}}] \\ \mu_{X|Z=z} = E^* \ell_z \end{cases} \qquad (5)$$

where $\mathsf{B}_2(\mathcal{G}, \mathcal{H})$ denotes the space of Hilbert-Schmidt operators[1] from $\mathcal{G}$ to $\mathcal{H}$. Given a dataset $\mathcal{D} = \{\mathbf{x}, \mathbf{z}\}$, this perspective allows to compute an estimate of the associated operator $E^* : \mathcal{G} \to \mathcal{H}$ as the solution to the regularised empirical least-squares problem as

$$\begin{cases} \hat{E}^* = \underset{C \in \mathsf{B}_2(\mathcal{G}, \mathcal{H})}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \|k_{x^{(i)}} - C\ell_{z^{(i)}}\|^2_{\mathcal{H}} + \gamma \|C\|^2_{\mathsf{B}_2} \\ \quad = \boldsymbol{k}_{\mathbf{x}}^\top (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \boldsymbol{\ell}_{\mathbf{z}} \\ \hat{\mu}_{X|Z=z} = \hat{E}^* \ell_z = \boldsymbol{k}_{\mathbf{x}}^\top (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \boldsymbol{\ell}_{\mathbf{z}}(z) \end{cases} \qquad (6)$$

where $\gamma > 0$, $\mathbf{L} = \ell(\mathbf{z}, \mathbf{z})$, $\boldsymbol{k}_{\mathbf{x}} = k(\mathbf{x}, \cdot)$ and $\boldsymbol{\ell}_{\mathbf{z}} = \ell(\mathbf{z}, \cdot)$. We refer the reader to (Muandet et al., 2017) for a comprehensive review of CMEs.

[1] i.e. bounded operators $A : \mathcal{G} \to \mathcal{H}$ such that $\mathrm{Tr}(A^* A) < \infty$. $\mathsf{B}_2(\mathcal{G}, \mathcal{H})$ has a Hilbert space structure for the inner product $\langle A, B \rangle_{\mathsf{B}_2} = \mathrm{Tr}(A^* B)$.

3. DAG inductive biases for regression

In this section, we aim to answer how knowledge of the causal graph of the underlying data generating process can help to perform regression. We start by reviewing the concept of Markov boundaries and how it is used for feature selection. We then show that even after feature selection has been performed, there is still residual information from colliders that is relevant for a regression problem.

3.1. Markov boundary for feature selection

Since we are focusing on regression, we are interested in how the DAG can inform us about $\mathbb{P}(Y|X)$. Suppose that for some vertex X_i, the DAG informs us that $Y \perp\!\!\!\perp X_i \mid X \setminus X_i$. Stated in terms of mutual information we have that[2] $I(Y;X) = I(Y;X \setminus X_i)$, therefore we can discard X_i from our set of covariates without any loss of probabilistic information for $\mathbb{P}(Y|X)$.

From a functional perspective, we can interpret this as incorporating the inductive bias that the regressor need only depend on $X \setminus X_i$, allowing us to learn simpler functions which should generalise better from the training set.

By repeating the process of removing features, we can iteratively construct a minimal set of necessary covariates that still retain all the probabilistic information about $\mathbb{P}(Y|X)$. This is known as feature selection (Dash & Liu, 1997).

Such a set, S, should satisfy $Y \perp\!\!\!\perp X \setminus S|S$ and we should not be able to remove a vertex from S without losing information about $\mathbb{P}(Y|X)$. A set of this form is known as the Markov boundary of Y (Statnikov et al., 2013), denoted by $\mathrm{Mb}(Y)$. If the only independences in the distribution are those implied by the DAG structure[3] then the Markov boundary is uniquely given by

$$\mathrm{Mb}(Y) = \mathrm{Pa}(Y) \cup \mathrm{Ch}(Y) \cup \mathrm{Sp}(Y), \tag{7}$$

where $\mathrm{Pa}(Y)$ are the parents of Y, $\mathrm{Ch}(Y)$ are the children of Y and $\mathrm{Sp}(Y)$ are the spouses of Y, i.e. the children's other parents. In Figure 2 the Markov boundary of Y is highlighted in blue.

3.2. Extracting inductive bias for regression

By construction the Markov boundary of Y cannot contain independence relationships of the form $Y \perp\!\!\!\perp X_i|X \setminus X_i$. However, it can still contain unused independence statements that involve Y, and therefore provides useful information about the conditional distribution $\mathbb{P}(Y|X)$.

For example, the graphical structure in Figure 2 gives

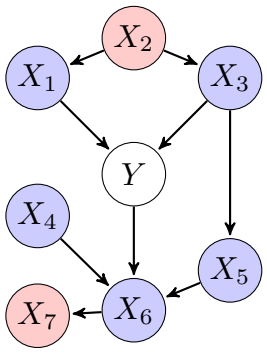

Figure 2. A causal graph with the Markov boundary of Y highlighted in blue and vertices outside the Markov boundary highlighted in red. Whilst Y and X_4 are marginally independent, the presence of the collider X_6 opens the path between Y and X_4.

that $Y \perp\!\!\!\perp X_4$ and $Y \perp\!\!\!\perp X_5 \mid X_3$. This implies that $\mathbb{P}(Y|X_4) = \mathbb{P}(Y)$ and $\mathbb{P}(Y|X_3, X_5) = \mathbb{P}(Y|X_3)$ which by marginalisation constrains $\mathbb{P}(Y|X)$ and so gives us extra information about it. The presence of these independence relationships inside $\mathrm{Mb}(Y)$ is only possible because a collider, X_6, has allowed for the spouses X_4 and X_5 to be within the Markov boundary without being adjacent to Y.

Hence, the presence of collider structures within the Markov boundary of Y provides additional independence relationships involving Y. The following proposition shows that the presence of a collider is not only a sufficient condition, but also necessary.

Proposition 3.1. *The Markov boundary of Y contains a collider if and only if there exists $Z \in \mathrm{Mb}(Y)$ and $S_Z \subset \mathrm{Mb}(Y)$ such that $Y \perp\!\!\!\perp Z \mid S_Z$.*

Proof. We have a conditional independence between two variables if and only if they are not adjacent (Lemma 3.1, 3.2 Koller & Friedman (2009)) and $\mathrm{Mb}(Y)$ contains a variable not adjacent to Y if and only if it contains a collider. □

The collider structures are thus the only graphical structures that provide conditional independence statement relevant to $\mathbb{P}(Y|X)$ within the Markov boundary. To the best of our knowledge, this information is currently left unused when addressing a regression problem.

However, unlike for the feature selection process, we cannot simply use these independence statements to discard covariates and reduce the set of features. This is because while the spouses of Y are uninformative on their own, they become informative in the presence of other covariates. Namely in Figure 2, while $Y \perp\!\!\!\perp X_4$ we have $Y \not\perp\!\!\!\perp X_4|X_6$ because X_6 is a collider. Therefore, we have that $I(Y;X) > I(Y;X \setminus X_4)$ and discarding X_4 would constitute a loss of information.

[2] This follows from $I(Y;X) = I(Y;X\setminus X_i)+I(Y;X_i|X\setminus X_i)$ and the conditional independence gives $I(Y;X_i|X \setminus X_i) = 0$.

[3] An assumption known as faithfulness (Meek, 1995) which we take throughout.

4. Collider Regression

In this section, we present a method for incorporating probabilistic inductive bias from a collider structure into a regression problem, and provide guarantees of improved generalisation error. For the sake of clarity, our exposition focuses on the simple collider structure depicted in Figure 3. We however emphasise this simplification does not harm the generality of our contribution and Section 5 shows how collider regression can be extended to more general DAGs.

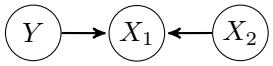

Figure 3. Simple collider structure

4.1. Simple collider regression setup

Let X_1, X_2, Y be random variables following the DAG structure in Figure 3 and taking values in $\mathcal{X}_1 \subseteq \mathbb{R}^{d_1}$, $\mathcal{X}_2 \subseteq \mathbb{R}^{d_2}$ and $\mathcal{Y} \subseteq \mathbb{R}$ respectively. Without loss of generality, we assume that $\mathbb{E}[Y] = 0$.

Under the squared loss, the optimal regressor is given by

$$f^*(x_1, x_2) = \mathbb{E}[Y|X_1 = x_1, X_2 = x_2]. \tag{8}$$

Since the collider gives the independence relationship $Y \perp\!\!\!\perp X_2$, we have that

$$\begin{aligned} \mathbb{E}[f^*(X_1, X_2)|X_2] &= \mathbb{E}\big[\mathbb{E}[Y|X_1, X_2] \mid X_2\big] \\ &= \mathbb{E}[Y|X_2] \\ &= \mathbb{E}[Y] \\ &= 0, \end{aligned} \tag{9}$$

where the second line comes from the tower property of the conditional expectation.

Hence, the optimal regressor f^* lies in the subspace of functions that have zero X_2-conditional expectation. To incorporate the knowledge from the DAG into our regression procedure, we should therefore ensure that our estimate $\hat{f}$ lies within the same subspace of functions, i.e. we want to satisfy the zero conditional expectation constraint

$$\hat{f} \in \big\{f \in \mathcal{F} \mid \mathbb{E}[f(X_1, X_2)|X_2] = 0\big\}. \tag{ZCE}$$

We propose to investigate how such a constraint can be enforced onto our hypothesis and how it benefits generalisation, starting by the general case of square-integrable functions. In what follows, we will use shorthand concatenated notations $X = (X_1, X_2)$, $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$, $x = (x_1, x_2) \in \mathcal{X}$ and $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2) \in \mathcal{X}^n$.

4.2. Respecting the collider structure in the hypothesis

Let $L^2(X)$ denote the space of square-integrable functions with respect to the probability measure induced by X and suppose $\mathcal{F} = L^2(X)$. Let $E : L^2(X) \to L^2(X)$ denote the conditional expectation operator defined by

$$Ef(x_1, x_2) = \mathbb{E}[f(X_1, X_2)|X_2 = \pi_2(x_1, x_2)], \tag{10}$$

where $\pi_2(x_1, x_2) = x_2$ is simply the mapping that discards the first component[4].

The operator E classically defines an orthogonal projection over the subspace of X_2-measurable functions. $L^2(X)$ thus orthogonally decomposes into its image, denoted $\mathrm{Range}(E)$, and its null-space, denoted $\mathrm{Ker}(E)$, as

$$L^2(X) = \mathrm{Ker}(E) \oplus \mathrm{Range}(E). \tag{11}$$

Using this notation, satisfying condition (ZCE) corresponds to having $\hat{f} \in \mathrm{Ker}(E)$. Alternatively, if we denote

$$P = \mathrm{Id} - E, \tag{12}$$

the orthogonal projection onto $\mathrm{Ker}(E)$, then we want to take $\mathcal{F} = \mathrm{Range}(P)$ as our hypothesis space.

In general, it may be hard to constrain the hypothesis space directly to be $\mathrm{Range}(P)$. However, the solution to the empirical risk minimisation problem (1) will always orthogonally decompose within $L^2(X)$ as

$$\hat{f} = P\hat{f} + E\hat{f}, \tag{13}$$

where only $P\hat{f} \in \mathrm{Range}(P)$ satisfies (ZCE). It turns out that discarding $E\hat{f}$ — the part that does not satisfy the constraint — will always yield generalisation benefits.

Proposition 4.1. *Let $h \in L^2(X)$ be any regressor from our hypothesis space. We have*

$$\Delta(h, Ph) = \|Eh\|^2_{L^2(X)}. \tag{14}$$

The generalisation gap is always greater than zero. Hence, for any given regressor $\hat{f}$, we can always improve its test performance by projecting it onto $\mathrm{Range}(P)$.

In practice, a simple estimator of $P\hat{f}$ can be obtained by subtracting an estimate of $\mathbb{E}[\hat{f}(X_1, X_2)|X_2]$ as

$$\hat{P}\hat{f}(x_1, x_2) = \hat{f}(x_1, x_2) - \hat{\mathbb{E}}[\hat{f}(X_1, X_2)|X_2 = x_2] \tag{15}$$

by following the procedure outlined in Algorithm 1.

It is worth noting that the second step of Algorithm 1 does not require observations from Y. As such, it naturally fits a semi-supervised setup where additional observations $\mathcal{D}' = \{\mathbf{x}'_1, \mathbf{x}'_2\}$ are available, and can be used to produce a better estimate of the conditional expectation $\mathbb{E}[\hat{f}(X_1, X_2)|X_2]$.

[4]This notation emphasises that Ef is formally a function of (x_1, x_2) and belongs in $L^2(X)$

Algorithm 1 General procedure to estimate $P\hat{f}$

1: Regress $(X_1, X_2) \to Y$
to get $(x_1, x_2) \mapsto \hat{f}(x_1, x_2)$
2: Regress $X_2 \to \hat{f}(X_1, X_2)$
to get $x_2 \mapsto \hat{\mathbb{E}}[\hat{f}(X_1, X_2)|X_2 = x_2]$
3: Take $\hat{P}\hat{f}(x_1, x_2) = \hat{f}(x_1, x_2) - \hat{\mathbb{E}}[\hat{f}(X_1, X_2)|X_2 = x_2]$

4.3. Theoretical guarantees in a RKHS

RKHSs are mathematically convenient functional spaces and under mild assumptions on the reproducing kernel, they can be proven to be dense in $L^2(X)$ (Sriperumbudur et al., 2011). This makes them a powerful tool for theoretical analysis and building intuition which can be expected to carry over to more general function spaces. For this reason, in this section we study the case where the hypothesis space is a RKHS $\mathcal{F} = \mathcal{H}$. We denote its inner product by $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and its reproducing kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$.

When solving the least-square regression problem in a RKHS, it is known that for Tikhonov regularisation $\Omega(f) = \|f\|^2_{\mathcal{H}}$, the solution to the empirical risk minimisation problem (1) in $\mathcal{H}$ enjoys a closed-form expression given by

$$\hat{f} = \mathbf{y}^\top \left(\mathbf{K} + \lambda \mathbf{I}_n\right)^{-1} \boldsymbol{k}_{\mathbf{x}}, \tag{16}$$

where $\mathbf{K} = k(\mathbf{x}, \mathbf{x})$ and $\boldsymbol{k}_{\mathbf{x}} = k(\mathbf{x}, \cdot)$.

