Approximate Bayesian Computation via Classification

Yuexi Wang, Tetsuya Kaji and Veronika Ročková *
Booth School of Business, University of Chicago

November 22, 2021

Abstract

Approximate Bayesian Computation (ABC) enables statistical inference in complex models whose likelihoods are difficult to calculate but easy to simulate from. ABC constructs a kernel-type approximation to the posterior distribution through an accept/reject mechanism which compares summary statistics of real and simulated data. To obviate the need for summary statistics, we directly compare empirical distributions with a Kullback-Leibler (KL) divergence estimator obtained via classification. In particular, we blend flexible machine learning classifiers within ABC to automate fake/real data comparisons. We consider the traditional accept/reject kernel as well as an exponential weighting scheme which does not require the ABC acceptance threshold. Our theoretical results show that the rate at which our ABC posterior distributions concentrate around the true parameter depends on the estimation error of the classifier. We derive limiting posterior shape results and find that, with a properly scaled exponential kernel, asymptotic normality holds. We demonstrate the usefulness of our approach on simulated examples as well as real data in the context of stock volatility estimation.

Keywords: *Approximate Bayesian Computation; Classification; Likelihood-free Inference; Kullback-Leibler Divergence; Posterior Concentration*

*The author gratefully acknowledges the support from the James S. Kemper Faculty Fund at the Booth School of Business and the National Science Foundation (Grant No. NSF DMS-1944740)

1 Introduction

Implicit in the deployment of many Bayesian methods is the assumption that the likelihood function can be numerically evaluated. More often than not, however, modern statistical models are prescribed by intractable likelihoods or are defined through an underlying generating mechanism. Fortunately, it may still be possible to simulate synthetic datasets from the model. The ability to simulate from the likelihood has opened up new opportunities for simulating from the posterior. For example, Approximate Bayesian Computation (ABC) (Pritchard et al., 1999; Beaumont et al., 2002) emerged as a default likelihood-free Bayesian inferential tool. It is an accept/reject posterior sampling mechanism which obviates likelihood evaluations. Each iteration proceeds by (1) simulating prior parameter guesses and fake data from the likelihood, and then (2) accepting those parameter values whose fake data were close to the observed data. A big challenge with ABC has been gauging the similitude between observed and fake data.

Measures of similarity between data sets have traditionally been based on summary statistics (see Blum et al. (2013) for an overview within the ABC context). In other words, two datasets are considered similar if their summary statistics are close. In the absence of expert knowledge, however, constructing effective summary statistics can be challenging (Joyce and Marjoram, 2008; Nunes and Balding, 2010; Blum et al., 2013) and one may need to resort to automated strategies. One possibility is regressing parameter values onto (functionals of) fake data in a pilot ABC run to train a flexible mapping which can be substituted for summary statistics (Fearnhead and Prangle, 2012; Jiang et al., 2017; Åkesson et al., 2020). Another possibility, related to indirect inference, is to construct summary statistics from an auxiliary model (Drovandi et al., 2011; Wood, 2010). One can also choose a subset of candidate summary statistics that satisfy some optimality criterion (Joyce and Marjoram, 2008; Nunes and Balding, 2010) or find an optimal projection of a set of summary statistics onto a lower-dimensional map (Fearnhead and Prangle, 2012). Alternatively, one can view the observed and synthetic data sets as empirical distributions and directly use a distance between them inside ABC (such as Kullback-Leibler (Jiang et al., 2018) or Wasserstein (Bernton et al., 2019) or Maximum Mean (MM) discrepancy (Park et al., 2016)). Our work fortifies this last point of view by focusing on the Kullback-Leibler discrepancy estimated via classification.

The KL divergence is one of the most widely used discrepancy metrics. It expresses the average information per observation to discriminate between two probabilistic models (Kullback,

1958). In large deviations, for example, it characterizes the exponential decay rate at which empirical measures converge to their probabilities (see Sanov's theorem in Den Hollander (2008)) and the rate of decay of the probability of error in a binary hypothesis testing problem (see Stein's Lemma in Cover and Thomas (1991)). KL also naturally connects to maximum likelihood estimation through its interpretation as the expectation of the log-likelihood ratio. There exist many methods for estimating the KL divergence. For example, Wang et al. (2009) proposed nearest-neighbor techniques to obtain a mean-square consistent estimator. Wang et al. (2005) proposed a histogram-based KL estimator based on partitioning of the space into statistically equivalent intervals. Silva and Narayanan (2007) and Silva and Narayanan (2010) went a step further and proposed using data-driven partitions (including multivariate recursive partitioning) and formulated sufficient consistency conditions. Alternatively, Nguyen et al. (2007) proposed a variational approach by turning KL estimation into a penalized convex risk minimization problem. Our work is different from the approaches above as we propose a KL estimator based on classification.

We suggest embedding a machine learning classifier inside ABC to determine whether or not fake and observed data are similar and, thereby, whether or not the underlying parameter value should be kept in the ABC reference table. The fundamental premise of this proposal is as follows: parameter values that yield indistinguishable simulated datasets can be deemed close. Bayesian inference via classification has been suggested before. Kaji and Rockova (2021) developed a version of the Metropolis-Hastings algorithm, called MHC, based on classification-based estimators of likelihood ratios. Gutmann et al. (2018) proposed a classification strategy related to ours using a different discrepancy metric. Our paper reframes the method of Gutmann et al. (2018) as a genuine ABC algorithm with a KL divergence discriminator and provides supporting theory which justifies its inferential potential.

In particular, we study statistical properties of the approximate posterior which, in part, depend on the properties of the KL divergence estimator. We consider the traditional accept/reject ABC version (with a uniform kernel) as well as an exponential kernel variant which does not require the ABC tolerance threshold. Similar to Frazier et al. (2018), we show that the choice of the ABC acceptance threshold ϵ plays a critical role in the convergence rate and in the limiting posterior shape. In practice, it is often not obvious what the optimal threshold ϵ should be. Motivated by the connections with the MHC algorithm of Kaji and Rockova (2021), we propose an exponential kernel which yields ABC posteriors that correspond to the stationary distribution

of MHC. Our ABC kernel method can be thus regarded as a parallelizable counterpart to the sequential MHC sampling, targeting the same posterior approximation. The concentration and asymptotic shape behavior of the ABC posterior, which can be derived from Kaji and Rockova (2021), theoretically justify our exponential weighting scheme. In addition, the exponential kernel can be motivated as an instantiation of General Bayesian Inference (GBI) (Bissiri et al., 2016) with the KL estimator as the loss function.

The remaining of the paper is structured as follows. In Section 2, we flesh out the basic idea of ABC and introduce our framework with classification. In Section 3, we investigate the posterior concentration and limiting shape behaviors of the ABC posteriors. Section 4 shows performance on simulated datasets and Section 5 further highlights the practical value of our approach on real data. In Section 6, we conclude with a discussion.

Notation. We employ the operator notation for expectation, e.g., $P_0 f = \int f dP_0$. The ϵ-bracketing number $N_{[]}(\epsilon, \mathcal{F}, d)$ of a set $\mathcal{F}$ with respect to a premetric d is the minimal number of ϵ-brackets in d needed to cover $\mathcal{F}$[1]. The δ-bracketing entropy integral of $\mathcal{F}$ with respect to d is $J_{[]}(\delta, \mathcal{F}, d) = \int_0^\delta \sqrt{1 + \log N_{[]}(\epsilon, \mathcal{F}, d)} d\epsilon$. Next, $K(f, g) = \int f \log(f/g) d\mu$ denotes the Kullback-Leibler divergence between two density functions and $V_2(f, g) = \int f \left|\log(f/g)\right|^2 d\mu$. For real-valued sequences $\{a_n\}_{n\geq 1}$ and $\{b_n\}_{n\geq 1}$, $a_n \lesssim b_n$ means that $a_n \leq C\, b_n$ for some generic constant $C > 0$, $a_n \asymp b_n$ means that $a_n \lesssim b_n \lesssim a_n$, and $a_n \gg b_n$ indicates a greater order of magnitude. For a sequence of random variables x_n, $x_n = o_P(a_n)$ if $\lim_{n\to\infty} P(|x_n/a_n| \geq C) = 0$ for every $C > 0$, and $x_n = O_P(a_n)$ if for every $C > 0$ there exists a finite $M > 0$ and a finite N such that $P(|x_n/a_n| \geq M) \leq C$ for all $n > N$. All limits are taken as $n \to \infty$. Take $\|\cdot\|$ to be the Euclidean norm.

2 ABC without Summary Statistics

Our framework consists of a collection of i.i.d. observations $\boldsymbol{X} = (X_1, \ldots, X_n)'$ where each $X_i \in \mathcal{X}$ is realized from a parametric model $\{P_{\theta_0} : \theta_0 \in \Theta \subset \mathbb{R}^d\}$. We assume that P_θ, for each $\theta \in \Theta$, admits a density p_θ. Hereafter, we will use the shorthand notation $p_0 = p_{\theta_0}$ and $P_0 = P_{\theta_0}$. We are interested in Bayesian inference about θ_0 based on the posterior distribution

$$\pi_n(\theta \mid \boldsymbol{X}) \propto p_\theta^{(n)}(\boldsymbol{X}) \pi(\theta) \tag{1}$$

[1] A premetric on $\mathcal{F}$ is a function $d : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ such that $d(f, f) = 0$ and $d(f, g) = d(g, f) \geq 0$.

prescribed by the likelihood $p_\theta^{(n)}(\boldsymbol{X})$ and the prior density $\pi(\theta)$. We are particularly interested in scenarios when the likelihood function cannot be directly expressed/evaluated (such as discretely observed diffusions (Sørensen, 2004) or generative models) but can be sampled from.

The now default ABC method for Bayesian likelihood-free inference constructs a nested kernel-type approximation to the posterior distribution. The first approximation occurs when the data is distilled into summary statistics to obtain $\pi(\theta \mid S_X) \propto \pi(S_X \mid \theta)\pi(\theta)$, where $S_X = S(\boldsymbol{X})$ is a vector of summary statistics. The quality of this approximation depends crucially on the informativeness of S_X. The actual ABC approximation to the posterior (1) is then constructed via a kernel function as $\pi_{ABC}(\theta \mid S_X) = \int \pi(\theta, S \mid S_X) dS$ with $\pi(\theta, S) \propto T_\epsilon(\|S - S_Y\|)\pi(S \mid \theta)\pi(\theta)$, where $T_\epsilon(\|u\|) = T(\|u\|/\epsilon)$ is a standard (smoothing) kernel with a scale parameter $\epsilon > 0$. The key two challenges with ABC are (1) deriving low-dimensional summary statistics with a minimal loss of information and (2) selecting the kernel and its the tolerance level ϵ. To remediate the reliance of ABC on summary statistics, we focus on viewing the observed and fake data as empirical distributions and gauge the distance between them. Regarding the choice of aggregation kernels, we consider the traditional uniform kernel yielding an accept/reject algorithm and a smoothing kernel free from ϵ-tuning.

2.1 ABC with KL Divergence

Instead of summary statistics, we use the estimated KL divergence inside the ABC algorithm. Our interest in the KL divergence as a discrepancy metric stems partially from the following connection to the generalized Bayesian inference (Bissiri et al., 2016). The posterior distribution (1) can be rewritten as a generalized posterior $\pi_n(\theta \mid \boldsymbol{X}) \propto \pi(\theta) \exp(-n \times KL(p_0^{(n)}, p_\theta^{(n)}))$ where the parameter θ is linked to data through the empirical Kullback-Liebler (KL) divergence $KL(p_0^{(n)}, p_\theta^{(n)}) \equiv \frac{1}{n}\sum_{i=1}^n \log(p_0/p_\theta)(X_i)$. For when the KL divergence cannot be easily evaluated, we consider various estimators in the next section. We denote a generic KL divergence estimator obtained from observed data $\boldsymbol{X} \sim P_0^{(n)}$ and pseudo-data $\tilde{\boldsymbol{X}}^\theta \sim P_\theta^{(n)}$ as $\hat{K}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^\theta)$. Adopting $\hat{K}(\cdot, \cdot)$ as the ABC summary statistics, we consider a simple accept/reject ABC mechanism detailed in Algorithm 1 below. While Jiang et al. (2017) used a nearest-neighbor estimator of the KL divergence, we devise a different estimator based on classification in Section 2.2.

Algorithm 1 simulates pairs of parameter values and pseudo-data $\{\theta, \tilde{\boldsymbol{X}}^\theta\}$ from the joint

Algorithm 1. KL-ABC with Accept-Reject

For a pre-determined tolerance level $\epsilon > 0$ repeat for $j = 1, \ldots, N$:

1. Simulate θ_j from $\pi(\theta)$.
2. Simulate $\tilde{\boldsymbol{X}}^{\theta_j} = (\tilde{X}_1^{\theta_j}, \ldots, \tilde{X}_m^{\theta_j})'$ through i.i.d sampling from the model p_{θ_j}.
3. Accept θ_j when $\hat{K}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^{\theta_j}) \leq \epsilon$.

posterior density

$$\hat{\pi}^{AR}\left(\theta, \tilde{\boldsymbol{X}}^{\theta} \mid \hat{K}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^{\theta}) \leq \epsilon\right) = \frac{\pi(\theta) p_{\theta}^{(n)}(\tilde{\boldsymbol{X}}^{\theta}) \mathbb{I}\big(\hat{K}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^{\theta}) \leq \epsilon\big)}{\int \pi(\theta) p_{\theta}^{(n)}(\tilde{\boldsymbol{X}}^{\theta}) \mathbb{I}\big(\hat{K}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^{\theta}) \leq \epsilon\big) \mathrm{d}\tilde{\boldsymbol{X}}^{\theta} \mathrm{d}\theta}, \tag{2}$$

which margins towards the following approximate (Accept/Reject) posterior density

$$\begin{aligned} \hat{\pi}_{\epsilon}^{AR}(\theta \mid \boldsymbol{X}) &= \int \hat{\pi}^{AR}\left(\theta, \tilde{\boldsymbol{X}}^{\theta} \mid \hat{K}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^{\theta}) \leq \epsilon\right) \mathrm{d}\tilde{\boldsymbol{X}}^{\theta} \\ &\equiv \frac{\pi(\theta) P_{\theta}^{(n)}\big(\hat{K}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^{\theta}) \leq \epsilon\big)}{\int \pi(\theta) P_{\theta}^{(n)}\big(\hat{K}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^{\theta}) \leq \epsilon\big) \mathrm{d}\theta}. \end{aligned} \tag{3}$$

The approximation (3) will be scrutinized theoretically later in Section 3.1. In particular, we will later see that the convergence rate of (3) around θ_0 depends on the choice of ϵ (Frazier et al., 2018). Algorithm 1 uses the uniform kernel which corresponds to the indicator function $T_{\epsilon}(\|u\|) = \mathbb{I}(\|u\| \leq \epsilon)$. In practice, it is difficult to balance out conflicting demands of smaller ϵ (yielding good approximability) and larger acceptance rates (yielding more posterior samples). As a remedy, we propose a way to aggregate the ABC samples through a scaled exponential kernel motivated by the connection between KL and the log-likelihood ratio. This ABC variant requires no ad-hoc thresholding and is summarized in Algorithm 2.

Algorithm 2. KL-ABC with Exponential Weighting

Repeat for $j = 1, \ldots, N$:

1. Simulate θ_j from $\pi(\theta)$.
2. Simulate $\tilde{\boldsymbol{X}}^{\theta_j} = (\tilde{X}_1^{\theta_j}, \ldots, \tilde{X}_m^{\theta_j})'$ through i.i.d sampling from the model p_{θ_j}.
3. Assign θ_j a weight proportional to $\exp\big(-n\hat{K}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^{\theta})\big)$.

Algorithm 2 generates draws for the pair $\{\theta, \tilde{\boldsymbol{X}}^\theta\}$ from a joint posterior density

$$\hat{\pi}^{EK}(\theta, \tilde{\boldsymbol{X}}^\theta \mid \boldsymbol{X}) = \frac{\pi(\theta) p_\theta^{(n)}(\tilde{\boldsymbol{X}}^\theta) \exp\big(-n\hat{K}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^\theta)\big)}{\int \pi(\theta) p_\theta^{(n)}(\tilde{\boldsymbol{X}}^\theta) \exp\big(-n\hat{K}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^\theta)\big) \mathrm{d}\tilde{\boldsymbol{X}}^\theta \mathrm{d}\theta}, \tag{4}$$

which leads to the approximated Bayesian posterior as

$$\hat{\pi}^{EK}(\theta \mid \boldsymbol{X}) = \int \hat{\pi}^{EK}(\theta, \tilde{\boldsymbol{X}}^\theta \mid \boldsymbol{X}) \mathrm{d}\tilde{\boldsymbol{X}}^\theta = \frac{\pi(\theta) P_\theta^{(n)} \exp\big(-n\hat{K}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^\theta)\big)}{\int \pi(\theta) P_\theta^{(n)} \exp\big(-n\hat{K}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^\theta)\big) \mathrm{d}\theta}. \tag{5}$$

Smooth kernels have been used inside ABC before to rescale the acceptance probability (e.g. Beaumont et al. (2002) employed the Epanechnikov kernel). Wilkinson (2013) and Sisson et al. (2018) provide a thorough overview and comparisons of the commonly used kernels. Our smoothed weights are directly interpretable due to their linkage between the KL divergence and the log-likelihood ratio. We later show in Section 3.1 that, with the scaled exponential kernel, the ABC posterior corresponds to the stationary distribution of the MHC algorithm of Kaji and Rockova (2021) and can be regarded as a posterior under a misspecified model.

