Batch simulations and uncertainty quantification in Gaussian process surrogate-based approximate Bayesian computation

Marko Järvenpää MARKO.J.JARVENPAA@AALTO.FI
Aki Vehtari AKI.VEHTARI@AALTO.FI
Pekka Marttinen PEKKA.MARTTINEN@AALTO.FI
Helsinki Institute for Information Technology HIIT, Department of Computer Science, Aalto University, Espoo, Finland

Abstract

Surrogate models can be used to accelerate approximate Bayesian computation (ABC). In one such framework the discrepancy between simulated and observed data is modelled with a Gaussian process. So far principled strategies have been proposed only for sequential selection of the simulation locations. To address this limitation, we develop Bayesian optimal design strategies to parallellise the expensive simulations. We also touch the problem of quantifying the uncertainty of the ABC posterior due to the limited budget of simulations.

1. Introduction

Approximate Bayesian computation (Marin et al., 2012; Lintusaari et al., 2017) is used for Bayesian inference when the analytic form of the likelihood function of a statistical model of interest is either unavailable or too costly to evaluate, but simulating the model is feasible. Unfortunately, many models e.g. in genomics and epidemiology (Numminen et al., 2013; Marttinen et al., 2015; McKinley et al., 2018) and climate science (Holden et al., 2018) are costly to simulate making sampling-based ABC inference algorithms infeasible. To increase sample-efficiency of ABC, various methods using surrogate models such as neural networks (Papamakarios and Murray, 2016; Papamakarios et al., 2019; Lueckmann et al., 2019; Greenberg et al., 2019) and Gaussian processes (Meeds and Welling, 2014; Wilkinson, 2014; Gutmann and Corander, 2016; Järvenpää et al., 2018, 2019a,b) have been proposed.

In one promising surrogate-based ABC framework the discrepancy between the observed and simulated data is modelled with a Gaussian process (GP) (Gutmann and Corander, 2016; Järvenpää et al., 2018, 2019a). Sequential *Bayesian experimental design* (or *active learning*) methods to select the simulation locations so as to maximise the sample-efficiency in this framework were proposed by Järvenpää et al. (2019a). However, one often has access to multiple computers to run some of the simulations in parallel. In this work, motivated by the related problem of batch Bayesian optimisation (Ginsbourger et al., 2010; Desautels et al., 2014; Shah and Ghahramani, 2015; Wu and Frazier, 2016) and the parallel GP-based method by Järvenpää et al. (2019b) for inference tasks where noisy and potentially expensive log-likelihood evaluations can be obtained, we resolve this limitation by developing principled batch simulation methods which considerably decrease the wall-time needed for ABC inference.

The posterior distribution is often summarised for further decision making using e.g. expectation and variance. When the computational resources for ABC inference are limited, it would be important to assess the accuracy of such summaries, but this has not been explicitly acknowledged in earlier work. We devise an approximate numerical method to propagate the uncertainty of the discrepancy, represented by the GP model, to the resulting ABC posterior summaries. We call our resulting framework as *Bayesian ABC* in analogy with the related problems of *Bayesian quadrature* (O'Hagan, 1991; Osborne et al., 2012; Briol et al., 2019) and *Bayesian optimisation* (BO) (Brochu et al., 2010; Shahriari et al., 2015).

2. Bayesian ABC framework

Let $\pi(\boldsymbol{\theta})$ denote the prior density of the (continuous) parameters $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$ of a statistical model of interest and $\pi(\mathbf{x}_{\text{obs}} \,|\, \boldsymbol{\theta})$ corresponding intractable likelihood function. Standard ABC algorithms such as the ABC rejection sampler target the approximate posterior

$$\pi_{\text{ABC}}(\boldsymbol{\theta}|\mathbf{x}_{\text{obs}}) \triangleq \frac{\pi(\boldsymbol{\theta}) \int_{\mathcal{X}} \mathbb{1}_{\Delta(\mathbf{x}_{\text{obs}},\mathbf{x}) \leq \varepsilon} \pi(\mathbf{x}|\boldsymbol{\theta}) \,\mathrm{d}\mathbf{x}}{\int_{\Theta} \pi(\boldsymbol{\theta}') \int_{\mathcal{X}} \mathbb{1}_{\Delta(\mathbf{x}_{\text{obs}},\mathbf{x}) \leq \varepsilon} \pi(\mathbf{x}'|\boldsymbol{\theta}') \,\mathrm{d}\mathbf{x}' \,\mathrm{d}\boldsymbol{\theta}'}, \tag{1}$$

where $\Delta : \mathcal{X}^2 \to \mathbb{R}_+$ is a discrepancy function used to compare the similarity between simulated data $\mathbf{x} \in \mathcal{X}$ and observed data $\mathbf{x}_{\text{obs}} \in \mathcal{X}$, and ε is a threshold parameter.

The main idea of Bayesian ABC is to explicitly use another layer of Bayesian inference to estimate the ABC posterior in Eq. 1. The previously obtained discrepancy-parameter-pairs are treated as data to learn a surrogate model, which will predict the discrepancy for a given parameter value. The surrogate model is further used to form an estimator for the ABC posterior in Eq. 1 and to adaptively acquire new data using Bayesian experimental design.

We make the assumption that the discrepancy evaluation Δ_i at $\boldsymbol{\theta}_i$ is generated as $\Delta_i = f(\boldsymbol{\theta}_i) + \nu_i$ with $\nu_i \sim_{\text{i.i.d.}} \mathcal{N}(0, \sigma_n^2)$, where $\sigma_n^2 > 0$ is the variance of the discrepancy. To encode the assumptions of smoothness and e.g. potential quadratic shape of the discrepancy $\Delta_{\boldsymbol{\theta}}$, its unknown mean function f is given a GP prior. Given $D_t \triangleq \{(\Delta_i, \boldsymbol{\theta}_i)\}_{i=1}^t$, we obtain $f \,|\, D_t \sim \mathcal{GP}(m_t(\boldsymbol{\theta}), c_t(\boldsymbol{\theta}, \boldsymbol{\theta}'))$. See Appendix A.1 for the details of the GP prior used and the formulas for $m_t(\boldsymbol{\theta})$ and $c_t(\boldsymbol{\theta}, \boldsymbol{\theta}')$. We define $s_t^2(\boldsymbol{\theta}) \triangleq c_t(\boldsymbol{\theta}, \boldsymbol{\theta})$ and $\Pi_{D_t}^f \triangleq \mathcal{GP}(m_t(\boldsymbol{\theta}), c_t(\boldsymbol{\theta}, \boldsymbol{\theta}'))$. If f and σ_n^2 were known, the ABC posterior could be obtained from Eq. 1 as

