# $\gamma$-ABC: Outlier-Robust Approximate Bayesian Computation Based on a Robust Divergence Estimator

**Masahiro Fujisawa**
The University of Tokyo
RIKEN AIP
fujisawa@ms.k.u-tokyo.ac.jp

**Takeshi Teshima**
The University of Tokyo
RIKEN AIP
teshima@ms.k.u-tokyo.ac.jp

**Issei Sato**
The University of Tokyo
RIKEN AIP
sato@k.u-tokyo.ac.jp

**Masashi Sugiyama**
RIKEN AIP
The University of Tokyo
sugi@k.u-tokyo.ac.jp

## Abstract

Approximate Bayesian computation (ABC) is a likelihood-free inference method that has been employed in various applications. However, ABC can be sensitive to outliers if a data discrepancy measure is chosen inappropriately. In this paper, we propose to use a nearest-neighbor-based $\gamma$-divergence estimator as a data discrepancy measure. We show that our estimator possesses a suitable theoretical robustness property called the *redescending* property. In addition, our estimator enjoys various desirable properties such as high flexibility, asymptotic unbiasedness, almost sure convergence, and linear-time computational complexity. Through experiments, we demonstrate that our method achieves significantly higher robustness than existing discrepancy measures.

## 1 Introduction

Approximate Bayesian computation (ABC) has been proposed as a "likelihood-free" inference scheme to approximately perform Bayesian inference when a complex model is used and it is impossible or difficult to compute its likelihood (see [53] for a general overview). Instead of investigating the explicit form of the likelihood function, ABC seeks parameters of a simulator-based model that can generate data that is close to the observed data under some discrepancy measure. ABC has been applied to many research fields, e.g., evolutionary biology [76], dynamic systems [83], economics [63], epidemiology [11]. aeronautics [18], and astronomy [15].

Rejection ABC [76, 65, 42], the basic ABC algorithm, proceeds as follows: (i) we draw an independent sample of the parameter $\theta$ from some prior $\pi$, (ii) we simulate data $Y^m = \{Y_j\}_{1:m}$ for each value of $\theta$, (iii) the parameter $\theta$ is discarded if the discrepancy $D(X^n, Y^m)$ between the observed data $X^n = \{X_i\}_{1:n}$ and the simulated data $Y^m = \{Y_j\}_{1:m}$ exceeds a tolerance threshold $\epsilon$. The accepted $\theta$ is used in the subsequent inference as a sample from an approximation to the posterior distribution called the *ABC posterior distribution*. Many studies have been performed to enhance the computational efficiency of the rejection ABC scheme, e.g., applying Markov Chain Monte Carlo (MCMC) [54, 80] or sequential Monte Carlo (SMC) [72, 24, 57].

The core element of ABC is the data discrepancy measure $D(X^n, Y^m)$ and the accuracy of parameters from the ABC posterior distribution crucially depends on its choice. While many discrepancies for ABC have been proposed, such as the distance between summary statistics [12, 25, 80], the maximum mean discrepancy (MMD) [60], the Wasserstein distance [7], a Kullback-Leibler (KL) divergence estimator [42], and the classification accuracy discrepancy (CAD) [35], these are often not robust to severe contamination of data [68, 49, 74]. Recently, two outlier-robust discrepancies have been proposed: one is a robust discrepancy based on MMD [49], and the other is using a robust M-estimator, e.g., Huber's estimation function [68]. However, these methods do not possess an ideal robust property for an extreme outlier, called the *redescending property* [55]. In addition, the former method has the cubic time cost $\mathcal{O}((n+m)^3/Q^2)$, where $Q$ is the number of blocks that divide the data. When ABC is applied to astronomy [47, 15], for example, we have to deal with noisy large-scale datasets, and the cubic time cost can be intractable. For the CAD, we may improve the robustness of the CAD by employing a robust classifier, such as robust LDA [20], robust FDA [40], or robust logistic regression [28]; however, its performance depends on the choice of the classifier, and its validity and robustness for heavy contamination data are not guaranteed in the ABC framework. Outside the ABC framework, recently, many robust inference schemes have been proposed, e.g., using robust divergences for *parametric* model inference [4, 29], Bayesian inference [45, 41, 58], variational

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. 

Table 1: Relationship between previous work and our work (*OR*: Outlier robustness, *RP*: Redescending property [55], *AU*: Asymptotically unbiasedness, *ASC*: Almost sure convergence, *QA*: ABC posterior analysis). In MONK [49], the number of blocks that divide the data is denoted by $Q$. The order of time costs for the $q$-Wasserstein distance is based on approximate optimization algorithms [21, 22]. The order of time costs for the CAD is based on logistic regression, where $d$ is the dimension of the observed and synthesized data. The symbol $(n \vee m)$ denotes $\max\{n, m\}$.

| Discrepancy measure | OR | RP | AU | ASC | QA | Time cost |
|---|---|---|---|---|---|---|
| MMD [60, 73] | - | - | - | - | - | $\mathcal{O}((n+m)^2)$ |
| $q$-Wasserstein distance [7] | - | - | - | - | - | $\mathcal{O}((n+m)^2)$ |
| CAD [35] | - | - | - | - | ✓ | $\mathcal{O}((n+m)d)$ |
| MONK BCD-Fast [49] | ✓ | - | - | - | - | $\mathcal{O}\left(\frac{(n+m)^3}{Q^2}\right)$ |
| KL-divergence estimator [61, 42] | - | - | ✓ | ✓ | ✓ | $\mathcal{O}((n \vee m)\log(n \vee m))$ |
| **$\gamma$-divergence estimator (ours)** | ✓ | ✓ | ✓ | ✓ | ✓ | $\mathcal{O}((n \vee m)\log(n \vee m))$ |

inference [30], and constructing Bayesian inference through a pseudo-likelihood via MMD [17]. Unfortunately, these methods cannot be used as a data discrepancy measure for ABC because these assume a *tractable* likelihood or *parametric* models. Therefore, there is no discrepancy measure that has both *well-guaranteed* robustness for an extreme outlier and reasonable time costs.

In this paper, we propose a novel outlier-robust and computationally-efficient discrepancy measure based on the $\gamma$-divergence [29]. Our discrepancy measure results in a robustness property of the ABC posterior called the *redescending property* [55], i.e., it automatically ignores extreme outliers in the observed data. Furthermore, we show that the $\gamma$-divergence estimation using a naive $k$-nearest neighbor density estimate has desirable asymptotic properties, which is not straightforward to prove unlike divergence estimators in $f$-divergence class, such as $\alpha$-divergence estimator [64]. Table 1 summarizes the relations among our method and major existing discrepancy measures. Our contributions are as follows.

- We construct a non-parametric and robust divergence estimator based on the $\gamma$-divergence (Section 3.3).
- We show that our method theoretically enjoys the robustness and validity (Sections 3.4 and 4.2).
- We show that our method has the asymptotic unbiasedness and the almost sure convergence property, which are mathematically much harder to show than those for the divergence estimators belonging to the $f$-divergence family (Section 4.1).
- Through experiments, we show that our estimator can significantly reduce the influence of the outliers even when the observed data have a large number of outliers (Section 5).

The rest of this paper is organized as follows. We summarize the ABC framework in Section 2. In Sections 3 and 4, we introduce the $k$-nearest neighborhood ($k$-NN) based density estimation, explain how to derive our divergence estimator based on $k$-NN, and conduct the theoretical analyses. Finally, we show the experimental results and the conclusion in Sections 5 and 6.

**Algorithm 1** Rejection ABC Algorithm [76, 65]

**Require:** Observed data $\{X_i\}_{i=1}^n$, prior $\pi(\theta)$ on the parameter space $\Theta$, tolerance threshold $\epsilon$, data discrepancy measure $D$

1: **Initialize:** $\epsilon$
2: **for** $t = 1$ to $T$ **do**
3:   **repeat:** propose $\theta \sim \pi(\theta)$ and draw $Y_1 \ldots, Y_m \overset{\text{i.i.d.}}{\sim} p_\theta$
4:   **until:** $D(X^n, Y^m) < \epsilon$
5:   Obtain $\theta^{(t)} = \theta$
6: **end for**
7: **return** $\{\theta^{(t)}\}_{t=1}^T$

## 2 Preliminaries

In this section, we give an overview of ABC. More detailed descriptions of the discrepancy measures which are often used in ABC can be found in Appendix E.

