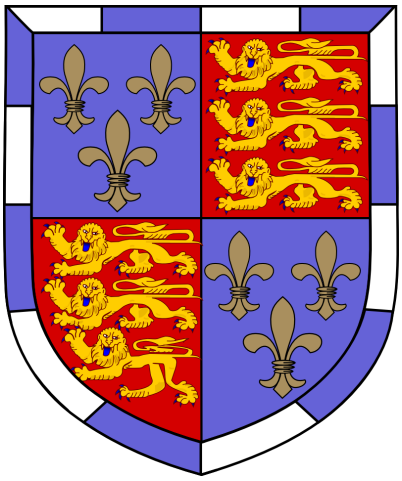
Approximate Bayesian Computation and Optimal Transport

Xing Liu
Supervised by Dr. Sergio Bacallado

This thesis is presented for the degree of
Master of Advanced Study.



Contents

Abstract

The complexity of many real-life data generating processes either defies the access to the likelihood function or renders it too expansive to be evaluated. In this case, standard Bayesian inference techniques, such as Markov chain Monte Carlo, can no longer be used. A popular roundabout is Approximate Bayesian Computation (ABC). ABC only assumes one has a generative model from which data can be drawn. It relies on a user-specified discrepancy metric that compares some summaries of the observation and the generated data. However, an improperly selected metric or summary may bias the discrimination between models. Optimal transport (OT) metrics have recently been proposed to remedy this issue. OT metrics are flexible, admit decent convergence properties and are often able to capture all differences between distributions. In this essay, we review and compare two OT metrics and one information-based measure that arose in the ABC literature, namely the Wasserstein distances, the maximum mean discrepancy (MMD) and the Kullback-Leibler (KL) divergence. We summarize the theoretical studies of their posterior concentration in the present literature, and discuss how these metrics can be adapted to large-scale data sets. We also compare these methods through four benchmark experiments, including a real-life study on ecological dynamic systems.

Acknowledgements

I would like to thank Dr. Sergio Bacallado for his valuable guidance on both this essay and my Part III study. His lectures on Modern Statistical Methods were amazing, and this essay topic he proposed has been particularly intriguing.

1 Introduction

Bayesian inference relies heavily on the likelihood function, which specifies connections between the data and some parameters governing the underlying model. When the likelihood function is known, classical Bayesian inference techniques such as Markov chain Monte Carlo is often used to draw posterior samples, and inference on the model parameters can then be made. However, the increasing complexity of many models in modern statistical research renders the likelihood function intractable or too costly to evaluate from. Approximate Bayesian computation (ABC) serves as an important tool to overcome this issue [Beaumont, 2019; Sisson et al., 2018; Marin et al., 2011]. ABC assumes that one has a *generative* model from which one can generate samples but has no access to its likelihood function. Its original idea was based on Pritchard et al. [2000], and a number of variants has been developed henceforth. ABC has been found useful in a range of applications, including model selection [Beaumont, 2006], archealogy [Wilkinson and Tavaré, 2009], ecological population study [Wood, 2010] and pathogen transmission [Tanaka et al., 2006].

Upon observing some data, the core idea of ABC is based on repetitive generation of *synthetic* (or *pseudo*) data from the generative model and only accepting those that are "close" to the observed one. Closeness is measured by some user-specified data discrepancy metric $\mathfrak{D}$ that discriminates the observed data from the synthetic ones. When numerical evaluation of the discrepancy between the full data is costly, one often adopts a low-dimensional transformation, known as a *summary statistic*. The quality of the ABC estimation depends heavily on both the data discrepancy and the summary statistic [Beaumont, 2019].

A main research theme in the past two decades has been devoted to providing guidelines on constructing suitable summary statistics (see e.g. Gutmann et al. [2018]; Fearnhead and Prangle [2012]). Another line of research that has recently gained popularity advertises the use of metrics on spaces of distributions as the data discrepancy, a celebrated example of which is the *optimal transport* (OT) metrics [Villani, 2009]. The latter approach is more flexible, circumvents the need to construct summary statistics and often exhibits a better posterior quality. Theoretical guarantees of the posterior concentration with specific OT metrics are also available in the current ABC literature, e.g. the Wasserstein ABC of Bernton et al. [2019a].

In this essay, we review and compare three ABC methods that use metrics on the space of probability measures, specifically, the Wasserstein distances, the maximum mean discrepancy (MMD) and the Kullback-Leibler (KL) divergence. The former two, in particular, are closely related to OT, while the latter is a famous divergence measure in information theory. This essay is based primarily on the papers Bernton et al. [2019a]; Park et al. [2016]; Bai Jiang [2018], which study respectively each of the three metrics in the context of ABC.

The rest of the essay is organized as follows. Section 1.1 introduces notations. Section 2 reviews the two fundamental ABC methods — the rejection and soft ABC — and provides illustrations via two toy examples. Section 3 reviews the Wasserstein distances, MMD and KL divergence in the context of ABC. Discussions on their posterior concentrations are provided in section 4. In particular, we extend the concentration guarantee for rejection ABC to a more general setting of soft ABC (Prop. 4.5), which, to our best knowledge, is not present elsewhere in the existing literature. In section 5, we apply these methods to four benchmark models, including a real-life example of ecological dynamic study, and comment on their performance. Section 6 concludes the essay.

1.1 Notation

We mostly follow the same notation in Bernton et al. [2019a]. Throughout, let $(\mathcal{X}, \mathcal{F})$ be a measurable space, where $\mathcal{X} \subset \mathbb{R}^d$ for some $d \in \mathbb{N}$ is endowed with some metric ρ. Denote by $\mathcal{P}(\mathcal{X})$ the set of

probability measures on $\mathcal{X}$. Each of the n observed data takes values in $\mathcal{Y} \subset \mathcal{R}^{d_y}$ for some $d_y \in \mathbb{N}$, and the random vector $y_{1:n} \coloneqq (y_1, \ldots, y_n)^\intercal \in \mathcal{Y}^n$ has distribution $\mu_*^{(n)} \in \mathcal{P}(\mathcal{Y}^n)$. We refer $\hat{\mu}_{y_{1:n}} \coloneqq n^{-1} \sum_{i=1}^n \delta_{y_i}$ to its empirical distribution even when y_i are non-i.i.d., where δ_{y_i} is the Dirac distribution with mass on y_i.

Let $\Theta \subset \mathbb{R}^{d_\theta}$ be the parameter space endowed with some metric ρ_Θ, where $d_\theta \in \mathbb{N}$. A model on $\mathcal{Y}^n$ is a collection of distributions parameterized by $\theta \in \Theta$ and is denoted by $\mathcal{M}(\mathcal{Y}^n) \coloneqq \{\mu_\theta^{(n)} : \theta \in \Theta\}$. A model is *identifiable* if $\mu_\theta = \mu_{\theta'}$ implies $\theta = \theta'$. Throughout, we only consider *purely generative models*, from which we can draw samples for any $\theta \in \Theta$, but whose likelihood is intractable.

Given $z_{1:n} \sim \mu_\theta^{(n)} \in \mathcal{M}(\mathcal{Y}^n)$, its empirical distribution is denoted as $\hat{\mu}_{\theta, z_{1:n}} \coloneqq n^{-1} \sum_{i=1}^n \delta_{z_i}$. We assume the sequences $\hat{\mu}_n \to \mu_*$ and $\hat{\mu}_{\theta,n} \to \mu_\theta$ as $n \to \infty$, for some $\mu_*, \mu_\theta \in \mathcal{P}(\mathcal{Y})$. Whenever we refer to μ_* and μ_θ, it is implicitly assumed that they exist.

Let $\mathfrak{D}$ be a discrepancy metric on $\mathcal{Y}^n$ quantifying the resemblance between the observed and synthetic data. It can be a function of the data themselves, or defined through empirical distributions formed by the data. Let s denote a summary statistic, i.e. a function of data mapping from $\mathcal{Y}^n$ to some subset of $\mathcal{Y}^n$. An identity function therefore corresponds to using the full data. With a slight abuse of notation, the composition $\mathfrak{D}(s(\cdot), s(\cdot))$ defines a dis-similarity metric for ABC algorithms. Whenever we refer to $\mathfrak{D}$ as the discrepancy metric, such composition is implicitly assumed. The fundamental question in ABC boils down to finding decent $\mathfrak{D}$ and s.

2 ABC with Summary Statistics

2.1 Rejection ABC

Suppose that π is a prior distribution on θ and both π and $\mu_\theta^{(n)}$ attain a density with respect to the Lebesgue measure. It follows from Bayes' rule that the posterior distribution of θ is given by

$$\pi(d\theta | y_{1:n}) = \frac{\mu_\theta^{(n)}(dy_{1:n})\pi(d\theta)}{\int_\Theta \mu_{\theta'}^{(n)}(dy_{1:n})\pi(d\theta')}.$$

In its most common form, the rejection ABC [Turner and Zandt, 2012; Sisson et al., 2018; Beaumont, 2019] proceeds by (i) sampling a candidate θ from π, (ii) generating synthetic data $z_{1:n}$ from $\mu_\theta^{(n)}$, and (iii) accepting θ if $\mathfrak{D}(y_{1:n}, z_{1:n}) \leq \epsilon$, where $\epsilon \geq 0$ is a similarity threshold (see Algorithm 1). In words, we keep θ if the model parametrized by θ generates data that are close to the true observation. This results in samples from the joint density

$$\pi(dz_{1:n}, d\theta \mid \mathfrak{D}(y_{1:n}, z_{1:n}) \leq \epsilon) \propto \mathbb{1}(\mathfrak{D}(y_{1:n}, z_{1:n}) \leq \epsilon)\mu_\theta^{(n)}(dz_{1:n})\pi(d\theta),$$

where $\mathbb{1}$ is the indicator function such that $\mathbb{1}(A) = 1$ if A is true and 0 otherwise. Marginalizing over $z_{1:n}$, we obtain the following *ABC posterior density*

$$\pi_{y_{1:n}}^\epsilon(d\theta) := \pi(d\theta \mid y_{1:n}; \mathfrak{D}, \epsilon) \propto \pi(d\theta) \int_{\mathcal{Y}^n} \mathbb{1}(\mathfrak{D}(y_{1:n}, z_{1:n}) \leq \epsilon)\mu_\theta^{(n)}(dz_{1:n}). \tag{1}$$

Given an observation $y_{1:n}$, it is known that, under mild conditions, the ABC posterior converges strongly to the true posterior $\pi(\cdot|y_{1:n})$ as ϵ tends to 0, see e.g. Bernton et al. [2019a]. Moreover, if one fix $\epsilon > 0$ and allows the number of observations to grow, the ABC posterior no longer converges to the true posterior, but only a conditional of it. See section 4 for detailed discussions on posterior convergence.

Algorithm 1: Rejection ABC

Input : Observed data $y_{1:n}$, discrepancy $\mathfrak{D}$, prior π, threshold ϵ.
Output: ABC posterior samples $\theta_j, j = 1, \ldots, M$.

Initialize $j = 0$;
while $j \leq M$ **do**
  Sample $\theta \sim \pi$;
  Sample $z_{1:n} \sim \hat{\mu}_\theta^{(n)}$;
  Accept $\theta_j = \theta$ if $\mathfrak{D}(y_{1:n}, z_{1:n}) \leq \epsilon$.
end while

Algorithm 2: Soft ABC

Input : Observed data $y_{1:n}$, discrepancy $\mathfrak{D}$, prior π, kernel κ, bandwidth ϵ.
Output: Empirical posterior distribution $\sum_{j=1}^{M} w_j \delta_{\theta_j}$.

for $j = 1, \ldots, M$ **do**
  Sample $\theta_j \sim \pi$;
  Sample $z_{1:n} \sim \hat{\mu}_{\theta_j}^{(n)}$;
  Set $\tilde{w}_j = \kappa_\epsilon(\mathfrak{D}(y_{1:n}, z_{1:n}))$;
end for
Normalize weights $w_j := \frac{\tilde{w}_j}{\sum_{k=1}^{M} \tilde{w}_k}$.

A major problem with the rejection-sampling paradigm of Algorithm 1 is that, for even moderately large n, the probability of sampling $z_{1:n}$ such that $\mathfrak{D}(y_{1:n}, z_{1:n}) \leq \epsilon$ can be vanishingly small. This motivates the use of a summary statistic s, i.e. accepting the proposed θ if $\mathfrak{D}(s(y_{1:n}), s(z_{1:n})) \leq \epsilon$. The rejection scheme then (marginally) samples θ from the density

$$\pi(d\theta \mid s(y_{1:n}); \mathfrak{D}, \epsilon) \propto \pi(d\theta) \int_{\mathcal{Y}^n} I\{\mathfrak{D}(s(y_{1:n}), s(z_{1:n})) \leq \epsilon\} \mu_\theta^{(n)}(dz_{1:n}).$$

Provided that s is *sufficient*, the likelihood $\mu_\theta^{(n)}$ depends on θ only through $s(z_{1:n})$. As a result, $\pi(\theta|s(y_{1:n}); \mathfrak{D}, \epsilon) = \pi(\theta|y_{1:n}; \mathfrak{D}, \epsilon)$ and this is equivalent to sampling with the full data with no "leakage" of information. If s is not sufficient, then ABC draws samples from a different distribution $\pi(\theta|s(y_{1:n}); \mathfrak{D}, \epsilon)$, thus resulting in inaccurate posterior samples. In practice, however, finding sufficient statistics is often a hard task, if not impossible. Where sufficient statistics are not available, common choices of summary statistics are sample moments and empirical quantiles [Fearnhead and Prangle, 2012].

2.2 Soft ABC

A moment's thought on Algorithm 1 reveals that it has the flavour of a kernel density estimation procedure. Indeed, the integrand in Eq. 1 can be viewed as the convolution between the approximate likelihood $\mu_\theta^{(n)}$ and a uniform kernel $\kappa_\epsilon(\mathfrak{D}(y_{1:n}, z_{1:n}))$, where $\kappa_\epsilon(u) = \epsilon^{-1}\mathbb{1}(u \leq \epsilon)$. Such a kernel imposes a hard threshold in the sense that it only keeps θ for which the associated $z_{1:n}$ is close to $y_{1:n}$, but it does not discriminate the rest. This is intuitively a waste of information. In light of this, *Soft ABC* [Sisson et al., 2018; Biau et al., 2015; Blum, 2010] allows a "soft" decision threshold by adopting a general kernel (see Algorithm 2).

More formally, we define a *kernel* to be a non-negative function $\kappa : \mathbb{R} \to [0, \infty)$ such that $\int \kappa(u)du = 1$, $\int u\kappa(u)du = 0$ and $\int u^2\kappa(u)du < \infty$. The *scaled kernel* is defined to be $\kappa_\epsilon(u) = \epsilon^{-1}\kappa(u/\epsilon)$, where

Kernel	$\kappa(u)$
Uniform	$\frac{1}{2}\mathbb{1}(\lvert u\rvert \leq 1)$
Gaussian	$\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}u^2)$
Laplacian	$\frac{4}{5}\exp(-\frac{1}{2}\lvert u\rvert)$
Triangular	$(1-\lvert u\rvert)\mathbb{1}(\lvert u\rvert \leq 1)$
Epanechnikov	$\frac{3}{4}(1-u^2)\mathbb{1}(\lvert u\rvert \leq 1)$

Table 1: The functional forms of some frequently used kernels.

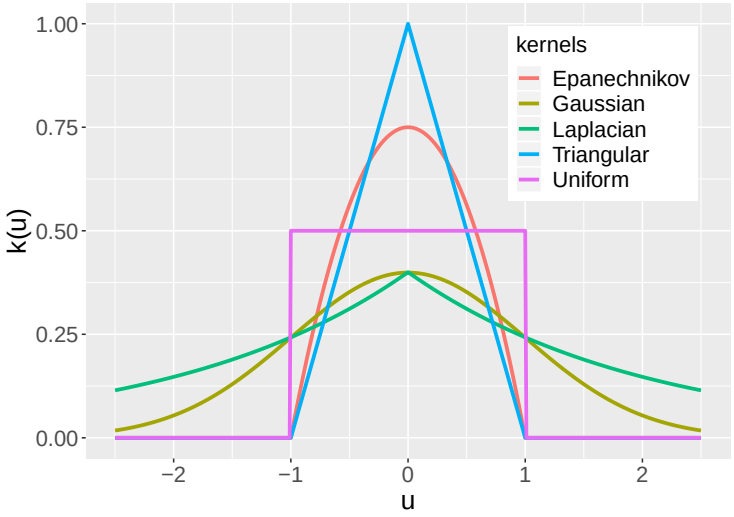


Table 2: Plot of kernels functions $\kappa(u)$ listed in Table 1.

$\epsilon > 0$ is called the *bandwidth* or *scale parameter*. Furthermore, it is often assumed that $\kappa_\epsilon(u)$ tends to a point mass at $u = 0$ as $\epsilon \to 0$. Popular choices of kernels are summarized in Table 1. In this essay, we focus primarily on the uniform and Gaussian kernels. Similarly to the nonparametric kernel density estimation, the choice of ϵ often has more impact on the performance of soft ABC than the choice of kernel [Sisson et al., 2018].

Instead of simple acceptance and rejection, soft ABC outputs a weighted sample $\{(\theta_j, w_j)\}_{j=1}^M$, where $M \in \mathbb{N}$ is the number of posterior samples, $w_j := \kappa_\epsilon(\mathfrak{D}(y_{1:n}, z_{1:n}^{(j)}))/\sum_{l=1}^M \kappa_\epsilon(\mathfrak{D}(y_{1:n}, z_{1:n}^{(l)}))$ are the weigths and $z_{1:n}^{(l)}$ is the l-th synthetic data. Posterior moments can hence be estimated directly: for a test function f, an unbiased estimator for $\int_\Theta f(\theta)\pi(d\theta|y_{1:n})$ is $\sum_{j=1}^M w_j f(\theta_j)$. The soft ABC posterior density generalizing Eq. 1 is

$$\pi_{y_{1:n}}^\epsilon(d\theta) \propto \pi(d\theta) \int_{\mathcal{Y}^n} \kappa_\epsilon(\mathfrak{D}(y_{1:n}, z_{1:n}))\mu_\theta^{(n)}(dz_{1:n}). \tag{2}$$

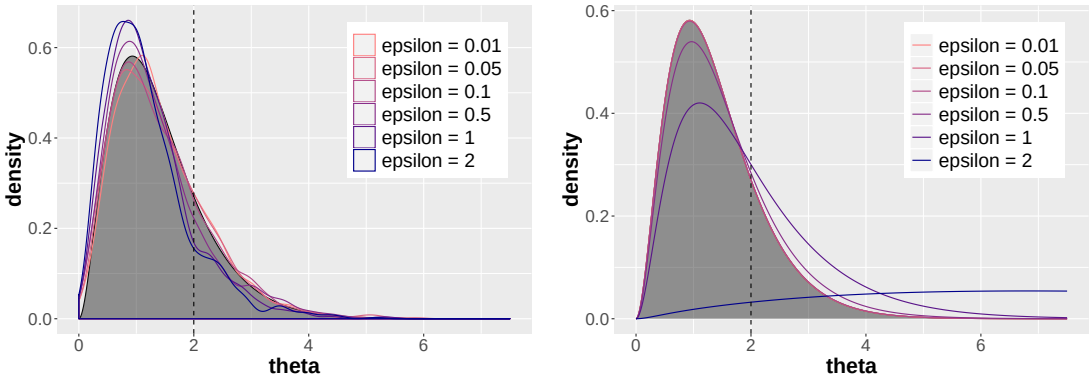
A similar expression with summary statistics can also be derived straightforwardly. In section 4, we show that, provided the kernel satisfies a so-called concentration condition, Eq. 2 converges strongly to the true posterior as ϵ tends to zero, similarly to the rejection ABC posterior.

2.3 Toy Examples

We run rejection and soft ABC on two concrete examples — an Exponential model with a Gamma prior and a normal model with a normal prior. They are designed in such a way that both the true and the ABC posteriors are available in close form so that the performance of the ABC methods can be compared directly. We use the Euclidean norm as the data discrepancy $\mathfrak{D}$ in both examples. For each threshold ϵ, $M = 1024$ posterior samples are drawn in producing the density plots.

2.3.1 Exponential Model

We consider one realization $y \in \mathcal{Y} = \mathbb{R}^+$ from $\text{Exp}(\theta_*)$ with $\text{Gamma}(\alpha_0, \beta_0)$ prior on parameter $\theta > 0$, where $\theta_*, \alpha_0, \beta_0 > 0$. It is known that the Gamma distribution is a conjugate prior for Exponential distributions; the resulting posterior is $\text{Gamma}(\alpha_0 + 1, \beta_0 + y)$ (see e.g. Casella and Berger [2001]). Furthermore, choosing $n = 1$ allows us to derive the rejection ABC posterior analytically from Eq. 1.



(a) Estimated ABC posteriors of θ.

(b) Exact ABC posteriors of θ.

Figure 1: Rejection ABC posteriors of the Exponential model with various acceptance thresholds. The true posterior density is shown by the grey areas. The approximation (left) worsens as ϵ increases. The exact ABC posterior tends to the true posterior as ϵ approaches 0. The true parameter $\theta_* = 2$ is shown by the black dotted line.