Remark 2.1 *(Generating Fake Data) We assume $\tilde{\boldsymbol{X}}^\theta = g_\theta(\tilde{\boldsymbol{X}})$, where $\tilde{\boldsymbol{X}} \in \mathbb{R}^m$ are random variables arriving from $\tilde{P}^{(m)}$ and where $g_\theta : \mathbb{R}^m \to \mathbb{R}^m$ is a deterministic mapping. Generating random variable draws by passing $\tilde{\boldsymbol{X}}$ through some mapping is commonly done in practice. For example, Gaussian random variables $\tilde{\boldsymbol{X}} = \{\tilde{X}_i^\theta\}_{i=1}^m$ that follow i.i.d. $N(\mu, \sigma^2)$ distribution can be obtained by transforming $\{\tilde{X}_i\}_{i=1}^m \overset{i.i.d.}{\sim} N(0,1)$ via $\tilde{X}_i^\theta = \mu + \sigma\tilde{X}_i$. In other cases, one can use uniform draws $\tilde{\boldsymbol{X}}$ and the inverse transform sampling.*

2.2 Estimating KL Divergence via Classification

We adopt the '$-\log D$' trick to estimate the KL divergence (Goodfellow et al., 2014). More precisely, a flexible discriminator D (such as a neural network or logistic regression) is trained to maximize

$$\mathbb{M}_{n,m}^\theta(D) = \mathbb{P}_n \log D + \mathbb{P}_m^\theta \log(1 - D), \tag{6}$$

where we employ the operator notation for expectation, e.g., $\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^n f(X_i)$ and $\mathbb{P}_m^\theta f = \frac{1}{m}\sum_{i=1}^m f(\tilde{X}_i^\theta)$. This can also be regarded as a classification problem where we label $\{X_i\}_{i=1}^n$ ('real' data) with 1 and $\{\tilde{X}_i^\theta\}_{i=1}^m$ ('fake' data) with 0. The oracle maximizer to (6) can be shown

to be (Goodfellow et al., 2014, Proposition 1)

$$D_\theta(X) = \frac{p_0(X)}{p_0(X) + p_\theta(X)}. \tag{7}$$

The functional form of the oracle solution in (7) naturally suggests the following KL estimator obtained from a trained discriminator $\hat{D}^\theta_{n,m}$

$$\hat{K}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^\theta) = \mathbb{P}_n \log \frac{\hat{D}^\theta_{n,m}}{1 - \hat{D}^\theta_{n,m}} = \frac{1}{n}\sum_{i=1}^{n} \log \frac{\hat{D}^\theta_{n,m}(X_i)}{1 - \hat{D}^\theta_{n,m}(X_i)}. \tag{8}$$

Next, we show that our classification-based KL estimator (8) converges to a well-defined limit counterpart $K(p_0, p_\theta)$ under mild conditions. We assume that the set of considered classifiers $\mathcal{D}$ resides in a sieve $\mathcal{D}_n$ that expands with the sample size and its size is measured by bracketing entropy $N_{[]}(\epsilon, \mathcal{F}, d)$. Kaji et al. (2020) prove the rate of convergence of such a classifier (under assumptions reviewed later) with a Hellinger-type distance defined as

$$d_\theta(D_1, D_2) = \sqrt{h_\theta(D_1, D_2)^2 + h_\theta(1 - D, 1 - D_\theta)^2},$$

where $h_\theta(D_1, D_2) = \sqrt{(P_0 + P_\theta)(\sqrt{D}_1 - \sqrt{D}_2)^2}$.

Assumption 1 *(Kaji et al., 2020, Assumption 3) Assume that n/m converges and that an estimator $\hat{D}^\theta_{n,m}$ exists that satisfies $\mathbb{M}_{n,m}(\hat{D}^\theta_{n,m}) \geq \mathbb{M}^\theta_{n,m}(D_\theta) - O_P(\delta_n^2)$ for a nonnegative sequence δ_n. Moreover, assume that the bracketing entropy integral satisfies $J_{[]}(\delta_n, \mathcal{D}^\theta_{n,\delta_n}, d_\theta) \lesssim \delta_n^2\sqrt{n}$ and that there exists $\alpha < 2$ such that $J_{[]}(\delta_n, \mathcal{D}^\theta_{n,\delta_n}, d_\theta)/\delta^\alpha$ is increasing in δ. Here $\mathcal{D}^\theta_{n,\delta_n} = \{D \in \mathcal{D}_n : d_\theta(D, D_\theta) \leq \delta\}$.*

Lemma 2.1 *(Kaji et al., 2020, Theorem 1') Under Assumption 1, $d_\theta(\hat{D}^\theta_{n,m}, D_\theta) = O^*_P(\delta_n)$.* [2]

In Assumption 1, we only assume that δ_n be a nonnegative sequence, not necessarily vanishing. To ensure that the approximate posterior converges, we require $\delta_n \to 0$ as $n \to \infty$. Kaji et al. (2020) show that if the true likelihood ratio has a low-dimensional representation and an appropriate neural network is used for the discriminator, the rate δ_n depends only on the underlying dimension and it vanishes when $n \to \infty$. This result can also be generalized to other network configurations, including Schmidt-Hieber (2020) and Yarotsky (2017). Hence, we implicitly assume $\delta_n = o(1)$ and note that δ_n will typically be slower than $n^{-1/2}$.

[2] We use P^* to denote outer expectation (definition see Section 1.2 of van der Vaart and Wellner (1996)), here is the expectation of a "a smallest measureable function g that dominates $d_\theta(\hat{D}^\theta_{n,m}, D_\theta)$".

To establish the rate of convergence of our approximated posterior, we have the following assumption.

Assumption 2 *(Kaji and Rockova, 2021, Assumption 2) There exists $\Lambda > 0$ such that for every $\theta \in \Theta$, $P_0(p_0/p_\theta)$ and $P_0(p_0/p_\theta)^2$ are bounded by Λ and*

$$\sup_{D \in \mathcal{D}^\theta_{n,\delta_n}} P_0\left(\frac{D_\theta}{D}\middle|\frac{D_\theta}{D} \geq \frac{25}{16}\right) < \Lambda, \quad \sup_{D \in \mathcal{D}^\theta_{n,\delta_n}} P_0\left(\frac{1-D_\theta}{1-D}\middle|\frac{1-D_\theta}{1-D} \geq \frac{25}{16}\right) < \Lambda,$$

for δ_n in Assumption 1. The brackets in Assumption 1 can be taken so that $P_0(\sqrt{u/l}-1)^2 = O(d_\theta(u,l)^2)$ and $P_0(\sqrt{(1-l)/(1-u)}-1)^2 = o(d_\theta(u,l))$.

The following theorem quantifies the rate of convergence of our estimator (8) towards the *empirical* KL divergence $\mathbb{P}_n \log \frac{p_0}{p_\theta}$.

Lemma 2.2 *(Convergence Rate of Estimation Errors) Under Assumptions 1 and 2,*

$$\left|\hat{K}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^\theta) - \mathbb{P}_n \log \frac{p_0}{p_\theta}\right| = O_{P^*}(\delta_n). \tag{9}$$

Proof. Since the estimation error can be rewritten as

$$\hat{K}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^\theta) - \mathbb{P}_n \log \frac{p_0}{p_\theta} = -\mathbb{P}_n\Big(\log \frac{1-\hat{D}^\theta_{n,m}}{1-D_\theta} - \log \frac{\hat{D}^\theta_{n,m}}{D_\theta}\Big),$$

it follows from Kaji and Rockova (2021, Theorem 4.1).

Remark 2.2 *(Uniform Convergence Rate) The rate of estimation in Lemma 2.1 and Lemma 2.2 is characterized as point-wise. While for each $\theta \in \Theta$ the estimation error is shrinking at the rate δ_n, the multiplication constant in front the rate could potentially depend on θ. Assuming a compact parameter support (which is not a-typical in deep learning models; see e.g. Schmidt-Hieber (2020) or Polson and Rockova (2018) and Wang and Rockova (2020)) and continuity of the multiplication constant, we can essentially regard the rate as uniform. Hereafter, we thereby abuse the notation and tacitly assume that δ_n is the worst rate over all $\theta \in \Theta \subset \mathbb{R}^d$, i.e. the rate with the largest multiplication constant.*

Next, we investigate convergence around the *actual* KL divergence $K(p_0, p_\theta)$. The next lemma will be utilized later in the proof of Theorem 3.1. However, it is of independent interest as it shows how the speed at which the joint error probability (accounting for randomness

of both the observed and fake data $(\boldsymbol{X}, \tilde{\boldsymbol{X}})$) decays in terms of the estimation error. Below, the probability P corresponds to $P_0^{(n)} \otimes \tilde{P}^{(m)}$, where $\tilde{P}^{(m)}$ is the measure for $\tilde{\boldsymbol{X}}$ (see Remark 2.1).

Theorem 2.1 *For a given $\theta \in \Theta$, we define for $u > 0$ and $\delta_n > 0$ as in Lemma 2.1 and for an arbitrarily slowly increasing sequence $C_n > 0$*

$$\rho_{n,\theta}(u; C_n; \delta_n) \equiv P\bigg(\Big| \hat{K}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^{\theta}) - K(p_0, p_\theta) \Big| > 2u, d_\theta(\hat{D}^{\theta}_{n,m}, D^{\theta}) \leq C_n \delta_n \bigg). \tag{10}$$

Under Assumptions 1 and 2, we then have

$$\rho_n(u; C_n; \delta_n) \equiv \sup_{\theta \in \Theta} \rho_{n,\theta}(u; C_n; \delta_n) = O\Big(\frac{C_n \delta_n}{u} + \frac{1}{nu^2} \Big).$$

The proof can be found in Appendix D.1. Note that the intersecting event above has probability converging to one, according to Lemma 2.1.

2.3 Other KL Estimators

Our ABC results, presented later in Section 3, can be extended to other types of KL estimators if similar estimation error results as in Theorem 2.1 can be shown. One example is the k-Nearest Neighbor (kNN) estimator proposed in Pérez-Cruz (2008). Wang et al. (2009) showed that this estimator is asymptotically unbiased and mean-square consistent and they propose a data-dependent choice of k which can improve the convergence speed. Jiang et al. (2018) assess data discrepancy inside ABC with the special case of 1-nearest neighbor, which is defined as

$$\hat{K}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^{\theta}) = \frac{d}{n} \sum_{i=1}^{n} \log \frac{\min_j \left\| X_i - \tilde{X}_j^{\theta} \right\|}{\min_{j \neq i} \| X_i - X_j \|} + \log \frac{m}{n-1}. \tag{11}$$

where d is the number of covariates in $\boldsymbol{X}$. Zhao and Lai (2020) provide estimation error rate results for this kNN estimator.

Another route to estimate the KL divergence is via (data-dependent) partitioning methods. Wang et al. (2005) proposed to estimate the Radon-Nikodym derivative $\mathrm{d}P_0/\mathrm{d}P_\theta$ using frequency counts on a statistically equivalent partition of $\mathbb{R}^d$. However, the computational complexity of their method is exponential in d and the estimation accuracy deteriorates quickly as the dimension increases. Silva and Narayanan (2010) further contributed to multivariate data-driven partition schemes by using a Barron-type histogram-based density estimate. They provide sufficient conditions on the partitions scheme to make the estimator strongly consistent.

Lastly, Nguyen et al. (2007) adopted a variational approach to estimate KL by reframing the estimation problem as a penalized convex risk minimization problem, where they restrict the estimate to a bounded subset of a Reproducing Kernel Hilbert Space (RKHS). Convergence rates are then obtained from empirical process theory on nonparametric M-estimators (van de Geer, 2000). In an independent contribution, Ghimire et al. (2020) used a discriminator in RKHS to estimate KL using a similar approach to ours. They showed that the estimator error bound is related to the complexity of the discriminator in RKHS.

2.4 Generalizations to GAN-style Discrepancy Metrics

Beyond the KL divergence, our classification framework allows us to consider other discrepancy metrics. Alternatively to (8), we could instead estimate the reversed KL divergence

$$\hat{K}_{\text{reverse}}(\tilde{\boldsymbol{X}}^\theta, \boldsymbol{X}) = \frac{1}{m}\sum_{i=1}^{m} \log \frac{1-\hat{D}^\theta_{n,m}}{\hat{D}^\theta_{n,m}}(\tilde{X}^\theta_i)$$

which converges to $K(p_\theta, p_0)$ and which is still uniquely minimized at $p_\theta = p_0$. One can show that the estimation error of this reversed KL estimator is still $O_{P^*}(\delta_n)$ by following the same techniques used in Lemma 2.2. The reversed KL divergence is widely used in variational inference (Jordan et al., 1999; Wainwright and Jordan, 2008). Forward and reversed KL's could perform differently when the function class inside the variational approach is not rich enough. The reversed KL is zero-forcing/mode-seeking, while the forward KL is mass-covering/mean-seeking (Bishop, 2006). Another related metric, deployed by Gutmann et al. (2018), is the classification accuracy (CA) defined as

$$\text{CA}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^\theta) = \frac{1}{n+m}\Big(\sum_{i=1}^{n} \hat{D}^\theta_{n,m}(X_i) + \sum_{i=1}^{m}\big(1-\hat{D}^\theta_{n,m}(\tilde{X}^\theta_i)\big)\Big). \tag{12}$$

Since $\hat{D}(\cdot)$ and $\log\frac{\hat{D}}{1-\hat{D}}(\cdot)$ are linked by a logistic transformation which is Lipschitz-continous, CA can be roughly regarded as a weighted average of the forward KL divergence and the reversed KL divergence. All three metrics are associated with Jensen-Shannon (JS) divergence, each having a slightly different emphasis.

There are many other divergences that can be estimated using different configurations of (6).

Given a set $\mathcal{F}$ of functions from $\mathcal{X}$ to $\mathbb{R}$, we can define

$$d_{\mathcal{F}}(p_0, p_\theta) = \sup_{f \in \mathcal{F}} \mathbb{P}_n f(X_i) - \mathbb{P}_m^\theta f(\tilde{X}_i^\theta), \tag{13}$$

as the Integral Probability Metric (IPM) (Müller, 1997) associated with the function class $\mathcal{F}$. Arjovsky et al. (2017) show that different choices of the function class $\mathcal{F}$ can dramatically change the topology, as we discuss below, and the regularity of $d_{\mathcal{F}}(p_0, p_\theta)$ as a loss function. Several widely used discrepancy metrics can be obtained from IPMs. For example, the 1-Wasserstein distance or the Earth-Mover distance (Villani, 2009) are obtained when $\mathcal{F}$ is the set of 1-Lipschitz functions. Arjovsky et al. (2017) proposed Wasserstein GANs, which approximate the 1-Wasserstein distance with gradient-clipped neural networks. Another example is the Total Variation (TV) distance, obtained when $\mathcal{F}$ is the set of all measurable functions bounded between -1 and 1. Next, the Maximum Mean (MM) discrepancy is obtained when $\mathcal{F} = \{f \in \mathcal{H} : \|f\|_\infty \leq 1\}$ for $\mathcal{H}$ some Reproducing Kernel Hilbert Space (RKHS) associated with a given kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Our classification-based GAN framework thus provides an alternative route towards implementing ABCs with the Wasserstein distance (Bernton et al., 2019) as well as MM (Park et al., 2016).

One important aspect of these various metrics is that they induce topologies of different strengths. A divergence makes it easier for a sequence of distributions to converge when it induces a weaker topology. Arjovsky et al. (2017) show that KL induces the strongest one, followed by JS and TV, with Wasserstein distance being the weakest. However, the computation costs of the Wasserstein distance are high, i.e. $O((n+m)^3 \log(n+m))$ to calculate it exactly (Burkard et al., 2012). The GAN version approximation also requires weight clipping to enforce the Lipschitz constraint which can be problematic to implement. Our KL estimator, on the other hand, is more scalable and places no constraints on the discriminator. The neural network discriminator can converge faster than traditional nonparametric methods if some low-dimensionality assumption is satisfied (Kaji et al., 2020; Bauer and Kohler, 2019). In addition, the KL estimator has a natural interpretation through its link to the log-likelihood ratio. The KL divergence appears in many theoretical results in Bayesian non-parametrics. For example, Schwartz's Theorem (Schwartz, 1965) requires that, for posterior consistency, the prior should assign positive probability to any KL neighborhood of p_0. By adopting the KL estimator within our framework, we can relate ABC to classical posterior convergence rate results.

3 Frequentist Analysis of the ABC Posterior

One way to assess the quality of the posterior distribution is through the speed at which it contracts around the truth θ_0 as $n \to \infty$. While the ABC posterior is ultimately an approximation, it might still concentrate about θ_0 at a reasonable rate. Concentration rates are typically quantified in terms of a prior concentration (measured in terms of a combination of the KL divergence and the KL variation) and the entropy of the model. We have a similar prior mass condition (see (8.4) in Section 8.2 in Ghosal and Van der Vaart (2017)). We denote the KL-neighborhood of p_0 by

$$B_2(p_0, \epsilon) = \{\theta : K(p_0, p_\theta) < \epsilon^2\}. \tag{14}$$

Assumption 3 *(Prior Mass) There exist some constants $\kappa > 0$ and $\xi > 0$ such that for every $0 < \epsilon < \xi$ and some constant $C > 0$, the prior probability satisfies $\Pi[B_2(p_0, \epsilon)] \geq C\epsilon^\kappa$.*

Next, we assume that the parameter θ is identifiable in the sense that the KL divergence is locally compatible with the Euclidean norm. This assumption is adopted from Assumption 3(ii) of Frazier et al. (2018).

Assumption 4 *(Identification) The density function p_θ is continuous in θ and for every θ in some open neighborhood of θ_0 satisfies*

$$\|\theta - \theta_0\| \leq L \times K(p_0, p_\theta)^\alpha$$

for some $L > 0$ and $\alpha > 0$.