### 2.1 Approximate Bayesian Computation

We define $\mathcal{X} \subset \mathbb{R}^d$ as the data space and $\Theta$ as the parameter space. The model $\{p_\theta : \theta \in \Theta\}$ is a family of probability distributions on $\mathcal{X}$ and has no explicit formula, but we assume that we can generate i.i.d. random samples from $p_\theta$ given the value of $\theta$. The purpose of ABC is to seek the model parameter $\theta$ by comparing the observed data $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} p_{\theta^*}$ and the synthetic data $Y_1, \ldots, Y_m \overset{\text{i.i.d.}}{\sim} p_\theta$, where $\theta^* \in \Theta$ is the true parameter. The criterion used to compare these datasets $X^n = \{X_1, \ldots, X_n\}$ and $Y^m = \{Y_1, \ldots, Y_m\}$ is the data discrepancy measure $D(X^n, Y^m)$ defined over $\mathcal{X}^n \times \mathcal{X}^m$.

A well known algorithm of ABC is the rejection ABC [76, 65, 42], which proceeds as follows: (i) we draw an independent sample of the parameter $\theta$ from some prior $\pi$, (ii) we simulate data $Y^m = \{Y_j\}_{1:m}$ for each value of $\theta$, (iii) the parameter $\theta$ is discarded if the discrepancy $D(X^n, Y^m)$ between the observed data $X^n = \{X_i\}_{1:n}$ and the simulated data $Y^m = \{Y_j\}_{1:m}$ exceeds a tolerance threshold $\epsilon$. The rejection ABC algorithm is shown in Algorithm 1. It enables us to obtain i.i.d. random samples $\{\theta^{(t)}\}_{t=1}^T$ from the ABC posterior distribution defined as follows.

**Definition 1** (ABC posterior distribution)**.** *Let $\epsilon$ be fixed. Then, the ABC posterior distribution is defined by*

$$\pi(\theta|X^n, D, \epsilon) \propto \int \pi(\theta)\mathbb{1}\{D(X^n, Y^m) < \epsilon\} p_\theta(Y^m)\mathrm{d}Y^m, \tag{1}$$

*where $\pi(\theta)$ is a prior over the parameter space $\Theta$, $\epsilon > 0$ is a tolerance threshold, and $p_\theta(Y^m) = \prod_{j=1}^m p_\theta(Y_j)$.*

We only focus on the rejection ABC [76, 65, 42] throughout this paper. The main reason is two-fold: (i) to make a fair comparison and followed the experimental setting of the recent paper [42] proposing discrepancy measures for ABC, and (ii) to give theoretical guarantees to the ABC posterior explicitly, e.g., Theorem 1 and Corollary 1 in this paper. The rejection ABC is a reasonable choice for this purpose. While there are many sophisticated ABC algorithms, they are often extensions of the rejection ABC [54, 80, 72, 24, 57]; therefore, we can easily combine our development of the rejection ABC with these algorithms.

### 2.2 Model of Data Contamination

In this paper, we assume that the *observed* data are contaminated with outliers and focus on Huber's contamination-by-outlier case [38], where observed data are sampled i.i.d. from the following distribution:

$$(1 - \eta)G(x) + \eta H(x), \tag{2}$$

where $G(x)$ is a distribution we are interested in, $H(x)$ is an arbitrary contamination distribution, and $\eta \in [0, 1]$ is the proportion of contamination. If $\eta$ is relatively high, e.g., $\eta = 0.2$, the observed data are highly contaminated by $H$.

Due to distribution contamination described above, severe bias occurs in parameter estimation. Many robust estimation methods have been proposed to reduce the estimation bias caused by outliers [38, 39, 82, 4]. However, these methods tend to exhibit undesirable behaviors both empirically and theoretically, for heavily contaminated data [29]. Furthermore, for *non-parametric* inference schemes such as ABC, many of such robust inference frameworks cannot be used because they normally assume that the likelihood is tractable. Although Lerasle et al. [49] and Ruli et al. [68] have proposed robust discrepancy measures that can be compatible with non-parametric inference, these methods also cannot deal with a heavy contamination and the former method has high time costs. To conduct a robust non-parametric inference for heavily contaminated data, it is necessary to construct an alternative discrepancy measure with both robustness for an extreme outlier and reasonable time costs.

## 3 $\gamma$-ABC and Its Robustness

In this section, we construct a *non-parametric* "likelihood-free" inference scheme based on the $\gamma$-divergence that has been used in robust parameter estimation from heavily contaminated data. In Section 3.1, we introduce the $\gamma$-divergence and explain why we choose a $k$-NN based density estimation to derive our estimator. Next, we overview a $k$-NN based density estimation in Section 3.2 and derive our estimator in Section 3.3. Finally, we guarantee the robustness of the ABC based on our estimator in Section 3.4.

### 3.1 $\gamma$-divergence and Its Estimation

To make a robust parameter estimation in the heavily contamination situation described in Section 2.2, Fujisawa and Eguchi [29] proposed the $\gamma$-*divergence*, which possesses strong robustness for heavily contaminated data.

**Definition 2** ($\gamma$-divergence [29])**.** *Let $p$ and $q$ be positive measurable functions from a measurable set $\mathcal{M}_0 \subseteq \mathbb{R}^d$ to $\mathbb{R}$. Let $\gamma > 0$. Then, the $\gamma$-divergence is defined as*

$$D_\gamma(p\|q) = \frac{1}{\gamma(1+\gamma)} \log \frac{\left(\int_{\mathcal{M}_0} p^{1+\gamma}(x)\mathrm{d}x\right)\left(\int_{\mathcal{M}_0} q^{1+\gamma}(x)\mathrm{d}x\right)^\gamma}{\left(\int_{\mathcal{M}_0} p(x)q^\gamma(x)\mathrm{d}x\right)^{1+\gamma}}. \tag{3}$$

To combine the $\gamma$-divergence with ABC, we need to estimate Eq. (3) from observed and synthesized data. A potential approach we can consider to estimating Eq. (3) is extending $f$-divergence estimation frameworks, e.g., kernel density estimation (KDE) [37] and direct density ratio estimation methods such as KLIEP [75] or uLSIF [44]. However, the former method suffers from high time cost due to a kernel function and the necessity to select appropriate kernels and its parameters. Furthermore, the latter methods need to construct a model directly for a density ratio; therefore, it is hard to use this approach in the estimation of Eq. (3) since $\gamma$-divergence, which is not included in the $f$-divergence class, is not expressed as a functional of the density ratio.

For these reasons, we consider using a $k$-NN based density estimation to estimate $\gamma$-divergence. This approach has only one hyper-parameter, $k$. Furthermore, this approach does not depend on any additional models.

### 3.2 $k$-Nearest Neighbor based Density Estimation

Let $X^n$ be an i.i.d. sample of size $n$ from a probability distribution with density $p$, and $Y^m$ be an i.i.d. sample of size $m$ from $q$. Furthermore, we define $\rho_k(i)$ as the Euclidean distance between the $i$-th sample $X_i$ of $X^n$ and its $k$-th nearest neighbor ($k$-NN) among $X^n \setminus X_i$. Similarly, we define $\nu_k(i)$ as the Euclidean distance between the $i$-th sample $X_i$ and its

$k$-NN among $Y^m$. Let $\mathcal{B}(x, R)$ be a closed ball with radius $R$ around $x \in \mathbb{R}^d$. Finally, $\mathcal{V}(\mathcal{B}(x, R)) = \bar{c}R^d$ is defined as its volume, where $\bar{c}$ is the volume of the $d$-dimensional unit ball.

Loftsgaarden and Quesenberry [51] constructed the density estimators of $p$ and $q$ at the $i$-th sample $X_i$ via $k$-NN as follows:

$$\hat{p}_k(x_i) = \frac{k}{(n-1)\mathcal{V}(\mathcal{B}(x_i, \rho_k(i)))} = \frac{k}{(n-1)\bar{c}\rho_k^d(i)}, \quad (4)$$

$$\hat{q}_k(x_i) = \frac{k}{m\mathcal{V}(\mathcal{B}(x_i, \nu_k(i))} = \frac{k}{m\bar{c}\nu_k^d(i)}. \quad (5)$$

These density estimators, Eqs. (4) and (5), are consistent estimators of the density only when the number of neighbors $k$ goes to infinity as the sample size $n$ goes to infinity. We use these estimators for constructing our robust divergence estimator. Hereafter, we fix the value of $k$ and show that our divergence estimator still has desirable asymptotic properties, including consistency.

