(Article begins on next page)

19 May 2024

ORIGINAL PAPER

Weighted approximate Bayesian computation via Sanov's theorem

Cecilia Viscardi[1] · **Michele Boreale[1]** · **Fabio Corradi[1]**

Abstract

We consider the problem of sample degeneracy in Approximate Bayesian Computation. It arises when proposed values of the parameters, once given as input to the generative model, rarely lead to simulations resembling the observed data and are hence discarded. Such "poor" parameter proposals do not contribute at all to the representation of the parameter's posterior distribution. This leads to a very large number of required simulations and/or a waste of computational resources, as well as to distortions in the computed posterior distribution. To mitigate this problem, we propose an algorithm, referred to as the Large Deviations Weighted Approximate Bayesian Computation algorithm, where, via Sanov's Theorem, strictly positive weights are computed for all proposed parameters, thus avoiding the rejection step altogether. In order to derive a computable asymptotic approximation from Sanov's result, we adopt the information theoretic "method of types" formulation of the method of Large Deviations, thus restricting our attention to models for i.i.d. discrete random variables. Finally, we experimentally evaluate our method through a proof-of-concept implementation.

Keywords ABC · Large deviation theory · Method of types · Sample degeneracy · ESS · Importance sampling

1 Introduction

Approximate Bayesian Computation (ABC) is a broad class of methods allowing Bayesian inference on parameters governing complex models. For such models, computing the likelihood, either analytically or numerically, is typically unfeasible. To

✉ Cecilia Viscardi
   cecilia.viscardi@unifi.it

1   Università di Firenze, DiSIA, Florence, Italy

 Springer

overcome this critical problem, ABC dispenses with exact likelihood computation, and only requires the ability of simulating pseudo-data by sampling observations from a *generative model*, as detailed in Sect. 2.

In the literature, a variety of ABC methods have been proposed, see Sisson et al (2018, Ch. 4), and for recent reviews Lintusaari et al. (2017), Karabatsos et al. (2018). In the vast majority of these methods, the approximate likelihood function takes positive values only when the distance between the simulated and the observed data is lower than a predefined threshold. In other words, most ABC schemes involve—implicitly or explicitly—a rejection step, which often leads to discarding a huge number of proposals. This results in a waste of computational resources and/or in an inadequate sample size, that is, in *sample degeneracy*. Sample degeneracy may also cause serious distortions in the form of the approximate posterior distribution, at least when the number of iterations is not large enough. Indeed, accepting poor parameter proposals, i.e., those producing simulated data very rarely resembling the observed data, is a rare event. In the lack of accepted values, the posterior probability of such proposals will be approximated just as zero, in turn resulting in a distortion in the tails. This may be especially problematic for posterior distributions with long tails.

Our idea is to mitigate the problem of sample degeneracy by improving the approximation of the likelihood function. In particular, we speculate that taking into account the positive, however small, probability of rare events, i.e., poor proposals leading to simulated data resembling the observed data, allows avoiding the rejection step altogether and weighting all parameter proposals. To this end, we resort to the theory of Large Deviations (LDT). Our aim is to show how LDT provides a convenient way to define an approximate likelihood, as well as guarantees of its convergence to the true likelihood as the size of pseudo-datasets goes to infinity. In order to make the incorporation of LDT into ABC as smooth as possible, we rely on one of the less general formulations of Sanov's theorem. Accordingly, we only consider models for discrete i.i.d. random variables which, despite their apparent simplicity, will be shown to be of interest in several applications of ABC. This allows adopting a straightforward information theoretic formulation of LDT known as the method of types (Csiszár 1998; Cover and Thomas 2006).

Related work In the literature, there have been many proposals aimed at improving the computational efficiency of basic ABC. Prangle (2016) proposed *Lazy* ABC, which saves computing time by abandoning simulations likely to lead to a poor match between the simulated and the observed data. To this end, at each iteration, the simulation is given up with a probability depending on the probability of its acceptance and on the expected required time for its completion. Unlike our method, Lazy ABC does not avoid rejection, but rather accelerates the process leading to discarding a proposal.

Another way to improve computational efficiency is to consider proposal distributions closer to the posterior on the parameter space, employing sophisticated sampling methods, such as MCMC (Marjoram et al. 2003), Population Monte Carlo (Beaumont et al. 2009) and Sequential Monte Carlo (Del Moral et al. 2012). In the same vein, Chiachio et al. (2014) proposed a sequential way of achieving computational efficiency by overcoming the difficulties in getting samples resembling the observed data. This latter was the first attempt to improve the acceptance rate by adopting a rare-event approach. In particular, Chiachio et al. (2014) combine the ABC scheme with a rare-

event sampler that draws conditional samples from a nested sequence of subdomains. However, even this method cannot completely avoid rejections, and only partially mitigates the sample degeneracy problem.

In order to tackle the problem more systematically, clever proposal distributions should be combined with better approximations to the likelihood. Accordingly, Prangle et al. (2018) also resorted to a sequential approach, but explicitly considering a likelihood estimate that takes into account the probability of rare events. As a comparison, our method evaluates the probabilities of rare events based on theoretical results (LDT), rather than on Monte Carlo estimates of tail probabilities. Moreover, they focus on continuous data by showing that extensions to discrete data can be challenging and require application-specific solutions; in contrast, the method of types provides a natural way of dealing with discrete random variables by summarizing data via empirical distributions, thus avoiding the common practice of summarizing data by selecting ad hoc summary statistics.

Other methods have been proposed avoiding the selection of summary statistics and relying on empirical distributions. In particular, Park et al. (2016) rely on the maximum mean discrepancy between the embeddings of the simulated and the observed empirical distributions. They avoid rejection by weighting each parameter proposal by means of a kernel function defined on a non-compact support. Other interesting methods involve the Wasserstein distance (Bernton et al. 2019) or the Kullback–Leibler divergence (Jiang 2018) as measures of the discrepancy between observed and simulated data. In particular, Jiang (2018) approximates the likelihood by means of an estimator of the Kullback–Leibler divergence between the unknown distribution of the data given the true parameter, and given the parameter sampled at the current iteration. Exploiting the fact that the maximum likelihood estimator is the one minimizing that Kullback–Leibler divergence, they prove that their approximate posterior distribution converges to a restriction of the prior distribution on the region in which the above mentioned divergence is smaller than a predefined threshold. Although most of the above mentioned methods apply to continuous data, we note that ABC applications to discrete data appear frequently in population genetics, epidemiology, ecology and system biology (see Beaumont (2010) for an overview of the applications of ABC in these fields). In particular, in population genetics, discrete (possibly i.i.d.) data representing the genotyping at a few loci of different (unrelated) individuals have often been summarized through their empirical distributions (Marjoram et al. 2003; Buzbas and Rosenberg 2015, among others).

A very different way of bypassing the selection of summary statistics relies on the random forest method (Raynal et al. 2019). Here, regression random forests are trained by using a training-set composed of a large number of parameter proposals and pseudo-datasets sampled from the prior distribution and the generative model, respectively. Since all the summary statistics are involved as covariates, summary selection is avoided. The output of the algorithm is the predicted expected value of an arbitrary function of interest on the parameter space, conditional on the observed data.

Structure of the paper The rest of this paper is structured as follows. Section 2 contains background and preliminaries on ABC, focusing on the importance sampling scheme

and the sample degeneracy issue. In Sect. 3 we introduce LDT by adopting the method of types. We also show how LDT allows poor parameter proposals to contribute to the representation of the approximate posterior distribution. Section 4 gives the LDW-ABC algorithm and compares it with R- ABC. In Sect. 5 we present the results of a toy example and an experiment conducted on a real world dataset. Section 6 contains some concluding remarks and ideas for future research. The Appendices contain the proofs, technical materials, and additional results from experiments.

2 Background on ABC

Let $x \in \mathcal{X}^n$ be the observed data, which will be assumed to be drawn from a probability distribution in the family $\mathcal{F} \stackrel{\triangle}{=} \{P(\cdot|\theta) : \theta \in \Theta\}$.

In principle, given a prior distribution $\pi(\theta)$ on Θ, the aim of Bayesian inference is to provide information about the uncertainty on θ by deriving the posterior distribution $\pi(\theta|x) \propto \pi(\theta)P(x|\theta)$ via Bayes' Theorem. When the likelihood function is intractable, ABC allows simulated inference providing a conversion of samples from the prior to samples from the posterior distribution, through comparisons between the observed data and the pseudo-datasets generated from a *simulator*. A simulator can be thought of as a probabilistic computer program taking as input a parameter value (or a vector thereof) $\theta \in \Theta$ and returning a sample from the distribution $P(\cdot|\theta)$. In general, no knowledge of the analytical form of the likelihood is necessary to write down such a program. More specifically, in the primal rejection sampling algorithm, whose origins can be traced back to Rubin (1984), Tavaré et al. (1997), Pritchard et al. (1999), the following actions are taken:

1. $S \geq 1$ parameter values from the prior distribution $\pi(\cdot)$ are generated;
2. for each $s \in \{1, \ldots, S\}$, given the parameter proposal $\theta^{(s)}$ as input, the simulator generates a realization of a random variable $Y \in \mathcal{X}^n$ distributed according to $P(\cdot|\theta^{(s)})$;
3. only parameter values leading to a pseudo-dataset equal to the observed data are accepted, thereby samples from the exact posterior are derived by conditioning on the event $\{Y = x\}$.

Introducing a twofold approximation scheme, as illustrated in Algorithm 1, might increase the efficiency of the algorithm outlined above. First, one introduces a summary statistic, $s(\cdot)$, which is a function from the sample space $\mathcal{X}^n \subseteq \mathbb{R}^n$ to a lower-dimensional space $\mathcal{S} \subset \mathbb{R}^k$, with $k \ll n$. Second, exact matching of the simulated and the observed data is relaxed to similarity, expressed in terms of a predefined distance function $d(\cdot, \cdot)$ and tolerance threshold $\epsilon > 0$.

Abbreviating $s(x)$ by s_x and $s(y)$ by s_y, the output of Algorithm 1 is a sample of pairs $(\theta^{(s)}, s_y^{(s)})$ from the following approximate joint posterior distribution

$$\tilde{\pi}(\theta, s_y|s_x) \propto \pi(\theta) \, P(s_y|\theta) \, \mathbb{1}\{d(s_y, s_x) \leq \epsilon\} \tag{1}$$

where $\mathbb{1}\{d(s_y, s_x) \leq \epsilon\}$, the indicator function assuming the value 1 if $d(s_y, s_x) \leq \epsilon$ and 0 otherwise, corresponds to the acceptance/rejection step. Marginalizing out s_y in

---

Algorithm 1 R- ABC

---

for $s = 1, \ldots, S$ **do**
    Draw $\theta^{(s)} \sim \pi$
    Generate $\mathbf{y} \sim P(\cdot|\theta^{(s)})$ from the simulator
    Accept the pair $(\theta^{(s)}, s_{\mathbf{y}}^{(s)})$ if $d(s_{\mathbf{y}}^{(s)}, s_{\mathbf{x}}) \leq \epsilon$
end for

---

(1), that is, ignoring the simulated summary statistics, the output of the algorithm becomes a sample from the marginal posterior distribution $\Pr(\theta|d(s_Y, s_{\mathbf{x}}) \leq \epsilon)$. Indeed, abbreviating s$(Y)$ by s$_Y$,

$$\tilde{\pi}(\theta|s_{\mathbf{x}}) \propto \int_S \pi(\theta)\, P(s_{\mathbf{y}}|\theta)\, \mathbb{1}\{d(s_{\mathbf{y}}, s_{\mathbf{x}}) \leq \epsilon\}\, ds_{\mathbf{y}}$$

$$= \pi(\theta) \int_S P(s_{\mathbf{y}}|\theta)\, \mathbb{1}\{d(s_{\mathbf{y}}, s_{\mathbf{x}}) \leq \epsilon\}\, ds_{\mathbf{y}}$$

$$= \pi(\theta) \cdot \Pr\big(d(s_Y, s_{\mathbf{x}}) \leq \epsilon|\theta\big)$$

$$\propto \Pr(\theta|d(s_Y, s_{\mathbf{x}}) \leq \epsilon).$$

Here $\Pr\big(d(s_Y, s_{\mathbf{x}}) \leq \epsilon|\theta\big)$ is called the ABC *approximate likelihood.*

Remark 1 (Marginal samplers) Some ABC sampling schemes (Sisson et al. 2007; Marjoram et al. 2003, among others) allow directly sampling from the approximate marginal posterior distribution $\tilde{\pi}(\theta|s_{\mathbf{x}})$. The key idea is that $\tilde{\pi}(\theta|s_{\mathbf{x}})$ can be estimated pointwise by

$$\pi(\theta^{(s)}) \cdot \frac{1}{M} \sum_{i=1}^{M} \mathbb{1}\{d(s_{\mathbf{y}}^{(i)}, s_{\mathbf{x}}) \leq \epsilon\} \qquad \forall s \in \{1, \ldots, S\} \tag{2}$$

by simulating M pseudo-datasets from $P(\cdot|\theta^{(s)})$ and computing $s_{\mathbf{y}}^{(i)}$ for $i \in 1, \ldots, M$ at each iteration s. As is apparent, the second term in (2) provides a Monte Carlo estimate of the ABC approximate likelihood.

Note that marginalizing the output of Algorithm 1 corresponds to the implementation of a marginal sampler with $M = 1$. In such a case, the indicator function represents a crude Monte Carlo estimate of the probability $\Pr\big(d(s_Y, s_{\mathbf{x}}) \leq \epsilon|\theta\big)$.