For $0 < \epsilon < \min(\beta_0, y)$ and $\theta > 0$, the ABC posterior density is

$$\pi_y^\epsilon(d\theta) \propto \pi(d\theta) \int_0^\infty \mathbb{1}(|y - z| \leq \epsilon)\theta e^{-\theta z} dz \propto \theta^{\alpha_0 - 1} e^{-\beta_0 \theta} \int_{y-\epsilon}^{y+\epsilon} \theta e^{-\theta z} dz \propto \theta^{\alpha_0 - 1} e^{-\beta_0 \theta} \left(e^{-(y-\epsilon)\theta} - e^{-(y+\epsilon)\theta} \right).$$

Integrating over $\Theta = \mathbb{R}^+$ and dividing by the normalizing constant yields the normalized ABC posterior density

$$\pi_y^\epsilon(d\theta) = \left(\frac{\Gamma(\alpha_0)}{(\beta_0 + y - \epsilon)^{\alpha_0}} - \frac{\Gamma(\alpha_0)}{(\beta_0 + y + \epsilon)^{\alpha_0}} \right)^{-1} \theta^{\alpha_0 - 1} e^{-(\beta_0 + y)\theta} \left(e^{\epsilon\theta} - e^{-\epsilon\theta} \right),$$

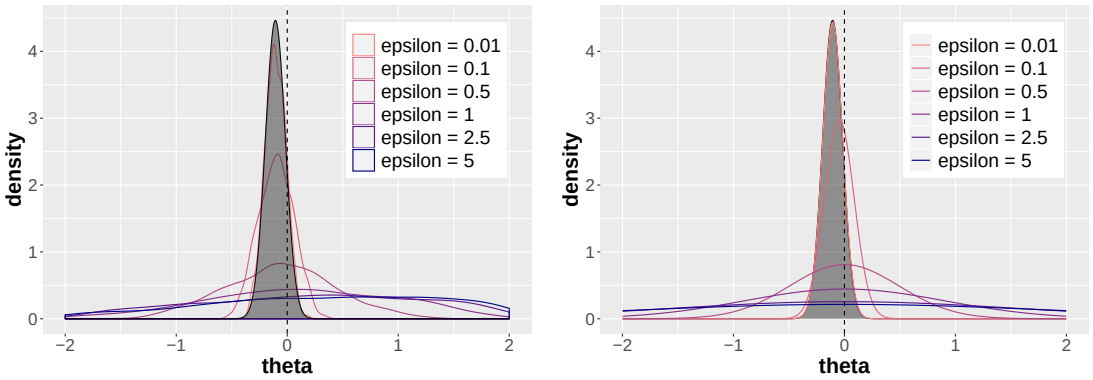
where Γ is the Gamma function. Keeping y fixed, a simple application of Taylor expansion on the normalizing constant yields

$$\begin{aligned}
\frac{\Gamma(\alpha_0)}{(\beta_0 + y - \epsilon)^{\alpha_0}} - \frac{\Gamma(\alpha_0)}{(\beta_0 + y + \epsilon)^{\alpha_0}} &= \frac{\Gamma(\alpha_0)}{(\beta_0 + y)^{\alpha_0}} \left(\frac{1}{(1 - \frac{\epsilon}{\beta_0 + y})^{\alpha_0}} - \frac{1}{(1 + \frac{\epsilon}{\beta_0 + y})^{\alpha_0}} \right) \\
&= \frac{\Gamma(\alpha_0)}{(\beta_0 + y)^{\alpha_0}} \left(1 + \frac{\alpha_0 \epsilon}{\beta_0 + y} - 1 + \frac{\alpha_0 \epsilon}{\beta_0 + y} + \mathcal{O}(\epsilon^3) \right) \\
&= \frac{\Gamma(\alpha_0)}{(\beta_0 + y)^{\alpha_0}} \left(\frac{2\alpha_0 \epsilon}{\beta_0 + y} + \mathcal{O}(\epsilon^3) \right).
\end{aligned}$$

Also, $\exp(\epsilon\theta) - \exp(-\epsilon\theta) = 2\epsilon\theta + \mathcal{O}(\epsilon^3)$. Combining the two gives

$$\pi_y^\epsilon(d\theta) = \theta^{\alpha_0 - 1} e^{-(\beta_0 + y)\theta} \cdot \frac{(\beta_0 + y)^{\alpha_0 + 1}(\epsilon\theta + \mathcal{O}(\epsilon^2))}{\alpha_0 \Gamma(\alpha_0)(\epsilon + \mathcal{O}(\epsilon^2))} = \frac{(\beta_0 + y)^{\alpha_0 + 1}}{\Gamma(\alpha_0 + 1)} \theta^{\alpha_0} e^{-(\beta_0 + y)\theta} \cdot \frac{\epsilon + \mathcal{O}(\epsilon^3)}{\epsilon + \mathcal{O}(\epsilon^3)},$$

which tends pointwise to a Gamma$(\alpha_0 + 1, \beta_0 + y)$ density as $\epsilon \to 0$, as expected. Fig 1 illustrates the true posterior as well as the exact and estimated ABC posterior densities for various choices of ϵ with $\theta_* = 2, \alpha_0 = 1$ and $\beta_0 = 1$. We can see that the estimation improves as ϵ decreases to 0, which matches the above analysis.



(a) Estimated ABC posteriors of θ.

(b) Exact ABC posteriors of θ.

Figure 2: Soft ABC posteriors of the univariate normal model using a Gaussian kernel and various bandwidths. The true posterior density is shown by the grey areas. The exact ABC posterior is a Gaussian density with a inflated variance compared with the true posterior, as shown by Eq. 4. Again, the ABC approximation improves as ϵ approaches 0. The true parameter $\theta_* = 0$ is shown by the black dotted line.

2.3.2 Univariate Normal Model

We consider i.i.d. samples y_i from $\mathcal{N}(\theta_*, \sigma_0^2)$, $i = 1, \ldots, n$, and a $\mathcal{N}(m_0, \tau_0^2)$ prior on the mean, with known $n \in \mathbb{N}, \sigma_0^2, \tau_0^2 > 0$ and m_0. Note first that the sample mean is a sufficient statistic for θ. Indeed, $\forall \theta \in \Theta = \mathbb{R}$,

$$\begin{aligned}
\mu_\theta^{(n)}(dz_{1:n}) &= \frac{1}{(2\pi\sigma_0^2)^{n/2}} \exp\left(-\frac{1}{2\sigma_0^2}\sum_{i=1}^n (z_i - \theta)^2\right) \\
&= \frac{1}{(2\pi\sigma_0^2)^{n/2}} \exp\left(-\frac{1}{2\sigma_0^2}\sum_{i=1}^n (z_i - \overline{z} + \overline{z} - \theta)^2\right) \\
&= \frac{1}{(2\pi\sigma_0^2)^{n/2}} \exp\left(-\frac{1}{2\sigma_0^2}\sum_{i=1}^n (z_i - \overline{z})^2\right) \exp\left(-\frac{n}{2\sigma_0^2}(\overline{z} - \theta)^2\right).
\end{aligned}$$

Sufficiency of $s(z_{1:n}) := \overline{z}_{1:n} = n^{-1}\sum_{i=1}^n z_i$ then follows from the Fisher-Neyman factorization theorem (see e.g. Casella and Berger [2001]). We perform soft ABC on s using a Gaussian kernel. The soft ABC approximation to the likelihood is given by

$$\begin{aligned}
\mu_\theta^{(n)}(\overline{z}_{1:n}) &\propto \int_{\mathbb{R}} \exp\left(-\frac{1}{2\epsilon}(\overline{z}_{1:n} - \overline{y}_{1:n})^2\right) \exp\left(-\frac{n}{2\sigma_0^2}(\overline{z}_{1:n} - \theta)^2\right) d\overline{z}_{1:n} \\
&\propto \int_{\mathbb{R}} \exp\left(-\frac{\epsilon^2 + \sigma_0^2/n}{2\epsilon^2\sigma_0^2/n}(\overline{z}_{1:n} - c)^2\right) \exp\left(-\frac{1}{2(\epsilon^2 + \sigma_0^2/n)}(\overline{y}_{1:n} - \theta)^2\right) d\overline{z}_{1:n} \\
&\propto \exp\left(-\frac{1}{2(\epsilon^2 + \sigma_0^2/n)}(\overline{y}_{1:n} - \theta)^2\right) \qquad (3)
\end{aligned}$$

for some constant c, where we have applied the identity $\phi_{\sigma_1}(x - m_1)\phi_{\sigma_2}(x - m_2) = \phi_{1/\sqrt{\sigma_1^2+\sigma_2^2}}(m_1 - m_2)\phi_{\sigma_1\sigma_2/\sqrt{\sigma_1^2+\sigma_2^2}}(x - (m_1\sigma_1^{-2} + m_2\sigma_2^{-2})/(\sigma_1^{-2} + \sigma_2^{-2}))$ to yield the second line, and $\phi_b(x - a)$ denotes a normal density with mean a and standard deviation b. Using a normal prior and applying this identity

again leads to the following soft ABC posterior density

$$\pi(d\theta|\overline{y}_{1:n};\mathfrak{D},\epsilon) \propto \exp\left(-\frac{\tau_0^{-2}+(\sigma_0^2/n+\epsilon^2)^{-1}}{2}\left(\theta-\frac{m_0\tau_0^{-2}+\overline{y}_{1:n}(\sigma_0^2/n+\epsilon^2)^{-1}}{\tau_0^{-2}+(\sigma_0^2/n+\epsilon^2)^{-1}}\right)^2\right). \tag{4}$$

Fig. 2 illustrates the results with $n = 500, \theta_* = 0, \sigma_0 = 2$, $m_0 = 1$ and $\tau_0 = 2$. We can see that for any fixed θ and $y_{1:n}$, $\pi(d\theta|\overline{y};\mathfrak{D},\epsilon) \to \pi(d\theta|y_{1:n})$, as $\epsilon \to 0$. This is not surprising given the sufficiency of the chosen summary statistic. This can also be anticipated by inspecting Eq. 3. Recall that $\overline{z}|\theta \sim \mathcal{N}(\theta, \sigma_0^2/n)$, so Eq. 3 can be viewed as the true likelihood with variance inflated by ϵ^2. The approximation should therefore be reasonable if ϵ^2 is dominated by σ_0^2/n.

3 ABC with Optimal Transport Metrics

Given the limitations of ABC methods that rely on summary statistics, various statistic-free approaches have been explored in recent literature. One line of research that has become particularly popular is to use dis-similarity measures on the space of distributions, which we refer to as OT metrics. Examples include the Wasserstein ABC (WABC) of Bernton et al. [2019a], which uses the Wasserstein distances. The recent work of Nadjahi et al. [2019] extends this idea to the sliced-Wasserstein distance [Kolouri et al., 2019] in order to alleviate the computational burden of the Wasserstein distances. Another ABC paradigm, termed K2-ABC [Park et al., 2016], projects the empirical distributions formed by the data into a reproducing kernel Hilbert space (RKHS) and uses the Hilbert space norm as the data discrepancy. Such a representation of distributions is known as the *kernel mean embedding* [Muandet et al., 2017], and the resulting metric is called the *maximum mean discrepancy* (MMD).

Another metric that has been studied in the context of ABC is the KL-divergence, which leads to the KL-ABC of Bai Jiang [2018]. It is less computationally demanding than WABC and K2-ABC, and discriminate models in a similar manner as using maximum likelihood estimates as the summary statistic.

Another potentially useful OT metric is the Sinkhorn divergence [Cuturi, 2013; Genevay et al., 2018]. It is an interpolation between the Wasserstein distances and the MMD, and can be computed in $\mathcal{O}(n^2)$ by the Sinkhorn's algorithm. Whether the Sinkhorn divergence satisfies the axioms of distance functions is still unknown, which leaves formal analysis on its behaviour within ABC an open problem. However, it has shown competitive performance in empirical experiments (e.g. in Bernton et al. [2019a]).

In this section, we review and compare WABC, K2-ABC and KL-ABC. We provide a brief background on each discrepancy metric, followed by discussions on its estimation and scalability to large-scale data.

3.1 Wasserstein ABC

Wasserstein ABC [Bernton et al., 2019a] replaces the arbitrary discrepancy metric in Algorithm 1 by the celebrated Wasserstein distances [Villani, 2009; Panaretos and Zemel, 2019]. The Wasserstein distances are metrics between probability measures and are inspired by the optimal transport problem. They have been applied to many areas in statistics and machine learning, including generative adversarial networks (GANs) [Arjovsky et al., 2017], goodness-of-fit testing [Ramdas et al., 2017; Panaretos and Zemel, 2019] and image retrieval [Grauman and Darrell, 2004]. We provide a brief review of Wasserstein distances and optimal transport before discussing its application to ABC.

3.1.1 Optimal Transport and Wasserstein Distances

The Wasserstein distance is a special case of the OT problem, which is one of the most fundamental optimization problems [Villani, 2009]. Its probabilistic formulation, in particular, asks how to find a

coupling that minimizes the cost between two random variables of known marginal distributions. The Wasserstein distance appears as the minimal value of the objective and serves naturally as a notion of distance between probability measures.

Using the same notations as before, let $\rho : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ be a ground distance on the data space $\mathcal{Y} \subset \mathbb{R}^{d_y}$, e.g. the Euclidean or L1 norm. For any $p \in [1, \infty)$, let $\mathcal{P}_p(\mathcal{Y})$ be the space of probability measures on $\mathcal{Y}$ with finite p-th moment, i.e. $\mathcal{P}_p(\mathcal{Y}) = \{\mu \in \mathcal{P}(\mathcal{Y}) : \int_{\mathcal{Y}} \rho(y_0, y)\mu(dy) < \infty \text{ for some } y_0 \in \mathcal{Y}\}$. This is known as the *Wasserstein space of order p*. It can be shown that this definition is in fact independent of the reference point y_0, as remarked on page 78 of Villani [2009].

Definition 1. The Wasserstein distance of order p between two distributions $\mu, \nu \in \mathcal{P}(\mathcal{Y})$ is

$$\mathcal{W}_p(\mu, \nu) = \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathcal{Y}\times\mathcal{Y}} \rho(y, z)^p \gamma(dy, dz) \right)^{1/p} = \inf_{Y\sim\mu, Z\sim\nu} (\mathbb{E}[\rho(Y, Z)^p])^{1/p},$$

where $\Gamma(\mu, \nu)$ is the set of measures with marginals μ and ν, known as the couplings of μ and ν.

It can be shown that the infimum is always attained [Villani, 2009, Theorem 4.1] and that $\mathcal{W}_p$ is finite on and *metrizes* $\mathcal{P}_p(\mathcal{Y})$ (definition of the metrizable property can be found in Appendix D.2; see also [Villani, 2009, page 84]). The special case $\mathcal{W}_1$ is also called the *Kantorovich-Rubinstein distance*. We focus primarily on $p = 1$ and on empirical distributions $\hat{\mu}_{y_{1:n}} = n^{-1} \sum_{i=1}^n \delta_{y_i}$ and $\hat{\mu}_{z_{1:m}} = m^{-1} \sum_{j=1}^m \delta_{z_j}$ formed by data $y_{1:n}$ and $z_{1:m}$. The Wasserstein distance between the empirical distributions takes the form

$$\mathfrak{D}_{\mathcal{W}_p}(\hat{\mu}_{y_{1:n}}, \hat{\mu}_{z_{1:m}}) \coloneqq \mathcal{W}_p(\hat{\mu}_{y_{1:n}}, \hat{\mu}_{z_{1:m}}) = \left(\inf_{\gamma \in \Gamma_{n,m}} \sum_{i=1}^n \sum_{j=1}^m \rho(y_i, z_j)^p \gamma_{ij} \right)^{1/p}, \tag{5}$$

where $\Gamma_{n,m} \coloneqq \{A \in \mathbb{R}^{n\times m} : A_{ij} \geq 0 \ \forall i, j, \text{ and } \sum_j A_{ij} = m^{-1} \ \forall i, \ \sum_i A_{ij} = n^{-1} \ \forall j\}$. In words, $\Gamma_{n,m}$ is the set of matrices with non-negative entries whose rows sum to m^{-1} and columns sum to n^{-1} [Villani, 2009, Introduction]. For simplicity, we shall write $\mathfrak{D}_{\mathcal{W}_p}(y_{1:n}, z_{1:m})$ for $\mathfrak{D}_{\mathcal{W}_p}(\hat{\mu}_{y_{1:n}}, \hat{\mu}_{z_{1:m}})$.

3.1.2 Fast WABC

When $m = n$, finding the optimal $\gamma^* \in \Gamma_{n,m}$ for Eq. 5 can be viewed as an assignment problem [Bernton et al., 2019b]. In one-dimensional cases, this can be done by first finding permutations σ_y and σ_z that respectively sort $y_{1:n}$ and $z_{1:n}$ in increasing order, then associating each y_i with $z_{\sigma_z \circ \sigma_y^{-1}(i)}$. The computational cost is therefore of the same order as a sorting problem ($\mathcal{O}(n \log n)$). However, for dimensions larger than 1, the cost is often of order n^3 using the state-of-the-art Hungarian algorithm [Munkres, 1957]. This can be prohibitive for large n. However, in some ABC settings, the cost of generating samples from $\mu_\theta^{(n)}$ would dominate the cost of evaluating Wasserstein distances whatsoever.

In our experiments, we follow the implementation in Bernton et al. [2019a], which adopts the *shortlist method* elaborated in Gottschlich and Schuhmacher [2014] and implemented by the R package `transport` [Schuhmacher et al., 2017]. This method has no guarantee of polynomial running times, but is often found to be subcubic in practice.

One way to speed up the computation of Eq. 5 is to modify the definition of Wasserstein distance by including a regularization term, which leads to the Sinkhorn divergence (SD) [Cuturi, 2013]. Depending on the regularization strength, the SD would tend to the Wasserstein distances (no regularization) or MMD (infinitely large regularization). Each evaluation of the SD requires $\mathcal{O}(n^2)$.

Bernton et al. [2019a] also proposed two alternatives to Wasserstein distances that enjoy a subcubic cost. The first one is the *Hilbert distance*, which generalizes the idea of solving for $\mathcal{W}_p$ by sorting in

the univariate case to multivariate distributions through projections via the Hilbert space-filling curve [Gerber and Chopin, 2014]. It upper-bounds $\mathcal{W}_p$, approximates $\mathcal{W}_p$ well for small d_y and discriminates between parameters in a similar fashion to $\mathcal{W}_p$. The cost of one call is $\mathcal{O}(n \log n)$.

A second metric they mentioned is called the *swapping distance*. It was originally proposed by Puccetti [2017], and is computed via a greedy swapping algorithm that finds an approximation to the optimal assignment σ. Beginning with the σ obtained by Hilbert sorting, the algorithm proceeds by checking, for all $1 \leq i < j \leq n$, whether swapping $\sigma(i)$ and $\sigma(j)$ would result in a decline in $\rho(y_i, z_{\sigma(i)})^p + \rho(y_i, z_{\sigma(j)})^p$. By construction, the swapping distance is bounded from below by $\mathcal{W}_p$ and from above by the Hilbert distance. It requires the same order of cost $\mathcal{O}(n^2)$ per evaluation as the Hilbert distance.

3.2 K2-ABC

K2-ABC, first presented by Park et al. [2016], is an ABC paradigm that uses the *maximum mean discrepancy* between the so-called *kernel mean embedding* (KME) of the empirical distributions formed by the data. Given a positive definite kernel (see Appendix A for its definition), the KME transforms a probabilistic measure to an element in the associated *reproducing kernel Hilbert space* (RKHS). The Hilbert space distance on the RKHS can then be used as a discrepancy between the embeddings. When the kernel possesses a so-called characteristic property, such an embedding mapping is injective. The embedding mapping plays the role of summary statistics and is able to capture all information about the data whenever a characteristic kernel is used [Muandet et al., 2017].

Before presenting the K2-ABC algorithm, we first define kernel mean embeddings on the space of probability measures, $\mathcal{P}(\mathcal{Y})$. With a slight abuse of notation, we will occasionally denote random variables by both capital and lower-case letters in this section. Whether a lower-case letter is random should be clear from context.

3.2.1 Kernel Mean Embedding

Definition 2 (Kernel mean embedding)**.** Given a probabilistic measure $\mu \in \mathcal{P}(\mathcal{X})$ and a positive definite kernel $k : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, the *kernel mean embedding*[1] of μ is $m_\mu \coloneqq \mathbb{E}_{Z\sim\mu}[k(\cdot, Z)]$.

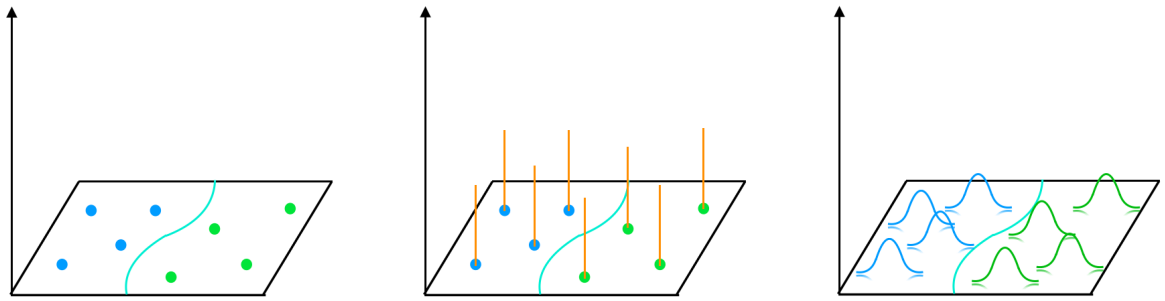
We remark that the positive definite kernel k is defined on the product domain of $\mathcal{Y}$ of an arbitrary dimension. This is not to be confused with the kernel κ introduced in section 2, which is only defined on $\mathbb{R}$ (see A of appendix).

The kernel mean embedding m_μ exists so long as k is bounded or, in the case where k is unbounded, μ satisfies a suitable moment condition with respect to k [Muandet et al., 2017, Lemma 3.1]. We shall give a proof of the former case, which holds for most commonly used kernels. When the embedding exists, it is an element of the RKHS $\mathcal{H}$ associated with kernel k.

The KME gives a representation property of distributions in a Hilbert space. Recall that a positive definite kernel k has the reproducing property where, for $y, z \in \mathcal{Y}$, $k(y, z) = \langle k(\cdot, y), k(\cdot, z)\rangle_{\mathcal{H}}$. A typical interpretation of the kernel is therefore a mapping of an element in $\mathcal{Y}$ to an inner product of some elements $z \mapsto k(\cdot, z)$ in $\mathcal{H}$. KMEs generalize this to a measurable space $(\mathcal{Y}, \mathcal{F})$ through a map $\mu \mapsto \int k(\cdot, z')d\mu(z')$. Indeed, if the measure is a Dirac-delta distribution δ_z with point mass at z, then we recover the usual interpretation, because any measurable and finite-valued function f is integrable with respect to δ_z, and its integral equals to $f(z)$ [Muandet et al., 2017]. This suggests that, similarly to the case of $\mathcal{Y}$ input, the KME can be viewed measure-theoretically as a representer of μ in the Hilbert space (see Fig. 3).

We now give conditions under which the kernel mean embeddings exist. Throughout the rest of this subsection, let $\mu \in \mathcal{P}(\mathcal{Y})$, k a positive definite kernel on $\mathcal{Y}$ and $\mathcal{H}$ its associated RKHS.

[1] We have used an unconventional notation for the kernel mean embedding to distinguish from the symbol for measures. In other literature, the common notation is μ_F for a distribution F.



(a) Ordinary RKHS representation: $z \mapsto k(\cdot, z) \in \mathcal{H}$

(b) Representation of Dirac measures: $\delta_z \mapsto m_{\delta_z} \in \mathcal{H}$

(c) Representation of general measures: $\mu \mapsto m_\mu \in \mathcal{H}$

Figure 3: Measure-theoretical interpretation of kernel mean embeddings. (a): Typical representation of data z by some high-dimensional feature map defined through the kernel k. (b): An embedding of data z into a high-dimensional feature space can be viewed measure-theoretically as an embedding of the Dirac distribution. (c): Generalising the embedding of Dirac distributions, we can extend the concept of feature maps to an arbitrary class of probability measures.

