Kernel Conditional Density Operators

Ingmar Schuster
Zalando Research
Berlin, Germany

Mattes Mollenhauer
Freie Universität Berlin
Berlin, Germany

Stefan Klus
Freie Universität Berlin
Berlin, Germany

Krikamol Muandet
MPI for Intelligent Systems
Tübingen, Germany

Abstract

We introduce a novel conditional density estimation model termed the *conditional density operator* (CDO). It naturally captures multivariate, multimodal output densities and shows performance that is competitive with recent neural conditional density models and Gaussian processes. The proposed model is based on 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 bounds for the estimation error in a standard density reconstruction scenario, independent of problem dimensionality. Interestingly, when a kernel is used that is also a probability density, the CDO allows us to both evaluate and sample the output density efficiently. We demonstrate the versatility and performance of the proposed model on both synthetic and real-world data.

1 Introduction

Conditional density estimation is an essential task in statistics and machine learning (Tsybakov, 2008; Dinh et al., 2017). Popular techniques for estimating conditional densities include a kernel density estimator (KDE, Tsybakov, 2008), Gaussian process (GP, Williams and Rasmussen, 2006), and deep neural networks (Dinh et al., 2017; Papamakarios et al., 2017). While a KDE is simple to use, it is known to suffer from the curse of dimensionality. While the GP is also a flexible model for conditional densities which enjoys a closed-form posterior thanks to the Gaussianity assumption, approximate inference is often required to model complex densities. Lastly, deep neural networks have recently been used to model complex densities. Despite a great representational power, they require large amount of training data and are prone to overfitting.

The conditional mean embedding (CME) has emerged as an alternative kernel-based nonparametric representation for complex conditional distributions (Song et al., 2009, 2013; Muandet et al., 2017). The CME can models complex distributions nonparametrically and can be estimated consistently from a finite sample. It is mathematically elegant and is less prone to the curse of dimensionality, see, e.g., Tolstikhin et al. (2017). However, one of the fundamental drawbacks of the CME is that a reconstruction of the associated conditional density becomes a non-trivial task. To recover densities, a common approach is to approximate them via a pre-image problem (Kanagawa and Fukumizu, 2014; Song et al., 2008), which requires parametric assumptions on the densities. For sampling, kernel herding can be used (Chen et al., 2010), which requires restrictive assumptions to ensure fast convergence.

In this paper, we present a novel kernel-based supervised learning model for estimating conditional densities, the *conditional density operator* (CDO). It has competitive performance with conditional density models based on deep neural networks (Dinh et al., 2017). To derive our model, we first present the problem of reconstructing a probability density from its associated kernel mean embedding (Muandet et al., 2017; Smola et al., 2007) and connect it to the estimation of Radon–Nikodym derivatives. While this very general problem has been tackled before in similar scenarios (Fukumizu et al., 2013a; Que and Belkin, 2013), we provide a characterization of conditions under which the density reconstruction as an inverse problem has a unique analytical solution. We show that in practical applications, this statistical inverse problem can be solved conveniently using Tikhonov regularization (Tikhonov and Arsenin, 1977; Tikhonov et al., 1995). Furthermore, we give finite sample concentration bounds for the stochastic reconstruction error of the Tikhonov solution. When applied to conditional density estimation, our approach yields solutions that can capture multivariate, multi-

Proceedings of the 23[rd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

modal and non-Gaussian conditional densities and is not constrained by a homoscedastic noise assumption. This compares favorably with standard GPs and is on par with neural conditional density models Williams and Rasmussen (2006); Dinh et al. (2017). 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 our contributions, we *(i)* derive conditions under which a density can be reconstructed in the RKHS, *(ii)* give a consistent estimator for the reconstructed density in the form of a statistical inverse problem, *(iii)* provide dimensionality-independent finite sample error bounds for the estimation error of reconstructed densities, *(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 literature. Our main theoretical results are presented in Sections 3 and 4, Section 5 discusses related work. Experiments on a toy dataset, rough terrain estimation and traffic prediction are reported in Section 6, while concluding remarks are presented in Section 7.

2 Preliminaries

Reproducing kernel Hilbert space (RKHS). We only state the important facts here and collect related results in the supplementary material (see also Steinwart and Christmann, 2008, Section 4.5). We consider a measurable space $(\mathbb{X}, \Sigma)$, where $\mathbb{X}$ is a topological space endowed with the Borel σ-algebra Σ. Let $k\colon \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ be a symmetric positive semidefinite kernel which induces an RKHS $H = \overline{\operatorname{span}\{k(x, \cdot) \mid x \in \mathbb{X}\}}$, 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*.

Assumption 1 (Separability). *The RKHS H is separable. Note that for a Polish space $\mathbb{X}$, the RKHSs induced by a continuous kernel $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ is also separable (see Steinwart and Christmann, 2008, Lemma 4.33). For a more general treatment of conditions implying separability, see Owhadi and Scovel (2017).*

The *reproducing property* $f(x) = \langle f, \phi(x) \rangle_H$ holds for all $f \in H$ and $x \in \mathbb{X}$. We fix a finite measure ρ on $\mathbb{X}$ such that $\int_{\mathbb{X}} \|\phi(x)\|_H^2 \, \mathrm{d}\rho(x) < \infty$. Then the *kernel mean embedding* $\mu_\rho := \int_{\mathbb{X}} \phi(x) \mathrm{d}\rho(x) \in H$ of the measure ρ exists (Sriperumbudur et al., 2011) and the (uncentered) *kernel covariance operator*[1] $C_\rho := \int_{\mathbb{X}} \phi(x) \otimes \phi(x) \mathrm{d}\rho(x)$ is well-defined as a positive self-adjoint Hilbert–Schmidt operator on H (Baker, 1973; Fukumizu et al., 2004; Muandet et al., 2017). Here, the map $\phi(x) \otimes \phi(x)\colon H \to H$, 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*.

