Bayesian Deconditional Kernel Mean Embeddings

Kelvin Hsu [1 2] **Fabio Ramos** [1 3]

Abstract

Conditional kernel mean embeddings form an attractive nonparametric framework for representing conditional means of functions, describing the observation processes for many complex models. However, the recovery of the original underlying function of interest whose conditional mean was observed is a challenging inference task. We formalize *deconditional kernel mean embeddings* as a solution to this inverse problem, and show that it can be naturally viewed as a nonparametric Bayes' rule. Critically, we introduce the notion of *task transformed Gaussian processes* and establish deconditional kernel means as their posterior predictive mean. This connection provides Bayesian interpretations and uncertainty estimates for deconditional kernel mean embeddings, explains their regularization hyperparameters, and reveals a marginal likelihood for kernel hyperparameter learning. These revelations further enable practical applications such as likelihood-free inference and learning sparse representations for big data.

1. Introduction

Observations of complex phenomena often lead to likelihoods that are described by a conditional mean. A widely applicable setting where this occurs is collecting observations under uncertain inputs, where the task is to learn a function $f : \mathcal{X} \to \mathbb{R}$ to model a real-valued response z as a function of inputs $x \in \mathcal{X}$ without being able to query or measure x directly to observe this phenomenon. Instead, another measured input $y \in \mathcal{Y}$ relates to x through $p(x|y)$. Consequently, given y, the response Z has mean $g(y) := \mathbb{E}[f(X)|Y = y]$, where g is called the conditional mean of f. Furthermore, $p(x|y)$ is often only available as sample pairs $\{x_i, y_i\}_{i=1}^n$, from simulations, algorithms, or separate experiments, making recovery of latent functions f from conditional means g a challenging inference task.

Our first contribution begins with formulating deconditional mean embeddings (DMEs) as solutions to this inference problem by building upon the framework of conditional mean embeddings (CMEs) (Song et al., 2013). We show that the DME can be established as a nonparametric Bayes' rule in the reproducing kernel Hilbert space (RKHS) and used for likelihood-free Bayesian inference. In contrast to kernel Bayes' rule (KBR) (Fukumizu et al., 2013) which uses third order tensors that can result in vanishing priors, DMEs use second order tensors and avoids this problem.

Together with CMEs and KBR, DMEs form a critical part of the kernel mean embedding (KME) (Muandet et al., 2017) framework, where probabilistic rules can be represented nonparametrically as operators that are linear in the RKHS. This greatly simplifies probabilistic inference without requiring parametrized distributions and compromising flexibility.

Despite this connection, there are elements unique to the KME framework that cannot be interpreted or solved via the parallel between probability rules and RKHS mean operations. Similar to empirical forms for KBR and CMEs, empirical DMEs are obtained by replacing expectations in its constituent operators with their empirical means, and introduce regularization for operator inverses to relax RKHS assumptions, instead of as the optimal solution to a particular loss. Setting regularization hyperparameters is difficult in practice without an appropriate loss for the inference task. Furthermore, similar to KBR, the nonparametric Bayes' rule provided by DMEs is a statement between observed (or simulated) variables and not on latent functions or quantities. Consequently, uncertainty estimation in inference of latent functions f still require a separate Bayesian formulation.

Our second contribution establishes a Bayesian view of DMEs as posterior predictive means of the task transformed Gaussian process (TTGP), a novel nonparametric Bayesian model that recover latent relationships between variables without observing them jointly. TTGPs are so named because we show that they are a type of transformed Gaussian process (Murray-Smith & Pearlmutter, 2005) where the transformations and noise covariances are learned, by transforming one Gaussian process (GP) task to another, rather than designed from expert knowledge. We use this connection to derive posterior and predictive uncertainty estimates for DMEs and explain their regularization hyperparameters

[1]University of Sydney [2]CSIRO, Sydney [3]NVIDIA, Seattle. Correspondence to: Kelvin Hsu <Kelvin.Hsu@sydney.edu.au>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

as a function of noise variance. Finally, we derive marginal likelihoods and their scalable computational forms to learn DME hyperparameters, which can also be applied to learn inducing points for sparse representations as a special case. All proofs are in the supplementary material.

2. Kernel Mean Embeddings

We begin with an overview of the KME framework from which DMEs are built upon. KMEs are an arsenal of techniques concerned with representations and transformations of function expectations under highly flexible distributions. They consider functions that lie within RKHSs $\mathcal{H}_k$ and $\mathcal{H}_\ell$, formed by positive definite kernels $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. The RKHSs $\mathcal{H}_k$ and $\mathcal{H}_\ell$ are the closure span of the features $\phi(x) = k(x, \cdot)$ and $\psi(y) = \ell(y, \cdot)$ across $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ respectively, endowed with the inner products $\langle \cdot, \cdot \rangle_k \equiv \langle \cdot, \cdot \rangle_{\mathcal{H}_k}$ and $\langle \cdot, \cdot \rangle_\ell \equiv \langle \cdot, \cdot \rangle_{\mathcal{H}_\ell}$.

The key object is the mean embedding of a distribution $\mu_X := \mathbb{E}[k(X, \cdot)] \in \mathcal{H}_k$. They encode function expectations in the sense that $\mathbb{E}[f(X)] = \langle \mu_X, f \rangle_k$, due to the reproducing property that $\langle k(x, \cdot), f \rangle_k = f(x)$ for all $f \in \mathcal{H}_k$.

Higher ordered mean embeddings are vital components of the framework. Specifically, second order mean embeddings such as $C_{YY} := \mathbb{E}[\ell(Y, \cdot) \otimes \ell(Y, \cdot)] \in \mathcal{H}_\ell \otimes \mathcal{H}_\ell$ and $C_{XY} := \mathbb{E}[k(X, \cdot) \otimes \ell(Y, \cdot)] \in \mathcal{H}_k \otimes \mathcal{H}_\ell$ can be identified as cross-covariance operators $C_{YY} : \mathcal{H}_\ell \to \mathcal{H}_\ell$ and $C_{XY} : \mathcal{H}_\ell \to \mathcal{H}_k$ that serve as building blocks of CMEs and DMEs.

In practical scenarios where only *iid* samples $\{x_i, y_i\}_{i=1}^n$ that are realizations of $(X_i, Y_i) \sim \mathbb{P}_{XY}$ for $i \in \{1, \dots, n\}$ are available, the KME framework becomes attractive for nonparametric inference because core objects only require expectations under distributions. Consequently, they can be estimated via empirical means as $\hat{\mu}_X := \frac{1}{n} \sum_{i=1}^n k(x_i, \cdot)$, $\hat{C}_{YY} := \frac{1}{n} \sum_{i=1}^n \ell(y_i, \cdot) \otimes \ell(y_i, \cdot)$, and $\hat{C}_{XY} := \frac{1}{n} \sum_{i=1}^n k(x_i, \cdot) \otimes \ell(y_i, \cdot)$ (Muandet et al., 2017).

For feature matrices, we stack features by columns $\Phi := \begin{bmatrix} \phi(x_1) & \cdots & \phi(x_n) \end{bmatrix}$ and $\Psi := \begin{bmatrix} \psi(y_1) & \cdots & \psi(y_n) \end{bmatrix}$. We write gram matrices as $K := \Phi^T \Phi$ and $L := \Psi^T \Psi$, where the (i, j)-th element of $A^T B$ is the inner product of the i-th column of A with the j-th column of B. That is, $K_{ij} = \phi(x_i)^T \phi(x_j)$ and $L_{ij} = \psi(y_i)^T \psi(y_j)$. When columns are elements of RKHSs such as when $\phi(x) = k(x, \cdot)$ in Φ and $\psi(y) = \ell(y, \cdot)$ in Ψ, the notation $(\cdot)^T(\cdot)$ is a shorthand for the corresponding RKHS inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ when it is clear from context what $\mathcal{H}$ is. For example, $f^T h$ is shorthand for $\langle f, h \rangle_k$ if $f, h \in \mathcal{H}_k$. Another common usage is $\Phi^T f = \{\phi(x_i)^T f\}_{i=1}^n = \{k(x_i, \cdot)^T f\}_{i=1}^n = \{\langle k(x_i, \cdot), f \rangle_k\}_{i=1}^n = \{f(x_i)\}_{i=1}^n =: \mathbf{f}$. For summing outer products, we write $\hat{C}_{YY} = \frac{1}{n} \Psi \Psi^T$ and $\hat{C}_{XY} = \frac{1}{n} \Phi \Psi^T$. Note that we use non-bold letters for single points x and y, even though they are often multivariate in practice.

3. Conditional Kernel Mean Embeddings

We now present CMEs in a fashion that focuses on their operator properties. By reviewing CMEs this way, parallels and contrast with DMEs in the subsequent section 4 become more apparent. Importantly, instead of defining CMEs via an explicit form, we begin by forming problem statements.

Definition 3.1 (Conditional Mean Problem Statement)**.** Given a function $f : \mathcal{X} \to \mathbb{R}$, infer the function $g : \mathcal{Y} \to \mathbb{R}$ such that $g(y) = \mathbb{E}[f(X)|Y = y] \equiv \mathbb{E}_{X|Y}[f](y)$. We call g the *conditional mean* of f with respect to $\mathbb{P}_{X|Y}$ and write the shorthand $g = \mathbb{E}_{X|Y}[f] = \mathbb{E}[f(X)|Y = \cdot]$.

This naturally leads to the notion of operators that map functions f to their conditional means $g = \mathbb{E}[f(X)|Y = \cdot]$.

Definition 3.2 (Conditional Mean Operators)**.** The conditional mean operator (CMO) $C_{X|Y} : \mathcal{H}_\ell \to \mathcal{H}_k$ corresponding to $\mathbb{P}_{X|Y}$ is the operator that satisfies

$$(C_{X|Y})^T f = \mathbb{E}[f(X)|Y = \cdot], \quad \forall f \in \mathcal{H}_k, \tag{3.1}$$

where $(C_{X|Y})^T : \mathcal{H}_k \to \mathcal{H}_\ell$ denotes the adjoint of $C_{X|Y}$.

Depending on the nature of ℓ, unique solutions exist.

Theorem 3.1 ((Fukumizu et al., 2004))**.** *Assume that $\ell(y, \cdot) \in \text{image}(C_{YY})$ for all $y \in \mathcal{Y}$. The conditional mean operator (CMO) $C_{X|Y}$ is unique and given by*

$$C_{X|Y} = C_{XY} C_{YY}^{-1}. \tag{3.2}$$

The assumption that $\ell(y, \cdot) \in \text{image}(C_{YY})$ for all $y \in \mathcal{Y}$ is commonly relaxed by introducing a regularization hyperparameter $\lambda > 0$ to the inverse, so that the CMO is replaced with $C_{XY}(C_{YY} + \lambda I)^{-1}$ (Song et al., 2013).

Contrary to definition 3.2, it is more common in the literature to define the CMO as the operator $C_{X|Y}$ that satisfies

$$C_{X|Y} \ell(y, \cdot) = \mathbb{E}[k(X, \cdot)|Y = y], \quad \forall y \in \mathcal{Y}, \tag{3.3}$$

while (3.1) is taken as an immediate property of CMOs (Fukumizu et al., 2004). However, due to lemma 3.2, we instead take definition 3.2 as the definition of CMOs, emphasizing CMOs as solutions to the conditional mean problem, and treat (3.3) as an immediate property.

Lemma 3.2. *Statements (3.1) and (3.3) are equivalent.*

The CME of $\mathbb{P}_{X|Y=y}$ is $\mu_{X|Y=y} := C_{X|Y} \ell(y, \cdot)$, equivalent to querying the CMO at a particular input y.

Consequently, $\langle \mu_{X|Y=y}, f \rangle_k = \langle C_{X|Y} \ell(y, \cdot), f \rangle_k = \langle \ell(y, \cdot), (C_{X|Y})^T f \rangle_\ell = \langle \ell(y, \cdot), \mathbb{E}_{X|Y}[f] \rangle_\ell = \mathbb{E}_{X|Y}[f](y)$.

Motivated by theorem 3.1, empirical CMOs and CMEs are defined by estimating their constituents by empirical means.

Definition 3.3 (Empirical Conditional Mean Operator)**.** The empirical CMO is $\hat{C}_{X|Y} := \hat{C}_{XY}(\hat{C}_{YY} + \lambda I)^{-1}$, $\lambda > 0$.

Theorem 3.3 ((Song et al., 2009)). *The nonparametric form for $\hat{C}_{X|Y}$ is*

$$\hat{C}_{X|Y} = \Phi(L + n\lambda I)^{-1}\Psi^T. \tag{3.4}$$

The empirical CME is then $\hat{\mu}_{X|Y=y} := \hat{C}_{X|Y}\ell(y,\cdot)$.

Consequently, with $\boldsymbol{\ell}(y) := \{\ell(y_i, y)\}_{i=1}^n$, an estimate for $\mathbb{E}_{X|Y}[f](y)$ is $\langle f, \hat{\mu}_{X|Y=y}\rangle_k = \langle f, \hat{C}_{X|Y}\ell(y,\cdot)\rangle_k = f^T\Phi(L + n\lambda I)^{-1}\Psi^T\ell(y,\cdot) = \mathbf{f}^T(L + n\lambda I)^{-1}\boldsymbol{\ell}(y)$.

Critically, while empirical CMOs (3.4) are estimated from joint samples from the joint distribution $\mathbb{P}_{XY}$, they only encode the conditional distribution $\mathbb{P}_{X|Y}$. This means that the empirical CMOs will encode the same conditional distribution even if the joint distribution $\mathbb{P}_{XY}$ changes but the conditional distribution $\mathbb{P}_{X|Y}$ stays the same. That is, the empirical CMO built from joint samples of $p(x, y) = p(x|y)p(y)$ and the empirical CMO built from joint samples of $q(x, y) := p(x|y)q(y)$ will encode the same conditional distribution $p(x|y)$ and converge to the same CMO.

4. Deconditional Kernel Mean Embeddings

We now present a novel class of KMEs referred to as deconditional mean embeddings (DMEs). They are natural counterparts to CMEs. The presentation of definitions and theorems in this section is mainly parallel to section 3. We define the *deconditional mean* problem as the task of recovering latent functions from their conditional means.

Definition 4.1 (Deconditional Mean Problem Statement)**.** Given a function $g : \mathcal{Y} \to \mathbb{R}$, infer a function $f : \mathcal{X} \to \mathbb{R}$ such that $g(y) = \mathbb{E}[f(X)|Y = y]$. We call f a *deconditional mean* of g with respect to $\mathbb{P}_{X|Y}$ and write the shorthand $f = \mathbb{E}^{\dagger}_{X|Y}[g]$.

The deconditional mean of a function g infers the function f whose conditional mean would be g with respect to $\mathbb{P}_{X|Y}$. The corresponding operator that encodes this transformation is the deconditional mean operator (DMO).

Definition 4.2 (Deconditional Mean Operators)**.** The deconditional mean operator (DMO) $C'_{X|Y} : \mathcal{H}_k \to \mathcal{H}_\ell$ corresponding to $\mathbb{P}_{X|Y}$ is the operator that satisfies

$$(C'_{X|Y})^T\mathbb{E}[f(X)|Y = \cdot] = f, \quad \forall f \in \mathcal{H}_k. \tag{4.1}$$

Depending on the nature of ℓ and k, unique solutions exist.

Theorem 4.1. *Assume that $\ell(y,\cdot) \in \text{image}(C_{YY})$ for all $y \in \mathcal{Y}$ and $k(x,\cdot) \in \text{image}(C_{X|Y}C_{YY}(C_{X|Y})^T)$ for all $x \in \mathcal{X}$. The deconditional mean operator (DMO) $C'_{X|Y}$ is unique and given by*

$$C'_{X|Y} = (C_{X|Y}C_{YY})^T(C_{X|Y}C_{YY}(C_{X|Y})^T)^{-1}. \tag{4.2}$$

Similar to the case with CMOs (Song et al., 2013), the assumption that $k(x,\cdot) \in \text{image}(C_{X|Y}C_{YY}(C_{X|Y})^T)$ for all $x \in \mathcal{X}$ can be relaxed by introducing a regularization hyperparameter $\epsilon > 0$ to the inverse, so that the DMO is replaced with $(C_{X|Y}C_{YY})^T(C_{X|Y}C_{YY}(C_{X|Y})^T + \epsilon I)^{-1}$.