As pointed out by Sisson et al. (2018, Ch. 1), the use of the indicator function does not enable one to discriminate between whether the pseudo-dataset $\mathbf{y}$ coincides with the observed data and whether the pseudo-dataset just is close enough. This may lead to a waste of information. For this reason, the indicator function in (1) is often replaced by a kernel function:

$$K_\epsilon\big(d(s_{\mathbf{y}}, s_{\mathbf{x}})\big) = \begin{cases} \kappa\big(d(s_{\mathbf{y}}, s_{\mathbf{x}})\big) & \text{if } d(s_{\mathbf{y}}, s_{\mathbf{x}}) \leq \epsilon \\ 0 & \text{if } d(s_{\mathbf{y}}, s_{\mathbf{x}}) > \epsilon \end{cases} \tag{3}$$

Algorithm 2 IS- ABC

---

for $s = 1, \ldots, S$ **do**

  Draw $\theta^{(s)} \sim q$

  Generate $\boldsymbol{y} \sim P(\cdot | \theta^{(s)})$ from the simulator

  Set the IS weight for $\theta^{(s)}$ to $\omega_s = K_\epsilon\big(d(s_{\boldsymbol{y}}, s_{\boldsymbol{x}})\big) \cdot \dfrac{\pi(\theta^{(s)})}{q(\theta^{(s)})}$

end for

---

where $\kappa(\cdot)$ is a kernel function (e.g., triangular, Epanechnikov, Gaussian, etc.) defined on a compact support and decaying continuously from 1 to 0, see e.g. Beaumont et al. (2002). Now the ABC approximate likelihood becomes the convolution of the true model with the kernel K_ϵ (Prangle et al. 2018):

$$\tilde{\mathcal{L}}_{\epsilon,d}(\theta; s_{\boldsymbol{x}}) = \int_{\mathcal{S}} P(s_{\boldsymbol{y}}|\theta)\, K_\epsilon(\,d(s_{\boldsymbol{y}}, s_{\boldsymbol{x}})\,)\, ds_{\boldsymbol{y}}. \tag{4}$$

Note that this general setting encompasses also the case when R- ABC employs the uniform kernel as $\kappa(\cdot)$. The accuracy of the posterior distribution approximation depends both on how much information about the parameters is preserved by the summary statistics and on the magnitude of the threshold ϵ. In fact, as $\epsilon \to 0$, the approximate likelihood $\tilde{\mathcal{L}}(\theta; \boldsymbol{x})$ converges to the true likelihood (Prangle et al. 2018, Appendix A) and, whenever sufficient summary statistics for θ have been chosen, the approximate posterior distribution $\tilde{\pi}(\cdot | s_{\boldsymbol{x}})$ converges to the true posterior $\pi(\cdot | \boldsymbol{x})$ (Sisson et al. 2018, Ch. 1). On the other hand, as $\epsilon \to \infty$, the probability $\Pr\big(d(s_{\boldsymbol{Y}}, s_{\boldsymbol{x}}) \leq \epsilon | \theta\big)$ approaches 1 and samples are generated from the prior distribution. This establishes a trade-off between the statistical bias and the computational efficiency (Lintusaari et al. 2017): as the tolerance level ϵ decreases, the error of the approximation of the ABC posterior vs. the true posterior decreases at the cost of higher computational effort.

2.1 Importance sampling ABC and sample degeneracy

In the ABC literature, a great variety of methods to sample from $\tilde{\pi}(\theta, s_{\boldsymbol{y}}|s_{\boldsymbol{x}})$ have been proposed that go beyond the rejection scheme. Hereafter, we will adopt an importance sampling scheme (IS- ABC) which, as outlined by Karabatsos et al. (2018), encompasses R- ABC and most of the other ABC algorithms.

Like the standard importance sampling, see Robert and Casella (2013, Ch. 3), IS- ABC consists of sampling pairs $(\theta^{(s)}, s_{\boldsymbol{y}}^{(s)})$ from an *importance distribution* and weighting each pair, avoiding the computation of the acceptance probabilities. More formally, let $h : (\Theta \times \mathcal{S}) \to \mathbb{R}$ be a function of interest and let $E_p[h(\theta, s_{\boldsymbol{Y}})]$ denote its expected value w.r.t. a probability distribution p over $\Theta \times \mathcal{S}$. Suppose that we are interested in estimating $E_{\tilde{\pi}}[h(\theta, s_{\boldsymbol{Y}})]$, where $\tilde{\pi}$ is our target distribution, i.e., the joint approximate posterior. In particular, by choosing $h(\cdot)$ to be a kernel function, this formulation also enables a kernel density estimation for the joint approximate posterior, a case which will be considered in Sect. 5.

Now it is a standard fact that

$$E_{\tilde{\pi}}[h(\theta, s_Y)] = E_q[\bar{\omega}(\theta, s_Y)h(\theta, s_Y)] \tag{5}$$

where $q(\theta, s_y)$ is the importance distribution on $\Theta \times S$ and $\bar{\omega}(\theta, s_y) = \dfrac{\tilde{\pi}(\theta, s_y|s_x)}{q(\theta, s_y)}$ are the importance weights. In particular, in the ABC framework, the importance distribution can be

$$q(\theta, s_y) = q(\theta)P(s_y|\theta)$$

and, denoting by Z the normalizing constant for the joint posterior, the resulting importance weights $\bar{\omega}(\theta^{(s)}, s_y^{(s)})$, $\bar{\omega}_s$ for short, are

$$\begin{aligned} \bar{\omega}_s &= \frac{\pi(\theta^{(s)}) \ P(s_y^{(s)}|\theta^{(s)}) \ K_\epsilon\big(d(s_y^{(s)}, s_x)\big)}{Z \, q(\theta^{(s)}) \ P(s_y^{(s)}|\theta^{(s)})} \\ &= \frac{K_\epsilon\big(d(s_y^{(s)}, s_x)\big)}{Z} \cdot \frac{\pi(\theta^{(s)})}{q(\theta^{(s)})} \qquad \forall s \in \{1, \ldots, S\} \end{aligned}$$

By computing, at each iteration s, the following unnormalized weight

$$\omega_s = K_\epsilon\big(d(s_y^{(s)}, s_x)\big) \cdot \frac{\pi(\theta^{(s)})}{q(\theta^{(s)})}, \tag{6}$$

an approximation for the constant Z is obtained:

$$\begin{aligned} Z &= \int_\Theta \int_S \pi(\theta) \ P(s_y|\theta) \ K_\epsilon\big(d(s_y, s_x)\big) \, ds_y d\theta \\ &= \int_\Theta \int_S \omega(\theta, s_y) \, q(\theta, s_y) \, ds_y d\theta \approx \frac{1}{S} \sum_{s=1}^{S} \omega_s \end{aligned}$$

where the second equality is obtained by multiplying and dividing by $q(\theta, s_y)$. It follows that from the output of Algorithm 2 we can estimate (5) as

$$\frac{1}{S Z} \sum_{s=1}^{S} \omega_s \, h(\theta^{(s)}, s_y^{(s)}) \approx \sum_{s=1}^{S} \tilde{\omega}_s \, h(\theta^{(s)}, s_y^{(s)}) \tag{7}$$

where each $\tilde{\omega}_s = \omega_s / \sum_{r=1}^{S} \omega_r$ is a normalized weight.

Unlike the standard importance sampling scheme, IS- ABC implicitly involves a rejection step. In fact, usually the kernel density function, $K_\epsilon(\cdot)$, is such that a strictly positive weight is given to a pair $(\theta^{(s)}, s_y^{(s)})$ only when $d(s_y^{(s)}, s_x) \le \epsilon$. For example, looking at Algorithm 1, it is apparent that the primal rejection scheme is a special

case of Algorithm 2 where the marginal importance distribution, $q(\theta)$, is the prior distribution and the resulting importance weights are

$$\omega_s = K_\epsilon\big(d(s_y^{(s)}, s_x)\big) = \mathbb{1}\{d(s_y^{(s)}, s_x)) \leq \epsilon\} \tag{8}$$

meaning that each pair is implicitly rejected or accepted depending on the value of $\omega_s \in \{0, 1\}$.

In order to evaluate the efficiency of an importance sampling method, a widespread "rule of thumb" is to evaluate the *Effective Sample Size* (ESS), see Liu (2008, Ch 2). ESS represents the number of samples from the target distribution needed to get a Monte Carlo estimate with the same variance as the IS estimate in (7) given a budget of S iterations, and is defined by

$$\text{ESS} \triangleq \frac{S}{1 + var[\bar{\omega}]}. \tag{9}$$

One of the major drawbacks of the importance sampling scheme is that the resulting Monte Carlo estimate in (7) is highly variable due to the problem of sample degeneracy, already mentioned previously, which in this context means that only a few of the proposed pairs (θ, s_y) have relatively high weights resulting in a small ESS. Generally speaking, sample degeneracy is caused by an importance distribution far from its target. In this case, parameter values from regions with a low target posterior density are very likely to be drawn under the importance distribution, so that they are often proposed and associated with very small weights.

In the ABC framework as well, an importance density $q(\theta)$ far from the marginal target $\tilde{\pi}(\theta|s_x)$ can lead to sample degeneracy. In this setting, an additional issue is that the weights also depend on the distance $d(s_Y, s_x)$, hence on the random variable[1] s_Y (Sisson et al. 2018). This implies that when a parameter θ^* is proposed such that $\Pr(s_Y = s_x|\theta^*)$ is close to zero, usually a very large number of zero-weighted pairs (θ^*, s_y) will be generated before a distance smaller than ϵ will be observed, especially when ϵ is small.

In the next two sections we propose a method to define a kernel $K_\epsilon(\cdot)$ that improves the efficiency of IS- ABC in terms of ESS.

3 ABC and the theory of large deviations

Recall that in R- ABC, at each iteration s, the indicator function represents a crude estimate for the probability $\Pr(d(s_Y, s_x) \leq \epsilon| \theta^{(s)})$ (see Remark 1). A possible approach to mitigate sample degeneracy is to provide a finer estimate for the ABC likelihood by evaluating that probability. In order to deal with rare events, we resort to LDT, which studies the exponential decay of the probability of rare events. We speculate that taking into account the positive probability of a large deviation event allows one

---

[1] As pointed out by Prangle (2016), in the ABC framework the sampling scheme is more rightly referred to as *Random Importance Sampling*: an importance sampling schemes in which the likelihood is evaluated by a random estimate.

to avoid rejection at all. This might provide a higher ESS, thus making the algorithm more efficient.

From now on we will confine our attention to discrete random variables, and adopt an information theoretic point of view based on the method of types (Csiszár 1998; Cover and Thomas 2006). In particular, we will assume that $\mathcal{X} = \{r_1, \ldots, r_{|\mathcal{X}|}\}$ is a finite, nonempty set. Moreover, $\mathcal{F} \stackrel{\triangle}{=} \{P(\cdot|\theta) : \theta \in \Theta\}$ is a family of distributions on $\mathcal{X}$, where each $P(\cdot|\theta) = P_\theta$ has full support: $\mathrm{supp}(P(\cdot|\theta)) \stackrel{\triangle}{=} \{r : P(r|\theta) > 0\} = \mathcal{X}$ for each $\theta \in \Theta$.

We will let $X^n = \{X_i\}_{i=1}^n$, $Y^m = \{Y_i\}_{i=1}^m$ and so on denote sequences of i.i.d. random variables, distributed according to an (intractable) probability distribution $P_\theta \in \mathcal{F}$.

3.1 LDT via the method of types

Let x^n be a sequence of n symbols drawn from $\mathcal{X}$, say $x^n = (x_1, \ldots, x_n)$. The method of types moves the focus from the sequence x^n itself to its type, defined as follows.

Definition 1 (Type) Let $x^n = (x_1, \ldots, x_n) \in \mathcal{X}^n$. The *type* of x^n, written T_{x^n}, is the probability distribution on $\mathcal{X}$ defined by

$$T_{x^n}(r) \stackrel{\triangle}{=} \frac{|\{i : x_i = r\}|}{n} \quad \forall r \in \mathcal{X}. \tag{10}$$

We let $\mathcal{T}^n$ denote the set of n-types, that is, types with denominator n.

Note that the superscript n keeps track of the length of the sequence, which is also the denominator of the type. As is apparent, *type* is a function summarizing the information included in the observed sequence x^n by mapping the n-dimensional observed sequence onto a $|\mathcal{X}|$-dimensional summary statistic.

The following quantities play a crucial role in the method of types. Below, we stipulate that $0 \cdot \log \frac{0}{r} \stackrel{\triangle}{=} 0$ and that $r \cdot \log \frac{r}{0} \stackrel{\triangle}{=} +\infty$ if $r > 0$, where log denotes the logarithm to base 2. Given two probability distributions on $\mathcal{X}$, P and Q, we consider

– the *entropy* of P, defined as

$$H(P) \stackrel{\triangle}{=} -\sum_{r \in \mathcal{X}} P(r) \log P(r) ;$$

– the *Kullback–Leibler divergence* between P and Q, defined as

$$D(P||Q) \stackrel{\triangle}{=} \sum_{r \in \mathcal{X}} P(r) \, \log \frac{P(r)}{Q(r)} .$$

With an abuse of notation, whenever the first argument of $D(\cdot||Q)$ is a set of probability distributions, say E, $D(E||Q)$ stands for $\inf_{P \in E} D(P||Q)$. When $P^* = \mathrm{argmin}_{P \in E} D(P||Q)$ exists, it is called the *information projection of Q onto E*.

Let $X^n = \{X_i\}_{i=1}^n$ be a sequence of i.i.d. random variables, distributed according to $P_\theta \overset{\triangle}{=} P(\cdot|\theta)$, for some $\theta \in \Theta$. In what follows, we let $\Pr(\cdot|\theta)$ be the probability measure on sequences induced by P_θ. The joint probability of n i.i.d. extractions x^n according to P_θ, can be written as

$$\Pr(X^n = x^n|\theta) = 2^{n\left(-D(T_{x^n}||P_\theta) - H(T_{x^n})\right)}. \tag{11}$$

See Cover and Thomas (2006, Ch.11) for a proof. It follows from the Neyman–Fisher theorem (Cox and Hinkley 1979, Ch. 2.2) that types are always sufficient statistics for θ, whatever P_θ.

