APPROXIMATE BAYESIAN COMPUTATION WITH THE SLICED-WASSERSTEIN DISTANCE

Kimia Nadjahi[1], Valentin De Bortoli[2], Alain Durmus[2], Roland Badeau[1], Umut Şimşekli[1]

[1] LTCI, Télécom Paris, Institut Polytechnique de Paris, France
[2] CMLA, École normale supérieure Paris-Saclay, CNRS, Université Paris-Saclay, France

ABSTRACT

Approximate Bayesian Computation (ABC) is a popular method for approximate inference in generative models with intractable but easy-to-sample likelihood. It constructs an approximate posterior distribution by finding parameters for which the simulated data are close to the observations in terms of summary statistics. These statistics are defined beforehand and might induce a loss of information, which has been shown to deteriorate the quality of the approximation. To overcome this problem, Wasserstein-ABC has been recently proposed, and compares the datasets via the Wasserstein distance between their empirical distributions, but does not scale well to the dimension or the number of samples. We propose a new ABC technique, called Sliced-Wasserstein ABC and based on the Sliced-Wasserstein distance, which has better computational and statistical properties. We derive two theoretical results showing the asymptotical consistency of our approach, and we illustrate its advantages on synthetic data and an image denoising task.

Index Terms— Likelihood-free inference, Approximate Bayesian Computation, Optimal transport, Sliced-Wasserstein distance

1. INTRODUCTION

Consider the problem of estimating the posterior distribution of some model parameters $\theta \in \mathbb{R}^{d_\theta}$ given n data points $\mathbf{y}_{1:n} \in \mathsf{Y}^n$. This distribution has a closed-form expression given by the Bayes' theorem up to a multiplicative constant: $\pi(\theta|\mathbf{y}_{1:n}) \propto \pi(\mathbf{y}_{1:n}|\theta)\pi(\theta)$. For many statistical models of interest, the likelihood $\pi(\mathbf{y}_{1:n}|\theta)$ cannot be numerically evaluated in a reasonable amount of time, which prevents the application of classical likelihood-based approximate inference methods. Nevertheless, in various settings, even if the associated likelihood is numerically intractable, one can still generate synthetic data given any model parameter value. This generative setting gave rise to an alternative framework of likelihood-free inference techniques. Among them, Approximate Bayesian Computation (ABC, [1, 2]) methods have become a popular choice and have proven useful in various practical applications (e.g., [3, 4, 5]). The core idea of ABC is to bypass calculation of the likelihood by using simulations: the exact posterior is approximated by retaining the parameter values for which the synthetic data are close enough to the observations. Closeness is usually measured with a discrepancy measure between the two datasets reduced to some 'summary statistics' (e.g., empirical mean or empirical covariance). While summaries allow a practical and efficient implementation of ABC, especially in high-dimensional data spaces, the quality of the approximate posterior distribution highly depends on them and constructing sufficient statistics is a non-trivial task. Summary statistics can be designed by hand using expert knowledge, which can be tedious especially in real-world applications, or in an automated way, for instance see [6].

Recently, discrepancy measures that view data sets as empirical probability distributions to eschew the construction of summary statistics have been proposed for ABC. Examples include the Kullback-Leibler divergence (KL, [7]), maximum mean discrepancy [8], and Wasserstein distance [9]. This latter distance emerges from the optimal transport (OT) theory and has attracted abundant attention in statistics and machine learning due to its strong theoretical properties and applications on many domains. In particular, it has the ability of making meaningful comparisons even between probability measures with non-overlapping supports, unlike KL. However, the computational complexity of the Wasserstein distance rapidly becomes a challenge when the dimension of the observations is large. Several numerical methods have been proposed during the past few years to speed-up this computation. Wasserstein-ABC (WABC, [9]) uses an approximation based on the Hilbert space-filling curve and termed the Hilbert distance, which is computationally efficient but accurate for small dimensions only. Besides, under a general setting, the Wasserstein distance suffers from a curse of dimensionality in the sense that the error made when approximating it from samples grows exponentially fast with the data space dimension [10]. These computational and statistical burdens can strongly affect the performance of WABC applied to high-dimensional data.

The Sliced-Wasserstein distance (SW, [11, 12]) is an alternative OT distance and leverages the attractive property that the Wasserstein distance between one-dimensional measures has an analytical form which can be efficiently approximated. SW is defined as an average of one-dimensional Wasserstein distances, and therefore has a significantly better computational complexity. Several recent studies have reported empirical success on generative modeling with SW [13, 14, 15] as well as nice asymptotic and statistical properties [16, 17, 18], making this alternative distance increasingly popular. In this paper, we develop a novel ABC framework that uses SW as the data discrepancy measure. This defines a likelihood-free method which does not require choosing summary statistics and is efficient even with high-dimensional observations. We derive asymptotical guarantees on the convergence of the resulting ABC posterior, and we illustrate the superior empirical performance of our methodology by applying it on a synthetical problem and an image denoising task.

2. BACKGROUND

Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with associated expectation operator $\mathbb{E}$, on which all the random variables are defined. Let $(Y_k)_{k \in \mathbb{N}^*}$ be a sequence of independent and identically distributed (i.i.d.) random variables associated with some observations $(\mathbf{y}_k)_{k \in \mathbb{N}^*}$ valued in $\mathsf{Y} \subset \mathbb{R}^d$. Denote by $\mu_\star$ the common distribution of

Pseudo-code 1: Vanilla ABC.