Therefore, if we now project $\hat{f}$ onto $\mathrm{Range}(P)$ as previously, the projected empirical risk minimiser writes

$$P\hat{f} = \mathbf{y}^\top \left(\mathbf{K} + \lambda \mathbf{I}_n\right)^{-1} P\boldsymbol{k}_{\mathbf{x}} \tag{17}$$

with notation abuse $P\boldsymbol{k}_{\mathbf{x}} = [Pk_{x^{(1)}} \dots Pk_{x^{(n)}}]^\top$.

Leveraging these analytical expressions, the following result establishes a strictly non-zero generalisation benefit from projecting $\hat{f}$. The proof techniques follows that of Elesedy (2021), but is adapted to our particular setup with relaxing assumptions about the projection orthogonality[5] and the form of the data generating process.

Theorem 4.2. *Suppose $M = \sup_{x \in \mathcal{X}} k(x, x) < \infty$ and $\mathrm{Var}(Y|X) \geq \eta > 0$. Then, the generalisation gap between $\hat{f}$ and $P\hat{f}$ satisfies*

$$\mathbb{E}[\Delta(\hat{f}, P\hat{f})] \geq \frac{\eta \mathbb{E}\left[\|\mu_{X|X_2}(X)\|^2_{L^2(X)}\right]}{\left(\sqrt{n}M + \lambda/\sqrt{n}\right)^2} \tag{18}$$

where $\mu_{X|X_2} = \mathbb{E}[k_X|X_2]$ is the CME of $\mathbb{P}(X|X_2)$.

This demonstrates that in a RKHS, projecting the empirical risk minimiser is strictly beneficial in terms of generalisation error. Specifically, if there exists a set with non-zero measure on which $Y \neq 0$ and $\mu_{X|X_2} \neq 0$ almost-everywhere, then the lower bound is strictly positive.

The magnitude of the lower bound depends on the variance of $\|\mu_{X|X_2}(X)\|_{L^2(X)}$ and the lower bound on $\mathrm{Var}(Y|X)$. This indicates that problems with more complex conditional distributions $\mathbb{P}(X|X_2)$ and $\mathbb{P}(Y|X)$ should enjoy a larger generalisation gap.

The theorem also suggests that the lower bound on the generalisation benefit decreases at the rate $\mathcal{O}(1/n)$ as the number of samples n grows. Since for the well-specified kernel ridge regression problem, the excess risk upper bound also decreases at rate $\mathcal{O}(1/n)$ (Bach, 2021; Caponnetto & De Vito, 2007), we have that $\mathbb{E}[\Delta(\hat{f}, P\hat{f})] = \Theta(1/n)$.

In a RKHS, $P\hat{f}$ can be rewritten using CMEs as

$$Pf(x_1, x_2) = f(x_1, x_2) - \langle f, \mu_{X|X_2=x_2} \rangle_{\mathcal{H}}. \tag{19}$$

Therefore, introducing a kernel $\ell : \mathcal{X}_2 \times \mathcal{X}_2 \to \mathbb{R}$, the CME estimate from (6) allows to devise an estimator of $P\hat{f}$ as:

$$\hat{P}\hat{f} = \mathbf{y}^\top (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \left(\boldsymbol{k}_{\mathbf{x}} - \mathbf{K}(\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \boldsymbol{\ell}_{\mathbf{x}_2}\right) \tag{20}$$

where $\mathbf{L} = \ell(\mathbf{x}_2, \mathbf{x}_2)$, $\boldsymbol{\ell}_{\mathbf{x}_2} = \ell(\mathbf{x}_2, \cdot)$ and $\gamma > 0$.

4.4. Respecting the collider structure in a RKHS

Similarly to the $L^2(X)$ case, the solution to the empirical risk minimisation problem in $\mathcal{H}$ will also decompose as $\hat{f} = P\hat{f} + E\hat{f}$. Thus, we can proceed similarly by simply discarding $E\hat{f}$ to improve performance. However, it turns out that using elegant functional properties of RKHSs, it is possible to take a step further and directly take $\mathcal{F} = \mathrm{Range}(P)$. In doing so, we can ensure that our hypothesis space only contains functions that satisfy constraint (ZCE).

Under assumptions detailed in Appendix C, we can view the projection P as a well-defined RKHS projection[5] $P : \mathcal{H} \to \mathcal{H}$. In particular, an important assumption is that the kernel takes the form

$$k\left(x, x'\right) = \left(r\left(x_1, x_1'\right) + 1\right) \ell\left(x_2, x_2'\right), \tag{21}$$

where $r : \mathcal{X}_1 \times \mathcal{X}_1 \to \mathbb{R}$ and $\ell : \mathcal{X}_2 \times \mathcal{X}_2 \to \mathbb{R}$ are also positive semi-definite kernels. This ensures that $\mathcal{H}$ contains functions that are constant with respect to x_1. Thus, the conditional expectation mapping $(x_1, x_2) \mapsto \mathbb{E}[f(X_1, X_2)|X_2 = x_2]$ belongs to the same RKHS.

If these assumptions are met, we denote $\mathcal{H}_P = \mathrm{Range}(P)$. The following result characterises $\mathcal{H}_P$ as a RKHS.

Proposition 4.3. *Let P^* be the adjoint operator of P in $\mathcal{H}$. Then $\mathcal{H}_P$ is also a RKHS with reproducing kernel*

$$k_P(x, x') = \langle P^* k_x, P^* k_{x'} \rangle_{\mathcal{H}} \tag{22}$$

with $P^* k_x = k_x - \mu_{X|X_2=\pi_2(x)}$.

[5] P is not necessarily orthogonal anymore as a projection of $\mathcal{H}$

[5] E then corresponds to what is referred to as a conditional mean operator in the kernel literature (Fukumizu et al., 2004).

Using the projected RKHS kernel k_P, it becomes possible to solve the least-square regression problem directly inside $\mathcal{F} = \mathcal{H}_P$. By taking $\Omega(f) = \|f\|^2_{\mathcal{H}_P}$, the empirical risk minimisation problem becomes a standard kernel ridge regression problem in $\mathcal{H}_P$ which admits closed-form solution

$$\hat{f}_P = \mathbf{y}^\top \left(\mathbf{K}_P + \lambda \mathbf{I}_n\right)^{-1} \boldsymbol{k}_{P,\mathbf{x}}, \tag{23}$$

where $\mathbf{K}_P = k_P(\mathbf{x}, \mathbf{x})$ and $\boldsymbol{k}_{P,\mathbf{x}} = k_P(\mathbf{x}, \cdot)$.

From a learning theory perspective, performing empirical risk minimisation inside $\mathcal{H}_P$ should provide tighter bounds on the generalisation error than on the entire space $\mathcal{H}$. This is because since $\mathcal{H}_P \subset \mathcal{H}$, the Rademacher complexity of $\mathcal{H}_P$ is smaller than that of $\mathcal{H}$.

It should be noted that k_P depends on the CME $\mu_{X|X_2=\pi_2(x)}$, which needs to be estimated. Therefore, in practice, our hypothesis will not lie in the true $\mathcal{H}_P$ but in an approximation of $\mathcal{H}_P$ and the approximation error will depend directly on the CME estimation error.

Algorithm 2 RKHS procedure to estimate $\hat{f}_P$

1: Let $\hat{P}^* k_x = k_x - \hat{\mu}_{X|X_2=\pi_2(x)}$
2: Let $\hat{k}_P(x, x') = \langle \hat{P}^* k_x, \hat{P}^* k_{x'} \rangle_{\mathcal{H}}$
3: Evaluate $\hat{\mathbf{K}}_P = \hat{k}_P(\mathbf{x}, \mathbf{x})$ and $\hat{\boldsymbol{k}}_{P,\mathbf{x}} = \hat{k}_P(\mathbf{x}, \cdot)$
4: Take $\hat{f}_{\hat{P}} = \mathbf{y}^\top (\hat{\mathbf{K}}_P + \lambda \mathbf{I}_n)^{-1} \hat{\boldsymbol{k}}_{P,\mathbf{x}}$

The estimation of (23) is again a two-stage procedure outlined in Algorithm 2. The distinction with the general $L^2(X)$ case is that we do not estimate the conditional expectation of any specific function. Instead, we estimate the conditional expectation operator through $\hat{\mu}_{X|X_2=x_2}$, and then use it through $\hat{P}^*$ to constrain the hypothesis space. This is possible because in a RKHS, the estimation of the conditional expectation operator can be achieved independently from the function it is applied to. Due to the assumption on the kernel introduced in equation 21 there are now alternative estimators for $\hat{\mu}_{X|X_2=x_2}$ which we provide details of in Appendix D.

The estimation of P^* in line 1 only requires observations from X_1, X_2. Thus, like in the $L^2(X)$ case, additional observations $\mathcal{D}' = \{\mathbf{x}'_1, \mathbf{x}'_2\}$ can help better estimate CMEs, and thus better approximate the projected RKHS $\mathcal{H}_P$.

5. Collider Regression on a more general DAG

We now return to a general Markov boundary. Any Markov boundary may be partitioned following Figure 5, where X_1 contains all direct children of Y, X_3 contains all parents of Y and all other variables are grouped in X_2. Furthermore, we assume that there exists no edge from a variable in X_1 to a variable in X_2.

This provides us with the probabilistic information that $Y \perp\!\!\!\perp X_2 \mid X_3$ but $Y \not\perp\!\!\!\perp X_2 \mid X_3, X_1$, which implies in expectation that $\mathbb{E}[Y|X_3] = \mathbb{E}[Y|X_2, X_3]$.

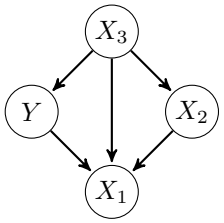

Figure 5. General Markov boundary collider structure.

If we now denote $X = (X_1, X_2, X_3)$ and $f_0(x) = \mathbb{E}[Y|X_3 = x_3]$, then the optimal least-square regressor $f^*(x) = \mathbb{E}[Y|X = x]$ satisfies

$$\begin{aligned} &\mathbb{E}\big[f^*(X) - f_0(X) \mid X_2, X_3\big] \\ &= \mathbb{E}\left[\mathbb{E}[Y|X] \mid X_2, X_3\right] - \mathbb{E}\left[\mathbb{E}[Y|X_3] \mid X_2, X_3\right] \\ &= \mathbb{E}[Y|X_2, X_3] - \mathbb{E}\left[\mathbb{E}[Y|X_2, X_3] \mid X_2, X_3\right] \\ &= 0. \end{aligned} \tag{24}$$

Therefore, if we center our hypothesis space on f_0, then like in Section 4.1, we want our centered estimate $\hat{f} - f_0$ to lie within the following subspace:

$$\hat{f} - f_0 \in \big\{f \in \mathcal{F} \mid \mathbb{E}\big[f(X) \mid X_2, X_3\big] = 0\big\}. \tag{25}$$

When $\mathcal{F} = L^2(X)$, this space can again be seen as the range of an orthogonal projection, this time defined by

$$P' = \mathrm{Id} - E' \tag{26}$$

where $E' : L^2(X) \to L^2(X)$ denotes the conditional expectation functional with respect to (X_2, X_3)

$$E' f(x_2, x_3) = \mathbb{E}[f(X)|X_2 = x_2, X_3 = x_3]. \tag{27}$$

While we focus in Section 4 on the simple collider structure for the sake of exposition, our result are stated for a general projection operator and still hold for P' — modulo a shift by f_0. Hence, we can still apply the techniques we have presented to encode probabilistic information from the general DAG in Figure 5 into a regression problem, with similar guarantees on the generalisation benefits.

Proposition 5.1. *Let $h \in L^2(X)$ be any regressor from our hypothesis space. We have*

$$\Delta(h, f_0 + P'h) = \|E'h - f_0\|^2_{L^2(X)}. \tag{28}$$

This means that, for any given regressor $\hat{f}$, we can always improve its test performance by first projecting it onto $\mathrm{Range}(P)$, and then shifting it by f_0.

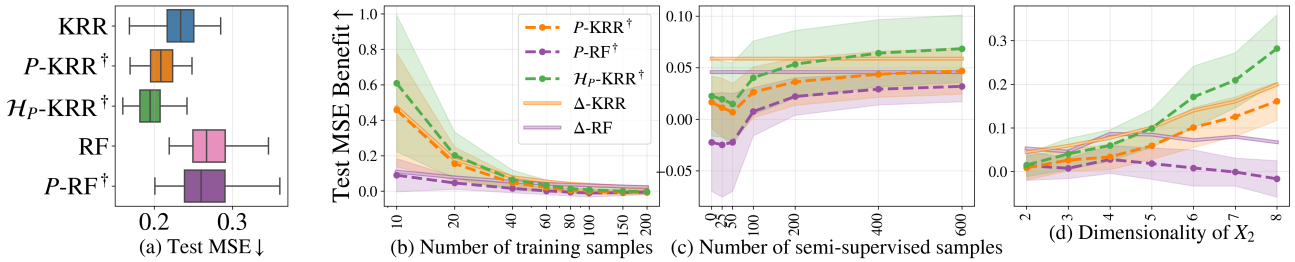

Figure 4. (a) : Test MSEs for the simulation experiment ; dataset is generated using $d_1 = 3$, $d_2 = 3$, $n = 50$ and 100 semi-supervised samples ; experiments is run for 100 datasets generated with different seeds ; statistical significance is confirmed in Appendix F ; (b, c, d) : Ablation study on the number of training samples, number of semi-supervised samples and dimensionality of X_2 ; experiments are run for 40 datasets generated with different seeds ; $\uparrow/\downarrow$ indicates higher/lower is better ; we report 1 s.d. ; † indicates our proposed methods.