Remark 2 (Types and ABC) While the number of sequences of length n is exponential in n, it is easy to show that the cardinality of $\mathcal{T}^n$ is polynomial in n; in fact, $|\mathcal{T}^n| \leq (n+1)^{|\mathcal{X}|}$, see Cover and Thomas (2006, Ch.11). From an ABC perspective, it follows that using types as summary statistics could mitigate the computational problems related to the comparison between the observed dataset and the pseudo-dataset, especially for large n. Furthermore, summarizing data through their empirical distributions is a way of overcoming the difficulties in finding sufficient statistics when P_θ is unknown (and $\Pr(\cdot|\theta)$ as well). Indeed, even when confined to discrete random variables, $P(\cdot|\theta)$ is an unknown model, not necessarily a Multinomial model, see Sect. 5 for examples. With no knowledge of the analytical form of the likelihood, finding a sufficient summary statistic for θ, the vector of parameters given as an input to the simulator, is a central issue. In the literature there are several examples of models for conditionally independent discrete data in which the likelihood is analytically intractable and the required ABC method concerns empirical distributions. Examples are the ABC methods proposed by Joyce et al. (2012) and Buzbas and Rosenberg (2015) to make inference on the mutation and selection parameters governing the Fisher–Wright model (Fisher 1930). There, despite the conditional independence and the discreteness of the observations, the likelihood function is difficult to evaluate since the normalizing constant depends on the parameters: for small values of the selection parameter, numerical solutions have been found by Genz and Joyce (2003), in other cases, likelihood-free methods are required.

Noting from (11) that the probability of the observed sequence decreases exponentially at a rate given by the Kullback–Leibler divergence between T_{x^n} and P_θ, we can say (informally) that a sequence x^n is *typical* if $D(T_{x^n}||P_\theta) < \delta$ for some small $\delta > 0$.

The Law of Large Numbers (LLN) states that as the length of a typical sequence goes to infinity, its type converges in probability to P_θ, see Cover and Thomas (2006, Ch 11.2.1) for formulation of the LLN in terms of the method of types presented below.

Theorem 1 (Law of Large Numbers) *Let* $X^n = \{X_i\}_{i=1}^n$ *be a sequence of i.i.d. random variables with* $X_i \sim P_\theta$. *Then for each* $\delta > 0$

$$\Pr\left(D(T_{X^n}||P_\theta) \leq \delta|\theta\right) \geq 1 - 2^{-n(\delta - |\mathcal{X}|\frac{\log(n+1)}{n})}. \tag{12}$$

Moreover, under $\Pr(\cdot|\theta)$, *as* $n \to \infty$, $D(T_{X^n}||P_\theta) \to 0$ *with probability 1.*

On the other hand, observing a sequence whose type is far from P_θ, called a *non-typical* sequence, is a rare event, and its probability obeys a fundamental result in LDT, Sanov's theorem; see Cover and Thomas (2006, Th.11.4.1).

Theorem 2 (Sanov's Theorem) *Let $X^n = \{X_i\}_{i=1}^n$ be i.i.d. random variables on X such that each $X_i \sim P_\theta$. Let $\Delta^{|X|-1}$ be the simplex of probability distributions over X and let $E \subseteq \Delta^{|X|-1}$. Then*

$$\Pr\left(T_{X^n} \in E \mid \theta\right) \leq (n+1)^{|X|} 2^{-nD(P^*||P_\theta)}, \tag{13}$$

where $P^* = \underset{P \in E}{\mathrm{argmin}}\, D(P||P_\theta)$ is the information projection of P_θ onto E. Furthermore, if E is the closure of its interior,

$$\lim_{n \to \infty} \frac{1}{n} \log \Pr(T_{X^n} \in E \mid \theta) = -D(E||P_\theta) = -D(P^*||P_\theta).$$

Suppose that E is composed of types of non-typical sequences. Then Sanov's theorem characterizes the exponential decrease rate of the probability of E. Taking into account this probability may provide a finer ABC approximation of the likelihood, as discussed in the next section.

3.2 LDT in ABC

In this section we provide a formal explanation of what is meant by poor parameter proposals and how they can contribute to the representation of the approximate posterior distribution by means of LDT. We are interested in obtaining an approximation of the posterior distribution, $\tilde{\pi}(\theta|x^n)$, via R- ABC or an equivalent IS- ABC by assuming as given: (a) the marginal importance density $q(\theta)$ to be the prior distribution on Θ; (b) $\epsilon > 0$ as a threshold; (c) types as summary statistics; (d) the Kullback–Leibler divergence as distance function. For the sake of simplicity, from now on we will also assume T_{x^n} to be full support.

Given a budget of S iterations, both R- ABC and IS- ABC generate a sequence of pairs $(T_{y^m}^{(s)}, \theta^{(s)})$ with $s \in \{1, \ldots, S\}$. Each $T_{y^m}^{(s)}$ is an m-type resulting from a sequence of i.i.d. random variables, $Y^m = \{Y_j\}_{j=1}^m$, distributed according to $P\left(\cdot \mid \theta^{(s)}\right)$. We stress that the length of the simulated sequence, m, need not be equal to n, the length of the observed data sequence. Note also that, because of the independence assumption, choosing $m = M \cdot n$ with $M \in \mathbb{N}$ means that the algorithm simulates M pseudo-datasets at each iteration, like a marginal sampler (see Remark 1).

Looking at Algorithms 1 and 2, being the whole pair $(\theta^{(s)}, T_{y^m}^{(s)})$ accepted or rejected, one can define the *joint acceptance region* for these algorithms on the space $\Theta \times \mathcal{T}^m$. However, as the acceptance rule is based only on the simulated type, regardless of the proposed parameter value, the acceptance region can be projected onto the probability simplex $\Delta^{|X|-1} \supset \mathcal{T}^m$.

Definition 2 (Acceptance region) Let $\Delta^{|\mathcal{X}|-1}$ be the simplex of probability distributions over $\mathcal{X}$ and let $T_{\boldsymbol{x}^n}$ be the type of the observed sequence $\boldsymbol{x}^n$. The *acceptance region* $\mathcal{B}_\epsilon(T_{\boldsymbol{x}^n})$, referred to as $\mathcal{B}_\epsilon$ for short, is defined for any $\epsilon \geq 0$, as

$$\mathcal{B}_\epsilon \stackrel{\triangle}{=} \left\{ P \in \Delta^{|\mathcal{X}|-1} : D(P || T_{\boldsymbol{x}^n}) \leq \epsilon \right\}.$$

Now we can define a poor parameter proposal as a parameter $\theta^{(s)}$ such that $T_{\boldsymbol{x}^n}$ and the other types in the acceptance region are types of non-typical sequences w.r.t. $P(\cdot | \theta^{(s)})$.

Accordingly, sampling a poor parameter means that there is a large divergence between $T_{\boldsymbol{x}^n}$ and $P(\cdot|\theta^{(s)})$. On the other hand, with m large enough, $T_{\boldsymbol{y}^m}^{(s)}$ is very likely to be close to $P(\cdot|\theta^{(s)})$, due to the Law of Large Numbers. Heuristically, this implies that the probability of simulating a sequence $\boldsymbol{y}^m$ whose type is in the acceptance region is very small. Recalling that in R- ABC and in IS- ABC outlined at the beginning of this section a crude Monte Carlo estimate of the probability $\Pr(T_{\boldsymbol{Y}^m} \in \mathcal{B}_\epsilon | \theta^{(s)})$ is given by the indicator function $\mathbb{1}\{D(T_{\boldsymbol{y}^m}^{(s)} || T_{\boldsymbol{x}^n}) \leq \epsilon\}$, the vast majority of the poor parameter proposals are discarded altogether. We propose to mitigate this problem by assigning strictly positive weights to each proposal $\theta^{(s)}$, even if $T_{\boldsymbol{y}^m}^{(s)}$ is outside the acceptance region. To this end, we want to replace the indicator function with a finer estimate of the probability $\Pr(T_{\boldsymbol{Y}^m} \in \mathcal{B}_\epsilon | \theta^{(s)})$.

In principle, Sanov's theorem implies that, for m large enough, that probability can be approximated at each iteration by

$$\Pr\left(T_{\boldsymbol{Y}^m} \in \mathcal{B}_\epsilon | \theta^{(s)}\right) \approx 2^{-m D(\mathcal{B}_\epsilon || P_{\theta^{(s)}})}. \tag{14}$$

By replacing the indicator function in (1) with (14), the approximate posterior becomes

$$\tilde{\pi}(\theta, T_{\boldsymbol{y}^m} | T_{\boldsymbol{x}^n}) \propto \pi(\theta) P(T_{\boldsymbol{y}^m} | \theta) 2^{-m D(\mathcal{B}_\epsilon || P_\theta)}. \tag{15}$$

Unfortunately, the computation of the probability in (14) is still not feasible when the model $\mathcal{F} = \{P_\theta : \theta \in \Theta\}$ is unknown, as we do not know how to compute $D(\mathcal{B}_\epsilon || P_{\theta^{(s)}})$. The following theorem provides an asymptotic approximation to circumvent the problem. A proof is provided in "Appendix A".

Theorem 3 *Let* $\boldsymbol{Y}^m = \{Y_j\}_{j=1}^m$ *be a sequence of i.i.d. random variables taking values on the finite set* $\mathcal{X} = \{r_1, \ldots, r_{|\mathcal{X}|}\}$, *with each* $Y_j \sim P_\theta$. *Then under the measure* $\Pr(\cdot|\theta)$

$$\lim_{m \to \infty} D(\mathcal{B}_\epsilon || T_{\boldsymbol{Y}^m}) = D(\mathcal{B}_\epsilon || P_\theta) \quad a.s. \tag{16}$$

In essence, this result says that, as m increases and the type $T_{\boldsymbol{y}^m}$ converges to the distribution P_θ that has generated $\boldsymbol{y}^m$, the information projection of $T_{\boldsymbol{y}^m}$ onto $\mathcal{B}_\epsilon$ converges to that of P_θ onto $\mathcal{B}_\epsilon$ (see Fig. 1). From (14) and Theorem 3, for m large enough, $2^{-m D(\mathcal{B}_\epsilon || T_{\boldsymbol{y}^m})}$ provides a feasible asymptotic estimate for the acceptance probability,
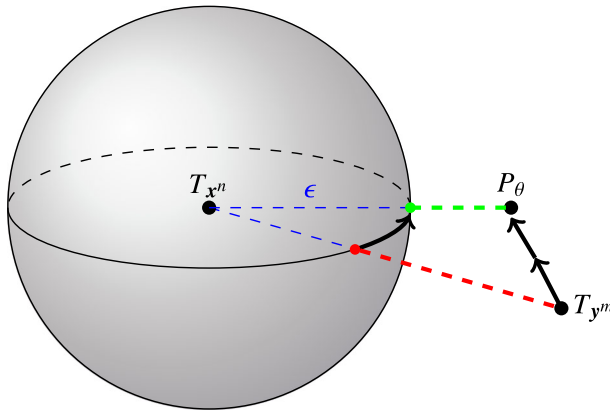
Fig. 1 Acceptance region, $\mathcal{B}_\epsilon$, types, T_{x^n} and T_{y^m}, and the probability distribution P_θ that generated y^m. Asymptotically (as $m \to \infty$) T_{y^m} converges to P_θ and the distance $D(\mathcal{B}_\epsilon || T_{y^m})$ (red) converges to $D(\mathcal{B}_\epsilon || P_\theta)$ (green) (color figure online)

$\Pr(T_{Y^m} \in \mathcal{B}_\epsilon \mid \theta)$. Replacing the indicator function in the ABC approximate posterior (1) with this estimate, we obtain the following new joint approximate posterior distribution:

$$\tilde{\pi}(\theta, T_{y^m}|T_{x^n}) \propto \pi(\theta) P(T_{y^m} \mid \theta) 2^{-mD(\mathcal{B}_\epsilon||T_{y^m})}. \tag{17}$$

4 Weighted approximate Bayesian computation

The discussion in the previous section indicates that IS- ABC can be improved by resorting to a better approximation for the likelihood. In particular, the (implicit) rejection step can be avoided by evaluating the positive probability of rare events via Sanov's theorem. Indeed, an easy way of sampling from (17) is a large deviations version of IS- ABC, which we will call the *weighted approximate Bayesian computation* (LDW- ABC).

Starting from the definition of an acceptance region satisfying the hypothesis of Sanov's theorem, as in Definition 2, a sample from the approximate posterior distribution $\tilde{\pi}(\theta, T_{y^m}|T_{x^n})$ can be obtained as described in Algorithm 3.

Algorithm 3 LDW- ABC

---
for $s = 1, \ldots, S$ **do**
   Draw $\theta^{(s)} \sim q$
   Generate $Y^m = \{Y_j\}_{j=1}^m$ with $Y_j \sim P(\cdot|\theta^{(s)})$ from the simulator
  if $D(T^{(s)}_{\mathbf{y}^m}||T_{\mathbf{x}^n}) \leq \epsilon$ **then**

     Set the IS weight for $(\theta^{(s)}, T^{(s)}_{\mathbf{y}^m})$ to $\omega_s = \dfrac{\pi(\theta^{(s)})}{q(\theta^{(s)})}$

  else

     Set the IS weights for $(\theta^{(s)}, T^{(s)}_{\mathbf{y}^m})$ to $\omega_s = 2^{-mD(\mathcal{B}_\epsilon||T^{(s)}_{\mathbf{y}^m})} \dfrac{\pi(\theta^{(s)})}{q(\theta^{(s)})}$

  end if
end for

Looking at Algorithm 3, it is apparent that LDW- ABC is a specialization of the more general IS- ABC. More specifically, the sufficient summary statistics involved are the *types*, the distance function is the Kullback–Leibler divergence, and the kernel density function is defined as follows:

$$K_{\epsilon,m}(T_{\mathbf{y}^m}) = \begin{cases} 1 & \text{if } D(T_{\mathbf{y}^m}||T_{\mathbf{x}^n}) \leq \epsilon \\ 2^{-mD(\mathcal{B}_\epsilon||T_{\mathbf{y}^m})} & \text{if } D(T_{\mathbf{y}^m}||T_{\mathbf{x}^n}) > \epsilon \end{cases}. \tag{18}$$

At each iteration a positive weight is assigned to the proposed $\theta^{(s)}$. More precisely, the weight equals 0 only when $D(\mathcal{B}_\epsilon||T_{\mathbf{y}^m}) = \infty$. Each ω_s is computed by approximating the divergence $D(\mathcal{B}_\epsilon||T_{\mathbf{y}^m})$ as described in "Appendix B".

