Improving ABC via Large Deviations Theory
Migliorare ABC *tramite la teoria delle Grandi Deviazioni*

Cecilia Viscardi, Michele Boreale and Fabio Corradi

Abstract *Sample degeneracy* in Approximate Bayesian Computation (ABC) is caused by the difficulty of simulating pseudo-data matching the observed data. In order to mitigate the resulting waste of computational resources and/or bias in the posterior distribution approximation, we propose to weight each parameter proposal by treating the generation of matching pseudo-data, given a "poor" parameter proposal, as a *rare event* in the sense of Sanov's Theorem. We experimentally evaluate our methodology through a proof-of-concept implementation.
Abstract *Il problema della* degenerazione del campione *in metodi ABC deriva dalla difficoltà di generare dati simili a quelli osservati. Al fine di evitare i conseguenti sforzi compuatazionali e/o distorsioni nell'approssimazione della distribuzione a posteriori, proponiamo di pesare ciascun parametro trattando la simulazione di dati uguali a quelli osservati come un evento raro nel senso del Teorema di Sanov. Si riportano i risultati di una valutazione empirica della metodologia proposta.*

Key words: ABC, Large Deviations, Sanov's Theorem, Sample Degeneracy.

1 Approximate Bayesian Computation and sample degeneracy

Let $x^n \in \mathscr{X}^n$ be a vector of observed data, which will be assumed to be drawn from a probability distribution in the family $\mathscr{F} \stackrel{\triangle}{=} \{P(\cdot|\theta) : \theta \in \Theta\}$. Suppose that our aim is to provide information about the uncertainty on θ by deriving the posterior distribution $\pi(\theta|x^n) \propto \pi(\theta)P(x^n|\theta)$ via Bayes' Theorem. When the likelihood function is analytically and numerically intractable, Approximate Bayesian Computation (ABC) allows for simulated inference by providing a conversion of samples from the prior into samples from the posterior distribution. This relies on compar-

Authors'affiliation: University of Florence - Department of Statistics and Computer Science (DiSIA) - e-mail: cecilia.viscardi@unifi.it, michele.boreale@unifi.it, fabio.corradi@unifi.it.

isons between the observed data and the pseudo-data generated from a *simulator*[1]. Algorithm 1 displays the rejection sampling scheme (R-ABC), whose origins can be traced back to [7, 4].

Algorithm 1 R-ABC

for $s = 1,...,S$ **do**
  Draw $\theta^{(s)} \sim \Pi$
  Generate $y \sim P(\cdot|\theta^{(s)})$
  Accept $(\theta^{(s)}, s_y^{(s)})$ if $d(s_y^{(s)}, s_x) < \varepsilon$
end for

Algorithm 2 IS-ABC

for $s = 1,...,S$ **do**
  Draw $\theta^{(s)} \sim q$
  Generate $y \sim P(\cdot|\theta^{(s)})$
  Set the IS weight for $\theta^{(s)}$ to

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

end for

Samples resulting from Alg.1 are not from the exact posterior distribution since a twofold approximation is introduced by: summarizing data through a statistic $s(\cdot)$ — i.e. a function from the sample space $\mathscr{X}^n \subseteq \mathbb{R}^n$ to a lower-dimensional space $\mathscr{S}$ — and assessing similarity via a distance function $d(\cdot,\cdot)$ and a tolerance threshold $\varepsilon > 0$.

Abbreviating $s(x^n)$ and $s(y^n)$ respectively as s_x and s_y, the output of the Alg.1 is a sample of pairs $(\theta^{(s)}, s_y^{(s)})$ from the following *approximated* joint posterior distribution

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

where $\mathbb{1}\{d(s_y, s_x) \leq \varepsilon\}$, the indicator function assuming value 1 if $d(s_y, s_x) \leq \varepsilon$ and 0 otherwise, corresponds to the acceptance step. Marginalizing out s_y in (1), that is ignoring the simulated summary statistics, the output of the algorithm becomes a sample from the following approximated marginal posterior distribution

$$\tilde{\pi}(\theta|s_x) \propto \int_{\mathscr{S}} \pi(\theta)\, P(s_y|\theta)\, \mathbb{1}\{d(s_y, s_x) \leq \varepsilon\}\, ds_y = \pi(\theta) \cdot \Pr\big(d(s_Y, s_x) \leq \varepsilon|\theta\big). \tag{2}$$