Whenever ρ is a probability measure, the kernel mean embedding admits the standard estimate $\hat{\mu}_\rho := M^{-1} \sum_{i=1}^M \phi(x_i)$ for i.i.d. samples $(x_i)_{i=1}^M \sim \rho$. For the covariance operator, we obtain the empirical estimate $\widehat{C}_\rho := M^{-1} \sum_{i=1}^M \phi(x_i) \otimes \phi(x_i)$ with samples as given above. Both $\hat{\mu}_\rho$ and $\widehat{C}_\rho$ converge with $\mathcal{O}(M^{-1/2})$ in probability in RKHS and Hilbert–Schmidt norm respectively (Muandet et al., 2017).

Inverse problems. The general theory of inverse problems, pseudoinverse operators and regularization has been well studied in the context of statistical learning over the last years (Rosasco et al., 2005; De Vito et al., 2005; Caponnetto and De Vito, 2007; Smale and Zhou, 2007; De Vito et al., 2006), we will therefore introduce these concepts only briefly. In general, the compact operator C_ρ can not be inverted on the whole space H. However, it admits a *pseudoinverse* $C_\rho^\dagger$, which is a (generally unbounded) operator with domain $\operatorname{range}(C_\rho) + \operatorname{range}(C_\rho)^\perp \subseteq H$. Note that $\operatorname{range}(C_\rho) + \operatorname{range}(C_\rho)^\perp = H$ if and only if $\operatorname{range}(C_\rho)$ is a closed subspace, which is equivalent to H being finite dimensional. The minimum norm solution to the inverse problem $C_\rho u = f$ with known right-hand side $f \in \operatorname{dom}(C_\rho^\dagger)$ is given by $u^\dagger := A^\dagger f$ 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 against perturbed right-hand sides $\tilde{f}$ and ensure that the solution is still well-defined even if $\tilde{f} \notin \operatorname{dom}(C_\rho^\dagger)$. Note that as $\alpha \to 0$ we have $\|u^\dagger - u_\alpha\|_H \to 0$. Convergence rates for Tikhonov regularization schemes have been derived in numerous settings depending on the problem and are usually connected to rate of decay of the eigenvalues of C_ρ. We refer the reader to the standard literature on inverse problems and regularization (Tikhonov and Arsenin, 1977; Engl and Groetsch, 1996; Tikhonov et al., 1995; Engl et al., 1996) for details.

Conditional mean embedding. The kernel mean embedding μ_ρ has been used extensively as a representation of the measure ρ (Muandet et al., 2017). We now extend this idea to conditional distributions (Song et al., 2009; Grünewälder et al., 2012; Song et al., 2013; Muandet et al., 2017). Note that (Song et al., 2009) formulates results in terms of (generally not existing) inverse operators under adequate regularity assumptions. We use pseudoinverses instead of inverses, which

[1] Note 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.

aligns with the classical theory of inverse problems. Assume we have a topological output space $\mathbb{Y}$ endowed with the Borel σ-algebra and a positive semidefinite kernel $\ell\colon \mathbb{Y}\times\mathbb{Y}\to\mathbb{R}$ inducing a separable RKHS F with feature map $\psi(y) := \ell(y,\cdot)$. All other assumptions we make for the space $\mathbb{X}$, its RKHS, and measures on $\mathbb{X}$ apply likewise for the output space $\mathbb{Y}$ and associated objects. We assume that 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 $C_{YX} := \int_{\mathbb{X}} \psi(y)\otimes\phi(x)\,\mathrm{d}\mathbb{P}_{XY}(x,y)$ 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 $\mathcal{U}_{Y|X} = C_{YX}C_X^{\dagger}\colon H\to F$ and satisfies the equation $\mu_{\mathbb{P}_y} = \mathcal{U}_{Y|X}\mu_{\mathbb{P}}$ for some distribution $\mathbb{P}$ on $\mathbb{X}$, where $\mathbb{P}_y(\cdot) = \int_{\mathbb{X}} \mathbb{P}_{Y|X=x}(\cdot)\,\mathrm{d}\mathbb{P}(x)$ (Song et al., 2009, 2013). 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 $\mu_{\mathbb{P}}\in\operatorname{dom}\mathcal{U}_{Y|X}$ under the condition that $\mathbb{E}[g(Y)\mid X=\cdot]\in H$ for all $g\in F$. This requirement is examined in (Fukumizu et al., 2004, 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 from kernel embeddings

Our strategy to develop the CDO now is as follows. In this section, we will derive a method to reconstruct densities from their mean embeddings. This methodology we can then apply to *conditional* mean embeddings and obtain the CDO as the aggregate of density reconstruction and conditonal mean operator.

