JOERI HERMANS

ADVANCES IN SIMULATION-BASED INFERENCE

University of Liège

Faculty of Applied Sciences
Department of Electrical Engineering and Computer Science

ADVANCES IN SIMULATION-BASED INFERENCE

TOWARDS THE AUTOMATION OF THE SCIENTIFIC METHOD THROUGH LEARNING ALGORITHMS

by

Joeri Hermans

Dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science.

EXAMINATION COMMITTEE

Prof. Dr. Pierre Geurts, University of Liége (president)
Prof. Dr. Gilles Louppe, University of Liége (promotor)
Prof. Dr. Louis Wehenkel, University of Liége
Dr. Dominique Sluse, University of Liége
Prof. Dr. Christophe Weniger, Universiteit van Amsterdam
Prof. Dr. Jakub Tomczak, Vrije Universiteit Amsterdam

Science is knowledge which we understand so well that we can teach it to a computer; and if we don't fully understand something, it is an art to deal with it. Since the notion of an algorithm or a computer program provides us with an extremely useful test for the depth of our knowledge about any given subject, the process of going from an art to a science means that we learn how to automate something.

— Donald E. Knuth [1]

Joeri Hermans: *Advances in Simulation-Based Inference*, Towards the Automation of the Scientific Method through Learning Algorithms.

WORK PERFORMED UNDER THE SUPERVISION OF
Prof. Dr. Gilles Louppe

Abstract

This dissertation presents several novel techniques and guidelines to advance the field of *simulation-based* inference. Simulation-based inference, or likelihood-free inference, refers to the process of statistical inference whenever *simulating* synthetic realizations $\boldsymbol{x}$ through detailed descriptions of their generating processes is possible, but evaluating the likelihood $p(\boldsymbol{x} \mid \boldsymbol{\vartheta})$ of parameters $\boldsymbol{\vartheta}$ tied to realizations $\boldsymbol{x}$ is *intractable*. What this effectively means is that while it is relatively simple to execute a computer simulation and collect samples from its generative process for various inputs $\boldsymbol{\vartheta}$, it is rather difficult to invert the process where one poses the question: "what set of parameters $\boldsymbol{\vartheta}$ could have been responsible producing $\boldsymbol{x}$ and what is their probability of doing that"?

An important point here is that the answer to this question only holds under the assumed scientific model – the model that the simulator implements.

The likelihood $p(\boldsymbol{x} \mid \boldsymbol{\vartheta})$ plays a central role in answering this question. However, for most scientific simulators, the direct evaluation of the (true and unknown) likelihood involves solving an inverse problem that rests on the integration of all possible forward realizations implicitly defined by the computer code of the simulator. This issue is the core reason why it is typically impossible to evaluate the likelihood model of a computer simulator: it requires us to integrate across all possible code paths for all inputs $\boldsymbol{\vartheta}$ that could have potentially led to the realization $\boldsymbol{x}$.

Classical statistical inference based on the likelihood is for this reason impractical. Nevertheless, approximate inference remains possible by relying on surrogates that produce estimates of key quantities necessary for statistical inference. This thesis introduces various techniques and guidelines to effectively construct such surrogates and demonstrates how these approximations should be applied **reliably**. We explicitly make the point that the dogma of data efficiency should not be central to the field. Rather, reliable approximations should if we ever are to deduce scientific results with the techniques we developed over the years. This point is strengthened by demonstrating that all techniques *can* produce approximations that are *not* reliable from a scientific point of view, that is, when one is interested in *constraining* parameters or models. We argue for novel protocols that provide theoretically backed reliability properties. To that end, this thesis introduces a novel algorithm that provides such guarantees in terms of

the binary classifier. In fact, the theoretical result is applicable to any binary classification problem.

Finally, these contributions are framed within the context of the automation of science. This thesis concerned itself with the automation of the last step of the scientific method, which is described as a recurrence over the sequence hypothesis, experiment, and conclusion. For the most part, the steps of hypothesis formation and experiment design remain however solely for the scientists to decide. Only occasionally are they explored, designed and automated through computer-assisted means. For these two steps, we provide research avenues and proof of concepts that could unlock their automation.

Usage

The contents of this thesis should be read in sequential order, as every part establishes dependencies in preceding chapters. To support the reproducability of this work, we provide all code on GitHub. Some results and demonstration are annotated with the clickable icon **</>**, which links to the code or Jupyter notebook used to generate it.

DIGITAL RESOURCES

The GitHub repository

`https://github.com/JoeriHermans/phd-thesis`

contains several resources:

1. *Software* in the form of demonstratory Jupyter notebooks and scripts to generate the figures and several experimental results. It should be noted that code specific to publications are not included in this repository. However, I have tried to make this thesis as self-contained as I could.
2. *This thesis* and its LaTeX source code.
3. *Presentations* associated with the contents of this thesis, including the slides of my doctoral defense.

Acknowledgments

The contents of this manuscript are financed by the *Fund for Research Training in Industry and Agriculture*, awarded by the *Fonds de la Recherche Scientifique*. This thesis would not have been possible if it wasn't for the generosity of many open-source software authors. Specifically, I would like to thank all developers of `pytorch` [2], `matplotlib` [3], `numpy` [4], the `jupyter` project [5], and all other open source tools I use daily. Thank you for building such thorough high-quality software.

Gilles, I would like to express my sincere gratitude and compliment you for your approach to guiding students, and of course, your near-infinite wisdom, creativity and playfulness. Thank you giving me the freedom to explore my own path, nudging me in the right direction along the way, and more importantly, allowing me to make mistakes and fail. Unknown to many, I have a moment of reflection at the end of every year. At that time, I classify the year as "successful" whenever I consider my historic self an idiot. Let me tell you, the gravity of the term idiot does not suffice to cover the past few years. It is insane how much I learned and the amount of skills I developed. For this reason alone, it was worth all the sweat. Although I probably do not often express my gratitude in public, let me tell you this is *very* sincere: Gilles, thank you.

Arnaud, Antoine, François, thank you in particular for your enthousiastic collaboration within the CASBI project (our internal name for everything related to conservative amortized simulation-based inference). There were a lot of insightful discussions that were necessary to make the theory work out. Thank you for your critical, bright and happy personalities. A big thank you goes to all physicists involved in these projects, thank you for your willingness to integrate and jointly innovate these machine learning workflows for better science! Finally, an applause for the friendly face at SEGI – Raphaël Philippart – for always finding the time to aid us whenever Alan needed maintenance, even during your holidays. Give that man a raise!

Volodimir and Misha, although our countries are effectively in a state of war, I'm sure we'll meet again at Lac Leman to celebrate the good times.

Kristof, Gaëtan, Wok and Nathan Ebel (whose identity remains elusive even after 8 years), thanks for all the support and cozy drinks over these years. We'll have another party with those Crypto profits soon!

Katelijn and Jens, thank you for entrusting me with being the godfather of Remi. It is a responsibility that I will carry until the end of my days. Let's spark his curiosity about this interesting but chaotic and unforgiving universe we live in. To all my grandparents and Annie, thank you for always having been there and pushing me to learn new things. I really miss those summer-evenings and -nights in the garden with the stars twinkling in the distance above us. Mom, dad, thank you for – always – being there for us, your unconditional love and support, no matter the context or magnitude of my stubbornness.

Finally but not least, Ellen, your smile, love and enthousiasm is surely the biggest indirect contributor to this manuscript. Thank you for motivating me to (eventually) finish this wall of text and just, you know, being there.

Contents

1

Introduction

In this dissertation we take a first step towards the complete automation of the scientific method, which can be summarized as a body of techniques whose goal is to acquire new knowledge about a phenomena of interest while at the same time incorporating previous domain insights and limiting the number of initial assumptions. The key to novel scientific knowledge lies in its ability to explain current observations and make precise predictions. Central to scientific knowledge is the concept of a natural law, a falsifiable generalization expressed through concise statements, supported by rigorous mathematical descriptions. Guided by this formalization, domain scientists compute the implied predictions or consequences of a law, and verify whether these are in agreement with nature by means of an experiment. Falsification of a law in this setup is decisive — if the computed predictions disagree with experiment, the law is wrong.

Although the computation of these consequences typically involves solving a set of equations, most are too complex to be studied analytically. The complexity does not necessarily arise from mathematical intricacies, but are rather driven by the implications of these equations and their dynamics. Meaning, even simple and straightforward descriptions can produce complex patterns that are challenging to study. Well-known instances of this class include the *Mandelbrot set* [6] whose boundary is a fractal curve, Sir Conway's *Game of Life* [7], and Dr. Wolfram's *Rule 30* [8].

Advances in computing technologies, programming languages, and a prevalent availability of high quality software libraries have jointly enabled domain scientists to express their scientific models through computer code with increasing fidelity. By expressing scientific models in the form of a computer program, or simulation model, the evaluation of a model's consequences simply amounts to the execution of said simulation model. Verifying whether the model is in agreement with nature therefore reduces to comparing the generated output of the simulation model with the data collected by the real experiment.

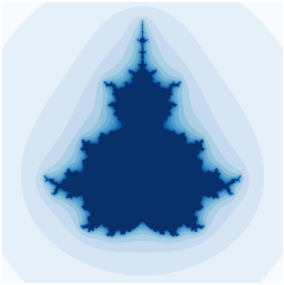

Figure 1.2: Visualization of the Mandelbrot set in the complex plane. A complex number c is part of the Mandelbrot set whenever the absolute value of the recursive computation of $z_{n+1} = z_n^2 + c$, for the sequence $n > 0$ and $z_0 = 0$ does not diverge. This mathematical description can easily be converted into an algorithm which – albeit approximately – decides whether c is an element of the Mandelbrot set. While the implementation of the decision problem is trivial, it generates geometrical structures of infinite richness. </>

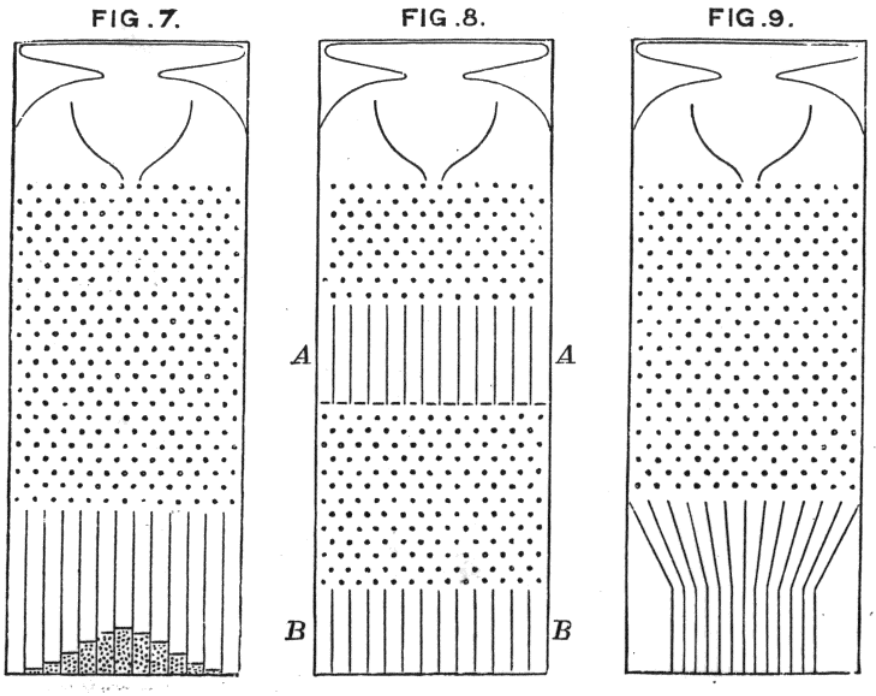

Figure 1.3: The Galton board as originally drawn by Sir Francis Galton.

Like most computer programs, simulation code and the scientific models they implement have configurable settings which influence the generated output — so-called model or free parameters $\boldsymbol{\vartheta}$. These parameters are of particular interest to domain scientists, as information about the value of these parameters automatically translates into scientific knowledge that could be applied to other models or theories. To *infer* the value of these parameters, domain scientists assume the scientific model to be true, and from there, determine plausible values such that the expected simulation output under the plausible set of values is in agreement with the experimentally observed data $\boldsymbol{x}$.

It should be noted however, that the aforementioned comparison is non-trivial, and we only mention it here to provide some insight on a conceptual level. The following chapters will treat approaches to this problem in detail.

Statistical inference plays a crucial role in determining the likelihood of parameter values or simulator configurations given the observed data. However, evaluating the likelihood implicitly defined through the computer code of the simulator requires us to solve an inverse problem that rests on the integration of all possible code paths, for all possible simulator configurations that could have potentially led to the observed data. Clearly, computing this quantity is infeasible or *intractable*. The increase in fidelity of modern computer simulations only amplified this problem.

To give the reader some additional intuition as to why the likelihood is intractable, let us briefly consider a metaphor that has been popularized by Kyle Cranmer and Gilles Louppe amongst others. It starts from the premises that the popular Galton board, as depicted in

Figure 1.3, could be viewed as a scientific simulator. The possible set of simulator outputs correspond to the various bins of the Galton board into which the beads can end up, whereas the simulator's configuration or free parameters relate to the position or bias of the various pegs. To simplify the discussion, let us make the assumption that there are $n + 1$ bins for n rows of pegs. Contrary to most simulators, the likelihood of a bead ending up in a particular bin *does* have a tractable likelihood whenever we consider an *idealized* Galton board. Under this assumed model, the probability of a bead ending up in bin k when counting from the left is defined as

$$\binom{n}{k} p^k (1-p)^{n-k}, \tag{1.1}$$

where p is the probability of a bead bouncing to the right. Recall that the evaluation of the likelihood depends on the integration of all possible code paths that could have produced the observed data. **If we view the bead traveling through the Galton board as an execution trace of a computer program with stochastic function calls**, then the number of possible paths the computer code can take to produce a bead in bin k is fully described by the binomial coefficient $\binom{n}{k}$. However, evaluating this likelihood analytically would *not* be possible if we were to change the position or bias of various pegs. In that case we would not be able to analytically describe the likelihood in the same way, but we would still be able to sample from the simulation model by simply dropping beads into the Galton board!

While the Galton board metaphor demonstrates that even for conceptual problems the computation of the likelihood quickly becomes impractical, the metaphor does not imply that statistical inference in these settings is impossible. In fact, one can still rely on *approximate* inference as long as it is **likelihood-free**. This is easier said than done as virtually all statistical inference relies on the likelihood in some way. However, the idea is that surrogates can be constructed **that do not rely on the direct evaluation of the likelihood** but rather produce estimates of key quantities necessary for statistical inference, be it numerically or otherwise. For instance, one such intractable quantity that is central to this dissertation is the Bayesian posterior

$$p(\boldsymbol{\vartheta} \mid \boldsymbol{x}) \triangleq p(\boldsymbol{\vartheta}) \frac{p(\boldsymbol{x} \mid \boldsymbol{\vartheta})}{p(\boldsymbol{x})}, \tag{1.2}$$

where the marginal model

$$p(\boldsymbol{x}) \triangleq \int \mathrm{d}\,\boldsymbol{\vartheta}\ \ p(\boldsymbol{\vartheta}) p(\boldsymbol{x} \mid \boldsymbol{\vartheta}), \tag{1.3}$$

for a given prior $p(\boldsymbol{\vartheta})$ quantifying the initial belief about the free parameters $\boldsymbol{\vartheta}$.

Simulation-based inference

In the literature, the problem setting outlined above is commonly referred to as likelihood-free - or **simulation-based inference**. The term is a common denominator for the process of **approximate** statistical inference whenever *simulating* synthetic realizations $\boldsymbol{x}$ through detailed descriptions of its generating processes is possible, but evaluating the likelihood $p(\boldsymbol{x} \mid \boldsymbol{\vartheta})$ of parameters $\boldsymbol{\vartheta}$ tied to realizations $\boldsymbol{x}$ is *intractable*.

This thesis will demonstrate that although the Bayesian posterior and other statistical quantities such as the likelihood ratio are not tractable in many problem domains and therefore not suitable for statistical inference, they can in fact be accurately approximated with modern *supervised* machine learning techniques. It should be noted that the use of supervised machine learning in this context does not imply that the desired target values, such as the posterior density function, are known beforehand. Rather, they are learned indirectly.

To that end, this dissertation contributes serveral techniques that are able to approximate these quantities in the aforementioned way.

1.1 Outline and thesis organization

Part i presents several inference protocol to effectively learn likelihood-free approximations with supervised machine learning techniques. Diagnostics are discussed to inspect the quality of these approximations and guidelines are put forward for the application of these algorithms to scientific problems. A complete demonstration of a simulation-based inference workflow within the context of inferring the properties of the Dark Matter particle is presented in Chapter 4.

Part ii tackles the problem of **reliable simulation-based inference**, which is critical to the goal of automated science. Unfortunately, but although expected, we provide experimental evidence that all (common) inference protocols can in fact produce unreliable estimates from a scientific point of view. That is, when the practitioner is concerned with constraining free parameters or models. Concretely, we show that the constraints these approximations produce – in expectation – are in fact stronger than their theoretical optima and are therefore *overconfident*. We motivate that likelihood-free approximations do not need to be exact. Rather, they should be conservative despite the available simulation budget and other hyper-parameters in order to be practically applicable. Since most algorithms do not reach their theoretical optima in practice anyway, we argue for theoretically motivated algorithms

that have guarantees *in practice*. To this end, we develop a reliability criterion that can be applied to any simulation-based inference protocol that relies on binary classifiers as a surrogate. In fact, the technique is applicable to any binary classification problem and opens the door for various applications outside of simulation-based inference.

Finally, Part iii summarizes the main conclusions of the dissertation and present several research avenues towards the complete automation of the scientific method.

1.1.1 *Publications*

The following publications form the core of this dissertation:

[9] Gilles Louppe, Joeri Hermans, and Kyle Cranmer. "Adversarial Variational Optimization of Non-Differentiable Simulators." In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Ed. by Kamalika Chaudhuri and Masashi Sugiyama. Vol. 89. Proceedings of Machine Learning Research. PMLR, 2019, pp. 1438–1447. URL: https://proceedings.mlr.press/v89/louppe19a.html

[10] Joeri Hermans, Volodimir Begy, and Gilles Louppe. "Likelihood-free MCMC with Amortized Approximate Ratio Estimators." In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 4239–4248. URL: https://proceedings.mlr.press/v119/hermans20a.html

[11] Joeri Hermans et al. "Towards constraining warm dark matter with stellar streams through neural simulation-based inference." In: *Monthly Notices of the Royal Astronomical Society* 507.2 (Aug. 2021), pp. 1999–2011. ISSN: 0035-8711. DOI: 10.1093/mnras/stab2181. eprint: https://academic.oup.com/mnras/article-pdf/507/2/1999/40078147/stab2181.pdf. URL: https://doi.org/10.1093/mnras/stab2181

[12] Joeri Hermans et al. "Averting A Crisis In Simulation-Based Inference." In: *arXiv e-prints*, arXiv:2110.06581 (Oct. 2021). arXiv: 2110.06581 [stat.ML]

Other manuscripts that are not included in this thesis, but were worked on during my scholarship:

[13] Joeri Hermans and Gilles Louppe. "Gradient Energy Matching for Distributed Asynchronous Gradient Descent." In: *arXiv e-prints*, arXiv:1805.08469 (May 2018). arXiv: 1805.08469 [cs.LG]

[14] Volodimir Begy et al. "Simulating Data Access Profiles of Computational Jobs in Data Grids." In: *2019 15th International Conference on eScience (eScience)*. 2019, pp. 394–402. DOI: 10.1109/eScience.2019.00051

[15] Johann Brehmer et al. "Mining for Dark Matter Substructure: Inferring Subhalo Population Properties from Strong Lenses with Machine Learning." In: *The Astrophysical Journal* 886.1 (2019), p. 49. DOI: 10.3847/1538-4357/ab4c41. URL: https://doi.org/10.3847/1538-4357/ab4c41

Part I

APPROXIMATING STATISTICAL QUANTITIES FOR SIMULATION-BASED INFERENCE

2

Adversarial Variational Optimization

The contents of this chapter are based on Louppe, Hermans, and Cranmer [9].

This chapter introduces *Adversarial Variational Optimization* (AVO), a simulation-based inference algorithm for fitting the parameters of a non-differentiable generative model to a set of observed data. The technique incorporates ideas from generative adversarial networks, variational optimization and empirical Bayes. In particular, we adapt the training procedure of generative adversarial networks by replacing the differentiable generative network with a non-differentiable domain-specific computer simulator. However, this particular formulation results in a non-differentiable minimax problem which is addressed by minimizing the variational upper bounds of two adversarial objectives designed such that AVO is able to adjust a parameterized proposal distribution over the simulator parameters. In doing so, AVO effectively minimizes the Jensen-Shannon divergence between the marginal distribution of the synthetic data generated by the simulator and the empirical distribution of the observed data.

2.1 INTRODUCTION

The inception of variational auto-encoders [16] and generative adversarial networks [17] initiated a vibrant research program around the learning *of* and *in* implicit generative models based on neural networks [18–24]. Like most scientific simulators, these generative models do not admit a tractable density. They are however, all differentiable by construction and can therefore be optimized by gradient descent. While generative models based on neural networks are highly parameterized, these parameters have no obvious interpretation. In contrast, scientific simulators can usually be thought of as highly regularized generative models because the models they implement can be described by relatively few parameters and are endowed with

some level of interpretation with a connection to the phenomena the scientific simulator is modelling.

2.2 PROBLEM STATEMENT

We consider a family of parameterized densities $p(\boldsymbol{x} \mid \boldsymbol{\vartheta})$ implicitely defined through a simulation model, where x is the observable and $\boldsymbol{\vartheta}$ are the model parameters of interest. The simulation may involve some complicated latent process where $z \in \mathcal{Z}$ is a latent variable providing an external source of randomness. Unlike implicit generative models based on neural networks, we do not assume z to be a fixed-size vector with a simple density. Instead, the dimension of z and the nature of its components (uniform, normal, discrete, continuous, etc.) are inherited from the control flow of the simulation code and may depend on $\boldsymbol{\vartheta}$ in some intricate way. Meaning, the execution of the simulation code for a specific parameterization $\boldsymbol{\vartheta}$ implicitely defines a distribution over x because the execution of code paths defined by the program is influenced by the values of the latent variables during the simulation.

FORMALISM We assume the stochastic generative process that implicitely defines $p(\boldsymbol{x} \mid \boldsymbol{\vartheta})$, is specified through a non-differentiable determinisitic function $g(\cdot) : \mathcal{Z} \to \mathbb{R}^d$, where d is the dimensionality of the observable x. Operationally, we consider a *synthetic* observable

$$x \sim p(\boldsymbol{x} \mid \boldsymbol{\vartheta}) \triangleq z \sim p(z \mid \boldsymbol{\vartheta}),\ x = g(z), \tag{2.1}$$

where the latent variable z *can* depend on the simulator parameter $\boldsymbol{\vartheta}$.

The dependence of the latent variable z on the simulator parameter $\boldsymbol{\vartheta}$ can be thought of in the following intuitive example: the value of $\boldsymbol{\vartheta}$ controls which specific code paths are executed. Depending on $\boldsymbol{\vartheta}$, there is effectively a distribution over the executed code, i.e., $p(z \mid \boldsymbol{\vartheta})$ by analogy.

OBJECTIVE Given a set of *observed* data $\{x_i\}_{i=1}^N$ drawn from the *unknown* true distribution $p_r(\boldsymbol{x})$, our goal is to estimate the parameter $\boldsymbol{\vartheta}^*$ which minimizes some divergence or distance ρ between $p_r(\boldsymbol{x})$ and the implicit model $p(\boldsymbol{x} \mid \boldsymbol{\vartheta})$. That is, we solve the following optimization problem

$$\boldsymbol{\vartheta}^* = \arg\min_{\boldsymbol{\vartheta}}\ \rho(p_r(\boldsymbol{x}) \mid\mid p(\boldsymbol{x} \mid \boldsymbol{\vartheta})), \tag{2.2}$$

which implies we are interested in *point-estimates* of $\boldsymbol{\vartheta}^*$.

2.3 BACKGROUND

2.3.1 *Generative adversarial networks*

Generative Adversarial Networks (GANS) were first proposed by Goodfellow et al. [17] as a way to build an implicit generative model capable of producing rich samples from random noise $z \sim p(z)$. The core principle of GANS is to pit a generative model $g_{\boldsymbol{\vartheta}}(\cdot) : \mathcal{Z} \to \mathbb{R}^d$ with

parameters $\boldsymbol{\vartheta}$ against an adversarial discriminator $d_{\boldsymbol{\phi}}(\cdot) : \mathbb{R}^d \to [0, 1]$ whose task is to recognize or *discriminate* empirically observed data $x_o \sim p_r(\boldsymbol{x})$ from synthetic data $x = g_{\boldsymbol{\vartheta}}(z),\ z \sim p(z)$. Both models are optimized simultaneously such that the generator $g_{\boldsymbol{\vartheta}}$ produces observables x for which $d_{\boldsymbol{\phi}}$ assigns the label "real", while $d_{\boldsymbol{\phi}}$ continuously adapts to infinitesimal changes in $g_{\boldsymbol{\vartheta}}$.

In practice however, the discriminator $d_{\boldsymbol{\phi}}$ and generator $g_{\boldsymbol{\vartheta}}$ are usually optimized with alternating stochastic gradient descent to respecitively minimize

$$\mathcal{L}_d(\boldsymbol{\phi}) \triangleq -\mathbb{E}_{p_r(\boldsymbol{x})}\left[\log d_{\boldsymbol{\phi}}(\boldsymbol{x})\right] - \mathbb{E}_{p(\boldsymbol{x}\,|\,\boldsymbol{\vartheta})}\left[\log(1 - d_{\boldsymbol{\phi}}(\boldsymbol{x}))\right], \tag{2.3}$$

$$\mathcal{L}_g(\boldsymbol{\vartheta}) \triangleq \mathbb{E}_{p(\boldsymbol{x}\,|\,\boldsymbol{\vartheta})}\left[\log(1 - d_{\boldsymbol{\phi}}(\boldsymbol{x}))\right], \tag{2.4}$$

where $\mathcal{L}_d(\boldsymbol{\phi})$ corresponds to the binary cross-entropy between the empirically observed and synthetic data produced by the implicit model $p(\boldsymbol{x}\,|\,\boldsymbol{\vartheta})$ defined by the generator $g_{\boldsymbol{\vartheta}}$.

Whenever $d_{\boldsymbol{\phi}}$ is trained to optimality before each infinitesimal small parameter update of the generator, it can be shown that the original adversarial learning procedure of Goodfellow et al. [17] amounts to minimizing the Jensen-Shannon divergence between the distributions $p_r(\boldsymbol{x})$ and $p(\boldsymbol{x}\,|\,\boldsymbol{\vartheta})$. Of course, this assumption is never met in practice and it often observed that the alternating optimization procedure of GANs does not lead to convergence. Research has therefore focused on finding optimization algorithms to promote stability in the optimization objective [25–29], and a better theoretical understanding of the training dynamics [30, 31].

2.3.2 *Computing gradients with respect to non-differentiable objectives*

Variational optimization [32, 33] and evolutionary strategies [34] are general optimization techniques that can be used to form a differentiable bound on the optima of a non-differentiable function. Given a function f to minimize and a proposal distribution $q_{\boldsymbol{\psi}}(\boldsymbol{\vartheta})$ parameterized by $\boldsymbol{\psi}$ over inputs $\boldsymbol{\vartheta}$, these techniques are based on the observation

$$\min_{\boldsymbol{\vartheta}} f(\boldsymbol{\vartheta}) \leq \mathbb{E}_{q_{\boldsymbol{\psi}}(\boldsymbol{\vartheta})}\left[f(\boldsymbol{\vartheta})\right]. \tag{2.5}$$

That is, the minimum of a set of function values is always less than or equal to any of their expectation. Provided the proposal distribution is sufficiently flexible, the parameters $\boldsymbol{\psi}$ can be updated to place the mass of the proposal distribution arbitrarily tight around the optimum

$$\boldsymbol{\vartheta}^* = \min_{\boldsymbol{\vartheta}} f(\boldsymbol{\vartheta}). \tag{2.6}$$

Under mild restrictions outlined by [32], the variational bound $U(\psi)$ is differentiable with respect to ψ. Using the log-likelihood trick, its gradient can be rewritten as

$$\begin{aligned}
\nabla_\psi \, U(\psi) &= \nabla_\psi \, \mathbb{E}_{q_\psi(\boldsymbol{\vartheta})} \left[f(\boldsymbol{\vartheta}) \right], && (2.7)\\
&= \nabla_\psi \int \mathrm{d}\,\boldsymbol{\vartheta} \;\; q_\psi(\boldsymbol{\vartheta}) \; f(\boldsymbol{\vartheta}), && (2.8)\\
&= \int \mathrm{d}\,\boldsymbol{\vartheta} \;\; \nabla_\psi \; q_\psi(\boldsymbol{\vartheta}) \; f(\boldsymbol{\vartheta}), && (2.9)\\
&= \int \mathrm{d}\,\boldsymbol{\vartheta} \;\; q_\psi(\boldsymbol{\vartheta}) \; \nabla_\psi \log \; q_\psi(\boldsymbol{\vartheta}), && (2.10)\\
&= \mathbb{E}_{q_\psi(\boldsymbol{\vartheta})} \left[\nabla_\psi \log \; q_\psi(\boldsymbol{\vartheta}) \right]. && (2.11)
\end{aligned}$$

Effectively, this means that provided the score function $\nabla_\psi \log q_\psi(\boldsymbol{\vartheta})$ of the proposal is known, and that one can evaluate $f(\boldsymbol{\vartheta})$ for any $\boldsymbol{\vartheta}$, then one can construct empirical estimates of Equation 2.7. These estimates can in turn be used to minimize $U(\psi)$ with stochastic gradient descent.

It should be noted that in the reinforcement learning literature, Equation 2.7 appears in the context of policy gradients, where $f(\boldsymbol{\vartheta})$ corresponds to a reward signal for the action $\boldsymbol{\vartheta}$ and the proposal $q_\psi(\boldsymbol{\vartheta})$ to a policy π_ψ we aim to optimize. In this context, empirical estimates of Equation 2.7 are better known as REINFORCE estimates [35].

2.4 METHOD

The alternating stochastic gradient descent on $\mathcal{L}_d(\phi)$ and $\mathcal{L}_g(\boldsymbol{\vartheta})$ in GANs implicitely assumes that the generator $g_{\boldsymbol{\vartheta}}$ is a differentiable function. This is in stark contrast to our specific setting, where it is not possible to compute the gradient $\nabla_{\boldsymbol{\vartheta}} \, g$ because differentiating through a scientific computer simulator, or any computer program for that matter, is typically not possible. As a result, the gradient $\nabla_{\boldsymbol{\vartheta}} \, \mathcal{L}_g$ cannot be computed and the optimization procedure cannot be carried out.

A SHORT INTERMEZZO ON DIFFERENTIABLE PROGRAMMING

There are various proposals [36–40] across the scientific community to integrate automatic differentiation packages inside scientfic simulators with the goal to effectively compute gradients with respect to domain-specific model parameters. Usually with the intended purpose to efficiently optimize the model parameters such that the best possible fit is obtained. Notable examples include Chianese et al. [41], in which a differentiable simulation of gravitational lensing is implemented using concepts from probabilistic programming combined with a differentiable surrogate of the underlaying physics simulation.

Although the above relies on a neural approximation of the physical system, fully differentiable simulations have the advantage that they obey dynamical laws by construction, i.e., concepts such as conservation of energy and momentum are embedded in the programming of the simulator. An aspect which needs to be specifically addressed in any approximation.

Closely related is the field of *physically based rendering* or *differentiable rendering* [42–48], in which the parameters of a 3D scene are optimized with respect to some target image. Recent approaches in differential rendering step away from automatic differentiation because of the vast memory requirements for complex scenes [46]. In addition, a prominent issue in differentiable rendering are (latent) boundary conditions, which causes the gradients to negatively infuence any gradient-based scene optimization procedure. On top of this is the problem of potential local minima and execution time of the involved physics simulation.

It is therefore not unthinkable that the same issues will arise in differentiable domain-specific simulators. The question becomes whether the benefits of implementing a *new* differentiable simulator, with the *potential* gains in statistical power, outweigh the cost of validating the simulation chain with respect to learning a sufficiently expressive surrogate. This has its own issues obviously, because the (Bayesian) inference results of the approximation still have to be validated in some way. However, it should be noted that a fully-differentiable simulator would still suffer from all common quirks attributed to gradient-based optimization.

In this work, we propose to address this problem by relying on variational optimization to minimize $\mathcal{L}_d$ and $\mathcal{L}_g$, thereby effectively bypassing the non-differentiability of the simulation model. We consider a proposal distribution $q_\psi(\boldsymbol{\vartheta})$ over the simulator parameters and alternately minimize the variational upper bounds

$$U_d(\phi) = \mathbb{E}_{q_\psi(\boldsymbol{\vartheta})}[\mathcal{L}_d(\phi)], \tag{2.12}$$

$$U_g(\psi) = \mathbb{E}_{q_\psi(\boldsymbol{\vartheta})}[\mathcal{L}_g(\boldsymbol{\vartheta})], \tag{2.13}$$

respectively over ϕ and ψ. The discriminator d_ϕ is therefore no longer pit against a single generator g, but instead against a family of generators induced by the proposal distribution.

When updating the discriminator parameters ϕ, unbiased estimates of $\nabla_\phi U_d$ can be obtained by directly evaluating the gradient of U_d over mini-batches of real and synthetic observables. When updating

the proposal parameters ψ, $\nabla_\psi U_g$ can be estimated as described in the previous section with $f(\boldsymbol{\vartheta}) = \mathcal{L}_g(\boldsymbol{\vartheta})$. That is,

$$\nabla_\psi U_g = \mathbb{E}_{p(\boldsymbol{x}\,|\,\boldsymbol{\vartheta})q_\psi(\boldsymbol{\vartheta})}\left[\nabla_\psi \log q_\psi(\boldsymbol{\vartheta}) \log(1 - d_\phi(\boldsymbol{x}))\right], \tag{2.14}$$

which we can approximate with mini-batches of synthetic observables produced by the simulation model. While the latter REINFORCE-like gradient estimator is unbiased, it is well-known that it suffers from high variance, which makes the optimization unstable and therefore tedious. A common remedy to this issue [35] is to make use of

$$\mathbb{E}_{q_\psi(\boldsymbol{\vartheta})}[\nabla_\psi \log q_\psi(\boldsymbol{\vartheta}) f(\boldsymbol{\vartheta})] = \mathbb{E}_{q_\psi(\boldsymbol{\vartheta})}[\nabla_\psi \log q_\psi(\boldsymbol{\vartheta})(f(\boldsymbol{\vartheta}) - b)] \tag{2.15}$$

for any constant b. The choice of a *baseline* b does not bias the gradient estimator, but it can however have an effect on its variance. For AVO, we pick the baseline which minimizes the variance of the empirical estimators of $\nabla_\psi U_g$, which is

$$b \triangleq \frac{\mathbb{E}\left[(\nabla_\psi \log q_\psi(\boldsymbol{\vartheta}))^2 (\log(1 - d_\phi(\boldsymbol{x}))^2\right]}{\mathbb{E}\left[(\nabla_\psi \log q_\psi(\boldsymbol{\vartheta}))^2\right]}. \tag{2.16}$$

For completeness, Algorithm 1 outlines the proposed procedure, as built on top of GANs with R_1 regularization [29]. Under suitable assumptions, this regularization term guarantees the (local) convergence of the training procedure, while keeping the original GAN algorithm otherwise unchanged.

