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

Kelvin Hsu and **Fabio Ramos** {Kelvin.Hsu, Fabio.Ramos}@sydney.edu.au
School of Computer Science, University of Sydney and *Data61, CSIRO, Sydney, Australia*

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 ϵ 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 approximate marginal likelihood. Our framework demonstrates improved accuracy and efficiency on challenging inference problems in ecology.

Keywords: Conditional kernel mean embeddings, approximate Bayesian inference, likelihood-free inference, intractable likelihoods, hyperparameter learning, posterior super-sampling

1. Introduction

Approximate Bayesian computation (ABC) methods are the state-of-the-art for simulation-based Bayesian inference (Marin et al., 2012). However, they can be expensive, and rely on discrepancy measures that are parametrized by hyperparameters such as ϵ, which are difficult to tune. 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. KELFI leverages likelihood smoothness within a reproducing kernel Hilbert space (RKHS) using conditional mean embeddings (CMEs), encoding conditional expectations empirically with only a few simulations.

2. Background: Likelihood-Free Inference

We begin with a forward simulator model $p(\mathbf{x}|\boldsymbol{\theta})$ which synthesizes simulations $\mathbf{x}$ given a parameters $\boldsymbol{\theta}$. Importantly, the simulator generates samples but not likelihood evaluations, making the likelihood intractable, or *likelihood-free*. Let $\boldsymbol{\theta} \in \vartheta$ denote realizations of parameters $\boldsymbol{\Theta}$. Let $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$ where $\mathcal{X}, \mathcal{Y} \subseteq \mathcal{D}$ denote realizations of simulations $\mathbf{X}$ and observations $\mathbf{Y}$ respectively. We posit a prior $p(\boldsymbol{\theta})$ and measure simulation discrepancy by an ϵ-kernel $p_{\epsilon}(\mathbf{y}|\mathbf{x}) = \kappa_{\epsilon}(\mathbf{y}, \mathbf{x})$, such as a Gaussian $\mathcal{N}(\mathbf{y}|\mathbf{x}, \epsilon^2 I)$ (Moreno et al., 2016). Based on

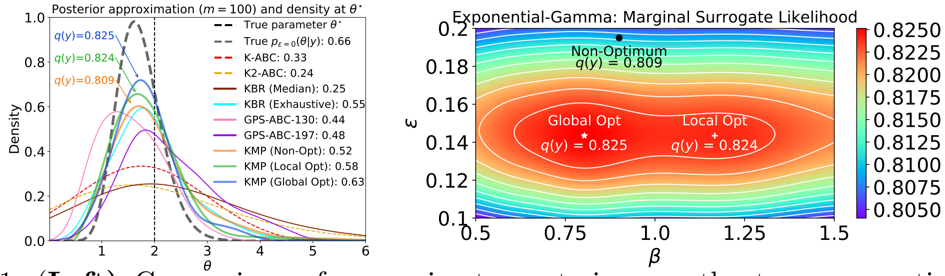

Figure 1: **(Left)** Comparison of approximate posteriors on the toy exponential-gamma problem and effect of hyperparameter learning on KMP. Density values at the true parameter $\theta^\star$ is recorded in the legend. **(Right)** The corresponding MKML surface $q(\mathbf{y})$ as a function of (ϵ, β) for fixed $\lambda = 10^{-4}$.

this formulation, the true full likelihood is $p_\epsilon(\mathbf{y}|\boldsymbol{\theta}) = \int_{\mathcal{X}} p_\epsilon(\mathbf{y}|\mathbf{x})p(\mathbf{x}|\boldsymbol{\theta})d\mathbf{x} = \mathbb{E}[\kappa_\epsilon(\mathbf{y}, \mathbf{X})|\boldsymbol{\Theta} = \boldsymbol{\theta}]$. The posterior is $p_\epsilon(\boldsymbol{\theta}|\mathbf{y}) = p_\epsilon(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})/p_\epsilon(\mathbf{y})$ where $p_\epsilon(\mathbf{y}) = \int_\vartheta p_\epsilon(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$.

3. Kernel Embedding Likelihood-Free Inference

We present KELFI in three stages – model, learning, and inference. We assume that the prior is an anisotropic Gaussian $p(\boldsymbol{\theta}) = \prod_{d=1}^{D} \mathcal{N}(\theta_d|\mu_d, \sigma_d^2)$. For most continuous priors, the ABC problem can be transformed into an equivalent ABC problem with a Gaussian prior. See section 14 for detail. We specifically focus on the case where 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 $\boldsymbol{\theta}_j$. KELFI uses 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 $\ell(\boldsymbol{\theta}, \boldsymbol{\theta}'; \boldsymbol{\beta}) = \exp\big(-\frac{1}{2}\sum_{d=1}^{D}(\theta_d - \theta'_d)^2/\beta_d^2\big)$ with length scales $\boldsymbol{\beta} = \{\beta_d\}_{d=1}^D$. For $f \in \mathcal{H}_k$, we construct an approximation to $\mathbb{E}[f(\mathbf{X})|\boldsymbol{\Theta} = \boldsymbol{\theta}]$ by $\langle f, \hat{\mu}_{\mathbf{X}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\rangle_{\mathcal{H}_k}$ 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 form $\mathbb{E}[f(\mathbf{X})|\boldsymbol{\Theta} = \boldsymbol{\theta}] \approx \mathbf{f}^T(L+m\lambda I)^{-1}\boldsymbol{\ell}(\boldsymbol{\theta})$, 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 (Song et al., 2009).

3.1. Model: Kernel Means Likelihood

Since the likelihood is an expectation under $p(\mathbf{x}|\boldsymbol{\theta})$, 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). By using $f = \kappa_\epsilon(\mathbf{y}, \cdot)$, $\boldsymbol{\kappa}_\epsilon(\mathbf{y}) := \{\kappa_\epsilon(\mathbf{y}, \mathbf{x}_j)\}_{j=1}^m$, and $\mathbf{v}(\mathbf{y}) := (L + m\lambda I)^{-1}\boldsymbol{\kappa}_\epsilon(\mathbf{y})$, the KML becomes

$$q(\mathbf{y}|\boldsymbol{\theta}) := \langle \kappa_\epsilon(\mathbf{y}, \cdot), \hat{\mu}_{\mathbf{X}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\rangle_{\mathcal{H}_k} = \boldsymbol{\kappa}_\epsilon(\mathbf{y})^T(L + m\lambda I)^{-1}\boldsymbol{\ell}(\boldsymbol{\theta}) = \sum_{j=1}^{m} v_j(\mathbf{y})\ell(\boldsymbol{\theta}_j, \boldsymbol{\theta}). \qquad (1)$$

The KML converges at the same rate as the CME. See theorem 3 for proof. 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

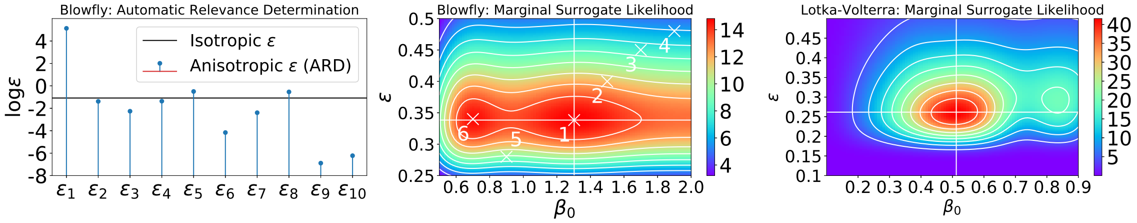

Figure 2: **(Left)** ARD on $\boldsymbol{\epsilon}$ for 10 summary statistics. **(Mid. & Right)** The MKML surface ($\times 10^5$) as a function of (ϵ, β_0) for fixed $\lambda = 10^{-4}$ where $\boldsymbol{\beta} = \beta_0 \boldsymbol{\sigma}$. White intersection indicate optimum. For Blowfly, the NMSE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

$k_\alpha = k_\epsilon$ 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.

3.2. Learning: Marginal Kernel Means Likelihood

The advantage of a surrogate likelihood model is that it readily provides a marginal surrogate likelihood quantity. We define the marginal kernel means likelihood (MKML) as

$$q(\mathbf{y}) := \int_\vartheta q(\mathbf{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} = \sum_{j=1}^{m} v_j(\mathbf{y}) \int_\vartheta \ell(\boldsymbol{\theta}_j, \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} = \sum_{j=1}^{m} v_j(\mathbf{y}) \mu_{\boldsymbol{\Theta}}(\boldsymbol{\theta}_j), \tag{2}$$

where $\mu_{\boldsymbol{\Theta}} := \int_\vartheta \ell(\boldsymbol{\theta}, \cdot) p(\boldsymbol{\theta}) d\boldsymbol{\theta}$. If we choose ℓ to be anisotropic Gaussian with length scales $\boldsymbol{\beta} = \{\beta_d\}_{d=1}^{D}$, then $\mu_{\boldsymbol{\Theta}}$ has closed-form $\mu_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) = \ell_{\boldsymbol{\nu}}(\boldsymbol{\theta}, \boldsymbol{\mu}) \prod_{d=1}^{D} \frac{\beta_d}{\nu_d}$ where with $\nu_d^2 := \beta_d^2 + \sigma_d^2$.