Input: observations $\mathbf{y}_{1:n}$, number of iterations T, data
discrepancy measure $\mathbf{D}$, summary statistics s,
tolerance threshold $\varepsilon > 0$.

for $t = 1, \ldots, T$ **do**
  repeat
    $\theta \sim \pi(\cdot)$ and $\mathbf{z}_{1:m} \sim \mu_\theta$ i.i.d.
  until $\mathbf{D}\big(s(\mathbf{y}_{1:n}), s(\mathbf{z}_{1:m})\big) \leq \varepsilon$;
  $\theta^{(t)} = \theta$
return $\theta^{(1)}, \ldots, \theta^{(T)}$

$(Y_k)_{k \in \mathbb{N}^*}$ and by $\mathcal{P}(\mathsf{Y})$ the set of probability measures on Y. For any $n \in \mathbb{N}^*$, $\hat{\mu}_n = (1/n) \sum_{i=1}^n \delta_{Y_i}$ denotes the empirical distribution corresponding to n observations. Consider a statistical model $\mathcal{M}_\Theta = \{\mu_\theta \in \mathcal{P}(\mathsf{Y}) \, : \, \theta \in \Theta\}$ parametrized by $\Theta \subset \mathbb{R}^{d_\theta}$. We focus on parameter inference for purely generative models: for any $\theta \in \Theta$, we can draw i.i.d. samples $(Z_k)_{k \in \mathbb{N}^*} \in \mathsf{Y}^{\mathbb{N}^*}$ from μ_θ, but the numerical evaluation of the likelihood is not possible or too expensive. For any $m \in \mathbb{N}^*$, $\hat{\mu}_{\theta,m} = (1/m) \sum_{i=1}^m \delta_{Z_i}$ is the empirical distribution of m i.i.d. samples generated by μ_θ, $\theta \in \Theta$. We assume that (a) Y, endowed with the Euclidean distance ρ, is a Polish space (*i.e.*, complete and separable), (b) Θ, endowed with the distance ρ_Θ, is a Polish space, (c) parameters are identifiable, *i.e.* $\mu_\theta = \mu_{\theta'}$ implies $\theta = \theta'$. $\mathcal{B}(\mathsf{Y})$ denotes the Borel σ-field of (Y, ρ).

Approximate Bayesian Computation. ABC methods are used to approximate the posterior distribution in generative models when the likelihood is numerically intractable but easy to sample from. The basic and simplest ABC algorithm is an acceptance-rejection method [1], which iteratively draws a candidate parameter θ' from a prior distribution π, and 'synthetic data' $\mathbf{z}_{1:m} = (\mathbf{z}_i)_{i=1}^m$ from $\mu_{\theta'}$, and keeps θ' if $\mathbf{z}_{1:m}$ is close enough to the observations $\mathbf{y}_{1:n} = (\mathbf{y}_i)_{i=1}^n$. Specifically, the acceptance rule is $\mathbf{D}\big(s(\mathbf{y}_{1:n}), s(\mathbf{z}_{1:m})\big) \leq \varepsilon$, where $\mathbf{D}$ is a data discrepancy measure taking non-negative values, ε is a tolerance threshold, and $s : \sqcup_{n \in \mathbb{N}^*} \mathsf{Y}^n \to \mathbb{R}^{d_s}$ with small d_s is a summary statistics. The algorithm is summarized in Pseudo-code 1 and returns samples of θ that are distributed from:

$$\pi_{\mathbf{y}_{1:n}}^\varepsilon(\theta) = \frac{\pi(\theta) \int_{\mathsf{Y}^m} \mathbb{1}\{\mathbf{D}\big(s(\mathbf{y}_{1:n}), s(\mathbf{z}_{1:m})\big) \leq \varepsilon\} \mathrm{d}\mu_\theta(\mathbf{z}_{1:m})}{\int_\Theta \mathrm{d}\pi(\theta) \int_{\mathsf{Y}^m} \mathbb{1}\{\mathbf{D}\big(s(\mathbf{y}_{1:n}), s(\mathbf{z}_{1:m})\big) \leq \varepsilon\} \mathrm{d}\mu_\theta(\mathbf{z}_{1:m})} \tag{1}$$

The choice of $s(\cdot)$ directly impacts the quality of the resulting approximate posterior: if the statistics are sufficient statistics, $\pi_{\mathbf{y}_{1:n}}^\varepsilon(\theta)$ converges to the true posterior $\pi(\theta|\mathbf{y}_{1:n})$ as $\varepsilon \to 0$, otherwise, the limiting distribution is at best $\pi(\theta|s(\mathbf{y}_{1:n}))$ [19, 20]. Wasserstein-ABC has been proposed to avoid this loss of information.

Wasserstein distance and ABC. For $p \geq 1$, consider $\mathcal{P}_p(\mathsf{Y}) = \{\mu \in \mathcal{P}(\mathsf{Y}) \, : \, \int_\mathsf{Y} \|y - y_0\|^p \mathrm{d}\mu(y) < +\infty\}$ for some $y_0 \in \mathsf{Y}$ and the Wasserstein distance of order p defined for any $\mu, \nu \in \mathcal{P}_p(\mathsf{Y})$ by,

$$\mathbf{W}_p^p(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu, \nu)} \left\{ \int_{\mathsf{Y} \times \mathsf{Y}} \|x - y\|^p \, \mathrm{d}\gamma(x, y) \right\} , \tag{2}$$

where $\Gamma(\mu, \nu)$ is the set of probability measures γ on $(\mathsf{Y} \times \mathsf{Y}, \mathcal{B}(\mathsf{Y}) \otimes \mathcal{B}(\mathsf{Y}))$ verifying: $\forall A \in \mathcal{B}(\mathsf{Y}), \gamma(A \times \mathsf{Y}) = \mu(A), \gamma(\mathsf{Y} \times A) = \nu(A)$.