### 3.3 Robust Divergence Estimator on $\gamma$-divergence

Now we derive a *non-parametric* $\gamma$-divergence estimator based on $k$-NN density estimation. In ABC settings, outliers could be included in the observed data $X^n$. To reduce the influence of outliers, we rewrite the term $\int_{\mathcal{M}_0} q^{1+\gamma}(x)\mathrm{d}x$ in Eq. (3) to be $k$-NN estimatable from $\mathcal{M}' \subseteq \mathbb{R}^d$, where $\mathcal{M}'$ is the support of $q$, i.e.,

$$\int_{\mathcal{M}'} q^{1+\gamma}(y)\mathrm{d}y.$$

We can use the same notion of Eq. (4) when we focus on the synthetic data $Y^m$; therefore, the density estimation for $q(y)$ based on $k$-NN can be written as

$$\hat{q}_k(y_j) = \frac{k}{(m-1)\mathcal{V}(\mathcal{B}(y_j, \bar{\rho}_k(j))} = \frac{k}{(m-1)\bar{c}\bar{\rho}_k^d(j)}, \quad (6)$$

where $\bar{\rho}_k(j)$ is the Euclidean distance between the $j$-th sample $Y_j$ of $Y^m$ and its $k$-NN among $Y^m \setminus Y_j$.

By plugging in the $k$-NN density estimator into Eqs. (4), (5), and (6), we derive the $k$-NN based $\gamma$-divergence estimator as

$$\widehat{D}_\gamma(X^n\|Y^m) = \frac{1}{\gamma(1+\gamma)} \times \left( \log \frac{\left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\bar{c}}{k}\hat{p}_k(x_i)\right)^\gamma\right)\left(\frac{1}{m}\sum_{j=1}^{m}\left(\frac{\bar{c}}{k}\hat{q}_k(y_j)\right)^\gamma\right)^\gamma}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\bar{c}}{k}\hat{q}_k(x_i)\right)^\gamma\right)^{1+\gamma}} \right). \quad (7)$$

The details of the derivation is in Appendix A.

The estimator in Eq. (7) involves $2n$ and $2m$ operations of nearest neighbor search. If we implement them by $KD$ trees [5, 52], the time cost of finding $\widehat{D}_\gamma(X^n\|Y^m)$ is $\mathcal{O}((n \vee m)\log(n \vee m))$, where $(n \vee m) = \max\{n, m\}$, which is among the fastest (up to log factors) of the existing robust discrepancy approximators (see Table 1). Furthermore, this estimator fortunately enjoys ideal asymptotic properties: *asymptotic unbiasedness* and *almost surely convergence* under mild assumptions. We will show them in Section 4.1.

### 3.4 Robustness Property of $\gamma$-ABC against Outliers

Here, we investigate the behavior of the *sensitivity curve* (SC), which is a finite-sample analogue of the *influence function* (IF), both of which are used in quantifying the robustness of statistics [36, 30, 68]. We fix the observed data $X^n$ and consider a contamination by an outlier $X_0$. We define the contaminated data as $X^n_{[X_0]} := (X_0, X_1, \ldots, X_n)$. Then, the SC is defined as follows.

**Definition 3** (Sensitivity curve [36, 2.1e]). *Let $\gamma, \epsilon > 0$. Let us define the (population) pseudo-posterior as $\hat{\pi}(\theta|X^n) := \pi(\theta|X^n, \widehat{D}_\gamma, \epsilon)$. The sensitivity curve of $\hat{\pi}$ is defined as*

$$\mathrm{SC}^\theta_{n+1}(X_0) := (n+1)\left(\hat{\pi}(\theta|X^n_{[X_0]}) - \hat{\pi}(\theta|X^n)\right).$$

We consider SC instead of IF for two reasons: (i) we are interested in the pseudo-posterior distribution $\hat{\pi}(\theta|X^n)$ with respect to a finite sample $X^n$, and (ii) the IF of the quantities based on the considered divergence estimator may not be even defined (a detailed explanation is in Remark 2 in Appendix B.3).

Under this definition and some additional assumptions, we obtain the following theorem. Our analysis is a finite-sample analogue of what is called the *redescending property* [55] in the context of IF analysis.

**Theorem 1** (Sensitivity curve analysis). *Assume $k < \min\{n, m\}$. Also assume that $F_\theta(\epsilon) := \int \mathbb{1}\{\widehat{D}_\gamma(X^n\|Y^m) < \epsilon\}p_\theta(Y^m)\mathrm{d}Y^m$ is $\beta$-Lipschitz continuous for all $\theta \in \Theta$. Then, we have*

$$\lim_{\|X_0\|\to\infty} \mathrm{SC}^\theta_{n+1}(X_0) \leq -\frac{\beta\pi(\theta)}{\Lambda_n(1+\gamma)}\log\left(1 - \frac{1}{n^2}\right)^{n+1},$$

*where $\Lambda_n := \int \pi(\theta')F_{\theta'}(\epsilon)\mathrm{d}\theta'$. Furthermore, if $\lim_{n\to\infty}\Lambda_n$ exists and is non-zero, then the right-hand side of the above inequality converges to 0.*

The proof is in Appendix B.2. Through Theorem 1, we can see that the influence of contamination is reduced when we have enough data, even if the magnitude of the outlier $X_0$ is very large. Intuitively, an estimator has the redescending property if its IF first ascends away from zero as outliers become more pronounced, while the IF "redescends" towards zero as outliers become increasingly extreme. Since our analysis is a finite-sample analogue of the redescending property in the context of IF, this result implies the robustness of our method that an extreme outlier is automatically ignored.

### 3.5 Robustness on Estimation Error of Discrepancy

In Section 3.4, we showed the theoretical robustness of our method. Here, we experimentally investigate the robustness based on the estimation error of discrepancy. Figure 1 shows the discrepancy estimation error. In this figure, we show the errors between robust MMD and true MMD, and the KL- or $\gamma$-divergence estimator and true KL-divergence. We considered the 1-dimensional standard normal distribution $\mathcal{N}(0,1)$ and the contaminated 1-dimensional standard normal distribution $(1-\eta)\mathcal{N}(0,1)+\eta\mathcal{N}(10,1)$. Whereas the outliers negatively affect the KL divergence estimator [61], MONK-BCD [49] and our $\gamma$-divergence estimator are robust to an increase in the contamination rate $\eta$. Furthermore, if we choose the hyper-parameter $\gamma$ properly, our discrepancy estimator achieves comparable accuracy to the KL-divergence estimator in the non-contaminated case.

## 4 Asymptotic Analysis on ABC

In this section, we elucidate essential asymptotic properties, such as asymptotic unbiasedness and almost sure convergence, for our divergence estimator defined by Eq. (7). Furthermore, we analyze an asymptotic behavior of the ABC posterior distributions built on our divergence estimator.

### 4.1 Theoretical Analysis for $\gamma$-divergence Estimator

To confirm the validity of the proposed estimator in Eq. (7), we show two properties: the asymptotic unbiasedness and the almost sure convergence.

We show the asymptotic unbiasedness by assuming that $\mathcal{M}$, i.e., the support of $p$, has the following mild regularity condition. These conditions are commonly used for investigating the asymptotic properties of divergence estimators, e.g., the $\alpha$-divergence estimator in Poczos and Schneider [64].

**Assumption 1** (Restrictions on the domain $\mathcal{M}$ [64]). *We assume*

$$\inf_{0<\delta<1}\inf_{x\in\mathcal{M}}\frac{\mathcal{V}(\mathcal{B}(x,\delta)\cap\mathcal{M})}{\mathcal{V}(\mathcal{B}(x,\delta))}:=r_{\mathcal{M}}>0.$$

Assumption 1 states that the intersection of $\mathcal{M}$ with an arbitrary small ball having the center in $\mathcal{M}$ has a volume that cannot be arbitrarily small relative to the volume of the ball. It intuitively means that almost all points of $\mathcal{M}$ are in its interior.

