
Kernel Conditional Density Operators

Ingmar Schuster
Zalando Research, Zalando SE
Berlin, Germany
ingmar.schuster@zalando.de

Mattes Mollenhauer
Freie Universität Berlin
Berlin, Germany
mattes.mollenhauer@fu-berlin.de

Stefan Klus
Freie Universität Berlin
Berlin, Germany
stefan.klus@fu-berlin.de

Krikamol Muandet
MPI for Intelligent Systems
Tübingen, Germany
krikamol@tuebingen.mpg.de

Abstract

We introduce a conditional density estimation model termed the *conditional density operator*. It naturally captures multivariate, multimodal output densities and is competitive with recent neural conditional density models and Gaussian processes. To derive the model, we propose a novel approach to the reconstruction of probability densities from their kernel mean embeddings by drawing connections to estimation of Radon–Nikodym derivatives in the reproducing kernel Hilbert space (RKHS). We prove finite sample error bounds which are independent of problem dimensionality. Furthermore, the resulting conditional density model is applied to real-world data and we demonstrate its versatility and competitive performance.

1 Introduction

We present a kernel-based supervised learning model for the estimation of conditional densities, the *conditional density operator* (CDO). It is competitive with conditional density models based on deep neural networks [13]. To derive the model, we will first focus on the problem of reconstructing a probability density from its associated kernel mean embedding [40, 51] and connect it to the estimation of Radon–Nikodym derivatives. While this very general problem has been tackled before in similar scenarios [24, 44], we provide a characterization of conditions that admit to formulate the density reconstruction as an inverse problem with a unique analytical solution. We show that in practical applications, the arising statistical inverse problem can be solved conveniently by using Tikhonov regularization [59, 60]. We furthermore give finite sample concentration bounds for the stochastic reconstruction error of the Tikhonov solution.

When applied to conditional density estimation, this density reconstruction approach yields solutions that can capture multivariate, multimodal and non-Gaussian conditional densities off the shelf. This compares favorably with standard Gaussian Processes and is on par with neural conditional density models [64, 13]. In a set of experiments on toy and real-world data, we demonstrate that these properties lead to state-of-the-art results in conditional density estimation. To summarize, we *(i)* derive conditions under which a density can be reconstructed in the RKHS, *(ii)* give a consistent estimator for the reconstructed density in terms of a statistical inverse problem, *(iii)* provide dimensionality-independent finite sample error bounds for the estimation error, *(iv)* introduce CDOs, a multivariate, multimodal kernel-based conditional density model.

The rest of this paper is structured as follows: In Section 2, we state assumptions and introduce some preliminaries from the RKHS literature. Our main theoretical results are presented in Section 3, Section 4 discusses related work. Experiments on a toy dataset, rough terrain estimation and traffic prediction are reported in Section 5, while concluding remarks are presented in Section 6.

2 Preliminaries and assumptions

We consider a measurable space (\mathbb{X}, Σ) , where \mathbb{X} is a compact metric space endowed with the Borel σ -algebra Σ . Let $k: \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}$ be a symmetric positive semidefinite kernel which is continuous and induces an RKHS $H = \overline{\text{span}\{k(x, \cdot) \mid x \in \mathbb{X}\}}$, where the closure is with respect to the inner product $k(x, x') = \langle \phi(x), \phi(x') \rangle_H$. Here, $\phi(x) := k(x, \cdot)$ is known as the *canonical feature map*. We will typically require that H is separable, which is justified under mild assumptions [58, Lemma 4.33]. The *reproducing property* $f(x) = \langle f, \phi(x) \rangle_H$ holds for all $f \in H$ and $x \in \mathbb{X}$. For any finite measure ρ on \mathbb{X} , we typically assume $\int_{\mathbb{X}} \sqrt{k(x, x)} d\rho(x) \leq \infty$ such that the *kernel mean embedding* $\mu_\rho := \int_{\mathbb{X}} \phi(x) d\rho(x) \in H$ of the measure ρ exists [56]. Whenever ρ is a probability measure, the kernel mean embedding admits the standard estimate $\hat{\mu}_\rho := M^{-1} \sum_{i=1}^M \phi(x_i)$ with the i.i.d. sample $(x_i)_{i=1}^M \sim \rho$. We additionally assume that $\int k(x, x) d\rho(x) < \infty$. In this case, the *covariance operator*¹ $C_\rho := \int_{\mathbb{X}} \phi(x) \otimes \phi(x) d\rho(x)$ is well-defined as a positive self-adjoint Hilbert–Schmidt operator on H [3, 23, 40]. Here, $\phi(x) \otimes \phi(x): H \rightarrow H$ given by $f \mapsto \phi(x) \langle f, \phi(x) \rangle_H = \phi(x) f(x)$ for all $f \in H$ is the rank-one *tensor product operator*. We obtain the empirical standard estimate $\hat{C}_\rho := M^{-1} \sum_{i=1}^M \phi(x_i) \otimes \phi(x_i)$ with data as given above. Both empirical estimates $\hat{\mu}_\rho$ and \hat{C}_ρ converge with $\mathcal{O}(M^{-1/2})$ in the measure ρ in both RKHS and Hilbert–Schmidt norm [40].

The $L_2(\rho)$ -*integral operator* associated with k is defined by $(\mathcal{E}_\rho f)(x') := \int k(x, x') f(x) d\rho(x)$ for all $f \in L_2(\rho)$. This operator is positive, self-adjoint, trace-class and therefore compact. Since both C_ρ and \mathcal{E}_ρ are compact, their relation can be expressed explicitly in terms of their eigendecompositions as a consequence of Mercer’s theorem [38]. We collect related results in the supplementary material and only state the important facts here [see 58, Section 4.5]. It is known that the operators \mathcal{E}_ρ and C_ρ share the same set of strictly positive eigenvalues $(\lambda_i)_{i \in I}$, where I is an at most countable index set and the λ_i are sorted in non-increasing order and form a zero sequence whenever I is not finite. The corresponding eigenfunctions $e_i \in L_2(\rho)$ form an orthonormal basis of $L_2(\rho)$. Additionally, every $L_2(\rho)$ eigenfunction class e_i can be identified with a unique continuous representative $\tilde{e}_i \in H$ such that $e_i = \tilde{e}_i$ holds ρ -almost everywhere and the rescaled versions $\sqrt{\lambda_i} \tilde{e}_i \in H$ are exactly the eigenfunctions of C_ρ corresponding to the eigenvalue λ_i . Furthermore, the functions $\sqrt{\lambda_i} \tilde{e}_i$ form an orthonormal system in H . In particular, the spaces $L_2(\rho)$ and H admit an isometric isomorphism defined by the componentwise map $e_i \mapsto \sqrt{\lambda_i} \tilde{e}_i$.

A measurable reproducing kernel $k: \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}$ is said to be *universal* on (\mathbb{X}, Σ) if the measure embedding map $\rho \mapsto \int_{\mathbb{X}} \phi(x) d\rho(x) = \mu_\rho \in H$ is injective on the set of finite signed measures on (\mathbb{X}, Σ) [56]. We remark that the original definition of universality is more complex than our simplified definition. See [57, 39, 56] for a general overview of this topic. As an example in our setting, the Gaussian kernel $k(x, x') = \exp(-\|x - x'\|_2^2 / 2\sigma^2)$ is universal on compact subsets of \mathbb{R}^d [39]. For any measure ν on \mathbb{X} , we write $\nu \ll \rho$ whenever ν is *absolutely continuous* with respect to ρ . We write the *Radon–Nikodym* derivative of ν with respect to ρ as $\frac{d\nu}{d\rho}$.

The general theory of inverse problems [15], pseudoinverse operators [6, 17, 18], and regularization [59, 60, 16] has been well studied in the context of statistical learning over the last years [45, 11, 8, 50, 12], we will therefore define these concepts only briefly. In general, the compact operator C_ρ cannot be inverted on the whole space H . However, it admits a *pseudoinverse* C_ρ^\dagger , which is a (generally unbounded) operator with domain $\text{range}(C_\rho) + \text{range}(C_\rho)^\perp$. The minimum norm solution to the inverse problem $C_\rho u = f$ with known right-hand side $f \in \text{dom}(C_\rho^\dagger)$ is given by $u^\dagger := C_\rho^\dagger f$ and is unique, but solutions of larger norm can exist in general. In practice, one can resort to the *Tikhonov-regularized* solution $u_\alpha := (C_\rho + \alpha \mathcal{I}_H)^{-1} f$ (for a regularization parameter $\alpha > 0$) to stabilize the problem with respect to perturbed right-hand sides f and ensure that the solution is still well-defined. Consistency results for regularization schemes $\alpha \rightarrow 0$ have been derived in numerous settings depending on the problem.