Since DMOs invert the results of CMOs, they can also be understood as pseudo-inverses of CMOs.

Theorem 4.2. *If the assumptions in theorem 4.1 hold and further $((C_{X|Y})^TC_{X|Y})^{-1}$ exists such that the pseudo-inverse $C^{\dagger}_{X|Y} := ((C_{X|Y})^TC_{X|Y})^{-1}(C_{X|Y})^T$ is defined, then DMOs are pseudo-inverses of CMOs $C'_{X|Y} = C^{\dagger}_{X|Y}$.*

The DME of $\mathbb{P}_{X=x|Y}$ is $\mu'_{X=x|Y} := C'_{X|Y}k(x,\cdot) \in \mathcal{H}_\ell$, equivalent to querying the DMO at a particular input x.

Consequently, $\langle \mu'_{X=x|Y}, g\rangle_\ell = \langle C'_{X|Y}k(x,\cdot), g\rangle_\ell = \langle k(x,\cdot), (C'_{X|Y})^Tg\rangle_k = \langle k(x,\cdot), f\rangle_k = f(x)$.

The form in (4.2) makes it evident that a DMO can be fully specified once $C_{X|Y}$ and C_{YY}, encoding the measures $\mathbb{P}_{X|Y}$ and $\mathbb{P}_Y$ respectively, are known. If densities exist, we write them as $p_{X|Y} \equiv p_{X|Y}(\cdot|\cdot)$ and $p_Y \equiv p_Y(\cdot)$, and drop the subscripts in density evaluations as $p(x|y)$ and $p(y)$ whenever the context is clear. Note that $\mathbb{P}_{X=x|Y}$ corresponds to $p_{X|Y}(x|\cdot)$ which is evaluated at x and now a function of y. This is in contrast with $\mathbb{P}_{X|Y=y}$ corresponding to $p_{X|Y}(\cdot|y)$ evaluated at y and now a function of x.

Consider the case where X and Y play the roles of observed and unobserved (latent) variables respectively. The DMO considers the conditional $p_{X|Y}$ and the marginal p_Y encoded as $C_{X|Y}$ and C_{YY} (theorem 4.1), and inverts the CMO $C_{X|Y}$ (theorem 4.2) with the help of the encoded marginal C_{YY}. This is analogous to the Bayes' rule, where the posterior $p_{Y|X}(\cdot|x) = \frac{p_{X|Y}(x|\cdot)p_Y(\cdot)}{\int_{\mathcal{Y}} p_{X|Y}(x|y)p_Y(y)dy}$ is fully specified by the likelihood $p_{X|Y}$ and prior p_Y. We can then interpret DMEs as querying the rule at the observed quantity x while leaving the rule as a function of y for inference. Consequently, we also refer to $C_{X|Y}$ and C_{YY} as the likelihood operator and the prior operator respectively.

The difference between the DMO (4.2) and CMO (3.2) equations is akin to writing $p_{Y|X}(\cdot|x)$ using Bayes' rule against using the conditional density rule. Compare the DMO decomposition (4.2) with the CMO decomposition $C_{Y|X} = C_{YX}C^{-1}_{XX} = (C_{XY})^TC^{-1}_{XX}$ in the other direction by reversing the roles of X and Y in (3.2), which would correspond to the posterior $\mathbb{P}_{Y|X}$. The CMO is composed of a *joint* operator C_{XY} and an *evidence* operator C_{XX} corresponding to the joint $\mathbb{P}_{XY}$ and evidence $\mathbb{P}_X$ distributions. Similarly, the DMO is also composed of a *joint* operator $C_{XY} = C_{X|Y}C_{YY} : \mathcal{H}_\ell \to \mathcal{H}_k$ and an *evidence* operator $C'_{XX} := C_{X|Y}C_{YY}(C_{X|Y})^T : \mathcal{H}_k \to \mathcal{H}_k$, but both specified from the likelihood and prior operators.

Motivated by this, we propose to estimate the likelihood and prior operators using separate and independently drawn

samples. The likelihood operator $C_{X|Y}$ is estimated as $\hat{C}_{X|Y}$ (definition 3.3) using *iid* samples $\{x_i, y_i\}_{i=1}^n$, also denoted as $\mathbf{x} := \{x_i\}_{i=1}^n$ and $\mathbf{y} := \{y_i\}_{i=1}^n$. Note that as the likelihood operator is a CMO, these joint samples can be from any joint distribution $\mathbb{Q}_{XY} \neq \mathbb{P}_{XY}$ as long as its conditional distribution is also $\mathbb{P}_{X|Y}$. The prior operator C_{YY} is estimated as $\tilde{C}_{YY} := \frac{1}{m}\sum_{j=1}^m \ell(\tilde{y}_j, \cdot) \otimes \ell(\tilde{y}_j, \cdot)$ using another set of *iid* samples $\tilde{\mathbf{y}} := \{\tilde{y}_j\}_{j=1}^m$ from $\mathbb{P}_Y$.

Definition 4.3 (Empirical Deconditional Mean Operator)**.** Let $\epsilon > 0$ be a regularization hyperparameter and define $\hat{C}_{X|Y}$ and $\tilde{C}_{YY}$ as above. The empirical DMO is

$$\bar{C}'_{X|Y} := (\hat{C}_{X|Y}\tilde{C}_{YY})^T(\hat{C}_{X|Y}\tilde{C}_{YY}(\hat{C}_{X|Y})^T + \epsilon I)^{-1}. \tag{4.3}$$

The accents notate the set of samples used for estimation. When both sets are used such as in the estimation of the DMO $C'_{X|Y}$, we denote it with a bar such as $\bar{C}'_{X|Y}$.

Theorem 4.3. *The nonparametric form for* $\bar{C}'_{X|Y}$ *is*

$$\bar{C}'_{X|Y} = \tilde{\Psi}\left[A^T K A + m\epsilon I\right]^{-1} A^T \Phi^T, \tag{4.4}$$

where $A := (L + n\lambda I)^{-1}\tilde{L}$, $\tilde{L} := \Psi^T\tilde{\Psi}$, *and* $\tilde{\Psi} := \begin{bmatrix}\psi(\tilde{y}_1) & \cdots & \psi(\tilde{y}_m)\end{bmatrix}$.

The empirical DME is then $\bar{\mu}'_{X=x|Y} := \bar{C}'_{X|Y}k(x, \cdot)$.

Consequently, with $\boldsymbol{k}(x) := \{k(x_i, x)\}_{i=1}^n$ and $\tilde{\mathbf{g}} := \{g(\tilde{y}_j)\}_{j=1}^m$, an estimate for $\mathbb{E}^\dagger_{X|Y}[g](x)$ is $\langle g, \bar{\mu}'_{X=x|Y}\rangle_\ell = \tilde{\mathbf{g}}^T\left[A^T K A + m\epsilon I\right]^{-1} A^T \boldsymbol{k}(x)$. This motivates the following definitions, where the notation $\tilde{\mathbf{g}}$ is replaced with $\tilde{\mathbf{z}}$, to be interpreted as target observations of g at $\tilde{\mathbf{y}}$.

Definition 4.4 (Nonparametric DME Estimator)**.** The nonparametric DME estimator, also called the kernel DME estimator or the DME estimator in function space view, is $\bar{f}(x) = \bar{\boldsymbol{\alpha}}^T\boldsymbol{k}(x) = \sum_{i=1}^n \bar{\alpha}_i k(x_i, x)$, where $\bar{\boldsymbol{\alpha}} := A\left[A^T K A + m\epsilon I\right]^{-1}\tilde{\mathbf{z}}$ and $A := (L + n\lambda I)^{-1}\tilde{L}$. Equivalently, $\bar{f}(x) = \tilde{\mathbf{z}}^T\left[A^T K A + m\epsilon I\right]^{-1} A^T\boldsymbol{k}(x)$. An alternative form is $\bar{f}(x) = \tilde{\mathbf{z}}^T A^T\left[KAA^T + m\epsilon I\right]^{-1}\boldsymbol{k}(x)$.

When features $\phi(x) \in \mathbb{R}^p$ and $\psi(y) \in \mathbb{R}^q$ are finite dimensional, we define the parametric DME estimator as follows by rewriting definition 4.4 using the Woodbury identity

Definition 4.5 (Parametric DME Estimator)**.** The parametric DME estimator, also called the feature DME estimator or the DME estimator in weight space view, is $\bar{f}(x) = \bar{\mathbf{w}}^T\phi(x)$, where $\bar{\mathbf{w}} = [\Phi A A^T\Phi^T + m\epsilon I]^{-1}\Phi A\tilde{\mathbf{z}}$ and $A := \Psi^T(\Psi\Psi^T + n\lambda I)^{-1}\tilde{\Psi}$. Equivalently, $\bar{f}(x) = \tilde{\mathbf{z}}^T A^T\Phi^T[\Phi A A^T\Phi^T + m\epsilon I]^{-1}\phi(x)$.

In definition 4.4 (resp. 4.5), computational complexity is dominated by inversions for $L + n\lambda I$ and $A^T K A + m\epsilon I$ (resp. $\Psi\Psi^T + n\lambda I$ and $\Phi A A^T\Phi^T + m\epsilon I$) at $O(n^3)$ and $O(m^3)$ (resp. $O(q^3)$ and $O(p^3)$). For the alternative form in definition 4.4, both inversions are $O(n^3)$, allowing for larger m at $O(m)$ without compromising tractability.

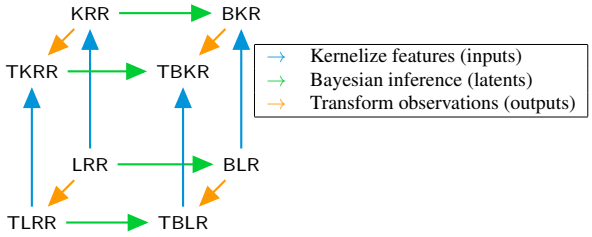

Figure 1. Three dimensions of model extensions (T: Transformed, B: Bayesian, K/L: Kernel/Linear, (R)R: (Ridge) Regression). Kernel extensions (blue) specify feature spaces implicitly through a kernel. Bayesian extensions (green) introduce notions of uncertainty on latent quantities (weights or functions). Finally, transformed extensions (orange) capture indirect function observations.

5. Task Transformed Gaussian Processes

DMEs are constructed as solutions to the task of inferring deconditional means, which are often real-valued functions. Regression problems also address inference of real-valued functions from data. This raises curiosity towards whether DMEs can be formulated as solutions to a regression-like problem, and what insights this connection would provide.

In this section, we formulate the task transformed regression problem to provide regression views of DMEs. To do this, we first briefly review transformed regression in section 5.1 before we present our contributions in section 5.2.

5.1. Transformed Regression

Standard regression models often assume a Gaussian full data likelihood $p(\mathbf{z}|\mathbf{f}) = \mathcal{N}(\mathbf{z}; \mathbf{f}, \sigma^2 I)$ with targets $\mathbf{z} := \{z_i\}_{i=1}^n \in \mathbb{R}^n$. In the generalized setting when observations of $\mathbf{f}$ at $\mathbf{x}$ are not available but observations of linear combinations thereof are, we can use $p(\tilde{\mathbf{z}}|\mathbf{f}) = \mathcal{N}(\tilde{\mathbf{z}}; M^T\mathbf{f}, \Sigma)$ for some transformation $M \in \mathbb{R}^{n\times m}$ and noise covariance Σ, where $\tilde{\mathbf{z}} := \{\tilde{z}_j\}_{j=1}^m \in \mathbb{R}^m$ are the available observations.

We refer to this setting as *transformed regression*. They can be seen as another dimension of modeling with linear ridge regression (LRR) as the base model (fig. 1). Kernel Ridge Regression (KRR) is obtained from LRR via the kernel trick and Woodbury identity, and they are maximum a posteriori (MAP) solutions or predictive means of Gaussian process regression (GPR) and Bayesian linear regression (BLR) respectively (Rasmussen & Williams, 2006). Consequently, we also refer to GPR as Bayesian kernel regression (BKR). Analogous relationships hold between transformed models.

5.2. Task Transformed Regression

We define *task transformed regression (TTR)* as the problem of learning to predict a target variable Z from features X when no direct sample pairs of X and Z are available but instead indirect samples $\{x_i, y_i\}_{i=1}^n$ and $\{\tilde{y}_j, \tilde{z}_j\}_{j=1}^m$ with a

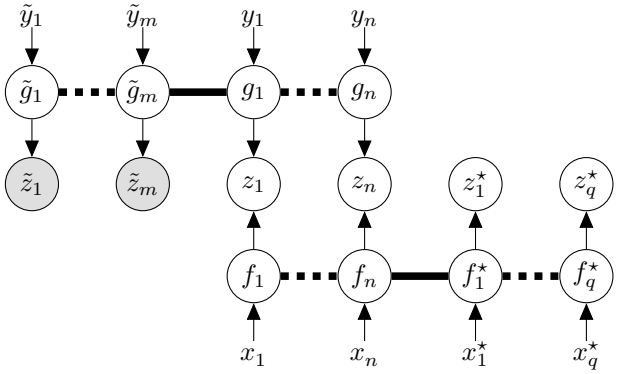

Figure 2. Graphical model (chain graph) for task transformed Gaussian process regression (TTGPR). Circles are random variables. Shaded circles are observed random variables. Undirected edges indicate the GP field, where all the random variables on the field are fully connected to each other (Rasmussen & Williams, 2006). The goal is to infer $\mathbf{f}^\star$ to predict $\mathbf{z}^\star$ at $\mathbf{x}^\star$, using only a *task* or *original* dataset $\{\tilde{y}_j, \tilde{z}_j\}_{j=1}^m$ and a *transformation* dataset $\{x_i, y_i\}_{i=1}^n$. To connect the two GPs, we posit that the unobserved targets $\mathbf{z}$ at $\mathbf{x}$ and at $\mathbf{y}$ would have been the same if they were observed. Note that like regular GPs, to TTGPs the inputs x and y are not modeled as random variables but treated as index variables instead.

mediating variable Y are available. The name illustrates the idea of transforming the task of regressing Z on Y to learn $g : \mathcal{Y} \to \mathbb{R}$, using the *task* or *original* dataset $\{\tilde{y}_j, \tilde{z}_j\}_{j=1}^m$, to the task of regressing Z on X to learn $f : \mathcal{X} \to \mathbb{R}$, by mediating the *task* dataset through the *transformation* dataset $\{x_i, y_i\}_{i=1}^n$. As the mediating variable Y links the two sets together, $\mathbf{y}$ and $\tilde{\mathbf{y}}$ control the task transformation.

DME as solution to chained loss We formulate losses for TTR, and establish DMEs as solutions. We begin with the parametric case with $f(x) = \mathbf{w}^T\phi(x)$ and $g(y) = \mathbf{v}^T\psi(y)$.

Theorem 5.1 (Task Transformed LRR (TTLRR))**.** *The weights of the parametric DME estimator $\bar{f}(x) = \bar{\mathbf{w}}^T\phi(x)$ (definition 4.5) solve chained regularized least square losses,*

$$\hat{\mathbf{v}}[\mathbf{w}] := \operatorname*{argmin}_{\mathbf{v}\in\mathbb{R}^q} \frac{1}{n}\sum_{i=1}^n (\mathbf{w}^T\phi(x_i) - \mathbf{v}^T\psi(y_i)))^2 + \lambda\|\mathbf{v}\|^2,$$
$$\bar{\mathbf{w}} := \operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{m}\sum_{j=1}^m (\tilde{z}_j - \hat{\mathbf{v}}[\mathbf{w}]^T\psi(\tilde{y}_j))^2 + \epsilon\|\mathbf{w}\|^2. \tag{5.1}$$

The notation $\hat{\mathbf{v}}[\mathbf{w}]$ explicitly denotes that $\hat{\mathbf{v}}$ depends on $\mathbf{w}$. Conceptually, in function space view the first optimization finds g so that f at $\mathbf{x}$ best matches with g at $\mathbf{y}$, leading to a solution $\hat{g}[f]$ that is dependent on f. The second finds f so that $\hat{g}[f]$ at $\tilde{\mathbf{y}}$ best matches targets $\tilde{\mathbf{z}}$. Using the kernel trick $k(x, x') = \phi(x)^T\phi(x')$, we obtain the nonparametric case.