Similarly as in (5.1) in Kleijn and van der Vaart (2006), Assumption 4 ensures posterior concentration around θ_0 when $K(p_0, p_\theta) \to 0$. This holds for many distributions. For example, for the exponential distribution with a rate parameter θ, we have $K(p_0, p_\theta) = \frac{\theta}{\theta_0} - \log(\frac{\theta}{\theta_0}) - 1$. Since $\log(1 + x) = x - \frac{x^2}{2} + o(x^2)$ when $x \to 0$, we have $K(p_0, p_\theta) \geq \frac{1}{2}\theta_0^{-2}(\theta - \theta_0)^2$. For multivariate normal distribution with a known variance Σ and an unknown location μ, we have $K(p_0, p_\theta) = \frac{1}{2}(\mu - \mu_0)\Sigma^{-1}(\mu - \mu_0) \geq \frac{1}{2}\rho(\Sigma)^{-1} \|\mu - \mu_0\|^2$, where $\rho(\Sigma)$ is the spectral radius, i.e., the largest eigenvalue, of a matrix Σ.

3.1 Posterior Concentration Rate

First, we focus on the uniform kernel $T_\epsilon(x) = \mathbb{I}(|x| \leq \epsilon)$ used in Algorithm 1. Recall (from Remark 2.1) that $\tilde{\boldsymbol{X}}^\theta = g_\theta(\tilde{\boldsymbol{X}})$ for some suitable mapping $g_\theta(\tilde{\boldsymbol{X}})$ where $\tilde{\boldsymbol{X}} \sim \tilde{P}^{(m)}$. The ABC

joint posterior (2) is a weighted aggregation of uniform kernels, i.e

$$\hat{\pi}^{AR}\left(\theta, \tilde{\boldsymbol{X}} \mid \hat{K}[\boldsymbol{X}, g_\theta(\tilde{\boldsymbol{X}})] \leq \epsilon\right) = \frac{\tilde{\pi}(\tilde{\boldsymbol{X}})\pi(\theta)\mathbb{I}\big(\hat{K}[\boldsymbol{X}, g_\theta(\tilde{\boldsymbol{X}})] \leq \epsilon\big)}{\int \tilde{\pi}(\tilde{\boldsymbol{X}})\pi(\theta)\mathbb{I}\big(\hat{K}[\boldsymbol{X}, g_\theta(\tilde{\boldsymbol{X}})] \leq \epsilon\big)\mathrm{d}\tilde{\boldsymbol{X}}\mathrm{d}\theta}, \tag{15}$$

which yields the following ABC posterior distribution

$$\hat{\Pi}^{AR}_{\epsilon_n}(A \mid \boldsymbol{X}) = \frac{\int_{\theta\in A}\pi(\theta)\tilde{P}^{(m)}\big(\hat{K}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^\theta) \leq \epsilon_n\big)\mathrm{d}\theta}{\int_\Theta \pi(\theta)\tilde{P}^{(m)}\big(\hat{K}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^\theta) \leq \epsilon_n\big)\mathrm{d}\theta} \quad \text{for a Borel-measurable } A \subset \Theta. \tag{16}$$

The following theorem (a modification of Theorem 1 in Frazier et al. (2018)) quantifies the concentration rate in terms of the tolerance threshold ϵ_n as well as the rate at which the classification-based KL estimator can estimate $\mathbb{P}_n \log(p_0/p_\theta)$ (as formulated in (9)).

Theorem 3.1 *Let Assumptions 1, 2 and 3 hold and take δ_n as in* (9) *in Lemma 2.2. Then, as $n \to \infty$ and with $\epsilon_n = o(1)$ such that $n\epsilon_n^2 \to \infty$ and $C_n\delta_n = o(\epsilon_n)$ for some arbitrarily slowly increasing sequence $C_n > 0$ we have*

$$P_0^{(n)}\Pi[K(p_0, p_\theta) > \lambda_n \mid \hat{K}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^\theta) \leq \epsilon_n] = o(1), \tag{17}$$

where $\lambda_n = \epsilon_n + M_nC_n\delta_n\epsilon_n^{-\kappa} + \sqrt{M_n}n^{-1/2}\epsilon_n^{-\kappa/2}$ for some arbitrarily slowly increasing sequence $M_n > 0$. Moreover, if Assumption 4 also holds, as $n \to \infty$, we have

$$P_0^{(n)}\Pi[\|\theta - \theta_0\| > L\lambda_n^\alpha \mid \hat{K}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^\theta) \leq \epsilon_n] = o(1). \tag{18}$$

The proof of the theorem is provided in Appendix D.2. Thus, the convergence rate of our ABC posterior depends on three components: the accept-reject threshold ϵ_n, the estimation error of the KL estimator δ_n and the rate of discrepancy $n^{-1/2}$ between the empirical and true KL divergence. Since δ_n will typically be greater than the parametric rate $n^{-1/2}$, the overall convergence rate is then driven by $\lambda_n = \epsilon_n + \widetilde{M}_n\delta_n\epsilon_n^{-\kappa}$, where $\widetilde{M}_n$ is an arbitrarily slowly increasing sequence.

In practice, it is unclear how to properly choose ϵ_n. In Algorithm 2, we proposed to weight the draws using a scaled exponential kernel $\exp(-n\hat{K}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^\theta))$. We denote the ABC posterior under the exponential kernel as $\hat{\Pi}^{EK}(\cdot \mid \boldsymbol{X})$ where

$$\hat{\Pi}^{EK}(A \mid \boldsymbol{X}) = \frac{\int_A \tilde{P}^{(m)}\exp\big(-n\hat{K}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^\theta)\big)\pi(\theta)\mathrm{d}\theta}{\int_\Theta \tilde{P}^{(m)}\exp\big(-n\hat{K}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^\theta)\big)\pi(\theta)\mathrm{d}\theta}. \tag{19}$$

To gain more insights into the ABC posterior behavior under the exponential kernel, we take

a closer look at the "likelihood function" above

$$\tilde{P}^{(m)} \exp\big(- n\hat{K}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^{\theta})\big) = \frac{p_{\theta}^{(n)}}{p_{0}^{(n)}} \tilde{P}^{(m)} \mathrm{e}^{u_{\theta}},$$

where $u_{\theta}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^{\theta}) = -n \times \Big(\hat{K}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^{\theta}) - \mathbb{P}_n \log \frac{p_0}{p_{\theta}}\Big)$. From the equations above, we can write

$$\hat{\pi}^{EK}(\theta \mid \boldsymbol{X}) \propto p_{\theta}^{(n)}(\boldsymbol{X}) \times \mathrm{e}^{\hat{u}_{\theta}(\boldsymbol{X})} \times \pi(\theta) \quad \text{with} \quad \hat{u}_{\theta}(\boldsymbol{X}) = \log \int \mathrm{e}^{u_{\theta}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^{\theta})} \mathrm{d}\tilde{P}^{(m)}(\tilde{\boldsymbol{X}}). \quad (20)$$

When $\hat{K}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^{\theta})$ is the classification-based estimator, $\hat{u}_{\theta}(\boldsymbol{X})$ can be related to the random generator setting of the Metropolis-Hastings MHC algorithm in Kaji and Rockova (2021) which has (20) as its stationary distribution. Similarly as in Appendix Section 5 of Kaji and Rockova (2021), we can regard the posterior approximation in (20) as a posterior $\hat{\pi}^{EK}(\theta \mid \boldsymbol{X}) \propto q_{\theta}^{(n)} \tilde{\pi}(\theta)$ under a misspecified likelihood

$$q_{\theta}^{(n)} = \frac{p_{\theta}^{(n)}(\boldsymbol{X}) \mathrm{e}^{\hat{u}_{\theta}(\boldsymbol{X})}}{C_{\theta}} \quad \text{where} \quad C_{\theta} = \int_{\mathcal{X}} p_{\theta}^{(n)}(\boldsymbol{X}) \mathrm{e}^{\hat{u}_{\theta}(\boldsymbol{X})} \mathrm{d}\boldsymbol{X} \quad (21)$$

and a modified prior $\tilde{\pi}(\theta) \propto \pi(\theta) C_{\theta}$. Since the likelihood is misspecified, the ABC posterior concentrates around a projection point θ^* defined as

$$\theta^* = \arg\min_{\theta \in \Theta} -P_0^{(n)} \log[q_{\theta}^{(n)} / p_0^{(n)}], \quad (22)$$

which corresponds to the mis-specified model that is closest to $P_0^{(n)}$ in the KL sense (Kleijn and van der Vaart, 2006). Kaji and Rockova (2021) study posterior concentration of (20). We recall their Theorem 4.5 in Appendix A. Unlike in Theorem 3.1, the posterior concentration rate here depends both on the estimation error δ_n and the actual concentration rate of the true posterior, not the acceptance threshold.

Remark 3.1 *(Vanishing Bias) To better understand the severity of the centering bias of the misspecified model, we note*

$$\begin{aligned} -P_0^{(n)} \log[q_{\theta^*}^{(n)} / p_0^{(n)}] \leq -P_0^{(n)} \log[q_{\theta_0}^{(n)} / p_0^{(n)}] &= P_0^{(n)} \log \frac{p_0^{(n)}}{p_0^{(n)} \mathrm{e}^{\hat{u}_{\theta_0}(\boldsymbol{X})} / C_{\theta_0}} \\ &= \log C_{\theta_0} - P_0^{(n)} \hat{u}_{\theta_0}(\boldsymbol{X}) = \log P_0^{(n)} \mathrm{e}^{\hat{u}_{\theta_0}(\boldsymbol{X})} - P_0^{(n)} \hat{u}_{\theta_0}(\boldsymbol{X}). \end{aligned}$$

This is essentially the Jensen gap. If we have this Jensen gap vanishing when $n \to \infty$, then we can conclude that the centering bias is also vanishing, and the ABC posterior in (19) *will*

eventually concentrate at the right location.

Remark 3.2 *(Connection to Generalized Bayesian Inference) The exponential kernel has an alternative interpretation as a generalized posterior within the Generalized Bayesian Inference (GBI) framework (Bissiri et al., 2016). Schmon et al. (2020) show that one can view the ABC posterior constructed from the scaled exponential kernel as*

$$p_l(\theta \mid \boldsymbol{X}) \propto \int \exp(-w \cdot l(\boldsymbol{X}, \tilde{\boldsymbol{X}}^\theta))\pi(\theta)\mathrm{d}\tilde{P}^{(m)}(\tilde{\boldsymbol{X}}),$$

where $l(\cdot, \cdot)$ is a "loss function" which can be any distance between the summary statistics or a data discrepancy metric used in ABC, and where the weight $w > 0$ can be chosen to adjust the influence of the loss on the posterior. The exponential kernel can thus be applied more generally to other summary statistics or data discrepancy measures.

3.2 Shape of the Limiting ABC Posterior Distribution

We now analyze the limiting shape of $\hat{\Pi}_{\epsilon_n}^{AR}(\cdot \mid \boldsymbol{X})$ defined in (16). We focus on the case when $\epsilon_n \gg \delta_n^{1/(\kappa+1)}$, where κ was defined in Assumption 3, since the posterior is not guaranteed to converge when the decision threshold ϵ_n is smaller than the estimation error δ_n of the KL estimator.

Assumption 5 *Assume that for every $\varepsilon > 0$, we have $\inf_{\|\theta-\theta_0\|>\varepsilon} K(p_0, p_\theta) > 0$. In addition, assume that $\log p_\theta$ is twice differentiable with respect to θ and that, for every θ in some neighborhood of θ_0, the remainder of the second order Taylor expansion of $K(p_0, p_\theta) = P_0 \log \frac{p_0}{p_\theta}(x)$ around θ_0 is comparatively small relative to the second-order term, i.e.*

$$\begin{aligned} K(p_0, p_\theta) &= \nabla_{\theta=\theta_0} K(p_0, p_\theta)(\theta - \theta_0) + \frac{1}{2}(\theta - \theta_0)' \nabla^2_{\theta=\theta_0} K(p_0, p_\theta)(\theta - \theta_0)\{1 + o(1)\} \\ &= \frac{1}{2}(\theta - \theta_0)' I(\theta_0)(\theta - \theta_0)\{1 + o(1)\}, \end{aligned}$$

where $I(\theta) = \nabla^2_\theta K(p_0, p_\theta) = P_0[(\nabla_\theta \log p_\theta)^2]$ is the Fisher information matrix.

When Assumption 5 is satisfied, the condition in Assumption 4 is immediately satisfied as well and the identification of θ_0 is guaranteed. In the open-neighborhood of θ_0, since we have

$$(\theta - \theta_0)' I(\theta_0)(\theta - \theta_0) = 2K(p_0, p_\theta)\{1 + o(1)\},$$

and $I(\theta_0)$ is positive definite, the convergence in the bilinear form $(\theta-\theta_0)'I(\theta_0)(\theta-\theta_0)$ ensures the convergence of the parameters.

Theorem 3.2 *Assume that the prior function $\pi(\cdot)$ is continuous around θ_0. Then, under Assumptions 1, 2, 3 and 5, if $\lim_n \delta_n/\epsilon_n^{\kappa+1} \to 0$, the average posterior distribution of $\epsilon_n^{-1/2}(\theta-\theta_0)$ converges to the uniform distribution over the ellipse $\{w : w'I(\theta_0)w \leq 2\}$ where $I(\theta)$ is the Fisher information matrix defined in Assumption 5. In particular, as $n \to 0$, we have*

$$P_0^{(n)} \int f\Big(\epsilon_n^{-1/2}(\theta-\theta_0)\Big)\hat{\Pi}_{\epsilon_n}^{AR}(\theta \mid \boldsymbol{X}) \to \int_{u'I(\theta_0)u\leq 2} f(u)\mathrm{d}u \Big/ \int_{u'I(\theta_0)u\leq 2} \mathrm{d}u$$

for every continuous and bounded function $f(\cdot) : \mathcal{X} \to \mathbb{R}$.

The proof is provided in Appendix D.3.

Remark 3.3 *Theorem 3.2 is adapted from the case (i) in Theorem 2 of Frazier et al. (2018). We only consider situations when $\epsilon_n \gg \delta_n^{1/(\kappa+1)}$ with the prior shrinkage parameter κ defined in Assumption 3. In other words, we assume that the ABC decision threshold ϵ_n is dominating both the estimation error δ_n and the asymptotic error $n^{-1/2}$ and, thereby, determines the posterior concentration rate. It is not entirely obvious how the posterior would behave when the threshold ϵ_n shrinks faster than the estimation error δ_n.*

For the asymptotic behavior of the ABC posterior (20) induced by the exponential kernel, we resort to BvM characterizations under misspefication in LAN models (Kleijn and van der Vaart, 2012). When the posterior (20) concentrates around θ^* in (22) at the rate ϵ_n^*, one can show under a suitable LAN condition that the approximate posterior converges to a sequence of normal distributions in total variation at the rate ϵ_n^*. The centering and the asymptotic covariance matrix both depend on θ^*. The formal statement and the proof is in Kaji and Rockova (2021).

4 Simulations

In this section, we illustrate our approach and make comparisons with other likelihood-free inference techniques. Within our KL-ABC framework, we include two types of KL estimators. One is obtained with the logit discriminator score, which we refer to as KL estimation via classification (KLC), and the other one is estimated via the kNN method (kNN) with $k = 1$ (Jiang et al., 2018). For both estimators, we aggregate ABC samples with the accept-reject kernel as in Algorithm

1 and the exponential kernel as in Algorithm 2. The latter will be denoted with a suffix 'exp', e.g. KLC-exp or kNN-exp. The discriminator used for each dataset will be specified later. We provide more discussions on discriminator calibrations in Appendix C. Besides the two examples presented here, we have another analysis on the g-and-k distribution in Appendix B.

The ABC discrepancy metrics we choose for comparisons are (1) the classification accuracy (CA) (Gutmann et al., 2018) defined as (12); (2) the Wasserstein (WA) distance under the Euclidean metric (Bernton et al., 2019) defined as $\text{WA}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^\theta) = \min_\gamma [\sum_{i=1}^n \sum_{j=1}^m \gamma_{ij} \left\| X_i - \tilde{X}_j^\theta \right\|^2]^{1/2}$ s.t. $\gamma' \mathbf{1}_m = \mathbf{1}_n, \gamma' \mathbf{1}_n = \mathbf{1}_m$ with $0 \leq \gamma_{ij} \leq 1$; (3) ℓ_2-distance between summary statistics (SS) and we use the semi-automatic (SA) method (Fearnhead and Prangle, 2012) if no candidate summary statistics are given; (4) auxiliary likelihood $AL = \frac{1}{m} \ln p_A(\tilde{\boldsymbol{X}}^\theta \mid \hat{\phi}(\tilde{\boldsymbol{X}}^\theta)) - \frac{1}{m} \ln p_A(\tilde{\boldsymbol{X}}^\theta \mid \hat{\phi}(\boldsymbol{X}))$ proposed by Drovandi et al. (2011), where $p_A(x \mid \phi)$ is a d-dimensional Gaussian distribution with ϕ being the sample mean and covariance. For the classification accuracy (CA), we use the same discriminator as the one in KL estimation. In each experiment, unless otherwise noted, we set the tolerance threshold ϵ adaptively such that $1\,000$ of $100\,000$ (i.e. the top 1%) proposed ABC samples are accepted.