The probability $\Pr\big(d(s_Y, s_x) \leq \varepsilon|\,\theta\big)$, where s_Y indicates $s(Y^n)$, is called the ABC *approximated likelihood.* As $\varepsilon \to 0$ the ABC likelihood converges to the true likelihood (see [3, Appendix A, p. 832]) and, whenever sufficient summary statistics for θ are chosen, $\tilde{\pi}(\cdot|s_x)$ converges to the true posterior $\pi(\cdot|x^n)$ (see [6, Ch. 1]). In practice the indicator function in (1) is often replaced by a kernel function $K_\varepsilon(\cdot)$ (e.g. triangular, Epanechnikov, Gaussian etc.) defined on the compact support $[0, \varepsilon]$ and providing a continuous decay from 1 to 0 (see e.g.[1]).

In the ABC literature a great variety of methods to sample from $\tilde{\pi}(\theta, s_y|s_x)$ have been proposed [2]. An example, is the importance sampling scheme IS-ABC reported as Alg. 2. Like the standard importance sampling, it suffers from *sample degeneracy*

[1] A simulator can be thought of as computer program taking as input a parameter value (or a vector thereof) $\theta^* \in \Theta$ and returning a sample from the distribution $P(\cdot|\theta^*)$.

[2] We refer the reader to [6, Ch 4] for an overview.

– i.e. only a small fraction of the proposed pairs has relatively high weights when the instrumental density $q(\cdot)$ is far from the target. Unlike the standard importance sampling, the IS-ABC implicitly involves a rejection when $\omega_s = 0$ — i.e. whenever a distance $d(s_y^{(s)}, s_x) > \varepsilon$ is observed [3]. Since they depend on the random variable s_Y through the distance $d(s_Y, s_x)$, when θ^* is such that $\Pr(s_Y = s_x|\theta^*)$ is close to zero, the importance weights will cause a huge number of rejections before a distance smaller than ε will be observed. This further aggravates the *sample degeneracy* issue. More sophisticated sampling schemes (e.g. MCMC-ABC, SMC-ABC, SIS-ABC, etc.) have been proposed to handle the issue of finding a good importance distribution, $q(\theta)$, but they completely ignore the effect of the kernel $K_\varepsilon(\cdot)$. In the next two sections we discuss how to define a kernel function $K_\varepsilon(\cdot)$ that improves the efficiency of ABC sampling schemes by avoiding rejections at all.

2 Large Deviations Theory in ABC

When a "poor" parameter proposal is given as an input to the generative model, simulating pseudo-data y^n such that $d(s_y, s_x) \leq \varepsilon$ can be treat as a *rare event*. This often leads to a shortage of accepted values mostly in regions of Θ with a low but positive (true) posterior density, in turn resulting in a bad approximation in the tails of the posterior distribution. A possible approach to mitigate those issues is to provide a finer estimate for the ABC likelihood allowing to avoid rejections at all. To this aim, we resort to *Large Deviations Theory* (LDT).

Let x^n be a sequence of n symbols drawn from $\mathscr{X}$ according to $P_\theta \stackrel{\triangle}{=} P(\cdot|\theta)$, say $x^n = (x_1, ..., x_n)$. The empirical distribution of x^n, written P_{x^n}, is the probability distribution on $\mathscr{X}$ defined by

$$P_{x^n}(r) \stackrel{\triangle}{=} \frac{|\{i : x_i = r\}|}{n} \quad \forall r \in \mathscr{X}. \tag{3}$$

Given a large n, observing a sequence whose empirical distribution is far from P_θ is a rare event, and its probability obeys to a fundamental result in LDT, Sanov's theorem (see [2, Th.11.4.1]).

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

$$\Pr\left(P_{X^n} \in E \,|\, \theta\right) \leq (n+1)^{|\mathscr{X}|} 2^{-nD(P^*||P_\theta)}. \tag{4}$$

where $D(\cdot||\cdot)$ is the Kullback-Leibler divergence and $P^* = \operatorname{argmin}_{P \in E} D(P||P_\theta)$ is the information projection of P_θ onto E. Furthermore, if E is the closure of its interior

[3] Note that Alg. 1 is a special case of the Alg.2 where the marginal importance distribution, $q(\theta)$, is the prior distribution and the resulting importance weights are $\omega_s \in \{0, 1\}$.

