A Supplementary material

A.1 Proofs for results in the main text

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

Proof of Proposition 3.3. 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 (Pinelis, 1992, 1994) to obtain the following concentration bounds (Rosasco et al., 2010; Schneider, 2016). For every $\epsilon_1, \epsilon_2 > 0$, we have

$$\Pr\left[\|\mu_{\mathbb{P}} - \hat{\mu}_{\mathbb{P}}\|_H \leq \epsilon_1\right] \geq 1 - 2\exp\left(-\frac{N\epsilon_1^2}{8c^2}\right) \tag{7}$$

as well as

$$\Pr\left[\left\|C_\rho - \widehat{C}_\rho\right\| \leq \epsilon_2\right] \geq 1 - 2\exp\left(-\frac{M\epsilon_2^2}{8c^4}\right), \tag{8}$$

where the estimates are based on N and M i.i.d. samples from $\mathbb{P}$ and ρ, respectively. Note that the bound (8) is first obtained in Hilbert–Schmidt norm and based on the fact the the operator norm is aways dominated by the Hilbert–Schmidt norm. 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

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

Using the fact that $\widehat{C}_\rho$ 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

$$(\star) \leq \frac{1}{\alpha^2}\left\|\widehat{C}_\rho - C_\rho\right\| \|\hat{\mu}_{\mathbb{P}}\|_H \leq \frac{\epsilon_2}{\alpha^2}\left(\|\mu_{\mathbb{P}}\|_H + \epsilon_1\right)$$

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

$$(\star\star) \leq \left\|(C_\rho + \alpha\mathcal{I}_H)^{-1}\right\| \|\mu_{\mathbb{P}} - \hat{\mu}_{\mathbb{P}}\|_H \leq \frac{\epsilon_1}{\alpha}.$$

Both bounds hold simultaneously with probability of at least

$$\left[1 - 2\exp\left(-\frac{N\epsilon_1^2}{8c^2}\right)\right]\left[1 - 2\exp\left(-\frac{M\epsilon_2^2}{8c^4}\right)\right]$$

as given by (7) and (8). Note that this implies $\|\hat{u} - u_\alpha\|_H \leq \frac{\epsilon_2}{\alpha^2}(\|\mu_{\mathbb{P}}\|_H + \epsilon_1) + \frac{\epsilon_1}{\alpha}$ 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 concentation bounds hold for arbitrary $\epsilon_1, \epsilon_2 > 0$, we can fix coefficients $0 < a < 1/2$ and $0 < b < 1/2$ and set $\epsilon_1 := N^{-a}$ and $\epsilon_2 := M^{-b}$, resulting in

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

with a probability of at least

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

A.2 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 $x' \in \mathbb{X}$.

We adopt the so-called *feature matrix notation* Muandet et al. (2017); Song et al. (2009) and define $\Phi = [k(x_1, \cdot), \ldots, k(x_N, \cdot)]$ and $\Psi = [\ell(y_1, \cdot), \ldots, \ell(y_N, \cdot)]$. 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 $\Gamma = [\ell(z_1, \cdot), \ldots, \ell(z_M, \cdot)]$ for $(z_i)_{i=1}^M \overset{\text{i.i.d.}}{\sim} \rho_y$. 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 Shawe-Taylor et al. (2002); Rosasco et al. (2010). In particular, we have the relation

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

where $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_r, 0, \ldots, 0) \in \mathbb{R}^{M\times M}$ 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 $\Lambda^{-1/2} = \operatorname{diag}(\lambda_1^{-1/2}, \ldots, \lambda_r^{-1/2}, 0, \ldots, 0)$.

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 $F = \operatorname{span}\Gamma \oplus (\operatorname{span}\Gamma)^\perp$. 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

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

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

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

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

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

Here, we make use of the estimate $\widehat{\mathcal{U}}_{Y|X} = \Psi(K_X + N\alpha I_N)^{-1}\Phi^\top$ derived in the literature (Muandet et al., 2017) and insert this expression of $\widehat{\mathcal{U}}_{Y|X}$ and the above derived expression for $(\widehat{C}_Z + \alpha' \mathcal{I}_F)^{-1}\Big|_{\operatorname{span}\Gamma}$ into $\widehat{\mathcal{A}}_{Y|X} = (\widehat{C}_Z + \alpha' \mathcal{I}_F)^{-1} \widehat{\mathcal{U}}_{Y|X}$. We discuss a potential bias induced by moving from $(\widehat{C}_Z + \alpha' \mathcal{I}_F)^{-1}$ to its restriction onto $\operatorname{span}\Gamma$ at the end of this subsection.
Inserting both terms yields