$$\pi_{\text{ABC}}^f(\boldsymbol{\theta}) \triangleq \frac{\pi(\boldsymbol{\theta}) \Phi\left((\varepsilon - f(\boldsymbol{\theta}))/\sigma_n\right)}{\int_{\Theta} \pi(\boldsymbol{\theta}') \Phi\left((\varepsilon - f(\boldsymbol{\theta}'))/\sigma_n\right) \mathrm{d}\boldsymbol{\theta}'}, \tag{2}$$

where $\Phi(\cdot)$ is the Gaussian cdf. As we have only access to observations D_t, our knowledge about f is represented by the Gaussian measure $f \sim \Pi_{D_t}^f$. The posterior of π_{ABC}^f in Eq. 2, describing its uncertainty due to the limited t simulations, is then obtained as the push-forward measure through the mapping $f \mapsto \pi_{\text{ABC}}^f$. While this is analytically intractable, the mean, variance and quantiles of the *unnormalised* ABC posterior $\tilde{\pi}_{\text{ABC}}^f(\boldsymbol{\theta}) \triangleq \pi(\boldsymbol{\theta}) \Phi((\varepsilon - f(\boldsymbol{\theta}))/\sigma_n)$, can be computed analytically allowing efficient computations, see Appendix A.1.

2.1. Parallel simulations

We aim to find informative simulation locations for obtaining the best possible estimate of the ABC posterior π_{ABC}^f given the postulated GP model. We here consider the (synchronous) batch setting where b simulations are simultaneously selected to be computed in parallel at each iteration of our algorithm. Consider a loss function $l : \mathscr{D}^2 \to \mathbb{R}_+$ so that $l(\pi_{\mathrm{ABC}}, d)$ quantifies the penalty of reporting $d \in \mathscr{D}$ as our ABC posterior when the true one is $\pi_{\mathrm{ABC}} \in \mathscr{D}$. Given D_t, the one-batch-ahead Bayes-optimal selection of the next batch of b evaluations $\boldsymbol{\theta}^{\mathrm{opt}} = [\boldsymbol{\theta}_1^{\mathrm{opt}}, \ldots, \boldsymbol{\theta}_b^{\mathrm{opt}}]$ is then obtained as $\boldsymbol{\theta}^{\mathrm{opt}} = \arg\min_{\boldsymbol{\theta}^* \in \Theta^b} L_t(\boldsymbol{\theta}^*)$, where

$$L_t(\boldsymbol{\theta}^*) = \mathbb{E}_{\boldsymbol{\Delta}^* \,|\, \boldsymbol{\theta}^*, D_t} \mathcal{L}(\Pi^f_{D_t \cup D^*}), \quad \mathcal{L}(\Pi^f_{D_t \cup D^*}) \triangleq \min_{d \in \mathscr{D}} \mathbb{E}_{f \,|\, D_t \cup D^*} l(\pi^f_{\mathrm{ABC}}, d). \tag{3}$$

In Eq. 3, an expectation over b future discrepancy evaluations $\boldsymbol{\Delta}^* = (\Delta_1^*, \ldots, \Delta_b^*)^\top$ at locations $\boldsymbol{\theta}^*$ needs to be computed assuming $\boldsymbol{\Delta}^*$ follows the current GP model. The expectation is taken of the *Bayes risk* $\mathcal{L}(\Pi^f_{D_t \cup D^*})$ resulting from the nested decision problem of choosing the estimator d, assuming $\boldsymbol{\Delta}^*$ are known and merged with current data D_t via $D^* \triangleq \{(\Delta_i^*, \boldsymbol{\theta}_i^*)\}_{i=1}^b$.

Using a loss function based on $\tilde{\pi}^f_{\mathrm{ABC}}$ instead of π^f_{ABC} allows tractable computations. If we choose L^2 loss function $\tilde{l}_2 \triangleq \int_\Theta (\tilde{\pi}^f_{\mathrm{ABC}}(\boldsymbol{\theta}) - \tilde{d}(\boldsymbol{\theta}))^2 \,\mathrm{d}\boldsymbol{\theta}$ between the unnormalised ABC posterior $\tilde{\pi}^f_{\mathrm{ABC}}$ and its estimator $\tilde{d}$, then the optimal estimator is the mean in Eq. 11. The resulting expected integrated variance (EIV) *acquisition function*, denoted as $L_t^{\mathrm{v}}(\boldsymbol{\theta}^*)$, is

$$L_t^{\mathrm{v}}(\boldsymbol{\theta}^*) = 2\int_\Theta \pi^2(\boldsymbol{\theta}) \left[T\left(a_t(\boldsymbol{\theta}), \frac{\sqrt{\sigma_n^2 + s_t^2(\boldsymbol{\theta}) - \tau_t^2(\boldsymbol{\theta};\boldsymbol{\theta}^*)}}{\sqrt{\sigma_n^2 + s_t^2(\boldsymbol{\theta}) + \tau_t^2(\boldsymbol{\theta};\boldsymbol{\theta}^*)}}\right) - T\left(a_t(\boldsymbol{\theta}), \frac{\sigma_n}{\sqrt{\sigma_n^2 + 2s_t^2(\boldsymbol{\theta})}}\right)\right] \mathrm{d}\boldsymbol{\theta}, \tag{4}$$

where $\tau_t^2(\boldsymbol{\theta};\boldsymbol{\theta}^*) = c_t(\boldsymbol{\theta},\boldsymbol{\theta}^*)[c_t(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*) + \sigma_n^2 \mathbf{I}]^{-1} c_t(\boldsymbol{\theta}^*,\boldsymbol{\theta})$ and T is Owen's T function. We use *greedy optimisation* as is also common in batch BO (see, e.g., Snoek et al., 2012; Wilson et al., 2018) and the integral over Θ is computed using importance sampling. We can also show that the corresponding L^1 loss function produces the marginal median in Eq. 13 of the Appendix as the optimal estimator. The resulting acquisition function, called expected integrated MAD (EIMAD), in addition to some heuristically-motivated batch methods used as baselines (called MAXV, MAXMAD), are developed in Appendix A.2.

