Sirio Legramanti

University of Bergamo

Concentration of discrepancy-based approximate Bayesian computation via Rademacher complexity

SISBayes

Padua, Italy – Sept. 5, 2025

Today's talk will be mainly based on

Legramanti, Durante, Alquier (2025),
Concentration of discrepancy-based approximate Bayesian computation via Rademacher complexity.
Annals of Statistics, 53(1) 37-60
https://doi.org/10.1214/24-AOS2453

Also available at:
https://arxiv.org/abs/2206.06991

Joint work with

Daniele Durante
Bocconi University, Milan

Pierre Alquier
ESSEC Asia–Pacific, Singapore

Bayesian (parametric) inference

Given

- an **observed dataset** $y_{1:n} = (y_1, \ldots, y_n) \overset{\text{i.i.d.}}{\sim} \mu^*$
- a **parametric model** $\{\mu_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$
- a **prior distribution** $\pi(\theta)$

we aim at the posterior distribution

$$\pi(\theta \mid y_{1:n}) \propto \pi(\theta)\ \mu_\theta^n(y_{1:n})$$

either in closed form, or by sampling

$$(\theta_1, \ldots, \theta_T) \overset{\text{i.i.d.}}{\sim} \pi(\theta \mid y_{1:n})$$

Approximate Bayesian Computation (ABC)

When the likelihood $\mu_\theta^n(y_{1:n})$ is intractable, Bayesian inference is still possible, as long as you can sample **synthetic data** from the model:

$$z_{1:m} = (z_1, \ldots, z_m) \overset{\text{i.i.d.}}{\sim} \mu_{\tilde{\theta}}$$

Rejection ABC

Iteratively:

- sample $\tilde{\theta}$ from the prior π
- sample **synthetic data** $z_{1:m} \overset{\text{i.i.d.}}{\sim} \mu_{\tilde{\theta}}$
- if $\Delta(z_{1:m}, y_{1:n}) \leq \varepsilon_n$, retain $\tilde{\theta}$ for your (approximate) posterior sample.

Note: as customary in theoretical studies of ABC, we set $m = n$.

ABC posterior

Rather than returning a sample from the **exact posterior**

$$\pi(\theta \mid y_{1:n}) \propto \pi(\theta)\ \mu_\theta^n(y_{1:n})$$

rejection ABC returns a sample from the **ABC posterior**

$$\pi_n^{(\varepsilon_n)}(\theta) \propto \pi(\theta) \int_{\mathcal{Y}^n} \mathbb{1}\{\Delta(z_{1:n}, y_{1:n}) \leq \varepsilon_n\}\ \mu_\theta^n(dz_{1:n})$$

whose properties clearly depend on the choice of the discrepancy $\Delta(\cdot, \cdot)$.

The choice of the discrepancy

This discrepancy was traditionally based on **summary statistics**

$$\Delta(z_{1:n}, y_{1:n}) = d(s(z_{1:n}), s(y_{1:n}))$$

but, unless such summaries are sufficient, this yields **information loss**.

This has motivated research on

- **selecting summaries**
  e.g. semi–automatically (Fearnhead and Prangle, 2012);
- **summary–free ABC**
  e.g. based on some discrepancy $\mathcal{D}$ among empirical distributions

$$\Delta(z_{1:n}, y_{1:n}) = \mathcal{D}(\hat{\mu}_{z_{1:n}}, \hat{\mu}_{y_{1:n}}).$$

Summary-free ABC

Popular choices for $\mathcal{D}$ in summary-free ABC are:

- **maximum mean discrepancy (MMD)**, i.e. "distance" in the RKHS, and the related energy distance (Park et al., 2016; Nguyen et al., 2020)
- **Kullback–Leibler (KL)** divergence (Jiang et al., 2018)
- **Wasserstein** distance (Bernton et al., 2019)
- **Hellinger and Cramer–von Mises** distances (Frazier, 2020)
- γ–**divergence** (Fujisawa et al., 2021)

MMD and Wasserstein-1 both belong to the class of
integral probability semimetrics (IPS)

Integral probability semimetrics (IPS)

Definition (IPS; Müller, 1997)

Let $\mathfrak{F}$ be a class of measurable functions $f : \mathcal{Y} \to \mathbb{R}$. Then the **integral probability semimetric** $\mathcal{D}_{\mathfrak{F}}$ among μ_1 and μ_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

- Wasserstein-1 distance
- maximum mean discrepancy (MMD)
- sup-distance among K summaries (e.g., moments)
- total variation (TV) distance
- Kolmogorov–Smirnov (KS) distance

What can go wrong?

