Bayesian Learning of Conditional Kernel Mean Embeddings for Automatic Likelihood-Free Inference

Kelvin Hsu[1 2] **Fabio Ramos**[1 3]
[1]School of Computer Science, The University of Sydney, Australia. [2]CSIRO, Australia. [3]NVIDIA, USA.

Abstract

In likelihood-free settings where likelihood evaluations are intractable, approximate Bayesian computation (ABC) addresses the formidable inference task to discover plausible parameters of simulation programs that explain the observations. However, they demand large quantities of simulation calls. Critically, hyperparameters that determine measures of simulation discrepancy crucially balance inference accuracy and sample efficiency, yet are difficult to tune. In this paper, we present kernel embedding likelihood-free inference (KELFI), a holistic framework that automatically learns model hyperparameters to improve inference accuracy given limited simulation budget. By leveraging likelihood smoothness with conditional mean embeddings, we nonparametrically approximate likelihoods and posteriors as surrogate densities and sample from closed-form posterior mean embeddings, whose hyperparameters are learned under its approximate marginal likelihood. Our modular framework demonstrates improved accuracy and efficiency on challenging inference problems in ecology.

1 Introduction

Scientific understanding of complex phenomena are deeply reliant on the study of probabilistic generative models and their match with real world data. Often, latent and convoluted interactions result in intractable likelihood evaluations, making the setting *likelihood-free*. Instead, generative models are expressed as a stochastic forward model simulator. Inference on latent variables in this setting is particularly challenging.

Proceedings of the 22[nd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

Approximate Bayesian computation (ABC) methods are the state-of-the-art in simulation-based Bayesian inference with intractable likelihoods (Marin et al., 2012). They infer posterior distributions of simulator parameters that aim to explain observed data. The posterior is of interest in its own right for understanding the complex phenomena, and also useful in forming predictions of future observations. They are popular due to their simplicity and applicability, and have been used extensively in the biological sciences (Beaumont, 2010; Toni et al., 2009). Nevertheless, complex models are often prohibitively expensive to simulate. Evolutionary processes of ecological systems, vibrational modes of a mechanical structure, and fluid flow across surfaces are all examples that result in formidable inference problems with demanding forward simulations. It is thus imperative for inference algorithms to perform under the constraint of limited simulation calls, posing an exceptionally challenging task.

Often, ABC methods rely on discrepancy measures between simulations and observations that are parametrized by hyperparameters such as ϵ. The resulting posterior approximation is highly sensitive to the choice of hyperparameters, yet appropriate hyperparameter tuning strategies remain to be established.

To address these issues, we present kernel embedding likelihood-free inference (KELFI), a holistic framework consisting of (1) a consistent surrogate likelihood *model* that modularizes queries from simulation calls, (2) a Bayesian *learning* objective for hyperparameters that improves inference accuracy, and (3) a posterior surrogate density and a super-sampling *inference* algorithm using its closed-form posterior mean embedding.

KELFI is based on approximating likelihoods with simulation samples using conditional mean embeddings (CMEs). CMEs encode conditional expectations empirically by leveraging smoothness within a reproducing kernel Hilbert space (RKHS) with only a small number of examples. This modularizes inference away from simulation calls. Consequently, scientists can proceed with posterior analysis after any number of simulations. Furthermore, KELFI infers both ap-

proximate posterior densities and samples. Critically, our learning algorithm tunes hyperparameters directly for the inference problem, including adapting ϵ to the number of simulations used. This removes the need for practitioners to arduously select hyperparameters. Finally, it can be extended to automatically learn the relevance and usefulness of each summary statistic.

2 Likelihood-Free Inference

In the likelihood-free setting, we begin with a stochastic forward model simulator which synthesizes simulations $\mathbf{x}$ given a parameter setting $\boldsymbol{\theta}$. Let $\boldsymbol{\theta} \in \vartheta$ denote a realization of the latent variable or parameter $\boldsymbol{\Theta}$, where we use upper cases to denote random variables. Let $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$ where $\mathcal{X}, \mathcal{Y} \subseteq \mathcal{D}$ denote realizations of simulation output $\mathbf{X}$ and observations $\mathbf{Y}$ respectively. We represent the simulator as $p(\mathbf{x}|\boldsymbol{\theta})$ from which we can only simulate or sample, but not query its density, making likelihood evaluations intractable, thus *likelihood-free*. To begin inference we posit a prior density $p(\boldsymbol{\theta})$ that encodes prior knowledge about plausible parameter settings to guide the inference. The goal is to infer a posterior distribution on the parameters $\boldsymbol{\theta}$ that could generate simulations $\mathbf{x}$ similar to our observations $\mathbf{y}$ by some comparison measure. This measure could be done by a standard ϵ-kernel or ABC kernel $p_\epsilon(\mathbf{y}|\mathbf{x}) = \kappa_\epsilon(\mathbf{y}, \mathbf{x})$, such as a Gaussian density $\mathcal{N}(\mathbf{y}|\mathbf{x}, \epsilon^2 I)$ (Price et al., 2017; Moreno et al., 2016).

Based on this formulation, the true full likelihood of our model can be written as follows,

$$p_\epsilon(\mathbf{y}|\boldsymbol{\theta}) = \int_{\mathcal{X}} p_\epsilon(\mathbf{y}|\mathbf{x}) p(\mathbf{x}|\boldsymbol{\theta}) d\mathbf{x} = \mathbb{E}[\kappa_\epsilon(\mathbf{y}, \mathbf{X})|\boldsymbol{\Theta} = \boldsymbol{\theta}]. \tag{2.1}$$

The corresponding posterior of interest is $p_\epsilon(\boldsymbol{\theta}|\mathbf{y}) = p_\epsilon(\mathbf{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) / p_\epsilon(\mathbf{y})$ where $p_\epsilon(\mathbf{y}) = \int_\vartheta p_\epsilon(\mathbf{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}$. Due to the presence of a non-zero ϵ, even a perfect approximation to the soft posterior $p_\epsilon(\boldsymbol{\theta}|\mathbf{y})$ will not be the exact posterior $p_{\epsilon=0}(\boldsymbol{\theta}|\mathbf{y})$ unless ϵ is annealed to zero. This is the necessary trade-off we make with limited simulations, where a non-zero ϵ is essential for tractable inference because no simulations will match the observations exactly in practice. If $\mathbf{y}$ is only available as a summary statistic, then this soft posterior $p_\epsilon(\boldsymbol{\theta}|\mathbf{y})$ that we are targeting is only an approximation to the posterior given the full data even with $\epsilon = 0$.

So far, the notation $\mathbf{x}$ and $\mathbf{y}$ denote either the full dataset or their summary statistics. This is because the summary operation can be appended to the simulator program to output summary statistics directly. In either case, we let the target posterior be $p_\epsilon(\boldsymbol{\theta}|\mathbf{y})$, so the inference problem remains structurally identical. For simplicity however, from here on $\mathbf{x}$ and $\mathbf{y}$ will denote summary statistics unless stated otherwise.

Since $p(\mathbf{x}|\boldsymbol{\theta})$ is intractable, so is the likelihood (2.1). Instead, approximations are required. Markov chain Monte Carlo (MCMC) ABC use empirical means, $p_\epsilon(\mathbf{y}|\boldsymbol{\theta}) \approx \frac{1}{S}\sum_{s=1}^S p_\epsilon(\mathbf{y}|\mathbf{x}^{(s)}) = \frac{1}{S}\sum_{s=1}^S \kappa_\epsilon(\mathbf{y}, \mathbf{x}^{(s)})$ (Andrieu et al., 2009). Synthetic likelihood ABC (SL-ABC) and adaptive SL-ABC (ASL-ABC) alternatively use Gaussian approximations $p_\epsilon(\mathbf{y}|\boldsymbol{\theta}) \approx \mathcal{N}(\mathbf{y}|\boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}} + \epsilon^2 I)$ and estimate the mean and covariance from simulations (Wood, 2010). These approaches require generating S new simulations $\{\mathbf{x}^{(s)}\}_{s=1}^S$ corresponding to $\boldsymbol{\theta}$ every time the likelihood is queried.

Not only are synthetic likelihoods parametric Gaussian approximations, they also approximate separately at each $\boldsymbol{\theta}$. Instead, surrogate likelihood approaches like KELFI use consistent nonparametric approximations so that (1) only one new simulation is required at each new parameter $\boldsymbol{\theta}$ and (2) likelihood queries do not need to be at parameters where simulations are available.

3 Kernel Embedding Likelihood-Free Inference

We present KELFI in three stages. In the model stage, we build a surrogate likelihood model by leveraging smoothness properties of CMEs. In the learning stage, we derive a differentiable marginal surrogate likelihood to drive hyperparameters learning. In the inference stage, we propose an algorithm to sample from the resulting mean embedding of the surrogate posterior.

When the prior is an anisotropic Gaussian $p(\boldsymbol{\theta}) = \prod_{d=1}^D \mathcal{N}(\theta_d|\mu_d, \sigma_d^2)$, closed form solutions for KELFI exists. We will present this setting since, for many common continuous priors, the likelihood-free inference (LFI) problem can be transformed into an equivalent problem that involves a Gaussian prior. See appendix F for more detail. When this is not possible or preferred, KELFI can be approximated arbitrarily well by using arbitrarily many prior samples.

3.1 Conditional Mean Embeddings

We begin with an overview of CMEs in the context of KELFI. Kernel mean embeddings (KMEs) are an arsenal of techniques used to represent distributions in a RKHS (Muandet et al., 2017). The key object is the mean embedding of a distribution $X \sim \mathbb{P}$ under a positive definite kernel k via $\mu_X := \int_{\mathcal{X}} k(x, \cdot) d\mathbb{P}(x) = \int_{\mathcal{X}} k(x, \cdot) p(x) dx \in \mathcal{H}_k$, where the last equality assumes a density p for $\mathbb{P}$ exists and $\mathcal{H}_k$ denotes the RKHS of k. They encode distributions in the sense that function expectations can be written as $\mathbb{E}[f(X)] = \langle \mu_X, f \rangle_{\mathcal{H}_k}$ if $f \in \mathcal{H}_k$. When μ_X can only be estimated empirically in some form denoted as $\hat{\mu}_X$, the expectation can be approximated by $\mathbb{E}[f(X)] \approx \langle \hat{\mu}_X, f \rangle_{\mathcal{H}_k}$.

CMEs are KMEs that encode conditional distributions. We specifically focus on their empirical estimates as we assume we only have the resource to obtain m sets of simulation data due to budget constraints. This results in joint samples $\{\boldsymbol{\theta}_j, \mathbf{x}_j\}_{j=1}^m$ from $p(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$ by sampling from a proposal prior π for $\boldsymbol{\theta}_j \sim \pi(\boldsymbol{\theta})$ and simulating $\mathbf{x}_j \sim p(\mathbf{x}|\boldsymbol{\theta}_j)$ at each $\boldsymbol{\theta}_j$. Note these samples are not necessarily from the original joint distribution $p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ if $\pi \neq p_{\boldsymbol{\Theta}}$.

We define positive definite and characteristic kernels (Sriperumbudur et al., 2010) $k : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ and $\ell : \vartheta \times \vartheta \to \mathbb{R}$. When relevant, we denote the hyperparameters of k and ℓ with α and β, and refer to them as $k_\alpha = k(\cdot, \cdot; \alpha)$ and $\ell_\beta = \ell(\cdot, \cdot; \beta)$. An useful example of such a kernel is an anisotropic Gaussian kernel $\ell(\boldsymbol{\theta}, \boldsymbol{\theta}'; \boldsymbol{\beta}) = \exp\big(-\frac{1}{2}\sum_{d=1}^D (\theta_d - \theta'_d)^2/\beta_d^2\big)$ whose hyperparameters are length scales $\boldsymbol{\beta} = \{\beta_d\}_{d=1}^D$ for each dimension $d \in [D] := \{1, \ldots, D\}$, and similarly for k.

For any function $f \in \mathcal{H}_k$, we construct an approximation to $\mathbb{E}[f(\mathbf{X})|\boldsymbol{\Theta} = \boldsymbol{\theta}]$ by the inner product $\langle f, \hat{\mu}_{\mathbf{X}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\rangle_{\mathcal{H}_k}$ with an empirical CME $\hat{\mu}_{\mathbf{X}|\boldsymbol{\Theta}=\boldsymbol{\theta}}$. Importantly, $\hat{\mu}_{\mathbf{X}|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ is estimated from the *joint* samples $\{\boldsymbol{\theta}_j, \mathbf{x}_j\}_{j=1}^m$, even though it is encoding the corresponding conditional distribution $p(\mathbf{x}|\boldsymbol{\theta})$. This approximation admits the following form (Song et al., 2009),

$$\mathbb{E}[f(\mathbf{X})|\boldsymbol{\Theta} = \boldsymbol{\theta}] \approx \mathbf{f}^T (L + m\lambda I)^{-1} \boldsymbol{\ell}(\boldsymbol{\theta}), \tag{3.1}$$

where $\mathbf{f} := \{f(\mathbf{x}_j)\}_{j=1}^m$, $L := \{\ell(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)\}_{i,j=1}^m$, $\boldsymbol{\ell}(\boldsymbol{\theta}) := \{\ell(\boldsymbol{\theta}_j, \boldsymbol{\theta})\}_{j=1}^m$, and $\lambda \geq 0$ is a regularization parameter. This approximation is known to converge at $O_p(m^{-\frac{1}{4}})$ if λ is chosen to decay at $O_p(m^{-\frac{1}{2}})$ or better under appropriate assumptions on $p(\mathbf{x}|\boldsymbol{\theta})$ (Song et al., 2013).

3.2 Model: Kernel Means Likelihood

We begin by presenting our surrogate likelihood model. Since the likelihood (2.1) is an expectation under $p(\mathbf{x}|\boldsymbol{\theta})$, we propose to approximate it via an inner product with the CME of $p(\mathbf{x}|\boldsymbol{\theta})$. Specifically, if we choose k such that $\kappa_\epsilon(\mathbf{y}, \cdot) \in \mathcal{H}_k$, then $p_\epsilon(\mathbf{y}|\boldsymbol{\theta})$ can be approximated by $q(\mathbf{y}|\boldsymbol{\theta}) := \langle \kappa_\epsilon(\mathbf{y}, \cdot), \hat{\mu}_{\mathbf{X}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\rangle_{\mathcal{H}_k}$. We refer to $q(\mathbf{y}|\boldsymbol{\theta})$ as the kernel means likelihood (KML). While the KML provides an asymptotically correct likelihood surrogate, for finitely many simulations it is not necessarily positive nor normalized. By using $f = \kappa_\epsilon(\mathbf{y}, \cdot)$ in (3.1) where $\boldsymbol{\kappa}_\epsilon(\mathbf{y}) := \{\kappa_\epsilon(\mathbf{y}, \mathbf{x}_j)\}_{j=1}^m$ and $\mathbf{v}(\mathbf{y}) := (L + m\lambda I)^{-1}\boldsymbol{\kappa}_\epsilon(\mathbf{y})$, the KML becomes

$$q(\mathbf{y}|\boldsymbol{\theta}) = \sum_{j=1}^m v_j(\mathbf{y})\ell(\boldsymbol{\theta}_j, \boldsymbol{\theta}). \tag{3.2}$$

The KML converges at the same rate as the CME. See theorem A.3 for proof. It is worthwhile to note that the assumption $\ell(\boldsymbol{\theta}, \cdot) \in \text{image}(C_{\boldsymbol{\Theta}\boldsymbol{\Theta}})$ is common for CMEs, and is not as restrictive as it may first appear, as it can be relaxed through introducing the regularization hyperparameter λ (Song et al., 2013).