2.2. Uncertainty quantification of the ABC posterior

Pointwise marginal uncertainty of the unnormalised ABC posterior $\tilde{\pi}^f_{\mathrm{ABC}}$ was used for selecting the simulation locations. However, knowing $\tilde{\pi}^f_{\mathrm{ABC}}$ and its marginal uncertainty in some individual $\boldsymbol{\theta}$-values is not very helpful for understanding the accuracy of the final estimate of π^f_{ABC}. Computing the distribution of e.g. ABC posterior expectation or marginals using π^f_{ABC} in Eq. 2 is clearly more intuitive. Unfortunately, such computations are difficult due to the nonlinear dependence on the infinite-dimensional quantity f. We propose a simulation-based approach where we combine drawing of GP sample paths and normalised importance sampling. For full details and an illustration, see Appendix A.3 and Fig. 3.

3. Experiments

We use two real-world simulation models to compare the performance of the sequential and synchronous batch versions of the acquisition methods. As a simple baseline, we consider random points (RAND) drawn from the prior. ABC-MCMC (Marjoram et al., 2003) with extensive simulations is used to compute the ground truth ABC posterior. Median total variation distance (TV) over 50 repeated simulations is used to measure the quality of approximations. See Appendix B for further details and C for additional results.

Lorenz model. This model describes the dynamics of slow weather variables and their dependence on unobserved fast weather variables. The model is represented by a coupled stochastic differential equation which can only be solved numerically resulting in an intractable likelihood. The model has two parameters $\boldsymbol{\theta} = (\theta_1, \theta_2)$ which we estimate from timeseries data. See Thomas et al. (2018) for full details and the experimental set-up that we also use, with the exception that we set $\boldsymbol{\theta} \sim \mathcal{U}([0,5] \times [0,0.5])$. The results are shown in Fig. 1(a). Furthermore, Fig. 1(b-c) demonstrates the uncertainty quantification of the expectation of the model-based ABC posterior. The effect of batch size is shown in Fig. 2(c).

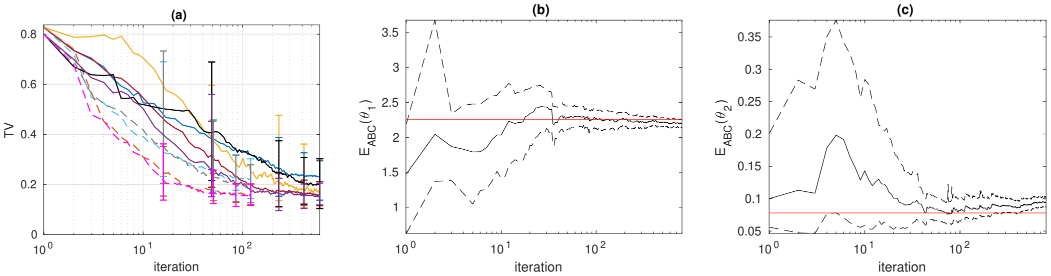

Figure 1: (a) Lorenz model. The intervals show the 90% variability. See Fig. 2(a) for the legend. (b-c) Black line is the mean and dashed black the 95% CI of the ABC posterior expectations. Red line shows the true value.

Bacterial infections model. This model describes transmission dynamics of bacterial infections in day care centers and features intractable likelihood function (Numminen et al., 2013). We estimate the internal, external and co-infection parameters $\beta \in [0, 11], \Lambda \in [0, 2]$ and $\theta \in [0, 1]$, respectively, using true data (Numminen et al., 2013) and uniform priors. The discrepancy is formed as in Gutmann and Corander (2016). The results with all methods are shown in Fig. 2(a) and Fig. 2(b) shows the effect of batch size for the two best methods.

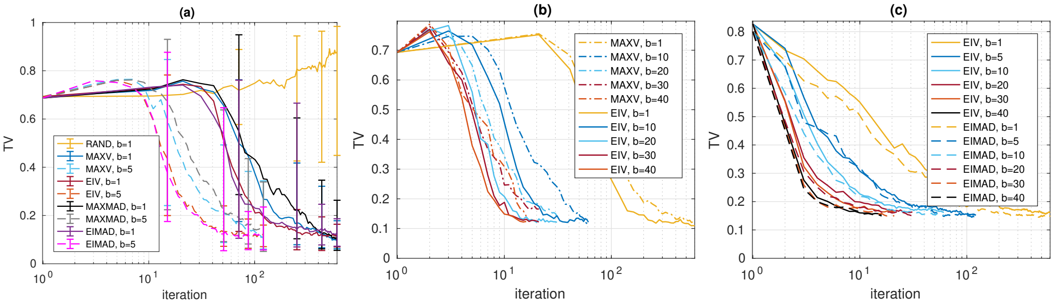

Figure 2: (a) Bacterial infections model. The intervals show the 90% variability. (b) Bacterial infections model. (c) Lorenz model. Recall that b denotes the batch size.

Discussion. We obtain reasonable posterior approximations considering the very limited budget of simulations. EIV and EIMAD tend to produce more stable and accurate ABC posterior estimates than MAXV and MAXMAD. Difference in approximation quality between

EIV and EIMAD, both based on the same Bayesian decision theoretic framework but different loss functions, was small. In all cases, our batch strategies produced similar evaluation locations as the corresponding sequential methods. This suggests that substantial improvements in wall-time can be obtained when the simulations are costly. The convergence of the uncertainty in the ABC posterior expectation in Fig. 1(b-c) is approximately towards the true ABC posterior expectation due to a slight GP misspecification. The ABC posterior marginals of the bacterial infection model in Appendix C contain some uncertainty after 600 iterations which our approach allows to rigorously quantify. Developing more effective (analytical) methods for computing these uncertainty estimates is an interesting topic for future work.

Acknowledgments

This work was funded by the Academy of Finland (grants no. 286607 and 294015 to PM). We acknowledge the computational resources provided by Aalto Science-IT project.

References

F.-X. Briol, C. J. Oates, M. Girolami, M. A. Osborne, and D. Sejdinovic. Probabilistic integration: A role in statistical computation? *Statistical Science*, 34(1):1–22, 02 2019.