Proposition 3.1 (Existance of kernel mean embeddings)**.** *If* $\mathbb{E}_{Z\sim\mu}[\sqrt{k(Z,Z)}] < \infty$*, then* $m_\mu \in \mathcal{H}$ *and* $\mathbb{E}_{Z\sim\mu}[f(Z)] = \langle f, m_\mu\rangle_{\mathcal{H}}$*,* $\forall f \in \mathcal{H}$*.*

A direct consequence of Prop. 3.1 is that the kernel mean embedding of μ exists so long as k is bounded on $\mathcal{Y}$. Most commonly used kernels, such as the Gaussian and the Laplace kernels, satisfy this condition (see A). On the contrary, non-constant polynomial kernels may violate this assumption if μ has an unbounded support. An example is a linear kernel with the Cauchy distribution.

Moreover, the second part of Prop. 3.1 says that the expectation with respect to μ of any f in the RKHS can be computed as an inner product between the function itself and the embedding m_μ. This can be viewed as a reproducing property of the expectation operator in the RKHS [Muandet et al., 2017].

We shall use the celebrated Riesz representation theorem to prove Prop. 3.1. The proof of Riesz representation Theorem lies beyond the scope of this essay (see e.g. Halmos [1982]).

Theorem 3.2 (Riesz representation)**.** *If* $\mathbf{L} : \mathcal{H} \to \mathbb{R}$ *is a bounded linear operator on Hilbert space* $\mathcal{H}$*, then there exists* $h \in \mathcal{H}$ *such that,* $\forall f \in \mathcal{H}$*,* $\mathbf{L}[f] = \langle f, h\rangle_{\mathcal{H}}$*.*

Proof of Prop. 3.1. Define the functional $\mathbf{L}_\mu : f \mapsto \mathbb{E}_{Z\sim\mu}[f(Z)]$, for all $f \in \mathcal{H}$. Then $\mathbf{L}_\mu$ is a linear operator. Moreover, for any $f \in \mathcal{H}$ and $z \in \mathcal{Y}$, the representation property gives $f(z) = \langle f, k(\cdot, z)\rangle_{\mathcal{H}}$. Therefore,

$$|\mathbf{L}_\mu[f]| = |\mathbb{E}_{Z\sim\mu}[f(Z)]| \leq \mathbb{E}_{Z\sim\mu}[|f(Z)|] = \mathbb{E}_{Z\sim\mu}[|\langle f, k(\cdot, Z)\rangle_{\mathcal{H}}|] \leq \mathbb{E}_{Z\sim\mu}\left[\|f\|_{\mathcal{H}}\sqrt{k(Z,Z)}\right],$$

where the first inequality follows from Jensen's inequality and the last step is due to the Cauchy-Schwarz inequality. Therefore, $\mathbf{L}_\mu$ is bounded if $\mathbb{E}_{Z\sim\mu}[\sqrt{k(Z,Z)}]$ is. By Riesz representation theorem, there exists $h \in \mathcal{H}$ such that $\mathbf{L}_\mu[f] = \langle f, h\rangle_{\mathcal{H}}$. Choose $f := k(\cdot, z)$ for any $z \in \mathcal{Y}$. Then $h(z) = \langle k(\cdot, z), h\rangle_{\mathcal{H}} = \mathbf{L}_\mu[k(\cdot, z)] = \int_{\mathcal{Y}} k(z, z')\mu(dz')$. That is, $h = \int_{\mathcal{Y}} k(\cdot, z')\mu(dz') = m_\mu$. ∎

3.2.2 Maximum Mean Discrepancy

Having constructed the kernel mean embedding, which maps a distribution to an element in the RKHS, we can use the RKHS distance to quantify the discrepancy between embeddings, thus also between distributions. This leads to the following maximum mean discrepancy.

Definition 3 (Maximum mean discrepancy)**.** Given two probability measures μ and ν, their (squared) *maximum mean discrepancy* (MMD) is the Hilbert space distance between their embeddings:

$$\begin{aligned}\mathrm{MMD}^2(\mu,\nu) &= \|m_\mu - m_\nu\|_{\mathcal{H}}^2 = \langle m_\mu - m_\nu, m_\mu - m_\nu\rangle_{\mathcal{H}} \\ &= \mathbb{E}_{Y\sim\mu}\mathbb{E}_{Y'\sim\mu}[k(Y,Y')] + \mathbb{E}_{Z\sim\mu}\mathbb{E}_{Z'\sim\mu}[k(Z,Z')] - 2\mathbb{E}_{Y\sim\mu}\mathbb{E}_{Z\sim\nu}[k(Y,Z)],\end{aligned}$$

where Y, Y', Z, Z' are independent.

The second line holds because of the reproducing property of kernels: $\langle k(\cdot,y), k(\cdot,z)\rangle_{\mathcal{H}} = k(y,z)$. Here, we only consider *characteristic* kernels, for which the kernel mean embedding is injective [Muandet et al., 2017; Fukumizu et al., 2008]. This therefore guarantees that $\mathrm{MMD}^2(\mu,\nu) = 0$ if and only if $\mu = \nu$, i.e. no information is lost after the mapping. This is the core of the idea of K2-ABC. The characteristic property also ensures that the MMD satisfies the three axioms of metrics on $\mathcal{P}(\mathcal{Y})$, where symmetry and triangular inequality follows from the fact that $\|\cdot\|_{\mathcal{H}}$ is a distance on the RKHS. Two examples of characteristic kernels are the Gaussian and Laplacian kernels. We refer the reader to Muandet et al. [2017] for details about the characteristic property.

The definition of MMD suggests a natural estimator. Given distributions μ, ν and mutually independent random samples $y_i \sim \mu$ and $z_j \sim \nu$, $i = 1,\dots,n$ and $j = 1,\dots,m$, an unbiased estimator for $\mathrm{MMD}^2(\mu,\nu)$ takes the form

$$\widehat{\mathrm{MMD}}^2(\mu,\nu) \coloneqq \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} k(y_i,y_j) + \frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j\neq i} k(z_i,z_j) - \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} k(y_i,z_j). \tag{6}$$

Unless otherwise mentioned, we assume samples are of equal size, so that $n = m$. Similar to the case of $\mathcal{W}_p$, we write $\mathfrak{D}_{\mathrm{MMD}}(y_{1:n}, z_{1:n})$ for $\widehat{\mathrm{MMD}}^2(\mu,\nu)$. That this estimator is unbiased is obvious from independence. Its form is in fact motivated from a more general family of the so-called U-statistics [Gretton et al., 2007, Lemma 5]. We remark that this estimator may take negative values, even though MMD is clearly non-negative.

MMD is closely related to the 1-Wasserstein distance. The kernel mean embedding (thus also MMD) can be defined alternatively as the supremum of some objective function over the unit ball in the RKHS (Gretton et al., 2012, p. 5; Muandet et al., 2017, p. 50), as outlined in Appendix B. In light of this, both MMD and the 1-Wasserstein distance are in fact special cases of a more general class of metrics on distributions, termed *integral probability metrics* [Müller, 1997]. The equivalence between the two definitions of MMD and of $\mathcal{W}_1$ is known as the *dual property*. Using this dual definition, Theorem 21 of Sriperumbudur et al. [2010] proves the inequality $\mathrm{MMD}^2(\mu,\nu) \leq \mathcal{W}_1(\mu,\nu)$, where the ground distance ρ is the Euclidean norm. In fact, $\mathcal{W}_1$ is in turn bounded by the Sinkhorn divergence. We shall see that this will help with the asymptotic analysis of the posterior concentration of MMD-ABC, which is defined in the next subsection.

3.2.3 K2-ABC

We are now ready to present the K2-ABC algorithm, which uses the MMD as the data discrepancy in Algorithm 2 of the soft ABC. Let $\hat{\mu}_{y_{1:n}}$ and $\hat{\mu}_{z_{1:n}}$ be the empirical distributions of the observation $y_{1:n}$ and synthetic data $z_{1:n}$, respectively. The K2-ABC algorithm in its original form employs a second kernel κ_ϵ on $\mathfrak{D}_{\mathrm{MMD}}(y_{1:n}, z_{1:n})$, typically of the form

$$\kappa_\epsilon(\mathfrak{D}_{\mathrm{MMD}}(y_{1:n}, z_{1:n})) = \exp\left(-\frac{\mathfrak{D}_{\mathrm{MMD}}(y_{1:n}, z_{1:n})}{2\epsilon}\right), \epsilon > 0.$$

It then uses the soft ABC paradigm to construct a set of weights to be assigned to each ABC posterior sample. Overall, two kernels are used, hence the "2" in its name. This idea is similar to

setting the empirical kernel embedding $\sum_{j=1}^{n} k(\cdot, y_i)$ to be the summary statistic. However, in that case, $\|s(y_{1:n}) - s(z_{1:n})\|_{\mathcal{H}}^2 = \mathrm{MMD}^2(\hat{\mu}_{y_{1:n}}, \hat{\mu}_{z_{1:n}})$ would have been biased. Moreover, we are reassured that all possible differences (i.e. moments) between $\hat{\mu}_{y_{1:n}}$ and $\hat{\mu}_{z_{1:n}}$ can be captured by choosing a characteristic k.

Obviously, the choice of the two kernels can heavily affect the effectiveness of K2-ABC. In fact, Sriperumbudur et al. [2010] pointed out that the characteristic property of the kernel k used for the embedding is the more crucial factor. They also proved that for any given characteristic kernel and bandwidth $\epsilon > 0$, there exist distributions $\mu \neq \nu$ which MMD cannot distinguish up to ϵ, i.e. $\mathrm{MMD}^2(\mu, \nu) \leq \epsilon$. In our implementation, we shall not use the K2-ABC in its original form of Eq. 3.2.3, but choose κ to be the uniform kernel so that the results are comparable with WABC and KL-ABC. We call this the MMD-ABC to distinguish from the original K2-ABC of Park et al. [2016].

3.2.4 Fast K2-ABC

The computation time of K2-ABC has a quadratic order in n and can scale badly with the data size. Indeed, each computation of $\mathfrak{D}_{\mathrm{MMD}}$ requires $\mathcal{O}(n^2)$, so Algorithm 2 with MMD has an order $\mathcal{O}(Mn^2)$ overall, where M is the number of posterior samples. Park et al. [2016] discussed two alternatives to speed it up — the linear-time MMD and approximations through random Fourier features.

The linear-time MMD simply replaces Eq. 6 with the following estimator that depends on the data size linearly:

$$\widehat{\mathrm{MMD}}_{\mathrm{L}}(\mu, \nu) \coloneqq \frac{1}{n-1}\sum_{i=1}^{n-1} k(y_i, y_{i+1}) + \frac{1}{m-1}\sum_{j=1}^{m-1} k(z_j, z_{j+1}) - \frac{2}{m}\sum_{j=1}^{m} k(y_j, z_j),$$

where we have assumed without loss of generality that $n \leq m$ and denoted $y_i \coloneqq y_{1+\mathrm{mod}(i-1,n)}$ as a cyclic shift. When $n = m$, the resulting K2-ABC has cost $\mathcal{O}(Mn)$.

Alternatively, one can use random Fourier features, which aims to construct a random feature map $\hat{\phi} : \mathcal{Y} \to \mathbb{R}^L$ such that $k(z, z') \approx \hat{\phi}(z)^\top \hat{\phi}(z')$, where L is the number of features. An estimator for

$$\begin{aligned}\mathrm{MMD}^2(\mu, \nu) &\approx \mathbb{E}_{Y\sim\mu}[\hat{\phi}(Y)]^\top \mathbb{E}_{Y'\sim\mu}[\hat{\phi}(Y')] + \mathbb{E}_{Z\sim\nu}[\hat{\phi}(Z)]^\top \mathbb{E}_{Z'\sim\nu}[\hat{\phi}(Z')] - 2\mathbb{E}_{Y\sim\mu}[\hat{\phi}(Y)]^\top \mathbb{E}_{Z\sim\nu}[\hat{\phi}(Z)] \\ &= \left\|\mathbb{E}_{Y\sim\mu}[\hat{\phi}(Y)] - \mathbb{E}_{Z\sim\nu}[\hat{\phi}(Z)]\right\|_2^2\end{aligned}$$

is then

$$\widehat{\mathrm{MMD}}^2_{\mathrm{rf}}(\mu, \nu) \coloneqq \left\|\frac{1}{n_y}\sum_{i=1}^{n_y} \hat{\phi}(y_i) - \frac{1}{n_z}\sum_{i=1}^{n_z} \hat{\phi}(z_i)\right\|_2^2,$$

where $y_i \overset{iid.}{\sim} \mu$ and $z_i \overset{iid.}{\sim} \nu$. This estimator is biased but leads to $\mathcal{O}(MLn)$ for K2-ABC. It therefore can reduce the computational burden when $n \gg L$. Park et al. [2016] showed through a number of experiments that the random Fourier approximation yields much more accurate estimation than the linear-time estimator in most cases.

To construct $\hat{\phi}$, one way is to use the Bochner's theorem (Rudin [2011]). It states that, for any translation-invariant kernel k (i.e. $k(z, z') = \tilde{k}(z - z')$, for some $\tilde{k}$), there exists a measure Λ and positive constant c such that

$$\begin{aligned}\tilde{k}(z - z') &= c\int \exp(iw^\top(z - z'))d\Lambda(w) = c\mathbb{E}_{W\sim\Lambda}[\cos(W^\top(z - z'))] \\ &= 2c\mathbb{E}_{U\sim\mathrm{Unif}(-\pi,\pi)}[\mathbb{E}_{W\sim\Lambda}\{\cos(W^\top z + U)\cos(W^\top z' + U)|U\}].\end{aligned}$$

Therefore, one can draw $W_j \sim \Lambda$ and $U_j \sim \text{Uniform}(-\pi, \pi)$, $j = 1, \ldots, L$, to approximate k. A Gaussian k of the form presented in Appendix A corresponds to $c = 1$ and $\mathcal{N}(0, \sigma^2 I_n)$ for Λ. It follows that, for $j = 1, \ldots, L$,

$$\hat{\phi}_j(z) = \sqrt{\frac{2}{L}} \cos(W_j^\top z + U_j).$$

3.3 KL-ABC

Another discrepancy metric that has been studied with ABC is the *Kullback-Leibler* (KL) divergence, also known as the information divergence or relative entropy [Kullback and Leibler, 1951; MacKay, 2003].

Definition 4 (KL divergence)**.** Given two probability measures $\mu, \nu \in \mathcal{P}(\mathcal{Y})$ that attain densities $\mu(dy), \nu(dz)$ with respect to the Lebesgue measure, the KL-divergence is defined to be

$$\text{KL}(\mu \| \nu) = \int \mu(dy) \log \frac{\mu(dy)}{\nu(dy)}. \tag{7}$$

It follows from a simple application of the Jensen's inequality that $\text{KL}(\mu\|\nu)$ is non-negative. Moreover, it is finite whenever μ is absolutely continuous with respect to ν, and zero if and only if $\mu = \nu$ [MacKay, 2003]. It is, however, not a metric in the strict sense, as it needs not obey symmetry and the triangular inequality. A consistent estimator for Eq. 7 is

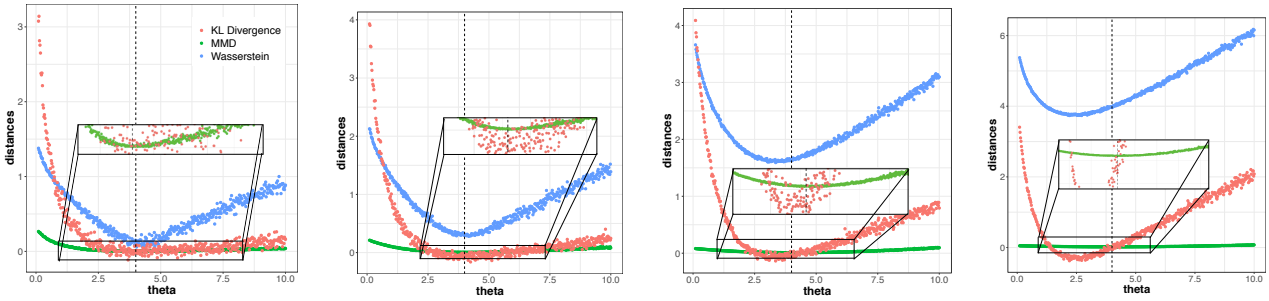
$$\mathfrak{D}_{\text{KL}}(y_{1:n}, z_{1:n}) := \frac{d_y}{n} \sum_{i=1}^{n} \log \left(\frac{\min_{j=1,\ldots,n} \|y_i - z_j\|_2}{\min_{\substack{j \neq i \\ j=1,\ldots,n}} \|y_i - y_j\|_2} \right) + \log \frac{n}{n-1}, \tag{8}$$

where $y_i \overset{iid.}{\sim} \mu$, $z_j \overset{iid.}{\sim} \nu$, and d_y is the dimension of data y_i as before. This estimator is a special case of Eq. 14 in Perez-Cruz [2008], which is motivated from a k-nearest neighbour density estimator. Perez-Cruz [2008] also established the almost sure convergence of this estimator. We defer this rather technical proof to the appendix (see C). The KL-ABC of Bai Jiang [2018] uses Eq. 8 as the discrepancy metric between the observed and synthetic data.

It is known that the (discretized) KL divergence between the empirical distributions of μ_* and μ_θ is minimized by the maximum likelihood estimator (MLE) [Bai Jiang, 2018]. KL-ABC shares the same idea and can be essentially thought as using MLEs as the summary statistic to quantify dis-similarity between distributions.

3.3.1 Computation

The estimation of KL-divergence by Eq. 8 boils down to finding the nearest neighbours. As a result, a call to the estimator Eq. 8 require $\mathcal{O}(n \log n)$, compared with $\mathcal{O}(n^3)$ for the Wasserstein distances and $\mathcal{O}(n^2)$ for the MMD. KL-ABC therefore has this computational benefit of admitting a sub-quadratic cost in data size. However, as discussed previously, the KL divergence essentially quantifies discrepancy only in terms of the MLEs, whereas the Wasserstein distances and the maximum mean embedding are able to capture all differences in moments up to some arbitrary order. As a result, KL-ABC has the caveat of "information leakage". Moreover, as mentioned on page 4 of Sriperumbudur et al. [2010] and the references therein, the estimator for the KL divergence exhibits arbitrarily slow convergence rates depending on the distribution, whereas $\mathcal{W}_1$ and MMD have good convergence behaviours.



(a) $d = 1$. (b) $d = 2$. (c) $d = 5$. (d) $d = 10$.

Figure 4: Discrepancy metrics evaluated at data generated from $\mathcal{N}(0, \theta I_{d_y})$ for varying $\theta \in [0.1, 10]$. Here, the observed data $y_{1:n}$ is generated with parameter $\theta_* = 4$, as marked by the black dotted lines.

	$\mathcal{W}_1$				MMD^2				KL divergence			
Dimensions	1	2	5	10	1	2	5	10	1	2	5	10
Time (milliseconds)	2920	4900	4690	4970	73.1	139	285	628	5.05	6.81	13.9	90.0

Table 3: Computational time (in milliseconds) for one evaluation of the three discrepancy metrics in different dimensions. Results are averages over 100 runs and rounded to 3 significant figures. There is a clear gap in the run times of $\mathcal{W}_1$ in univariate and multivariate cases.

3.4 Sequential Sampling of the ABC Posterior

As illustrated in section 2.3, the acceptance threshold ϵ is a crucial parameter that trades-off computational effort and the quality of the posterior samples. A computationally effective approach to anneal ϵ is through a sequential Monte Carlo (SMC) algorithm introduced by Del Moral et al. [2012] and implemented in Bernton et al. [2019a]. Specifically, this algorithm relies on a transition kernel to select a sequence of thresholds $(\epsilon_t)_{0 \le t \le T}$. It begins by drawing M posterior samples with $\epsilon_0 = \infty$, for which the ABC posterior is effectively the prior. Upon choosing ϵ_{t-1}, it determines ϵ_t such that αM samples from the previous step are accepted, for some user-specified $\alpha \in (0, 1)$. A resampling and rejuvenation step is then performed to obtain M samples using the transition kernel and the current threshold ϵ_t, and the algorithm proceeds. The output is M posterior samples associated with ϵ_T.

Although any MCMC kernel can be used as the transition kernel, in our implementation, we adopt the r-hit kernel of Lee [2012], following the implementation in Bernton et al. [2019a]. In all experiments in section 5, we set $M = 1024$ and $\alpha = 50\%$. We choose $r = 2$ in the r-hit kernel and a mixture of multivariate normal distributions with 5 components as the proposal of the MCMC steps. This framework is shown to be more advantageous than vanilla rejection ABC [Lee and Łatuszyński, 2014].

3.5 Comparing Discrepancy Metrics

We provide a concrete example where we compare the three discrepancy metrics introduced earlier between different distributions. We consider the model $\mathcal{N}(0, \theta I_{d_y})$, where $d_y \in \mathbb{N}$ is the dimension of the data and $\theta \in \mathbb{R}^+$ is the parameter. $n = 1000$ i.i.d. observations are drawn with $\theta_* = 4$. We then generate $z_{1:n}$ from the model using a grid of 500 values of θ equi-spaced on interval $[0.1, 10]$. We plot θ against $\mathfrak{D}_{\mathcal{W}_1}(y_{1:n}, z_{1:n})$, $\mathfrak{D}_{\text{MMD}}(y_{1:n}, z_{1:n})$ and $\mathfrak{D}_{\text{KL}}(y_{1:n}, z_{1:n})$ in Fig. 4 and list their required computational times in Table 3 (recall that $\mathfrak{D}_{\text{MMD}}(y_{1:n}, z_{1:n})$ denotes the squared MMD estimator 6). All computational times are estimated by averaging over 100 runs using R [R Core Team, 2019] and an 1.3 GHz Intel Core I5 processor.

We remark that the (squared) MMD is indeed always no larger than $\mathcal{W}_1$ as expected. The minima of the the metrics shift away from the true parameter $\theta_* = 4$ (black dotted lines) as dimension increases, which is an example of the curse of dimensionality. This is particularly not surprising for the KL divergence, as it is known that the KL divergence is minimized at the MLE, which biased for θ_* in this example. Finally, the 1-Wasserstein distance and MMD are indeed much more computationally demanding than KL divergence.

4 Asymptotic Behaviours

In this section, we investigate the convergence of the ABC posterior as (i) $n \to \infty$ when ϵ is fixed, and (ii) $\epsilon \to 0$ when $y_{1:n}$ is fixed, which we call the *large-sample* and *small-tolerance* asymptotics, respectively. The latter is a generalization of Proposition 2 in Bernton et al. [2019a] to soft ABC posteriors with kernels satisfying some regularity conditions. We also present an upper bound on the concentration rate of WABC when $n \to \infty$ and ϵ decreases sufficiently fast, which was proven by Bernton et al. [2019a], and argue how a similar bound for MMD-ABC can be obtained as a corollary.

4.1 Large-Sample Asymptotic

The following result establishes the convergence of the ABC posterior as we obtain more and more data while keeping the threshold fixed. Later in the section, we shall show that this convergence is also achieved with a sequence of thresholds that converges to some positive ϵ from above (see Cor. 4.3).