If $\mathcal{D}_{\mathfrak{F}} = TV$ and both μ^* and μ_θ are continuous, then

$$\mathcal{D}_{\mathfrak{F}}(\hat{\mu}_{y_{1:n}}, \hat{\mu}_{z_{1:n}}) = 1, \qquad \textbf{almost surely.}$$

This implies that

- if $\varepsilon < 1$, you **never** accept any θ from the prior
- if $\varepsilon \geq 1$, you **always** accept any θ from the prior

$\rightarrow$ the ABC posterior is either undefined or equal to the prior.

Research question

Which discrepancies $\mathcal{D}_{\mathfrak{F}}$ (i.e., families $\mathfrak{F}$) work well for ABC?

Rademacher complexity

The key element turns out to be the **richness** of $\mathfrak{F}$, measured via

Definition (Rademacher complexity)

Given $x_{1:n} = (x_1, \dots, x_n) \overset{\text{i.i.d.}}{\sim} \mu \in \mathcal{P}(\mathcal{Y})$ and a class $\mathfrak{F}$ of measurable functions $f : \mathcal{Y} \to \mathbb{R}$, 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 also define $\qquad \mathfrak{R}_n(\mathfrak{F}) := \sup_{\mu \in \mathcal{P}(\mathcal{Y})} \mathfrak{R}_{\mu,n}(\mathfrak{F})$.

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

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

[an additional setting with fixed ε in the paper]

Assumptions

(C1) the observed data $y_{1:n}$ are i.i.d. from μ^*; [relaxed in the Suppl.]

(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) $||f||_\infty \leq b, \quad \forall f \in \mathfrak{F}$;

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

A key lemma

Since $\mathcal{D}_{\mathfrak{F}}$ is a semimetric,

$$\mathcal{D}_{\mathfrak{F}}(\mu_\theta, \mu^*) \leq \mathcal{D}_{\mathfrak{F}}(\hat{\mu}_{z_{1:n}}, \mu_\theta) + \mathcal{D}_{\mathfrak{F}}(\hat{\mu}_{z_{1:n}}, \hat{\mu}_{y_{1:n}}) + \mathcal{D}_{\mathfrak{F}}(\hat{\mu}_{y_{1:n}}, \mu^*).$$

Lemma (Theorem 4.10 and Proposition 4.12 in Wainwright (2019))

Let $x_{1:n} \overset{i.i.d.}{\sim} \mu$. Then, if $\mathfrak{F}$ satisfies (C3), for any $n \geq 1$ and any $\delta \geq 0$,

$$\mathbb{P}_{x_{1:n}}\left[\mathcal{D}_{\mathfrak{F}}(\hat{\mu}_{x_{1:n}}, \mu) \leq 2\mathfrak{R}_{\mu,n}(\mathfrak{F}) + \delta\right] \geq 1 - e^{-n\delta^2/2b^2},$$

$$\mathbb{P}_{x_{1:n}}\left[\mathcal{D}_{\mathfrak{F}}(\hat{\mu}_{x_{1:n}}, \mu) \geq \mathfrak{R}_{\mu,n}(\mathfrak{F})/2 - \sup_{f\in\mathfrak{F}}|\mathbb{E}(f)|/2n^{1/2} - \delta\right] \geq 1 - e^{-n\delta^2/2b^2}.$$

Without (C4), $\mathcal{D}_{\mathfrak{F}}(\hat{\mu}_{z_{1:n}}, \mu_\theta)$ and $\mathcal{D}_{\mathfrak{F}}(\hat{\mu}_{y_{1:n}}, \mu^*)$ remain large with pr. > 0 and a small $\mathcal{D}_{\mathfrak{F}}(\hat{\mu}_{z_{1:n}}, \hat{\mu}_{y_{1:n}})$ does not guarantee a small $\mathcal{D}_{\mathfrak{F}}(\mu_\theta, \mu^*)$.

Note: $z_{1:n}$ are i.i.d. by construction, and $y_{1:n}$ are i.i.d. by (C1).

Which IPS satisfy (C3) and (C4)?

- When $\mathcal{Y} \subset \mathbb{R}^d$ is bounded, **Wasserstein-1 distance** satisfies (C3)–(C4) with no contraints on μ. When $\mathcal{Y}$ is unbounded, restrictions on μ^* and μ_θ (or a variable transformation) are required.

- **MMD with bounded kernels** (e.g. Gaussian, Laplace) satisfies (C3)–(C4) with no constraints on $\mathcal{Y}$, μ^* and μ_θ.

- **MMD with unbounded kernels** requires constraints on μ^* and μ_θ.

- **Summary-based distances** can be seen as a special case of MMD with either bounded or unbounded kernels.

- **KS** satisfies (C3)–(C4).

- **TV** satisfies (C3) but generally not (C4).