E. Brochu, V. M. Cora, and N. de Freitas. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. 2010. Available at https://arxiv.org/abs/1012.2599.

H. R. Chai and R. Garnett. Improving quadrature for constrained integrands. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2751–2759, 2019.

T. Desautels, A. Krause, and J. W. Burdick. Parallelizing exploration-exploitation tradeoffs in Gaussian process bandit optimization. *Journal of Machine Learning Research*, 15: 4053–4103, 2014.

D. Ginsbourger, R. Le Riche, and L. Carraro. *Kriging Is Well-Suited to Parallelize Optimization*, pages 131–162. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

D. Greenberg, M. Nonnenmacher, and J. Macke. Automatic posterior transformation for likelihood-free inference. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2404–2414, 2019.

T. Gunter, M. A. Osborne, R. Garnett, P. Hennig, and S. J. Roberts. Sampling for inference in probabilistic models with fast Bayesian quadrature. In *Advances in Neural Information Processing Systems 27*, pages 2789–2797. 2014.

M. U. Gutmann and J. Corander. Bayesian optimization for likelihood-free inference of simulator-based statistical models. *Journal of Machine Learning Research*, 17(125):1–47, 2016.

H. Haario, M. Laine, A. Mira, and E. Saksman. DRAM: Efficient adaptive MCMC. *Statistics and Computing*, 16(4):339–354, 2006.

J. Hakkarainen, A. Ilin, A. Solonen, M. Laine, H. Haario, J. Tamminen, E. Oja, and H. Järvinen. On closure parameter estimation in chaotic systems. *Nonlinear Processes in Geophysis*, 19:127–143, 2012.

J. M. Hernández-Lobato, M. W. Hoffman, and Z. Ghahramani. Predictive Entropy Search for Efficient Global Optimization of Black-box Functions. *Advances in Neural Information Processing Systems 28*, pages 1–9, 2014.

P. Holden, N. Edwards, and R. Wilkinson. *ABC for climate: dealing with expensive simulators*, pages 569–95. 2018. In The Handbook of ABC.

M. Järvenpää, M. U. Gutmann, A. Vehtari, and P. Marttinen. Gaussian process modelling in approximate Bayesian computation to estimate horizontal gene transfer in bacteria. *The Annals of Applied Statistics*, 12(4):2228–2251, 2018.

M. Järvenpää, M. U. Gutmann, A. Pleska, A. Vehtari, and P. Marttinen. Efficient acquisition rules for model-based approximate Bayesian computation. *Bayesian Analysis*, 14(2): 595–622, 2019a.

M. Järvenpää, M. U. Gutmann, A. Vehtari, and P. Marttinen. Parallel Gaussian process surrogate method to accelerate likelihood-free inference. Available at https://arxiv.org/abs/1905.01252, 2019b.

J. Lintusaari, M. U. Gutmann, R. Dutta, S. Kaski, and J. Corander. Fundamentals and Recent Developments in Approximate Bayesian Computation. *Systematic biology*, 66(1): e66–e82, 2017.

J.-M. Lueckmann, G. Bassetto, T. Karaletsos, and J. H. Macke. Likelihood-free inference with emulator networks. In *Proceedings of The 1st Symposium on Advances in Approximate Bayesian Inference*, volume 96 of *Proceedings of Machine Learning Research*, pages 32–53, 2019.

J. M. Marin, P. Pudlo, C. P. Robert, and R. J. Ryder. Approximate Bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180, 2012.

P. Marjoram, J. Molitor, V. Plagnol, and S. Tavare. Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America*, 100(26):15324–8, 2003.

P. Marttinen, M. U. Gutmann, N. J. Croucher, W. P. Hanage, and J. Corander. Recombination produces coherent bacterial species clusters in both core and accessory genomes. *Microbial Genomics*, 1(5), 2015.

T. J. McKinley, I. Vernon, I. Andrianakis, N. McCreesh, J. E. Oakley, R. N. Nsubuga, M. Goldstein, and R. G. White. Approximate Bayesian computation and simulation-based inference for complex stochastic epidemic models. *Statistical Science*, 33(1):4–18, 2018.

E. Meeds and M. Welling. GPS-ABC: Gaussian Process Surrogate Approximate Bayesian Computation. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*, 2014.

E. Numminen, L. Cheng, M. Gyllenberg, and J. Corander. Estimating the transmission dynamics of streptococcus pneumoniae from strain prevalence data. *Biometrics*, 69(3): 748–757, 2013.

A. O'Hagan. Bayes-Hermite quadrature. *Journal of Statistical Planning and Inference*, 1991.

A. O'Hagan and J. F. C. Kingman. Curve fitting and optimal design for prediction. *Journal of the Royal Statistical Society. Series B (Methodological)*, 40(1):1–42, 1978.

M. A. Osborne, D. Duvenaud, R. Garnett, C. E. Rasmussen, S. J. Roberts, and Z. Ghahramani. Active Learning of Model Evidence Using Bayesian Quadrature. *Advances in Neural Information Processing Systems 26*, pages 1–9, 2012.

D. B. Owen. Tables for computing bivariate normal probabilities. *The Annals of Mathematical Statistics*, 27(4):1075–1090, 12 1956.

G. Papamakarios and I. Murray. Fast e-free inference of simulation models with Bayesian conditional density estimation. In *Advances in Neural Information Processing Systems 29*, 2016.

G. Papamakarios, D. Sterratt, and I. Murray. Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pages 837–848, 2019.

M. Patefield and D. Tandy. Fast and accurate calculation of Owen's T function. *Journal of Statistical Software*, 5(5):1–25, 2000.

G. Pleiss, J. Gardner, K. Weinberger, and A. G. Wilson. Constant-time predictive distributions for Gaussian processes. In *Proceedings of the 35th International Conference on Machine Learning*, pages 4114–4123, 2018.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20*, pages 1177–1184. 2008.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning.* The MIT Press, 2006.

J. Riihimäki and A. Vehtari. Laplace approximation for logistic Gaussian process density estimation and regression. *Bayesian Analysis*, 9(2):425–448, 2014.

A. Shah and Z. Ghahramani. Parallel Predictive Entropy Search for Batch Global Optimization of Expensive Objective Functions. In *Advances in Neural Information Processing Systems 28*, 2015.

B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas. Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, 104(1), 2015.