2.4.1 *Empirical Bayes through Variational Inference*

The variational objectives in Equations 2.12-2.13 effectively replace the modeled data distribution $p(\boldsymbol{x}\,|\,\boldsymbol{\vartheta})$ with the parameterized marginal distribution of the generated data

$$q_\psi(\boldsymbol{x}) = \int \mathrm{d}\,\boldsymbol{\vartheta}\;\; p(\boldsymbol{x}\,|\,\boldsymbol{\vartheta}) q_\psi(\boldsymbol{\vartheta}). \tag{2.17}$$

We can think of $q_\psi(\boldsymbol{x})$ as a *variational program* as described by Ranganath et al. [49], although more complicated compared to a simple reparameterization of normally distributed noise z through a differentiable function. In our case, the variational program is a marginalized, non-differentiable program with an *intractable* density, i.e., the simulator. Nevertheless, it can generate samples x whose expectations are differentiable with respect to ψ through AVO. Operationally, we sample from this marginal model via

$$x \sim q_\psi(\boldsymbol{x}) \triangleq \boldsymbol{\vartheta} \sim q_\psi(\boldsymbol{\vartheta}),\; z \sim p(z|\,\boldsymbol{\vartheta}),\; x = g(z). \tag{2.18}$$

The optimization of $q_\psi(\boldsymbol{x})$ with respect to ψ can therefore be viewed through the lens of *Empirical Bayes*, where samples from $p_r(\boldsymbol{x})$ are used to optimize the prior within the family $q_\psi(\boldsymbol{\vartheta})$.

Algorithm 1 Adversarial Variational Optimization (AVO)

Inputs:	Observed data $\{x_i \sim p_r(\boldsymbol{x})\}_{i=1}^N$, simulator g.
Outputs:	Proposal distribution $q_\psi(\boldsymbol{\vartheta}) \approx p_r(\boldsymbol{x})$.
Hyper-parameters:	Training iterations k of the discriminator d_ϕ, Batch-size M (default = 32), R_1 regularization coefficient λ (default: $\lambda = 0$), Baseline strategy b in REINFORCE estimates, Entropy penalty coefficient γ (default: $\gamma = 0$).

$q_\psi(\boldsymbol{\vartheta}) \leftarrow$ prior on θ (with differentiable and known density)
while $q_\psi(\boldsymbol{\vartheta})$ has not converged **do**
  for $i = 1$ to k **do** ▷ Update d_ϕ
    $\mathcal{L}_d \leftarrow -\mathbb{E}_{p_r(\boldsymbol{x})}\left[\log d_\phi(\boldsymbol{x})\right] - \mathbb{E}_{p(\boldsymbol{x}\,|\,\boldsymbol{\vartheta})q_\psi(\boldsymbol{\vartheta})}\left[\log(1 - d_\phi(\boldsymbol{x}))\right]$
    $R_1 \leftarrow \mathbb{E}_{p_r(\boldsymbol{x})}\left[||\nabla_\phi\, d_\phi(\boldsymbol{x})||^2\right]$
    $\phi \leftarrow \text{OPTIMIZER}(\nabla_\phi \mathcal{L}_d + \lambda \nabla_\phi R_1)$
  end for
  Simulate $\{\boldsymbol{\vartheta}_m \sim q_\psi(\boldsymbol{\vartheta}),\ x \sim p(\boldsymbol{x}\,|\,\boldsymbol{\vartheta}_m)\}_{m=1}^M$ ▷ Update $q_\psi(\boldsymbol{\vartheta})$
  $\nabla_\psi U_g \leftarrow \frac{1}{M}\sum_{m=1}^M \nabla_\psi \log q_\psi(\boldsymbol{\vartheta})(\log(1 - d_\phi(\boldsymbol{x}_m)) - b)$
  $\nabla_\psi H \leftarrow -\frac{1}{M}\sum_{m=1}^M \nabla_\psi\ q_\psi(\boldsymbol{\vartheta}_m) \log q_\psi(\boldsymbol{\vartheta}_m)$
  $\psi \leftarrow \text{OPTIMIZER}(\nabla_\psi U_g + \gamma \nabla_\psi\ H)$
end while

Whenever the simulator is misspecified, AVO will smear $q_\psi(\boldsymbol{\vartheta})$ over the parameter space such that the marginal model $q_\psi(\boldsymbol{x})$ is as close as possible to $p_r(\boldsymbol{x})$ because the GAN-inspired procedure effectively minimizes the Jensen-Shannon divergence between the observed data distribution $p_r(\boldsymbol{x})$ and the marginal distribution $q_\psi(\boldsymbol{x})$. However, if the simulator is well-specified, then $q_\psi(\boldsymbol{\vartheta})$ will concentrate its mass around the true data generating parameter $\boldsymbol{\vartheta}^*$ whenever a sufficient amount of observables are available.

In order to effectively target point estimates of $\boldsymbol{\vartheta}^*$ through the maximum likelihood estimator of $q_\psi(\boldsymbol{\vartheta})$, we can augment the variational objective with an entropic regularization term $H_\psi(\boldsymbol{\vartheta})$ such that

$$U_g \triangleq \mathbb{E}_{p(\boldsymbol{x}\,|\,\boldsymbol{\vartheta})q_\psi(\boldsymbol{\vartheta})}\left[\mathcal{L}_g\right] + \gamma H\left[q_\psi(\boldsymbol{\vartheta})\right], \tag{2.19}$$

where $\gamma \in \mathbb{R}^+$ is a hyper-parameter controlling the trade-off between the generator object and the Shannon entropy H of the proposal distribution. For small values of γ, proposal distributions with large entropy are not penalized. On the other hand, for large values of γ, the procedure is encouraged to fit a proposal distribution with low entropy, which has the effect of concentrating its density.

The hyper-parameter γ effectively controls the "sharpness" of the proposal distribution.

Finally, we note that very large penalties might eventually make the optimization procedure unstable, as the variance of $\nabla_\psi \log q_\psi(\boldsymbol{\vartheta})$ typically increases as the entropy of the proposal distribution decreases.

Depending on the proposal, it may also be possible to always arbitrarily decrease the entropy, without necessarily producing accurate parameter point estimates. In this case, properly controlling for γ and the number of training epochs is critical. It should be noted however, that in practice this is not desired, as practicioners are interested in plausible values of the simulator parameters which are consistent with empirical observations.

Behaviour of the proposal parameterization ψ in practice

An important point to consider when applying AVO in practice is the choice of the proposal $q_\psi(\boldsymbol{\vartheta})$ and its parameterization ψ. In fact, for particular choices of $q_\psi(\boldsymbol{\vartheta})$, the parameterization ψ can have a significant effect on AVO's *stability* during the optimization procedure.

Intuition of the problem Consider the case where $q_\psi(\boldsymbol{\vartheta})$ is a normal distribution whose mean μ and standard deviation σ are embedded in ψ, i.e., $\psi \triangleq (\mu, \sigma)$. Because AVO effectively applies gradient descent on ψ, it remains numerically possible for the standard deviation σ to be smaller than 0. Such invalid configurations are in practice mainly reached due to *(i)* the settings of the optimization procedure (e.g., the learning rate), or *(ii)* optimizer effects such as *momentum* [50, 51].

Possible solutions After every optimization step, a procedure can be implemented to verify the integrity of ψ. Although various procedures could be implemented which are specific to the family of the proposal distributions in question, a simple strategy could be to compute the absolute value of the standard deviation or the diagonal of the covariance matrix whenever the proposal distribution was a normal or multivariate normal distribution respectively. While such an operation might in itself introduce instabilities in AVO's optimization procedure, the approach proved to be sufficient at a minimal implementation cost.

Alternatively, one could avoid hardcoding ad-hoc constraints like the above by parameterizing σ through some auxilliary free variable σ' such that $\psi \triangleq (\mu, \sigma')$ and $\sigma' = \log(1 + \exp(\sigma))$. By parameterizing ψ in this way, we ensure that invalid configurations of σ cannot be reached. Our reference implementation in `hypothesis`, and therefore the experiments in this manuscript, apply this procedure.

2.5 EXPERIMENTS

2.5.1 *Illustrative example*

As a first illustrative demonstration, we evaluate the inference performance of AVO on a discrete Poisson distribution with unknown rate $\boldsymbol{\vartheta}^*$. We artificially consider the distribution as the likelihood model of some scientific simulator, from which we can *only* generate data. That is, we consider evaluating the likelihood of the parameters with respected to the observed data – samples from $p_r(\boldsymbol{x})$ – to be intractable.

The discrete observed data is sampled from a Poisson distribution with rate $\boldsymbol{\vartheta}^* = 1.0$. We initialize the proposal distribution $q_\phi(\boldsymbol{\vartheta})$ as $\mathcal{N}(\mu = 8.0,\ \sigma = 2.5)$, whose density is completely described the mean $\mu = 8.0$ and the standard deviation $\sigma = 2.5$, such that $\psi \triangleq (\mu,\ \sigma)$. We execute AVO for 1000 iterations with mini-batches of sizes 32 using the default configuration specified in Algorithm 1.

The top row in Figure 2.1 illustrates the effects of updating the proposal distribution $q_\psi(\boldsymbol{\vartheta})$ after executing AVO for 250 iterations. In particular, the left subplot shows how the empirical distribution $p_r(\boldsymbol{x})$ compares against the estimated marginal model $q_\psi(\boldsymbol{x})$, while the right subplot shows the evolution of the proposal distribution compared to its initial state. The bottom row – the results after having ran AVO for 1000 iterations – shows the final result. It should be noted the proposal distribution correctly concentrates its density around the true generating parameter value $\boldsymbol{\vartheta}^* = 1.0$, yielding in this case more precise inference as the uncertainty – directly related to the entropy of $q_\psi(\boldsymbol{\vartheta})$ – reduces. In addition, as we expect theoretically from adversarial training, we see that the marginal model $q_\psi(\boldsymbol{x})$ aligns with the (true) empirical distribution $p_r(\boldsymbol{x})$, a direct result from the fact that AVO minimizes the Jensen-Shannon divergence $p_r(\boldsymbol{x})$ and $q_\psi(\boldsymbol{x})$. These results highlight that AVO is effective despite the discreteness of the data and the lack of access to the density $p(\boldsymbol{x} \mid \boldsymbol{\vartheta})$ or its gradient.

2.5.2 *Detector calibration in High-Energy Physics*

As a more challenging demonstration, we now turn to a particle physics inference problem. We consider the `pythia` simulator [52] for high-energy particle collisions. In particular, electron-positron collisions at a center-of-mass energy of 91.2 GeV are simulated, in which a Z boson is produced which subsequently decays to quarks. Our particle detector emulates a 32×32 spherical uniform grid in pseudorapidity η and in azimuthal angle ϕ, covering $(\eta,\ \phi)$. The detector itself is parameterized by an offset parameter $\boldsymbol{\vartheta}$ in the z-axis relative to the beam crossing point [53]. An offset of $\boldsymbol{\vartheta} = 0$ implies

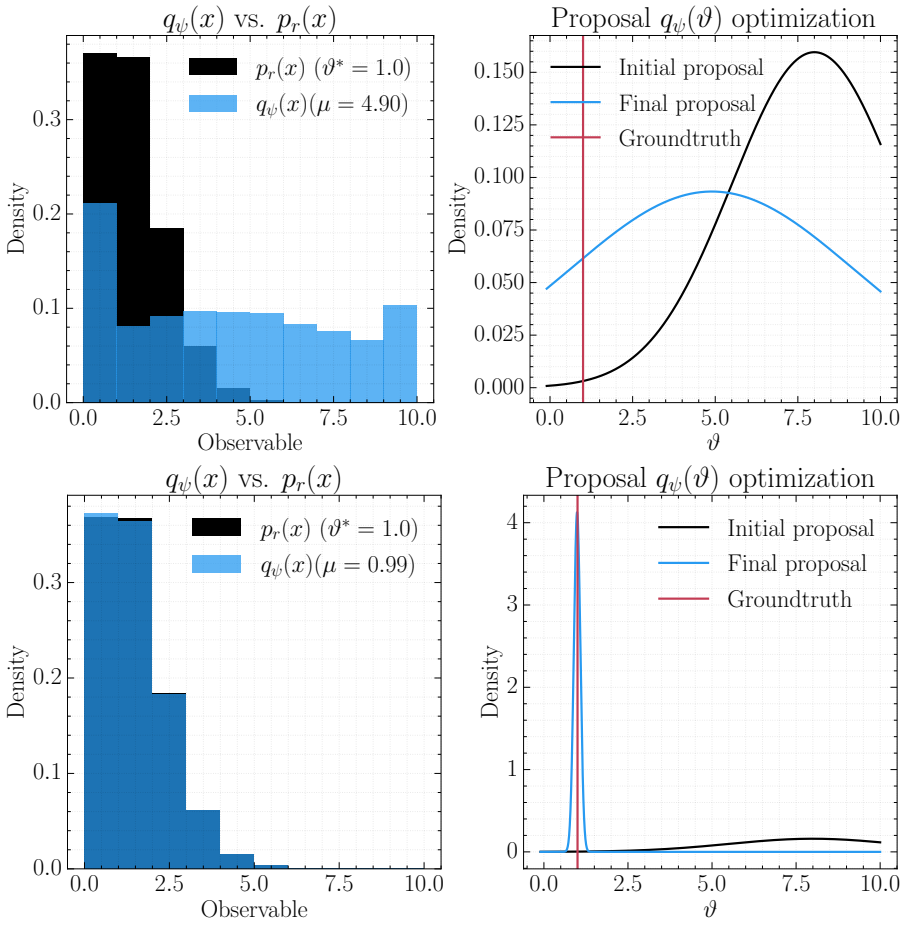

Figure 2.1: Discrete Poisson model with an unkown rate ϑ^*. In this case, the proposal $q_\psi(\vartheta)$ is a normal distribution which is initialized with mean $\mu = 8.0$ and standard-deviation $\sigma = 2.5$. The proposal is therefore completely described by $\psi \triangleq (\mu,\ \sigma)$. *(Top left)* Direct comparison of the $p_r(\boldsymbol{x})$ against the estimated marginal model $q_\psi(\boldsymbol{x})$ after 250 iterations of AVO. *(Top right)* Evolution of the proposal distribution (in blue) with respect to its initial state (in black). It should be noted that at this stage of the optimization procedure, AVO decreases the entropy of the proposal distribution as an initial approach to maximize the likelihood of the parameters given emperical samples of $p_r(\boldsymbol{x})$. *(Bottom left)* Final estimated marginal model $q_\psi(\boldsymbol{x})$ after 1000 iterations of AVO, this figure illustrates that AVO minimizes the Jensen-Shannon divergence between $p_r(\boldsymbol{x})$ and $q_\psi(\boldsymbol{x})$, a result which directly stems from the adversarial training procedure included in AVO. *(Bottom right)* Final optimized proposal, which concentrates its density *tightly* around the true generating parameter ϑ^*. </>

that spherical detector is centered at the collision point, while $\vartheta = 1$ leads to a shift of roughly one pixel.

The inference problem we are concerned with is the estimation of the detector offset parameter $\boldsymbol{\vartheta}$ from a set of 32×32-dimensional observables which represent the pixels in the particle detector. Although the problem at hand is simplified by assuming a simple detector structure, **this task is representative of calibration and alignment tasks, which are critical in experimental particle physics as they have significant impact on the accoracy of the collision reconstruction algorithms.** AVO could therefore provide an *automated* calibration strategy of an experimental setup in High-Energy Physics.

Figure 2.2 shows the average detector response of detector offsets $\boldsymbol{\vartheta} = 0$ and $\boldsymbol{\vartheta} = 1$. The figures highlight the challenging difficulty of the inference problem: the difference between the average detector responses is barely noticeble, even when these response constitue of 10 000 individual $e^- e^+$ collisions. These samples also stress the critical role of a relevant summary statistic on such high-dimensional observables, which is essential to alternative likelihood-free inference methodologies such as *Approximate Bayesian Computation* (ABC), where such summary statistics have to be handcrafted.

For this problem setting, we consider observables drawn from the empirical data distribution $p_r(\boldsymbol{x})$ to be simulated at the nominal parameter $\boldsymbol{\vartheta}^* = 1$, which means particle detector is offset by 1 unit from the collising crossing point of the particles. We initialize the proposal distribution as $\mathcal{N}(\mu = 0.0,\ \sigma = 2.5)$, reflecting our belief in our expectation that the particle detector has been aligned correctly. As before, we use the defaults of AVO as specified in Algorithm 1 and run the optimization procedure for 1000 steps. The results are summarized in Figure 2.3.

2.5.3 *On benchmarking* AVO

In contrast to most related works in the simulation-based inference literature, AVO takes a unique position: whereas most works concern themselves with inferring model parameters tied to a *single* observable x, AVO specializes in dealing with a (large) sample of observables drawn from the data distribution $p_r(\boldsymbol{x})$. For this reason, we cannot consider common likelihood-free benchmarks such as the M/G/1 queueing model, or the Lotka-Volterra population model, which are all defined as inference problems with single observables. **Our proposed method, avo, is less appropriate in these settings** as the discriminator d is not expected to provide a good learning signal for fitting the simulator parameters, because the discriminator will easily overfit and therefore not properly guide $q_\psi(\boldsymbol{\vartheta})$ to a suitable solution. Figure 2.4 illustrates the behaviour of various dataset sizes within the detector calibration problem.

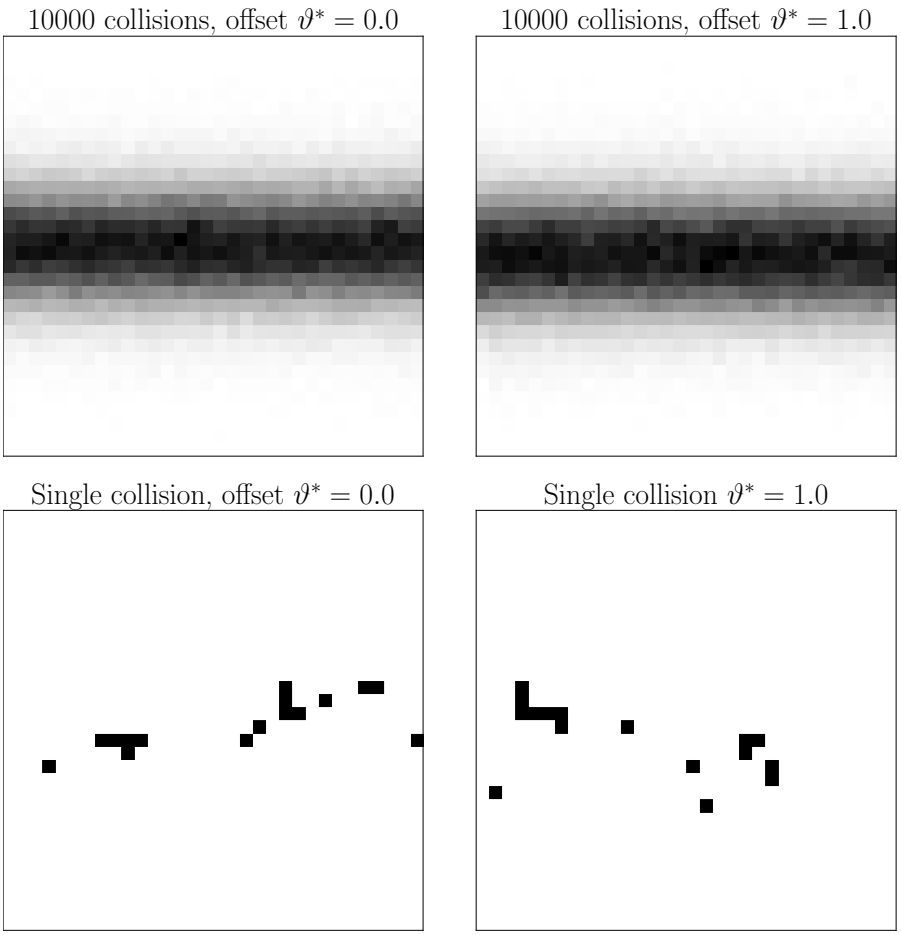

Figure 2.2: Detector responses at various detector offsets. Every pixel in the particle detector "activates" whenever a particle hits the corresponding area, this constitutes a (binary) activation of such a cell and are highlighted in black. The top row shows the average detector response after 10000 particle collisions. A difference between the detector offsets *only* becomes visually apparent after a large number of particle collisions. Detector calibration is therefore associated with a significant computational cost, something we would like to prevent with AVO by directly performing gradient descent on the detector offset to minimize the Jensen-Shannon divergence between the emperical data distribution $p_r(\boldsymbol{x})$ – collected by the real experiment – and the marginal model $q_\psi(\boldsymbol{x})$. The bottom row shows individual particle collisions, demonstrating the complexity and and wide range of variability of the individual particle collisions, emphasizing the difficulty of the problem. </>

2.6 RELATED WORK

This work sits at the intersection of several lines of research related to likelihood-free inference, Approximate Bayesian Computation (ABC),

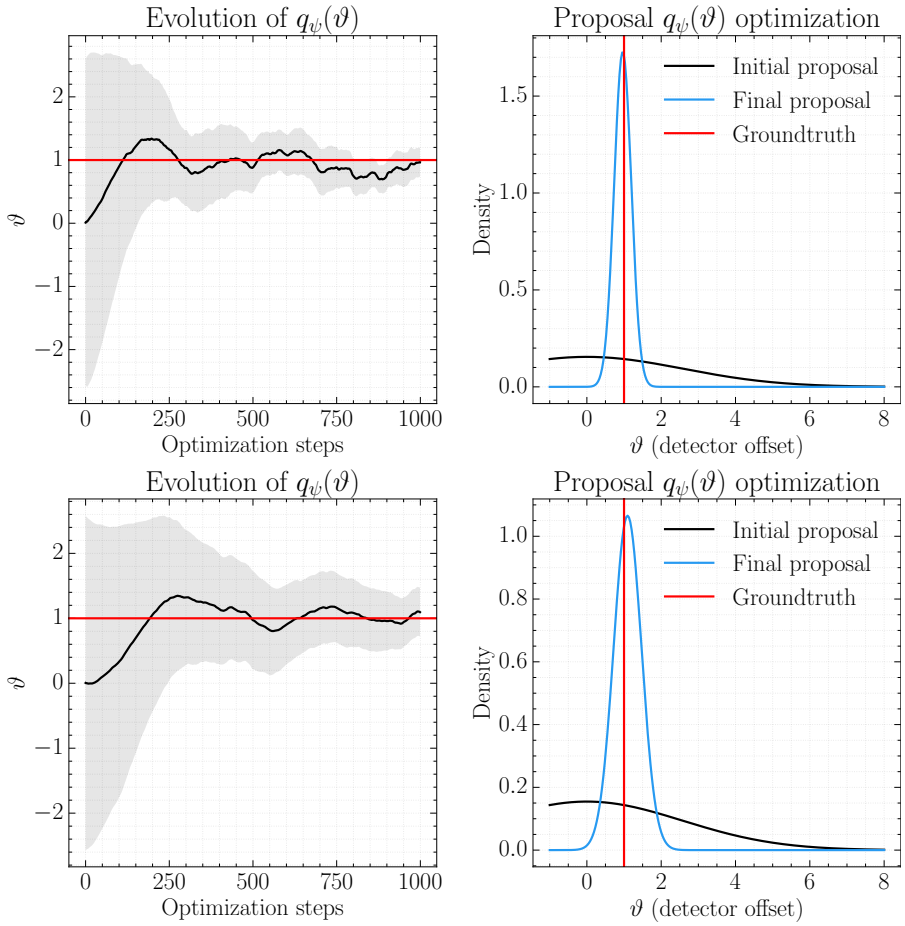

Figure 2.3: Results of the particle detector alignment experiment where the proposal is a normal distribution, where ψ describes its mean (solid black line) and standard deviation (semi-transparent gray area). *(Top)* Experimental setup with a proposal learning rate of 0.01. *(Bottom)* Proposal learning rate of 0.005. </>

implicit generative models, and variational inference. Viewed from the literature around implicit generative models based on neural networks, the proposed method can be considered as a direct adaptation of generative adversarial networks [17] to non-differentiable simulators using variational optimization [32]. From the point of view of likelihood-free inference, where non-differentiable simulators are the norm, our contributions are threefold.

First is the process of lifting the expectation with respect to the non-differentiable simulator $\mathbb{E}_{p(\boldsymbol{x}\mid\boldsymbol{\vartheta})}$ to a differentiable expectation with respect to the variational program $\mathbb{E}_{q_\psi(\boldsymbol{x})}$. Secondly, is the introduction of a novel form of variational inference that works in a likelihood-free setting. Thirdly, AVO can be viewed as a form of Empirical Bayes where the prior is optimized based on the data.

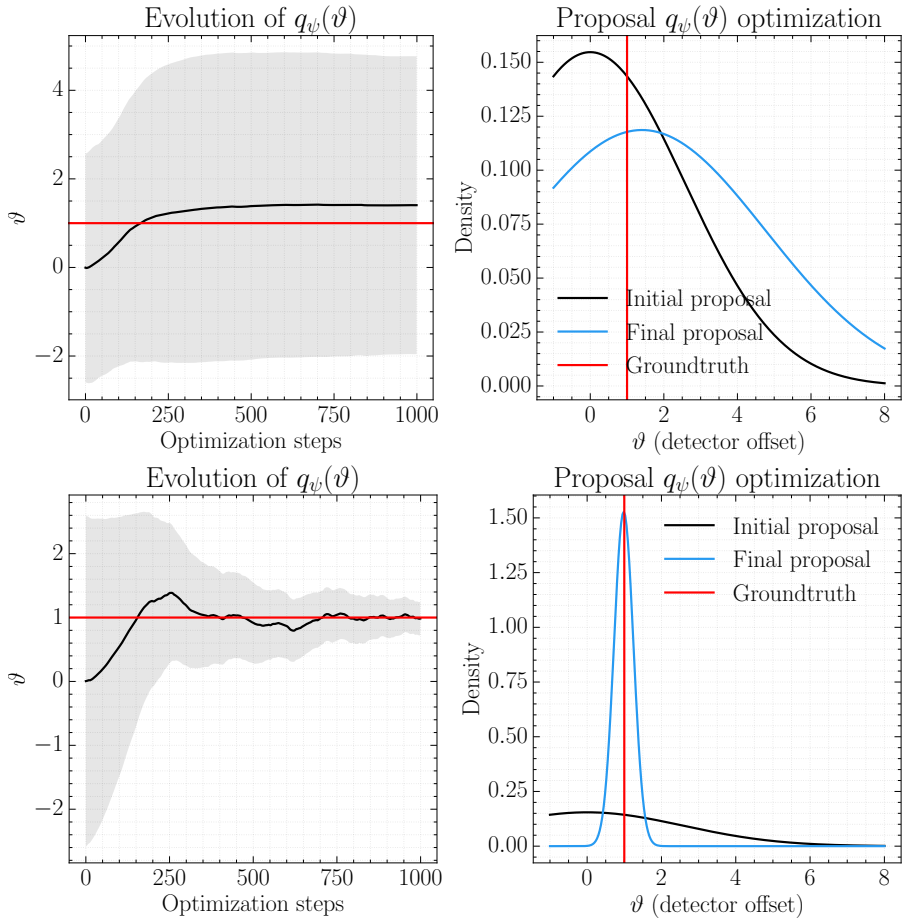

Figure 2.4: Demonstration that AVO requires a sufficiently large dataset to provide a good learning signal to the discriminator in order to fit the simulator parameters. *(Top)* Dataset size of 100. Although overfitting might play a very important role here, we selected the proposal parameterization for which the discriminator validation loss was minimized. It should be noted that in this regime AVO actually increased its uncertainty about the detector offset parameter ϑ to minimize the Jensen-Shannen divergence. In addition, this highlights the difficulty of the - although simplified - task based on individual particle collisions. *(Bottom)* Dataset size of 10 000. </>

As for many likelihood-free inference algirhtms, AVO is tied to a class of algorithms which can be framed as density estimation-by-comparison, as reviewed in Mohamed and Lakshminarayanan [19]. In most case, these inference algorithms are formulated as an iterative two-step process where the model distribution is first compared to the true data distribution and subseuquently updated to make said distribution more comparable to the latter. Relevant work in this direction includes those that rely on a classifier to estimate the discrepancy between the observed data and the model distributions [54–59].

Of direct relevance to the likelihood-free setup, Hamiltonian ABC estimates gradients with respect to $\boldsymbol{\vartheta}$ through finite difference from multiple forward passes of the simulation with variance reduction strategies based on controlling the source of randomness used for the latent variable $\boldsymbol{\vartheta}$. Sharing similar foundational principles as AVO, the SPIRAL [60] makes use of the Wasserstein GAN [27] objective and variants of REINFORCE gradient estimates to adversarially train an agent to synthesize programs controlling a non-differentiable graphics engine to reconstruct target images, or perform unconditional generation.

Likewise, AVO closely relates to recent extensions of GANs, such as ALI [61], adversarial feature learning (BIGAN) [62], α-GAN [63], AVB [64], and the `PC-Adv` algorithm [65], which add an inference network to the generative model. Each of these assume a tractable density $p(\boldsymbol{x} \mid \boldsymbol{\vartheta})$ that is differentiable with respect to $\boldsymbol{\vartheta}$, which is per-definition impossible in the likelihood-free setting.

In AVO, lifing the likelihood model of the non-differentiable simulator $p(\boldsymbol{x} \mid \boldsymbol{\vartheta})$ to the variational program $q_\psi(\boldsymbol{x})$ provides the ability to differentiate expectations with respect to ψ. However, the *marginal* model density $q_\psi(\boldsymbol{x})$ is still intractable. Moreover, we do not attempt to define a recognition model $q_\psi(z, \boldsymbol{\vartheta})$ as the latent space of many real-world simulators is highly complex and not amenable to a neural recognition model. Although recently Baydin et al. [66] has succesfully approached this problem in a High-Energy Physics setting.

This work has also many connections to work on variational inference, in which the goal is to optimize the recognition model $q_\psi(z, \boldsymbol{\vartheta})$ so that it is close the the true posterior $p(z, \boldsymbol{\vartheta} \mid x)$. There have been efforts to extend variational inference to intractable likelihoods; however, many require restrictive assumptions. In [67], the authors consider Variational Bayes with an Intractable Likelihood (VBIL). In that approach "the only requirement is that the intractable likelihood can be estimated unbiasedly." In the case of simulators, they propose to use ABC (Approximate Bayesian Computation) approximated-likelihood with an ϵ-kernel. This likelihood **is only unbiased as $\epsilon \to 0$ and the summary statistic used to reduce the dimensionality of the observable is sufficient**, thus this method inherits the drawbacks of the ABC including the choice of summary statistics and the inefficiency in evaluating the ABC likelihood for high-dimensional data and small ϵ. More recently, Tran, Ranganath, and Blei [68] adapted variational inference to hierarchical implicit models defined on simulators. In this work, the authors step around the intractable likelihoods by reformulating the optimization of the Evidence Lower Bound (ELBO) in terms of a neural and differentiable approximation r of the log-likelihood ratio $\log \frac{p}{q}$, thereby effectively using the same core principle as used in GANs

[19]. With a similar objective, McCarthy, Rodriguez, and Minchole [69] adapt variational inference to a non-differentiable cardiac simulator by maximizing the ELBO using Bayesian optimization, hence bypassing altogether the need for gradient estimates.

2.7 SUMMARY & DISCUSSION

In this work, se developed a likelihood-free inference methodology for non-differentiable implicit generative models. The algorithm combines ideas from adversarial training and variational optimization to minimize variational upper bounds on otherwise non-differentiable adversarial objectives. AVO enables Emperical Bayes through variational inference in the likelihood-free setting and does not incur the inefficiences of an ABC-like rejection sampler nor the disadvantages of likelihood-free inference algorithms which rely on ad-hoc hyperparamters and *handcrafted* summary statistics, which can potentially lead to biased inference results due to the (possibly naive) human element. AVO does not suffer from these limitations, as the discriminator automatically learns an internal representation directly from the presented data. Whenever the simulation model is well-specified, AVO provides point estimates for the generative model, which asymptotically corresponds to the data generating parameters.

We expect AVO to shine in inference settings where a large amount of observables are available, as the discriminator will have a proper learning signal because of the sufficiently large dataset size. This is typically the case in population studies, where a large set of observables are jointly being examined. After having learned some proposal $q_\psi(\boldsymbol{\vartheta})$ with AVO over the model parameters of interest, $q_\psi(\boldsymbol{\vartheta})$ could in fact serve as a prior for more in-depth analyses of a specific observable. Some caution needs to be applied however, as we do not guarantee that the prior is conservative, i.e., it *never* excludes viable solutions with respect to the assumed simulation model. To combat this from a practical point of view, we suggest to tune the entropic regularizer to force the proposal distribution to be more uncertain, and run AVO multiple times with distinct initial parameterizations ψ to prevent the optimization procedure being stuck in local minima and limitations that arise due to the selected family of the proposal distribution.

3

Approximating Posteriors with Amortized Approximate Ratio Estimators

The contents of this chapter are based on Hermans, Begy, and Louppe [10].

This chapter presents an approach to address the intractability of the likelihood and the marginal model in a Bayesian analysis concerned with approximating a posterior given a *single* observable. This is achieved by learning a flexible estimator which directly approximates the likelihood-to-evidence ratio for any observable supported by the marginal model.

The resulting *amortized* ratio estimator is subsequently embedded in Markov chain Monte Carlo samplers such as Metropolis-Hastings and Hamiltonian Monte Carlo to approximate the likelihood-ratio between consecutive states in the Markov chain, allowing us to draw samples from the unknown intractable posterior. Techniques are presented to improve the numerical stability. We demonstrate our approach on a variety of benchmarks and compare against well-established approximate inference techniques. Scientific applications in high energy and astrophysics with high-dimensional observations demonstrate the applicability of the presented methodology.

3.1 INTRODUCTION

In a Bayesian analysis, domain scientists are interested in the posterior

$$p(\boldsymbol{\vartheta}|\,\boldsymbol{x}) = \frac{p(\boldsymbol{\vartheta})p(\boldsymbol{x}\,|\boldsymbol{\vartheta})}{p(\boldsymbol{x})}, \tag{3.1}$$

which relates the parameters $\boldsymbol{\vartheta}$ of a model or theory under some prior $p(\boldsymbol{\vartheta})$ to an observation x. Although Bayesian inference is natural for

such settings, the implied computation is generally not. Often the marginal model

$$p(\boldsymbol{x}) = \int p(\boldsymbol{\vartheta}) p(\boldsymbol{x} \,|\, \boldsymbol{\vartheta}) \,\mathrm{d}\, \boldsymbol{\vartheta} \tag{3.2}$$

cannot be evaluated directly due to the associated computational cost or the absence of an analytical expression. Thereby making posterior inference by direct evaluation of Bayes' rule impractical. Methods such as Markov chain Monte Carlo (MCMC), as described by Metropolis et al. [70] and Hastings [71], bypass the dependency and evaluation of the marginal model by evaluating some form of the likelihood ratio between consecutive states in the Markov chain. This allows the posterior to be approximated numerically, provided the likelihood $p(\boldsymbol{x} \,|\, \boldsymbol{\vartheta})$ and the prior $p(\boldsymbol{\vartheta})$ are tractable.

This chapter considers an equally common and more challenging setting in which the likelihood cannot be evaluated in a reasonable amount of time, or has no closed-form expression (intractable). However, drawing samples from the forward or simulation model remains possible. The prevalence of this problem gave rise to a large body of research typically referred to as "simulation-based" or "likelihood-free" inference.

This chapter introduces an approach to perform likelihood-free posterior inference and a technique to draw samples from the approximated posteriors using MCMC. Our method relies on an amortized ratio estimator that can be trained on presimulated samples from the joint $p(\boldsymbol{\vartheta}, \boldsymbol{x})$ to approximate the likelihood-to-evidence ratio

$$r(\boldsymbol{x} \,|\, \boldsymbol{\vartheta}) \triangleq \frac{p(\boldsymbol{x} \,|\, \boldsymbol{\vartheta})}{p(\boldsymbol{x})}, \tag{3.3}$$

and subsequently the posterior

$$p(\boldsymbol{\vartheta} \,|\, \boldsymbol{x}) = p(\boldsymbol{\vartheta}) r(\boldsymbol{x} \,|\, \boldsymbol{\vartheta}). \tag{3.4}$$

In addition, the amortized ratio estimator can be used to compute the acceptance probability in Metropolis-Hastings [70, 71]. Whenever the ratio estimator is differentiable – typically the case whenever the estimator is a neural network – we derive the score $\nabla_{\boldsymbol{\vartheta}} \log p(\boldsymbol{x} \,|\, \boldsymbol{\vartheta})$, making the proposed method also applicable to Hamiltonian Monte Carlo [72, 73].