Lemma 5.2 (Task Transformed KRR (TTKRR))**.** *The weights of the nonparametric DME estimator $\bar{f}(x) = \bar{\boldsymbol{\alpha}}^T\boldsymbol{k}(x)$ (definition 4.4) satisfies $\bar{\mathbf{w}} = \Phi\bar{\boldsymbol{\alpha}}$ (the kernel trick).*

DME as posterior predictive mean of TTGP We extend TTLRR and TTKRR to the Bayesian case. This connection reveals that TTR models are transformed regression models with transformations and noise covariances that are *learned*.

In the parametric case, we have task transformed BLR (TTBLR). We place separate independent Gaussian priors $p(\mathbf{v}) = \mathcal{N}(\mathbf{v}; \mathbf{0}, \beta^2 I)$ and $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}; \mathbf{0}, \gamma^2 I)$ for g and f respectively. As z is not observed directly from f but only for g, we include noise only for observing g to arrive at likelihoods $p(z|\mathbf{v}) = \mathcal{N}(z; \mathbf{v}^T\psi(y), \sigma^2)$ and $p(z|\mathbf{w}) = \mathcal{N}(z; \mathbf{w}^T\phi(x), 0)$ for g and f respectively.

In the nonparametric case, we have task transformed BKR (TTBKR). We place GP priors $g \sim \mathcal{GP}(0, \ell)$ and $f \sim \mathcal{GP}(0, k)$ on the functions directly. Consequently, TTBKR is also referred to as task transformed Gaussian process regression (TTGPR). Similar to TTBLR, the likelihoods are $p(z|g) = \mathcal{N}(z; g(y), \sigma^2)$ and $p(z|f) = \mathcal{N}(z; f(x), 0)$.

The graphical model for TTGPR is shown in fig. 2. The two GPs for g and f are linked by constraining their targets to be the same at $\mathbf{y}$ and $\mathbf{x}$ respectively. The GP for g is used to infer the predictive distribution $p(\tilde{\mathbf{z}}|\mathbf{z})$, which in turn specifies the overall likelihood $p(\tilde{\mathbf{z}}|\mathbf{f})$ used to infer f. Detailed derivations are provided in the proof of theorem 5.3.

Theorem 5.3 (Task Transformed BLR (TTBLR) and Task Transformed BKR (TTBKR))**.** ***(1)** The TTBLR is a transformed BLR (TBLR) with $M = \Psi^T(\Psi\Psi^T + \frac{\sigma^2}{\beta^2}I)^{-1}\tilde{\Psi}$ and $\Sigma = \sigma^2\tilde{\Psi}^T(\Psi\Psi^T + \frac{\sigma^2}{\beta^2}I)^{-1}\tilde{\Psi} + \sigma^2 I$ as the transformation and noise covariance. **(2)** The TTBKR is a transformed BKR (TBKR) with transformation $M = (L + \sigma^2 I)^{-1}\tilde{L}$ and noise covariance $\Sigma = \tilde{\tilde{L}} + \sigma^2 I - \tilde{L}^T(L + \sigma^2 I)^{-1}\tilde{L}$. **(3)** The TTBLR and TTBKR marginal likelihoods are $p(\tilde{\mathbf{z}}) = \mathcal{N}(\tilde{\mathbf{z}}; \mathbf{0}, [\Sigma^{-1} - \Sigma^{-1}A^T\Phi^T C\Phi A\Sigma^{-1}]^{-1})$ where $C = [\Phi A\Sigma^{-1}A^T\Phi^T + \frac{1}{\gamma^2}I]^{-1}$ and $p(\tilde{\mathbf{z}}) = \mathcal{N}(\tilde{\mathbf{z}}; \mathbf{0}, A^T K A + \Sigma)$ respectively. **(4)** For both models, when the posterior for g is approximated via MAP, the covariance becomes $\Sigma = \sigma^2 I$. In this case, the parametric (resp. nonparametric) DME estimator (definitions 4.4 and 4.5) is the predictive mean of a TTBLR (resp. TTBKR) with $\lambda = \frac{\sigma^2}{n\beta^2}$ and $\epsilon = \frac{\sigma^2}{m\gamma^2}$ (resp. $\lambda = \frac{\sigma^2}{n}$ and $\epsilon = \frac{\sigma^2}{m}$). An alternative TTBKR marginal likelihood is $p(\tilde{\mathbf{z}}) = \mathcal{N}(\tilde{\mathbf{z}}; \mathbf{0}, \sigma^2[I - A^T(KAA^T + \sigma^2 I)^{-1}KA]^{-1})$. **(5)** When both posteriors for g and f are approximated via MAP, TTBLR and TTBKR becomes TTLRR and TTKRR respectively with λ and ϵ from (4).*

Importantly, our end goal is to infer f. While this involves inferring g, g is not of direct interest. A simpler alternative is to only perform Bayesian inference on f and approximate g with its MAP solution, simplifying the noise covariance via (4) of theorem 5.3. This establishes a Bayesian interpretation for DMEs as MAP estimates of TTGPs. Critically, by maximizing the TTGP marginal likelihood, we can learn DME hyperparameters of kernels k and ℓ, and also λ and

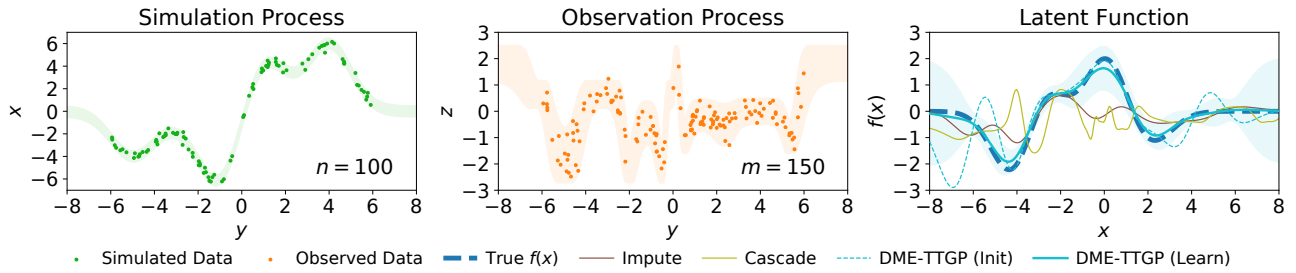

Figure 3. Illustration of latent function recovery with a task transformed Gaussian process (TTGP) on non-trivial simulation and observation processes $p(x|y)$ and $p(z|y)$. (Left) The simulation process $p(x|y)$. (Center) The observation process $p(z|y)$. (Right) The true latent function f, the naive solutions using cascaded regressors and imputed data, and the mean and uncertainty bounds of the TTGP, also the Bayesian DME, with initial and learned hyperparameters. All bounds are 2 standard deviations from the mean.

ϵ. Furthermore, the computational complexity for alternative marginal likelihood is dominated by inversions that are $O(n^3)$ only, again allowing for larger m at $O(m)$.

In summary, we first establish the DME as a nonparametric solution to the TTR problem under chained regularized least squares losses that make learning f dependent on learning g. While λ and ϵ are previously seen as numerical adjustments to relax RKHS assumptions and stabilize matrix inversions in the KME framework, they can now be seen as controlling the amount of function regularization under this loss. Secondly, we present TTGPs as nonparametric Bayesian solutions to this regression problem and show that DMEs are their posterior predictive means. Again, inference of f is dependent on the inference of g, allowing GP uncertainties to propagate through. This connection provides Bayesian interpretations of DMEs and enable uncertainty estimation in inferring deconditional means. Critically, we use this to derive marginal likelihoods for hyperparameter learning.

6. Nonparametric Bayes' Rule

While DMOs were constructed as solutions to the deconditional mean problem, they also resemble Bayes' rule when we focus solely on considering the encoded relationship between X and Y. This was motivated by theorem 4.1, which revealed that the DMO can be fully specified by the CMO $C_{X|Y}$ and the second order mean embedding C_{YY} that encoded the likelihood $\mathbb{P}_{Y|X}$ and prior $\mathbb{P}_Y$ respectively. To establish this view, we investigate the conditions for which the DMO $C'_{X|Y}$ coincide with the CMO $C_{Y|X}$ that encodes the posterior $\mathbb{P}_{Y|X}$, leading to a nonparametric Bayes' rule.

While first class citizens of probability rules are density evaluations, first class citizens of the KME framework are expectations. Consequently, instead of relating density evaluations, rules under the KME framework relate mean embeddings of distributions at various orders. Importantly, while a distribution $Y \sim \mathbb{P}_Y$ has one simple density evaluation $p_Y(y)$, it can have different RKHS representations at different orders such as μ_Y and C_{YY} or higher.

A nonparametric Bayes' rule is a rule which translates Bayes' rule into the RKHS, where distributions are represented as RKHS operators, alleviating limitations from parametric assumptions such as Gaussian posteriors. It computes a posterior operator $C_{Y|X}$ when given only likelihood operators (e.g. $C_{X|Y}$) and prior operators (e.g. C_{YY}). The DMO is appealing as all operators involved are of second order and the same second order likelihood and prior operators are used for both the joint and evidence operator.

However, because C'_{XX} is not necessarily the same as C_{XX}, the DMO $C'_{X|Y} = (C_{XY})^T (C'_{XX})^{-1}$ is not necessarily the posterior operator $C_{Y|X} = (C_{XY})^T (C_{XX})^{-1}$. Nevertheless, under certain conditions they coincide with each other.

Theorem 6.1. *If $C_{X|Y} C_{Y|X} C_{XX} = C_{XX}$, then $C'_{XX} = C_{XX}$ and $C'_{X|Y} = C_{Y|X}$.*

A special instance where the assumptions are met is when $X = r(Y)$ where r is not necessarily invertible. Importantly, for empirical DMOs, having $x_i = x_j$ for any $y_i = y_j$ suffices, which can be achieved if all y_i are unique. Furthermore, empirical DMOs $\bar{C}'_{X|Y}$ can be seen as generalizations of empirical CMOs $\hat{C}_{Y|X}$ in the other direction.

Theorem 6.2. *If $m = n$ and $\tilde{y}_i = y_i$ for all $i \in \{1, \ldots, n\}$, then the empirical DMO corresponding to $\mathbb{P}_{X|Y}$ becomes the empirical CMO corresponding to $\mathbb{P}_{Y|X}$ for $\lambda \to 0^+$,*

$$\lim_{\lambda \to 0^+} \bar{C}'_{X|Y} = \Psi \left[K + n\epsilon I\right]^{-1} \Phi^T = \hat{C}_{Y|X}. \tag{6.1}$$

Intuitively, suppose $\{x_i, y_i\}_{i=1}^n$ are from $p(x, y) := p(x|y)p(y)$ and $\{\tilde{y}_j\}_{j=1}^m = \{y_i\}_{i=1}^n$ is from $p(y)$, then the DMO of $p(x|y)$ is equivalent to the CMO of $p(y|x) = p(x|y)p(y) / \int_{\mathcal{Y}} p(x|y)p(y)dy$ in the other direction, as per theorem 6.2. In general, however, if $\{x_i, y_i\}_{i=1}^n$ are from $q(x, y) := p(x|y)q(y)$ and $\{\tilde{y}_j\}_{j=1}^m$ is from $p(y)$, then using the joint samples from $q(x, y)$ only to build the CMO will yield the CMO of $q(y|x) = p(x|y)q(y) / \int_{\mathcal{Y}} p(x|y)q(y)dy$, while using both the joint samples from $q(x, y)$ and marginal samples from $p(y)$ to build the DMO will yield the CMO corresponding to $p(y|x) = p(x|y)p(y) / \int_{\mathcal{Y}} p(x|y)p(y)dy$. Appendix E details further parallels with probabilistic rules.

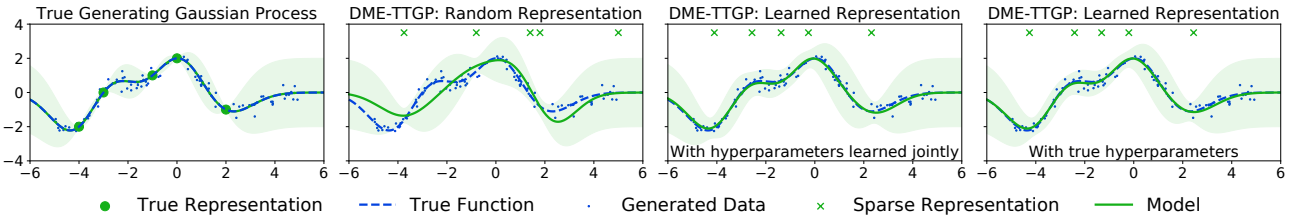

Figure 4. Sparse representation learning on a dataset of 100 points generated by using the toy process from Rasmussen & Williams (2006) as ground truth. (Left) The true function, represented exactly by a GP mean using 5 points. (Left Center) DME using 5 random points. (Right Center) DME with the 5 points and all hyperparameters learned jointly via its marginal likelihood. (Right) DME with the 5 points learned via its marginal likelihood under true hyperparameters. The vertical position of the sparse representation has no meaning.

7. Related Work

DMOs have strong connections to KBR. Both provide a nonparametric Bayes' rule under the KME framework. In contrast to DMOs where both likelihood and prior operators are of second order, both KBR(a) and KBR(b) (Song et al., 2013; Fukumizu et al., 2013) use a third order likelihood operator $C_{XX|Y}$ and a first order prior embedding μ_Y for the evidence operator $C_{XX} = C_{XX|Y}\mu_Y$. KBR(b) further uses a different third order likelihood operator $C_{XY|Y}$ for the joint operator $C_{XY} = C_{XY|Y}\mu_Y$. Consequently, KBR becomes sensitive to inverse regularizations and effects of prior samples $\tilde{\mathbf{y}}$ can vanish. For instance, when $\epsilon \to 0^+$, KBR(b) degenerate to $\bar{C}_{Y|X} = \Psi K^{-1}\Phi^T$, which is a CMO that no longer depend on $\tilde{\mathbf{y}}$. Instead, DMOs degenerate to $\bar{C}'_{X|Y} = \tilde{\Psi}A^T\left[AA^T\right]^{-1}K^{-1}\Phi^T$, retaining their original structure. Detailed comparisons are provided in appendix F.

Viewing KMEs as regressors provides valuable insights and interpretations to the framework. CMOs can be established as regressors where the vector-valued targets are also kernel induced features (Grünewälder et al., 2012). In contrast, we establish DMOs as solutions to task transformed regressors which recover latent functions that, together with a likelihood, governs interactions between three variables.

The TTR problem describes the setting of learning from conditional distributions in the extreme case where only one sample of x_i is available for each y_i to describe $p(x|y)$. Dual KMEs (Dai et al., 2017) formulate this setting as a saddle point problem, and employ stochastic approximations to efficiently optimize over the function space. However, without connections to Bayesian models such as TTGPs that admit a marginal likelihood, hyperparameter selection often require inefficient grid search.

Hyperparameter learning of marginal embeddings have been investigated by placing GP priors on the embedding itself to yield a marginal likelihood objective (Flaxman et al., 2016). However, it is unclear how this can be extended to CMEs. Our marginal likelihoods (theorem 5.3 and theorem G.1) provide such objective for DMEs and, due to theorem 6.2, it can also be applied to CMEs as a special case.

8. Applications and Experiments

While DMEs are developed to complement the theoretical framework of KMEs, in this section we describe and demonstrate some of their practical applications with experiments.

8.1. Hyperparameter Learning for TTR

We first illustrate in fig. 3 the TTR problem, the primary application of TTGPs and DMEs. While X and Y are multivariate in general, we use 1D examples to enable visualizations. Although this is a 1D problem, the simulation process $p(x|y)$ and observation process $p(z|y)$ are governed by non-trivial relationships where successful recovery of f requires dealing with difficult multi-modalities in $p(y|x)$. To generate the data, we choose non-trivial functions r and f and generate $X_i = r(Y_i) + \eta_i$ and $\tilde{Z}_j = f(r(\tilde{Y}_j) + \tilde{\eta}_j) + \tilde{\xi}_j$, where $Y_i, \tilde{Y}_j \sim U(-6, 6)$ and $\eta_i, \tilde{\eta}_j, \tilde{\xi}_j \sim \mathcal{N}(0, 0.25^2)$ for all $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, m\}$. In this way, $p(x|y) = \mathcal{N}(x; r(y), 0.25^2)$, $p(z|x) = \mathcal{N}(z; f(x), 0.25^2)$, and $\mathbb{E}[Z|Y = y] = \mathbb{E}[f(r(y) + \tilde{\eta}) + \tilde{\xi}] = \mathbb{E}[f(X)|Y = y]$.