In practice, the estimation strategies introduced in Section 4 can still be applied to obtain an estimate of $P'\hat{f}$. An additional procedure to estimate f_0 will however be needed. This can be achieved by regressing Y onto X_3. We provide corresponding algorithms and estimators in Appendix E.

6. Experiments

This section provides empirical evidence that incorporating probabilistic causal knowledge into a regression problem benefits performance. First, we demonstrate our method on an illustrative simulation example. We conduct an ablation study on the number of training samples, the dimensionality of X_2 and the use of additional semi-supervised samples. Then, we address a challenging climate science problem that respects the collider structure. Our results underline the benefit of enforcing constraint (ZCE) onto the hypothesis. Code and data are made available[6].

Models We compare five models:

1. *RF*: A baseline random forest model.
2. *P-RF*: The baseline RF model projected following Algorithm 1 and using a linear regression to estimate $\hat{\mathbb{E}}[\hat{f}(X_1, X_2)|X_2 = x_2]$.
3. *KRR*: A baseline kernel ridge regression.
4. *P-KRR*: The KRR model projected following (20).
5. *$\mathcal{H}_P$-KRR*: A kernel ridge regression model fitted directly in the projected RKHS following Algorithm 2.

For both KRR and RF, we use Proposition 4.1 to compute Monte Carlo estimates of the expected generalisation gap $\mathbb{E}[\Delta(\hat{f}, P\hat{f})]$, which we denote as Δ-KRR and Δ-RF respectively. This provides an indicator of the greatest achievable generalisation gain if we had access to the exact projection P. Hyperparameters are tuned using a cross-validated grid search and model details are specified in Appendix F.

[6] https://github.com/shahineb/collider-regression.

6.1. Simulation example

Data generating process We propose the following construction that follows the simple collider structure from Figure 3. Let $d_1, d_2 \geq 1$ denote respectively the dimensionalities of X_1 and X_2. We first generate a fixed positive definite matrix Σ of size $(d_1 + d_2 + 1)$ which has zero off-diagonals on the $(d_1 + d_2)^{\text{th}}$ row and column . We then follow the generating process described in Algorithm 3 and generate a dataset of n observations $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}\}$. The zero off-diagonal terms in Σ ensure that we satisfy $Y \perp\!\!\!\perp X_2$ and g_1, g_2 are nontrivial mappings that introduce a non-linear dependence (details in Appendix F).

Algorithm 3 Data generating process simulation example

Input: $\Sigma \succcurlyeq 0$, $\sigma > 0$, $g_1 : \mathbb{R}^{d_1} \to \mathbb{R}^{d_1}$, $g_2 : \mathbb{R}^{d_2} \to \mathbb{R}^{d_2}$
$\begin{bmatrix} X_1 & X_2 & Y \end{bmatrix}^\top \sim \mathcal{N}(0, \Sigma)$, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$
$X_1 \leftarrow g_1(X_1) + \varepsilon$
$X_2 \leftarrow g_2(X_2)$
return X_1, X_2, Y

Results Figure 4(a) provides empirical evidence that, for both KRR and RF, incorporating probabilistic inductive biases from the collider structure in the hypothesis benefits the generalisation error.

In addition, Figure 4(b)(c)(d) shows that the empirical generalisation benefit is greatest when : fewer training samples are available, semi-supervised samples can be easily obtained and the dimensionality of X_2 is larger. This is in keeping with Theorem 4.2 which predicts the benefit will be larger when we have fewer labeled samples and a more complicated relationship between X_2 and X_1.

Because the decision nodes learnt by RF largely rely on X_1 and the early dimensions of X_2, increasing the dimensionality of X_2 has little to negative effect as shown in Figure 4(d).

6.2. Aerosols radiative forcing

Background The radiative forcing is defined as the difference between incoming and outgoing flux of energy in the Earth system. At equilibirum, the radiative forcing should be of 0 W m^{-2}. Carbon dioxide emissions from human activity contribute a positive radiative forcing of +1.89 W m^{-2} which causes warming of the Earth (Bellouin et al., 2020).

Aerosols are microscopic particles suspended in the atmosphere (e.g. dust, sea salt, black carbon) that contribute a negative radiative forcing by helping reflect solar radiation, which cools the Earth. However, the magnitude of their forcing represents the largest uncertainty in assessments of global warming, with uncertainty bounds that could offset global warming or double its effects. It is thus critical to obtain better estimate of the aerosol radiative forcing.

The carbon dioxide and aerosol forcings are independent factors[7] that contribute to the observed global temperatures. Hence, by setting $Y =$ "aerosol forcing", $X_2 =$"CO_2 forcing" and $X_1 =$ "global temperature", this problem has a collider structure and observations from global temperature and CO_2 forcing can be used to regress the aerosol forcing.

Data generating process FaIR (for Finite amplitude Impulse Response) is a deterministic model that proposes a simplified low-order representation of the climate system (Millar et al., 2017; Smith et al., 2018). Surrogate climate models like FaIR — referred to as *emulators* — have been widely used, notably in reports of the Intergovernmental Panel on Climate Change (Masson-Delmotte et al., 2021), because they are fast and inexpensive to compute.

We use a modified version of FaIRv2.0.0 (Leach et al., 2021) where we introduce variability by adding white noise on the forcing to account for climate internal variability (Hasselmann, 1976; Cummins et al., 2020). To generate a sample, we run the emulator over historical greenhouse gas and aerosol emission data and retain scalar values for $y =$ "aerosol forcing in 2020", $x_2 =$ "CO_2 forcing in 2020" and $x_1 =$ "global temperature anomaly in 2020". We perform this n times to generate dataset $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}\}$.

Results Results are reported in Table 1. We observe that the incorporation of inductive bias from the collider resulted in consistently improved performance for both RF and KRRs. This shows that while the proposed methodology is only formulated in terms of squared error, it can also improve performance for other metrics.

[7] this is because whilst human activity can confound CO_2 and aerosol emissions, the timescale on which CO_2 and aerosol forcing operate (century vs week) are so different that the forcings at a given time can be considered independent.

Table 1. MSE, signal-to-noise ratio (SNR) and correlation on test data for the aerosol radiative forcing experiment ; $n = 50$ and 200 semi-supervised samples ; statistical significance is confirmed in Appendix F ; experiments is run for 100 datasets generated with different seeds ; $\uparrow/\downarrow$ indicates higher/lower is better ; we report 1 standard deviation ; † indicates our proposed methods.

	MSE ↓	SNR ↑	Correlation ↑
RF	0.90±0.04	0.44±0.19	0.32±0.08
P-RF†	**0.89**±0.03	**0.49**±0.15	**0.34**±0.07
KRR	0.88±0.04	0.58±0.17	0.37±0.05
P-KRR†	**0.86**±0.03	**0.65**±0.13	**0.40**±0.01
$\mathcal{H}_P$-KRR†	**0.86**±0.03	**0.65**±0.14	**0.40**±0.01

7. Discussion and Related Work

Regression and Causal Inference Currently causal inference is most commonly used in regression problems when reasoning about invariance (Peters et al., 2016; Arjovsky et al., 2019). These methods aim to use the causal structure to guarantee the predictors will transfer to new environments (Gulrajani & Lopez-Paz, 2020) and recent work discusses how causal structure plays a role in the effectiveness of these methods (Wang & Veitch, 2022). Our work takes a complimentary route in asking how causal structure can benefit in regression, and, in contrast to prior work, focuses on a fixed environment.

Causal and Anti-causal learning Our work is closely related to work on anti-causal learning (Schölkopf et al., 2012) which argues that $\mathbb{P}(X)$ will only provide additional information about $\mathbb{P}(Y|X)$ if we are working in an anti-causal prediction problem $Y \to X$. This leads the authors to hypothesise that additional unlabelled semi-supervised samples will be most helpful in the anti-causal direction. In our work, we go further and prove a concrete generalisation benefit from using additional samples from $\mathbb{P}(X)$ when the data generating process follows a collider, a graphical structure which is inherently anti-causal as it relies on Y having shared children with another vertex.

Independence Regularisation and Fair Learning Our work is related to the large body of recent work aiming to force conditional independence constraints, either for fairness (Kamishima et al., 2011) or domain generalisation (Pogodin et al., 2022). However, it is important to note that if Y satisfies a conditional independence this does not mean that the optimal least-square regressor $\mathbb{E}[Y|X]$ will satisfy the same conditional independence. For example, let

$$\begin{cases} Y, X_2 \sim \mathcal{N}(0,1) \text{ with } Y \perp\!\!\!\perp X_2 \\ X_1 = Y\mathbb{1}\{X_2 > 0\}. \end{cases} \tag{29}$$

Then we have $\mathbb{E}[Y|X_1, X_2] = X_1\mathbb{1}\{X_2 > 0\}$, hence $\mathbb{E}[Y|X_1, X_2]$ is constant when $X_2 < 0$ but not otherwise. Therefore $\mathbb{E}[Y|X_1, X_2] \not\perp\!\!\!\perp X_2$, even though $Y \perp\!\!\!\perp X_2$.

Therefore, our methodology is more similar to ensuring independence in expectation. Specifically, the RKHS methodology is related to work on fair kernel learning (Pérez-Suay et al., 2017; Li et al., 2022b). However, in contrast to the work on fair kernel learning where regularisation terms for encouraging independence are proposed, we go further by enforcing the mean independence constraint directly onto the hypothesis space.

Availability of DAG as prior knowledge Our work is based on the premise of having exact knowledge of the DAG underlying the data generating structure. This knowledge typically comes from domain expertise, with examples in genetics (Day et al., 2016) or in the aerosol radiative forcing experiment we present. However, when domain expertise is insufficient, we may need causal discovery methods to uncover the causal relationships. These methods can be expensive to run at large scale and can provide a DAG with missing or extra edges when compared to the true DAG. If collider regression is run with a partially incorrect DAG, it is likely that it would degrade the performance, as such a setting would amount to introducing incorrect prior information in the model. However, if the estimated DAG is "close" to the true DAG in the sense of the independence relationships they induce, then there may still be benefit in the finite sample regime.

Generality of proposed method Two aspects of the methodology introduced in Section 5 need to be caveated. First, it is important to require there exists no edge from children of Y to spouses of Y, otherwise that would break the conditional independence $Y \perp\!\!\!\perp X_2|X_3$. Second, whilst this is a general procedure that provides a useful inductive bias and helps restrict the hypothesis class, this procedure may not account for all the possible inductive biases that arise from the DAG at its most granular level. The procedure accounts for the collider constraint that arises from grouping variables together, not for every collider structure that might exist in the DAG. Encoding more granular collider structure would require additional regression steps, and a systematic way to perform such additional steps remains an interesting avenue for further research.

8. Conclusion

In this work we have demonstrated that collider structures within causal graphs constitute a useful form of inductive bias for regression that benefits generalisation performance. Whilst we focused on least-square regression, we expect that the collider regression framework should benefit a wider range of machine learning problems that aim to make inferences about $\mathbb{P}(Y|X)$. For example, a natural extension of this work should investigate collider regression for classification or quantile regression tasks.

Acknowledgements

The authors would like to thank Bryn Elesedy, Dimitri Meunier, Siu Lun Chau, Jean-François Ton, Christopher Williams, Duncan Watson-Parris, Eugenio Clerico[8] and Arthur Gretton for many helpful discussions and valuable feedbacks. Shahine Bouabid receives funding from the European Union's Horizon 2020 research and innovation programme under Marie Skłodowska-Curie grant agreement No 860100. Jake Fawkes receives funding from the EPSRC.

References

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

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

Bach, F. Learning theory from first principles. *Draft of a book, version of Sept*, 6:2021, 2021.

Bellouin, N., Davies, W., Shine, K. P., Quaas, J., Mülmenstädt, J., Forster, P. M., Smith, C., Lee, L., Regayre, L., Brasseur, G., Sudarchikova, N., Bouarar, I., Boucher, O., and Myhre, G. Radiative forcing of climate change from the copernicus reanalysis of atmospheric composition. *Earth System Science Data*, 12(3):1649–1677, 2020. doi: 10.5194/essd-12-1649-2020. URL https://essd.copernicus.org/articles/12/1649/2020/.

Berlinet, A. and Thomas-Agnan, C. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

Caponnetto, A. and De Vito, E. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

[8] and Tyler Farghly

Chau, S. L., Bouabid, S., and Sejdinovic, D. Deconditional downscaling with gaussian processes. *Advances in Neural Information Processing Systems*, 34:17813–17825, 2021.

Cummins, D. P., Stephenson, D. B., and Stott, P. A. Optimal estimation of stochastic energy balance model parameters. *Journal of Climate*, 2020.

Dash, M. and Liu, H. Feature selection for classification. *Intelligent data analysis*, 1(1-4):131–156, 1997.

Day, F. R., Loh, P.-R., Scott, R. A., Ong, K. K., and Perry, J. R. A robust example of collider bias in a genetic association study. *The American Journal of Human Genetics*, 98(2):392–393, 2016.

Elesedy, B. Provably strict generalisation benefit for invariance in kernel methods. *Advances in Neural Information Processing Systems*, 34:17273–17283, 2021.

Fawkes, J., Hu, R., Evans, R. J., and Sejdinovic, D. Doubly robust kernel statistics for testing distributional treatment effects even under one sided overlap. *arXiv preprint arXiv:2212.04922*, 2022.

Fukumizu, K., Bach, F. R., and Jordan, M. I. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *Journal of Machine Learning Research*, 2004.

Fukumizu, K., Song, L., and Gretton, A. Kernel bayes' rule: Bayesian inference with positive definite kernels. *The Journal of Machine Learning Research*, 14(1):3753–3783, 2013.

Gardner, J., Pleiss, G., Weinberger, K. Q., Bindel, D., and Wilson, A. G. Gpytorch: Blackbox matrix-matrix Gaussian process inference with GPU acceleration. In *Advances in Neural Information Processing Systems*, 2018.

Grünewälder, S., Lever, G., Baldassarre, L., Patterson, S., Gretton, A., and Pontil, M. Conditional Mean Embeddings as Regressors. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, 2012.

Gulrajani, I. and Lopez-Paz, D. In search of lost domain generalization. In *International Conference on Learning Representations*, 2020.