As a special case of the general IS- ABC, the output of Algorithm 3 is a weighted sample from the following approximate joint posterior distribution:

$$\tilde{\pi}(\theta, T_{\mathbf{y}^m}|T_{\mathbf{x}^n}) \propto \pi(\theta)\, K_{\epsilon,m}(T_{\mathbf{y}^m})\, P_\theta(T_{\mathbf{y}^m}) \tag{19}$$

which, by marginalizing out simulated types, becomes

$$\tilde{\pi}(\theta|T_{\mathbf{x}^n}) \propto \pi(\theta) \sum_{T_{\mathbf{y}^m} \in \mathcal{T}^m} K_{\epsilon,m}(T_{\mathbf{y}^m})\, P_\theta(T_{\mathbf{y}^m}) \tag{20}$$

where $\mathcal{T}^m$ denotes the set of the m-types. Hence, the likelihood approximated by LDW- ABC is

$$\tilde{\mathcal{L}}_{\epsilon,m}(\theta; T_{\mathbf{x}^n}) \overset{\triangle}{=} \sum_{T_{\mathbf{y}^m} \in \mathcal{T}^m} K_{\epsilon,m}(T_{\mathbf{y}^m})\, P_\theta(T_{\mathbf{y}^m}). \tag{21}$$

Note that the quality of the approximation depends both on the threshold ϵ and on the size of pseudo-dataset, m. More precisely, the *adjustment* w.r.t. the likelihood approximate by R- ABC,[2] here denoted by $\tilde{\mathcal{L}}^R_{\epsilon,m}(\theta; T_{\mathbf{x}^n}) \overset{\triangle}{=} \sum_{T_{\mathbf{y}^m} \in \mathcal{B}_\epsilon} P_\theta(T_{\mathbf{y}^m})$, depends

---

[2] Here we refer to an R- ABC involving types as summary statistics, the Kullback–Leibler divergence as distance function, and the same tuning parameters, m and ϵ, as in the corresponding LDW- ABC.

on m and ϵ. In fact, from (18) and Definition 2, the approximate likelihood in (21) can be written as

$$\tilde{\mathcal{L}}_{\epsilon,m}(\theta; T_{\boldsymbol{x}^n}) = \sum_{T_{\boldsymbol{y}^m} \in \mathcal{B}_\epsilon} P_\theta(T_{\boldsymbol{y}^m}) + \sum_{T_{\boldsymbol{y}^m} \in \mathcal{B}_\epsilon^c} 2^{-mD(\mathcal{B}_\epsilon \| T_{\boldsymbol{y}^m})} P_\theta(T_{\boldsymbol{y}^m})$$

$$= \tilde{\mathcal{L}}_{\epsilon,m}^R(\theta; T_{\boldsymbol{x}^n}) + \alpha_{\epsilon,m}(\theta)$$

where the term $0 \le \alpha_{\epsilon,m}(\theta) \le 1$ is the adjustment. The following lemma gives an upper bound for that adjustment $\alpha_{\epsilon,m}(\theta)$, in two cases depending on P_θ. The proof is in "Appendix A".

Proposition 1 (The adjustment upper bound) *Let* $\alpha_{\epsilon,m}(\theta) = \tilde{\mathcal{L}}_{\epsilon,m}(\theta; T_{\boldsymbol{x}^n}) - \tilde{\mathcal{L}}_{\epsilon,m}^R(\theta; \mathcal{B}_\epsilon)$ *be the difference between the two likelihood functions approximated by* LDW- ABC *and* R- ABC. *Let* $\mathcal{B}_\epsilon$ *be the* ABC *acceptance region and* $\mathring{\mathcal{B}}_\epsilon$ *its interior. We have the following upper bounds, depending on* θ, *which hold for all* $m \ge 1$.

(a) $P_\theta \in \mathring{\mathcal{B}}_\epsilon$. *Then* $D(\mathcal{B}_\epsilon^c \| P_\theta) > 0$ *and* $\alpha_{\epsilon,m}(\theta) \le (m+1)^{|\mathcal{X}|} 2^{-mD(\mathcal{B}_\epsilon^c \| P_\theta)}$;

(b) $P_\theta \in \mathcal{B}_\epsilon^c$. *Let* $\gamma \triangleq D(\mathcal{B}_\epsilon \| P_\theta) > 0$. *Then there exists* $0 < \delta < \gamma$ *s.t.* $\alpha_{\epsilon,m}(\theta) \le (m+1)^{|\mathcal{X}|} 2^{-m\delta}$.

From Proposition 1 it follows that as m goes to infinity, $\alpha_{\epsilon,m}(\theta) \to 0$ for almost all $\theta \in \Theta$. Therefore, the approximate likelihood from LDW- ABC achieves the approximate likelihood from R- ABC and preserves its asymptotic properties. Moreover, we speculate that LDW- ABC improves the efficiency by mitigating the sample degeneracy. An evaluation of ESS might be a way of appreciating the improvement induced by avoiding the implicit rejection.

Since the term ESS in (9) involves the evaluation of the variance of the normalized weights, $var[\bar{\omega}]$, its exact computation is infeasible, as it depends on the unknown target normalizing constant. For this reason, we adopt the following estimate derived by Kong (1992) and Elvira et al. (2018):

$$\widehat{\text{ESS}} \triangleq \frac{\left(\sum_{s=1}^S \omega_s \right)^2}{\sum_{s=1}^S \omega_s^2} \tag{22}$$

(with the proviso that $\widehat{\text{ESS}} \triangleq 0$ if all ω_s's are zero). Let $\widehat{\text{ESS}}_{IS}$ and $\widehat{\text{ESS}}_{LD}$ be, respectively, the value of ESS achieved by S iterations of IS- ABC and LDW- ABC by setting the same tuning parameters, distance function and importance density $q(\theta)$. Explicitly, let us assume that the kernel function for IS- ABC is 1 within the acceptance region $\mathcal{B}_\epsilon$ and 0 outside. Heuristically, adding positive weights increases the numerator more than the denominator in (22), suggesting that a non null weight assigned by LDW- ABC to a parameter proposal rejected by IS- ABC is enough to have $\widehat{\text{ESS}}_{LD} > \widehat{\text{ESS}}_{IS}$. This is confirmed by the following simple result, whose proof is given in "Appendix A".

Proposition 2 (Empirical ESS) *It holds that $\widehat{ESS}_{LD} \geq \widehat{ESS}_{IS}$. Moreover this inequality is strict, provided that in at least one iteration of the algorithm there is generated a full support $T_{\mathbf{y}^m}$ falling outside $\mathcal{B}_\epsilon$.*

A tedious but straightforward analysis shows in fact that the event mentioned in the statement, upon which strict inequality holds, occurs with probability 1 as $S \longrightarrow +\infty$. The above result will be empirically validated in the experiments of Sect. 5, thus providing further evidence that LDW- ABC achieves an improvement in terms of efficiency.

Below, we sum up the technical development so far with a discussion of the role of the parameters m and ϵ.

Remark 3 (On the role of the tuning parameters) Concerning the role of m, the size of pseudo-dataset, and ϵ, the tolerance, we can sum up the content of Propositions 1 and 2 as follows:

1. large m and small ϵ point to low $\widehat{ESS}$ and low $\alpha_{\epsilon,m}$;
2. small m and large ϵ point to high $\widehat{ESS}$ and high $\alpha_{\epsilon,m}$.

If one regards $\widehat{ESS}$ as a measure of efficiency, and $\alpha_{\epsilon,m}$ as a measure of (lack of) accuracy w.r.t. the R-ABC likelihood, (but see also below), 1 and 2 above indicate how to trade off one for the other.

In particular, as Theorem 3 requires a relatively large m in order to get a good approximation for the posterior probability, 1 above says we can increase the tolerance ϵ to mitigate the resulting inefficiency. On the other hand, in cases where a small tolerance parameter ϵ is required, 2 above offers room to mitigate the resulting inefficiency by decreasing m.

Note, however, that when considering accuracy w.r.t. the target posterior density $\pi(\theta|T_{\mathbf{x}^n})$, the adjustment $\alpha_{\epsilon,m}$ cannot simply be regarded as a measure of imprecision: rather, it represents a compensation for those θ's that would be assigned a too low probability by pure R-ABC. In this case, a sounder measure of precision can be obtained by directly comparing a kernel-estimated density (obtained with LDW-ABC weights) and the target posterior density, e.g., in terms of the mean integrated squared error (MISE). This measure is, however, impossible to evaluate analytically, since its calculation presupposes the knowledge of the target posterior density. From a more empirical point of view, further discussion of the consequences of different choices of ϵ and m on the performances of the posterior estimators is presented in Sect. 5, illustrated by a number of examples.

5 Experiments

In order to evaluate the performance of the proposed method, we have put a proof-of-concept implementation of LDW- ABC at work on two examples. We compare the results obtained from LDW- ABC with those obtained from R- ABC. For both examples, there is a MCMC method for sampling from the exact posterior distribution, and the resulting posterior inference is taken as a reference for comparison.

5.1 Example 1: mixture of binomial distributions

Let $X^n = \{X_i\}_{i=1}^n$ be a sequence of i.i.d. discrete random variables distributed according to the following parametric finite mixture model:

$$\lambda Bin(\theta_1, N = 4) + (1 - \lambda)Bin(\theta_2, N = 4). \tag{23}$$

Here we assume a uniform prior distribution on the mixture weight λ and that (θ_1, θ_2) are uniformly distributed on the set $\{(\theta_1, \theta_2) : 0 \leq \theta_2 \leq \theta_1 \leq 1\}$ by imposing the following *identifiability constraint*:

$$\theta_1 \geq \theta_2.$$

An analytical computation of the posterior distribution requires the evaluation of the likelihood

$$P(x^n | \lambda, \theta_1, \theta_2)$$
$$= \prod_{i=1}^n \binom{N}{x_i} \left[\lambda \theta_1^{x_i} (1 - \theta_1)^{N - x_i} + (1 - \lambda)\theta_2^{x_i} (1 - \theta_2)^{N - x_i} \right]. \tag{24}$$

The direct computation of (24) is infeasible, as even with a few hundred observations, it involves the expansion of the likelihood into 2^n terms. In the literature, there are several methods to deal with this problem, which allow sampling from the parameters' posterior distributions, see Marin et al. (2005). A widespread method is a Gibbs Sampling handling the finite mixtures issue as a missing data problem, see Diebolt and Robert (1994). Samples from the joint posterior distribution are obtained by means of a hierarchical model involving a vector of latent random variables, $Z^n = \{Z_i\}_{i=1}^n$, where each $Z_i \sim Bernoulli(1 - \lambda)$ indicates to which component the i-th observation belongs:

$$\begin{cases} X_i \sim Bin(\theta_1, N) & \text{if } z_i = 0 \\ X_i \sim Bin(\theta_2, N) & \text{if } z_i = 1. \end{cases} \tag{25}$$

Here the generative model consists of simulating each of the n values from one of the two binomials according to the result of a Bernoulli(1-λ) experiment. The same generative model has been run by a plug-in of the true values of the parameters displayed in Table 1 (LHS) to obtain the observed data. We ran Algorithm 4 as detailed in Table 1 (RHS), and after burn-in and thinning we got 5000 values for each parameter regarded as drawn independently from the true posterior distributions. The posterior means and variances are displayed in Table 2.

In order to compare performance of LDW- ABC with that of R- ABC, the marginal importance distributions are set equal to the prior distributions.

We ran Algorithms 1 and 3 with $S = 100,000$ and with four different pairs (m, ϵ). Since our speculation is that introducing the evaluation of the probability of rare events provides a better approximation in the tails of the distributions, we are interested in

---

Algorithm 4 GIBBS SAMPLING

Require: x^n

  Initialize $p^{(0)} = p_1^{(0)}, \ldots, p_n^{(0)}$

  for $s = 1, \ldots, S$ **do**

    Draw $Z_i^{(s)} \sim Ber(p_i^{(s-1)}) \quad \forall i \in \{1, \ldots, n\}$

    Draw $\theta_1^{(s)} \sim TruncatedBeta\big(1 + \sum_{i=1}^n x_i \mathbb{1}\{z_i = 0\}, 1 + \sum_{i=1}^n (N - x_i)\mathbb{1}\{z_i = 0\}, \theta_2^{(s-1)}, 1\big)$

    Draw $\theta_2^{(s)} \sim TruncatedBeta\big(1 + \sum_{i=1}^n x_i \mathbb{1}\{z_i = 1\}, (1 + \sum_{i=1}^n N - x_i)\mathbb{1}\{z_i = 1\}, 0, \theta_1^{(s)}\big)$

    Draw $\lambda^{(s)} \sim Beta\big(1 + n - \sum_{i=1}^n z_i^{(s)}, 1 + \sum_{i=1}^n z_i^{(s)}\big)$

    Compute $p_i^{(s)} = \dfrac{(1 - \lambda^{(s)})(\theta_2^{(s)})^{x_i}(1 - \theta_2^{(s)})^{N - x_i}}{(1 - \lambda^{(s)})(\theta_2^{(s)})^{x_i}(1 - \theta_1^{(s)})^{N - x_i} + \lambda^{(s)}(\theta_1^{(s)})^{x_i}(1 - \theta_2^{(s)})^{N - x_i}}$

  end for

Table 1 Details for the simulation of the dataset and for the Gibbs implementation

θ_1^{true}	θ_2^{true}	λ^{true}	N	n	S	Burn-in	Thinning
0.9	0.2	0.8	4	100	100,000	50,000	10

Table 2 Posterior estimates derived via Gibbs sampling

	MCMC posterior estimates		
	θ_1	θ_2	λ
Mean	0.8998	0.1556	0.8281
Variance	0.0004	0.0036	0.0018

comparing the shape of the posterior distributions obtained by LDW- ABC and by R- ABC taking the Gibbs posterior distributions as reference. Accordingly, besides the posterior estimates of the means and variances, we also reconstruct the posterior densities by means of a Gaussian kernel density estimation.