Proposition 4.1 (Large-sample asymptotic)**.** *Given a metric $\mathfrak{D}$ on $\mathcal{P}(\mathcal{Y})$, let $\mathfrak{D}(\hat{\mu}_{y_{1:n}}, \hat{\mu}_{\theta,z_{1:n}})$ be an estimator for $\mathfrak{D}(\mu_*, \mu_\theta)$ that is consistent for $\pi-$almost all $\theta \in \Theta$. Fix $\epsilon > 0$. Assume $\mathbb{P}(\mathfrak{D}(\mu_*, \mu_\theta) \leq \epsilon) > 0$ and $\mathbb{P}(\mathfrak{D}(\mu_*, \mu_\theta) = \epsilon) = 0$. Then for all $h \in L^1(\pi) \coloneqq \{h : \Theta \to \mathbb{R} : \ h \text{ is measurable and } \int_\Theta |h(\theta)|\pi(d\theta) < \infty\}$, we have*

$$\mathbb{E}[h(\theta)|\mathfrak{D}(\hat{\mu}_n, \hat{\mu}_{\theta,z_{1:n}}) \leq \epsilon] \to \mathbb{E}[h(\theta)|\mathfrak{D}(\mu_*, \mu_\theta) \leq \epsilon]. \tag{9}$$

In particular, $\pi^\epsilon_{y_{1:n}}$ converges strongly to $\pi(\cdot|\mathfrak{D}(\mu_*, \mu_\theta) \leq \epsilon)$, i.e. $\pi^\epsilon_{y_{1:n}}(A) \to \pi(A|\mathfrak{D}(\mu_*, \mu_\theta) \leq \epsilon)$ for all measurable $A \subset \Theta$.

This result implies that the ABC posterior for a fixed $\epsilon > 0$ does *not* converge to the true posterior, but only to a restricted one. This aligns with the analysis of the Exponential model in section 2.3.1.

When the distributions are continuous, the conditions that $\mathbb{P}(\mathfrak{D}(\mu_*, \mu_\theta) \leq \epsilon) > 0$ and $\mathbb{P}(\mathfrak{D}(\mu_*, \mu_\theta) = \epsilon) = 0$ are mild and hold for many standard metrics, e.g. Wasserstein, MMD and KL divergence. One caveat, however, is that $\mathfrak{D}(\hat{\mu}_{y_{1:n}}, \hat{\mu}_{z_{1:n}})$ needs to be consistent for $\mathfrak{D}(\mu_*, \mu_\theta)$. If the data discrepancy $\mathfrak{D}$ *metrizes* the space of distributions, one can prove that the distance between two convergent sequences of distributions tends to the distance between their limiting distributions. This is known as the continuity in the metric $\mathfrak{D}$; see D.2 for detailed discussions. Wasserstein distances obey this property, as it is known that $\mathcal{W}_p$ metrizes $\mathcal{P}_p(\mathcal{Y})$, the Wasserstein space of order p [Villani, 2009, Theorem 6.8]. For MMD, Theorem 12 of Simon-Gabriel and Schölkopf [2018] implies that, with a continuous, characteristic kernel such as the Gaussian and Laplacian kernels, MMD metrizes $\mathcal{P}(\mathcal{Y})$.

The proof of Prop. 4.1 is a direct application of Lemma S3.1 from Miller and Dunson [2019]. We present a more general version with a convergent sequence of random variables $(V_n)_{n\geq 1}$ instead of a single one V. An alternative proof, which replies on the Lévy's upward theorem, is given by Theorem 1 of Bai Jiang [2018].

Lemma 4.2 (Miller and Dunson [2019])**.** *Let $U, (U_n)_{n\geq 1}, V, (V_n)_{n\geq 1}, W \in \mathbb{R} \cup \{\infty\}$ be random variables such that $\mathbb{P}(U \leq V) > 0$, $\mathbb{P}(U = V) = 0$, $U_n \xrightarrow[n\to\infty]{a.s.} U$, $V_n \xrightarrow[n\to\infty]{a.s.} V$ and $\mathbb{E}[|W|] < \infty$. It follows that $\mathbb{E}[W|U_n \leq V_n] \to \mathbb{E}[W|U \leq V]$, as $n \to \infty$.*

Proof of Lemma 4.2. Since $\mathbb{P}(U = V) = 0$, $U_n \xrightarrow{\text{a.s.}} U$ and $V_n \xrightarrow{\text{a.s.}} V$, there exists a set of probability one on which $U \neq V$, $U_n \to U$ and $V_n \to V$. On this set, $\mathbb{1}(U_n \leq V_n) \to \mathbb{1}(U \leq V)$. Hence, $W\mathbb{1}(U_n \leq V_n) \xrightarrow{\text{a.s.}} W\mathbb{1}(U \leq V)$. Note also that $|\mathbb{1}(U \leq V)| \leq 1$, $|W\mathbb{1}(U \leq V)| \leq |W|$ and $\mathbb{E}[|W|] < \infty$ by assumption. Applying dominated convergence theorem [Bartle, 1995] yields,

$$\mathbb{E}[\mathbb{1}(U_n \leq V_n)] \to \mathbb{E}[\mathbb{1}(U \leq V)],$$
$$\mathbb{E}[W\mathbb{1}(U_n \leq V_n)] \to \mathbb{E}[W\mathbb{1}(U \leq V)].$$

By assumption, $\mathbb{P}(U \leq V) > 0$. Hence, $\mathbb{E}[\mathbb{1}(U_n \leq V_n)] = \mathbb{P}(U_n \leq V_n) > 0$ for n large enough, and

$$\mathbb{P}(W|U_n \leq V_n) = \frac{\mathbb{E}[W\mathbb{1}(U_n \leq V_n)]}{\mathbb{E}[\mathbb{1}(U_n) \leq V_n]} \to \frac{\mathbb{E}[W\mathbb{1}(U \leq V)]}{\mathbb{E}[\mathbb{1}(U) \leq V]} = \mathbb{P}(W|U \leq V).$$

■

Proof of Prop. 4.1. Apply Lemma 4.2 with $U_n = \mathfrak{D}(\hat{\mu}_{y_{1:n}}, \hat{\mu}_{\theta,z_{1:n}}), U = \mathfrak{D}(\mu_*, \mu_\theta), V_n = V = \epsilon$ and $W = h$. By assumption, $\mathbb{P}(U \leq \epsilon) > 0$, $\mathbb{P}(U = \epsilon) = 0$, $U_n \xrightarrow{\text{a.s.}} U$ as $n \to \infty$ and $\mathbb{E}[|W|] < \infty$. The result follows. ■

We remark that, although we have chosen a fixed threshold in the above proof, it can be easily relaxed to a monotonically decreasing sequence of thresholds $(\epsilon_n)_{n\geq 1}$ with limit $\epsilon > 0$.

Corollary 4.3 (Large-sample asymptotic with annealing thresholds)**.** *Under the same assumptions in Prop. 4.1, let $(\epsilon_n)_{n\geq 1}$ be a monotonically decreasing sequence of thresholds that converges to a positive limit ϵ. It follows that $\pi^{\epsilon_n}_{y_{1:n}}(A) \to \pi(A|\mathfrak{D}(\mu_*, \mu_\theta) \leq \epsilon)$ for all measurable $A \in \Theta$.*

Proof. Following the proof of Prop. 4.1 with $V_n = \epsilon_n$ and $V = \epsilon$, we have the desired result. ■

4.2 Small-Tolerance Asymptotic

We now present the asymptotic behaviour as ϵ decreases to 0 while $y_{1:n}$ is kept fixed. This result is similar to Proposition 2 of Bernton et al. [2019a] but is slightly more general — instead of restricting to rejection ABC, we consider the posterior resulted from soft ABC that uses a general kernel satisfying some mild conditions. We first present the result for rejection ABC with discrepancy $\mathfrak{D}$.

Proposition 4.4 (Proposition 2 of Bernton et al. [2019a])**.** *Let $\mu^{(n)}_\theta$ attain a continuous (in metric $\mathfrak{D}$) density $f^{(n)}_\theta$. Assume*

1. $\sup_{\theta\in\Theta\setminus\mathcal{N}_\Theta} f^{(n)}_\theta(y_{1:n}) < \infty$, *for some $\mathcal{N}_\Theta \subset \Theta$ such that $\pi(\mathcal{N}_\Theta) = 0$.*

2. *There exists $\bar{\epsilon} > 0$ such that $\sup_{\theta\in\Theta\setminus\mathcal{N}_\Theta} \sup_{z_{1:n}\in\mathcal{B}_{\bar{\epsilon}}(y_{1:n})} f^{(n)}_\theta(z_{1:n}) < \infty$, where $\mathcal{B}_{\bar{\epsilon}}(y_{1:n}) := \{z_{1:n} \in \mathcal{Y}^n : \mathfrak{D}(y_{1:n}, z_{1:n}) \leq \bar{\epsilon}\}$.*

Suppose also that $\mathfrak{D}$ is continuous in the sense that $\mathfrak{D}(y_{1:n}, z_{1:n}) \to \mathfrak{D}(y_{1:n}, z'_{1:n})$ whenever $z_{1:n} \to z'_{1:n}$ component-wise in metric ρ. If either

3.1. *$f^{(n)}_\theta$ is $n-$exchangeable, i.e. $f^{(n)}_\theta(y_{1:n}) = f^{(n)}_\theta(y_{\sigma(1:n)})$, for all permutations $\sigma \in S_n$, and $\mathfrak{D}(y_{1:n}, z_{1:n}) = 0$ if and only if $y_{1:n} = z_{\sigma(1:n)}$ for some $\sigma \in S_n$, or*

3.2. *$\mathfrak{D}(y_{1:n}, z_{1:n}) = 0$ if and only if $y_{1:n} = z_{1:n}$,*

then, for $y_{1:n}$ fixed, the ABC posterior $\pi^{\epsilon}_{y_{1:n}}$ converges strongly to the posterior $\pi(\cdot|y_{1:n})$.

A consequence of Prop. 4.4 is that $\pi^{\epsilon}_{y_{1:n}}$ converges to the true posterior as ϵ decreases to 0, when $y_{1:n}$ is kept fixed. It however does *not* give any information about the rate of convergence. Neither does it guarantees that the sequential sampling scheme of section 3.4 with adaptively selected thresholds will be successful in a reasonable time.

Conditions 1 and 2 in Prop. 4.4 assert that, for $\pi-$almost all $\theta \in \Theta$, $f^{(n)}_{\theta}$ is bounded at $y_{1:n}$ and in a neighbourhood around $y_{1:n}$. They may be hard to check in practice, since we typically do not have access to the functional form of the data generating process. They are, however, rather mild if we restrict Θ to a bounded set. Conditions 3.1 and 3.2 place assumptions on the ability of $\mathfrak{D}$ to distinguish between $y_{1:n}$ and $z_{1:n}$. The former applies to Wasserstein, while the latter holds for MMD and KL-divergence.

As also remarked previously, this result can be generalized to soft ABC posteriors. An additional assumption required is that the kernel satisfies the so-called concentration condition (definition 5). This is similar to but different from the concentration condition (Condition K) of Rubio and Johansen [2013] in that they assume a stronger condition where $\kappa_{\epsilon}(y_{1:n}, \cdot)$ is zero outside the ball $\mathcal{B}_{\epsilon}(y_{1:n}) = \{z_{1:n} \in \mathcal{Y}^n : \mathfrak{D}(y_{1:n}, z_{1:n}) \leq \epsilon\}$, whereas we allow it to decrease fast enough to 0.

We shall prove the more general result first then prove Prop. 4.4 as a corollary.

Definition 5 (Concentration condition)**.** A kernel $\kappa_{\epsilon} = \epsilon^{-1}\kappa(\cdot/\epsilon)$ with bandwidth $\epsilon > 0$ is said to satisfy the *concentration condition* with respect to some metric $\mathfrak{D}$ on the data space if

$$\sup_{z_{1:n} \in \mathcal{Y}^n \setminus \mathcal{B}_{\epsilon}(y_{1:n})} \kappa_{\epsilon}(\mathfrak{D}(y_{1:n}, z_{1:n})) = \sup_{z_{1:n} \in \mathcal{Y}^n \setminus \mathcal{B}_{\epsilon}(y_{1:n})} \kappa_{\epsilon}(\mathfrak{D}(z_{1:n}, y_{1:n})) \to 0,$$

as $\epsilon \to 0$.

Proposition 4.5 (Small-tolerance asymptotic)**.** *Let κ_{ϵ} be a kernel satisfying the concentration condition 5. Suppose the same assumptions in Prop. 4.4 hold. If, instead of assumption 2, we have*

4. $\sup_{\theta \in \Theta \setminus \mathcal{N}_{\Theta}} \sup_{z_{1:n} \in \mathcal{Y}^n} f^{(n)}_{\theta}(z_{1:n}) < \infty,$

then the soft ABC posterior $\pi^{\epsilon}_{y_{1:n}}$ with kernel κ_{ϵ} converges strongly to the posterior $\pi(\cdot|y_{1:n})$.

Proof of Prop. 4.5. Fix $y_{1:n}$ and let $\bar{\epsilon}$ be as in assumption 2. Pick $0 < \epsilon < \bar{\epsilon}$, and denote the normalized "quasi-likelihood" as

$$q^{\epsilon}(\theta) \coloneqq \frac{\int_{\mathcal{Y}^n} \kappa_{\epsilon}(\mathfrak{D}(y_{1:n}, z_{1:n})) f^{(n)}_{\theta}(z_{1:n}) dz_{1:n}}{\int_{\mathcal{Y}^n} \kappa_{\epsilon}(\mathfrak{D}(y_{1:n}, z'_{1:n})) dz'_{1:n}}$$

The intuition behind the definition of $q^{\epsilon}(\theta)$ is that it behaves as a likelihood function in the soft ABC posterior formula (Eq. 2). Write

$$\overline{K}_{\epsilon}(y_{1:n}, z_{1:n}) \coloneqq \frac{\kappa_{\epsilon}(\mathfrak{D}(y_{1:n}, z_{1:n}))}{\int_{\mathcal{Y}^n} \kappa_{\epsilon}(\mathfrak{D}(y_{1:n}, z'_{1:n})) dz'_{1:n}},$$

so that $q^{\epsilon}(\theta) = \int_{\mathcal{Y}^n} \overline{K}_{\epsilon}(y_{1:n}, z_{1:n}) f^{(n)}_{\theta}(z_{1:n}) dz_{1:n}$. Note that if a uniform kernel is used as in rejection ABC, then κ_{ϵ} is the indicator function on the ball $\mathcal{B}_{\epsilon}(y_{1:n})$, and $\overline{K}_{\epsilon}$ corresponds to the uniform density on $\mathcal{B}_{\epsilon}(y_{1:n})$. To proceed, we shall prove (i) $q^{\epsilon}(\theta) \xrightarrow{\text{a.s.}} f^{(n)}_{\theta}(y_{1:n})$ in θ, and (ii) q^{ϵ} is bounded almost surely, then we shall apply the dominated convergence theorem to conclude.

For (i), note first that $\mathcal{B}_\epsilon(y_{1:n})$ is compact by continuity of $\mathfrak{D}$. Now, $\forall \theta \in \Theta \backslash \mathcal{N}_\Theta$,

$$\left|q^\epsilon(\theta) - f_\theta^{(n)}(y_{1:n})\right| = \left|\int_{\mathcal{Y}^n} \overline{K}_\epsilon(y_{1:n}, z_{1:n})(f_\theta^{(n)}(z_{1:n}) - f_\theta^{(n)}(y_{1:n}))dz_{1:n}\right| \tag{10}$$

$$\leq \int_{\mathcal{B}_\epsilon(y_{1:n})} \overline{K}_\epsilon(y_{1:n}, z_{1:n}) \left|f_\theta^{(n)}(z_{1:n}) - f_\theta^{(n)}(y_{1:n})\right| dz_{1:n}$$

$$+ \int_{\mathcal{Y}^n \backslash \mathcal{B}_\epsilon(y_{1:n})} \overline{K}_\epsilon(y_{1:n}, z_{1:n}) \left|f_\theta^{(n)}(z_{1:n}) - f_\theta^{(n)}(y_{1:n})\right| dz_{1:n}, \tag{11}$$

where in Eq. 10 we used that $\overline{K}_\epsilon(y_{1:n}, \cdot)$ integrates to 1 on $\mathcal{Y}^n$. The first integral in Eq. 11 can be bounded as follows

$$\int_{\mathcal{B}_\epsilon(y_{1:n})} \overline{K}_\epsilon(y_{1:n}, z_{1:n}) \left|f_\epsilon^{(n)}(z_{1:n}) - f_\theta^{(n)}(y_{1:n})\right| dz_{1:n} \leq \sup_{z_{1:n} \in \mathcal{B}_\epsilon(y_{1:n})} \left|f_\theta^{(n)}(z_{1:n}) - f_\theta^{(n)}(y_{1:n})\right|$$

$$= \left|f_\theta^{(n)}(z_{1:n}^\epsilon) - f_\theta^{(n)}(y_{1:n})\right|,$$

for some $z_{1:n}^\epsilon \in \mathcal{B}_\epsilon(y_{1:n})$ by compactness of $\mathcal{B}_\epsilon(y_{1:n})$. The first inequality holds again because $\overline{K}_\epsilon(y_{1:n}, \cdot)$ is non-negative and integrates to 1 on $\mathcal{Y}^n$. By continuity of $f_\theta^{(n)}$, either $\cap_{\epsilon \in \mathbb{Q}} \mathcal{B}_\epsilon(y_{1:n}) = \{y_{\sigma(1:n)} : \sigma \in S_n\}$ under condition 3.1 in Prop. 4.4, or $\cap_{\epsilon \in \mathbb{Q}} \mathcal{B}_\epsilon(y_{1:n}) = \{y_{1:n}\}$ under 3.2. In both cases, $|f_\theta^{(n)}(z_{1:n}^\epsilon) - f_\theta^{(n)}(y_{1:n})| \to 0$ as $\epsilon \to 0$, by continuity of $f_\theta^{(n)}$.

For the second integral, recall that, by assumption, $f_\theta^{(n)}(z_{1:n}) \leq \sup_{\theta \in \Theta \backslash \mathcal{N}_\Theta} \sup_{z_{1:n} \in \mathcal{Y}^n} f_\theta^{(n)}(z_{1:n}) := C < \infty$ for any $z_{1:n} \in \mathcal{Y}^n$. Therefore,

$$\int_{\mathcal{Y}^n \backslash \mathcal{B}_\epsilon(y_{1:n})} \overline{K}_\epsilon(y_{1:n}, z_{1:n}) \left|f_\theta^{(n)}(z_{1:n}) - f_\theta^{(n)}(y_{1:n})\right| dz_{1:n} \leq 2C \int_{\mathcal{Y}^n} \overline{K}_\epsilon(y_{1:n}, z_{1:n}) \mathbb{1}(\mathfrak{D}(y_{1:n}, z_{1:n}) > \epsilon) dz_{1:n}.$$

Now, for any $z_{1:n}$, $\left|\overline{K}_\epsilon(y_{1:n}, z_{1:n}) \mathbb{1}(\mathfrak{D}(y_{1:n}, z_{1:n}) > \epsilon)\right|$ is bounded from above by $\overline{K}_\epsilon(y_{1:n}, z_{1:n})$, which is integrable over $\mathcal{Y}^n$. The concentration condition of κ_ϵ implies that $\kappa_\epsilon(\mathfrak{D}(y_{1:n}, z_{1:n})) \mathbb{1}(\mathfrak{D}(y_{1:n}, z_{1:n}) > \epsilon) \to 0$ pointwise in $z_{1:n}$, as $\epsilon \to 0$. Hence, $\overline{K}_\epsilon(y_{1:n}, z_{1:n}) \mathbb{1}(\mathfrak{D}(y_{1:n}, z_{1:n}) > \epsilon) \to 0$ pointwise as $\epsilon \to 0$. By dominated convergence theorem,

$$\lim_{\epsilon \to 0} \int_{\mathcal{Y}^n} \overline{K}_\epsilon(y_{1:n}, z_{1:n}) \mathbb{1}(\mathfrak{D}(y_{1:n}, z_{1:n}) > \epsilon) dz_{1:n} = \int_{\mathcal{Y}^n} \lim_{\epsilon \to 0} \overline{K}_\epsilon(y_{1:n}, z_{1:n}) \mathbb{1}(\mathfrak{D}(y_{1:n}, z_{1:n}) > \epsilon) dz_{1:n} = 0.$$

The second integral therefore converges to 0. We conclude from Eq. 11 that $q^\epsilon \xrightarrow{\text{a.s.}} f_\theta^{(n)}(y_{1:n})$. Now, $\forall \epsilon \leq \bar{\epsilon}$,

$$\sup_{\theta \in \mathcal{N}_\Theta} q^\epsilon(\theta) = \sup_{\theta \in \mathcal{N}_\Theta} \int_{\mathcal{Y}^n} \overline{K}_\epsilon(y_{1:n}, z_{1:n}) f_\theta^{(n)}(z_{1:n}) dz_{1:n} \leq \sup_{\theta \in \mathcal{N}_\Theta} \sup_{z_{1:n} \in \mathcal{B}_\epsilon(y_{1:n})} f_\theta^{(n)}(z_{1:n}) \leq C < \infty,$$

where the first inequality holds because $\int_{\mathcal{Y}^n} \overline{K}_\epsilon(y_{1:n}, z_{1:n}) dz_{1:n} = 1$. Hence, by dominated (or bounded) convergence theorem again, for any measurable $A \subset \Theta$,

$$\int_A q^\epsilon(\theta) \pi(d\theta) \to \int_A f_\theta^{(n)}(y_{1:n}) \pi(d\theta).$$

In particular, $\int_\Theta q^\epsilon(\theta') \pi(d\theta') > 0$ for ϵ small enough, and

$$\lim_{\epsilon \to 0} \int_A \pi_{y_{1:n}}^\epsilon(d\theta) = \frac{\lim_{\epsilon \to 0} \int_A q^\epsilon(\theta) \pi(d\theta)}{\lim_{\epsilon \to 0} \int_\Theta q^\epsilon(\theta') \pi(d\theta')} = \frac{\int_A f_\theta^{(n)}(y_{1:n}) \pi(d\theta)}{\int_\Theta f_\theta^{(n)}(y_{1:n}) \pi(d\theta)} = \int_A \pi(d\theta | y_{1:n}).$$

■

Proof of Prop. 4.4. We follow the same idea in the proof for Prop. 4.5, with a simplification that the second integral in Eq.11 is now zero, because $\kappa_\epsilon(\mathfrak{D}(y_{1:n}, z_{1:n})) = \mathbb{1}(\mathfrak{D}(y_{1:n}, z_{1:n}) \leq \epsilon)$ is supported only on $\mathcal{B}_\epsilon(y_{1:n})$. ■

4.3 Posterior Concentration Rates

Prop. 4.5 says that the ABC posterior concentrates on the true posterior as $\epsilon \to 0$ for fixed $y_{1:n}$, but gives no quantitative information on the concentration rate. Bernton et al. [2019a] provided a quantitative upper bound on the rate of convergence of the WABC posterior under mild assumptions on the data generating process and the prior. We summarize this result and its assumptions below, and argue how a similar bound on the rate of convergence of the MMD-ABC posterior can be obtained similarly .