J. Snoek, H. Larochelle, and R. P. Adams. Practical Bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems 25*, pages 1–9, 2012.

O. Thomas, R. Dutta, J. Corander, S. Kaski, and M. U. Gutmann. Likelihood-free inference by ratio estimation. Available at https://arxiv.org/abs/1611.10242, 2018.

Z. Wang and S. Jegelka. Max-value entropy search for efficient Bayesian optimization. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3627–3635, 2017.

R. D. Wilkinson. Accelerating ABC methods using Gaussian processes. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, 2014.

J. Wilson, F. Hutter, and M. Deisenroth. Maximizing acquisition functions for Bayesian optimization. In *Advances in Neural Information Processing Systems 31*, pages 9906–9917. 2018.

J. Wu and P. Frazier. The parallel knowledge gradient method for batch Bayesian optimization. In *Advances in Neural Information Processing Systems 29*, pages 3126–3134. 2016.

Appendix A. Technical details

A.1. GP model for the discrepancy

We place the following hierarchical GP prior for f:

$$f \,|\, \boldsymbol{\gamma} \sim \mathcal{GP}(m_0(\boldsymbol{\theta}), k_{\boldsymbol{\phi}}(\boldsymbol{\theta}, \boldsymbol{\theta}')), \quad m_0(\boldsymbol{\theta}) \triangleq \sum_{i=1}^{r} \gamma_i h_i(\boldsymbol{\theta}), \quad \boldsymbol{\gamma} \sim \mathcal{N}(\mathbf{b}, \mathbf{B}), \tag{5}$$

where $k_{\boldsymbol{\phi}} : \Theta^2 \to \mathbb{R}$ is a covariance function with hyperparameters $\boldsymbol{\phi}$ and $h_i : \Theta \to \mathbb{R}$ are basis functions. We marginalise $\boldsymbol{\gamma}$ in Eq. 5, as in O'Hagan and Kingman (1978), and Riihimäki and Vehtari (2014), to obtain the equivalent GP prior

$$f \sim \mathcal{GP}(\mathbf{h}(\boldsymbol{\theta})^\top \mathbf{b}, k_{\boldsymbol{\phi}}(\boldsymbol{\theta}, \boldsymbol{\theta}') + \mathbf{h}(\boldsymbol{\theta})^\top \mathbf{B} \mathbf{h}(\boldsymbol{\theta}')), \tag{6}$$

where $\mathbf{h}(\boldsymbol{\theta}) \in \mathbb{R}^r$ is a column vector consisting of the basis functions h_i evaluated at $\boldsymbol{\theta}$. We assume the GP hyperparameters $\boldsymbol{\psi} \triangleq (\sigma_n^2, \boldsymbol{\phi})$ are fixed and omit $\boldsymbol{\psi}$ from our notation for brevity.

The mean and covariance functions of the GP posterior for f, when conditioned on data D_t, are

$$m_t(\boldsymbol{\theta}) \triangleq k_t(\boldsymbol{\theta}) \mathbf{K}_t^{-1} \boldsymbol{\Delta}_t + \mathbf{R}_t^\top(\boldsymbol{\theta}) \bar{\gamma}_t, \tag{7}$$

$$c_t(\boldsymbol{\theta}, \boldsymbol{\theta}') \triangleq k(\boldsymbol{\theta}, \boldsymbol{\theta}') - k_t(\boldsymbol{\theta}) \mathbf{K}_t^{-1} k_t^\top(\boldsymbol{\theta}') + \mathbf{R}_t^\top(\boldsymbol{\theta}) [\mathbf{B}^{-1} + \mathbf{H}_t \mathbf{K}_t^{-1} \mathbf{H}_t^\top]^{-1} \mathbf{R}_t(\boldsymbol{\theta}'), \tag{8}$$

where $[\mathbf{K}_t]_{ij} \triangleq k(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) + \mathbb{1}_{i=j}\sigma_n^2$, $k_t(\boldsymbol{\theta}) \triangleq (k(\boldsymbol{\theta}, \boldsymbol{\theta}_1), \ldots, k(\boldsymbol{\theta}, \boldsymbol{\theta}_t))^\top$, $\boldsymbol{\Delta}_t \triangleq (\Delta_1, \ldots, \Delta_t)^\top$ and

$$\bar{\gamma}_t \triangleq [\mathbf{B}^{-1} + \mathbf{H}_t \mathbf{K}_t^{-1} \mathbf{H}_t^\top]^{-1} (\mathbf{H}_t \mathbf{K}_t^{-1} \boldsymbol{\Delta}_t + \mathbf{B}^{-1} \mathbf{b}), \tag{9}$$

$$\mathbf{R}_t(\boldsymbol{\theta}) \triangleq \mathbf{H}(\boldsymbol{\theta}) - \mathbf{H}_t \mathbf{K}_t^{-1} k_t^\top(\boldsymbol{\theta}). \tag{10}$$

Above $\bar{\gamma}_t$ is the generalised least-squares estimate, $\mathbf{H}_t$ is the $r \times t$ matrix whose columns consist of basis function values evaluated at $\boldsymbol{\theta}_{1:t}$, $\boldsymbol{\theta}_{1:t}$ is a $p \times t$ matrix, and $\mathbf{H}(\boldsymbol{\theta}) \in \mathbb{R}^r$ is the corresponding vector of test point $\boldsymbol{\theta}$. For further details of GP regression, see e.g. Rasmussen and Williams (2006).

Formulas for the mean, median and variance of $\tilde{\pi}^f_{\text{ABC}}$ were derived by Järvenpää et al. (2019a) in the case of a zero mean GP prior. It is easy to see that these formulas hold also for our more general GP model. For example,

$$\mathbb{E}_{f\,|\,D_t}(\tilde{\pi}^f_{\text{ABC}}(\boldsymbol{\theta})) = \pi(\boldsymbol{\theta})\Phi(a_t(\boldsymbol{\theta})), \tag{11}$$

$$a_t(\boldsymbol{\theta}) \triangleq (\varepsilon - m_t(\boldsymbol{\theta}))/\sqrt{\sigma_n^2 + s_t^2(\boldsymbol{\theta})}, \tag{12}$$

$$\text{med}_{f\,|\,D_t}(\tilde{\pi}^f_{\text{ABC}}(\boldsymbol{\theta})) = \pi(\boldsymbol{\theta})\Phi\left((\varepsilon - m_t(\boldsymbol{\theta}))/\sigma_n\right), \tag{13}$$

$$\mathbb{V}_{f\,|\,D_t}(\tilde{\pi}^f_{\text{ABC}}(\boldsymbol{\theta})) = \pi^2(\boldsymbol{\theta})\left[\Phi(a_t(\boldsymbol{\theta}))\Phi(-a_t(\boldsymbol{\theta})) - 2T\left(a_t(\boldsymbol{\theta}), \sigma_n/\sqrt{\sigma_n^2 + 2s_t^2(\boldsymbol{\theta})}\right)\right], \tag{14}$$

where med denotes the marginal (i.e. elementwise) median.