We now give a brief overview of conditional mean embeddings [53, 28, 54, 40], which are the foundation of our model for conditional density estimation. Note that the original work formulates results in terms of (generally not existing) inverse operators under adequate regularity assumptions. We use pseudoinverses instead of inverses, which is equivalent under the given regularity assumptions

¹Note that technically, the term *covariance operator* is misleading when ρ is not a probability measure. Since we will require ρ to be finite, we will nevertheless use this term to reflect the standard definition.

and aligns with the literature on inverse problems. Assume we have a compact metric output space \mathbb{Y} endowed with the Borel σ -algebra and a continuous, positive semidefinite kernel $\ell: \mathbb{Y} \times \mathbb{Y} \rightarrow \mathbb{R}$ inducing a separable RKHS F with feature map $\psi(y) := \ell(y, \cdot)$. All other assumptions we made for the space \mathbb{X} , its RKHS, and measures on \mathbb{X} apply likewise for the output space \mathbb{Y} and associated objects. Assume two random variables X and Y with sample spaces \mathbb{X} and \mathbb{Y} follow the joint distribution \mathbb{P}_{XY} with marginals $\mathbb{P}_X, \mathbb{P}_Y$ and induced conditional distribution $\mathbb{P}_{Y|X}$. Let $C_{YX} := \int_{\mathbb{X}} \psi(y) \otimes \phi(x) d\mathbb{P}_{XY}(x, y)$ be the induced cross-covariance operator from H to F and C_X the covariance operator on H , respectively. Then the *conditional mean operator* (CMO) is defined as $\mathcal{U}_{Y|X} = C_{YX} C_X^\dagger: H \rightarrow F$ and satisfies the equation $\mu_{\mathbb{P}_y} = \mathcal{U}_{Y|X} \mu_{\mathbb{P}}$ for some distribution \mathbb{P} on \mathbb{X} , where $\mathbb{P}_y(\cdot) = \int_{\mathbb{X}} \mathbb{P}_{Y|X=x}(\cdot) d\mathbb{P}(x)$ [53, 54]. In particular, if \mathbb{P} is the Dirac measure on $x' \in \mathbb{X}$, this yields $\mu_{\mathbb{P}_{Y|X=x'}} = \mathcal{U}_{Y|X} k(x', \cdot)$.² Note that the CMO is in general not a globally defined bounded operator. It is defined pointwise as $\mu_{\mathbb{P}_y} = \mathcal{U}_{Y|X} \mu_{\mathbb{P}} \in F$ for $\mu_{\mathbb{P}} \in \text{range}(C_\rho)$ under the condition that $\mathbb{E}[g(Y) | X = \cdot] \in H$ for all $g \in F$. This requirement is examined in [23, Appendix A.1]. In practical applications, the pseudoinverse C_X^\dagger is usually replaced with its Tikhonov-regularized analogue, ensuring that $\mathcal{U}_{Y|X}$ is globally defined and bounded.

3 Density reconstruction and conditional density operators

We now show how the density reconstruction from a kernel mean embedding can be formulated in terms of an inverse problem admitting a unique analytical solution. We additionally prove that under verifiable conditions on the regularization parameter, the popular Tikhonov approximation of the pseudoinverse solution yields consistent estimates of this solution. Finally, we derive finite sample error bounds for the estimation error.

From now on, we will assume that Radon–Nikodym derivatives of the probability measure of interest \mathbb{P} with respect to a positive finite reference measure ρ are elements of $L_2(\rho)$. We first present our theoretical main result under the assumption that the $L_2(\rho)$ equivalence class of functions associated with the density admits a continuous representative in the RKHS. Note that this imposes some restrictions, although similar assumptions are typical in the scenario of kernel embeddings [see 23, 53, 25]. The key insight leading to our main result is the observation that whenever an RKHS function $\tilde{p} \in H$ is a Radon–Nikodym derivative of a distribution \mathbb{P} with respect to reference ρ , the kernel embedding $\mu_{\mathbb{P}} \in H$ can be expressed by applying C_ρ to \tilde{p} , as noted for example in [25]. We now show that under regularity assumptions, the corresponding inverse problem yields a unique solution of the density reconstruction problem.

Proposition 3.1 (Uniqueness of reconstructed densities). *Let $k: \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}$ be universal on (\mathbb{X}, Σ) . Let furthermore ρ be a finite positive measure with full support on \mathbb{X} and \mathbb{P} be a probability measure on (\mathbb{X}, Σ) such that $\mathbb{P} \ll \rho$ and $p := \frac{d\mathbb{P}}{d\rho} \in L_2(\rho)$. Additionally assume that p admits a continuous representative in H . Let $\mu_{\mathbb{P}} \in H$ be the mean embedding of \mathbb{P} . Then the inverse problem*

$$C_\rho u = \mu_{\mathbb{P}}, \quad u \in H, \quad (1)$$

has a unique solution $u^\dagger \in H$ such that $u^\dagger = p$ holds ρ -almost everywhere. That is, $u^\dagger \in H$ is the continuous representative of the equivalence class $p \in L_2(\rho)$.

Proof. Let $\tilde{p} \in H$ be the representative of $p \in L_2(\rho)$. By construction, we have

$$C_\rho \tilde{p} = \int_{\mathbb{X}} \phi(x) \tilde{p}(x) d\rho(x) = \int_{\mathbb{X}} \phi(x) d\mathbb{P}(x) = \mu_{\mathbb{P}},$$

so therefore $\tilde{p} \in H$ is a solution to the inverse problem. Assume that there exists a second solution $u' \in H$ of (1) with $u' \neq p$ (in the $L_2(\rho)$ sense). Then u' induces a finite signed measure ν on (\mathbb{X}, Σ) through the relation $\nu(A) := \int_A u'(x) d\rho(x)$ for all $A \in \Sigma$ with $\rho \neq \nu$. By construction, we have $u' = \frac{d\nu}{d\rho}$. We immediately see that in a similar fashion, we obtain the embedding of the measure ν by applying C_ρ to u' . That is, $C_\rho u' = \int_{\mathbb{X}} \phi(x) u'(x) d\rho(x) = \int \phi(x) d\nu(x) = \mu_\nu$. This implies that $\mu_{\mathbb{P}} = \mu_\nu$, which is a contradiction to the kernel being universal. ■

²The latter is how the CMO is usually introduced, while $\mu_{\mathbb{P}_y} = \mathcal{U}_{Y|X} \mu_{\mathbb{P}}$ is referred to as *kernel sum rule* in the literature.

We now give a sufficient condition for $L_2(\rho)$ function classes to admit a continuous RKHS representative based on the spectral decomposition of \mathcal{E}_ρ .

Lemma 3.2. *Let ρ be a positive finite measure with full support on \mathbb{X} . Let \mathbb{P} be a probability measure on \mathbb{X} such that $\mathbb{P} \ll \rho$. Furthermore, let $(\lambda_i, e_i)_{i \in I}$ be the eigenvalue/eigenfunction pairs of \mathcal{E}_ρ and assume $p := \frac{d\mathbb{P}}{d\rho} \in L_2(\rho)$ is the Radon–Nikodym derivative of \mathbb{P} with respect to ρ . If $\left(\langle p, e_i \rangle_{L_2(\rho)} \lambda_i^{-1/2} \right)_{i \in I} \in \ell_2(I)$, then the equivalence class of $p \in L_2(\rho)$ has a continuous representative $\tilde{p} \in H$ such that $p = \tilde{p}$ holds ρ -almost everywhere.*

The above result is a direct consequence of Mercer’s theorem, its proof can be found in the supplementary material. The requirement that the reweighted basis coefficients of a Radon–Nikodym derivative are in $\ell_2(I)$ can hardly be assessed in the very general case. However, the eigenvalue/eigenfunction pairs $(\lambda_i, e_i)_{i \in I}$ of \mathcal{E}_ρ are known for specific kernels. For the Gaussian kernel, the eigenfunctions e_i can be written in terms of Hermite polynomials [64]. The analytical examination of function classes which fulfill the above assumption is beyond the scope of this paper and subject to future research. From now on, we assume that the Radon–Nikodym derivatives of interest admit a representative in the RKHS H .

Proposition 3.1 in combination with the spectral decomposition of C_ρ shows that the reconstruction of the embedded density is given by $u^\dagger := C_\rho^\dagger \mu_\mathbb{P} = \sum_i \lambda_i^{-1/2} \langle \sqrt{\lambda_i} \tilde{e}_i, \mu_\mathbb{P} \rangle_H \sqrt{\lambda_i} \tilde{e}_i$. Although C_ρ^\dagger is in general not a globally defined bounded operator, Proposition 3.1 ensures that the reconstructed density $u^\dagger \in H$ is well-defined, unique and coincides with $\frac{d\mathbb{P}}{d\rho}$ ρ -almost everywhere. We may therefore tackle the density reconstruction problem with the classical toolset for inverse problems. In what follows, we use Proposition 3.1 to reconstruct densities from the embeddings of conditional distributions, giving rise to the *conditional density operator* (CDO). The CDO allows to estimate a conditional density over an output domain given an input value or a distribution over the input domain. Assume in what follows that we have fixed a finite positive reference measure ρ_y on \mathbb{Y} , such that C_{ρ_y} is a well-defined positive self-adjoint Hilbert–Schmidt operator on F . From now on, densities on \mathbb{Y} will be Radon–Nikodym derivatives with respect to ρ_y . The following result is a direct consequence of Proposition 3.1.