Theorem 3.1. *Assume $\ell(\boldsymbol{\theta}, \cdot) \in \text{image}(C_{\boldsymbol{\Theta}\boldsymbol{\Theta}})$. The kernel means likelihood (KML) $q(\mathbf{y}|\boldsymbol{\theta})$ converges to the likelihood $p_\epsilon(\mathbf{y}|\boldsymbol{\theta})$ uniformly at rate $O_p((m\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}})$ as a function of $\boldsymbol{\theta} \in \vartheta$ and $\mathbf{y} \in \mathcal{Y}$.*

To satisfy $\kappa_\epsilon(\mathbf{y}, \cdot) \in \mathcal{H}_k$, we choose the standard Gaussian ϵ-kernel $\kappa_\epsilon(\mathbf{y}, \mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{x}, \epsilon^2 I)$ and let $k_\alpha = k_\epsilon$ be a Gaussian kernel with length scale $\alpha = \epsilon$. Since $\kappa_\epsilon(\mathbf{y}, \mathbf{x})$ and $k_\epsilon(\mathbf{y}, \mathbf{x})$ are scalar multiples of each other, we have that $\kappa_\epsilon(\mathbf{y}, \cdot) \in \mathcal{H}_k$. In fact, any positive definite kernel κ_ϵ can be used, since we can simply choose k_α to be its scalar multiple to form the RKHS.

When the raw data is *iid* and no sufficient summary statistics are available, we can employ a kernel on the empirical distributions of the two datasets via $\kappa_{\epsilon,\alpha}(\mathbf{y}, \mathbf{x}) \propto k_{\epsilon,\alpha}(\mathbf{y}, \mathbf{x}) = \exp\big(-\frac{1}{2\epsilon^2}\|\hat{\mu}_{\mathbf{Y}} - \hat{\mu}_{\mathbf{X}}\|_{\mathcal{H}_k}^2\big)$, where $\hat{\mu}_{\mathbf{Y}} = \frac{1}{n}\sum_{i=1}^n \bar{k}_\alpha(\mathbf{y}_i, \cdot)$, $\hat{\mu}_{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^n \bar{k}_\alpha(\mathbf{x}_i, \cdot)$ are empirical mean embeddings of the observed and simulated raw data. Here $\bar{k}$ is another kernel with hyperparameters α. This was also used in double kernel ABC (K2-ABC) (Park et al., 2016) and distribution regression ABC (DR-ABC) (Mitrovic et al., 2016) to remove the requirement of summary statistics.

3.3 Learning: Hyperparameter Learning with Marginal Kernel Means Likelihood

We now propose a hyperparameter learning algorithm for our surrogate likelihood model. The main advantage of using an approximate surrogate likelihood surrogate model is that it readily provides a marginal surrogate likelihood quantity that lends itself to a hyperparameter learning algorithm. We define the marginal kernel means likelihood (MKML) as follows,

$$q(\mathbf{y}) := \int_\vartheta q(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} = \sum_{j=1}^m v_j(\mathbf{y})\mu_{\boldsymbol{\Theta}}(\boldsymbol{\theta}_j), \tag{3.3}$$

where $\mu_{\boldsymbol{\Theta}} := \int_\vartheta \ell(\boldsymbol{\theta}, \cdot)p(\boldsymbol{\theta})d\boldsymbol{\theta}$ is the mean embedding of $p_{\boldsymbol{\Theta}}$. If we choose ℓ to be an anisotropic Gaussian kernel with length scales $\boldsymbol{\beta} = \{\beta_d\}_{d=1}^D$, then $\mu_{\boldsymbol{\Theta}}$ is closed-form for anisotropic Gaussian priors $p(\boldsymbol{\theta}) = \prod_{d=1}^D \mathcal{N}(\theta_d|\mu_d, \sigma_d^2)$. Let $\nu_d^2 := \beta_d^2 + \sigma_d^2$, then we have

$$\mu_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) = \ell_{\boldsymbol{\nu}}(\boldsymbol{\theta}, \boldsymbol{\mu}) \prod_{d=1}^D \frac{\beta_d}{\nu_d}. \tag{3.4}$$

Similar to the KML, the MKML converges at the same rate as the CME. See theorem A.4 for proof.

Theorem 3.2. *Assume $\ell(\boldsymbol{\theta}, \cdot) \in \text{image}(C_{\boldsymbol{\Theta}\boldsymbol{\Theta}})$. The marginal kernel means likelihood (MKML) $q(\mathbf{y})$ converges to marginal likelihood $p_\epsilon(\mathbf{y})$ uniformly at rate $O_p((m\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}})$ as a function of $\mathbf{y} \in \mathcal{Y}$.*

Algorithm 3.1 KELFI: Kernel Embedding Likelihood-Free Inference

1: **Input:** Data $\mathbf{y}$, simulations $\{\boldsymbol{\theta}_j, \mathbf{x}_j\}_{j=1}^m \sim p(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$, query parameters $\{\boldsymbol{\theta}_r^\star\}_{r=1}^R$, KML hyperparameters $(\boldsymbol{\epsilon}, \boldsymbol{\beta}, \lambda)$, prior hyperparameters $(\boldsymbol{\mu}, \boldsymbol{\sigma})$ or samples $\{\tilde{\boldsymbol{\theta}}_t\}_{t=1}^T$, number of samples S, kernel ℓ and ϵ-kernel κ
2: Compute $\mathbf{v} \leftarrow (L + m\lambda I)^{-1}\boldsymbol{\kappa}_{\boldsymbol{\epsilon}}(\mathbf{y})$ where $L \leftarrow \{\ell_{\boldsymbol{\beta}}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)\}_{i,j=1}^m$ and $\boldsymbol{\kappa}_{\boldsymbol{\epsilon}}(\mathbf{y}) \leftarrow \{\kappa_{\boldsymbol{\epsilon}}(\mathbf{y}, \mathbf{x}_j)\}_{j=1}^m$
3: Compute $q(\mathbf{y}) \leftarrow \mathbf{v}^T\boldsymbol{\mu}_{\boldsymbol{\Theta}}$ where $\boldsymbol{\mu}_{\boldsymbol{\Theta}} \leftarrow \{\mu_{\boldsymbol{\Theta}}(\boldsymbol{\theta}_j)\}_{j=1}^m$ using (3.4) or $\boldsymbol{\mu}_{\boldsymbol{\Theta}} \leftarrow \frac{1}{T}\tilde{L}\mathbf{1}_T$ where $\tilde{L} \leftarrow \{\ell_{\boldsymbol{\beta}}(\boldsymbol{\theta}_j, \tilde{\boldsymbol{\theta}}_t)\}_{j,t=1}^{m,T}$
4: Compute $H \leftarrow \{h(\boldsymbol{\theta}_j, \boldsymbol{\theta}_r^\star)\}_{j=1,r=1}^{m,R}$ using (3.6) or $H \leftarrow \frac{1}{T}\tilde{L}\tilde{L}^\star$ where $\tilde{L}^\star = \{\ell_{\boldsymbol{\beta}}(\tilde{\boldsymbol{\theta}}_t, \boldsymbol{\theta}_r^\star)\}_{t,r=1}^{T,R}$
5: Compute posterior mean embedding $\boldsymbol{\mu} \leftarrow H^T\mathbf{v}/q(\mathbf{y}) \in \mathbb{R}^R$ and initialize $\mathbf{a} \leftarrow \mathbf{0} \in \mathbb{R}^R$
6: **for** $s \in \{1, \ldots, S\}$ **do**
7: Obtain super-sample $\hat{\boldsymbol{\theta}}_s \leftarrow \tilde{\boldsymbol{\theta}}_{r^\star}$ where $r^\star \leftarrow \operatorname{argmax}_{r\in\{1,\ldots,R\}} \mu_r - (a_r/s)$
8: Update kernel sum $\mathbf{a} \leftarrow \mathbf{a} + \{\ell_{\boldsymbol{\beta}}(\boldsymbol{\theta}_r^\star, \hat{\boldsymbol{\theta}}_s)\}_{r=1}^R$
9: **end for**
10: **Output:** Posterior super-samples $\{\hat{\boldsymbol{\theta}}_s\}_{s=1}^S$

Consequently, the MKML $q(\mathbf{y}) = q(\mathbf{y}; \boldsymbol{\epsilon}, \boldsymbol{\beta}, \lambda)$ approximates the true marginal likelihood $p_\epsilon(\mathbf{y})$ of the inference problem defined by our likelihood-prior pair. It is a function of the hyperparameters $(\boldsymbol{\epsilon}, \boldsymbol{\beta}, \lambda)$ of the ϵ-kernel and KML model. As $p_\epsilon(\mathbf{y})$ is unavailable, we instead maximize the MKML for hyperparameter learning. Furthermore, prior hyperparameters $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ can also be included and learned jointly. Since the map $(\boldsymbol{\epsilon}, \boldsymbol{\beta}, \lambda) \mapsto q(\mathbf{y}; \boldsymbol{\epsilon}, \boldsymbol{\beta}, \lambda)$ is differentiable, optimization can be done in an auto-differentiation environment. The learning objective to be optimized is computed in line 3 of algorithm 3.1. Each automatic gradient update has complexity dominated by $O(m^3)$ due to the Cholesky decomposition in line 2. However, since we are addressing scenarios where simulations are limited so that m is small, this optimization is relatively fast.

Importantly, if we use an anisotropic Gaussian density for the ϵ-kernel $\kappa_{\boldsymbol{\epsilon}}$ where $\boldsymbol{\epsilon} = \{\epsilon_i\}_{i=1}^n$ are the length scales corresponding to each summary statistic $\mathbf{y} = \{y_i\}_{i=1}^n$, we can perform automatic relevance determination (ARD) to learn the relevance and usefulness of each summary statistic, where a small length scale indicate high relevance for that statistic. This is because $\boldsymbol{\epsilon}$ are also the length scales of the kernel k which defines the RKHS $\mathcal{H}_k$. Since the anistropic Gaussian kernel is learned, we also refer to it as an ARD kernel. We can also learn the length scales $\boldsymbol{\beta} = \{\beta_d\}_{d=1}^D$ for the kernel $\ell_{\boldsymbol{\beta}}$ on $\boldsymbol{\theta}$, although we found that it is more useful to let $\boldsymbol{\beta} = \beta_0\boldsymbol{\sigma}$ where $\boldsymbol{\sigma} = \{\sigma_d\}_{d=1}^D$ are the standard deviations of the Gaussian prior. By doing this, we make better use of the scale differences within $\boldsymbol{\theta}$ from the prior, and let β_0 learn the overall scale that is most useful for the KML.

For general non-Gaussian kernels and priors, $\mu_{\boldsymbol{\Theta}}$ in (3.3) can be approximated using T independent prior samples $\tilde{\boldsymbol{\theta}}_t \sim p(\boldsymbol{\theta})$, $t \in [T]$, as $\tilde{\mu}_{\boldsymbol{\Theta}} = \frac{1}{T}\sum_{t=1}^T \ell(\tilde{\boldsymbol{\theta}}_t, \cdot)$.

By formulating a learning objective directly for the inference problem, KELFI provides a way to automatically tune ϵ and its own model hyperparameters.

3.4 Inference: Kernel Means Posterior and Posterior Embedding Super-Sampling

We finally present an approach for posterior inference by super-sampling directly from the equivalent posterior mean embedding defined by the KML model and the prior. Our approach begins by defining a surrogate density to approximate the posterior $p_\epsilon(\boldsymbol{\theta}|\mathbf{y})$ in analogy to the Bayes' rule, $q(\boldsymbol{\theta}|\mathbf{y}) := q(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})/q(\mathbf{y})$. We refer to $q(\boldsymbol{\theta}|\mathbf{y})$ as the kernel means posterior (KMP). Importantly, $q(\boldsymbol{\theta}|\mathbf{y})$ is unaffected even if κ_ϵ is unnormalized, so that ϵ-kernels on distributions can be readily used. The KMP has the following convergence properties. See theorem A.5 for proof.

Theorem 3.3. *Assume $\ell(\boldsymbol{\theta}, \cdot) \in \operatorname{image}(C_{\boldsymbol{\Theta}\boldsymbol{\Theta}})$ and that there exists $\delta > 0$ such that $q(\mathbf{y}) \geq \delta$ for all $m \geq M$ where $M \in \mathbb{N}_+$. The kernel means posterior (KMP) $q(\boldsymbol{\theta}|\mathbf{y})$ converges pointwise to the posterior $p_\epsilon(\boldsymbol{\theta}|\mathbf{y})$ at rate $O_p((m\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}})$ as a function of $\boldsymbol{\theta} \in \vartheta$ and $\mathbf{y} \in \mathcal{Y}$. If $\sup_{\boldsymbol{\theta}\in\vartheta} p(\boldsymbol{\theta}) < \infty$ and $\sup_{\boldsymbol{\theta}\in\vartheta} p_\epsilon(\mathbf{y}|\boldsymbol{\theta}) < \infty$, then the convergence is uniform in $\boldsymbol{\theta} \in \vartheta$. If $\sup_{\mathbf{y}\in\mathcal{Y}} p_\epsilon(\boldsymbol{\theta}|\mathbf{y}) < \infty$, then the convergence is uniform in $\mathbf{y} \in \mathcal{Y}$.*

Importantly, the requirement for a $\delta > 0$ such that $q(\mathbf{y}) \geq \delta$ for all $m \geq M$ where $M \in \mathbb{N}_+$ provides an intuition for why high MKML values are favorable for learning a good approximate posterior. This requirement is an reflection on the capability of the simulator to recreate the observations $\mathbf{y}$ relative to the scale ϵ. Intuitively, the more capable the simulator $p(\mathbf{x}|\boldsymbol{\theta})$ is at generating simulations $\mathbf{x}$ that is close to $\mathbf{y}$ with respect to ϵ, the higher $p_\epsilon(\mathbf{y}) > 0$ will be relatively. Since theorem 3.2 guarantees that, for large $m > M$, $q(\mathbf{y})$ will be close to $p_\epsilon(\mathbf{y})$, we have that $q(\mathbf{y}) > 0$ for all large $m > M$ with increasing probability. In this situation, theorem 3.3 guarantees that the KMP will converge to the posterior of interest. However, consider the case when the simulator is ill-designed to recreate $\mathbf{y}$ such that the true marginal likelihood $p_\epsilon(\mathbf{y}) \approx 0$ is small. As $q(\mathbf{y})$ tends to $p_\epsilon(\mathbf{y}) \approx 0$ due to theorem 3.2, it may

struggle to always stay strictly positive even for large $m > M$ since it is stochastically converging to approximately zero. In this case, convergence is difficult since the simulator was ill-designed. However, by learning ϵ through maximizing $q(\mathbf{y})$, we adapt the threshold ϵ to make $p_\epsilon(\mathbf{y})$ as high as possible, leading to a more stable posterior $p_\epsilon(\boldsymbol{\theta}|\mathbf{y})$ for the KMP to converge to.

We now define kernel means posterior embedding (KMPE), the mean embedding of the KMP, as $\tilde{\mu}_{\boldsymbol{\Theta}|\mathbf{Y}=\mathbf{y}}(\boldsymbol{\theta}^\star) := \int_\vartheta \ell(\boldsymbol{\theta}, \boldsymbol{\theta}^\star) q(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}$. This becomes

$$\tilde{\mu}_{\boldsymbol{\Theta}|\mathbf{Y}=\mathbf{y}}(\boldsymbol{\theta}^\star) = \frac{1}{q(\mathbf{y})} \sum_{j=1}^{m} v_j(\mathbf{y}) h(\boldsymbol{\theta}_j, \boldsymbol{\theta}^\star), \tag{3.5}$$

where $h(\boldsymbol{\theta}, \boldsymbol{\theta}^\star) := \int_\vartheta \ell(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) \ell(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^\star) p(\tilde{\boldsymbol{\theta}}) d\tilde{\boldsymbol{\theta}}$. Importantly, since the KMPE is constructed from the CME used to form the KML, it converges in RKHS norm at the same rate. See theorem A.6 for proof.