The MKML converges at the same rate as the CME. See theorem 4 for proof. The MKML $q(\mathbf{y}) = q(\mathbf{y}; \boldsymbol{\epsilon}, \boldsymbol{\beta}, \lambda)$ is a differentiable function of the hyperparameters $(\boldsymbol{\epsilon}, \boldsymbol{\beta}, \lambda)$. We can thus perform approximate maximum marginal likelihood for hyperparameter learning.

3.3. Inference: Kernel Means Posterior Embedding and Super-Sampling

Finally, inference begins by defining a posterior surrogates for $p_\epsilon(\boldsymbol{\theta}|\mathbf{y})$ in analogy to the Bayes' rule, $q(\boldsymbol{\theta}|\mathbf{y}) := q(\mathbf{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) / q(\mathbf{y})$, referred as the kernel means posterior (KMP), which converges under well behaved marginal surrogate likelihoods. See theorem 5 for proof. We now define kernel means posterior embedding (KMPE), the mean embedding of the KMP, as $\tilde{\mu}_{\boldsymbol{\Theta}|\mathbf{Y}=\mathbf{y}}(\boldsymbol{\theta}^\star) := \int_\vartheta \ell(\boldsymbol{\theta}, \boldsymbol{\theta}^\star) q(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}$. Importantly, it converges in RKHS norm at the same CME rate. See theorem 6 for proof. With $h(\boldsymbol{\theta}_j, \boldsymbol{\theta}^\star) := \int_\vartheta \ell(\boldsymbol{\theta}, \boldsymbol{\theta}^\star) \ell(\boldsymbol{\theta}_j, \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}$,

$$\tilde{\mu}_{\boldsymbol{\Theta}|\mathbf{Y}=\mathbf{y}}(\boldsymbol{\theta}^\star) = \frac{1}{q(\mathbf{y})} \sum_{j=1}^{m} v_j(\mathbf{y}) h(\boldsymbol{\theta}_j, \boldsymbol{\theta}^\star), \tag{3}$$

If we choose ℓ to be anisotropic Gaussian with length scales $\boldsymbol{\beta} = \{\beta_d\}_{d=1}^{D}$, h becomes

$$h(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) = \prod_{d=1}^{D} \frac{s_d}{\sigma_d} \exp\left[-\frac{1}{2 s_d^2} \left(\frac{\theta_d^2 + \tilde{\theta}_d^2 + \gamma_d^2 \mu_d^2}{2 + \gamma_d^2} - \left(\frac{\theta_d + \tilde{\theta}_d + \gamma_d^2 \mu_d}{2 + \gamma_d^2} \right)^2 \right) \right], \tag{4}$$

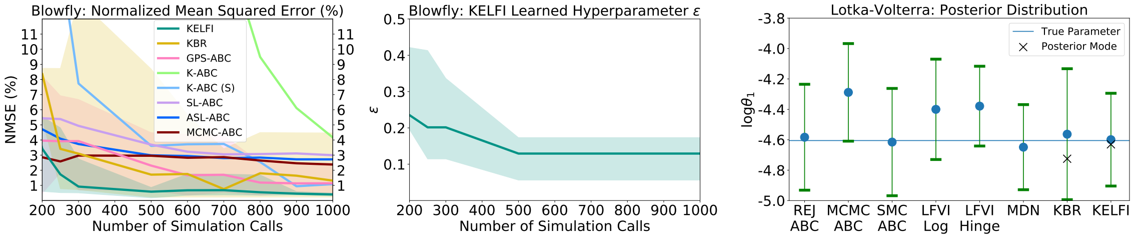

Figure 3: **(Left)** Average NMSE (in %) under posteriors v.s. simulations. Shaded regions show variability for KELFI, KBR, and GPS-ABC. **(Mid.)** Learned ϵ under maximum MKML. **(Right)** Marginal posterior of $\log \theta_1$ within 2 standard deviations.

where $\gamma_d^2 := \beta_d^2/\sigma_d^2$ and $s_d^{-2} := 2\beta_d^{-2} + \sigma_d^{-2}$. Note that $q(\cdot|\mathbf{y})$ is bounded and normalized but potentially non-positive. Consequently, we can see it as a surrogate density corresponding to a signed measure. This suggests that the map $q(\cdot|\mathbf{y}) \mapsto \tilde{\mu}_{\mathbf{\Theta}|\mathbf{Y}=\mathbf{y}}$ is injective for characteristic kernels ℓ, analogous to mean embeddings (Sriperumbudur et al., 2011). Furthermore, as the integral (3) is a linear operator on $\ell(\boldsymbol{\theta}^\star, \cdot)$, the surrogate posterior mean embedding $\tilde{\mu}_{\mathbf{\Theta}|\mathbf{Y}=\mathbf{y}} \in \mathcal{H}_\ell$ is in the RKHS of ℓ. Hence, we can apply kernel herding (Chen et al., 2010) on $\tilde{\mu}_{\mathbf{\Theta}|\mathbf{Y}=\mathbf{y}}$ (3) using kernel ℓ to obtain S super samples $\{\boldsymbol{\theta}^{(s)}\}_{s=1}^S$ from the surrogate density $q(\boldsymbol{\theta}|\mathbf{y})$. That is, for each $s \in \{1, \dots, S\}$, we have $\boldsymbol{\theta}^{(s)} = \mathrm{argmax}_{\boldsymbol{\theta} \in \vartheta}\, \tilde{\mu}_{\mathbf{\Theta}|\mathbf{Y}=\mathbf{y}}(\boldsymbol{\theta}) - \frac{1}{s}\sum_{i=1}^{s-1} \ell(\boldsymbol{\theta}^{(i)}, \boldsymbol{\theta})$. The inference algorithm is presented in algorithm 1.

4. Experiments

We compare KELFI on a standard toy problem in fig. 1 and two challenging chaotic ecological systems in fig. 2 and fig. 3 – Blowfly and Lotka-Volterra. Figure 1 shows that KMP under KELFI outperforms other surrogate based methods in both posterior approximation and density at true parameter (left), and inference improves as we learn hyperparameters under by maximizing the MKML (right with left). Figure 2 show a similar accuracy improvement on the challenging Blowfly simulator (middle), and also using automatic relevance determination (ARD) to determine that the last two summary statistics are most useful which agrees with the intuition from Wood (2010) (left). Figure 3 demonstrate that KELFI outperforms other methods consistently in both Blowfly (left) and Lotka-Volterra (right), and learns an ϵ-decay schedule automatically (middle). For more detail see section 7.

5. Conclusion and Future Work

KELFI 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. Since the samples $\boldsymbol{\theta}_j \sim \pi(\boldsymbol{\theta})$ do not have the be from the prior $p(\boldsymbol{\theta})$, to further reduce simulation requirements it is possible to choose or adapt π during the simulation process in a way that focuses on high likelihood regions. As future work, this can potentially be achieved by applications of Bayesian optimization.

References

Claudio Carmeli, Ernesto De Vito, Alessandro Toigo, and Veronica Umanitá. Vector valued reproducing kernel hilbert spaces and universality. *Analysis and Applications*, 8(01):19–61, 2010.

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

Seth Flaxman, Dino Sejdinovic, John P Cunningham, and Sarah Filippi. Bayesian learning of kernel embeddings. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, pages 182–191. AUAI Press, 2016.

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

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

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

Takafumi Kajihara, Keisuke Yamazaki, Motonobu Kanagawa, and Kenji Fukumizu. Kernel Recursive ABC: Point Estimation with Intractable Likelihood. *arXiv preprint arXiv:1802.08404*, 2018.

Motonobu Kanagawa, Yu Nishiyama, Arthur Gretton, and Kenji Fukumizu. Filtering with state-observation examples via kernel monte carlo filter. *Neural computation*, 28(2):382–444, 2016.

Jean-Michel Marin, Pierre Pudlo, Christian P Robert, and Robin J Ryder. Approximate Bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180, 2012.

Paul Marjoram, John Molitor, Vincent Plagnol, and Simon Tavaré. Markov chain monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26): 15324–15328, 2003.

Edward Meeds and Max Welling. GPS-ABC: Gaussian process surrogate approximate Bayesian computation. *arXiv preprint arXiv:1401.2838*, 2014.