Main result

Theorem 1 (Concentration)

Let $\mathcal{D}_{\mathfrak{F}}$ be an IPS, $\bar{\varepsilon}_n \to 0$ 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), the ABC posterior with threshold $\varepsilon_n = \varepsilon^* + \bar{\varepsilon}_n$ satisfies

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

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

Hence, the ABC posterior asymptotically concentrates around those

$$\{\theta \in \Theta : \mathcal{D}_{\mathfrak{F}}(\mu_\theta, \mu^*) \leq \varepsilon^*\}$$

Note: if the model is well-specified, then $\varepsilon^* = \inf_{\theta \in \Theta} \mathcal{D}_{\mathfrak{F}}(\mu_\theta, \mu^*) = 0$.

Simulation study

Huber-contaminated data

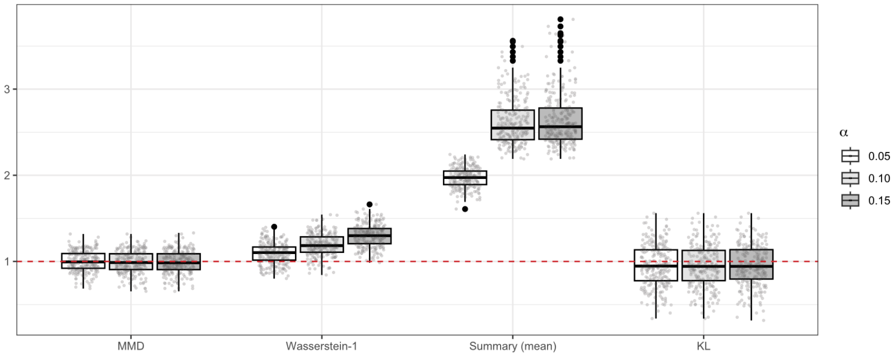
$$y_{1:100} \overset{\text{i.i.d.}}{\sim} \mu^* = (1-\alpha)\mu_{\theta_0} + \alpha\mu_C, \qquad \alpha \in \{0.05, 0.10, 0.15\}$$

- $\mu_{\theta_0} = t(\theta_0(1,1)^T, \Sigma_0, \nu_0 = 3)$, with $\theta_0 = 1$
- $\mu_C = t(\theta_C(1,1)^T, \Sigma_0, \nu_0 = 3)$, with $\theta_c = 20$
- $\mathcal{Y} = \mathbb{R}^2$, hence unbounded

Gaussian model	$\mu_\theta = N_2(\theta(1,1)^T, \Sigma_0)$ (misspecified even for $\alpha = 0$)
Gaussian prior	$\theta \sim N(0,1)$

Our theory ensures concentration also around the uncontaminated μ_{θ_0}

ABC posterior for a single simulated dataset



- the mean summary statistic yields strong bias even with $\alpha = 0.05$;
- Wasserstein-1 yields smaller but increasing bias as α grows;
- KL (not an IPS) stays almost unbiased but at lower concentration;
- MMD with Gaussian (bounded) kernel is robust even as α grows.

MSE averaged over 50 simulated datasets

$$MSE = E_{\pi_n^{(\varepsilon_n)}}(\theta - \theta_0)^2$$

	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.15$
(IPS) **MMD with Gaussian kernel**	**0.024**	**0.027**	**0.031**
(IPS) Wasserstein-1	0.027	0.067	0.122
(IPS) Summary (mean)	0.841	2.648	2.835
(non–IPS) KL	0.073	0.076	0.077

- at $\alpha = 0.05$, both MMD and Wasserstein-1 perform well;
- as α grows, Wasserstein-1 deteriorates while MMD stays robust;
- ABC using the mean as a summary statistic suffers significantly from location contamination, even at $\alpha = 0.05$;
- KL performs worse than MMD but is consistent as α grows.

Conclusions and future directions

We built **a bridge between ABC and Rademacher complexity** for the broad IPS class, which include MMD and Wasserstein-1.

Possible extensions include:

- **beyond IPS:** e.g., f–divergences (like KL and Hellinger distance) via unified treatment with IPS (Agrawal and Horel, 2021; Birrell et al., 2022);
- **beyond i.i.d. and β-mixing data**;
- **beyond ABC**: e.g., generalized likelihood–free Bayesian inference and discrepancy–based pseudo–posteriors (Miller and Dunson, 2019; Matsubara et al., 2022; Dellaporta et al., 2022)

Working paper (with Marta Catalano, Luiss)

Is Wasserstein doomed for ABC? Spoiler: not quite (stay tuned!)

Thanks for your attention

For further questions, 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.

C. Dellaporta, J. Knoblauch, T. Damoulas, and F.-X. Briol. Robust Bayesian inference for simulator-based models via the MMD posterior bootstrap. In *International Conference on Artificial Intelligence and Statistics*, pages 943–970. PMLR, 2022.

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*, 74(3):419–474, 2012.

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.

References II

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.

T. Matsubara, J. Knoblauch, F.-X. Briol, and C. J. Oates. Robust generalised Bayesian inference for intractable likelihoods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(3):997–1022, 2022.

J. W. Miller and D. B. Dunson. Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, 2019.

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.

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