Theorem 3.4. *Assume $\ell(\boldsymbol{\theta}, \cdot) \in \text{image}(C_{\boldsymbol{\Theta}\boldsymbol{\Theta}})$ and that there exists $\delta > 0$ such that $q(\mathbf{y}) \geq \delta$ for all $m \geq M$ where $M \in \mathbb{N}_+$. The kernel means posterior embedding (KMPE) $\tilde{\mu}_{\boldsymbol{\Theta}|\mathbf{Y}=\mathbf{y}}$ converges in RKHS norm to the posterior mean embedding $\mu_{\boldsymbol{\Theta}|\mathbf{Y}=\mathbf{y}}$ at rate $O_p((m\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}})$.*

If we choose ℓ to be an anisotropic Gaussian kernel with length scales $\boldsymbol{\beta} = \{\beta_d\}_{d=1}^D$, h exhibits the following closed-form under anisotropic Gaussian priors,

$$h(\boldsymbol{\theta}, \boldsymbol{\theta}^\star) = \prod_{d=1}^{D} \frac{s_d}{\sigma_d} \exp\Big[-\frac{1}{2s_d^2}\Big(a_d - b_d^2\Big)\Big], \tag{3.6}$$

where $a_d := (\theta_d^2 + {\theta_d^\star}^2 + \gamma_d^2 \mu_d^2)/(2+\gamma_d^2)$, $b_d := (\theta_d + \theta_d^\star + \gamma_d^2 \mu_d)/(2+\gamma_d^2)$, $\gamma_d^2 := \beta_d^2/\sigma_d^2$ and $s_d^{-2} := 2\beta_d^{-2} + \sigma_d^{-2}$. For general non-Gaussian kernels and priors, h can be approximated as $\tilde{h}(\boldsymbol{\theta}, \boldsymbol{\theta}^\star) = \frac{1}{T} \sum_{t=1}^{T} \ell(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}_t) \ell(\tilde{\boldsymbol{\theta}}_t, \boldsymbol{\theta}^\star)$.

The KMP $q(\cdot|\mathbf{y})$ is bounded and normalized but potentially non-positive. Consequently, it can be seen as a surrogate density corresponding to a signed measure. This suggests that the map $q(\cdot|\mathbf{y}) \mapsto \tilde{\mu}_{\boldsymbol{\Theta}|\mathbf{Y}=\mathbf{y}}$ is injective for characteristic kernels ℓ, analogous to mean embeddings (Sriperumbudur et al., 2011). Furthermore, as the integral (3.5) is a linear operator on $\ell(\boldsymbol{\theta}^\star, \cdot)$, the surrogate posterior mean embedding $\tilde{\mu}_{\boldsymbol{\Theta}|\mathbf{Y}=\mathbf{y}} \in \mathcal{H}_\ell$ is in the RKHS of ℓ. With a surrogate embedding that is injective to our surrogate posterior and in the RKHS, we can apply kernel herding (Chen et al., 2010) on $\tilde{\mu}_{\boldsymbol{\Theta}|\mathbf{Y}=\mathbf{y}}$ (3.5) using kernel ℓ to obtain S super-samples $\{\hat{\boldsymbol{\theta}}_s\}_{s=1}^S$ from the surrogate density $q(\boldsymbol{\theta}|\mathbf{y})$. That is, for each $s \in [S]$, the samples are obtained by

$$\hat{\boldsymbol{\theta}}_s = \underset{\boldsymbol{\theta} \in \vartheta}{\text{argmax}}\, \tilde{\mu}_{\boldsymbol{\Theta}|\mathbf{Y}=\mathbf{y}}(\boldsymbol{\theta}) - \frac{1}{s} \sum_{s'=1}^{s-1} \ell(\hat{\boldsymbol{\theta}}_{s'}, \boldsymbol{\theta}). \tag{3.7}$$

The inference algorithm is presented in algorithm 3.1.

4 Related Work

The simplest ABC algorithm is arguably the rejection ABC (REJ-ABC) sampler (Pritchard et al., 1999). It posits a set of prior parameters and rejects those whose simulations do not match the observations within a fixed threshold $\epsilon > 0$ under a distance measure.

Instead of sampling from the prior, MCMC-ABC and sequential Monte Carlo ABC (SMC-ABC) sample from proposal distributions iteratively and carefully accepts or discards each proposal stochastically based on approximate likelihood ratios (Sisson et al., 2007; Marjoram et al., 2003). They can however suffer from slow mixing, where it is difficult to escape a lucky sample with a high likelihood. They also do not leverage likelihood smoothness and thus require new simulations every iteration, which are then discarded and may still not result in an accepted sample.

Another branch of study include stochastic variational inference (SVI) approaches to ABC, which treats the likelihood approximation as another source of stochasticity in the stochastic gradient. This includes AV-ABC (Moreno et al., 2016), VBIL (Tran et al., 2017b), and VBSL (Ong et al., 2018). In contrast, likelihood-free variational inference (LFVI) (Tran et al., 2017a) uses density ratio estimation to approximate the variational objective, emphasizing inference on local latent variables. Nevertheless, SVI approaches posit parametric approximations that may have asymptotic bias.

Kernel-based approaches that leverage likelihood smoothness have been studied recently to reduce simulation requirements. The philosophy is that simulations of close-by parameters are informative, thus past results should not be discarded but remembered, even if this introduces model bias. Kernel ABC (K-ABC) (Nakagome et al., 2013), kernel recursive ABC (KR-ABC) (Kajihara et al., 2018), and kernel Bayes' rule (KBR) (Fukumizu et al., 2013) also employ CMEs to reduce simulation requirements. They differ to KELFI in the three aspects of model, learning, and inference. (Model) While they build posterior mean embeddings directly, KELFI builds likelihood surrogates first and make use of the full prior density to further leverage prior information before building posterior surrogates, which are then embedded into closed-form posterior mean embeddings. In contrast, the prior only appears as samples from $p(\boldsymbol{\theta})$ in K-ABC, KR-ABC, and KBR. This both limits the prior knowledge leveraged and prohibit the use of proposal prior samples. (Learning) KELFI crucially addresses hyperparameter learning in reference to the inference problem directly which was not straightforward previously. (Inference) K-ABC and KBR primarily infer posterior expectations, while KR-ABC produce point estimates. Instead, we de-

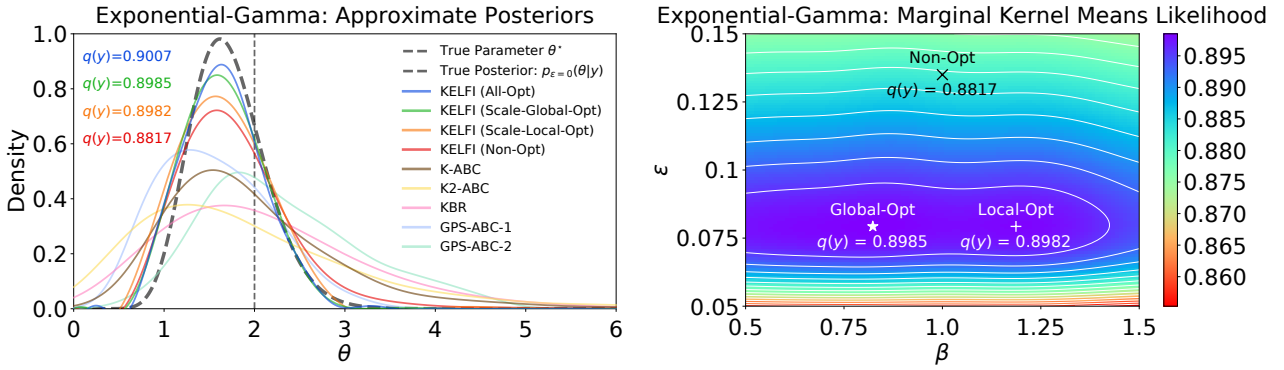

Figure 1: **(Left)** Comparison of approximate posteriors obtained from surrogate methods on the toy exponential-gamma problem. **(Right)** The corresponding MKML surface $q(\mathbf{y})$ as a function of (ϵ, β) with $\lambda = 10^{-3}\beta$.

sign a posterior sampling algorithm, which subsumes inferring posterior expectation. We further provide approximate posterior density KMP, which can both produce point estimates and quantify uncertainty.

As a consequence of theorem 3.4, the KMPE converges at rate $O_p(m^{-\frac{1}{4}})$ in RKHS norm if the regularization hyperparameter λ is chosen to decay at rate $O_p(m^{-\frac{1}{2}})$. Notably, this is faster than the convergence rate of KBR at $O_p(m^{-\frac{8}{27}\alpha})$ where $0 < \alpha \leq \frac{1}{2}$, which also requires other assumptions on the cross-covariance operators and for its two regularization hyperparameters to be decayed appropriately (Fukumizu et al., 2013).

Finally, we highlight that hyperparameter learning is a crucial aspect and differentiator of KELFI. This is especially true for learning ϵ, which tunes the critical balance between an accurate posterior $p_\epsilon(\boldsymbol{\theta}|\mathbf{y}) \approx p_0(\boldsymbol{\theta}|\mathbf{y})$ with small ϵ requiring high numbers of simulation calls, or a less accurate posterior with large ϵ relaxing the number of simulations required. This has been a challenging issue to address in the ABC literature in reference to the inference problem, even though its selection is often pivotal to the performance of the algorithm.

In the Gaussian process (GP) literature, hyperparameter learning through maximum marginal likelihood plays an important role in the success of a GP regressor (GPR). GP surrogate ABC (GPS-ABC) (Meeds and Welling, 2014) and GP-accelerated ABC (GPA-ABC) (Wilkinson, 2014) model the summary statistics surface and log likelihood surface respectively via a GP surrogate. In contrast, the KML model is equivalent to placing a GP surrogate on the likelihood surface itself. This removes the assumption that summary statistics are independent and Gaussian distributed as in GPS-ABC. Importantly, while GPS-ABC and GPA-ABC apply the GP marginal likelihood to learn their surrogate hyperparameters, it cannot learn ϵ or other hyperparameters since they are not part of the surrogate. This is because both approaches maximize the marginal likelihood for the GPR problem on the their respective target surfaces, but not the marginal likelihood for the overall inference problem, thus excluding other hyperparameters in the process.

5 Experiments

The goal of the experiments is to demonstrate the inference accuracy of KELFI under limited simulation budget and the effectiveness of MKML hyperparameter learning. We begin with isotropic ϵ and anisotropic $\boldsymbol{\beta} = \beta_0\boldsymbol{\sigma}$, and learn (ϵ, β_0) by maximizing the MKML (3.3) while keeping $\lambda = 10^{-3}\beta_0$ fixed for simplicity.

5.1 Toy Problem: Exponential-Gamma

The toy exponential-gamma problem is a standard benchmark for likelihood-free inference, since the true posterior $p_\epsilon(\boldsymbol{\theta}|\mathbf{y})$ is known and tractable even for $\epsilon = 0$.

To stress-test each method, we compare inference accuracy under very limited simulations of $m = 100$. We focus on comparing surrogate approaches, since other methods such as REJ-ABC, MCMC-ABC, SL-ABC, and ASL-ABC have reported simulation requirements several orders higher than 100 on this problem (Meeds and Welling, 2014). We use datasets of $n = 15$ for both observations and simulations, with their sample means as the summary statistic.

For GPS-ABC only we set a simulation budget of $m \leq 200$ and run it until 10000 posterior samples are generated. The hyperparameters of the GP surrogate itself are learned by maximizing the marginal likelihood of the GPR (Rasmussen and Williams, 2006). For the remaining hyperparameters that are not part of the surrogate, several configurations are compared and the results of the best two are shown, which used $m = 130$ and $m = 197$ simulations. For K-ABC, K2-ABC, and KBR, we use the median heuristic to set their length scale hyperparameters and manually

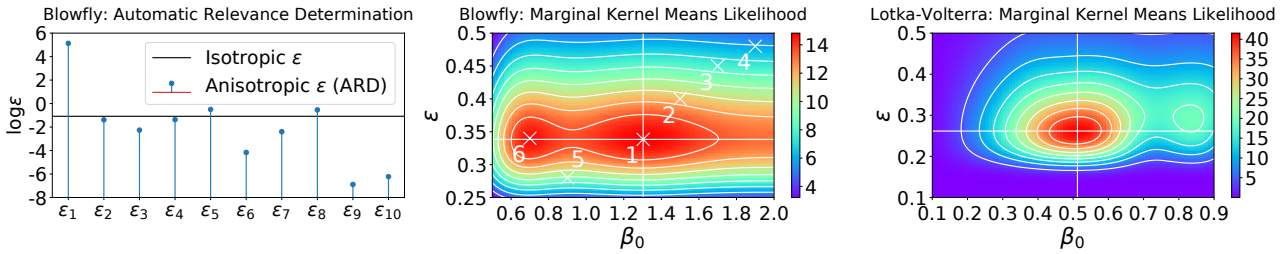

Figure 2: **(Left)** Blowfly: ARD on $\boldsymbol{\epsilon}$ for 10 summary statistics. **(Mid. & Right)** The MKML surface ($\times 10^5$) as a function of (ϵ, β_0) for fixed $\lambda = 10^{-3}\beta_0$ where $\boldsymbol{\beta} = \beta_0 \boldsymbol{\sigma}$. White intersection indicate optimum. For Blowfly, the NMSE (in %) for the indicated hyperparameter choices are: (1)0.72 ± 0.02, (2)1.10 ± 0.01, (3)2.07 ± 0.01, (4)2.15 ± 0.02, (5)1.11 ± 0.02, (6)1.11 ± 0.03. At $(\epsilon, \beta_0) = (10, 10)$ (outside the plot) the NMSE (in %) is 6.28 ± 0.03.

search for the most appropriate regularization hyperparameters. We use kernel density estimation (KDE) to visualize the posterior density from the unweighted samples of GPS-ABC and normalized weighted samples of K-ABC, K2-ABC, and KBR in fig. 1 (left).

For KELFI, we show the KMPs directly in fig. 1 (left). We first demonstrate the case when all hyperparameters $(\epsilon, \beta, \lambda)$ are learned (All-Opt). To enable visualization in 2D, we also present the case when the regularization hyperparameter λ is set to $10^{-3}\beta$ and only length scale hyperparameters (ϵ, β) are learned. In this case, we show KMPs under globally optimal (Scale-Global-Opt), locally optimal (Scale-Local-Opt), and arbitrarily chosen hyperparameters (Non-Opt). The corresponding MKML surface is shown in fig. 1 (right).

In fig. 1 we compare approximate posteriors from each algorithm against the true posterior $p_{\epsilon=0}(\boldsymbol{\theta}|\mathbf{y})$. While $\epsilon = 0$ for $p_{\epsilon=0}(\boldsymbol{\theta}|\mathbf{y})$, with only 100 simulations $\epsilon > 0$ is required for most LFI methods. Furthermore, except for K2-ABC, they only make use of summary statistics without further knowledge of the dataset size n. Consequently, most LFI methods produce approximations wider than $p_{\epsilon=0}(\boldsymbol{\theta}|\mathbf{y})$. Intuitively, there is not enough simulations and thus information to justify a more confident and peaked posterior. Nevertheless, by learning hyperparameters under the MKML, KELFI determines an appropriate scale ϵ for 100 simulations. As a result, KMPs are the closest to the true posterior $p_{\epsilon=0}(\boldsymbol{\theta}|\mathbf{y})$, with higher MKML $q(\mathbf{y})$ leading to more accurate KMPs $q(\boldsymbol{\theta}|\mathbf{y})$. This demonstrates the effectiveness of MKML as a hyperparameter learning objective for improving inference accuracy. In contrast, the two instances of GPS-ABC reveals that varying hyperparameters lead to significant changes in the resulting approximate posterior, yet without a similar objective like MKML it is unclear which one to use without ground truth. This is further emphasized by the wider posterior approximations obtained from K-ABC, K2-ABC, and KBR, which use the median heuristic to set hyperparameters. This is often sub-optimal since the heuristic makes no reference to the inference problem.

5.2 Chaotic Ecological Systems: Blowfly

The Blowfly simulator describes the complex population dynamics of adult blowflies. Across a range of parameters it exhibits chaotic behavior that have distinct discrepancies from real observations, resulting in a challenging inference problem. We follow the setup of Wood (2010). There are 6 model parameters from which the simulator generates a time series of 180 data points that is then summarized into 10 statistics as described in Meeds and Welling (2014), Moreno et al. (2016), and Park et al. (2016). We similarly place a broad diagonal Gaussian prior on log parameters.

The standard Blowfly problem has no ground truth parameters, only a set of observations. We therefore measure inference accuracy by considering mean squared errors (MSEs) between statistics generated using the posterior and the observed statistics. We normalize the MSE of each statistic by the corresponding MSE achieved under the prior, and average across the 10 statistics into a final normalized MSE (NMSE). As simulations are expensive, in fig. 3 (left) we record average NMSE against simulations used to understand inference efficiency. Each method is repeated 10 times with randomized simulations before their NMSE is averaged. Appendix D provides further details.