Furthermore, we define the following function:

$$H(x,p,\delta,\omega):=\sum_{j=0}^{k-1}\left(\frac{1}{j!}\right)^{\omega}\Gamma(\kappa+j\omega)\left(\frac{p(x)+\delta}{p(x)-\delta}\right)^{j\omega}\times(p(x)-\delta)^{-\gamma}((1-\delta)\omega)^{-\kappa-j\omega}, \tag{8}$$

where $\Gamma(\cdot)$ is the gamma function defined as $\Gamma(z)=\int_0^\infty t^{z-1}\exp(-t)\mathrm{d}t$. Poczos and Schneider [64] used Assumption 1 to show a uniform variant of Lebesgue's lemma and to show that Definition 4 below is well-defined. The function in Eq. (8) appears in the upper bound of the moment for the $\omega$-powered conditional distribution function (CDF) when Assumption 1 holds (see Theorem 37 in Poczos and Schneider [64]). In addition, we impose some reasonable assumptions, which are also assumed in Poczos and Schneider [64].

**Definition 4** (Uniformly Lebesgue-approximable function [64]). *Denote by $L_1(\mathcal{M})$ the set of Lebesgue integrable functions on $\mathcal{M}$ and let $g\in L_1(\mathcal{M})$. The function $g$ is uniformly Lebesgue approximable on $\mathcal{M}$ if for any series $R_n\to 0$ and any $\delta>0$, there exists $n_0=n_0(\delta)\in\mathbb{Z}^+$ such that if $n>n_0$, then for almost all $x\in\mathcal{M}$,*

$$g(x)-\delta<\frac{\int_{\mathcal{B}(x,R_n)\cap\mathcal{M}}g(t)\mathrm{d}t}{\mathcal{V}(\mathcal{B}(x,R_n)\cap\mathcal{M})}<g(x)+\delta. \tag{9}$$

**Assumption 2** (Condition for $p$ and $q$ from Poczos and Schneider [64]). *The positive functions $p$ and $q$ are bounded away from zero and uniformly Lebesgue approximable. Furthermore, the expectations of the $l_2$-norm powered by $\kappa$ over $p$ and $q$ are bounded, i.e.,*

$$\int_{\mathcal{M}}\|x-y\|^{\kappa}p(y)\mathrm{d}y<\infty,\ \int_{\mathcal{M}}\|x-y\|^{\kappa}q(y)\mathrm{d}y<\infty,$$

*for almost all $x\in\mathcal{M}$. Furthermore, the following conditions hold:*

$$\int\int_{\mathcal{M}^2}\|x-y\|^{\kappa}p(y)p(x)\mathrm{d}y\mathrm{d}x<\infty,$$
$$\int\int_{\mathcal{M}^2}\|x-y\|^{\kappa}q(y)p(x)\mathrm{d}y\mathrm{d}x<\infty.$$

**Assumption 3** (Condition for powered CDF in $\mathcal{M}$ [64]). *The expectations of $H(x,p,\delta,1)$ and $H(x,q,\delta,1)$ are bounded as follows: $\exists\delta_0$ s.t. $\forall\delta\in(0,\delta_0)$,*

$$\int_{\mathcal{M}}H(x,q,\delta,1)p(x)\mathrm{d}x<\infty,\int_{\mathcal{M}}H(x,p,\delta,1)p(x)\mathrm{d}x<\infty.$$

This assumption indicates that the expectations of the $\omega$-powered CDFs of $p$ and $q$ are bounded, respectively. The expectations appear in the upper bound of the moment of the $\omega$-powered CDFs.

Since our method has a term involving an expectation with respect to $q$, we set the following assumption that is almost the same as the condition for the support of $p$.

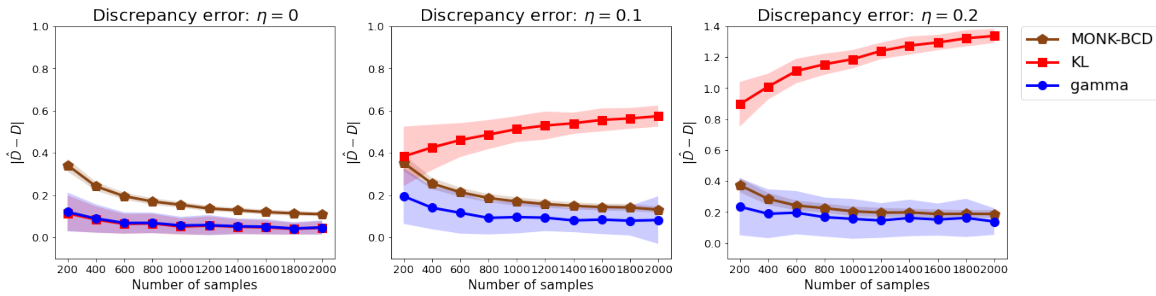


Figure 1: Experimental results for the absolute error of discrepancy. We compared the robustness of the discrepancy measure based on robust MMD (MONK-BCD) and divergence-based discrepancies (KL and ours). The mean $\pm$ std. values of each error are plotted (solid lines are mean values and shaded areas represent the range of mean $\pm$ std values). Lower values are better. The true discrepancy $D$ was estimated by $10^4$ Monte Carlo samples. We conducted experiments on each discrepancy for 100 times for various contamination rates ($\eta = 0, 0.1, 0.2$) and sample sizes ($200, 400, \ldots, 2000$). We set the hyper-parameter candidates of our $\gamma$-divergence estimator as $\gamma = (0.1, 0.2, 0.25, 0.4, 0.5, 0.6, 0.75, 0.9)$ and the one with the smallest mean score among them is displayed.

**Assumption 4** (Extra condition for powered CDF in $\mathcal{M}'$)**.** *The expectation of $H(y, q, \delta, 1)$ is bounded as*

$$\exists \delta_0 \ \text{s.t.} \ \forall \delta \in (0, \delta_0), \ \int_{\mathcal{M}'} H(y, q, \delta, 1) q(y) \mathrm{d}y < \infty.$$

This assumption means that the expectation of the $\omega$-powered CDF with respect to $q$ is bounded. The above expectation appears in the upper bound of the moment for the $\omega$-powered CDFs. Under these assumptions, the following theorem holds.

**Theorem 2** (Asymptotic unbiasedness)**.** *Let $0 < \gamma < k$ or $-k < \gamma < 0$. Suppose that Assumption 2 holds with $\kappa = \gamma$ and that Assumptions 3 and 4 hold. Also assume that $q$ is bounded from above. Then, $\widehat{D}_\gamma(X^n \| Y^m)$ defined in Eq. (7) is asymptotically unbiased, i.e.,*

$$\lim_{n,m \to \infty} \mathbb{E}\left[\widehat{D}_\gamma(X^n \| Y^m)\right] = D_\gamma(p \| q).$$

From this result, we can see that the asymptotic unbiasedness holds even if we set $\gamma$ as $-k < \gamma < 0$ (see Theorems 8 and 9 in Appendix D.1).

Next, we establish the almost sure convergence of our estimator.

**Theorem 3** (Almost sure convergence)**.** *Let $\gamma < k$. Suppose that Assumption 2 holds with $\kappa = \gamma$ and that Assumptions 3 and 4 hold. Also assume that $p$ and $q$ are bounded from above. Let $k(n)$ denote the number of neighbors applied at sample size $n$ such that $\lim_{n\to\infty} k(n) = \infty$, $\lim_{n\to\infty} n/k(n) = \infty$, $\lim_{m\to\infty} k(m) = \infty$ and $\lim_{m\to\infty} m/k(m) = \infty$. Then, our estimator converges almost surely to $D_\gamma(p\|q)$, that is,*

$$\widehat{D}_\gamma(X^n \| Y^m) \overset{a.s.}{\to} D_\gamma(p \| q).$$

The proofs for these theorems are in Appendices D.1 and D.2. Note that in the proofs of Theorems 2 and 3, we cannot reuse the known theoretical results for the divergence estimators in the $f$-divergence class [61, 64] and it required us to newly show several asymptotic properties specifically for $\gamma$-divergence estimation, which are given in Appendices C and D.

