Investigating model misspecification in simulation-based inference

Supervision:

- Pedro L. C. Rodrigues, Statify, Inria (pedro.rodrigues@inria.fr)
- Michael N. Arbel, Thoth, Inria (michael.arbel@inria.fr)
- Florence Forbes, Statify, Inria (florence.forbes@inria.fr)
- Julien Mairal, Thoth, Inria (julien.mairal@inria.fr)

Context Bayesian inference is a powerful framework for non-linear inverse problems. It yields a posterior probability density function that can be used to investigate which parameters $\boldsymbol{\theta}$ are the most likely to have generated a given experimental observation $\boldsymbol{x}$. Unfortunately, directly applying the Bayesian framework within modern complex scientific models is, in general, a difficult task, because their likelihood functions are often intractable or simply unavailable.

A modern approach for bypassing such difficulties is to use simulation-based inference (SBI), where one can learn an approximation to the likelihood or the posterior distribution based on several simulations over different parameters [1] similarly to what is classically done in approximate Bayesian computation (ABC) [2]. SBI leverages the recent advances in deep generative modeling and probabilistic programming and was demonstrated to be more sample efficient compared to ABC, to easily handle data-driven summary statistics, and avoid requiring any adjustment of ad-hoc parameters such as the ε-resolution hyper-parameter used in ABC.

> In simulation-based inference, we sample parameters $\boldsymbol{\theta}_i \sim p(\boldsymbol{\theta})$ from a prior distribution and generate a set of corresponding data points $\boldsymbol{x}_i \sim p(\boldsymbol{x} \mid \boldsymbol{\theta}_i)$ with a simulator describing the model of interest. The dataset $\{(\boldsymbol{\theta}_i, \boldsymbol{x}_i)\}$ is then used to learn an approximation $q_\psi(\boldsymbol{x} \mid \boldsymbol{\theta})$ of the model's likelihood or an approximation $q_\phi(\boldsymbol{\theta} \mid \boldsymbol{x})$ of the true posterior distribution $p(\boldsymbol{\theta} \mid \boldsymbol{x})$. Common classes of approximators are mixture density networks [3] and normalizing flows [4].

Challenges Despite the many recent applications of SBI to different scientific domains, such as neuroscience [5, 6], cosmology [7], and particle physics [8], there still remains a considerable gap between theoretical understanding on toy illustrative models and actual implementations for real noisy experimental data. One of the main difficulties comes from the fact that "all models are wrong" and, therefore, the simulations used to approximate the posterior distribution often don't exactly match the data obtained in real experiments. This important drawback is commonly known as *model misspecification* and has been often handled with heuristic approaches [6, 9].

> The main difficulty with model misspecification is that the statistical distribution of the data points $(\boldsymbol{\theta}_i, \boldsymbol{x}_i)$ used to train $q_\phi(\boldsymbol{\theta} \mid \boldsymbol{x})$ can be very different from that of the actual data points $\boldsymbol{x}_0$ obtained in real experiments, due to many factors, such as observation noise, nuisance variables, imperfections in the model, etc. In such situations, the approximation $q_\phi(\boldsymbol{\theta} \mid \boldsymbol{x}_0)$ can be very far from the true posterior distribution and lead to erroneous conclusions.

Another important challenge in SBI is its application to *large-scale models* such as those studied in climatology [10] and computational neuroscience [11]. In such cases, the simulations can be very time consuming, making the usual "simulate-then-train" approach inviable. To counter this difficulty, one can build an emulator capable of approximating the simulator based on simpler and/or faster calculations – see [10] for an example in climatology. The posterior distribution can then be approximated using the data generated by the emulator. Note, however, that there will very likely exist a mismatch between the data from the true simulator and from its emulator, i.e. the model used to train the posterior approximator will be misspecified.

In this work, we will investigate and better understand the effects of model misspecification for the approximation of posterior distributions in the SBI framework. For carrying out this investigation, the doctoral student will begin his work with getting acquainted to the mathematical and computational tools of SBI (e.g. normalizing flows, automatic differentiation, probabilistic programming, etc) and then investigate the effects of model misspecification in simple toy models – the main references for this initial part are [9] and [2]. To this end, the student will rely on recent tools developed in the Statify team for checking the quality and calibration of posterior approximators [12]. He/she will then study the role of model misspecification in the context of emulators for large-scale models and investigate how the predefined degree of simplification of the emulator can be used to control the quality of the posterior approximations.