As new simulations become available, we relearn and update the hyperparameters for KELFI by maximizing the MKML. Figure 2 (center) shows an instance of the MKML surface used to learn the hyperparameters for KELFI when using $m = 280$ simulations. For KBR and K-ABC we update hyperparameters by the median length heuristic. For K-ABC we also report the case where the heurstic is scaled by a constant denoted with (S), which achieved significantly better accuracy and confirms that the heuristic is often sub-optimal.

Overall, the top three performers are KELFI, KBR, and GPS-ABC. Across a range of simulation calls, KELFI achieves the lowest error. It is also the only method that achieved less than 1% average NMSE

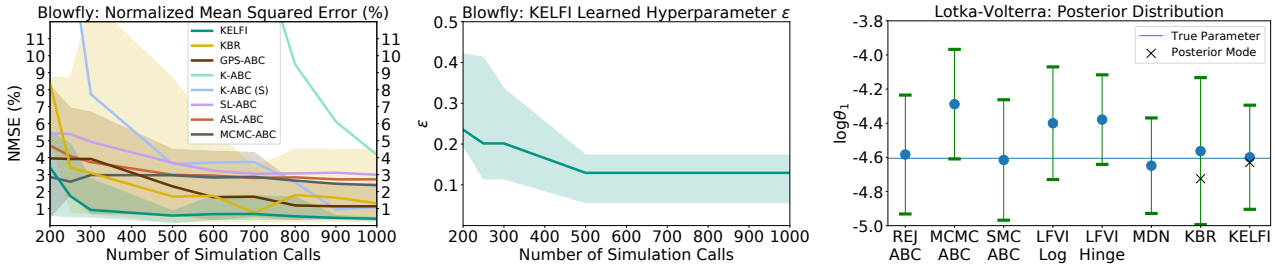

Figure 3: (**Left**) Blowfly: Average NMSE (in %) under posteriors against simulation calls. Shaded regions show NMSE variability for KELFI, KBR, and GPS-ABC. (**Mid.**) Blowfly: Learned ϵ value under maximum MKML. (**Right**) Lotka-Volterra: The middle 95% credible interval of the marginal posterior distribution of $\log \theta_1$.

within 1000 simulations and achieves this as early as 300 simulations. The most competitive methods to KELFI are KBR and GPS-ABC. For these three methods, we also show their variability from best to worst case NMSEs out of the 10 repeats to visualize their sensitivity to the stochasticity in randomized simulations. This reveals that KELFI is a stable outperformer with comparatively less variability across randomized runs.

We proceed to demonstrate and emphasize the effectiveness and suitability of MKML as a hyperparameter learning objective, using the case with 280 simulations as an example. Figure 2 (center) illustrates that hyperparameters with a higher MKML (3.3) result in lower NMSE consistently. Notably, even with suboptimal hyperparameter choices, KELFI still achieves competitive average NMSE scores of less than 2.2%. At 280 simulations, the next best average NMSE score is almost 3% by MCMC-ABC as shown in fig. 3 (left).

Figure 3 (center) suggests that learning the scale ϵ under MKML reveals an automatic decay schedule which does not have to be set a-priori. As ϵ controls the scale within which discrepancies between simulations and observations are measured, it is expected that this scale decays as more simulation data is available. Without the MKML, both the initialization of ϵ and its decay schedule are not straight forward to determine.

In fig. 2 (left), we show that we can perform ARD on the ABC ϵ-kernel $\kappa_{\boldsymbol{\epsilon}}$, and hence the kernel $k_{\boldsymbol{\epsilon}}$, by using a different ϵ_i for each of the 10 statistics. We do this by initializing each ϵ_i to the isotropic solution in fig. 2 (center) and further optimize the MKML to learn all ϵ_i jointly. In particular, the first summary statistic describes the average log population numbers nears its troughs (first quartile), and is determined to be comparatively irrelevant (high ϵ_i). Meanwhile, the last two statistics describe the number of peaks at two thresholds, and are determined to be comparatively relevant (low ϵ_i). This agrees with the intuition that Blowfly population dynamics are highly characterized by its peaks, instead of i its troughs (Wood, 2010).

5.3 Predator-Prey Dynamics: Lotka-Volterra

The Lotka-Volterra simulator describes the time evolution of the populations within a predator-prey system. Only for a small set of parameters does the model simulate a realistic scenario with oscillatory behavior, making the inference task formidably challenging. We follow the exact setup as described in Papamakarios and Murray (2016). 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 show in fig. 3 (right) the marginal posterior distribution for $\log \theta_1$.

KELFI achieves competitive performance using only 2500 simulations, with both posterior mean and mode close to the true value. The MKML for hyperparameter learning is shown in fig. 2 (right). Posterior mode is obtained by maximizing the KMP. Meanwhile, the three ABC methods used up to 100000 simulations. While confident, LFVI (Tran et al., 2017a) tends to have a biased posterior mean. For direct comparison, both KELFI and mixture density network (MDN) (Papamakarios and Murray, 2016) use the original prior as the proposal prior. KELFI achieves slightly higher accuracy than MDN which used 10000 simulations, 4 times that used for KELFI. Finally, we also similarly use 2500 simulations for KBR. With the same number of simulations, KELFI achieves higher accuracy in both mean and mode with higher confidence.

6 Conclusion

KELFI provides a holistic framework for automatic likelihood-free inference. It is a stable outperformer compared to state-of-the-art methods, while producing interpretable automatic relevance determination of summary statistics and automatic decay schedules for ϵ. By optimizing an approximate Bayesian marginal likelihood, it automatically learns and adapts hyperparameters including the ϵ-kernel to improve inference accuracy when limited simulations are available.

References

Andrieu, C., Roberts, G. O., et al. (2009). The pseudo-marginal approach for efficient monte carlo computations. *The Annals of Statistics*, 37(2):697–725.

Beaumont, M. A. (2010). Approximate bayesian computation in evolution and ecology. *Annual review of ecology, evolution, and systematics*, 41:379–406.

Carmeli, C., De Vito, E., Toigo, A., and Umanitá, V. (2010). Vector valued reproducing kernel hilbert spaces and universality. *Analysis and Applications*, 8(01):19–61.

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

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

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

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

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

Kajihara, T., Kanagawa, M., Yamazaki, K., and Fukumizu, K. (2018). Kernel recursive abc: Point estimation with intractable likelihood. In *International Conference on Machine Learning*, pages 2405–2414.

Kanagawa, M., Nishiyama, Y., Gretton, A., and Fukumizu, K. (2016). Filtering with state-observation examples via kernel monte carlo filter. *Neural computation*, 28(2):382–444.

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

Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003). Markov chain monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328.

Meeds, E. and Welling, M. (2014). Gps-abc: Gaussian process surrogate approximate bayesian computation. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, pages 593–602. AUAI Press.

Mitrovic, J., Sejdinovic, D., and Teh, Y. W. (2016). DR-ABC: Approximate Bayesian Computation with kernel-based distribution regression.

Moreno, A., Adel, T., Meeds, E., Rehg, J. M., and Welling, M. (2016). Automatic variational abc. *stat*, 1050:28.

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

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

Ong, V. M., Nott, D. J., Tran, M.-N., Sisson, S. A., and Drovandi, C. C. (2018). Variational Bayes with synthetic likelihood. *Statistics and Computing*, 28(4):971–988.

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

Park, M., Jitkrittum, W., and Sejdinovic, D. (2016). K2-ABC: Approximate Bayesian computation with kernel embeddings.

Price, L. F., Drovandi, C. C., Lee, A., and Nott, D. J. (2017). Bayesian synthetic likelihood. *Journal of Computational and Graphical Statistics*, pages 1–11.

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., and Feldman, M. W. (1999). Population growth of human y chromosomes: a study of y chromosome microsatellites. *Molecular biology and evolution*, 16(12):1791–1798.

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

Sisson, S. A., Fan, Y., and Tanaka, M. M. (2007). Sequential monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 104(6):1760–1765.

Snoek, J., Larochelle, H., and Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959.

Song, L., Fukumizu, K., and Gretton, A. (2013). Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference

in graphical models. *IEEE Signal Processing Magazine*, 30(4):98–111.

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

Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. R. (2010). On the relation between universality, characteristic kernels and RKHS embedding of measures. In *AISTATS*, pages 773–780.

Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. R. (2011). Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12(Jul):2389–2410.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, M. P. (2009). Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6(31):187–202.

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

Tran, M.-N., Nott, D. J., and Kohn, R. (2017b). Variational Bayes with intractable likelihood. *Journal of Computational and Graphical Statistics*, 26(4):873–882.

Wilkinson, R. (2014). Accelerating abc methods using gaussian processes. In *Artificial Intelligence and Statistics*, pages 1015–1023.

Wood, S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466(7310):1102.

A Theoretical Guarantees on Convergence

We provide theoretical guarantees that establish convergence of the kernel embedding likelihood-free inference (KELFI) framework. Appendix A.1 begins by summarizing the properties of kernels used in KELFI and introducing relevant quantities. Appendices A.2 and A.3 provide an overview of conditional mean embeddings (CMEs) and their empirical estimates respectively in the context of KELFI. Appendix A.4 establishes general convergence theorems for estimators based on the CME. Using these results, we prove convergence guarantees for the kernel means likelihood (KML), marginal kernel means likelihood (MKML), kernel means posterior (KMP), and kernel means posterior embedding (KMPE) in appendices A.5, A.6, A.7 and A.8 respectively.

A.1 Kernel Properties

The KELFI framework uses a data kernel $k : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ where $\mathcal{X}, \mathcal{Y} \subseteq \mathcal{D}$. We do not assume that $\mathcal{X}$ and $\mathcal{Y}$ are necessarily the same. For example, it is possible to record observation $\mathbf{y}$ in which the simulator $p(\mathbf{x}|\boldsymbol{\theta})$ can never generate or fully recover, such as when $\mathcal{X} \subset \mathcal{Y}$. Conversely, it is also possible that the simulator $p(\mathbf{x}|\boldsymbol{\theta})$ can generate a larger variety of simulations $\mathbf{x}$ than that is possible to observe, such as when $\mathcal{Y} \subset \mathcal{X}$. It can also be neither of such cases such as when $\mathcal{X}$ an $\mathcal{Y}$ only have some overlap. However, since we assume $\mathcal{X}, \mathcal{Y} \subseteq \mathcal{D}$, the kernel k is able to measure the similarity between simulated data $\mathbf{x} \in \mathcal{X} \subseteq \mathcal{D}$ and observed data $\mathbf{y} \in \mathcal{Y} \subseteq \mathcal{D}$.

The KELFI framework employs bounded symmetric positive definite kernels ℓ and k. Because they are bounded, we can explicitly denote the following upper bounds to their RKHS norm,

$$\bar{\ell} := \sup_{\boldsymbol{\theta} \in \vartheta} \|\ell(\boldsymbol{\theta}, \cdot)\|_{\mathcal{H}_\ell} = \sup_{\boldsymbol{\theta} \in \vartheta} \sqrt{\ell(\boldsymbol{\theta}, \boldsymbol{\theta})}, \tag{A.1}$$

$$\bar{k} := \sup_{\mathbf{d} \in \mathcal{D}} \|k(\mathbf{d}, \cdot)\|_{\mathcal{H}_k} = \sup_{\mathbf{d} \in \mathcal{D}} \sqrt{k(\mathbf{d}, \mathbf{d})}. \tag{A.2}$$

When ℓ and k are stationary, we have $\bar{\ell} = \sqrt{\ell(\mathbf{0}, \mathbf{0})}$ and $\bar{k} = \sqrt{k(\mathbf{0}, \mathbf{0})}$.

In the KELFI framework, we first select the ϵ-kernel κ_ϵ. Based on this the choice of the ϵ-kernel, we then select the kernel k to satisfy

$$\kappa_\epsilon(\mathbf{y}, \mathbf{x}) = c_\epsilon k(\mathbf{y}, \mathbf{x}), \tag{A.3}$$

where $c_\epsilon > 0$ is a scaling constant to ensure that $\kappa_\epsilon(\mathbf{y}, \mathbf{x}) = p_\epsilon(\mathbf{y}|\mathbf{x})$ is a normalized density on $\mathcal{Y}$. In contrast, the kernel k has no such restriction. Since it is a scaled version of k, κ_ϵ is also bounded symmetric positive definite as a function of $\mathbf{x}$ and $\mathbf{y}$. In this way, $\kappa_\epsilon(\mathbf{d}, \cdot) \in \mathcal{H}_k$ is always in the RKHS $\mathcal{H}_k$ characterized by k for all $\mathbf{d} \in \mathcal{D}$. As a consequence, ϵ is also a hyperparameter of k, although this is not explicitly notated for brevity.

Since $\mathbf{y} \in \mathcal{Y} \subseteq \mathcal{D}$, we have $\kappa_\epsilon(\mathbf{y}, \cdot) \in \mathcal{H}_k$. We can then find its RKHS norm,

$$\|\kappa_\epsilon(\mathbf{y}, \cdot)\|_{\mathcal{H}_k} = c_\epsilon \|k(\mathbf{y}, \cdot)\|_{\mathcal{H}_k} = c_\epsilon \sqrt{k(\mathbf{y}, \mathbf{y})} = \sqrt{c_\epsilon} \sqrt{c_\epsilon k(\mathbf{y}, \mathbf{y})} = \sqrt{c_\epsilon} \sqrt{\kappa_\epsilon(\mathbf{y}, \mathbf{y})}, \tag{A.4}$$

which is different to $\|\kappa_\epsilon(\mathbf{y}, \cdot)\|_{\mathcal{H}_{\kappa_\epsilon}} = \sqrt{\kappa_\epsilon(\mathbf{y}, \mathbf{y})}$. Therefore, while the KELFI algorithm only requires κ_ϵ to be specified and k is not explicitly used, this subtle difference is a reminder that k is the underlying kernel that defines the RKHS, not κ_ϵ. As a consequence, we have that the upper bound to the RKHS norm of κ_ϵ satisfies

$$\bar{\kappa}_\epsilon := \sup_{\mathbf{d} \in \mathcal{D}} \|\kappa_\epsilon(\mathbf{d}, \cdot)\|_{\mathcal{H}_k} = \sqrt{c_\epsilon} \sup_{\mathbf{d} \in \mathcal{D}} \sqrt{\kappa_\epsilon(\mathbf{d}, \mathbf{d})}. \tag{A.5}$$

Furthermore, if κ_ϵ is stationary, then $\kappa_\epsilon(\mathbf{d}, \mathbf{d}) = \kappa_\epsilon(\mathbf{0}, \mathbf{0})$ for all $\mathbf{d} \in \mathcal{D}$. A typical example is the Gaussian density $\kappa_\epsilon(\mathbf{y}, \mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{x}, \epsilon^2 I)$. In this case, $c_\epsilon = 1/(\sqrt{2\pi}\epsilon)^n$ and $\kappa_\epsilon(\mathbf{y}, \mathbf{y}) = 1/(\sqrt{2\pi}\epsilon)^n$ are the same, and thus $\|\kappa_\epsilon(\mathbf{y}, \cdot)\|_{\mathcal{H}_k} = 1/(\sqrt{2\pi}\epsilon)^n = c_\epsilon$. The corresponding kernel k is the isotropic Gaussian kernel

When $\mathcal{D} = \mathbb{R}^n$, the most commonly used kernel for the KELFI framework is the anisotropic Gaussian kernel where each dimension uses a potentially different length scale σ_i. When its length scales are learned via some hyperparameter learning algorithm, it is also referred to as the ARD kernel. This kernel has the following form,

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{1}{2} \sum_{i=1}^{n} \left(\frac{x_i - x_i'}{\sigma_i} \right)^2 \right). \tag{A.6}$$

Since $\kappa_\epsilon(\mathbf{y}, \mathbf{x}) = c_\epsilon k(\mathbf{y}, \mathbf{x})$, this means that the length scales are simply the ABC tolerance $\sigma_i = \epsilon_i$ for $i \in [n]$, and that there can be a separate tolerance for each dimension of the data or summary statistic. Similarly, when $\vartheta = \mathbb{R}^D$, we also often employ the ARD kernel for ℓ, but we use β_d, $d \in [D]$, to denote the length scales.