### 4.2 Asymptotic Property of ABC Posterior Distributions with $\gamma$-divergence Estimator

Now we analyze whether the ABC posterior based on our robust discrepancy measure can accurately estimate the parameter $\theta$ with small exact $\gamma$-divergence $D_\gamma(p_{\theta^*} \| p_\theta)$ asymptotically.

According to Theorem 1 in [42], the asymptotic ABC posterior is a restriction of the prior $\pi$ to the region $\{\theta \in \Theta : D(p_{\theta^*} \| p_\theta) < \epsilon\}$ with appropriate scaling. Combining this with the almost sure convergence of $\widehat{D}_\gamma(X^n \| Y^m)$ established in Theorem 3, we can obtain the following corollary.

**Corollary 1** (Asymptotic ABC posterior with $\gamma$-divergence estimator)**.** *Suppose that Assumptions 2-4 are satisfied with $\kappa = \gamma$. Let $n \to \infty$ and $m/n \to \alpha > 0$. Let $\pi(\theta | D_\gamma(p_{\theta^*} \| p_\theta) < \epsilon)$ be the posterior under $D_\gamma(p_{\theta^*} \| p_\theta) < \epsilon$. If $\widehat{D}_\gamma(X^n \| Y^m)$ is used as the data discrepancy measure in Algorithm 1, the ABC posterior distribution satisfies*

$$\pi(\theta | X^n; \widehat{D}_\gamma, \epsilon) \to \pi(\theta | D_\gamma(p_{\theta^*} \| p_\theta) < \epsilon),$$

*almost surely, and therefore*

$$\lim_{n,m\to\infty} \pi(\theta | X^n; \widehat{D}_\gamma, \epsilon) \propto \pi(\theta) \mathbb{1}\{D_\gamma(p_{\theta^*} \| p_\theta) < \epsilon\},$$

*almost surely.*

*Proof sketch.* In the same way as Jiang et al. [42], we use Lévy's upward theorem (enabled by Theorem 3; see Theorem 4 in Appendix C) to $Z_n = \mathbb{1}\{\widehat{D}_\gamma(X^n \| Y^\infty) < \epsilon\}$ and apply the dominated convergence theorem [78] to complete the proof. □

Corollary 1 shows that the ABC posterior based on our estimator converges to the maximum likelihood estimator minimizes the *exact* $\gamma$-divergence between the empirical distribution of $p_{\theta^*}$ and $p_\theta$. Thus, ABC with our $\gamma$-divergence estimator asymptotically collects the $\theta$ with small $\gamma$-divergence.

# 5 Experiments

In this section, we report the performance of our estimator through five benchmark experiments of ABC. Here, we confirm that the ABC with our discrepancy measure has immunity against heavily contaminated data.

## 5.1 Settings

We set $n = m$ following Jiang et al. [42] to prevent the resulting ABC posterior of the indirect method to be overprecise [25] and to avoid arbitrariness in the experiments. The tolerance threshold $\epsilon$ was adaptively initialized so that $0.5\%$ of proposed parameters $\theta$ were accepted in each experiment by calculating each discrepancy measure $10^3$ times. Furthermore, we artificially generated the $n$ i.i.d. observed data from $G(X_i)$ and replaced them by some outliers from $\mathcal{N}(10, 1)$, where $G(X_i)$ is an observed data distribution. In short, the contaminated data can be expressed as $(1-\eta)G(X_i) + \eta\mathcal{N}(10, 1)$ in each dimension. In addition, we varied the contamination level $\eta$ in $\{0, 0.1, 0.2\}$ to confirm the robustness. The hyper-parameter $\gamma$ was selected from $\{0.1, 0.2, 0.25, 0.4, 0.5, 0.6, 0.75, 0.9\}$ for our method.

We measured the accuracy by the simulation error based on the *energy distance*, which is a standard metric for distributions in statistics and has been used in the ABC literature, e.g., Kajihara et al. [43]. This allows us to directly compare the distributions between the non-contaminated observed data and the synthesized data simulated with the estimated parameter. We approximated the MAP estimator $\hat{\theta}_{\mathrm{MAP}}$ of the ABC posterior by kernel density estimation with the Gaussian kernel with the bandwidth parameter $n^{-1/(d+4)}$, that is known as Scott's Rule [70].

From each of the 10 different models, we sampled the data and performed the ABC (Algorithm 1) with $T = 10^5$. We repeated the procedure independently for 10 times, and reported the average results. The results with the standard errors are reported in Appendix G.2.6, and the results of the mean-squared error (MSE) between $\hat{\theta}_{\mathrm{MAP}}$ and the true parameter are also reported in Appendix G.

For our method, we conducted experiments independently for several $\gamma$ and displayed the one with the smallest mean score of the energy distance and the MSE among them. The full results are reported in Figures 6–10 and 13–17 in Appendix G. Furthermore, we compared the ABC posteriors of our method and that of the second-best method. These results are reported in Figures 18–22 in Appendix G.3.

## 5.2 Models and Results

Here, we summarize the model settings and the results of each experiment. The details of the baseline discrepancies and the model architectures are shown in Appendices E and F.

**Gaussian Mixture Model (GM):** The univariate Gaussian mixture model is the most basic benchmark setup in the ABC literature [81, 42]. We adopted a bivariate Gaussian mixture model with the true parameters $p^* = 0.3$, $\mu_0^* = (0.7, 0.7)$ and $\mu_1^* = (-0.7, -0.7)$, where $p^*$ is the mixture weight and $\mu_0^*, \mu_1^*$ are the means of the component distributions. The variances are fixed as $0.5I - 0.3I^\top$ and $0.25I$, where $I$ is the identity matrix of size $(2, 2)$.

From the experimental results in Figure 2, we can see that our method achieves a better performance when the observed data are contaminated, whereas the other methods fail to give good scores. In addition, in terms of the MSE, our method outperforms the baseline methods (see Appendix G) under contamination. From the results in Figure 4, we can confirm that the ABC posterior with our method places high density around the ground-truth parameter, whereas the baseline method fails to do so.

***M/G/*1-queuing Model (MG1):** Queuing models are an example of stochastic models which are easy to sample from but have intractable likelihoods [27]. The $M/G/1$-queuing model has been often used in the ABC literature [27, 42]. This model has three parameters: $\theta = (\theta_1, \theta_2, \theta_3)$. We adopted this model with the true parameter $\theta^* = (1, 5, 0.2)$.

From the experimental results in Figure 2, we can see that our method outperforms the other methods even if the data has no contamination. In addition, our method also achieves better performance in terms of the MSE scores than the baseline methods (see Appendix G). Figure 5 indicates that the ABC posterior with our method places high density around the ground-truth parameter, e.g., for $\theta_2$. On the other hand, the CAD via boosting places higher density around the wrong parameter than our method, e.g., for $\theta_3$.

**Bivariate Beta Model (BB):** The bivariate beta model can be used to model data sets exhibiting positive or negative correlation [2]. This model was originally proposed as a model with 8 parameters $\theta = (\theta_1, \ldots, \theta_8)$ by Arnold and Ng [2], and Crackel and Flegal [19] later reconsidered its 5-parameter sub-model by restricting to $\theta_3, \theta_4, \theta_5 =$

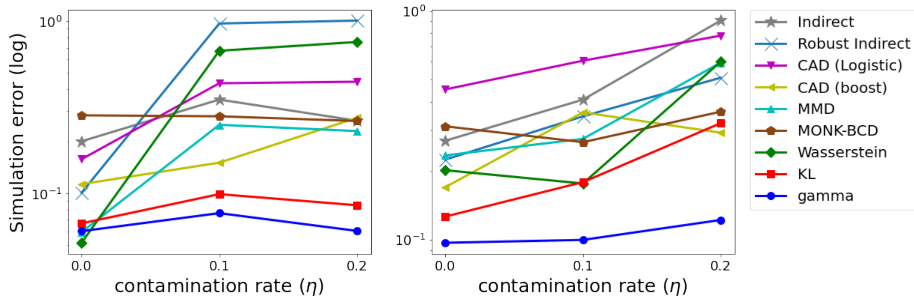


Figure 2: Simulation error for the GM (left) and the MG1 model (right) experiments.

