Sirio Legramanti

University of Bergamo

Concentration and robustness of discrepancy–based ABC via Rademacher complexity

2022 ISBA World Meeting

June 26th - July 1st, Montreal (Canada)

Coauthors

Daniele Durante
Bocconi, Milan

Pierre Alquier
RIKEN, Tokyo

Context and goal

ABC allows Bayesian inference on a **parameter** θ even when the likelihood is intractable, as long as synthetic data can be sampled from the **model** μ_θ

ABC outputs a sample from an **approximate posterior**, made of those θs that produced synthetic data that are "close" to the observed data

"Closeness" was traditionally measured through **summary statistics** but, unless such summaries are sufficient, this yields an **information loss**

This has motivated research on

- selecting summaries (e.g. semi–automatically)
- **summary–free ABC** (e.g. discrepancy among empirical distributions)

We study **concentration** and **robustness** of discrepancy–based ABC via the concept of **Rademacher complexity**

Setup and notation

Observed data: $y_{1:n} = (y_1, \ldots, y_n) \overset{\text{i.i.d.}}{\sim} \mu^*$

Statistical model: $\{\mu_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$

Prior distribution: $\pi(\theta)$

Rejection ABC

Iteratively:

- sample θ from π,
- sample **synthetic data** $z_{1:m} = (z_1, \ldots, z_m) \overset{\text{i.i.d.}}{\sim} \mu_\theta$,
- if $\Delta(z_{1:m}, y_{1:n}) \leq \varepsilon_n$, retain θ

Output: a sample $(\theta_1, \ldots, \theta_T)$ from the ABC posterior $\pi_n^{(\varepsilon_n)}(\theta)$

Setup and notation

Observed data: $y_{1:n} = (y_1, \ldots, y_n) \overset{\text{i.i.d.}}{\sim} \mu^*$

Statistical model: $\{\mu_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$

Prior distribution: $\pi(\theta)$

Rejection ABC

Iteratively:

- sample θ from π,
- sample **synthetic data** $z_{1:m} = (z_1, \ldots, z_m) \overset{\text{i.i.d.}}{\sim} \mu_\theta$,
- if $\Delta(z_{1:m}, y_{1:n}) \leq \varepsilon_n$, retain θ

Output: a sample $(\theta_1, \ldots, \theta_T)$ from the ABC posterior $\pi_n^{(\varepsilon_n)}(\theta)$

Following common practice in theoretical studies of ABC, we set $m = n$

Setup and notation

Observed data: $y_{1:n} = (y_1, \ldots, y_n) \stackrel{\text{i.i.d.}}{\sim} \mu^*$

Statistical model: $\{\mu_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$

Prior distribution: $\pi(\theta)$

Rejection ABC

Iteratively:

- sample θ from π,
- sample **synthetic data** $z_{1:n} = (z_1, \ldots, z_n) \stackrel{\text{i.i.d.}}{\sim} \mu_\theta$,
- if $\Delta(z_{1:n}, y_{1:n}) \leq \varepsilon_n$, retain θ

Output: a sample $(\theta_1, \ldots, \theta_T)$ from the ABC posterior $\pi_n^{(\varepsilon_n)}(\theta)$

Setup and notation

Observed data: $y_{1:n} = (y_1, \ldots, y_n) \overset{\text{i.i.d.}}{\sim} \mu^*$

Statistical model: $\{\mu_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$

Prior distribution: $\pi(\theta)$

Rejection ABC

Iteratively:

- sample θ from π,
- sample **synthetic data** $z_{1:n} = (z_1, \ldots, z_n) \overset{\text{i.i.d.}}{\sim} \mu_\theta$,
- if $\Delta(z_{1:n}, y_{1:n}) \leq \varepsilon_n$, retain θ

Output: a sample $(\theta_1, \ldots, \theta_T)$ from the ABC posterior $\pi_n^{(\varepsilon_n)}(\theta)$

$\Delta(z_{1:n}, y_{1:n})$ was traditionally induced by some distance among summaries

Setup and notation

Observed data: $y_{1:n} = (y_1, \ldots, y_n) \overset{\text{i.i.d.}}{\sim} \mu^*$

Statistical model: $\{\mu_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$

Prior distribution: $\pi(\theta)$

Rejection ABC

Iteratively:

- sample θ from π,
- sample **synthetic data** $z_{1:n} = (z_1, \ldots, z_n) \overset{\text{i.i.d.}}{\sim} \mu_\theta$,
- if $\mathcal{D}(\hat{\mu}_{z_{1:n}}, \hat{\mu}_{y_{1:n}}) \leq \varepsilon_n$, retain θ