3.2 BACKGROUND

3.2.1 *Markov chain Monte Carlo*

In Bayesian analyses, Markov Chain Monte Carlo (MCMC) methods are generally applied to *(i)* draw samples from a posterior probability

distribution $p(\boldsymbol{\vartheta}|\boldsymbol{x})$ with an intractable marginal model, or *(ii)* to estimate expectations under various functions $f(\boldsymbol{\vartheta})$ of the form

$$\mathbb{E}_{p(\boldsymbol{\vartheta}|\boldsymbol{x})}[f(\boldsymbol{\vartheta})] = \int_{\boldsymbol{\vartheta}} \mathrm{d}\,\boldsymbol{\vartheta}\; p(\boldsymbol{\vartheta}|\boldsymbol{x}) f(\boldsymbol{\vartheta}), \tag{3.5}$$

for which point-wise evaluations of the likelihood are possible [70, 71, 74].

These two problems are useful for approximating the target distribution $p(\boldsymbol{\vartheta}|\boldsymbol{x})$ itself, because posterior samples can be drawn from the posterior $p(\boldsymbol{\vartheta}|\boldsymbol{x})$ by collecting a set of dependent states $\boldsymbol{\vartheta}_{0:T}$ from a Markov chain. Although the mechanism for transitioning from $\boldsymbol{\vartheta}_t$ to the next state $\boldsymbol{\vartheta}'$ depends on the algorithm at hand, the acceptance of a transition $\boldsymbol{\vartheta}_t \to \boldsymbol{\vartheta}'$ for $\boldsymbol{\vartheta}'$ sampled from a proposal mechanism $q(\boldsymbol{\vartheta}'|\boldsymbol{\vartheta}_t)$ is usually determined by evaluating some form of the posterior ratio

$$\frac{p(\boldsymbol{\vartheta}'|x)}{p(\boldsymbol{\vartheta}_t|x)} = \frac{\frac{p(\boldsymbol{\vartheta}')p(\boldsymbol{x}|\boldsymbol{\vartheta}')}{p(\boldsymbol{x})}}{\frac{p(\boldsymbol{\vartheta}_t)p(\boldsymbol{x}|\boldsymbol{\vartheta}_t)}{p(\boldsymbol{x})}} = \frac{p(\boldsymbol{\vartheta}')p(\boldsymbol{x}|\boldsymbol{\vartheta}')}{p(\boldsymbol{\vartheta}_t)p(\boldsymbol{x}|\boldsymbol{\vartheta}_t)}. \tag{3.6}$$

From this formulation it is directly evident that *(i)* the normalizing constant $p(\boldsymbol{x})$ cancels out within the ratio, thereby bypassing the need for its intractable evaluation, and *(ii)* how necessary the likelihood ratio is in assessing the quality of a candidate state $\boldsymbol{\vartheta}'$ against the current state $\boldsymbol{\vartheta}_t$ of the Markov chain.

3.2.1.1 *Metropolis-Hastings Markov Chain Monte Carlo*

The Metropolis-Hastings (MH) [70, 71] Markov Chain Monte Carlo sampler is a straightforward implementation of Equation 3.6. The proposal mechanism $q(\boldsymbol{\vartheta}'|\boldsymbol{\vartheta}_t)$ is typically a tractable distribution, whose conditional likelihood can be evaluated efficiently and for which it is easy to draw samples from. For a given state $\boldsymbol{\vartheta}_t$ in the Markov chain, these components are subsequently combined to generate a proposal sample

$$\boldsymbol{\vartheta}' \sim q(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}_t), \tag{3.7}$$

and to compute the acceptance probability ρ of a transition $\boldsymbol{\vartheta}_t \to \boldsymbol{\vartheta}'$:

$$\rho = \min\left(1,\; \frac{p(\boldsymbol{\vartheta}')p(\boldsymbol{x}|\boldsymbol{\vartheta}')}{p(\boldsymbol{\vartheta}_t)p(\boldsymbol{x}|\boldsymbol{\vartheta}_t)} \frac{q(\boldsymbol{\vartheta}'|\boldsymbol{\vartheta}_t)}{q(\boldsymbol{\vartheta}_t|\boldsymbol{\vartheta}')}\right). \tag{3.8}$$

The choice of an appropriate transition distribution is important to maximize the effective sample size (sampling efficiency) and reduce the autocorrelation.

On transition distributions and their parameterizations

Consider the case where the proposal or transition distribution is a normal distribution with a relatively small standard deviation and whose mean is parameterized by $\boldsymbol{\vartheta}_t$. That is $q(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}_t) \triangleq \mathcal{N}(\boldsymbol{\vartheta}_t; \sigma)$. The sample efficiency of the resulting Markov chain will be very low due to the large autocorrelation between the *accepted* samples, driven by the small standard deviation of the proposal distribution. To reduce the autocorrelation, one can subsample accepted states from the chain and thereby increase its sample efficiency. However, this comes at a cost because subsampling reduces the number of samples in the Markov chain. To retain a certain *effective* sample size, one can simply increase the number of sampling steps, or tune the parameterization of the proposal distribution to reduce the autocorrelation.

In addition, for a single Markov chain with for the same proposal distribution it would be impossible to properly approximate a bimodal posterior whose modes are clearly seperated in the parameter space, since the proposal distribution will have an extremely low probability of generating a proposal state in the other mode of the posterior. As before, this can be addressed by tuning the parameterization of the transition distribution. As before, autocorrelation needs to be monitored. Note that in this specific instance, a large standard deviation negatively affects the autocorrelation as well due to rejected Markov chain transitions. **A certain balance concerning the hyperparameters of the transition distribution is therefore required**.

Other approaches are possible however, one such instance is the usage of *ensembles*: n walkers or Markov chains are evolved simultaneously [75], such that the proposal of a single walker is defined by the *current state* of the remaining $n - 1$ Markov chains. This approach reduces the autocorrelation since it draws proposal states from a *symmetric* proposal distribution which is implicitely defined through an ensemble of Markov chains covering distinct parts of the parameter space.

Algorithm 2 summarizes the implementation of a Metropolis-Hastings sampler in a Bayesian context for completeness.

3.2.1.2 *Hamiltonian Markov Chain Monte Carlo*

Hamiltonian Monte Carlo (HMC) [72, 73, 76] improves upon the sampling efficiency of Metropolis-Hastings by reducing the autocorrelation of the Markov chain. Improving the sampling efficiency is especially useful in scenarios where designing a proper transition distribution is

Algorithm 2 Metropolis-Hastings Markov Chain Monte Carlo

Inputs: Initial sample $\boldsymbol{\vartheta}_0$
Prior $p(\boldsymbol{\vartheta})$
Likelihood $p(\boldsymbol{x}\,|\boldsymbol{\vartheta})$
Transition distribution $q(\boldsymbol{\vartheta})$
Observable x
Outputs: Markov chain $\boldsymbol{\vartheta}_{0:T}$
Hyperparameters: Markov chain transitions (steps) T

$t \leftarrow 0$
$\boldsymbol{\vartheta}_t \leftarrow \boldsymbol{\vartheta}_0$
for $t < T$ **do**
  $\boldsymbol{\vartheta}' \sim q(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}_t)$
  $\rho \leftarrow \min\left(1,\ \frac{p(\boldsymbol{\vartheta}')p(\boldsymbol{x}\,|\boldsymbol{\vartheta}')}{p(\boldsymbol{\vartheta}_t)p(\boldsymbol{x}\,|\boldsymbol{\vartheta}_t)}\frac{q(\boldsymbol{\vartheta}'|\boldsymbol{\vartheta}_t)}{q(\boldsymbol{\vartheta}_t|\boldsymbol{\vartheta}')}\right)$
  $\boldsymbol{\vartheta}_{t+1} \leftarrow \begin{cases} \boldsymbol{\vartheta}' & \text{with probability } \rho \\ \boldsymbol{\vartheta}_t & \text{with probability } 1-\rho \end{cases}$
  $t \leftarrow t+1$
end for
return $\boldsymbol{\vartheta}_{0:T}$

difficult. This is achieved by modeling the target density $p(\boldsymbol{\vartheta}|\,\boldsymbol{x})$ as a potential energy function

$$U(\boldsymbol{\vartheta}) \triangleq -\log p(\boldsymbol{\vartheta}|\,\boldsymbol{x}), \tag{3.9}$$
$$\triangleq -\log p(\boldsymbol{\vartheta}) - \log p(\boldsymbol{x}\,|\boldsymbol{\vartheta}) + \log p(\boldsymbol{x}), \tag{3.10}$$

and attributing some kinetic energy,

$$K(m) \triangleq \frac{1}{2}m^2 \tag{3.11}$$

with momentum $m \sim p(m)$ to the current state $\boldsymbol{\vartheta}_t$.

A new state $\boldsymbol{\vartheta}'$ can subsequently be proposed by simulating the Hamiltonian dynamics of $\boldsymbol{\vartheta}_t$. This is achieved by leapfrog integration of $\nabla_{\boldsymbol{\vartheta}}\, U(\boldsymbol{\vartheta})$ over a fixed number of steps with initial momentum m. Afterwards, the acceptance ratio

$$\rho \triangleq \min\left(1,\ \exp\left(U(\boldsymbol{\vartheta}') - U(\boldsymbol{\vartheta}_t) + K(m') - K(m)\right)\right) \tag{3.12}$$

is computed to assess the quality of the candidate $\boldsymbol{\vartheta}'$. Note that in our earlier definition, $U(\boldsymbol{\vartheta})$ describes the energy potential of the target density. Given the Bayesian lens of this chapter, this target density is a posterior with an intractable marginal model. Like Metropolis-Hastings however, the energy potential only requires the prior and the likelihood

$$U(\boldsymbol{\vartheta}) \triangleq -\log p(\boldsymbol{\vartheta}) - \log p(\boldsymbol{x}\,|\boldsymbol{\vartheta}), \tag{3.13}$$

because the evidence term $-\log p(\boldsymbol{x})$ cancels out in Equation 3.12. Algorithm 3 summarizes the complete procedure

While the simulation of the Hamiltonian dynamics should in principle concentrate the available computational resources on the typical set [76] – the region which only contributes significantly to the approximation of the posterior or some expectation – there are various practical issues. First and foremost is the *critical* hyperparameterization of the HMC sampler. It requires a properly specified momentum distribution $q(m)$ to efficiently let the Hamiltonian dynamics postively affect the autocorrelation [76]. Another crucial hyperparameter is the number of leapfrog integration steps and stepsize. Both are able to significantly affect performance of the sampler, and the sample efficiency of the resulting Markov chain as well. To that end, the *No-U-Turn Sampler* [77] (NUTS) has been proposed as an extension to HMC. The sampler automatically adjusts the number of leapfrog-integration steps, and *emperically performs at least as efficiently compared to a* ***well-tuned*** HMC application [77] without requiring corrections from the end-user.

3.2.2 *Approximating likelihood ratios*

The most powerful test-statistic to compare two hypotheses $\boldsymbol{\vartheta}_0$ and $\boldsymbol{\vartheta}_1$ for an observation x is the likelihood ratio [78]

$$r(\boldsymbol{x}\,|\boldsymbol{\vartheta}_0, \boldsymbol{\vartheta}_1) \triangleq \frac{p(\boldsymbol{x}\,|\boldsymbol{\vartheta}_0)}{p(\boldsymbol{x}\,|\boldsymbol{\vartheta}_1)}. \tag{3.14}$$

Previous work [55] shows that it is possible to express the test-statistic through a change of variables $(\dot{x})\colon \mathbb{R}^m \mapsto [0,1]$, where m is the dimensionality of the observable x.

This observation can be used in a supervised setting to train a classifier $d(x)$ to distinguish samples $x \sim p(\boldsymbol{x}\,|\boldsymbol{\vartheta}_0)$ with class label $y = 1$ from $x \sim p(\boldsymbol{x}\,|\boldsymbol{\vartheta}_1)$ labeled $y = 0$. In this case, the decision function modeled by the optimal classifier [55] or discriminator $d(x)$ is

$$d(x) = p(y = 1|\,\boldsymbol{x}) = \frac{p(\boldsymbol{x}\,|\boldsymbol{\vartheta}_0)}{p(\boldsymbol{x}\,|\boldsymbol{\vartheta}_0) + p(\boldsymbol{x}\,|\boldsymbol{\vartheta}_1)}, \tag{3.15}$$

thereby obtaining the likelihood ratio

$$r(\boldsymbol{x}\,|\boldsymbol{\vartheta}_0, \boldsymbol{\vartheta}_1) = \frac{d(x)}{1 - d(x)}. \tag{3.16}$$

This approach of density ratio estimation by classification, also known as the "likelihood ratio trick" (LRT), is well-established in the literature [19, 55, 57, 68, 79, 80], especially in the area of Generative Adversarial Networks (GANs) [81–84] and variational inference [85].

Algorithm 3 Hamiltonian Markov Chain Monte Carlo

Inputs: Initial parameter $\boldsymbol{\vartheta}_0$
Differentiable prior $p(\boldsymbol{\vartheta})$
Momentum distribution $q(m)$
Differentiable likelihood $p(\boldsymbol{x}\,|\boldsymbol{\vartheta})$
Observable x
Outputs: Markov chain $\boldsymbol{\vartheta}_{0:T}$
Hyperparameters: Steps T.
Leapfrog-integration steps l and stepsize η.

$t \leftarrow 0$
$\boldsymbol{\vartheta}_t \leftarrow \boldsymbol{\vartheta}_0$
for $t < T$ **do**
  $m_t \sim q(m)$
  $k \leftarrow 0$
  $m_k \leftarrow m_t$
  $\boldsymbol{\vartheta}_k \leftarrow \boldsymbol{\vartheta}_t$
  $U(\boldsymbol{\vartheta}_t) \leftarrow -\log p(\boldsymbol{\vartheta}_t) - \log p(\boldsymbol{x}\,|\boldsymbol{\vartheta}_t)$
  for $k < l$ **do**
    $m_k \leftarrow m_k - \frac{\eta}{2}\nabla_{\boldsymbol{\vartheta}}\, U(\boldsymbol{\vartheta}_k)$
    $\boldsymbol{\vartheta}_k \leftarrow \boldsymbol{\vartheta}_k + \eta \cdot m_k$
    $U(\boldsymbol{\vartheta}_k) \leftarrow -\log p(\boldsymbol{\vartheta}_k) - \log p(\boldsymbol{x}\,|\boldsymbol{\vartheta}_k)$
    $m_k \leftarrow m_k + \frac{\eta}{2}\nabla_{\boldsymbol{\vartheta}}\, U(\boldsymbol{\vartheta}_k)$
    $k \leftarrow k + 1$
  end for
  $\rho \leftarrow \min\left[1,\ \frac{\exp\left(U(\boldsymbol{\vartheta}_k) + K(m_k)\right)}{\exp\left(U(\boldsymbol{\vartheta}_t) + K(m_t)\right)}\right]$
  $\boldsymbol{\vartheta}_{t+1} \leftarrow \begin{cases} \boldsymbol{\vartheta}_k & \text{with probability } \rho \\ \boldsymbol{\vartheta}_t & \text{with probability } 1-\rho \end{cases}$
  $t \leftarrow t + 1$
end for
return $\boldsymbol{\vartheta}_{0:T}$

Because domain scientists are often interested in computing the likelihood ratio between arbitrary hypotheses, training $d(x)$ for every possible pair of hypotheses becomes impractical. A solution proposed by [55, 86] is to, in addition to x, parameterize the classifier d with $\boldsymbol{\vartheta}$ and train $d(\boldsymbol{\vartheta}, x)$ to distinguish between samples from $p(\boldsymbol{x}\,|\boldsymbol{\vartheta})$ and samples from a mathematically arbitrary (but fixed) reference density $p(\boldsymbol{x}\,|\boldsymbol{\vartheta}_{\text{ref}})$ as described above. In this case, the decision modeled by the *optimal* classifier [55] is

$$d(\boldsymbol{\vartheta}, x) = \frac{p(\boldsymbol{x}\,|\boldsymbol{\vartheta})}{p(\boldsymbol{x}\,|\boldsymbol{\vartheta}) + p(\boldsymbol{x}\,|\boldsymbol{\vartheta}_{\text{ref}})}, \tag{3.17}$$

thereby defining the likelihood-to-reference ratio

$$r(\boldsymbol{x}\,|\boldsymbol{\vartheta}) \triangleq r(\boldsymbol{x}\,|\boldsymbol{\vartheta}, \boldsymbol{\vartheta}_{\text{ref}}) = \frac{d(\boldsymbol{\vartheta}, \boldsymbol{x})}{1 - d(\boldsymbol{\vartheta}, \boldsymbol{x})}. \tag{3.18}$$

The likelihood ratio between arbitrary hypotheses $\boldsymbol{\vartheta}_0$ and $\boldsymbol{\vartheta}_1$ of choice can subsequently be computed through the *amortized* ratio estimator $r(\boldsymbol{x}\,|\boldsymbol{\vartheta}_0)$ as

$$r(\boldsymbol{x}\,|\boldsymbol{\vartheta}_0, \boldsymbol{\vartheta}_1) = \frac{r(\boldsymbol{x}\,|\,\boldsymbol{\vartheta}_0)}{r(\boldsymbol{x}\,|\,\boldsymbol{\vartheta}_1)}. \tag{3.19}$$

3.3 METHOD

We now introduce a new method that is able to

1. draw samples from a posterior with an intractable likelihood and marginal model,
2. and directy evaluate estimates of the posterior density function,

and does so for arbitrary observables $\boldsymbol{x} \sim p(\boldsymbol{x})$ without retraining.

3.3.1 *Drawing samples from an intractable posterior without a likelihood*

As noted above, MCMC samplers rely on the likelihood ratio to compute the acceptance probability between consequative states in the Markov chain. We propose to remove the dependency on the intractable likelihoods $p(\boldsymbol{x}\,|\boldsymbol{\vartheta}')$ and $p(\boldsymbol{x}\,|\boldsymbol{\vartheta}_t)$ by directly modeling their ratio using an amortized ratio *estimator*

$$\hat{r}(x|\boldsymbol{\vartheta}', \boldsymbol{\vartheta}_t) = \frac{\hat{r}(x|\boldsymbol{\vartheta}')}{\hat{r}(x|\boldsymbol{\vartheta}_t)}. \tag{3.20}$$

We call this method amortized approximate likelihood ratio MCMC (AALR-MCMC).

3.3.1.1 *Likelihood-free Metropolis-Hastings*

Adapting Metropolis-Hastings to the likelihood-free setup with likelihood-ratio estimators is achieved by replacing the computation of the intractable acceptance probability in Equation 3.8 with

$$\rho = \min\left(1, \frac{p(\boldsymbol{\vartheta}')\hat{r}(x|\boldsymbol{\vartheta}')}{p(\boldsymbol{\vartheta}_t)\hat{r}(x|\boldsymbol{\vartheta}_t)} \frac{q(\boldsymbol{\vartheta}'|\boldsymbol{\vartheta}_t)}{q(\boldsymbol{\vartheta}_t|\boldsymbol{\vartheta}')}\right). \tag{3.21}$$

The algorithm remains otherwise unchanged. The full procedure is summarized in Algorithm 4.

Algorithm 4 Likelihood-free Metropolis-Hastings MCMC

Inputs: Initial sample $\boldsymbol{\vartheta}_0$
Prior $p(\boldsymbol{\vartheta})$
Likelihood $p(\boldsymbol{x}\,|\boldsymbol{\vartheta})$
Ratio estimator $\hat{r}(x|\boldsymbol{\vartheta})$
Transition distribution $q(\boldsymbol{\vartheta})$
Observable x
Outputs: Markov chain $\boldsymbol{\vartheta}_{0:T}$
Hyperparameters: Markov chain transitions (steps) T

$t \leftarrow 0$
$\boldsymbol{\vartheta}_t \leftarrow \boldsymbol{\vartheta}_0$
for $t < T$ **do**
  $\boldsymbol{\vartheta}' \sim q(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}_t)$
  $\rho = \min\left(1, \frac{p(\boldsymbol{\vartheta}')\hat{r}(x|\boldsymbol{\vartheta}')}{p(\boldsymbol{\vartheta}_t)\hat{r}(x|\boldsymbol{\vartheta}_t)} \frac{q(\boldsymbol{\vartheta}'|\boldsymbol{\vartheta}_t)}{q(\boldsymbol{\vartheta}_t|\boldsymbol{\vartheta}')}\right)$
  $\boldsymbol{\vartheta}_{t+1} \leftarrow \begin{cases} \boldsymbol{\vartheta}' & \text{with probability } \rho \\ \boldsymbol{\vartheta}_t & \text{with probability } 1-\rho \end{cases}$
  $t \leftarrow t+1$
end for
return $\boldsymbol{\vartheta}_{0:T}$

3.3.1.2 *Likelihood-free Hamiltonian Monte Carlo*

The first step in making HMC likelihood-free, is by showing that $U(\boldsymbol{\vartheta}_t) - U(\boldsymbol{\vartheta}')$ primarily reduces to the log-likelihood ratio,

$$\begin{aligned} U(\boldsymbol{\vartheta}_t) - U(\boldsymbol{\vartheta}') &= \log \frac{p(\boldsymbol{x}\,|\boldsymbol{\vartheta}')}{p(\boldsymbol{x}\,|\boldsymbol{\vartheta}_t)} + \log \frac{p(\boldsymbol{\vartheta}')}{p(\boldsymbol{\vartheta}_t)}, \\ &= \log r(\boldsymbol{x}\,|\boldsymbol{\vartheta}', \boldsymbol{\vartheta}_t) + \log \frac{p(\boldsymbol{\vartheta}')}{p(\boldsymbol{\vartheta}_t)}. \end{aligned} \tag{3.22}$$

Furthermore, to simulate the Hamiltonian dynamics of $\boldsymbol{\vartheta}_t$, we require a likelihood-free definition of $\nabla_{\boldsymbol{\vartheta}} U(\boldsymbol{\vartheta})$. Within our framework, $\nabla_{\boldsymbol{\vartheta}} U(\boldsymbol{\vartheta})$ can be expressed as

$$\nabla_{\boldsymbol{\vartheta}} U(\boldsymbol{\vartheta}) = -\frac{\nabla_{\boldsymbol{\vartheta}} r(\boldsymbol{x}|\boldsymbol{\vartheta})}{r(\boldsymbol{x}|\boldsymbol{\vartheta})} - \nabla_{\boldsymbol{\vartheta}} p(\boldsymbol{\vartheta}). \tag{3.23}$$

This form can be recovered by a differentiable ratio estimator $r(\boldsymbol{x}|\boldsymbol{\vartheta})$, as expanding $r(\boldsymbol{x}|\boldsymbol{\vartheta})$ in Equation 3.23 yields

$$-\frac{\nabla_{\boldsymbol{\vartheta}} p(\boldsymbol{x}|\boldsymbol{\vartheta})}{p(\boldsymbol{x}|\boldsymbol{\vartheta})} = -\frac{\nabla_{\boldsymbol{\vartheta}} r(\boldsymbol{x}|\boldsymbol{\vartheta})}{r(\boldsymbol{x}|\boldsymbol{\vartheta})} = -\nabla_{\boldsymbol{\vartheta}} \log p(\boldsymbol{x}|\boldsymbol{\vartheta}). \tag{3.24}$$

Having likelihood-free alternatives for $U(\boldsymbol{\vartheta}) - U(\boldsymbol{\vartheta}')$ and $\nabla_{\boldsymbol{\vartheta}} U(\boldsymbol{\vartheta})$, we can replace these components in HMC to obtain a likelihood-free HMC sampler. This procedure is summarized in Algorithm 5. While likelihood-free HMC does not rely on the intractable likelihood, it still depends on the computation of $\nabla_{\boldsymbol{\vartheta}} \hat{r}(x|\boldsymbol{\vartheta})$ to recover $\nabla_{\boldsymbol{\vartheta}} U(\boldsymbol{\vartheta})$. This can be a costly operation depending on the architecture of the ratio estimator. Similar to HMC, the sampler requires careful tuning to maximize the sampling efficiency. Ideas such as neural proposals [87], or a likelihood-free NUTS sampler could aid here.

3.3.2 *Improving the ratio estimator $\hat{r}(x|\boldsymbol{\vartheta})$ by directly estimating the posterior probability density function*

Simply relying on the previously defined amortized likelihood-to-reference ratio estimator $\hat{r}(x|\boldsymbol{\vartheta})$ does not yield satisfactory results, even when considering simple toy problems. Experiments indicate that the choice of the although mathematically arbitrary reference hypothesis $\boldsymbol{\vartheta}_{\text{ref}}$ does have a significant effect on the approximated likelihood ratios in practice. Other independent investigations [57] observe similar issues and conclude, like us, that the reference hypothesis $\boldsymbol{\vartheta}_{\text{ref}}$ is a sensitive hyper-parameter which requires careful tuning for the problem at hand.

We find that poor inference results occur in the absence of support between $p(\boldsymbol{x}|\boldsymbol{\vartheta})$ and $p(\boldsymbol{x}|\boldsymbol{\vartheta}_{\text{ref}})$, as illustrated in Figure 3.1. In this example, the evaluation of the approximate ratio $\hat{r}(x|\boldsymbol{\vartheta})$ for an observable $x \sim p(\boldsymbol{x}|\boldsymbol{\vartheta}^*)$ is undefined when the observation x does not have density in $p(\boldsymbol{x}|\boldsymbol{\vartheta})$ and $p(\boldsymbol{x}|\boldsymbol{\vartheta}_{\text{ref}})$, or either of the densities is numerically negligible. Therefore, the continuous decision function modeled by the optimal classifier $d(\boldsymbol{\vartheta}, x)$ outside of the support of $p(\boldsymbol{x}|\boldsymbol{\vartheta})$ and $p(\boldsymbol{x}|\boldsymbol{\vartheta}_{\text{ref}})$ is undefined. Practically, this implies that the ratio estimator $\hat{r}(x|\boldsymbol{\vartheta})$ can take on an arbitrary value, which is detrimental to the inference procedure. In this case, the value of $\hat{r}(x|\boldsymbol{\vartheta})$

Algorithm 5 Likelihood-free Hamiltonian MCMC

Inputs: Initial parameter $\boldsymbol{\vartheta}_0$
Differentiable prior $p(\boldsymbol{\vartheta})$
Momentum distribution $q(m)$
Differentiable ratio estimator $\hat{r}(x|\boldsymbol{\vartheta})$
Differentiable likelihood $p(\boldsymbol{x}\,|\boldsymbol{\vartheta})$
Observable x
Outputs: Markov chain $\boldsymbol{\vartheta}_{0:T}$
Hyperparameters: Steps T.
Leapfrog-integration steps l and stepsize η.

$t \leftarrow 0$
$\boldsymbol{\vartheta}_t \leftarrow \boldsymbol{\vartheta}_0$
for $t < T$ **do**
  $m_t \sim q(m)$
  $k \leftarrow 0$
  $m_k \leftarrow m_t$
  $\boldsymbol{\vartheta}_k \leftarrow \boldsymbol{\vartheta}_t$
  $U(\boldsymbol{\vartheta}_t) \leftarrow -\log p(\boldsymbol{\vartheta}_t) - \log \hat{r}(x|\boldsymbol{\vartheta}_t)$
  for $k < l$ **do**
    $m_k \leftarrow m_k - \frac{\eta}{2} \nabla_{\boldsymbol{\vartheta}}\, U(\boldsymbol{\vartheta}_k)$
    $\boldsymbol{\vartheta}_k \leftarrow \boldsymbol{\vartheta}_k + \eta \cdot m_k$
    $U(\boldsymbol{\vartheta}_k) \leftarrow -\log p(\boldsymbol{\vartheta}_k) - \log \hat{r}(x|\boldsymbol{\vartheta}_k)$
    $m_k \leftarrow m_k + \frac{\eta}{2} \nabla_{\boldsymbol{\vartheta}}\, U(\boldsymbol{\vartheta}_k)$
    $k \leftarrow k + 1$
  end for
  $\rho \leftarrow \min\left[1,\ \frac{\exp\left(U(\boldsymbol{\vartheta}_k) + K(m_k)\right)}{\exp\left(U(\boldsymbol{\vartheta}_t) + K(m_t)\right)}\right]$
  $\boldsymbol{\vartheta}_{t+1} \leftarrow \begin{cases} \boldsymbol{\vartheta}_k & \text{with probability } \rho \\ \boldsymbol{\vartheta}_t & \text{with probability } 1-\rho \end{cases}$
  $t \leftarrow t + 1$
end for
return $\boldsymbol{\vartheta}_{0:T}$

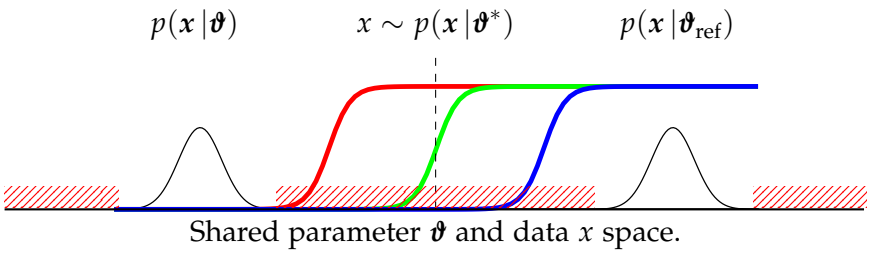

Figure 3.1: Consider having access to an optimal classifier $d(\boldsymbol{\vartheta}, \boldsymbol{x})$ modeling $r(\boldsymbol{x}\,|\boldsymbol{\vartheta})$ with $\boldsymbol{x} \sim p(\boldsymbol{x}\,|\boldsymbol{\vartheta}^*)$. This ratio is undefined for x as neither $p(\boldsymbol{x}\,|\boldsymbol{\vartheta})$ nor $p(\boldsymbol{x}\,|\boldsymbol{\vartheta}_{\text{ref}})$ puts numerically non-negligible density on x. This implies that $\hat{r}(x|\boldsymbol{\vartheta})$ and its decision function $d(\boldsymbol{\vartheta}, \boldsymbol{x})$ can take on arbitrary values in regions not covered by $p(\boldsymbol{x}\,|\boldsymbol{\vartheta})$ or $p(\boldsymbol{x}\,|\boldsymbol{\vartheta}_{\text{ref}})$ (striped areas) because no such training data exists, or the availability of such data is sparse. The red, green and blue lines depict optimal decision functions as they all minimize the criterion which captures the ability to classify between samples from $p(\boldsymbol{x}\,|\boldsymbol{\vartheta})$ and $p(\boldsymbol{x}\,|\boldsymbol{\vartheta}_{\text{ref}})$. However, the functions all have different approximations of $\hat{r}(x|\boldsymbol{\vartheta})$.

might depend on architectural choices in $d(\boldsymbol{\vartheta}, \boldsymbol{x})$ or stochastic aspects of the training procedure.

To overcome the issues associated with a fixed reference hypothesis, we propose to train the classifier to distinguish samples from $p(\boldsymbol{x}\,|\boldsymbol{\vartheta})$ (numerator) and the marginal model $p(\boldsymbol{x})$ (denominator) such that

$$r(\boldsymbol{x}\,|\boldsymbol{\vartheta}) \triangleq \frac{p(\boldsymbol{x}\,|\boldsymbol{\vartheta})}{p(\boldsymbol{x})}. \tag{3.25}$$

This modification ensures that *(i)* the likelihood-to-evidence ratio will always be defined everywhere it needs to be evaluated, as the likelihood $p(\boldsymbol{x}\,|\boldsymbol{\vartheta})$ is consistently supported by the marginal model $p(\boldsymbol{x})$, and *(ii)* enables the direct evaluation of the posterior density function because

$$p(\boldsymbol{\vartheta}|\,\boldsymbol{x}) = p(\boldsymbol{\vartheta})\frac{p(\boldsymbol{x}\,|\boldsymbol{\vartheta})}{p(\boldsymbol{x})}, \tag{3.26}$$

$$\approx p(\boldsymbol{\vartheta})\hat{r}(x|\boldsymbol{\vartheta}). \tag{3.27}$$

We summarize the procedure for learning the associated discriminator $d(\boldsymbol{\vartheta}, \boldsymbol{x})$ and the corresponding ratio estimator $\hat{r}(x|\boldsymbol{\vartheta})$ in Algorithm 6.

Proposition 1. *The decision function modeled by the optimal discriminator $d(\boldsymbol{\vartheta}, x)$ trained under a prior $p(\boldsymbol{\vartheta})$ to distinguish samples from the joint $p(\boldsymbol{\vartheta}, \boldsymbol{x})$ and product of the marginals $p(\boldsymbol{\vartheta})p(\boldsymbol{x})$ is*

$$d(\boldsymbol{\vartheta}, \boldsymbol{x}) = \frac{p(\boldsymbol{\vartheta}, \boldsymbol{x})}{p(\boldsymbol{\vartheta}, \boldsymbol{x}) + p(\boldsymbol{\vartheta})p(\boldsymbol{x})} = \frac{p(\boldsymbol{x}\,|\boldsymbol{\vartheta})}{p(\boldsymbol{x}\,|\boldsymbol{\vartheta}) + p(\boldsymbol{x})}. \tag{3.28}$$

Algorithm 6 Optimization of $d_\psi(\boldsymbol{\vartheta}, \boldsymbol{x})$ to obtain a ratio estimator.

Inputs:	Criterion ℓ (e.g., BCE, or binary cross-entropy)
	Implicit generative model (simulator) $p(\boldsymbol{x}\,|\boldsymbol{\vartheta})$
	Prior $p(\boldsymbol{\vartheta})$
Outputs:	Discriminator $d_\psi(\boldsymbol{\vartheta}, x)$
Hyperparameters:	Batch-size M

1: **while not converged do**
2: **Sample** $\boldsymbol{\vartheta} \leftarrow \{\boldsymbol{\vartheta}_m \sim p(\boldsymbol{\vartheta})\}_{m=1}^M$
3: **Sample** $\boldsymbol{\vartheta}^{'} \leftarrow \{\boldsymbol{\vartheta}_m^{'} \sim p(\boldsymbol{\vartheta})\}_{m=1}^M$
4: **Simulate** $x \leftarrow \{x_m \sim p(\boldsymbol{x}\,|\boldsymbol{\vartheta}_m)\}_{m=1}^M$
5: $\mathcal{L}\left[d_\psi(\boldsymbol{\vartheta}, \boldsymbol{x})\right] \leftarrow \ell(d_\psi(\boldsymbol{\vartheta}, \boldsymbol{x}),\ 1) + \ell(d_\psi(\boldsymbol{\vartheta}', \boldsymbol{x}),\ 0)$
6: $\psi \leftarrow$ **optimizer**$(\psi,\ \nabla_\psi \mathcal{L}\left[d_\phi(\boldsymbol{\vartheta}, \boldsymbol{x})\right])$
7: **end while**
8: **return** d_ψ

Proof. The core of our contribution rests on the proper estimation of the likelihood-to-evidence ratio. This proof demonstrates the minimization of the binary cross-entropy (BCE) loss of a classifier tasked to distinguish between *dependent* input pairs $(\boldsymbol{\vartheta}, \boldsymbol{x}) \sim p(\boldsymbol{\vartheta}, \boldsymbol{x})$ (samples drawn the joint) and independent input pairs $(\boldsymbol{\vartheta}, \boldsymbol{x}) \sim p(\boldsymbol{\vartheta})p(\boldsymbol{x})$ (samples drawn from the product of the marginals) results in an optimal classifier $d(\boldsymbol{\vartheta}, \boldsymbol{x})$ *when minimized*.