Assume we are given the mean embedding $\mu_{\mathbb{P}}$ of a target probability distribution $\mathbb{P}$. We now show how to reconstruct a Radon–Nikodym derivative $\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\rho}$ with respect to a chosen finite positive *reference measure* ρ on $\mathbb{X}$ that satisfies $\mathbb{P}\ll\rho$ and Assumptions 2 and 3.

Assumption 2 (RKHS representative)**.** *We assume that the $L_1(\rho)$ equivalence class of the Radon–Nikodym derivative $\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\rho}$ admits a representative which is an element of H. For simplicity, we will write $\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\rho}\in H$ for this representative.*

We note that Assumption 2 is not always satisfied in practice and is essentially a model assumption. However, the approximative qualities of RKHSs in terms of their "size" with respect to other function spaces such as $C(\mathbb{X})$ or $L_p(\rho)$, $p\in[1,\infty)$ are well examined – for a lot of kernels it can be shown that H is dense in these spaces Micchelli et al. (2006); Steinwart and Christmann (2008); Sriperumbudur et al. (2008).

Assumption 3 (Injective covariance operator)**.** *The kernel covariance operator C_ρ exists (i.e. $\int_{\mathbb{X}}\|\phi(x)\|_H^2\,\mathrm{d}\rho(x)<\infty$) and is injective. Note that for example when k is continuous on $\mathbb{X}\times\mathbb{X}$ and ρ has full support on $\mathbb{X}$, the covariance operator is always injective (Fukumizu et al., 2013b).*

The theoretical background used in the derivation of the following results has appeared in a similar form in Fukumizu et al. (2013b). We apply it in the context of density reconstruction and provide a formal mathematical setting in terms of a statistical inverse problem which can elegantly be used in practice. The following result characterizes the Radon–Nikodym derivative $\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\rho}\in H$ as the solution of an inverse problem.

Proposition 3.1 (Radon–Nikodym derivatives)**.** *Let Assumption 2 and 3 be satisfied. Then the inverse problem*

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

has the unique solution $u^{\dagger} := C_\rho^{\dagger}\mu_{\mathbb{P}} = \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\rho}\in H$.

Proof. We have $C_\rho\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\rho} = \int\limits_{\mathbb{X}}\phi(x)\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\rho}(x)\,\mathrm{d}\rho(x) = \int\limits_{\mathbb{X}}\phi(x)\,\mathrm{d}\mathbb{P}(x) = \mu_{\mathbb{P}}$. Uniqueness of the solution follows directly since C_ρ is injective. ■

Densities in the classical sense are Radon–Nikodym derivatives with respect to Lebesgue measure. This immediately gives the following special case.

Corollary 3.2 (Density reconstruction)**.** *Let $\mathbb{X}\subseteq\mathbb{R}^d$ be compact and the kernel k continuous. Let ρ be Lebesgue measure on $\mathbb{X}$ and $\mu_{\mathbb{P}}$ be the kernel mean embedding of a probability distribution $\mathbb{P}$ on $\mathbb{X}$. If $\mathbb{P}$ admits a density and Assumption 2 is satisfied, then the density is given by $C_\rho^{\dagger}\mu_{\mathbb{P}}$.*

Whenever we are given an analytical mean embedding $\mu_{\mathbb{P}}$ in the setting of Corollary 3.2, we can compute the unique solution $C_\rho^{\dagger}\mu_{\mathbb{P}}$ and reconstruct the density of $\mathbb{P}$. In practice, we are usually given $\mu_{\mathbb{P}}$ in terms of an empirical estimate $\hat{\mu}_{\mathbb{P}}$, for example as an output of a mean embedding-based statistical model. We will now address the consistency and statistical details for the typical case that $\mu_{\mathbb{P}}$ is given in terms of its standard estimate $\hat{\mu}_{\mathbb{P}}$ and we can sample from the reference measure ρ. We emphasize that (1) can in theory be solved with any kind of numerical scheme for integral equations in the classical setting of inverse problems.

[2]The 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 (Fukumizu et al., 2013b).

3.1 Consistency and convergence rate of the Tikhonov-regularized solution

In practice, we cannot access C_ρ and $\mu_{\mathbb{P}}$ analytically. The idea is now to estimate C_ρ from an i.i.d. random sample. For the general case, this can be achieved by importance sampling. For the special case that ρ is the Lebesgue measure and $\rho(\mathbb{X}) = 1$, we can simply sample from it under the assumption that $\mathbb{X}$ is of convenient shape. We can then use the standard estimate for $\widehat{C}_\rho$ (see Section 2). Additionally, we will assume for now that $\mu_{\mathbb{P}}$ is also given in terms of its standard estimate $\hat{\mu}_{\mathbb{P}}$. Note that $\hat{\mu}_{\mathbb{P}}$ might instead be an estimate of a conditional mean embedding (as we will later consider) or an output of other model such as kernel Bayes rule (Fukumizu et al., 2013a). 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 Tikohonov-regularized solution

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

for a regularization parameter $\alpha > 0$. We examine this problem under the assumption that $\widehat{C}_\rho$ is a standard empirical estimate based on M i.i.d. ρ-samples and $\hat{\mu}_{\mathbb{P}}$ is estimated from N i.i.d. $\mathbb{P}$-samples. Next, we show that the reconstruction error $\left\|u^\dagger - \hat{u}\right\|_H$ vanishes in probability as $M, N \to \infty$ for an appropriately chosen positive regularization scheme $\alpha \to 0$ depending on sample sizes. We define the regularized solution $u_\alpha := (C_\rho + \alpha \mathcal{I}_H)^{-1} \mu_{\mathbb{P}}$ and decompose the total error:

$$\left\|u^\dagger - \hat{u}\right\|_H \leq \left\|u^\dagger - u_\alpha\right\|_H + \left\|u_\alpha - \hat{u}\right\|_H. \tag{3}$$

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.