Theorem 3.3 (Conditional density operator). *Assume $\mathbb{P}_y(\cdot) = \int_{\mathbb{X}} \mathbb{P}_{Y|X=x}(\cdot) d\mathbb{P}(x)$ admits a Radon–Nikodym derivative $p_y \in L_2(\rho_y)$ with respect to the reference measure ρ_y , such that the assumptions of Proposition 3.1 are satisfied. Additionally assume that the conditional mean operator $\mathcal{U}_{Y|X} = C_{YX} C_X^\dagger$ for $\mathbb{P}_{Y|X}$ exists. Then $\mathcal{A}_{Y|X} \mu_\mathbb{P} := C_{\rho_y}^\dagger \mu_\mathbb{P}_y = C_{\rho_y}^\dagger \mathcal{U}_{Y|X} \mu_\mathbb{P} = C_{\rho_y}^\dagger C_{YX} C_X^\dagger \mu_\mathbb{P} \in F$ exists and satisfies*

$$p_y \stackrel{\rho_y \text{ a.e.}}{=} \mathcal{A}_{Y|X} \mu_\mathbb{P}.$$

If \mathbb{P} is the Dirac measure on x' , this entails $\frac{d\mathbb{P}_{Y|X=x'}}{d\rho_y} \stackrel{\rho_y \text{ a.e.}}{=} \mathcal{A}_{Y|X} k(x', \cdot)$.

We call $\mathcal{A}_{Y|X} = C_{\rho_y}^\dagger C_{YX} C_X^\dagger$ the *conditional density operator* (CDO). Conditional density operators have several advantages over Gaussian Processes (GPs), the mainstream kernel method for conditional density estimation [64]. In particular, they allow for density estimation in arbitrary output dimensions, unlike standard GPs, which estimate a 1d density (see the literature on multi-output GPs for a remedy, e.g. [1, 7]). Also, multiple modes in the output are captured by CDOs. Though this might be achieved with GP mixtures, CDOs allow for more flexibility with respect to the mixture components. A CDOs output could be a mixture of multivariate Laplace or Student- t densities; any universal kernel that is also a probability density gives rise to CDOs where the output density can not only be evaluated, but also trivially sampled from. See Figure 1 for plots illustrating reconstructed multivariate, multimodal conditional densities, and Section 5.1 for a description of the data generating process.

3.1 Consistency and convergence rate of the Tikhonov-regularized solution

In practical applications, we cannot access C_ρ analytically. The idea is now to choose the reference measure ρ to be normalized on \mathbb{X} , i.e., as a probability measure, such that C_ρ can be estimated. Additionally, $\mu_\mathbb{P}$ is also only given in terms of an estimate $\hat{\mu}_\mathbb{P}$. Instead of computing the analytical density reconstruction $u^\dagger = C_\rho^\dagger \mu_\mathbb{P}$, we construct an empirical estimate of u^\dagger by defining the empirical Tikhonov-regularized solution

$$\hat{u} := (\hat{C}_\rho + \alpha \mathcal{I}_H)^{-1} \hat{\mu}_\mathbb{P} \quad (2)$$

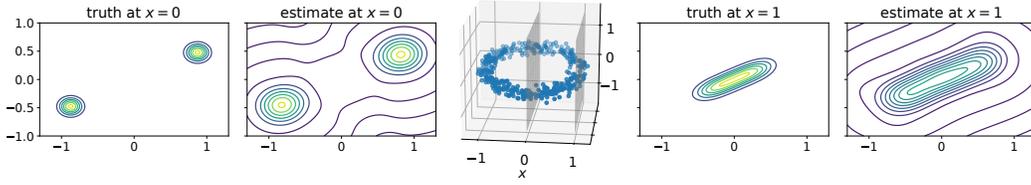


Figure 1: Cross sections of a donut shaped density and estimates using a CDO.

for a regularization parameter $\alpha > 0$. We examine this problem under the assumption that \widehat{C}_ρ is the standard estimate based on M i.i.d. ρ -samples and $\widehat{\mu}_\mathbb{P}$ is estimated from N i.i.d. \mathbb{P} -samples. Note that in practice $\widehat{\mu}_\mathbb{P}$ might instead be an output of another model. In what follows, we show that the reconstruction error $\|u^\dagger - \widehat{u}\|_H$ vanishes in probability as $M, N \rightarrow \infty$ for an appropriately chosen positive regularization scheme $\alpha \rightarrow 0$. We define the regularized analytic solution $u_\alpha := (C_\rho + \alpha \mathcal{I}_H)^{-1} \mu_\mathbb{P}$ and decompose the total error:

$$\|u^\dagger - \widehat{u}\|_H \leq \|u^\dagger - u_\alpha\|_H + \|u_\alpha - \widehat{u}\|_H. \quad (3)$$

The first error term is deterministic and depends only on the analytical nature of the problem based on the decay of the eigenvalues of C_ρ . For $\alpha \rightarrow 0$, we always have $\|u^\dagger - u_\alpha\|_H \rightarrow 0$ by standard results from inverse problem theory [16]. While the convergence rate can be given in specific cases when the eigenvalue decay of C_ρ is known, the convergence can be arbitrarily slow in general [48]. The next result is based on a Hilbert space version of Hoeffding's inequality [42, 43] and gives a general concentration bound for the estimation error term $\|u_\alpha - \widehat{u}\|_H$.

Proposition 3.4 (Finite sample bound of estimation error). *Let $\sup_x \sqrt{k(x, x)} = \sup_x \|\phi(x)\|_H = c < \infty$ and $\alpha > 0$ be a fixed regularization parameter. Let $0 < a < 1/2$ and $0 < b < 1/2$ be fixed. If $\widehat{C}_\rho = M^{-1} \sum_{i=1}^M \phi(x_i) \otimes \phi(x_i)$ with $(x_i)_{i=1}^M \stackrel{\text{i.i.d.}}{\sim} \rho$ and $\widehat{\mu}_\mathbb{P} = N^{-1} \sum_{j=1}^N \phi(x'_j)$ with $(x'_j)_{j=1}^N \stackrel{\text{i.i.d.}}{\sim} \mathbb{P}$ and both sets of samples are independent, then we have*

$$\begin{aligned} \Pr \left[\|u_\alpha - \widehat{u}\|_H \leq \frac{M^{-2b}}{\alpha^2} (\|\mu_\mathbb{P}\|_H + N^{-2a}) + \frac{N^{-2a}}{\alpha} \right] \\ \geq \left[1 - 2 \exp \left(-\frac{N^{1-2a}}{8c^2} \right) \right] \left[1 - 2 \exp \left(-\frac{M^{1-2b}}{8c^4} \right) \right]. \end{aligned} \quad (4)$$

The proof can be found in the supplementary material. We emphasize that the above error bound does not depend on the dimensionality of the data. By combining the convergence of the deterministic error and the convergence in probability given by Proposition 3.4, we can obtain a regularization scheme which ensures that \widehat{u} is a consistent estimate of u^\dagger .

Corollary 3.5 (Consistency). *Let $\alpha = \alpha(M, N)$ be a regularization scheme such that $\alpha(M, N) \rightarrow 0$ as well as*

$$\frac{M^{-2b}}{\alpha(M, N)^2} \rightarrow 0 \quad \text{and} \quad \frac{N^{-2a}}{\alpha(M, N)} \rightarrow 0 \quad (5)$$

as $M, N \rightarrow \infty$. Then the empirical solution \widehat{u} obtained from (2) regularized with the scheme $\alpha(M, N)$ converges in probability to the analytical solution u^\dagger .

In practice, after picking a, b as above and $c' \in (0, 1)$, the scheme $\alpha(M, N) = \max(M^{-bc'}, N^{-2ac'})$ guarantees consistency. Note that Proposition 3.4 gives bounds for the case that $\widehat{\mu}_\mathbb{P}$ is given in terms of its standard estimate. In the CDO, $\widehat{\mu}_\mathbb{P}$ is given as a conditional mean embedding. The proof of Proposition 3.4 can be easily adapted to work with arbitrary concentration bounds for $\widehat{\mu}_\mathbb{P}$. Additionally, the proof can be modified to obtain concentration bounds for the estimation of conditional mean embeddings [53], since the conditional mean embedding operator amounts to a similar regularized inverse of a covariance operator and additional composition with a cross-covariance operator.

3.2 Estimation of conditional density operators

The CDO is defined pointwise when the assumptions of Theorem 3.3 are satisfied. Analogously to the empirical inverse problem above, we replace the pseudoinverses of both C_x and C_{ρ_y} with

their regularized inverses for the empirical version of the CDO. That is, from the analytical version $\mathcal{A}_{Y|X} = C_{\rho_y}^\dagger C_{YX} C_X^\dagger$, we obtain $\widehat{\mathcal{A}}_{Y|X} = (\widehat{C}_{\rho_y} + \alpha' \mathcal{I}_F)^{-1} \widehat{C}_{YX} (\widehat{C}_X + \alpha \mathcal{I}_H)^{-1}$. Note that in contrast to the analytical version, the empirical version of the CDO is a globally defined bounded operator.