Hasselmann, K. Stochastic climate models part i. theory. *Tellus*, 1976.

Hsu, K. and Ramos, F. Bayesian deconditional kernel mean embeddings. In *International Conference on Machine Learning*, pp. 2830–2838. PMLR, 2019.

Kamishima, T., Akaho, S., and Sakuma, J. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pp. 643–650. IEEE, 2011.

Kanagawa, M., Hennig, P., Sejdinovic, D., and Sriperumbudur, B. K. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*, 2018.

Klebanov, I., Schuster, I., and Sullivan, T. J. A rigorous theory of conditional mean embeddings. *SIAM Journal on Mathematics of Data Science*, 2(3):583–606, 2020.

Koller, D. and Friedman, N. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Leach, N. J., Jenkins, S., Nicholls, Z., Smith, C. J., Lynch, J., Cain, M., Walsh, T., Wu, B., Tsutsui, J., and Allen, M. R. Fairv2.0.0: a generalized impulse response model for climate uncertainty and future scenario exploration. *Geoscientific Model Development*, 2021.

Li, Z., Meunier, D., Mollenhauer, M., and Gretton, A. Optimal rates for regularized conditional mean embedding learning. *arXiv preprint arXiv:2208.01711*, 2022a.

Li, Z., Perez-Suay, A., Camps-Valls, G., and Sejdinovic, D. Kernel dependence regularizers and gaussian processes with applications to algorithmic fairness. *Pattern Recognition*, pp. 108922, 2022b.

Lun Chau, S., Hu, R., Gonzalez, J., and Sejdinovic, D. Rkhs-shap: Shapley values for kernel methods. *Advances in neural information processing systems*, 36, 2022.

Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S. L., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M., et al. Climate change 2021: the physical science basis. *Contribution of working group I to the sixth assessment report of the intergovernmental panel on climate change*, 2, 2021.

Meek, C. Strong completeness and faithfulness in bayesian networks. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pp. 411–418, 1995.

Millar, R. J., Nicholls, Z. R., Friedlingstein, P., and Allen, M. R. A modified impulse-response representation of the global near-surface air temperature and atmospheric concentration response to carbon dioxide emissions. *Atmospheric Chemistry and Physics*, 2017.

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

Mori, T. Comments on" a matrix inequality associated with bounds on solutions of algebraic riccati and lyapunov equation" by jm saniuk and ib rhodes. *IEEE transactions on automatic control*, 33(11):1088, 1988.

Muandet, K., Sriperumbudur, B., Fukumizu, K., Gretton, A., and Schölkopf, B. Kernel mean shrinkage estimators. *Journal of Machine Learning Research*, 17, 2016.

Muandet, K., Fukumizu, K., Sriperumbudur, B., Schölkopf, B., et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

Park, J. and Muandet, K. A measure-theoretic approach to kernel conditional mean embeddings. *Advances in neural information processing systems*, 33:21247–21259, 2020.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*. 2019.

Paulsen, V. I. and Raghupathi, M. *An introduction to the theory of reproducing kernel Hilbert spaces*, volume 152. Cambridge university press, 2016.

Pearl, J. Evidential reasoning using stochastic simulation of causal models. *Artificial intelligence*, 32(2):245–257, 1987.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 2011.

Pérez-Suay, A., Laparra, V., Mateo-García, G., Muñoz-Marí, J., Gómez-Chova, L., and Camps-Valls, G. Fair kernel learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 339–355. Springer, 2017.

Peters, J., Bühlmann, P., and Meinshausen, N. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

Pogodin, R., Deka, N., Li, Y., Sutherland, D. J., Veitch, V., and Gretton, A. Efficient conditionally invariant representation learning. *arXiv preprint arXiv:2212.08645*, 2022.

Rasmussen, C. and Williams, C. Gaussian Processes for Machine Learning, 2005.

Särkkä, S. Linear operators and stochastic partial differential equations in gaussian process regression. In *International Conference on Artificial Neural Networks*, pp. 151–158. Springer, 2011.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. M. On causal and anticausal learning. In *ICML*, 2012.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. Toward causal representation learning. *Proceedings of the IEEE*, 109(5): 612–634, 2021.

Shalit, U., Johansson, F. D., and Sontag, D. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pp. 3076–3085. PMLR, 2017.

Smith, C. J., Forster, P. M., Allen, M., Leach, N., Millar, R. J., Passerello, G. A., and Regayre, L. A. Fair v1.3: a simple emissions-based impulse response and carbon cycle model. *Geoscientific Model Development*, 2018.

Song, L., Huang, J., Smola, A., and Fukumizu, K. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009.

Song, L., Gretton, A., Bickson, D., Low, Y., and Guestrin, C. Kernel belief propagation. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 707–715. JMLR Workshop and Conference Proceedings, 2011.

Song, L., Fukumizu, K., and Gretton, A. Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. *IEEE Signal Processing Magazine*, 2013.

Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. R. Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12 (7), 2011.

Statnikov, A., Lytkin, N. I., Lemeire, J., and Aliferis, C. F. Algorithms for discovery of multiple markov boundaries. *Journal of machine learning research: JMLR*, 14:499, 2013.

Steinwart, I. and Christmann, A. *Support vector machines*. Springer Science & Business Media, 2008.

Szabó, Z. and Sriperumbudur, B. K. Characteristic and universal tensor product kernels. *J. Mach. Learn. Res.*, 18:233–1, 2017.

Ton, J.-F., Lucian, C., Teh, Y. W., and Sejdinovic, D. Noise contrastive meta-learning for conditional density estimation using kernel mean embeddings. In *International Conference on Artificial Intelligence and Statistics*, pp. 1099–1107. PMLR, 2021.

Wang, Z. and Veitch, V. A unified causal view of domain invariant representation learning. In *ICML 2022: Workshop on Spurious Correlations, Invariance and Stability*, 2022. URL https://openreview.net/forum?id=-l9cpeEYwJJ.

Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*, 2012.

A. Notations and useful Results

A.1. Notations

Let $\mathcal{X}$ be a Borel space, $\mathcal{Y} \subseteq \mathbb{R}$ and let X and Y be random variables valued in $\mathcal{X}$ and $\mathcal{Y}$. We denote $(L^2(X), \langle \cdot, \cdot \rangle_{L^2(X)})$ the Hilbert space of functions from $\mathcal{X}$ to $\mathbb{R}$ which are square-integrable with respect to the pushforward measure induced by X, i.e. $\mathbb{P}_X = \mathbb{P} \circ X^{-1}$.

Let $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ be a RKHS of functions from $\mathcal{X}$ to $\mathbb{R}$ with reproducing kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. We denote its canonical feature map as k_x for any $x \in \mathcal{X}$.

Let $\mathbf{A} \in \mathbb{R}^{n \times n}$, we denote $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ the smallest and largest eigenvalues of $\mathbf{A}$ respectively.

A.2. Useful results

Theorem A.1 (Theorem 3.11, (Paulsen & Raghupathi, 2016)). *Let $\mathcal{H}$ be a RKHS on $\mathcal{X}$ with reproducing kernel k and let $f : \mathcal{X} \to \mathbb{R}$. Then the following are equivalent:*

(i) $f \in \mathcal{H}$

(ii) there exists $c \geq 0$ such that $c^2 k(x, y) - f(x) f(y)$ is kernel function

Lemma A.2 (Corollary 5.5, (Paulsen & Raghupathi, 2016)). *Let $\mathcal{H}_1, \mathcal{H}_2$ be RKHS on $\mathcal{X}$ with reproducing kernels k_1, k_2. If $\mathcal{H}_1 \cap \mathcal{H}_2 = \{0\}$, then $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2$ is a RKHS with reproducing kernel $k = k_1 + k_2$ and $\mathcal{H}_1, \mathcal{H}_2$ are orthogonal subspaces of $\mathcal{H}$.*

Proposition A.3. *Let $(\mathcal{V}, \langle \cdot, \cdot \rangle_{\mathcal{V}})$ be a Hilbert space, $\varphi : \mathcal{X} \to \mathcal{V}$ be a mapping function and*

$$k(x, y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{V}}, \quad x, y \in \mathcal{X} \tag{30}$$

the kernel function induced by φ. Then the RKHS induced by k is given by

$$\mathcal{H} = \{x \mapsto \langle v, \varphi(x) \rangle_{\mathcal{V}} \mid v \in \mathcal{V}\}. \tag{31}$$

Proof. The proof follows from the application of the Pull-back Theorem [Theorem 5.7](Paulsen & Raghupathi, 2016) to the linear kernel $L : \mathcal{V} \times \mathcal{V} \to \mathbb{R}, (v, v') \mapsto \langle v, v' \rangle_{\mathcal{V}}$ composed with the feature map $\varphi : \mathcal{X} \to \mathcal{V}$. □

Lemma A.4. *Suppose $\mathbf{A} \in \mathbb{R}^{n \times n}$ is symmetric, let Z be a random variable, $\mathbf{x} \in \mathbb{R}^n$ be a random vector.*

$$\mathbb{E}[\mathbf{x}^\top \mathbf{A} \mathbf{x} \mid Z] = \operatorname{Tr}\left(\mathbf{A} \operatorname{Var}(\mathbf{x}|Z)\right) + \mathbb{E}[\mathbf{x}|Z]^\top \mathbf{A} \mathbb{E}[\mathbf{x}|Z]. \tag{32}$$

Proof.

$$\begin{aligned}
\mathbb{E}[\mathbf{x}^\top \mathbf{A} \mathbf{x} \mid Z] &= \mathbb{E}\left[\operatorname{Tr}\left(\mathbf{A} \mathbf{x} \mathbf{x}^\top\right) \mid Z\right] && (33)\\
&= \operatorname{Tr}\left(\mathbf{A} \mathbb{E}[\mathbf{x}\mathbf{x}^\top | Z]\right) && (34)\\
&= \operatorname{Tr}\left(\mathbf{A}\left(\operatorname{Var}(\mathbf{x}|Z) + \mathbb{E}[\mathbf{x}|Z]\mathbb{E}[\mathbf{x}|Z]^\top\right)\right) && (35)\\
&= \operatorname{Tr}\left(\mathbf{A} \operatorname{Var}(\mathbf{x}|Z)\right) + \operatorname{Tr}\left(\mathbf{A}\mathbb{E}[\mathbf{x}|Z]\mathbb{E}[\mathbf{x}|Z]^\top\right) && (36)\\
&= \operatorname{Tr}\left(\mathbf{A} \operatorname{Var}(\mathbf{x}|Z)\right) + \mathbb{E}[\mathbf{x}|Z]^\top \mathbf{A}\mathbb{E}[\mathbf{x}|Z]. && (37)
\end{aligned}$$

□

Lemma A.5 ((Mori, 1988)). *Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times n}$ and suppose $\mathbf{A}$ symmetric and $\mathbf{B}$ positive semi-definite, then*

$$\operatorname{Tr}(\mathbf{A}\mathbf{B}) \geq \lambda_{\min}(\mathbf{A}) \operatorname{Tr}(\mathbf{B}). \tag{38}$$

Lemma A.6 (Lemma B.3, (Elesedy, 2021)). *Let $\mathbf{A} \in \mathbb{R}^{n \times n}$, then*

$$\lambda_{\max}(\mathbf{A}) \leq n \max_{i,j} |\mathbf{A}_{ij}|. \tag{39}$$

B. Supporting proofs

B.1. Notations

We start by introducing measure-theoretic notations which will be of use in the supporting proofs.

Let $(\Omega, \mathfrak{F}, \mathbb{P})$ denote a probability space, we denote $L^2(\Omega, \mathfrak{F}, \mathbb{P})$ the space of random variables with finite variance, which we will denote $L^2(\Omega)$ for conciseness when the σ-algebra is $\mathfrak{F}$. Endowed with inner product $\langle Z, Z' \rangle_{L^2(\Omega)} = \mathbb{E}[ZZ']$, $L^2(\Omega)$ has a Hilbert structure. For any random variable Z, we denote $\sigma(Z) \subset \mathfrak{F}$ the σ-algebra generated by Z.

B.2. Proofs of Proposition 4.1

Proposition 4.1. *Let $h \in L^2(X)$ be any regressor from our hypothesis space. We have*

$$\Delta(h, Ph) = \|Eh\|^2_{L^2(X)}. \tag{40}$$

Proof. The conditional expectation $\Pi : Z \in L^2(\Omega) \mapsto \mathbb{E}[Z|X_2]$ defines an orthogonal projection onto the space of X_2-measurable random variables with finite variance $L^2(\Omega, \sigma(X_2), \mathbb{P})$. Thus, its range and null space are orthogonal in $L^2(\Omega)$.

Let $h \in L^2(X)$. We have $Eh(X) = \mathbb{E}[h(X)|X_2] = \Pi h(X)$ hence $Eh(X)$ is in the range of Π. On the other hand,

$$\mathbb{E}[Ph(X)|X_2] = \mathbb{E}[h(X)|X_2] - \mathbb{E}[Eh(X)|X_2] = \mathbb{E}[h(X)|X_2] - \mathbb{E}[h(X)|X_2] = 0, \tag{41}$$

therefore $Ph(X)$ is in the null space of Π. Finally, because $Y \perp\!\!\!\perp X_2$ we have $\mathbb{E}[Y|X_2] = \mathbb{E}[Y] = 0$ by assumption, therefore Y is also in the null space of Π.

Hence, adopting this random variable view, the desired result simply follows from $L^2(\Omega)$ orthogonality:

$$\begin{aligned}
\Delta(h, Ph) &= \mathbb{E}[(Y - h(X))^2] - \mathbb{E}[(Y - Ph(X))^2] \\
&= \|Y - h(X)\|^2_{L^2(\Omega)} - \|Y - Ph(X)\|^2_{L^2(\Omega)} \\
&= \|Y - Ph(X) - Eh(X)\|^2_{L^2(\Omega)} - \|Y - Ph(X)\|^2_{L^2(\Omega)} \\
&= \|Y - Ph(X)\|^2_{L^2(\Omega)} + \|Eh(X)\|^2_{L^2(\Omega)} - \|Y - Ph(X)\|^2_{L^2(\Omega)} \\
&= \mathbb{E}[Eh(X)^2] \\
&= \|Eh\|^2_{L^2(X)}.
\end{aligned}$$

□

B.3. Proofs of Proposition 4.3

Proposition 4.3. *Let P^* be the adjoint operator of P in $\mathcal{H}$. Then $\mathcal{H}_P$ is also a RKHS with reproducing kernel*

$$k_P(x, x') = \langle P^* k_x, P^* k_{x'} \rangle_{\mathcal{H}} \tag{42}$$

with $P^* k_x = k_x - \mu_{X|X_2=\pi_2(x)}$.