A.2. Other acquisition functions

The EIMAD acquisition function, denoted as $L_t^{\text{m}}(\boldsymbol{\theta}^*)$ can be shown to be

$$L_t^{\text{m}}(\boldsymbol{\theta}^*) = 2\int_\Theta \pi(\boldsymbol{\theta})T\left(a_t(\boldsymbol{\theta}), \frac{\sqrt{s_t^2(\boldsymbol{\theta}) - \tau_t^2(\boldsymbol{\theta};\boldsymbol{\theta}^*)}}{\sqrt{\sigma_n^2 + \tau_t^2(\boldsymbol{\theta};\boldsymbol{\theta}^*)}}\right)\mathrm{d}\boldsymbol{\theta}, \tag{15}$$

where, similarly as for EIV in Eq. 4, T is Owen's T function (Owen, 1956) and a_t is given by Eq. 12. MAD stands for mean absolute deviation (around median). We do not show a detailed derivation of EIV and EIMAD acquisition functions here but only note that these can be obtained using similar computations as in Järvenpää et al. (2019a,b).

We consider also a heuristic acquisition function which evaluates where the pointwise uncertainty of $\tilde{\pi}^f_{\text{ABC}}(\boldsymbol{\theta})$ is highest. Such intuitive strategy is sometimes called as *uncertainty sampling* and used, e.g., by Gunter et al. (2014), Järvenpää et al. (2019a) and Chai and Garnett (2019). When variance is used as the measure of uncertainty of $\tilde{\pi}^f_{\text{ABC}}(\boldsymbol{\theta})$, we call the method as MAXV and when MAD is used, we obtain an alternative strategy called analogously MAXMAD. The resulting acquisition functions can be computed analytically. Specifically, the variance is computed using Eq. 14. A similar formula can be derived for MAD.

Finally, we propose a heuristic approach from BO (Snoek et al., 2012) to parallellise MAXV and MAXMAD strategies: The first point in the batch is chosen as in the sequential case. The further points are iteratively selected as the locations where the expected variance (or MAD), taken with respect to the discrepancy values of the pending points, that is points that have been already chosen to the current batch, is highest. The resulting acquisition functions are immediately obtained as the integrands of Eq. 4 and 15.

A.3. Uncertainty quantification of the ABC posterior

Fig. 3 demonstrates the GP modelling and the uncertainty quantification of the ABC posterior in a simple 1D scenario.

To access the posterior of π^f_{ABC}, it would be possible to fix a sample path $f^{(i)} \sim \Pi^f_{D_t}$, then use it to fix a realisation of the ABC posterior $\pi^{f^{(i)}}_{\text{ABC}}$ using Eq. 2 and finally use e.g. MCMC

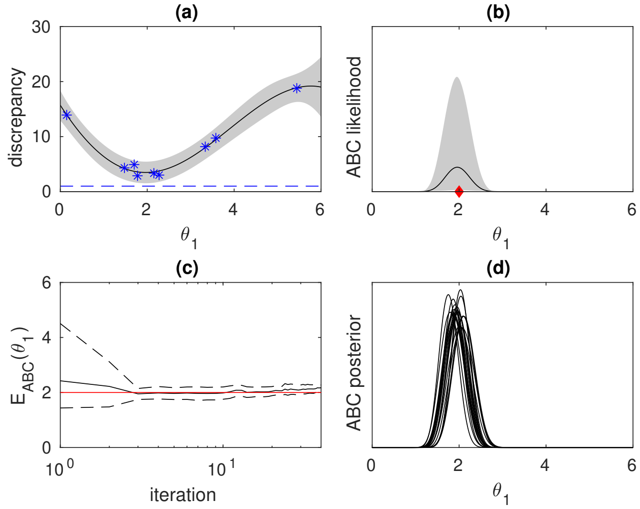

Figure 3: Demonstration of ABC posterior uncertainty quantification using Lorenz model with parameter θ_2 fixed. (a) GP model for Δ_{θ_1} (blue dashed line ε, blue stars 10 discrepancy evaluations), (b) uncertainty of unnormalised ABC posterior $\tilde{\pi}^f_{\text{ABC}}$, (c) evolution of model-based ABC posterior expectation (black line) and its 95% CI (dashed black) for 40 iterations, (d) uncertainty of ABC posterior π^f_{ABC}.

to sample from $\pi^{f^{(i)}}_{\text{ABC}}$. This would be repeated s times and the resulting set of samples $\{\{\boldsymbol{\theta}^{(i,j)}\}_{j=1}^{n}\}_{i=1}^{s}$ (where n is the length of the MCMC chain for each posterior realisation $i = 1, \ldots, s$) approximately describes the posterior of π^f_{ABC} given D_t. The uncertainty of GP hyperparameters $\boldsymbol{\psi}$ could also be taken into account by drawing $\boldsymbol{\psi}^{(i)} \sim \pi(\boldsymbol{\psi} \,|\, D_t)$ as the very first step but we here consider $\boldsymbol{\psi}$ as known for simplicity although this causes some underestimation of the uncertainty of π^f_{ABC}. The outlined approach involves a major computational challenge as evaluating the s sample paths at n different sets of test points scales[1] as $\mathcal{O}(s(nt^2 + tn^2) + sn^3)$.