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

which for given $x' \in \mathbb{X}$ can be evaluated as $\widehat{\mathcal{A}}_{Y|X} k(x', \cdot) = \sum_{i=1}^{M} \beta_i \ell(z_i, \cdot)$ with the coefficient vector $\beta = M^{-2}(L_Z + \alpha' I_M)^{-2} L_{ZY}(K_X + N\alpha I_N)^{-1}[k(x_1, x'), \ldots, k(x_N, x')]^\top \in \mathbb{R}^M$. 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 $\operatorname{span} \Gamma$ in the analytical version of the estimate $\widehat{\mathcal{A}}_{Y|X} = (\widehat{C}_Z + \alpha' \mathcal{I}_F)^{-1} \widehat{\mathcal{U}}_{Y|X}$. This is because $\operatorname{range}(\mathcal{U}_{Y|X}) = \operatorname{range}(\widehat{C}_{YX}) = \operatorname{span} \Psi$ is not necessarily contained in $\operatorname{span} \Gamma$, 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 $\operatorname{span} \Gamma$ in numerical scenarios. By assuming that the reference samples are covering the domain $\mathbb{X}$ 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 $\operatorname{span} \Gamma$ 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.2.1 Closed form expression for mean and variance

Let $\widehat{u} = \sum_{i=1}^{M} \beta_i \ell(z_i, \cdot)$ 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 $\widehat{u}$ 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.3 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.3.1 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

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

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 jth such sample. In feature matrix notation (see A.2), this is equivalent to $C_{YX} \approx N^{-1} \Psi \Phi^\top$ for $\Phi = [k(x_1, \cdot), \ldots, k(x_N, \cdot)]$ and $\Psi = [n_1^{-1} \sum_{j=1}^{n_1} \ell(y_{1,j}, \cdot), \ldots, n_N^{-1} \sum_{j=1}^{n_N} \ell(y_{N,j}, \cdot)]$. 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.3.2 Trick for high dimensions using Kronecker structure of Gram matrices

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

$$\ell([y_1, y_2, \ldots, y_d]^\top, [y_1', y_2', \ldots, y_d']^\top) = \prod_{i=1}^{d} \ell_i(y_i, y_i')$$

where $\ell_1, \ldots, \ell_d$ 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 $L = L_1 \otimes \cdots \otimes L_d$ and by properties of the Kronecker product, we have $L^{-1} = L_1^{-1} \otimes \cdots \otimes L_d^{-1}$.
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 Wilson and Nickisch (2015); Flaxman et al. (2015); Nickson et al. (2015); Evans and Nair (2018).

B Related work: Least squares conditional density estimator

We also experimented with the *least squares conditional density estimator* (LSCDE, Sugiyama et al., 2010). This method uses a signed mixture as a least-squares approximation to the conditional density of interest. To get an unsigned density, mixture components with a negative weight are subsequently clipped to weight zero. We suspect that the bad performance of LSCDE estimates results from the clipping operation. While the necessity of getting unsigned estimates is clear, no justification for clipping rather than another method is given in the paper (Sugiyama et al., 2010). A principled way would be to use gradient based optimization rather than the closed form solution to minimize the LSCDE objective, which would enable enforcing the nonnegativity constraint. Another would be to compute an unsigned mixture density that is closest to the closed form solution in a vector space norm. See Table 3 and Figures 4, 5 for experimental results including LSCDE estimates.

Table 3: 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
LSCDE	0.8497 ± 0.0006

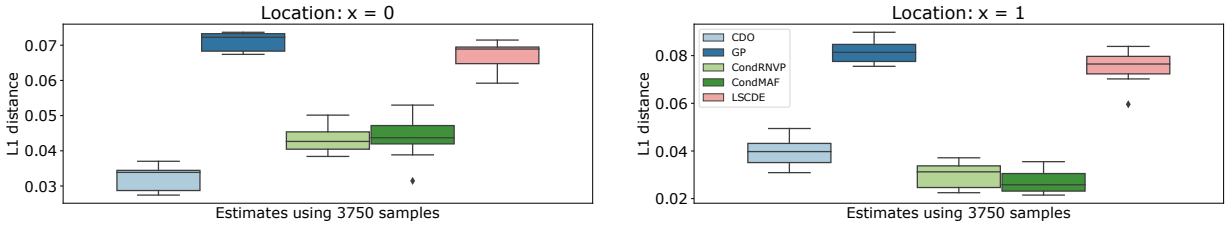

Figure 4: Errors of conditional density estimation for the Gaussian donut in $L_1(\rho_y)$-norm.

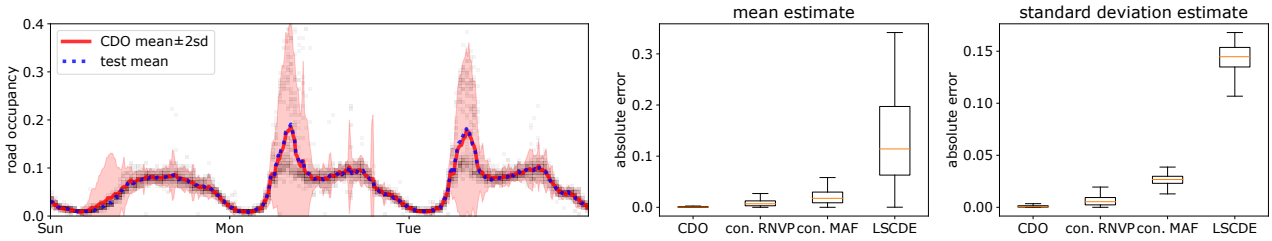

Figure 5: Road occupancy prediction experiment. *Left:* Histogram of test data for three days in black, test data mean and prediction. *Right:* Boxplots of scaled absolute errors with respect to test data.