More formally, we say a sequence of distributions $\pi_{y_{1:n}}$ on Θ depending on data $y_{1:n}$ is consistent at some θ_* if, for all $\delta > 0$, $\mathbb{E}[\pi_{y_{1:n}}(\{\theta \in \Theta : \rho_\Theta(\theta, \theta_*) > \delta\})] \to 0$ as $n \to \infty$, where the expectation is taken over $\mu_*^{(n)}$, the distribution of $y_{1:n}$. The rate of convergence of $\pi_{y_{1:n}}$ is said to be upper bounded by a sequence $(\delta_n)_{n\geq 1}$ if $\mathbb{E}[\pi_{y_{1:n}}(\{\theta \in \Theta : \rho_\Theta(\theta, \theta_*) > \delta_n\})] \to 0$. The fastest decaying sequence $(\delta_n)_{n\geq 1}$ is called the rate of convergence of $\pi_{y_{1:n}}$.

Proposition 3 in Bernton et al. [2019a] establishes an upper bound of the rate of convergence of the WABC posterior around $\theta_* \in \operatorname{argmin}_{\theta\in\Theta} \mathcal{W}_p(\mu_\theta, \mu_*)$ as $n \to \infty$ and as ϵ decreases to $\epsilon_* := \mathcal{W}_p(\mu_{\theta_*}, \mu_*)$. It replies on the following assumptions.

Assumption 1. *$\mathcal{W}_p(\hat{\mu}_{y_{1:n}}, \mu_*) \to 0$ in $\mathbb{P}$-probability as $n \to \infty$.*

Assumption 2. *For all $\epsilon > 0$, $\mu_\theta^{(n)}(\mathcal{W}_p(\mu_\theta, \hat{\mu}_{\theta,z_{1:n}}) > \epsilon) \leq c(\theta) f_n(\epsilon)$, for some sequence of functions $f_n(\epsilon)$ that are strictly decreasing in ϵ for fixed n and $f_n(\epsilon) \to 0$ as $n \to \infty$ for fixed ϵ, and $c : \Theta \to \mathbb{R}^+$ is π-integrable such that there exists some $c_0, \delta_0 > 0$ satisfying*

$$\sup_{\{\theta:\mathcal{W}_p(\mu_\theta,\mu_*)\leq\delta_0+\epsilon_*\}} c(\theta) \leq c_0.$$

Assumption 3. *There exists some $L, c_\pi > 0$ such that for ϵ small enough,*

$$\pi(\{\theta \in \Theta : \mathcal{W}_p(\mu_\theta, \mu_*) \leq \epsilon + \epsilon_*\}) \geq c_\pi \epsilon^L.$$

Assumption 1 asserts that the observed data $y_{1:n}$ form an empirical distribution $\hat{\mu}_{y_{1:n}}$ that converges to the true likelihood μ_* as $n \to \infty$ in the discrepancy metric, in this case, the Wasserstein distances. Assumption 2 places a condition on how fast $\hat{\mu}_{\theta,z_{1:n}}$ concentrates around μ_θ in $\mathcal{W}_p$ as $n \to \infty$, and assumption 3 asserts that the prior puts enough mass on those $\theta \in \Theta$ that yields μ_θ close to μ_*. The main result is the following [Bernton et al., 2019a, Proposition 3]:

Proposition 4.6. *Under assumptions $1 - 3$, let $(\epsilon_n)_{n\geq 0}$ be a sequence of positive numbers such that $\epsilon_n \to 0$, $f_n(\epsilon_n) \to 0$ and $\mathbb{P}(\mathcal{W}_p(\hat{\mu}_{y_{1:n}}, \mu_*) \leq \epsilon_n) \to 1$, as $n \to \infty$. Then there exists $C \in (0, \infty)$ such that, for any $R \in (0, \infty)$, the WABC posterior with threshold $\epsilon_n + \epsilon_*$ satisfies*

$$\pi_{y_{1:n}}^{\epsilon_n+\epsilon_*}(\{\theta \in \Theta : \mathcal{W}_p(\mu_\theta, \mu_*) > \epsilon_* + 4\epsilon_n/3 + f_n^{-1}(\epsilon_n^L/R)\}) \leq \frac{C}{R},$$

with $\mathbb{P}$-probability going to 1 as $n \to \infty$.

Before presenting its proof, we make a few remarks about the implications of Prop. 4.6. Firstly, the inverse $f_n^{-1}(\epsilon)$ is well-defined due to the strict monotonicity of $f_n(\epsilon)$. If we can find $f_n(\epsilon)$ such that $f_n^{-1}(\epsilon_n^L/R) \to 0$ for all $R > 0$, then Prop. 4.6 implies that, with probability going to 1, the WABC posterior probability of the set of θ that yields μ_θ far from the "best" μ_{θ_*} is arbitrarily small, where "best" is in the sense that μ_{θ_*} is the distribution in the model that is closest to the true data generating process μ_* in terms of $\mathcal{W}_p$. The exact bound will depend on the $c(\theta)$ and f_n in assumption 2.

To find an upper bound on the concentration rate of WABC, Bernton et al. [2019a] make the following two additional assumptions. Assumption 4 says that $\mathcal{W}_p$ *well-separates* θ_* in the parameter space, while assumption 5 can be viewed as a "Hölder continuity" condition for ρ_Θ in the discrepancy metric $\mathcal{W}_p$.

Assumption 4. $\theta_* = \operatorname{argmin}_{\theta\in\Theta} \mathcal{W}_p(\mu_\theta, \mu_*)$ *exists, and is well-separated. That is,* $\forall \delta > 0$*, there exists* $\delta' > 0$ *such that*

$$\inf_{\theta\in\Theta:\rho_\Theta(\theta,\theta_*)>\delta} \mathcal{W}_p(\mu_\theta, \mu_*) > \mathcal{W}_p(\mu_{\theta_*}, \mu_*) + \delta' = \epsilon_* + \delta'.$$

Assumption 5. *The parameters* θ *are identifiable, and there exists* $K, \alpha > 0$ *and an open neighbourhood* $U \subset \Theta$ *containing* θ_* *such that,* $\forall \theta \in U$*,* $\rho_\Theta(\theta, \theta_*) \leq K(\mathcal{W}_p(\mu_\theta, \mu_*) - \epsilon_*)^\alpha$.

Corollary 4.7. *Under assumptions* $1-5$*, let* $(\epsilon_n)_{n\geq 1}$ *be a sequence of positive numbers such that* $\epsilon_n \to 0$*,* $f_n(\epsilon_n) \to 0$*,* $f_n^{-1}(\epsilon_n^L) \to 0$ *and* $\mathbb{P}(\mathcal{W}_p(\hat{\mu}_{y_{1:n}}, \mu_*) \leq \epsilon_n) \to 1$*, as* $n \to \infty$*. Then there exists* $C \in (0, \infty)$ *such that, for any* $R \in (0, \infty)$*, the WABC posterior with threshold* $\epsilon_n + \epsilon_*$ *satisfies*

$$\pi_{y_{1:n}}^{\epsilon_n+\epsilon_*}(\{\theta \in \Theta : \rho_\Theta(\theta, \theta_*) > K(4\epsilon_n/3 + f_n^{-1}(\epsilon_n^L/R))^\alpha\}) \leq \frac{C}{R},$$

with $\mathbb{P}$*-probability going to 1 as* $n \to \infty$.

Proofs of Prop. 4.6 and Cor. 4.7 can be found in appendix D. As remarked in Bernton et al. [2019a], assumptions $1-5$ seem technical at first sight and can be difficult to verify in practice. We now discuss their applicability to two specific metrics — the 1-Wasserstein distance and the MMD — when the observed data are i.i.d.

4.4 WABC with Independent and Identically Distributed Data

The case of WABC with i.i.d. data was also discussed by Bernton et al. [2019a]. We shall follow their approach to verify assumptions 1 and 2. The other ones can be check in the same way as for MMD, thus are deferred to section 4.5. Throughout, let $p \in [1, \infty)$ and $\mathcal{P}_p(\mathcal{Y})$ be the associated Wasserstein space of order p.

When $y_i \overset{iid.}{\sim} \mu_*$, assumption 1 essentially follows from the fact that $\mathcal{W}_1$ metrizes the Wasserstein space $\mathcal{P}_p(\mathcal{Y})$ (see Appendix D.2). More specifically, Theorem 3 in Varadarajan [1958] shows the existence of a set E_1 with $\mathbb{P}(E_1) = 1$ on which $\hat{\mu}_{y_{1:n}}$ converges to μ_* in distribution. Moreover, by the strong law of large numbers, there exists $\mathbb{P}$-a.s. E_2 and $y_0 \in \mathcal{Y}$ such that $\int_{\mathcal{Y}} \rho(y_0, z)^p \hat{\mu}_{y_{1:n}}(w)(dz) \to \int_{\mathcal{Y}} \rho(y_0, z)^p \mu_*(dz)$ for all $w \in E_2$. It follows that $\hat{\mu}_{y_{1:n}}$ *converges weakly to* μ_* on $E_1 \cap E_2$ by the definition of weak convergence (see definition 9 in Appendix). Now, Theorem 6.8 of Villani [2009] shows that weak convergence implies $\mathcal{W}_p(\hat{\mu}_{y_{1:n}}, \mu_*) \to 0$ in $\mathcal{P}_p(\mathcal{Y})$ (metrizable property of $\mathcal{W}_1$). Hence, assumption 1 holds on $E_1 \cap E_2$.

For assumption 2, by Theorem 2 in Fournier and Guillin [2015], when $\epsilon < 1$, there exists $c(\theta), C(\theta) > 0$ such that $\mu_\theta^{(n)}(\mathcal{W}_p(\mu_\theta, \hat{\mu}_{\theta,n}) > \epsilon) \leq C(\theta) \exp(-c(\theta) n \epsilon^k)$, where $k = \max(d_y, 2p)$. To proceed, we assume as in Bernton et al. [2019a] that $c(\theta), C(\theta)$ can be replace by constants $c, C > 0$ and define $f_n : \epsilon \mapsto \exp(-cn\epsilon^k)$ so that Assumption 2 holds. Assuming the rate of convergence of $\hat{\mu}_{y_{1:n}}$ to μ_* is no slower than that of $\hat{\mu}_{\theta,z_{1:n}}$ to μ_θ, the condition on ϵ_n in Prop. 4.6 holds with $\epsilon_n := c_\epsilon((\log n)/n)^{1/k}$ for some $c_\epsilon > 0$. Indeed, note that

1. $\epsilon_n \to 0$,
2. $f_n(\epsilon_n) = \exp(-cn\frac{c_\epsilon^k \log n}{n}) = n^{-cc_\epsilon^k} \to 0$, and
3. For any $R > 0$,

$$f_n^{-1}(\epsilon^L/R) = \left[-\frac{1}{cn}\log\left(\frac{c_\epsilon^L}{R}\left(\frac{\log n}{n}\right)^{L/k}\right)\right]^{1/k} = \left[-\frac{1}{cn}\log\left(\frac{c_\epsilon^L}{R}\right) - \frac{L}{kc}\frac{\log(\log n)}{n} + \frac{L}{kc}\frac{\log n}{n}\right]^{1/k} \to 0.$$

For large n, the concentration rate in Prop. 4.6 and Cor. 4.7, $K(4\epsilon_n/3 + f_n^{-1}(\epsilon_n^L/R))$, behaves as $(\log n/n)^{\alpha/k}$. We remark that the dimension d_y comes in k and this bound worsens rapidly as d_y increases — an example of the curse of dimensionality.

4.5 MMD-ABC with Independent and Identically Distributed Data

We show a similar bound on the concentration rate for the K2-ABC posterior. In addition to assumptions 1−5, the proofs of Prop. 4.6 and Cor. 4.7 only require that the data discrepancy (in that case $\mathcal{W}_p$) satisfies the triangular inequality and symmetry; see D.1 for more detailed discussions. A direct consequence is that any other symmetric metric satisfying assumptions $1-5$ and the triangular inequality yields the same upper bound. Recall from section 3.2.2 that MMD with a characteristic kernel is a metric. Hence, it remains to check whether assumptions 1, 2, 4 and 5 also hold for MMD (assumption 3 relies on the user-specified prior).

For assumption 1, Theorem 12 of Simon-Gabriel and Schölkopf [2018] shows that, provided the kernel is continuous and characteristic (e.g. the Gaussian and Laplacian kernels), MMD metrizes $\mathcal{P}(\mathcal{Y})$. That is, a sequence of probability measures converges in MMD if and only if it converges weakly (see definition 8 in Appendix). In particular, by the same argument in the case of $\mathcal{W}_p$, assumption 1 holds on some $E_1 \cap E_2$ of $\mathbb{P}$-probability one.

For assumption 2, we begin with the same choice of $f_n : \epsilon \mapsto \exp(-cn\epsilon^k)$ as in the case of WABC, so that $\mu_\theta^{(n)}(\mathcal{W}_p(\mu_\theta, \hat{\mu}_{\theta,z_{1:n}}) > \epsilon) \leq C\exp(-cn\epsilon^k)$. Recall that MMD lower-bounds $\mathcal{W}_1$. Hence,

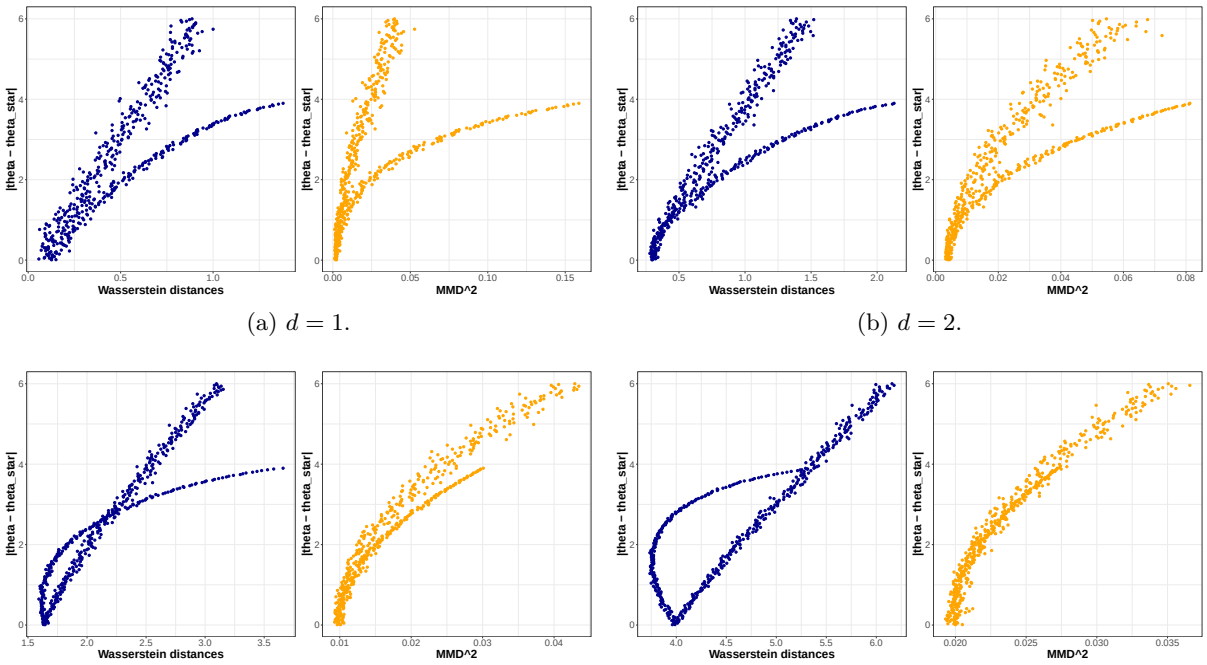
$$\mu_\theta^{(n)}(\mathrm{MMD}(\mu_\theta, \hat{\mu}_{\theta,z_{1:n}}) > \epsilon) \leq \mu_\theta^{(n)}(\mathcal{W}_1(\mu_\theta, \hat{\mu}_{\theta,z_{1:n}}) > \epsilon) \leq C\exp(-cn\epsilon^k),$$

and assumption 2 holds with the same choice of f_n.

Assumption 4 holds under further assumptions on the continuity of the metric in terms of ρ_Θ, as discussed in the supplementary material of Bernton et al. [2019a]. We generalize their result to give sufficient conditions under which assumption 4 holds for an arbitrary metric $\mathfrak{D}$ (see appendix D.1). Both the Wasserstein distances and MMD satisfy these conditions.

To check assumption 5, we suggest numerical evidence to be provided in a case-specific way. As a concrete example, we consider the same setup in section 3.5, where the model is $\mathcal{N}(0, \theta I_d)$ and the parameter is $\theta \in \mathbb{R}^+$. The metric on the parameter space is chosen to be the Euclidean norm, i.e. $\rho_\Theta(\theta_1, \theta_2) = \|\theta_1 - \theta_2\|$. For each synthetic data $z_{1:n}$ generated, $\mathrm{MMD}(\hat{\mu}_{y_{1:n}}, \hat{\mu}_{z_{1:n}})$ and $\mathcal{W}_1(\hat{\mu}_{y_{1:n}}, \hat{\mu}_{z_{1:n}})$ are computed and plotted against $\|\theta - \theta_*\|$. This is repeated for dimensions $d = 1, 2, 5, 10$, as shown in Fig. 5. The curves exhibit a linear or sublinear shape, which implies that one can find constants k and α for which the inequality in assumption 5 holds for both MMD and $\mathcal{W}_1$.

Finally, if we assume that the rate of convergence of $\hat{\mu}_n$ to μ_* is at least as fast as that of $\hat{\mu}_{\theta,z_{1:n}}$ to μ_θ, then the condition on ϵ_n in Prop. 4.6 holds with $\epsilon_n = c_\epsilon((\log n)/n)^{1/k}$ for some $c_\epsilon > 0$, by the same reasoning as in WABC with i.i.d. data. We therefore have the same bound on rate of convergence in Prop. 4.6 and Cor. 4.7 for MMD.



(a) $d = 1$. (b) $d = 2$.

(c) $d = 5$. (d) $d = 10$.

Figure 5: $\mathcal{W}_1$ and MMD^2 between empirical distributions formed by observed and synthetic data generated from $\mathcal{N}(0, \theta I_d)$ with different θ, plotted against $\rho_\Theta(\theta, \theta_*)$. The observed data is drawn with $\theta_* = 4$. Synthetic data $z_{1:n}$ are drawn with θ equi-spaced on $[0.1, 10]$.

5 Experiments

We run experiments on four benchmark models using WABC, MMD-ABC and KL-ABC. As a baseline, we also include the vanilla rejection ABC with case-specific summary statistics and the Euclidean norm as the data discrepancy, which we call Euclidean ABC.

For each method, we use the SMC framework introduced in section 3.4 to tune the thresholds ϵ. Unless otherwise mentioned, we use a fixed budget of 10^5 model simulations and 2048 particles in the SMC sampler. 1024 posterior samples yielding the smallest distances are kept for the posterior plots.

All experiments were based on R [R Core Team, 2019] and an Intel Core i5 (1.3 GHz). For the WABC and the SMC algorithm, we relied on the package `winference`[2] of Bernton et al. [2019a]. We have also built upon their source code for the `cpp` implementation of MMD. In particular, a Gaussian kernel is used for the embedding and the bandwidth is set to be the median of $\{\|y_i - y_j\|_1 : i, j = 1, \ldots, n\}$, as suggested in Park et al. [2016] (see appendix A). The `KLx.divergence` function from the `FNN` package [Beygelzimer et al., 2019] was used to estimate the KL divergence. Code for reproducing all experiments is available at `https://github.com/XingLLiu/approximate-bayesian-computation`.

5.1 Bivariate Gaussian Mixture Model

We begin with a bivariate Gaussian mixture model. 500 i.i.d. data are drawn from $y_i|u_i = 0 \sim \mathcal{N}((\mu_{01}, \mu_{02})^\intercal, \Sigma_0)$ and $y_i|u_i = 1 \sim \mathcal{N}((\mu_{11}, \mu_{12})^\intercal, \Sigma_1)$, where $u_i \sim \text{Bernoulli}(p)$, and the known co-

[2] `winference` codebase: `https://github.com/pierrejacob/winference`.

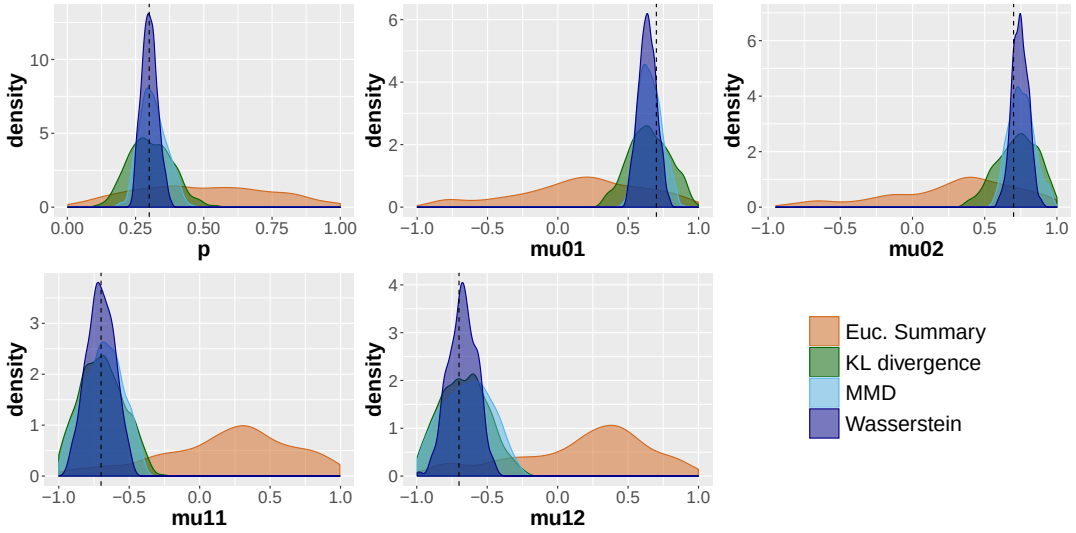


Figure 6: Posteriors of the five parameters of the bivariate Gaussian mixture model. The true parameters $p = 0.3, \mu_{01} = \mu_{02} = 0.7, \mu_{11} = \mu_{12} = -0.7$ are shown by the black dotted lines.

variance matrices are

$$\Sigma_0 = \begin{pmatrix} 0.5 & -0.3 \\ -0.3 & 0.5 \end{pmatrix}, \qquad \Sigma_1 = \begin{pmatrix} 0.25 & 0 \\ 0 & 0.25 \end{pmatrix}.$$

The parameters of interest $\theta = (p, \mu_{01}, \mu_{02}, \mu_{11}, \mu_{12})^\intercal$ consist of the mixture ratio and the means. We choose $\theta_* = (0.3, 0.7, 0.7, -0.7, -0.7)^\intercal$ to be the true parameters and assign independent priors $p \sim \text{Uniform}(0, 1)$ and $\mu_{jk} \sim \text{Uniform}(-1, 1)$, for $j = 0, 1$ and $k = 1, 2$. For the Euclidean ABC, we use a 5-dimensional summary statistic consisting of the two marginal sample means, the two marginal sample variances and the sample covariance. Results are shown in Fig 6.