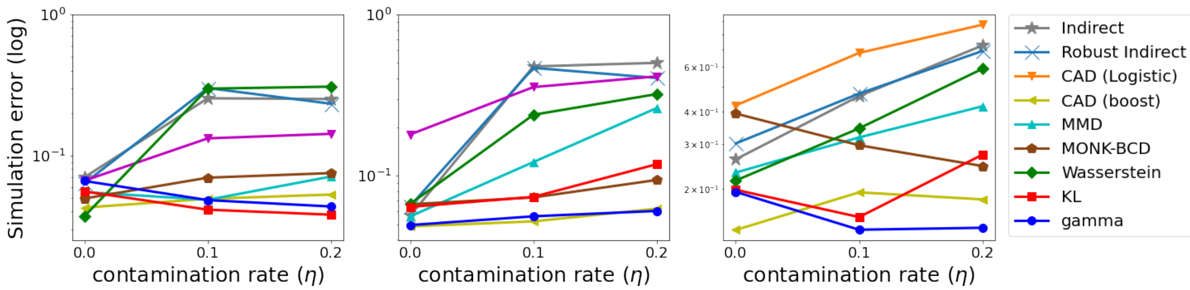


Figure 3: Simulation error for the BB (left), the MA2 (center), and the GK model (right) experiments.

$0$. Jiang et al. [42] used the $5$-parameter models for ABC experiments and therefore we also adopted this with the true parameter $\theta^* = (3, 2.5, 2, 1.5, 1)$ as a benchmark model.

From the experimental results in Figure 3, the KL- and the $\gamma$-divergence based methods achieve better performances than those of the baseline methods when the observed data are heavily contaminated. The Wasserstein method achieves a better performance than the others when the data has no contamination; however, the performance becomes significantly worse when the contamination occurs. In addition, the CAD with boosting method achieves a better performance in terms of the MSE; however, the simulation error is worse than the KL- and the $\gamma$-divergence based methods (see Appendix G). Figure 5 shows that the ABC posterior with our method places higher density around the ground-truth parameter than the KL method. In addition, the KL method sometimes places high density around the wrong parameter, e.g., for $\theta_8$. However, in this experiment, the simulation error of the KL method is slightly better than that of our method. This indicates that carefully tuning the hyperparameter $\gamma$ is important in our method.

**Moving-average Model of Order 2 (MA2):** The moving-average model is often used for modeling univariate time series. Marin et al. [53] used the moving-average model of order $2$ as a benchmark model for ABC. We adopted this model with $10$-length time series. For the unobserved noise distribution, we used Student's t-distribution with $5$ degrees of freedom. We set the true parameter $\theta^* = (0.6, 0.2)$.

From the experimental results in Figure 3, our method and the CAD with boosting outperform the other methods. The CAD with boosting achieves comparable performance to our method even if the observed data are contaminated; however, in terms of the time cost, our method is better than that of this method because the gradient boosting has $\mathcal{O}(Kd(n+m)+(n+m)\log B)$ time cost, where $K$ is the total number of trees and $B$ is s the maximum number of rows in each block (see Chen and Guestrin [16]). In addition, in terms of the MSEs, our method achieves better performance when the observed data are heavily contaminated (see Appendix G). In Figure 4, we found that the ABC posterior with our method places high density around the ground-truth parameter. On the other hand, the CAD via boosting places higher density around the wrong parameter than our method.

**Multivariate $g$-and-$k$ Distribution (GK):** The univariate $g$-and-$k$ distribution is a generalization of the standard normal distribution with extra parameters: the *skewness* and the *kurtosis*. This distribution is known to have no analytical form of the density function, and the numerical evaluation of the likelihood function is costly [66]. Thus, it is a model for which ABC is specifically suited [27, 1]. Some studies [24, 50] also considered the multivariate $g$-and-$k$ distribution. We adopted the multivariate model proposed by Drovandi and Pettitt [24] with the true parameters $A^* = 3$, $B^* = 1$, $g^* = 2$, $k^* = 0.5$ and $\rho^* = -0.3$, where $A^*, B^*, g^*, k^*$ control the location, the scale, the skewness and the kurtosis, respectively.

From the experimental results in Figure 3, we can see that

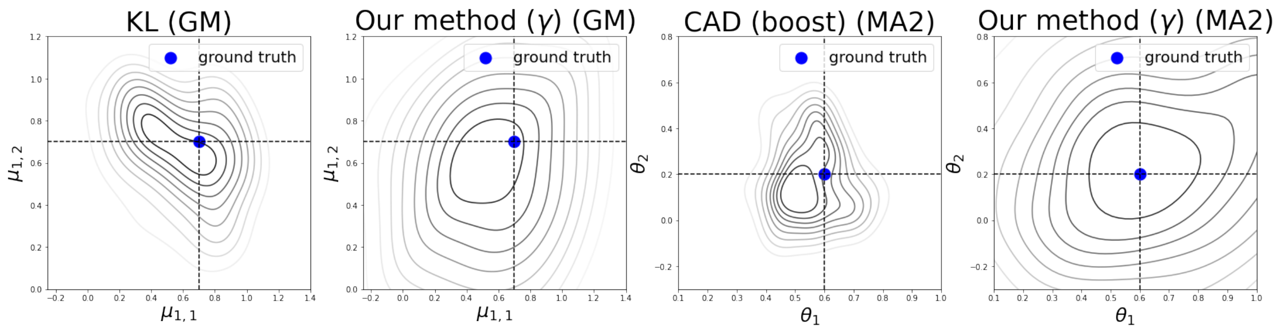


Figure 4: ABC posterior distributions of the GM and MA2 model experiments for $\eta = 0.2$ (excerpted).

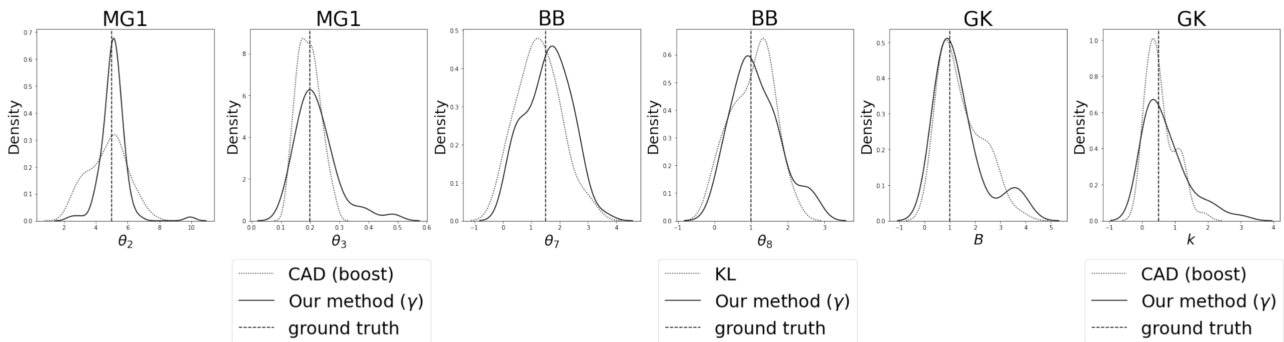


Figure 5: ABC posterior distributions of the MG1, BB, and GK model experiments for $\eta = 0.2$ (excerpted).

our method achieves better performance even if the observed data are contaminated, although the other baseline methods fail to give good scores. In addition, in terms of the MSE, our method outperforms the other baseline methods when the observed data have heavy contamination (see Appendix G). Figure 5 shows that the ABC posterior with our method places slightly higher density on the ground-truth parameter than that of the CAD via boosting.

## 6 Conclusion and Discussion

We have proposed a $\gamma$-divergence estimator and used it as a robust data discrepancy for ABC. We theoretically have guaranteed its robustness against outliers and its desirable asymptotic properties, i.e., the asymptotic unbiasedness and the almost sure convergence to the approximate posterior. In addition, we have shown the redescending property of the ABC posterior of our method, indicating the high robustness of our method against extreme outliers. Through the experiments on benchmark models, we empirically confirmed that our method is robust against heavy contamination by outliers.

Our work has two limitations: (i) our method can become statistically inefficient in high-dimensional cases (the curse of dimensionality) due to the $k$-NN based density estimation, and therefore (ii) its performance has only been confirmed in some low-dimensional cases. To overcome these limitations, in the future, we will consider extending our method to a non-parametric estimation that can handle high dimensions and conduct experiments for more realistic high-dimensional cases. Furthermore, on the basis of our idea, we plan to develop outlier-robust methods for other ABC approaches, e.g., ABC without discrepancies [59, 32, 77].