Evaluating the Wasserstein distance between multi-dimensional probability measures turns out to be numerically intractable in general, and solving (2) between empirical distributions over n samples leads to computational costs in $\mathcal{O}(n^3 \log(n))$ [21]. Nevertheless, $\mathbf{W}_p$ between one-dimensional measures $\mu, \nu \in \mathcal{P}_p(\mathbb{R})$ has a closed-form

expression [22, Theorem 3.1.2.(a)], given by:

$$\mathbf{W}_p^p(\mu, \nu) = \int_0^1 \left| F_\mu^{-1}(t) - F_\nu^{-1}(t) \right|^p \mathrm{d}t , \tag{3}$$

where F_μ^{-1} and F_ν^{-1} denote the quantile functions of μ and ν respectively. For empirical one-dimensional distributions, (3) can be efficiently approximated by simply sorting the n samples drawn from each distribution and computing the average cost between the sorted samples. This amounts to $\mathcal{O}(n \log(n))$ operations at worst.

Wasserstein-ABC [9] is a variant of ABC (1) that uses $\mathbf{W}_p, p \geq 1$ between the empirical distributions of the observed and synthetic data, in place of the discrepancy measure $\mathbf{D}$ between summaries. To make this method scalable to any dataset size, [9] introduces a new approximation of (2), the Hilbert distance, which extends the idea behind the computation of $\mathbf{W}_p$ in 1D to higher dimensions, by sorting samples according to their projection obtained via the Hilbert space-filling curve. This alternative can be computed in $\mathcal{O}(n \log(n))$, but yields accurate approximations only for low dimensions [9]. They also use a second approximation, the swapping distance, based on an iterative greedy swapping algorithm. However, each iteration requires n^2 operations, and there is no guarantee of convergence to $\mathbf{W}_p$.

3. SLICED-WASSERSTEIN ABC

Sliced-Wasserstein distance. The analytical expression of $\mathbf{W}_p$ in (3) motivates the formulation of an alternative OT distance, called the Sliced-Wasserstein distance [11, 12]. SW is obtained by reducing multi-dimensional distributions to one-dimensional representations through linear projections, and then by averaging 1D Wasserstein distances between these projected distributions. More formally, we denote by $\mathbb{S}^{d-1} = \{u \in \mathbb{R}^d \, : \, \|u\|_2 = 1\}$ the d-dimensional unit sphere, and by $\langle \cdot, \cdot \rangle$ the Euclidean inner-product. For any $u \in \mathbb{S}^{d-1}$, $u^\star$ is the linear form associated with u such that for any $y \in \mathsf{Y}$, $u^\star(y) = \langle u, y \rangle$. For $p \geq 1$, the Sliced-Wasserstein distance of order p between $\mu, \nu \in \mathcal{P}_p(\mathsf{Y})$ is defined as,

$$\mathbf{SW}_p^p(\mu, \nu) = \int_{\mathbb{S}^{d-1}} \mathbf{W}_p^p(u_\sharp^\star \mu, u_\sharp^\star \nu) \mathrm{d}\boldsymbol{\sigma}(u) , \tag{4}$$

where $\boldsymbol{\sigma}$ is the uniform distribution on $\mathbb{S}^{d-1}$ and for any measurable function $f : \mathsf{Y} \to \mathbb{R}$ and $\zeta \in \mathcal{P}(\mathsf{Y})$, $f_\sharp \zeta$ is the push-forward measure of ζ by f, defined as: $\forall A \in \mathcal{B}(\mathbb{R}), f_\sharp \zeta(A) = \zeta(f^{-1}(A))$, with $f^{-1}(A) = \{y \in \mathsf{Y} \, : \, f(y) \in A\}$. $\mathbf{SW}_p$ is a distance on $\mathcal{P}_p(\mathsf{Y})$ [16] with significantly lower computational requirements than the Wasserstein distance: in practice, the integration in (4) is approximated with a finite-sample average, using a simple Monte Carlo (MC) scheme.

SW also seems to have better statistical properties than the Wasserstein distance and its approximations. We illustrate it with the task of estimating the scaling factor of the covariance matrix in a multivariate Normal model, as in the supplementary material of [9]. For any $\sigma > 0$, denote by μ_σ the d-dimensional Gaussian distribution with zero-mean and covariance matrix $\sigma^2 \mathbf{I}_d$. Observations are assumed i.i.d. from $\mu_{\sigma_\star}$ with $\sigma_\star^2 = 4$, and we draw the same number of i.i.d. data from μ_σ for 100 values of σ^2 equispaced between 0.1 and 9. We then compute $\mathbf{W}_2$ and $\mathbf{SW}_2$ between the empirical distributions of the samples, and the swapping and Hilbert approximations presented in [9], for $d \in \{2, 10, 100\}$ and 1000 observations. $\mathbf{W}_2$ between two Gaussian measures has an analytical formula, which boils down in our setting to: $\mathbf{W}_2^2(\mu_{\sigma_\star}, \mu_\sigma) = d(\sigma_\star - \sigma)^2$, and we approximate the exact SW using a MC approximation of:

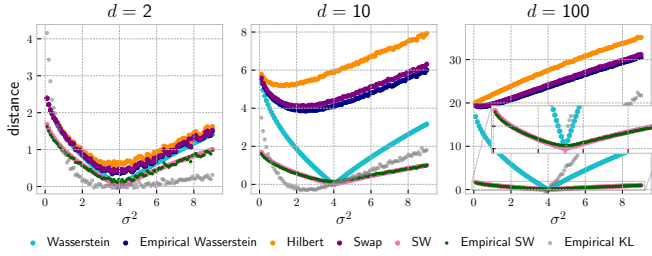$$\mathbf{SW}_2^2(\mu_{\sigma_\star}, \mu_\sigma) = \mathbf{W}_2^2(\mu_{\sigma_\star}, \mu_\sigma) \int_{\mathbb{S}^{d-1}} u^T u \, \mathrm{d}\boldsymbol{\sigma}(u) ,$$

Fig. 1: Comparison of OT distances and KL between data generated from d-dimensional Gaussian distributions μ_σ vs. $\mu_{\sigma_\star}$, $\sigma_\star^2 = 4$, with 1000 i.i.d draws. SW is approximated with 100 random projections.

This formula is derived from (4) and the exact $\mathbf{W}_2$ between one-dimensional Gaussian distributions. We also compute KL with the estimator proposed for KL-based ABC (KL-ABC, [7]). Figure 1 shows the distances plotted against σ^2 for each d. When the dimension increases, we observe that (i) as pointed out in [9], the quality of the approximation of empirical Wasserstein returned by Hilbert and swapping rapidly deteriorates, and (ii) SW, approximated using densities or samples, is the only approximate metric that attains its minimum at $\sigma_\star^2$. This curse of dimensionality can be a limiting factor for the performance of WABC and KL-ABC in high dimensions.

Sliced-Wasserstein ABC. Motivated by the practical success of SW regardless of the dimension value in the previous experiment, we propose a variant of ABC based on SW, referred to as the Sliced-Wasserstein ABC (SW-ABC). Our method is similar to WABC in the sense that it compares empirical distributions, but instead of $\mathbf{W}_p$, we choose the discrepancy measure to be $\mathbf{SW}_p$, $p \geq 1$. The usage of SW allows the method to scale better to the data size and dimension. The resulting posterior distribution, called the SW-ABC posterior, is thus defined in (1) with $\mathbf{D}$ replaced by $\mathbf{SW}_p$.

4. THEORETICAL STUDY

In this section, we analyze the asymptotic behavior of the SW-ABC posterior under two different regimes. Our first result concerns the situation where the observations $\mathbf{y}_{1:n}$ are fixed, and ε goes to zero. We prove that the SW-ABC posterior is asymptotically consistent in the sense that it converges to the true posterior, under specific assumptions on the density used to generate synthetic data.

Proposition 1. *Let $p \geq 1$. Suppose that μ_θ has a density f_θ w.r.t. the Lebesgue measure such that f_θ is continuous and there exists $\mathcal{N}_\Theta \subset \Theta$ satisfying $\sup_{\theta \in \Theta \setminus \mathcal{N}_\Theta} f_\theta(\mathbf{y}_{1:n}) < \infty$ and $\pi(\mathcal{N}_\Theta) = 0$. In addition, assume that there exists $\bar{\varepsilon} > 0$ such that $\sup_{\theta \in \Theta \setminus \mathcal{N}_\Theta} \sup_{\mathbf{z}_{1:m} \in \mathsf{A}^{\bar{\varepsilon}}} f_\theta(\mathbf{z}_{1:m}) < \infty$, where $\mathsf{A}^{\bar{\varepsilon}} = \{\mathbf{z}_{1:m} : \mathbf{SW}_p(\mathbf{y}_{1:n}, \mathbf{z}_{1:m}) \leq \bar{\varepsilon}\}$. Then, with $\mathbf{y}_{1:n}$ fixed, the SW-ABC posterior converges to the true posterior as ε goes to 0, in the sense that, for any measurable $\mathsf{B} \subset \Theta$, $\lim_{\varepsilon \to 0} \pi^\varepsilon_{\mathbf{y}_{1:n}}(\mathsf{B}) = \pi(\mathsf{B}|\mathbf{y}_{1:n})$, where $\pi^\varepsilon_{\mathbf{y}_{1:n}}$ is defined by (1).*

Proof of Proposition 1. The proof consists in applying [9, Proposition 3.1], which establishes the conditions for the data discrepancy measure to yield an ABC posterior that converges to the true posterior in the asymptotic regime we consider. This amounts to verify that:

(i) For any $\mathbf{y}_{1:n}$ and $\mathbf{z}_{1:m}$, with respective empirical distributions $\hat{\mu}_n$ and $\hat{\mu}_{\theta,m}$, $\mathbf{SW}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m}) = 0$ if and only if $\hat{\mu}_n = \hat{\mu}_{\theta,m}$.

(ii) $\mathbf{SW}_p$ is continuous in the sense that, if $(\mathbf{z}_{1:m}^k)_{k \in \mathbb{N}}$ converges to $\mathbf{z}_{1:m}$ in the metric ρ, then, for any empirical distribution $\hat{\mu}_n$,

$\lim_{k \to \infty} \mathbf{SW}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m}^k) = \mathbf{SW}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m})$, where $\hat{\mu}_{\theta,m}^k$ is the empirical measure of $\mathbf{z}_{1:m}^k$.