4.1 M/G/1-queuing Model

Because queuing models are usually easy to simulate from, but have no tractable likelihoods, they have been frequently used as test cases in the ABC literature, see e.g. Fearnhead and Prangle (2012) and Bernton et al. (2019). Here, we choose the same setup as in Jiang et al. (2017). Each datum is a 5-dimensional vector consisting of the first five inter-departure times $x_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5})'$. In the model, the service times u_{ik} follow a uniform distribution $U[\theta_1, \theta_2]$, and the arrival times w_{ik} are exponentially distributed with the rate θ_3. We only observe the interdeparture times x_i, given by the process $x_{ik} = u_{ik} + \max(0, \sum_{j=1}^k w_{ij} - \sum_{j=1}^{k-1} x_{ij})$. We perform ABC on $n = 500$ observed samples which are generated from the true parameter $\theta_0 = (1, 5, 0.2)$. The prior on $(\theta_1, \theta_2 - \theta_1, \theta_3)$ is uniform on $[0, 10]^2 \times [0, 0.5]$. [3]

We train the discriminator as a ℓ_1-penalized logistic regressor onto degree-2 polynomials of the data, including quadratic and interaction terms. The penalty parameter λ is selected via 5-fold cross-validation. We implement the discriminator with the R package `glmnet`.

The shape of the ABC posteriors is given in Figure 1. The plot reveals that, among the three

[3] We place the uniform prior on $\theta_2 - \theta_1$ instead of θ_2, since θ_2 must be larger than θ_1. This is used in Jiang et al. (2017).

Figure 1: ABC reconstructed posteriors under M/G/1-queuing model with $\theta_0 = (1, 5, 0.2)'$. The vertical black lines mark the true values. The blue curves in kNN and KLC boxes mark a smoothed density calculated from the exponential kernel.

parameters, θ_1 is the hardest one to estimate where all methods, except for the semi-automatic ABC (SA), gave relatively flat posterior estimates. Regarding θ_2, our methods, SA and W2 seem to center well around the truth and W2 gives the tightest estimation. Regarding θ_3, it is obvious that SA does not provide enough information. Next, we repeat the experiments on 10 different datasets, and summarize the average squared estimation errors and the width of the 95% credible intervals in Table 1. Overall, we see that although SA captures θ_1 and θ_2 well, it fails to identify the correct location of θ_3. Our KLC method outperforms the kNN estimator in estimating all three parameters. The two variants of KLC and CA give relatively good and robust estimates for all three parameters. These methods tend to be similar in both the point estimate and credible intervals, which is not entirely unexpected since they are derived from the same discriminators.

4.2 Lotka-Volterra Model

The Lotka-Volterra (LV) predator-prey model (Wilkinson, 2018) describes population evolutions in ecosystems where predators interact with prey. It is one of the classical stochastic kinetic network model examples. The state of the population is prescribed deterministically via a system

Method	$\theta_1 = 1$		$\theta_2 = 5$		$\theta_3 = 0.2$	
	$(\hat{\theta}_1 - \theta_1)^2$	95% CI width	$(\hat{\theta}_2 - \theta_2)^2$	95% CI width	$(\hat{\theta}_3 - \theta_3)^2$	95%CI width
KLC	0.197	3.116	0.217	4.599	0.308	0.064
KLC-exp	**0.169**	2.851	0.312	3.708	**0.234**	**0.030**
kNN	0.525	3.135	0.106	3.986	3.659	0.094
kNN-exp	1.057	2.664 (0.8)	0.431	**3.331** (0.9)	2.634	0.072
AL	0.259	3.423	0.057	5.651	0.575	0.044
CA	0.191	3.107	0.235	4.602	0.280	0.064
SA	0.180	**2.457**	0.355	3.514	45.297	0.446
W2	0.595	3.631	**0.039**	4.871	3.846	0.052 (0.8)

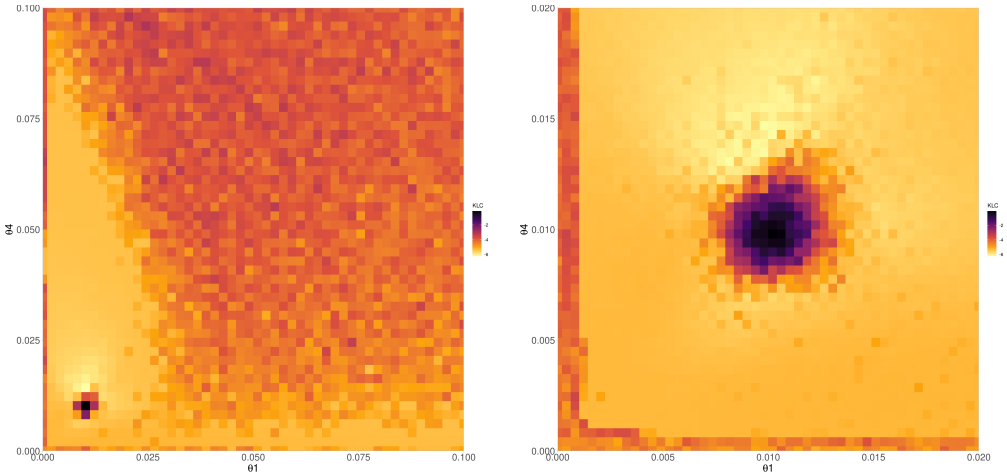
Table 1: ABC performance on the M/G/1 queuing model over 10 repetitions with top 1% ABC samples selected. Most of the 95% CIs have full coverage with the rest having their coverage marked in brackets. The bold fonts mark the best model in each metric.

of ordinary differential equations (ODEs). Inference for such models is challenging because the transition density is intractable. However, simulation from the model is possible, which makes it a natural candidate for ABC methods.

The model monitors population sizes of predators X_t and prey Y_t over time t. The changes in states are determined by four parameters $\theta = (\theta_1, \ldots, \theta_4)'$ controlling: (1) the rate $r_1^t = \theta_1 X_t Y_t$ of a predator being born; (2) the rate $r_2^t = \theta_2 X_t$ of a predator dying; (3) the rate $r_3^t = \theta_3 Y_t$ of a prey being born; (4) the rate $r_4^t = \theta_4 X_t Y_t$ of a prey dying. Given the initial population sizes (X_0, Y_0) at time $t = 0$, the dynamics can be simulated using the Gilliespie algorithm (Gillespie, 1977). The algorithm samples times to an event from an exponential distribution (with a rate $\sum_{j=1}^4 r_j^t$) and picks one of the four reactions with probabilities proportional to their individual rates r_j^t.

We use the same simulation setup as Kaji and Rockova (2021). Each simulation is started at $X_0 = 50$ and $Y_0 = 100$ and state observations are recorded every 0.1 time units for a period of 20 time units, resulting in a series of $T = 201$ observations each. The real data ($n = 20$ time series) are generated with true values $\theta_0 = (0.01, 0.5, 1, 0.01)'$. The predator-prey interaction dynamic is very sensitive to parameter changes. For example, Figure 4 of Kaji and Rockova (2021) shows that a slight perturbation in θ_2 leads to significant changes in the population renewal cycle. We rely on the ability of the discriminator to tell such different patterns apart. The sensitivity of the model to minor parameter changes is confirmed with heat-map plots of the estimated KL divergence as a function of $(\theta_1, \theta_4)'$ in Figure 2. Figure 2(a) provides a plot of the estimated KL over the region $[0, 0.1]^2$ where, apparently, the majority of the region is flat and uninformative

with a sharp spike around the true values at $\theta_1 = \theta_4 = 0.01$. We thus narrow the investigation down to a smaller region $[0, 0.02]^2$ in Figure 2(b). Again, the curvature in the estimated KL around the truth is quite steep. This may pose some issues for Metropolis-Hasting algorithms, since the majority of the prior region is uninformative and improper initialization could lead to extremely slow convergence.



(a) $(\theta_1, \theta_4) \in (0, 0.1)^2$ (b) $(\theta_1, \theta_4) \in (0, 0.02)^2$

Figure 2: Estimated $\hat{K}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^{\theta})$ (via Classification) over a grid of (θ_1, θ_4) values. The other two parameters are fixed as $\theta_2 = 0.5, \theta_3 = 1$.

Previous ABC analyses of this model suggested various summary statistics including the mean, log-variance, autocorrelation (at lag 1 and 2) of each series as well as their cross-correlation (Papamakarios and Murray, 2016). For the discriminator of our method, we choose the ℓ_1-penalized (LASSO) logistic regression classifier with $m = n$ and with a penalty λ selected via 5-fold cross-validation (as implemented in the R package `glmnet`).

Similar to Kaji and Rockova (2021), we use an informative prior $\theta \in U(\Xi)$ with a restricted domain $\Xi = [0, 0.1] \times [0, 1] \times [0, 2] \times [0, 0.1]$ so that the computation is more economic and efficient. A typical snapshot of the ABC posteriors is plotted in Figure 3. From the plot, we can see that the difficulty in estimating different parameters varies a lot and θ_3 is the most challenging among all. While other methods give relatively flat posteriors for θ_3, our method (KLC) combined with the scaled exponential kernel identifies the correct location of the parameter with a much tighter posterior. The average performance of ABC methods under this model is summarized in Table 2. We can see that KLC with the exponential kernel gives the tightest CI most of the time, while maintaining relatively small estimation errors.

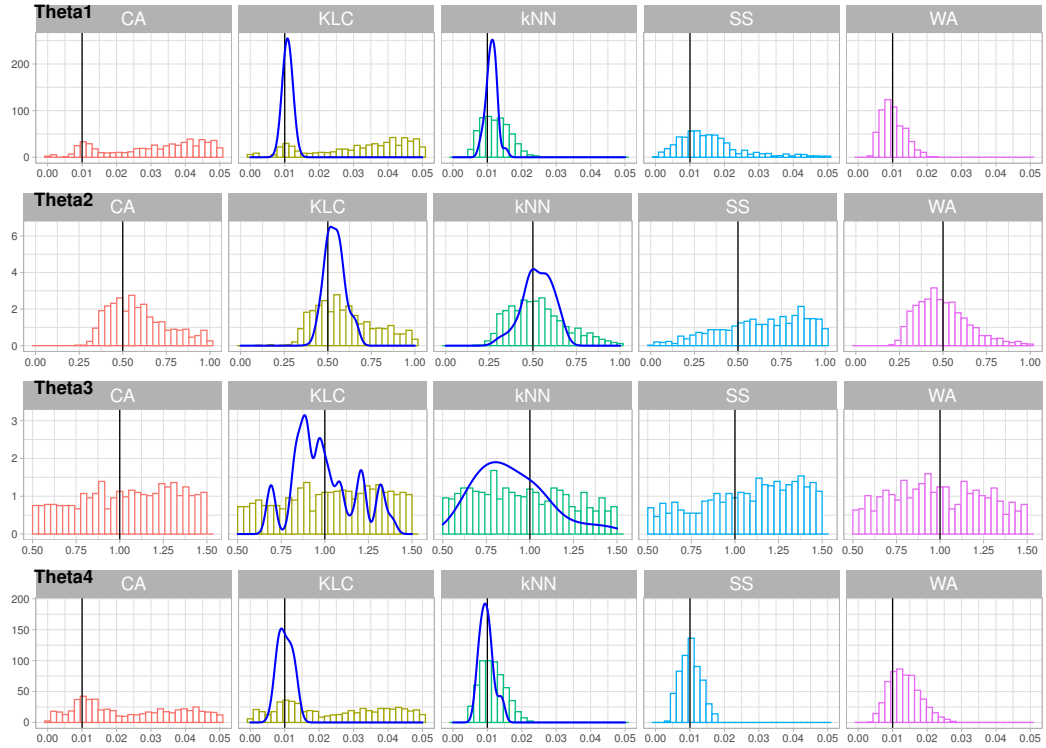


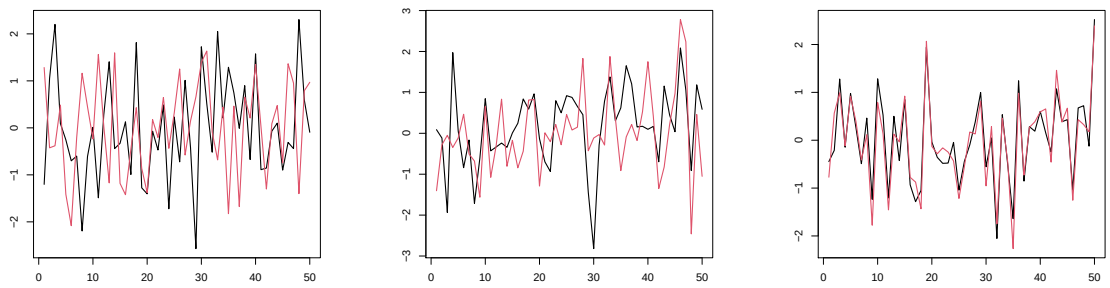
Figure 3: ABC posterior densities for the Lotka-Volterra model with $\theta_0 = (0.01, 0.5, 1, 0.01)'$. The black vertical lines mark the true parameter values. The blue curves in kNN and KLC boxes represent the smoothed density calculated from the exponential kernel. The ABC posteriors are plotted from the top 1% out of 10^5 samples.

Method (scale)	$\theta_1 = 0.01$ $(\hat{\theta}_1 - \theta_1)^2$ (10^{-5})	95% CI width	$\theta_2 = 0.5$ $(\hat{\theta}_2 - \theta_2)^2$	95% CI width	$\theta_3 = 1$ $(\hat{\theta}_3 - \theta_3)^2$ (10^{-2})	95% CI width	$\theta_4 = 0.01$ $(\hat{\theta}_4 - \theta_4)^2$ (10^{-5})	95% CI width
KLC	42.949	0.0447	0.0169	0.626	0.119	0.945	17.159	0.0472
KLC-exp	0.066	**0.0035**	0.0025	**0.285**	0.405	**0.643**	**0.021**	**0.0063**
kNN	0.563	0.0146	0.0011	0.602	0.054	0.934	0.203	0.0133
kNN-exp	**0.049**	0.0054 (0.9)	**0.0004**	0.243 (0.9)	0.346	**0.599 (0.9)**	0.066	**0.0062 (0.9)**
CA	36.477	0.0438	0.0160	0.629	0.118	0.945	15.684	0.0462
Stats	1.639	0.0178	0.0226	0.721	1.019	0.901	3.767	0.0166
W2	1.265	0.0167	0.0048	0.599	**0.011**	0.917	4.080	0.0205

Table 2: ABC performance evaluations on the Lotka-Volterra Model, averaged over 10 repetitions, with the top 1% ABC samples selected. Most of the 95% CIs have full coverage with the rest having their coverage marked in brackets. The bold fonts mark the best model in each metric.

5 Empirical Analysis

We further demonstrate our approach on the nontrivial problem of estimating stock volatility using merely daily observations on high, low and closing prices. All of these price observations are typically available to investors. We use a similar data generating process as in Magdon-Ismail and Atiya (2003), assuming that the assets follow a Brownian motion with a constant drift and volatility. In particular, suppose that the log-price processes $X_j(t), i = 1, \ldots, d$, are correlated



(a) $\sigma_{12} = 0.05$ (b) $\sigma_{12} = 0.5$ (c) $\sigma_{12} = 0.95$

Figure 4: Closing prices time series for three choices of σ_{12} ($\mu_1 = \mu_2 = 0, \sigma_{11} = \sigma_{22} = 1$).

Brownian motions, that is $E[X_i(s)X_j(t)] = \sigma_{ij} \min\{s, t\}$, and that the joint movement of the log-price processes $\boldsymbol{X}(t) = (X_1(t), \ldots, X_d(t))'$ follows a multivariate Brownian motion as

$$d\boldsymbol{X}(t) = \boldsymbol{\mu} dt + \Sigma d\boldsymbol{W}(t), \tag{23}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)'$ and $\Sigma = [\sigma_{ij}]_{1 \leq i,j \leq d}$ denote the drift and the volatility of the log processes, respectively. We write

$$H_j = \max_{0 \leq t \leq 1} X_j(t), \qquad L_j = \min_{0 \leq t \leq 1} X_j(t), \qquad S_j = X_j(1),$$

for the high, low and final log price, respectively, over a fixed time interval $[0, 1]$. We want to estimate the drift $\boldsymbol{\mu}$ and the volatility matrix Σ merely from observing these three prices over a period of time.

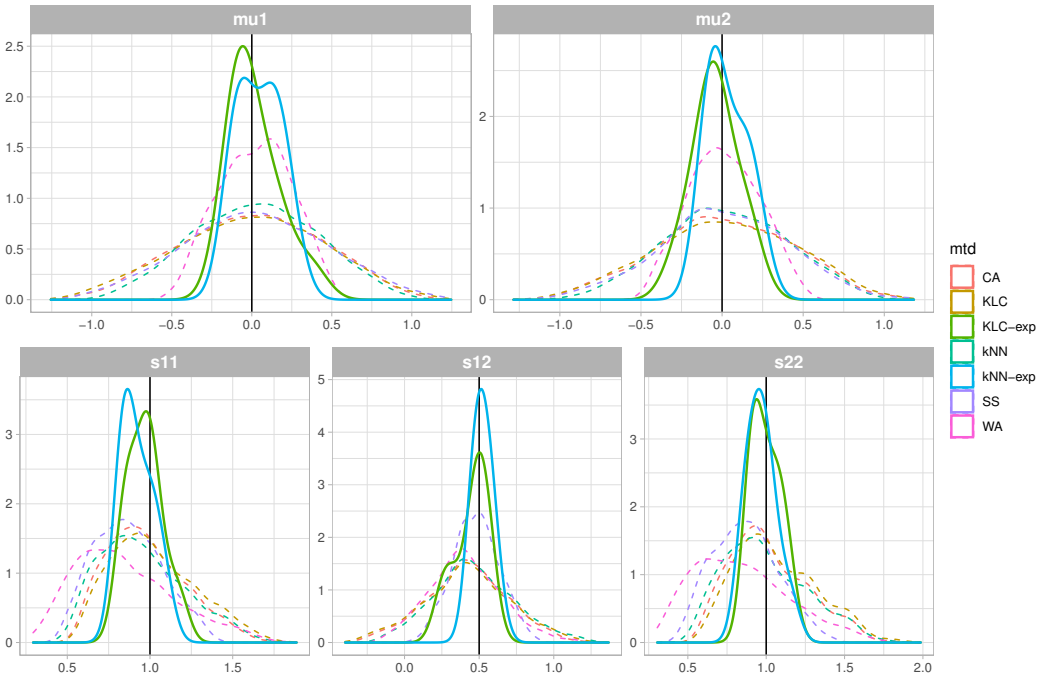
We impose a normal-inverse-Wishart prior $(\boldsymbol{\mu}, \Sigma) \sim NIW(\boldsymbol{\mu}_0, \lambda, \Phi, \nu)$. This distribution can be sampled from in two steps: (1) sample Σ from an inverse Wishart distribution $\Sigma \mid \Phi, \nu \sim W^{-1}(\Phi, \nu)$; (2) sample $\boldsymbol{\mu}$ from a multivariate normal distribution $\boldsymbol{\mu} \mid \boldsymbol{\mu}_0, \lambda, \Sigma \sim N(\boldsymbol{\mu}_0, \frac{1}{\lambda}\Sigma)$. Since Σ is a semi-positive definite matrix, we model the parameters through its Cholesky root $\Sigma^{1/2}$. Without loss of generality, we only consider the case $W(0) = (0, \ldots, 0)'$, since the closing prices on the previous day or the opening prices of today are usually known.