Assume that we have samples $(x_i, y_i)_{i=1}^N \stackrel{\text{i.i.d.}}{\sim} \mathbb{P}_{XY}$ such that the \mathbb{P}_{XY} -induced conditional distribution $\mathbb{P}_{Y|X}$ is the distribution of interest. Assume that we furthermore have samples $(z_i)_{i=1}^M \stackrel{\text{i.i.d.}}{\sim} \rho_y$, where ρ_y is the reference measure on \mathbb{Y} which we use to reconstruct the desired conditional density. The density over \mathbb{Y} induced by fixing the input at $x' \in \mathbb{X}$ is approximated as

$$\widehat{\mathcal{A}}_{Y|X} k(x', \cdot) \approx \sum_{i=1}^M \beta_i \ell(z_i, \cdot) \quad (6)$$

with $\beta = M^{-2} (L_Z + \alpha' I_M)^{-2} L_{ZY} (K_X + N \alpha I_N)^{-1} [k(x_1, x'), \dots, k(x_N, x')]^\top \in \mathbb{R}^M$, where we use the kernel matrices $K_X = [k(x_i, x_j)]_{i,j} \in \mathbb{R}^{N \times N}$, $L_Y = [\ell(y_i, y_j)]_{i,j} \in \mathbb{R}^{M \times M}$ as well as $L_{ZY} = [\ell(z_i, y_j)]_{i,j} \in \mathbb{R}^{M \times N}$ and the corresponding identity matrices $I_N \in \mathbb{R}^{N \times N}$, $I_M \in \mathbb{R}^{M \times M}$. If one is interested in the conditional distribution of y for $x \sim \mathbb{P}$, the $k(x_j, x')$ are replaced by $\mu_{\mathbb{P}}(x_j)$ in the expression for β . The derivation of the representation in (6) builds upon a similar derivation of the conditional mean embedding estimate and can be found in the supplementary material. Detailed convergence rates and error bounds for this empirical estimate are beyond the scope of this paper, but consistency can easily be derived by combining results from Section 3.1 with various convergence results for the conditional mean embedding [53, 28, 22].

4 Related work

Finding the pre-image of a feature vector in the RKHS induced by the kernel is a classical problem in kernel methods [35, 4]. In the context of this work, we aim to reconstruct a probability density p from the kernel mean embedding $\mu_{\mathbb{P}}$ of some distribution \mathbb{P} . In the literature, two popular approaches for recovering information from kernel mean embeddings exist, namely distributional pre-image learning [52, 31] and kernel herding [9].

Given an empirical kernel mean $\hat{\mu}_{\mathbb{P}}$, the idea of distributional pre-image learning is to pick a family of densities $\mathcal{P} = \{p_\theta : \theta \in \Theta\}$ and then find $\theta^* = \arg \min_{\theta \in \Theta} \|\hat{\mu}_{\mathbb{P}} - \mu_{p_\theta}\|_H^2$ [52]. The drawback of this approach is that it requires parametric assumptions on the family of densities \mathcal{P} . Moreover, it requires solving a constrained non-convex optimization problem. On the other hand, our method provides an analytic solution which only requires that \mathbb{P} is absolutely continuous with respect to the reference measure ρ .

Kernel herding, on the other hand, aims to greedily generate a representative set of T pseudo-samples from \mathbb{P} in a deterministic fashion using the estimate $\hat{\mu}_{\mathbb{P}}$ [9]. The advantage of herding is that it is shown to exhibit an integration error of order $\mathcal{O}(T^{-1})$. Similarly, our method gives rise to a probability density from which random samples can be easily generated. While our work also relates to the literature of kernel-based density ratio estimation [32, 44], our goal is not to estimate a density ratio, but an actual density with respect to a specific reference measure ρ . Furthermore, unlike previous work, we provide a rigorous treatment of the L_2 properties of our estimates and good choices for regularization constants.

The kernel mean embedding has recently been applied to learn high-dimensional *implicit* density models such as generative adversarial networks (GANs) [14, 37, 36] and autoencoders [62]. It would also be interesting to extend our results to this area of research.

Classical methods for (conditional) density estimation [5, 30] are known to suffer from slow convergence in high dimensions, see, e.g., Tsybakov [63, Chap. 1]. Some methods propose estimators that are similar to the CDO, although not making use of RKHS arguments and not proving consistency [5]. An advantage of the CDO is that it is less prone to the curse of dimensionality. Concretely, the convergence rate of Proposition 3.4 and regularizing scheme from Corollary 3.5 do not depend on the problem dimension. Nevertheless, it might affect the deterministic error which could converge arbitrarily slowly, see, e.g., Tolstikhin et al. [61]. Neural density models can also scale gracefully with increasing dimensions, as demonstrated empirically especially in the image generation domain [34, 13]. However, little theory exists to confirm this observation and understand under which conditions on the problem and the network architecture it applies. Standard neural density models can easily be extended to include conditioning on an input variable. However, conditioning on a

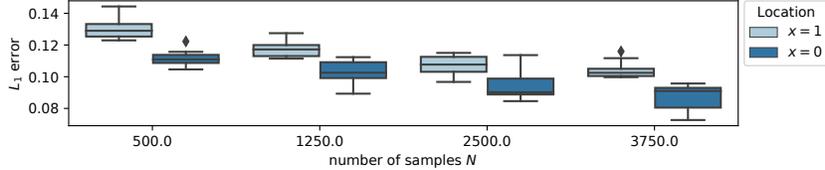


Figure 2: Errors of conditional density estimation for the Gaussian donut in $L_1(\rho_y)$ -norm.

distribution over the input variable is non-trivial, unlike in the CDO setting.

In the RKHS setting, infinite dimensional exponential families (IDEF) assume the log-likelihood of a density to be an RKHS function and where recently extended to conditional density estimation [55, 2]. Fitting such a model is solved by using an optimization approach, while CDOs allow closed form solutions. Sampling from IDEF approximations requires MCMC techniques rather than ordinary Monte Carlo, as possible with our approach. Sampling is necessary to estimate predictive mean and variance in IDEF models, while closed form expressions exist for CDOs, see A.3.1.

5 Experiments

In this section, we report results on one toy and two real-world datasets, showing competitive performance of the CDO in conditional density estimation tasks in comparison to recent state-of-the-art approaches. We use a computational trick for large datasets described, along with a trick for high-dimensional output spaces, in the supplementary material A.4.

5.1 Gaussian donut

For this toy example, we pick 50 points on a circle in the (x, y) plane, embed them into a 3D ambient space and slightly rotate them around the y axis. Each of the points is the mean of an isotropic Gaussian distribution, and each mean has equal probability, giving rise to a mixture that we call a Gaussian donut. We draw 50 samples from the isotropic Gaussian noise distribution per mean to form the training data for a CDO that estimates the density on y, z coordinates given x . The reference measure ρ_y is taken to be the uniform distribution on a zero-centered square with side length 4. See Figure 1 for the ground truth density and CDO estimate at x equal to 0 and 1, respectively. We report numerical errors in density approximation in $L_1(\rho_y)$ -norm, i.e., $\|\hat{u} - p_{Y,Z|X=x}\|_{L_1}$, for an increasing number of samples per mean, resulting in N overall samples from the Gaussian donut. The uniform reference measure is represented by a regular grid of $M = \lfloor \sqrt{N} \rfloor^2$ points. The procedure is repeated 10 times for different random seeds. See Figure 2 for a plot of the L_1 error.

5.2 Rough terrain reconstruction

Rough terrain reconstruction is used in robotics and navigation [29, 27]. Given measurements of longitude, latitude, and altitude, the task is to estimate the altitude for unobserved coordinates on a map. We reproduce an experiment from [19], considering around 23 million non-uniformly sampled measurements of Mount St. Helens, binned into a 120×117 grid. We randomly chose 90% of the data as training, the rest as test data. We fit an exact Gaussian Process by optimizing the length scale of a Gaussian kernel with respect to marginal likelihood of the training data and compute the scaled mean absolute error (SMAE) for the test locations. For the conditional density operator, we pick a Gaussian kernel for input and output domains. The input length scale is chosen using the median heuristic, i.e., $\sigma^2 = \text{median}\{\|x_i - x_j\|_2^2 : 1 \leq i < j \leq n\}$. The output domain is chosen as an interval based on the minimum and maximum of the training output data, with a uniform reference measure represented by equidistant grid points, the output length scale based on the distance between adjacent grid points. Using this procedure, for the GP we obtain an SMAE of 0.0358 ± 0.00062 , whereas with the conditional density operator we get 0.0269 ± 0.00055 . This hints at the fact that our method is competitive with other kernel-based learning algorithms. We conjecture that added flexibility is a reason for this. While the output distribution of the GP is a Gaussian, in the CDO used here it is a mixture of Gaussians. A related possibility is that we use a homoscedastic likelihood in the GP, leading to a certain minimum amount of assumed noise, while the CDO does not do this.

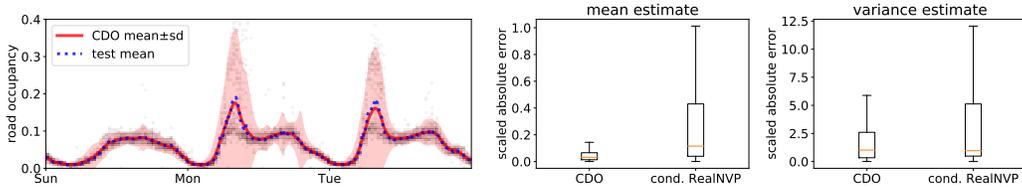


Figure 3: Road occupancy prediction experiment. *Left*: Histogram of test data for three days in black, test data mean and prediction. *Right*: Boxplots of scaled absolute errors with respect to test data.