Proof of Proposition 4.3. Let $\mathcal{H}_P$ denote the reproducing kernel with k_P. We start by showing that $P\mathcal{H} \subseteq \mathcal{H}_P$.

Let $f \in P\mathcal{H}$, then it admits a pre-image $w_f \in \mathcal{H}$ such that $f = Pw_f$. Hence for any $x \in \mathcal{X}$, we get that

$$f(x) = \langle f, k_x \rangle_{\mathcal{H}} = \langle Pw_f, k_x \rangle_{\mathcal{H}} = \langle w_f, P^* k_x \rangle_{\mathcal{H}}. \tag{43}$$

Hence, f writes as an element of the RKHS induced by the feature map $x \mapsto P^* k_x$ and by Proposition A.3 $f \in \mathcal{H}_P$.

Reciprocally, let us now show that $\mathcal{H}_P \subseteq P\mathcal{H}$. Let $f \in \mathcal{H}_P$, again by Proposition A.3 there exists $w_f \in \mathcal{H}$ such that for any $x \in \mathcal{X}$,

$$f(x) = \langle w_f, P^* k_x \rangle_{\mathcal{H}} = Pw_f(x). \tag{44}$$

This proves that $f \in P\mathcal{H}$ which concludes the proof. □

B.4. Proofs of Theorem 4.2

Theorem 4.2. *Suppose* $M = \sup_{x\in\mathcal{X}} k(x,x) < \infty$ *and* $\mathrm{Var}(Y|X) \geq \eta > 0$. *Then, the generalisation gap between* $\hat{f}$ *and* $P\hat{f}$ *satisfies*

$$\mathbb{E}[\Delta(\hat{f}, P\hat{f})] \geq \frac{\eta \mathbb{E}\left[\|\mu_{X|X_2}(X)\|^2_{L^2(X)}\right]}{\left(\sqrt{n}M + \lambda/\sqrt{n}\right)^2} \tag{45}$$

where $\mu_{X|X_2} = \mathbb{E}[k_X|X_2]$ *is the CME of* $\mathbb{P}(X|X_2)$.

Proof of Theorem 4.2. Let $\Pi = \mathbb{E}[\cdot|X_2]$ be the $L^2(\Omega)$ orthogonal projection onto the subspace of X_2-measurable random variables. For any $h \in L^2(X)$, we verify that $Eh(X) = \mathbb{E}[h(X)|X_2] = \Pi[h(X)]$ hence $Eh(X) \in \mathrm{Range}(\Pi)$. Furthermore, because $Y \perp\!\!\!\perp X_2$ we have $\Pi[Y] = \mathbb{E}[Y|X_2] = \mathbb{E}[Y] = 0$ by assumption, hence $Y \in \mathrm{Ker}(\Pi)$.

Now let

$$\mathbf{x} = \begin{bmatrix} X^{(1)} \\ \vdots \\ X^{(n)} \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} Y^{(1)} \\ \dots \\ Y^{(n)} \end{bmatrix} \tag{46}$$

denote vectors of n independent copies of X and Y and let

$$j(x,x') = \langle Ek_x, Ek_{x'}\rangle_{L^2(X)} = \mathbb{E}[Ek_x(X)Ek_{x'}(X)] \;\; \forall x, x' \in \mathcal{X}. \tag{47}$$

be the positive definite kernel induced by $L^2(X)$ inner product of Ek_x and

$$\mathbf{J} = j(\mathbf{x},\mathbf{x}) = \left[j(X^{(i)}, X^{(j)})\right]_{1\leq i,j\leq n} \tag{48}$$

the resulting Gram-matrix.

Using notations from Section 4.3, we know the solution of the kernel ridge regression problem in $\mathcal{H}$ takes the form

$$\hat{f} = \mathbf{y}^\top \left(\mathbf{K} + \lambda \mathbf{I}_n\right)^{-1} \boldsymbol{k}_{\mathbf{x}}. \tag{49}$$

Hence, by linearity of the projection, we have

$$E\hat{f} = \mathbf{y}^\top \left(\mathbf{K} + \lambda \mathbf{I}_n\right)^{-1} E\boldsymbol{k}_{\mathbf{x}} \tag{50}$$

with notation abuse $E\boldsymbol{k}_{\mathbf{x}} = [Ek_{X^{(1)}} \dots Ek_{X^{(n)}}]^\top$.

Therefore, we can write

$$\begin{aligned}
\Delta(\hat{f}, \hat{P}f) &= \|E\hat{f}\|^2_{L^2(X)} && (51)\\
&= \mathbb{E}_X[E\hat{f}(X)^2] && (52)\\
&= \mathbb{E}_X\left[\left(\mathbf{y}^\top \left(\mathbf{K} + \lambda \mathbf{I}_n\right)^{-1} E\boldsymbol{k}_{\mathbf{x}}(X)\right)^2\right] && (53)\\
&= \mathbb{E}_X\left[\mathbf{y}^\top \left(\mathbf{K} + \lambda \mathbf{I}_n\right)^{-1} E\boldsymbol{k}_{\mathbf{x}}(X) E\boldsymbol{k}_{\mathbf{x}}(X)^\top (\mathbf{K} + \lambda \mathbf{I}_n)^{-1}\mathbf{y}\right] && (54)\\
&= \mathbf{y}^\top \left(\mathbf{K} + \lambda \mathbf{I}_n\right)^{-1} \mathbb{E}_X[E\boldsymbol{k}_{\mathbf{x}}(X) E\boldsymbol{k}_{\mathbf{x}}(X)^\top](\mathbf{K} + \lambda \mathbf{I}_n)^{-1}\mathbf{y} && (55)\\
&= \mathbf{y}^\top \left(\mathbf{K} + \lambda \mathbf{I}_n\right)^{-1} \mathbf{J}(\mathbf{K} + \lambda \mathbf{I}_n)^{-1}\mathbf{y}. && (56)
\end{aligned}$$

Let us now denote for conciseness $\mathbf{A} = \left(\mathbf{K} + \lambda \mathbf{I}_n\right)^{-1} \mathbf{J}(\mathbf{K} + \lambda \mathbf{I}_n)^{-1}$. We have by Lemma XX,

$$\begin{aligned}
\mathbb{E}_{\mathbf{y}}[\Delta(\hat{f}, P\hat{f}) \mid \mathbf{x}] &= \mathbb{E}_{\mathbf{y}}[\mathbf{y}^\top \mathbf{A}\mathbf{y} \mid \mathbf{x}] && (57)\\
&= \mathrm{Tr}(\mathbf{A}\,\mathrm{Var}(\mathbf{y}|\mathbf{x})) + \mathbb{E}[\mathbf{y}|\mathbf{x}]^\top \mathbf{A}\mathbb{E}[\mathbf{y}|\mathbf{x}] \qquad \text{Lemma A.4} && (58)\\
&\geq \mathrm{Tr}(\mathbf{A}\,\mathrm{Var}(\mathbf{y}|\mathbf{x})), && (59)
\end{aligned}$$

where the conditional variance is the diagonal matrix given by

$$\mathrm{Var}(\mathbf{y}|\mathbf{x}) = \begin{bmatrix} \mathrm{Var}(Y^{(1)}|X^{(1)}) & & \\ & \ddots & \\ & & \mathrm{Var}(Y^{(n)}|X^{(n)}) \end{bmatrix} \tag{60}$$

because the copies of (X, Y) are mutually independent.

We therefore obtain,

$$\begin{aligned}
\mathbb{E}_{\mathbf{y}}[\Delta(\hat{f}, P\hat{f}) \mid \mathbf{x}] &\geq \mathrm{Tr}(\mathbf{A}\,\mathrm{Var}(\mathbf{y}|\mathbf{x})) && (61)\\
&\geq \min_i \mathrm{Var}(Y^{(i)}|X^{(i)})\,\mathrm{Tr}(\mathbf{A}) && (62)\\
&\geq \eta\,\mathrm{Tr}(\mathbf{A}) && (63)\\
&= \eta\,\mathrm{Tr}\left((\mathbf{K}+\lambda\mathbf{I}_n)^{-1}\mathbf{J}(\mathbf{K}+\lambda\mathbf{I}_n)^{-1}\right) && (64)\\
&\geq \eta\lambda_{\min}((\mathbf{K}+\lambda\mathbf{I}_n)^{-1})^2\,\mathrm{Tr}(\mathbf{J}) && \text{Lemma A.5} \quad (65)\\
&\geq \eta\frac{\mathrm{Tr}(\mathbf{J})}{(Mn+\lambda)^2} && \text{Lemma A.6.} \quad (66)
\end{aligned}$$

Finally taking the expectation against $\mathbf{x}$, we get

$$\begin{aligned}
\mathbb{E}[\Delta(\hat{f}, P\hat{f})] &\geq \frac{\mathbb{E}_{\mathbf{x}}[\eta\,\mathrm{Tr}(\mathbf{J})]}{(Mn+\lambda)^2} && (67)\\
&= \frac{\eta\sum_{i=1}^n \mathbb{E}_{X^{(i)}}[j(X^{(i)}, X^{(i)})]}{(Mn+\lambda)^2} && (68)\\
&= \frac{n\eta\mathbb{E}[j(X,X)]}{(Mn+\lambda)^2} && (69)\\
&= \frac{\eta\mathbb{E}[j(X,X)]}{(M\sqrt{n}+\lambda/\sqrt{n})^2} && (70)\\
&= \frac{\eta\mathbb{E}\left[\|Ek_X\|^2_{L^2(X)}\right]}{(M\sqrt{n}+\lambda/\sqrt{n})^2} && (71)\\
& && (72)
\end{aligned}$$

Now, for our particular choice of projection E, we have for any $x \in \mathcal{X}$ that

$$\begin{aligned}
Ek_X(x) &= \mathbb{E}_{X'}[k_X(X')|X_2 = \pi_2(x)] && (73)\\
&= \mathbb{E}_{X'}[k(X, X')|X_2 = \pi_2(x)] && (74)\\
&= \langle k_X, \mathbb{E}_{X'}[k_{X'}|X_2 = \pi_2(x)]\rangle_{\mathcal{H}} && (75)\\
&= \langle k_X, \mu_{X|X_2=\pi_2(x)}\rangle_{\mathcal{H}} && (76)\\
&= \mu_{X|X_2=\pi_2(x)}(X) && (77)
\end{aligned}$$

Therefore using the measure-theoretical CME notation from (Park & Muandet, 2020), we have

$$\|Ek_X\|^2_{L^2(X)} = \mathbb{E}_{X'}[Ek_X(X')^2] = \mathbb{E}_{X'}[\mu_{X|X_2=\pi_2(X')}(X)^2] = \|\mu_{X|X_2}(X)\|^2_{L^2(X)} \tag{78}$$

which concludes the proof. □

B.5. Proof of Proposition 5.1

Proposition 5.1. *Let $h \in L^2(X)$ be any regressor from our hypothesis space. We have*

$$\Delta(h, f_0 + P'h) = \|E'h - f_0\|^2_{L^2(X)}. \tag{79}$$

Proof. This proof follows the same structure than the proof of Proposition 4.1.

Let $\Pi = \mathbb{E}[\cdot|X_2, X_3]$ be the $L^2(\Omega)$ orthogonal projection onto the subspace of (X_2, X_3)-measurable random variables with finite variance $L^2(\Omega, \sigma(X_2, X_3), \mathbb{P})$. We have that

$$\begin{aligned}
\Pi[Y - f_0(X)] &= \mathbb{E}[Y|X_2, X_3] - \mathbb{E}[f_0(X)|X_2, X_3] && (80)\\
&= \mathbb{E}[Y|X_2, X_3] - \mathbb{E}[\mathbb{E}[Y|X_3]|X_2, X_3] && (81)\\
&= \mathbb{E}[Y|X_2, X_3] - \mathbb{E}[\mathbb{E}[Y|X_2, X_3]|X_2, X_3] \qquad (Y \perp\!\!\!\perp X_2|X_3) && (82)\\
&= 0, && (83)
\end{aligned}$$

therefore $Y - f_0(X) \in \mathrm{Ker}(\Pi)$. On the other hand, we can easily verify that for any $h \in L^2(X)$, we have $E'h(X) \in \mathrm{Range}(\Pi)$ and $P'h(X) \in \mathrm{Ker}(\Pi)$.

Therefore, it follows by $L^2(\Omega)$ orthogonality that for any $h \in L^2(X)$

$$\begin{aligned}
\|Y - h(X)\|^2_{L^2(\Omega)} &= \|(Y - f_0(X)) - (h(X) - f_0(X))\|^2_{L^2(\Omega)}\\
&= \|(Y - f_0(X)) - P'(h - f_0)(X)\|^2_{L^2(\Omega)} + \|E'(h - f_0)(X)\|^2_{L^2(\Omega)}\\
&= \|Y - (f_0(X) + P'h(X))\|^2_{L^2(\Omega)} + \|E'h - f_0\|^2_{L^2(X)} \qquad (f_0 \in \mathrm{Range}(E') = \mathrm{Ker}(P')).
\end{aligned}$$

Which allows to conclude that

$$\begin{aligned}
\Delta(h, f_0 + P'h) &= \mathbb{E}[(Y - h(X))^2] - \mathbb{E}[\big(Y - (f_0(X) + P'h(X))\big)^2]\\
&= \|Y - h(X)\|^2_{L^2(\Omega)} - \|Y - (f_0(X) + P'h(X))\|^2_{L^2(\Omega)}\\
&= \|E'h - f_0\|^2_{L^2(X)}.
\end{aligned}$$

□

C. Conditions for $P : \mathcal{H} \to \mathcal{H}$ to be well-defined

Let $\mathcal{H}$ be a RKHS of real-valued functions over $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$ with reproducing kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. In this section, we discuss conditions under which the orthogonal projection $P : L^2(X) \to L^2(X)$ can be seen as a well-defined projection over $\mathcal{H} \subset L^2(X)$.