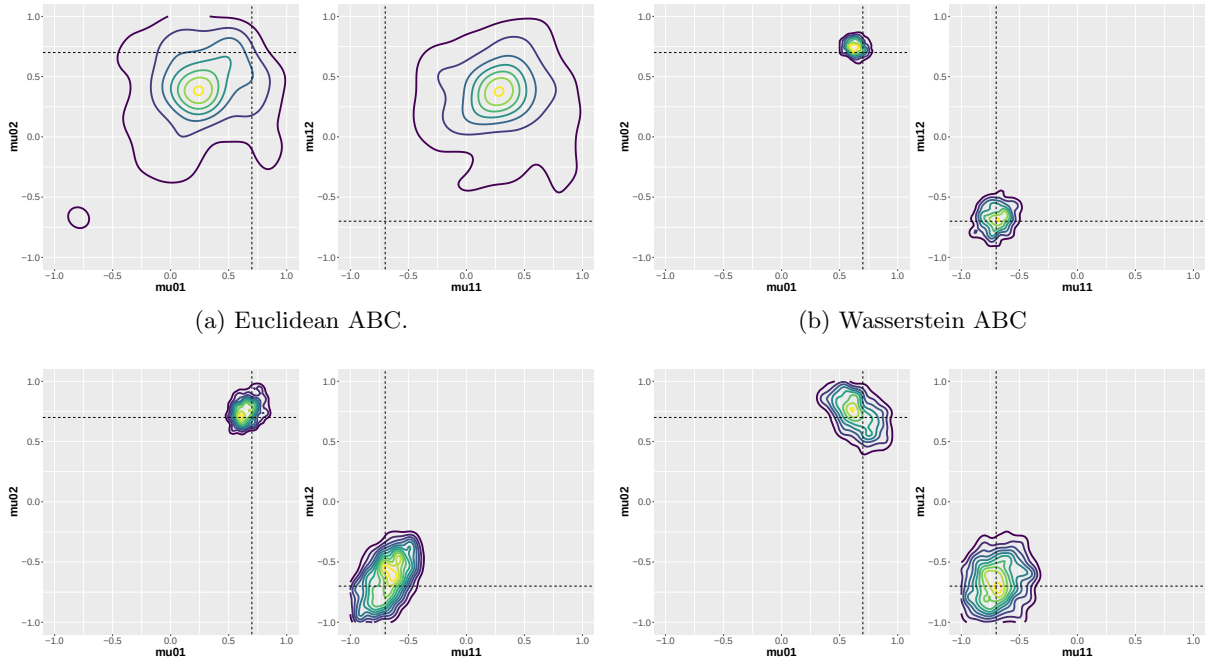
Using a total budget of 10^5 simulations, all except the Euclidean ABC successfully picked up the true parameters and exhibit a single mode (see Fig. 7). This is not surprising given the simple set-up of this problem. In particular, Wasserstein ABC had the best concentration for all parameters, followed by MMD-ABC, which yields the second best estimation. However, in this multivariate example, evaluating the Wasserstein distance ($\mathcal{O}(n^3)$) and MMD ($\mathcal{O}(n^2)$) is much more expansive than KL-divergence ($\mathcal{O}(n \log n)$). On an Intel Core i5 (1.3 GHz) and over 1000 repetitions, the average run time per call was $0.433s$ for the Wasserstein and $0.0330s$ for the MMD, whereas one evaluation of the KL-divergence only took $0.00240s$. One call of the summary statistics and the Euclidean discrepancy combined was order-of-magnitude faster ($9.36 \times 10^{-5}s$).

5.2 Univariate g-and-k Distribution

The univariate g-and-k distribution is an example of quantile distributions, which is a class of flexible distributions defined in terms of its inverse CDF or quantile function [Drovandi et al., 2011]. It was first proposed by Rayner and Macgillivray [2002] and has been widely studied in the context of ABC [Prangle, 2017; Bernton et al., 2019a]. Its quantile function is defined through five parameters as

$$F^{-1}(r) = a + b\left(1 + c\frac{1 - \exp(-gz(r))}{1 + \exp(-gz(r))}\right)(1 + z(r)^2)^k z(r),$$

for $r \in (0, 1)$, where z refers to the r-th quantile of $\mathcal{N}(0, 1)$ and $c = 0.8$ by convention (see Rayner and Macgillivray [2002] for a justification). The other 4 parameters $\theta = (a, b, g, k)^\intercal$ controls its location,



(a) Euclidean ABC.

(b) Wasserstein ABC

(c) MMD-ABC.

(d) KL-ABC.

Figure 7: Contours of the estimated posterior densities for each method in the bivariate Gaussian mixture model. The black dotted lines mark the true mean values $(\mu_{01}, \mu_{02}) = (0.7, 0.7), (\mu_{11}, \mu_{12}) = (-0.7, -0.7)$. Except the Euclidean ABC, all methods managed to identify the parameters with a high quality.

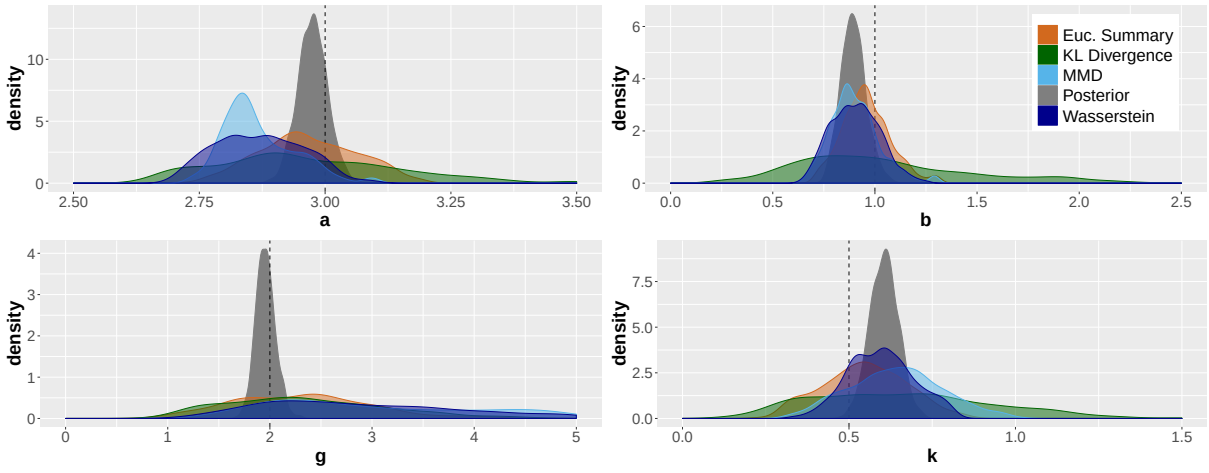


Figure 8: Posteriors of the four parameters in the univariate g-and-k model. MH approximation of the true posterior in shown by the grey area. The true parameters $\theta_* = (3, 1, 2, 0.5)^\intercal$ are shown by the black dotted lines.

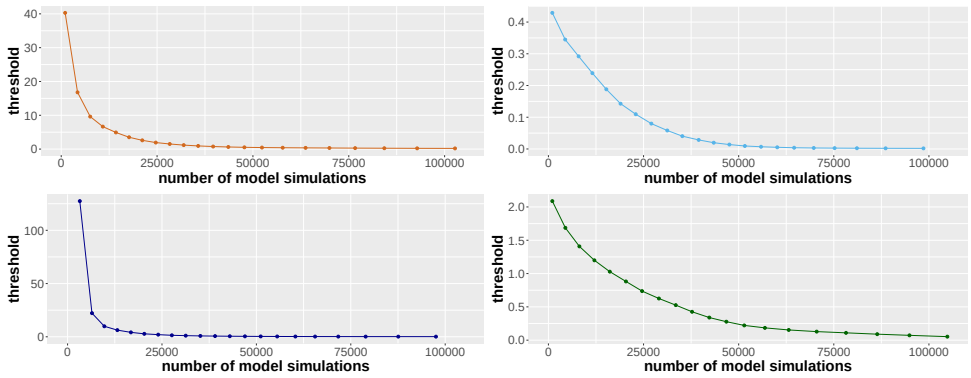


Figure 9: Sequentially selected thresholds for the g-and-k model using a fixed budget of 10^5 simulations. Top left: Euclidean ABC; top right: MMD-ABC; bottom left: WABC; bottom right: KL-ABC. All methods converged within $6,000$ model simulations, with WABC showing the fastest convergence.

scale, skewness and kurtosis, respectively. Simulating from g-and-k distributions is straightforward with the inversion method: one first simulates a standard normal random variate $z(r)$, then substitutes it into $F^{-1}(r)$. Unfortunately, its density function has no closed form. Numerical estimation of the likelihood is possible through finite difference, but this is very costly.

We choose the standard setting with $n = 1000, \theta_* = (3, 1, 2, 0.5)^\intercal$ and a uniform prior on $[0, 10]^4$. This set-up was also studied in Bai Jiang [2018]; Sisson et al. [2018]; Prangle [2017]. To estimate the true posterior, we run four Metropolis-Hastings chains [Hastings, 1970], discard the initial $50,000$ ones as burnin and keep once every 10 samples in the rest to yield 1024 samples. The resulting posterior densities are shown in Fig. 8.

For the Euclidean ABC, the summary statistics is chosen to be the four statistics recommended by Sisson et al. [2018]:

$$s(y_{1:n}) = (E_4, E_6 - E_2, (E_6 + E_2 - 2E_4)/(E_6 - E_2), (E_7 - E_5 + E_3 - E1)/(E_6 - E_2))^\intercal,$$

where $E_1 \leq \ldots \leq E_8$ are the octiles of $y_{1:n}$. Because of this clever choice of summary statistics, Euclidean ABC showed superior performance over the others. We remark that choosing good summary statistics itself can be a non-trivial task in more complicated examples (e.g. multivariate g-and-k distributions). Nonetheless, WABC beat Euclidean ABC in estimating k and MMD-ABC was the most superior in estimating b. WABC also exhibited a faster convergence, as shown by the trace plot of the adaptively selected thresholds (Fig. 9). WABC, however, failed to concentrate on the true parameters for a and g. The estimation of KL-ABC was rather poor. None of the methods successfully identified a and g within the budget.

5.3 M/G/1 Queuing Model

The M/G/1 queuing model, introduced by Heggland and Frigessi [2004], has been a popular example of non-i.i.d. data in the ABC literature. It models a service system with a single server where the service times are uniformly distributed on interval $[\theta_1, \theta_2]$ and the inter-arrival times w_i of customers follow an Exponential distribution with rate θ_3. The inter-departure times $(y_i)_{i\geq 1}$ are modelled through the process $y_i = \max(0, \sum_{j=1}^{i} w_j - \sum_{j=1}^{i-1} y_j)$. Bayesian inference on $\theta = (\theta_1, \theta_2, \theta_3)^\intercal$ is done by assuming that the inter-departure times are the only quantities observed. Since the inter-arrival times are unobserved, numerical evaluations of the likelihood is expansive [Blum and François, 2010].

We consider the set-up in Blum and François [2010]: we generate $n = 50$ samples $y_{1:n}$ with $\theta_* = (1, 4, 0.2)^\intercal$. For the priors, we assign independent Uniform(0, 10) to θ_1 and $(\theta_2 - \theta_1)$, and Uniform$(0, 1/3)$

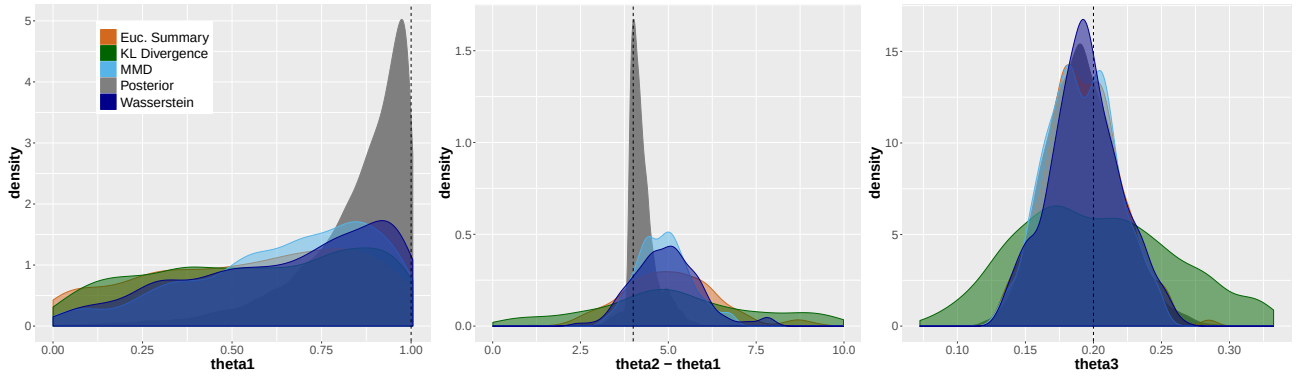


Figure 10: Posteriors of the three parameters in the M/G/1 queuing model. PMMH approximation of the true posterior in shown by the grey area. The true parameters $\theta_1 = 1, \theta_2 - \theta_1 = 3$ and $\theta_3 = 0.2$ are shown by the black dotted lines.

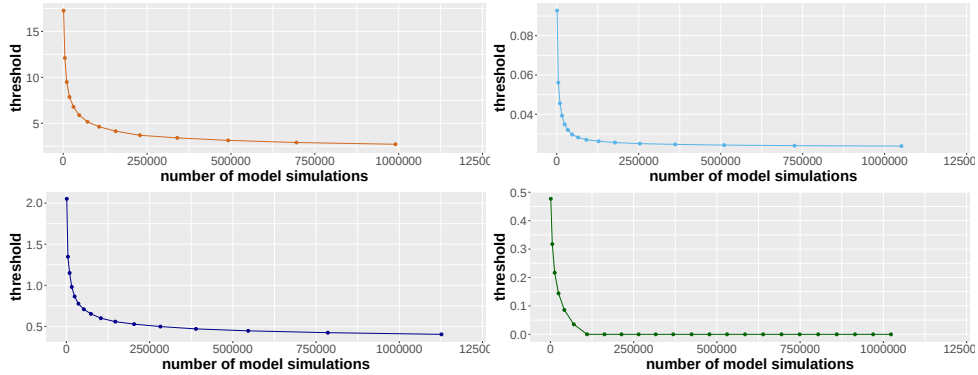


Figure 11: Sequentially selected thresholds for the queuing model. Top left: Euclidean ABC; top right: MMD-ABC; bottom left: WABC; bottom right: KL-ABC. KL-ABC exhibited slightly slower convergence than the others.

to θ_3. Note that, upon observing $y_{1:n}$, θ_1 cannot be larger than $\min_{i=1,\dots,n} y_i$. This can be encoded by placing a $\text{Uniform}(0, 10 \wedge y_{(1)})$ prior on θ_1, where $y_{(1)}$ denotes the minimum of y_i and $a \wedge b := \min(a, b)$.

To estimate the true posteriors, we follow the implementation in Bernton et al. [2019a], which used a *particle marginal Metropolis-Hastings* (PMMH) algorithm of Andrieu et al. [2010] with 4096 particles and 100,000 iterations. The estimated posteriors together with the ABC posteriors are shown in Fig. 10. The summary statistic for the Euclidean ABC was chosen to be a 20-dimensional vector consisting of the 5%-quantiles. We remark that, within a budget of 10^6 simulations, all methods successfully identified $\theta_2 - \theta_1$ and θ_3, but struggled to produce decent estimation for θ_1. WABC and MMD-ABC outperformed the other two methods in all cases, with the former showing slightly better approximations for θ_1 and θ_3. KL-ABC, despite being able to pick up the true values, exhibited slower and unsatisfactory concentration.

5.4 Ecological Dynamic Systems

As an example of real-life applications of ABC on ecological dynamic systems. We consider a data set studied by Wood [2010] and Fearnhead and Prangle [2012] that contains $T = 180$ observations on adult blowfly populations over time. The population N_{t+1} at time $t+1$ is modelled by a discretized differential

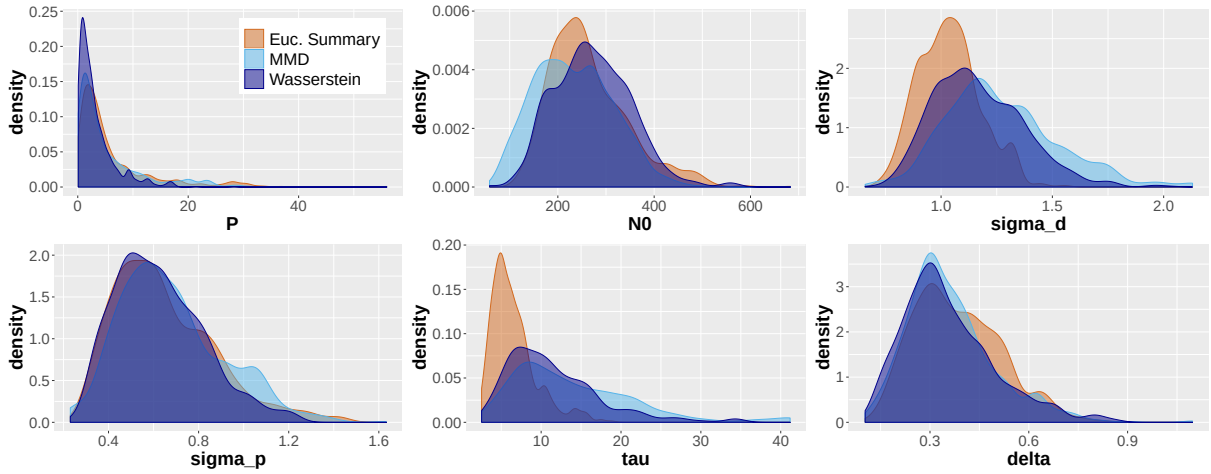


Figure 12: Posteriors of the six parameters $\theta = (P, N_0, \sigma_d, \sigma_p, \tau, \delta)^\intercal$ in the blowfly model with the real data set, where the true parameters are unknown.

equation:

$$N_{t+1} = PN_{t-\tau} \exp\left(-\frac{N_{t-\tau}}{N_0}\right) e_t + N_t \exp\left(-\delta\epsilon_t\right),$$

where N_{t+1} is determined by the time-lagged observations N_t and $N_{t-\tau}$ and independent Gamma noises $e_t \sim \text{Gamma}(1/\sigma_p^2, \sigma_p^2)$ and $\epsilon_t \sim \text{Gamma}(1/\sigma_d^2, \sigma_d^2)$. We are interested in performing inference on parameters $\theta = (P, N_0, \sigma_d, \sigma_p, \tau, \delta)^\intercal$ upon observing $N_{1:T}$. As remarked in Wood [2010], this model is *near-chaotic* in the sense that a small change in the parameters would drastically vary the realizations generated. Evaluation of the likelihood function is therefore extremely costly.

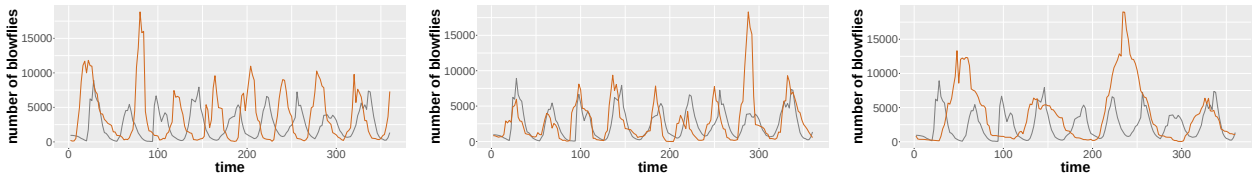
Following Park et al. [2016], we place independent Gaussian priors on the logarithms of the parameters:

$$\log(P) \sim \mathcal{N}(2, 4), \quad \log(N_0) \sim \mathcal{N}(5, 0.25), \quad \log(\sigma_d) \sim \mathcal{N}(-0.5, 1),$$
$$\log(\sigma_p) \sim \mathcal{N}(-0.5, 1), \quad \log(\tau) \sim \mathcal{N}(2, 1), \quad \log(\delta) \sim \mathcal{N}(-1, 0.16).$$

Eight summary statistics are chosen for the Euclidean ABC: 25% quantiles of $N_{1:T}/1000$ (4 statistics) and of its first order differences (4 statistics). Resulting posteriors with a budget of 5×10^6 simulations are plotted in Fig. 12. The results for KL-divergence are not shown for illustration purpose as they did not concentrate within the limited budget.

We can see from Fig. 12 that all three methods concentrated on similar values for all parameters. Euclidean ABC showed better concentration for N_0, σ_d and τ, whereas WABC and MMD-ABC exhibited better concentration for P and δ. We remark that since the data are **not** i.i.d., the discussions in section 4.4 and 4.5 to verify the assumptions for concentration rates no longer apply. Using the posterior means of these parameters from each method, we can draw one set of realizations from the model to visualize the quality of estimation, as shown in Fig. 13.

Except the realization produced by MMD-ABC, which showed apparent deviation from the real data (grey curves), it is hard to tell whether Euclidean ABC is better than WABC solely from their realizations. Instead, we can repeat the same experiment on a model with known parameters and compare their concentration to argue which method is superior. The parameters are chosen to be the posterior means reported in the code given by Park et al. [2016]: $P = 29, \delta = 0.2, N_0 = 260, \sigma_d = 0.6, \sigma_p = 0.3$ and $\tau = 7$. The posterior densities are shown in Fig. 15. We can see that, in fact, Wasserstein had competitive performance for most parameters apart from σ_d and τ, for which it is outperformed by the



(a) Euclidean ABC. (b) Wasserstein ABC (c) MMD-ABC.

Figure 13: Realizations from the blowfly model with parameters equal to the posterior means of Euclidean ABC, Wasserstein ABC and MMD-ABC. The blowfly population (y-axis) is plotted against time (x-axis). Grey lines are the real data reported in Wood [2010]. The realization produced with Wasserstein ABC resembles the real data most.