The next result is based on a Hilbert space version of Hoeffding's inequality (Pinelis, 1992, 1994) and gives a general concentration bound for the estimation error term $\left\|u_\alpha - \hat{u}\right\|_H$.

Proposition 3.3 (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 constants. If $\widehat{C}_\rho = M^{-1} \sum_{i=1}^M \phi(x_i) \otimes \phi(x_i)$ with $(x_i)_{i=1}^M \overset{\text{i.i.d.}}{\sim} \rho$ and $\hat{\mu}_{\mathbb{P}} = N^{-1} \sum_{j=1}^N \phi(x_j')$ with $(x_j')_{j=1}^N \overset{\text{i.i.d.}}{\sim} \mathbb{P}$ and both sets of samples are independent, then we have*

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

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 3.3, we can obtain a regularization scheme which ensures that $\hat{u}$ is a consistent estimate of $u^\dagger$.

Corollary 3.4 (Consistency and regularization choice). *Let $\alpha = \alpha(M, N)$ be a regularization scheme such that $\alpha(M, N) \to 0$ as well as*

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

as $M, N \to \infty$. Then the empirical solution $\hat{u}$ obtained from (2) regularized with the scheme $\alpha(M, N)$ converges in probability to the analytical solution $u^\dagger$. One such choice, given $c' \in (0, 1)$, is $\widetilde{\alpha}(M, N) = \max(M^{-b}, N^{-2a})^{c'}$.

Small values for c' imply larger bias and smaller variance (tighter bounds on the stochastic error), while large values for c' imply smaller bias and larger variance. Note that Proposition 3.3 gives bounds only for the case that $\hat{\mu}_{\mathbb{P}}, \widehat{C}_\rho$ are given in terms of their standard empirical estimates.

4 Conditional density operators

In this section, we use Corollary 3.2 to define the *conditional density operator* (CDO), which directly results in a conditional density for an output variable given an input variable or a distribution over the input variable. This is achieved by combining the density reconstruction method derived in the last section with conditional mean operators.

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, injective, and positive self-adjoint Hilbert–Schmidt operator on F. Moreover, densities on $\mathbb{Y}$ are assumed to be Radon–Nikodym derivatives with respect to ρ_y (we get densities in the usual sense if ρ_y is Lebesgue measure). The following result is a direct consequence of Corollary 3.2.

Theorem 4.1 (Conditional density operator). *Assume $\mathbb{P}_y(\cdot) = \int_{\mathbb{X}} \mathbb{P}_{Y|X=x}(\cdot)\, \mathrm{d}\mathbb{P}(x)$ admits a density $p_y \in F$ with respect to reference measure ρ_y, such that the assumptions of Corollary 3.2 are satisfied. Additionally assume that the conditional mean operator $\mathcal{U}_{Y|X} = C_{YX} C_X^\dagger$ for $\mathbb{P}_{Y|X}$ exists and $\mu_{\mathbb{P}} \in \operatorname{dom}(\mathcal{U}_{Y|X})$. 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

$$p_y = \mathcal{A}_{Y|X}\mu_{\mathbb{P}}.$$

If $\mathbb{P}$ is the Dirac measure on x', this results in the density of Y given $X = x'$

$$p_{Y|X=x'} = \mathcal{A}_{Y|X} k(x', \cdot).$$

We call the operator $\mathcal{A}_{Y|X} = C_{\rho_y}^{\dagger} C_{YX} C_X^{\dagger}$ mapping from H to F the *conditional density operator* (CDO). The CDO has several advantages over GPs, the mainstream kernel method for conditional density estimation (Williams and Rasmussen, 2006). In particular, it allows 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. Alvarez et al., 2012; Boyle and Frean, 2005). Moreover, multiple modes in the output can be captured by the CDO. Though this might be achieved with GP mixtures, the CDO allows for more flexibility as it requires no parametric assumptions on the mixture components. Heteroscedastic noise on the output is accounted for by standard CDOs, but nontrivial to include in GP models. Interestingly, any output kernel that is also a probability density gives rise to CDOs where the output density can be both evaluated and sampled efficiently. Also, CDOs allow uncertain inputs with any distribution, while closed form predictions for GPs are only possible when the input uncertainty is Gaussian. Conditional densities estimated by a CDO are illustrated in Figure 1; see Section 6.1 for a description of the data generating process. Conditional densities estimated by a CDO are illustrated in Figure 1; see Section 6.1 for a description of the data generating process. Model selection and cross validation schemes for inferred conditional densities may for example be derived by using an empirical least squares error similar to the objective function used by Sugiyama et al. (2010) on validation data. We will address this topic and potential connections to operator-valued regression problems (Grünewälder et al., 2012) in future work.