Jovana Mitrovic, Dino Sejdinovic, and Yee Whye Teh. DR-ABC: Approximate Bayesian Computation with kernel-based distribution regression. 2016.

Alexander Moreno, Tameem Adel, Edward Meeds, James M Rehg, and Max Welling. Automatic variational ABC. *arXiv preprint arXiv:1606.08549*, 2016.

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

Victor MH Ong, David J Nott, Minh-Ngoc Tran, Scott A Sisson, and Christopher C Drovandi. Variational Bayes with synthetic likelihood. *Statistics and Computing*, 28 (4):971–988, 2018.

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

Mijung Park, Wittawat Jitkrittum, and Dino Sejdinovic. K2-ABC: Approximate Bayesian computation with kernel embeddings. 2016.

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

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian processes for machine learning.* The MIT Press, 2006.

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

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012.

Le Song, Jonathan Huang, Alex Smola, and Kenji Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 961–968. ACM, 2009.

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

Bharath K Sriperumbudur, Kenji Fukumizu, and Gert RG Lanckriet. On the relation between universality, characteristic kernels and RKHS embedding of measures. In *AISTATS*, pages 773–780, 2010.

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

Dustin Tran, Rajesh Ranganath, and David M Blei. Hierarchical Implicit Models and Likelihood-free Variational Inference. *arXiv preprint arXiv:1702.08896*, 2017a.

Minh-Ngoc Tran, David J Nott, and Robert Kohn. Variational Bayes with intractable likelihood. *Journal of Computational and Graphical Statistics*, 26(4):873–882, 2017b.

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

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

Algorithm 1 KELFI: Kernel Embedding Likelihood-Free Inference

```
1: Input: Data $\mathbf{y}$, simulations $\{\boldsymbol{\theta}_j, \mathbf{x}_j\}_{j=1}^m \sim p(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$, parameter samples $\{\tilde{\boldsymbol{\theta}}_i\}_{i=1}^N$, KML hyperparameters $(\boldsymbol{\epsilon}, \boldsymbol{\beta}, \lambda)$, prior hyperparameters $(\boldsymbol{\mu}, \boldsymbol{\sigma})$, desired number of super samples $S$, Gaussian kernel $\ell$ and $\epsilon$-kernel $\kappa$
2: Compute $\mathbf{v} \leftarrow (L + m\lambda I)^{-1} \boldsymbol{\kappa}_{\boldsymbol{\epsilon}}(\mathbf{y})$, $L \leftarrow \{\ell_{\boldsymbol{\beta}}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)\}_{i,j=1}^m$, $\boldsymbol{\kappa}_{\boldsymbol{\epsilon}}(\mathbf{y}) \leftarrow \{\kappa_{\boldsymbol{\epsilon}}(\mathbf{y}, \mathbf{x}_j)\}_{j=1}^m$
3: Compute marginal surrogate likelihood $q(\mathbf{y}) \leftarrow \mathbf{v}^T \boldsymbol{\mu}_{\boldsymbol{\Theta}}$ where $\boldsymbol{\mu}_{\boldsymbol{\Theta}} := \{\mu_{\boldsymbol{\Theta}}(\boldsymbol{\theta}_j)\}_{j=1}^m$
4: Compute $H \leftarrow \{h(\tilde{\boldsymbol{\theta}}_i, \boldsymbol{\theta}_j)\}_{i=1,j=1}^{N,m}$ using (4)
5: Compute posterior mean embedding $\boldsymbol{\mu} \leftarrow H\mathbf{v}/q(\mathbf{y}) \in \mathbb{R}^N$ and initialize $\mathbf{m} \leftarrow \mathbf{0} \in \mathbb{R}^N$
6: for $s \in \{1, \dots, S\}$ do
7:     Obtain super sample $\boldsymbol{\theta}^{(s)} \leftarrow \tilde{\boldsymbol{\theta}}_i$ where $i \leftarrow \mathrm{argmax}_{i \in \{1,\dots,N\}} \mu_i - (m_i/s)$
8:     Update kernel sum $\mathbf{m} \leftarrow \mathbf{m} + \{\ell_{\boldsymbol{\beta}}(\tilde{\boldsymbol{\theta}}_i, \boldsymbol{\theta}^{(s)})\}_{i=1}^N$
9: end for
10: Output: Posterior super samples $\{\boldsymbol{\theta}^{(s)}\}_{s=1}^S$
```

6. Clarifications

Algorithm Algorithm 1 show the inference algorithm. To perform hyperparameter learning, optimize the MKML objective computed on line 3.

Likelihood-Free Inference 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 $p_{\epsilon=0}(\boldsymbol{\theta}|\mathbf{y})$. 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 $\epsilon = 0$.

Simulator Samples The simulator samples are not necessarily from the original joint distribution $p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ if $\pi \neq p_{\boldsymbol{\Theta}}$.

Kernel Means Likelihood It is worthwhile to note that the assumption on the input kernels $\ell(\boldsymbol{\theta}, \cdot) \in \mathrm{image}(C_{\boldsymbol{\Theta}\boldsymbol{\Theta}})$ 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).

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.

Marginal Kernel Means Likelihood Each automatically computed gradient step takes $O(m^3)$ to compute due to the Cholesky decomposition of the regularized gram matrix in line 2. Since we are addressing the scenario where simulations are limited, m is usually in the hundreds or at most thousands, making this optimization 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 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 prior information regarding the scale differences within $\boldsymbol{\theta}$, and let β_0 learn the overall scale that is most useful.

Kernel Means Posterior 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 4 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 5 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 4, 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.

Importantly, $q(\boldsymbol{\theta}|\mathbf{y})$ is unaffected even if κ_ϵ is unnormalized, so that ϵ-kernels on distributions can be readily used.

7. Experiments

The goal of the experiments is to demonstrate the inference accuracy of KELFI under limited simulation budget and the usefulness 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 (2) while keeping $\lambda = 10^{-4}$ fixed for simplicity.

7.1. 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 $\epsilon = 0$.

We first describe the experiment. To stress-test our algorithm, we use only $m = 100$ simulations to emphasize the inference quality under the setting with limited simulations. We focus on comparing surrogate approaches. Other methods such as rejection ABC (REJ-ABC), Markov chain Monte Carlo (MCMC)-ABC, synthetic likelihood ABC (SL-ABC), and adaptive SL-ABC (ASL-ABC) have reported simulation requirements several orders higher than 100 on this problem (Meeds and Welling, 2014). For GP surrogate ABC (GPS-ABC) only we set a simulation budget $m \leq 200$ and run it until either 10000 posterior samples are generated or the simulation budget is reached, and plot its posterior density using kernel density estimation (KDE) since it only produces samples. For hyperparameters, we used standard median heuristic for kernel ABC (K-ABC), K2-ABC, and kernel Bayes' rule (KBR). We also find optimal hyperparameters for KBR by exhaustive search. The hyperparameters of the Gaussian process (GP) surrogate itself used in GPS-ABC are learned by maximizing the marginal likelihood of the GP regressor (GPR) (Rasmussen and Williams, 2006). However, for hyperparameters of GPS-ABC that are not part of the surrogate, we select them based on the original paper (Meeds and Welling, 2014). We then report its best two results which used $m = 130$ and $m = 197$ simulations. Finally, for KELFI we show KMPs using globally, locally, and non-optimal hyperparameters under $q(\mathbf{y})$.

In fig. 1 we compare the approximate posteriors from each method with the true posterior $p_{\epsilon=0}(\boldsymbol{\theta}|\mathbf{y})$. While $\epsilon = 0$ for the true posterior, with only 100 simulations $\epsilon > 0$ is required for posterior approximations. Consequently, most methods produce approximations wider than $p_{\epsilon=0}(\boldsymbol{\theta}|\mathbf{y})$. Intuitively, this is because there is not enough simulations or 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. Consequently, KMP is the closest to the true posterior $p_{\epsilon=0}(\boldsymbol{\theta}|\mathbf{y})$ in spread, with higher MKML $q(\mathbf{y})$ leading to a more accurate KMP $q(\boldsymbol{\theta}|\mathbf{y})$. This demonstrates the effectiveness of a marginal surrogate likelihood objective for hyperparameter learning to improve inference. In contrast, the two instances for each of GPS-ABC and KBR show 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 K-ABC and K2-ABC and KBR (median) which uses the median heuristic to set hyperparameters. Since this heuristic makes no reference to the inference problem, the resulting hyperparameters do not achieve accurate posteriors.

Finally, we compare the density values at the true parameter $\boldsymbol{\theta}^\star$ for each method, which reveals that KELFI outperform remaining methods. In particular, while KBR with expensive and exhaustive hyperparameter search was able to outperform KMP with non-optimal hyperparameters, both globally and locally optimal KMPs under KELFI significantly outperform KBR.