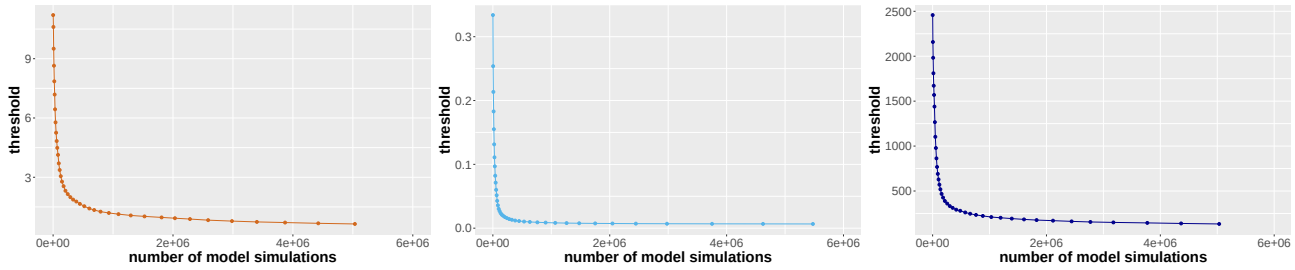


Figure 14: Sequentially selected thresholds for the blowfly model. Top left: Euclidean ABC; top right: MMD-ABC; bottom left: WABC; bottom right: KL-ABC. All methods converge within 4×10^5 model simulations.

Euclidean ABC. In the case of limited computational budget, we argue that either the Euclidean ABC or Wasserstein ABC could be used for this problem. However, if one affords more computational power, we suggest to run both methods and use a combination of the results by choosing the posterior means of Euclidean ABC for σ_d and τ and of Wasserstein ABC for the rest.

For the computational times, on average of 1000 runs, each call of the Wasserstein distance took $1.90 \times 10^{-4}s$, while each evaluation of the summary statistic and the Euclidean discrepancy took $7.83 \times 10^{-4}s$. The cost per call of the MMD was remarkably larger ($2.91 \times 10^{-3}s$). We emphasize that computing Wasserstein distances in one dimension is resolved to a sorting problem (see section 3.1.2), which is likely to have caused the difference in the run times between Wasserstein and MMD.

6 Discussion

How to choose a suitable summary statistic and a discrepancy metric lies at the heart of ABC problems. We have reviewed three discrepancy metrics on the space of probability measures — Wassertsein distances, maximum mean discrepancy and KL-divergence — which are popular in optimal transport and information theory, and bypass the need to select summary statistics.

As shown through a number of benchmark experiments, ABC with Wasserstein distances and MMD, despite being more computationally involved, generally outperform KL-ABC. In particular, KL-ABC exhibits the slowest posterior concentration rate in most experiments. In comparison, WABC is able to identify the true parameters and achieves decent concentration in most experiments. The performance of MMD is only slightly worse than WABC, but it has the advantage of having a cost that is only quadratic in the data size, making it more suitable for large-scale data sets. We hope this essay has provided guidance for practitioners who are interested in using these metrics in their ABC paradigm.

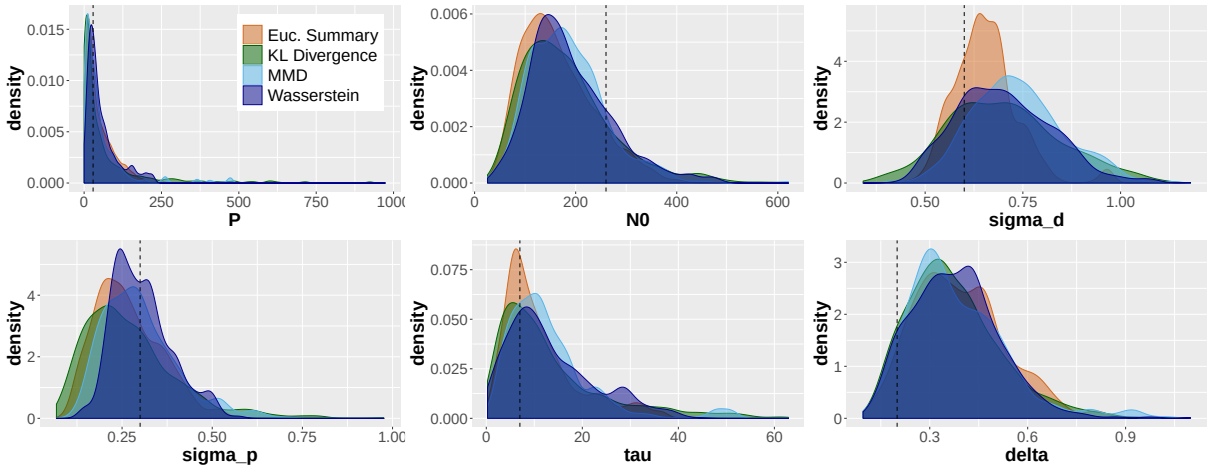


Figure 15: Posteriors of the six parameters in the blowfly model with simulated data and known true parameters: $P = 29, \delta = 0.2, N_0 = 260, \sigma_d = 0.6, \sigma_p = 0.3$ and $\tau = 7$. The true parameters are marked by black dotted lines.

We have also reviewed the asymptotic results presented in Bernton et al. [2019a]. In particular, the large-sample asymptotic (Prop. 4.1) shows that, for a fixed threshold ϵ, the ABC posterior as we obtain more and more data converges not to the true posterior, but to a restricted version of it. Moreover, we generalize the small-tolerance asymptotic (Proposition 2 of Bernton et al. [2019a]; Prop. 4.4 of this essay) to any soft ABC posterior whose associated kernel satisfies the concentration condition (definition 5). This guarantees the convergence of the ABC posterior to the true posterior as the threshold tends to zero. Furthermore, one can obtain quantitative bounds on the rate of posterior concentration by assuming mild conditions on the data discrepancy, the data generating process and the parameter space (assumption 1 – 5). Sufficient conditions for these assumption to hold for $\mathcal{W}_p$ and MMD are provided.

Directions of future research may include finding the optimal rates of concentration of WABC and MMD-ABC posteriors. In particular, it would be beneficial to generalize the convergence properties of ABC posteriors with an arbitrary member of the family of integral probability metrics Müller [1997], to which Wasserstein and MMD belong.

Furthermore, ABC with the aforementioned Sinkhorn divergence, which is an interpolation between the Wasserstein distance and MMD, is still under-explored and deserves further study. This, together with the sliced-Wasserstein distance mentioned in Nadjahi et al. [2019], are also linked to the scalability of ABC, which remains an ongoing research topic in the present literature.

Finally, the fact that the KL divergence is not a proper metric defies any theoretical guarantees on the posterior concentration of KL-ABC. More justification on this aspect is needed. This may also shed light on more informed use of another broad family of metrics within ABC, termed the *f-divergence* [Csiszár and Shields, 2004], especially because it has been proven successful in many other applications, such as variational inference and generative adversarial networks [Nowozin et al., 2016].

A Positive Definite Kernels

We provide the definition of positive definite kernels discussed in section 3.2.

Definition 6 (Positive definite kernels)**.** A symmetric function $k : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is a positive definite kernel if, for any $y_1, \ldots, y_n \in \mathcal{Y}$ and any $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(y_i, y_j) \geq 0.$$

This is not to be confused with the kernels introduced in section 2.2, which are functions on $\mathbb{R}$. Three examples of positive definite kernels are:

1. **Polynomial**: $k(y, z) = (y^\intercal z + c)^p$, where $c > 0$ and $p \in \mathbb{N}$.
2. **Gaussian**: $k(y, z) = \exp(-\|y - z\|^2/(2\sigma^2))$, where $\sigma > 0$.
3. **Laplacian**: $k(y, z) = \exp(-\|y - z\|_1/(2\sigma))$, where $\sigma > 0$.

Due to the characteristic property (see e.g. Definition 3.2 of Muandet et al. [2017]) of the Gaussian and Laplacian kernels, they are commonly used in kernel mean embedding. Also note that these two kernels are bounded, so it follows from Prop. 3.1 that the embedding of any distribution is well-defined. The parameter σ is known as the bandwidth, similarly to the case of kernel density estimation.

Park et al. [2016] suggest to use the Gaussian kernel with the median of $\{\|y_i - y_j\|_1 : i, j = 1, \ldots, n\}$ as the bandwidth.

B An Alternative Definition of Maximum Mean Embedding

Definition 7 (Sriperumbudur et al. [2010])**.** Let $\mathcal{F}$ be a collection of real-valued bounded measurable functions on $\mathcal{Y}$ and let μ, ν be two probability measures. The *integral probability metric* is defined as

$$\gamma[\mathcal{F}; \mu, \nu] = \sup_{f \in \mathcal{F}} \left\{ \int f(x) d\mu(x) - \int f(x) d\nu(x) \right\}.$$

It can be shown that [Dudley, 2002, Theorem 11.8.2; Gretton et al., 2012, Lemma 4 Muandet et al., 2017, p. 50]:

1. If $\mathcal{F}$ is the set of all 1-Lipschitz functions with respect to a metric ρ on $\mathcal{Y}$, i.e. $\mathcal{F} = \{f : |f(x) - f(y)| \leq \rho(x, y)\}$, then $\gamma[\mathcal{F}; \mu, \nu] = \mathcal{W}_1(\mu, \nu)$.
2. If $\mathcal{F}$ is the set of functions in the unit ball of some RKHS $\mathcal{H}$, i.e. $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \leq 1\}$, then $\gamma[\mathcal{F}; \mu, \nu] = \mathrm{MMD}^2(\mu, \nu)$.

Formally, 1 follows from the *Kantorovich-Rubinstein duality*. We can therefore see that WABC and K2-ABC are closely related in the sense that they are both using special cases of integral probability metrics on $\mathcal{P}(\mathcal{Y})$ as data discrepancy.

C Almost Sure Convergence of the 1NN Estimator for KL Divergence

We present the proof that the estimator Eq. 8 for the KL divergence is consistent. A more general result for an estimator using k-nearest-neighbours can be derived with the same idea (see, e.g. Perez-Cruz, 2008, Theorem 2). The proof relies on the following lemma.

Lemma C.1. *Assume μ and ν are absolute continuous probability measures and μ is absolutely continuous with respect to ν. Denote*

$$\begin{cases} r(y_i) \coloneqq \min\limits_{j=1,\dots,n} \|y_i - z_j\|_2, \\ s(y_i) \coloneqq \min_{j\neq i; j=1,\dots,n} \|y_i - y_j\|_2, \end{cases} \qquad \begin{cases} \hat{p}(y_i) \coloneqq \frac{1}{n-1} \cdot \frac{\Gamma(\frac{d}{2}+1)}{\pi^{d/2} r(y_i)}, \\ \hat{q}(y_i) \coloneqq \frac{1}{n} \cdot \frac{\Gamma(\frac{d}{2}+1)}{\pi^{d/2} s(y_i)}, \end{cases}$$

where y_i and z_j are i.i.d. samples from μ and ν, respectively. It follows that $\mu(dy)/\hat{p}(y)$ is an exponential random variable with unit rate for any y in the support of $\mu(dy)$.

The assumption that μ is absolutely continuous with respect to ν guarantees that the KL divergence is finite. $\hat{p}$ and $\hat{q}$ can be thought intuitively as estimators for the densities $\mu(dy)$ and $\nu(dy)$. The almost sure convergence of the estimator 8 for KL-divergence then follows from the following result.

Theorem C.2. *Under the same assumptions in Lemma C.1, $\mathfrak{D}_{KL}(y_{1:n}, z_{1:n}) \overset{a.s.}{\to} KL(\mu\|\nu)$ as $n \to \infty$.*

Proof of Lemma C.1. Note first that, by definition of $r(y_i)$, $\mathbb{P}(r(y) > \epsilon) \to 0$ as $n \to \infty$ for any positive ϵ and any y in the support of $\mu(dy)$. Therefore, as n approaches infinity, we can consider y and its nearest neighbour to be drawn from a uniform distribution μ. A consequence is that we can assume without loss of generality that μ is a $d-$dimensional uniform distribution on its support.

Suppose μ is uniform as described. Let the set $S_{y,R} \coloneqq \{y_i : \|y_i - y\|_2 \leq R\}$ be the set of all samples contained within the ball of radius R centred at y, denoted $\mathcal{B}_{y,R}$. If $\mathcal{B}_{y,R}$ lies in the support of $\mu(dy)$, then $\{\|y_i - y\|_2^d : y_i \in S_{y,R}\}$ are uniformly distributed between 0 and R^d. Recall the standard result that the waiting time from the origin to the first arrival of N uniform random variable, where N is a Poisson random variable, is exponentially distributed (see e.g. [Balakrishnan, 1996, Chapter 33.2.3]). Therefore, $r(y)^d = \min_{y_i \in S_{y,R}} \|y_i - y\|_2^d$ is an exponential random variable. Now,

$$\frac{\mu(dy)}{\hat{p}(y)} = \frac{\mu(dy)(n-1)\pi^{d/2} r(y_i)^d}{\Gamma(d/2+1)}.$$

Since $\pi^{d/2}/\Gamma(d/2+1)$ is the volume of a d-dimensional unit ball, $\mu(dy)(n-1)\pi^{d/2}/\Gamma(d/2+1)$ is the mean number of samples contained in a unit ball centred at y. It follows by scaling that $\mu(dy)/\hat{p}(y)$ is exponentially distributed with rate 1. ∎

Proof of Theorem C.2. We begin by rewritting Eq. 8 as

$$\mathfrak{D}_{KL}(y_{1:n}, z_{1:n}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\hat{p}(y_i)}{\hat{q}(y_i)} = \frac{1}{n}\sum_{i=1}^{n} \log \frac{\mu(dy_i)}{\nu(dy_i)} - \frac{1}{n}\sum_{i=1}^{n} \log \frac{\mu(dy_i)}{\hat{p}(y_i)} + \frac{1}{n}\sum_{i=1}^{n} \log \frac{\nu(dy_i)}{\hat{q}(y_i)}.$$

The first term converges almost surely to the KL divergence between μ and ν. For the second term, recall from Lemma C.1 that $\mu(dy_i)/\hat{p}(y_i)$ is Exponentially distributed with unit rate. The second term hence converges almost surely to $\mathbb{E}_X[\log(X)]$, where $X \sim$ Exponential(1). By the same argument, $\nu(dy_i)/\hat{q}(y_i)$ also follows Exponential(1). The third term again converges almost surely to $\mathbb{E}_X[\log(X)]$ and cancels with the second.

Finally, since a finite sum of almost surely convergent terms converges almost surely, the result follows. ∎

D Proofs of the Posterior Concentration Rates

We present the proofs of Prop. 4.6 and of Cor. 4.7, which are provided in the appendix of Bernton et al. [2019a].

Proof of Prop. 4.6. We begin by considering the probability $\pi^{\epsilon+\epsilon_n}_{y_{1:n}}(\mathcal{W}_p(\mu_\theta, \mu_*) > \delta)$ for some $\delta, \epsilon > 0$. By Bayes' formula,

$$\pi^{\epsilon+\epsilon_n}_{y_{1:n}}(\mathcal{W}_p(\mu_\theta, \mu_*) > \delta) = \frac{\mathbb{P}_{\theta,z_{1:n}}(\mathcal{W}_p(\mu_\theta, \mu_*) > \delta, \mathcal{W}_p(\hat{\mu}_{y_{1:n}}, \hat{\mu}_{\theta,z_{1:n}}) \leq \epsilon + \epsilon_*)}{\mathbb{P}_{\theta,z_{1:n}}(\mathcal{W}_p(\hat{\mu}_{y_{1:n}}, \hat{\mu}_{\theta,z_{1:n}}) \leq \epsilon + \epsilon_*)}, \tag{12}$$

where $\mathbb{P}_{\theta,z_{1:n}}$ denotes the joint distribution of parameter θ and synthetic data $z_{1:n} \sim \mu^{(n)}_\theta$. We will proceed to upper-bound Eq. 12 by upper-bounding the numerator and lower-bounding the denominator.

Denote the event on the numerator by $\Omega = \{\mathcal{W}_p(\mu_\theta, \mu_*) > \delta, \mathcal{W}_p(\hat{\mu}_{y_{1:n}}, \hat{\mu}_{\theta,z_{1:n}}) \leq \epsilon + \epsilon_*\}$. On Ω, we have

$$\begin{aligned} \delta < \mathcal{W}_p(\mu_\theta, \mu_*) &\leq \mathcal{W}_p(\mu_*, \hat{\mu}_{y_{1:n}}) + \mathcal{W}_p(\hat{\mu}_{y_{1:n}}, \hat{\mu}_{\theta,z_{1:n}}) + \mathcal{W}_p(\hat{\mu}_{\theta,z_{1:n}}, \mu_\theta) \qquad (13) \\ &\leq \mathcal{W}_p(\mu_*, \hat{\mu}_{y_{1:n}}) + \mathcal{W}_p(\hat{\mu}_{\theta,z_{1:n}}, \mu_\theta) + \epsilon + \epsilon_*, \end{aligned}$$

where Eq. 13 follows from triangular inequality of $\mathcal{W}_p$. Let $A(n, \epsilon) := \{y_{1:n} \in \mathcal{Y}^n : \mathcal{W}_p(\hat{\mu}_{y_{1:n}}, \mu_*) \leq \epsilon/3\}$ and assume henceforth that $y_{1:n} \in A(n, \epsilon)$. We have

$$\delta < \mathcal{W}_p(\mu_\theta, *) \leq \mathcal{W}_p(\hat{\mu}_{\theta,z_{1:n}}, \mu_\theta) + 4\epsilon/3 + \epsilon_*.$$

Reparametrizing $\xi := \delta - 4\epsilon/3 - \epsilon_*$, we can use the above inequality to bound the numerator by

$$\pi^{\epsilon+\epsilon_*}_{y_{1:n}}(\mathcal{W}_p(\mu_\theta, \mu_*) > 4\epsilon/3 + \epsilon_* + \xi) \leq \frac{\mathbb{P}_{\theta,z_{1:n}}(\mathcal{W}_p(\hat{\mu}_{\theta,z_{1:n}}, \mu_\theta) > \xi)}{\mathbb{P}_{\theta,z_{1:n}}(\mathcal{W}_p(\hat{\mu}_{y_{1:n}}, \hat{\mu}_{\theta,z_{1:n}}) \leq \epsilon + \epsilon_*)}. \tag{14}$$

The rest of the proof proceeds by further bounding the above fraction using the three assumptions. Focusing first on the numerator, we have by assumption 2 that

$$\begin{aligned} \mathbb{P}_{\theta,z_{1:n}}(\mathcal{W}_p(\hat{\mu}_{\theta,z_{1:n}}, \mu_\theta) > \xi) &= \int_\Theta \mu^{(n)}_\theta(\mathcal{W}_p(\hat{\mu}_{\theta,z_{1:n}}, \mu_\theta) > \xi)\pi(d\theta) \\ &\leq \int_\Theta c(\theta) f_n(\xi)\pi(d\theta) = c_1 f_n(\xi), \end{aligned}$$

where $c_1 := \int_\Theta c(\theta)\pi(d\theta) \leq \infty$. For the numerator,

$$\begin{aligned} &\mathbb{P}_{\theta,z_{1:n}}(\mathcal{W}_p(\hat{\mu}_{y_{1:n}}, \hat{\mu}_{\theta,z_{1:n}}) \leq \epsilon + \epsilon_*) = \int_\Theta \mu^{(n)}_\theta(\mathcal{W}_p(\hat{\mu}_{y_{1:n}}, \hat{\mu}_{\theta,z_{1:n}}) \leq \epsilon + \epsilon_*)\pi(d\theta) \\ &\geq \int_{\mathcal{W}_p(\mu_\theta,\mu_*)\leq\epsilon/3+\epsilon_*} \mu^{(n)}_\theta(\mathcal{W}_p(\hat{\mu}_{y_{1:n}}, \hat{\mu}_{\theta,z_{1:n}}) \leq \epsilon + \epsilon_*)\pi(d\theta) \\ &\geq \int_{\mathcal{W}_p(\mu_\theta,\mu_*)\leq\epsilon/3+\epsilon_*} \mu^{(n)}_\theta(\mathcal{W}_p(\hat{\mu}_{y_{1:n}}, \mu_*) + \mathcal{W}_p(\mu_*, \mu_\theta) + \mathcal{W}_p(\mu_\theta, \hat{\mu}_{\theta,z_{1:n}}))\pi(d\theta) \qquad (15) \\ &\geq \int_{\mathcal{W}_p(\mu_\theta,\mu_*)\leq\epsilon/3+\epsilon_*} \mu^{(n)}_\theta(\mathcal{W}_p(\mu_\theta, \hat{\mu}_{\theta,z_{1:n}}) \leq \epsilon/3)\pi(d\theta) \\ &\qquad\qquad \text{(as } \mathcal{W}_p(\mu_*, \mu_\theta) \leq \epsilon/3 + \epsilon_* \text{ and } \mathcal{W}_p(\hat{\mu}_{y_{1:n}}, \mu_*) \leq \epsilon/3\text{)} \\ &= \pi(\mathcal{W}_p(\mu_\theta, \mu_*) \leq \epsilon/3 + \epsilon_*) - \int_{\mathcal{W}_p(\mu_\theta,\mu_*)\leq\epsilon/3+\epsilon_*} \mu^{(n)}_\theta(\mathcal{W}_p(\mu_\theta, \hat{\mu}_{\theta,z_{1:n}}) > \epsilon/3)\pi(d\theta) \\ &\geq \pi(\mathcal{W}_p(\mu_\theta, \mu_*) \leq \epsilon/3 + \epsilon_*) - \int_{\mathcal{W}_p(\mu_\theta,\mu_*)\leq\epsilon/3+\epsilon_*} c(\theta) f_n(\epsilon/3)\pi(d\theta) \qquad \text{(by assumption 2,)} \end{aligned}$$

where Eq. 15 follows from the triangular inequality. By the condition in assumption 2, choosing $\epsilon > 0$ small enough such that $\epsilon/3 \leq \delta_0$ yields $c(\theta) \leq c_0$ for some $c_0 > 0$. The above line is then bounded below by $\pi(\mathcal{W}_p(\mu_*, \mu_\theta) \leq \epsilon/3 + \epsilon_*)(1 - c_0 f_n(\epsilon/3))$.

Now, replacing ϵ by ϵ_n such that $f_n(\epsilon_n) \to 0$ implies that $c_0 f_n(\epsilon/3) \leq 1/2$ for n sufficiently large. Therefore, we have the following lower bound fort he denominator

$$\frac{1}{2}\pi(\mathcal{W}_p(\mu_*, \mu_\theta) \leq \epsilon_n/3 + \epsilon_*) \leq c_\pi \epsilon_n^L,$$

for n large enough, by assumption 3. Combining this with the bound on the numerator, we have

$$\pi_{y_{1:n}}^{\epsilon_n+\epsilon_*}(\mathcal{W}_p(\mu_*, \mu_\theta) > 4\epsilon_n/3 + \epsilon_* + \xi) \leq C f_n(\xi)\epsilon_n^{-L},$$

where $C \coloneqq c_1/c_\pi$. Now, since f_n is strictly decreasing, the inverse f_n^{-1} is well-defined at ϵ_n^L/R, and we can choose $\xi_n = f_n^{-1}(\epsilon_n^L/R)$ to yield

$$\pi_{y_{1:n}}^{\epsilon_n+\epsilon_*}(\mathcal{W}_p(\mu_*, \mu_\theta) > 4\epsilon/3 + \epsilon_* + f_n^{-1}(\epsilon_n^L/R)) \leq C/R.$$

Finally, $\mathbb{P}(\{\omega : y_{1:n}(\omega) \in A(n, \epsilon_n)\}) \to 1$ as $n \to \infty$ by assumption. Hence, the above inequality holds with probability going to 1. ■

Proof of Cor. 4.7. Let $\delta > 0$ such that $\{\theta \in \Theta : \rho_\Theta(\theta, \theta_*) \leq \delta\} \subset U$, where U is the set described in assumption 5. By assumption 4, there exists $\delta' > 0$ such that $\rho_\Theta(\theta, \theta_*) > \delta$ implies $\mathcal{W}_p(\mu_\theta, \mu_*) - \epsilon_* > \delta'$.