4.1 Consistency of the conditional density operator

We can use the results from Section 3.1 to assess the consistency of the CDO. The CDO is defined pointwise when the assumptions of Theorem 4.1 are satisfied. Analogously to the empirical inverse problem in Section 3.1, we replace the pseudoinverses of both C_X and C_{ρ_y} with their regularized inverses for the empirical version of the CDO. From the (unbounded) analytical version $\mathcal{A}_{Y|X} = C_{\rho_y}^{\dagger} C_{YX} C_X^{\dagger}$, we obtain $\widehat{\mathcal{A}}_{Y|X} = (\widehat{C}_{\rho_y} + \alpha' \mathcal{I}_F)^{-1} \widehat{C}_{YX} (\widehat{C}_X + \alpha \mathcal{I}_H)^{-1}$ which is a globally defined bounded operator.

The proof of Proposition 3.3 in the supplementary material can directly be modified to see that whenever $\left\|\widehat{\mu}_{\mathbb{P}_y} - \mu_{\mathbb{P}_y}\right\|_F \to 0$ for a suitable regularization scheme $\alpha > 0$, we obtain a consistent regularized empirical solution of the CDO when $\alpha' > 0$ is chosen appropriately. We will leave the statistical details to future work but want to emphasize that the proof of Proposition 3.3 can also be used to obtain bounds for the conditional mean embedding by simply performing an additional composition with a cross-covariance operator. See Song et al. (2009); Fukumizu et al. (2013b); Fukumizu (2017) for asymptotic consistency results of the conditional mean embedding and appropriate regularization schemes.

4.2 Numerical representation of the conditional density operator

Assume that we have an i.i.d. sample $(x_i, y_i)_{i=1}^N \sim \mathbb{P}_{XY}$ such that the $\mathbb{P}_{XY}$-induced conditional distribution $\mathbb{P}_{Y|X}$ is the distribution of interest and another i.i.d. sample $(z_i)_{i=1}^M \sim \rho_y$, 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

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

with $\beta = M^{-2}(L_Z + \alpha' I_M)^{-2} L_{ZY} (K_X + N\alpha I_N)^{-1} [k(x_1, x'), \ldots, k(x_N, x')]^\top \in \mathbb{R}^M$, where we use the kernel matrices $K_X = [k(x_i, x_j)]_{ij} \in \mathbb{R}^{N \times N}$, $L_Y = [\ell(y_i, y_j)]_{ij} \in \mathbb{R}^{M \times M}$ as well as $L_{ZY} = [\ell(z_i, y_j)]_{ij} \in \mathbb{R}^{M \times N}$ 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 marginal distribution of Y when integrating out $x' \sim \mathbb{P}$, 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.

5 Related work

Finding the pre-image of a feature vector in the RKHS is a classical problem in kernel methods (Kwok and Tsang, 2003; Bakir et al., 2004). In this work, our goal is to reconstruct a density p from the kernel mean embedding $\mu_{\mathbb{P}}$ of some distribution $\mathbb{P}$. There exist two popular approaches in the literature for recovering information from $\mu_{\mathbb{P}}$, namely distributional pre-image learning (Song et al., 2008; Kanagawa and Fukumizu, 2014) and kernel herding (Chen et al., 2010). Given an empirical kernel mean $\hat{\mu}_{\mathbb{P}}$, the idea of the former is to pick a family of densities $\mathcal{P} = \{p_\theta : \theta \in \Theta\}$ and then find $\theta^* = \arg\min_{\theta\in\Theta} \|\hat{\mu}_{\mathbb{P}} - \mu_{p_\theta}\|_H^2$ (Song et al., 2008). The drawback of this approach is the parametric assumptions on the family of densities $\mathcal{P}$. Moreover, it requires solving a constrained non-convex optimization problem. A related approach is that of using (Ton

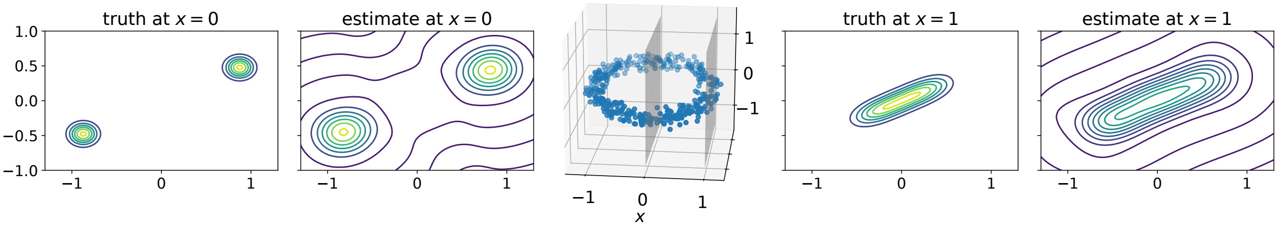

Figure 1: Cross sections of a donut shaped density and estimates using a CDO.