It is worth mentioning that there have been several studies on general losses within Bayesian procedures [10, 46], and recent studies have connected the ideas of these studies with ABC [69]. Following these studies, it would be interesting to see whether our method stands in the framework of generalized approximate Bayesian inference under the condition of ABC with a general loss function [69].

In addition, research on the consistency and robustness of Bayesian estimation against model misspecification has attracted attention recently. For example, Cherief-Abdellatif and Alquier [17] has proposed Bayesian estimation based on a pseudo-likelihood by using MMD and has theoretically shown that it is effective in this problem setting. We also explore the potential of our method in the context of model misspecification.

### Acknowledgements

MF and TT were supported by RIKEN Junior Research Associate Program. MF and TT were supported by Toyota/Dwango AI scholarship. MF was supported by JST CREST including AIP challenge program. TT was supported by Masason Foundation. MS was supported by KAKENHI 17H00757. We appreciate Dr. Ikko Yamane, Kento Nozawa, Dr. Yoshihiro Nagano, and Han Bao for their kind effort to maintain the experimental environment.

## References

[1] D. Allingham, Robert King, and K. Mengersen. Bayesian estimation of quantile distributions. *Statistics and Computing*, 19:189–201, 2009.

[2] B. C. Arnold and H. K. T. Ng. Flexible Bivariate Beta Distributions. *Journal of Multivariate Analysis*, 102 (8):1194–1202, 2011.

[3] Franz Aurenhammer and Rolf Klein. Voronoi Diagrams. In *Handbook of Computational Geometry*, pages 201–290. Elsevier, 2000. ISBN 978-0-444-82537-7. doi: 10.1016/B978-044482537-7/50006-1. URL https://linkinghub.elsevier.com/retrieve/pii/B9780444825377500061.

[4] Ayanendranath Basu, Ian R. Harris, Nils L. Hjort, and M. C. Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85 (3):549–559, 09 1998.

[5] Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Commun. ACM*, 18:509–517, September 1975. ISSN 0001-0782.

[6] Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Space in Probability and Statistics*. Springer Science & Business Media, 2004.

[7] Espen Bernton, Pierre Jacob, Mathieu Gerber, and Christian Robert. Inference in generative models using the Wasserstein distance. *arXiv preprint arXiv:1701.05146*, abs/1701.05146, 2017.

[8] Gérard Biau and Luc Devroye. The k-nearest neighbor density estimate. In Gérard Biau and Luc Devroye, editors, *Lectures on the Nearest Neighbor Method*, Springer Series in the Data Sciences, pages 25–32. Springer International Publishing, Cham, 2015. ISBN 978-3-319-25388-6. doi: 10.1007/978-3-319-25388-6_3. URL https://doi.org/10.1007/978-3-319-25388-6_3.

[9] Patrick Billingsley. *Probability and Measure*. Wiley, third edition, 1995. ISBN 0-471-00710-2. URL http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0471007102. Published: Hardcover.

[10] P. G. Bissiri, C. C. Holmes, and S. G. Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130, 2016.

[11] M. G. B. Blum and V. C. Tran. HIV with contact-tracing: a case study in Approximate Bayesian Computation. *Biostatistics*, 11(4):644–660, 2010.

[12] M. G. B. Blum, M. A. Nunes, D. Prangle, and S. A. Sisson. A Comparative Review of Dimension Reduction Methods in Approximate Bayesian Computation. *Statistical Science*, 28(2):189–208, 2013.

[13] Michael Blum and Olivier François. Non-linear regression models for Approximate Bayesian Computation. *Statistics and Computing*, 20:63–73, 2010.

[14] Rainer Burkard, Mauro Dell'Amico, and Silvano Martello. *Assignment Problems*. Society for Industrial and Applied Mathematics, 2009.

[15] E. Cameron and A. N. Pettitt. Approximate Bayesian Computation for astronomical model analysis: a case study in galaxy demographics and morphological transformation at high redshift. *Monthly Notices of the Royal Astronomical Society*, 425(1):44–65, 2012.

[16] Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, page 785–794, 2016.

[17] Badr-Eddine Cherief-Abdellatif and Pierre Alquier. MMD-Bayes: Robust Bayesian Estimation via Maximum Mean Discrepancy. In *Proceedings of The 2nd Symposium on Advances in Approximate Bayesian Inference*, volume 118, pages 1–21, 2020.

[18] Jason Christopher, Caelan Lapointe, Nicholas Wimer, Torrey Hayden, Ian Grooms, Gregory Rieker, and Peter Hamlington. Parameter Estimation for a Turbulent Buoyant Jet using Approximate Bayesian Computation. In *55th AIAA Aerospace Sciences Meeting*, 2017.

[19] Roberto Crackel and James Flegal. Bayesian inference for a flexible class of bivariate beta distributions. *Journal of Statistical Computation and Simulation*, 87(2): 295–312, 2017.

[20] Christophe Croux and Catherine Dehon. Robust linear discriminant analysis using S-estimators. *Canadian Journal of Statistics*, 29(3):473–493, 2001.

[21] Marco Cuturi. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In *Advances in Neural Information Processing Systems 26 (NeurIPS)*, pages 2292–2300, 2013.

[22] Marco Cuturi and Arnaud Doucet. Fast Computation of Wasserstein Barycenters. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pages 685–693, 2014.

[23] Nguyen Viet Dang. Complex powers of analytic functions and meromorphic renormalization in QFT. *arXiv:1503.00995 [math-ph]*, March 2015. URL http://arxiv.org/abs/1503.00995.

[24] Christopher C. Drovandi and Tony Pettitt. Estimation of parameters for macroparasite population evolution using approximate Bayesian computation. *Biometrics*, 67(1):225–233, 2011.

[25] Christopher C. Drovandi, Anthony Pettitt, and Anthony Lee. Bayesian Indirect Inference Using a Parametric Auxiliary Model. *Statistical Science*, 30(1): 72–95, 2015.

[26] Herbert Edelsbrunner, Joseph O'Rourke, and Raimund Seidel. Constructing arrangements of lines and hyperplanes with applications. *SIAM Journal on Computing*, 15(2), 1986. URL https://research-explorer.app.ist.ac.at/record/4105.

[27] Paul Fearnhead and Dennis Prangle. Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation: Semi-automatic Approximate Bayesian Computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):419–474, 2012.

[28] Jiashi Feng, Huan Xu, Shie Mannor, and Shuicheng Yan. Robust Logistic Regression and Classification. In *Advances in Neural Information Processing Systems 27 (NeurIPS)*, pages 253–261, 2014.

[29] Hironori Fujisawa and Shinto Eguchi. Robust Parameter Estimation with a Small Bias Against Heavy Contamination. *Journal of Multivariate Analysis*, 99 (9):2053–2081, October 2008.

[30] Futoshi Futami, Issei Sato, and Masashi Sugiyama. Variational Inference based on Robust Divergences. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.

[31] A. Gleim and C. Pigorsch. Approximate Bayesian computation with indirect summary statistics. *Draft paper: http://ect-pigorsch.mee.uni-bonn.de/data/research/papers*, 2013.

[32] David Greenberg, Marcel Nonnenmacher, and Jakob Macke. Automatic posterior transformation for likelihood-free inference. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

[33] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A Kernel Two-Sample Test . *Journal of Machine Learning Research*, 13:723–773, 2012.

[34] G.R. Grimmett and D.R. Stirzaker. *Probability and random processes*. Oxford university press, 2001.

[35] Michael U. Gutmann, Ritabrata Dutta, Samuel Kaski, and Jukka Corander. Likelihood-free inference via classification. *Statistics and Computing*, 28(2):411–425, 2018.

[36] Frank R. Hampel, editor. *Robust Statistics: The Approach Based on Influence Functions*. Wiley Series in Probability and Mathematical Statistics. Wiley, New York, digital print edition, 2005. ISBN 978-0-471-73577-9. OCLC: 255133771.

[37] Wolfgang Karl Härdle, Marlene Müller, Stefan Sperlich, and Axel Werwatz. *Nonparametric and Semiparametric Models*. Springer, 01 2006. ISBN 978-3-642-62076-8.