Environment The Ph.D. will take place in Inria Grenoble, in a joint collaboration between the Statify and Thoth teams. Statify specializes in the statistical modeling of systems involving data with a complex structure, aiming towards strong methodological developments for general applications in data science, such as neuroscience, environmental risk analysis, and geosciences. Thoth aims at developing machine learning models and methods that are both robust and efficient. Its main research directions range from computer vision using limited annotations and data to statistical machine learning and optimization. The doctoral student is expected to work closely with his supervisors as well as other researchers from both the Statify and Thoth teams. We also plan to collaborate with colleagues from the DataMove team (Bruno Raffin) due to their recent interest in applying SBI to large-scale models.

Main activities The PhD student is expected to perform the usual research activities that are essential to the success of a PhD program:

- Performing a rigorous and broad literature review.
- Formulating a problem, a hypothesis and proposing a new approach for solving/testing the problem or hypothesis.
- Writing scientific papers.
- Participating to research events: seminars, reading groups, workshops, conferences, etc.
- Dissertation writing and thesis defense.

Given the nature of the research topic with both theory and application components, the student will also have the opportunity to make the following contributions to the research community:

- Novel methodological developments at the frontier of the field.
- Efficient algorithms and their software implementations in commonly used machine learning frameworks.

Finally, any other activities required by Inria and the doctoral school such as course requirements will be part of the position's main requirements.

Requirements/Skills

- Strong mathematical background, specially advanced knowledge in probability and statistics.
- Good working knowledge in scientific programming; experience with a deep learning library (e.g. `pyTorch` or `tensorflow`) will be a plus. The applicant should show a strong interest in conducting rigorous machine learning experiments using these libraries.
- A successful candidate should have or be willing to develop a rigorous approach in their research, be receptive to feedback while also having critical thinking and willing to collaborate with researchers from different backgrounds and disciplines.
- Language: excellent English skills both written and spoken.

How to apply Please provide a CV, a cover letter and your Master's grades. A reference letter from the Master course supervisors (or equivalent) is also required.

References

[1] K. Cranmer, J. Brehmer, and G. Louppe, "The frontier of simulation-based inference," *Proceedings of the National Academy of Sciences*, vol. 117, no. 48, pp. 30055–30062, 2020.

[2] D. Frazier, C. Robert, and J. Rousseau, "Model misspecification in approximate Bayesian computation: consequences and diagnostics," *Journal of the Royal Statistical Society: Series B*, vol. 82, no. 2, pp. 421–444, 2019.

[3] F. Forbes, H. D. Nguyen, T. T. Nguyen, and J. Arbel, "Summary statistics and discrepancy measures for ABC via surrogate posteriors," *Statistics and Computing*, vol. 32, no. 85, 2022.

[4] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan, "Normalizing flows for probabilistic modeling and inference," *arXiv preprint arXiv:1912.02762*, 2019.

[5] P. L. Coelho Rodrigues, T. Moreau, G. Louppe, and A. Gramfort, "HNPE: Leveraging Global Parameters for Neural Posterior Estimation," in *NeurIPS 2021*, (Sydney (Online), Australia), Dec. 2021.

[6] Y. Bernaerts, M. Deistler, P. J. Gonçalves, J. Beck, M. Stimberg, F. Scala, A. S. Tolias, J. Macke, D. Kobak, and P. Berens, "Combined statistical-mechanistic modeling links ion channel genes to physiology of cortical neuron types," *bioRxiv*, 2023.

[7] P. Lemos, M. Cranmer, M. Abidi, C. Hahn, M. Eickenberg, E. Massara, D. Yallup, and S. Ho, "Robust simulation-based inference in cosmology with bayesian neural networks," *Machine Learning: Science and Technology*, vol. 4, p. 01LT01, feb 2023.

[8] J. Brehmer, "Simulation-based inference in particle physics," *Nature Reviews Physics*, vol. 3, pp. 305–305, Mar. 2021.

[9] D. Ward, P. Cannon, M. Beaumont, M. Fasiolo, and S. M. Schmon, "Robust neural posterior estimation and statistical model criticism," *CoRR*, vol. abs/2210.06564, 2022.

[10] M. F. Kasim, D. Watson-Parris, L. Deaconu, S. Oliver, P. Hatfield, D. H. Froula, G. Gregori, M. Jarvis, S. Khatiwala, J. Korenaga, J. Topp-Mugglestone, E. Viezzer, and S. M. Vinko, "Building high accuracy emulators for scientific simulations with deep neural architecture search," *Machine Learning: Science and Technology*, vol. 3, p. 015013, dec 2021.

[11] N. Tolley, P. L. C. Rodrigues, A. Gramfort, and S. Jones, "Methods and considerations for estimating parameters in biophysically detailed neural models with simulation based inference," *bioRxiv*, 2023.

[12] J. Linhart, A. Gramfort, and P. L. C. Rodrigues, "Validation diagnostics for sbi algorithms based on normalizing flows," 2022.