Formally, let $\iota : \mathcal{H} \to L^2(X)$ denote the inclusion operator that maps elements of the RKHS $\mathcal{H} \ni f \mapsto [f]_\sim$ to their equivalence class in $L^2(X)$. Saying that P is well-defined as a projection over $\mathcal{H}$ means that

$$P\iota f \in \iota\mathcal{H} \ \ \forall f \in \mathcal{H}. \tag{84}$$

Such construction however raises two issues

1. Since $P = \mathrm{Id} - E$ and $E\iota f : x \mapsto \mathbb{E}[\iota f(X)|X_2 = x_2]$ is a function of x_2 only, for $E\iota f$ to lie in RKHS it is necessary for $\mathcal{H}$ to contain functions that are constant with respect to x_1.
2. If $f \in \mathcal{H}$, there is no guarantee that $E\iota f = \mathbb{E}[\iota f(X)|X_2 = \pi(\cdot)]$ will also lie in $\mathcal{H}$. In fact, this will often not be true — e.g. when $\mathcal{X}$ is a continuous domain (Song et al., 2009) — and we only have $P\iota\mathcal{H} \subset L^2(X)$.

In what follows, we permit ourselves to drop the ι notation.

C.1. Issue 1 : $\mathcal{H}$ must contain functions constant wrt x_1

In general, it is not guaranteed that a RKHS will contain constant functions. In fact, this is not the case for generic RKHSs such as the RKHSs induced by Gaussian or Matérn kernels (Steinwart & Christmann, 2008). To overcome this issue, we propose a particular form for the reproducing kernel that will ensure the RKHS contains constant functions with respect to x_1.

Proposition C.1. *Let $r : \mathcal{X}_1 \times \mathcal{X}_1 \to \mathbb{R}$ and $\ell : \mathcal{X}_2 \times \mathcal{X}_2 \to \mathbb{R}$ be kernel functions. Then the RKHS with reproducing kernel*

$$k = (r + 1) \otimes \ell \tag{85}$$

contains functions that are constant with respect to the first variable x_1.

Proof. Let $r : \mathcal{X}_1 \times \mathcal{X}_1 \to \mathbb{R}$ be a kernel function on $\mathcal{X}_1$ and consider the kernel defined by $r^+ = r + 1$ with RKHS $\mathcal{H}_{r^+}$. Let $c \in \mathbb{R}$ and consider the constant function $g(x_1) = c \ \ \forall x_1 \in \mathcal{X}_1$.

Then for any $x_1, x_1' \in \mathcal{X}_1$ we have

$$c^2 r^+(x_1, x_1') - g(x_1)g(x_1') = c^2 r(x_1, x_1') + c^2 - c^2 \tag{86}$$

$$= c^2 r(x_1, x_1') \tag{87}$$

which is a kernel function. By Theorem A.1 we conclude that $\mathcal{H}_{r^+}$ contains constant functions.

We now consider a second kernel $\ell : \mathcal{X}_2 \times \mathcal{X}_2 \to \mathbb{R}$ with RKHS $\mathcal{H}_\ell$ and we propose to take $\mathcal{H}$ as the tensor product RKHS

$$\mathcal{H} = \mathcal{H}_{r^+} \otimes \mathcal{H}_\ell, \tag{88}$$

which will have reproducing kernel

$$k = r^+ \otimes \ell. \tag{89}$$

Functions from $\mathcal{H}$ now contain functions which are the product of functions from $\mathcal{H}_{r^+}$ and $\mathcal{H}_\ell$ and are therefore allowed to be constant with respect to x_1 since $\mathcal{H}_{r^+}$ contains constant functions. □

Note that while this structural assumption may appear to limit the generality of the proposed methodology, tensor product RKHSs are a widely used form of RKHS (Szabó & Sriperumbudur, 2017; Pogodin et al., 2022; Lun Chau et al., 2022) that preserve universality of kernels from individual dimension and provide a rich function space.

Recall now the expression of the finite sample P^* estimate used in (20),

$$\hat{P}^* = \mathrm{Id} - \boldsymbol{k}_{\mathbf{x}}^\top (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \boldsymbol{\ell}_{\mathbf{x}_2}. \tag{90}$$

This allows to estimate the projected kernel k_P following

$$\begin{aligned}
\hat{k}_P(x, x') &= \langle \hat{P}^* k_x, \hat{P}^* k_x \rangle_{\mathcal{H}} \\
&= \left\langle k_x - \boldsymbol{k}_{\mathbf{x}}^\top (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \boldsymbol{\ell}_{\mathbf{x}_2}(x_2), k_{x'} - \boldsymbol{k}_{\mathbf{x}}^\top (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \boldsymbol{\ell}_{\mathbf{x}_2}(x_2') \right\rangle_{\mathcal{H}} \\
&= k(x, x') \\
&- \boldsymbol{\ell}_{\mathbf{x}_2}(x_2)^\top (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \boldsymbol{k}_x(x') \\
&- \boldsymbol{\ell}_{\mathbf{x}_2}(x_2')^\top (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \boldsymbol{k}_x(x) \\
&- \boldsymbol{\ell}_{\mathbf{x}_2}(x_2)^\top (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \mathbf{K} (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \boldsymbol{\ell}_{\mathbf{x}_2}(x_2').
\end{aligned}$$

However, for the above derivation to be correct, we need that evaluations of the second kernel ℓ can be obtained by taking an inner product in $\mathcal{H}$. Namely, we need that

$$\ell(x_2, x_2') = \langle \ell_{x_2}, \ell_{x_2'} \rangle_{\mathcal{H}_\ell} = \langle \ell_{x_2}, \ell_{x_2'} \rangle_{\mathcal{H}}. \tag{91}$$

The following result shows that a sufficient condition for this to hold is that $\mathcal{H}_r$ itself does not contain constant functions. As mentioned above, this is a condition satisfied by generic RKHSs such as the RKHSs of the Gaussian kernel or the Matérn kernels (Steinwart & Christmann, 2008) — which is the RKHS we work with in our experiments.

Proposition C.2. *Let $\mathcal{H} = \mathcal{H}_{r^+} \otimes \mathcal{H}_\ell$ where $r^+ = r + 1$. If $\mathcal{H}_r$ does not contain constant functions, then we have that $\ell(x_2, x_2') = \langle \ell_{x_2}, \ell_{x_2'} \rangle_{\mathcal{H}}$.*

Proof. The kernel $r^+(x_1, x_1') = r(x_1, x_1') + 1$ here induces a RKHS $\mathcal{H}_{r^+}$ (of functions from $\mathcal{X}_1$ to $\mathbb{R}$) which does contain constant functions, e.g., $e \in \mathcal{H}_{r^+}$, where $e(x_1) = 1, \forall x_1 \in \mathcal{X}_1$.

This choice of kernel ensures that $\ell_{x_2} \in \mathcal{H}$ when viewed as a function on $\mathcal{X}_1 \times \mathcal{X}_2$, i.e. we can write it as $e \otimes \ell_{x_2}$, so it is clear that it belongs to $\mathcal{H} = \mathcal{H}_{r^+} \otimes \mathcal{H}_\ell$, since $e \in \mathcal{H}_{r^+}$ and $\ell_{x_2} \in \mathcal{H}_\ell$.

Furthermore, we have

$$\begin{aligned}
\langle \ell_{x_2}, \ell_{x_2'} \rangle_{\mathcal{H}} &= \langle e \otimes \ell_{x_2}, e \otimes \ell_{x_2'} \rangle_{\mathcal{H}} \\
&= \langle e, e \rangle_{\mathcal{H}_{r^+}} \langle \ell_{x_2}, \ell_{x_2'} \rangle_{\mathcal{H}_\ell} \\
&= \langle e, e \rangle_{\mathcal{H}_{r^+}} \ell(x_2, x_2').
\end{aligned}$$

However,

$$\begin{aligned}
\langle e, e \rangle_{\mathcal{H}_{r^+}} &= \langle e, e + r_{x_1} \rangle_{\mathcal{H}_{r^+}} - \langle e, r_{x_1} \rangle_{\mathcal{H}_{r^+}} \\
&= \langle e, r_{x_1}^+ \rangle_{\mathcal{H}_{r^+}} - \langle e, r_{x_1} \rangle_{\mathcal{H}_{r^+}} \\
&= e(x_1) - \langle e, r_{x_1} \rangle_{\mathcal{H}_{r^+}} \\
&= 1 - \langle e, r_{x_1} \rangle_{\mathcal{H}_{r^+}}.
\end{aligned}$$

Now if $\mathcal{H}_r$ does not contain constant functions, we have $\mathrm{Span}(\{e\}) \cap \mathcal{H}_r = \{0\}$. Hence, by Lemma A.2 we obtain that e and r_{x_1} are orthogonal in $\mathcal{H}_{r^+}$ which in turn gives that

$$\langle e, r_{x_1} \rangle_{\mathcal{H}_{r^+}} = 0 \Rightarrow \langle e, e \rangle_{\mathcal{H}_{r^+}} = 1. \tag{92}$$

Therefore, if $\mathcal{H}_r$ does not contain constant functions we have that

$$\langle \ell_{x_2}, \ell_{x_2'} \rangle_{\mathcal{H}} = \langle e, e \rangle_{\mathcal{H}_{r^+}} \ell(x_2, x_2') = \ell(x_2, x_2'). \tag{93}$$

□

C.2. Issue 2 : P is not necessarily closed as an operator on $\mathcal{H}$

Too Long; Didn't Read We make the assumption that $\mathbb{E}[f(X)|X_2 = \cdot] \in \mathcal{H}$ for $f \in \mathcal{H}$.

Too Short; Want More It is possible to choose the reproducing kernel k such that $\mathcal{H}$ is dense in $L^2(X)$. This property is called L^2-universality (Sriperumbudur et al., 2011). Whilst this might suggest that the assumption $Ef \in \mathcal{H}$ for $f \in \mathcal{H}$ could be reasonable when L^2-universality is met, in practice no explicit case is provided in the literature where it is easy to verify that $\mathbb{E}[f(X)|X_2 = \cdot] \in \mathcal{H}$ for $f \in \mathcal{H}$.

In fact, a classic counter example given by Fukumizu et al. (2013) is the case where $\mathcal{H}$ is the RKHS of the Gaussian kernel on $\mathcal{X}$ and $X \perp\!\!\!\perp Z$. Then, $\mathbb{E}[f(X)|Z = \cdot]$ is constant for any $f \in \mathcal{H}$ but $\mathcal{H}$ does not contain constant functions (Steinwart & Christmann, 2008). In the context of our work, we do not have $(X_1, X_2) \perp\!\!\!\perp X_2$ but it nonetheless remains difficult to verify whether $\mathbb{E}[f(X)|X_2 = \cdot] \in \mathcal{H}$.

Efforts to study this nontrivial research direction must be highlighted : Mollenhauer & Koltai (2020) show that under denseness assumptions, it is possible to approximate the conditional expectation operator $E : L^2(X) \to L^2(X)$ with a Hilbert-Schmidt operator on $\mathcal{H}$ with arbitrary precision. Klebanov et al. (2020) propose a rigorous RKHS-friendly construction of E that only assumes that Ef lies a constant away from $\mathcal{H}$. Most recently, Li et al. (2022a) consider the weaker assumption that for $f \in \mathcal{H}$, Ef lies in an interpolation space between $\mathcal{H}$ and $L^2(X)$ and prove optimal learning rates for its estimator.

The theoretical intricacies of such considerations tend however to undermine more "practical"-driven work. For this reason, it is common to defer such consideration to theoretical research and make the assumption that $\mathbb{E}[f(X)|X_2 = \cdot] \in \mathcal{H}$ (Fukumizu et al., 2004; Song et al., 2011; Muandet et al., 2016; Hsu & Ramos, 2019; Ton et al., 2021; Chau et al., 2021; Fawkes et al., 2022). Since the RKHS theory is not central to our motivations but only a tool we use to demonstrate the benefits of collider regression, we propose to make a similar assumption and delegate this theoretical consideration for future work.