By optimizing the marginal likelihood in theorem 5.3 (3), we see that the DME is able to adapt from its initial hyperparameters to learn the latent function accurately.

We compare this to two naive solutions that one may propose when faced with a TTR problem. The *cascade* method trains separate regressors from X to Y, with the transformation set, and from Y to Z, with the task set. They use the former to predict $y^\star$ from $x^\star$ and the latter to predict $z^\star$ from $y^\star$. The *impute method* trains a regressor from Y to Z with the task set and predicts $\mathbf{z}_{\text{fake}}$ at locations $\mathbf{y}$, and trains a regressor on the dataset $(\mathbf{x}, \mathbf{z}_{\text{fake}})$ to predict $z^\star$ from a new $x^\star$. We use GPR means (KRR) for all such regressors. Both methods suffer because uncertainty propagation is lost by training regressors separately. The cascade method suffers further because $p(y|x)$ is usually highly multi-modal such as in this example, so unimodal regressors like GPR from X to Y are unsuitable. This also highlights that while DMEs provides unimodal Gaussian uncertainty on function evaluations and thus Z, they capture multi-modality as a nonparametric Bayes' rule between X and Y.

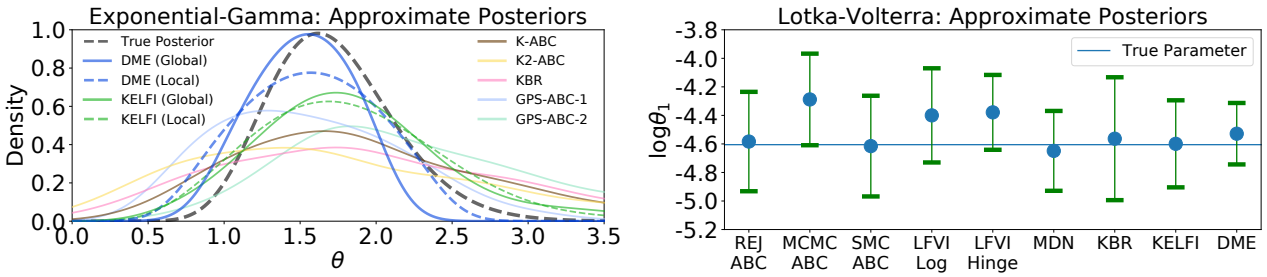

Figure 5. Application to LFI. (Left) Approximate posteriors under kernel based LFI methods for the toy exponential-gamma problem using 100 simulations. 'Global' and 'Local' refer to the optimality of model hyperparameters with respect to their respective approximate marginal likelihoods. (Right) Approximate posteriors of the first (log) parameter. Error bars represent the middle 95% credible interval.

8.2. Sparse Representation Learning with TTGP

A special case of the TTR problem is to learn sparse representations for big data with trainable inducing points. Continuing with the notations used so far, we are given a large original dataset $(\tilde{\mathbf{y}}, \tilde{\mathbf{z}})$ of size m, with inputs Y and target Z. We let the transformation dataset be a set of n inducing points $\mathbf{x} = \mathbf{y} = \mathbf{u}$ for Y where $n << m$. That is, we degenerate to $X = Y$ and $\mathcal{X} = \mathcal{Y}$. We maximize the alternative marginal likelihood in theorem 5.3 (4) with respect to the inducing points and learn the TTGP hyperparameters jointly. For predictive mean we use the alternative computational form in definition 4.4. Similar form exists for the covariance. These alternative forms are suitable for this application because n is small for its $O(n^3)$ inversions and dependence on m is only $O(m)$. We illustrate this process in fig. 4.

8.3. Likelihood-Free Inference with DME

As a nonparametric Bayes' rule, DMEs can be used for likelihood-free inference (LFI) (Marin et al., 2012) where likelihood evaluations are intractable but sampling from a simulator $\mathbf{x} \sim p(\mathbf{x}|\boldsymbol{\theta})$ is possible. The simulator takes parameters $\boldsymbol{\theta}$ and stochastically generates simulated data that are often summarized into statistics $\mathbf{x}$. Observed data are also summarized into statistics $\mathbf{y}$, and discrepancies with $\mathbf{x}$ are often measured by an ϵ-kernel $\kappa_\epsilon(\mathbf{y}, \mathbf{x}) = p_\epsilon(\mathbf{y}|\mathbf{x})$ such that $p_\epsilon(\mathbf{y}|\boldsymbol{\theta}) = \int p_\epsilon(\mathbf{y}|\mathbf{x}) p(\mathbf{x}|\boldsymbol{\theta}) d\boldsymbol{\theta}$. This ϵ is not to be confused with the regularization used for C'_{XX}, which we denote as δ for this section only. After selecting a prior $p(\boldsymbol{\theta})$, the goal is to approximate the posterior $p_\epsilon(\boldsymbol{\theta}|\mathbf{y})$.

Translating notations into the LFI setting, we have $x_i \to \mathbf{x}_i$, $y_i \to \boldsymbol{\theta}_i$, $\tilde{y}_j \to \tilde{\boldsymbol{\theta}}_j$, and $x \to \mathbf{y}$. We first simulate $\mathbf{x}_i \sim p(\mathbf{x}|\boldsymbol{\theta}_i)$ on parameters $\{\boldsymbol{\theta}_i\}_{i=1}^n \sim \pi(\boldsymbol{\theta})$ not necessarily from the prior to get $\{\boldsymbol{\theta}_i, \mathbf{x}_i\}_{i=1}^n$ for the likelihood, and sample $\{\tilde{\boldsymbol{\theta}}_j\}_{j=1}^m \sim p(\boldsymbol{\theta})$ for the prior. We then build the DME $\bar{\mu}_{\boldsymbol{\Theta}|\mathbf{X}=\mathbf{y}}$ and sample it with kernel herding (Chen et al., 2010) for posterior super-samples. This is described in algorithm 8.1. We also provide the approximate marginal likelihood objective $\bar{q}$ to maximize for hyperparameter learning of the DME. Derivations are detailed in appendix G.

Algorithm 8.1 Deconditional Mean Embeddings for LFI

1: **Input:** Data $\mathbf{y}$, simulations $\{\boldsymbol{\theta}_i, \mathbf{x}_i\}_{i=1}^n \sim p(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$, prior samples $\{\tilde{\boldsymbol{\theta}}_j\}_{j=1}^m \sim p(\boldsymbol{\theta})$, query points $\{\boldsymbol{\theta}_r^\star\}_{r=1}^R$, kernels k, κ_ϵ, ℓ, and ℓ', regularization λ and δ
2: $L \leftarrow \{\ell(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)\}_{i,j=1}^{n,n}$, $\tilde{L} \leftarrow \{\ell\boldsymbol{\theta}_i, \tilde{\boldsymbol{\theta}}_j)\}_{i,j=1}^{n,m}$
3: $A \leftarrow (L + n\lambda I)^{-1}\tilde{L}$, $\tilde{L}^\star \leftarrow \{\ell(\tilde{\boldsymbol{\theta}}_j, \boldsymbol{\theta}_r^\star)\}_{j,r=1}^{m,R}$,
4: $K \leftarrow \{k(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^{n,n}$, $\boldsymbol{k}(\mathbf{y}) \leftarrow \{k(\mathbf{x}_i, \mathbf{y})\}_{i=1}^n$
5: DME: $\boldsymbol{\mu} \leftarrow (\tilde{L}^\star)^T A^T \left[KAA^T + m\delta I\right]^{-1} \boldsymbol{k}(\mathbf{y}) \in \mathbb{R}^R$
6: **for** $s \in \{1, \ldots, S\}$ with $\mathbf{a} \leftarrow \mathbf{0} \in \mathbb{R}^R$ initialized **do**
7: $\quad \hat{\boldsymbol{\theta}}_s \leftarrow \boldsymbol{\theta}_{r^\star}^\star$ where $r^\star \leftarrow \operatorname{argmax}_r \mu_r - (a_r/s)$
8: $\quad \mathbf{a} \leftarrow \mathbf{a} + \{\ell'(\boldsymbol{\theta}_r^\star, \hat{\boldsymbol{\theta}}_s)\}_{r=1}^R$
9: **end for**
10: **Output:** Posterior super-samples $\{\hat{\boldsymbol{\theta}}_s\}_{s=1}^S$
11: **Learning:** $\bar{q} \leftarrow \operatorname{mean}(A^T \boldsymbol{\kappa}_\epsilon)$, $\boldsymbol{\kappa}_\epsilon \leftarrow \{\kappa_\epsilon(\mathbf{y}, \mathbf{x}_i)\}_{i=1}^n$

Figure 5 demonstrates algorithm 8.1 on two standard benchmarks. For the toy exponential-gamma problem we compare directly with other kernel approaches. As simulations are usually very expensive, we show the case with very limited simulations $(n = 100)$, leading to most methods producing posteriors wider than the ground truth. Nevertheless, by optimizing $\bar{q}$ in line 11, DMEs can adapt their kernel length scales accordingly. For Lotka-Volterra, the ABC methods used more than 100000 simulations, while MDN used 10000 simulations. To achieve competitive accuracy, kernel approaches such as DMEs, KELFI (Hsu & Ramos, 2019), and KBR used 2000, 2500, and 2500 simulations. Acronyms and experimental details are described in appendix G.

9. Conclusion

The connections of DMEs with CMEs and GPs produce useful insights towards the KME framework, and are important steps towards establishing Bayesian views of KMEs. DMEs provide novel solutions to a class of nonparametric Bayesian regression problems and enable applications such as sparse representation learning and LFI. For future work, relaxing assumptions required for DMOs as a nonparametric Bayes' rule can have fruitful theoretical and practical implications.

References

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Chen, Y., Welling, M., and Smola, A. Super-samples from kernel herding. In *The Twenty-Sixth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-10)*, pp. 109–116. AUAI Press, 2010.

Dai, B., He, N., Pan, Y., Boots, B., and Song, L. Learning from Conditional Distributions via Dual Embeddings. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1458–1467. PMLR, 20–22 Apr 2017.

Flaxman, S., Sejdinovic, D., Cunningham, J. P., and Filippi, S. Bayesian learning of kernel embeddings. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, pp. 182–191. AUAI Press, 2016.

Fukumizu, K., Bach, F. R., and Jordan, M. I. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 5 (Jan):73–99, 2004.

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

Grünewälder, S., Lever, G., Baldassarre, L., Patterson, S., Gretton, A., and Pontil, M. Conditional mean embeddings as regressors. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*, volume 2, pp. 1823–1830, 2012.

Hastings, W. K. Monte carlo sampling methods using markov chains and their applications. 1970.

Higham, N. J. *Accuracy and stability of numerical algorithms*. SIAM, 2002.

Hsu, K. and Ramos, F. Bayesian Learning of Conditional Kernel Mean Embeddings for Automatic Likelihood-Free Inference. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 2631–2640. PMLR, 16–18 Apr 2019.

Marin, J.-M., Pudlo, P., Robert, C. P., and Ryder, R. J. Approximate Bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180, 2012.

Meeds, E. and Welling, M. GPS-ABC: Gaussian process surrogate approximate Bayesian computation. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, pp. 593–602. AUAI Press, 2014.

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

Murray-Smith, R. and Pearlmutter, B. A. Transformations of gaussian process priors. In *Deterministic and Statistical Methods in Machine Learning*, pp. 110–123. Springer, 2005.

Nakagome, S., Fukumizu, K., and Mano, S. Kernel approximate Bayesian computation in population genetic inferences. *Statistical applications in genetics and molecular biology*, 12(6):667–678, 2013.

Papamakarios, G. and Murray, I. Fast ε-free inference of simulation models with bayesian conditional density estimation. In *Advances in Neural Information Processing Systems*, pp. 1028–1036, 2016.

Park, M., Jitkrittum, W., and Sejdinovic, D. K2-ABC: Approximate Bayesian Computation with Kernel Embeddings. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pp. 398–407. PMLR, 09–11 May 2016.

Rasmussen, C. E. and Williams, C. K. I. *Gaussian processes for machine learning*. The MIT Press, 2006.

Song, L., Huang, J., Smola, A., and Fukumizu, K. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 961–968. ACM, 2009.

Song, L., Fukumizu, K., and Gretton, A. Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. *IEEE Signal Processing Magazine*, 30(4):98–111, 2013.

Tran, D., Ranganath, R., and Blei, D. Hierarchical implicit models and likelihood-free variational inference. In *Advances in Neural Information Processing Systems*, pp. 5523–5533, 2017.

A. Supporting Proofs for Section 3

Proof of Theorem 3.1. Let $f \in \mathcal{H}_k$ and $g(y) := \mathbb{E}[f(X)|Y = y]$. Assuming $g \in \mathcal{H}_\ell$, then $C_{YY}g = C_{YX}f$ (Fukumizu et al., 2004), so that

$$\begin{aligned} g &= C_{YY}^{-1}C_{YX}f \\ &= ((C_{YX})^T C_{YY}^{-1})^T f \\ &= (C_{XY}C_{YY}^{-1})^T f, \end{aligned} \tag{A.1}$$

where the inverse C_{YY}^{-1} exists because $\ell(y, \cdot)$ is assumed to be in the image of C_{YY} so that any $g \in \mathcal{H}_\ell$ is also in the image. Hence, $C_{XY}C_{YY}^{-1}$ satisfies the definition of a CMO. □

Proof of Lemma 3.2. Each of the following statements are equivalent to each other.

$$\begin{aligned} & (C_{X|Y})^T f = \mathbb{E}[f(X)|Y = \cdot], \quad \forall f \in \mathcal{H}_k \\ \iff & \langle \ell(y, \cdot), (C_{X|Y})^T f\rangle_{\mathcal{H}_\ell} = \langle \ell(y, \cdot), \mathbb{E}[f(X)|Y = \cdot]\rangle_{\mathcal{H}_\ell}, \quad \forall f \in \mathcal{H}_k, \quad \forall y \in \mathcal{Y} \\ \iff & \langle C_{X|Y}\ell(y, \cdot), f\rangle_{\mathcal{H}_k} = \mathbb{E}[f(X)|Y = y] = \langle \mathbb{E}[k(X, \cdot)|Y = y], f\rangle_{\mathcal{H}_k}, \quad \forall f \in \mathcal{H}_k, \quad \forall y \in \mathcal{Y} \\ \iff & C_{X|Y}\ell(y, \cdot) = \mathbb{E}[k(X, \cdot)|Y = y], \quad \forall y \in \mathcal{Y}. \end{aligned} \tag{A.2}$$

Consequently, the first and last statements are equivalent. □

Proof of Theorem 3.3. We show that the empirical CMO can be written as (3.4). We use a special case of the Woodbury identity (Higham, 2002), $B(CB + \lambda I)^{-1} = (BC + \lambda I)^{-1}B$, where B and C are appropriately defined operators, such matrices with the correct shapes. Using the empirical forms for the cross-covariance operators, we have

$$\begin{aligned} \hat{C}_{X|Y} &:= \hat{C}_{XY}(\hat{C}_{YY} + \lambda I)^{-1} \\ &= \frac{1}{n}\Phi\Psi^T(\frac{1}{n}\Psi\Psi^T + \lambda I)^{-1} \\ &= \Phi\Psi^T(\Psi\Psi^T + n\lambda I)^{-1} \\ &= \Phi(\Psi^T\Psi + n\lambda I)^{-1}\Psi^T \\ &= \Phi(L + n\lambda I)^{-1}\Psi^T. \end{aligned} \tag{A.3}$$