et al., 2019), which suggests to use conditional means as an input for a neural density model. On the other hand, our method provides an analytic solution for conditional densities which only requires that $\mathbb{P}$ is absolutely continuous with respect to the reference measure ρ. Alternatively, kernel herding aims to greedily generate a representative set of T pseudo-samples from $\mathbb{P}$ in a deterministic fashion using the estimate $\hat{\mu}_{\mathbb{P}}$ (Chen et al., 2010). The advantage of herding is an integration error of order $\mathcal{O}(T^{-1})$ under some assumptions. Similarly, our method gives rise to a probability density from which random samples can be easily generated. Note that while our work also relates to the literature of kernel-based density ratio estimation (Kanamori et al., 2012; Que and Belkin, 2013), our goal is not to estimate a density ratio. Furthermore, unlike previous work, we provide a rigorous treatment of the error bounds of our estimates and good choices for regularization constants. Lastly, the kernel mean embedding has recently been applied to fit high-dimensional *implicit* density models such as generative adversarial networks (GANs) (Dziugaite et al., 2015; Li et al., 2015, 2017) and autoencoders (Tolstikhin et al., 2018). It would also be interesting to extend our results to this area of research.

Classical methods for (conditional) density estimation (Bashtannyk and Hyndman, 2001; Hall et al., 2004) are known to suffer from slow convergence in high dimensions, see, e.g., Tsybakov (2008, Chap. 1). Some methods propose estimators that are similar to the CDO, although not making use of RKHS arguments and not proving consistency (Bashtannyk and Hyndman, 2001). An advantage of the CDO is that it is less prone to the curse of dimensionality. Concretely, the convergence rate of Proposition 3.3 and regularizing scheme from Corollary 3.4 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. (2017). Neural density models can also scale gracefully with increasing dimensions, as demonstrated empirically especially in the image generation domain (Kingma and Dhariwal, 2018; Dinh et al., 2017). 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 distribution over the input variable is non-trivial, unlike in the CDO setting.

In the RKHS setting, infinite dimensional exponential families (IDEF) and their conditional extension, kernel conditional exponential families (KCEF) assume the log-likelihood of a (conditional) density to be an RKHS function (Sriperumbudur et al., 2017; Arbel and Gretton, 2018). Fitting such a model is solved by using an optimization approach, while CDOs allow closed form solutions. Furthermore, CDOs allow for trivial normalization of the estimated densities unlike kernel exponential family approaches. Sampling from IDEF and KCEF approximations requires MCMC techniques rather than ordinary Monte Carlo as with our approach. Sampling is necessary to estimate predictive mean and variance in IDEF models, while closed form expressions exist for CDOs, see A.2.1.

6 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 which is described, along with a trick for high-dimensional output spaces, in the supplementary material A.3. For regularization of the CDO in the experiments that follow, we always use Corollary 3.4 with $c' = 0.99999$. Neural density models used as baseline methods where implemented in PyTorch and optimized using ADAM (Kingma and Ba, 2014) with a learning rate tuned to achive the best training log likelihood. Hidden layers contained 100 ReLUs each. The optimal number of hidden layers differed and is stated per experiment.

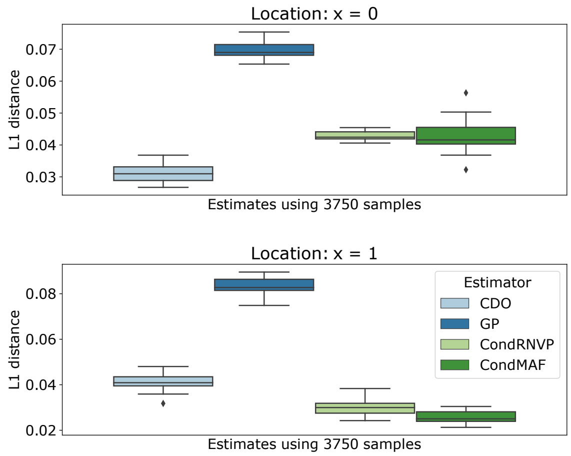

Figure 2: Errors of conditional density estimation for the Gaussian donut in $L_1(\rho_y)$-norm.

6.1 Gaussian donut

For this toy example, a unit circle in the (x, y) plane is embedded into a 3D ambient space and rotated around the y axis. We pick 50 equidistant points on this circle. Each of the points is the mean of an isotropic Gaussian distribution, giving rise to an equal weights mixture that we call a Gaussian donut. We draw 50 samples per mixture component to form the training data for estimating the density on y, z coordinates given x. The uniform on a zero-centered square with side length 4 is used as reference ρ_y. See Figure 1 for the ground truth and CDO estimate at x equal to 0 and 1, respectively. We report numerical density approximation errors in $L_1(\rho_y)$-norm, i.e., $\|\widehat{u} - p_{Y,Z|X=x}\|_{L_1}$ in Figure 2. Input and output kernels where Laplace and Gaussian with lengthscale resulting from the median heuristic (Garreau et al., 2017). The procedure is repeated 10 times for different random seeds. For comparison, independent GPs on each output dimension are used with a Gaussian kernel and lengthscale optimized for highest marginal likelihood with GPyTorch (Gardner et al., 2018). Furthermore, we use conditional versions of RealNVP and masked autoregressive flows (MAF) with 5 hidden layers, see Dinh et al. (2017); Papamakarios et al. (2017). The KCEF method (Arbel and Gretton, 2018) could not be used because it yields unnormalized density estimates and we could not get reliable estimates of the normalizing constant. Figure 2 shows that the CDO provides competitive conditional density estimates on this simple dataset.