Table 1: SMAE on the test set of road occupancy data

	CDO	Conditional Real NVP
mean	0.05 ± 0.06	0.32 ± 0.41
variance	2.58 ± 4.36	6.40 ± 22.64

5.3 Traffic density prediction from time features

In this experiment, we predict the occupancy rate of different locations on freeways in the San Francisco bay area based on a given day of week and time of day.³ The occupancy rate is encoded as a number between 0 and 1 for 963 different locations. The measurements are sampled every 10 minutes, resulting in 144 measurements per day (i.e., times of day). See Figure 3 for example histograms at one particular location.

In the training dataset, each day of week occurred 32 times (discarding measurements to get a balanced dataset), resulting in $32 \times 144 \times 7 = 32\,256$ input-output pairs. In the test set, each day of week occurred 20 times. The task is to get a predictive density for the locations occupancy given time of day and day of week as inputs.

We fit a conditional density operator using Gaussian kernels on the output and Laplacian kernels on the input domain. Laplacians are chosen because they result in smoother estimates, while Gaussians showed more oscillations for the output density estimates. The bandwidth is chosen based on a scaled median heuristic for the input. Samples for the uniform reference measure on the output domain are taken to be a regular grid between minimum and maximum occurring values. The bandwidth for the output kernel is chosen as the Euclidean distance between adjacent grid points. The regularization parameter for input and output domains is fixed using the practical scheme suggested in Section 3.1 using $a = b = 0.49$ and $c' = 0.9$. For comparison, we use a RealNVP deep neural network [13] written in PyTorch and fine-tuned for the dataset, conditioned on the input and trained with Adam [33]. We estimate the expectation (w.r.t. model predictive distribution) of the absolute error when estimating test set mean and variance, i.e., scaled mean absolute error (SMAE), as well as absolute error standard deviation. The conditional density operator allows to estimate these in closed form. As this is not possible for the RealNVP model, we draw 2000 samples for estimation. Even though the dataset is rather large, the CDO can be fitted in under 5 minutes using a scheme outlined in supplementary material A.4.1. We report scaled mean absolute errors for both models in Table 1. Clearly, our conditional density operator outperforms the RealNVP model.

6 Conclusion

In this paper, we show that the reconstruction of square-integrable densities from kernel mean embeddings can be formulated as an inverse problem under some regularity assumptions. In particular, we draw the connection to estimation of Radon–Nikodym derivatives with respect to a reference measure, for which the solution is shown to be exact almost everywhere under certain conditions. We prove that the popular Tikhonov approach to solving the inverse problem is consistent and allows for finite sample bounds on the stochastic error independent of the dimensionality of the data. However, we want to point out that the proposed solution scheme is only one possible approach which could be replaced by, e.g., conjugate gradient. We focus on the conditional density operator as one application.

³Data available at <https://archive.ics.uci.edu/ml/datasets/PEMS-SF>, detailed description in [10].

The CDO is closely related to the conditional mean embedding, can model multivariate, multimodal conditional distributions and performs competitively in our experiments.

In future work, numerical routines for scaling the method up to even larger datasets will be of interest. One way to do this might be conjugate gradient algorithms and making use of Toeplitz and Kronecker structure in the kernels, as recently done in fitting GPs [26, 65]. Theoretical avenues to take might be to alleviate the compactness conditions on the involved data spaces and finding rigorously justified ways of choosing good kernels and kernel parameters.

Acknowledgments

We would like to thank Kashif Rasul for providing the conditional RealNVP implementation for the traffic dataset and Ilja Klebanov and Tim Sullivan for helpful discussions and pointing out relevant references. Partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germanys Excellence Strategy – MATH+: The Berlin Mathematics Research Center, EXC-2046/1 – project ID: 390685689.

References

- [1] M. A. Alvarez, L. Rosasco, N. D. Lawrence, et al. Kernels for vector-valued functions: A review. *Foundations and Trends® in Machine Learning*, 4(3):195–266, 2012.
- [2] M. Arbel and A. Gretton. Kernel conditional exponential family. In *International Conference on Artificial Intelligence and Statistics*, pages 1337–1346, 2018.
- [3] C. Baker. Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society*, 186:273–289, 1973.
- [4] G. H. Bakir, J. Weston, and B. Schölkopf. Learning to find pre-images. In *Advances in Neural Information Processing Systems 16*, pages 449–456. MIT Press, 2004.
- [5] D. M. Bashtannyk and R. J. Hyndman. Bandwidth selection for kernel conditional density estimation. *Computational Statistics & Data Analysis*, 36(3):279–298, 2001.
- [6] F. Beutler. The operator theory of the pseudo-inverse. *J. Math. Anal. Appl.*, 10:451–470, 1965.
- [7] P. Boyle and M. Frean. Dependent Gaussian processes. In *Advances in Neural Information Processing Systems*, pages 217–224, 2005.
- [8] A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.
- [9] Y. Chen, M. Welling, and A. Smola. Super-samples from kernel herding. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, pages 109–116. AUAI Press, 2010.
- [10] M. Cuturi. Fast global alignment kernels. In *Proceedings of the 28th International Conference on Machine Learning*, pages 929–936, 2011.
- [11] E. De Vito, L. Rosasco, A. Caponnetto, U. De Giovannini, and F. Odone. Learning from examples as an inverse problem. *Journal of Machine Learning Research*, 6:883–904, 2005.
- [12] E. De Vito, L. Rosasco, and A. Caponnetto. Discretization Error Analysis for Tikhonov Regularization. *Analysis and Applications*, 04(01):81–99, 2006.
- [13] L. Dinh, J. Sohl-Dickstein, and S. Bengio. Density estimation using Real NVP. In *International Conference on Learning Representations*, 2017.
- [14] G. K. Dziugaite, D. M. Roy, and Z. Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. In *UAI*, 2015.
- [15] H. Engl and C. W. Groetsch. *Inverse and Ill-Posed Problems*. Academic Press, 1996.
- [16] H. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Kluwer, 1996.
- [17] I. Erdelyi. A generalized inverse for arbitrary operators between Hilbert spaces. *Proc. Camb. Phil. Soc.*, 71(43):43–50, 1972.
- [18] I. Erdelyi and I. Ben-Israel. Extremal solutions of linear equations and generalized inversion between hilbert spaces. *J. Math. Anal. Appl.*, 39:298–313, 1972.

- [19] D. Eriksson, K. Dong, E. Lee, D. Bindel, and A. G. Wilson. Scaling Gaussian process regression with derivatives. In *Advances in Neural Information Processing Systems*, pages 6867–6877, 2018.
- [20] T. Evans and P. Nair. Scalable gaussian processes with grid-structured eigenfunctions (gp-grief). In *International Conference on Machine Learning*, pages 1416–1425, 2018.
- [21] S. Flaxman, A. Wilson, D. Neill, H. Nickisch, and A. Smola. Fast kronecker inference in gaussian processes with non-gaussian likelihoods. In *International Conference on Machine Learning*, pages 607–616, 2015.
- [22] K. Fukumizu. Nonparametric bayesian inference with kernel mean embedding. In G. Peters and T. Matsui, editors, *Modern Methodology and Applications in Spatial-Temporal Modeling*. 2017.
- [23] K. Fukumizu, F. R. Bach, and M. I. Jordan. Dimensionality reduction for supervised learning with Reproducing Kernel Hilbert Spaces. *Journal of Machine Learning Research*, 5:73–99, 2004.
- [24] K. Fukumizu, L. Song, and A. Gretton. Kernel bayes’ rule: Bayesian inference with positive definite kernels. *Journal of Machine Learning Research*, 14(1):3753–3783, 2013.
- [25] K. Fukumizu, L. Song, and A. Gretton. Kernel Bayes’ rule: Bayesian inference with positive definite kernels. *Journal of Machine Learning Research*, 14:3753–3783, 2013.
- [26] J. Gardner, G. Pleiss, K. Q. Weinberger, D. Bindel, and A. G. Wilson. Gpytorch: Black-box matrix-matrix gaussian process inference with gpu acceleration. In *Advances in Neural Information Processing Systems*, pages 7576–7586, 2018.
- [27] D. Gingras, T. Lamarche, J.-L. Bedwani, and É. Dupuis. Rough terrain reconstruction for rover motion planning. In *2010 Canadian Conference on Computer and Robot Vision*, pages 191–198. IEEE, 2010.
- [28] S. Grünewälder, G. Lever, L. Baldassarre, S. Patterson, A. Gretton, and M. Pontil. Conditional mean embeddings as regressors. In *International Conference on Machine Learning*, volume 5, 2012.
- [29] R. Hadsell, J. A. Bagnell, D. Huber, and M. Hebert. Space-carving kernels for accurate rough terrain estimation. *The International Journal of Robotics Research*, 29(8):981–996, 2010.
- [30] P. Hall, J. Racine, and Q. Li. Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association*, 99(468):1015–1026, 2004.
- [31] M. Kanagawa and K. Fukumizu. Recovering Distributions from Gaussian RKHS Embeddings. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pages 457–465. PMLR, 2014.
- [32] T. Kanamori, T. Suzuki, and M. Sugiyama. Statistical analysis of kernel-based least-squares density-ratio estimation. *Machine Learning*, 86(3):335–367, 2012.
- [33] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference for Learning Representations*, 2014.
- [34] D. P. Kingma and P. Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, pages 10215–10224, 2018.
- [35] J. T. Kwok and I. W.-H. Tsang. The pre-image problem in kernel methods. *IEEE Transactions on Neural Networks*, 15:1517–1525, 2003.
- [36] C.-L. Li, W.-C. Chang, Y. Cheng, Y. Yang, and B. Póczos. MMD GAN: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems*, pages 2203–2213, 2017.
- [37] Y. Li, K. Swersky, and R. Zemel. Generative moment matching networks. In *International Conference on Machine Learning*, pages 1718–1727, 2015.
- [38] J. Mercer. Functions of Positive and Negative Type, and their Connection with the Theory of Integral Equations. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 209(441-458):415–446, 1909.
- [39] C. A. Micchelli, Y. Xu, and H. Zhang. Universal kernels. *Journal of Machine Learning Research*, 7:2651–2667, 2006.