□

B. Supporting Proofs for Section 4

Proof of Theorem 4.1. Let $f \in \mathcal{H}_k$ and $g(y) := \mathbb{E}[f(X)|Y = y]$, then from definition 3.2 we have

$$\begin{aligned} g &= (C_{X|Y})^T f \\ C_{XY} g &= C_{XY}(C_{X|Y})^T f \\ C_{X|Y} C_{YY} g &= C_{X|Y} C_{YY} (C_{X|Y})^T f \\ (C_{X|Y} C_{YY} (C_{X|Y})^T)^{-1} C_{X|Y} C_{YY} g &= f \\ ((C_{X|Y} C_{YY})^T (C_{X|Y} C_{YY} (C_{X|Y})^T)^{-1})^T g &= f, \end{aligned} \tag{B.1}$$

where the inverse $(C_{X|Y} C_{YY} (C_{X|Y})^T)^{-1}$ exists because $C_{XY} g \in \mathcal{H}_k$ and $k(x, \cdot) \in \text{image}(C_{X|Y} C_{YY} (C_{X|Y})^T)$ so that $C_{XY} g$ for any $g \in \mathcal{H}_\ell$ is also in the image. In the last line we also used the fact that $(C_{X|Y} C_{YY} (C_{X|Y})^T)^T = C_{X|Y} C_{YY} (C_{X|Y})^T$ is symmetric since $(C_{YY})^T = C_{YY}$. Hence, $(C_{X|Y} C_{YY})^T (C_{X|Y} C_{YY} (C_{X|Y})^T)^{-1}$ satisfies the definition of a DMO. The assumption $\ell(y, \cdot) \in \text{image}(C_{YY})$ is required so that the original CMO exists and is unique. □

Proof of Theorem 4.2. Since $\ell(y, \cdot) \in \text{image}(C_{YY})$ for all $y \in \mathcal{Y}$ and $k(x, \cdot) \in \text{image}(C_{X|Y} C_{YY} (C_{X|Y})^T)$ for all $x \in \mathcal{X}$, we have that C_{YY}^{-1} exists so that $C_{X|Y}$ is unique and $(C_{X|Y} C_{YY} (C_{X|Y})^T)^{-1}$ exists so that $C'_{X|Y}$ is unique. Due to theorem 4.1 we have $C'_{X|Y} = (C_{X|Y} C_{YY})^T (C_{X|Y} C_{YY} (C_{X|Y})^T)^{-1}$. Since $C_{X|Y} C_{YY} (C_{X|Y})^T$ is at least positive semi-definite and invertible we can write $(C_{X|Y} C_{YY} (C_{X|Y})^T)^{-1} = \lim_{\epsilon \to 0^+} (C_{X|Y} C_{YY} (C_{X|Y})^T + \epsilon I)^{-1}$,

$$\begin{aligned} C'_{X|Y} &= \lim_{\epsilon \to 0^+} (C_{X|Y} C_{YY})^T (C_{X|Y} C_{YY} (C_{X|Y})^T + \epsilon I)^{-1} \\ &= \lim_{\epsilon \to 0^+} (C_{X|Y} C_{YY})^T (C_{X|Y} (C_{X|Y} C_{YY})^T + \epsilon I)^{-1} \\ &= \lim_{\epsilon \to 0^+} (C_{XY})^T (C_{X|Y} (C_{XY})^T + \epsilon I)^{-1} \\ &= \lim_{\epsilon \to 0^+} C_{YX} (C_{X|Y} C_{YX} + \epsilon I)^{-1} \\ &= \lim_{\epsilon \to 0^+} (C_{YX} C_{X|Y} + \epsilon I)^{-1} C_{YX} \\ &= \lim_{\epsilon \to 0^+} (C_{YY} C_{YY}^{-1} C_{YX} C_{X|Y} + \epsilon C_{YY} C_{YY}^{-1})^{-1} C_{YX} \\ &= \lim_{\epsilon \to 0^+} (C_{YY}^{-1} C_{YX} C_{X|Y} + \epsilon C_{YY}^{-1})^{-1} C_{YY}^{-1} C_{YX} \\ &= \lim_{\epsilon \to 0^+} (C_{YY}^{-1} C_{YX} C_{X|Y} + \epsilon C_{YY}^{-1})^{-1} C_{YY}^{-1} C_{YX} \\ &= \lim_{\epsilon \to 0^+} ((C_{XY} C_{YY}^{-1})^T C_{X|Y} + \epsilon C_{YY}^{-1})^{-1} (C_{XY} C_{YY}^{-1})^T \\ &= \lim_{\epsilon \to 0^+} ((C_{X|Y})^T C_{X|Y} + \epsilon C_{YY}^{-1})^{-1} (C_{X|Y})^T \\ &= ((C_{X|Y})^T C_{X|Y})^{-1} (C_{X|Y})^T =: C_{X|Y}^\dagger. \end{aligned} \tag{B.2}$$

In line 6 we used the Woodbury identity (Higham, 2002). In the last line, the limit exists as $((C_{X|Y})^T C_{X|Y})^{-1}$ exists. □

Proof of Theorem 4.3. We show that the empirical DMO can be written as (4.4). From definition 3.3 and theorem 3.3, the likelihood operator is estimated from $\{x_i, y_i\}_{i=1}^n$ as

$$\hat{C}_{X|Y} := \hat{C}_{XY} (\hat{C}_{XX} + \lambda I)^{-1} = \Phi (L + n\lambda I)^{-1} \Psi^T. \tag{B.3}$$

The prior operator corresponding to the marginal $\mathbb{P}_Y$ is estimated from $\{\tilde{y}_j\}_{j=1}^m$ as

$$\tilde{C}_{YY} = \frac{1}{m} \tilde{\Psi} \tilde{\Psi}^T. \tag{B.4}$$

Let $A := (L + n\lambda I)^{-1} \tilde{L}$, the joint operator is estimated as

$$\hat{C}_{X|Y} \tilde{C}_{YY} = \frac{1}{m} \Phi (L + n\lambda I)^{-1} \Psi^T \tilde{\Psi} \tilde{\Psi}^T = \frac{1}{m} \Phi (L + n\lambda I)^{-1} \tilde{L} \tilde{\Psi}^T = \frac{1}{m} \Phi A \tilde{\Psi}^T. \tag{B.5}$$

The evidence operator is estimated as

$$\begin{aligned}\hat{C}_{X|Y}\tilde{C}_{YY}(\hat{C}_{X|Y})^T &= \frac{1}{m}\Phi(L+n\lambda I)^{-1}\tilde{L}\tilde{\Psi}^T\Psi(L+n\lambda I)^{-1}\Phi^T \\ &= \frac{1}{m}\Phi(L+n\lambda I)^{-1}\tilde{L}\tilde{L}^T(L+n\lambda I)^{-1}\Phi^T \\ &= \frac{1}{m}\Phi AA^T\Phi^T.\end{aligned} \tag{B.6}$$

Finally, by definition 4.3, the DMO is estimated as

$$\begin{aligned}\bar{C}'_{X|Y} &= (\hat{C}_{X|Y}\tilde{C}_{YY})^T(\hat{C}_{X|Y}\tilde{C}_{YY}(\hat{C}_{X|Y})^T+\epsilon I)^{-1} \\ &= \left[\frac{1}{m}\Phi A\tilde{\Psi}^T\right]^T\left[\frac{1}{m}\Phi AA^T\Phi^T+\epsilon I\right]^{-1} \\ &= \left[\Phi A\tilde{\Psi}^T\right]^T\left[\Phi AA^T\Phi^T+m\epsilon I\right]^{-1} \\ &= \tilde{\Psi}A^T\Phi^T\left[\Phi AA^T\Phi^T+m\epsilon I\right]^{-1} \\ &= \tilde{\Psi}\left[A^T\Phi^T\Phi A+m\epsilon I\right]^{-1}A^T\Phi^T \\ &= \tilde{\Psi}\left[A^TKA+m\epsilon I\right]^{-1}A^T\Phi^T.\end{aligned} \tag{B.7}$$

□

C. Supporting Proofs for Section 5

Proof of Theorem 5.1. Each optimization is a standard regularized least squares problem. The first optimization over $\mathbf{v}$ can be written as

$$\hat{\mathbf{v}}[\mathbf{w}] = \operatorname*{argmin}_{\mathbf{v}\in\mathbb{R}^q} \|\Phi^T\mathbf{w} - \Psi^T\mathbf{v}\|^2 + n\lambda\|\mathbf{v}\|^2, \tag{C.1}$$

where $\mathbf{f} = \Phi^T\mathbf{w}$ is the target and Ψ is the feature matrix. This gives the solution $\hat{\mathbf{v}}[\mathbf{w}] = (\Psi\Psi^T + n\lambda I)^{-1}\Psi(\Phi^T\mathbf{w})$. Therefore, The second optimization over $\mathbf{w}$ can be written as

$$\begin{aligned}
\bar{\mathbf{w}} &= \operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^p} \|\tilde{\mathbf{z}} - \tilde{\Psi}^T\hat{\mathbf{v}}[\mathbf{w}]\|^2 + m\epsilon\|\mathbf{w}\|^2 \\
&= \operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^p} \|\tilde{\mathbf{z}} - \tilde{\Psi}^T(\Psi\Psi^T + n\lambda I)^{-1}\Psi\Phi^T\mathbf{w}\|^2 + m\epsilon\|\mathbf{w}\|^2 \\
&= \operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^p} \|\tilde{\mathbf{z}} - A^T\Phi^T\mathbf{w}\|^2 + m\epsilon\|\mathbf{w}\|^2 \\
&= \operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^p} \|\tilde{\mathbf{z}} - \Theta^T\mathbf{w}\|^2 + m\epsilon\|\mathbf{w}\|^2,
\end{aligned} \tag{C.2}$$

where we used $A := \Psi^T(\Psi\Psi^T + n\lambda I)^{-1}\tilde{\Psi}$ as per definition 4.5 and we define $\Theta := \Phi A$. This is now a regularized least squares problem with $\tilde{\mathbf{z}}$ as the target and $\Theta := \Phi A$ as the feature matrix. This gives the solution $\bar{\mathbf{w}} = (\Theta\Theta^T + m\epsilon I)^{-1}\Theta\tilde{\mathbf{z}} = (\Phi AA^T\Phi^T + m\epsilon I)^{-1}\Phi A\tilde{\mathbf{z}}$, which yields the parametric DME estimator in definition 4.5. □

Proof of Lemma 5.2. We first establish that the transformation matrix in definition 4.4 $A = (L + n\lambda I)^{-1}\tilde{L}$ is the same as the transformation matrix in definition 4.5 $A = \Psi^T(\Psi\Psi^T + n\lambda I)^{-1}\tilde{\Psi}$ via a special case of the Woodbury identity $B(CB + \delta I)^{-1} = (BC + \delta I)^{-1}B$ for appropriately sized matrices or operators B and C (Higham, 2002). Consequently, $(L + n\lambda I)^{-1}\tilde{L} = (\Psi^T\Psi + n\lambda I)^{-1}\Psi^T\tilde{\Psi} = \Psi^T(\Psi\Psi^T + n\lambda I)^{-1}\tilde{\Psi}$.

From definition 4.4 we have $\bar{\boldsymbol{\alpha}} := A\big[A^TKA + m\epsilon I\big]^{-1}\tilde{\mathbf{z}}$ so that

$$\begin{aligned}
\Phi\bar{\boldsymbol{\alpha}} &= \Phi A\big[A^TKA + m\epsilon I\big]^{-1}\tilde{\mathbf{z}} \\
&= \Phi A\big[A^T\Phi^T\Phi A + m\epsilon I\big]^{-1}\tilde{\mathbf{z}} \\
&= \big[\Phi AA^T\Phi^T + m\epsilon I\big]^{-1}\Phi A\tilde{\mathbf{z}} \\
&= \bar{\mathbf{w}}.
\end{aligned} \tag{C.3}$$

This relationship is a direct consequence of the kernel trick, where we used $k(x, x') = \phi(x)^T\phi(x')$ such that $K = \Phi^T\Phi$. □

Proof of Theorem 5.3 Part 1 – *Task Transformed BLR (TTBLR).* In this proof we provide the derivations for task transformed BLR (TTBLR). We first reiterate the priors and likelihoods used.

Priors We first place priors on the weights of our linear models $g(y) = \mathbf{v}^T\psi(y)$ and $f(x) = \mathbf{w}^T\psi(x)$,

$$\begin{aligned}
p(\mathbf{v}) &\sim \mathcal{N}(\mathbf{v}; \mathbf{0}, \beta^2 I), \\
p(\mathbf{w}) &\sim \mathcal{N}(\mathbf{w}; \mathbf{0}, \gamma^2 I).
\end{aligned} \tag{C.4}$$

Likelihoods As we only observe from g and never from f directly, there is no need to add noise from $f(x)$ to z and we degenerate the likelihood to $z = f(x)$. The likelihood for g is the regular Gaussian likelihood due to observational noise. Together, we have

$$\begin{aligned}
p(z|\mathbf{v}) &= \mathcal{N}(z; \mathbf{v}^T\psi(y), \sigma^2), \\
p(z|\mathbf{w}) &= \mathcal{N}(z; \mathbf{w}^T\phi(x), 0).
\end{aligned} \tag{C.5}$$

Prior for g The prior on the weights of g is

$$p(\mathbf{v}) = \mathcal{N}(\mathbf{v}; \mathbf{0}, \beta^2 I). \tag{C.6}$$

Likelihood for g In task transformed learning, the pairs $(\mathbf{y}, \mathbf{z})$ are used to learn g, and $(\tilde{\mathbf{y}}, \tilde{\mathbf{z}})$ are the query points for g. Although $\mathbf{z}$ is not directly available, they are propagated through from f. We also refer to $\mathbf{z}$ as the pseudo-training targets. This leads to the following likelihood,

$$\begin{aligned} p(\mathbf{z}|\mathbf{v}) &= \mathcal{N}(\mathbf{z}; \Psi^T \mathbf{v}, \sigma^2 I), \\ p(\tilde{\mathbf{z}}|\mathbf{v}) &= \mathcal{N}(\tilde{\mathbf{z}}; \tilde{\Psi}^T \mathbf{v}, \sigma^2 I). \end{aligned} \tag{C.7}$$

Marginal Likelihood for g The marginal likelihood of observing the pseudo-training targets $\mathbf{z}$ is

$$\begin{aligned} p(\mathbf{z}) &= \int_{\mathbb{R}^q} p(\mathbf{z}|\mathbf{v}) p(\mathbf{v}) d\mathbf{v} \\ &= \mathcal{N}(\mathbf{z}; \mathbf{0}, \beta^2 \Psi^T \Psi + \sigma^2 I). \end{aligned} \tag{C.8}$$

Posterior for g The posterior of the weights given the pseudo-training targets $\mathbf{z}$ is

$$\begin{aligned} p(\mathbf{v}|\mathbf{z}) &= \frac{p(\mathbf{z}|\mathbf{v}) p(\mathbf{v})}{p(\mathbf{z})} \\ &= \mathcal{N}\left(\mathbf{v}; \left(\Psi\Psi^T + \frac{\sigma^2}{\beta^2} I\right)^{-1} \Psi \mathbf{z}, \sigma^2 \left(\Psi\Psi^T + \frac{\sigma^2}{\beta^2} I\right)^{-1}\right). \end{aligned} \tag{C.9}$$

Predictive distribution for g The posterior predictive distribution of $\tilde{\mathbf{z}}$ given the pseudo-training targets $\mathbf{z}$ is

$$\begin{aligned} p(\tilde{\mathbf{z}}|\mathbf{z}) &= \int_{\mathbb{R}^q} p(\tilde{\mathbf{z}}|\mathbf{v}) p(\mathbf{v}|\mathbf{z}) d\mathbf{v} \\ &= \mathcal{N}\left(\tilde{\mathbf{z}}; \tilde{\Psi}^T \left(\Psi\Psi^T + \frac{\sigma^2}{\beta^2} I\right)^{-1} \Psi \mathbf{z}, \sigma^2 \tilde{\Psi}^T \left(\Psi\Psi^T + \frac{\sigma^2}{\beta^2} I\right)^{-1} \tilde{\Psi} + \sigma^2 I\right) \\ &= \mathcal{N}(\tilde{\mathbf{z}}; A^T \mathbf{z}, \Sigma), \end{aligned} \tag{C.10}$$

where $A = \Psi^T (\Psi\Psi^T + \frac{\sigma^2}{\beta^2} I)^{-1} \tilde{\Psi}$ and $\Sigma = \sigma^2 \tilde{\Psi}^T (\Psi\Psi^T + \frac{\sigma^2}{\beta^2} I)^{-1} \tilde{\Psi} + \sigma^2 I$.