A.2 Conditional Mean Embedding

To construct a conditional mean operator $\mathcal{U}_{\mathbf{X}|\mathbf{\Theta}}$ corresponding to the distribution $p(\mathbf{x}|\boldsymbol{\theta})$, we first choose a kernel $\ell : \vartheta \times \vartheta \to \mathbb{R}$ for domain ϑ and another kernel $k : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ for domain $\mathcal{D}$. These kernels ℓ and k each describe how similarity is measured within their respective domains, and are bounded symmetric positive definite such that they uniquely define the RKHS $\mathcal{H}_\ell$ and $\mathcal{H}_k$.

The conditional mean operator $\mathcal{U}_{\mathbf{X}|\mathbf{\Theta}} : \mathcal{H}_\ell \to \mathcal{H}_k$ is defined by the equation $\mu_{\mathbf{X}|\mathbf{\Theta}=\boldsymbol{\theta}} = \mathcal{U}_{\mathbf{X}|\mathbf{\Theta}}\ell(\boldsymbol{\theta}, \cdot)$, where $\mu_{\mathbf{X}|\mathbf{\Theta}=\boldsymbol{\theta}}$ is the CME defined by

$$\mu_{\mathbf{X}|\mathbf{\Theta}=\boldsymbol{\theta}} := \mathbb{E}[k(\mathbf{X}, \cdot)|\mathbf{\Theta} = \boldsymbol{\theta}]. \tag{A.7}$$

In this sense, $\mathcal{U}_{\mathbf{X}|\mathbf{\Theta}}$ sweeps out a family of CMEs $\mu_{\mathbf{X}|\mathbf{\Theta}=\boldsymbol{\theta}} \in \mathcal{H}_k$, each indexed by $\boldsymbol{\theta} \in \vartheta$.

We then define cross covariance operators $C_{\mathbf{X}\mathbf{\Theta}} := \mathbb{E}[k(\mathbf{X}, \cdot) \otimes \ell(\mathbf{\Theta}, \cdot)] : \mathcal{H}_\ell \to \mathcal{H}_k$ and $C_{\mathbf{\Theta}\mathbf{\Theta}} := \mathbb{E}[\ell(\mathbf{\Theta}, \cdot) \otimes \ell(\mathbf{\Theta}, \cdot)] : \mathcal{H}_\ell \to \mathcal{H}_\ell$. Alternatively, they can be seen as elements within the tensor product space $C_{\mathbf{X}\mathbf{\Theta}} \in \mathcal{H}_k \otimes \mathcal{H}_\ell$ and $C_{\mathbf{\Theta}\mathbf{\Theta}} \in \mathcal{H}_\ell \otimes \mathcal{H}_\ell$. That is, they are second order mean embeddings.

Under the assumption that $\ell(\boldsymbol{\theta}, \cdot) \in \text{image}(C_{\mathbf{\Theta}\mathbf{\Theta}})$, it can be shown that $\mathcal{U}_{\mathbf{X}|\mathbf{\Theta}} = C_{\mathbf{X}\mathbf{\Theta}}(C_{\mathbf{\Theta}\mathbf{\Theta}})^{-1}$. While this assumption is satisfied for finite domains ϑ with a characteristic kernel ℓ, it does not necessarily hold when ϑ is a continuous domain (Fukumizu et al., 2004). Instead, in this case $C_{\mathbf{X}\mathbf{\Theta}}(C_{\mathbf{\Theta}\mathbf{\Theta}})^{-1}$ becomes only an approximation to $\mathcal{U}_{\mathbf{X}|\mathbf{\Theta}}$, and we instead regularize the inversion with a regularization hyperparameter $\lambda \geq 0$ and use $\mathcal{U}_{\mathbf{X}|\mathbf{\Theta}} = C_{\mathbf{X}\mathbf{\Theta}}(C_{\mathbf{\Theta}\mathbf{\Theta}} + \lambda I)^{-1}$, which also serves to avoid overfitting (Song et al., 2013). This relaxation can be applied to all subsequent results and theorems.

A.3 Empirical Estimate for the Conditional Mean Embedding

Suppose $\{\boldsymbol{\theta}_j, \mathbf{x}_j\} \sim p(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$ are *iid* across $j \in [m]$. The conditional mean operator $\mathcal{U}_{\mathbf{X}|\mathbf{\Theta}}$ is estimated by

$$\hat{\mathcal{U}}_{\mathbf{X}|\mathbf{\Theta}} = \Phi(L + m\lambda I)^{-1}\Psi^T, \tag{A.8}$$

where $\Phi := \begin{bmatrix} k(\mathbf{x}_1, \cdot) & \cdots & k(\mathbf{x}_m, \cdot) \end{bmatrix}$, $\Psi := \begin{bmatrix} \ell(\boldsymbol{\theta}_1, \cdot) & \cdots & \ell(\boldsymbol{\theta}_m, \cdot) \end{bmatrix}$, and $L := \{\ell(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)\}_{i,j=1}^m$. The CME can then be estimated by

$$\hat{\mu}_{\mathbf{X}|\mathbf{\Theta}=\boldsymbol{\theta}} = \hat{\mathcal{U}}_{\mathbf{X}|\mathbf{\Theta}}\ell(\boldsymbol{\theta}, \cdot) = \Phi(L + m\lambda I)^{-1}\boldsymbol{\ell}(\boldsymbol{\theta}) \tag{A.9}$$

where $\boldsymbol{\ell}(\boldsymbol{\theta}) := \{\ell(\boldsymbol{\theta}_j, \boldsymbol{\theta})\}_{j=1}^m$ (Song et al., 2009).

For any function $f \in \mathcal{H}_k$, the conditional expectation of f under $p(\mathbf{x}|\boldsymbol{\theta})$, or $g(\boldsymbol{\theta}) := \mathbb{E}[f(\mathbf{X})|\mathbf{\Theta} = \boldsymbol{\theta}]$, can be approximated by the inner product $\hat{g}(\boldsymbol{\theta}) := \langle f, \hat{\mu}_{\mathbf{X}|\mathbf{\Theta}=\boldsymbol{\theta}} \rangle_{\mathcal{H}_k}$ by using an empirical CME $\hat{\mu}_{\mathbf{X}|\mathbf{\Theta}=\boldsymbol{\theta}}$. Letting $\mathbf{f} := \{f(\mathbf{x}_j)\}_{j=1}^m$, this approximation admits the following form,

$$\hat{g}(\boldsymbol{\theta}) = \mathbf{f}^T(L + m\lambda I)^{-1}\boldsymbol{\ell}(\boldsymbol{\theta}). \tag{A.10}$$

Importantly, $\hat{\mu}_{\mathbf{X}|\mathbf{\Theta}=\boldsymbol{\theta}}$ is estimated from *joint* samples $\{\boldsymbol{\theta}_j, \mathbf{x}_j\}_{j=1}^m$, even though it is encoding the corresponding conditional distribution $p(\mathbf{x}|\boldsymbol{\theta})$. It is this fact that allows for an arbitrary choice $\pi(\boldsymbol{\theta})$ on the marginal distribution of $\mathbf{\Theta}$, which does not necessarily need to be the same as $p(\boldsymbol{\theta})$.

Under the assumption that $\ell(\boldsymbol{\theta}, \cdot) \in \text{image}(C_{\mathbf{\Theta}\mathbf{\Theta}})$, the empirical CME $\hat{\mu}_{\mathbf{X}|\mathbf{\Theta}=\boldsymbol{\theta}}$ converges to the true CME $\mu_{\mathbf{X}|\mathbf{\Theta}=\boldsymbol{\theta}}$ in RKHS norm at rate $O_p((m\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}})$ (Song et al., 2009, Theorem 6). That is,

$$\begin{aligned} &\forall \boldsymbol{\theta} \in \vartheta,\ \forall \epsilon > 0,\ \exists M_\epsilon > 0 \quad s.t. \\ &\mathbb{P}\Big[\big\|\hat{\mu}_{\mathbf{X}|\mathbf{\Theta}=\boldsymbol{\theta}} - \mu_{\mathbf{X}|\mathbf{\Theta}=\boldsymbol{\theta}}\big\|_{\mathcal{H}_k} > M_\epsilon\Big((m\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}}\Big)\Big] < \epsilon. \end{aligned} \tag{A.11}$$

Consequently, the empirical CME converges at rate $O_p(m^{-\frac{1}{4}})$ if λ is chosen to decay at rate $O_p(m^{-\frac{1}{2}})$, and often better convergence rates can be achieved under appropriate assumptions on $p(\mathbf{x}|\boldsymbol{\theta})$ (Song et al., 2013). Again, the regularization hyperparameter λ relaxes the assumption that $\ell(\boldsymbol{\theta}, \cdot) \in \text{image}(C_{\mathbf{\Theta}\mathbf{\Theta}})$.

Finally, since $\hat{\mu}_{\mathbf{X}|\mathbf{\Theta}=\boldsymbol{\theta}} = \hat{\mathcal{U}}_{\mathbf{X}|\mathbf{\Theta}}\ell(\boldsymbol{\theta}, \cdot)$ convergences to $\mu_{\mathbf{X}|\mathbf{\Theta}=\boldsymbol{\theta}} = \mathcal{U}_{\mathbf{X}|\mathbf{\Theta}}\ell(\boldsymbol{\theta}, \cdot)$ in RKHS norm at rate $O_p((m\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}})$ for all $\boldsymbol{\theta} \in \vartheta$ and $\ell(\boldsymbol{\theta}, \cdot)$ does not depend on m, we also have that $\hat{\mathcal{U}}_{\mathbf{X}|\mathbf{\Theta}}$ converges to $\mathcal{U}_{\mathbf{X}|\mathbf{\Theta}}$ in Hilbert Schmidt (HS) norm at the same rate. That is,

$$\begin{aligned} &\forall \epsilon > 0,\ \exists M_\epsilon > 0 \quad s.t. \\ &\mathbb{P}\Big[\big\|\hat{\mathcal{U}}_{\mathbf{X}|\mathbf{\Theta}} - \mathcal{U}_{\mathbf{X}|\mathbf{\Theta}}\big\|_{HS} > M_\epsilon\Big((m\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}}\Big)\Big] < \epsilon. \end{aligned} \tag{A.12}$$

A.4 General Convergence Theorems

We now establish some general convergence theorems for estimators based on inner products with the CME. The aim is to provide a sense of the stochastic convergence of any estimator $\hat{a}$ to its true quantity a with respect to some metric $d(\hat{a}, a)$. We do this by showing that either $\|\hat{\mu}_{\mathbf{X}|\boldsymbol{\Theta}=\boldsymbol{\theta}} - \mu_{\mathbf{X}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\|_{\mathcal{H}_k}$ or $\|\hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}}\|_{HS}$ is an upper bound of $d(\hat{a}, a)$ up to a scaling constant.

Lemma A.1. *Suppose that $\ell(\boldsymbol{\theta}, \cdot) \in \text{image}(C_{\boldsymbol{\Theta}\boldsymbol{\Theta}})$ and that there exists $0 \leq \gamma < \infty$ such that for some estimator $\hat{a}$, target a, and metric $d(\hat{a}, a)$,*

$$d(\hat{a}, a) \leq \gamma \left\| \hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}} \right\|_{HS}, \tag{A.13}$$

then the estimator $\hat{a}$ converges to the target a with respect to the metric d at rate $O_p((m\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}})$.

Proof. Suppose that there exists $0 \leq \gamma < \infty$ such that (A.13) is satisfied. That is, the inequality (A.13) holds for all possible data observations $\{\boldsymbol{\theta}_j, \mathbf{x}_j\}_{j=1}^m$. For any constant C, the implication statement $\left\| \hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}} \right\|_{HS} \leq C \implies d(\hat{a}, a) \leq C\gamma$ holds for all possible observation events $\omega \in \Omega$. Writing this explicitly in event space translates this to a statement of probability inequality,

$$\begin{aligned} &\{\omega \in \Omega : \left\| \hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}} \right\|_{HS} \leq C\} \subseteq \{\omega \in \Omega : d(\hat{a}, a) \leq C\gamma\} \\ &\implies \mathbb{P}\Big[\left\| \hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}} \right\|_{HS} \leq C\Big] \leq \mathbb{P}\Big[d(\hat{a}, a) \leq C\gamma\Big]. \end{aligned} \tag{A.14}$$

Since we assume that $\ell(\boldsymbol{\theta}, \cdot) \in \text{image}(C_{\boldsymbol{\Theta}\boldsymbol{\Theta}})$, statement (A.11) is valid. By letting $C = M_\epsilon((m\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}})$ in (A.14), we immediately have that the probability inequality in statement (A.12) is also true if we replace $\left\| \hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}} \right\|_{HS}$ with $d(\hat{a}, a)$ and M_ϵ with γM_ϵ,

$$\begin{aligned} &\mathbb{P}\Big[\left\| \hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}} \right\|_{HS} > M_\epsilon\Big((m\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}}\Big)\Big] < \epsilon \\ \implies &1 - \mathbb{P}\Big[\left\| \hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}} \right\|_{HS} \leq M_\epsilon\Big((m\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}}\Big)\Big] < \epsilon \\ \implies &\mathbb{P}\Big[\left\| \hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}} \right\|_{HS} \leq M_\epsilon\Big((m\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}}\Big)\Big] > 1 - \epsilon \\ \implies &\mathbb{P}\Big[d(\hat{a}, a) \leq \gamma M_\epsilon\Big((m\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}}\Big)\Big] > 1 - \epsilon \\ \implies &1 - \mathbb{P}\Big[d(\hat{a}, a) \leq \gamma M_\epsilon\Big((m\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}}\Big)\Big] < \epsilon \\ \implies &\mathbb{P}\Big[d(\hat{a}, a) > \gamma M_\epsilon\Big((m\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}}\Big)\Big] < \epsilon, \end{aligned} \tag{A.15}$$

where we employed statement (A.14) between the third and fourth line for $C = M_\epsilon((m\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}})$. Therefore, since M_ϵ is arbitrary, define $\tilde{M}_\epsilon := \gamma M_\epsilon$ so that the following statement holds,

$$\forall \epsilon > 0,\ \exists \tilde{M}_\epsilon > 0 \quad s.t. \quad \mathbb{P}\Big[d(\hat{a}, a) > \tilde{M}_\epsilon\Big((m\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}}\Big)\Big] < \epsilon. \tag{A.16}$$

In other words, the estimator $\hat{a}$ stochastically converges to a at a rate of at least $O_p((n\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}})$ with respect to the metric d. □

Lemma A.2. *Suppose that $\ell(\boldsymbol{\theta}, \cdot) \in \text{image}(C_{\boldsymbol{\Theta}\boldsymbol{\Theta}})$ and that there exists $0 \leq \gamma < \infty$ such that for some estimator $\hat{a}$, target a, and metric $d(\hat{a}, a)$,*

$$d(\hat{a}, a) \leq \gamma \left\| \hat{\mu}_{\mathbf{X}|\boldsymbol{\Theta}=\boldsymbol{\theta}} - \mu_{\mathbf{X}|\boldsymbol{\Theta}=\boldsymbol{\theta}} \right\|_{\mathcal{H}_k}, \tag{A.17}$$

then the estimator $\hat{a}$ converges to the target a with respect to the metric d at rate $O_p((m\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}})$.

Proof. The proof is identical to the proof for lemma A.1, where $\left\| \hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}} \right\|_{HS}$ is replaced with $\left\| \hat{\mu}_{\mathbf{X}|\boldsymbol{\Theta}=\boldsymbol{\theta}} - \mu_{\mathbf{X}|\boldsymbol{\Theta}=\boldsymbol{\theta}} \right\|_{\mathcal{H}_k}$ throughout. Alternatively, since $\left\| \hat{\mu}_{\mathbf{X}|\boldsymbol{\Theta}=\boldsymbol{\theta}} - \mu_{\mathbf{X}|\boldsymbol{\Theta}=\boldsymbol{\theta}} \right\|_{\mathcal{H}_k} = \left\| (\hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}})\ell(\boldsymbol{\theta}, \cdot) \right\|_{\mathcal{H}_k} \leq \left\| \hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}} \right\|_{HS} \left\| \ell(\boldsymbol{\theta}, \cdot) \right\|_{\mathcal{H}_\ell} = \left\| \hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}} \right\|_{HS} \sqrt{\ell(\boldsymbol{\theta}, \boldsymbol{\theta})}$, $\forall \boldsymbol{\theta} \in \vartheta$, we have $d(\hat{a}, a) \leq \gamma \ell(\boldsymbol{\theta}, \boldsymbol{\theta}) \left\| \hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}} \right\|_{HS} \leq \gamma(\sup_{\boldsymbol{\theta} \in \vartheta} \sqrt{\ell(\boldsymbol{\theta}, \boldsymbol{\theta})}) \left\| \hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}} \right\|_{HS} = \gamma\bar{\ell} \left\| \hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}} \right\|_{HS}$, $\forall \boldsymbol{\theta} \in \vartheta$. Since $\gamma\bar{\ell}$ is finite and does not depend on m, we apply lemma A.1 to arrive at lemma A.2. □

With lemmas A.1 and A.2, we are now equipped to show the convergence of various estimators based on CMEs.