7.2. 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 MSEs 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. Section 12 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 example 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 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 usefulness 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 (2) 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, and not its troughs (Wood, 2010).

7.3. 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, likelihood-free variational inference (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.

8. Related Work

The simplest ABC method is arguably the REJ-ABC sampler (Pritchard et al., 1999). This posits a set of prior parameters and rejects those whose simulations do not match the observations within a 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, 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 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. K-ABC (Nakagome et al., 2013), kernel recursive ABC (KR-ABC) (Kajihara et al., 2018), and 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 current CME based approaches. 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 design a posterior sampling algorithm, which subsumes expectation inference. We further provide approximate posterior density KMP, which can both produce point estimates and quantify uncertainty.

As a consequence of theorem 6, 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 $p_\epsilon(\boldsymbol{\theta}|\mathbf{y}) \approx p(\boldsymbol{\theta}|\mathbf{y})$ 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 GP literature, hyperparameter learning through maximum marginal likelihood plays an important role in the success of a GPR. 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.

9. Theoretical Guarantees on Convergence

We provide theoretical guarantees that establish convergence of the kernel embedding likelihood-free inference (KELFI) framework. Section 9.1 begins by summarizing the properties of kernels used in KELFI and introducing relevant quantities. Sections 9.2 and 9.3 provide an overview of conditional mean embeddings (CMEs) and their empirical estimates respectively in the context of KELFI. Section 9.4 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 sections 9.5, 9.6, 9.7 and 9.8 respectively.

9.1. 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 an observed data $\mathbf{y}$ 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 data $\mathbf{x}$ than that is possible to observe, such as when $\mathcal{Y} \subset \mathcal{X}$. 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,

$$\bar{\ell} := \sup_{\boldsymbol{\theta} \in \vartheta} \|\ell(\boldsymbol{\theta}, \cdot)\|_{\mathcal{H}_\ell} = \sup_{\boldsymbol{\theta} \in \vartheta} \sqrt{\ell(\boldsymbol{\theta}, \boldsymbol{\theta})}, \tag{5}$$

$$\bar{k} := \sup_{\mathbf{d} \in \mathcal{D}} \|k(\mathbf{d}, \cdot)\|_{\mathcal{H}_k} = \sup_{\mathbf{d} \in \mathcal{D}} \sqrt{k(\mathbf{d}, \mathbf{d})}. \tag{6}$$

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

$$\kappa_\epsilon(\mathbf{y}, \mathbf{x}) = c_\epsilon k(\mathbf{y}, \mathbf{x}), \tag{7}$$

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 $\kappa_\epsilon(\mathbf{d}, \cdot) \in \mathcal{H}_k$, we can find its RKHS norm,

$$\|\kappa_\epsilon(\mathbf{y}, \cdot)\|_{\mathcal{H}_k} = c_\epsilon \|k(\mathbf{y}, \cdot)\|_{\mathcal{H}_k} = c_\epsilon \sqrt{k(\mathbf{y}, \mathbf{y})} = \sqrt{c_\epsilon}\sqrt{c_\epsilon k(\mathbf{y}, \mathbf{y})} = \sqrt{c_\epsilon}\sqrt{\kappa_\epsilon(\mathbf{y}, \mathbf{y})}, \tag{8}$$

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.

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 $\|\kappa_\epsilon(\mathbf{y}, \cdot)\|_{\mathcal{H}_k} = 1/(\sqrt{2\pi}\epsilon)^n = c_\epsilon$.

As a consequence, we have that the upper bound to the RKHS norm of κ_ϵ satisfies

$$\bar{\kappa}_\epsilon := \sup_{\mathbf{d} \in \mathcal{D}} \|\kappa_\epsilon(\mathbf{d}, \cdot)\|_{\mathcal{H}_k} = \sqrt{c_\epsilon} \sup_{\mathbf{d} \in \mathcal{D}} \sqrt{\kappa_\epsilon(\mathbf{d}, \mathbf{d})}. \tag{9}$$

When $\mathcal{D} = \mathbb{R}^n$, the most commonly used kernel for the KELFI framework is the Gaussian anisotropic, or ARD, kernel, where each dimension uses a potentially different length scale σ_i,

$$k(\mathbf{x}, \mathbf{x}') = \exp\bigg(-\frac{1}{2} \sum_{i=1}^{n} \Big(\frac{x_i - x_i'}{\sigma_i} \Big)^2 \bigg). \tag{10}$$

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 Gaussian ARD kernel for ℓ, but we use β_d, $d \in [D]$ to denote the length scales.

9.2. 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

$$\mu_{\mathbf{X}|\mathbf{\Theta}=\boldsymbol{\theta}} := \mathbb{E}[k(\mathbf{X}, \cdot)|\mathbf{\Theta} = \boldsymbol{\theta}]. \tag{11}$$

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 $C_{\mathbf{\Theta}\mathbf{\Theta}} \in \mathcal{H}_\ell \otimes \mathcal{H}_\ell$.

Under the assumption $\ell(\boldsymbol{\theta}, \cdot) \in \mathrm{image}(C_{\mathbf{\Theta}\mathbf{\Theta}})$, 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 theorems.

9.3. 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*, $j \in \{1, \dots, m\}$. The conditional mean operator $\mathcal{U}_{\mathbf{X}|\mathbf{\Theta}}$ is estimated by

$$\hat{\mathcal{U}}_{\mathbf{X}|\mathbf{\Theta}} = \Phi(L + m\lambda I)^{-1}\Psi^T, \tag{12}$$

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

$$\hat{\mu}_{\mathbf{X}|\boldsymbol{\Theta}=\boldsymbol{\theta}} = \hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} \ell(\boldsymbol{\theta}, \cdot) = \Phi (L + m\lambda I)^{-1} \boldsymbol{\ell}(\boldsymbol{\theta}) \tag{13}$$

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

$$\hat{g}(\boldsymbol{\theta}) = \mathbf{f}^T (L + m\lambda I)^{-1} \boldsymbol{\ell}(\boldsymbol{\theta}). \tag{14}$$

Importantly, $\hat{\mu}_{\mathbf{X}|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ 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 $\boldsymbol{\Theta}$, which does not necessarily need to be the same as $p(\boldsymbol{\theta})$.

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

$$\begin{aligned} &\forall \boldsymbol{\theta} \in \vartheta, \ \forall \epsilon > 0, \ \exists M_\epsilon > 0 \quad s.t. \\ &\mathbb{P}\Big[\big\|\hat{\mu}_{\mathbf{X}|\boldsymbol{\Theta}=\boldsymbol{\theta}} - \mu_{\mathbf{X}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\big\|_{\mathcal{H}_k} > M_\epsilon \Big((m\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}}\Big)\Big] < \epsilon. \end{aligned} \tag{15}$$

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 $\ell(\boldsymbol{\theta}, \cdot) \in \mathrm{image}(C_{\boldsymbol{\Theta}\boldsymbol{\Theta}})$.

Finally, since $\hat{\mu}_{\mathbf{X}|\boldsymbol{\Theta}=\boldsymbol{\theta}} = \hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} \ell(\boldsymbol{\theta}, \cdot)$ convergences to $\mu_{\mathbf{X}|\boldsymbol{\Theta}=\boldsymbol{\theta}} = \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}} \ell(\boldsymbol{\theta}, \cdot)$ 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 $\hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}}$ converges to $\mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}}$ in Hilbert Schmidt (HS) norm at the same rate. That is,

$$\begin{aligned} &\forall \epsilon > 0, \ \exists M_\epsilon > 0 \quad s.t. \\ &\mathbb{P}\Big[\big\|\hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}}\big\|_{HS} > M_\epsilon \Big((m\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}}\Big)\Big] < \epsilon. \end{aligned} \tag{16}$$

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

$$d(\hat{a}, a) \leq \gamma \big\|\hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}}\big\|_{HS}, \tag{17}$$

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 Suppose that there exists $0 \leq \gamma < \infty$ such that (17) is satisfied. That is, the inequality (17) holds for all possible data observations $\{\boldsymbol{\theta}_j, \mathbf{x}_j\}_{j=1}^{m}$. For any constant C, the implication statement $\big\|\hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}}\big\|_{HS} \leq C \implies d(\hat{a}, a) \leq C\gamma$ holds for all possible observation events $\omega \in \Omega$. Writing this explicitly in event space translates this to a probability statement,

$$\begin{aligned} \{\omega \in \Omega : \big\|\hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}}\big\|_{HS} \leq C\} &\subseteq \{\omega \in \Omega : d(\hat{a}, a) \leq C\gamma\} \\ \implies \mathbb{P}\Big[\big\|\hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}}\big\|_{HS} \leq C\Big] &\leq \mathbb{P}\Big[d(\hat{a}, a) \leq C\gamma\Big]. \end{aligned} \tag{18}$$