Importantly, the MAP solution for learning g amount to just taking the posterior mean $\hat{\mathbf{v}} = (\Psi\Psi^T + \frac{\sigma^2}{\beta^2} I)^{-1} \Psi \mathbf{z}$ as a point estimate. In this case, the predictive covariance would simplify to $\Sigma = \sigma^2 I$.

Prior for f The prior on the weights of f is

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}; \mathbf{0}, \gamma^2 I). \tag{C.11}$$

Likelihood for f As targets z are never directly observed from f, the likelihood is a noiseless Gaussian likelihood,

$$\begin{aligned} p(\mathbf{z}|\mathbf{w}) &= \mathcal{N}(\mathbf{z}; \Phi^T \mathbf{w}, 0I), \\ p(\mathbf{z}^\star|\mathbf{w}) &= \mathcal{N}(\mathbf{z}^\star; (\Phi^\star)^T \mathbf{w}, 0I). \end{aligned} \tag{C.12}$$

Propagating this likelihood through the predictive distribution of g, we have

$$\begin{aligned} p(\tilde{\mathbf{z}}|\mathbf{w}) &= \int_{\mathbb{R}^n} p(\tilde{\mathbf{z}}|\mathbf{z}) p(\mathbf{z}|\mathbf{w}) d\mathbf{z} \\ &= \mathcal{N}(\tilde{\mathbf{z}}; A^T \Phi^T \mathbf{w}, \Sigma). \end{aligned} \tag{C.13}$$

The above prior-likelihood pair describes a TBLR with $M = A = \Psi^T (\Psi\Psi^T + \frac{\sigma^2}{\beta^2} I)^{-1} \tilde{\Psi}$ as the transformation matrix and $\Sigma = \sigma^2 \tilde{\Psi}^T (\Psi\Psi^T + \frac{\sigma^2}{\beta^2} I)^{-1} \tilde{\Psi} + \sigma^2 I$ as the noise covariance. As such, the remaining distributions exhibit the same forms as shown in table C.1.

Table C.1. Summary of TBLR and TBKR, where $C := [\Phi M \Sigma^{-1} M^T \Phi^T + \frac{1}{\gamma^2} I]^{-1}$, $\mathbf{m} := C\Phi M \Sigma^{-1} \tilde{\mathbf{z}}$, and $S := M^T K M + \Sigma$.

Density	Transformed Bayesian Linear Regression	Transformed Bayesian Kernel Regression
Prior	$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}; \mathbf{0}, \gamma^2)$	$p(\mathbf{f}) = \mathcal{N}(\mathbf{f}; \mathbf{0}, K)$
Likelihood	$p(\tilde{\mathbf{z}}\|\mathbf{w}) = \mathcal{N}(\tilde{\mathbf{z}}; M^T \Phi^T \mathbf{w}, \Sigma)$	$p(\tilde{\mathbf{z}}\|\mathbf{f}) = \mathcal{N}(\tilde{\mathbf{z}}; M^T \mathbf{f}, \Sigma)$
Evidence	$p(\tilde{\mathbf{z}}) = \mathcal{N}(\tilde{\mathbf{z}}; \mathbf{0}, [\Sigma^{-1} - \Sigma^{-1} M^T \Phi^T C \Phi M \Sigma^{-1}]^{-1})$	$p(\tilde{\mathbf{z}}) = \mathcal{N}(\tilde{\mathbf{z}}; \mathbf{0}, M^T K M + \Sigma)$
Posterior	$p(\mathbf{w}\|\tilde{\mathbf{z}}) = \mathcal{N}(\mathbf{w}; \mathbf{m}, C)$	$p(\mathbf{f}\|\tilde{\mathbf{z}}) = \mathcal{N}(\mathbf{f}; KMS^{-1}\tilde{\mathbf{z}}, K - KMS^{-1}M^T K)$
Predictive	$p(\mathbf{z}^\star\|\tilde{\mathbf{z}}) = \mathcal{N}(\mathbf{z}^\star; \Phi^{\star T}\mathbf{m}, \Phi^{\star T} C \Phi^\star)$	$p(\mathbf{z}^\star\|\tilde{\mathbf{z}}) = \mathcal{N}(\mathbf{z}^\star; K^{\star T} M S^{-1} \tilde{\mathbf{z}}, K^{\star\star} - K^{\star T} M S^{-1} M^T K^\star)$

Marginal Likelihood for f The marginal likelihood for the observed targets $\tilde{\mathbf{z}}$ is

$$\begin{aligned} p(\tilde{\mathbf{z}}) &= \int_{\mathbb{R}^p} p(\tilde{\mathbf{z}}|\mathbf{w}) p(\mathbf{w}) d\mathbf{w} \\ &= \mathcal{N}(\tilde{\mathbf{z}}; \mathbf{0}, \gamma^2 A^T \Phi^T \Phi A + \Sigma) \\ &= \mathcal{N}(\tilde{\mathbf{z}}; \mathbf{0}, [\Sigma^{-1} - \Sigma^{-1} A^T \Phi^T C \Phi A \Sigma^{-1}]^{-1}), \end{aligned} \tag{C.14}$$

where $C = [\Phi A \Sigma^{-1} A^T \Phi^T + \frac{1}{\gamma^2} I]^{-1}$. The last line is an alternative form that is more computationally efficient when the number of features is less than $p < m$ where p is the dimensionality of the feature $\phi(x)$ for f.

Posterior for f The posterior of the weights $\mathbf{w}$ given the observed targets $\tilde{\mathbf{z}}$ is

$$\begin{aligned} p(\mathbf{w}|\tilde{\mathbf{z}}) &= \frac{p(\tilde{\mathbf{z}}|\mathbf{w}) p(\mathbf{w})}{p(\tilde{\mathbf{z}})} \\ &= \mathcal{N}(\mathbf{w}; \mathbf{m}, C), \end{aligned} \tag{C.15}$$

where $\mathbf{m} := C\Phi A \Sigma^{-1} \tilde{\mathbf{z}}$.

Predictive distribution for f Finally, the overall predictive distribution of query targets $\mathbf{z}^\star$ given the observed targets $\tilde{\mathbf{z}}$ is

$$\begin{aligned} p(\mathbf{z}^\star|\tilde{\mathbf{z}}) &= \int_{\mathbb{R}^p} p(\mathbf{z}^\star|\mathbf{w}) p(\mathbf{w}|\tilde{\mathbf{z}}) d\mathbf{w} \\ &= \mathcal{N}(\mathbf{z}^\star; \Phi^{\star T}\mathbf{m}, \Phi^{\star T} C \Phi^\star). \end{aligned} \tag{C.16}$$

Consider the posterior mean $\mathbf{m} := C\Phi A \Sigma^{-1} \tilde{\mathbf{z}} = [\Phi A \Sigma^{-1} A^T \Phi^T + \frac{1}{\gamma^2} I]^{-1} \Phi M \Sigma^{-1} \tilde{\mathbf{z}}$, which would also be the MAP solution for f. Using the MAP solution for learning g such that $\Sigma = \sigma^2 I$, we have $\mathbf{m} := [\Phi A A^T \Phi^T + \frac{\sigma^2}{\gamma^2} I]^{-1} \Phi A \tilde{\mathbf{z}}$. This is the same form as the weights $\tilde{\mathbf{w}}$ of the parametric DME estimator (definition 4.5) with $\lambda = \frac{\sigma^2}{n\beta^2}$ and $\epsilon = \frac{\sigma^2}{m\gamma^2}$. □

***Proof of Theorem 5.3 Part 2** – Task Transformed BKR (TTBKR).* In this proof we provide the derivations for task transformed BKR (TTBKR), also named task transformed Gaussian process regression (TTGPR), whose graphical model is provided in fig. 2. We first reiterate the priors and likelihoods used.

Priors We place GP priors on the functions g and f directly,

$$\begin{aligned} g &\sim \mathcal{GP}(0, \ell), \\ f &\sim \mathcal{GP}(0, k). \end{aligned} \tag{C.17}$$

Likelihoods As we only observe from g and never from f directly, there is no need to add noise from $f(x)$ to z and we degenerate the likelihood to $z = f(x)$. The likelihood for g is the regular Gaussian likelihood due to observational noise. Together, we have

$$\begin{aligned} p(z|g) &= \mathcal{N}(z; g(y), \sigma^2), \\ p(z|f) &= \mathcal{N}(z; f(x), 0). \end{aligned} \tag{C.18}$$

Prior for g The prior of g at $\mathbf{y}$ is

$$p(\mathbf{g}) = \mathcal{N}(\mathbf{g}; \mathbf{0}, L). \tag{C.19}$$

Likelihood for g In task transformed learning, the pairs $(\mathbf{y}, \mathbf{z})$ are used to learn g, and $(\tilde{\mathbf{y}}, \tilde{\mathbf{z}})$ are the query points for g. Although $\mathbf{z}$ is not directly available, they are propagated through from f. We also refer to $\mathbf{z}$ as the pseudo-training targets. The likelihood of observing $\mathbf{z}$ at $\mathbf{y}$ is,

$$p(\mathbf{z}|\mathbf{g}) = \mathcal{N}(\mathbf{z}; \mathbf{g}, \sigma^2 I). \tag{C.20}$$

Marginal Likelihood for g The marginal likelihood of observing the psuedo-training targets $\mathbf{z}$ is:

$$\begin{aligned} p(\mathbf{z}) &= \int_{\mathbb{R}^n} p(\mathbf{z}|\mathbf{g}) p(\mathbf{g}) d\mathbf{g} \\ &= \mathcal{N}(\mathbf{z}; \mathbf{0}, L + \sigma^2 I). \end{aligned} \tag{C.21}$$

Posterior for g The posterior of the latent function evaluations $\mathbf{g}$ at $\mathbf{y}$ given the pseudo-training targets $\mathbf{z}$ is

$$\begin{aligned} p(\mathbf{g}|\mathbf{z}) &= \frac{p(\mathbf{z}|\mathbf{g}) p(\mathbf{g})}{p(\mathbf{z})} \\ &= \mathcal{N}(\mathbf{g}; L(L + \sigma^2 I)^{-1}\mathbf{z}, L - L(L + \sigma^2 I)^{-1} L). \end{aligned} \tag{C.22}$$

Predictive distribution for g To obtain the predictive distribution, we first condition the GP field for on the latent function evaluations $\mathbf{g}$ at $\mathbf{y}$ and to obtain the conditional distribution for $\tilde{\mathbf{g}}$ at $\tilde{\mathbf{y}}$ given $\mathbf{g}$ at $\mathbf{y}$,

$$p(\tilde{\mathbf{g}}|\mathbf{g}) = \mathcal{N}(\tilde{\mathbf{g}}; \tilde{L}^T L^{-1} \mathbf{g}, \tilde{\tilde{L}} - \tilde{L}^T L^{-1} \tilde{L}). \tag{C.23}$$

where $\tilde{\tilde{L}} := \tilde{\Psi}^T \tilde{\Psi}$. Now, marginalize the conditional field against the posterior,

$$\begin{aligned} p(\tilde{\mathbf{g}}|\mathbf{z}) &= \int_{\mathbb{R}^n} p(\tilde{\mathbf{g}}|\mathbf{g}) p(\mathbf{g}|\mathbf{z}) d\mathbf{g} \\ &= \mathcal{N}(\tilde{\mathbf{g}}; \tilde{L}^T (L + \sigma^2 I)^{-1} \mathbf{z}, \tilde{\tilde{L}} - \tilde{L}^T (L + \sigma^2 I)^{-1} \tilde{L}) \end{aligned} \tag{C.24}$$

Finally, marginalize the likelihood $p(\tilde{\mathbf{g}}|\mathbf{z})$ with the predictive distribution of the latent evaluations $\tilde{\mathbf{g}}$ to get the final predictive distribution of the observations $\tilde{\mathbf{z}}$,

$$\begin{aligned} p(\tilde{\mathbf{z}}|\mathbf{z}) &= \int_{\mathbb{R}^m} p(\tilde{\mathbf{z}}|\tilde{\mathbf{g}}) p(\tilde{\mathbf{g}}|\mathbf{z}) d\tilde{\mathbf{g}} \\ &= \mathcal{N}(\tilde{\mathbf{z}}; \tilde{L}^T (L + \sigma^2 I)^{-1} \mathbf{z}, \tilde{\tilde{L}} + \sigma^2 I - \tilde{L}^T (L + \sigma^2 I)^{-1} \tilde{L}) \\ &= \mathcal{N}(\tilde{\mathbf{z}}; A^T \mathbf{z}, \Sigma), \end{aligned} \tag{C.25}$$

where $A = (L + \sigma^2 I)^{-1} \tilde{L}$ and $\Sigma = \tilde{\tilde{L}} + \sigma^2 I - \tilde{L}^T (L + \sigma^2 I)^{-1} \tilde{L}$.

Importantly, the MAP solution for learning g amount to just taking the posterior mean $\tilde{\mathbf{g}} = \tilde{L}^T (L + \sigma^2 I)^{-1} \mathbf{z}$ as a point estimate. In this case, the predictive covariance would simplify to $\Sigma = \sigma^2 I$.

Prior for f The prior of f at $\mathbf{x}$ is

$$p(\mathbf{f}) = \mathcal{N}(\mathbf{f}; \mathbf{0}, K). \tag{C.26}$$

Likelihood for f As targets z are never directly observed from f, the likelihood is a noiseless Gaussian likelihood,

$$p(\mathbf{z}|\mathbf{f}) = \mathcal{N}(\mathbf{z}; \mathbf{f}, 0I). \tag{C.27}$$

Propagating this likelihood through the predictive distribution of g, we have

$$\begin{aligned} p(\tilde{\mathbf{z}}|\mathbf{f}) &= \int_{\mathbb{R}^n} p(\tilde{\mathbf{z}}|\mathbf{z}) p(\mathbf{z}|\mathbf{f}) d\mathbf{z} \\ &= \mathcal{N}(\tilde{\mathbf{z}}; A^T \mathbf{f}, \Sigma). \end{aligned} \tag{C.28}$$

The above prior-likelihood pair describes a TBKR with $M = A = (L + \sigma^2 I)^{-1}\tilde{L}$ as the transformation matrix and $\Sigma = \tilde{\tilde{L}} + \sigma^2 I - \tilde{L}^T(L + \sigma^2 I)^{-1}\tilde{L}$ as the noise covariance. As such, the remaining distribution exhibit the same forms as shown in table C.1.