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

In order to show how to make use of the Sanov's result in the approximated likelihood computation, from now on we will assume: a) $\varepsilon>0$ as a threshold; b) empirical distributions as summary statistics; c) the Kullback-Leibler divergence as distance function; d) the empirical distribution P_{x^n} to be full support. The IS-ABC generates pairs $(P^{(s)}_{y^m},\theta^{(s)})$ with $s\in\{1,...,S\}$. Each $P^{(s)}_{y^m}$ is an empirical distribution resulting from a sequence of i.i.d. random variables, $Y^m=\{Y_j\}_{j=1}^m$, distributed according to $P\big(\cdot\,|\,\theta^{(s)}\big)$. We want to stress that the length of the simulated sequence, m, need not be equal to n, the length of the observed data sequence.

Under our assumptions, each $\theta^{(s)}$ is accepted or rejected depending on the divergence $D(P_{y^m}||P_{x^n})$. Thus, we can define the following acceptance region:

Definition 1 (Acceptance region). Let $\Delta^{|\mathscr{X}|-1}$ be the simplex of probability distributions over $\mathscr{X}$ and let P_{x^n} be the empirical distribution of the observed sequence x^n. The *acceptance region* $\mathscr{B}_\varepsilon(P_{x^n})$, shortly $\mathscr{B}_\varepsilon$, is defined for any $\varepsilon\geq 0$, as

$$\mathscr{B}_\varepsilon\stackrel{\triangle}{=}\big\{P\in\Delta^{|\mathscr{X}|-1}:D(P||P_{x^n})\leq\varepsilon\big\}.$$

Sanov's result, for m large enough, allows to approximate the probability of simulating pseudo-data whose summary statistic, P_{y^m}, is in the acceptance region even when a "poor" parameter in proposed:

$$\Pr\big(P_{Y^m}\in\mathscr{B}_\varepsilon\,|\,\theta^{(s)}\big)\approx 2^{-mD(\mathscr{B}_\varepsilon||P_{\theta^{(s)}})}. \tag{5}$$

Unfortunately, the computation of the probability in (5) is still not feasible when the model $\mathscr{F}=\{P_\theta:\theta\in\Theta\}$ is unknown, as we do not know how to compute $D(\mathscr{B}_\varepsilon||P_{\theta^{(s)}})$. However, one we can prove that

$$\lim_{m\to\infty}D(\mathscr{B}_\varepsilon||P_{Y^m})=D(\mathscr{B}_\varepsilon||P_\theta)\quad a.s. \tag{6}$$

According to (5) and (6) we propose the following kernel function:

$$K_{\varepsilon,m}\big(P_{y^m}\big)=\begin{cases}1 & \text{if } D(P_{y^m}||P_{x^n})\leq\varepsilon\\ 2^{-mD(\mathscr{B}_\varepsilon||P_{y^m})} & \text{if } D(P_{y^m}||P_{x^n})>\varepsilon\end{cases} \tag{7}$$

By resorting to this kernel the joint and the marginal posterior distributions are characterized by the following equations:

$$\tilde{\pi}(\theta,P_{y^m}|P_{x^n})\propto\pi(\theta)\,K_{\varepsilon,m}\big(P_{y^m}\big)\,P_\theta(P_{y^m}) \tag{8}$$

$$\tilde{\pi}(\theta|P_{x^n})\propto\pi(\theta)\sum_{P_{y^m}\in\mathscr{P}^m}K_{\varepsilon,m}\big(P_{y^m}\big)\,P_\theta(P_{y^m}) \tag{9}$$

where $\mathscr{P}^m\subset\Delta^{|\mathscr{X}|-1}$ denotes the set of the empirical distributions with denominator m. Accordingly, the ABC likelihood is defined as follows

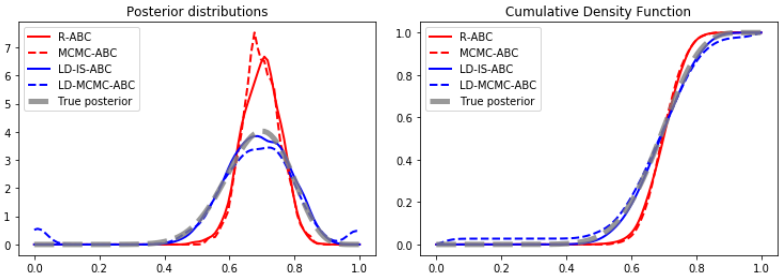

Fig. 1 Parameter posterior distributions (LHS) and posterior cumulative density functions (RHS) derived through IS and MCMC schemes with the uniform kernel and with the proposed kernel.