Using calculus of variations and reproducing the structure of Algorithm 6, we define the loss functional $\mathcal{L}[d(\boldsymbol{\vartheta}, \boldsymbol{x})]$

$$
\begin{aligned}
&= \int d\boldsymbol{\vartheta} \int dx \int d\boldsymbol{\vartheta}' p(\boldsymbol{\vartheta}) p(\boldsymbol{x}\,|\boldsymbol{\vartheta}) p(\boldsymbol{\vartheta}') \Big[-\log d(\boldsymbol{\vartheta}, \boldsymbol{x}) - \log(1 - (d(\boldsymbol{\vartheta}'), \boldsymbol{x}) \Big] \\
&= \int d\boldsymbol{\vartheta} \int dx \int \underbrace{p(\boldsymbol{\vartheta}) p(\boldsymbol{x}\,|\boldsymbol{\vartheta}) \Big[-\log d(\boldsymbol{\vartheta}, \boldsymbol{x}) \Big] + p(\boldsymbol{\vartheta}) p(\boldsymbol{x}) \Big[-\log(1 - d(\boldsymbol{\vartheta}, \boldsymbol{x})) \Big]}_{F(d(\boldsymbol{\vartheta}, \boldsymbol{x}))}.
\end{aligned}
$$

This loss functional is minimized for a function $d(\boldsymbol{\vartheta}, \boldsymbol{x})$ such that

$$
0 = \left. \frac{\delta F}{\delta d} \right|_d = p(\boldsymbol{\vartheta}) p(\boldsymbol{x}\,|\boldsymbol{\vartheta}) \Big[-\frac{1}{d(\boldsymbol{\vartheta}, \boldsymbol{x})} \Big] + p(\boldsymbol{\vartheta}) p(\boldsymbol{x}) \Big[\frac{1}{1 - d(\boldsymbol{\vartheta}, \boldsymbol{x})} \Big]. \quad (3.29)
$$

As long as $p(\boldsymbol{\vartheta}) > 0$, this is equivalent to

$$
p(\boldsymbol{x}\,|\boldsymbol{\vartheta}) \frac{1}{d(\boldsymbol{\vartheta}, \boldsymbol{x})} = p(\boldsymbol{x}) \frac{1}{1 - d(\boldsymbol{\vartheta}, \boldsymbol{x})}, \quad (3.30)
$$

and

$$
p(\boldsymbol{\vartheta}, \boldsymbol{x}) \frac{1}{d(\boldsymbol{\vartheta}, \boldsymbol{x})} = p(\boldsymbol{\vartheta}) p(\boldsymbol{x}) \frac{1}{1 - d(\boldsymbol{\vartheta}, \boldsymbol{x})}. \quad (3.31)
$$

Which implies that

$$
d(\boldsymbol{\vartheta}, \boldsymbol{x}) = \frac{p(\boldsymbol{\vartheta}, \boldsymbol{x})}{p(\boldsymbol{\vartheta}, \boldsymbol{x}) + p(\boldsymbol{\vartheta}) p(\boldsymbol{x})} = \frac{p(\boldsymbol{x}\,|\boldsymbol{\vartheta})}{p(\boldsymbol{x}\,|\boldsymbol{\vartheta}) + p(\boldsymbol{x})}. \quad (3.32)
$$

Therefore, the optimal discriminator models the likelihood-to-evidence ratio

$$r(\boldsymbol{x}\,|\boldsymbol{\vartheta}) \triangleq \frac{p(\boldsymbol{\vartheta},\boldsymbol{x})}{p(\boldsymbol{\vartheta})p(\boldsymbol{x})} = \frac{p(\boldsymbol{x}\,|\boldsymbol{\vartheta})}{p(\boldsymbol{x})}. \tag{3.33}$$

□

Although the usage of the marginal model instead of an arbitrary reference hypothesis vastly improves the accuracy of $\hat{r}(x|\boldsymbol{\vartheta})$, obtaining the likelihood-to-evidence ratio $\hat{r}(x|\boldsymbol{\vartheta})$ by transforming the output of $d(\boldsymbol{\vartheta},\boldsymbol{x})$ can still be susceptible to numerical errors. This may happen in the saturating regime where the classifier $d(\boldsymbol{\vartheta},\boldsymbol{x})$ is able to (almost) perfectly discriminate samples from $p(\boldsymbol{x}\,|\boldsymbol{\vartheta})$ and $p(\boldsymbol{x})$. We prevent this issue by extracting $\log \hat{r}(x|\boldsymbol{\vartheta})$ from the neural network before applying the sigmoidal projection in the output layer, since $\log \hat{r}(x|\boldsymbol{\vartheta})$ is the logit of $d(\boldsymbol{\vartheta},\boldsymbol{x})$. This choice also mitigates a vanishing gradient due to the sigmoidal output when computing $\nabla_{\boldsymbol{\vartheta}} \log \hat{r}(x|\boldsymbol{\vartheta})$ or $\nabla_x \log \hat{r}(x|\boldsymbol{\vartheta})$. Which is benificial in any application of likelihood-free HMC.

Finally, approximating the likelihood-to-evidence ratio also enables the direct estimation of the posterior probability density function because $\hat{p}(\boldsymbol{\vartheta}|\,\boldsymbol{x}) = p(\boldsymbol{\vartheta})\hat{r}(x|\boldsymbol{\vartheta})$. This is especially useful in low-dimensional model parameter spaces, where scanning is a reasonable strategy and much more efficient compared to MCMC. In addition, scanning does not rely on tunable parameters.

3.3.3 *Assessing the quality of the ratio estimates*

Likelihood-free computations are challenging to verify as the quantity which drives the inference procedure, the likelihood, is by definition intractable. A robust strategy is therefore *crucial* to verify the quality of any approximation before making any scientific conclusions. Inspired by Cranmer, Pavez, and Louppe [55], one can identify issues in our ratio estimator $\hat{r}(x|\boldsymbol{\vartheta})$ by evaluating the identity

$$p(\boldsymbol{x}\,|\boldsymbol{\vartheta}) = p(\boldsymbol{x})r(\boldsymbol{x}\,|\boldsymbol{\vartheta}). \tag{3.34}$$

Whenever the *estimator* $\hat{r}(x|\boldsymbol{\vartheta})$ is exact, then a classifier should not be able to distinguish between samples from $p(\boldsymbol{x}\,|\boldsymbol{\vartheta})$ and the reweighted marginal model $p(\boldsymbol{x})\hat{r}(x|\boldsymbol{\vartheta})$. The discriminative performance of the classifier can be assessed by means of a Receiver Operating Characteristic (ROC) curve. A diagonal ROC (with Area Under Curve = 0.5) curve indicates that a classifier is insensitive and $\hat{r}(x|\boldsymbol{\vartheta}) = r(\boldsymbol{x}\,|\boldsymbol{\vartheta})$. It should be noted however this result could also be obtained by a classifier that is insenstive to differences between samples from $p(\boldsymbol{x}\,|\,\boldsymbol{\vartheta})$ and the reweighted marginal model. Figure 3.2 provides an illustration of this diagnostic.

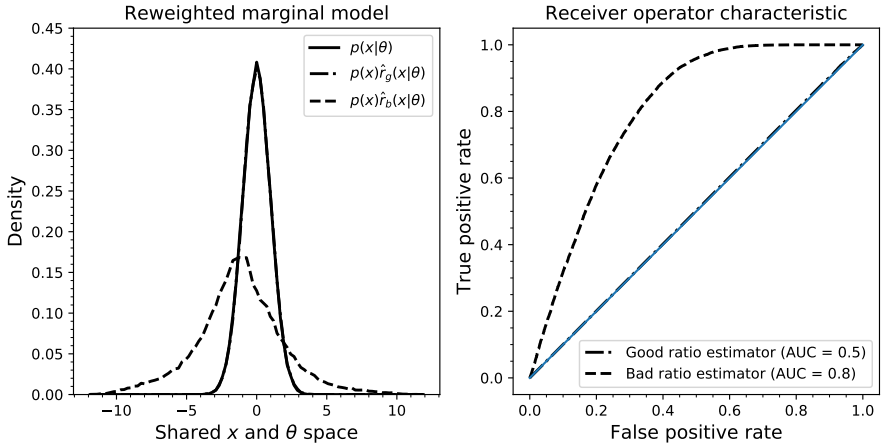

Figure 3.2: This figure demonstrates the diagnostic presented in Section 3.3.3. We train two ratio estimators. The first approximates the ratio $r(\boldsymbol{x} \mid \boldsymbol{\vartheta})$ well, while the other does not. We denote these estimators as $\hat{r}_g(\boldsymbol{x} \mid \boldsymbol{\vartheta})$ and $\hat{r}_b(\boldsymbol{x} \mid \boldsymbol{\vartheta})$ respectively. The test diagnostic is applied to a single test hypothesis $\boldsymbol{\vartheta} = 0$. *(Left):* Marginal model reweighted using $\hat{r}_g(\boldsymbol{x} \mid \boldsymbol{\vartheta})$ and $\hat{r}_b(\boldsymbol{x} \mid \boldsymbol{\vartheta})$. It is clear that $\hat{r}_b(\boldsymbol{x} \mid \boldsymbol{\vartheta})$ does not properly approximate $r(\boldsymbol{x} \mid \boldsymbol{\vartheta})$, as the reweighted marginal model is distinguishable from the test hypothesis $p(\boldsymbol{x} \mid \boldsymbol{\vartheta} = 0)$. *(Right):* A classifier is trained to distinguish between samples from the test hypothesis and the reweighted marginal models. The ROC curve indicates that the classifier could not extract any predictive features for samples $\boldsymbol{x} \sim p(\boldsymbol{x})$ reweighted by $\hat{r}_g(\boldsymbol{x} \mid \boldsymbol{\vartheta})$, indicating a good approximation of $r(\boldsymbol{x} \mid \boldsymbol{\vartheta})$ by $\hat{r}_g(\boldsymbol{x} \mid \boldsymbol{\vartheta})$.

While this particular approach is able to detect defects in the ratio estimators for a specific $\boldsymbol{\vartheta}$, it does not give an indication about the ratio estimator's performance across the prior $p(\boldsymbol{\vartheta})$. To compute some estimate of the ratio estimator's reliability across the prior $p(\boldsymbol{\vartheta})$, there are 2 distinct approaches. The first simply repeats the procedure above for various $\boldsymbol{\vartheta} \sim p(\boldsymbol{\vartheta})$. The second however, relies on the observation that the Bayes optimal ratio estimator

$$r(\boldsymbol{x} \mid \boldsymbol{\vartheta}) = \frac{p(\boldsymbol{x} \mid \boldsymbol{\vartheta})}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{\vartheta}, \boldsymbol{x})}{p(\boldsymbol{\vartheta})p(\boldsymbol{x})}. \tag{3.35}$$

This allows us to determine the exactness of any approximate ratio estimator $\hat{r}(\boldsymbol{x} \mid \boldsymbol{\vartheta})$ over the prior, by evaluating the identity

$$p(\boldsymbol{\vartheta}, \boldsymbol{x}) = r(\boldsymbol{x} \mid \boldsymbol{\vartheta})p(\boldsymbol{\vartheta})p(\boldsymbol{x}). \tag{3.36}$$

In contrast to the identity in Equation 3.34, this formulation requires a classifier that accepts both $\boldsymbol{\vartheta}$ and $\boldsymbol{x}$ as inputs to discriminate between samples from the joint and the reweighted product of marginals. It should be noted however that the task of discriminating between these two densities is harder compared to the previous formulation. Because of this, special consideration should be given to the fact that the classifier might be insensitive to discriminate the joint and the reweighted product of the marginals.

3.4 RELATED WORK

Algorithms such as ABC [88–91] tackle the problem of Bayesian inference by collecting proposal states $\boldsymbol{\vartheta} \sim p(\boldsymbol{\vartheta})$ whenever an observation $\boldsymbol{x}$ produced by the forward model $\boldsymbol{x} \sim p(\boldsymbol{x} \mid \boldsymbol{\vartheta})$ resembles an observation $\boldsymbol{x}_o$. Formally, a proposal state $\boldsymbol{\vartheta}$ is accepted whenever a compressed observation $\sigma(\boldsymbol{x})$ (low-dimensional summary statistic) satisfies $d(\sigma(\boldsymbol{x}), \sigma(\boldsymbol{x}_o)) < \epsilon$ for some distance function and acceptance threshold ϵ. *The resulting approximation of the posterior will only be exact whenever the summary statistic is sufficient and* $\epsilon \to 0$ [90]. Several procedures have been proposed to improve the acceptance rate by guiding simulations based on previously accepted states [92–94]. Other works investigated learning summary statistics [95–97]. Contrary to these methods, AALR-MCMC does not actively use the simulator during inference and learns a direct mapping from data and parameter space to likelihood-to-evidence ratios.

Other approaches take the perspective to cast inference as an optimization problem [98, 99]. In variational inference, a parameterized posterior over parameters of interest is optimized [100]. Amortized variational inference [101, 102] expands on this idea by using generative models to capture inference mappings. Recent work in [9]

proposes a novel form of variational inference by introducing an adversary in combination with REINFORCE-estimates [103, 104] to optimize a parameterized prior. Others have investigated meta-learning to learn parameter updates [105]. However, these works only provide point-estimates.

Sequential approaches such as SNPE-A [106], SNPE-B [107] and APT/SNPE-C [108] iteratively adjust an approximate posterior parameterized as a mixture density network or a normalizing flow. Instead of learning the posterior directly, SNL [109] makes use of autoregressive flows to model an approximate likelihood. AALR-MCMC mirrors SNL as the trained conditional density estimator is plugged into MCMC samplers to bypass the intractable marginal model. This allows SNL to approximate the posterior numerically. Contrary to our approach, SNL cannot directly provide estimates of the posterior posterior density function.

The usage of ratios is explored in several studies. CARL [110] models likelihood ratios for frequentist tests. As shown in Section 3.3.2, CARL does not produce accurate results in some cases. LFIRE [57] models a likelihood-to-evidence ratio by logistic regression and relies on the usage of summary statistics. Unlike us, they require samples from the marginal model and a specific (reference) likelihood, while we only require samples from the joint $p(\boldsymbol{\vartheta}, \boldsymbol{x})$. Therefore, LFIRE requires retraining for every evaluation of different $\boldsymbol{\vartheta}$.

Finally, an important concern of likelihood-free inference is minimizing the number of simulation calls. Active simulation strategies such as BOLFI [111] and others [112, 113] achieve this through Bayesian optimization. Emulator networks [114] exploit the uncertainty within an ensemble to guide simulations. Recent works [80, 115] significantly reduce the amount of required simulations, provided joint likelihood ratios and scores can be extracted from the simulator.

Other approaches attempt to learn approximate versions of the simulator [114] which is used to perform efficient inference, adapting a similar strategy as in world models for reinforcement learning [116]. Our method relies on the training of an amortized likelihood-to-evidence ratio estimator using samples from the joint $p(\boldsymbol{\vartheta}, \boldsymbol{x})$. Thereby directly modeling all posteriors, unlike Lueckmann et al. [114], which learns a global likelihood model, but does not provide a concrete density estimator to achieve this. Our ratio estimator could enable this.

3.5 EXPERIMENTS

3.5.1 *Setup*

We compare AALR-MCMC using our likelihood-to-evidence ratio estimator against classical ABC [117] and modern posterior approximation techniques such as SNPE-A [106], SNPE-B [107], and APT [108]. All methods have a simulation budget of one million samples. Sequential approaches such as SNPE-A, SNPE-B, and APT spread this budget equally across 50 rounds. These rounds are used to iteratively improve the approximation of the posterior. Our evaluations consider the posterior estimate of the final round. By default, our evaluations use the likelihood-free Metropolis-Hastings sampler unless stated otherwise. The experiments are repeated 10 times.

3.5.1.1 *Benchmark problems*

The accuracy and robustness of AALR-MCMC will be assessed by comparing AALR-MCMC against ABC, SNPE-A, SNPE-B and APT on the following benchmarks:

TRACTABLE PROBLEM Given a model parameter sample $\boldsymbol{\vartheta} \in \mathbb{R}^5$, the forward generative process is defined as:

$$\mu_{\boldsymbol{\vartheta}} = (\boldsymbol{\vartheta}_0, \boldsymbol{\vartheta}_1), \tag{3.37}$$

$$s_1 = \boldsymbol{\vartheta}_2^2, \quad s_2 = \boldsymbol{\vartheta}_3^2, \quad \rho = \tanh(\boldsymbol{\vartheta}_4), \tag{3.38}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{\vartheta}} = \begin{bmatrix} s_1^2 & \rho s_1 s_2 \\ \rho s_1 s_2 & s_2^2 \end{bmatrix}, \tag{3.39}$$

$$\text{with } \boldsymbol{x} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_4) \text{ where } \boldsymbol{x}_i \sim \mathcal{N}(\mu_{\boldsymbol{\vartheta}}, \boldsymbol{\Sigma}_{\boldsymbol{\vartheta}}) \tag{3.40}$$

The likelihood $p(\boldsymbol{x}\,|\boldsymbol{\vartheta}) = \prod_{i=1}^{4} \mathcal{N}(\boldsymbol{x}_i \,|\, \mu_{\boldsymbol{\vartheta}}, \boldsymbol{\Sigma}_{\boldsymbol{\vartheta}})$ with prior $p(\boldsymbol{\vartheta}) \triangleq \mathcal{U}(-3, 3)$. The resulting posterior is non-trivial due to the squaring operation, which is responsible for generating multiple modes. An observation x_o is generated by conditioning the forward model on $\boldsymbol{\vartheta}^* = (0.7, -2.9, -1.0, -0.9, 0.6)$ as in [108, 109].

DETECTOR CALIBRATION We are interested in determining the offset $\boldsymbol{\vartheta} \in \mathbb{R}$ of a particle detector from the collision point given a detector response $\boldsymbol{x}_o$. Our particle detector emulates a 32×32 spherical uniform grid such that $x \in \mathbb{R}^{1024}$. Every detector pixel is able to measure the momentum of the particles passing through the detector material. The `pythia` simulator [52] generates electron-positron ($e^- e^+$) collisions and is configured according to the parameters derived by the Monash tune [118]. The resulting collision products and their

momenta are processed by `pythiamill` [53] to compute the response of the detector by simulating the interaction of the collision products with the detector material. We consider a prior $p(\boldsymbol{\vartheta}) \triangleq \mathcal{U}(-30, 30)$. An observable $\boldsymbol{x}_o$ is generated at the collision point $\boldsymbol{\vartheta}^* = 0$.

POPULATION MODEL The Lotka-Volterra model [119] describes the evolution of predator-prey populations. The population dynamics are driven by a set of differential equations with parameters $\boldsymbol{\vartheta} \in \mathbb{R}^4$. An observation describes the population counts of both groups over time. Simulations are typically compressed into a summary statistic $\bar{\boldsymbol{x}} \in \mathbb{R}^9$ [108, 109]. We also follow this approach to remain consistent. The prior $p(\boldsymbol{\vartheta}) \triangleq \mathcal{U}(-10, 2)$ (log-scale). We generate an observable from the narrow oscillating regime $\boldsymbol{\vartheta}^* = (-4.61, -0.69, 0, -4.61)$.

M/G/1 QUEUING MODEL This model describes a queuing system of continuously arriving jobs at a single server and is described by a model parameter $\boldsymbol{\vartheta} \in \mathbb{R}^3$. The time it takes to process every job is uniformly distributed in the interval $[\boldsymbol{\vartheta}_1, \boldsymbol{\vartheta}_2]$. The arrival time between two consecutive jobs is exponentially distributed according to the rate $\boldsymbol{\vartheta}_3$. An observation x are 5 equally evenly spaced percentiles of interdeparture times, i.e., the 0th, 25th, 50th, 75th and 100th percentiles. To generate the observation $\boldsymbol{x}_o$, we draw a sample from the forward model using the generating parameter $\boldsymbol{\vartheta}^* = (1.0, 5.0, 0.2)$. We consider the uniform prior $p(\boldsymbol{\vartheta}) \triangleq \mathcal{U}(0, 10) \times \mathcal{U}(0, 10) \times \mathcal{U}(0, 0.333)$.

3.5.2 *Results*

Table 3.1 shows the posterior log probabilities of the generating parameter $\boldsymbol{\vartheta}^*$ for an observation $\boldsymbol{x}_o$. Our ROC diagnostic reports AUC = 0.5 for the detector calibration and M/G/1 benchmarks, and AUC = 0.55 for the population evolution model. These results demonstrate that the proposed ratio estimator provides accurate and consistent ratio estimates.

If we assess the quality of a method exclusively based on the log probabilities in Table 3.1, we could argue that SNPE-A, SNPE-B and APT are close in terms of approximation. This is potentially misleading as it does not take the structure of the posterior into account. To demonstrate the accuracy of AALR-MCMC in this regard, we focus on the tractable problem. We conduct two distinct quantitative analysis, the first computes the Maximum Mean Discrepancy (MMD) [120] between samples of the true posterior and the approximated posterior, while the latter trains a classifier to compute the ROC AUC between samples of the approximate posterior and the MCMC groundtruth. Results are summarized in Table 3.2. Figure 3.3 shows the approximations of

Algorithm	Tractable problem	Detector calibration	Population model	M/G/1
ABC ($\epsilon =$ large)	-8.686 ± 0.000	-3.087 ± 0.000	N/A	N/A
ABC ($\epsilon =$ intermediate)	-7.620 ± 0.000	-2.491 ± 0.000	N/A	N/A
ABC ($\epsilon =$ small)	-6.668 ± 0.000	-2.180 ± 0.000	N/A	N/A
APT	-4.441 ± 0.487	-2.004 ± 0.753	6.366 ± 0.432	-2.741 ± 3.356
SNPE-A	-6.141 ± 1.227	-1.775 ± 1.775	$\mathbf{7}.024 \pm 0.515$	1.177 ± 0.937
SNPE-B	-5.693 ± 0.809	-1.075 ± 0.226	-0.632 ± 0.843	1.105 ± 0.384
AALR-MCMC (ours)	-4.126 ± 0.004	-1.005 ± 0.074	6.482 ± 0.214	$\mathbf{2}.302 \pm 0.189$

Table 3.1: Posterior log probability $p(\boldsymbol{\vartheta} = \boldsymbol{\vartheta}^* | \boldsymbol{x} = x_o)$ for generating parameters $\boldsymbol{\vartheta}^*$ and observable $\boldsymbol{x}_o$. For SNPE-A, SNPE-B and APT we directly extracted the posterior log probability from the mixture of Gaussians. Since the proposed ratio estimator models the log likelihood-to-evidence ratio, we have to add the log prior probability of the generating parameters to obtain the posterior log probability.

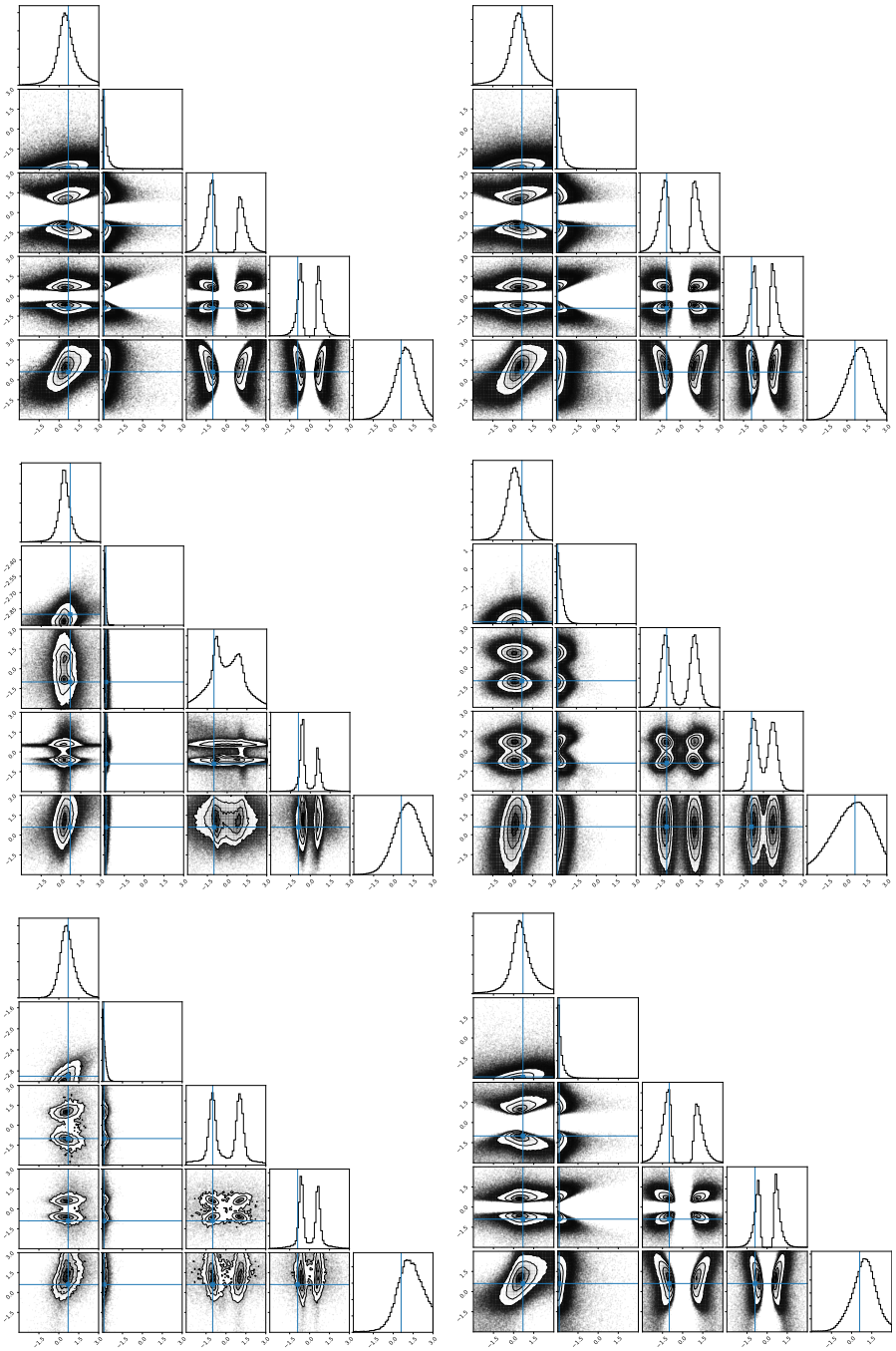

Figure 3.3: Posteriors from the tractable benchmark. **From left to right**, the MCMC ground truth, the proposed method, SNPE-C, SNPE-B, APT and SNL. The experiments are repeated 25 times and the approximate posteriors are subsampled from those runs. An objective visual assessment can be made: AALR-MCMC shares the same structure with the MCMC truth, demonstrating its accuracy. Some runs of the other methods were not consistent, contributing to the variance observed in Table 3.2.

Algorithm	MMD	ROC AUC
AALR-MCMC (ours)	$\mathbf{0}.05 \pm 0.005$	$\mathbf{0}.59 \pm 0.0010$
ABC ($\epsilon = 32$)	0.51 ± 0.001	0.99 ± 0.0001
ABC ($\epsilon = 16$)	0.50 ± 0.003	0.99 ± 0.0002
ABC ($\epsilon = 8$)	0.39 ± 0.001	0.99 ± 0.0003
ABC ($\epsilon = 4$)	0.29 ± 0.004	0.98 ± 0.0007
APT	0.17 ± 0.036	0.86 ± 0.0008
AALR-MCMC (LRT)	0.53 ± 0.004	0.99 ± 0.0001
SNPE-A	0.21 ± 0.070	0.97 ± 0.0098
SNPE-B	0.20 ± 0.061	0.92 ± 0.0181

Table 3.2: AALR-MCMC outperforms all other methods. Numerical errors introduced by MCMC might have contributed to these results. The MMD scores are in agreement with [108].

AALR-MCMC, SNPE-A, SNPE-B and APT against the MCMC groundtruth. AALR-MCMC's accuracy is especially apparent when comparing SNPE-A, SNPE-B and APT against the groundtruth. While AALR-MCMC accurately models the true posterior, SNPE-A, SNPE-B and APT fail to do so. The discrepancy between the LRT and the proposed ratio estimator indicate that the improvements from Section 3.3.2 are *critical*.

In addition to comparing the final approximations, we evaluate the accuracy of the approximations with respect to a given simulation budget. In doing so we challenge our method even further, as sequential approaches are specifically designed to be simulation efficient. We expect sequential approaches to obtain more accurate approximations with less simulations. The results of this evaluation are shown in Figure 3.4. With the exception of SNL which produces results comparable to ours, we unexpectedly find that the sequential approaches were not able to outperform our method on this (toy) problem, even though AALR-MCMC and its ratio estimator tackle the harder task of amortized inference. This demonstrates the accuracy and robustness of our method.

Finally, because the other methodologies are not amortized, i.e., they cannot approximate arbitrary posteriors, we note that experiments consider a single and fixed observation only. General conclusions should therefore be drawn with caution.

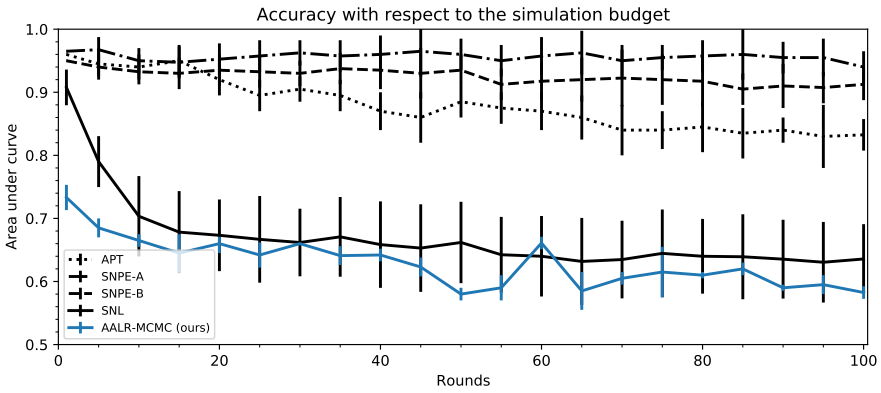

Figure 3.4: We evaluate the accuracy of the approximations with respect to different simulation budgets on the tractable benchmark. The accuracy is obtained by computing the ROC AUC between samples from the approximation and the MCMC groundtruth. Except for SNL which yields comparable results, sequential approaches are not able to outperform AALR-MCMC.

3.5.3 *Demonstrations: strong gravitational lensing*

The following demonstrations will showcase several aspects of our method while considering the problem of strong gravitational lensing. We use `autolens` [121] to simulate the telescope optics, imaging sensors and physics governing strong lensing. The simulation black-box encapsulates these components. The output of the simulation is a high-dimensional observation $\boldsymbol{x} \in \mathbb{R}^{128\times128}$ with uninformative data dimensions. We use a ratio estimator based on RESNET-18 [122] parameterized by $\boldsymbol{\vartheta}$ in the fully connected trunk.

The simulation model consists of 4 main components. The first involves the telescope optics. We model the PSF (point spread function) as a Gaussian with standard deviation 0.5 in a 3×3 pixel kernel. The CCD sensor is set to an exposure time of 1000 seconds, background sky level = 0.1 and CCD noise is added. The mass distribution of the foreground galaxy is modeled as an elliptical isothermal [123] at redshift $z = 0.5$ with axis ratio = 0.99, a random orientation-angle and an Einstein radius sampled from the prior. We do not model galaxy foreground light for the marginalization problem. For the Bayesian model selection problem, we model the foreground light of the lensing galaxy as an elliptical sersic with a random orientation angle and a sersic index sampled from $\mathcal{U}(.5, 1.5)$. For every source galaxy, we only model the light profile and their relative positions with respect to the lens. Source galaxies have an assumed redshift of $z = 2$. We assume the Plack15 cosmology. Table 3.3 describes the parameters and

respective distributions we sampled from to generate a light profile for a single source galaxy.

Parameter	Distribution
Axis ratio	$\mathcal{U}(0.1, 0.9)$
Effective radius	$\mathcal{U}(0.1, 0.4)$
Intensity (flux)	$\mathcal{U}(0.1, 0.5)$
Location x	$\mathcal{U}(-1.0, 1.0)$
Location y	$\mathcal{U}(-1.0, 1.0)$
Axis orientation	$\mathcal{U}(0, 360)$
Sersic index	$\mathcal{U}(0.5, 3.0)$

Table 3.3: A complete description of the parameters describing the light profile is described in the `autolens` documentation.

3.5.3.1 *Marginalization*

Often scientists are aphetic about a posterior describing all model parameters. Rather, they are interested in a posterior in which nuisance parameters have been marginalized out. This is easily achieved within our framework by including all parameters (including nuisance parameters) to the simulation model, but only presenting the parameters of interest to the ratio estimator during training. The training procedure remains otherwise unchanged. This problem focuses on recovering the Einstein radius $\boldsymbol{\vartheta} \in \mathbb{R}$ of a gravitational lens. We are not interested in the parameters describing the source and foreground galaxy (15 parameters). Figure 3.5 depicts our posterior approximation, ROC diagnostic and observation $\boldsymbol{x}_o$ with $\boldsymbol{\vartheta}^* = 1.66$ and prior $p(\boldsymbol{\vartheta}) \triangleq \mathcal{U}(0.5, 3.0)$.

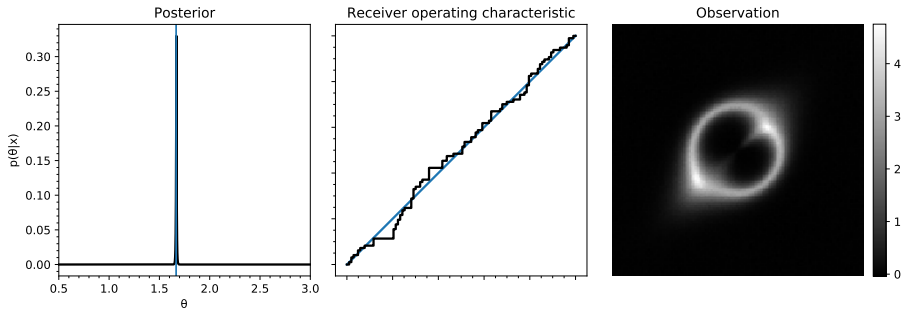

Figure 3.5: *(Left):* Approximation of the posterior. *(Middle):* Diagonal ROC diagnostic, indicating a good approximation of the posterior. *(Right):* Observation associated with the posterior.