5.1 Synthetic Data

To compare various likelihood-free estimators, we first generate synthetic data for 1 000 trading days. For each day (of length $t = 1$), we simulate the Brownian motion using 500 time steps

to obtain the high, low and closing price data for each particular window. We first restrict our attention to the case of just two assets, which leaves us with 5 parameters to estimate: μ_1, μ_2 and the upper triangular root of Σ, denoted with $L = \begin{bmatrix} l_{11} & 0 \\ l_{12} & l_{22} \end{bmatrix}$. We first illustrate how the covariance parameter σ_{12} impacts the co-movement of asset prices. Holding $\mu_1 = \mu_2 = 0$ and $\sigma_{11} = \sigma_{22} = 1$ fixed, we plot time series realizations of the closing prices for three particular choices of σ_{12} in Figure 4. The patterns are as expected where the prices tend to co-fluctuate when σ_{12} is closer to one. The success of our method depends on how well the discriminator can tell apart these trajectories.

For our simulation, we set $\mu = (0, 0)'$ and $\sigma_{11} = \sigma_{22} = 1, \sigma_{12} = 0.5$. We choose relatively non-informative prior hyper-parameters $\boldsymbol{\mu}_0 = (0, 0)', \lambda = 1, \Phi = I_d$ and $\nu = d$. Posterior distributions reconstructed with different ABC methods are given in Figure 5 and the averaged performance over 10 repetitions is summarized in Table 3. It is clear that our exponential kernel methodsplace more mass around the true location of the parameters, and that the shape of kNN-exp is slightly less regular than KLC-exp. Although KLC-exp induces a larger bias in estimating the drift (μ_1, μ_2), it does a better job at capturing the correct location of the volatilities.



(a) ABC

Figure 5: Posterior densities on simulated volatility data ($d = 2$)

		KLC	KLC-exp	kNN	kNN-exp	CA	SS	WA
$\mu_1 = 0$	MSE($\times 10^{-4}$)	2.004	2.113	0.638	**0.159**	1.497	1.216	0.170
	95% CI width	1.760	**0.331**	1.480	0.442	1.660	1.734	0.831
$\mu_2 = 0$	MSE($\times 10^{-4}$)	0.734	7.439	0.363	0.259	0.142	**0.021**	0.277
	95% CI width	1.628	**0.431**	1.435	0.538	1.564	1.569	0.813
$\sigma_{11} = 1$	MSE($\times 10^{-3}$)	0.311	0.111	0.167	0.460	**0.002**	1.404	2.634
	95% CI width	0.963	0.371	0.954	**0.329**	0.896	0.763	1.140
$\sigma_{12} = 0.5$	MSE($\times 10^{-3}$)	1.035	**0.103**	0.427	0.107	0.861	0.109	1.528
	95% CI width	1.193	0.355	1.092	**0.324**	1.122	0.628	0.988
$\sigma_{22} = 1$	MSE($\times 10^{-3}$)	0.416	0.304	0.124	0.018	**0.007**	3.687	4.244
	95% CI width	0.959	**0.252**	0.974	0.314	0.920	0.783	1.157

Table 3: Performance on the stock volatility estimation example, averaged over 10 repetitions, with top 1% selected. All 95% CIs have full coverage of the true parameters. The bold fonts mark the best model in each row.

5.2 Real Data Analysis

Following the example in Rogers and Zhou (2008), we examine a small data set of stock prices focusing on two stocks: Boeing (BA) and Proctor & Gamble (PG). The prices were obtained from NYSE (Yahoo Finance), starting from 3rd January 2011 and consisting of $1\,000$ trading days. Since the off-market trades follow a different mechanism than the market trade, we only model price changes from the opening price to the closing price each day where the log price differences $X(t)$ are all computed based on the opening prices of that day.

We observe that the fluctuations in the prices are much smaller than in our simulated time series and we thereby choose the hyperparameters based on the mean and covariance of the closing prices. Figure 6 gives the ABC posterior distributions estimated from different methods and the corresponding summaries of the distributions are provided in Table 4. We again observe that our methods with the exponential kernels return much narrower posterior distributions. The estimates on the volatilities are quite close except for the ABC with summary statistics (SS). The drifts of both BA and PG are not significantly different from zero.

6 Discussion

This paper develops an ABC variant using a classification-based KL estimator as a discrepancy measure. By deploying a flexible classifier, the empirical KL divergence can be estimated with a vanishing error. In addition, inspired by the connection between the KL divergence and the log-likelihood ratio, we propose a scaled exponential kernel to aggregate ABC samples. This

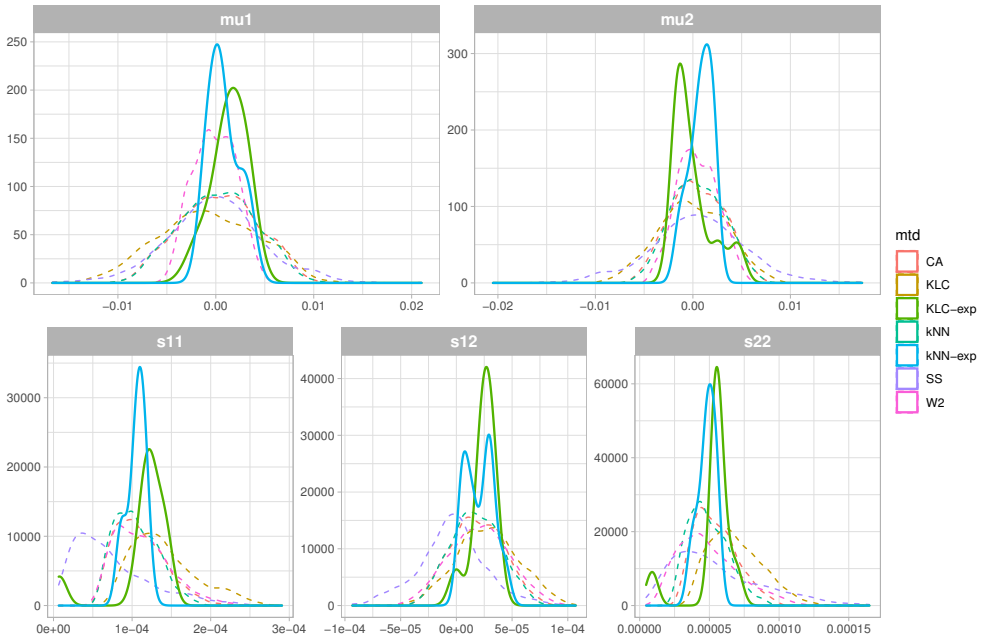


Figure 6: Posterior densities estimated from log-prices of BA and PG

		KLC	KLC-exp	kNN	kNN-exp	CA	SS	W2
μ_1 $(\times 10^{-3})$	$\bar{\mu}_1$	-0.774	1.354	0.250	0.779	0.342	-0.071	-0.126
	l	-10.380	-1.981	-6.816	-1.145	-7.099	-9.532	-3.987
	u	8.202	3.514	7.229	3.385	7.395	10.121	3.736
μ_2 $(\times 10^{-3})$	$\bar{\mu}_2$	0.009	-0.039	0.491	0.822	0.445	0.450	0.453
	l	-6.591	-1.718	-4.329	-1.406	-4.604	-9.831	-3.491
	u	6.531	4.563	5.503	2.166	5.538	10.171	4.450
$\sigma_{11}(\times 10^{-4})$	$\bar{\sigma}_{11}$	1.406	1.142	1.049	1.054	1.101	0.738	1.130
	l	0.824	0.076	0.633	0.844	0.634	0.155	0.626
	u	2.362	1.444	1.640	1.202	1.819	1.976	1.938
$\sigma_{12}(\times 10^{-5})$	$\bar{\sigma}_{12}$	2.811	2.446	1.767	2.084	1.906	-0.376	2.052
	l	-2.379	-0.37	-2.359	0.494	-2.390	-6.660	-3.137
	u	8.226	3.629	6.037	4.298	6.442	6.065	7.253
$\sigma_{22}(\times 10^{-4})$	$\bar{\sigma}_{22}$	0.692	0.526	0.492	0.482	0.547	0.552	0.495
	l	0.407	0.089	0.277	0.365	0.315	0.133	0.173
	u	1.105	0.698	0.802	0.559	0.883	1.285	0.984

Table 4: Posterior estimates on analysis of BA and PG. For each parameter, we report three summary statistics, the posterior means, the lower limit of the 95% CI intervals (l) and the upper limit of the 95% CI intervals (u).

smoothing variant avoids the need for choosing the ad hoc threshold ϵ_n and fully utilizes information returned from all ABC samples. Under mild conditions, we show that the posterior concentration rate of the accept-reject ABC depends on the estimation error δ_n and the accept-reject threshold ϵ_n, while the rate of the smooth version depends on the estimation error δ_n and the contraction rate of the actual posterior distribution. Our methodology can also be related to many other likelihood-free inference methods, including ABC with Classification Accuracy

(Gutmann et al., 2018), Wasserstein distance ABC (Bernton et al., 2019), and Generalized Posteriors (Schmon et al., 2020). Along with our theoretical investigations, we demonstrate competitive performance of our methods on benchmark examples. Our theoretical analysis provides theoretical justifications for the method of Gutmann et al. (2018).

References

Åkesson, M., Singh, P., Wrede, F., and Hellander, A. (2020). Convolutional neural networks as summary statistics for approximate Bayesian computation. *arXiv*.

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR.

Bauer, B. and Kohler, M. (2019). On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *The Annals of Statistics*, 47(4):2261–2285.

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035.

Bernton, E., Jacob, P. E., Gerber, M., and Robert, C. P. (2019). Approximate Bayesian computation with the Wasserstein distance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(2):235–269.

Bishop, C. M. (2006). *Pattern recognition and Machine Learning*. Springer-Verlag, Berlin, Heidelberg.

Bissiri, P. G., Holmes, C. C., and Walker, S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 78(5):1103.

Blum, M. G., Nunes, M. A., Prangle, D., and Sisson, S. A. (2013). A comparative review of dimension reduction methods in approximate Bayesian computation. *Statistical Science*, 28(2):189–208.

Burkard, R., Dell'Amico, M., and Martello, S. (2012). *Assignment problems: revised reprint*. SIAM.

Cover, T. M. and Thomas, J. A. (1991). Entropy, relative entropy and mutual information. *Elements of information theory*, 2(1):12–13.

Den Hollander, F. (2008). *Large deviations*, volume 14. American Mathematical Soc.

Drovandi, C. C., Pettitt, A. N., and Faddy, M. J. (2011). Approximate Bayesian computation using indirect inference. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 60(3):317–337.

Fearnhead, P. and Prangle, D. (2012). Constructing summary statistics for approximate bayesian computation: Semi-automatic approximate bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):419–474.

Frazier, D. T., Martin, G. M., Robert, C. P., and Rousseau, J. (2018). Asymptotic properties of approximate bayesian computation. *Biometrika*, 105(3):593–607.

Ghimire, S., Gyawali, P. K., and Wang, L. (2020). Reliable estimation of kullback-leibler divergence by controlling discriminator complexity in the reproducing kernel hilbert space. *arXiv*.

Ghosal, S. and Van der Vaart, A. (2017). *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry*, 81(25):2340–2361.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680.

Gutmann, M. U., Dutta, R., Kaski, S., and Corander, J. (2018). Likelihood-free inference via classification. *Statistics and Computing*, 28(2):411–425.

Jiang, B., Wu, T.-Y., and Wong, W. H. (2018). Approximate Bayesian computation with Kullback-Leibler divergence as data discrepancy. In *International Conference on Artificial Intelligence and Statistics*, pages 1711–1721. PMLR.

Jiang, B., Wu, T.-y., Zheng, C., and Wong, W. H. (2017). Learning summary statistic for approximate Bayesian computation via deep neural network. *Statistica Sinica*, pages 1595–1618.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233.

Joyce, P. and Marjoram, P. (2008). Approximately sufficient statistics and Bayesian computation. *Statistical applications in genetics and molecular biology*, 7(1).

Kaji, T., Manresa, E., and Pouliot, G. (2020). An adversarial approach to structural estimation. *arXiv*.

Kaji, T. and Rockova, V. (2021). Metropolis-hastings via classification. *arXiv*.

Kleijn, B. J. and van der Vaart, A. W. (2006). Misspecification in infinite-dimensional bayesian statistics. *The Annals of Statistics*, 34(2):837–877.

Kleijn, B. J. and van der Vaart, A. W. (2012). The bernstein-von-mises theorem under misspecification. *Electronic Journal of Statistics*, 6:354–381.

Kullback, S. (1958). *Information theory and statistics*. Wiley, New York.

Magdon-Ismail, M. and Atiya, A. F. (2003). A maximum likelihood approach to volatility estimation for a Brownian motion using high, low and close price data. *Quantitative Finance*, 3(5):376.

Müller, A. (1997). Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, pages 429–443.

Nguyen, X., Wainwright, M. J., and Jordan, M. I. (2007). Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, pages 1089–1096.

Nunes, M. A. and Balding, D. J. (2010). On optimal selection of summary statistics for approximate Bayesian computation. *Statistical applications in genetics and molecular biology*, 9(1).

Papamakarios, G. and Murray, I. (2016). Fast ε-free inference of simulation models with bayesian conditional density estimation. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 1036–1044.

Park, M., Jitkrittum, W., and Sejdinovic, D. (2016). K2-ABC: Approximate Bayesian computation with kernel embeddings. In *Artificial Intelligence and Statistics*, pages 398–407. PMLR.

Pérez-Cruz, F. (2008). Kullback-Leibler divergence estimation of continuous distributions. In *2008 IEEE international symposium on information theory*, pages 1666–1670. IEEE.

Polson, N. G. and Rockova, V. (2018). Posterior concentration for sparse deep learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 938–949.

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

Rogers, L. C. and Zhou, F. (2008). Estimating correlation from high, low, opening and closing prices. *The Annals of Applied Probability*, 18(2):813–823.

Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics*, 48(4):1875–1897.

Schmon, S. M., Cannon, P. W., and Knoblauch, J. (2020). Generalized posteriors in approximate Bayesian computation. In *Third Symposium on Advances in Approximate Bayesian Inference*.

Schwartz, L. (1965). On Bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 4(1):10–26.

Silva, J. and Narayanan, S. (2007). Universal consistency of data-driven partitions for divergence estimation. In *2007 IEEE International Symposium on Information Theory*, pages 2021–2025. IEEE.

Silva, J. and Narayanan, S. S. (2010). Information divergence estimation based on data-dependent partitions. *Journal of Statistical Planning and Inference*, 140(11):3180–3198.

Sisson, S. A., Fan, Y., and Beaumont, M. A. (2018). Overview of abc. In *Handbook of approximate Bayesian computation*, pages 3–54. Chapman and Hall/CRC.

Sørensen, H. (2004). Parametric inference for diffusion processes observed at discrete points in time: a survey. *International Statistical Review*, 72(3):337–354.

van de Geer, S. A. (2000). *Empirical Processes in M-estimation*, volume 6. Cambridge university press.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes: with applications to statistics*. Springer.

Villani, C. (2009). *Optimal transport: old and new*, volume 338. Springer.

Wainwright, M. J. and Jordan, M. I. (2008). *Graphical models, exponential families, and variational inference*. Now Publishers Inc.

Wang, Q., Kulkarni, S. R., and Verdú, S. (2005). Divergence estimation of continuous distributions based on data-dependent partitions. *IEEE Transactions on Information Theory*, 51(9):3064–3074.

Wang, Q., Kulkarni, S. R., and Verdú, S. (2009). Divergence estimation for multidimensional densities via k-nearest-neighbor distances. *IEEE Transactions on Information Theory*, 55(5):2392–2405.

Wang, Y. and Rockova, V. (2020). Uncertainty quantification for sparse deep learning. In *International Conference on Artificial Intelligence and Statistics*, pages 298–308. PMLR.

Wilkinson, D. J. (2018). *Stochastic modelling for systems biology*. CRC press.

Wilkinson, R. D. (2013). Approximate bayesian computation (abc) gives exact results under the assumption of model error. *Statistical applications in genetics and molecular biology*, 12(2):129–141.

Wood, S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466(7310):1102–1104.

Yarotsky, D. (2017). Error bounds for approximations with deep relu networks. *Neural Networks*, 94:103–114.

Zhao, P. and Lai, L. (2020). Analysis of k nearest neighbor KL divergence estimation for continuous distributions. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 2562–2567. IEEE.