- [40] K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10(1–2): 1–141, 2017.
- [41] T. Nickson, T. Gunter, C. Lloyd, M. A. Osborne, and S. Roberts. Blitzkriging: Kronecker-structured stochastic gaussian processes. *arXiv preprint arXiv:1510.07965*, 2015.
- [42] I. Pinelis. An approach to inequalities for the distributions of infinite-dimensional martingales. *Probability in Banach Spaces: Proceedings of the Eighth International Conference*, 8:128–134, 01 1992.
- [43] I. Pinelis. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, 22(4):1679–1706, 1994.
- [44] Q. Que and M. Belkin. Inverse density as an inverse problem: The fredholm equation approach. In *Advances in neural information processing systems*, pages 1484–1492, 2013.
- [45] L. Rosasco, A. Caponnetto, E. D. Vito, F. Odone, and U. D. Giovannini. Learning, regularization and ill-posed inverse problems. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 1145–1152. MIT Press, 2005.
- [46] L. Rosasco, M. Belkin, and E. D. Vito. On learning with integral operators. *Journal of Machine Learning Research*, 11:905–934, 2010.
- [47] M. Schneider. Probability inequalities for kernel embeddings in sampling without replacement. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 66–74. PMLR, 2016.
- [48] E. Sock. On the asymptotic order of accuracy of Tikhonov regularization. *Journal of Optimization Theory and Applications*, 44(1):95–104, Sep 1984.
- [49] J. Shawe-Taylor, C. Williams, N. Cristianini, and J. Kandola. On the eigenspectrum of the gram matrix and its relationship to the operator eigenspectrum. In N. Cesa-Bianchi, M. Numao, and R. Reischuk, editors, *Algorithmic Learning Theory*, pages 23–40, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg.
- [50] S. Smale and D.-X. Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26(2):153–172, 2007.
- [51] A. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *Proceedings of the 18th International Conference on Algorithmic Learning Theory*, pages 13–31. Springer-Verlag, 2007.
- [52] L. Song, X. Zhang, A. Smola, A. Gretton, and B. Schölkopf. Tailoring density estimation via reproducing kernel moment matching. In *Proceedings of the 25th International Conference on Machine Learning*, pages 992–999, New York, NY, USA, 2008. ACM.
- [53] L. Song, J. Huang, A. Smola, and K. Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 961–968, 2009. doi: 10.1145/1553374.1553497.
- [54] L. Song, K. Fukumizu, and A. Gretton. Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. *IEEE Signal Processing Magazine*, 30(4):98–111, 2013.
- [55] B. Sriperumbudur, K. Fukumizu, A. Gretton, A. Hyvärinen, and R. Kumar. Density estimation in infinite dimensional exponential families. *The Journal of Machine Learning Research*, 18(1): 1830–1888, 2017.
- [56] B. K. Sriperumbudur, K. Fukumizu, and G. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12(Jul):2389–2410, 2011.
- [57] I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2001.
- [58] I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, 2008.
- [59] A. N. Tikhonov and V. Y. Arsenin. *Solutions of Ill Posed Problems*. W. H. Winston, 1977.
- [60] A. N. Tikhonov, V. V. S. A. V. Goncharsky, and A. G. Yagola. *Numerical methods for the solution of ill-posed problems*. Kluwer, 1995.

- [61] I. Tolstikhin, B. K. Sriperumbudur, and K. Muandet. Minimax estimation of kernel mean embeddings. *Journal of Machine Learning Research*, 18(1):3002–3048, 2017.
- [62] I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schoelkopf. Wasserstein auto-encoders. In *International Conference on Learning Representations*, 2018.
- [63] A. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008.
- [64] C. K. Williams and C. E. Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT Press Cambridge, MA, 2006.
- [65] A. Wilson and H. Nickisch. Kernel interpolation for scalable structured gaussian processes (kiss-gp). In *International Conference on Machine Learning*, pages 1775–1784, 2015.

A Supplementary material

A.1 Mercer's theorem and the isomorphism between $L_2(\rho)$ and H

We refer to [58] for a detailed derivation and proofs of the following results.

Proposition A.1. *Let \mathcal{X} be a compact metric space endowed with the Borel σ -algebra and $k: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ be a continuous kernel. Let ρ be a positive finite measure on (\mathcal{X}, Σ) with full support on \mathcal{X} , such that $\int k(x, x) d\rho(x) < \infty$. Then the following statements hold:*

- (i) *The integral operator \mathcal{E}_ρ has eigenvalue/eigenfunction pairs $(\lambda_i, e_i)_{i \in I} \subseteq \mathbb{R} \times L_2(\rho)$, where I is a finite or countably infinite index set and the eigenvalues λ_i are positive and non-increasing:*

$$\mathcal{E}_\rho = \sum_{i \in I} \lambda_i e_i \otimes e_i.$$

Additionally, when I is not finite then $(\lambda_i)_{i \in I}$ is a zero sequence. The eigenfunctions e_i form a complete orthonormal system in $L_2(\rho)$.

- (ii) *Every eigenfunction $e_i \in L_2(\rho)$ can be interpreted as a function in H , i.e., there exists a unique function $\tilde{e}_i \in H$ such that $e_i = \tilde{e}_i$ ρ -almost everywhere.*

- (iii) **Mercer's theorem** [38]: *The kernel k can be expressed as*

$$k(x, x') = \sum_{i \in I} \lambda_i \tilde{e}_i(x) \tilde{e}_i(x')$$

for all $x, x' \in \mathcal{X}$. The convergence of the above series is absolute and uniform.

- (iv) *Given the eigendecomposition of \mathcal{E}_ρ , the covariance operator $C_\rho: H \rightarrow H$ has the eigenvalue/eigenfunction pairs $(\lambda_i, \sqrt{\lambda_i} \tilde{e}_i)_{i \in I} \subseteq \mathbb{R} \times H$, where $\tilde{e}_i \in H$ is given as described above:*

$$C_\rho = \sum_{i \in I} \lambda_i (\sqrt{\lambda_i} \tilde{e}_i) \otimes (\sqrt{\lambda_i} \tilde{e}_i).$$

In particular, \mathcal{E}_ρ and C_ρ share the same eigenvalues. Additionally, the eigenfunctions $\sqrt{\lambda_i} \tilde{e}_i$ form a complete orthonormal system in H .

- (v) *As a consequence, we have an isometric isomorphism from $L_2(\rho)$ to H defined by the componentwise map $e_i \mapsto \sqrt{\lambda_i} \tilde{e}_i$ for all $i \in I$.*

A.2 Proofs

Here we provide the proofs which were omitted in the main text due to the page limitation.

Proof of Lemma 3.2. Let $p := \frac{d\mathbb{P}}{d\rho} \in L_2(\rho)$. Since $(e_i)_{i \in I}$ is a complete orthonormal system in $L_2(\rho)$, it admits the series expansion $p = \sum_{i \in I} \langle p, e_i \rangle_{L_2(\rho)} e_i$. We know that for each $e_i \in L_2(\rho)$, there exists a corresponding $\tilde{e}_i \in H$ such that $e_i = \tilde{e}_i$ holds ρ -almost everywhere and $(\sqrt{\lambda_i} \tilde{e}_i)_{i \in I}$ is a complete orthonormal system in H . We now construct $\tilde{p} := \sum_{i \in I} \langle p, e_i \rangle_{L_2(\rho)} \tilde{e}_i = \sum_{i \in I} \left(\langle p, e_i \rangle_{L_2(\rho)} \lambda_i^{-1/2} \right) \lambda_i^{1/2} \tilde{e}_i \in H$. Note that the $\lambda_i^{1/2} \tilde{e}_i$ form a complete orthonormal system in H . Parseval's identity yields

$$\|\tilde{p}\|_H^2 = \sum_{i \in I} \left(\langle p, e_i \rangle_{L_2(\rho)} \lambda_i^{-1/2} \right)^2 < \infty$$

by the initial assumption and therefore $\tilde{p} \in H$. Note that by construction, $p = \tilde{p}$ holds ρ -almost everywhere. ■

Proof of Proposition 3.4. Since we assume $\sup_x \sqrt{k(x, x)} = \sup_x \|\phi(x)\|_H = c < \infty$, we can apply a Hilbert space version of Hoeffding's inequality [42, 43] to obtain the following concentration bounds [46, 47]. For every $\delta, \gamma > 0$, we have

$$\Pr[\|\mu_{\mathbb{P}} - \hat{\mu}_{\mathbb{P}}\|_H \leq \delta] \geq 1 - 2 \exp\left(-\frac{N\delta^2}{8c^2}\right) \quad (7)$$

as well as

$$\Pr \left[\left\| C_\rho - \widehat{C}_\rho \right\|_H \leq \gamma \right] \geq 1 - 2 \exp \left(-\frac{M\gamma^2}{8c^4} \right), \quad (8)$$

where the estimates are based on N and M i.i.d. samples from \mathbb{P} and ρ , respectively. We assume that (7) and (8) hold independently. We remark that every alternative concentration bound for the above estimation errors can be used in the same way below, leading to analogue results.