The point estimates are compared via the MSE, and the kernel density estimates via the MISE. In particular, the corresponding estimates, $\widehat{\text{MSE}}$ and $\widehat{\text{MISE}}$, are computed by averaging over 100 runs the squared errors and the integrated squared errors w.r.t. the output of the Gibbs sampler. The results are summarized in Table 3.

First, we note that both the $\widehat{\text{MSE}}$ and the $\widehat{\text{MISE}}$ achieved by LDW- ABC are always lower for LDW- ABC than for R- ABC. Hence, in our example, taking into account the probability of large deviation events has improved both the point estimates and the approximation of the posterior distributions. Moreover, as already pointed out in Sect. 4, LDW- ABC mitigates the sample degeneracy by achieving an $\widehat{\text{ESS}}$ up to more than five times that achieved by R- ABC (see Table 4).

In order to evaluate the sample degeneracy, Table 4 (RHS) also displays the *normalized perplexity*, which equals $2^{H(\tilde{\omega})}/S$, where $H(\tilde{\omega})$ denotes the entropy of the normalized weights. Cappé et al. (2008) show that the normalized perplexity represents an estimate of $2^{-D\big(\tilde{\pi}(\theta, T_{y^m}|T_{x^n})||q(\theta, T_{y^m})\big)}$, meaning that when the perplexity is larger, the sample degeneracy is smaller.

Table 3 Squared errors and integrated squared errors w.r.t. the output of the Gibbs sampler averaged over 100 runs

	$m = 500$, $\epsilon = 0.005$			$m = 500$, $\epsilon = 0.01$		
	θ_1	θ_2	λ	θ_1	θ_2	λ
$\widehat{\text{MSE}}_{mean}$						
LD	0.0121×10^{-4}	0.1516×10^{-4}	0.1444×10^{-4}	0.015×10^{-4}	0.0581×10^{-4}	0.0251×10^{-4}
R	0.0679×10^{-4}	2.4437×10^{-4}	0.9420×10^{-4}	0.0288×10^{-4}	1.6023×10^{-4}	0.681×10^{-4}
$\widehat{\text{MSE}}_{var}$						
LD	0.0000×10^{-4}	0.0012×10^{-4}	0.0004×10^{-4}	0.0000×10^{-4}	0.0079×10^{-4}	0.0013×10^{-4}
R	0.0006×10^{-4}	0.055×10^{-4}	0.0135×10^{-4}	0.0003×10^{-4}	0.0277×10^{-4}	0.0065×10^{-4}
$\widehat{\text{MISE}}$						
LD	0.1445	0.0479	0.0799	0.4662	0.2019	0.2344
R	2.9162	0.8656	1.6831	0.8509	0.2744	0.4634

	$m = 5000$, $\epsilon = 0.005$			$m = 5000$, $\epsilon = 0.01$		
	θ_1	θ_2	λ	θ_1	θ_2	λ
$\widehat{\text{MSE}}_{mean}$						
LD	0.0189×10^{-4}	2.7694×10^{-4}	0.9609×10^{-4}	0.0184×10^{-4}	1.489×10^{-4}	0.6666×10^{-4}
R	0.0281×10^{-4}	3.8095×10^{-4}	1.1787×10^{-4}	0.024×10^{-4}	2.3092×10^{-4}	0.9049×10^{-4}
$\widehat{\text{MSE}}_{var}$						
LD	0.0006×10^{-4}	0.0639×10^{-4}	0.0148×10^{-4}	0.0003×10^{-4}	0.029×10^{-4}	0.0068×10^{-4}
R	0.0009×10^{-4}	0.0854×10^{-4}	0.0196×10^{-4}	0.0005×10^{-4}	0.0495×10^{-4}	0.0115×10^{-4}
$\widehat{\text{MISE}}$						
LD	3.2921	1.0344	1.6270	0.7410	0.2175	0.3856
R	7.3482	2.3676	3.4212	1.8753	0.5775	0.9733

Each column contains results for one of the model parameters both for LDW- ABC and R- ABC
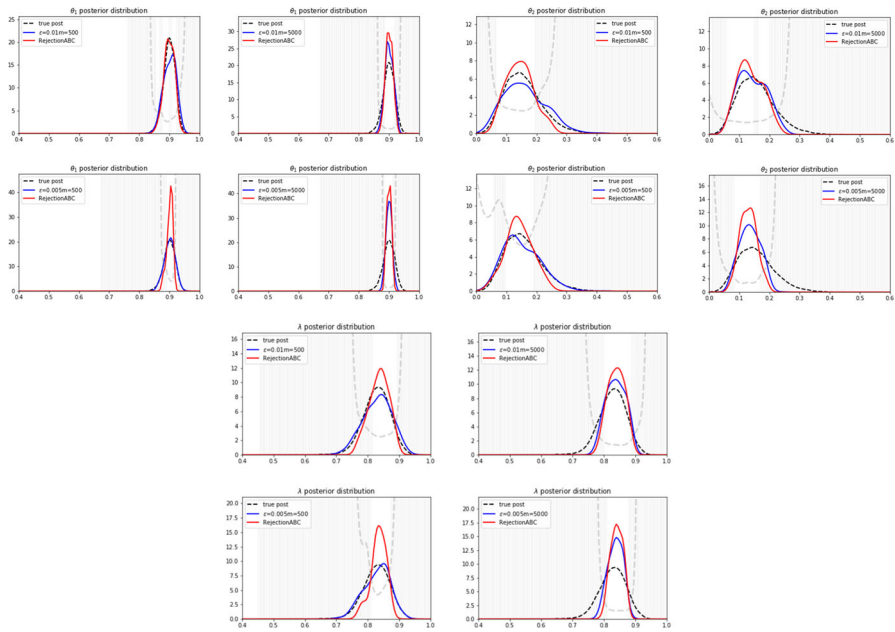
Fig. 2 Posterior distributions corresponding to four different pairs of tuning parameters (m, ϵ). Each panel refers to one of the three model parameters. Red lines represent the posterior density estimates provided via R- ABC. The blue lines represent the estimates provided via LDW- ABC. The dashed black lines are the output of the Gibbs sampler. The gray dashed lines are the ratios $\tilde{\mathcal{L}}_{\epsilon,m}(\theta; T_{\boldsymbol{x}^n})/\tilde{\mathcal{L}}^R_{\epsilon,m}(\theta; T_{\boldsymbol{x}^n})$ providing a representation of the adjustment $\alpha_{\epsilon,m}$ (color figure online)

Table 4 ESS and normalized perplexity averaged over 100 runs for each pair of tuning parameters

	Effective sample size			Normalized perplexity	
	$\epsilon = 0.005$	$\epsilon = 0.01$		$\epsilon = 0.005$	$\epsilon = 0.01$
$m = 500$					
LD	261	445	LD	0.0034	0.0055
R	25	81	R	0.0002	0.0008
$m = 5000$					
LD	71	168	LD	0.0008	0.0018
R	31	94	R	0.0003	0.0009

The following comments are consistent with Remark 3. Concerning the role of the tuning parameters, m and ϵ, we note that by fixing a large m (e.g., 5000), as ϵ increases both $\widehat{\text{ESS}}$ and the perplexity increase. Moreover, both $\widehat{\text{MSE}}$ and $\widehat{\text{MISE}}$ decrease. The same happens by reducing m with ϵ fixed to a small value (e.g., 0.005). This provides guidance on how to set the tuning parameters. In Fig. 2 three matrices of plots, one for each parameter, show the posterior densities: the size of the pseudo-dataset, m, equals 500 in the plots on the LHS of each panel, and 5, 000 on the RHS. The topmost plots show the approximate distributions with $\epsilon = 0.01$, the others the distributions
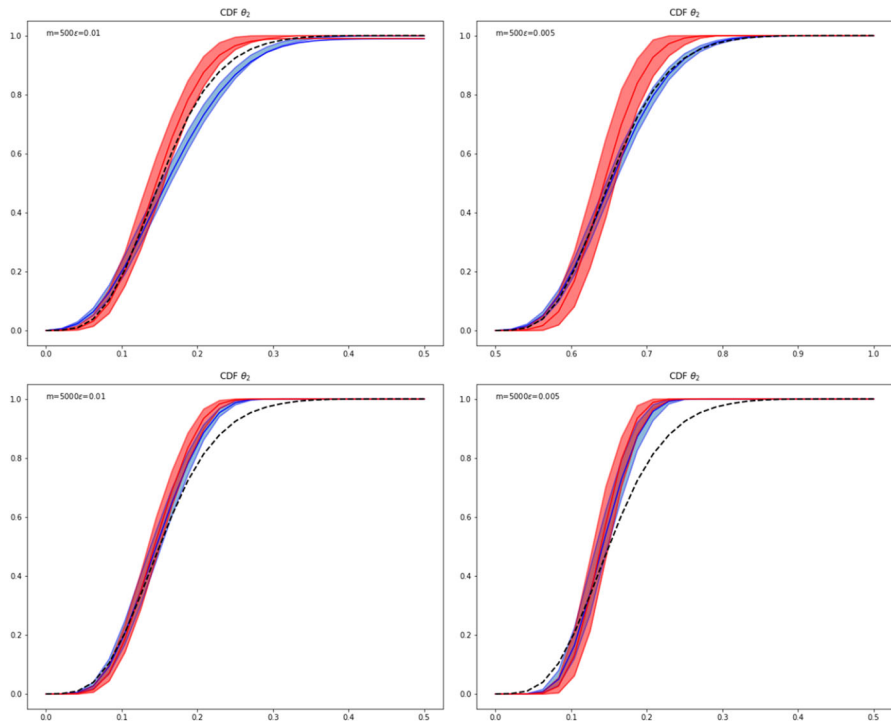
Fig. 3 Posterior cumulative density functions for θ_2. Each plot shows in blue the output of LDW- ABC, in red the output of R- ABC and in black the true cumulative density function for a pair (m, ϵ). For $\theta_2 > 0.5$ both the cumulative density functions equal 1. 90% intervals over 100 runs of each algorithm are also shown (color figure online)

corresponding to $\epsilon = 0.005$. According to Remark 3, we note that as m increases the blue lines (LDW- ABC) overlap the red ones (R- ABC). In principle, we would expect that both the algorithms achieve a better approximation of the posterior shapes with $\epsilon = 0.005$ than $\epsilon = 0.01$. However, in the case of R- ABC, we see a deviation from the true posterior distributions (dotted lines), when moving from the first to the second row of each matrix. The same deviation occurs for LDW- ABC, but only in the second column, when $m = 5,000$. This suggests that the quality of the R- ABC approximation is affected by a low value of the ESS which is in turn determined by a too exacting value of ϵ and m. In fact, when $m = 500$, LDW- ABC manages to mitigate the effect of a small ϵ, but it fails when a large value of m causes a too small ESS for the LDW- ABC as well. In the figures we also superimposed the ratio $\tilde{\mathcal{L}}_{\epsilon,m}(\theta; T_{x^n})/\tilde{\mathcal{L}}^R_{\epsilon,m}(\theta; T_{x^n}) = 1 + \alpha_{\epsilon,m}(\theta; T_{x^n})/\tilde{\mathcal{L}}^R_{\epsilon,m}(\theta; T_{x^n})$ evaluated pointwise and shown by the gray dashed lines. This quantity depends on the contribution of the adjustment w.r.t. the R- ABC likelihood and shows how the adjustment acts in modifying this latter when the R-ABC posterior density is underestimated (gray areas). Figure 3 shows the posterior cumulative density functions for θ_2. The posterior cumulative density functions for the other two parameters are given in "Appendix C", Fig. 4. We also show the 90% credible intervals for the estimated cumulative density functions. The red areas are

Table 5 Cleartext and anonymized tables

ID	Nat.	ZIP	Dis.	GID	Nat.	ZIP	Dis.
(a) Original table				(b) Anatomized table			
1	Malaysia	45501	Heart	1	Japan	45502	Heart
2	Japan	45502	Flu	1	Malaysia	45501	Flu
3	Japan	55503	Flu	2	Japan	55504	Flu
4	Japan	55504	Stomach	2	Japan	55503	Stomach
5	China	66601	HIV	3	Japan	66601	HIV
6	Japan	66601	Diabetes	3	China	66601	Diabetes
7	India	77701	Flu	4	Malaysia	77701	Flu
8	Malaysia	77701	Heart	4	India	77701	Heart

always larger than the blue areas, meaning that the estimates provided by R- ABC exhibit greater variability. This is more significant when $\epsilon = 0.005$, due to the small acceptance probability.

To wrap up, as suggested by the $\widehat{\text{MISE}}$'s, the posterior distributions approximated by LDW- ABC appear more faithful to the true shapes. Moreover, the ESS and the variability of the estimates are less sensitive to small values of ϵ.

5.2 Example 2: learning from anonymized data

The second example is aimed at comparing LDW- ABC and R- ABC at work on a real-world dataset. We consider a scenario in which the dataset contains microdata that have been anonymized in order to protect the privacy of the individuals involved. More specifically, our data are a subset of 5692 rows from the Adult dataset extracted by Barry Becker from the 1994 US Census database and available from the UCI machine learning repository (Kohavi and Becker 1996). The anonymizaton method we have adopted, Anatomy (Xiao and Tao 2006), is a group based anonymization scheme. Given a dataset consisting of a collection of rows, each one corresponding to an individual and containing his/her sensitive (e.g., disease, income) and nonsensitive attributes (e.g., gender, nationality, ZIP code), a group-based anonymization algorithm produces an obfuscated version of itself by partitioning the rows into groups. The idea is to process the set of rows in each group so that even knowing the nonsensitive attribute of an individual, one cannot identify his/her sensitive values. To reach this goal, Anatomy vertically and randomly permutes the nonsensitive features within each group, thus breaking the link between the sensitive and nonsensitive attributes.