A Frequentist's analysis on the exponential kernel

We thus study the concentration in terms of a KL neighborhood around $Q_{\theta^*}^{(n)}$ defined as

$$B(\epsilon, Q_{\theta^*}^{(n)}; P_0^{(n)}) = \{Q_{\theta^*}^{(n)} \in \mathcal{Q}^{(n)} : \tilde{K}(\theta^*, \theta) \leq n\epsilon^2, \tilde{V}(\theta^*, \theta) \leq n\epsilon^2\}, \tag{24}$$

where $\tilde{K}(\theta^*, \theta) \equiv P_0^{(n)} \log \frac{q_{\theta^*}^{(n)}}{q_\theta^{(n)}}$ and $\tilde{V}(\theta^*, \theta) \equiv P_0^{(n)} \left| \log \frac{q_{\theta^*}^{(n)}}{q_\theta^{(n)}} - \tilde{K}(\theta^*, \theta) \right|^2$.

The following corollary is directly adopted from Theorem 5.1 of Kaji and Rockova (2021).

Corollary A.1 *Denote with $\widetilde{Q}_\theta^{(n)}$ a measure defined through $\mathrm{d}\widetilde{Q}_\theta^{(n)} = (p_0^{(n)}/q_{\theta^*}^{(n)})\mathrm{d}P_\theta^{(n)}$ and let $d(\cdot,\cdot)$ be a semi-metric on $\mathcal{P}^{(n)}$. Suppose that there exists a sequence $\epsilon_n > 0$ satisfying $\epsilon_n \to 0$ and $n\epsilon_n^2 \to \infty$ such that for every $\epsilon > \epsilon_n$ there exists a test ϕ_n (depending on ϵ) such that for every $J \in \mathbb{N}_0$*

$$P_0^{(n)}\phi_n \lesssim e^{-n\epsilon^2/4} \quad \textit{and} \quad \sup_{Q_\theta^{(n)}: d(Q_\theta^{(n)}, P_{\theta^*}^{(n)}) > J\epsilon} \widetilde{Q}_\theta^{(n)}(1-\phi_n) \leq e^{-nJ^2\epsilon^2/4}. \tag{25}$$

Let $B(\epsilon, Q_{\theta^*}^{(n)}; P_0^{(n)})$ be as in (24) *and let $\tilde{\Pi}_n(\theta)$ be a prior distribution with a density $\tilde{\pi}(\theta) \propto C_\theta \pi(\theta)$ with C_θ as in* (21)*. Assume that there exists a constant $L > 0$ such that, for all n and $j \in \mathbb{N}$,*

$$\frac{\tilde{\Pi}_n\left(\theta \in \Theta : j\epsilon_n < d(Q_\theta^{(n)}, P_{\theta^*}^{(n)}) \leq (j+1)\epsilon_n\right)}{\tilde{\Pi}_n\left(B(\epsilon_n, Q_{\theta^*}^{(n)}; P_0^{(n)})\right)} \leq e^{n\epsilon_n^2 j^2/8}. \tag{26}$$

Then for every sufficiently large constant M, as $n \to \infty$,

$$P_0^{(n)}\Pi^{EK}\left(Q_\theta^{(n)} : d(Q_\theta^{(n)}, P_{\theta^*}^{(n)}) \geq M\epsilon_n \mid X^{(n)}\right) \to 0. \tag{27}$$

Next, we want to show the shape of the posterior is actually asymptotically gaussian around θ^*. The following corollary follows from Theorem 2.1 of Kleijn and van der Vaart (2012) and Lemma 8.1 of Kaji and Rockova (2021).

Corollary A.2 *(Bernstein von-Mises) Assume that the posterior* (21) *concentrates around θ^* at the rate ϵ_n^* and that for every compact $K \in \mathbb{R}^d$*

$$\sup_{h \in K} \left| \log \frac{q_{\theta^* + \epsilon_n^* h}^{(n)}(\boldsymbol{X})}{q_{\theta^*}^{(n)}(\boldsymbol{X})} - h'\tilde{V}_{\theta^*}\tilde{\delta}_{n,\theta^*} - \frac{1}{2}h'\tilde{V}_{\theta^*}h \right| \to 0 \tag{28}$$

for some random vector $\tilde{\delta}_{n,\theta^*}$ and a non-singular matrix $\tilde{V}_{\theta^*}$. Then the approximated posterior

$\Pi^{EK}(\cdot)$ *converges to a sequence of normal distributions in total variation at the rate* ϵ_n^**, i.e.*

$$\sup_B \left| \Pi^{EK}\left({\epsilon_n^*}^{-1}(\theta - \theta^*) \in B \mid \boldsymbol{X} \right) - N_{\tilde{\delta}_{n,\theta^*}, \tilde{V}_{\theta^*}^{-1}(B)} \right| \to 0 \quad \textit{in } P_0^{(n)} - \textit{probability}. \tag{29}$$

B Simulation: Univariate g-and-k Distributions

Another classical example in the ABC literature is the univariate g-and-k distribution. It is defined implicitly by its inverse distribution function

$$F^{-1}(x) = A + B\Big[1 + 0.8\frac{1 - \mathrm{e}^{-gz_x}}{1 + \mathrm{e}^{-gz_x}}\Big](1 + z_x^2)^k z_x,$$

where z_x is the x-th quantile of the standard normal distribution, and parameters A, B, g, k are related to location, scale, skewness and kurtosis. The probability density function has no analytical form but can be numerically calculated with high precision since it only involves one-dimensional inversions and differentiations of the quantile function. Therefore, Bayesian inference can be carried out.

We generate $n = 500$ observations from the model by letting $A = 3, B = 1, g = 2, k = 0.5$. Among the four parameters, g is the hardest to identify, as observed in previous analyses. From a pilot run, narrow the range of our uniform priors to $A \sim U([2, 4]), B \sim U([0, 0.25]), g \sim U([0, 10])$ and $k \sim U([0, 2.5])$.

We estimate the KL divergence by a neural network with two hidden layers with 10 nodes each. The first layer is activated with the rectified linear unit (ReLU) function, while the second layer is equipped with a hyperbolic tangent function (tanh). In addition to observed data $\boldsymbol{X}$, we also include higher-order terms $\boldsymbol{X}^2$ and $\boldsymbol{X}^3$ in the input.[4] We compare the posteriors obtained from different ABC methods in Figure 7. We can see that, indeed, the parameter g is the most difficult one to estimate. While the majority of methods return a flat posterior, KLC and kNN with the exponential kernel still place most of the posterior mass around the true value. AL and SA misidentify the location of one or two parameters. While CA and KLC give relatively similar ranking on the ABC sampled parameters, KLC with the exponential kernel can utilize all the samples and re-adjust the shape according to weights. In addition, KLC-exp returns more spiky posteriors than kNN-exp. This is also reflected in the performance summaries in Table 5 which

[4]We are trying to manually create more features for this univariate data. We also explore even higher order terms and find the including up to $\boldsymbol{X}^3$ works the best in practice.

show that KLC-exp has a narrower CI than kNN-exp on average in terms of estimating A, B and k. We observe that kNN could be advantageous than KLC in estimating g-and-k distributions, since the data is univariate and the use of classifiers may not be necessary.

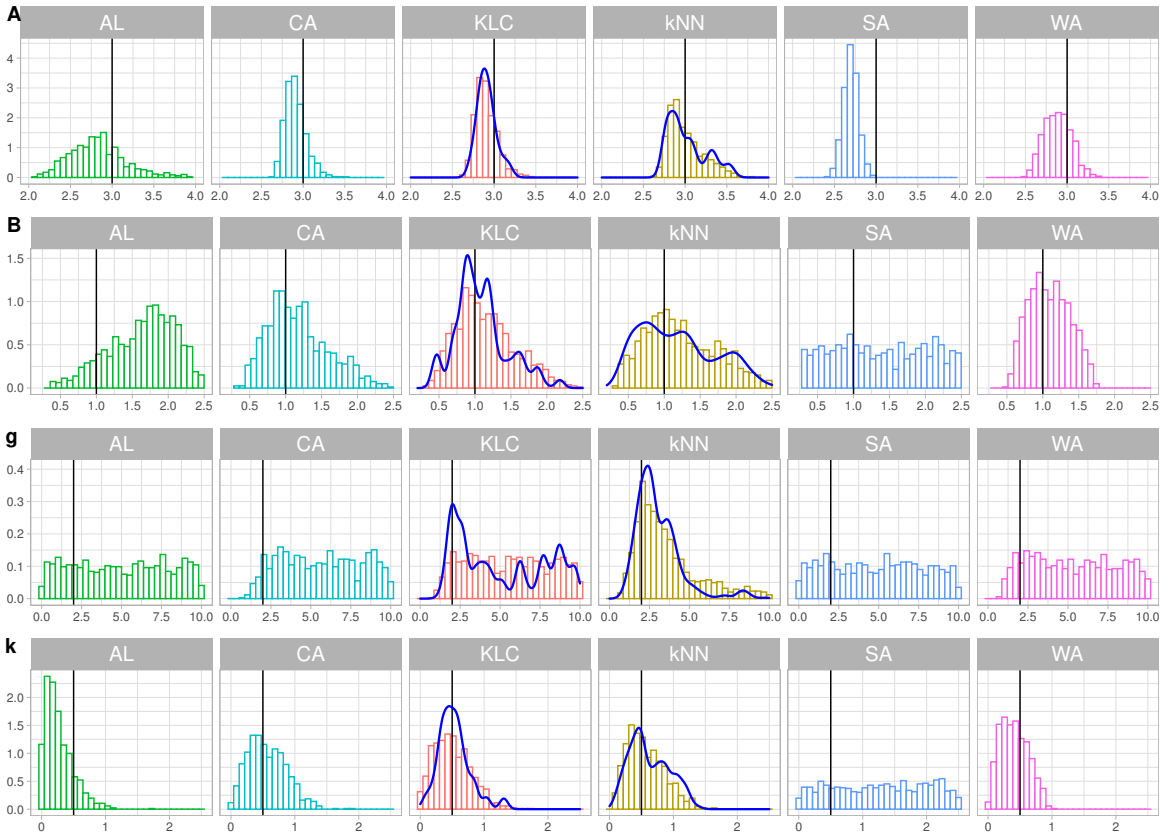


Figure 7: Posterior densities for the g-and-k distributions $(A = 3, B = 1, g = 2, k = 0.5)$. The black vertical lines mark the true parameter values. The blue curves in kNN and KLC boxes represent the weighted density calculated from the exponential kernel. The ABC posteriors were plotted with the top 1% of 10^5 samples.

	$A = 3$		$B = 1$		$g = 2$		$k = 0.5$	
Method	$(\hat{A} - A)^2$	95% CI width	$(\hat{B} - B)^2$	95% CI width	$(\hat{g} - g)^2$	95% CI width	$(\hat{k} - k)^2$	95 %CI width
KLC	0.0143	0.535	0.0369	1.528	12.840	8.356	**0.00341**	1.044
KLC-exp	0.0156	**0.455**	0.0339	1.229	12.516	8.191	0.00800	0.784
kNN	**0.0024**	0.843	0.0486	1.770	2.695	7.558	0.00421	1.086
kNN-exp	0.0032	0.610	0.0410	1.333	**1.673**	**4.109**	0.00480	0.805
AL	0.0176	1.325	0.2211	1.593	8.704	9.480	0.04451	0.835
CA	0.0125	0.503	0.0466	1.568	13.077	8.162	0.00574	1.080
SA	0.1191	0.424 (0)	0.0598	2.379	8.800	9.478	0.56399	2.379
WA	0.0145	0.627	**0.0240**	**1.034**	11.621	8.566	0.01060	**0.776**

Table 5: Performance summaries on g-and-k distributions over 10 repetitions with top 1% ABC samples selected. Most of the 95% CIs have full coverage with the rest having their coverage marked after the CI width. The bold fonts mark the best model in each metric.

C Discriminator calibration

We provide more intuition for how the choice of the discriminator, including the m/n ratio and the number of sets of latent variables needed to approximate the KL divergence, affects our results. We also discuss possible modifications we make to the KL estimation to eliminate bias. We use the M/G/1-queuing model as an example.

First, we want to investigate how the choice of the discriminator and the m/n ratio could impact the estimation of the KL divergence. We consider three discriminators: (1) logistic regression on degree-2 polynomial terms of $\boldsymbol{X}$ (LRD); (2) a neural network with one layer of 10 ReLu nodes and two layers of 10 tanh nodes with input $\boldsymbol{X}$ (NND1); (3) a neural network with one layer of 10 ReLu nodes and one layer of 10 tanh nodes with input as the degree-2 polynomial terms of $\boldsymbol{X}$ (NND2). For each discriminator, we do a grid search over $m/n \in \{1, 2, 3, 5\}$. We plot the estimated KL divergence (rescaled) in Figure 8. From the plot we can see it is quite obvious that LRD best identifies the location of the true values, and the fluctuations/variance in its estimations are the smallest. It also worth noting that the difficulties in estimating different parameters are significantly different here. All discriminators are able to learn the location of θ_3 with little variance, but only LRD is able to learn θ_1 and θ_2 well. The m/n ratio does not seem to have significant impact for LRD, but we observe for NND1, NND2 and some other models that the variance in the estimation tends to get smaller when $m/n > 1$ but not overly large to avoid bias. We continue our experiments with LRD and $m/n = 3$.

Next, we want to know how many different sets of latent variables are needed to control the variance in estimation. We focus on the best discriminator we have previously found (i.e. LRD with $m/n = 3$). We include the cases where $nlatent \in \{1, 5, 10\}$ here. The results are shown in Figure 9. It is expected that with more sets of latent variables to average over, the variance in estimation gets smaller. The main challenge for the M/G/1 model lies in estimating θ_1 correctly and stably. With only one or five sets of latent variable, it seems hopeless to identify θ_1. To trade off variance and computation costs, we set nlatent $= 10$ in our experiments.

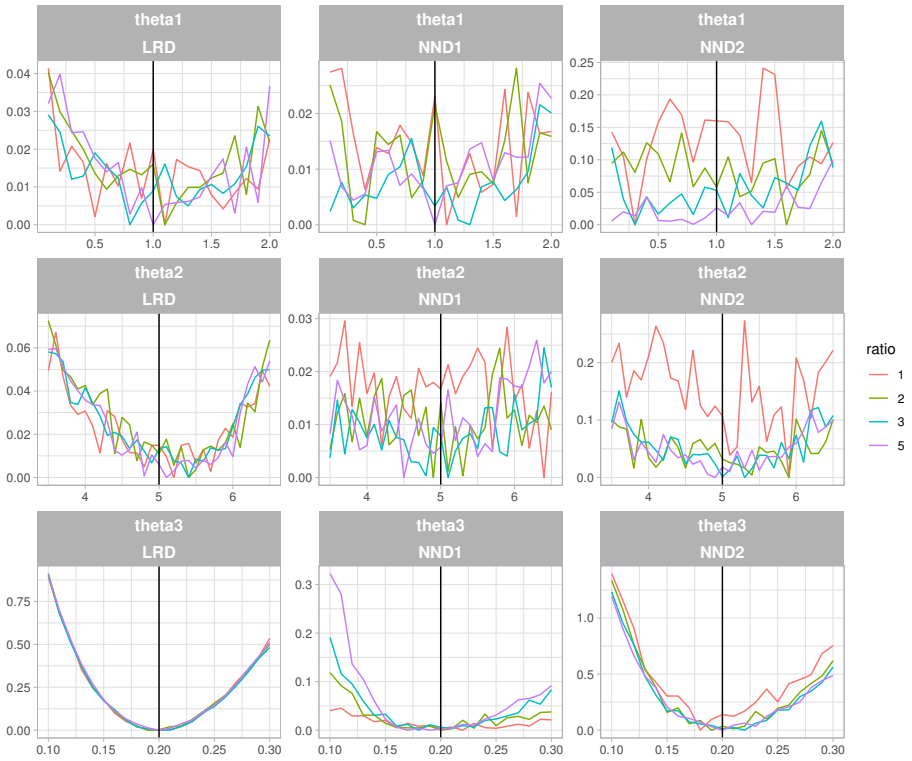


Figure 8: Estimated KL divergence (rescaled) under different discriminators (nlatent=10). We plot the marginal changes in estimated KL divergence with respect to different parameters and different m/n ratios here. The estimated values are rescaled by subtracting the minimum in each scenario so the changes are more comparable.

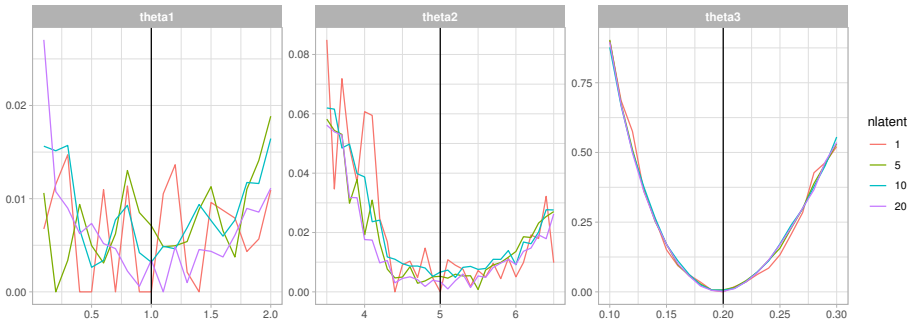


Figure 9: Estimated KL divergence (rescaled) using LRD and $m/n = 3$. We plot the marginal changes in estimated KL divergence with respect to different parameters and different number of sets of latent variables. The estimated values are rescaled by subtracting the minimum in each scenario so the changes are more comparable.