Since we assume that $\ell(\boldsymbol{\theta}, \cdot) \in \mathrm{image}(C_{\boldsymbol{\Theta}\boldsymbol{\Theta}})$, statement (15) is valid. By letting $C = M_{\epsilon}((m\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}})$ in (18), we immediately have that the probability inequality in statement (16) is also true if we replace $\big\|\hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}}\big\|_{HS}$ with $d(\hat{a}, a)$ and M_{ϵ} with γM_{ϵ},

$$\begin{aligned} \mathbb{P}\Big[\big\|\hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}}\big\|_{HS} > M_{\epsilon}\Big((m\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}}\Big)\Big] &< \epsilon \\ \implies 1 - \mathbb{P}\Big[\big\|\hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}}\big\|_{HS} \leq M_{\epsilon}\Big((m\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}}\Big)\Big] &< \epsilon \\ \implies \mathbb{P}\Big[\big\|\hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}}\big\|_{HS} \leq M_{\epsilon}\Big((m\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}}\Big)\Big] &> 1 - \epsilon \\ \implies \mathbb{P}\Big[d(\hat{a}, a) \leq \gamma M_{\epsilon}\Big((m\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}}\Big)\Big] &> 1 - \epsilon \\ \implies 1 - \mathbb{P}\Big[d(\hat{a}, a) \leq \gamma M_{\epsilon}\Big((m\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}}\Big)\Big] &< \epsilon \\ \implies \mathbb{P}\Big[d(\hat{a}, a) > \gamma M_{\epsilon}\Big((m\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}}\Big)\Big] &< \epsilon, \end{aligned} \tag{19}$$

where we employed statement (18) 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,

$$\forall \epsilon > 0, \; \exists \tilde{M}_{\epsilon} > 0 \quad s.t. \quad \mathbb{P}\Big[d(\hat{a}, a) > \tilde{M}_{\epsilon}\Big((m\lambda)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}}\Big)\Big] < \epsilon. \tag{20}$$

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

$$d(\hat{a}, a) \leq \gamma \big\|\hat{\mu}_{\mathbf{X}|\boldsymbol{\Theta}=\boldsymbol{\theta}} - \mu_{\mathbf{X}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\big\|_{\mathcal{H}_k}, \tag{21}$$

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 theorem 1, where $\big\|\hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}}\big\|_{HS}$ is replaced with $\big\|\hat{\mu}_{\mathbf{X}|\boldsymbol{\Theta}=\boldsymbol{\theta}} - \mu_{\mathbf{X}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\big\|_{\mathcal{H}_k}$ throughout. Alternatively, since $\big\|\hat{\mu}_{\mathbf{X}|\boldsymbol{\Theta}=\boldsymbol{\theta}} - \mu_{\mathbf{X}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\big\|_{\mathcal{H}_k} = \big\|(\hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}})\ell(\boldsymbol{\theta}, \cdot)\big\|_{\mathcal{H}_k} \leq \big\|\hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}}\big\|_{HS} \big\|\ell(\boldsymbol{\theta}, \cdot)\big\|_{\mathcal{H}_\ell} \leq \big\|\hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}}\big\|_{HS} \ell(\boldsymbol{\theta}, \boldsymbol{\theta})$, $\forall \boldsymbol{\theta} \in \vartheta$, we have $d(\hat{a}, a) \leq \gamma \ell(\boldsymbol{\theta}, \boldsymbol{\theta}) \big\|\hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}}\big\|_{HS} \leq \gamma (\sup_{\boldsymbol{\theta} \in \vartheta} \ell(\boldsymbol{\theta}, \boldsymbol{\theta})) \big\|\hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}}\big\|_{HS}$,

$\forall \boldsymbol{\theta} \in \vartheta$. Since $\gamma \sup_{\boldsymbol{\theta} \in \vartheta} \ell(\boldsymbol{\theta}, \boldsymbol{\theta})$ is finite and does not depend on m, we apply theorem 1 to arrive at theorem 2. ■

With theorems 1 and 2, we are now equipped to show the convergence of various estimators based on the CME.

9.5. 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 relevant hyperparameters, even though this is not explicitly notated.

Theorem 3 *Assume* $\ell(\boldsymbol{\theta}, \cdot) \in \mathrm{image}(C_{\boldsymbol{\Theta}\boldsymbol{\Theta}})$. *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 $p_\epsilon(\mathbf{y}|\boldsymbol{\theta})$,

$$\begin{aligned}
|q(\mathbf{y}|\boldsymbol{\theta}) - p_\epsilon(\mathbf{y}|\boldsymbol{\theta}))| =& |\langle \kappa_\epsilon(\mathbf{y}, \cdot), \hat{\mu}_{\mathbf{X}|\boldsymbol{\Theta}=\boldsymbol{\theta}} \rangle_{\mathcal{H}_k} - \langle \kappa_\epsilon(\mathbf{y}, \cdot), \mu_{\mathbf{X}|\boldsymbol{\Theta}=\boldsymbol{\theta}} \rangle_{\mathcal{H}_k}| \\
=& |\langle \kappa_\epsilon(\mathbf{y}, \cdot), \hat{\mu}_{\mathbf{X}|\boldsymbol{\Theta}=\boldsymbol{\theta}} - \mu_{\mathbf{X}|\boldsymbol{\Theta}=\boldsymbol{\theta}} \rangle_{\mathcal{H}_k}| \\
\leq& \|\kappa_\epsilon(\mathbf{y}, \cdot)\|_{\mathcal{H}_k} \|\hat{\mu}_{\mathbf{X}|\boldsymbol{\Theta}=\boldsymbol{\theta}} - \mu_{\mathbf{X}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\|_{\mathcal{H}_k} \\
\leq& \bar{\kappa}_\epsilon \|\hat{\mu}_{\mathbf{X}|\boldsymbol{\Theta}=\boldsymbol{\theta}} - \mu_{\mathbf{X}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\|_{\mathcal{H}_k} \\
\leq& \bar{\kappa}_\epsilon \ell(\boldsymbol{\theta}, \boldsymbol{\theta}) \big\|\hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}}\big\|_{HS} \\
\leq& \bar{\kappa}_\epsilon \bar{\ell} \big\|\hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}}\big\|_{HS}.
\end{aligned} \tag{22}$$

■

Since $\gamma = \bar{\kappa}_\epsilon \bar{\ell}$ is independent of m, we apply theorem 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 section 11.

9.6. Convergence Guarantees for Marginal Kernel Means Likelihood

Theorem 4 *Assume* $\ell(\boldsymbol{\theta}, \cdot) \in \mathrm{image}(C_{\boldsymbol{\Theta}\boldsymbol{\Theta}})$. *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 $p_\epsilon(\mathbf{y})$,

$$\begin{aligned}
|q(\mathbf{y}) - p_\epsilon(\mathbf{y})| &= |\mathbb{E}[q(\mathbf{y}|\mathbf{\Theta}) - p(\mathbf{y}|\mathbf{\Theta})]| \\
&\leq \mathbb{E}[|q(\mathbf{y}|\mathbf{\Theta}) - p(\mathbf{y}|\mathbf{\Theta})|] \\
&\leq \bar{\kappa}_\epsilon \mathbb{E}[\|\hat{\mu}_{\mathbf{X}|\mathbf{\Theta}=\mathbf{\Theta}} - \mu_{\mathbf{X}|\mathbf{\Theta}=\mathbf{\Theta}}\|_{\mathcal{H}_k}] \\
&= \bar{\kappa}_\epsilon \mathbb{E}[\|(\hat{\mathcal{U}}_{\mathbf{X}|\mathbf{\Theta}} - \mathcal{U}_{\mathbf{X}|\mathbf{\Theta}})\ell(\mathbf{\Theta}, \cdot)\|_{\mathcal{H}_k}] \\
&\leq \bar{\kappa}_\epsilon \mathbb{E}[\|\hat{\mathcal{U}}_{\mathbf{X}|\mathbf{\Theta}} - \mathcal{U}_{\mathbf{X}|\mathbf{\Theta}}\|_{HS}\|\ell(\mathbf{\Theta}, \cdot)\|_{\mathcal{H}_\ell}] \\
&= \bar{\kappa}_\epsilon \mathbb{E}[\|\hat{\mathcal{U}}_{\mathbf{X}|\mathbf{\Theta}} - \mathcal{U}_{\mathbf{X}|\mathbf{\Theta}}\|_{HS}\sqrt{\ell(\mathbf{\Theta}, \mathbf{\Theta})}] \\
&= \bar{\kappa}_\epsilon \mathbb{E}[\sqrt{\ell(\mathbf{\Theta}, \mathbf{\Theta})}]\|\hat{\mathcal{U}}_{\mathbf{X}|\mathbf{\Theta}} - \mathcal{U}_{\mathbf{X}|\mathbf{\Theta}}\|_{HS} \\
&\leq \bar{\kappa}_\epsilon \mathbb{E}[\bar{\ell}]\|\hat{\mathcal{U}}_{\mathbf{X}|\mathbf{\Theta}} - \mathcal{U}_{\mathbf{X}|\mathbf{\Theta}}\|_{HS} \\
&= \bar{\kappa}_\epsilon \bar{\ell}\|\hat{\mathcal{U}}_{\mathbf{X}|\mathbf{\Theta}} - \mathcal{U}_{\mathbf{X}|\mathbf{\Theta}}\|_{HS}
\end{aligned} \tag{23}$$