Condition (i) follows from the fact that $\mathbf{SW}_p$ is a distance [16, Proposition 5.1.2]. Now, let $\mathbf{y}' \in \mathsf{Y}$ and $\psi : \mathsf{Y} \to \mathbb{R}$ be a continuous function such that for any $\mathbf{y} \in \mathsf{Y}$, $|\psi(\mathbf{y})| \leq K(1 + \rho(\mathbf{y}', \mathbf{y})^p)$ with $K \in \mathbb{R}$. Since $(\mathbf{z}_{1:m}^k)_{k \in \mathbb{N}}$ converges to $\mathbf{z}_{1:m}$ in the metric ρ and ψ is continuous, we get that $\lim_{k \to \infty} \int \psi \, \mathrm{d}\hat{\mu}_{\theta,m}^k = \int \psi \, \mathrm{d}\hat{\mu}_{\theta,m}$. This implies that $\hat{\mu}_{\theta,m}^k$ weakly converges to $\hat{\mu}_{\theta,m}$ in $\mathcal{P}_p(\mathsf{Y})$ [23, Definition 6.7], which, by [23, Theorem 6.8], is equivalent to $\lim_{k \to \infty} \mathbf{W}_p(\hat{\mu}_{\theta,m}^k, \hat{\mu}_{\theta,m}) = 0$. By applying the triangle inequality and [16, Proposition 5.1.3], there exists $C \geq 0$ such that, for any empirical measure $\hat{\mu}_n$,

$$|\mathbf{SW}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m}^k) - \mathbf{SW}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m})| \leq \mathbf{SW}_p(\hat{\mu}_{\theta,m}^k, \hat{\mu}_{\theta,m})$$
$$\leq C^{1/p} \, \mathbf{W}_p(\hat{\mu}_{\theta,m}^k, \hat{\mu}_{\theta,m}) \ .$$

We conclude that $\lim_{k \to \infty} \mathbf{SW}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m}^k) = \mathbf{SW}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m})$, making condition (ii) applicable. $\square$

Next, we study the limiting SW-ABC posterior when the value of ε is fixed and the number of observations increases, *i.e.* $n \to \infty$. We suppose the size m of the synthetic dataset grows to αn with $\alpha > 0$, such that m can be written as a function of n, $m(n)$, satisfying $\lim_{n \to \infty} m(n) = \infty$. We show that, under this setting and appropriate additional conditions, the resulting approximate posterior converges to the prior distribution on θ restricted to the region $\{\theta \in \Theta : \mathbf{SW}_p(\mu_{\theta_\star}, \mu_\theta) \leq \varepsilon\}$.

Proposition 2. *Let $p \geq 1$, $\varepsilon > 0$ and $(m(n))_{n \in \mathbb{N}^*}$ be an increasing sequence satisfying $\lim_{n \to \infty} m(n)/n = \alpha$, for $\alpha > 0$. Assume that the statistical model $\mathscr{M}_\Theta$ is well specified, i.e. there exists $\theta_\star \in \Theta$ such that $\mu_\star = \mu_{\theta_\star}$, and that almost surely the following holds:*

$$\lim_{n \to \infty} \mathbf{SW}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m(n)}) = \mathbf{SW}_p(\mu_{\theta_\star}, \mu_\theta) \ , \qquad (5)$$

where $\hat{\mu}_n$, $\hat{\mu}_{\theta,m(n)}$ denote the empirical distributions of the observations $\mathbf{y}_{1:n}$ and synthetic data $\mathbf{z}_{1:m(n)}$ respectively. Then, the SW-ABC posterior converges to the restriction of the prior π on the region $\{\theta \in \Theta : \mathbf{SW}_p(\mu_{\theta_\star}, \mu_\theta) \leq \varepsilon\}$ as $n \to \infty$, i.e. for any $\theta \in \Theta$,

$$\lim_{n \to \infty} \pi^\varepsilon_{\mathbf{y}_{1:n}}(\theta) = \pi(\theta|\mathbf{SW}_p(\mu_{\theta_\star}, \mu_\theta) \leq \varepsilon)$$
$$\propto \pi(\theta)\mathbb{1}\{\mathbf{SW}_p(\mu_{\theta_\star}, \mu_\theta) \leq \varepsilon\} \ .$$

Proof of Proposition 2. The result follows from the application of [7, Theorem 1] to $\mathbf{SW}_p$ and the required conditions. $\square$

Remark 1. *Condition (5) is a mild assumption, e.g. is fulfilled if Y is compact and separable: in this case, for any $\nu \in \mathcal{P}_p(\mathsf{Y})$ and its empirical instantiation $\hat{\nu}_n$, $\lim_{n \to \infty} \mathbf{W}_p(\nu, \hat{\nu}_n) = 0$ ν-almost surely [10], then $\lim_{n \to \infty} \mathbf{SW}_p(\nu, \hat{\nu}_n) = 0$ ν-almost surely [16, Proposition 5.1.3], and (5) follows by applying the triangle inequality.*

5. EXPERIMENTS

Multivariate Gaussians. As a first set of experiments, we investigate the performance of SW-ABC on a synthetical setting where the posterior distribution is analytically available. We consider $n = 100$ observations $(\mathbf{y}_i)_{i=1}^n$ i.i.d. from a d-dimensional Gaussian $\mathcal{N}(\mathbf{m}_\star, \sigma_\star^2 \mathbf{I}_d)$, with $\mathbf{m}_\star \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and $\sigma_\star^2 = 4$. The parameter θ is σ^2 for which the prior distribution is assigned to be an inverse gamma distribution $\mathcal{IG}(1, 1)$. Therefore, the posterior distribution of σ^2 given $(\mathbf{y}_i)_{i=1}^n$ and $\mathbf{m}_\star$ is $\mathcal{IG}(1 + n \cdot d/2, 1 + 2^{-1} \sum_{i=1}^n \|\mathbf{y}_i - \mathbf{m}_\star\|^2)$. We compare