A.5 Convergence Guarantees for Kernel Means Likelihood

In all subsequent theorems and proofs, recall that the approximate surrogate densities q depend on m and ϵ, as well as other kernel and regularization hyperparameters, even though this is not explicitly notated.

Theorem A.3. *Assume $\ell(\boldsymbol{\theta}, \cdot) \in \text{image}(C_{\boldsymbol{\Theta}\boldsymbol{\Theta}})$. The kernel means likelihood (KML) $q(\mathbf{y}|\boldsymbol{\theta})$ converges to the likelihood $p_\epsilon(\mathbf{y}|\boldsymbol{\theta})$ uniformly at rate $O_p((m\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}})$ as a function of $\boldsymbol{\theta} \in \vartheta$ and $\mathbf{y} \in \mathcal{Y}$.*

Proof. Consider the absolute difference between the KML $q(\mathbf{y}|\boldsymbol{\theta})$ and the likelihood $p_\epsilon(\mathbf{y}|\boldsymbol{\theta})$,

$$\begin{aligned}
|q(\mathbf{y}|\boldsymbol{\theta}) - p_\epsilon(\mathbf{y}|\boldsymbol{\theta}))| =& |\langle \kappa_\epsilon(\mathbf{y}, \cdot), \hat{\mu}_{\mathbf{X}|\boldsymbol{\Theta}=\boldsymbol{\theta}} \rangle_{\mathcal{H}_k} - \langle \kappa_\epsilon(\mathbf{y}, \cdot), \mu_{\mathbf{X}|\boldsymbol{\Theta}=\boldsymbol{\theta}} \rangle_{\mathcal{H}_k}| \\
=& |\langle \kappa_\epsilon(\mathbf{y}, \cdot), \hat{\mu}_{\mathbf{X}|\boldsymbol{\Theta}=\boldsymbol{\theta}} - \mu_{\mathbf{X}|\boldsymbol{\Theta}=\boldsymbol{\theta}} \rangle_{\mathcal{H}_k}| \\
\leq& \|\kappa_\epsilon(\mathbf{y}, \cdot)\|_{\mathcal{H}_k} \|\hat{\mu}_{\mathbf{X}|\boldsymbol{\Theta}=\boldsymbol{\theta}} - \mu_{\mathbf{X}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\|_{\mathcal{H}_k} \\
\leq& \bar{\kappa}_\epsilon \|\hat{\mu}_{\mathbf{X}|\boldsymbol{\Theta}=\boldsymbol{\theta}} - \mu_{\mathbf{X}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\|_{\mathcal{H}_k} \\
=& \bar{\kappa}_\epsilon \|(\hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}})\ell(\boldsymbol{\theta}, \cdot)\|_{\mathcal{H}_k} \\
\leq& \bar{\kappa}_\epsilon \left\|\hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}}\right\|_{HS} \left\|\ell(\boldsymbol{\theta}, \cdot)\right\|_{\mathcal{H}_\ell} \\
=& \bar{\kappa}_\epsilon \sqrt{\ell(\boldsymbol{\theta}, \boldsymbol{\theta})} \left\|\hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}}\right\|_{HS} \\
\leq& \bar{\kappa}_\epsilon \bar{\ell} \left\|\hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}}\right\|_{HS}.
\end{aligned} \tag{A.18}$$

□

Since $\gamma = \bar{\kappa}_\epsilon \bar{\ell}$ is independent of m, we apply lemma A.1 to establish the convergence. Since this upper bound does not depend on $\boldsymbol{\theta} \in \vartheta$ or $\mathbf{y} \in \mathcal{Y}$ and the metric is the absolute difference, this convergence is uniform as a function of both $\boldsymbol{\theta} \in \vartheta$ and $\mathbf{y} \in \mathcal{Y}$.

Alternatively, convergence guarantees for the KML can be established by its connection to the form of a GP regressor (GPR), leveraging frameworks and properties from a regression perspective. This connection is discussed briefly in appendix C.

A.6 Convergence Guarantees for Marginal Kernel Means Likelihood

Theorem A.4. *Assume $\ell(\boldsymbol{\theta}, \cdot) \in \text{image}(C_{\boldsymbol{\Theta}\boldsymbol{\Theta}})$. The marginal kernel means likelihood (MKML) $q(\mathbf{y})$ converges to the marginal likelihood $p_\epsilon(\mathbf{y})$ uniformly at rate $O_p((m\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}})$ as a function of $\mathbf{y} \in \mathcal{Y}$.*

Proof. We begin by writing the marginalization operation as an expectation over $p(\boldsymbol{\theta})$. This gives us $q(\mathbf{y}) := \int_\vartheta q(\mathbf{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} = \mathbb{E}[q(\mathbf{y}|\boldsymbol{\Theta})]$ and $p_\epsilon(\mathbf{y}) := \int_\vartheta p_\epsilon(\mathbf{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} = \mathbb{E}[p_\epsilon(\mathbf{y}|\boldsymbol{\Theta})]$. Consider the absolute difference between the MKML $q(\mathbf{y})$ and the marginal likelihood $p_\epsilon(\mathbf{y})$,

$$\begin{aligned}
|q(\mathbf{y}) - p_\epsilon(\mathbf{y})| &= |\mathbb{E}[q(\mathbf{y}|\boldsymbol{\Theta}) - p(\mathbf{y}|\boldsymbol{\Theta})]| \\
&\leq \mathbb{E}[|q(\mathbf{y}|\boldsymbol{\Theta}) - p(\mathbf{y}|\boldsymbol{\Theta})|] \\
&\leq \bar{\kappa}_\epsilon \mathbb{E}[\|\hat{\mu}_{\mathbf{X}|\boldsymbol{\Theta}=\boldsymbol{\Theta}} - \mu_{\mathbf{X}|\boldsymbol{\Theta}=\boldsymbol{\Theta}}\|_{\mathcal{H}_k}] \\
&= \bar{\kappa}_\epsilon \mathbb{E}[\|(\hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}})\ell(\boldsymbol{\Theta}, \cdot)\|_{\mathcal{H}_k}] \\
&\leq \bar{\kappa}_\epsilon \mathbb{E}[\|\hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}}\|_{HS} \|\ell(\boldsymbol{\Theta}, \cdot)\|_{\mathcal{H}_\ell}] \\
&= \bar{\kappa}_\epsilon \mathbb{E}[\|\hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}}\|_{HS} \sqrt{\ell(\boldsymbol{\Theta}, \boldsymbol{\Theta})}] \\
&= \bar{\kappa}_\epsilon \mathbb{E}[\sqrt{\ell(\boldsymbol{\Theta}, \boldsymbol{\Theta})}] \|\hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}}\|_{HS} \\
&\leq \bar{\kappa}_\epsilon \mathbb{E}[\bar{\ell}] \|\hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}}\|_{HS} \\
&= \bar{\kappa}_\epsilon \bar{\ell} \|\hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}}\|_{HS}
\end{aligned} \tag{A.19}$$

Since $\gamma = \bar{\kappa}_\epsilon \bar{\ell}$ is independent of m, we apply lemma A.1 to establish the convergence. Since this upper bound does not depend on $\mathbf{y} \in \mathcal{Y}$ and the metric is the absolute difference, this convergence is uniform as a function of $\mathbf{y} \in \mathcal{Y}$. □

A.7 Convergence Guarantees for Kernel Means Posterior

Theorem A.5. *Assume $\ell(\boldsymbol{\theta}, \cdot) \in \text{image}(C_{\boldsymbol{\Theta}\boldsymbol{\Theta}})$ and that there exists $\delta > 0$ such that $q(\mathbf{y}) \geq \delta$ for all $m \geq M$ where $M \in \mathbb{N}_+$. The kernel means posterior (KMP) $q(\boldsymbol{\theta}|\mathbf{y})$ converges pointwise to the posterior $p_\epsilon(\boldsymbol{\theta}|\mathbf{y})$ at rate $O_p((m\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}})$ as a function of $\boldsymbol{\theta} \in \vartheta$ and $\mathbf{y} \in \mathcal{Y}$. If $\sup_{\boldsymbol{\theta}\in\vartheta} p(\boldsymbol{\theta}) < \infty$ and $\sup_{\boldsymbol{\theta}\in\vartheta} p_\epsilon(\mathbf{y}|\boldsymbol{\theta}) < \infty$, then the convergence is uniform in $\boldsymbol{\theta} \in \vartheta$. If $\sup_{\mathbf{y}\in\mathcal{Y}} p_\epsilon(\boldsymbol{\theta}|\mathbf{y}) < \infty$, then the convergence is uniform in $\mathbf{y} \in \mathcal{Y}$.*

Proof. First, consider the density ratio between the approximate and true densities for the likelihood and marginal likelihood,

$$\left|\frac{q(\mathbf{y}|\boldsymbol{\theta})}{p_\epsilon(\mathbf{y}|\boldsymbol{\theta})} - 1\right| \leq \frac{1}{p_\epsilon(\mathbf{y}|\boldsymbol{\theta})}\left|q(\mathbf{y}|\boldsymbol{\theta}) - p_\epsilon(\mathbf{y}|\boldsymbol{\theta})\right| \leq \frac{\bar{\kappa}_\epsilon \bar{\ell}}{p_\epsilon(\mathbf{y}|\boldsymbol{\theta})}\left\|\hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}}\right\|_{HS}, \tag{A.20}$$

$$\left|\frac{q(\mathbf{y})}{p_\epsilon(\mathbf{y})} - 1\right| \leq \frac{1}{p_\epsilon(\mathbf{y})}\left|q(\mathbf{y}) - p_\epsilon(\mathbf{y})\right| \leq \frac{\bar{\kappa}_\epsilon \bar{\ell}}{p_\epsilon(\mathbf{y})}\left\|\hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}}\right\|_{HS}. \tag{A.21}$$

Now, consider the absolute difference between the KMP $q(\boldsymbol{\theta}|\mathbf{y})$ and the posterior $p_\epsilon(\boldsymbol{\theta}|\mathbf{y})$ for all $m > M$.

$$\begin{aligned}
\left|q(\boldsymbol{\theta}|\mathbf{y}) - p_\epsilon(\boldsymbol{\theta}|\mathbf{y})\right| =& \left|\frac{q(\mathbf{y}|\boldsymbol{\theta})}{q(\mathbf{y})} - \frac{p_\epsilon(\mathbf{y}|\boldsymbol{\theta})}{p_\epsilon(\mathbf{y})}\right| p(\boldsymbol{\theta}) \\
=& \left|\frac{q(\mathbf{y}|\boldsymbol{\theta})}{p_\epsilon(\mathbf{y}|\boldsymbol{\theta})} - \frac{q(\mathbf{y})}{p_\epsilon(\mathbf{y})}\right| \frac{p_\epsilon(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{|q(\mathbf{y})|} \\
=& \left|\left(\frac{q(\mathbf{y}|\boldsymbol{\theta})}{p_\epsilon(\mathbf{y}|\boldsymbol{\theta})} - 1\right) - \left(\frac{q(\mathbf{y})}{p_\epsilon(\mathbf{y})} - 1\right)\right| \frac{p_\epsilon(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{|q(\mathbf{y})|} \\
\leq& \left(\left|\frac{q(\mathbf{y}|\boldsymbol{\theta})}{p_\epsilon(\mathbf{y}|\boldsymbol{\theta})} - 1\right| + \left|\frac{q(\mathbf{y})}{p_\epsilon(\mathbf{y})} - 1\right|\right) \frac{p_\epsilon(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{|q(\mathbf{y})|} \\
\leq& \left(\frac{\bar{\kappa}_\epsilon \bar{\ell}}{p_\epsilon(\mathbf{y}|\boldsymbol{\theta})}\left\|\hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}}\right\|_{HS} + \frac{\bar{\kappa}_\epsilon \bar{\ell}}{p_\epsilon(\mathbf{y})}\left\|\hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}}\right\|_{HS}\right) \frac{p_\epsilon(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{|q(\mathbf{y})|} \\
\leq& \left(\bar{\kappa}_\epsilon \bar{\ell} p(\boldsymbol{\theta})\left\|\hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}}\right\|_{HS} + \bar{\kappa}_\epsilon \bar{\ell} p_\epsilon(\boldsymbol{\theta}|\mathbf{y})\left\|\hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}}\right\|_{HS}\right) \frac{1}{|q(\mathbf{y})|} \\
\leq& \bar{\kappa}_\epsilon \bar{\ell}\big(p(\boldsymbol{\theta}) + p_\epsilon(\boldsymbol{\theta}|\mathbf{y})\big)\left\|\hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}}\right\|_{HS} \frac{1}{|q(\mathbf{y})|} \\
\leq& \frac{\bar{\kappa}_\epsilon \bar{\ell}\big(p(\boldsymbol{\theta}) + p_\epsilon(\boldsymbol{\theta}|\mathbf{y})\big)}{\delta}\left\|\hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}}\right\|_{HS}.
\end{aligned} \tag{A.22}$$

Since $\gamma = \frac{\bar{\kappa}_\epsilon \bar{\ell}}{\delta}\big(p(\boldsymbol{\theta}) + p_\epsilon(\boldsymbol{\theta}|\mathbf{y})\big)$ is independent of m and the upper bound holds for all $m > M$, we apply lemma A.1 to establish the convergence. Since this upper bound does depend on $\boldsymbol{\theta} \in \vartheta$ and $\mathbf{y} \in \mathcal{Y}$ and the metric is the absolute difference, this convergence is pointwise as a function of $\boldsymbol{\theta} \in \vartheta$ and $\mathbf{y} \in \mathcal{Y}$.

Furthermore, if $\bar{p}_{\boldsymbol{\Theta}} := \sup_{\boldsymbol{\theta}\in\vartheta} p(\boldsymbol{\theta}) < \infty$ and $\bar{p}_{\mathbf{Y}|\boldsymbol{\Theta}} := \sup_{\boldsymbol{\theta}\in\vartheta} p_\epsilon(\mathbf{y}|\boldsymbol{\theta}) < \infty$, then

$$\begin{aligned}
p(\boldsymbol{\theta}) + p_\epsilon(\boldsymbol{\theta}|\mathbf{y}) \leq& \sup_{\boldsymbol{\theta}\in\vartheta}\big(p(\boldsymbol{\theta}) + p_\epsilon(\boldsymbol{\theta}|\mathbf{y})\big) \leq \sup_{\boldsymbol{\theta}\in\vartheta} p(\boldsymbol{\theta}) + \sup_{\boldsymbol{\theta}\in\vartheta} p_\epsilon(\boldsymbol{\theta}|\mathbf{y}) \\
\leq& \sup_{\boldsymbol{\theta}\in\vartheta} p(\boldsymbol{\theta}) + \frac{\sup_{\boldsymbol{\theta}\in\vartheta} p_\epsilon(\mathbf{y}|\boldsymbol{\theta}) \sup_{\boldsymbol{\theta}\in\vartheta} p(\boldsymbol{\theta})}{p_\epsilon(\mathbf{y})} \\
=& \bar{p}_{\boldsymbol{\Theta}} + \frac{\bar{p}_{\mathbf{Y}|\boldsymbol{\Theta}}\bar{p}_{\boldsymbol{\Theta}}}{p_\epsilon(\mathbf{y})}.
\end{aligned} \tag{A.23}$$

So, $\left|q(\boldsymbol{\theta}|\mathbf{y}) - p_\epsilon(\boldsymbol{\theta}|\mathbf{y})\right| \leq \frac{\bar{\kappa}_\epsilon \bar{\ell}}{\delta}\big(\bar{p}_{\boldsymbol{\Theta}} + \frac{\bar{p}_{\mathbf{Y}|\boldsymbol{\Theta}}\bar{p}_{\boldsymbol{\Theta}}}{p_\epsilon(\mathbf{y})}\big)\left\|\hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}}\right\|_{HS}$. Since the upper bound does not depend on $\boldsymbol{\theta} \in \vartheta$, the convergence is uniform as a function of in $\boldsymbol{\theta} \in \vartheta$.