Marginal Likelihood for f The marginal likelihood for the observed targets $\tilde{\mathbf{z}}$ is

$$\begin{aligned} p(\tilde{\mathbf{z}}) &= \int_{\mathbb{R}^n} p(\tilde{\mathbf{z}}|\mathbf{f})p(\mathbf{f})d\mathbf{f} \\ &= \mathcal{N}(\tilde{\mathbf{z}}; \mathbf{0}, A^T K A + \Sigma). \end{aligned} \tag{C.29}$$

Posterior for f The posterior of the function evaluations $\mathbf{f}$ at $\mathbf{x}$ given the observed targets $\tilde{\mathbf{z}}$ is

$$\begin{aligned} p(\mathbf{f}|\tilde{\mathbf{z}}) &= \frac{p(\tilde{\mathbf{z}}|\mathbf{f})p(\mathbf{f})}{p(\tilde{\mathbf{z}})} \\ &= \mathcal{N}(\mathbf{f}; KA(A^T K A + \Sigma)^{-1}\tilde{\mathbf{z}}, K - KA(A^T K A + \Sigma)^{-1}A^T K). \end{aligned} \tag{C.30}$$

Predictive distribution for f Finally, to obtain the predictive distribution we first condition the GP field on the latent function evaluations $\mathbf{f}$ to obtain the conditional distribution for $\mathbf{f}^\star$ at $\mathbf{x}^\star$ given $\mathbf{f}$ at $\mathbf{x}$.

$$p(\mathbf{f}^\star|\mathbf{f}) = \mathcal{N}(\mathbf{f}^\star; (K^\star)^T K^{-1}\mathbf{f}, K^{\star\star} - (K^\star)^T K^{-1} K^\star). \tag{C.31}$$

Now, marginalize the conditional field against the posterior,

$$\begin{aligned} p(\mathbf{f}^\star|\tilde{\mathbf{z}}) &= \int_{\mathbb{R}^n} p(\mathbf{f}^\star|\mathbf{f})p(\mathbf{f}|\tilde{\mathbf{z}})d\mathbf{f} \\ &= \mathcal{N}(\mathbf{f}^\star; (K^\star)^T A(A^T K A + \Sigma)^{-1}\tilde{\mathbf{z}}, K^{\star\star} - (K^\star)^T A(A^T K A + \Sigma)^{-1}A^T K^\star). \end{aligned} \tag{C.32}$$

Finally, the overall predictive distribution of query targets $\mathbf{z}^\star$ given the observed targets $\tilde{\mathbf{z}}$ is

$$\begin{aligned} p(\mathbf{z}^\star|\tilde{\mathbf{z}}) &= \int_{\mathbb{R}^n} p(\mathbf{z}^\star|\mathbf{f}^\star)p(\mathbf{f}^\star|\tilde{\mathbf{z}})d\mathbf{f}^\star \\ &= \mathcal{N}(\mathbf{z}^\star; (K^\star)^T A(A^T K A + \Sigma)^{-1}\tilde{\mathbf{z}}, K^{\star\star} - (K^\star)^T A(A^T K A + \Sigma)^{-1}A^T K^\star). \end{aligned} \tag{C.33}$$

Consider the posterior predictive mean at a particular query point x, $\bar{f}(x) = (\mathbf{k}(x))^T A(A^T K A + \Sigma)^{-1}\tilde{\mathbf{z}} = \tilde{\mathbf{z}}^T(A^T K A + \Sigma)^{-1}A^T\mathbf{k}(x)$. Using the MAP solution for learning g such that $\Sigma = \sigma^2 I$, we have $\bar{f}(x) = \tilde{\mathbf{z}}^T(A^T K A + \sigma^2 I)^{-1}A^T\mathbf{k}(x)$. This is the same form as the nonparametric DME estimator (definition 4.5) with $\lambda = \frac{\sigma^2}{n}$ and $\epsilon = \frac{\sigma^2}{m}$. □

D. Supporting Proofs for Section 6

Proof of Theorem 6.1. We first factorize the joint operator $C_{YX} = (C_{XY})^T$ in both directions,

$$C_{Y|X}C_{XX} = C_{YX} = (C_{XY})^T = (C_{X|Y}C_{YY})^T = C_{YY}(C_{X|Y})^T. \tag{D.1}$$

This is analogous to the equation $p(y|x)p(x) = p(y,x) = p(x,y) = p(x|y)p(y) = p(y)p(x|y)$.

Since $C_{X|Y}C_{Y|X}C_{XX} = C_{XX}$, we then apply $C_{X|Y}$ on both sides to cancel out $C_{Y|X}$ and obtain the equation for C_{XX},

$$C_{XX} = (C_{X|Y}C_{Y|X})C_{XX} = C_{X|Y}(C_{Y|X}C_{XX}) = C_{X|Y}C_{YY}(C_{X|Y})^T. \tag{D.2}$$

This is analogous to the equation $p(x) = \int_{\mathcal{Y}} p(y|x)dyp(x) = \int_{\mathcal{Y}} p(y|x)p(x)dy = \int_{\mathcal{Y}} p(y)p(x|y)dy$.

Hence,

$$C'_{XX} := C_{X|Y}C_{YY}(C_{X|Y})^T = C_{XX}. \tag{D.3}$$

Finally, from theorem 4.1 we have

$$C'_{X|Y} = (C_{X|Y}C_{YY})^T(C_{X|Y}C_{YY}(C_{X|Y})^T)^{-1} = C_{YX}C_{XX}^{-1} = C_{Y|X}. \tag{D.4}$$

□

Proof of Theorem 6.2. Since $m = n$ and $\tilde{\mathbf{y}} = \mathbf{y}$, we have that $\tilde{L} = L$, $\tilde{\Psi} = \Psi$. Consequently, $\lim_{\lambda\to 0^+} A = \lim_{\lambda\to 0^+}(L + n\lambda I)^{-1}L = I$. Substituting this into (3.4) we have

$$\lim_{\lambda\to 0^+} \bar{C}'_{X|Y} = \lim_{\lambda\to 0^+} \tilde{\Psi}\big[A^TKA + m\epsilon I\big]^{-1}A^T\Phi^T \quad = \Psi\big[I^TKI + n\epsilon I\big]^{-1}I^T\Phi^T = \Psi\big[K + n\epsilon I\big]^{-1}\Phi^T. \tag{D.5}$$

Reversing the roles of X and Y in (3.4) and replacing the notation λ with ϵ, we have that $\hat{C}_{Y|X} = \Psi\big[K + n\epsilon I\big]^{-1}\Phi^T$. This concludes the proof. □

Table E.1. Mean embeddings and their encoded expectations. Switch $X \leftrightarrow Y$ for all combinations. Since the bottom two rows do not apply for the first column, additional equivalences for the last column are provided instead. The kernels $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ are positive definite and characteristic on their respective spaces $\mathcal{X}$ and $\mathcal{Y}$. They define the RKHS $\mathcal{H}_k$ and $\mathcal{H}_\ell$ respectively. We define $\mathcal{H}_{\ell\ell} := \mathcal{H}_\ell \otimes \mathcal{H}_\ell$ and $\mathcal{H}_{k\ell} := \mathcal{H}_k \otimes \mathcal{H}_\ell$ and let $g, g' \in \mathcal{H}_\ell$ and $f \in \mathcal{H}_k$ be generic example functions within each RKHS.

Random Variable	Y $: \Omega \to \mathcal{Y}$	(Y, Y) $: \Omega \to \mathcal{Y} \times \mathcal{Y}$	(X, Y) $: \Omega \to \mathcal{X} \times \mathcal{Y}$	$X \vert Y = y$ $: \Omega \to \mathcal{X}$
Density Function	$p_Y \in \mathcal{P}_\mathcal{Y}$ $p_Y(y) \in \mathbb{R}^+$	$p_{YY} \in \mathcal{P}_{\mathcal{Y}\times\mathcal{Y}}$ $p_{YY}(y, y') \in \mathbb{R}^+$	$p_{XY} \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$ $p_{XY}(x, y) \in \mathbb{R}^+$	$p_{X\vert Y=y} \in \mathcal{P}_\mathcal{X}$ $p_{X\vert Y=y}(x) \in \mathbb{R}^+$
Mean Map Definition	$\mu_Y :=$ $\mathbb{E}[\ell(Y, \cdot)]$	$\mu_{YY} :=$ $\mathbb{E}[\ell(Y, \cdot)\ell(Y, \cdot)^T]$	$\mu_{XY} :=$ $\mathbb{E}[k(X, \cdot)\ell(Y, \cdot)^T]$	$\mu_{X\vert Y=y} =$ $\mathbb{E}[k(X, \cdot)\vert Y = y]$
Mean Embedding	$\mu_Y \in \mathcal{H}_\ell$ $\mu_Y(y) \in \mathbb{R}$	$\mu_{YY} \in \mathcal{H}_{\ell\ell}$ $\mu_{YY}(y, y') \in \mathbb{R}$	$\mu_{XY} \in \mathcal{H}_{k\ell}$ $\mu_{XY}(x, y) \in \mathbb{R}$	$\mu_{X\vert Y=Y} \in \mathcal{H}_k$ $\mu_{X\vert Y=y}(x) \in \mathbb{R}$
Encoded Expectation	$\langle \mu_Y, g \rangle_{\mathcal{H}_\ell}$ $= \mathbb{E}[g(Y)]$	$\langle \mu_{YY}, g'g^T \rangle_{\mathcal{H}_{\ell\ell}}$ $= \mathbb{E}[g'(Y)g(Y)]$	$\langle \mu_{XY}, fg^T \rangle_{\mathcal{H}_{k\ell}}$ $= \mathbb{E}[f(X)g(Y)]$	$\langle \mu_{X\vert Y=y}, f \rangle_{\mathcal{H}_k}$ $= \mathbb{E}[f(X)\vert Y = y]$
Operator Definition	$C_{X\vert Y}C_{YY}$ $= C_{XY}$	$C_{YY} := \mu_{YY}$ $(C_{YY})^T = C_{YY}$	$C_{XY} := \mu_{XY}$ $(C_{XY})^T = C_{YX}$	$C_{X\vert Y}\ell(y, \cdot) :=$ $\mu_{X\vert Y=y}$
Encoded Expectation	$f^T C_{XY} =$ $g^T C_{YY}$	$\langle g', C_{YY}g \rangle_{\mathcal{H}_\ell}$ $= \mathbb{E}[g'(Y)g(Y)]$	$\langle f, C_{XY}g \rangle_{\mathcal{H}_k}$ $= \mathbb{E}[f(X)g(Y)]$	$(C_{X\vert Y})^T f = g :=$ $\mathbb{E}[f(X)\vert Y = \cdot]$

E. Parallels between Probabilistic Rules and Mean Operators for Section 6

Both the usual and nonparametric Bayes' rule are derived to reverse the relationship specified by the likelihood (density or operator, resp.) by matching the joint. In both cases, the prior (density or operator, resp.) is inevitably required to perform this computation.

Consider the derivation for Bayes' rule. When given a *forward* density $p(x|y)$ and a marginal density on its conditioned variable $p(y)$ which specifies a joint $p(x, y) = p(x|y)p(y)$, we seek a *backward* density $q(y|x)$ and a marginal density $q(x)$ that would yield the same joint $q(y|x)q(x) = p(x, y) = p(x|y)p(y)$. It is only when applying $\int_\mathcal{Y} \cdot dy$ on both sides, requiring that $q(y|x)$ is a density, that we have $q(x) = \int_\mathcal{Y} p(x|y)p(y)dy$ and thus Bayes' rule.

Similarly, when given a *forward* CMO $C_{X|Y} : \mathcal{H}_\ell \to \mathcal{H}_k$ and a symmetric operator $C_{YY} : \mathcal{H}_\ell \to \mathcal{H}_\ell$ on its conditioned variable which specifies a joint $C_{XY} = C_{X|Y}C_{YY}$, we seek a *backward* operator $D_{Y|X} : \mathcal{H}_k \to \mathcal{H}_\ell$ and a symmetric operator $D_{XX} : \mathcal{H}_k \to \mathcal{H}_k$ that would yield the same joint $D_{Y|X}D_{XX} = C_{YX} = (C_{XY})^T = (C_{X|Y}C_{YY})^T$. Without further requirement we see that $D_{Y|X} = C'_{X|Y}$ (4.2) and $D_{XX} = C'_{XX}$ is one solution. It is only when applying $C_{X|Y}$ on both sides, requiring the assumption of theorem 6.1, that we have $D_{Y|X} = C_{Y|X}$ and $D_{XX} = C_{XX}$ and thus a nonparametric Bayes' rule.

Importantly, it is only when DMOs and KBR are viewed as a statement for relationship between X and Y that they are seen as nonparametric versions of the Bayes' rule. However, DMOs and KBR are not Bayesian models with respect to the task of inferring deconditional mean or conditional means. This is because both models only infer point estimates for the deconditional or conditional mean, and no measure of uncertainty in the inferred function is provided.

Table E.1 review mean embeddings and their encoded expectations, providing probabilistic interpretations to RKHS embeddings and operators (Song et al., 2013).

Table F.1. Empirical estimators for DMO and KBR. We use the shorthand $A := (L + n\lambda I)^{-1}\tilde{L}$ and $D := \mathrm{diag}(A\mathbf{1})$.

Method	Joint Operator $\bar{C}_{XY}$	Evidence Operator $\bar{C}_{XX}$ or $\bar{C}'_{XX}$	Posterior Operator $\bar{C}_{Y\|X}$ or $\bar{C}'_{X\|Y}$	Computational Form $\bar{C}_{Y\|X}$ or $\bar{C}'_{X\|Y}$
DMO	$\hat{C}_{X\|Y}\tilde{C}_{YY}$	$\hat{C}_{X\|Y}\tilde{C}_{YY}(\hat{C}_{X\|Y})^T$	$(\bar{C}_{XY})^T(\bar{C}'_{XX} + \epsilon I)^{-1}$	$\tilde{\Psi}\left[A^T K A + m\epsilon I\right]^{-1} A^T \Phi^T$
DMO(W)	$\hat{C}_{X\|Y}\tilde{C}_{YY}$	$\hat{C}_{X\|Y}\tilde{C}_{YY}(\hat{C}_{X\|Y})^T$	$(\bar{C}_{XY})^T(\bar{C}'_{XX} + \epsilon I)^{-1}$	$\tilde{\Psi}A^T\left[KAA^T + m\epsilon I\right]^{-1}\Phi^T$
KBR(a)-I	$\hat{C}_{X\|Y}\tilde{C}_{YY}$	$\hat{C}_{XX\|Y}\tilde{\mu}_Y$	$(\bar{C}_{XY})^T(\bar{C}_{XX} + \epsilon I)^{-1}$	$\tilde{\Psi}A^T\left[KD + m\epsilon I\right]^{-1}\Phi^T$
KBR(a)-II	$\hat{C}_{X\|Y}\tilde{C}_{YY}$	$\hat{C}_{XX\|Y}\tilde{\mu}_Y$	$(\bar{C}_{XY})^T(\bar{C}^2_{XX} + \epsilon I)^{-1}\bar{C}_{XX}$	$\tilde{\Psi}A^T\left[(KD)^2 + m^2\epsilon I\right]^{-1} KD\Phi^T$
KBR(b)-I	$\hat{C}_{XY\|Y}\tilde{\mu}_Y$	$\hat{C}_{XX\|Y}\tilde{\mu}_Y$	$(\bar{C}_{XY})^T(\bar{C}_{XX} + \epsilon I)^{-1}$	$\Psi D\left[KD + m\epsilon I\right]^{-1}\Phi^T$
KBR(b)-II	$\hat{C}_{XY\|Y}\tilde{\mu}_Y$	$\hat{C}_{XX\|Y}\tilde{\mu}_Y$	$(\bar{C}_{XY})^T(\bar{C}^2_{XX} + \epsilon I)^{-1}\bar{C}_{XX}$	$\Psi D\left[(KD)^2 + m^2\epsilon I\right]^{-1} KD\Phi^T$

F. Connections between the Deconditional Mean Operator and Kernel Bayes' Rule for Section 7

Bayesian inference often requires computation of the posterior $\mathbb{P}_{Y|X}$ when given the likelihood $\mathbb{P}_{X|Y}$ and the prior $\mathbb{P}_Y$. When density evaluations exist, the Bayes' rule provides their relationship as $p_{Y|X}(\cdot|x) = \frac{p_{X|Y}(x|\cdot)p_Y(\cdot)}{\int_{\mathcal{Y}} p_{X|Y}(x|y)p_Y(y)dy}$.

Nevertheless, several levels of intractability may arise. The first is when both likelihood and prior density evaluations are tractable but the evidence integral $\int_{\mathcal{Y}} p_{X|Y}(x|y)p_Y(y)dy$ is intractable, leading to literatures such as variational inference (VI) (Blei et al., 2017) and Markov chain Monte Carlo (MCMC) (Hastings, 1970). The next is when only likelihood evaluations are intractable but sampling is possible, leading literatures such as LFI and approximate Bayesian computation (ABC) (Marin et al., 2012). More rarely, only prior evaluations are intractable but available via sampling, leading to literatures in implicit priors. The last is when both the likelihood and prior evaluations are intractable but available via sampling, leading to newer literatures such as implicit generative models.

While there are many approaches that addresses each of these scenarios, the underlying limitation is that Bayes' rule requires density evaluations that are difficult to approximate in high dimensions from samples. Instead, if relationships between the posterior, likelihood, and prior can be captured without using density evaluations, but directly by using samples, this issue could be more naturally sidestepped. Both DMOs and KBR provide such a nonparametric Bayes' rule.

Table F.1 compares all four forms of KBR (Song et al., 2013) with DMO. This table illustrates the different ways each method estimates the joint and evidence operators from likelihood and prior operators, the type of regularization used for inverting the evidence operator, and the final computational form. For KBR, (a) and (b) differ in the joint operator, and I and II differ in the type of regularization used for inverting the evidence operator. Via the Woodbury identity, for DMO we also show an alternative computational form DMO(W) that better illustrate its contrast with KBR(a)-I and KBR(b)-I. Note that unlike the four types of KBR, DMO(W) is the same model as DMO, just with a different computational form.