3.5.3.2 *Amortization enables population studies*

Consider a set of n independent and identically distributed observations $\mathcal{X} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$. The amortization of the ratio estimator allows additional observations to be included in the computation of the posterior $p(\boldsymbol{\vartheta} \,|\mathcal{X})$ without requiring new simulations or retraining. This allows us to efficiently undertake population studies. Bayes' rule tells us

$$\begin{aligned} p(\boldsymbol{\vartheta} \,|\mathcal{X}) &= \frac{p(\boldsymbol{\vartheta}) \prod_{\boldsymbol{x} \in \mathcal{X}} p(\boldsymbol{x} \,|\, \boldsymbol{\vartheta})}{\int p(\boldsymbol{\vartheta}) \prod_{\boldsymbol{x} \in \mathcal{X}} p(\boldsymbol{x} \,|\, \boldsymbol{\vartheta}) d\,\boldsymbol{\vartheta}'} \\ &\approx \frac{p(\boldsymbol{\vartheta}) \prod_{\boldsymbol{x} \in \mathcal{X}} \hat{r}(\boldsymbol{x} \,|\, \boldsymbol{\vartheta})}{\int p(\boldsymbol{\vartheta}) \prod_{\boldsymbol{x} \in \mathcal{X}} \hat{r}(\boldsymbol{x} \,|\, \boldsymbol{\vartheta}) d\,\boldsymbol{\vartheta}}. \end{aligned} \tag{3.41}$$

The denominator can efficiently be approximated by Monte Carlo sampling using the ratio estimator $\hat{r}(\boldsymbol{x} \,|\, \boldsymbol{\vartheta})$. However, with MCMC the denominator cancels out within the ratio between consecutive states $\boldsymbol{\vartheta}_t \to \boldsymbol{\vartheta}'$. Thereby obtaining

$$\frac{\hat{p}(\boldsymbol{\vartheta}' | \mathcal{X})}{\hat{p}(\boldsymbol{\vartheta}_t | \mathcal{X})} = \frac{\frac{p(\boldsymbol{\vartheta}') \prod_{\boldsymbol{x} \in \mathcal{X}} \hat{r}(\boldsymbol{x} \,|\, \boldsymbol{\vartheta}')}{\int p(\boldsymbol{\vartheta}) \prod_{\boldsymbol{x} \in \mathcal{X}} \hat{r}(\boldsymbol{x} \,|\, \boldsymbol{\vartheta}) d\,\boldsymbol{\vartheta}}}{\frac{p(\boldsymbol{\vartheta}_t) \prod_{\boldsymbol{x} \in \mathcal{X}} \hat{r}(\boldsymbol{x} \,|\, \boldsymbol{\vartheta}_t)}{\int p(\boldsymbol{\vartheta}) \prod_{\boldsymbol{x} \in \mathcal{X}} \hat{r}(\boldsymbol{x} \,|\, \boldsymbol{\vartheta}) d\,\boldsymbol{\vartheta}}}. \tag{3.42}$$

We consider the same simulation model as in Section 3.5.3.1, with the exception that the Einstein radius used to simulate a gravitational lens is not $\boldsymbol{\vartheta}$, but instead drawn from $\mathcal{N}(\boldsymbol{\vartheta}, 0.25)$. We reduce the uncertainty about the generating parameter $\boldsymbol{\vartheta}^* = 2$ by modeling the posterior $\hat{p}(\boldsymbol{\vartheta} \,|\mathcal{X})$. This is demonstrated in Figure 3.6. All individual posteriors (dotted lines) are derived using the same pretrained ratio estimator. The posterior $\hat{p}(\boldsymbol{\vartheta} \,|\mathcal{X})$ is approximated using the formalism described above.

3.5.3.3 *Bayesian model selection*

Until now we only considered posteriors with continuous model parameters. We turn to a setting in which scientists are interested in a discrete space of models. *In essence casting classification as Bayesian model selection, allowing us the quantify the uncertainty among models (classes) with respect to an observation.* where every model m_i has a parameter space $\boldsymbol{\vartheta} \in \mathbb{R}^{d_i}$ of dimensionality d_i. Bayesian model selection is achieved by computing the Bayes factor b of two models m_i and m_j with parameter vectors $\boldsymbol{\vartheta}_i$ and $\boldsymbol{\vartheta}_j$:

$$\begin{aligned} b &= \frac{\int p(m_i, \boldsymbol{\vartheta}_i) p(\boldsymbol{x} \,|m_i, \boldsymbol{\vartheta}_i) d\,\boldsymbol{\vartheta}_i \,/ \int p(\boldsymbol{x}, \boldsymbol{\vartheta}_i) d\,\boldsymbol{\vartheta}_i}{\int p(m_j, \boldsymbol{\vartheta}_j) p(\boldsymbol{x} \,|m_j, \boldsymbol{\vartheta}_j) d\,\boldsymbol{\vartheta}_j \,/ \int p(\boldsymbol{x}, \boldsymbol{\vartheta}_j) d\,\boldsymbol{\vartheta}_j}, \\ &= \frac{p(m_i) p(\boldsymbol{x} \,|m_i)}{p(m_j) p(\boldsymbol{x} \,|m_j)} \approx \frac{p(m_i) \hat{r}(\boldsymbol{x} \,|m_i)}{p(m_j) \hat{r}(\boldsymbol{x} \,|m_j)}, \end{aligned} \tag{3.43}$$

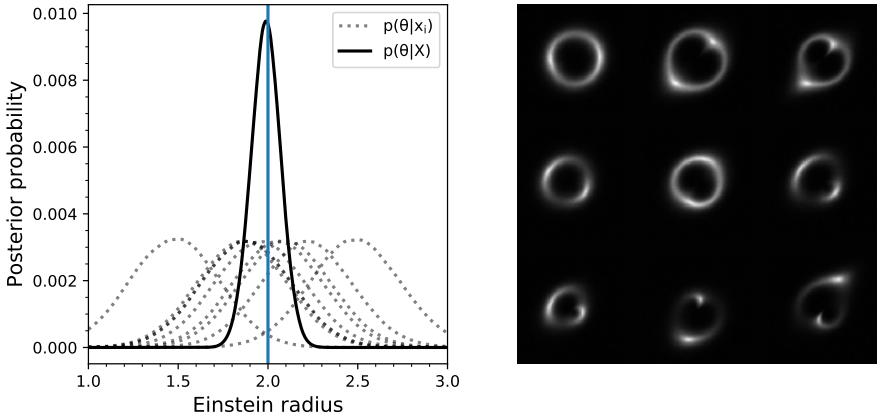

Figure 3.6: *(Left)*: The dotted lines represent the posteriors $\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x} = \boldsymbol{x}_i)$ for every independent and identically distributed observation $\boldsymbol{x}_i$, while the solid line depicts the posterior $\hat{p}(\boldsymbol{\vartheta} \mid \mathcal{X})$. All posteriors are derived using the same pretrained ratio estimator. *(Right)*: Observations sampled from $p(\boldsymbol{x} \mid \boldsymbol{\vartheta} = \boldsymbol{\vartheta}^*)$.

where a one-hot encoded model m_i is supplied to the ratio estimator during training. We demonstrate the task of model selection by computing the posterior $\hat{p}(m \mid \boldsymbol{x})$ across a space of 10 models $\mathcal{M} = \{m_0, \ldots, m_9\}$. The index i of a model m_i corresponds to the number of source galaxies present in the lensing system. The categorical prior $p(m)$ is uniform. Figure 3.7 shows $\hat{p}(m \mid \boldsymbol{x})$ and the associated diagnostic for different observations. Both posteriors were computed using the same ratio estimator.

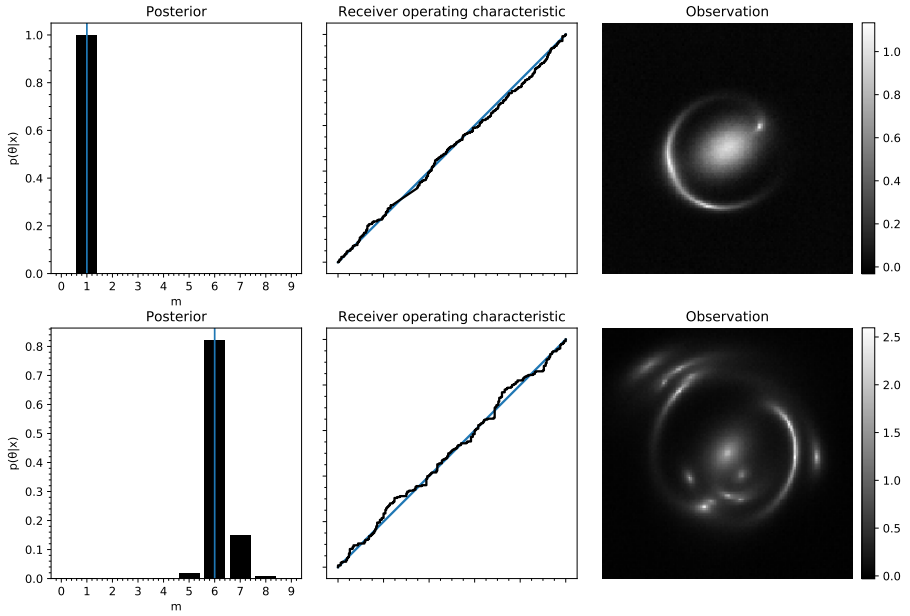

Figure 3.7: Posterior $\hat{p}(m|\boldsymbol{x})$ over the model space $\mathcal{M}$. Both diagnostics are diagonal. *(Top)*: Lensing system with a single source galaxy. *(Bottom):* Lensing system with 6 different source galaxies. The MAP of the posterior $\hat{p}(m|\boldsymbol{x})$ identifies the correct number of source galaxies, despite abundant lensing artifacts.

3.5.4 *Estimator capacity and sequential ratio estimation*

The amortization of our ratio estimator requires sufficient representational capacity to accurately approximate $r(\boldsymbol{x}\,|\,\boldsymbol{\vartheta})$, which of course directly depends on the complexity of the task at hand. Whenever the capacity of the ratio estimator is not sufficient, the quality of inference will be impacted.

However, increasing the capacity of a ratio estimator to match the complexity of the inference problem is not always a viable strategy, nor easy to determine beforehand. We observe that for a trained classifier $\hat{d}(vtheta, \boldsymbol{x})$ with insufficient capacity (AUC > 0.5) the posterior $\hat{p}(\boldsymbol{\vartheta}\,|\,\boldsymbol{x} = \boldsymbol{x}_o)$ is typically larger compared to the true posterior. **However, we make no statements as to whether this is always the case.**

Due to this observation, we can run a sequential ratio estimation procedure in which the posterior for $\boldsymbol{x} = \boldsymbol{x}_o$ is refined iteratively across a series of rounds. Starting with the initial prior $p_0(\boldsymbol{\vartheta}) := p(\boldsymbol{\vartheta})$, we improve the posterior by setting as prior for the next round, $p_{t+1}(\boldsymbol{\vartheta})$, the posterior $\hat{p}_t(\boldsymbol{\vartheta}\,|\,\boldsymbol{x} = \boldsymbol{x}_o)$ obtained at the previous round. At each iteration, the training procedure is repeated and eventually terminates based on the ROC diagnostic (AUC = 0.5).

To demonstrate this sequential ratio estimation procedure, let us assume the population model setting. Our ratio estimator is a low-capacity MLP with 3 layers and 50 hidden units. In every round t, 10,000 sample-parameter pairs are drawn from the joint $p(\boldsymbol{x}, \boldsymbol{\vartheta})$ with prior $p_t(\boldsymbol{\vartheta})$ for training. The following AUC scores were obtained: .99, .92, .54, and finally .50, terminating the algorithm.

Let us finally note that some time after the first version of this work, Durkan, Murray, and Papamakarios [124] identified that the sequential ratio estimation procedure outlined here is strongly related to APT/SNPE-C, in the sense that both approaches can actually be viewed as instances of a more general and unified constrastive learning scheme.

Regarding the naming of our method

It is widely known the hardest problem in Computer Science is in fact naming things. While the method put forword here is referred to as AALR, it should be noted that in the broader simulation-based inference literature this algorithm is referred to as NRE (neural ratio estimation). Equally, the sequential version of AALR presented above is commonly referred to as (S)NRE (sequential neural ratio estimation). As these names are objectively more suitable, we recommend the NRE naming scheme.

3.6 SUMMARY AND DISCUSSION

This work introduces a novel approach for Bayesian inference.We achieve this by replacing the intractable evaluation of the likelihood ratio in MCMC with an amortized likelihood ratio estimator. We demonstrate that a straightforward application of the likelihood ratio trick to MCMC is insufficient. We solve this by modeling the likelihood-to-evidence ratio for arbitrary observations $\boldsymbol{x}$ and model parameters $\boldsymbol{\vartheta}$. This implies that a pretrained ratio estimator can be used to infer the posterior density function of arbitrary observations. A theoretical argument demonstrates that the training procedure yields the optimal ratio estimator. The accuracy of an approximation can easily be verified by the proposed diagnostic. No summary statistics are required, as the technique directly learns mappings from observations and model parameters to likelihood-to-evidence ratios. Our framework allows for the usage of off-the-shelf neural architectures such as RESNET [122]. Experiments highlight the accuracy and robustness of our method.

SIMULATION EFFICIENCY We take the point of view that accuracy of the approximation is preferred over simulation cost. This is the case in many scientific disciplines which seek to reduce the uncertainty over a parameter of interest. Despite the experimental handicap, we have shown that existing simulation efficient approaches are not able to outperform our method in terms of accuracy with respect to a certain (and small) simulation budget.

4

Constraining Dark Matter with Stellar Streams and Machine Learning

The contents of this chapter are based on Hermans et al. [11].

A statistical analysis of the observed perturbations in the density of stellar streams can in principle set stringent constraints on the mass function of dark matter subhaloes, which in turn can be used to constrain the mass of the dark matter particle. However, the likelihood of a stellar density with respect to the stream and subhaloes parameters involves solving an intractable inverse problem which rests on the integration of all possible forward realisations implicitly defined by the simulation model. In order to infer the subhalo abundance, previous analyses have relied on Approximate Bayesian Computation (ABC) together with domain-motivated but handcrafted summary statistics. Here, we introduce a likelihood-free Bayesian inference pipeline based on the technique presented in Chapter 3. In particular, we will apply Amortized Approximate Likelihood Ratios (AALR, or NRE), previously introduced in Chapter 3, to automatically learn a mapping between the data and the simulator parameters. Thereby obviating the need to handcraft a possibly insufficient summary statistic. We apply the method to the simplified case where stellar streams are only perturbed by dark matter subhaloes, thus neglecting baryonic substructures, and describe several diagnostics that demonstrate the effectiveness of the new method and the statistical quality of the learned estimator.

4.1 INTRODUCTION

Cold Dark Matter (CDM) models [125, 126] predict a hierarchical collapse in which large haloes form through the merging of smaller dark matter clumps [127–129]. This process is driven by CDM's scale-free halo mass function [130, 131] and depends on the initial conditions of the matter power spectrum, which in turn anticipates the existence of dark matter haloes down to 10^{-4} $M_\odot$ [132]. Warm Dark Matter

(WDM) models [133–135] on the other hand, in which the dark matter particle is much lighter, influence structure formation down to the scale of dwarf galaxies. While at large scales the collapse in WDM is hierarchical as well, it becomes strongly suppressed below the half-mode mass scale of the corresponding dark matter model, where the non-negligible velocity dispersion of dark matter particles prevents haloes to form and smooths the density field instead [136]. Therefore, a powerful method of probing the particle nature of dark matter is to measure the abundances of the lowest mass subhaloes in our galaxy. While higher mass subhaloes will eventually initiate star formation and manifest themselves as dwarf galaxies, detecting low mass subhaloes ($\lesssim 10^9$ $M_\odot$) remains particularly hard since they either have very few faint stars or none at all.

Cold stellar streams that formed due to the tidal disruption of globular clusters by the Milky Way potential are a powerful probe for detecting and measuring the abundances of these low mass subhaloes [137–142]. When a subhalo flies past a stellar stream, it gravitationally perturbs the orbit of the stream stars around the point of closest approach, which leaves a visible imprint in the form of a region of low stellar density or a *gap*. Such gaps can be individually analyzed to infer the properties of a single subhalo perturber [142]. However, a stream is expected to encounter many subhalo impacts over its dynamical age, leading to complicated density structures that can be hard to separate into individual gaps. Therefore, a more pragmatic approach is to study the full stream density and statistically infer the subhalo abundance within the galactocentric radius of the stream [143].

Stream-subhalo encounters are described by various quantities such as the impact parameter, the flyby velocity of the subhalo, mass and size of the subhalo, and the time and angle of the subhalo impact. While simulating stream-subhalo encounters and their effects on the stellar density through these parameters is relatively straightforward, the forward model does not easily lend itself to statistical inference. The reason for this is that the likelihood of a stellar density with respect to these parameters involves solving an intractable inverse problem which rests on the integration of all possible forward realisations implicitly defined by the simulation model. It remains however possible to infer the underlying probabilities by relying on likelihood-free approximations [144]. From this perspective, Bovy, Erkal, and Sanders [143] applied Approximate Bayesian Computation (ABC) [117] to infer subhalo abundance using the power spectrum and bispectrum of the stream density as a summary statistic. With the same ABC technique, Banik et al. [145] and Banik et al. [146] applied the stream density power spectrum as a summary statistic to infer the particle mass of thermal relic dark matter.

It should be noted that ABC posteriors are *only* exact whenever the handcrafted summary statistic is *sufficient*, and the distance function chosen to express the similarity between observed and simulated data tends to 0, which in practice is never achievable. We address this issue by introducing a likelihood-free Bayesian inference pipeline based on amortized approximate likelihood ratios (AALR) [147], which automatically learns a mapping between the data and the simulator parameters by solving a tractable minimization problem. Afterwards, the learned estimator is able to accurately approximate the posterior density function of arbitrary stellar streams supported by the simulation model. By automatically learning this relation from data, we obviate the need to handcraft a possibly insufficient summary statistic, therefore enabling domain-scientists to pivot from solving the intractable inverse problem to the more natural forward modeling. In addition, we describe several diagnostics to inspect the statistical quality of the learned estimators with respect to the simulation model. We demonstrate the effectiveness of this method by inferring the particle mass of dark matter within the stellar stream framework.

4.2 MODELING OF STELLAR STREAMS

We use the `streampepperdf` simulator[1] that is based within the `galpy` framework [148] to model stream-subhalo interactions. Baryonic structures in our galaxy, namely, the bar, spiral arms and the Giant Molecular clouds can induce stream density variations similar to those caused by subhalo impacts [149–152]. However, owing to its retrograde orbit and a perigalacticon of ~ 14 kpc, the effect of the baryonic structures on the GD-1 stream [153] is expected to be subdominant compared to that by a CDM like population of subhalos. Therefore, we have used the GD-1 stream for our analyses and ignored the effects from the baryonic structures. Since the location of GD-1's progenitor is not known, we adopt the model presented in Webb and Bovy [154], which proposes that the progenitor cluster disrupted in its entirety approximately 500 Myr ago and resulted in the gap at the observed stream coordinate $\phi = -40^\circ$. The dynamical age of the GD-1 stream is also unknown and so following the arguments in [155], we consider all stream models in the range of 3-7 Gyr.

Our simulation model samples subhaloes in the sub-dwarf-galaxy mass range $[10^5 - 10^9]$ $M_\odot$, since density perturbations due to subhaloes less massive than 10^5 $M_\odot$ are below the level of noise in the current data. Warm Dark Matter (WDM) is modeled as a thermal relic candidate which is completely described by its particle mass. The

1 Available at https://github.com/jobovy/streamgap-pepper .

implementation of the subhaloes follows the same procedure as in [143, 145, 155].

We summarize the salient steps of the forward model for completeness. The WDM mass function is modeled following Lovell et al. [156]:

$$\left(\frac{dn}{dM}\right)_{\rm WDM} = \left(1+\gamma\frac{M_{\rm hm}}{M}\right)^{-\beta}\left(\frac{dn}{dM}\right)_{\rm CDM}, \tag{4.1}$$

where $\gamma = 2.7$, $\beta = 0.99$ and $\left(\frac{dn}{dM}\right)_{\rm CDM} \propto M^{-1.9}$. Here, $M_{\rm hm}$ is the half-mode mass that quantifies the scale below which the mass function is strongly suppressed. Both the CDM and WDM mass functions were obtained by fitting the subhaloes within a Milky Way like analogue from the Aquarius cosmological simulations [157]. Being dark matter only simulations, these mass functions do not account for the disruption of subhaloes due to baryonic structures such as the disk, which has been shown to be capable of destroying around $\sim 10-50\%$ of the subhaloes within the galactocentric radius of the GD-1 stream and in the mass range $10^{6.5}-10^{8.5}$ $M_\odot$ [158–162]. Moreover, the disrupted fraction of WDM subhaloes is expected to be even higher due to their lower concentrations. In this paper we have ignored subhalo disruptions due to baryonic effects and defer a full analysis to a future publication.

For each simulated stream density, we consider the region $-34^\circ < \phi < 10^\circ$ in the observed coordinate frame, and normalize the stream density by dividing it by the mean density. The latter step is different from what was done in [143, 155], where the authors normalize the stream density by dividing it by a 3$^{\rm rd}$ order polynomial fit. We tested that both normalization procedures gave similar results. This was also demonstrated in Bovy, Erkal, and Sanders [143], where they found that changing the order of the smoothing polynomial did not significantly affect the (ABC) posterior. Finally, noise is added to every simulated stream density by sampling a Gaussian realisation of the noise from the observed GD-1 data from Boer, Erkal, and Gieles [163].

4.3 METHOD

4.3.1 *Statistical formalism*

This work considers two inference scenarios. In the first we jointly infer the WDM mass $m_{\rm WDM}$ and the stream age $t_{\rm age}$. The second scenario solely considers $m_{\rm WDM}$ while marginalizing the stream age. Because our methodology generalizes to various domains, we ease the discussion by simplifying the nomenclature into the following concepts:

Target parameters $\boldsymbol{\vartheta}$ denote the main parameters of our simulation model. Depending on the inference scenario at hand, $\boldsymbol{\vartheta} \triangleq (m_{\text{WDM}}, t_{\text{age}})$ or $\boldsymbol{\vartheta} \triangleq (m_{\text{WDM}})$. Given the Bayesian perspective of this analysis, we define the priors over the WDM mass m_{WDM} and stream age t_{age} to be $\texttt{uniform}(1,\ 50)$ keV and $\texttt{uniform}(3,\ 7)$ billion years (Gyr) respectively. The upper bound of 50 keV is justified since it corresponds to a half-mode mass of $\sim 4 \times 10^4$ $\text{M}_\odot$, which is below the sensitivity of stellar streams given current observational uncertainties.

Observables $\boldsymbol{x}$ encapsulate the simulated stellar density of mock streams and the *observed* GD-1 density. An observable is encoded as a 66-dimensional vector along the linear angle ϕ between -34 and 10 degrees.

Nominal value $\boldsymbol{\vartheta}^*$ or groundtruth used to simulate the observable x of a mock stream, i.e., $\boldsymbol{x} \sim p(\boldsymbol{x} \mid \boldsymbol{\vartheta}^*)$.

Nuisance parameters η such as the impact angle and subhalo mass are not of direct interest, but their (random) effects must be accounted for to infer $\boldsymbol{\vartheta}$ [164]. However, this leaves us with the likelihood function $p(\boldsymbol{x} \mid \boldsymbol{\vartheta}, \eta)$. Given the Bayesian perspective of this work, we incorporate nuisance parameter uncertainty [165] by integration. The priors associated with the nuisance parameters are implicitly defined through the simulation model.

4.3.2 *Motivation*

Our multi-faceted simulation model induces an extensive space of possible execution paths, which, for example, correspond to randomly sampled dark matter haloes that impact the stellar stream throughout its evolution. The evaluation of the likelihood $p(\boldsymbol{x} \mid \boldsymbol{\vartheta})$ of an observable $\boldsymbol{x}$ therefore involves amongst others the integration over a large variety of possible collision histories that are consistent with ϑ. Given the high-dimensional nature of this integral, directly evaluating data likelihoods is intractable.

A common Bayesian approach to address the intractability of the likelihood is to reduce the dimensionality of an observable $\boldsymbol{x}$ by means of a summary statistic $s(\boldsymbol{x})$. The reduction in dimensionality allows the posterior to be approximated numerically by collecting samples $\boldsymbol{\vartheta} \sim p(\boldsymbol{\vartheta})$ for which observables produced by the forward model $s(\boldsymbol{x}) \sim p(\boldsymbol{x} \mid \boldsymbol{\vartheta})$ are similar, in terms of some distance, to the compressed representation of the observed data $s(\boldsymbol{x}_o)$. This rejection sampling scheme is commonly referred to as *Approximate Bayesian Computation* [117] (ABC) and is, as the name indicates, *approximate*. Although the compression of $\boldsymbol{x}$ into a summary statistic makes the numerical

approximation of the posterior tractable, it may reduce the statistical power of an analysis because the selected summary statistic often destroys relevant information. In fact, ABC is *only* exact whenever the summary statistic is *sufficient* and the distance function chosen to express the similarity between between $s(\boldsymbol{x})$ and $s(\boldsymbol{x}_o)$ tends to 0. This is in practice never achievable because for a given simulation budget *(i)* a small acceptance threshold severely impacts the rate at which proposed samples are accepted, affecting the approximation of the posterior density function, and *(ii)* the *assumed* sufficiency of the summary statistic is virtually never thoroughly demonstrated in practice. Despite these shortcomings, ABC has been fruitfully applied in cosmology to constrain dark matter models within the context of stellar streams [145, 146, 166], and more recently gravitational lensing [167].

This work tackles the *intractability* of the likelihood from a different perspective. Instead of manually crafting a summary statistic and a distance function with a specific acceptance threshold, we propose to learn an amortized mapping from target parameters $\boldsymbol{\vartheta}$ and observables $\boldsymbol{x}$ to posterior densities by solving a *tractable* minimization problem. The learned mapping has the potential to increase the statistical power of an analysis since the procedure, in contrast to ABC, *automatically* attempts to learn an internal sufficient summary statistic of the data. The automated procedure therefore enables domain-experts to solely focus on the forward modeling of the phenomena of interest, because the method does not require any consideration whether synthetic observables are compressible into low-dimensional summary statistics. Although the proposed method treats the simulation model as a black box, we would like to point out that it is possible to improve the efficiency of the minimization problem, provided that latent information can be extracted from the simulation model [15, 168, 169], albeit at some implementation cost.

4.3.3 *Inference*

The Bayesian paradigm finds model parameters compatible with observation by computing the *posterior*

$$p(\boldsymbol{\vartheta} \mid \boldsymbol{x}) = \frac{p(\boldsymbol{\vartheta})p(\boldsymbol{x} \mid \boldsymbol{\vartheta})}{p(\boldsymbol{x})}. \tag{4.2}$$

Evaluating the posterior density for a given target parameter $\boldsymbol{\vartheta}$ and an observable x in our setting is not possible because the likelihood $p(\boldsymbol{x} \mid \boldsymbol{\vartheta})$ is per definition intractable. To enable the tractable evaluation of the posterior, we have to rely on likelihood-free surrogates for key components in Bayes' rule. Note that Equation 4.2 can be factorized

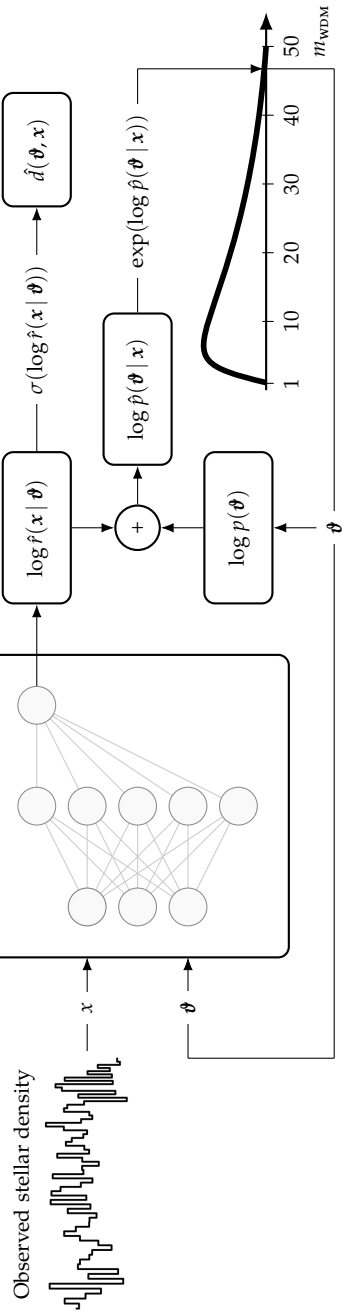

Figure 4.1: Graphical representation of the inference procedure after training the ratio estimator (neural network). The ratio estimator accepts a target parameter $\boldsymbol{\vartheta}$ and an observable $\boldsymbol{x}$ as inputs, which are subsequently used to approximate the likelihood-to-evidence ratio $\hat{r}(\boldsymbol{x} \mid \boldsymbol{\vartheta})$. The discriminator output $\hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x})$ — the sigmoidal projection $\sigma(\cdot)$ of $\log \hat{r}(\boldsymbol{x} \mid \boldsymbol{\vartheta})$ — is only used during training. Given that the log prior probability of $\boldsymbol{\vartheta}$ is a tractable quantity, we can easily approximate the log posterior probability $\log \hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})$ by adding the approximated log likelihood-to-evidence ratio. Taking the exponent of the produced quantity results in a direct estimate of the posterior density. This procedure can be repeated for arbitrary target parameters $\boldsymbol{\vartheta}$ supported by the prior. It should be noted that the neural network depicted here is an abstract representation. Our technique does not put any constraints on the architecture of the neural network. It is therefore possible to use of-the-shelf architectures of arbitrary complexity available in the Machine Learning literature.

into the product of the tractable prior probability and the intractable likelihood-to-evidence ratio $r(\boldsymbol{x} \mid \boldsymbol{\vartheta})$:

$$p(\boldsymbol{\vartheta} \mid \boldsymbol{x}) = p(\boldsymbol{\vartheta}) \frac{p(\boldsymbol{x} \mid \boldsymbol{\vartheta})}{p(\boldsymbol{x})} = p(\boldsymbol{\vartheta}) \frac{p(\boldsymbol{\vartheta}, \boldsymbol{x})}{p(\boldsymbol{\vartheta}) p(\boldsymbol{x})} = p(\boldsymbol{\vartheta}) r(\boldsymbol{x} \mid \boldsymbol{\vartheta}). \tag{4.3}$$

[147] show that an amortized estimator $\hat{r}(\boldsymbol{x} \mid \boldsymbol{\vartheta})$ of the intractable likelihood-to-evidence ratio can be obtained by training a discriminator $d(\boldsymbol{\vartheta}, \boldsymbol{x})$ with inputs $\boldsymbol{\vartheta}$ and x, to distinguish between samples from the joint $p(\boldsymbol{\vartheta}, \boldsymbol{x})$ with class label 1 and samples from the product of marginals $p(\boldsymbol{\vartheta})p(\boldsymbol{x})$ with class label 0 using a discriminative criterion such as the binary cross entropy. Whenever the training criterion is minimized, the authors theoretically demonstrate that the optimal discriminator $d(\boldsymbol{\vartheta}, \boldsymbol{x})$ models the Bayes optimal decision function

$$d(\boldsymbol{\vartheta}, \boldsymbol{x}) = \frac{p(\boldsymbol{\vartheta}, \boldsymbol{x})}{p(\boldsymbol{\vartheta}, \boldsymbol{x}) + p(\boldsymbol{\vartheta}) p(\boldsymbol{x})}. \tag{4.4}$$

Subsequently, given a model parameter $\boldsymbol{\vartheta}$ and an observable x, we can use the discriminator as a density *ratio estimator* to compute the likelihood-to-evidence ratio

$$r(\boldsymbol{x} \mid \boldsymbol{\vartheta}) = \frac{1 - d(\boldsymbol{\vartheta}, \boldsymbol{x})}{d(\boldsymbol{\vartheta}, \boldsymbol{x})} = \frac{p(\boldsymbol{\vartheta}, \boldsymbol{x})}{p(\boldsymbol{\vartheta}) p(\boldsymbol{x})} = \frac{p(\boldsymbol{x} \mid \boldsymbol{\vartheta})}{p(\boldsymbol{x})}. \tag{4.5}$$

However, the computation of this formulation suffers from significant numerical issues in the saturating regime where the output of the discriminator tends to 0. Considering that $\log r(\boldsymbol{x} \mid \boldsymbol{\vartheta}) = \operatorname{logit}(d(\boldsymbol{\vartheta}, \boldsymbol{x}))$ for classifiers with a *sigmoidal* projection at the output, it is possible to directly obtain $\log r(\boldsymbol{x} \mid \boldsymbol{\vartheta})$ from the classifier by extracting the quantity before the sigmoidal operation. This strategy ensures that the approximation of $\log r(\boldsymbol{\vartheta} \mid \boldsymbol{x})$ is numerically stable. In addition, randomly shuffling $\boldsymbol{\vartheta}$ in a batch $\boldsymbol{\vartheta}, \boldsymbol{x} \sim p(\boldsymbol{\vartheta}, \boldsymbol{x})$ instead of drawing a new samples from the product of marginals significantly aids the convergence rate of the discriminator. After training, estimates of the posterior probability density function can be approximated for arbitrary (without retraining) target parameters $\boldsymbol{\vartheta}$ and observables $\boldsymbol{x}$ by computing

$$\log p(\boldsymbol{\vartheta} \mid \boldsymbol{x}) \approx \log p(\boldsymbol{\vartheta}) + \log \hat{r}(\boldsymbol{x} \mid \boldsymbol{\vartheta}), \tag{4.6}$$

provided that $\boldsymbol{\vartheta}$ and $\boldsymbol{x}$ are supported by the prior $p(\boldsymbol{\vartheta})$ and the marginal model $p(\boldsymbol{x})$ respectively, thereby enabling consistent and fast likelihood-free posterior inference. Figure 4.1 provides a graphical overview. We refer the reader to [147] or our GitHub repository for implementation details.

The ratio estimator can likewise be adapted to compute a credible region (CR) at a desired level of uncertainty α by constructing a region Θ in the model parameter space which satisfies

$$\int_{\Theta} p(\boldsymbol{\vartheta}) r(\boldsymbol{x} \mid \boldsymbol{\vartheta}) \; \mathrm{d}\, \boldsymbol{\vartheta} = 1 - \alpha. \tag{4.7}$$

Since many such regions Θ exist, we select the highest posterior density region, which is the smallest credible region.

Although our analysis focuses on the Bayesian paradigm, it is possible use the ratio estimator in a frequentist setting [15, 55]. The likelihood-ratio $\lambda(\boldsymbol{x} \mid \boldsymbol{\vartheta}_0, \boldsymbol{\vartheta}_1)$ between two hypotheses $\boldsymbol{\vartheta}_0$ and $\boldsymbol{\vartheta}_1$ can easily be computed with the ratio estimator as the denominators of $r(\boldsymbol{x} \mid \boldsymbol{\vartheta}_0)$ and $r(\boldsymbol{x} \mid \boldsymbol{\vartheta}_1)$ cancel out, i.e.,

$$\lambda(\boldsymbol{x} \mid \boldsymbol{\vartheta}_0, \boldsymbol{\vartheta}_1) = \frac{p(\boldsymbol{x} \mid \boldsymbol{\vartheta}_0)}{p(\boldsymbol{x} \mid \boldsymbol{\vartheta}_1)} = \frac{r(\boldsymbol{x} \mid \boldsymbol{\vartheta}_0)}{r(\boldsymbol{x} \mid \boldsymbol{\vartheta}_1)}. \tag{4.8}$$

The same strategy applies to the likelihood-ratio [170] test statistic for a specific observable $\boldsymbol{x}$

$$-2 \log \lambda(\boldsymbol{\vartheta}) = -2 \log \frac{p(\boldsymbol{x} \mid \boldsymbol{\vartheta})}{p(\boldsymbol{x} \mid \hat{\boldsymbol{\vartheta}})}, \tag{4.9}$$

where the maximum likelihood estimate $\hat{\boldsymbol{\vartheta}}$ is

$$\hat{\boldsymbol{\vartheta}} = \arg\max_{\boldsymbol{\vartheta}} r(\boldsymbol{x} \mid \boldsymbol{\vartheta}). \tag{4.10}$$

The test statistic can thus be expressed [55] as

$$-2 \log \lambda(\boldsymbol{\vartheta}) = -2 \left[\log r(\boldsymbol{x} \mid \boldsymbol{\vartheta}) - \log r(\boldsymbol{x} \mid \hat{\boldsymbol{\vartheta}}) \right]. \tag{4.11}$$

As a result of Wilks' theorem [171], we can directly convert the test statistic into a confidence level (CL) under the assumption that the statistic is $\mathcal{X}^2_k$-distributed with k degrees of freedom (in function of $\boldsymbol{\vartheta}$'s dimensionality).

4.3.4 *Diagnostics*

Before making any scientific conclusion, it is crucial to verify the result of the involved statistical computation. This is especially challenging in the likelihood-free setting because evaluating the likelihood is intractable. The following subsections describe several diagnostics to assess the quality of the amortized ratio estimates. No additional training or fine-tuning is applied as this would change the statistical properties of the ratio estimator.