Output: a sample $(\theta_1, \ldots, \theta_T)$ from the ABC posterior $\pi_n^{(\varepsilon_n)}(\theta)$

where $\hat{\mu}_{x_{1:n}} = n^{-1}\sum_{i=1}^n \delta_{x_i}$ is the empirical distribution of a sample $x_{1:n}$

Discrepancy–based ABC

Popular choices for $\mathcal{D}$ are

- maximum mean discrepancy (MMD) (Park et al., 2016),
- Kullback–Leibler (KL) divergence (Jiang et al., 2018),
- Wasserstein distance (Bernton et al., 2019),
- energy statistic (Nguyen et al., 2020),
- Hellinger and Cramer–von Mises distances (Frazier, 2020),
- γ–divergence (Fujisawa et al., 2021)

MMD, Wasserstein distance and energy statistic belong to the class of **integral probability semimetrics (IPS)**

Integral probability semimetrics (IPS)

Definition (Müller, 1997)

Let $(\mathcal{Y}, \mathscr{A})$ be a measure space, $g : \mathcal{Y} \to [1, \infty)$ a measurable function, and $\mathfrak{B}_g$ the set of measurable functions $f : \mathcal{Y} \to \mathbb{R}$ such that $||f||_g := \sup_{y \in \mathcal{Y}} |f(y)|/g(y) < \infty$. Then, for a chosen $\mathfrak{F} \subseteq \mathfrak{B}_g$, an integral probability semimetric $\mathcal{D}_{\mathfrak{F}}$ among $\mu_1, \mu_2 \in \mathcal{P}(\mathcal{Y})$ is defined as

$$\mathcal{D}_{\mathfrak{F}}(\mu_1, \mu_2) := \sup_{f \in \mathfrak{F}} \left| \int f d\mu_1 - \int f d\mu_2 \right|.$$

For different choices of $\mathfrak{F}$, we get

- total variation (TV) distance
- Kolmogorov–Smirnov (KS) distance
- Wasserstein distance
- maximum mean discrepancy (MMD) & energy statistic
- sup-distance among K linear summaries

Rademacher complexity

The properties of an IPS, $\mathcal{D}_{\mathfrak{F}}$, and consequently of IPS–ABC, crucially depend on the **richness** of family $\mathfrak{F}$, which can be measured by...

Definition (Rademacher complexity)

Given an i.i.d. sample $x_{1:n} = (x_1, \dots, x_n) \in \mathcal{Y}^n$ from $\mu \in \mathcal{P}(\mathcal{Y})$,
and a class $\mathfrak{F}$ of real–valued measurable functions,
the Rademacher complexity of $\mathfrak{F}$ with respect to μ is defined as

$$\mathfrak{R}_{\mu,n}(\mathfrak{F}) = \mathbb{E}_{x_{1:n},\epsilon_{1:n}} \left[\sup_{f \in \mathfrak{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(x_i) \right| \right]$$

where $\epsilon_{1:n}$ are i.i.d. Rademacher r.v.'s, i.e. $\mathbb{P}(\epsilon_i = 1) = \mathbb{P}(\epsilon_i = -1) = 1/2$.

We will be mostly interested in its *supremum* over distributions

$$\mathfrak{R}_n(\mathfrak{F}) := \sup_{\mu \in \mathcal{P}(\mathcal{Y})} \mathfrak{R}_{\mu,n}(\mathfrak{F})$$

Assumptions

- We consider the **scenario**

$$n \to \infty \quad \text{and} \quad \varepsilon_n \to \varepsilon^* = \inf_{\theta \in \Theta} \mathcal{D}_{\mathfrak{F}}(\mu_\theta, \mu^*)$$

or equivalently $\quad \varepsilon_n = \varepsilon^* + \bar{\varepsilon}_n \quad$ with $\quad \bar{\varepsilon}_n \to 0$

- We **assume**

(C1) the observed data $y_{1:n}$ are i.i.d. from μ^*

(C2) there exist some positive L and c_π such that, for $\bar{\varepsilon}$ small enough,

$$\pi\left(\{\theta \in \Theta : \mathcal{D}_{\mathfrak{F}}(\mu_\theta, \mu^*) \leq \varepsilon^* + \bar{\varepsilon}\}\right) \geq c_\pi \bar{\varepsilon}^L$$

(C3) $\mathfrak{F}$ is a family of b–uniformly bounded functions, i.e.

$$||f||_\infty \leq b \qquad \text{for all} \quad f \in \mathfrak{F}$$

(C4) $\mathfrak{R}_n(\mathfrak{F}) \to 0$ as $n \to \infty$

Which/when IPS satisfy (C3) and (C4)