D Proofs

D.1 Proof of Theorem 2.1

Denote $K_n = \mathbb{P}_n \log \frac{p_0}{p_\theta}$. Using Chebyshev's inequality and Assumption 2 as

$$P_0^{(n)}[|K_n - K(p_0, p_\theta)| > u] = P_0^{(n)}\Big(\Big|(\mathbb{P}_n - P_0)\log \frac{p_0}{p_\theta}\Big| > u \Big) \le \frac{1}{u^2} P_0^{(n)}\Big[\Big|(\mathbb{P}_n - P_0)\log \frac{p_0}{p_\theta}\Big|^2 \Big]$$
$$= \frac{1}{u^2} P_0^{(n)}\Big[\Big|\mathbb{P}_n\big(\log \frac{p_0}{p_\theta} - P_0 \log \frac{p_0}{p_\theta}\big)\Big|^2 \Big] = \frac{1}{nu^2} P_0\Big[\Big|\log \frac{p_0}{p_\theta} - P_0 \log \frac{p_0}{p_\theta}\Big|^2 \Big]$$

where the last inequality follows from Lemma 2.1 (iii) of Kaji and Rockova (2021). Next, note

$$\hat{K}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^\theta) - K_n = -\mathbb{P}_n\Big(\log \frac{1-\hat{D}^\theta_{n,m}}{1-D_\theta} - \log \frac{\hat{D}^\theta_{n,m}}{D_\theta} \Big).$$

Recall the outer expectation P^* in Lemma 2.1 and the set of classifiers $D^\theta_{n,\delta}$ defined in Assumption 1, we can bound

$$\begin{aligned}
&P\bigg(\Big| \mathbb{P}_n\Big(\log \frac{1-\hat{D}^\theta_{n,m}}{1-D_\theta} - \log \frac{\hat{D}^\theta_{n,m}}{D_\theta} \Big) \Big| > u, d_\theta(\hat{D}^\theta_{n,m}, D_\theta) \le C_n\delta_n \bigg) \\
&\qquad \le P^*\Big(\sup_{D \in \mathcal{D}^\theta_{C_n\delta_n}} \Big| \mathbb{P}_n\Big(\log \frac{1-D}{1-D_\theta} - \log \frac{D}{D_\theta} \Big) \Big| > u \Big) \\
&\qquad \le \frac{1}{u}\mathbb{E}^* \sup_{D \in \mathcal{D}^\theta_{C_n\delta_n}} \Big| \mathbb{P}_n\Big(\log \frac{1-D}{1-D_\theta} - \log \frac{D}{D_\theta} \Big) \Big|
\end{aligned}$$

by Markov's inequality. The proof of Theorem 4.1 of Kaji and Rockova (2021) shows that the expectation is $O(C_n\delta_n)$. Using the triangle inequality and the Bonferroni inequality, since $h(p_0, p_\theta) \le 2$, we can then write

$$\begin{aligned}
&P\bigg(\Big| \hat{K}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^\theta) - K(p_0, p_\theta) \Big| > 2u, d_\theta(\hat{D}^\theta_{n,m}, D^\theta) \le C_n\delta_n \bigg) \\
&\quad \le P\bigg(\Big| \hat{K}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^\theta) - K_n \Big| + |K_n - K(p_0, p_\theta)| > 2u, d_\theta(\hat{D}^\theta_{n,m}, D_\theta) \le C_n\delta_n \bigg) \\
&\quad \le P\bigg(\Big| \hat{K}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^\theta) - K_n \Big| > u, d_\theta(\hat{D}^\theta_{n,m}, D_\theta) \le C_n\delta_n \bigg) + P_0^{(n)}[|K_n - K(p_0, p_\theta)| > u] \\
&\quad \le O\Big(\frac{C_n\delta_u}{u} \Big) + \frac{32(2+\Lambda)}{nu^2}.
\end{aligned}$$

D.2 Proof of Theorem 3.1

Throughout, we continue to assume that $\tilde{\boldsymbol{X}}^\theta = g_\theta(\tilde{\boldsymbol{X}})$ and we denote with $P = P_0^{(n)} \otimes \tilde{P}^{(m)}$ the joint measure for $(\boldsymbol{X}, \tilde{\boldsymbol{X}})$. Below, we will be using the notation $\Pi(\cdot)$ to denote the generic probability, i.e. for $(\theta, \tilde{\boldsymbol{X}})$ or for the conditional probability θ given $\tilde{\boldsymbol{X}}$. Later, we will define a high-probability event $\Omega_n(C, \varepsilon_n)$ such that $P_0^{(n)}[\Omega_n(C, \varepsilon_n)^c] = o(1)$ for some $C \in (0, 1)$. Given $\delta_n > 0$ from our assumptions, we can write for every $\lambda_n > 0$ and $\epsilon_n > 0$ and for every arbitrarily

slowly increasing sequence $C_n > 0$

$$P_0^{(n)}\Pi\left(K(p_0,p_\theta) > \lambda_n \middle| \hat{K}(\boldsymbol{X},\tilde{\boldsymbol{X}}^\theta) \le \epsilon_n\right) \le \Pi_1 + o(1) + \\ P_0^{(n)}\Pi\left(K(p_0,p_\theta) > \lambda_n \middle| \hat{K}(\boldsymbol{X},\tilde{\boldsymbol{X}}^\theta) \le \epsilon_n, d(\hat{D}^\theta_{n,m}, D_\theta) \le C_n\delta_n\right) \mathbb{I}[\Omega_n(C,\varepsilon_n)], \quad (30)$$

where, using Lemma 2.1 and the fact that the rate δ_n is uniform (see Remark 2.2),

$$\Pi_1 \equiv P_0^{(n)}\Pi(d(\hat{D}^\theta_{n,m}, D_\theta) > C_n\delta_n) \le \sup_{\theta\in\Theta} P(d(\hat{D}^\theta_{n,m}, D_\theta) > C_n\delta_n) = o(1).$$

Consider the joint event, for some $\delta' > 0$,

$$A_{\epsilon_n}(\delta') = \{(\tilde{\boldsymbol{X}},\theta) : \hat{K}[\boldsymbol{X}, g_\theta(\tilde{\boldsymbol{X}})] \le \epsilon_n\} \cap \{K(p_0,p_\theta) > \delta'\}.$$

For every $(\tilde{\boldsymbol{X}},\theta) \in A_{\epsilon_n}(\delta')$ we have

$$K(p_0,p_\theta) \le \hat{K}(\boldsymbol{X}, g_\theta(\tilde{\boldsymbol{X}})) + \left|\hat{K}(\boldsymbol{X}, g_\theta(\tilde{\boldsymbol{X}})) - K(p_0,p_\theta)\right| \le \epsilon_n + \left|\hat{K}(\boldsymbol{X}, g_\theta(\tilde{\boldsymbol{X}})) - K(p_0,p_\theta)\right|.$$

Hence $(\tilde{\boldsymbol{X}},\theta) \in A_{\epsilon_n}(\delta')$ implies that

$$\left|\hat{K}(\boldsymbol{X},\tilde{\boldsymbol{X}}^\theta) - K(p_0,p_\theta)\right| > \delta' - \epsilon_n,$$

and choosing $\delta' \ge \epsilon_n + t_\epsilon$ leads to

$$\Pi[A_{\epsilon_n}(\delta')] \le \int_\Theta \tilde{P}^{(m)}\left[\left|\hat{K}(\boldsymbol{X},\tilde{\boldsymbol{X}}^\theta) - K(p_0,p_\theta)\right| > t_\epsilon\right] d\Pi(\theta).$$

Using (30), we now focus on the conditional probability, given $d(\hat{D}^\theta_{n,m}, D_\theta) \le C_n\delta_n$,

$$\Pi\left(K(p_0,p_\theta) > \epsilon_n + t_\epsilon \middle| \hat{K}(\boldsymbol{X},\tilde{\boldsymbol{X}}^\theta) \le \epsilon_n, d(\hat{D}^\theta_{n,m}, D_\theta) \le C_n\delta_n\right) \\ \le \frac{\int_\Theta \tilde{P}^{(m)}\left[\left|\hat{K}(\boldsymbol{X},\tilde{\boldsymbol{X}}^\theta) - K(p_0,p_\theta)\right| > t_\epsilon \mid d(\hat{D}^\theta_{n,m}, D_\theta) \le C_n\delta_n\right] d\Pi(\theta)}{\int_\Theta \tilde{P}^{(m)}\left[\hat{K}(\boldsymbol{X},\tilde{\boldsymbol{X}}^\theta) \le \epsilon_n \mid d(\hat{D}^\theta_{n,m}, D_\theta) \le C_n\delta_n\right] d\Pi(\theta)}. \quad (31)$$

We now find a lower bound for the denominator. Recall the KL neighborhood $B_2(p_0,\epsilon_n)$ defined in (14). Since

$$\hat{K}(\boldsymbol{X},\tilde{\boldsymbol{X}}^\theta) \le K(p_0,p_\theta) + \left|\hat{K}(\boldsymbol{X},\tilde{\boldsymbol{X}}^\theta) - K(p_0,p_\theta)\right| \le \epsilon_n/2 + K(p_0,p_\theta),$$

provided that $\left|\hat{K}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^{\theta}) - K(p_0, p_\theta)\right| \leq \epsilon_n/2$. The denominator can be then bounded by

$$\begin{aligned}
&\int_{\Theta} \tilde{P}^{(m)}[\hat{K}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^{\theta}) \leq \epsilon_n \mid d(\hat{D}^{\theta}_{n,m}, D_\theta) \leq C_n\delta_n] d\Pi(\theta) \\
&\quad \geq \int_{B_2(p_0,\epsilon_n/2)} \tilde{P}^{(m)}\left[\left|\hat{K}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^{\theta}) - K(p_0, p_\theta)\right| \leq \epsilon_n/2 \mid d(\hat{D}^{\theta}_{n,m}, D_\theta) \leq C_n\delta_n\right] d\Pi(\theta) \\
&\quad \geq \Pi[B_2(p_0, \epsilon_n/2)] - \int_{B_2(p_0,\epsilon_n/2)} \tilde{P}^{(m)}\left[\left|\hat{K}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^{\theta}) - K(p_0, p_\theta)\right| > \epsilon_n/2 \mid d(\hat{D}^{\theta}_{n,m}, D_\theta) \leq C_n\delta_n\right] d\Pi(\theta).
\end{aligned}$$

Denoting

$$Z(\boldsymbol{X}) \equiv \int_{B_2(p_0,\epsilon_n/2)} \tilde{P}^{(m)}\left[\left|\hat{K}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^{\theta}) - K(p_0, p_\theta)\right| > \epsilon_n/2 \mid d(\hat{D}^{\theta}_{n,m}, D_\theta) \leq C_n\delta_n\right] d\Pi(\theta)$$

we can write, for every $C > 0$, using Fubini's theorem and Markov's inequality

$$\begin{aligned}
P_0^{(n)}\left(Z(\boldsymbol{X}) > C\right) &\leq \frac{1}{C}\int_{B_2(p_0,\epsilon_n/2)} P\left[\left|\hat{K}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^{\theta}) - K(p_0, p_\theta)\right| > \epsilon_n/2 \mid d(\hat{D}^{\theta}_{n,m}, D_\theta) \leq C_n\delta_n\right] d\Pi(\theta) \\
&= \frac{\Pi(B_2(p_0, \epsilon_n/2))}{C} \frac{\sup_{\theta\in\Theta} P\left[\left|\hat{K}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^{\theta}) - K(p_0, p_\theta)\right| > \epsilon_n/2, d(\hat{D}^{\theta}_{n,m}, D_\theta) \leq C_n\delta_n\right]}{\sup_{\theta\in\Theta} P\left[d(\hat{D}^{\theta}_{n,m}, D_\theta) \leq C_n\delta_n\right]} \\
&= \frac{\Pi(B_2(p_0, \epsilon_n/2))}{C(1+o(1))} \sup_{\theta\in\Theta} \rho_{n,\theta}(\varepsilon_n/2; C_n; \delta_n),
\end{aligned}$$

where we have used Theorem 2.1 and the fact that $P(d(\hat{D}^{\theta}_{n,m}, D_\theta) \leq C_n\delta_n) = 1 + o(1)$ for every $\theta \in \Theta$ from Lemma 2.1. We now define an event, for some $0 < C < 1$ and $\epsilon_n > 0$,

$$\Omega_n(C, \epsilon_n) = \left\{\boldsymbol{X} : Z(\boldsymbol{X}) \leq C \times \Pi(B_2(p_0, \epsilon_n/2))\right\}.$$

Using Theorem 2.1, we have

$$\rho_{n,\theta}(\epsilon_n/2; C_n; \delta_n) = O\Big(\frac{C_n\delta_n}{\epsilon_n} + \frac{1}{n\epsilon_n^2}\Big) \quad \text{for every} \quad \theta \in \Theta.$$

Choosing $\epsilon_n > 0$ such that $\epsilon_n = o(1)$ and $n\epsilon_n^2 \to \infty$ and $C_n\delta_n = o(\epsilon_n)$ we have $P_0^{(n)}[\Omega_n(C, \epsilon_n)^c] = o(1)$ for every $C \in (0, 1)$. On the event $\Omega_n(C, \epsilon_n)$, for some $0 < C < 1$, we can lower-bound the denominator with

$$\int_{\Theta} \tilde{P}^{(m)}[\hat{K}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^{\theta}) \leq \epsilon_n \mid d(\hat{D}^{\theta}_{n,m}, D_\theta) \leq C_n\delta_n] d\Pi(\theta) > (1 - C) \times \Pi(B_2(p_0, \epsilon_n/2).$$

Using this bound and applying Fubini's theorem, we can further write

$$P_0^{(n)}\Pi\left(K(p_0,p_\theta) > \epsilon_n + t_\epsilon \mid \hat{K}(\boldsymbol{X},\tilde{\boldsymbol{X}}^\theta) \le \epsilon_n, d(\hat{D}^\theta_{n,m}, D_\theta) \le C_n\delta_n\right)\mathbb{I}[\Omega_n(C,\varepsilon_n)]$$
$$\le \frac{\int_\Theta P\left[\left|\hat{K}(\boldsymbol{X},\tilde{\boldsymbol{X}}^\theta) - K(p_0,p_\theta)\right| > t_\epsilon \mid d(\hat{D}^\theta_{n,m}, D_\theta) \le C_n\delta_n\right] d\Pi(\theta)}{(1-C)\times\Pi(B_2(p_0,\epsilon_n/2)} \quad (32)$$

Using Theorem 2.1 again, we obtain an upper bound for the display above with

$$\frac{\rho_n(t_\epsilon; C_n; \delta_n)}{(1-C)\times\Pi(B_2(p_0,\epsilon_n/2)(1+o(1))}.$$

Using the prior Assumption 3, we can choose t_ϵ such that

$$\Big(\frac{C_n\delta_n}{t_\epsilon} + \frac{1}{nt_\epsilon^2}\Big)/\epsilon_n^\kappa = 1/M_n$$

for some arbitrarily slowly increasing sequence $M_n > 0$. We choose $t_\epsilon = M_nC_n\delta_n/\epsilon_n^\kappa + \sqrt{M}_n n^{-1/2}/\epsilon_n^{\kappa/2}$. Since $\delta_n \gtrsim n^{-1/2}$ and $\epsilon_n^{-\kappa} \ge \epsilon_n^{-\kappa/2}$, the overall rate is then driven by $\epsilon_n + t_\epsilon = \epsilon_n + \widetilde{M}_n\delta_n\epsilon_n^{-\kappa}$, where $\widetilde{M}_n = M_nC_n$.

D.3 Proof of Theorem 3.2

Recall from Theorem 3.1 that with the accept-reject strategy, the true KL divergence $K(p_0,p_\theta)$ is contracting at the rate $\lambda_n = \epsilon_n + \widetilde{M}_n\delta_n\epsilon_n^{-\kappa}$, where $\widetilde{M}_n$ is a slowly increasing sequence that diverges faster than C_n. Consider the case when $\epsilon_n \gg \widetilde{M}_n\delta_n\epsilon_n^{-\kappa}$ or, equivalently, $\epsilon_n \gg \delta_n^{1/(\kappa+1)}$. Denote

$$x(\theta) = \epsilon_n^{-1}K(p_0,p_\theta) \quad \text{and} \quad f_n(\theta-\theta_0) = f\big(\epsilon_n^{-1/2}(\theta-\theta_0)\big).$$

We express the ABC posterior expectation of $f_n(\theta-\theta_0)$ for a non-negative and bounded

function $f_n(\cdot)$ by

$$
\begin{aligned}
P_0^{(n)} E_{\hat{\Pi}_{\epsilon_n}^{AR}}\Big[f_n(\theta-\theta_0)\Big] &= P_0^{(n)} \int f_n(\theta-\theta_0)\mathrm{d}\hat{\Pi}_{\epsilon_n}^{AR}(\theta \mid \boldsymbol{X}) \\
&= \underbrace{P_0^{(n)} \int f_n(\theta-\theta_0)\mathbb{I}[K(p_0,p_\theta)\le\lambda_n, d(\hat{D}_{n,m}^{\theta}, D_\theta)\le C_n\delta_n)]\mathrm{d}\hat{\Pi}_{\epsilon_n}^{AR}(\theta \mid \boldsymbol{X})}_{\text{(I)}} \\
&\quad + \underbrace{P_0^{(n)} \int f_n(\theta-\theta_0)\mathbb{I}[K(p_0,p_\theta)\le\lambda_n, d(\hat{D}_{n,m}^{\theta}, D_\theta)> C_n\delta_n)]\mathrm{d}\hat{\Pi}_{\epsilon_n}^{AR}(\theta \mid \boldsymbol{X})}_{\text{(II)}} \\
&\quad + \underbrace{P_0^{(n)} \int f_n(\theta-\theta_0)\mathbb{I}[K(p_0,p_\theta)> \lambda_n]\mathrm{d}\hat{\Pi}_{\epsilon_n}^{AR}(\theta \mid \boldsymbol{X})}_{\text{(III)}}
\end{aligned}
$$

where the term (III) can be controlled using Fubini's theorem and the concentration result in Theorem 3.1 as follows

$$
\begin{aligned}
\text{(III)} &= \int f_n(\theta-\theta_0) P_0^{(n)}\mathbb{I}[K(p_0,p_\theta)>\lambda_n]\mathrm{d}\hat{\Pi}_{\epsilon_n}^{AR}(\theta \mid \boldsymbol{X}) \\
&\le \|f\|_\infty P_0^{(n)}\Pi\Big(K(p_0,p_\theta)>\lambda_n \mid \hat{K}(\boldsymbol{X},\tilde{\boldsymbol{X}}^\theta)\le\epsilon_n\Big) = o(1).
\end{aligned}
$$

The second term (II) can be bounded similarly as

$$
\begin{aligned}
\text{(II)} &= P_0^{(n)} \int_{K(p_0,p_\theta)\le\lambda_n} f_n(\theta-\theta_0)\mathbb{I}[d(\hat{D}_{n,m}^{\theta}, D_\theta)> C_n\delta_n)]\mathrm{d}\hat{\Pi}_{\epsilon_n}^{AR}(\theta \mid \boldsymbol{X}) \\
&\le \|f\|_\infty \int_{K(p_0,p_\theta)\le\lambda_n} P(d(\hat{D}_{n,m}^{\theta}, D_\theta)> C_n\delta_n)\mathrm{d}\Pi(\theta) \\
&\le \|f\|_\infty \sup_{\theta\in\Theta} P(d(\hat{D}_{n,m}^{\theta}, D_\theta)> C_n\delta_n) = o(1)
\end{aligned}
$$

where we use the fact that $\sup_{\theta\in\Theta} P(d(\hat{D}_{n,m}^{\theta}, D_\theta)> C_n\delta_n) = o(1)$ from Lemma 2.1.