4.3.4.1 *Proper probability density*

A ratio estimator $\hat{r}(x \mid \boldsymbol{\vartheta})$ which correctly models the true likelihood-to-evidence ratio should satisfy

$$\int_{\boldsymbol{\vartheta}} p(\boldsymbol{\vartheta}) \hat{r}(x \mid \boldsymbol{\vartheta}) \, \mathrm{d}\, \boldsymbol{\vartheta} \approx 1 \; \forall x. \tag{4.12}$$

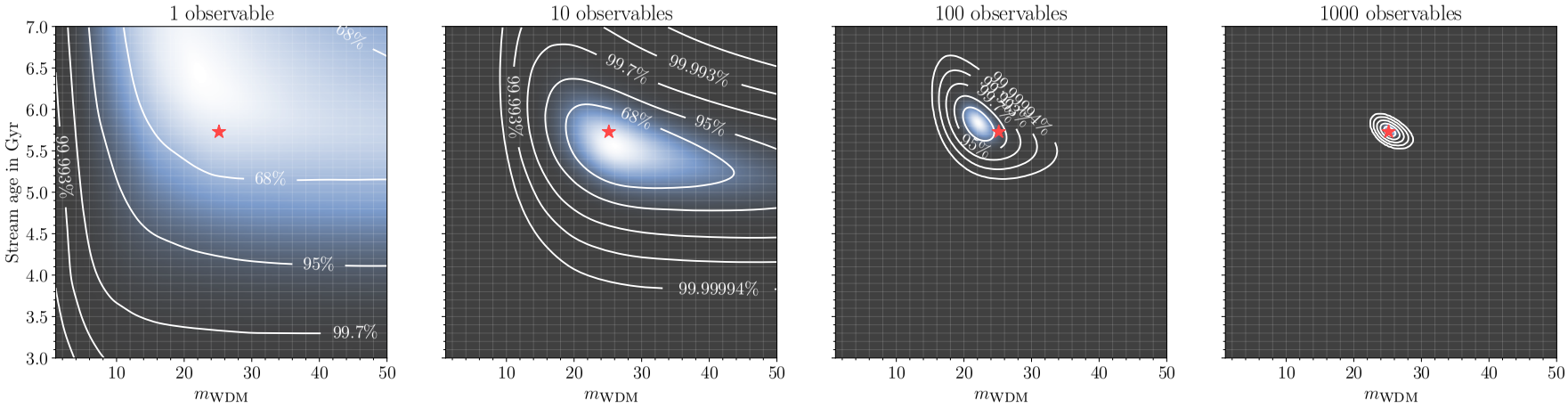

Figure 4.2: Demonstration of the mode convergence diagnostic described in Section 4.3.4.3. The figures show, from left to right, the posteriors for 1, 10, 100 and 1000 independent and identically distributed mock GD-1 observables. Every figure adopts the same nominal value or groundtruth, which is highlighted by the red star. As the amount of observables increases, the posteriors are becoming increasingly more tight around the nominal value. This indicates that the individual posteriors do not, in expectation, introduce significant bias for independent and identically distributed observables. </>

The diagnostic should be applied to observables x of an evaluation dataset *and* real observables x_o. Passing the diagnostic on the evaluation dataset, while failing on x_o indicates that x_o is not supported by the marginal model $p(\boldsymbol{x})$, because ratio estimates in this regime are undefined and can therefore take on arbitrary values.

4.3.4.2 *Coverage*

Coverage quantifies the reliability of a statistical method to reconstruct the nominal value [172–175]. The approximation of the ratio estimator can thus be assessed by determining whether the empirical coverage probability matches the nominal coverage probability, which corresponds to the confidence level $1 - \alpha$. The empirical coverage probability is estimated using samples from a (large) presimulated evaluation dataset. This evaluation dataset consists of samples $\boldsymbol{\vartheta}, \boldsymbol{x} \sim p(\boldsymbol{\vartheta}, \boldsymbol{x})$. For every pair $(\boldsymbol{\vartheta}, \boldsymbol{x})$ in the evaluation dataset, we compute the corresponding credible or confidence interval. The fraction of samples for which the groundtruth was contained within the interval corresponds to the empirical coverage probability. If the empirical coverage probability $\geq 1 - \alpha$, then the ratio estimator passes the diagnostic. It is of course desirable that the empirical coverage probability of the ratio estimator converges to the confidence level. A substantially larger empirical coverage probability corresponds to intervals which are overly conservative. This implies that the ratio estimates are wrong, *but*, that in expectation the estimated posteriors are conservative, which is not an undesirable property. It should be noted that coverage can only be computed efficiently because our ratio estimator amortizes the estimation of the likelihood-to-evidence ratio. An equivalent study for ABC would have a significant computational cost.

4.3.4.3 *Convergence of the mode towards the nominal value*

The diagnostic is based on the idea that the maximum a posteriori (MAP) estimate converges towards the nominal value $\boldsymbol{\vartheta}^*$ for an increasing number of independent and identically distributed observables $\boldsymbol{x} \sim p(\boldsymbol{x} \mid \boldsymbol{\vartheta}^*)$. If the approximation of $\hat{r}(\boldsymbol{x} \mid \boldsymbol{\vartheta})$ is correct, the MAP should in the limit coincide with the nominal value $\boldsymbol{\vartheta}^*$. Let $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$

be a set of i.i.d. observables. To compute the MAP, we need $p(\boldsymbol{\vartheta}\,|\mathcal{X})$. As noted by Brehmer et al. [15], Bayes' rule can be rewritten as

$$p(\boldsymbol{\vartheta}\,|\mathcal{X}) = \frac{p(\boldsymbol{\vartheta})\prod_{x\in\mathcal{X}} p(\boldsymbol{x}\,|\,\boldsymbol{\vartheta})}{\int p(\boldsymbol{\vartheta}')\prod_{x\in\mathcal{X}} p(\boldsymbol{x}\,|\,\boldsymbol{\vartheta}')\ \mathrm{d}\,\boldsymbol{\vartheta}'} \tag{4.13}$$

$$= p(\boldsymbol{\vartheta})\left[\int p(\boldsymbol{\vartheta}')\prod_{x\in\mathcal{X}}\frac{p(\boldsymbol{x}\,|\,\boldsymbol{\vartheta}')}{p(\boldsymbol{x}\,|\,\boldsymbol{\vartheta})}\ \mathrm{d}\,\boldsymbol{\vartheta}'\right]^{-1} \tag{4.14}$$

$$\approx p(\boldsymbol{\vartheta})\left[\int p(\boldsymbol{\vartheta}')\prod_{x\in\mathcal{X}}\frac{\hat{r}(\boldsymbol{x}\,|\,\boldsymbol{\vartheta}')}{\hat{r}(\boldsymbol{x}\,|\,\boldsymbol{\vartheta})}\ \mathrm{d}\,\boldsymbol{\vartheta}'\right]^{-1}. \tag{4.15}$$

The integral can be estimated through Monte Carlo sampling. By checking whether the MAP concurs with the nominal value, we effectively probe the bias. Ideally, this diagnostic should be repeated for various groundtruths to inspect the behaviour of the ratio estimator over the complete model parameter space. In some settings however, the posterior may be multi-modal. In such scenarios the convergence of the mode(s) instead of the MAP should be assessed. A trial of the diagnostic is shown in Figure 4.2.

4.3.4.4 *Receiver operating characteristic*

We note that $\hat{r}(\boldsymbol{x}\,|\,\boldsymbol{\vartheta})$ is only exact whenever

$$p(\boldsymbol{x})\frac{p(\boldsymbol{x}\,|\,\boldsymbol{\vartheta})}{p(\boldsymbol{x})} = p(\boldsymbol{x})\hat{r}(\boldsymbol{x}\,|\,\boldsymbol{\vartheta}) = p(\boldsymbol{x}\,|\,\boldsymbol{\vartheta}), \tag{4.16}$$

is satisfied for all $\boldsymbol{\vartheta}$ and x. Although $p(\boldsymbol{x})$ and $p(\boldsymbol{x}\,|\,\boldsymbol{\vartheta})$ cannot be evaluated directly, it remains possible to sample from these densities. Given samples from the reweighted marginal model $p(\boldsymbol{x})\hat{r}(\boldsymbol{x}\,|\,\boldsymbol{\vartheta})$, and from a specific likelihood $p(\boldsymbol{x}\,|\,\boldsymbol{\vartheta})$, the idea is that $\hat{r}(\boldsymbol{x}\,|\,\boldsymbol{\vartheta})$ can only be equivalent to $r(\boldsymbol{x}\,|\,\boldsymbol{\vartheta})$ whenever a classifier tasked to distinguish between $p(\boldsymbol{x})\hat{r}(\boldsymbol{x}\,|\,\boldsymbol{\vartheta})$ and $p(\boldsymbol{x}\,|\,\boldsymbol{\vartheta})$, cannot extract any predictive features. The discriminative performance of this classifier can be assessed by means of a *Receiver Operating Characteristic* (ROC) curve. A diagonal ROC, which has an *Area Under Curve* (AUC) of 0.5, corresponds to a classifier which is insensitive. In that case, the ratio estimator passes the diagnostic. We emphasize that the ratio estimator can incorrectly pass the diagnostic whenever the classifier is not sufficiently expressive.

4.3.4.5 *Alternative diagnostics*

Our list of diagnostics is not exhaustive. Some diagnostics are specific to our ratio estimator and can only be computed efficiently because ratio estimates are amortized. In fact, the development of diagnostics for the simulation-based inference literature is an active area of research. For more recent methodologies we refer the reader to Talts et al. [176] and Dalmasso et al. [177].

4.3.5 *Overview of the proposed recipe*

1. Simulate a train and test dataset by sampling from the joint $p(\boldsymbol{\vartheta}, \boldsymbol{x})$. This is done by drawing samples $\boldsymbol{\vartheta} \sim p(\boldsymbol{\vartheta})$ and conditioning the simulation model on $\boldsymbol{\vartheta}$ to generate observables $\boldsymbol{x} \sim p(\boldsymbol{x} \mid \boldsymbol{\vartheta})$. These simulations can be parallelised arbitrarily because the samples are drawn independently. The effective number of simulations depends on the problem at hand. In practice additional simulations were added whenever the ratio estimators did not pass the coverage diagnostic, or, if we found over-fitting to be a significant issue during training.

2. Train several discriminators $d(\boldsymbol{\vartheta}, \boldsymbol{x})$ on the previously simulated dataset. This has several uses. First, the ratio estimators can be ensembled to reduce the variance of the approximation. Secondly, as there is only a single true likelihood-to-evidence ratio $r(\boldsymbol{x} \mid \boldsymbol{\vartheta})$, the variability of ratio estimates within the ensemble can be used to quickly assess the convergence. A significant deviation in the ratio estimates is indicative of a ill-tuned optimization procedure.

3. Probe the trained ratio estimators for flaws with the diagnostics. Afterwards, apply the diagnostic described in Section 4.3.4.1 to the observable(s) $\boldsymbol{x}_o$.

4. Compute the posterior $\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x}_o) = p(\boldsymbol{\vartheta})\hat{r}(\boldsymbol{x}_o \mid \boldsymbol{\vartheta})$ and the desired credible or confidence intervals.

4.4 EXPERIMENTS AND RESULTS

We demonstrate the usage of our technique on various synthetic realisations of GD-1. Diagnostics are applied to probe the statistical quality of the approximated posteriors under the specified simulation model. By comparing our technique against ABC, we highlight the gain in statistical power our technique can bring to the scientific community. We compute *preliminary* constraints on m_{WDM} based on observations of GD-1 by *Gaia* proper motions [178, 179] and *Pan-STARRS* photometry [180]. It should be noted these constraints only hold under the assumed simulation model. An analysis of (simulation) model misspecification is outside the scope of this work.

4.4.1 *Setup*

Empirical coverage probability						
Architecture	68% CR	95% CR	99.7% CR	68% CL	95% CL	99.7% CL
$\hat{r}(\boldsymbol{x} \mid \boldsymbol{\vartheta})$ with $\boldsymbol{\vartheta} \triangleq (m_{\text{WDM}})$						
MLP	0.685 $\pm$0.004	0.954 $\pm$0.002	0.997 $\pm$0.001	0.750 $\pm$0.004	0.968 $\pm$0.002	0.999 $\pm$0.000
MLP-BN	0.687 $\pm$0.006	0.951 $\pm$0.002	0.997 $\pm$0.000	0.760 $\pm$0.003	0.970 $\pm$0.002	0.999 $\pm$0.000
RESNET-18	0.667 $\pm$0.004	0.943 $\pm$0.002	0.996 $\pm$0.001	0.721 $\pm$0.005	0.960 $\pm$0.002	0.997 $\pm$0.000
RESNET-18-BN	0.672 $\pm$0.004	0.945 $\pm$0.001	0.996 $\pm$0.001	0.736 $\pm$0.003	0.961 $\pm$0.002	0.998 $\pm$0.000
RESNET-50	0.671 $\pm$0.005	0.947 $\pm$0.003	0.996 $\pm$0.001	0.726 $\pm$0.005	0.963 $\pm$0.000	0.998 $\pm$0.001
RESNET-50-BN	0.678 $\pm$0.004	0.949 $\pm$0.004	0.996 $\pm$0.001	0.743 $\pm$0.002	0.966 $\pm$0.001	0.998 $\pm$0.000
$\hat{r}(\boldsymbol{x} \mid \boldsymbol{\vartheta})$ with $\boldsymbol{\vartheta} \triangleq (m_{\text{WDM}}, t_{\text{age}})$						
MLP	0.685 $\pm$0.005	0.953 $\pm$0.002	0.998 $\pm$0.000	0.752 $\pm$0.003	0.968 $\pm$0.001	0.999 $\pm$0.000
MLP-BN	0.685 $\pm$0.004	0.952 $\pm$0.003	0.997 $\pm$0.000	0.758 $\pm$0.003	0.970 $\pm$0.002	0.999 $\pm$0.000
RESNET-18	0.666 $\pm$0.005	0.945 $\pm$0.002	0.995 $\pm$0.001	0.724 $\pm$0.005	0.961 $\pm$0.002	0.998 $\pm$0.000
RESNET-18-BN	0.671 $\pm$0.003	0.945 $\pm$0.003	0.996 $\pm$0.001	0.736 $\pm$0.004	0.961 $\pm$0.002	0.998 $\pm$0.000
RESNET-50	0.674 $\pm$0.006	0.944 $\pm$0.002	0.996 $\pm$0.001	0.740 $\pm$0.004	0.970 $\pm$0.002	0.999 $\pm$0.000
RESNET-50-BN	0.677 $\pm$0.004	0.947 $\pm$0.003	0.997 $\pm$0.000	0.738 $\pm$0.004	0.970 $\pm$0.002	0.999 $\pm$0.000

Table 4.1: Results of the overage diagnostic. Architectures with the BN suffix make use of Batch Normalization. For all ratio estimator architectures, we report Bayesian credible regions and frequentist confidence intervals. Although credible regions do not necessarily have a frequentist interpretation, they are in fact much closer to the nominal coverage probability compared to the confidence intervals. On the contrary, the confidence intervals have coverage, but are slightly conservative. Our analyses will therefore focus on constraints based on confidence intervals. </>

SIMULATIONS We follow the simulation formalism described in the section above using the priors defined in Section 4.3.1. 10 million pairs $(\boldsymbol{\vartheta}, \boldsymbol{x}) \sim p(\boldsymbol{\vartheta}, \boldsymbol{x})$ are drawn from the simulation model for training, and 100,000 for testing. The simulations in the training dataset are reused in our ABC analyses.

RATIO ESTIMATOR TRAINING All architectures are trained with identical hyperparameter settings. No exhaustive hyperparameter optimization or architecture-search was conducted. Options such as weight-decay and batch-normalization (BN) [181] were evaluated to reduce over-fitting. All ratio estimators use SELU [182] activations and were trained using the ADAMW [183] optimizer for 50 epochs with a batch-size of 4096. We found that larger batch-sizes, for our setting, generalized better. We empirically found SELU and ELU activations to be preferable over RELU-like activations, because the approximation of the posterior density function was generally smoother. Nevertheless, architectural aspects should be evaluated on a per-problem basis. This work considers 3 main architectures; *(i)* a simple feedforward MLP, and variants to RESNET [122] such as *(ii)* RESNET-18 and *(iii)* RESNET-50. Both use 1 dimensional convolutions without dilations since the usage of dilated convolutions did not yield any significant improvements in terms of test loss. Because our methodology treats $\boldsymbol{\vartheta}$ as an input feature, we cannot easily condition the convolutional layers of the RESNET-based architectures on $\boldsymbol{\vartheta}$. This would require conditional convolutions [184] or hypernetworks [185] to generate specialized kernels for a given $\boldsymbol{\vartheta}$. To retain the simplicity of our architecture, we inject the dependency on $\boldsymbol{\vartheta}$ in the fully connected trunk of the convolutional ratio estimators. Other architectural considerations were not explored.

The same hyperparameters are used across all architectures. We did not explore specific settings for every individual architecture, demonstrating the robustness of our technique. A learning rate of 0.0001 with a batch-size of 4096 and a weight-decay factor of 0.1 was used during training. The ratio estimators do not use dropout [186]. The remaining hyperparameters (e.g., of Batch Normalization) were set to the *PyTorch* defaults. </>.

REJECTION APPROXIMATE BAYESIAN COMPUTATION Instead of using the stream density power as summary statistics as in Bovy, Erkal, and Sanders [143] and Banik et al. [155], we construct a summary statistics based on the stream density itself. We divide the synthetic observable x (with $n = 66$ bins) by the observable of interest $\boldsymbol{x}_o$ to

obtain the bin-wise stellar density ratio $d = \boldsymbol{x} \,/\, \boldsymbol{x}_o$. Our summary statistic and distance function are jointly expressed as

$$s(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} (d_i - \bar{d})^2, \tag{4.17}$$

where $\bar{d}$ is the mean stellar density ratio. Ideally, if both observables match perfectly, then $s(\boldsymbol{x}) = 0$. The acceptance threshold is tuned such that for any given observable of interest $\boldsymbol{x}_o$, the number of accepted posterior samples is 0.1% of the simulation budget, therefore yielding the smallest threshold with respect to the specified acceptance rate. This corresponds to approximately 10,000 posterior samples. Our goal is to highlight generic aspects of ABC with respect to the proposed method in terms of tuning of the analyses, and its statistical quality *for the given simulation budget*. We emphasize that several scheduling and threshold strategies for ABC exist in the literature, see e.g. [187, 188]. We opted here for a method that is based on the same number of simulations used for training the neural network. The threshold was chosen heuristically to obtain sufficiently smooth posteriors across the entire parameter space, and was not tuned depending on the WDM mass and stream age. This is different from the targeted convergence check and simulation strategy in previous ABC-based streams analyses [143, 145, 146, 166]. We cannot exclude that the ABC results shown here could further improve by significantly increasing the number of simulations beyond what was needed for the neural network training. This is beyond the scope of the current work.

4.4.2 *Statistical quality*

We now assess the statistical properties of the trained ratio estimators. For every architecture, we select the weights which achieved the smallest validation-loss.

PROPER PROBABILITY DENSITY The computational cost of the integration does not allow us to do an exhaustive analysis. Instead, we apply the diagnostic to 1000 randomly sampled observables. As before, we repeat the experiment 10 times. The following results were obtained: MLP (1.023 ± 0.11), MLP-BN (1.037 ± 0.09), RESNET-18 (1.00 ± 0.02), RESNET-18-BN (0.973 ± 0.03), RESNET-50 (0.993 ± 0.03), and RESNET-50-BN (1.001 ± 0.04) </>. Although the average integrated area under the approximated posterior density functions approaches 1 for all ratio estimator architectures, the results suggest that the approximations of the RESNET-based architectures are more robust. A more careful analysis of the integrated areas, presented in Figure 4.3, confirms this. Interestingly, the integrated areas for RESNET

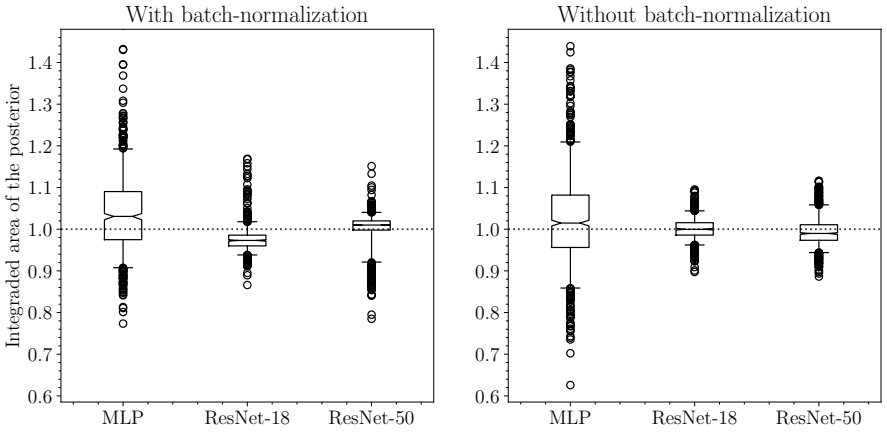

Figure 4.3: Result of the proper probability density diagnostic. As expected, high-capacity models (RESNET) have tighter approximations compared to the MLP architectures. An interesting discrepancy between the usage of with and without batch normalization is observed. *(Left)* With batch-normalization. *(Right)* Without batch-normalization. </>

architectures *with* batch-normalization have a larger spread compared to their counterparts without batch-normalization. Our evaluations on GD-1 will therefore focus on RESNET-based architectures without batch-normalization.

COVERAGE Table 4.1 summarizes the empirical expected *frequentist* coverage probability of the ratio estimators. For every ratio estimator, we compute the credible and confidence intervals as described in Section 4.3.3. For both paradigms, we evaluate the interval construction on 10,000 observables, which is repeated 10 times. The empirical coverage probability of a ratio estimator is therefore based on approximately 100,000 observables in total. We empirically find that MLP-based architectures have coverage under both Bayesian credible and frequentist confidence intervals. This is not the case for RESNET-based architectures. It is noteworthy that the empirical coverage probability of the credible regions are much closer to the nominal coverage probabilities compared to their frequentist counterparts. Additional statistical power could therefore be extracted if the credible regions could be tuned to sufficiently cover the groundtruth at a given nominal coverage probability.

RECEIVER OPERATING CHARACTERISTIC We now directly probe the correctness of the approximated likelihood-to-evidence ratios. Every ratio estimator is evaluated on 10 uniformly sampled test-hypotheses. 10,000 observables are drawn from every test-hypothesis.

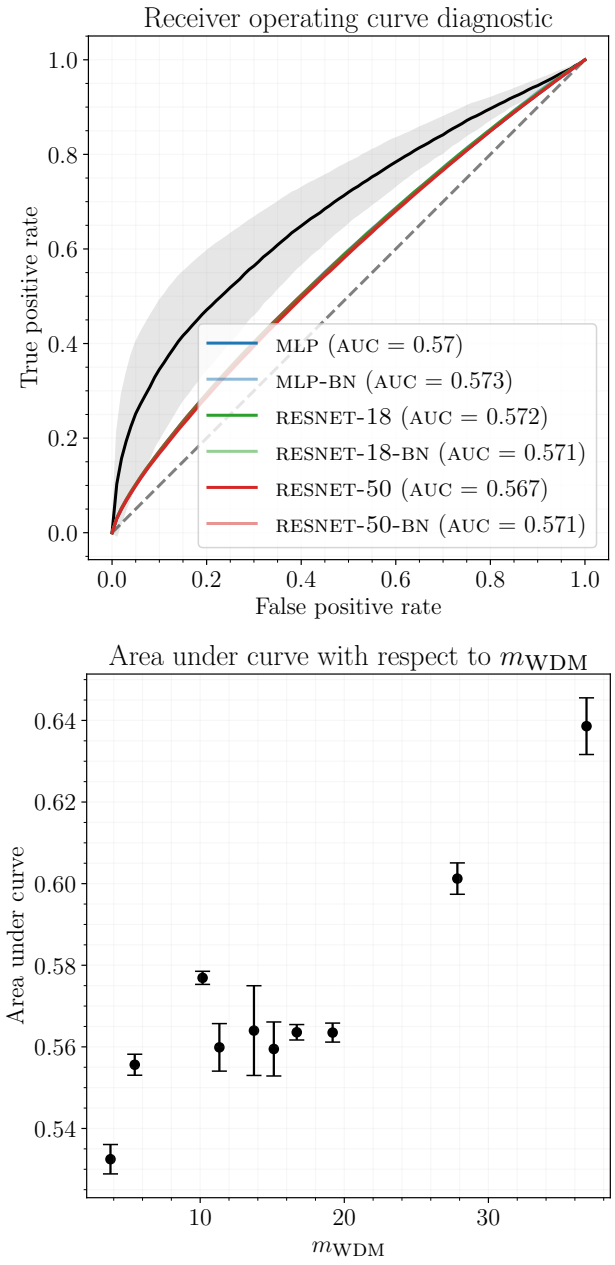

Figure 4.4: Summary of the receiver operating curve diagnostic. *(Top)* Area Under Curve (AUC) for all test-hypotheses. A baseline measurement, indicated by the black line, does not reweigh the marginal model. Although the ratio estimators perform significantly better compared to the baseline, the diagnostic indicates that all ratio estimators do not perfectly approximate the likelihood-to-evidence ratio (since AUC $\neq$ 0.5). This is not necessarily an issue, because the coverage diagnostic demonstrates that the confidence intervals are conservative. *(Bottom)* Average AUC of the test-hypotheses under consideration. Larger values of m_{WDM} are associated with a degraded quality of the ratio estimates. </>

For every test-hypothesis, we repeat the computation of the area under curve 10 times to account for the stochastic training of the classifier tasked to distinguish between samples from the reweighted marginal model and samples from the test-hypothesis. Figure 4.4 summarizes the results. In general, we find that all ratio estimators are unable to perfectly approximate $r(\boldsymbol{x} \mid \boldsymbol{\vartheta})$. This result is not unexpected, because the coverage diagnostic indicates that the confidence intervals are conservative, which implies that our estimates of the *true* likelihood-to-evidence ratio must be wrong. Incorrect, but conservative estimates of the posterior are not a significant issue because we mainly seek to constrain m_{WDM}.

We additionally find that the quality of the ratio estimates degrades for larger values of m_{WDM} across all architectures. Several strategies could be applied to address this. First, more expressive architectures could be explored which potentially make more efficient use of the available data. Second, by using additional observables could be simulated to aid the approximation of the underlying densities. In our specific case, a straightforward application of this strategy would be to simulate additional observables for $\boldsymbol{\vartheta} \gtrapprox 20$ keV. We would like to emphasize that increasing the size of the training dataset by simulating additional observables at specific target parameters $\boldsymbol{\vartheta}$ *should not be done*, because this implicitly changes the prior and therefore the underlying marginal model. Instead, additional observables should only be simulated by sampling from the joint $p(\boldsymbol{\vartheta}, \boldsymbol{x})$.

4.4.3 *Evaluation*

The performance of both methods is assessed on various randomly sampled GD-1 mock simulations with distinct nominal target parameters. A compact overview of the computed posteriors is shown in Figure 4.5. A full overview is shown in Figures 4.7 and 4.8. We find the proposed methodology to be preferable over the ABC analysis regarding the reconstruction of the nominal target parameters, and with respect to our stronger, statistically tested, confidence intervals.

In conjunction with the foregoing statistical validation of the ratio estimators, these results highlight the fact that ABC requires a carefully crafted summary statistic; a problem that is absent, or effectively automatised, in the proposed method. As mentioned earlier, an ABC posterior is only exact whenever the summary statistic is sufficient *and* the acceptance threshold tends to 0. If these conditions are not met, the posterior is possibly inaccurate or biased. The necessity of a sufficient summary statistic underlines an important issue with ABC in practice; the *assumed* sufficiency. Determining the statistical validity of an ABC analysis is computationally demanding and often not feasible. Our

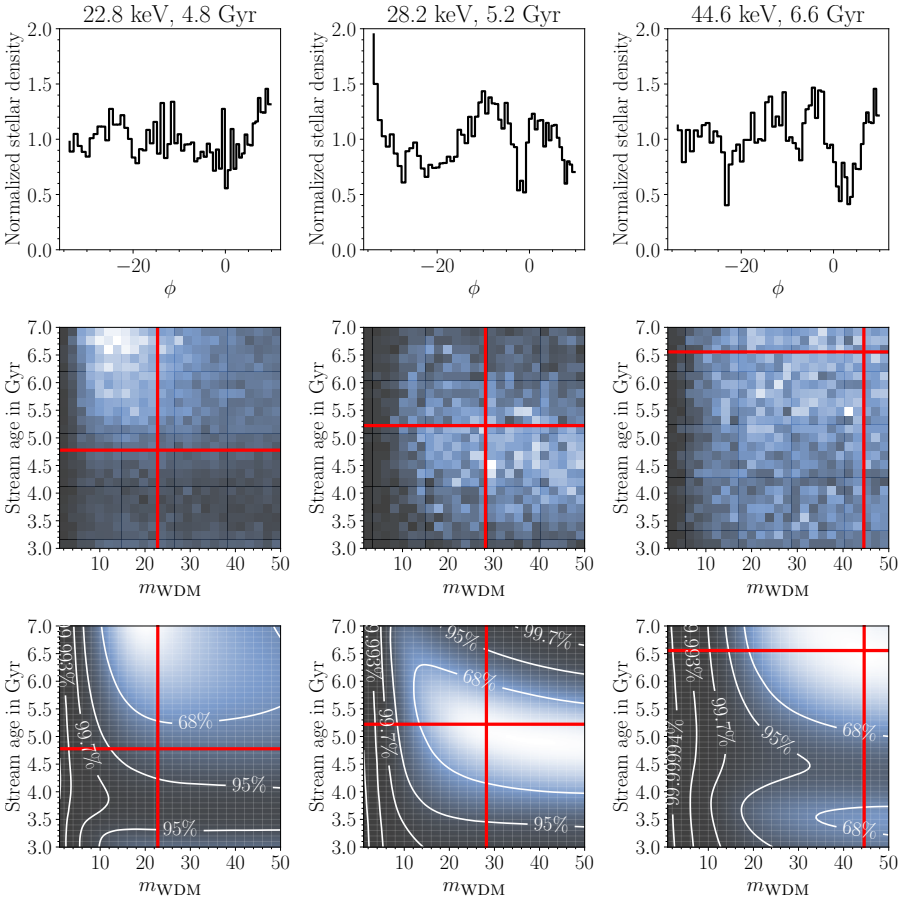

Figure 4.5: Compact summary of comparisons against ABC. Every column relates to a single mock simulation. The rows show, from top to bottom, the observable, the approximate posterior ABC, and our method respectively. The red cross indicates the groundtruth. ABC and our method are in agreement for most mock simulations. </>

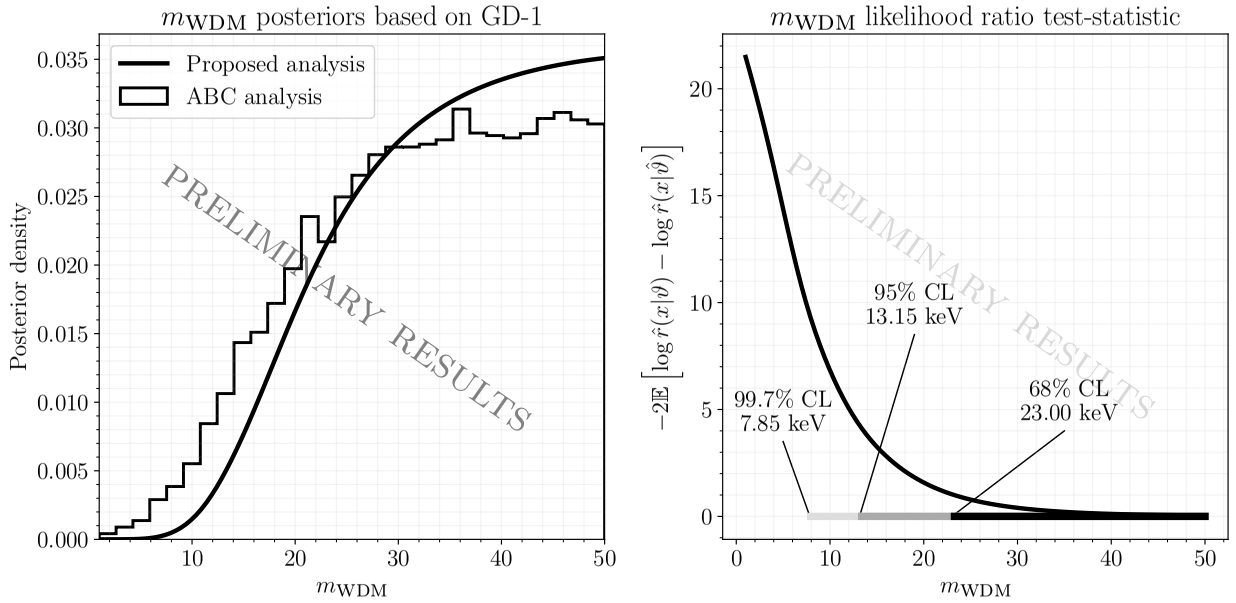

Figure 4.6: Age-marginalized results based on the observed stellar density variations of GD-1. *The results shown here illustrate the power of the proposed methodology, but should be considered as preliminary, since e.g. baryonic effects are not yet fully included in the simulation model. (Left)* Direct comparison of the reference ABC and the proposed analysis. Both posteriors indicate a preference for CDM over WDM within the assumed simulation model. We find that the proposed method is able to put stronger constraints on m_{WDM}. *(Right)* Likelihood ratio test-statistic used to derive the lower limit confidence intervals. </>

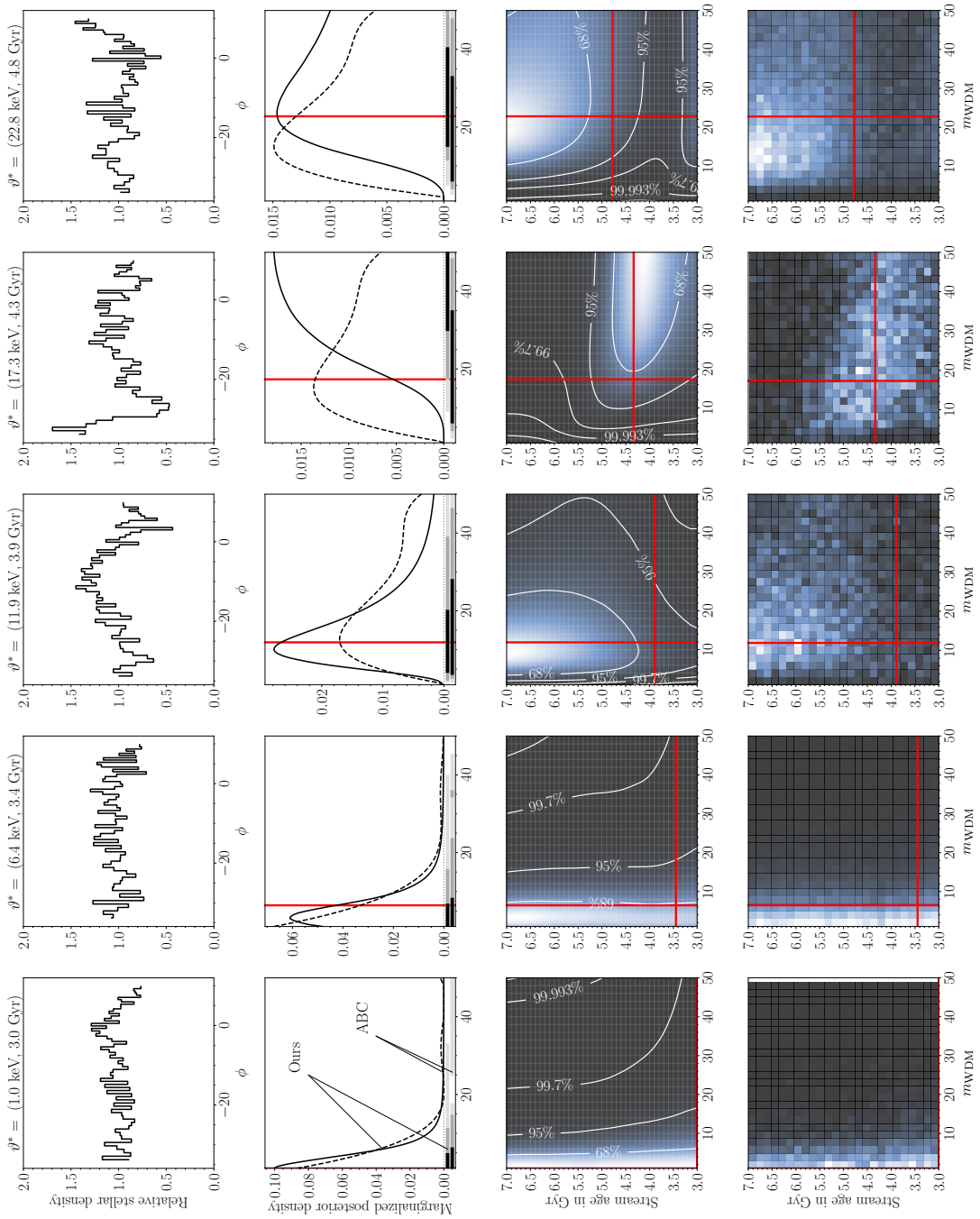

Figure 4.7: Direct comparison of ABC against the proposed method. The top row shows the observable. The second row the marginalized posteriors for both methods. Finally, row 3 and 4 show the joint posterior for our method and ABC respectively. The nominal target parameter is indicated by the red line. It is visually apparent that the proposed methodology produces stronger constraints of the groundtruth compared to ABC. ‹/›