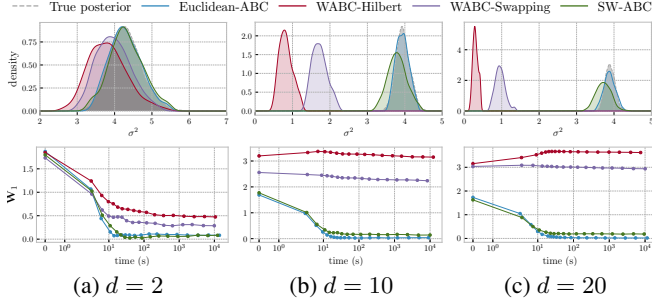(a) $d = 2$ (b) $d = 10$ (c) $d = 20$

Fig. 2: Comparison of SMC-ABC strategies in the multivariate Gaussians problem. Each strategy uses 1000 particles and are run for 3 hours max. First row shows ABC and true posteriors of σ^2, second row reports $\mathbf{W}_1$-distance to true posterior vs. time. SW is approximated with its MC estimate over 100 random projections.

SW-ABC against ABC using the Euclidean distance between sample variances (Euclidean-ABC), WABC with the Hilbert distance, WABC with the swapping distance and KL-ABC. Each ABC approximation was obtained using the sequential Monte Carlo sampler-based ABC method [24], which is more computationally efficient than vanilla ABC (1) and implemented in the package pyABC [25]. We provide our code in [26]. Figure 2 reports for $d \in \{2, 10, 20\}$, the resulting ABC posteriors and $\mathbf{W}_1$ to the true posterior vs. time. Due to the poor performance of the estimator of KL between two empirical distributions proposed in [7] (see Fig. 1), KL-ABC fails at approximating well the posterior in these experiments. Hence, we excluded those results from Fig. 2 for clarity. Euclidean-ABC provides the most accurate approximation as expected since it relies on statistics that are sufficient in our setting. WABC performs poorly with high-dimensional observations, contrary to SW-ABC, which approximates well the posterior for each dimension value and is as fast.

Image denoising. We now evaluate our approach on a real application, namely image denoising. We consider a widely used algorithm for this task, the Non-Local Means algorithm (NL-means, [27]), and we present a novel variant of it derived from SW-ABC.

We formally define the denoising problem: let $\mathbf{u} \in \mathbb{R}^{M \times N}$, denote a clean gray-level image. We observe a corrupted version of this image, $\mathbf{v} = \mathbf{u} + \mathbf{w}$, where $\mathbf{w}$ is some noise in $\mathbb{R}^{M \times N}$. The goal is to restore $\mathbf{u}$ from $\mathbf{v}$. We focus on denoising methods that consider 'patch-based representations' of images, e.g. NL-means. Specifically, let $r \in \mathbb{N}$ be a patch size and $I = \{1, \dots, M\} \times \{1, \dots, N\}$ the set of pixel positions. For $\mathbf{i} \in I$, $\mathbf{u}'(\mathbf{i})$ denotes the pixel value at position $\mathbf{i}$ in image $\mathbf{u}'$, and $P_{\mathbf{i}}$ is a $(2r+1) \times (2r+1)$ window in $\mathbf{v}$ centered at $\mathbf{i}$: for $\mathbf{k} \in \{-r, \dots, r\}^2$, $P_{\mathbf{i}}(\mathbf{k}) = \mathbf{v}(\mathbf{i} + \mathbf{k})$, where $\mathbf{v}$ is extended to $\mathbb{Z}^2$ by periodicity. Let $D \subset I$ be a dictionary of positions, and $\phi : I \to D$ such that, for $\mathbf{i} \in I$, $\phi(\mathbf{i}) = \arg\min_{\mathbf{j} \in D} \|P_{\mathbf{i}} - P_{\mathbf{j}}\|_2$, *i.e.* $\phi(\mathbf{i})$ is the position in D of the most similar patch to $P_{\mathbf{i}}$. For $\mathbf{j} \in D$, an estimator of $P_{\mathbf{j}}$ is given by $\hat{P}_{\mathbf{j}} = \mathbb{E}_{\pi(\mathbf{i}|(P_{\mathbf{k}})_{\mathbf{k} \in \phi^{-1}(\mathbf{j})})\tilde{\pi}(\mathbf{l})}[P_{\mathbf{i}+\mathbf{l}}]$, $\tilde{\pi}$ being the uniform distribution on $\phi^{-1}(\mathbf{j})$. In practice, it is approximated with a Monte Carlo scheme:

$$\hat{P}_{\mathbf{j}} \approx (Tn)^{-1} \sum_{t=1}^{T} \sum_{s=1}^{S} P_{\mathbf{i}^{(t)}+\mathbf{l}^{(s)}}, \qquad (6)$$