An example of group based anonymization is in Table 5, adapted from Wong et al. (2011). The LHS table is the original table collecting medical data from eight individuals; here, *Disease* is considered as the only sensitive attribute. The RHS table is an example of an application of the Anatomy scheme: within each group, the nonsensitive part of the rows is vertically and randomly permuted, thus breaking the link between the sensitive and nonsensitive values. Generally speaking, the obfuscated table can

be seen as the output of a generative mechanism: given the population parameters as input, the mechanism first generates a cleartext table by drawing a number of rows i.i.d., then applies the anonymization algorithm to this table and outputs the result. One is interested in the posterior distribution of the population parameters, given an observation of the obfuscated table. Clearly, the likelihood function involved in this mechanism is highly nontrivial, and also depends on the details of the anonymization algorithm. Below we make this precise by adopting the model proposed by Boreale et al. (2020).

Let a row of the original (cleartext) dataset be a pair $(s, r) \in \mathcal{S} \times \mathcal{R}$, for finite, nonempty sets $\mathcal{S}$ and $\mathcal{R}$. Here, s and r represent the sensitive and nonsensitive attributes (or vectors of attributes). Given a multiset of n rows, $d = \{(s_1, r_1), \ldots, (s_n, r_n)\}$, Anatomy will first arrange d into a sequence of *groups*, $\boldsymbol{x} = g_1, \ldots, g_k$, the *cleartext table*. Each group in turn is a sequence of n_i rows, $g_i = (s_{i,1}, r_{i,1}), \ldots, (s_{i,n_i}, r_{i,n_i})$. The *obfuscated table* is then obtained as the sequence $\boldsymbol{x}^* = g_1^*, \ldots, g_k^*$, where the obfuscation of each group g_i is a pair $g_i^* = (\sigma_i, \rho_i)$. Here, each $\sigma_i = s_{i,1}, \ldots, s_{i,n_i}$ is the sequence of *sensitive* values occurring in g_i; each ρ_i, called the *generalized nonsensitive value*, is the multiset of g_i's nonsensitive values: $\rho_i = \{r_{i,1}, \ldots, r_{i,n_i}\}$—i.e., ρ_i includes all those and only those values, with multiplicities, found in g_i.

The model consists of the following random variables:

- $\boldsymbol{\theta} = (\theta_S, \boldsymbol{\theta}_{R|S})$, where $\boldsymbol{\theta}_{R|S} = \{\theta_{R|s} : s \in \mathcal{S}\}$. Here, $\boldsymbol{\theta}$ takes values on the set of full support probability distributions $\mathcal{D}$ over $\mathcal{S} \times \mathcal{R}$ and represents the joint probability distribution of the sensitive and nonsensitive attributes in the population.
- $X = G_1, \ldots, G_k$, which takes values in the set of cleartext tables $\mathcal{X}$. Each group G_i is in turn a sequence of $n_i \geq 1$ consecutive rows in X, $G_i = (S_{i,1}, R_{i,1}), \ldots, (S_{i,n_i}, R_{i,n_i})$ with $S_{i,j} \sim \theta_S$ and $R_{i,j} \sim \theta_{R|s_{i,j}}$. The number of groups k is not fixed, but depends on the anonymization scheme and on the specific tuples in d.
- $X^* = G_1^*, \ldots, G_k^*$, which takes values in the set of obfuscated tables $\mathcal{X}^*$.

We assume that X^* solely depends on the table X and the underlying obfuscation algorithm, thus the above three random variables form a Markov chain:

$$\boldsymbol{\theta} \longrightarrow X \longrightarrow X^*. \tag{26}$$

Here, our aim is to derive the posterior distribution for the population parameters $\boldsymbol{\theta}$ by observing an instance of the anonymized table, $\boldsymbol{x}^*$. There is no tractable analytical expression for the likelihood $\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{x}^*)$. In Kifer (2009) and Boreale et al. (2020) this problem is circumvented by defining an MCMC scheme for sampling from the joint posterior $\pi(\boldsymbol{\theta}, \boldsymbol{x}|\boldsymbol{x}^*)$ and then discarding the cleartext tables (further details on the MCMC scheme are in Boreale et al. (2020, Section 5). Here we pursue an alternative solution based on ABC. Specifically, we simulate the anonymized tables $\boldsymbol{y}^*$ according to the following generative model:

1. Generate a table of n i.i.d. rows $(s, r) \in \mathcal{S} \times \mathcal{R}$ distributed according to the $\boldsymbol{\theta}$ given as input;
2. Partition the table into k groups of dimensions $\{n_i\}_{i=1}^k$;

Table 6 The leftmost table shows the squared errors integrated over the 3-simplex and averaged over 100 runs of ABC

	MISE				$\widehat{\text{ESS}}$	
	Government	Self-employed	Private	Without pay		
R	0.6372	0.9185	16.3660	0.1477	R	16,704
LD	0.4147	0.5814	7.3125	0.106	LD	35,231

Each column corresponds to an element of $\{\boldsymbol{\theta}_{R|s} : s \in$ {Government, Self-employed, Private, Without pay}}. The rightmost table shows the ESS achieved by R- ABC and LDW- ABC averaged over 100 runs

3. Randomly permute the values of the nonsensitive attributes, $r_1, \ldots, r_{n_i}$, within each group g_i.

Here, n is the number of rows and k is the number of groups in the observed anonymized table $\boldsymbol{x}^*$, while n_i is the number of rows in g_i^*. In our example, the observed data $\boldsymbol{x}^*$ consists of an obfuscated table composed of $k = 1423$ groups, with $n_i = 4$ for each group g_i. We take *race* (four possible values) as a nonsensitive attribute and *workclass* (four possible values) as the sensitive attribute. As in Boreale et al. (2020), we assume that θ_S and the $\theta_{R|s}$'s are independently distributed according to non-informative Dirichlet prior distributions. The output of the ABC algorithms will be a sample from approximate joint posterior distribution $\tilde{\pi}(\boldsymbol{\theta}_{R|S}, T_{y^*}|T_{x^*})$ since the sensitive part is not changed by the anonymization algorithm and the posterior distribution for θ_S exists in closed form.[3] Note that in order to satisfy the assumptions of Sanov's theorem, the m simulated rows must be independent and identically distributed. Recalling that the m pairs (s, r)'s are generated independently and that the permutation is completely at random, we conjecture that the i.i.d. assumption is satisfied. We have positively verified this assumption empirically via the permutation test based on the periodicity test statistic described in the National Institute of Standards and Technology Special Publication 800-90B (Turan et al. 2018).

As in the previous experiment, we consider the output of 100,000 MCMC runs as a reference, in order to compute the $\widehat{\text{MSE}}$'s and $\widehat{\text{MISE}}$'s, and thus comparing the accuracy of LDW- ABC and R- ABC. The posterior means derived via MCMC are displayed in "Appendix C" (Table 8).

By setting $m = 100$ and $\epsilon = 1$, as far as point estimations are concerned, both LDW- ABC and R- ABC perform quite similarly. The results are displayed in "Appendix C" (Table 9). Nevertheless, the $\widehat{\text{MSE}}$'s achieved by LDW- ABC are almost always smaller than the ones achieved by R- ABC. Concerning the approximations of the multivariate posterior distributions, looking at $\widehat{\text{MISE}}$ we can conclude that LDW- ABC outperforms R- ABC (see Table 6). Moreover, by focusing on the improvement in efficiency, we note that the value of $\widehat{\text{ESS}}$ for LDW- ABC is more than twice that for R- ABC.

---

[3] The posterior distribution of θ_s is simply a Dirichlet distribution where the parameters are updated by the frequency counts of each $s \in \mathcal{S}$.

6 Conclusions and future research

We have put forward an approach to address sample degeneracy in methods of approximate Bayesian computation (ABC). Our proposal consists in the definition of a convenient kernel function which, via the theory of large deviations, takes into account the probability of rare events—a poor parameter proposal generating pseudo-data close to those observed. By adopting the information theoretic method of types, which involves summarizing data via their empirical distributions, we also by-pass the issue of selecting summary statistics. The proposed kernel function, being defined on a non-compact support, avoids any implicit or explicit rejection step, thus effectively increasing the effective sample size, as empirically verified in Sect. 5. Moreover, the resulting approximate ABC likelihood leads to a better approximation of the tails of the posterior distributions, that is, poor parameter proposals are assigned small but nonzero probability. We also provide formal guarantees of the convergence of our ABC approximate likelihood to the true likelihood.

The proposed method deals with the inefficiency in ABC algorithms by focusing on the kernel function. Although a variety of ABC sampling schemes addressing the same problem have been proposed, most of them (e.g., MCMC- ABC, SMC- ABC, PMC- ABC, SIS- ABC, etc.) handle the problem of finding a good marginal importance distribution, $q(\theta)$, completely ignoring the choice of the kernel $K_\epsilon(\cdot)$. We speculate that these two approaches can be combined by adopting those sampling schemes rather than the involved IS- ABC. Finally, further research is called for in order to apply the proposed method to sequences of dependent random variables, such as Markov chains, and to continuous data. This will in turn require considering more sophisticated versions of the theory of large deviations, where the i.i.d. assumption is relaxed. Such extensions are required to make the algorithm applicable to more complex situations in which no other ways of sampling from the posterior distribution are available. However we want to emphasize that at the current stage, the method is already applicable in contexts in which the other available sampling methods (e.g., MCMC) can be computationally demanding. For example, an efficient ABC algorithm may be usefully applied to models involving high-dimensional latent variables even when an MCMC algorithm exists. In fact, ABC algorithms can be run in parallel, leading to a computational gain when many of the existing MCMC algorithms suffer from a slow mixing of the chain.

A Proofs

In what follows, we will make use of a few basic notions and facts about the method of types and information projections, for which we refer the reader to (Csiszár et al. 2004, Chap. 1). The simplex of the distributions over $\mathcal{X}$, given a subset $\Delta^{|\mathcal{X}|-1} \subseteq \mathbb{R}^{|\mathcal{X}|}$, inherits the standard topology from $\mathbb{R}^{|\mathcal{X}|}$. W.r.t. this topology, the function $D(P||Q)$ is lower semi-continuous in the pair of arguments (P, Q), and continuous at (P, Q) whenever Q has full support, that is, whenever $\mathrm{supp}(Q) \stackrel{\triangle}{=} \{r \in \mathcal{X} : Q(r) > 0\} = \mathcal{X}$. Convergence to Q in KL divergence, $D(Q_n||Q) \to 0$, implies convergence in the standard topology, $Q_n \to Q$. As a function of P, $D(P||Q)$ is strictly convex, and continuous whenever Q is full support. Hence for any convex and closed set $E \subseteq \Delta^{|\mathcal{X}|-1}$ the information projection of Q onto E, $P^* = \mathrm{argmin}_{P \in E} D(P||Q)$, exists and is unique. The following is a fundamental result about information projections. The support of E is defined as $\mathrm{supp}(E) \stackrel{\triangle}{=} \bigcup_{P \in E} \mathrm{supp}(P)$.

Theorem A.1 (Pythagorean inequality, Csiszár et al. (2004) Th.3.1) *Let E be a closed and convex set and Q be full support. Let $P^* = \mathrm{argmin}_{P \in E} D(P||Q)$. Then $\mathrm{supp}(P^*) = \mathrm{supp}(E)$. Moreover, for each $P \in E$, $D(P||Q) \geq D(P||P^*) + D(P^*||Q)$.*

Proof of Theorem 3 Fix an infinite sequence $\tau \in \mathcal{X}^\infty$, $\tau = (y_1, y_2, y_3, \ldots)$. For each $m \geq 1$, let $T_{\mathbf{y}^m}(\tau)$, $T_{\mathbf{y}^m}$ for short, denote the type of the first m symbols of τ, the sequence $(y_1, \ldots, y_m)$. Assume τ is such that $T_{\mathbf{y}^m} \to P_\theta$ as $m \to +\infty$. Note that, since P_θ is full support, this implies that for all sufficiently large m, $T_{\mathbf{y}^m}$ is full support as well. Define P^* and, for any such sufficiently large m, P_m^* as follows:

$$P^* \stackrel{\triangle}{=} \mathrm{argmin}_{P \in \mathcal{B}_\epsilon} D(P||P_\theta) \quad \text{and} \quad P_m^* \stackrel{\triangle}{=} \mathrm{argmin}_{P \in \mathcal{B}_\epsilon} D(P||T_{\mathbf{y}^m}).$$

Note that as $T_{\mathbf{y}^m}$ is full support and $\mathcal{B}_\epsilon = \{P : D(P||T_{\mathbf{x}^n}) \leq \epsilon\}$ is convex and closed, the projection P_m^* exists and is unique. Moreover, as $T_{\mathbf{x}^n}$ is by assumption full support, it is easily seen that $\mathrm{supp}(\mathcal{B}_\epsilon) = \mathcal{X}$: hence, by the first part of Theorem A.1, the projection P_m^* is full support as well.