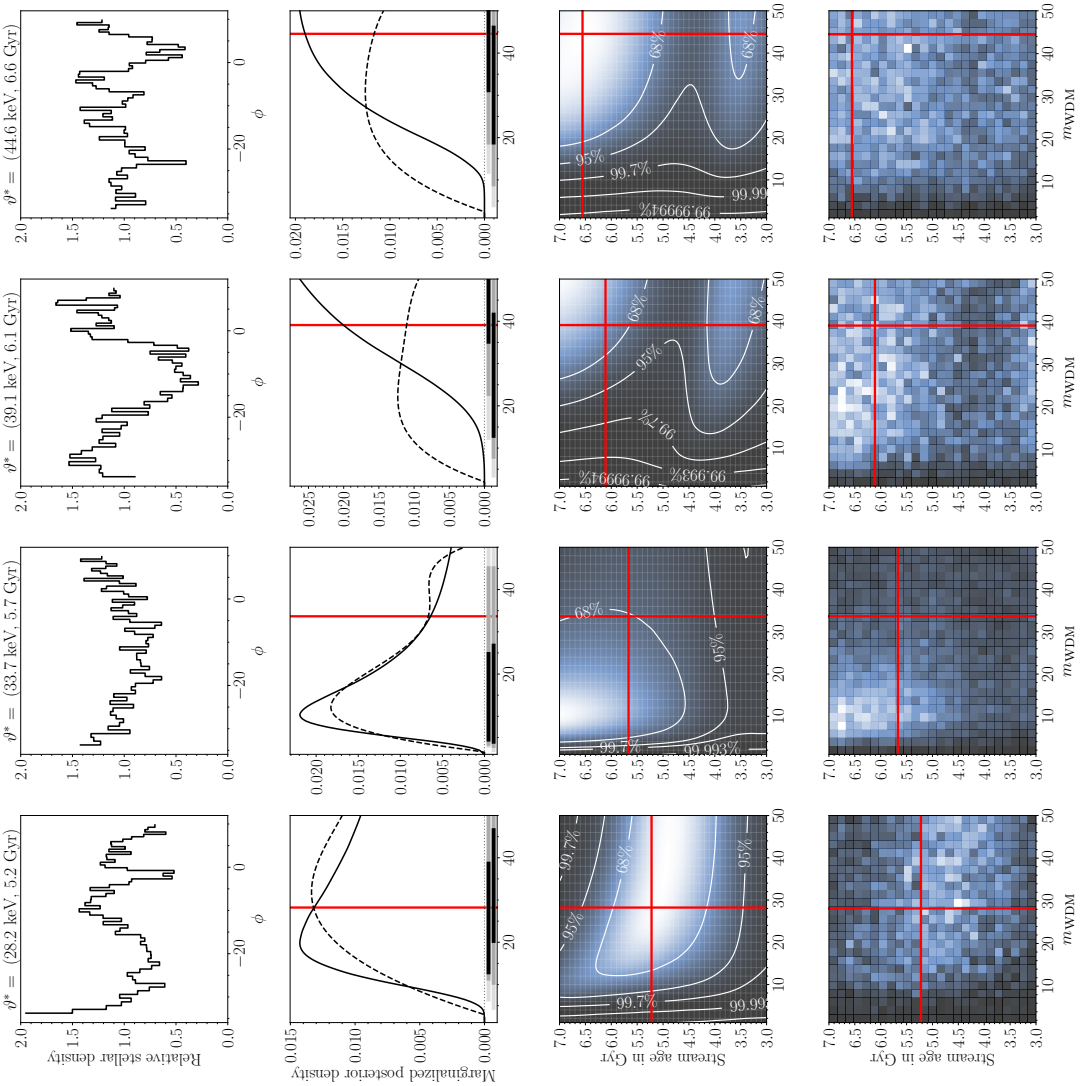

Figure 4.8: Direct comparison of ABC against the proposed method. Refer to Figure 4.7 for the initial results. </>

method does not suffer from this issue, because the estimation of the posterior density is amortized.

4.4.4 *Towards constraining m_{WDM} with GD-1*

We now apply our methodology to obtain a *preliminary* constraint on $m_{\rm WDM}$, based on the observed stellar density along the GD-1 stream. The posteriors in this section are computed using the previously trained and statistically validated RESNET-50 ratio estimator. We would like to remind the reader that the coverage diagnostic indicates that the derived confidence intervals are slightly conservative. Our results suggest a strong preference for CDM over WDM. The posteriors and credible intervals at various confidence levels are shown in Figure 4.6. We find the integrated area under the approximated posterior to be $(0.96 \pm 0.011$ </>). After marginalizing the stream age, the proposed methodology yields $m_{\rm WDM} \geq 17.5$ keV (95% CR) and $m_{\rm WDM} \geq 10.5$ keV (99.7% CR). No significant constraints can be put on the age of GD-1, although an older stream is preferred. A frequentist perspective based on likelihood ratio limits finds $m_{\rm WDM} \geq 13.15$ keV (95% CL) and $m_{\rm WDM} \geq 7.85$ keV (99.7% CL) after marginalizing the stream age. Assuming the posterior approximated by ABC is exact, we find $m_{\rm WDM} \geq 10.8$ keV (95% CL) and $m_{\rm WDM} \geq 3.5$ keV (99.7% CL). We emphasize that *our simulation model does not account for baryonic effects, disturbances caused by massive ($\gtrsim 10^9\ M_\odot$) subhaloes, and effects induced by variations in the Milky Way potential.*

However, our results are promising. We expect that the proposed method will enable an optimal discrimination between dark matter and baryonic effects (provided the latter can be convincingly modeled). It thus constitutes a powerful probe for constraining the mass of thermal or sterile neutrino dark matter [134, 189–192] (although a discrimination between such WDM models might be challenging).

4.5 SUMMARY AND DISCUSSION

This work proposes a general recipe for the usage of neural simulation-based inference in the natural sciences. Although the procedure generalizes to many domains, we apply our methodology in the stellar stream framework to determine the nature of the dark matter particle. We summarize our findings as follows:

- Bayesian inference based on Amortized Approximate Likelihood Ratios (AALR) is a powerful and convenient analysis framework to study the statistical properties of density variations of stellar streams. In Figure 4.5 we demonstrate that (at least in the absence

of the uncertainties from the baryonic effects), GD-1-like streams could be used to simultaneously constrain the mass of thermal relic dark matter and the age of the stream.

- AALR, in contrast to ABC, does not require handcrafted summary statistics and tuned acceptance thresholds. Our out-of-the-box AALR analysis are expected to be at least as good as any ABC implementation, and to often significantly outperform ABC, as evident in Figure 4.6.
- The amortized posterior estimation in AALR allows for a variety of diagnostics, including coverage and bias tests, which are computationally demanding and often infeasible for ABC. We explicitly demonstrate that posteriors approximated by AALR are unbiased and that the corresponding confidence intervals have coverage, as show in in see Figure 4.2 and Table 4.1 respectively.
- AALR provides a convenient and robust simulation-based inference without painfully hand-crafting summary statistics and tuning inference algorithms.

Finally, our preliminary results for GD-1 are promising as they indicate that AALR is an excellent and versatile method to probe the nature of dark matter with stellar streams. At face value, we can probe WDM masses up to 17.5 keV (95% credible lower limit for a GD-1-like stream). *We emphasize however that our simulation codes do not account for baryonic effects, which are expected to significantly impact the results.* In upcoming analyses we plan to use the improved statistical power achieved through AALR to obtain more statistically robust and tighter constraints on the particle mass of dark matter. However, we do expect some loss in sensitivity when including baryonic effects, because we expect the task of discriminating between CDM and WDM impacted streams to be harder for AALR.

Part II

RELIABLE SIMULATION-BASED INFERENCE

5

Averting A Potential Crisis in Simulation-Based Inference

The contents of this chapter are based on Hermans et al. [12].

We present extensive empirical evidence showing that current Bayesian simulation-based inference algorithms are inadequate for the falsificationist methodology of scientific inquiry. Our results collected through months of experimental computations show that all benchmarked algorithms – (s)NPE, (s)NRE, SNL and variants of ABC – may produce overconfident posterior approximations, which makes them demonstrably unreliable and dangerous if one's scientific goal is to constrain parameters of interest. We believe that failing to address this issue will lead to a well-founded trust crisis in simulation-based inference. For this reason, we argue that research efforts should now consider theoretical and methodological developments of conservative approximate inference algorithms and present promising research directions towards this objective. In this regard, we show empirical evidence that ensembles are consistently more reliable.

5.1 INTRODUCTION

While simulation-based inference targets domain sciences, advances in the field are mainly driven from a machine learning perspective. The field, therefore, inherits the quality assessments [193] customary to the machine learning literature, such as the minimization of classical divergence criteria. Despite recent developments of post hoc diagnostics to inspect the quality of likelihood-free approximations [55, 176, 193–197], assessing whether approximate inference results are sufficiently reliable for scientific inquiry remains largely unanswered whenever fitting criteria are not globally optimized or whenever the data is limited. In fact, domain sciences, and more specifically the physical sciences, are not necessarily interested in the *exactness* of an approximation. Instead, in the tradition of Popperian falsification, they

often seek to **constrain parameters** of interest as much as possible at a given confidence level. Scientific examples include frequentist confidence intervals on the mass of the Higgs boson [198], Bayesian credible regions on cosmological parameters [199, 200], or constraints on the intrinsic parameters of binary black hole coalescences [201]. From a Bayesian perspective, this implies that statistical approximations in simulation-based inference should ideally come with *conservative* guarantees to not produce credible regions smaller than they should be, even if it would incur a loss in statistical power. Wrongly constraining model parameters would otherwise impede scientific inquiry.

In this chapter, we measure the quality of the credible regions computed through various Bayesian techniques in simulation-based inference. We frame our main contribution as the collection of extensive empirical evidence that required months of computation. Our results demonstrate that all benchmarked techniques can produce non-conservative credible regions, highlighting the need for a new class of conservative approximate inference algorithms. The structure of the paper is outlined as follows. Section 5.2 describes the statistical formalism, necessary background and includes a thorough motivation for coverage. Section 5.3 highlights our main results. Section 5.4 presents several avenues of future research to **enable drawing reliable scientific conclusions** with simulation-based inference. All code related to this manuscript is available at

`github.com/montefiore-ai/averting-a-crisis-in-sbi`.

5.2 BACKGROUND

5.2.1 *Statistical formalism*

We evaluate posterior estimators that produce approximations $\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})$ with the following semantics.

Target parameters $\boldsymbol{\vartheta}$ denote the parameters of interest of a simulation model, and are sometimes referred to as *free* or *model* parameters. The precise definition of $\boldsymbol{\vartheta}$ depends on the problem setting. We make the reasonable assumption that the prior $p(\boldsymbol{\vartheta})$ is tractable.

An **observable** $\boldsymbol{x}$ denotes a synthetic realization of the simulator. Observed data $\boldsymbol{x}_o$ is the observable we would like to do inference on, under the assumption that the simulation model is correctly specified.

The **likelihood** model $p(\boldsymbol{x} \mid \boldsymbol{\vartheta})$ implicitly defined by the simulator's computer code. While we cannot evaluate the density $p(\boldsymbol{x} \mid \boldsymbol{\vartheta})$, we *can* simulate samples.

The **ground truth** $\boldsymbol{\vartheta}^*$ specified to the simulation model whose forward evaluation produced the observable $\boldsymbol{x}_o$, i.e., $\boldsymbol{x}_o \sim p(\boldsymbol{x} \mid \boldsymbol{\vartheta} = \boldsymbol{\vartheta}^*)$.

A **credible region** is a space Θ within the target parameter domain that satisfies

$$\int_\Theta p(\boldsymbol{\vartheta} \mid \boldsymbol{x} = \boldsymbol{x}_o) \, \mathrm{d}\, \boldsymbol{\vartheta} = 1 - \alpha \tag{5.1}$$

for some observable $\boldsymbol{x}_o$ and confidence level $1 - \alpha$. Because many such regions exist, we compute the credible region with the *smallest* volume. In the literature this credible region is known as the *highest posterior density region* [202, 203].

5.2.2 *Statistical quality assessment*

Common metrics for evaluating the quality of a posterior surrogate include the Classifier Two-sample Test [204, 205] and Maximum Mean Discrepancy [206–208]. The main problem with these metrics is that

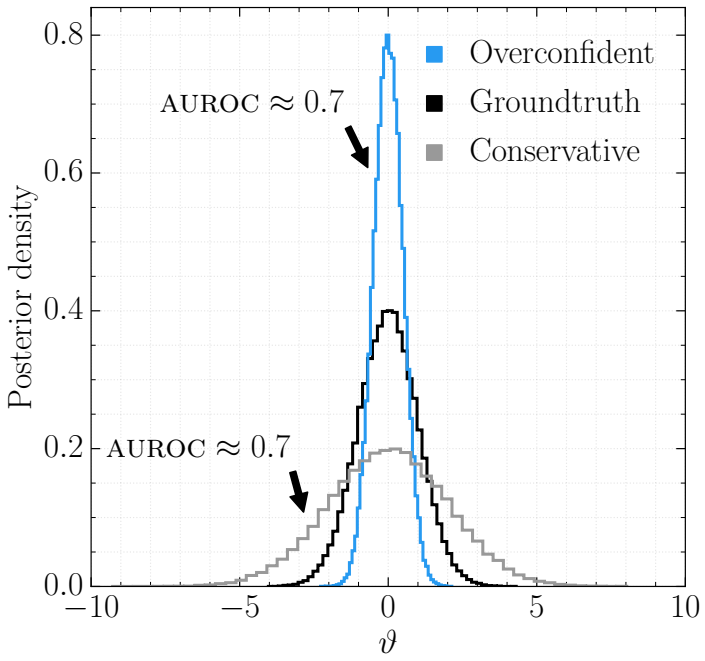

Figure 5.1: A classifier-based metric measures the divergence between posterior approximations and a ground truth by means of evaluating the classifier's discriminative performance through Area Under the Receiver Operating Characteristics curve (AUROC). In this case, the metric argues that both the conservative and overconfident approximations are equally accurate as it yields AUROC = 0.7 for both approximations. From an inference perspective however, the conservative approximation is more suitable because it produces credible regions larger than they should be.

they assess *exactness* of an approximation through a divergence with respect to a posterior that is intractable in practice. Even if such evaluations would hypothetically be possible, there are no criteria to what constitutes an *acceptable* estimator. Moreover, it is not possible to be certain whether the classifier or kernel used to measure the divergence are expressive enough to differentiate between the true posterior and its approximation.

To clarify these points, consider the demonstration in Figure 5.1. A binary classifier is trained to discriminate between samples from a posterior approximation and the true posterior. The discriminative performance of the classifier is expressed through Area Under the Receiver Operating Characteristics curve (AUROC) and serves as a measure for divergence between both densities. An AUROC = 0.5 suggests an approximation that is indistinguishable from the true posterior, while AUROC = 1.0 implies that both distributions do not overlap. Although both approximations in our demonstration are equally accurate according to the AUROC, the *overconfident* approximation illustrates the potential trust crisis in simulation-based inference: producing credible regions that are biased or smaller than they should be, potentially leading to erroneous scientific conclusions. For this reason, we take the position that posterior approximations should, irrespective of the available simulation budget, produce inflated credible regions and do not have to closely match the true posterior to draw meaningful inferences.

Instead of measuring exactness of approximations with respect to an intractable posterior, this work directly probes the quality of credible regions through the notion of *expected* coverage, which determines whether posterior approximations are well-calibrated with respect to the specified prior. It is a quantity that can be estimated in practice and has a threshold to determine whether a posterior estimator is acceptable.

Definition 1. *The **expected coverage probability** of the $1-\alpha$ highest posterior density regions derived from the posterior estimator $\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})$ is*

$$\mathbb{E}_{p(\boldsymbol{\vartheta},\boldsymbol{x})}\left[\mathbb{1}\left[\boldsymbol{\vartheta} \in \Theta_{\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})}(1-\alpha)\right]\right], \tag{5.2}$$

where the function $\Theta_{\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})}(1-\alpha)$ yields the $1-\alpha$ highest posterior density region of $\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})$.

Note that Equation 5.2 can be expressed as either

$$\mathbb{E}_{p(\boldsymbol{\vartheta})}\mathbb{E}_{p(\boldsymbol{x} \mid \boldsymbol{\vartheta})}\left[\mathbb{1}\left[\boldsymbol{\vartheta} \in \Theta_{\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})}(1-\alpha)\right]\right], \tag{5.3}$$

which is the *expected* frequentist coverage probability, or alternatively as the expected Bayesian credibility

$$\mathbb{E}_{p(\boldsymbol{x})}\mathbb{E}_{p(\boldsymbol{\vartheta} \mid \boldsymbol{x})}\left[\mathbb{1}\left[\boldsymbol{\vartheta} \in \Theta_{\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})}(1-\alpha)\right]\right], \tag{5.4}$$

whose inner expectation reduces to $1 - \alpha$ whenever the posterior estimator $\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})$ is well-calibrated.

Definition 2. *The **empirical expected coverage probability** of the $1 - \alpha$ highest posterior density regions derived from the posterior estimator $\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})$ given a set of n i.i.d. samples $(\boldsymbol{\vartheta}_i^*, \boldsymbol{x}_i) \sim p(\boldsymbol{\vartheta}, \boldsymbol{x})$ is*

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{1} \left[\boldsymbol{\vartheta}_i^* \in \Theta_{\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x}_i)}(1 - \alpha) \right]. \tag{5.5}$$

Definition 3. *The **nominal expected coverage probability** is the expected coverage probability of the true posterior and is equal to the confidence level.*

Definition 4. *A posterior estimator is deemed acceptable if it **has coverage** at the confidence level of interest, i.e., whenever the empirical expected coverage probability is **larger or equal** to the nominal expected coverage probability.*

Definition 5. *A **conservative posterior estimator** has coverage for **all** confidence levels.*

While coverage is a necessary metric to assess conservativeness, it is limited in its ability to determine the information gain a posterior (approximation) has over its prior. To clarify this point, consider an estimator whose posteriors are identical to the prior. In this case, there is no gain in information and the empirical expected coverage probability is equal to the nominal expected coverage probability. For this reason, a complete analysis should be complemented with measures such as the mutual information or expected information gain $\mathbb{E}_{p(\boldsymbol{\vartheta}, \boldsymbol{x})} [\log p(\boldsymbol{\vartheta} \mid \boldsymbol{x}) - \log p(\boldsymbol{\vartheta})]$. This work is however concerned with **conservative inference** and will therefore limit the analysis to the evaluation of expected coverage. Finally, it should be noted that expected coverage is a statement about the credible regions in expectation and therefore does not make any statement about the quality of an individual posterior.

5.3 EXPERIMENTAL OBSERVATIONS

This section covers our main contribution: the collection of empirical evidence to determine whether approaches in simulation-based inference are conservative by nature. We are particularly interested in determining *if* certain approaches should be favoured over others. We do so by measuring the coverage of estimators attained by these approaches across a broad range of hyperparameters and benchmarks of varied complexity, including two *real* problems from the field of astronomy. As in real use-cases, the true posteriors associated with these benchmarks are unknown.

We make the distinction between two paradigms. Non-amortized approaches are designed to approximate a *single* posterior, while amortized methods aim to learn a general purpose estimator that attempts to approximate *all* posteriors supported by the prior.

5.3.0.1 *Amortized*

NEURAL POSTERIOR ESTIMATION (NPE) is concerned with directly learning an amortized posterior estimator $\hat{p}_\psi(\boldsymbol{\vartheta} \mid \boldsymbol{x})$ with normalizing flows. Normalizing flows define a class of probability distributions $p_\psi(\cdot)$ built from neural network based bijective transformations [209, 210] parameterized by ψ and are usually optimized via

$$\arg\min_{\psi} \mathbb{E}_{p(\boldsymbol{x})} \left[\mathrm{KL}(p(\boldsymbol{\vartheta} \mid \boldsymbol{x}) \mid\mid \hat{p}_\psi(\boldsymbol{\vartheta} \mid \boldsymbol{x}) \right], \tag{5.6}$$

which is equivalent to $\arg\max_\psi \mathbb{E}_{p(\boldsymbol{\vartheta},\boldsymbol{x})} \left[\log \hat{p}_\psi(\boldsymbol{\vartheta} \mid \boldsymbol{x}) \right]$. Once trained, the density of the modeled distribution can be evaluated *and* sampled from.

NEURAL RATIO ESTIMATION (NRE) is an established approach in the simulation-based inference literature both from a frequentist [55] and Bayesian [147, 211] perspective. In a Bayesian analysis, an amortized estimator $\hat{r}(\boldsymbol{x} \mid \boldsymbol{\vartheta})$ of the *intractable* likelihood-to-evidence ratio $r(\boldsymbol{x} \mid \boldsymbol{\vartheta})$ can be learned by training a binary discriminator $\hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x})$ to distinguish between samples of the joint $p(\boldsymbol{\vartheta}, \boldsymbol{x})$ with class label 1 and samples of the product of marginals $p(\boldsymbol{\vartheta})p(\boldsymbol{x})$ with class label 0 using a criterion such as the binary cross entropy. Similar to the density-ratio trick [55, 81, 147, 212], the Bayes optimal discriminator $d(\boldsymbol{\vartheta}, \boldsymbol{x})$ models

$$\frac{p(\boldsymbol{\vartheta}, \boldsymbol{x})}{p(\boldsymbol{\vartheta}, \boldsymbol{x}) + p(\boldsymbol{\vartheta})p(\boldsymbol{x})} = \sigma\left(\log \frac{p(\boldsymbol{\vartheta}, \boldsymbol{x})}{p(\boldsymbol{\vartheta})p(\boldsymbol{x})} \right), \tag{5.7}$$

where $\sigma(\cdot)$ is the sigmoid function. Given a target parameter $\boldsymbol{\vartheta}$ and observable $\boldsymbol{x}$ supported by $p(\boldsymbol{\vartheta})$ and $p(\boldsymbol{x})$ respectively, the learned discriminator $\hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x})$ approximates the log likelihood-to-evidence ratio $\log r(\boldsymbol{x} \mid \boldsymbol{\vartheta})$ through the logit function because

$$\log \hat{r}(\boldsymbol{x} \mid \boldsymbol{\vartheta}) = \texttt{logit}\left(\hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x}) \right) \approx \log \frac{p(\boldsymbol{\vartheta}, \boldsymbol{x})}{p(\boldsymbol{\vartheta})p(\boldsymbol{x})}. \tag{5.8}$$

The log posterior density function is approximated as $\log \hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x}) = \log p(\boldsymbol{\vartheta}) + \log \hat{r}(\boldsymbol{x} \mid \boldsymbol{\vartheta})$.

ENSEMBLES of models constitute a standard method to improve predictive performance. In this work, we consider an ensemble model that *averages* the approximated posteriors of n independently trained posterior estimators. While this formulation is natural for NPE, averaging likelihood-to-evidence ratios is equivalent, as $p(\boldsymbol{\vartheta}) \frac{1}{n} \sum_{i=1}^{n} \hat{r}_i(\boldsymbol{x} \mid \boldsymbol{\vartheta}) = \frac{1}{n} \sum_{i=1}^{n} \hat{p}_i(\boldsymbol{\vartheta} \mid \boldsymbol{x})$.

5.3.0.2 *Non-amortized*

APPROXIMATE BAYESIAN COMPUTATION (ABC) [213, 214] numerically estimates a *single* posterior by collecting samples $\boldsymbol{\vartheta} \sim p(\boldsymbol{\vartheta})$ whenever $\boldsymbol{x} \sim p(\boldsymbol{x} \mid \boldsymbol{\vartheta})$ is *similar* to $\boldsymbol{x}_o$. Similarity is expressed by means of a *distance function* ρ. For high-dimensional observables, the probability density of simulating an observable $\boldsymbol{x}$ such that $\boldsymbol{x} = \boldsymbol{x}_o$ is extremely small. For this reason, ABC uses a *summary statistic* s and an *acceptance threshold* ϵ. Using these components, ABC accepts samples into the approximate posterior whenever $\rho(s(\boldsymbol{x}), s(\boldsymbol{x}_o)) \leq \epsilon$. In our experiments we use the identity function as a sufficient summary statistic. Finally, we emphasize that ABC approximations are *only exact* whenever the summary statistic is sufficient *and* the acceptance threshold ϵ tends to 0 [215].

Sequential methods aim to approximate a single posterior by *iteratively* improving a posterior approximation. These methods alternate between a simulation and exploitation phase. The latter being designed to take *current* knowledge into account such that subsequent simulations can be focused on parameters that are more likely to produce observables x similar to x_o.

SEQUENTIAL MONTE-CARLO ABC (SMC-ABC) [216–218] iteratively updates a set of proposal states to match the posterior distribution. At each iteration, accepted proposals are ranked by distance. The rankings determine whether a proposal is propagated to the next iteration. New candidate proposals are generated by perturbing the selected ranked proposals.

SEQUENTIAL NEURAL POSTERIOR ESTIMATION (SNPE) [219–221] directly models the posterior. Our evaluations will specifically use the SNPE-C [221] variant.

SEQUENTIAL NEURAL LIKELIHOOD (SNL) [222] models the likelihood $p(\boldsymbol{x} \mid \boldsymbol{\vartheta})$. A numerical approximation of the posterior is obtained by plugging the learned likelihood estimator into a Markov Chain Monte Carlo (MCMC) sampler as a surrogate likelihood.

SEQUENTIAL NEURAL RATIO ESTIMATION (SNRE) [147, 223] iteratively improves the modelled likelihood-to-evidence ratio.

5.3.1 *Benchmarks*

Our evaluations consider 7 benchmarks, ranging from toy problems to real applications in astrophysics.

The *SLCP* simulator models a fictive problem with 5 parameters. The observable $\boldsymbol{x}$ is composed of 8 scalars which represent the 2D-coordinates of 4 points. The coordinate of each point is sampled from the same multivariate Gaussian whose mean and covariance matrix are parametrized by $\boldsymbol{\vartheta}$. We consider an alternative version of the original task [222] by inferring the marginal posterior density of 2 of those parameters. In contrast to its original formulation, the likelihood is not tractable due to the marginalization.

The *Weinberg* problem [224] concerns a simulation of high energy particle collisions $e^+e^- \to \mu^+\mu^-$. The angular distributions of the particles can be used to measure the Weinberg angle $\boldsymbol{x}$ in the standard model of particle physics. From the scattering angle, we are interested in inferring Fermi's constant $\boldsymbol{\vartheta}$.

The *Spatial SIR* model generates a grid-world $\boldsymbol{x}$ of susceptible, infected, and recovered individuals. This information is encoded in 3 channels. Based on the initial state of $\boldsymbol{x}$ and the infection and recovery rate $\boldsymbol{\vartheta}$, the model describes the evolution of an infection through this grid-like world. The disease spreads spatially.

M/G/1, originally introduced by Papamakarios, Sterratt, and Murray [222], models a processing and arrival queue. The problem is described by 3 parameters $\boldsymbol{\vartheta}$ that influence the time it takes to serve a customer, and the time between their arrivals. The observable $\boldsymbol{x}$ is composed of 5 equally spaced quantiles of inter-departure times.

The *Lotka-Volterra* population model [119, 225] describes a process of interactions between a predator and a prey species. The model is conditioned by 4 parameters $\boldsymbol{\vartheta}$ which influence the reproduction and mortality rate of the predator and prey species. We infer the marginal posterior of the 2 predator's parameters. Our implementation, described by a Markov process, generates 2 time-series of 1001 samples representing the evolution of the prey and predator populations over time.

Stellar *Streams* form due to the disruption of spherically packed clusters of stars by the Milky Way. Because of their distance from the galactic center and other visible matter, *distant* stellar streams are considered to be ideal probes to detect gravitational interactions with

dark matter. The simulation model [145, 148, 196] evolves the stellar density $\boldsymbol{x}$ of a stream over several billion years, while perturbing the stream over its evolution through gravitational interactions with dark matter subhaloes parameterized through the dark matter mass $\boldsymbol{\vartheta}$.

Gravitational Waves (GW) are ripples in space-time emitted during events such as the collision of two black-holes. They can be detected through interferometry measurements $\boldsymbol{x}$ and convey information about celestial bodies, unlocking new ways to study the universe. We consider inferring the masses $\boldsymbol{\vartheta}$ of two black-holes colliding through the observation of the gravitational wave as measured by LIGO's dual detectors [226, 227].

5.3.2 *Setup*

Our evaluations consider simulation budgets ranging from 2^{10} up to 2^{17} samples and confidence levels from 0.05 up to 0.95. Within the *amortized* setting we train, for every simulation budget, 5 posterior estimators for 100 epochs. The empirical coverage probability is computed on at least 5,000 unseen samples from the joint $p(\boldsymbol{\vartheta}, \boldsymbol{x})$ and for all confidence levels under consideration. In addition, we repeat the coverage evaluation for ensembles of 5 estimators as well.

Special care for *non-amortized* approaches is necessary because they only approximate a *single* posterior and can therefore not evaluate coverage. Our experiments estimate coverage of these methods *by proxy* by repeating the inference procedure on 300 distinct observables for a given simulation budget (2100 times per method per benchmark). The empirical coverage probabilities are estimated using the 300 approximated posteriors. Our experiments with NPE, SNPE, SNL, SNRE, REJ-ABC and SMC-ABC rely on the implementation in the `sbi` package [228], while we use a custom implementation for NRE.

COMPUTATIONAL COST We would like to emphasize the computational requirements necessary to generate our main contribution: the experimental observations, whose generation took months. The bulk of the cost was associated with the repeated optimization procedure of non-amortized methods and the constant resampling of the simulator. Totalling in the order of 3000 CPU days to compute the results, simulations included (200 days amortized vs. 2800 days non-amortized).

5.3.3 *Results*

Observation 1 All benchmarked algorithms produce non-conservative posterior approximations. This pathology tends to be accentuated when using a small simulation-budget in both amortized and non-amortized approaches.

Observation 2 For a given simulation budget, amortized approaches have the tendency to be more conservative than non-amortized approaches.

Observation 3 The empirical coverage probability of an ensemble model is larger than the expected individual model. The ensemble size positively affects the empirical coverage probability as well.

Observation 4 Amortized methods are not subordinate with respect to the simulation-efficiency of non-amortized sequential methods, especially when taking hyper-parameter tuning and the evaluation of the coverage diagnostic into account.

Figures 5.2 and 5.3 highlight our main results. Through these plots we can directly assess the *conservativeness* at a given confidence level and simulation budget. The figures should be interpreted as follows: a *perfectly calibrated* posterior has an empirical coverage probability equal to the nominal coverage probability. Plotting this relation produces a diagonal line. Conservative estimators on the other hand produce curves *above* the diagonal and overconfident models underneath. The plots highlight an unsettling observation: **all** benchmarked approaches produce non-conservative posterior approximations. In general, this pathology is especially prominent in non-amortized approaches with a small simulation budget; a regime they have been specifically designed for. A large simulation budget does not guarantee conservativeness either.

In sequential approaches, this behaviour could be explained by the alternating exploitation and simulation phase. One potential failure mode is that a non-conservative posterior approximation at a previous iteration forces the next simulation phase to not produce observables that should be associated with a higher posterior density, causing the estimator to increase its non-conservativeness at each iteration.

Despite all ABC approaches use a sufficient summary statistic (the identity function), our results demonstrate that this alone is no guarantee for conservative posterior approximations. In fact, using a sufficient summary statistic with $\epsilon > 0$ does not always correspond to conservative approximations. In such cases, ABC accepts samples with larger distances, permitting the procedure to shift the mass of the approxi-

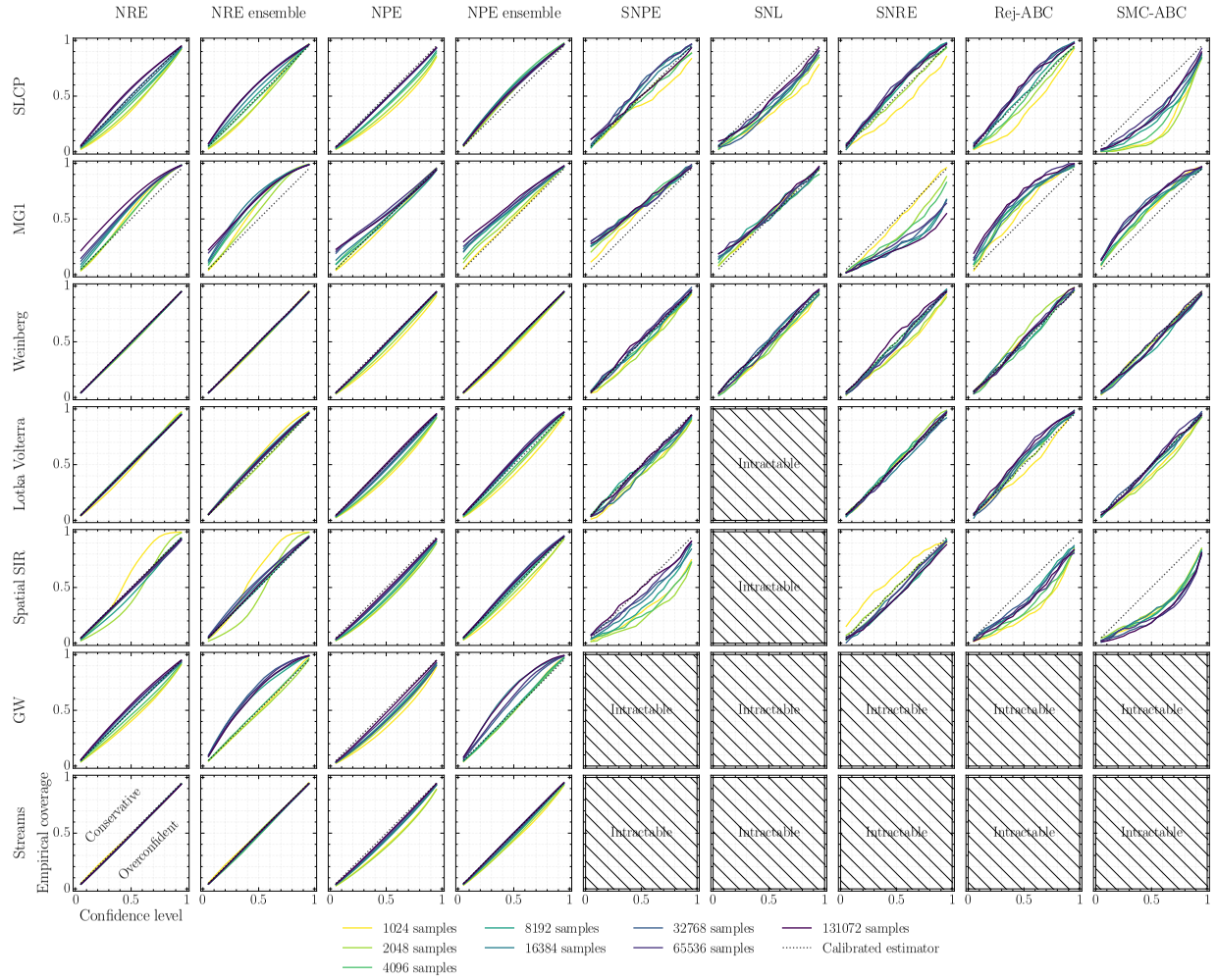

Figure 5.2: Evolution of the coverage w.r.t simulation budget. A perfectly calibrated posterior has an empirical coverage probability equal to the nominal coverage probability and produces a diagonal line. Conservative estimators on the other hand produce curves *above* the diagonal and overconfident models underneath. All algorithms can lead to non-conservative estimators, this pathology tends to be accentuated for small simulation budgets and non-amortized methods. Finally, the intractable results indicate that the computational requirements did not allow for a coverage analysis. In the case of SNL, this was mostly due to the high dimensional observables. We did not train an embedding network as that is outside of the scope of this work. For the astronomy applications, the simulation model was simply too expensive to reasonable evaluate coverage for non-amortized methods.