- **TV** satisfies (C3) by definition, since $\mathfrak{F}_{TV} = \{f : ||f||_\infty \leq 1\}$, but generally not (C4), e.g. when $\mathcal{Y} = \mathbb{R}$ and μ is continuous
- **KS** satisfies (C3) by definition, since $\mathfrak{F}_{KS} = \{\mathbb{1}_{(-\infty,a]}\}_{a\in\mathbb{R}}$, and also (C4) since $\mathfrak{R}_{\mu,n}(\mathfrak{F}) \leq 2[\log(n+1)/n]^{1/2}$ (Wainwright, 2019)
- **Wasserstein** dist. is induced by a $\mathfrak{F}$ that is not b–uniformly bounded, but its value does not change if $\mathfrak{F}$ is constrained to be that way in order to satisfy (C3). (C4) is more problematic: no general upper bound for the Rademacher compl. of Wass. dist. is available, but it is when $\mathcal{Y} \subset \mathbb{R}^d$ and is **bounded** (Sriperumbudur et al., 2010, 2012)
- **MMD with bounded kernels** (e.g. Gaussian, Laplace) satisfy both (C3) and (C4), since $\mathfrak{R}_{\mu,n}(\mathfrak{F}) \leq [\mathbb{E}_{x\sim\mu} k(x,x)/n]^{1/2}$, and $|f(x)| \leq [k(x,x)]^{1/2}||f||_{\mathcal{H}}$. We also cover unbounded kernels, but under alternative assumptions

Concentration

Theorem 1 (Concentration)

Let $\mathcal{D}_{\mathfrak{F}}$ be a semimetric, $\bar{\varepsilon}_n \to 0$ as $n \to \infty$, $n\bar{\varepsilon}_n^2 \to \infty$ and $\bar{\varepsilon}_n / \mathfrak{R}_n(\mathfrak{F}) \to \infty$. If (C1)–(C4) then, for any sequence $M_n > 1$, the IPS–ABC posterior with threshold $\varepsilon_n = \varepsilon^* + \bar{\varepsilon}_n$ satisfies

$$\pi_n^{(\varepsilon^* + \bar{\varepsilon}_n)}\left(\left\{\theta : \mathcal{D}_{\mathfrak{F}}(\mu_\theta, \mu^*) > \varepsilon^* + \frac{4}{3}\bar{\varepsilon}_n + 2\mathfrak{R}_n(\mathfrak{F}) + \left[\frac{2b^2}{n}\log\left(\frac{M_n}{\bar{\varepsilon}_n^L}\right)\right]^{1/2}\right\}\right) \leq \frac{2 \cdot 3^L}{c_\pi M_n}$$

with $\mathbb{P}_{y_{1:n}}$–probability going to 1 as $n \to \infty$.

For MMD with unbounded kernel, which does not satisfy (C3), we provide a similar bound under alternative assumptions, satisfied e.g. by the polynomial kernel in $\mathbb{R}^d$, i.e. $k(x, x') = (1 + a\langle x, x' \rangle)^q$

Robustness

Assume that the obs. data are i.i.d. from a **Huber contamination model**

$$\mu^* = (1 - \alpha_n)\mu_{\theta^*} + \alpha_n \mu_C \tag{1}$$

We are now interested in concentration around μ_{θ^*} rather than μ^*

Theorem 2 (Robustness)

Consider the Huber contamination model in (1).Then, under the same assumptions of Theorem 1 and for the same choice of $\bar{\varepsilon}_n$, we have that, for any $M_n > 1$, any $\alpha_n \in [0, 1)$ and any μ_C:

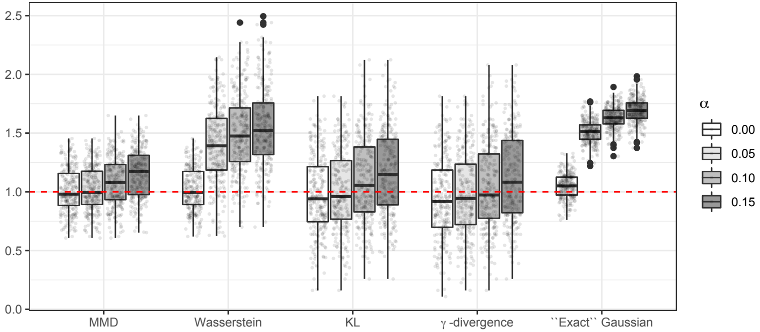
$$\pi_n^{(\varepsilon^* + \bar{\varepsilon}_n)}\left(\left\{\theta : \mathcal{D}_{\mathfrak{F}}(\mu_\theta, \mu_{\theta^*}) > 4b\alpha_n + \frac{4}{3}\bar{\varepsilon}_n + 2\mathfrak{R}_n(\mathfrak{F}) + \left[\frac{2b^2}{n}\log\left(\frac{M_n}{\bar{\varepsilon}_n^L}\right)\right]^{1/2}\right\}\right) \leq \frac{2 \cdot 3^L}{c_\pi M_n}$$

with $\mathbb{P}_{y_{1:n}}$–probability going to 1 as $n \to \infty$.