Choose n large enough such that $4\epsilon_n/3 + f_n^{-1}(\epsilon_n^L/R) < \delta'$. For all $\theta \in \Theta$ such that $\mathcal{W}_p(\mu_\theta, \mu_*) \leq 4\epsilon_n/3 + f_n^{-1}(\epsilon_n^L/R) + \epsilon_*$, we have $\mathcal{W}_p(\mu_\theta, \mu_*) - \epsilon_* \leq \delta'$, which implies $\rho_\Theta(\theta, \theta_*) \leq \delta$. It follows that $\{\theta \in \Theta : \mathcal{W}_p(\mu_\theta, \mu_*) - \epsilon_* \leq 4\epsilon_n/3 + f_n^{-1}(\epsilon_n^L/R)\} \subset U$.

Therefore, the inequality in assumption 5 implies

$$\begin{aligned}\pi_{y_{1:n}}^{\epsilon_n+\epsilon_*}(\rho_\Theta(\theta, \theta_*) \leq K(4\epsilon_n/4 + f_n^{-1}(\epsilon_n^L/R))^\alpha) &\geq \pi_{y_{1:n}}^{\epsilon_n+\epsilon_*}(\mathcal{W}_p(\mu_\theta, \mu_*) - \epsilon_* \leq 4\epsilon_n/3 + f_n^{-1}(\epsilon_n^L/R)) \\ &\geq 1 - C/R,\end{aligned}$$

with probability going to 1, where the second line follows from Prop. 4.6. ■

D.1 Generalization to Other Metrics

We remark that, in the above two proofs, the only assumptions on the discrepancy metric $\mathcal{W}_p$, apart from assumptions 1-5, are that

(i) it satisfies the triangular inequality (Eq. 13 and Eq. 15), and

(ii) it is symmetric.

In other words, any metric $\mathfrak{D}$ satisfying these two conditions and assumptions 1-5 yields the same bounds on the concentration rate as in Prop. 4.6 and Cor. 4.7, by replacing $\mathcal{W}_p$ with $\mathfrak{D}$ and following the same argument in the two proofs. In particular, since the MMD is a proper metric, we have shown that MMD-ABC has the same bound on the concentration rate. In contrast, the proofs would not work for KL-ABC, as it is well known that the KL divergence does *not* satisfies the triangular inequality and is *not* symmetric [MacKay, 2003].

D.2 Weak Convergence and the Metrizable Property

To verify assumption 4 with i.i.d. data, Bernton et al. [2019a] gave in their supplementary material sufficient conditions in the case of Wasserstein distances. Here, we generalize them to an arbitrary metric $\mathfrak{D}$ that metrizes the weak convergence in $\mathcal{P}(\mathcal{Y})$. We begin by defining the weak convergence and the metrizable property [Villani, 2009, Definition 6.7].

Definition 8 (Weak convergence in $\mathcal{P}(\mathcal{Y})$)**.** A sequence of probability measures $(\mu_n)_{n\geq 1}$ in $\mathcal{P}(\mathcal{Y})$ is said to converge weakly to some μ in $\mathcal{P}(\mathcal{Y})$ if $\mathbb{E}_{\mu_n}[f] \to \mathbb{E}_{\mu}[f]$, for all bounded, continuous function f.

Definition 9 (Weak convergence in $\mathcal{P}_p(\mathcal{Y})$)**.** Let $p \in [1, \infty)$ and $\mathcal{P}_p(\mathcal{Y})$ be the Wasserstein space of order p (see section 3.1.1). A sequence of probability measures $(\mu_n)_{n\geq 1}$ in $\mathcal{P}_p(\mathcal{Y})$ is said to converge weakly to some μ in $\mathcal{P}_p(\mathcal{Y})$ if $\mu_n \to \mu$ in distribution, and there exists $y_0 \in \mathcal{Y}$ such that

$$\int_{\mathcal{Y}} \rho(y_0, y)^p d\mu_n(y) \to \int_{\mathcal{Y}} \rho(y_0, y)^p d\mu(y).$$

Definition 10. A metric $\mathfrak{D}$ is said to metrize the weak convergence in a space of probability measures $\mathcal{P}(\mathcal{Y})$ if a sequence of probability measures in $\mathcal{P}(\mathcal{Y})$ converges in $\mathfrak{D}$ if and only if it converges weakly.

It is known that $\mathcal{W}_p$ for $p \in [1, \infty)$ metrizes $\mathcal{P}_p(\mathcal{Y})$ [Villani, 2009, Theorem 6.8] and the MMD associated with a continuous and characteristic kernel metrizes $\mathcal{P}(\mathcal{Y})$ [Simon-Gabriel and Schölkopf, 2018, Theorem 12]. Assuming $\mathfrak{D}$ metrizes the weak convergence in some space of probability measures $\mathcal{P}(\mathcal{Y})$ and the data are i.i.d., the following conditions imply assumption 4.

Assumption 6. *For any $\theta_n, \theta \in \Theta$, $n \in \mathbb{N}$, $\rho_\Theta(\theta_n, \theta) \to 0$ implies $\mathfrak{D}(\mu_{\theta_n}, \mu_\theta) \to 0$.*

Proposition D.1. *Let $\mathfrak{D}$ be a metric that metrizes the weak convergence in $\mathcal{P}(\mathcal{Y})$ and assume 6 holds. Suppose that there exists a connected and compact $S \subset \Theta$ with positive Lebesgue measure such that*

$$\inf_{\theta \in \Theta \setminus S} \mathfrak{D}(\mu_*, \mu_\theta) > \inf_{\theta \in \Theta} \mathfrak{D}(\mu_*, \mu_\theta). \tag{16}$$

Then $\theta \mapsto \mathfrak{D}(\mu_*, \mu_\theta)$ attains its minimum at some θ_*. Furthermore, if θ_* is unique, then it is well-separated.

Proof of Prop. D.1. Assumption 6 and the metrizable property of $\mathfrak{D}$ gives that $\theta \mapsto \mathfrak{D}(\mu_*, \mu_\theta)$ is continuous. It therefore attains a minimum θ_* in compact S. This is also a global minimum by the condition Eq. 16.

Now, assume θ_* is unique. We shall show that it is well-separated, i.e. $\forall \epsilon > 0$, $\exists \delta > 0$ such that

$$\inf_{\theta \in \Theta : \mathfrak{D}(\mu_\theta, \mu_*) \geq \epsilon} > \mathfrak{D}(\mu_*, \mu_{\theta_*}) + \delta.$$

Fix $\epsilon > 0$. If $\{\theta \in \Theta : \mathfrak{D}(\mu_\theta, \mu_*) \geq \epsilon\} \subset \Theta \setminus S$, then by Eq. 16,

$$\inf_{\{\theta \in \Theta : \mathfrak{D}(\mu_\theta, \mu_{\theta_*}) \geq \epsilon\}} \mathfrak{D}(\mu_*, \mu_\theta) \geq \inf_{\theta \in \Theta \setminus S} \mathfrak{D}(\mu_*, \mu_\theta) > \inf_{\theta \in \Theta} \mathfrak{D}(\mu_*, \mu_\theta),$$

so well-separation follows. If $\{\theta \in \Theta : \mathfrak{D}(\mu_\theta, \mu_*) \geq \epsilon\} \cap S \neq \emptyset$, then it is compact. Indeed, since S is compact, there exists $\epsilon' > \epsilon$ such that $S \subset \{\theta \in \Theta : \epsilon' \geq \mathfrak{D}(\mu_\theta, \mu_{\theta_*})\}$. Hence, $\{\theta \in \Theta : \mathfrak{D}(\mu_\theta, \mu_{\theta_*}) \geq \epsilon\} \cap S = \{\theta \in \Theta : \epsilon' \geq \mathfrak{D}(\mu_\theta, \mu_{\theta_*}) \geq \epsilon\} \cap S$. Since $\{\theta \in \Theta : \epsilon' \geq \mathfrak{D}(\mu_\theta, \mu_{\theta_*}) \geq \epsilon\}$ is compact, this intersection of 2 compact sets is also compact. It follows by continuity of $\theta \mapsto \mathfrak{D}(\mu_\theta, \mu_\theta)$ that it attains an infimum on $\{\theta \in \Theta : \mathfrak{D}(\mu_\theta, \mu_{\theta_*}) \geq \epsilon\} \cap S$. Note that θ_* is not contained in this set, so this infimum cannot be θ_* by uniqueness. Hence, well-separation follows. ■

Assumption 6 can be checked in special cases where explicit relationship between ρ_Θ and $\mathcal{W}_p$ can be derived. An example is well-specified location models with ρ_Θ being the Euclidean distance, in which case Bernton et al. [2019a] showed in their supplementary material that $\mathcal{W}_p(\mu_\theta, \mu_{\theta_*}) = \rho_\Theta(\theta, \theta_*)$.

References

C. Andrieu, a Doucet, and R. Holenstein. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B*, 72(3):269 – 342, 6 2010. ISSN 0035-9246. doi: 10.1111/j.1467-9868.2009.00736.x. Publisher: Wiley.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/arjovsky17a.html.

W. H. W. Bai Jiang, Tung-Yu Wu. Approximate Bayesian computation with Kullback-Leibler divergence as data discrepancy. In A. Storkey and F. Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1711–1721, Playa Blanca, Lanzarote, Canary Islands, 09–11 Apr 2018. PMLR. URL http://proceedings.mlr.press/v84/jiang18a.html.

K. Balakrishnan. *Exponential Distribution: Theory, Methods and Applications*. Taylor & Francis, 1996. ISBN 9782884491921. URL https://books.google.co.uk/books?id=sQfpNOKzDjcC.

R. G. Bartle. *The Elements of Integration and Lebesgue Measure*. John Wiley & Sons, New York, 1995.

M. A. Beaumont. Joint determination of topology, divergence time and immigration in population trees 1. *Simulations, Genetics and Human Prehistory*, pages 135 – 154, 01 2006.

M. A. Beaumont. Approximate Bayesian Computation. *Annual Review of Statistics and Its Application*, 6 (1):379–403, 2019. doi: 10.1146/annurev-statistics-030718-105212. URL https://doi.org/10.1146/annurev-statistics-030718-105212.

E. Bernton, P. Jacob, M. Gerber, and C. Robert. Approximate Bayesian computation with the Wasserstein distance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81, 02 2019a. doi: 10.1111/rssb.12312. URL https://rss.onlinelibrary.wiley.com/doi/epdf/10.1111/rssb.12312.

E. Bernton, P. E. Jacob, M. Gerber, and C. P. Robert. On parameter estimation with the Wasserstein distance. *Information and Inference: A Journal of the IMA*, 8(4):657–676, 10 2019b. ISSN 2049-8772. doi: 10.1093/imaiai/iaz003. URL https://doi.org/10.1093/imaiai/iaz003.

A. Beygelzimer, S. Kakadet, J. Langford, S. Arya, D. Mount, and S. Li. *FNN: Fast Nearest Neighbor Search Algorithms and Applications*, 2019. URL https://CRAN.R-project.org/package=FNN. R package version 1.1.3.

G. Biau, F. Cérou, and A. Guyader. New insights into approximate Bayesian computation. *Ann. Inst. H. Poincaré Probab. Statist.*, 51(1):376–403, 02 2015. doi: 10.1214/13-AIHP590. URL https://doi.org/10.1214/13-AIHP590.

M. Blum and O. François. Non-linear regression models for approximate Bayesian computation. *Statistics and Computing*, 20:63–73, 03 2010. doi: 10.1007/s11222-009-9116-0.

M. G. B. Blum. Approximate Bayesian computation: A nonparametric perspective. *Journal of the American Statistical Association*, 105(491):1178–1187, 2010. doi: 10.1198/jasa.2010.tm09448. URL https://doi.org/10.1198/jasa.2010.tm09448.

G. Casella and R. Berger. *Statistical Inference*. Duxbury Resource Center, June 2001. ISBN 0534243126.

I. Csiszár and P. Shields. Information theory and statistics: A tutorial. *Foundations and Trends in Communications and Information Theory*, 1(4):417–528, 2004. ISSN 1567-2190. doi: 10.1561/0100000004. URL http://dx.doi.org/10.1561/0100000004.

M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 2292–2300, Red Hook, NY, USA, 2013. Curran Associates Inc.

P. Del Moral, A. Doucet, and A. Jasra. On adaptive resampling strategies for sequential monte carlo methods. *Bernoulli*, 18(1):252–278, 02 2012. doi: 10.3150/10-BEJ335. URL https://doi.org/10.3150/10-BEJ335.

C. Drovandi, A. Pettitt, and M. Faddy. Approximate Bayesian computation using indirect inference. *Journal of the Royal Statistical Society Series C*, 60:317–337, 05 2011. doi: 10.2307/41262278.

R. M. R. M. Dudley. *Real analysis and probability / R.M. Dudley.* Cambridge studies in advanced mathematics ; no. 74. Cambridge University Press, second edition. edition, 2002. ISBN 9780511755347.

P. Fearnhead and D. Prangle. Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate bayesian computation [with discussion]. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 74(3):419–474, 2012. ISSN 13697412, 14679868. URL http://www.jstor.org/stable/41674639.

N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707, Aug. 2015. URL https://hal.archives-ouvertes.fr/hal-00915365.

K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 489–496. Curran Associates, Inc., 2008. URL http://papers.nips.cc/paper/3340-kernel-measures-of-conditional-dependence.pdf.

A. Genevay, G. Peyre, and M. Cuturi. Learning Generative Models with Sinkhorn Divergences. In A. Storkey and F. Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1608–1617, Playa Blanca, Lanzarote, Canary Islands, 09–11 Apr 2018. PMLR. URL http://proceedings.mlr.press/v84/genevay18a.html.

M. Gerber and N. Chopin. Sequential quasi-Monte Carlo. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77, 02 2014. doi: 10.1111/rssb.12104.

C. Gottschlich and D. Schuhmacher. The shortlist method for fast computation of the earth mover's distance and finding optimal solutions to transportation problems. *PLoS ONE*, 9:e110214, 10 2014. doi: 10.1371/journal.pone.0110214.

K. Grauman and T. Darrell. Fast contour matching using approximate earth mover's distance. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 1, pages I–I, June 2004. doi: 10.1109/CVPR.2004.1315035.

A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two-sample-problem. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 513–520. MIT Press, 2007. URL http://papers.nips.cc/paper/3110-a-kernel-method-for-the-two-sample-problem.pdf.

A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13:723–773, 03 2012.

M. Gutmann, R. Dutta, S. Kaski, and J. Corander. Likelihood-free inference via classification. *Statistics and Computing*, 28(2):411–425, 3 2018. ISSN 0960-3174. doi: 10.1007/s11222-017-9738-6.

P. Halmos. *A Hilbert Space problem book*. Graduate Texts in Mathematics. Springer, 1982. ISBN 9780387906850. URL https://books.google.co.uk/books?id=S57XLkgbf0oC.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970. ISSN 00063444. URL http://www.jstor.org/stable/2334940.

K. Heggland and A. Frigessi. Estimating functions in indirect inference. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 66(2):447–462, 2004. ISSN 13697412, 14679868. URL http://www.jstor.org/stable/3647536.

S. Kolouri, K. Nadjahi, U. Simsekli, R. Badeau, and G. K. Rohde. Generalized sliced Wasserstein distances. *Annual Conference on Neural Information Processing Systems*, abs/1902.00434:261–272, 2019.

S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86, 03 1951. doi: 10.1214/aoms/1177729694. URL https://doi.org/10.1214/aoms/1177729694.

A. Lee. On the choice of MCMC kernels for approximate bayesian computation with SMC samplers. In *Proceedings of the 2012 Winter Simulation Conference (WSC)*, pages 1–12, 2012.

A. Lee and K. Łatuszyński. Variance bounding and geometric ergodicity of Markov chain Monte Carlo kernels for approximate Bayesian computation. *Biometrika*, 101(3):655–671, 08 2014. ISSN 0006-3444. doi: 10.1093/biomet/asu027. URL https://doi.org/10.1093/biomet/asu027.

D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Copyright Cambridge University Press, 2003.

J.-M. Marin, P. Pudlo, C. Robert, and R. Ryder. Approximate Bayesian computational methods. *Statistics and Computing*, 22, 01 2011. doi: 10.1007/s11222-011-9288-2.

J. W. Miller and D. B. Dunson. Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, 114(527):1113–1125, 2019. doi: 10.1080/01621459.2018.1469995. URL https://doi.org/10.1080/01621459.2018.1469995. PMID: 31942084.

K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf. *Kernel mean embedding of distributions: a review and beyond*. now, 2017. ISBN 9781680832891. URL https://ieeexplore.ieee.org/document/8187176.

A. Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997. ISSN 00018678. URL http://www.jstor.org/stable/1428011.

J. Munkres. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, 5(1):32–38, 1957. ISSN 03684245. URL http://www.jstor.org/stable/2098689.

K. Nadjahi, V. D. Bortoli, A. Durmus, R. Badeau, and U. Şimşekli. Approximate Bayesian computation with the sliced-Wasserstein distance, 2019.

S. Nowozin, B. Cseke, and R. Tomioka. F-gan: Training generative neural samplers using variational divergence minimization. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 271–279, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

V. M. Panaretos and Y. Zemel. Statistical aspects of Wasserstein distances. *Annual Review of Statistics and Its Application*, 6(1):405–431, 2019. doi: 10.1146/annurev-statistics-030718-104938. URL https://doi.org/10.1146/annurev-statistics-030718-104938.

M. Park, W. Jitkrittum, and D. Sejdinovic. K2-abc: Approximate Bayesian computation with kernel embeddings. In A. Gretton and C. C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 398–407, Cadiz, Spain, 09–11 May 2016. PMLR. URL http://proceedings.mlr.press/v51/park16.html.

F. Perez-Cruz. Kullback-Leibler divergence estimation of continuous distributions. In *2008 IEEE International Symposium on Information Theory*, pages 1666 – 1670, 08 2008. doi: 10.1109/ISIT.2008.4595271.

D. Prangle. Adapting the ABC Distance Function. *Bayesian Anal.*, 12(1):289–309, 03 2017. doi: 10.1214/16-BA1002. URL https://doi.org/10.1214/16-BA1002.

J. Pritchard, M. Seielstad, A. Perez-Lezaun, and M. Feldman. Population growth of human y chromosomes: A study of y chromosome microsatellites. *Molecular biology and evolution*, 16:1791–8, 01 2000. doi: 10.1093/oxfordjournals.molbev.a026091.

G. Puccetti. An algorithm to approximate the optimal expected inner product of two vectors with given marginals. *Journal of Mathematical Analysis and Applications*, 451, 02 2017. doi: 10.1016/j.jmaa.2017.02.003.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019. URL https://www.R-project.org/.

A. Ramdas, N. Trillos, and M. Cuturi. On Wasserstein Two-Sample Testing and Related Families of Nonparametric Tests. *Entropy*, 19(2):47, Jan 2017. ISSN 1099-4300. doi: 10.3390/e19020047. URL http://dx.doi.org/10.3390/e19020047.

G. Rayner and H. Macgillivray. Numerical maximum likelihood estimation for the g-and-k and generalized g-and-h distributions. *Statistics and Computing*, 12:57–75, 01 2002. doi: 10.1023/A:1013120305780.

F. J. Rubio and A. M. Johansen. A simple approach to maximum intractable likelihood estimation. *Electron. J. Statist.*, 7:1632–1654, 2013. doi: 10.1214/13-EJS819. URL https://doi.org/10.1214/13-EJS819.

W. Rudin. *Fourier analysis on groups: interscience tracts in pure and applied Mathematics*. John Wiley & Sons, Ltd, 2011. ISBN 9781118165621. doi: 10.1002/9781118165621.fmatter. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118165621.fmatter.

D. Schuhmacher, B. Bhre, C. Gottschlich, and F. Heinemann. *Transport: optimal transport in various forms.*, 2017. R package version 0.8-2.

C.-J. Simon-Gabriel and B. Schölkopf. Kernel distribution embeddings: Universal kernels, characteristic kernels and kernel metrics on distributions. *Journal of Machine Learning Research*, 19(44):1–29, 2018. URL http://jmlr.org/papers/v19/16-291.html.

S. A. Sisson, Y. Fan, and M. A. Beaumont. Overview of Approximate Bayesian Computation. *arXiv e-prints*, art. arXiv:1802.09720, Feb 2018.

B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. Lanckriet. Hilbert space embeddings and metrics on probability measures. *J. Mach. Learn. Res.*, 11:1517–1561, Aug. 2010. ISSN 1532-4435.

M. Tanaka, A. Francis, F. Luciani, and S. Sisson. Using Approximate Bayesian Computation to Estimate Tuberculosis Transmission Parameters From Genotype Data. *Genetics*, 173:1511–20, 08 2006. doi: 10.1534/genetics.106.055574.

B. M. Turner and T. V. Zandt. A tutorial on Approximate Bayesian Computation. *Journal of Mathematical Psychology*, 56(2):69 – 85, 2012. ISSN 0022-2496. doi: https://doi.org/10.1016/j.jmp.2012.02.005. URL http://www.sciencedirect.com/science/article/pii/S0022249612000272.

V. S. Varadarajan. On a problem in measure-spaces. *Ann. Math. Statist.*, 29(4):1275–1278, 12 1958. doi: 10.1214/aoms/1177706461. URL https://doi.org/10.1214/aoms/1177706461.

C. Villani. *Optimal transport : old and new*. Grundlehren der mathematischen Wissenschaften ; 338. Springer, 2009. ISBN 3540710493.

R. D. Wilkinson and S. Tavaré. Estimating primate divergence times by using conditioned birth-and-death processes. *Theoretical Population Biology*, 75(4):278 – 285, 2009. ISSN 0040-5809. doi: https://doi.org/10.1016/j.tpb.2009.02.003. URL http://www.sciencedirect.com/science/article/pii/S0040580909000215. Sam Karlin: Special Issue.

S. Wood. Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466(7310):1102–1104, 8 2010. ISSN 0028-0836. doi: 10.1038/nature09319.