6.2 Rough terrain reconstruction

We reproduce an experiment from Eriksson et al. (2018), considering non-uniformly sampled measurements of longitude, latitude, and altitude of Mount St. Helens binned into a 120×117 grid. The task is to estimate the altitude for unobserved coordinates. We randomly chose 80% of the data as training, the rest as test data. We fit an exact GP 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. Furthermore, we fit neural conditional density models based on RealNVP and MAF (Dinh et al., 2017; Papamakarios et al., 2017), where 10 hidden layers gave the best training log likelihood. For the CDO, we pick a Gaussian kernel for input and output domains. The input length scale is chosen using the median heuristic (Garreau et al., 2017). 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 11232 equidistant grid points (matching the dataset size), the output length scale based on the distance between adjacent grid points. See Table 1 for a summary of the SMAEs reached by each method. The result again suggests that our method is competitive with other kernel-based learning algorithms and recent neural density models. We conjecture that added flexibility is a reason for for the better performance compared to a GP. 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.

Table 1: Test Set SMAEs rough terrain

Estimator	SMAE
CDO	0.0269 ± 0.0006
GP	0.0358 ± 0.0006
Cond. Real NVP	0.0373 ± 0.0380
Cond. MAF	0.0309 ± 0.0395

6.3 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 occured

[3]Detailed description in Cuturi (2011), data available at https://archive.ics.uci.edu/ml/datasets/PEMS-SF.

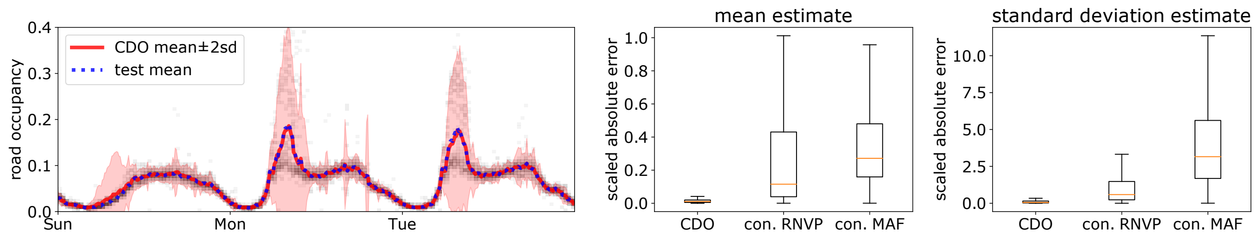

Figure 3: Road occupancy prediction experiment. *Left:* Histogram of test data for three days in black, test data mean and prediction. *Right:* Boxplots of scaled absolute errors with respect to test data.

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 occured 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 CDO 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. Samples for the uniform reference measure on the output domain are taken to be a regular grid between minimum and maximum occuring values. Bandwidth for both kernels is chosen based on the median heuristic. For comparison, we use both RealNVP and MAF deep neural networks (Dinh et al., 2017; Papamakarios et al., 2017), where 5 hidden layers gave the best training log likelihood. 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), and its standard deviation. Mean and variance are chosen because closed form estimates of these exist under the CDO. As this is not the case for the neural models, we draw 2000 samples for estimation. Even though the dataset is rather large, the CDO can be fitted in under one minute on a modern laptop using a scheme outlined in A.3.1. Because we could not adapt this scheme to KCEF (Arbel and Gretton, 2018), it was impossible to fit this alternative kernel conditional density model, because memory requirements could not be satisfied even on a large compute server. Errors are summarized in Table 2 and plotted in Figure 3. Clearly, our CDO outperforms the neural models. While we also fitted GPs using the GPyTorch package, the errors where huge because this problem necessitates heteroscedastic likelihood noise, which is unavailable in currently maintained GP packages.

Table 2: Test Set SMAEs Road Occupancy Data

	mean	sd
CDO	0.02 ± 0.05	0.36 ± 1.21
Cond. Real NVP	0.32 ± 0.41	1.19 ± 1.65
Cond. MAF	0.52 ± 0.81	4.50 ± 4.40

7 Conclusion

In this paper, we show that the reconstruction of densities from kernel mean embeddings can be formulated as an inverse problem under some regularity assumptions. In particular, we draw connections to the estimation of Radon–Nikodym derivatives with respect to a Lebesgue reference measure, for which the solution is shown to be unique. We prove that the popular Tikhonov approach to solving the inverse problem is consistent and allows for finite sample bounds on the estimation error independent of the dimensionality of the data. However, we want to point out that the proposed Tikhonov scheme is only one possible approach for finding a solution. We focus on the conditional density operator as an straight forward application of the density reconstruction result. 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 (Gardner et al., 2018; Wilson and Nickisch, 2015). Theoretical avenues to take might be to find rigorously justified ways of choosing good kernels and kernel parameters.

Acknowledgments

We thank Kashif Rasul provided the conditional normalizing flow implementation and Ilja Klebanov and Tim Sullivan for helpful discussions and pointing out 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

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.

M. Arbel and A. Gretton. Kernel conditional exponential family. In *International Conference on Artificial Intelligence and Statistics*, pages 1337–1346, 2018.