Simulation study

- Observed data $y_{1:100}$ from a $N(1,1)$ with different levels $\alpha \in \{0.00, 0.05, 0.10, 0.15\}$ of **Cauchy–tail contamination**
- **Model** for y: $\mu_\theta = N(\theta, 1)$
- **Prior** for θ: $N(0,1)$
- Discrepancy–based rejection ABC with $m = n$, as implemented in the Python library **ABCpy** (Dutta et al., 2021)
- Instead of specifying the rejection thresholds for each discrepancy, we set a common **computational budget** of 2500 simulations and keep the 10% of θs yielding synthetic data closest to the obs. data

Simulation results



- The closed–form posterior concentrates at $\theta^* = 1$, but gets shifted away by even a small fraction α of contaminated data
- Wasserstein–ABC posterior partially replicates this lack of robustness
- KL and γ–divergence exhibit good robustness, but also inflated variance
- MMD with Gaussian kernel shows the best concentration–robustness tradeoff

Conclusions

- We proved **concentration and robustness** results for discrepancy–based ABC under a broad class of semimetrics, **IPS**, which include MMD and Wasserstein distance among others
- We built **a bridge** between such properties and the Rademacher complexity associated with the chosen discrepancy
- Our framework is ready to leverage **new bounds** on Rad. complexity e.g. for non–i.i.d. processes (Mohri and Rostamizadeh, 2008) or for Wasserstein distance in unbounded spaces (not available yet)
- KL and Hellinger distance are not IPS but rather **f–divergences**: they may be tackled through unified treatments of these two classes (Agrawal and Horel, 2021; Birrell et al., 2022)
- Our results could be extended to generalized likelihood–free Bayesian inference via discrepancy–based **pseudo–posteriors**

Thanks for your attention!

Feel free to contact me at

sirio.legramanti@unibg.it

References I

R. Agrawal and T. Horel. Optimal bounds between f-divergences and integral probability metrics. *Journal of Machine Learning Research*, 22:1–59, 2021.

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)*, 81(2):235–269, 2019.

J. Birrell, P. Dupuis, M. A. Katsoulakis, Y. Pantazis, and L. Rey-Bellet. (f, γ)-divergences: Interpolating between f-divergences and integral probability metrics. *Journal of Machine Learning Research*, 23(39):1–70, 2022.

R. Dutta, M. Schoengens, L. Pacchiardi, A. Ummadisingu, N. Widmer, P. Künzli, J.-P. Onnela, and A. Mira. ABCpy: A high-performance computing perspective to approximate Bayesian computation. *Journal of Statistical Software*, 100(7):1–38, 2021.

D. T. Frazier. Robust and efficient approximate Bayesian computation: A minimum distance approach. *arXiv:2006.14126*, 2020.

M. Fujisawa, T. Teshima, I. Sato, and M. Sugiyama. γ-ABC: Outlier-robust approximate Bayesian computation based on a robust divergence estimator. In *International Conference on Artificial Intelligence and Statistics*, pages 1783–1791. PMLR, 2021.

B. Jiang, T.-Y. Wu, and W. H. Wong. Approximate Bayesian computation with Kullback-Leibler divergence as data discrepancy. In *International Conference on Artificial Intelligence and Statistics*, pages 1711–1721. PMLR, 2018.

References II

M. Mohri and A. Rostamizadeh. Rademacher complexity bounds for non-i.i.d. processes. In *Advances in Neural Information Processing Systems*, volume 21, pages 1097–1104. Curran Associates, Inc., 2008.

A. Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

H. D. Nguyen, J. Arbel, H. Lü, and F. Forbes. Approximate Bayesian computation via the energy statistic. *IEEE Access*, 8:131683–131698, 2020.

M. Park, W. Jitkrittum, and D. Sejdinovic. K2-ABC: Approximate Bayesian computation with kernel embeddings. In *International Conference on Artificial Intelligence and Statistics*, pages 398–407. PMLR, 2016.

B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. G. Lanckriet. Non-parametric estimation of integral probability metrics. In *2010 IEEE International Symposium on Information Theory*, pages 1428–1432, 2010.

B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6: 1550–1599, 2012.

M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48. Cambridge University Press, 2019.