where $\mathbf{i}^{(t)} \sim \pi(\mathbf{i}^{(t)}|(P_{\mathbf{k}})_{\mathbf{k} \in \phi^{-1}(\mathbf{j})})$, $\mathbf{l}^{(s)} \sim \tilde{\pi}(\mathbf{l})$, and $\mathbf{i}, \mathbf{l}$ are mutually independent. Finally, we construct an estimate $\hat{\mathbf{u}}$ of $\mathbf{u}$ as follows: for any $\mathbf{i} \in I$, $\hat{\mathbf{u}}(\mathbf{i}) = \sum_{\mathbf{k}, \|\mathbf{k}-\mathbf{i}\|_\infty \le r} \hat{P}_{\phi(\mathbf{k})}(\mathbf{i} - \mathbf{k}) (2r+1)^{-2}$. The classical NL-means estimator corresponds to the case where $D = I$ (thus $\phi = \mathrm{Id}$) and for any $\mathbf{i} \in I$ and $P \in \mathbb{R}^{(2r+1) \times (2r+1)}$, $\pi(\mathbf{i}, P) \propto \mathbb{1}_W(\mathbf{i}) e^{-\|P - P_{\mathbf{i}}\|^2/(2\sigma^2)}$, where W is a search window.

Table 1: Comparison of NL-means and SW-ABC on the image denoising task in terms of average PSNR (dB). For each σ, we fine-tuned the hyperparameters of NL-means and reported the best result.

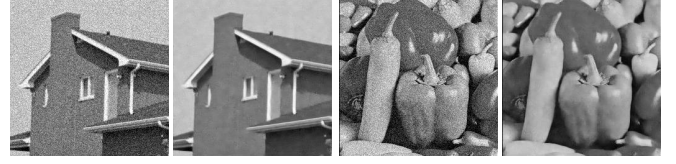|	$\sigma = 10$	$\sigma = 20$	$\sigma = 30$	$\sigma = 50$
NL-means	**30.43**	**26.32**	24.22	21.99
SW-ABC	27.09	26.26	**24.86**	**22.56**

Fig. 3: Visualization of the results. For each couple, the left one is the noisy image ($\sigma = 20$) and the right one is the output of SW-ABC.

In our work, we assume that the likelihood $\pi(P|\mathbf{i})$ is not available, but we observe for $j \in D$, $(P_{\mathbf{k}_\ell})_{\ell=1}^{m}$ ($\mathbf{k}_\ell \in \phi^{-1}(\mathbf{j})$) i.i.d. from $\pi(\cdot|\mathbf{i})$. By replacing $\pi(\mathbf{i}|(P_{\mathbf{k}_\ell})_{\ell=1}^{m})$ in (6) by the SW-ABC posterior, we obtain the proposed denoising method, called the SW-ABC NL-means algorithm. We provide our Python implementation in [26].

We compare our approach with the classical NL-means. We consider one of the standard image denoising datasets [28], called CBSD68 [29] and consisting of 68 colored images of size 321×481. We first convert the images to gray scale, then manually corrupt each of them with a Gaussian noise with standard deviation σ, and try to recover the clean image. The quality of the output images is evaluated with the Peak Signal to Noise Ratio (PSNR) measure: $\mathrm{PSNR} = -10 \log_{10}\{\|\mathbf{u} - \hat{\mathbf{u}}\|_2^2/(255^2 NM)\}$. In our experiments, we use a dictionary of 1000 patches picked uniformly at random, we set $T = S = m = 10$, $r = 3$, $W = \{-10, \dots, 10\}^2$, $\varepsilon = (2r+1)^2$, and we compute SW with a MC scheme over $L = 100$ projections.

We report the average PSNR values for different values of the noise level σ in Table 1. We observe that for small σ, NL-means provides more accurate results, whereas when σ becomes larger SW-ABC outperforms NL-means, thanks to the patch representation and the use of SW. On the other hand, another important advantage of SW-ABC becomes prominent in the computation time: the proposed approach takes $\approx 6s$ on a standard laptop computer per image whereas the classical NL-means algorithm takes $\approx 30s$. Indeed, the computational complexity of SW-ABC NL-means is upper-bounded by $|D| TSC_{\mathrm{SW}}$, with $C_{\mathrm{SW}} = Lm \log(m)$ is the cost of computing SW, whereas it is $NM|W|(2r+1)^2$ for the naïve implementation of NL-means, where $|A|$ denotes the cardinal number of the set A. We can observe that SW-ABC has a lower computational complexity since $|D| \ll NM$ in practice. We note that the computation time of NL-means can be improved by certain acceleration techniques, which can be directly used to improve the speed of SW-ABC NL-means as well. Finally, in Figure 3, we illustrate the performance of SW-ABC on two 512×512 images for visual inspection. The results show that the injected noise is successfully removed by the proposed approach.

6. CONCLUSION

We proposed a novel ABC method, SW-ABC, based on the Sliced-Wasserstein distance. We derived asymptotic guarantees for the convergence of the SW-ABC posterior to the true posterior under different regimes, and we evaluated our approach on a synthetical and an image denoising problem. Our results showed that SW-ABC provides an accurate approximation of the posterior, even with high-dimensional data spaces and a small number of samples. Future work includes extending SW-ABC to generalized SW distances [30].

5473

7. REFERENCES

[1] S. Tavaré, D. J. Balding, R. C. Griffiths, and P. Donnelly, "Inferring coalescence times from dna sequence data," *Genetics*, vol. 145, no. 2, pp. 505–518, 1997.

[2] M. A. Beaumont, W. Zhang, and D. J. Balding, "Approximate bayesian computation in population genetics," *Genetics*, vol. 162, no. 4, pp. 2025–2035, 2002.

[3] G. Peters and S. Sisson, "Bayesian inference, monte carlo sampling and operational risk," *Journal of Operational Risk*, vol. 1, 09 2006.