As $\mathcal{B}_\epsilon$ is closed and $D(\cdot||P_\theta)$ is continuous, $P^* \in \mathcal{B}_\epsilon$. We can now apply the Pythagorean Inequality, considering P_m^* as a projection and $P^* \in E = \mathcal{B}_\epsilon$, and obtain

$$D(P^*||T_{\mathbf{y}^m}) \geq D(P^*||P_m^*) + D(P_m^*||T_{\mathbf{y}^m}). \tag{27}$$

As P_θ is assumed to be full support, $D(\cdot||\cdot)$ as a function of its second argument is continuous at P_θ, hence

$$\lim_{m \to \infty} D(P^*||T_{\mathbf{y}^m}) = D(P^*||P_\theta). \tag{28}$$

Assuming $\{P_m^*\}$ converges, let $P^{**} \stackrel{\triangle}{=} \lim_{m \to \infty} P_m^*$, where clearly $P^{**} \in \mathcal{B}_\epsilon$; if $\{P_m^*\}$ does not converge, we can equivalently take any convergent subsequence of it. Taking

lim inf on both sides of (27), and exploiting (28) on the left-hand side, and lower semi-continuity on the right-hand side, we can write

$$
\begin{aligned}
D(P^*||P_\theta) &= \lim_{m\to\infty} D(P^*||T_{\mathbf{y}^m}) \\
&\geq \liminf_{m\to\infty} \left(D(P^*||P_m^*) + D(P_m^*||T_{\mathbf{y}^m}) \right) \\
&\geq \liminf_{m\to\infty} D(P^*||P_m^*) + \liminf_{m\to\infty} D(P_m^*||T_{\mathbf{y}^m}) \\
&\geq D(P^*||P^{**}) + D(P^{**}||P_\theta).
\end{aligned}
$$

Summing up

$$
D(P^*||P_\theta) \geq D(P^*||P^{**}) + D(P^{**}||P_\theta). \tag{29}
$$

Recalling that P^* is the information projection of P_θ onto $\mathcal{B}_\epsilon$, that $P^{**} \in \mathcal{B}_\epsilon$ and that $D(\cdot||\cdot)$ is nonnegative, the only possibility for (29) to hold is that $D(P^*||P^{**}) = 0$, which implies $P^* = P^{**}$. In other words

$$
\lim_{m\to\infty} P_m^* = P^*. \tag{30}
$$

This way, we have shown that $(P_m^*, T_{\mathbf{y}^m}) \to (P^*, P_\theta)$. Under $D(\cdot||\cdot)$ this limit becomes, by continuity at (P^*, P_θ):

$$
\lim_{m\to\infty} D(P_m^*||T_{\mathbf{y}^m}) = D(P^*||P_\theta). \tag{31}
$$

We have shown that (31) holds for any sequence $\tau \in \mathcal{X}^\infty$ such that $T_{\mathbf{y}^m} = T_{\mathbf{y}^m}(\tau) \to P_\theta$. Now let $\Pr(\cdot|\theta)$ be the probability measure on $\mathcal{X}^\infty$ induced by P_θ. The LLN (Theorem 1) says that, under $\Pr(\cdot|\theta)$, the set of such τ's has probability 1.

Hence (31) under $\Pr(\cdot|\theta)$ holds with probability 1, that is, almost surely. □

Recall that, for each Q and $\delta \geq 0$, $\mathcal{B}_\delta(Q) \subseteq \Delta^{|\mathcal{X}|-1}$ denotes the ball of radius δ centered at Q:

$$
\mathcal{B}_\delta(Q) \triangleq \{P : D(P||Q) \leq \delta\}.
$$

Lemma A.1 *Let $E \subseteq \Delta^{|\mathcal{X}|-1}$ be a convex and closed set. Let $Q \in \Delta^{|\mathcal{X}|-1}$ be such that $\gamma \triangleq D(E||Q) > 0$. Then for each $0 < \gamma' < \gamma$ there is $\delta > 0$ such that for each $Q' \in \mathcal{B}_\delta(Q)$ one has $D(E||Q') \geq \gamma'$.*

Proof The fact that E is closed and convex ensures that the projection $D(E||Q)$ exists and is finite. Consider the strictly descending chain of balls of radius $\delta_n = 1/n$ centered at Q: $\mathcal{B}_{\delta_1}(Q) \supseteq \mathcal{B}_{\delta_2}(Q) \supseteq \cdots \supseteq \mathcal{B}_{\delta_n}(Q) \supseteq \cdots$.

By contradiction, assume that there exists $0 < \gamma' < \gamma$ such that for each $\delta > 0$, there is $Q' \in \mathcal{B}_\delta(Q)$ such that $D(E||Q_n') < \gamma'$. In particular, we then have that

$$
\text{for each } n \geq 1 \text{ there is } Q_n' \in \mathcal{B}_{\delta_n}(Q) \text{ s.t. } D(E||Q_n') < \gamma'. \tag{32}
$$

We can therefore assume without loss of generality that

$$\lim_{n \to \infty} D(E || Q'_n) < \gamma'. \tag{33}$$

(if not, we can anyway extract from $\{D(E || Q'_n)\}$ a subsequence with the desired property). On the other hand, being $\lim_{n \to \infty} D(Q'_n || Q) = 0$, we have $\lim_{n \to \infty} Q'_n = Q$. Being $D(\cdot||\cdot)$ lower semi-continuous, we obtain

$$\liminf_{n \to \infty} D(E || Q'_n) \geq D(E || Q) = \gamma > \gamma'. \tag{34}$$

But this contradicts (32). □

Proof of Proposition 1 Let us consider the two cases separately, $P_\theta \in \mathring{\mathcal{B}}_\epsilon = \{P \in \Delta^{|\mathcal{X}|-1} : D(P||T_{x^n}) < \epsilon\}$ and $P_\theta \in \mathcal{B}_\epsilon^c = \{P \in \Delta^{|\mathcal{X}|-1} : D(P||T_{x^n}) > \epsilon\}$.

– $P_\theta \in \mathring{\mathcal{B}}_\epsilon$.

$$\alpha_{\epsilon,m} \leq \sum_{T_{y^m} \in \mathcal{B}_\epsilon^c} P_\theta(T_{y^m}) \leq (m+1)^{|\mathcal{X}|} 2^{-m D(\mathcal{B}_\epsilon^c || P_\theta)}$$

where the last inequality follows from a direct application of Sanov's Theorem.

– $P_\theta \in \mathcal{B}_\epsilon^c$. Choose any $0 < \gamma' < \gamma \overset{\triangle}{=} D(\mathcal{B}_\epsilon^c || T_{x^n})$ (note that $\gamma > 0$) and apply Lemma A.1 with $E = \mathcal{B}_\epsilon$ and $Q = P_\theta$ to obtain $\delta > 0$ such that $D(\mathcal{B}_\epsilon || Q') \geq \gamma'$ for each $Q' \in \mathcal{B}_\delta(P_\theta)$. We can assume without loss of generality that $\delta \leq \gamma'$. It follows that

$$\alpha_{\epsilon,m} = \sum_{T_{y^m} \in \mathcal{B}_\epsilon^c} 2^{-m D(\mathcal{B}_\epsilon || T_{y^m})} P_\theta(T_{y^m}) \tag{35}$$

$$= \sum_{T_{y^m} \in \mathcal{B}_\epsilon^c \cap \mathcal{B}_\delta^c} 2^{-m D(\mathcal{B}_\epsilon || T_{y^m})} P_\theta(T_{y^m}) + \sum_{T_{y^m} \in \mathcal{B}_\delta} 2^{-m D(\mathcal{B}_\epsilon || T_{y^m})} P_\theta(T_{y^m}) \tag{36}$$

$$\leq \sum_{T_{y^m} \in \mathcal{B}_\epsilon^c \cap \mathcal{B}_\delta^c} P_\theta(T_{y^m}) + \sum_{T_{y^m} \in \mathcal{B}_\delta} 2^{-m D(\mathcal{B}_\epsilon || T_{y^m})} \tag{37}$$

$$\leq \sum_{T_{y^m} \in \mathcal{B}_\epsilon^c \cap \mathcal{B}_\delta^c} 2^{-m D(T_{y^m} || P_\theta)} + \sum_{T_{y^m} \in \mathcal{B}_\delta} 2^{-m D(\mathcal{B}_\epsilon || T_{y^m})} \tag{38}$$

$$\leq \sum_{T_{y^m} \in \mathcal{B}_\epsilon^c \cap \mathcal{B}_\delta^c} 2^{-m\delta} + \sum_{T_{y^m} \in \mathcal{B}_\delta} 2^{-m\gamma'} \tag{39}$$

$$\leq \sum_{T_{y^m} \in \mathcal{B}_\epsilon^c} 2^{-m\delta} \tag{40}$$

$$\leq (m+1)^{|\mathcal{X}|} 2^{-m\delta} \tag{41}$$

where (38) follows from (11) and the last step follows from an upper bound for the size of $\mathcal{T}^m$ (see (Cover and Thomas 2006, Ch. 11, Th. 11.1.1)).

Proof of Proposition 2 Let us consider $\widehat{\text{ESS}} : \mathbb{R}^S \to \mathbb{R}$ as a function of S variables, $\widehat{\text{ESS}}(x_1, \ldots, x_S)$, defined for nonnegative reals x_i's, not all zero, representing the weights. The partial derivative of $\widehat{\text{ESS}}$ w.r.t. x_i has the form

$$\frac{\partial}{\partial x_i}\widehat{\text{ESS}}(x_1, \ldots, x_S) = C \cdot \sum_{j \neq i}(x_j^2 - x_i x_j)$$

for a function C that is > 0 in the domain of definition of $\widehat{\text{ESS}}$. Therefore, $\frac{\partial}{\partial x_i}\widehat{\text{ESS}}$ is nonnegative when evaluated at any point $(x_1, \ldots, x_S)$ in the domain of $\widehat{\text{ESS}}$ with the following property: for each $j \neq i$ s.t. $x_j > 0$, one has $0 \leq x_i \leq x_j$. If additionally at least one $j \neq i$ exists s.t. $x_j > x_i$, then $\frac{\partial}{\partial x_i}\widehat{\text{ESS}}$ is strictly positive.

An execution of the IS algorithm consists of $S \geq 1$ independent iterations of the main loop: let us denote by ω_s and ρ_s the unnormalized weights (6) generated using the LDW- ABC and IS- ABC kernel functions, respectively, at iteration $s = 1, \ldots, S$, and by $\omega = (\omega_1, \ldots, \omega_S)$ and $\rho = (\rho_1, \ldots, \rho_S)$ the resulting sequences. By definition, the set of indices $s = 1, \ldots, S$ can be partitioned into three subsets: the subset A where $\rho_s = \omega_s > 0$, the subset B where $\rho_s = 0$ and $\omega_s > 0$, and the subset C where $\rho_s = \omega_s = 0$. Moreover, for each $s \in A$ and $s' \in B$, $\omega_s > \omega_{s'}$. For notational simplicity, assume $A = \{1, \ldots, h\}$, $B = \{h + 1, \ldots, S'\}$ and $C = \{S' + 1, \ldots, S\}$, for some $0 \leq h \leq S' \leq S$. Also assume, again only for notational simplicity, that $\omega_{h+1} \geq \omega_{h+2} \geq \cdots$.

If $S' = 0$, then $h = 0$ and by definition $\widehat{\text{ESS}}_{LD} = \widehat{\text{ESS}}_{IS} = 0$, hence assume $S' > 0$. If $h = S$, then $S' = S$ and $\omega = \rho$, hence the inequality in the statement again holds trivially as equality. Consider now a case where $0 < h < S$, that is $\omega \neq \rho$. For each $i = h + 1, \ldots, S'$, consider a point $\rho_i(x) \overset{\triangle}{=} (\omega_1, \ldots, \omega_{i-1}, x, 0, \ldots 0)$, with $0 \leq x \leq \omega_i$. The fact that $0 \leq x \leq \omega_j$ for each $j < i$, and moreover that $\omega_1 > x_i$, by the above considerations entails the strict positivity of $\frac{\partial}{\partial x_i}\widehat{\text{ESS}}$ when evaluated at $\rho_i(x)$, for $0 \leq x \leq \omega_i$. Therefore, considering $i = h + 1, \ldots, S'$ in turn, we have

$$
\begin{aligned}
\widehat{\text{ESS}}_{IS} &= \widehat{\text{ESS}}(\omega_1, \ldots, \omega_h, 0, \ldots, 0) \\
&< \widehat{\text{ESS}}(\omega_1, \ldots, \omega_h, \omega_{h+1}, 0, \ldots, 0) \\
&< \cdots \\
&< \widehat{\text{ESS}}(\omega_1, \ldots, \omega_h, \omega_{h+1}, \ldots, \omega_{S'}, 0, \ldots, 0) \\
&= \widehat{\text{ESS}}_{LD}.
\end{aligned}
$$

□

B Minimization of the KL divergence

In the proposed LDW- ABC, the minimization of the KL divergence between the acceptance region and the simulated type poses a computational difficulty. This is a

Table 7 Summaries of the empirical distributions of the relative errors of the approximate distances

Min	Mean	Max	s.d.
0.0052×10^{-14}	1.0719×10^{-6}	0.0052	5.3075×10^{-7}

constrained minimization problem on a space of dimension $|\mathcal{X}|$. As $|\mathcal{X}|$ grows, this problem can rapidly become intractable.

A practical work-around to this problem can be found by considering a suitable path from $T_{\boldsymbol{y}^m}$ to $T_{\boldsymbol{x}^n}$, passing through P^*. In *Information Geometry*, this path is represented by a linear interpolation on the logarithmic scale, the *exponential geodesic* (Nielsen 2018).

Definition 3 (Exponential geodesic) Let P_1 and P_2 be two probability distributions over $\mathcal{X}$ and let P_ξ be the probability distribution such that for each $r \in \mathcal{X}$

$$\log P_\xi(r) = \xi \, \log P_1(r) + (1 - \xi) \, \log P_2(r) + \log c$$

where $\xi \in [0, 1]$ and c is a proper normalizing constant. The *exponential geodesic* between P_1 and P_2 is the following set of distributions

$$\gamma_e\big(P_1, P_2\big) \triangleq \{P_\xi : \xi \in [0, 1]\}. \tag{42}$$

Our approach when minimizing the KL divergence between $\mathcal{B}_\epsilon(T_{\boldsymbol{x}^n})$ and $T_{\boldsymbol{y}^m}$ is to focus on a path between the observed and the simulated type, that is the exponential geodesic $\gamma_e(T_{\boldsymbol{x}^n}, T_{\boldsymbol{y}^m})$. We search in this path the information projection P^*, or an approximation of it. This reduces the dimension of the minimization problem from $|\mathcal{X}|$ to 1, that of the parameter ξ. Specifically, let $P_{\xi^*} \in \mathcal{B}_\epsilon(T_{\boldsymbol{x}^n})$ be the element of $\gamma_e(T_{\boldsymbol{x}^n}, T_y)$ defined as

$$P_{\xi^*}(r) \triangleq T_{\boldsymbol{x}^n}(r)^{\xi^*} \cdot T_{\boldsymbol{y}^m}(r)^{1-\xi^*} c^* \quad (r \in \mathcal{X})$$
$$\text{where } \xi^* \triangleq \underset{\xi \in [0,1]\,:\, T_\xi \in \mathcal{B}_\epsilon}{\operatorname{argmin}} \ D(P_\xi || T_{\boldsymbol{y}^m}).$$

We have empirically verified that $D(P_{\xi^*} || T_{\boldsymbol{y}^m})$ approximates with very good accuracy $D(\mathcal{B}_\epsilon || T_{\boldsymbol{y}^m})$.