Similarly, if $\bar{p}_{\boldsymbol{\Theta}|\mathbf{Y}} := \sup_{\mathbf{y}\in\mathcal{Y}} p_\epsilon(\boldsymbol{\theta}|\mathbf{y}) < \infty$, then $\left|q(\boldsymbol{\theta}|\mathbf{y}) - p_\epsilon(\boldsymbol{\theta}|\mathbf{y})\right| \leq \frac{\bar{\kappa}_\epsilon \bar{\ell}}{\delta}\big(p(\boldsymbol{\theta}) + \bar{p}_{\boldsymbol{\Theta}|\mathbf{Y}}\big)\left\|\hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}}\right\|_{HS}$. Since the upper bound does not depend on $\mathbf{y} \in \mathcal{Y}$, the convergence is uniform as a function of in $\mathbf{y} \in \mathcal{Y}$. □

A.8 Convergence Guarantees for Kernel Means Posterior Embedding

Theorem A.6. *Assume $\ell(\boldsymbol{\theta},\cdot) \in \text{image}(C_{\boldsymbol{\Theta}\boldsymbol{\Theta}})$ and that there exists $\delta > 0$ such that $q(\mathbf{y}) \geq \delta$ for all $m \geq M$ where $M \in \mathbb{N}_+$. The kernel means posterior embedding (KMPE) $\tilde{\mu}_{\boldsymbol{\Theta}|\mathbf{Y}=\mathbf{y}}$ converges in RKHS norm to the posterior mean embedding $\mu_{\boldsymbol{\Theta}|\mathbf{Y}=\mathbf{y}}$ at rate $O_p((m\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}})$.*

Proof. Since ℓ is a bounded kernel, let $\bar{\bar{\ell}} := \sup_{\boldsymbol{\theta}\in\vartheta} \sup_{\boldsymbol{\theta}'\in\vartheta} \ell(\boldsymbol{\theta},\boldsymbol{\theta}') > 0$. Note that this is not necessarily the same as $\bar{\ell} := \sup_{\boldsymbol{\theta}\in\vartheta} \ell(\boldsymbol{\theta},\boldsymbol{\theta})$. Consider the RKHS norm of the difference between KMPE $\tilde{\mu}_{\boldsymbol{\Theta}|\mathbf{Y}=\mathbf{y}}$ and the posterior mean embedding $\mu_{\boldsymbol{\Theta}|\mathbf{Y}=\mathbf{y}}$ for all $m > M$,

$$
\begin{aligned}
&\left\| \tilde{\mu}_{\boldsymbol{\Theta}|\mathbf{Y}=\mathbf{y}} - \mu_{\boldsymbol{\Theta}|\mathbf{Y}=\mathbf{y}} \right\|^2_{\mathcal{H}_\ell} \\
&= \left\| \int_\vartheta \ell(\boldsymbol{\theta},\cdot) q(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} - \int_\vartheta \ell(\boldsymbol{\theta},\cdot) p_\epsilon(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \right\|^2_{\mathcal{H}_\ell} \\
&= \left\| \int_\vartheta \ell(\boldsymbol{\theta},\cdot) \Big(q(\boldsymbol{\theta}|\mathbf{y}) - p_\epsilon(\boldsymbol{\theta}|\mathbf{y}) \Big) d\boldsymbol{\theta} \right\|^2_{\mathcal{H}_\ell} \\
&= \left\langle \int_\vartheta \ell(\boldsymbol{\theta},\cdot) \Big(q(\boldsymbol{\theta}|\mathbf{y}) - p_\epsilon(\boldsymbol{\theta}|\mathbf{y}) \Big) d\boldsymbol{\theta}, \int_\vartheta \ell(\boldsymbol{\theta}',\cdot) \Big(q(\boldsymbol{\theta}'|\mathbf{y}) - p_\epsilon(\boldsymbol{\theta}'|\mathbf{y}) \Big) d\boldsymbol{\theta}' \right\rangle_{\mathcal{H}_\ell} \\
&= \int_\vartheta \int_\vartheta \langle \ell(\boldsymbol{\theta},\cdot), \ell(\boldsymbol{\theta}',\cdot) \rangle_{\mathcal{H}_\ell} \Big(q(\boldsymbol{\theta}|\mathbf{y}) - p_\epsilon(\boldsymbol{\theta}|\mathbf{y}) \Big) \Big(q(\boldsymbol{\theta}'|\mathbf{y}) - p_\epsilon(\boldsymbol{\theta}'|\mathbf{y}) \Big) d\boldsymbol{\theta} d\boldsymbol{\theta}' \\
&= \int_\vartheta \int_\vartheta \ell(\boldsymbol{\theta},\boldsymbol{\theta}') \Big(q(\boldsymbol{\theta}|\mathbf{y}) - p_\epsilon(\boldsymbol{\theta}|\mathbf{y}) \Big) \Big(q(\boldsymbol{\theta}'|\mathbf{y}) - p_\epsilon(\boldsymbol{\theta}'|\mathbf{y}) \Big) d\boldsymbol{\theta} d\boldsymbol{\theta}' \\
&= \left| \int_\vartheta \int_\vartheta \ell(\boldsymbol{\theta},\boldsymbol{\theta}') \Big(q(\boldsymbol{\theta}|\mathbf{y}) - p_\epsilon(\boldsymbol{\theta}|\mathbf{y}) \Big) \Big(q(\boldsymbol{\theta}'|\mathbf{y}) - p_\epsilon(\boldsymbol{\theta}'|\mathbf{y}) \Big) d\boldsymbol{\theta} d\boldsymbol{\theta}' \right| \\
&\leq \int_\vartheta \int_\vartheta \Big| \ell(\boldsymbol{\theta},\boldsymbol{\theta}') \Big| \Big| q(\boldsymbol{\theta}|\mathbf{y}) - p_\epsilon(\boldsymbol{\theta}|\mathbf{y}) \Big| \Big| q(\boldsymbol{\theta}'|\mathbf{y}) - p_\epsilon(\boldsymbol{\theta}'|\mathbf{y}) \Big| d\boldsymbol{\theta} d\boldsymbol{\theta}' \\
&\leq \int_\vartheta \int_\vartheta \bar{\bar{\ell}}^2 \Big| q(\boldsymbol{\theta}|\mathbf{y}) - p_\epsilon(\boldsymbol{\theta}|\mathbf{y}) \Big| \Big| q(\boldsymbol{\theta}'|\mathbf{y}) - p_\epsilon(\boldsymbol{\theta}'|\mathbf{y}) \Big| d\boldsymbol{\theta} d\boldsymbol{\theta}' \\
&= \bar{\bar{\ell}}^2 \int_\vartheta \Big| q(\boldsymbol{\theta}|\mathbf{y}) - p_\epsilon(\boldsymbol{\theta}|\mathbf{y}) \Big| d\boldsymbol{\theta} \int_\vartheta \Big| q(\boldsymbol{\theta}'|\mathbf{y}) - p_\epsilon(\boldsymbol{\theta}'|\mathbf{y}) \Big| d\boldsymbol{\theta}' \\
&= \bar{\bar{\ell}}^2 \left(\int_\vartheta \Big| q(\boldsymbol{\theta}|\mathbf{y}) - p_\epsilon(\boldsymbol{\theta}|\mathbf{y}) \Big| d\boldsymbol{\theta} \right)^2 .
\end{aligned}
\tag{A.24}
$$

We now employ inequality (A.22) that was derived within the proof of theorem A.5,

$$
\begin{aligned}
\left\| \tilde{\mu}_{\boldsymbol{\Theta}|\mathbf{Y}=\mathbf{y}} - \mu_{\boldsymbol{\Theta}|\mathbf{Y}=\mathbf{y}} \right\|_{\mathcal{H}_\ell} &\leq \bar{\bar{\ell}} \int_\vartheta \Big| q(\boldsymbol{\theta}|\mathbf{y}) - p_\epsilon(\boldsymbol{\theta}|\mathbf{y}) \Big| d\boldsymbol{\theta} \\
&\leq \bar{\bar{\ell}} \int_\vartheta \frac{\bar{\kappa}_\epsilon \bar{\ell} \big(p(\boldsymbol{\theta}) + p_\epsilon(\boldsymbol{\theta}|\mathbf{y}) \big)}{\delta} \left\| \hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}} \right\|_{HS} d\boldsymbol{\theta} \\
&= \bar{\bar{\ell}} \left(\int_\vartheta \big(p(\boldsymbol{\theta}) + p_\epsilon(\boldsymbol{\theta}|\mathbf{y}) \big) d\boldsymbol{\theta} \right) \frac{\bar{\kappa}_\epsilon \bar{\ell}}{\delta} \left\| \hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}} \right\|_{HS} \\
&= \frac{2\bar{\kappa}_\epsilon \bar{\ell} \bar{\bar{\ell}}}{\delta} \left\| \hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}} \right\|_{HS} .
\end{aligned}
\tag{A.25}
$$

Since $\gamma = \frac{2\bar{\kappa}_\epsilon \bar{\ell} \bar{\bar{\ell}}}{\delta}$ is independent of m and the upper bound holds for all $m > M$, we apply lemma A.1 to establish the convergence under the RKHS norm. □

B Surrogate Densities

Instead of modeling the posterior mean embedding directly in a fashion similar to K-ABC, KR-ABC, and KBR, our approach begins by using CMEs to approximate the full likelihood (2.1) first as a surrogate likelihood, the KML. While the KML provides an asymptotically correct surrogate for the likelihood, for finitely many simulations the KML is not necessarily positive nor normalized. To make the KML compatible with MCMC-based or variational approaches would require further amendments to the KML, ranging from simple clipping $[q(\mathbf{y}|\boldsymbol{\theta})]^+$ or a positivity constraint in the empirical least-squares problem for the CME weights, since CMEs can be seen as the solution to a vector valued regression problem in the RKHS (Grünewälder et al., 2012). These amendments would however introduce further bias to the already biased likelihood approximation. While these biases vanishes asymptotically as the KML approaches a valid density due to theorem 3.1, the asymptotic behavior is rarely reached under limited simulations, which is the scenario of interest. Instead, KELFI performs inference by considering the surrogate posterior and its mean embedding defined directly from the KML.

Constructed from the KML, the KMP is also a surrogate density, although it is normalized. While the KMP is useful for finding maximum a posteriori (MAP) solutions and visualizing posterior uncertainties, we cannot directly sample from a surrogate density that is possibly non-positive. To address this, KELFI is motivated by super-sampling of general CMEs with kernel herding (Chen et al., 2010). Although mean embeddings are strictly positive for strictly positive kernels, when they are estimated from empirical CMEs, the resulting mean embedding may not be strictly positive (Song et al., 2009). Nevertheless, kernel herding can still obtain super-samples from CME estimates which effectively minimizes the maximum mean discrepancy (MMD) discrepancy between the original CME estimate and the new embedding formed from super-samples. This idea has been used to sample from conditional distributions through its empirical CME representation in kernel Monte Carlo filter (KMCF) (Kanagawa et al., 2016) and KR-ABC (Kajihara et al., 2018). Furthermore, super-samples are more informative than random samples, in the sense that empirical expectations under super-samples can potentially converge faster at $O(S^{-1})$ for S samples instead of $O(S^{-\frac{1}{2}})$ for random samples.

In general, surrogate densities can be seen as the "density" of a signed measure. Most of the properties of KMEs, including injectivity between mean embeddings and distributions, remain valid for signed measures. By defining an analogous form of mean embeddings for surrogate densities, KELFI arrives at a novel posterior mean embedding that is associated with a marginal surrogate likelihood for hyperparameter learning.

In all experiments we found that we did not need to clip the KML or KMP even though they are not guaranteed a-priori to be strictly positive. This is because we used an universal kernel such as a Gaussian kernel on both ϑ and $\mathcal{D}$ so that their RKHS is dense in their respective L^2 spaces (Carmeli et al., 2010). Because densities and likelihoods are often square-integrable, accurate estimations can be achieved. Finally, since we use kernel herding to super-sample the KMPE, the KMPE is not required to be positive to begin with.

C Connections and Future Work

The KML enables approximate likelihood queries at any $\boldsymbol{\theta} \in \vartheta$, even if simulation data is not available at the corresponding $\boldsymbol{\theta}$. By using the KML as a surrogate model for the true likelihood and accepting some modeling bias, we avoid requiring multiple expensive simulations at each query $\boldsymbol{\theta}$ that is used by many MCMC-based ABC approaches. In fact, as a function of $\boldsymbol{\theta}$ the KML $q(\mathbf{y}|\cdot)$ is the predictive mean of a GPR (Rasmussen and Williams, 2006) trained on observations $\{\boldsymbol{\theta}_j, \kappa_\epsilon(\mathbf{y}, \mathbf{x}_j)\}_{j=1}^m$ with a GP prior $\mathcal{GP}(0, \ell)$ and Gaussian likelihood $\mathcal{N}(\mathbf{0}, m\lambda I)$, since they admit the same resulting form. This connection could provide uncertainty estimates in the KML approximation of the likelihood via the GP predictive variance. It is possible to then use Bayesian optimization (BO) (Snoek et al., 2012) or active learning methods to guide the proposal prior π in a sequential learning fashion that will result in the more accurate KML approximations for a fixed number m of simulations.

While our posterior mean embedding (3.5) is closed-form and thus exact for the surrogate density $q(\boldsymbol{\theta}|\mathbf{y})$, it is an approximation to the mean embedding $\mu_{\boldsymbol{\Theta}|\mathbf{Y}=\mathbf{y}} := \int_\vartheta \ell(\boldsymbol{\theta}, \cdot) p_\epsilon(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}$ of the true soft posterior $p_\epsilon(\boldsymbol{\theta}|\mathbf{y}) \equiv p^{(\epsilon)}_{\boldsymbol{\Theta}|\mathbf{Y}}(\boldsymbol{\theta}|\mathbf{y})$, and converges in RKHS norm at the same rate as the KML. This is different in a subtle way to the CME of the posterior used by K-ABC, KR-ABC, and KBR, which in fact is an approximation to $\mu_{\boldsymbol{\Theta}|\mathbf{X}=\mathbf{y}} := \int_\vartheta \ell(\boldsymbol{\theta}, \cdot) p_{\boldsymbol{\Theta}|\mathbf{X}}(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}$, the mean embedding of $p_{\boldsymbol{\Theta}|\mathbf{X}}(\boldsymbol{\theta}|\mathbf{y})$, which avoids using the ϵ-kernel. A key difference is that there is no known associated marginal likelihood or approximations thereof for the direct posterior mean embedding, so cross validation is required for selecting the remaining kernel hyperparameters in K-ABC, KR-

ABC, and KBR. K-ABC also do not address sampling, although kernel herding can be readily applied in the same way. Kernel herding is applied to KBR in KMCF (Kanagawa et al., 2016) for resampling distributions represented as a CME. We believe it would be an interesting direction to investigate the relationships between the original empirical posterior mean embedding and the surrogate posterior mean embedding.

With regards to hyperparameter learning, in the KME literature, Bayesian learning of hyperparameters in marginal mean embeddings have been addressed through a different marginal likelihood approach by placing a GP prior on the embedding (Flaxman et al., 2016). However, a general approach for learning CME hyperparameters in a Bayesian framework remains an open question. Our simple surrogate density approach can be an alternative solution to the CME Bayesian hyperparameter learning problem, and may lead to interesting connections.

With regards to sampling, by super-sampling the surrogate posterior mean embedding, the number of posterior samples is decoupled from the number of simulations. This is unlike likelihood-free MCMC methods for which the algorithm guides the simulator queries at parameter values that is not necessarily drawn from the prior, but rather from proposals of a Markov chain. This avoids the problem of slow mixing that is inherent in MCMC methods, and make KELFI more suitable for multi-modal posteriors, which remains to be experimented upon.