Since $\gamma = \bar{\kappa}_\epsilon \bar{\ell}$ is independent of m, we apply theorem 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 $\mathbf{y} \in \mathcal{Y}$. ■

9.7. Convergence Guarantees for Kernel Means Posterior

Theorem 5 *Assume $\ell(\boldsymbol{\theta}, \cdot) \in \text{image}(C_{\mathbf{\Theta}\mathbf{\Theta}})$ 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 between the likelihood and marginal likelihood,

$$\left|\frac{q(\mathbf{y}|\boldsymbol{\theta})}{p_\epsilon(\mathbf{y}|\boldsymbol{\theta})} - 1\right| \leq \frac{1}{p_\epsilon(\mathbf{y}|\boldsymbol{\theta})}\left|q(\mathbf{y}|\boldsymbol{\theta}) - p_\epsilon(\mathbf{y}|\boldsymbol{\theta})\right| \leq \frac{\bar{\kappa}_\epsilon \bar{\ell}}{p_\epsilon(\mathbf{y}|\boldsymbol{\theta})}\left\|\hat{\mathcal{U}}_{\mathbf{X}|\mathbf{\Theta}} - \mathcal{U}_{\mathbf{X}|\mathbf{\Theta}}\right\|_{HS}, \tag{24}$$

$$\left|\frac{q(\mathbf{y})}{p_\epsilon(\mathbf{y})} - 1\right| \leq \frac{1}{p_\epsilon(\mathbf{y})}\left|q(\mathbf{y}) - p_\epsilon(\mathbf{y})\right| \leq \frac{\bar{\kappa}_\epsilon \bar{\ell}}{p_\epsilon(\mathbf{y})}\left\|\hat{\mathcal{U}}_{\mathbf{X}|\mathbf{\Theta}} - \mathcal{U}_{\mathbf{X}|\mathbf{\Theta}}\right\|_{HS}. \tag{25}$$

Finally, 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}
\Big|q(\boldsymbol{\theta}|\mathbf{y}) - p_\epsilon(\boldsymbol{\theta}|\mathbf{y})\Big| =& \bigg|\frac{q(\mathbf{y}|\boldsymbol{\theta})}{q(\mathbf{y})} - \frac{p_\epsilon(\mathbf{y}|\boldsymbol{\theta})}{p_\epsilon(\mathbf{y})}\bigg| p(\boldsymbol{\theta}) \\
=& \bigg|\frac{q(\mathbf{y}|\boldsymbol{\theta})}{p_\epsilon(\mathbf{y}|\boldsymbol{\theta})} - \frac{q(\mathbf{y})}{p_\epsilon(\mathbf{y})}\bigg| \frac{p_\epsilon(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{|q(\mathbf{y})|} \\
=& \bigg|\Big(\frac{q(\mathbf{y}|\boldsymbol{\theta})}{p_\epsilon(\mathbf{y}|\boldsymbol{\theta})} - 1\Big) - \Big(\frac{q(\mathbf{y})}{p_\epsilon(\mathbf{y})} - 1\Big)\bigg| \frac{p_\epsilon(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{|q(\mathbf{y})|} \\
\leq& \bigg(\Big|\frac{q(\mathbf{y}|\boldsymbol{\theta})}{p_\epsilon(\mathbf{y}|\boldsymbol{\theta})} - 1\Big| + \Big|\frac{q(\mathbf{y})}{p_\epsilon(\mathbf{y})} - 1\Big|\bigg) \frac{p_\epsilon(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{|q(\mathbf{y})|} \\
\leq& \bigg(\frac{\bar{\kappa}_\epsilon \bar{\ell}}{p_\epsilon(\mathbf{y}|\boldsymbol{\theta})} \Big\|\hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}}\Big\|_{HS} + \frac{\bar{\kappa}_\epsilon \bar{\ell}}{p_\epsilon(\mathbf{y})} \Big\|\hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}}\Big\|_{HS}\bigg) \frac{p_\epsilon(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{|q(\mathbf{y})|} \\
\leq& \bigg(\bar{\kappa}_\epsilon \bar{\ell} p(\boldsymbol{\theta}) \Big\|\hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}}\Big\|_{HS} + \bar{\kappa}_\epsilon \bar{\ell} p_\epsilon(\boldsymbol{\theta}|\mathbf{y}) \Big\|\hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}}\Big\|_{HS}\bigg) \frac{1}{|q(\mathbf{y})|} \\
\leq& \bar{\kappa}_\epsilon \bar{\ell} \big(p(\boldsymbol{\theta}) + p_\epsilon(\boldsymbol{\theta}|\mathbf{y})\big) \Big\|\hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}}\Big\|_{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} \Big\|\hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}}\Big\|_{HS}.
\end{aligned} \tag{26}$$

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 theorem 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

$$\begin{aligned}
p(\boldsymbol{\theta}) + p_\epsilon(\boldsymbol{\theta}|\mathbf{y}) \leq \sup_{\boldsymbol{\theta} \in \vartheta} \big(p(\boldsymbol{\theta}) + p_\epsilon(\boldsymbol{\theta}|\mathbf{y})\big) &\leq \sup_{\boldsymbol{\theta} \in \vartheta} p(\boldsymbol{\theta}) + \sup_{\boldsymbol{\theta} \in \vartheta} p_\epsilon(\boldsymbol{\theta}|\mathbf{y}) \\
&\leq \sup_{\boldsymbol{\theta} \in \vartheta} p(\boldsymbol{\theta}) + \frac{\sup_{\boldsymbol{\theta} \in \vartheta} p_\epsilon(\mathbf{y}|\boldsymbol{\theta}) \sup_{\boldsymbol{\theta} \in \vartheta} p(\boldsymbol{\theta})}{p_\epsilon(\mathbf{y})} \\
&= \bar{p}_{\boldsymbol{\Theta}} + \frac{\bar{p}_{\mathbf{Y}|\boldsymbol{\Theta}} \bar{p}_{\boldsymbol{\Theta}}}{p_\epsilon(\mathbf{y})}.
\end{aligned} \tag{27}$$

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

9.8. Convergence Guarantees for Kernel Means Posterior Embedding

Theorem 6 *Assume $\ell(\boldsymbol{\theta}, \cdot) \in \mathrm{image}(C_{\boldsymbol{\Theta}\boldsymbol{\Theta}})$ 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}
&\Big\|\tilde{\mu}_{\boldsymbol{\Theta}|\mathbf{Y}=\mathbf{y}} - \mu_{\boldsymbol{\Theta}|\mathbf{Y}=\mathbf{y}}\Big\|_{\mathcal{H}_\ell}^2 \\
&= \Big\| \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} \Big\|_{\mathcal{H}_\ell}^2 \\
&= \Big\| \int_\vartheta \ell(\boldsymbol{\theta}, \cdot) \Big(q(\boldsymbol{\theta}|\mathbf{y}) - p_\epsilon(\boldsymbol{\theta}|\mathbf{y}) \Big) d\boldsymbol{\theta} \Big\|_{\mathcal{H}_\ell}^2 \\
&= \Big\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}' \Big\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}' \\
&= \Big| \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}' \Big| \\
&\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 \Big(\int_\vartheta \Big| q(\boldsymbol{\theta}|\mathbf{y}) - p_\epsilon(\boldsymbol{\theta}|\mathbf{y}) \Big| d\boldsymbol{\theta} \Big)^2 .
\end{aligned} \tag{28}$$