Thus, the asymptotic behavior is mainly determined by the term (I). Combined with the continuity of $\pi(\theta)$ at θ_0 we can re-write (I) as

$$
\begin{aligned}
\text{(I)} &= P_0^{(n)} \frac{\int_{K(p_0,p_\theta)\le\lambda_n} \pi(\theta) f_n(\theta-\theta_0)\tilde{P}^{(m)}\big[\hat{K}(\boldsymbol{X},\tilde{\boldsymbol{X}}^\theta)\le\epsilon_n, d(\hat{D}_{n,m}^{\theta}, D_\theta)\le C_n\delta_n\big]\mathrm{d}\theta}{\int_{K(p_0,p_\theta)\le\lambda_n} \pi(\theta)\tilde{P}^{(m)}\big[\hat{K}(\boldsymbol{X},\tilde{\boldsymbol{X}}^\theta)\le\epsilon_n, d(\hat{D}_{n,m}^{\theta}, D_\theta)\le C_n\delta_n\big]\mathrm{d}\theta} \\
&= P_0^{(n)} \frac{\int_{K(p_0,p_\theta)\le\lambda_n} f_n(\theta-\theta_0)\tilde{P}^{(m)}\big[\hat{K}(\boldsymbol{X},\tilde{\boldsymbol{X}}^\theta)\le\epsilon_n, d(\hat{D}_{n,m}^{\theta}, D_\theta)\le C_n\delta_n\big]\mathrm{d}\theta}{\int_{K(p_0,p_\theta)\le\lambda_n} \tilde{P}^{(m)}\big[\hat{K}(\boldsymbol{X},\tilde{\boldsymbol{X}}^\theta)\le\epsilon_n, d(\hat{D}_{n,m}^{\theta}, D_\theta)\le C_n\delta_n\big]\mathrm{d}\theta}(1+o(1))
\end{aligned}
$$

We have $x(\theta) \geq 0$ for all $\theta \in \Theta$ and since

$$\hat{K}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^{\theta}) = \hat{K}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^{\theta}) - K(p_0, p_\theta) + \epsilon_n x(\theta),$$

we can write

$$\begin{aligned}
&\tilde{P}^{(m)}\big[\hat{K}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^{\theta}) \leq \epsilon_n, d(\hat{D}^{\theta}_{n,m}, D_\theta) \leq C_n\delta_n\big] \\
&\qquad \geq \tilde{P}^{(m)}\left[\left|\hat{K}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^{\theta}) - K(p_0, p_\theta)\right| \leq \epsilon_n(1 - x(\theta)), d(\hat{D}^{\theta}_{n,m}, D_\theta) \leq C_n\delta_n\right] \\
&\qquad = 1 - \tilde{P}^{(m)}\left[\left|\hat{K}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^{\theta}) - K(p_0, p_\theta)\right| > \epsilon_n(1 - x(\theta)), d(\hat{D}^{\theta}_{n,m}, D_\theta) \leq C_n\delta_n\right]. \quad (33)
\end{aligned}$$

Denoting with

$$\tilde{Z}(\boldsymbol{X}) = \int_{x(\theta) \leq 1 - \frac{M_n C_n \delta_n}{\epsilon_n}} \tilde{P}^{(m)}\left[\left|\hat{K}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^{\theta}) - K(p_0, p_\theta)\right| > \epsilon_n\big(1 - x(\theta)\big), d(\hat{D}^{\theta}_{n,m}, D_\theta) \leq C_n\delta_n\right] \mathrm{d}\theta.$$

Using Markov's inequality and Fubini's theorem we have, for every $\tilde{C} > 0$,

$$\begin{aligned}
&P_0^{(n)}(\tilde{Z}(\boldsymbol{X}) > \tilde{C}) \\
&\quad \leq \frac{1}{\tilde{C}} \int\limits_{x(\theta) \leq 1 - \frac{M_n C_n \delta_n}{\epsilon_n}} P\left[\left|\hat{K}(\boldsymbol{X}, \tilde{\boldsymbol{X}}^{\theta}) - K(p_0, p_\theta)\right| > \epsilon_n\big(1 - x(\theta)\big), d(\hat{D}^{\theta}_{n,m}, D_\theta) \leq C_n\delta_n\right] \mathrm{d}\theta \\
&\quad \leq \frac{1}{\tilde{C}} \int_{x(\theta) \leq 1 - \frac{M_n C_n \delta_n}{\epsilon_n}} \rho_n\big(\epsilon_n(1 - x(\theta)); C_n; \delta_n\big) \mathrm{d}\theta \\
&\quad \leq \frac{1}{\tilde{C}} \times \int_{x(\theta) \leq 1 - \frac{M_n C_n \delta_n}{\epsilon_n}} \mathrm{d}\theta \times \sup_{x(\theta) \leq 1 - \frac{M_n C_n \delta_n}{\epsilon_n}} \rho_n\big(\epsilon_n(1 - x(\theta)); C_n; \delta_n\big) \\
&\quad \leq \frac{\int_{x(\theta) \leq 1 - \frac{M_n C_n \delta_n}{\epsilon_n}} \mathrm{d}\theta}{\tilde{C}} \times \rho_n(M_n C_n \delta_n; C_n; \delta_n) \\
&\quad = \frac{\int_{x(\theta) \leq 1 - \frac{M_n C_n \delta_n}{\epsilon_n}} \mathrm{d}\theta}{\tilde{C}} \times O\left(\frac{1}{M_n} + \frac{1}{n M_n^2 C_n^2 \delta_n^2}\right).
\end{aligned}$$

where $\rho_n(\cdot; C_n; \delta_n)$ is defined in Theorem 2.1. Thus, we can define a set $\tilde{\Omega}_n(\tilde{C}_n)$, for some for some arbitrarily slowly increasing sequence $\tilde{C}_n > 0$, and $\tilde{C}_n = O(1/M_n)$, as

$$\tilde{\Omega}_n(\tilde{C}_n) = \left\{\boldsymbol{X} : \tilde{Z}(\boldsymbol{X}) \leq \tilde{C}_n \times \int_{x(\theta) \leq 1 - \frac{M_n C_n \delta_n}{\epsilon_n}} \mathrm{d}\theta\right\}.$$

Then we have that $P_0^{(n)}(\tilde{\Omega}_n(\tilde{C}_n)^c) = o(1)$. Recall the inequality in (33), on $\tilde{\Omega}_n(\tilde{C}_n)$, we have that

$$\int_{x(\theta)\le 1-\frac{M_nC_n\delta_n}{\epsilon_n}} \tilde{P}^{(m)}\big[\hat{K}(\boldsymbol{X},\tilde{\boldsymbol{X}}^\theta)\le\epsilon_n, d(\hat{D}^\theta_{n,m},D_\theta)\le C_n\delta_n\big]\mathrm{d}\theta \\ \ge \int_{x(\theta)\le 1-\frac{M_nC_n\delta_n}{\epsilon_n}} \mathrm{d}\theta - \tilde{Z}(\boldsymbol{X}) = \int_{x(\theta)\le 1-\frac{M_nC_n\delta_n}{\epsilon_n}} \mathrm{d}\theta \times (1-o(1))$$

And this quantity is upper-bounded by $\int_{x(\theta)\le 1-\frac{M_nC_n\delta_n}{\epsilon_n}} \mathrm{d}\theta$. Therefore, we can conclude that

$$\int_{x(\theta)\le 1-\frac{M_nC_n\delta_n}{\epsilon_n}} \tilde{P}^{(m)}\big[\hat{K}(\boldsymbol{X},\tilde{\boldsymbol{X}}^\theta)\le\epsilon_n, d(\hat{D}^\theta_{n,m},D_\theta)\le C_n\delta_n\big]\mathrm{d}\theta = \int_{x(\theta)\le 1-\frac{M_nC_n\delta_n}{\epsilon_n}} \mathrm{d}\theta\big(1+o(1)\big).$$

Note that $x(\theta)\le 1-M_nC_n\delta_n/\epsilon_n$ implies that $K(p_0,p_\theta)\le\epsilon_n-M_nC_n\delta_n$. On this set $\tilde{\Omega}_n(\tilde{C}_n)$, we can further lower bound the denominator as

$$\begin{aligned}
&\int_{K(p_0,p_\theta)\le\lambda_n} \tilde{P}^{(m)}\big[\hat{K}(\boldsymbol{X},\tilde{\boldsymbol{X}}^\theta)\le\epsilon_n, d(\hat{D}^\theta_{n,m},D_\theta)\le C_n\delta_n\big]\mathrm{d}\theta \\
&\quad= \underbrace{\int_{K(p_0,p_\theta)\le\epsilon_n-M_nC_n\delta_n} \mathrm{d}\theta(1+o(1))}_{D_1} \\
&\qquad+ \underbrace{\int_{\epsilon_n-M_nC_n\delta_n<K(p_0,p_\theta)\le\lambda_n} \tilde{P}^{(m)}\big[\hat{K}(\boldsymbol{X},\tilde{\boldsymbol{X}}^\theta)\le\epsilon_n, d(\hat{D}^\theta_{n,m},D_\theta)\le C_n\delta_n\big]\mathrm{d}\theta}_{D_2}.
\end{aligned}$$

Next, we show that the second term D_2 is $o(D_1)$. Let $u=\epsilon_n^{-1/2}(\theta-\theta_0)$. Under Assumption 5, we have $K(p_0,p_\theta)=\frac{1}{2}(\theta-\theta_0)'I(\theta_0)(\theta-\theta_0)\{1+o(1)\}$, which is $x(\theta_0+\sqrt{\epsilon}_n u)=\frac{1}{2}u'I(\theta_0)u$. Since $I(\theta_0)$ is positive definite we can write

$$\frac{D_2}{D_1} \le \frac{\int_{1-M_nC_n\delta_n/\epsilon_n<x(\theta)\le 1+\widetilde{M}_n\delta_n/\epsilon_n^{\kappa+1}}\mathrm{d}\theta}{\int_{x(\theta)\le 1-M_nC_n\delta_n/\epsilon_n}\mathrm{d}\theta} \tag{34}$$

$$\le \frac{\int_{2(1-M_nC_n\delta_n/\epsilon_n)\le u'I(\theta_0)u\le 2(1+\widetilde{M}_n\delta_n/\epsilon_n^{\kappa+1})}\mathrm{d}\theta}{\int_{u'I(\theta_0)u\le 2(1-M_nC_n\delta_n/\epsilon_n)}\mathrm{d}\theta} \lesssim \frac{M_nC_n\delta_n}{\epsilon_n}+\frac{\widetilde{M}_n\delta_n}{\epsilon_n^{\kappa+1}}=o(1). \tag{35}$$

where the approximation follows from the fact that $\epsilon_n \gg \delta_n^{1/(\kappa+1)}$ and because the denominator is approximating the integral $\int_{u'I(\theta_0)u\le 2}\mathrm{d}\theta$ and the numerator is the length of shrinking intervals. Combining the above results, we find that, on the set $\tilde{\Omega}_n(\tilde{C}_n)$, the denominator can be lower-bounded by $\int_{x(\theta)\le 1-\frac{M_nC_n\delta_n}{\epsilon_n}}\mathrm{d}\theta\{1+o(1)\}$.

This implies

$$(\mathbf{I}) = \frac{\int_{K(p_0,p_\theta)\leq\lambda_n} f_n(\theta-\theta_0)P\big[\hat{K}(\boldsymbol{X},\tilde{\boldsymbol{X}}^\theta)\leq\epsilon_n, d(\hat{D}^\theta_{n,m},D_\theta)\leq C_n\delta_n\big]\mathrm{d}\theta}{(1+o(1))\int_{x(\theta)\leq 1-\frac{M_nC_n\delta_n}{\epsilon_n}}\mathrm{d}\theta}(1+o(1))+o(1)$$

$$= (N_1+N_2)\{1+o(1)\}+o(1). \tag{36}$$

with

$$N_1 \equiv \frac{\int\limits_{x(\theta)\leq 1-\frac{M_nC_n\delta_n}{\epsilon_n}} f\Big(\epsilon_n^{-1/2}(\theta-\theta_0)\Big)\mathrm{d}\theta}{\int\limits_{x(\theta)\leq 1-\frac{M_nC_n\delta_n}{\epsilon_n}}\mathrm{d}\theta} \tag{37}$$

and

$$N_2 \equiv \frac{\int\limits_{K(p_0,p_\theta)\leq\lambda_n} \mathbb{I}\left[x(\theta)>1-\frac{M_nC_n\delta_n}{\epsilon_n}\right] f\Big(\epsilon_n^{-1/2}(\theta-\theta_0)\Big)P\Big[\hat{K}(\boldsymbol{X},\tilde{\boldsymbol{X}}^\theta)\leq\epsilon_n, d(\hat{D}^\theta_{n,m},D_\theta)\leq C_n\delta_n\Big]\mathrm{d}\theta}{\int\limits_{x(\theta)\leq 1-\frac{M_nC_n\delta_n}{\epsilon_n}}\mathrm{d}\theta}, \tag{38}$$

where the second equality follows from the fact that $x(\theta)\leq 1-\frac{M_nC_n\delta_n}{\epsilon_n}$ leads to $K(p_0,p_\theta)\leq \epsilon_n - M_nC_n\delta_n$ and then $K(p_0,p_\theta)\leq\lambda_n$ is trivially satisfied and where the last $o(1)$ comes from the set $\tilde{\Omega}_n(\tilde{C}_n)^c$.

Since we have $\epsilon_n \gg \delta_n$, with $u=\epsilon_n^{-1/2}(\theta-\theta_0)$, the first term is approximately equal to

$$N_1 = \frac{\int_{K(p_0,p_{\theta_0+\sqrt{\epsilon_n}u})\leq\epsilon_n} f(u)\mathrm{d}u}{\int_{K(p_0,p_{\theta_0+\sqrt{\epsilon_n}u})\leq\epsilon_n}\mathrm{d}u}\left(1+o(1)\right).$$

Using Assumption 5 again, we have

$$\int_{K(p_0,p_{\theta_0+\sqrt{\epsilon_n}u})\leq\epsilon_n}\mathrm{d}u = \int_{\frac{1}{2}\sqrt{\epsilon_n}u'I(\theta)\sqrt{\epsilon_n}u\leq\epsilon_n}\mathrm{d}u+o(1) = \int_{u'I(\theta_0)u\leq 2}\mathrm{d}u+o(1).$$

This leads to

$$N_1 = \frac{\int_{u'I(\theta_0)u\leq 2} f(u)\mathrm{d}u}{\int_{u'I(\theta_0)u\leq 2}\mathrm{d}u}\left(1+o(1)\right).$$

Next, we show that the fraction N_2 converges to 0. The numerator can be simplified as an integral

over $1 - \frac{M_n C_n \delta_n}{\epsilon_n} < x(\theta) \leq 1 + \frac{\widetilde{M}_n \delta_n}{\epsilon_n^{\kappa+1}}$, which can be bounded by (34) as

$$\begin{aligned} N_2 &\leq \frac{\|f\|_\infty \int_{1-\frac{M_n C_n \delta_n}{\epsilon_n} < x(\theta) \leq 1+\frac{\widetilde{M}_n \delta_n}{\epsilon_n^{\kappa+1}}} \mathrm{d}\theta}{\int_{x(\theta) \leq 1-\frac{M_n C_n \delta_n}{\epsilon_n}} \mathrm{d}\theta} \\ &\leq \|f\|_\infty \frac{D_2}{D_1} = o(1). \end{aligned}$$

Since we have $\epsilon_n^{\kappa+1} \gg \delta_n$, putting all the terms together, we obtain that the $P_0^{(n)}$-averaged ABC posterior distribution of $\epsilon_n^{-1/2}(\theta - \theta_0)$ is asymptotically uniform over the ellipsoid $\{u : u' I(\theta_0) u \leq 2\}$.