D Experimental Details for Blowfly

Our experimental setup follows that of Wood (2010). We adopted the 10 summary statistics used in Meeds and Welling (2014), Moreno et al. (2016), and Park et al. (2016), which are the log of the mean of each quartile of $\{N_t/1000\}_{t=1}^T$ (4 statistics), the mean of each quartile of first-order differences of $\{N_t/1000\}_{t=1}^T$ (4 statistics), and the maximal peaks of smoothed $\{N_t\}_{t=1}^T$ with two different thresholds (2 statistics). We also use a diagonal Gaussian prior on $\log \boldsymbol{\theta}$ with means $[2, -1.5, 6, -1, -1, \log(15)]$ and standard deviations $[2, 0.5, 0.5, 1, 1, \log(5)]$. Notice that we have slightly modified the standard deviation to be broader to make the problem more challenging.

We describe the NMSE metric that is used to compare algorithms in our experiments. Before the experiments, we first obtain 10000 parameter samples from the prior and simulate summary statistics from each of them. We then calculate the MSEs of each simulated summary statistics against the observed summary statistic, and average them cross the 10000 samples. This is now a vector of 10 numbers, since we have an average MSE value for each summary statistic. Those are now the MSEs achieved under the prior. We chose 10000 parameter samples because at this point the MSEs for the prior has stabilized without much variance.

During each experiment, we compute the MSEs by averaging MSEs scores across 1000 simulations under the posterior mean or mode obtained from the algorithm. This also produces a vector of 10 numbers. We then divide the MSE of each statistic from the posterior by that from the prior computed earlier. This results in a vector of 10 numbers which is now the NMSE for the 10 summary statistics. Since now all 10 numbers are normalized errors with respect to the prior, we average these NMSE scores across the statistics for a final single NMSE score.

In this way, each statistic is normalized in the final average and a NMSE of 100% correspond to the performance of the prior. Hence, the NMSE measures the error as a percentage of the error achieved by the prior.

Note that this is the NMSE score for a particular experiment. For each algorithm, we further repeat the experiment and thus this calculation process 10 times and show the average and the deviations in fig. 3.

For all algorithms except KBR, we evaluate their performance by simulating from their posterior mean. For KBR only, we simulate from its posterior mode. This is because we noticed that KBR posterior mode decoding consistently outperformed KBR posterior mean for the Blowfly problem. Using the posterior mode will present KBR in its best light.

We now detail the hyperparameter choices for each algorithm other than KELFI, since most algorithms do not have a hyperparameter learning algorithm for the inference problem. Refer to their respective papers for a description of the meaning of each hyperparameter. For algorithms that use a MCMC proposal distribution, we choose a Gaussian proposal distribution with a proposal standard deviations that are 10% of the prior standard deviations. For MCMC-ABC, we used $\epsilon = 5$. For SL-ABC, we used $\epsilon = 0.5$ and $S = 10$. For ASL-ABC, we used $S_0 = 10$, $\epsilon = 0.5$, $\xi = 0.3$, $m = 10$, and $\Delta S = 10$. For GPS-ABC, we used $S_0 = 20$ samples from ASL-ABC to initialize the GP surrogate, and choose $\epsilon = 2$, $\xi = 0.05$, $m = 10$, and $\Delta S = 5$. For K-ABC and KBR, we used median length heuristic to set length scale hyperparameters, and choose $\lambda = 10^{-4}$. Note that KBR uses two kernels on both the parameter and the summary statistics and have two regularization hyperparameters.

E Experimental Details for Lotka-Volterra

For the Lotka-Volttera problem, our setup follows exactly as described in Papamakarios and Murray (2016). We simulate data using the ground truth parameters and treat this as the observational data, and use it across all experiments and algorithms.

In particular, the problem places a uniform prior over $\log \boldsymbol{\theta}$. Since the parameters are independent from each other in the prior, transforming the ABC task into one with a Gaussian prior is straight forward by doing it separately for each parameter. To convert from $\log \boldsymbol{\theta}$ to $\mathbf{z}$, denoting a realization of a Gaussian random variable, we first offset and scale it to a uniform in $[0, 1]$ then apply the standard normal quantile function. To convert it back, which is required before we pass our parameter query to the simulator or to present our results, we apply the standard normal cumulative distribution function and scale and offset the uniform back to its original ranges. Similar to the other experiments, we do not learn the prior hyperparameters in this paper to enable benchmarking against other methods with the same prior, so the transformed prior stay as a standard normal.

To apply the closed-form solutions for KELFI, we transform the prior samples into a standard Gaussian distributed samples, apply KELFI, and transform the posterior samples back to the original space for $\log \boldsymbol{\theta}$.

With a uniform prior and a complex intractable likelihood, the posterior is unlikely to be a Gaussian. KELFI does not assume that the posterior is a Gaussian and thus can provide more flexible and accurate posteriors. After learning appropriate hyperparameters for KELFI under MKML, we draw 10000 super-samples from the KMPE to compute the posterior mean, and maximize the KMP to compute the posterior mode. Finally, to compute the 95% credible interval, we compute the empirical 2.5% and 97.5% quantile using the 10000 super-samples.

F Gaussian Prior Transformations for Likelihood-Free Inference Problems

Under certain non-exhaustive conditions, we can always transform a particular LFI problem into another LFI problem that involves a Gaussian prior without loss of generality. These assumptions are that $p_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) = \prod_{d=1}^{D} p_{\Theta_d}(\theta_d)$ is a continuous probability density function (PDF) whose entries are independent, and that its inverse marginal cumulative distribution functions (CDFs) $P_{\Theta_d}^{-1}$ exists and is tractable.

In terms of notation, we denote the parameters as $\boldsymbol{\theta} = \{\theta_d\}_{d=1}^{D} \in \vartheta$ for D parameters. For this section only, multiple *iid* copies will be indexed by a superscript $\boldsymbol{\theta}^{(j)}$ for $j \in [m]$. Hence, the d-th parameter of the j-th parameter values is $\theta_d^{(j)}$. For densities, we use the corresponding random variable as the subscript to denote which distribution we are referring to. For example, we used $p(\boldsymbol{\theta})$ as the shorthand for the more formal notation of $p_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$ in the rest of the paper, but here we will keep the subscript to make this explicit.

Suppose the original prior $p_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$ is not necessarily Gaussian, but satisfies the aforementioned assumptions. Let $\mathbf{Z}$ be a random variable of the same dimensionality as $\boldsymbol{\Theta}$ with realization $\mathbf{z} \in \mathcal{Z}$. Let $p_{\mathbf{Z}}(\mathbf{z}) = \prod_{d=1}^{D} p_{Z_d}(z_d)$, where $p_{Z_d}(z_d) = \mathcal{N}(\mu_d, \sigma_d^2)$ so that its density is a multivariate anisotropic Gaussian. Convenient choices that simplify transformations are $\mu_d = 0$ and $\sigma_d = \sigma$ for all $d \in [D]$, although the general methodology remains.

Below we outline the general procedure for transforming a LFI problem into another LFI problem that involves a Gaussian prior.

1. Generate Gaussian samples $\mathbf{z}^{(j)} \sim p_{\mathbf{Z}}(\mathbf{z})$ for $j \in [m]$.

2. Convert Gaussian samples $\mathbf{z}$ into uniform samples $\mathbf{u}$ through $u_d^{(j)} = P_{Z_d}(z_d^{(j)})$ for $j \in [m]$ and $d \in [D]$.
   That is, $\mathbf{u}^{(j)} \sim U(0,1)^D$ for $j \in [m]$.

3. Convert uniform samples $\mathbf{u}$ into prior samples through $\theta_d^{(j)} = P_{\Theta_d}^{-1}(u_d^{(j)})$ for $j \in [m]$ and $d \in [D]$.
   The overall forward transformation is $\mathbf{T}(\mathbf{z}) := \{T_d(z_d)\}_{d=1}^{D}$ where $T_d(z_d) = P_{\Theta_d}^{-1}(P_{Z_d}(z_d))$.
   Since $P_{Z_d}^{-1}$ exists, the inverse transformation is $\mathbf{T}^{-1}(\boldsymbol{\theta}) = \{T_d^{-1}(\theta_d)\}_{d=1}^{D}$ where $T_d^{-1}(\theta_d) = P_{Z_d}^{-1}(P_{\Theta_d}(\theta_d))$.
   Hence, we have $\boldsymbol{\theta}^{(j)} = \mathbf{T}(\mathbf{z}^{(j)})$ for $j \in [m]$.

4. Run the simulator at the parameter samples $\mathbf{x}^{(j)} \sim p_{\mathbf{X}|\boldsymbol{\Theta}}(\cdot|\boldsymbol{\theta}^{(j)}) = p_{\mathbf{X}|\boldsymbol{\Theta}}(\cdot|T(\mathbf{z}^{(j)})) = p_{\mathbf{X}|\mathbf{Z}}(\cdot|\mathbf{z}^{(j)})$. We now have joint samples $\{\mathbf{z}^{(j)}, \mathbf{x}^{(j)}\}_{j=1}^{m}$.

5. Use the KELFI framework to approximate the posterior $p_{\mathbf{Z}|\mathbf{Y}}(\mathbf{z}|\mathbf{y})$ using the simulation pairs $\{\mathbf{z}^{(j)}, \mathbf{x}^{(j)}\}_{j=1}^m$. Either we obtain the KMP $q_{\mathbf{Z}|\mathbf{Y}}(\mathbf{z}|\mathbf{y})$, or we obtain KMPE super-samples $\{\hat{\mathbf{z}}_s\}_{s=1}^S$.

6. If we have samples $\{\hat{\mathbf{z}}_s\}_{s=1}^S$, then to obtain the corresponding samples for $q_{\mathbf{\Theta}|\mathbf{Y}}(\boldsymbol{\theta}|\mathbf{y})$, we simply pass the samples $\{\hat{\mathbf{z}}_s\}_{s=1}^S$ through the transformation $\mathbf{T}$ so that $\hat{\boldsymbol{\theta}}_s = \mathbf{T}(\hat{\mathbf{z}}_s)$ for $s \in [S]$.

7. If we have the KMP, then to obtain the corresponding posterior density we use the standard change of variable transformation $q_{\mathbf{\Theta}|\mathbf{Y}}(\boldsymbol{\theta}|\mathbf{y}) = q_{\mathbf{Z}|\mathbf{Y}}(\mathbf{T}^{-1}(\boldsymbol{\theta})|\mathbf{y})|\det J_{\mathbf{T}^{-1}}(\boldsymbol{\theta})|$.

   The Jacobian of $\mathbf{T}^{-1}$ is a $D \times D$ matrix whose (i,j)-th entry is $(J_{\mathbf{T}^{-1}}(\boldsymbol{\theta}))_{ij} := \frac{\partial T_i^{-1}}{\partial \theta_j}(\boldsymbol{\theta})$.

   Since the transformations of each parameter is done independently from each other, T_i^{-1} does not depend on θ_j if $i \neq j$. Consequently, the Jacobian is diagonal.

   The diagonal entries are $\frac{\partial T_i^{-1}}{\partial \theta_i}(\theta_i) = \frac{\partial}{\partial \theta_i} P_{Z_i}^{-1}(P_{\Theta_i}(\theta_i)) = (P_{Z_i}^{-1})'(P_{\Theta_i}(\theta_i))p_{\Theta_i}(\theta_i) = [p_{Z_i}(P_{Z_i}^{-1}(P_{\Theta_i}(\theta_i)))]^{-1}p_{\Theta_i}(\theta_i) = [p_{Z_i}(T_i^{-1}(\theta_i))]^{-1}p_{\Theta_i}(\theta_i)$. In the second last equality we made use of the fact that the computation of the derivative of the quantile function requires only the knowledge of the density and the quantile function itself, since $(P^{-1})'(u) = (P'(P^{-1}(u)))^{-1}$. Thus, the determinant of the Jacobian is $\det J_{\mathbf{T}^{-1}}(\boldsymbol{\theta}) = \prod_{i=1}^d [p_{Z_i}(T_i^{-1}(\theta_i))]^{-1}p_{\Theta_i}(\theta_i) = p_{\mathbf{\Theta}}(\boldsymbol{\theta})\big[\prod_{i=1}^d p_{Z_i}(T_i^{-1}(\theta_i))\big]^{-1} = p_{\mathbf{\Theta}}(\boldsymbol{\theta})\big[p_{\mathbf{Z}}(\mathbf{T}^{-1}(\boldsymbol{\theta}))\big]^{-1}$.

   The change of variable transformation becomes

$$q_{\mathbf{\Theta}|\mathbf{Y}}(\boldsymbol{\theta}|\mathbf{y}) = q_{\mathbf{Z}|\mathbf{Y}}(\mathbf{T}^{-1}(\boldsymbol{\theta})|\mathbf{y})\frac{p_{\mathbf{\Theta}}(\boldsymbol{\theta})}{p_{\mathbf{Z}}(\mathbf{T}^{-1}(\boldsymbol{\theta}))}. \tag{F.1}$$

   Finally, the form simplifies when the form of the KMP $q_{\mathbf{Z}|\mathbf{Y}}(\mathbf{T}^{-1}(\boldsymbol{\theta})|\mathbf{y})$ is substituted back in,

$$q_{\mathbf{\Theta}|\mathbf{Y}}(\boldsymbol{\theta}|\mathbf{y}) = \frac{q_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y}|\mathbf{T}^{-1}(\boldsymbol{\theta}))p_{\mathbf{Z}}(\mathbf{T}^{-1}(\boldsymbol{\theta}))}{q_{\mathbf{Y}}(\mathbf{y})}\frac{p_{\mathbf{\Theta}}(\boldsymbol{\theta})}{p_{\mathbf{Z}}(\mathbf{T}^{-1}(\boldsymbol{\theta}))} = \frac{q_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y}|\mathbf{T}^{-1}(\boldsymbol{\theta}))p_{\mathbf{\Theta}}(\boldsymbol{\theta})}{q_{\mathbf{Y}}(\mathbf{y})}. \tag{F.2}$$

   Note that the MKML $q_{\mathbf{Y}}(\mathbf{y})$ is still marginalized over the simpler Gaussian distribution,

$$q_{\mathbf{Y}}(\mathbf{y}) = \int_{\mathcal{Z}} q_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y}|\mathbf{z})p_{\mathbf{Z}}(\mathbf{z})d\mathbf{z}. \tag{F.3}$$

In this way, we simplify the LFI problem into another LFI problem which involves a Gaussian prior such that KELFI solutions are closed-form under Gaussian kernels. Once KELFI solutions have been computed in the new parameter space $\mathcal{Z}$, the solutions can be easily transformed back into the original parameter space ϑ as above.

This process is possible since the likelihood is intractable already. Hence, transformations $\mathbf{T}$ of variables $\mathbf{z}$ into simulator parameters $\boldsymbol{\theta}$ can be included as part of the simulator without changing the nature of the problem.

If simulation pairs $\{\boldsymbol{\theta}^{(j)}, \mathbf{x}^{(j)}\}_{j=1}^m$ in the original space are already provided, parameters $\boldsymbol{\theta}^{(j)}$ can be converted into Gaussian variables via $\mathbf{z}^{(j)} = \mathbf{T}^{-1}(\boldsymbol{\theta}^{(j)})$ for $j \in [m]$ so that the pairs $\{\mathbf{z}^{(j)}, \mathbf{x}^{(j)}\}_{j=1}^m$ can be used to proceed.

As an extension, instead of transforming the LFI problem with a general continuous prior into one with a Gaussian prior, if the prior is fundamentally multi-modal, we can also transform it into one with a Gaussian mixture model as the prior. Since the prior density is a linear combination of Gaussians, all derivations remain closed-form from a linear combination of the results with each Gaussian component.

Finally, it is important to recognize that while there is no loss of generality to the inference problem when performing this prior transform, the transformation do change the interpretation of the hyperparameters learned with the MKML. Since the kernel ℓ is now placed in the $\mathcal{Z}$ space, the hyperparameters of ℓ cannot be interpreted directly for the original parameter space ϑ unless the transformation between $\mathcal{Z}$ and ϑ is simple enough to translate the interpretation. Nevertheless, hyperparameters can still be learned by optimizing the MKML.