We now employ inequality (26) that was derived within the proof of theorem 5,

$$\begin{aligned}
\Big\|\tilde{\mu}_{\boldsymbol{\Theta}|\mathbf{Y}=\mathbf{y}} - \mu_{\boldsymbol{\Theta}|\mathbf{Y}=\mathbf{y}}\Big\|_{\mathcal{H}_\ell} &\leq \bar{\bar{\ell}} \int_\vartheta \Big| q(\boldsymbol{\theta}|\mathbf{y}) - p_\epsilon(\boldsymbol{\theta}|\mathbf{y}) \Big| d\boldsymbol{\theta} \\
&\leq \bar{\bar{\ell}} \int_\vartheta \frac{\bar{\kappa}_\epsilon \bar{\ell} \big(p(\boldsymbol{\theta}) + p_\epsilon(\boldsymbol{\theta}|\mathbf{y}) \big)}{\delta} \Big\| \hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}} \Big\|_{HS} d\boldsymbol{\theta} \\
&= \bar{\bar{\ell}} \Big(\int_\vartheta \big(p(\boldsymbol{\theta}) + p_\epsilon(\boldsymbol{\theta}|\mathbf{y}) \big) d\boldsymbol{\theta} \Big) \frac{\bar{\kappa}_\epsilon \bar{\ell}}{\delta} \Big\| \hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}} \Big\|_{HS} \\
&= \frac{2 \bar{\kappa}_\epsilon \bar{\ell} \bar{\bar{\ell}}}{\delta} \Big\| \hat{\mathcal{U}}_{\mathbf{X}|\boldsymbol{\Theta}} - \mathcal{U}_{\mathbf{X}|\boldsymbol{\Theta}} \Big\|_{HS} .
\end{aligned} \tag{29}$$

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 theorem 1 to establish the convergence under the RKHS norm. ■

10. 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 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, the asymptotic behavior is rarely reached under limited simulation data, which is the scenario of interest. We therefore present an inference approach by considering the surrogate posterior defined directly using the KML in its original form.

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, we cannot directly sample from a surrogate density that is possibly non-positive. To address this, our solution is motivated by super sampling with kernel herding (Chen et al., 2010) on general CMEs. 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 with super samples 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 kernel mean embeddings (KMEs), including injectivity between mean embeddings and distributions, remain valid for signed measures. By defining an analogous form of mean embeddings for surrogate densities, we arrive at a novel posterior mean embedding that is associated with a marginal surrogate likelihood.

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.

11. Connection to Other Models

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) 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 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 and KBR. K-ABC also do not address sampling, although kernel herding can be readily applied in the same way, where as for KBR kernel herding is applied 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.

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.

12. Experimental Details for the Blowfly Problem

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. Note that 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.

In the experiments, 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.

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.

13. Experimental Details for the Lotka-Volterra Problem

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 the $\log \boldsymbol{\theta}$. 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 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 the uniform back to its original ranges. Similar to the other experiments, we do not learn the prior hyperparameters in this paper, 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% interval, we compute the empirical 2.5% quantile and 97.5% quantile using the 10000 super-samples.

14. Transforming Priors into a Gaussian Prior

Under certain assumptions, we can always transform our ABC problem into another ABC problem with a Gaussian prior without loss of generality. These assumptions are that $p_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) = \prod_{i=1}^{d} p_{\Theta_i}(\theta_i)$ is a continuous probability density function (PDF) whose entries are independent, and that its inverse marginal cumulative distribution functions (CDFs) $P_{\Theta_i}^{-1}$ exists and is tractable.

In this section only, we will use a bold face to denote unobserved variables $\boldsymbol{\theta}$, in order to make explicit of the vector nature of this quantity. In other sections of the paper, the formulation was general enough that θ could be any type of quantity other than scalars and vectors. However, here we will perform the derivations specifically for the case there $\boldsymbol{\theta}$ is a vector quantity. Each entry of $\boldsymbol{\theta}$ will be denoted as θ_i and the dimensionality of this vector will be denoted as d. If we generate multiple realizations of θ, we will use a superscript to denote this, such as $\{\boldsymbol{\theta}^{(j)}\}_{j=1}^{m}$, so the i-th entry of the j-th simulation is $\theta_i^{(j)}$.

We will also 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 our original prior $p_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$ is a general distribution that is not necessarily Gaussian, but satisfies the aforementioned assumptions.

Suppose we want to simplify our ABC problem into one which uses a Gaussian prior.

The following is an outline for how this can be done. Let $\mathbf{Z}$ be a random variable of the same dimensionality as $\boldsymbol{\Theta}$ and let $p_{\mathbf{Z}}(\mathbf{z}) = \prod_{i=1}^{d} p_{Z_i}(z_i)$, where $p_{Z_i}(z_i) = \mathcal{N}(\mu_i, \sigma_i^2)$ so that its density is a multivariate isotropic Gaussian. For convenience, we usually choose $\mu_i = 0$ and $\sigma_i = \sigma$ for all $i \in [d]$, although we can keep this general for now.

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

2. Convert them to uniform samples through $u_i^{(j)} = P_{Z_i}(z_i^{(j)})$ for $j \in [m]$ and $i \in [d]$, in this way we know that $\mathbf{u}^{(j)} \sim U(0,1)^d$ for $j \in [m]$.

3. Convert them to samples from the prior through $\theta_i^{(j)} = P_{\Theta_i}^{-1}(u_i^{(j)})$ for $j \in [m]$ and $i \in [d]$.

   In effect, if we define define $\mathbf{T}(\mathbf{z}) := \{T_i(z_i)\}_{i=1}^{d}$ and $T_i(z_i) = P_{\Theta_i}^{-1}(P_{Z_i}(z_i))$, then $\boldsymbol{\theta}^{(j)} = \mathbf{T}(\mathbf{z}^{(j)})$ for $j \in [m]$.

   Also, notice that the inverse $\mathbf{T}^{-1}(\boldsymbol{\theta}) = \{T_i^{-1}(\theta_i)\}_{i=1}^{d}$ where $T_i^{-1}(\theta_i) = P_{Z_i}^{-1}(P_{\Theta_i}(\theta_i))$ exists.

4. Run the simulator $\mathbf{x}^{(j)} \sim p_{\mathbf{X}|\boldsymbol{\Theta}}(\mathbf{x}|\boldsymbol{\theta}^{(j)}) = p_{\mathbf{X}|\boldsymbol{\Theta}}(\mathbf{x}|T(\mathbf{z}^{(j)}))$. 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 posterior samples $\{\mathbf{z}^{(l)}\}_{l=1}^{r}$ to represent the posterior empirically.

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

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

   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 T_i^{-1} does not depend on θ_j if $i \neq j$, the Jacobian is diagonal, and 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 $\det J_{\mathbf{T}^{-1}}(\boldsymbol{\theta}) = \prod_{i=1}^{d} [p_{Z_i}(T_i^{-1}(\theta_i))]^{-1} p_{\Theta_i}(\theta_i) = p_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) \big[\prod_{i=1}^{d} p_{Z_i}(T_i^{-1}(\theta_i)) \big]^{-1} = p_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) [p_{\mathbf{Z}}(\mathbf{T}^{-1}(\boldsymbol{\theta}))]^{-1}$. So,

$$q_{\boldsymbol{\Theta}|\mathbf{Y}}(\boldsymbol{\theta}|\mathbf{y}) = q_{\mathbf{Z}|\mathbf{Y}}(\mathbf{T}^{-1}(\boldsymbol{\theta})|\mathbf{y}) \frac{p_{\boldsymbol{\Theta}}(\boldsymbol{\theta})}{p_{\mathbf{Z}}(\mathbf{T}^{-1}(\boldsymbol{\theta}))}. \tag{30}$$

Finally, the form simplifies when the form of the KMP is substituted back in,

$$q_{\mathbf{\Theta}|\mathbf{Y}}(\boldsymbol{\theta}|\mathbf{y}) = \frac{q_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y}|\mathbf{T}^{-1}(\boldsymbol{\theta}))p_{\mathbf{Z}}(\mathbf{T}^{-1}(\boldsymbol{\theta}))}{q_{\mathbf{Y}}(\mathbf{y})} \frac{p_{\mathbf{\Theta}}(\boldsymbol{\theta})}{p_{\mathbf{Z}}(\mathbf{T}^{-1}(\boldsymbol{\theta}))} = \frac{q_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y}|\mathbf{T}^{-1}(\boldsymbol{\theta}))p_{\mathbf{\Theta}}(\boldsymbol{\theta})}{q_{\mathbf{Y}}(\mathbf{y})}, \tag{31}$$

where the MKML is still marginalized over the simpler Gaussian distribution,

$$q_{\mathbf{Y}}(\mathbf{y}) = \int_{\mathcal{Z}} q_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y}|\mathbf{z})p_{\mathbf{Z}}(\mathbf{z})d\mathbf{z}. \tag{32}$$

In this way, we can always simplify the ABC problem by another ABC problem with a Gaussian prior. If the original simulation samples $\{\boldsymbol{\theta}^{(j)}, \mathbf{x}^{(j)}\}_{j=1}^{m}$ are already provided, simply convert the $\boldsymbol{\theta}^{(j)}$ back to a normal distribution by $\mathbf{z}^{(j)} = \mathbf{T}^{-1}(\boldsymbol{\theta}^{(j)})$ for $j \in [m]$ and use $\{\mathbf{z}^{(j)}, \mathbf{x}^{(j)}\}_{j=1}^{m}$ to proceed. Note that this is possible in general only for ABC problems, because if the likelihood is intractable anyway and we only have a simulator for it, we can prepend transformations $\mathbf{T}$ before the simulator and the problem does not become any more harder or simpler. If we were to do this in the case when we have a nice tractable likelihood, then doing this may make the complexity of the posterior learning problem much harder as we have modified the simple relationship between the unobserved and the observed variable.