[4] M. Tanaka, A. Francis, F. Luciani, and S. Sisson, "Using approximate bayesian computation to estimate tuberculosis transmission parameters from genotype data," *Genetics*, vol. 173, pp. 1511–20, 08 2006.

[5] S. N. Wood, "Statistical inference for noisy nonlinear ecological dynamic systems," *Nature*, vol. 466, no. 7310, pp. 1102–1104, 8 2010.

[6] P. Fearnhead and D. Prangle, "Constructing summary statistics for approximate bayesian computation: semi-automatic approximate bayesian computation," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 74, no. 3, pp. 419–474, 2012.

[7] B. Jiang, T.-Y. Wu, and W.-H. Wong, "Approximate bayesian computation with kullback-leibler divergence as data discrepancy," in *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, Amos Storkey and Fernando Perez-Cruz, Eds., Playa Blanca, Lanzarote, Canary Islands, 09–11 Apr 2018, vol. 84 of *Proceedings of Machine Learning Research*, pp. 1711–1721, PMLR.

[8] M. Park, W. Jitkrittum, and D. Sejdinovic, "K2-abc: Approximate bayesian computation with kernel embeddings," in *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, Arthur Gretton and Christian C. Robert, Eds., Cadiz, Spain, 09–11 May 2016, vol. 51 of *Proceedings of Machine Learning Research*, pp. 398–407, PMLR.

[9] E. Bernton, P. E. Jacob, M. Gerber, and C. P. Robert, "Approximate bayesian computation with the wasserstein distance," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 81, no. 2, pp. 235–269, 2019.

[10] J. Weed and F. Bach, "Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance," *Bernoulli*, vol. 25, no. 4A, pp. 2620–2648, 11 2019.

[11] J. Rabin, G. Peyré, J. Delon, and M. Bernot, "Wasserstein barycenter and its application to texture mixing," in *Scale Space and Variational Methods in Computer Vision*, Berlin, Heidelberg, 2012, pp. 435–446, Springer Berlin Heidelberg.

[12] N. Bonneel, J. Rabin, G. Peyré, and H. Pfister, "Sliced and Radon Wasserstein barycenters of measures," *Journal of Mathematical Imaging and Vision*, vol. 51, no. 1, pp. 22–45, 2015.

[13] I. Deshpande, Z. Zhang, and A. G. Schwing, "Generative modeling using the sliced Wasserstein distance," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3483–3491.

[14] A. Liutkus, U. Şimşekli, S. Majewski, A. Durmus, and F.-R. Stoter, "Sliced-Wasserstein flows: Nonparametric generative modeling via optimal transport and diffusions," in *International Conference on Machine Learning*, 2019.

[15] S. Kolouri, P. E. Pope, C. E. Martin, and G. K. Rohde, "Sliced Wasserstein auto-encoders," in *International Conference on Learning Representations*, 2019.

[16] N. Bonnotte, *Unidimensional and Evolution Methods for Optimal Transportation*, Ph.D. thesis, Paris 11, 2013.

[17] I. Deshpande, Y.-T. Hu, R. Sun, A. Pyrros, N. Siddiqui, S. Koyejo, Z. Zhao, D. Forsyth, and A. Schwing, "Max-Sliced Wasserstein distance and its use for GANs," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[18] K. Nadjahi, A. Durmus, U. Şimşekli, and R. Badeau, "Asymptotic Guarantees for Learning Generative Models with the Sliced-Wasserstein Distance," in *Advances in Neural Information Processing Systems*, 2019.

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

[20] D. T. Frazier, G. M. Martin, C.P. Robert, and J. Rousseau, "Asymptotic properties of approximate Bayesian computation," *Biometrika*, vol. 105, no. 3, pp. 593–607, 06 2018.

[21] G. Peyré, M. Cuturi, et al., "Computational optimal transport," *Foundations and Trends® in Machine Learning*, vol. 11, no. 5-6, pp. 355–607, 2019.

[22] S. T. Rachev and L. Rüschendorf, *Mass transportation problems. Vol. I*, Probability and its Applications (New York). Springer-Verlag, New York, 1998, Theory.

[23] C. Villani, *Optimal Transport: Old and New*, Grundlehren der mathematischen Wissenschaften. Springer, 2009 edition, Sept. 2008.

[24] T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M. Stumpf, "Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems," *Journal of the Royal Society, Interface / the Royal Society*, vol. 6, pp. 187–202, 03 2009.

[25] E. Klinger, D. Rickert, and J. Hasenauer, "pyABC: distributed, likelihood-free inference," *Bioinformatics*, vol. 34, no. 20, pp. 3591–3593, 05 2018.

[26] K. Nadjahi, V. De Bortoli, A. Durmus, R. Badeau, and U. Şimşekli, "Sw-abc software implementation," https://github.com/kimiandj/slicedwass_abc, https://vdeborto.github.io/publication/sw_abc.

[27] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. IEEE, 2005, vol. 2, pp. 60–65.

[28] L. Fan, F. Zhang, H. Fan, and C. Zhang, "Brief review of image denoising techniques," *Visual Computing for Industry, Biomedicine, and Art*, vol. 2, no. 1, pp. 7, 2019.

[29] D. Martin, C. Fowlkes, D. Tal, J. Malik, et al., "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," Iccv Vancouver:, 2001.

[30] S. Kolouri, K. Nadjahi, U. Simsekli, R. Badeau, and G. K. Rohde, "Generalized Sliced Wasserstein Distances," in *Advances in Neural Information Processing Systems*, 2019.