For every fixed $\alpha > 0$ and corresponding solution to the regularized empirical and analytical problem ($\hat{u} = (\widehat{C}_\rho + \alpha \mathcal{I}_H)^{-1} \hat{\mu}_\mathbb{P}$ and $u_\alpha = (C_\rho + \alpha \mathcal{I}_H)^{-1} \mu_\mathbb{P}$, respectively), we have

$$\begin{aligned} \|\hat{u} - u_\alpha\|_H &= \left\| (\widehat{C}_\rho + \alpha \mathcal{I}_H)^{-1} \hat{\mu}_\mathbb{P} - (C_\rho + \alpha \mathcal{I}_H)^{-1} \mu_\mathbb{P} \right\|_H \\ &\leq \underbrace{\left\| (\widehat{C}_\rho + \alpha \mathcal{I}_H)^{-1} \hat{\mu}_\mathbb{P} - (C_\rho + \alpha \mathcal{I}_H)^{-1} \hat{\mu}_\mathbb{P} \right\|_H}_{(\star)} + \underbrace{\left\| (C_\rho + \alpha \mathcal{I}_H)^{-1} \hat{\mu}_\mathbb{P} - (C_\rho + \alpha \mathcal{I}_H)^{-1} \mu_\mathbb{P} \right\|_H}_{(\star\star)}. \end{aligned}$$

Using the fact that \widehat{C}_ρ and C_ρ are both self-adjoint and positive, we have $\left\| (\widehat{C}_\rho + \alpha \mathcal{I}_H)^{-1} \right\| \leq \frac{1}{\alpha}$ as well as $\left\| (C_\rho + \alpha \mathcal{I}_H)^{-1} \right\| \leq \frac{1}{\alpha}$. Together with the identity $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$ for all bounded linear operators A and B , we get

$$\left\| (\widehat{C}_\rho + \alpha \mathcal{I}_H)^{-1} - (C_\rho + \alpha \mathcal{I}_H)^{-1} \right\| \leq \frac{1}{\alpha^2} \left\| \widehat{C}_\rho - C_\rho \right\|.$$

We use the above inequality to bound the term (\star) as

$$(\star) \leq \frac{1}{\alpha^2} \left\| \widehat{C}_\rho - C_\rho \right\| \|\hat{\mu}_\mathbb{P}\|_H \leq \frac{\gamma}{\alpha^2} (\|\mu_\mathbb{P}\|_H + \delta)$$

and the term $(\star\star)$ as

$$(\star\star) \leq \left\| (C_\rho + \alpha \mathcal{I}_H)^{-1} \right\| \|\mu_\mathbb{P} - \hat{\mu}_\mathbb{P}\|_H \leq \frac{\delta}{\alpha}.$$

Both bounds hold simultaneously with probability of at least

$$\left[1 - 2 \exp \left(-\frac{N\delta^2}{8c^2} \right) \right] \left[1 - 2 \exp \left(-\frac{M\gamma^2}{8c^4} \right) \right]$$

as given by (7) and (8). Note that this implies $\|\hat{u} - u_\alpha\|_H \leq \frac{\gamma}{\alpha^2} (\|\mu_\mathbb{P}\|_H + \delta) + \frac{\delta}{\alpha}$ with the same probability by the inequalities above. We now express the resulting bound in terms of sample sizes M and N . Since the above concentration bounds hold for arbitrary $\delta, \gamma > 0$, we can fix coefficients $0 < a < 1/2$ and $0 < b < 1/2$ and set $\delta := N^{-a}$ and $\gamma := M^{-b}$, resulting in

$$\|\hat{u} - u_\alpha\|_H \leq \frac{M^{-2b}}{\alpha^2} (\|\mu_\mathbb{P}\|_H + N^{-2a}) + \frac{N^{-2a}}{\alpha}.$$

with probability of at least

$$\left[1 - 2 \exp \left(-\frac{N^{1-2a}}{8c^2} \right) \right] \left[1 - 2 \exp \left(-\frac{M^{1-2b}}{8c^4} \right) \right]. \quad \blacksquare$$

A.3 Numerical representation of $\widehat{\mathcal{A}}_{Y|X}$ based on training data

In what follows, we derive a closed form expression for $\widehat{\mathcal{A}}_{Y|X} = (\widehat{C}_z + \alpha' \mathcal{I}_F)^{-1} \widehat{\mathcal{U}}_{Y|X}$ which can be approximated numerically given a fixed input $x' \in \mathcal{X}$.

We adopt the so-called *feature matrix notation* [40, 53] and define $\Phi = [k(x_1, \cdot), \dots, k(x_N, \cdot)]$ and $\Psi = [\ell(y_1, \cdot), \dots, \ell(y_N, \cdot)]$. We express the Gram matrix for X as $K_X = \Phi^\top \Phi$. Then we have the standard estimates $C_{YX} \approx \widehat{C}_{YX} = N^{-1} \Psi \Phi^\top$ and $\widehat{C}_X = N^{-1} \Phi \Phi^\top$. Assume additionally that we have drawn samples from ρ_y and let $\Gamma = [\ell(z_1, \cdot), \dots, \ell(z_M, \cdot)]$ for $(z_i)_{i=1}^M \stackrel{\text{i.i.d.}}{\sim} \rho_y$. Let Z be a ρ_y -distributed random variable. This implies $C_{\rho_y} \approx \widehat{C}_Z = M^{-1} \Gamma \Gamma^\top$.

It is well known that $M^{-1}L_Z = M^{-1}\Gamma^\top\Gamma \in \mathbb{R}^{M \times M}$ and the empirical covariance operator \widehat{C}_Z share the same nonzero eigenvalues and their eigenvectors/eigenfunctions can be related. This fact has been examined a lot in various scenarios, see for example [49, 46]. In particular, we have the relation

$$M^{-1}L_Z = V\Lambda V^\top \Leftrightarrow \widehat{C}_Z = \sum_{i=1}^r \lambda_i (\lambda_i^{-1/2}\Gamma v_i) \otimes (\lambda_i^{-1/2}\Gamma v_i) = (\Gamma V\Lambda^{-1/2})\Lambda(\Gamma V\Lambda^{-1/2})^\top,$$

where $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_r, 0, \dots, 0) \in \mathbb{R}^{M \times M}$ contains the $r \leq M$ nonzero eigenvalues λ_i of $M^{-1}L_Z$ corresponding to unit norm eigenvectors $v_i \in \mathbb{R}^M$ and $\Lambda^{-1/2} = \text{diag}(\lambda_1^{-1/2}, \dots, \lambda_r^{-1/2}, 0, \dots, 0)$.

Hence, the F -normalized eigenfunctions of \widehat{C}_Z are given by $\lambda_i^{-1/2}\Gamma v_i = \lambda_i^{-1/2} \sum_{j=1}^M v_i^{(j)} \ell(z_j, \cdot)$. Note that $F = \text{span } \Gamma \oplus (\text{span } \Gamma)^\perp$. For a closed subspace $U \subseteq F$, let P_U denote the orthogonal projection operator onto U . Based on the eigendecomposition of \widehat{C}_Z , we naturally have

$$\widehat{C}_Z + \alpha' \mathcal{I}_F = (\Gamma V\Lambda^{-1/2})(\Lambda + \alpha' I_M)(\Gamma V\Lambda^{-1/2})^\top + \alpha' P_{(\text{span } \Gamma)^\perp}$$

for any fixed regularization parameter $\alpha' > 0$. As an immediate consequence, we obtain

$$\begin{aligned} (\widehat{C}_Z + \alpha' \mathcal{I}_F)^{-1} &= (\Gamma V\Lambda^{-1/2})(\Lambda + \alpha' I_M)^{-1}(\Gamma V\Lambda^{-1/2})^\top + \alpha'^{-1} P_{(\text{span } \Gamma)^\perp} \\ &= \Gamma V(\Lambda^{-1/2}\Lambda^{-1/2})(\Lambda + \alpha' I_M)^{-1}V^\top \Gamma^\top + \alpha'^{-1} P_{(\text{span } \Gamma)^\perp} \\ &= \Gamma V(\Lambda^{-1/2}\Lambda^{-1/2})V^\top V(\Lambda + \alpha' I_M)^{-1}V^\top \Gamma^\top + \alpha'^{-1} P_{(\text{span } \Gamma)^\perp} \\ &= M^{-2}\Gamma L_Z^\dagger (L_Z + \alpha' I_M)^{-1}\Gamma^\top + \alpha'^{-1} P_{(\text{span } \Gamma)^\perp}, \end{aligned}$$

Where we use that $\Lambda^{-1/2}$ and $(\Lambda + \alpha' I_M)^{-1}$ are diagonal and therefore commute with every $M \times M$ matrix and the fact that $V(\Lambda^{-1/2}\Lambda^{-1/2})V^\top V(\Lambda + \alpha' I_M)^{-1}V^\top = M^{-2}L_Z^\dagger (L_Z + \alpha' I_M)^{-1}$.