Hence, whatever $|\mathcal{X}|$, $D(\mathcal{B}_\epsilon || T_{\boldsymbol{y}^m})$ is approximate by means of a minimization with respect to a single parameter, ξ. Table 7 summarizes the distribution of the distances approximation relative errors w.r.t. the true distance, over the $S = 100{,}000$ simulations in the experiment in Sect. 5.1, with $m = 500$ and $\epsilon = 0.005$.
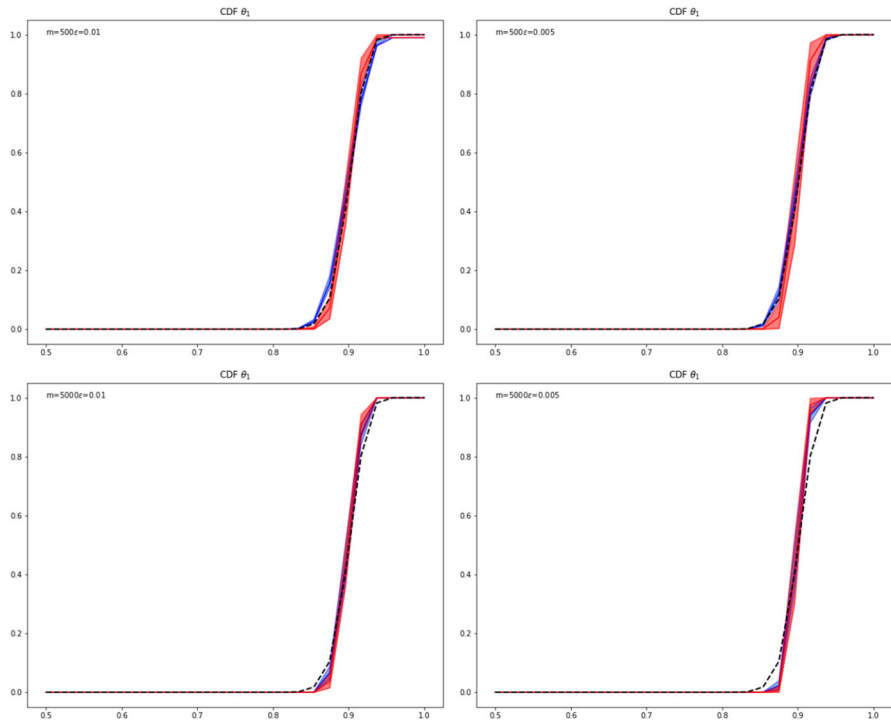
Fig. 4 Posterior cumulative density functions for θ_1. Each plot shows in blue the output of LDW- ABC, in red the output of R- ABC and in black the true cumulative density function for a pair (m, ϵ). For $\theta_1 < 0.5$ both the cumulative density functions are equal to 0. The 90% intervals over 100 runs of each algorithm are also shown (color figure online)

C Additional results from the experiments
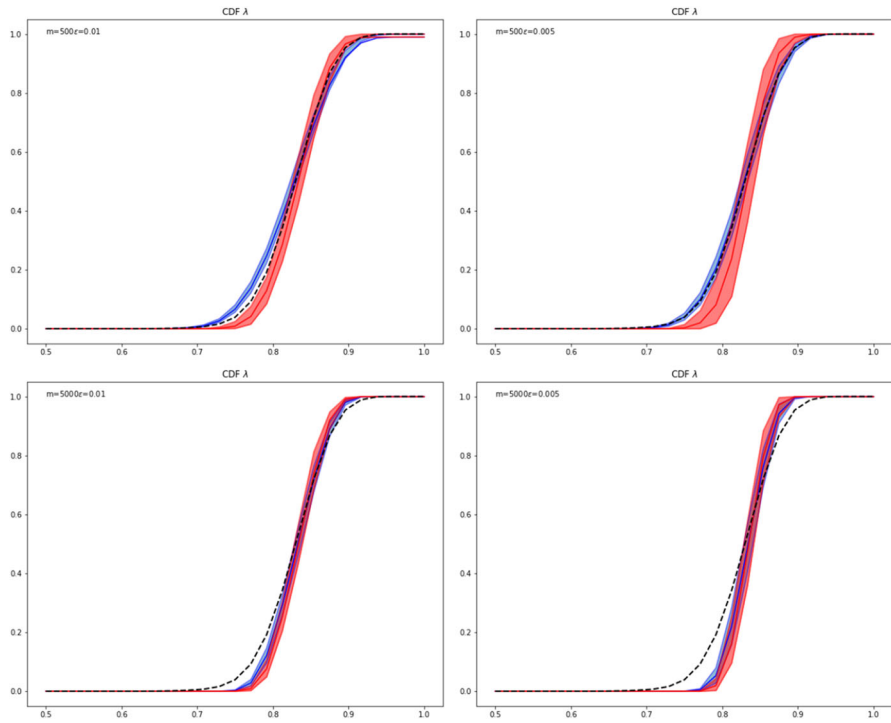
C.1 Example 1

Fig. 5 Posterior cumulative density functions for λ. Each plot shows in blue the output of LDW- ABC, in red the output of R- ABC and in black the true cumulative density function for a pair (m, ϵ). For $\lambda > 0.5$ both the cumulative density functions equal 1. The 90% intervals over 100 runs of each algorithm are also shown (color figure online)

Table 8 Posterior means via MCMC. Each column corresponds to the vector of posterior means for an element of $\left\{ \boldsymbol{\theta}_{R|s} : s \in \{\text{Government, Self-employed, Private, Without pay}\} \right\}$

| | Posterior Means | | | |
	Gov.	Self-emp	Private	Without pay
White				
R	0.3991	0.3854	0.3859	0.2507
MCMC	0.249	0.2494	0.2505	0.2501
LD	0.3909	0.3774	0.3805	0.2389
Asian-Pac-Islander				
R	0.1968	0.2015	0.1918	0.2507
MCMC	0.2512	0.2495	0.2501	0.2501
LD	0.1999	0.2041	0.1938	0.2530
Black				
R	0.2428	0.2375	0.2527	0.2486
MCMC	0.2502	0.2519	0.2492	0.2496
LD	0.2438	0.2389	0.2505	0.2492
Other				
R	0.1613	0.1756	0.1696	0.2500
MCMC	0.2496	0.2492	0.2501	0.2502
LD	0.1654	0.1796	0.1727	0.2438

C.2. Example 2

Table 8 shows the posterior means for each $\theta_{R|s}$ estimated via MCMC. Such estimates are used as a benchmark in the computation of the $\widehat{\text{MSE}}$'s shown in Table 9.

Table 9 Squared errors averaged over 100 runs of ABC. Each column corresponds to an element of $\{\boldsymbol{\theta}_{R|s} : s \in \{\text{Government, Self-employed, Private, Without pay}\}\}$

	MSE Government	Self-emp	Private	Without pay
White				
LD	1.997×10^{-3}	3.121×10^{-3}	3.958×10^{-2}	1.088×10^{-6}
R	3.141×10^{-3}	5.056×10^{-3}	7.893×10^{-2}	2.343×10^{-6}
Asian-Pac-Islander				
LD	2.976×10^{-4}	3.908×10^{-4}	6.109×10^{-3}	8.338×10^{-7}
R	4.576×10^{-4}	6.416×10^{-4}	1.147×10^{-2}	1.902×10^{-6}
Black				
LD	3.489×10^{-6}	1.194×10^{-5}	3.555×10^{-4}	1.212×10^{-6}
R	2.410×10^{-6}	2.628×10^{-6}	1.782×10^{-3}	2.651×10^{-6}
Other				
LD	6.701×10^{-4}	1.078×10^{-3}	1.039×10^{-2}	1.02×10^{-6}
R	1.241×10^{-3}	2.038×10^{-3}	1.733×10^{-2}	2.353×10^{-6}

References

Beaumont MA (2010) Approximate bayesian computation in evolution and ecology. Annu Rev Ecol Evol Syst 41:379–406

Beaumont MA, Zhang W, Balding DJ (2002) Approximate bayesian computation in population genetics. Genetics 162(4):2025–2035

Beaumont MA, Cornuet JM, Marin JM, Robert CP (2009) Adaptive approximate Bayesian computation. Biometrika 96(4):983–990

Bernton E, Jacob PE, Gerber M, Robert CP (2019) Approximate Bayesian computation with the Wasserstein distance. J R Stat Soc Ser B (Stat Methodol) 81(2):235–269

Boreale M, Corradi F, Viscardi C (2020) Relative privacy threats and learning from anonymized data. IEEE Trans Inf Forensics Secur 15:1379–1393

Buzbas EO, Rosenberg NA (2015) Aabc: approximate approximate Bayesian computation for inference in population-genetic models. Theor Popul Biol 99:31–42

Cappé O, Douc R, Guillin A, Marin JM, Robert CP (2008) Adaptive importance sampling in general mixture classes. Stat Comput 18(4):447–459

Chiachio M, Beck JL, Chiachio J, Rus G (2014) Approximate Bayesian computation by subset simulation. SIAM J Sci Comput 36(3):A1339–A1358

Cover TM, Thomas JA (2006) Elements of information theory. Wiley

Cox DR, Hinkley DV (1979) Theoretical statistics. Chapman and Hall/CRC

Csiszár I (1998) The method of types [information theory]. IEEE Trans Inf Theory 44(6):2505–2523

Csiszár I, Shields PC et al (2004) Information theory and statistics: a tutorial. Found Trends Commun Inf Theory 1(4):417–528

Del Moral P, Doucet A, Jasra A (2012) An adaptive sequential monte Carlo method for approximate Bayesian computation. Stat Comput 22(5):1009–1020

Diebolt J, Robert CP (1994) Estimation of finite mixture distributions through Bayesian sampling. J Roy Stat Soc Ser B (Methodol) 56(2):363–375

Elvira V, Martino L, Robert CP (2018) Rethinking the effective sample size. arXiv preprint arXiv:1809.04129

Fisher RA (1930) The genetical theory of natural selection. The Clarendon Press

Genz A, Joyce P (2003) Computation of the normalization constant for exponentially weighted Dirichlet distribution integrals. Comput Sci Stat 35:557–563

Jiang B (2018) Approximate Bayesian computation with Kullback–Leibler divergence as data discrepancy. In: International conference on artificial intelligence and statistics, pp 1711–1721

Joyce P, Genz A, Buzbas EO (2012) Efficient simulation and likelihood methods for non-neutral multi-allele models. J Comput Biol 19(6):650–661

Karabatsos G, Leisen F et al (2018) An approximate likelihood perspective on ABC methods. Stat Surv 12:66–104

Kifer D (2009) Attacks on privacy and Definetti's theorem. In: Proceedings of the 2009 ACM SIGMOD international conference on management of data, pp 127–138

Kohavi R, Becker B (1996) Uci machine learning repository: adult data set

Kong A (1992) A note on importance sampling using standardized weights. University of Chicago, Dept of Statistics, Tech Rep, p 348

Lintusaari J, Gutmann MU, Dutta R, Kaski S, Corander J (2017) Fundamentals and recent developments in approximate Bayesian computation. Syst Biol 66(1):e66–e82

Liu JS (2008) Monte Carlo strategies in scientific computing. Springer

Marin JM, Mengersen K, Robert CP (2005) Bayesian modelling and inference on mixtures of distributions. Handb Stat 25:459–507

Marjoram P, Molitor J, Plagnol V, Tavaré S (2003) Markov chain monte Carlo without likelihoods. Proc Nat Acad Sci 100(26):15324–15328

Nielsen F (2018) What is...an information projection. Not AMS 65(3):321–324

Park M, Jitkrittum W, Sejdinovic D (2016) K2-abc: approximate Bayesian computation with kernel embeddings. In: Proceedings of machine learning research

Prangle D (2016) Lazy abc. Stat Comput 26(1–2):171–185

Prangle D, Everitt RG, Kypraios T (2018) A rare event approach to high-dimensional approximate Bayesian computation. Stat Comput 28(4):819–834

Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW (1999) Population growth of human y chromosomes: a study of y chromosome microsatellites. Mol Biol Evol 16(12):1791–1798

Raynal L, Marin JM, Pudlo P, Ribatet M, Robert CP, Estoup A (2019) Abc random forests for Bayesian parameter inference. Bioinformatics 35(10):1720–1728

Robert C, Casella G (2013) Monte Carlo statistical methods. Springer

Rubin DB (1984) Bayesianly justifiable and relevant frequency calculations for the applies statistician. Ann Stat 12:1151–1172

Sisson SA, Fan Y, Tanaka MM (2007) Sequential monte Carlo without likelihoods. Proc Nat Acad Sci 104(6):1760–1765

Sisson SA, Fan Y, Beaumont M (2018) Handbook of approximate Bayesian computation. Chapman and Hall/CRC

Tavaré S, Balding DJ, Griffiths RC, Donnelly P (1997) Inferring coalescence times from DNA sequence data. Genetics 145(2):505–518

Turan MS, Barker E, Kelsey J, McKay KA, Baish ML, Boyle M (2018) Recommendation for the entropy sources used for random bit generation. NIST Spec Publ 800(90B)

Wong RCW, Fu AWC, Wang K, Yu PS, Pei J (2011) Can the utility of anonymized data be used for privacy breaches? ACM Trans Knowl Discov Data (TKDD) 5(3):1–24

Xiao X, Tao Y (2006) Anatomy: simple and effective privacy preservation. In: Proceedings of the 32nd international conference on Very large data bases, pp 139–150