As an extension, instead of transforming the ABC 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. In practice, transformations which are close to singular will also introduce numerical errors when applying the gradient based learning algorithm when optimizing the MKML. Nevertheless, those are potential drawbacks that are inherent in transforming distributions and not specific to the KELFI framework.

15. Extension: Kernel Means Likelihood for Spatio-Temporal data

In this section we specifically address a scenario that is common for ABC type problems but is never really treated separately to the standard *iid* dataset scenario. This is the scenario where are dataset is in the form of a time series. In general, the approach we will discuss applies to datasets formed from observing any random field, but for the purpose of simplicity and clarity, we will focus on a observations from stochastic process that has one input dimension. This input dimension is usually thought of as time, hence our observations and thus dataset is in the form of time series.

Again, the reason why this is not usually treated separately from the standard scenario with *iid* datasets scenario is the assumption that a good sufficient statistic exists to summarize the time series data. By using this assumption, the problem has now been reduced down to the same problem as the standard scenario with *iid* observations, which traditionally also assumes the existence and availability of sufficient statistics. As we mentioned before, we did not rely on this in our approach for the scenario with *iid* data, and in this current discussion we will also formulate a corresponding approach for the scenario with time series data without relying on using sufficient statistics.

Constructing ϵ-kernels with Gaussian processes. Suppose observed spatio-temporal data comes in the form of $\{t_i, y_i\}_{i=1}^{n}$ and a simulation of such data comes in the form of $\{s_i, x_i\}_{i=1}^{n'}$. We will assume that $x_i, y_i \in \mathbb{R}$ for simplicity, although this framework can be easily extended for vector outputs. In this case, t_i and s_i can be time stamps, in which they are scalar values, or spatio-temporal locations, in which they are vector inputs. In either case, we will denote the space they lie in as $\mathcal{T} = \mathcal{S}$. For simplicity, the reader can think of these are scalar time stamps, although we do not restrict this to be the case in our derivations. We will denote $\mathbf{y} = \{y_i\}_{i=1}^{n}$, $\mathbf{x} = \{x_i\}_{i=1}^{n'}$, $\mathbf{t} = \{t_i\}_{i=1}^{n}$, and $\mathbf{s} = \{s_i\}_{i=1}^{n'}$ (not to be confused with the summary statistics).

We first define a positive definite kernel $h : \mathcal{T} \times \mathcal{T} \to \mathbb{R}_{+}$ to measure the similarity between the spatio-temporal inputs. Then, we construct an ϵ-kernel by using the full predictive distribution of a GPR,

$$\begin{aligned} \kappa_{\epsilon}((\mathbf{t}, \mathbf{y}), (\mathbf{s}, \mathbf{x})) &= p(\mathbf{y}|\mathbf{x}, \mathbf{t}, \mathbf{s}) \\ &= \mathcal{N}(\mathbf{y}|H_{\mathbf{st}}^{T}(H_{\mathbf{ss}} + \epsilon^{2} I_{n'})^{-1}\mathbf{x}, H_{\mathbf{tt}} + \epsilon^{2} I_{n} - H_{\mathbf{s},\mathbf{t}}^{T}(H_{\mathbf{ss}} + \epsilon^{2} I_{n'})^{-1} H_{\mathbf{s},\mathbf{t}}) \end{aligned} \tag{33}$$

where $H_{\mathbf{s},\mathbf{t}} = \{H(s_i, t_j)\}_{i=1,j=1}^{n',n}$, $H_{\mathbf{s},\mathbf{s}} = \{H(s_i, s_j)\}_{i=1,j=1}^{n',n'}$, $H_{\mathbf{t},\mathbf{t}} = \{H(t_i, t_j)\}_{i=1,j=1}^{n,n}$.

That is, $p(\mathbf{y}|\mathbf{x}, \mathbf{t}, \mathbf{s})$ is the full predictive distribution of a GPR trained on $\{s_i, x_i\}_{i=1}^{n'}$ and evaluated at the query points $\{t_i, y_i\}_{i=1}^{n}$. The derivation of this full predictive distribution is given in Rasmussen and Williams (2006).

This was a modeling choice to leverage the spatio-temporal relationship between each data point. So, $\{s_i, x_i\}_{i=1}^{n'}$ and $\{t_i, y_i\}_{i=1}^{n}$ are modeled as noisy realizations (with noise level ϵ) from a GP. The ϵ-kernel then measures what is the probability of observing our observed spatio-temporal data $\{t_i, y_i\}_{i=1}^{n}$ given our simulation data $\{s_i, x_i\}_{i=1}^{n'}$.

With the above ϵ-kernel, the new data likelihood estimate is

$$q(\mathbf{y}|\mathbf{t}, \theta) = \boldsymbol{\kappa}_{\epsilon}(\mathbf{t}, \mathbf{y})^{T}(L + m\lambda I)^{-1}\boldsymbol{\ell}(\theta), \tag{34}$$

where $\boldsymbol{\kappa}_{\epsilon}(\mathbf{t}, \mathbf{y}) := \{\kappa_{\epsilon}((\mathbf{t}, \mathbf{y}), (\mathbf{s}_j, \mathbf{x}_j))\}_{j=1}^{m}$. We call this the spatio-temporal kernel means likelihood (ST-KML).

Notice that we use the GP in a very different way to other ABC approaches which uses Gaussian processes (GPs). For example, GPS-ABC (Meeds and Welling, 2014) models the generation of the simulator as a GP, which is an entirely different purpose to ours, which is to model the spatio-temporal data itself. For example, the input to their GP is a parameter $\theta \in \vartheta$, whereas the input to our GP is a spatio-temporal coordinate $t \in \mathcal{T}$ or $s \in \mathcal{S}$. On the other hand, GPA-ABC (Wilkinson, 2014) uses their GP to model the log likelihood density directly, instead of generated samples from the simulator. Again, the input to their GP is a parameter $\theta \in \vartheta$.

Furthermore, unlike GPS-ABC where the likelihood is assumed to be Gaussian, in our case use the GP to model the *residual* process between $x(s)$ and $y(t)$. The Gaussian assumption for the residual is as justified as using a Gaussian noise assumption for the ϵ-kernel in the usual ABC setting, where the inference becomes exact when $\epsilon \to 0$ and when $\epsilon > 0$ we trade off exact inference for tractability (to make use of close enough simulations instead of having to wait for simulations to exactly match the data).

This is the first work in our knowledge that specifically address spatio-temporal data by defining a GP-based ϵ-kernel to capture the non-*iid* aspects of the data. Specifically, we leverage the smoothness properties in such a data, where y_i and y_j would be more related in t_i and t_j are close.

It is worthwhile to point out again that the GP and CME are modeling entirely different relationships between different pairs of spaces. Specifically, the GP is modeling the relationship between the output space $\mathcal{Y} = \mathcal{X}$ and the input space $\mathcal{T} = \mathcal{S}$ which represents the ϵ-kernel, where the input is treated as deterministic and the input distribution is not modeled. Meanwhile, the CME is modeling the relationship between the output space $\mathcal{X}$ and the input space ϑ which represents the likelihood relationship, where the input distribution is explicitly modeled, usually by the prior $p(\theta)$.

16. Predictive Surrogate Density

Since the KML is a likelihood surrogate, we can also query the likelihood on a different set of data $\mathbf{y}^\star$ by $q(\mathbf{y}^\star|\boldsymbol{\theta})$. This enables us to compute the posterior predictive distribution as a surrogate density,

$$q(\mathbf{y}^\star|\mathbf{y}) := \int_\vartheta q(\mathbf{y}^\star|\boldsymbol{\theta}) q(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} = \sum_{j=1}^{m} v_j(\mathbf{y}^\star) \int_\vartheta \ell(\boldsymbol{\theta}_j, \boldsymbol{\theta}) q(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} = \sum_{j=1}^{m} v_j(\mathbf{y}^\star) \tilde{\mu}_{\boldsymbol{\Theta}|\mathbf{Y}=\mathbf{y}}(\boldsymbol{\theta}_j). \tag{35}$$

This equips scientists with the ability to quantify how likely another set of observed data $\mathbf{y}^\star$ was generated from the same models that also could have generated the observed data $\mathbf{y}$.