[38] Peter J. Huber. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35(1):73–101, 1964.

[39] P.J. Huber, J. Wiley, and W. InterScience. *Robust statistics*. Wiley New York, 1981.

[40] Seung jean Kim, Alessandro Magnani, and Stephen Boyd. Robust Fisher Discriminant Analysis. In *Advances in Neural Information Processing Systems 18 (NeurIPS)*, pages 659–666, 2006.

[41] Jack Jewson, Jim Q. Smith, and Chris Holmes. Principles of Bayesian Inference Using General Divergence Criteria. *Entropy*, 20(6), 2018. ISSN 1099-4300.

[42] Bai Jiang, Tung-Yu Wu, and Wing Hung Wong. Approximate Bayesian Computation with Kullback-Leibler Divergence as Data Discrepancy. In *Proceedings of the 21th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 84, 2018.

[43] Takafumi Kajihara, Motonobu Kanagawa, Keisuke Yamazaki, and Kenji Fukumizu. Kernel Recursive ABC: Point Estimation with Intractable Likelihood. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 2400–2409, 2018.

[44] Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A Least-Squares Approach to Direct Importance Estimation. *Journal of Machine Learning Research*, 10:1391–1445, 2009.

[45] Jeremias Knoblauch, Jack E Jewson, and Theodoros Damoulas. Doubly Robust Bayesian Inference for Non-Stationary Streaming Data with $\beta$-Divergences. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[46] Jeremias Knoblauch, J. Jewson, and T. Damoulas. Generalized Variational Inference: Three arguments for deriving new Posteriors. *arXiv preprint arXiv:1904.02063*, abs/1904.02063, 2019.

[47] Jan Kremer, Kristoffer Stensbo-Smidt, Fabian Gieseke, Kim Pedersen, and Christian Igel. Big Universe, Big Data: Machine Learning and Image Analysis for Astronomy. *IEEE Intelligent Systems*, 32:16–22, 03 2017.

[48] Nikolai Leonenko, Luc Pronzato, and Vippal Savani. A class of Rényi information estimators for multidimensional densities. *Annals of Statistics*, 36(5):2153–2182, 2008.

[49] Matthieu Lerasle, Zoltan Szabo, Timothée Mathieu, and Guillaume Lecue. MONK Outlier-Robust Mean Embedding Estimation by Median-of-Means. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pages 3782–3793, 2019.

[50] Jingjing Li, David Nott, Yanan Fan, and Scott Sisson. Extending approximate Bayesian computation methods to high dimensions via Gaussian copula. *Computational Statistics & Data Analysis*, 106:77–89, 2015.

[51] D. O. Loftsgaarden and C. P. Quesenberry. A Nonparametric Estimate of a Multivariate Density Function. *The Annals of Mathematical Statistics*, 36(3): 1049–1051, 06 1965.

[52] Songrit Maneewongvatana and David M. Mount. On the Efficiency of Nearest Neighbor Searching with Data Clustered in Lower Dimensions. In *Computational Science – ICCS 2001*, pages 842–851, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg.

[53] Jean Michel Marin, Pierre Pudlo, Christian P. Robert, and Robin J. Ryder. Approximate Bayesian Computational methods. *Statistics and Computing*, 22: 1167–1180, 2012.

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

[55] Ricardo A. Maronna. *Robust Statistics: Theory and Methods (with R)*. Wiley Series in Probability and Statistics. WIley, Hoboken, NJ, second edition, 2019. ISBN 978-1-119-21467-0 978-1-119-21466-3.

[56] Boris Mityagin. The Zero Set of a Real Analytic Function. *arXiv:1512.07276 [math]*, December 2015. URL http://arxiv.org/abs/1512.07276.

[57] Pierre Moral, Arnaud Doucet, and Ajay Jasra. An Adaptive Sequential Monte Carlo Method for Approximate Bayesian Computation. *Statistics and Computing*, 22(5):1009–1020, 2012.

[58] Tomoyuki Nakagawa and Shintaro Hashimoto. Robust Bayesian inference via $\gamma$-divergence. *Communications in Statistics - Theory and Methods*, 49(2):343–360, 2020.

[59] George Papamakarios and Iain Murray. Fast $\epsilon$-free inference of simulation models with bayesian conditional density estimation. In *Advances in Neural Information Processing Systems 29 (NeurIPS)*, pages 1028–1036, 2016.

[60] Mijung Park, Wittawat Jitkrittum, and Dino Sejdinovic. K2-ABC: Approximate Bayesian Computation with Kernel Embeddings. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 51 of *Proceedings of Machine Learning Research*, pages 398–407, 2016.

[61] Fernando Pérez-Cruz. Kullback-Leibler divergence estimation of continuous distributions. *2008 IEEE International Symposium on Information Theory*, pages 1666–1670, 2008.

[62] G.W. Peters and S.A. Sisson. Bayesian inference, Monte Carlo sampling and operational risk. *Journal of Operational Risk*, 1(3):27–50, 2006.

[63] G.W. Peters, S.A. Sisson, and Y. Fan. Likelihood-free Bayesian inference for $\alpha$-stable models. *Computational Statistics & Data Analysis*, 56(11):3743–3756, 2012.

[64] Barnabas Poczos and Jeff Schneider. On the Estimation of $\alpha$-Divergences. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 15, 2011.

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

[66] G. Rayner and H. Macgillivray. Numerical maximum likelihood estimation for the g-and-k and generalized g-and-h distributions. *Statistics and Computing*, 12(1): 57–75, 2002.

[67] Walter Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, 1976.

[68] Erlis Ruli, Nicola Sartori, and Laura Ventura. Robust approximate Bayesian inference. *Journal of Statistical Planning and Inference*, 205:10–22, 2020.

[69] Sebastian M Schmon, Patrick W Cannon, and Jeremias Knoblauch. Generalized Posteriors in Approximate Bayesian Computation. In *Third Symposium on Advances in Approximate Bayesian Inference*, 2021.

[70] David W Scott. *Multivariate density estimation: theory, practice, and visualization; 2nd ed.* Wiley series in probability and statistics. Wiley, Hoboken, NJ, 2015. doi: 10.1002/9781118575574.

[71] S. A. Sisson, Y. Fan, and Mark M. Tanaka. Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 104(6):1760–1765, 2007.

[72] S. A. Sisson, Y. Fan, and Mark M. Tanaka. Correction for Sisson et al., "Sequential Monte Carlo without likelihoods". *Proceedings of the National Academy of Sciences*, 106(39):16889–16889, 2009.

[73] A. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert Space Embedding for Distributions. *Algorithmic Learning Theory: 18th International Conference (ALT 2007)*, pages 13–31, 2007.

[74] Guillaume Staerman, Pierre Laforgue, Pavlo Mozharovskyi, and Florence d'Alché Buc. When

OT meets MoM: Robust estimation of Wasserstein Distance. *arXiv preprint arXiv:2006.10325*, abs/2006.10325, 2020.

[75] Masashi Sugiyama, Taiji Suzuki, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60:699–746, 02 2008.

[76] Simon Tavare, David J Balding, Robert C Griffiths, and Peter Donnelly. Inferring Coalescence Times from DNA Sequence Data. *Genetics*, 162(2):505–518, 1997.

[77] O. Thomas, R. Dutta, J. Corander, S. Kaski, and M.U. Gutmann. Likelihood-Free Inference by Ratio Estimation. *Bayesian Analysis*, 2020.

[78] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[79] Q. Wang, S. R. Kulkarni, and S. Verdu. Divergence Estimation for Multidimensional Densities Via $k$-Nearest-Neighbor Distances. *IEEE Transactions on Information Theory*, 55(5):2392–2405, 2009.

[80] Daniel Wegmann, Christoph Leuenberger, and Laurent Excoffier. Efficient Approximate Bayesian Computation Coupled With Markov Chain Monte Carlo Without Likelihood. *Genetics*, 182(4):1207–1218, 2009.

[81] Richard Wilkinson. Approximate Bayesian Computation (ABC) gives exact results under the assumption of model error. *Statistical applications in genetics and molecular biology*, 12:1–13, 2013.

[82] Michael P Windham. Robustifying model fitting. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 599–609, 1995.

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