D. Collider Regression on a simple DAG: estimators

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $\ell : \mathcal{X}_2 \times \mathcal{X}_2 \to \mathbb{R}$ be positive definite kernel. In what follows, we adopt notations from the Section 4.3. $\hat{f} = \mathbf{y}^\top (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \boldsymbol{k}_{\mathbf{x}}$ denotes the solution to the kernel ridge regression problem in $\mathcal{H}$. We abuse notation and denote the pairwise inner product of feature maps as

$$\langle \boldsymbol{k}_{\mathbf{x}}, \boldsymbol{k}_{\mathbf{x}} \rangle_{\mathcal{H}} = \left[\langle k_{x_i}, k_{x_j} \rangle_{\mathcal{H}} \right]_{1 \leq i,j \leq n} = \left[k(x_i, x_j) \right]_{1 \leq i,j \leq n} = \mathbf{K}. \tag{94}$$

D.1. For a general choice of kernel k

D.1.1. Estimating $\mu_{X|X_2=x_2}$

We are interested in estimating the CME $\mu_{X|X_2=x_2}$. Using the CME estimate from (6), we obtain

$$\boxed{\hat{\mu}_{X|X_2=x_2} = \boldsymbol{k}_{\mathbf{x}}^\top (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \boldsymbol{\ell}_{\mathbf{x}_2}(x_2).} \tag{95}$$

D.1.2. Estimating $P\hat{f}$

Writing out

$$P\hat{f}(x_1, x_2) = \hat{f}(x_1, x_2) - \langle \hat{f}, \mu_{X|X_2=x_2} \rangle_{\mathcal{H}} \tag{96}$$

$$= \mathbf{y}^\top (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \boldsymbol{k}_{\mathbf{x}}(x_1, x_2) - \mathbf{y}^\top (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \langle \boldsymbol{k}_{\mathbf{x}}, \mu_{X|X_2=x_2} \rangle_{\mathcal{H}}, \tag{97}$$

it appears we can obtain an estimate of $P\hat{f}$ by substituting $\mu_{X|X_2=x_2}$ with its estimate in the above. We obtain

$$\hat{P}\hat{f}(x_1, x_2) = \mathbf{y}^\top (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \boldsymbol{k}_{\mathbf{x}}(x_1, x_2) - \mathbf{y}^\top (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \underbrace{\langle \boldsymbol{k}_{\mathbf{x}}, \boldsymbol{k}_{\mathbf{x}} \rangle_{\mathcal{H}}}_{\mathbf{K}} (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \boldsymbol{\ell}_{\mathbf{x}_2}(x_2) \tag{98}$$

$$= \mathbf{y}^\top (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \left(\boldsymbol{k}_{\mathbf{x}}(x_1, x_2) - \mathbf{K} (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \boldsymbol{\ell}_{\mathbf{x}_2}(x_2) \right), \tag{99}$$

or in functional form

$$\boxed{\hat{P}\hat{f} = \mathbf{y}^\top (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \left(\boldsymbol{k}_{\mathbf{x}} - \mathbf{K} (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \boldsymbol{\ell}_{\mathbf{x}_2} \right).} \tag{100}$$

D.2. When $k = (r+1) \otimes \ell$

In Section 4.4, a sufficient assumption for the projection to be well-defined is that the kernel takes the form

$$k = (r+1) \otimes \ell, \tag{101}$$

where $r : \mathcal{X}_1 \times \mathcal{X}_1 \to \mathbb{R}$ is a positive definite kernel. When we choose this particular form of kernel, alternative estimators can be devised.

In what follow, we denote $r^+ = r + 1, \boldsymbol{r}^+_{\mathbf{x}_1} = r^+(\mathbf{x}_1, \cdot)$ and $\mathbf{R}^+ = r^+(\mathbf{x}_1, \mathbf{x}_1)$.

D.2.1. Estimating $\mu_{X|X_2=x_2}$

Going back to the definition of CMEs, we can write

$$\mu_{X|X_2=x_2} = \mathbb{E}[k_X | X_2 = x_2] = \mathbb{E}[r^+_{X_1} \otimes \ell_{X_2} | X_2 = x_2] = \mathbb{E}[r^+_{X_1} | X_2 = x_2] \otimes \ell_{x_2} = \mu_{X_1|X_2=x_2} \otimes \ell_{x_2}. \tag{102}$$

Therefore, it is sufficient to obtain an estimate of $\mu_{X_1|X_2=x_2}$, which we can get as

$$\hat{\mu}_{X_1|X_2=x_2} = \boldsymbol{r}^{+\top}_{\mathbf{x}_1} (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \boldsymbol{\ell}_{\mathbf{x}_2}(x_2), \tag{103}$$

and take as a CME estimator

$$\boxed{\hat{\mu}_{X|X_2=x_2} = \left[\boldsymbol{r}^{+\top}_{\mathbf{x}_1} (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \boldsymbol{\ell}_{\mathbf{x}_2}(x_2) \right] \ell_{x_2}(\cdot)} \tag{104}$$

D.2.2. ESTIMATING $P\hat{f}$

Following the similar derivations than in the general case, we obtain

$$\hat{P}\hat{f}(x_1, x_2) = \mathbf{y}^\top(\mathbf{K} + \lambda \mathbf{I}_n)^{-1}\boldsymbol{k}_{\mathbf{x}}(x_1, x_2) - \mathbf{y}^\top(\mathbf{K} + \lambda \mathbf{I}_n)^{-1}\langle \boldsymbol{r}^+_{\mathbf{x}_1}, \boldsymbol{r}^+_{\mathbf{x}_1}\rangle_{\mathcal{H}_{r^+}} \langle \ell_{\mathbf{x}_2}, \ell_{x_2}\rangle_{\mathcal{H}_\ell}(\mathbf{L} + \gamma \mathbf{I}_n)^{-1}\boldsymbol{\ell}_{\mathbf{x}_2}(x_2) \quad (105)$$

$$= \mathbf{y}^\top(\mathbf{K} + \lambda \mathbf{I}_n)^{-1}\left[\boldsymbol{k}_{\mathbf{x}}(x_1, x_2) - \mathrm{Diag}(\ell_{\mathbf{x}_2}(x_2))\mathbf{R}^+(\mathbf{L} + \gamma \mathbf{I}_n)^{-1}\boldsymbol{\ell}_{\mathbf{x}_2}(x_2)\right], \quad (106)$$

where $\mathrm{Diag}(\ell_{\mathbf{x}_2}(x_2))$ is the diagonal matrix that has the vector $\ell_{\mathbf{x}_2}(x_2) = \ell(\mathbf{x}_2, x_2)$ as its diagonal. Written in functional form we obtain

$$\boxed{\hat{P}\hat{f} = \mathbf{y}^\top(\mathbf{K} + \lambda \mathbf{I}_n)^{-1}\left[\boldsymbol{k}_{\mathbf{x}} - \mathrm{Diag}(\ell_{\mathbf{x}_2}(\cdot))\mathbf{R}^+(\mathbf{L} + \gamma \mathbf{I}_n)^{-1}\boldsymbol{\ell}_{\mathbf{x}_2}\right]} \quad (107)$$

D.2.3. ESTIMATING k_P

Writing out,

$$k_P(x, x') = \langle P^* k_x, P^* k_{x'}\rangle_{\mathcal{H}} \quad (108)$$

$$= \langle k_x - \mu_{X|X_2=x_2}, k_{x'} - \mu_{X|X_2=x'_2}\rangle_{\mathcal{H}} \quad (109)$$

$$= \langle k_x, k_{x'}\rangle_{\mathcal{H}} \quad (110)$$

$$- \langle \mu_{X_1|X_2=x_2} \otimes \ell_{x_2}, k_{x'}\rangle_{\mathcal{H}} \quad (111)$$

$$- \langle k_x, \mu_{X_1|X_2-x'_2} \otimes \ell_{x'_2}\rangle_{\mathcal{H}} \quad (112)$$

$$+ \langle \mu_{X_1|X_2=x_2} \otimes \ell_{x_2}, \mu_{X_1|X_2-x'_2} \otimes \ell_{x'_2}\rangle_{\mathcal{H}} \quad (113)$$

$$= r^+(x_1, x'_1)\ell(x_2, x'_2) \quad (114)$$

$$- \langle \mu_{X_1|X_2=x_2}, r^+_{x'_1}\rangle_{\mathcal{H}_{r^+}} \ell(x_2, x'_2) \quad (115)$$

$$- \langle r^+_{x_1}, \mu_{X_1|X_2=x'_2}\rangle_{\mathcal{H}_{r^+}} \ell(x_2, x'_2) \quad (116)$$

$$+ \langle \mu_{X_1|X_2=x_2}, \mu_{X_1|X_2=x'_2}\rangle_{\mathcal{H}_{r^+}} \ell(x_2, x'_2) \quad (117)$$

$$= \ell(x_2, x'_2)\left[r^+(x_1, x'_1) - \langle \mu_{X_1|X_2=x_2}, r^+_{x'_1}\rangle_{\mathcal{H}_{r^+}} - \langle r^+_{x_1}, \mu_{X_1|X_2=x'_2}\rangle_{\mathcal{H}_{r^+}} + \langle \mu_{X_1|X_2=x_2}, \mu_{X_1|X_2=x'_2}\rangle_{\mathcal{H}_{r^+}}\right] \quad (118)$$

Therefore, substituting $\mu_{X_1|X_2=x_2}$ with its estimate, we obtain

$$\boxed{\begin{aligned}
\hat{k}_P(x, x') &= \ell(x_2, x'_2) && (119a)\\
&\times \left[r^+(x_1, x'_1)\right. && (119b)\\
&- \boldsymbol{\ell}_{\mathbf{x}_2}(x_2)^\top(\mathbf{L} + \gamma \mathbf{I}_n)^{-1}\boldsymbol{r}^+_{\mathbf{x}_1}(x'_1) && (119c)\\
&- \boldsymbol{\ell}_{\mathbf{x}_2}(x'_2)^\top(\mathbf{L} + \gamma \mathbf{I}_n)^{-1}\boldsymbol{r}^+_{\mathbf{x}_1}(x_1) && (119d)\\
&- \left.\boldsymbol{\ell}_{\mathbf{x}_2}(x_2)^\top(\mathbf{L} + \gamma \mathbf{I}_n)^{-1}\mathbf{R}^+(\mathbf{L} + \gamma \mathbf{I}_n)^{-1}\boldsymbol{\ell}_{\mathbf{x}_2}(x'_2)\right]. && (119e)
\end{aligned}}$$

E. Collider Regression on a general DAG: algorithms and estimators

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $r : \mathcal{X}_1 \times \mathcal{X}_1 \to \mathbb{R}$ and $\ell : (\mathcal{X}_2 \times \mathcal{X}_3) \times (\mathcal{X}_2 \times \mathcal{X}_3) \to \mathbb{R}$ be psd kernels. We follow the same notation convention that in the case of a simple collider, except that now ℓ is a kernel over $\mathcal{X}_2 \times \mathcal{X}_3$. Define $f_0(x) = \mathbb{E}[Y|X_3 = x_3]$. Here $g^* = f^* - f_0$ must live in the appropriate subspace of functions which have zero conditional expectation on (X_2, X_3).

E.1. Algorithms

Algorithm 4 General procedure to estimate $f_0 + P'\hat{g}$

1: Regress $X_3 \to Y$ to get $x_3 \mapsto \hat{f}_0(x_3)$
2: Take $\tilde{Y} = Y - \hat{f}_0(X_3)$
3: Regress $(X_1, X_2, X_3) \to \tilde{Y}$ to get $(x_1, x_2, x_3) \mapsto \hat{g}(x_1, x_2, x_3)$
4: Regress $(X_2, X_3) \to \hat{g}(X_1, X_2, X_3)$ to get $(x_2, x_3) \mapsto \hat{\mathbb{E}}[\hat{g}(X_1, X_2, X_3)|X_2 = x_2, X_3 = x_3]$
5: Take $\hat{P}'\hat{g}(x_1, x_2, x_3) = \hat{g}(X_1, X_2, X_3) - \hat{\mathbb{E}}[\hat{g}(x_1, x_2, x_3)|X_2 = x_2, X_3 = x_3]$
6: **return** $\hat{f}_0 + \hat{P}'\hat{g}$

Algorithm 5 RKHS procedure to estimate $f_0 + P'\hat{g}$

1: Estimate $\hat{\mu}_{X|X_2=x_2,X_3=x_3}$
2: Regress $X_3 \to Y$ to get $x_3 \mapsto \hat{f}_0(x_3)$
3: Take $\tilde{\mathbf{y}} = \mathbf{y} - \hat{f}_0(\mathbf{x}_3)$
4: Take $\hat{g} = \tilde{\mathbf{y}}^\top (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \boldsymbol{k}_{\mathbf{x}}$
5: Let $\hat{P}'\hat{g} = \hat{g} - \langle \hat{g}, \hat{\mu}_{X|X_2=\cdot,X_3=\cdot} \rangle_{\mathcal{H}}$
6: **return** $\hat{f}_0 + \hat{P}'\hat{g}$

Algorithm 6 RKHS procedure to estimate $f_0 + \hat{g}_{P'}$

1: Estimate $\hat{\mu}_{X|X_2=x_2,X_3=x_3}$
2: Regress $X_3 \to Y$ to get $x_3 \mapsto \hat{f}_0(x_3)$
3: Take $\tilde{\mathbf{y}} = \mathbf{y} - \hat{f}_0(\mathbf{x}_3)$
4: Let $\hat{P}'^* k_x = k_x - \hat{\mu}_{X|X_2=x_2,X_3=x_3}$
5: Let $\hat{k}_{P'}(x, x') = \langle \hat{P}'^* k_x, \hat{P}'^* k_{x'} \rangle_{\mathcal{H}}$
6: Evaluate $\hat{\mathbf{K}}_{P'} = \hat{k}_{P'}(\mathbf{x}, \mathbf{x})$ and $\hat{\boldsymbol{k}}_{P',\mathbf{x}} = \hat{k}_{P'}(\mathbf{x}, \cdot)$
7: Take $\hat{g}_{P'} = \tilde{\mathbf{y}}^\top (\hat{\mathbf{K}}_{P'} + \lambda \mathbf{I}_n)^{-1} \hat{\boldsymbol{k}}_{P',\mathbf{x}}$
8: **return** $\hat{f}_0 + \hat{g}_{P'}$

E.2. Estimators for a general kernel k

E.2.1. ESTIMATING $\mu_{X|X_2=x_2,X_3=x_3}$

$$\boxed{\hat{\mu}_{X|X_2=x_2,X_3=x_3} = \boldsymbol{k}_{\mathbf{x}}^\top (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \boldsymbol{\ell}_{\mathbf{x}_2,\mathbf{x}_3}(x_2, x_3).} \tag{120}$$

E.2.2. ESTIMATING $P'\hat{g}$

$$\boxed{\hat{P}'\hat{g} = \tilde{\mathbf{y}}^\top (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \left(\boldsymbol{k}_{\mathbf{x}} - \mathbf{K} (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \boldsymbol{\ell}_{\mathbf{x}_2,\mathbf{x}_3} \right).} \tag{121}$$

E.3. Estimators when $k = (r+1) \otimes \ell$

E.3.1. Estimating $\mu_{X|X_2=x_2,X_3=x_3}$

$$\boxed{\hat{\mu}_{X|X_2=x_2} = \left[\boldsymbol{r}_{\mathbf{x}_1}^{+\top}(\mathbf{L}+\gamma\mathbf{I}_n)^{-1}\boldsymbol{\ell}_{\mathbf{x}_2,\mathbf{x}_3}((x_2,x_3))\right]\ell_{x_2,x_3}(\cdot)} \tag{122}$$