We propose the following approach: In small dimensions, when $p \leq 2$, we evaluate each sample path $f^{(i)}, i = 1, \ldots, s$ at $\bar{n}^p$ fixed grid points and compute the required integrations numerically. This approach scales as $\mathcal{O}(\bar{n}^p t^2 + \bar{n}^{2p}(t+s) + \bar{n}^{3p})$. If $p > 2$, then self-normalised importance sampling is used. We draw n samples from instrumental density, the α-quantile of $\tilde{\pi}^f_{\text{ABC}}$ interpreted as a pdf with $\alpha = 0.95$. The samples are thinned and the resulting $\tilde{n} \ll n$ representative samples $\{\boldsymbol{\theta}^{(j)}\}_{j=1}^{\tilde{n}}$ are used to compute the normalised importance weights $\omega^{(i,j)}$ for each sampled posterior $i = 1, \ldots, s$. The output is a set of weighted sample sets $\{\{(\omega^{(i,j)}, \boldsymbol{\theta}^{(j)})\}_{j=1}^{\tilde{n}}\}_{i=1}^{s}$ from which moments and marginal densities can be computed using standard Monte Carlo estimators for each $i = 1, \ldots, s$. This approach requires only

[1] Approximations such as random Fourier features (RFF) (Rahimi and Recht, 2008) and those by Pleiss et al. (2018) can be used to reduce this cost, e.g. Hernández-Lobato et al. (2014) and Wang and Jegelka (2017) used RFF to approximately optimise GP sample paths. However, this produces tradeoff between exact GP but small n v.s. large n but inexact GP which we do not analyse in this paper.

one MCMC sampling from the instrumental density and scales as $\mathcal{O}(nt^2)$ so that n can be large. Total cost is $\mathcal{O}((n+\tilde{n})t^2 + \tilde{n}^2(t+s) + \tilde{n}^3)$.

Appendix B. Additional details of implementation and experiments

We briefly describe some key details of our algorithm and the experiments. Locations for fitting the initial GP model are sampled from the uniform prior in all cases. We take 10 initial points for 2D and 20 for 3D cases. We use $\mathbf{b} = \mathbf{0}$, $B_{ij} = 10^2 \mathbb{1}_{i=j}$ and include basis functions of the form $1, \theta_i, \theta_i^2$. The discrepancy $\Delta_{\boldsymbol{\theta}}$ is assumed smooth and we use the squared exponential covariance function $k(\boldsymbol{\theta}, \boldsymbol{\theta}') = \sigma_f^2 \exp(-\frac{1}{2}\sum_{i=1}^{p}(\theta_i - \theta_i')^2/l_i^2)$. GP hyperparameters $\boldsymbol{\psi} = (\sigma_n^2, l_1, \ldots, l_p, \sigma_f^2)$ are given weakly informative priors and their values are obtained using MAP estimation at each iteration. Owen's T function values are computed using a C-implementation of the algorithm by Patefield and Tandy (2000).

For simplicity and to ensure meaningful comparisons to ground-truth, we fix ε to certain small predefined values although, in practice, its value is set adaptively (Järvenpää et al., 2019a) or based on pilot runs. We compute the estimate of the unnormalised ABC posterior using the Eq. 11 for MAXV, EIV, RAND and Eq. 13 for MAXMAD, EIMAD. Adaptive MCMC (Haario et al., 2006) is used to sample from the resulting ABC posterior estimates and from instrumental densities needed for IS approximations. TV denotes the median total variation distance between the estimated ABC posterior and the true one (2D) or the average TV between their marginal TV values (3D) computed numerically over 50 repeated runs. Iteration (i.e. number of batches chosen) serves as a proxy for wall-time. The number of simulations i.e. the maximum value of t is fixed in all experiments and the batch methods thus finish earlier.

Mahalanobis distance was used as the discrepancy for Lorenz model. The simulation model was run 500 times to estimate the covariance matrix of the six summary statistics by Hakkarainen et al. (2012) at the true parameter and the the inverse of the covariance matrix was used in the Mahalanobis distance. Of course, such discrepancy is unavailable in practice because the true parameter is unknown and the computational budget limited. However, as the main goal of this paper is to approximate any given ABC posterior with a limited simulation budget, we chose our target ABC posterior this way.

Gutmann and Corander (2016) defined a discrepancy for the bacterial infections model by summing four L^1-distances computed between certain individual summaries. For details, see example 7 in Gutmann and Corander (2016). We used the same discrepancy except that we further took square root of their discrepancy function. We obtained a similar ABC posterior as the original article (Numminen et al., 2013) where ABC-PMC algorithm and slightly different approach for comparing the data sets were used.

Appendix C. Additional illustrations and experiments

Fig. 4 shows the evaluation locations and the resulting estimates of the ABC posteriors after 110 simulations (corresponding 100 iterations in the sequential case and 20 iteration is the batch-sequential case $b = 5$) for a synthetic 2D model called 'Banana'. This test problem was taken from Järvenpää et al. (2019a).

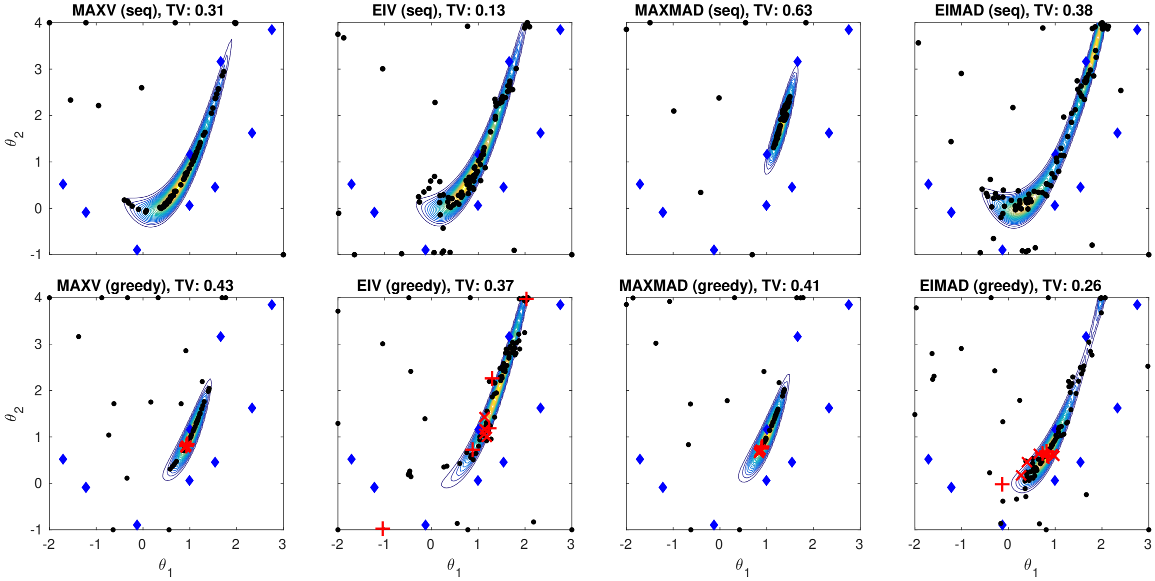

Figure 4: Illustration of evaluation locations. The first row shows the sequential methods and the second row the corresponding greedy batch methods. The blue diamonds show the 10 initial points and the black dots 100 additional points selected using each acquisition function (the last two batches in the second row are however highlighted by red plus-signs and crosses). The TV value in the title shows the total variation distance between the true and estimated ABC posteriors for each particular case.