$$\tilde{\mathscr{L}}_{\varepsilon,m}(\theta;P_{x^n}) \stackrel{\triangle}{=} \sum_{P_{y^m}\in\mathscr{P}^m} K_{\varepsilon,m}\left(P_{y^m}\right) P_\theta\left(P_{y^m}\right). \tag{10}$$

Note that, now Alg.2 gives a positive weight to each $\theta^{(s)}$. More precisely, the weight equals 0 only when $D(\mathscr{B}_\varepsilon||P_{y^m}) = \infty$. In the next section we empirically demonstrate the improvements achieved by resorting to the proposed kernel.

3 A toy example

Let x^{20} be a sample from i.i.d. Bernoulli random variables with parameter θ. Suppose that x^{20} has empirical distribution $P_{x^n} = [0.3, 0.7]$. Assuming an uniform prior distribution, the posterior distribution, $\pi(\theta|x^{20})$, is a Beta distribution with parameters $\alpha = 15$ and $\beta = 7$.

We ran $S = 10000$ iterations of IS-ABC both with the uniform kernel and the proposed kernel. Note that in the first case the algorithm corresponds to a R-ABC. We also implemented the MCMC-ABC sampling scheme (see [6, Ch. 4]). For the sake of simplicity we adopt the abbreviation LD, standing for Large Deviations, to indicate that the employed kernel function is (7).

Fig.1 shows the posterior distributions and cumulative density functions (CDF) approximated by each algorithm. As it is apparent, the LD algorithms (blue lines) approximate better the true posterior (dashed grey line). Looking at the CDF's, we can see that using the uniform kernel (red lines) results in a worse approximation in the tails.

We evaluate the posterior mean point estimates and the posterior density estimates through the Squared Error and the Integrated Squared Error respectively. We also consider the Effective Sample Size as a measure of the degree of sample degeneracy. From Tab.1, we can see that, despite the quality of the point estimations is almost the same, the proposed kernel function leads to clear improvements in terms of density estimations and *ESS*, both for the IS-ABC and for the MCMC-ABC.

Table 1 Squared Errors, Integrated Squared Errors and Effective Sample Sizes with $\varepsilon = 0.01$, $m = 100$.

Algorithm	*SE*	*ISE*	*ESS*
R-ABC	0.0002	0.6428	1279
LD-IS-ABC	0.0003	0.0096	3060
MCMC-ABC	0.0002	0.8597	655
LD-MCMC-ABC	0.0002	0.041	1929

4 Conclusions

We have put forward an approach to address sample degeneracy in ABC. Our proposal consists in the definition of a convenient kernel function which, via Large Deviations Theory, takes into account the probability of rare events. Being defined on a non-compact support, the proposed kernel allows to avoid rejections, thus mitigating the effects of the sample degeneracy. We have also evaluated our methodology on a simple example, showing that it provides a better approximation of the posterior density and increases the Effective Sample Size.

References

1. Beaumont, M. A., Zhang, W., Balding, D. J.: Approximate Bayesian computation in population genetics. Genetics,**96**(4), 2025–2035 (2002).
2. Cover, T.M and Thomas, J. A. Element of information theory, John Wiley & Sons (2006).
3. Prangle, D., Everitt, R. G., Kypraios, T.: A rare event approach to high-dimensional approximate Bayesian computation. Statistics and Computing, **28**(4), 819–834 (2018).
4. Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., Feldman, M. W.: Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Molecular biology and evolution, **16**(12), 1791–1798 (1999).
5. Rubin, D.B. : Bayesianly justifiable and relevant frequency calculations for the applies statistician. The Annals of Statistics, 1151–1172 (1984).
6. Sisson, S. A., Fan, Y., & Beaumont, M. Handbook of approximate Bayesian computation. Chapman and Hall& CRC (2018).
7. Tavaré, S., Balding, D. J., Griffiths, R. C. & Donnelly, P. : Inferring coalescence times from DNA sequence data. Genetics, **145**(2), 505–518 (1997).