E.3.2. Estimating $P'\hat{g}$

$$\boxed{\hat{P'}\hat{g} = \tilde{\mathbf{y}}^\top(\mathbf{K}+\lambda\mathbf{I}_n)^{-1}\left[\boldsymbol{k}_{\mathbf{x}} - \mathrm{Diag}(\ell_{\mathbf{x}_2,\mathbf{x}_3}(\cdot))\mathbf{R}^+(\mathbf{L}+\gamma\mathbf{I}_n)^{-1}\boldsymbol{\ell}_{\mathbf{x}_2,\mathbf{x}_3}\right]} \tag{123}$$

E.3.3. Estimating $k_{P'}$

$$\begin{aligned}
\hat{k}_{P'}(x,x') &= \ell\left((x_2,x_3),(x_2',x_3')\right) && \text{(124a)}\\
&\times \left[r^+(x_1,x_1')\right. && \text{(124b)}\\
&- \boldsymbol{\ell}_{\mathbf{x}_2,\mathbf{x}_3}((x_2,x_3))^\top(\mathbf{L}+\gamma\mathbf{I}_n)^{-1}\boldsymbol{r}^+_{\mathbf{x}_1}(x_1') && \text{(124c)}\\
&- \boldsymbol{\ell}_{\mathbf{x}_2,\mathbf{x}_3}((x_2',x_3'))^\top(\mathbf{L}+\gamma\mathbf{I}_n)^{-1}\boldsymbol{r}^+_{\mathbf{x}_1}(x_1) && \text{(124d)}\\
&- \left.\boldsymbol{\ell}_{\mathbf{x}_2,\mathbf{x}_3}((x_2,x_3))^\top(\mathbf{L}+\gamma\mathbf{I}_n)^{-1}\mathbf{R}^+(\mathbf{L}+\gamma\mathbf{I}_n)^{-1}\boldsymbol{\ell}_{\mathbf{x}_2,\mathbf{x}_3}((x_2',x_3'))\right]. && \text{(124e)}
\end{aligned}$$

F. Details on experiments

F.1. Models

➢ **RF** We use the scikit-learn (Pedregosa et al., 2011) `sklearn.ensemble.RandomForestRegressor` implementation which we tune for

- `n_estimators`
- `max_depth`
- `min_samples_split`
- `min_samples_leaf`

using a cross-validated grid search over an independently generated validation set.

➢ P**-RF** Once RF has been fitted as $\hat{f}$, we estimate $\mathbb{E}[\hat{f}(X_1, X_2)|X_2]$ by fitting a linear regression model of X_2 onto $\hat{f}(X_1, X_2)$.

➢ **KRR** We implement our own kernel ridge regression in PyTorch (Paszke et al., 2019). The kernel is taken as

$$k\big((x_1, x_2), (x_1', x_2')\big) = \big(\kappa_{\theta_1}(x_1, x_1') + 1\big)\kappa_{\theta_2}(x_2, x_2'), \tag{125}$$

where κ_θ denotes the Gaussian kernel with lengthscale $\theta > 0$

$$\kappa_\theta(u, u') = \exp\left(-\frac{\|u - u'\|_2^2}{\theta}\right). \tag{126}$$

The kernel lengthscales θ_1, θ_2 and the regularisation weight $\lambda > 0$ are tuned using a cross-validated grid search on an independently generated validation set.

➢ P**-KRR** Once KRR has been fitted as $\hat{f} = \mathbf{y}^\top (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \boldsymbol{k}_{\mathbf{x}}$, we estimate the CME and use it to estimate $P\hat{f}(x_1, x_2) = \hat{f}(x_1, x_2) - \langle \hat{f}, \mu_{X|X_2=x_2} \rangle_{\mathcal{H}}$ following

$$\hat{\mu}_{X|X_2=x_2} = \boldsymbol{k}_{\mathbf{x}}^\top (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \boldsymbol{\ell}_{\mathbf{x}_2}(x_2) \tag{127}$$

$$\Rightarrow \hat{P} = \mathrm{Id} - \hat{\mu}_{X|X_2=\cdot} \tag{128}$$

$$= \mathrm{Id} - \boldsymbol{k}_{\mathbf{x}}^\top (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \boldsymbol{\ell}_{\mathbf{x}_2} \tag{129}$$

$$\Rightarrow \hat{P}\hat{f} = \hat{f} - \hat{f}\boldsymbol{k}_{\mathbf{x}}^\top (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \boldsymbol{\ell}_{\mathbf{x}_2} \tag{130}$$

$$= \mathbf{y}^\top (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \boldsymbol{k}_{\mathbf{x}} - \mathbf{y}^\top (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \mathbf{K} (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \boldsymbol{\ell}_{\mathbf{x}_2} \tag{131}$$

$$= \mathbf{y}^\top (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \big(\boldsymbol{k}_{\mathbf{x}} - \mathbf{K} (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \boldsymbol{\ell}_{\mathbf{x}_2}\big) \tag{132}$$

The kernel on $\mathcal{X}_2$ is taken as $\ell = \kappa_{\theta_2}$. The CME regularisation weight $\gamma > 0$ is tuned using a cross-validated grid search on an independently generated validation set.

➢ $\mathcal{H}_P$**-KRR** We use the same base kernel as for KRR with again $\ell = \kappa_{\theta_2}$. We implement our estimator of the projected kernel k_P is GPyTorch (Gardner et al., 2018)[9]. The kernel lengthscales and regularisation weights are tuned using a cross-validated grid search on an independently generated validation set.

[9] which can be readily incorporated into GP regression pipelines.

F.2. Simulation example

Data generating process Algorithm 7 outlines the procedure we use to generate a positive definite matrix Σ that encodes independence between X_2 and Y.

Algorithm 7 Procedure to generate Σ

Input: $d_1 \geq 1, d_2 \geq 1$
`# Generate a` $4 \times (d_1 + d_2 + 1)$ `random matrix`
for $i \in \{1, \ldots, d_1 + d_2 + 1\}$ **do**
  $M_i \sim \mathcal{N}(0, \mathbf{I}_4)$
  $M_i \leftarrow M_i \,/\, \|M_i\|_2$
end for
`# Make` Y `column orthogonal to all` X_2 `columns`
$M_Y \leftarrow M_{d_1+d_2+1}$
for $i \in \{d_1 + 1, \ldots, d_1 + d_2\}$ **do**
  $M_i \leftarrow M_i - (M_i^\top M_Y) M_Y$
end for
$M \leftarrow \left[M_1 \mid \ \ldots \ \mid M_{d_1+d_2} \mid \ M_Y\right] \in \mathbb{R}^{4 \times (d_1+d_2+1)}$
$\Sigma \leftarrow M^\top M + 0.01 * \mathbf{I}_{d_1+d_2+1}$
`# Normalise variances to 1`
$\Lambda \leftarrow \mathrm{Diag}(\Sigma)$
$\Sigma \leftarrow \Lambda^{-1/2} \Sigma \Lambda^{-1/2}\ \Sigma$

Non-linear mappings The mappings g_1 and g_2 are applied to each component of the input vectors and are given by

$$g_1(u) = u + 0.1\ \cos(2\pi u^2) \tag{133}$$

$$g_2(u) = u + 0.1\ \sin(2\pi u^2). \tag{134}$$

Statistical significance table

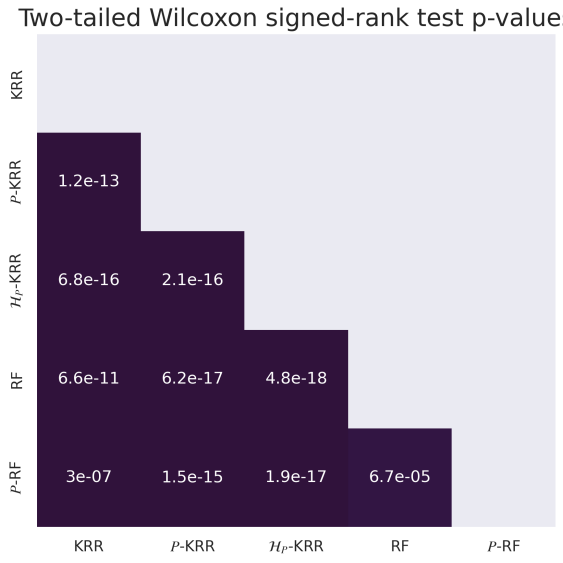

Figure 6. p-values from a two-tailed Wilcoxon signed-rank test between all pairs of methods for the test MSE of the simulation example. The null hypothesis is that scores samples come from the same distribution. We only present the lower triangular matrix of the table for clarity of reading.

F.3. Aerosol radiative forcing

Statistical significance table

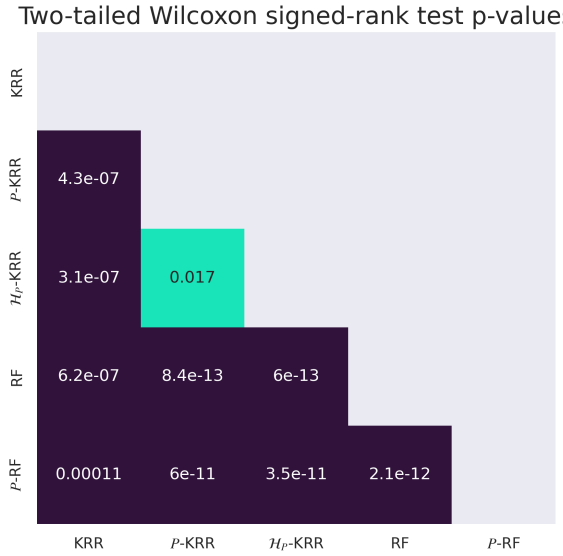

Figure 7. p-values from a two-tailed Wilcoxon signed-rank test between all pairs of methods for the test **MSE** of the aerosol radiative forcing experiment. The null hypothesis is that scores samples come from the same distribution. We only present the lower triangular matrix of the table for clarity of reading.

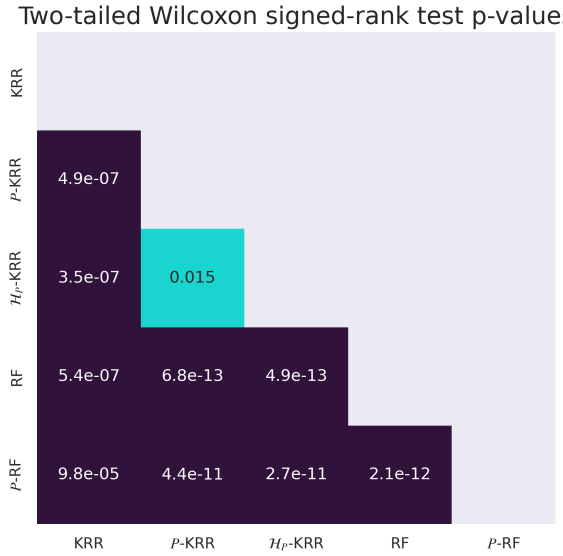

Figure 8. p-values from a two-tailed Wilcoxon signed-rank test between all pairs of methods for the test **SNR** of the aerosol radiative forcing experiment. The null hypothesis is that scores samples come from the same distribution. We only present the lower triangular matrix of the table for clarity of reading.

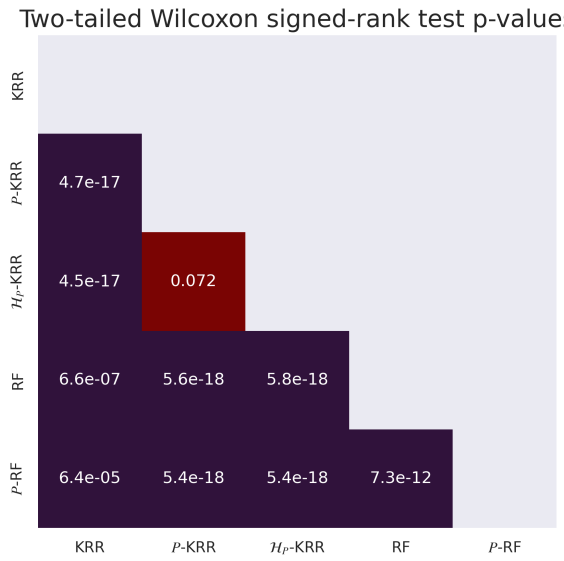

Figure 9. p-values from a two-tailed Wilcoxon signed-rank test between all pairs of methods for the test **correlation** of the aerosol radiative forcing experiment. The null hypothesis is that scores samples come from the same distribution. We only present the lower triangular matrix of the table for clarity of reading.

G. Future direction

G.1. Extension to Gaussian processes

Extension to Gaussian processes The methodology presented can naturally be extended to the Bayesian counterpart of kernel ridge regression, Gaussian processes (GPs) (Rasmussen & Williams, 2005). One can either apply the projection operator $P : L^2(X) \to L^2(X)$ to the GP prior (or posterior), or use the projected kernel k_P to specify the covariance function[10].

However, such approach raises important questions from a theoretical perspective. If $f \sim \mathrm{GP}(0, k)$, the application of the $L^2(X)$ projection to f will result in a linearly transformed GP $Pf \sim \mathrm{GP}(0, PkP^*)$ (Särkkä, 2011) and its draws will lie in the range of P. In contrast, since draws from a GP almost surely lie outside the RKHS associated with its covariance (Kanagawa et al., 2018), draws from $f \sim \mathrm{GP}(0, k_P)$ will almost surely lie outside $\mathcal{H}_P$. It is therefore unclear whether these draws will lie in the range of the projection and satisfy the desired constraint for f. On the other hand, the posterior mean of the GP will always lie in $\mathcal{H}_P$.

Furthermore, the projection is targeted at improving performance in mean square error. Because this metric is not necessarily adequate to evaluate GPs, it is unclear whether applying the projection would result in a performance improvement on more commonly used metrics for GPs such as maximum likelihood.

[10] Our implementation of $\hat{k}_P$ is available in GPyTorch (Gardner et al., 2018) and can be readily incorporated into GP regression pipelines.