Fig. 5 and 6 show typical estimated ABC posterior densities of the Lorenz and bacterial infections models, respectively. These results are shown to demonstrate the accuracy obtainable with very limited simulations. These particular results were obtained with the sequential EIV method using 600 iterations corresponding to 610 simulations (Lorenz model) or 620 simulations (bacterial infections model).

Fig. 7 illustrates the ABC posterior uncertainty quantification for the bacterial infections model. Sequential EIV method was used and one typical case is shown. The results suggest that while the ABC posterior is well estimated at the last iteration, there is some uncertainty left about its exact shape.

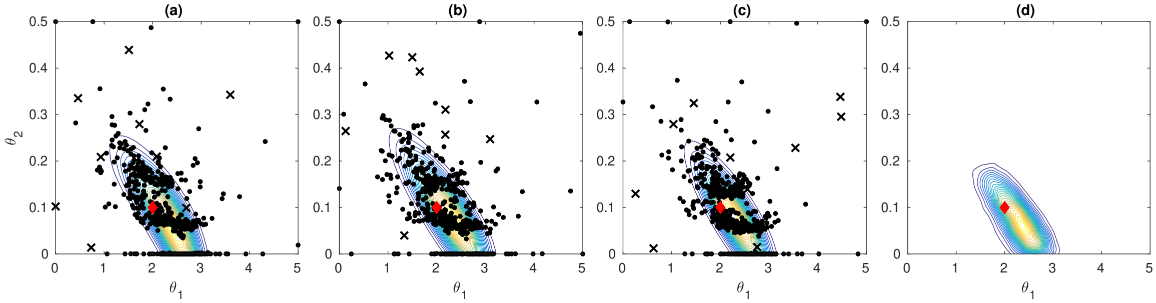

Figure 5: Estimated ABC posteriors for the Lorenz model. (a-c) Three typical estimates of the ABC posterior with corresponding simulation locations. Initial locations are shown as black crosses and the ones selected using EIV acquisition function are shown as black dots. The true parameter value used to generate the data is marked with the red diamond. (d) The true ABC posterior computed using ABC-MCMC with extensive simulations.

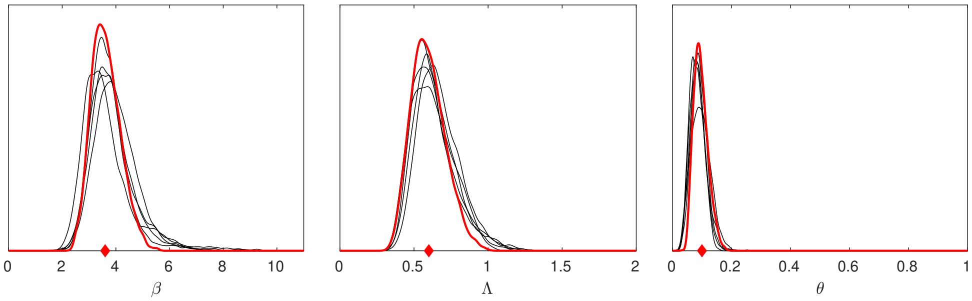

Figure 6: Estimated marginal ABC posteriors for the bacterial infections model. Red lines show the true ABC posterior computed using ABC-MCMC with extensive simulations. Black lines show five typical estimated ABC posteriors resulting from different simulation model realisations and the sets of initial simulation locations. The true parameter value used to generate the data is marked with the red diamond.

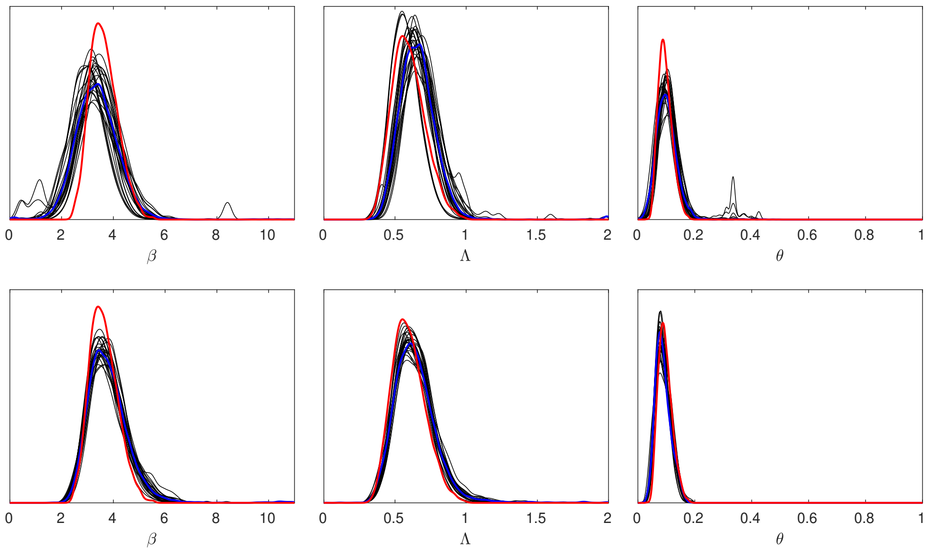

Figure 7: Uncertainty quantification for the ABC posterior marginals of the bacterial infections model at the 100th iteration corresponding $t = 120$ simulations (top row) and at the last iteration corresponding $t = 620$ simulations (bottom row). Red line shows the true ABC posterior, blue line shows the estimate based on Eq. 11 and the black lines show some sampled ABC marginal posteriors that (approximately) represent the uncertainty due to the limited number of simulations t.