C. Baker. Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society*, 186:273–289, 1973.

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.

D. M. Bashtannyk and R. J. Hyndman. Bandwidth selection for kernel conditional density estimation. *Computational Statistics & Data Analysis*, 36(3):279–298, 2001.

P. Boyle and M. Frean. Dependent Gaussian processes. In *Advances in Neural Information Processing Systems*, pages 217–224, 2005.

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

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.

M. Cuturi. Fast global alignment kernels. In *Proceedings of the 28th International Conference on Machine Learning*, pages 929–936, 2011.

E. De Vito, L. Rosasco, A. Caponetto, U. De Giovannini, and F. Odone. Learning from examples as an inverse problem. *Journal of Machine Learning Research*, 6:883–904, 2005.

E. De Vito, L. Rosasco, and A. Caponnetto. Discretization Error Analysis for Tikhonov Regularization. *Analysis and Applications*, 04(01):81–99, 2006.

L. Dinh, J. Sohl-Dickstein, and S. Bengio. Density estimation using Real NVP. In *International Conference on Learning Representations*, 2017.

G. K. Dziugaite, D. M. Roy, and Z. Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. In *UAI*, 2015.

H. Engl and C. W. Groetsch. *Inverse and Ill-Posed Problems*. Academic Press, 1996.

H. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Kluwer, 1996.

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.

T. Evans and P. Nair. Scalable gaussian processes with grid-structured eigenfunctions (gp-grief). In *International Conference on Machine Learning*, pages 1416–1425, 2018.

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.

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.

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.

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, 2013a.

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

J. Gardner, G. Pleiss, K. Q. Weinberger, D. Bindel, and A. G. Wilson. Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. In *Advances in Neural Information Processing Systems*, pages 7576–7586, 2018.

D. Garreau, W. Jitkrittum, and M. Kanagawa. Large sample analysis of the median heuristic. *arXiv preprint arXiv:1707.07269*, 2017.

S. Grünewälder, G. Lever, L. Baldassarre, S. Patterson, A. Gretton, and M. Pontil. Conditional mean embeddings as regressors. In *International Conference on Machine Learing*, volume 5, 2012.

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.

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.

T. Kanamori, T. Suzuki, and M. Sugiyama. Statistical analysis of kernel-based least-squares density-ratio estimation. *Machine Learning*, 86(3):335–367, 2012.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference for Learning Representations*, 2014.

D. P. Kingma and P. Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, pages 10215–10224, 2018.

J. T. Kwok and I. W.-H. Tsang. The pre-image problem in kernel methods. *IEEE Transactions on Neural Networks*, 15:1517–1525, 2003.

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.

Y. Li, K. Swersky, and R. Zemel. Generative moment matching networks. In *International Conference on Machine Learning*, pages 1718–1727, 2015.

C. A. Micchelli, Y. Xu, and H. Zhang. Universal kernels. *Journal of Machine Learning Research*, 7:2651–2667, 2006.

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.

T. Nickson, T. Gunter, C. Lloyd, M. A. Osborne, and S. Roberts. Blitzkriging: Kronecker-structured stochastic gaussian processes. *arXiv preprint arXiv:1510.07965*, 2015.

H. Owhadi and C. Scovel. Separability of reproducing kernel spaces. *Proc. Amer. Math. Soc.*, 145:2131–2138, 2017.

G. Papamakarios, T. Pavlakou, and I. Murray. Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems*, pages 2338–2347, 2017.

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.

I. Pinelis. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, 22(4):1679–1706, 1994.

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.

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.

L. Rosasco, M. Belkin, and E. D. Vito. On learning with integral operators. *Journal of Machine Learning Research*, 11:905–934, 2010.

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.

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.

S. Smale and D.-X. Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26(2):153–172, 2007.

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.

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.

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.

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.

B. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf. Injective Hilbert space embeddings of probability measures. In *The 21st Annual Conference on Learning Theory*, pages 111–122. Omnipress, 2008.

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.

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.

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, 2008.

M. Sugiyama, I. Takeuchi, T. Suzuki, T. Kanamori, H. Hachiya, and D. Okanohara. Conditional density estimation via least-squares density ratio estimation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 781–788. PMLR, 2010.

A. N. Tikhonov and V. Y. Arsenin. *Solutions of Ill Posed Problems*. W. H. Winston, 1977.

A. N. Tikhonov, V. V. S. A. V. Goncharsky, and A. G. Yagola. *Numerical methods for the solution of ill-posed problems*. Kluwer, 1995.

I. Tolstikhin, B. K. Sriperumbudur, and K. Muandet. Minimax estimation of kernel mean embeddings. *Journal of Machine Learning Research*, 18(1):3002–3048, 2017.

I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schoelkopf. Wasserstein auto-encoders. In *International Conference on Learning Representations*, 2018.

J.-F. Ton, L. Chan, Y. W. Teh, and D. Sejdinovic. Noise Contrastive Meta-Learning for Conditional Density Estimation using Kernel Mean Embeddings. 2019.

A. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008.

C. K. Williams and C. E. Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT Press Cambridge, MA, 2006.

A. Wilson and H. Nickisch. Kernel interpolation for scalable structured gaussian processes (kiss-gp). In *International Conference on Machine Learning*, pages 1775–1784, 2015.