In particular, the diagonal matrix $D := \mathrm{diag}(A\mathbf{1})$ arises from the use of third order operators. This can make estimators sensitive to regularizations on inverse operators. This is best seen in the degenerate case of $\epsilon \to 0^+$, shown in table F.2, where for KBR(b) the effect of $\tilde{\mathbf{y}}$ vanishes, even though ϵ does not correspond to regularizations from the prior.

Furthermore, the original computational form of DMOs involves the inverse of a positive definite matrix. This however is not true for KBR(a) and KBR(b) since KD is not symmetric and thus the resulting matrix to be inverted cannot be positive definite. For KBR(b), by using $D = D^{\frac{1}{2}}D^{\frac{1}{2}}$ and the Woodbury identity, KBR(b)-I and KBR(b)-II can be written in forms with symmetric matrix inverses as $\bar{C}_{Y|X} = \Psi D^{\frac{1}{2}}\left[D^{\frac{1}{2}}KD^{\frac{1}{2}} + m\epsilon I\right]^{-1}D^{\frac{1}{2}}\Phi^T$ and $\bar{C}_{Y|X} = \Psi D^{\frac{1}{2}}\left[D^{\frac{1}{2}}KDKD^{\frac{1}{2}} + m^2\epsilon I\right]^{-1}D^{\frac{1}{2}}KD\Phi^T$ respectively. However, it is difficult to interpret this form.

Finally, similar to theorem 6.2, for the other degenerate case where $m = n$, $\tilde{\mathbf{y}} = \mathbf{y}$, and $\lambda \to 0^+$, all estimators revert to a CME $\hat{C}_{Y|X} = \Psi(K + n\epsilon I)^{-1}\Phi^T$.

Table F.2. Degenerate case for empirical estimators when $\epsilon = 0$

DMO(W)	KBR(a)	KBR(b)
$\tilde{\Psi}A^T\left[AA^T\right]^{-1}K^{-1}\Phi^T$	$\tilde{\Psi}A^T D^{-1}K^{-1}\Phi^T$	$\Psi K^{-1}\Phi^T$

G. Theorems and Experiment Details for Section 8

G.1. Hyperparameter Learning of Deconditional Kernel Mean Embeddings for Likelihood Free Inference

To learn hyperparameters, we maximize the following objective function which approximate the marginal likelihood of the inference problem.

Theorem G.1 (Approximate Marginal Likelihood for LFI)**.** *Assume $\kappa_\epsilon(\mathbf{y},\cdot) \in \mathcal{H}_k$ and that $\hat{C}_{\mathbf{X}|\mathbf{\Theta}}$ is a bounded operator for all n. Denote $\boldsymbol{\kappa}_\epsilon(\mathbf{y}) = \{\kappa_\epsilon(\mathbf{y},\mathbf{x}_i)\}_{i=1}^n$ and $\mathbf{1}_m = \{1\}_{j=1}^m$, then $\bar{q}(\mathbf{y}) := \langle \kappa_\epsilon(\mathbf{y},\cdot), \hat{C}_{\mathbf{X}|\mathbf{\Theta}}\tilde{\mu}_{\mathbf{\Theta}}\rangle_{\mathcal{H}_k} = \frac{1}{m}\boldsymbol{\kappa}_\epsilon^T A \mathbf{1}_m$ is an estimator to the marginal likelihood $p_\epsilon(\mathbf{y})$ and converge at $O_p(m^{-\frac{1}{2}} + (n\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}})$.*

Proof of Theorem G.1. Consider the absolute difference between $\bar{q}(\mathbf{y})$ and $p_\epsilon(\mathbf{y})$,

$$|\bar{q}(\mathbf{y}) - p_\epsilon(\mathbf{y})| \leq |\bar{q}(\mathbf{y}) - q(\mathbf{y})| + |q(\mathbf{y}) - p_\epsilon(\mathbf{y})|. \tag{G.1}$$

where $q(\mathbf{y}) := \langle \kappa_\epsilon(\mathbf{y},\cdot), \hat{C}_{\mathbf{X}|\mathbf{\Theta}}\mu_{\mathbf{\Theta}}\rangle_{\mathcal{H}_k} = \mathbb{E}[\langle \kappa_\epsilon(\mathbf{y},\cdot), \hat{C}_{\mathbf{X}|\mathbf{\Theta}}\ell(\mathbf{\Theta},\cdot)\rangle_{\mathcal{H}_k}]$. The first term is

$$\begin{aligned}
|\bar{q}(\mathbf{y}) - q(\mathbf{y})| &= |\langle \kappa_\epsilon(\mathbf{y},\cdot), \hat{C}_{\mathbf{X}|\mathbf{\Theta}}(\tilde{\mu}_{\mathbf{\Theta}} - \mu_{\mathbf{\Theta}})\rangle_{\mathcal{H}_k}| = |\langle (\hat{C}_{\mathbf{X}|\mathbf{\Theta}})^T\kappa_\epsilon(\mathbf{y},\cdot), (\tilde{\mu}_{\mathbf{\Theta}} - \mu_{\mathbf{\Theta}})\rangle_{\mathcal{H}_\ell}| \\
&\leq \|(\hat{C}_{\mathbf{X}|\mathbf{\Theta}})^T\kappa_\epsilon(\mathbf{y},\cdot)\|_{\mathcal{H}_\ell}\|(\tilde{\mu}_{\mathbf{\Theta}} - \mu_{\mathbf{\Theta}})\|_{\mathcal{H}_\ell} \\
&\leq c\|(\tilde{\mu}_{\mathbf{\Theta}} - \mu_{\mathbf{\Theta}})\|_{\mathcal{H}_\ell}.
\end{aligned} \tag{G.2}$$

for some constant c since $\hat{C}_{\mathbf{X}|\mathbf{\Theta}}$ is a bounded operator for all n. Hence, $|\bar{q}(\mathbf{y}) - q(\mathbf{y})|$ decays at $O(m^{-\frac{1}{2}})$.

For the second term, we have $p_\epsilon(\mathbf{y}) = \mathbb{E}[p_\epsilon(\mathbf{y}|\mathbf{\Theta})] = \mathbb{E}[\langle \kappa_\epsilon(\mathbf{y},\cdot), \mu_{\mathbf{X}|\mathbf{\Theta}=\mathbf{\Theta}}\rangle_{\mathcal{H}_k}] = \mathbb{E}[\langle \kappa_\epsilon(\mathbf{y},\cdot), C_{\mathbf{X}|\mathbf{\Theta}}\ell(\mathbf{\Theta},\cdot)\rangle_{\mathcal{H}_k}]$, similar to $q(\mathbf{y}) = \mathbb{E}[\langle \kappa_\epsilon(\mathbf{y},\cdot), \hat{C}_{\mathbf{X}|\mathbf{\Theta}}\ell(\mathbf{\Theta},\cdot)\rangle_{\mathcal{H}_k}]$. Since we use bounded kernels, define $\bar{\ell} := \sup_{\boldsymbol{\theta}} \|\ell(\boldsymbol{\theta},\cdot)\|_{\mathcal{H}_\ell}$ and $\bar{\kappa}_\epsilon := \sup_{\mathbf{y}} \|\kappa_\epsilon(\mathbf{y},\cdot)\|_{\mathcal{H}_k}$. The second term becomes

$$\begin{aligned}
|q(\mathbf{y}) - p_\epsilon(\mathbf{y})| &= |\mathbb{E}[\langle \kappa_\epsilon(\mathbf{y},\cdot), \hat{C}_{\mathbf{X}|\mathbf{\Theta}}\ell(\mathbf{\Theta},\cdot)\rangle_{\mathcal{H}_k}] - \mathbb{E}[\langle \kappa_\epsilon(\mathbf{y},\cdot), C_{\mathbf{X}|\mathbf{\Theta}}\ell(\mathbf{\Theta},\cdot)\rangle_{\mathcal{H}_k}]| \\
&\leq \mathbb{E}[|\langle \kappa_\epsilon(\mathbf{y},\cdot), \hat{C}_{\mathbf{X}|\mathbf{\Theta}}\ell(\mathbf{\Theta},\cdot)\rangle_{\mathcal{H}_k}] - \mathbb{E}[\langle \kappa_\epsilon(\mathbf{y},\cdot), C_{\mathbf{X}|\mathbf{\Theta}}\ell(\mathbf{\Theta},\cdot)\rangle_{\mathcal{H}_k}|] \\
&= \mathbb{E}[|\langle \kappa_\epsilon(\mathbf{y},\cdot), (\hat{C}_{\mathbf{X}|\mathbf{\Theta}} - C_{\mathbf{X}|\mathbf{\Theta}})\ell(\mathbf{\Theta},\cdot)\rangle_{\mathcal{H}_k}|] \\
&\leq \mathbb{E}[\|\kappa_\epsilon(\mathbf{y},\cdot)\|_{\mathcal{H}_k}\|(\hat{C}_{\mathbf{X}|\mathbf{\Theta}} - C_{\mathbf{X}|\mathbf{\Theta}})\ell(\mathbf{\Theta},\cdot)\|_{\mathcal{H}_k}] \\
&= \|\kappa_\epsilon(\mathbf{y},\cdot)\|_{\mathcal{H}_k}\mathbb{E}[\|(\hat{C}_{\mathbf{X}|\mathbf{\Theta}} - C_{\mathbf{X}|\mathbf{\Theta}})\ell(\mathbf{\Theta},\cdot)\|_{\mathcal{H}_k}] \\
&= \bar{\kappa}_\epsilon\mathbb{E}[\|(\hat{C}_{\mathbf{X}|\mathbf{\Theta}} - C_{\mathbf{X}|\mathbf{\Theta}})\ell(\mathbf{\Theta},\cdot)\|_{\mathcal{H}_k}] \\
&\leq \bar{\kappa}_\epsilon\mathbb{E}[\|\hat{C}_{\mathbf{X}|\mathbf{\Theta}} - C_{\mathbf{X}|\mathbf{\Theta}}\|_{HS}\|\ell(\mathbf{\Theta},\cdot)\|_{\mathcal{H}_\ell}] \\
&= \bar{\kappa}_\epsilon\mathbb{E}[\|\hat{C}_{\mathbf{X}|\mathbf{\Theta}} - C_{\mathbf{X}|\mathbf{\Theta}}\|_{HS}\sqrt{\ell(\mathbf{\Theta},\mathbf{\Theta})}] \\
&= \bar{\kappa}_\epsilon\mathbb{E}[\sqrt{\ell(\mathbf{\Theta},\mathbf{\Theta})}]\|\hat{C}_{\mathbf{X}|\mathbf{\Theta}} - C_{\mathbf{X}|\mathbf{\Theta}}\|_{HS} \\
&\leq \bar{\kappa}_\epsilon\mathbb{E}[\bar{\ell}]\|\hat{C}_{\mathbf{X}|\mathbf{\Theta}} - C_{\mathbf{X}|\mathbf{\Theta}}\|_{HS} \\
&= \bar{\kappa}_\epsilon\bar{\ell}\|\hat{C}_{\mathbf{X}|\mathbf{\Theta}} - C_{\mathbf{X}|\mathbf{\Theta}}\|_{HS}.
\end{aligned} \tag{G.3}$$

Hence, in the worst case $|q(\mathbf{y}) - p_\epsilon(\mathbf{y})|$ decays at the rate $\|\hat{C}_{\mathbf{X}|\mathbf{\Theta}} - C_{\mathbf{X}|\mathbf{\Theta}}\|_{HS}$ decays, which is $O_p((n\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}})$. Together with the first term, we have the claimed convergence rate.

Finally, the empirical form is obtained from substituting the empirical forms for the likelihood CMO and prior embedding, $\bar{q}(\mathbf{y}) := \langle \kappa_\epsilon(\mathbf{y},\cdot), \hat{C}_{\mathbf{X}|\mathbf{\Theta}}\tilde{\mu}_{\mathbf{\Theta}}\rangle_{\mathcal{H}_k} = \langle \kappa_\epsilon(\mathbf{y},\cdot), (\Phi(L + n\lambda I)^{-1}\Psi^T)(\frac{1}{m}\tilde{\Psi}\mathbf{1}_m)\rangle_{\mathcal{H}_k} = \frac{1}{m}\boldsymbol{\kappa}_\epsilon^T A\mathbf{1}_m$. □

To satisfy $\kappa_\epsilon(\mathbf{y},\cdot) \in \mathcal{H}_k$, we use $\kappa_\epsilon(\mathbf{y},\mathbf{x}) = p_\epsilon(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y};\mathbf{x},\epsilon^2 I)$ and Gaussian kernel for k with length scale ϵ, so that κ_ϵ is just the normalized version of the reproducing kernel k.

Importantly, while the approximate marginal likelihood $\bar{q}(\mathbf{y})$ depends on the hyperparameters of the kernels k and ℓ and the regularization λ, it does not depend on ϵ. At first, it seems that this objective cannot help us learn ϵ. Nevertheless, due to points (4) and (5) of theorem 5.3, we have that a good proxy for setting ϵ once λ is learned is $\epsilon = \frac{n}{m}\lambda$.

Nevertheless, for simplicity in our experiments we optimize all kernel hyperparameters and keep the regularization hyperparameters fixed, which has already achieved sufficiently accurate results.

G.2. Exponential-Gamma Experiment

The toy exponential-gamma problem is a standard benchmark for likelihood-free inference, where the true posterior $p_\epsilon(\boldsymbol{\theta}|\mathbf{y})$ is known and tractable even for $\epsilon = 0$. We follow the experimental setup of Meeds & Welling (2014).

Of the kernel based methods that we have benchmarked against, kernel ABC (K-ABC) (Nakagome et al., 2013), double kernel ABC (K2-ABC) (Park et al., 2016), KBR (Fukumizu et al., 2013), and kernel embedding likelihood-free inference (KELFI) (Hsu & Ramos, 2019) are also LFI methods based on the KME framework. Consequently, they are very suitable for comparisons towards DME. For all these methods, we apply kernel herding on their posterior embeddings to get posterior samples, and plot the approximate posterior density in fig. 5 (left) using kernel density estimation (KDE) on the posterior samples. In contrast, Gaussian process surrogate ABC (GPS-ABC) (Meeds & Welling, 2014) has its own adaptive MCMC based sampling algorithm. We set a simulation budget of 200 simulations and run it until either 10000 posterior samples are generated or the simulation budget is reached. For hyperparameters, we used standard median heuristic for K-ABC, K2-ABC, and KBR. In contrast, DME and KELFI have their own marginal likelihoods for hyperparameter learning. For both cases, we find global and local optimums of the marginal likelihood for the hyperparameters and show their results, emphasizing that maximizing the marginal likelihood objective produces better inference results. The hyperparameters of the GP surrogate itself used in GPS-ABC are learned by maximizing the marginal likelihood of the GPR (Rasmussen & Williams, 2006). However, for hyperparameters of GPS-ABC that are not part of the surrogate, we select them based on the original paper (Meeds & Welling, 2014). We then report its best two results.

G.3. Lotka-Volterra Experiment

The Lotka-Volterra simulator describes the population dynamics of a well known predator-prey system. For most parameters, the simulation produces chaotic behavior. Realistic scenarios with oscillatory behavior appears only for a small set of parameters. Consequently, inference on the Lotka-Volterra simulator is extremely challenging.

We follow the setup of Papamakarios & Murray (2016) and Tran et al. (2017). There are 4 parameters and 9 normalized summary statistics. We place the same uniform prior on the log parameters and use the same ground truth parameters. After performing inference on all four parameters, we similarly show in fig. 5 (right) the marginal posterior distribution for $\log \theta_1$ in the same format as Papamakarios & Murray (2016) and Tran et al. (2017).

For KBR (Fukumizu et al., 2013), KELFI (Hsu & Ramos, 2019), and DME, we again sample their posterior mean embeddings with kernel herding to get 10000 posterior samples. Finally, to compute the 95% interval, we compute the empirical 2.5% quantile and 97.5% quantiles on marginal samples of $\log \theta_1$ from the 10000 posterior samples. For mixture density network (MDN) (Papamakarios & Murray, 2016) and the two likelihood-free variational inference (LFVI) methods (Tran et al., 2017), we report the results from the original source, as well as their results for rejection ABC (REJ-ABC), Markov chain Monte carlo ABC (MCMC-ABC), and sequential Monte carlo ABC (SMC-ABC).