For stability reasons, we can additionally replace L_Z^\dagger in the above expression with its regularized inverse and end up with

$$(\widehat{C}_Z + \alpha' \mathcal{I}_F)^{-1} \Big|_{\text{span } \Gamma} = M^{-2}\Gamma (L_Z + \alpha' I_M)^{-2}\Gamma^\top. \quad (9)$$

Here, we make use of the estimate $\widehat{U}_{Y|X} = \Psi(K_X + N\alpha I_N)^{-1}\Phi^\top$ derived in the literature [40] and insert this expression of $\widehat{U}_{Y|X}$ and the above derived expression for $(\widehat{C}_Z + \alpha' \mathcal{I}_F)^{-1} \Big|_{\text{span } \Gamma}$ into $\widehat{A}_{Y|X} = (\widehat{C}_Z + \alpha' \mathcal{I}_F)^{-1}\widehat{U}_{Y|X}$. We discuss a potential bias induced by moving from $(\widehat{C}_Z + \alpha' \mathcal{I}_F)^{-1}$ to its restriction onto $\text{span } \Gamma$ at the end of this subsection.

Inserting both terms yields

$$\widehat{A}_{Y|X} \approx M^{-2}\Gamma (L_Z + \alpha' I_M)^{-2}\Gamma^\top \Psi(K_X + N\alpha I_N)^{-1}\Phi^\top,$$

which for given $x' \in \mathcal{X}$ can be evaluated as $\widehat{A}_{Y|X}k(x', \cdot) = \sum_{i=1}^M \beta_i \ell(z_i, \cdot)$ with the coefficient vector $\beta = M^{-2}(L_Z + \alpha' I_M)^{-2}L_{ZY}(K_X + N\alpha I_N)^{-1}[k(x_1, x'), \dots, k(x_N, x')]^\top \in \mathbb{R}^M$. The latter is the form presented in the main text.

In general, we introduce a bias by replacing $(\widehat{C}_Z + \alpha' \mathcal{I}_F)^{-1}$ with its restriction to $\text{span } \Gamma$ in the analytical version of the estimate $\widehat{A}_{Y|X} = (\widehat{C}_Z + \alpha' \mathcal{I}_F)^{-1}\widehat{U}_{Y|X}$. This is because $\text{range}(\mathcal{U}_{Y|X}) = \text{range}(\widehat{C}_{YX}) = \text{span } \Psi$ is not necessarily contained in $\text{span } \Gamma$, so information can get “lost”. We note that in this general scenario, this cannot be avoided since $(\widehat{C}_Z + \alpha' \mathcal{I}_F)^{-1}$ is always of infinite range when F is infinite dimensional – however, we must approximate $(\widehat{C}_Z + \alpha' \mathcal{I}_F)^{-1}$ on the finite-dimensional subspace $\text{span } \Gamma$ in numerical scenarios. By assuming that the reference samples are covering the domain \mathcal{X} in a sufficient way such that this loss of information becomes arbitrarily small, replacing $(\widehat{C}_Z + \alpha' \mathcal{I}_F)^{-1}$ with its restriction to $\text{span } \Gamma$ also introduces an arbitrarily small error since $(\widehat{C}_Z + \alpha' \mathcal{I}_F)^{-1}$ is bounded. The detailed analysis of this phenomenon will be covered in future work.

A.3.1 Closed form expression for mean and variance

Let $\hat{u} = \sum_{i=1}^M \beta_i \ell(z_i, \cdot)$ be the RKHS approximation of a density and $\ell(z_i, \cdot)$ be not only a psd kernel evaluated in one argument, but also a probability density with variance v_ℓ . Then the mean of \hat{u} this is given by $m_u = \sum_{i=1}^M \beta_i z_i$ and the variance by $v_u = \sum_{i=1}^M \beta_i z_i^2 - m_u^2 + v_\ell$.

A.4 Computational tricks

In this section, we will detail two tricks that can help fitting large datasets or using density reconstruction when the output domain is high-dimensional.

A.4.1 Trick for large datasets using factorization of the joint probability

We fitted the training data of 32 256 input-output pairs for the traffic prediction experiment in under 5 minutes by observing that the dataset only had 1008 distinct inputs and 32 output samples per input. The following general method takes advantage of this, reducing the involved real matrices from size $32\,256^2$ to 1008^2 . Note that the cross-covariance operator can be written as

$$C_{YX} = \int_{\mathcal{X}} \psi(y) \otimes \phi(x) d\mathbb{P}_{XY}(x, y) = \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} \psi(y) d\mathbb{P}_{Y|X=x}(y) \right) \otimes \phi(x) d\mathbb{P}_X(x),$$

which suggests the empirical estimate $C_{YX} \approx N^{-1} \sum_{i=1}^N \left(n_i^{-1} \sum_{j=1}^{n_i} \psi(y_{i,j}) \right) \otimes \phi(x_i)$, where n_i is the number of output samples for input sample x_i and $y_{i,j}$ is the j th such sample. In feature matrix notation (see A.3), this is equivalent to $C_{YX} \approx N^{-1} \Psi \Phi^\top$ for $\Phi = [k(x_1, \cdot), \dots, k(x_N, \cdot)]$ and $\Psi = [n_1^{-1} \sum_{j=1}^{n_1} \ell(y_{1,j}, \cdot), \dots, n_N^{-1} \sum_{j=1}^{n_N} \ell(y_{N,j}, \cdot)]$. For simplicity, consider the conditional mean operator estimate resulting from this. This will be given by $\mathcal{U}_{Y|X} \approx \Psi(G_\Phi + \alpha N I_N)^{-1} \Phi^\top$, where $\Phi^\top \Phi = G_\Phi \in \mathbb{R}^{N \times N}$ is the Gram matrix induced by Φ . Thus we have to compute the inverse of an $N \times N$ real matrix, while in the standard method a $\left(\sum_{i=1}^N n_i \right) \times \left(\sum_{i=1}^N n_i \right)$ matrix has to be inverted, reducing the complexity from $\mathcal{O}(N^3)$ to $\mathcal{O}\left(\left(\sum_{i=1}^N n_i \right)^3 \right)$. When solving the system of equations instead of computing a matrix inverse, we also get computational savings from this trick, even if slightly less so. Also, the trick is applicable if there are multiple inputs per output by using the factorizing $\mathbb{P}_{XY}(x, y) = \mathbb{P}_{X|Y=y}(x) \mathbb{P}_Y(y)$ instead.

A.4.2 Trick for high dimensions using Kronecker structure of Gram matrices

Assume we have a positive definite kernel ℓ over \mathbb{R}^d such that

$$\ell([y_1, y_2, \dots, y_d]^\top, [y'_1, y'_2, \dots, y'_d]^\top) = \prod_{i=1}^d \ell_i(y_i, y'_i)$$

where ℓ_1, \dots, ℓ_d are positive definite kernels, i.e., ℓ factorizes. Choose $M \in \mathbb{N}_+$ such that $\sqrt[d]{M}$ is an integer. Furthermore, let L_i be the Gram matrix computed on $\sqrt[d]{M}$ samples from the uniform covering the support of the data distribution in dimension j . Then $L = L_1 \otimes \dots \otimes L_d$ and by properties of the Kronecker product, we have $L^{-1} = L_1^{-1} \otimes \dots \otimes L_d^{-1}$.

Thus, by inverting d gram matrices of size $\sqrt[d]{M} \times \sqrt[d]{M}$ and computing Kronecker products, we can get the inverse of an $M \times M$ gram matrix. The inversion has computational complexity $\mathcal{O}(dM^{3/d})$,

while the Kronecker products have complexity $\mathcal{O}\left(\left(\sqrt[d]{M} \right)^{2d} \right) = \mathcal{O}(M^2)$. Assuming $d \geq 2$ and

$\sqrt[d]{M} > 2$, the $\mathcal{O}(M^2)$ complexity of the Kronecker products will dominate. This is a significant improvement from the $\mathcal{O}(M^3)$ computational complexity it would take to invert L directly. The d -dimensional points for which L is the Gram matrix uniformly cover a d -dimensional box. Thus, this trick will be useful with a Lebesgue (i.e., uniform) reference measure on this box. Another advantage is that the computation of Kronecker products is vectorized in most linear algebra packages and trivial to parallelize across dimensions, and further computation could be saved by taking advantage of the symmetry of Gram matrices when computing Kronecker products. Similar tricks have been used in the literature on scalable Gaussian Processes, see for example [65, 21, 41, 20].