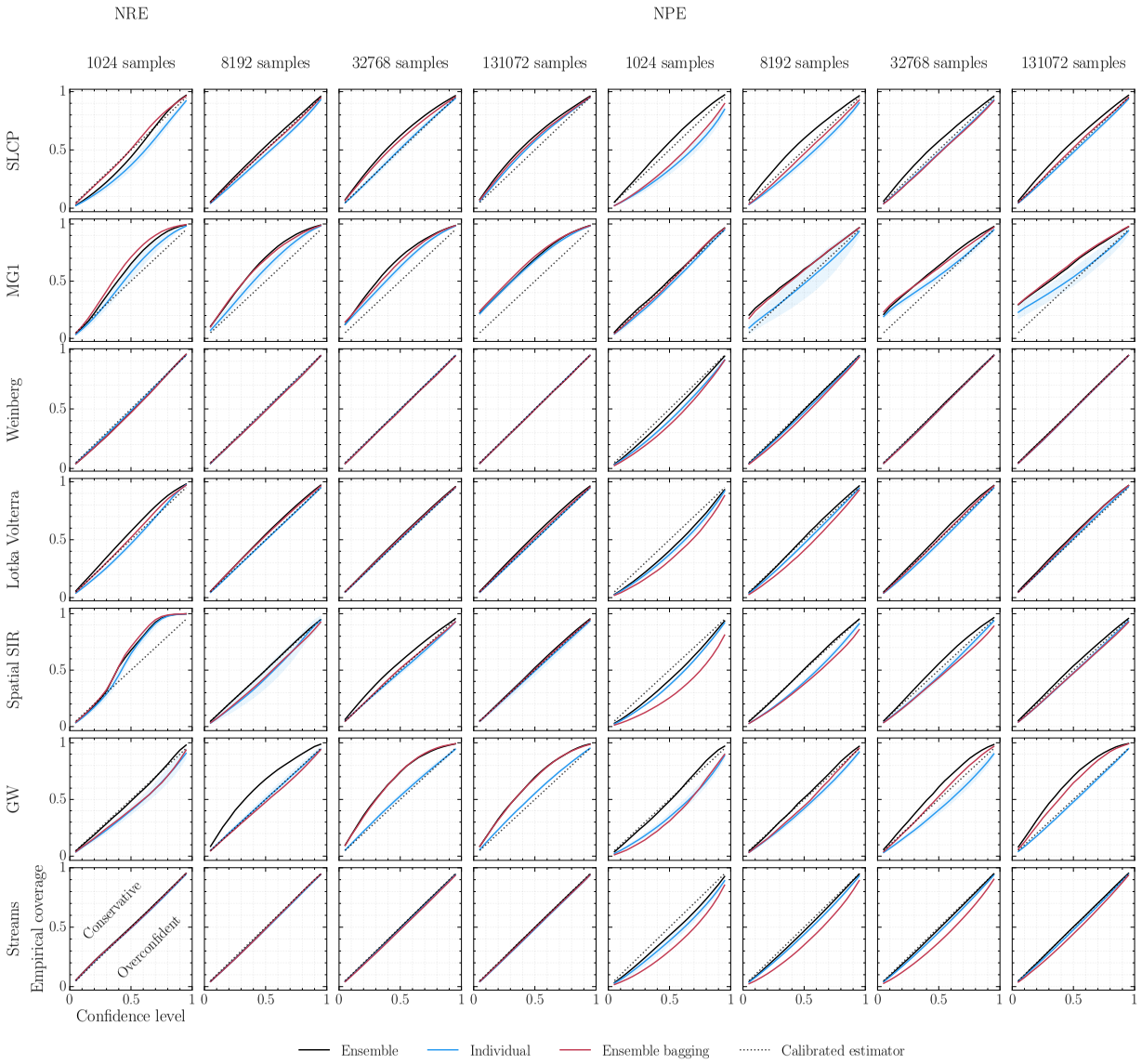

Figure 5.3: Analysis of coverage between ensemble and individual models w.r.t various simulation budgets. The blue line represents the mean coverage of individual models over 5 runs, the shaded area represents its standard deviation. The black line represents the coverage of a single ensemble composed of 5 models. We observe that ensembles consistently have a higher empirical coverage probability compared to the average individual model. A similar effect is not always observed with bagging, indicated by the red line. Ensembles are only considered on amortized approached due to computational requirements inherent to non-amortized approaches.

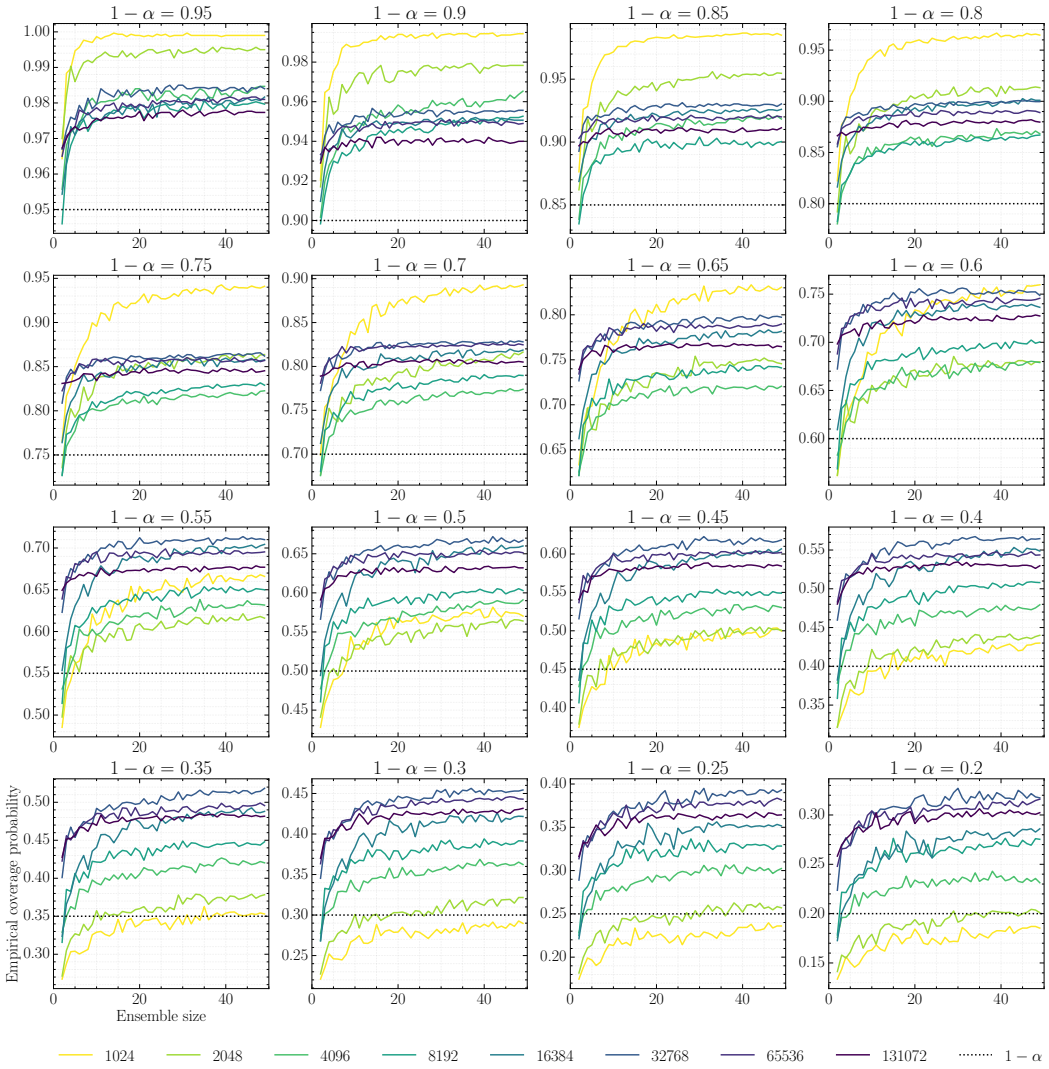

Figure 5.4: Evolution of the empirical coverage probability with respect to ensemble size for various confidence levels. The results are obtained by training 100 ratio estimators (NRE) on the SLCP benchmark. A positive effect is observed in terms of empirical coverage probability and ensemble size, i.e., a larger ensemble size correlates with a larger empirical coverage probability. This is not unsurprising, because a larger ensemble is expected to capture more of the uncertainty that stems from the training procedure.

mated posterior elsewhere. In addition, a limited number of posterior samples can negatively affect the quality of the credible regions, e.g., when approximating the posterior density function with kernel density estimation. Both cases could cause the observed behaviour. Scientific applications should therefore be cautious. Even though a handcrafted, albeit sufficient, summary statistic provides some insight into the approximated posterior, it does not imply that ABC approximations are conservative whenever $\epsilon > 0$.

In Figure 5.3 we observe that the empirical coverage probability of ensemble models is consistently larger than the empirical coverage probability of the expected individual posterior estimator. Current applications of simulation-based inference can therefore rely on ensembling to build more conservative posterior estimators. However, the ensemble model can still be non-conservative. We hypothesize that the increase in coverage is linked to the added uncertainty captured by the ensemble model, leading to inflated credible regions. In fact, individual estimators only captures aleatoric uncertainty, i.e., uncertainty that is linked to stochasticity of the data generation procedure, while an ensemble is expected to capture *part* of the epistemic uncertainty as well, i.e., uncertainty issued by the lack of knowledge. Surprisingly, we find that ensembles built using bagging do not always produce higher coverage than individual models while they should also capture part of the epistemic uncertainty. Although a deeper understanding of this effect is required, this behaviour could be explained by the fact that bagging reduces the effective dataset size used to train each member of the ensemble. Additionally, Figure 5.4 illustrates a positive effect with respect to ensemble size.

Not evident from Figures 5.2 and 5.3 are the computational consequences of a coverage analysis on non-amortized methods. Although the figures mention a certain simulation budget, the *total* number of simulations for non-amortized methods should be multiplied by the number of approximated posteriors (300) to estimate coverage, highlighting the simulation cost associated with diagnosing non-amortized approaches. This issue is not limited to coverage. Simulation-Based Calibration (SBC) [176] relies on samples of arbitrary posterior approximations. Diagnosing non-amortized estimators with SBC therefore requires a similar approach as we have taken in our coverage analyses. In fact, Lueckmann et al. [193] also mention that SBC is computationally prohibitive for non-amortized approaches and therefore restrain from evaluating it.

Our results illustrate a clear distinction between the amortized and non-amortized paradigms. Amortized methods do not require retraining or new simulations to determine the empirical coverage

probability of a posterior estimator, while non-amortized methods do. Moreover, a coverage analysis of non-amortized approaches only measures the quality of the training procedure. In contrast to amortized approaches, where the posterior estimator is diagnosed. This has severe implications on the applicability of non-amortized methods, because their reliability cannot be practically determined. In addition, non-amortized approaches have to repeat the approximation procedure whenever architectural or hyperparameter changes are made, while amortized methods reuse previously simulated datasets. In particular, sequential methods cannot do this as new simulations depend on the posterior approximation at a previous state. This is often overlooked in studies on simulation efficiency and raises questions about whether sequential approaches should still be considered simulation efficient over their amortized counterparts. Especially because our results indicate that for a given simulation budget, amortized approaches produce trustworthier posterior approximations in expectation. Moreover, estimators obtained through amortized methods can be calibrated [55, 194, 196] after training, while such procedures are impossible with non-amortized approaches.

All of the above leads us to conclude that *currently*, amortization should be favoured over non-amortized approaches because their reliability cannot practically be determined. Moreover, our results suggest that even for small simulation budgets amortized methods, on average, produce more conservative estimators compared to non-amortized methods; a striking result, given that non-amortized, and sequential methods in particular, dedicate the available simulation budget to accurately approximate a single posterior.

5.4 DISCUSSION

As demonstrated empirically, simulation-based inference can be unreliable, especially when its approximations cannot be diagnosed. The problem of determining whether a posterior approximation is correct is in fact not restricted to simulation-based inference specifically, the concern occurs in all of approximate Bayesian inference. The MCMC literature deals with this exact same problem in the form of determining whether a set of Markov chain samples have converged to the target distribution [229, 230]. In this regard, empirical diagnostic tools have been proposed over the years [176, 231–234] and have helped practitioners using MCMC properly. Nonetheless, there is currently no clear solution to determine convergence with absolute certainty [235, 236], even if the likelihood function is here tractable.

We are of the opinion that theoretical and methodological advances within the field of simulation-based inference will strengthen its re-

liability and promote its applicability in sciences. First, although all benchmarked algorithms recover the true posterior under specific optimal conditions, it is generally not possible to know whether those conditions are satisfied in practice. Therefore, the study of new objective functions that would force posterior estimators to always be conservative, regardless of their optimal conditions, is worth investigating. From a Bayesian perspective, Rozet and Louppe [237] propose using the focal and the peripheral losses to weigh down easily classified samples as a means to tune the conservativeness of a posterior estimator. Dalmasso et al. [238] consider the frequentist setting and introduce a theoretically-grounded algorithm for the construction of confidence intervals that are guaranteed to have perfect coverage, regardless of the quality of the used statistic. Second, in light of our results that ensembles produce more conservative posteriors, model averaging constitutes another promising direction of study, as a simple and efficient method to produce reliable posterior estimators. However, a deeper understanding of the behaviour we observe is certainly first required to further develop these methods. Third, post-training calibration can be used to improve the reliability of posterior estimators and should certainly be considered as a way towards more conservative inference. To some extent, this has already been considered for amortized methods [55, 194, 196] and would be worth exploring further, especially for non-amortized approaches.

In summary, we show that current algorithms for simulation-based inference may all produce overconfident posterior approximations, making them demonstrably unreliable if one's scientific goal is to constrain parameters of interest or reject theoretical models. Nevertheless, we remain confident and optimistic and advocate that this result is only a stepping stone towards more reliable simulation-based inference, its wider adoption, and eventually better science.

6

Towards Reliable Simulation-Based Inference with Binary Classification

The contents of this chapter are based on unpublished and incomplete work.

Our study focuses on a specific family of likelihood-free techniques whose surrogate can be parameterized through a binary classifier or discriminator. Although our technique generalizes to any binary classification problem, we mainly consider the Bayesian likelihood-free paradigm to contain the scope of the discussion. In particular, we focus on NRE, introduced in Chapter 3.

The results of Chapter 5 demonstrate that likelihood-free approximations *can* produce non-conservative or overconfident posteriors, regardless of the problem setting or available simulation budget. Determining the reliability of the learned posterior estimators through the evaluation of diagnostics such as expected coverage or Simulation-Based Calibration (SBC) [176] is therefore critical for scientific applications. Unfortunately, computing these diagnostics is infeasible for *non-amortized* inference protocols due to their intrinsic computational requirements. Although amortized protocols can in fact be diagnosed, they are still susceptible to the same issue. More importantly, as the general reliability of likelihood-free approximations cannot be determined and theoretical guarantees are lacking, it leaves practicioners uncertain about the applicability of the learned estimators.

If we are to gather scientific knowledge with simulation-based inference, reliability of its approximations are key. Contrary to the current school of thought which is mainly interested in *exact* or *accurate* approximations, domain sciences such as physics rather seek to *constrain* parameters $\boldsymbol{\vartheta}$ through frequentist confidence intervals or Bayesian credible regions (constraints) at a given confidence or credibility level respectively. Exactness of an approximation is not a primary objective. In these fields it is of the upmost importance that approximated constraints pass the (expected) coverage diagnostic, while at the same producing constraints for the target parameters $\boldsymbol{\vartheta}$ that are as strong

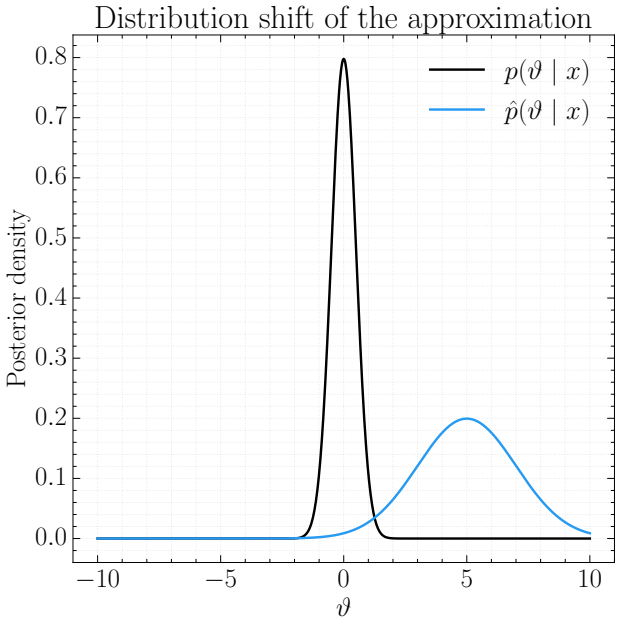

Figure 6.1: In this scenario, the posterior approximation $\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})$ is said to be *shifted* with respect to the true posterior $p(\boldsymbol{\vartheta} \mid \boldsymbol{x})$. A posterior approximation that assigns a larger uncertainty to a parameter with respect to the true posterior is not necessarily conservative!

as possible. An overconfident estimator could potentially impede scientific inquiry by steering research to abandon plausible models. It is for this reason alarming that various simulation-based inference protocols *can* produce constraints that are smaller than they should be [12], despite exactness guarantees these algorithms carry whenever their fitting criteria are globally optimized.

Given that the expected coverage diagnostic measures the reliability of an estimator, a conservative approximation should have two properties to pass this test. First, the estimated constraints should not be smaller than they should be. Second, the distribution shift, as illustrated in Figure 6.1, of the approximated posterior with respect to the true posterior and by transitivity their constraints, should not be significant. Meaning, it is be possible to have constraints that are larger than they should be, which is good, but whose approximated posteriors are in fact *shifted* with respect to the true posterior. This particular effect translates to an estimator that could fail the expected coverage diagnostic, even though its constraints are larger than they should be. Jointly, these two conditions contribute to a conservative approximation.

Motivated by these issues, this chapter seeks to increase the reliability of simulation-based inference *in practice* by proposing a technique that *(i)* establishes a premise for conservative approximations *(ii)*

comes with theoretical guarantees and *(iii)* is directly applicable, independent of the problem setting or hyperparameters such as simulation budget. These properties provide an answer to the question: **how can we learn reliable approximations whenever fitting criteria are *not* globally optimized?**

> REMARK
>
> Due to the incomplete state of this work, we will not consider a study on the potential distribution shift of the posterior.

6.1 AN INITIAL ATTEMPT

Because the size of estimated constraints are mainly driven by the uncertainty related to the target parameter $\boldsymbol{\vartheta}$, it is natural to assume that the differential entropy of the (approximated) posterior contributes to their size. Assuming there is no distribution shift with respect to the true posterior, this would imply that for an approximate posterior $\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x}) \triangleq p(\boldsymbol{\vartheta})\hat{r}(\boldsymbol{x} \mid \boldsymbol{\vartheta})$ to be conservative for a specific observable $\boldsymbol{x}$, the differential posterior entropy would need to be smaller or equal to the posterior entropy *of the approximation*, or

$$-\mathbb{E}_{p(\boldsymbol{\vartheta} \mid \boldsymbol{x})}\left[\log p(\boldsymbol{\vartheta} \mid \boldsymbol{x})\right] \leq -\mathbb{E}_{\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})}\left[\log \hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})\right]. \tag{6.1}$$

Ideally, this should be true for every observable $\boldsymbol{x}$. However, we can relax this constraint slightly and target this requirement in expectation.

$$\mathbb{E}_{p(\boldsymbol{x})}\left[-\mathbb{E}_{p(\boldsymbol{\vartheta} \mid \boldsymbol{x})}\left[\log p(\boldsymbol{\vartheta} \mid \boldsymbol{x})\right]\right] \leq \mathbb{E}_{p(\boldsymbol{x})}\left[-\mathbb{E}_{\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})}\left[\log \hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})\right]\right]. \tag{6.2}$$

Note that these quantities are actually the conditional entropy $H(\boldsymbol{\vartheta} \mid \boldsymbol{x})$ of the true posterior $p(\boldsymbol{\vartheta} \mid \boldsymbol{x})$ and the conditional entropy $\hat{H}(\boldsymbol{\vartheta} \mid \boldsymbol{x})$ of the posterior approximation $\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})$ respectively;

$$-\mathbb{E}_{p(\boldsymbol{\vartheta}, \boldsymbol{x})}\left[\log p(\boldsymbol{\vartheta} \mid \boldsymbol{x})\right] \leq -\mathbb{E}_{\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})p(\boldsymbol{x})}\left[\log \hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})\right], \tag{6.3}$$

$$H(\boldsymbol{\vartheta} \mid \boldsymbol{x}) \leq \hat{H}(\boldsymbol{\vartheta} \mid \boldsymbol{x}). \tag{6.4}$$

Where the conditional entropy can be expressed as

$$H(\boldsymbol{\vartheta} \mid \boldsymbol{x}) = H(\boldsymbol{\vartheta}) - I(\boldsymbol{\vartheta}, \boldsymbol{x}), \tag{6.5}$$

where $H(\boldsymbol{\vartheta})$ is the differential entropy of the prior $p(\boldsymbol{\vartheta})$ and $I(\boldsymbol{\vartheta}, \boldsymbol{x})$ the mutual information. Because $H(\boldsymbol{\vartheta})$ appears in both $H(\boldsymbol{\vartheta} \mid \boldsymbol{x})$ and $\hat{H}(\boldsymbol{\vartheta} \mid \boldsymbol{x})$, we have

$$H(\boldsymbol{\vartheta}) - I(\boldsymbol{\vartheta}, \boldsymbol{x}) \leq H(\boldsymbol{\vartheta}) - \hat{I}(\boldsymbol{\vartheta}, \boldsymbol{x}), \tag{6.6}$$

$$-I(\boldsymbol{\vartheta}, \boldsymbol{x}) \leq -\hat{I}(\boldsymbol{\vartheta}, \boldsymbol{x}), \tag{6.7}$$

$$I(\boldsymbol{\vartheta}, \boldsymbol{x}) \geq \hat{I}(\boldsymbol{\vartheta}, \boldsymbol{x}), \tag{6.8}$$

where the mutual information as computed through $\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})p(x)$ is

$$\hat{I}(\boldsymbol{\vartheta}, \boldsymbol{x}) = \int\int \mathrm{d}\,\boldsymbol{\vartheta}\,\mathrm{d}\,\boldsymbol{x}\;\; \hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})p(x) \log \hat{r}(\boldsymbol{x} \mid \boldsymbol{\vartheta}), \tag{6.9}$$

$$= \mathbb{E}_{p(\boldsymbol{\vartheta})p(\boldsymbol{x})} \left[\hat{r}(\boldsymbol{x} \mid \boldsymbol{\vartheta}) \log \hat{r}(\boldsymbol{x} \mid \boldsymbol{\vartheta})\right]. \tag{6.10}$$

The above directly implies that for posterior approximations to be conservative, we have to enforce $\hat{I}(\boldsymbol{\vartheta}, \boldsymbol{x})$ to be upper-bounded by $I(\boldsymbol{\vartheta}, \boldsymbol{x})$. While this seems we could rely on techniques such as MINE [239], there are however a few issues. First, there are currently no guarantees that $\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})$ for any observable $\boldsymbol{x}$ supported by $p(\boldsymbol{x})$ is a proper probability density, i.e.,

$$\int \mathrm{d}\,\boldsymbol{\vartheta}\;\; \hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x}) = 1. \tag{6.11}$$

Second, the theoretical arguments presented for the mutual information estimator in Belghazi et al. [239] does not hold for $\hat{I}(\boldsymbol{\vartheta}, \boldsymbol{x})$, but instead for the mutual information estimator under the *true* joint $\mathbb{E}_{p(\boldsymbol{\vartheta}, \boldsymbol{x})} \left[\log \hat{r}(\boldsymbol{x} \mid \boldsymbol{\vartheta})\right]$. While the intuition sketched above clearly connects with the posterior approximation $\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})$, we are not able to extract any meaningful criteria that we can enforce during training to ensure $H(\boldsymbol{\vartheta} \mid \boldsymbol{x}) \leq \hat{H}(\boldsymbol{\vartheta} \mid \boldsymbol{x})$.

Observation

It is worth noting that whenever,

$$\mathbb{E}_{p(\boldsymbol{\vartheta})p(\boldsymbol{x})} \left[\hat{r}(\boldsymbol{x} \mid \boldsymbol{\vartheta})\right] \leq 1, \tag{6.12}$$

we equivalently have

$$\mathbb{E}_{p(\boldsymbol{\vartheta})p(\boldsymbol{x})r(\boldsymbol{x} \mid \boldsymbol{\vartheta})} \left[\frac{\hat{r}(\boldsymbol{x} \mid \boldsymbol{\vartheta})}{r(\boldsymbol{x} \mid \boldsymbol{\vartheta})}\right] \leq 1. \tag{6.13}$$

Through Jensen's inequality, we obtain

$$\log \mathbb{E}_{p(\boldsymbol{\vartheta})p(\boldsymbol{x})r(\boldsymbol{x} \mid \boldsymbol{\vartheta})} \left[\frac{\hat{r}(\boldsymbol{x} \mid \boldsymbol{\vartheta})}{r(\boldsymbol{x} \mid \boldsymbol{\vartheta})}\right] \leq \log 1, \tag{6.14}$$

$$\mathbb{E}_{p(\boldsymbol{\vartheta})p(\boldsymbol{x})r(\boldsymbol{x} \mid \boldsymbol{\vartheta})} \left[\log \frac{\hat{r}(\boldsymbol{x} \mid \boldsymbol{\vartheta})}{r(\boldsymbol{x} \mid \boldsymbol{\vartheta})}\right] \leq \log \mathbb{E}_{p(\boldsymbol{\vartheta})p(\boldsymbol{x})r(\boldsymbol{x} \mid \boldsymbol{\vartheta})} \left[\frac{\hat{r}(\boldsymbol{x} \mid \boldsymbol{\vartheta})}{r(\boldsymbol{x} \mid \boldsymbol{\vartheta})}\right], \tag{6.15}$$

and therefore

$$\mathbb{E}_{p(\boldsymbol{\vartheta})p(\boldsymbol{x})r(\boldsymbol{x} \mid \boldsymbol{\vartheta})} \left[\log \frac{\hat{r}(\boldsymbol{x} \mid \boldsymbol{\vartheta})}{r(\boldsymbol{x} \mid \boldsymbol{\vartheta})}\right] \leq 0, \tag{6.16}$$

$$\mathbb{E}_{p(\boldsymbol{\vartheta}, \boldsymbol{x})} \left[\log \frac{\hat{r}(\boldsymbol{x} \mid \boldsymbol{\vartheta})}{r(\boldsymbol{x} \mid \boldsymbol{\vartheta})}\right] \leq 0. \tag{6.17}$$

Proving the inequality $\mathbb{E}_{p(\boldsymbol{\vartheta})p(\boldsymbol{x})} \left[\hat{r}(\boldsymbol{x} \mid \boldsymbol{\vartheta})\right] \leq 1$ implies the estimator $\mathbb{E}_{p(\boldsymbol{\vartheta}, \boldsymbol{x})} \left[\log \hat{r}(\boldsymbol{x} \mid \boldsymbol{\vartheta})\right]$ is a lower-bound to $I(\boldsymbol{\vartheta}, \boldsymbol{x})$. Practically, this means

that we could enforce $\mathbb{E}_{p(\boldsymbol{\vartheta})p(\boldsymbol{x})}\left[\hat{r}(\boldsymbol{x} \mid \boldsymbol{\vartheta})\right] = 1$ during the minimization of the loss functional. While this formulation leads to conservative credible regions (at least experimentally), we found that there was a significant deficit in the ability of the posterior estimator to constrain the target parameters. We are as of this moment unable to exactly determine the source of this effect, as the Bayes optimal discriminator shares the property that $\mathbb{E}_{p(\boldsymbol{\vartheta})p(\boldsymbol{x})}\left[r(\boldsymbol{x} \mid \boldsymbol{\vartheta})\right] = 1$, because

$$\mathbb{E}_{p(\boldsymbol{\vartheta})p(\boldsymbol{x})}\left[r(\boldsymbol{x} \mid \boldsymbol{\vartheta})\right] = \mathbb{E}_{p(\boldsymbol{\vartheta})p(\boldsymbol{x})}\left[\frac{p(\boldsymbol{\vartheta}, \boldsymbol{x})}{p(\boldsymbol{\vartheta})p(\boldsymbol{x})}\right], \tag{6.18}$$

$$= \int\int \mathrm{d}\,\boldsymbol{\vartheta}\,\mathrm{d}\,\boldsymbol{x}\;\; p(\boldsymbol{\vartheta})p(\boldsymbol{x}) = 1. \tag{6.19}$$

6.2 THE BALANCING CONDITION

Binary classification is the backbone of many likelihood-free protocols, including NRE presented in Chapter 3, due to its ability to effectively approximate density ratios. Whenever a binary classifier or discriminator is trained to distinguish between densities $p(\boldsymbol{x})$ with class-label 1 and $q(\boldsymbol{x})$ with class-label 0, the Bayes optimal discriminator can be expressed as

$$d(\boldsymbol{x}) = \frac{p(\boldsymbol{x})}{p(\boldsymbol{x}) + q(\boldsymbol{x})}, \tag{6.20}$$

and consequently, the density-ratio

$$r(x) = \frac{d(x)}{1 - d(x)} = \frac{p(x)}{q(x)}. \tag{6.21}$$

The combination of this result and the modern machine learning toolbox, makes binary classification a powerful candidate to solve intractable inference problems. However, given that we are interested in conservative approximations, we have to ask ourselves the question: **what does it mean for a binary classifier to be conservative and how does this connect to a posterior approximation?**

To answer this question we have to look at the decision function of an arbitrary approximate discriminator with respect to its Bayes optimal version. First, note that whenever the discriminator output increases, so does the ratio since

$$d(x) = \frac{r(x)}{r(x) + 1}. \tag{6.22}$$

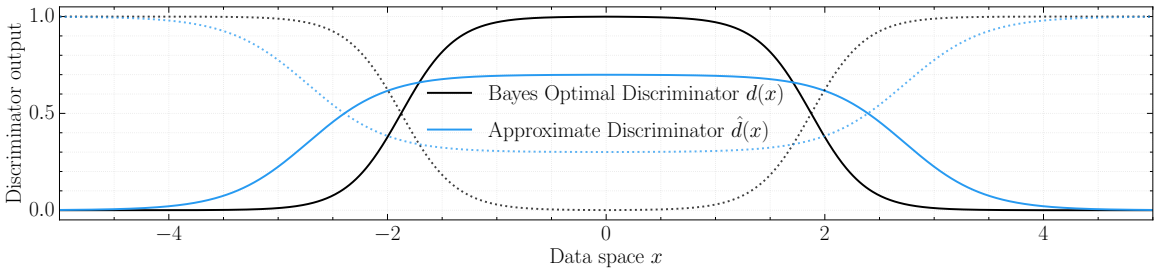

Figure 6.2: Decision functions of the Bayes optimal discriminator $d(x)$ (in black) and an arbitrary approximate discriminator $\hat{d}(x)$ (in blue). The dotted lines show their inverses, i.e. $1 - d(x)$ and $1 - \hat{d}(x)$ respectively in the same color scheme. For an approximate discriminator to be reliable or robust, it means that its discriminator output must be larger compared to its Bayes optimal version for samples with class-label 0, and the reverse for samples with class-label 1.

Connecting this with the discriminator $d(\boldsymbol{\vartheta}, \boldsymbol{x})$, where $p(\boldsymbol{\vartheta}, \boldsymbol{x})$ and $p(\boldsymbol{\vartheta})p(\boldsymbol{x})$ take the role of $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ respectively,it means that $d(\boldsymbol{\vartheta}, \boldsymbol{x})$ can be re-expressed as

$$\frac{r(\boldsymbol{x} \mid \boldsymbol{\vartheta})}{r(\boldsymbol{x} \mid \boldsymbol{\vartheta}) + 1} = \frac{p(\boldsymbol{\vartheta}) r(\boldsymbol{x} \mid \boldsymbol{\vartheta})}{p(\boldsymbol{\vartheta}) r(\boldsymbol{x} \mid \boldsymbol{\vartheta}) + p(\boldsymbol{\vartheta})} = \frac{p(\boldsymbol{\vartheta} \mid \boldsymbol{x})}{p(\boldsymbol{\vartheta} \mid \boldsymbol{x}) + p(\boldsymbol{\vartheta})}. \tag{6.23}$$

The above shows that the discriminator output $d(\boldsymbol{\vartheta}, \boldsymbol{x})$ is directly tied to the posterior density function, albeit in a non-linear fashion. The same holds for any approximate discriminator $\hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x})$.

From this discriminator perspective, we would be able to increase the differential entropy of the posterior approximation by "smearing" the decision function of $\hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x})$ with respect to $d(\boldsymbol{\vartheta}, \boldsymbol{x})$, as illustrated in Figure 6.2. More formally, this would mean that the approximate discriminator $\hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x})$ should satisfy

$$\mathbb{E}_{p(\boldsymbol{\vartheta})p(\boldsymbol{x})} \left[\hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x}) \right] \geq \mathbb{E}_{p(\boldsymbol{\vartheta})p(\boldsymbol{x})} \left[d(\boldsymbol{\vartheta}, \boldsymbol{x}) \right] \tag{6.24}$$

for samples with class-label 0 (from $p(\boldsymbol{\vartheta})p(\boldsymbol{x})$), but at the same time for samples with class-label 1 (from $p(\boldsymbol{\vartheta}, \boldsymbol{x})$)

$$\mathbb{E}_{p(\boldsymbol{\vartheta}, \boldsymbol{x})} \left[\hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x}) \right] \leq \mathbb{E}_{p(\boldsymbol{\vartheta}, \boldsymbol{x})} \left[d(\boldsymbol{\vartheta}, \boldsymbol{x}) \right]. \tag{6.25}$$

However, these two conditions are not sufficient. Decision functions could exist, at least in principle, where

$$\mathbb{E}_{p(\boldsymbol{\vartheta})p(\boldsymbol{x})} \left[\hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x}) \right] \geq \mathbb{E}_{p(\boldsymbol{\vartheta}, \boldsymbol{x})} \left[\hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x}) \right]. \tag{6.26}$$

This inequality would force credible regions away from the joint $p(\boldsymbol{\vartheta}, \boldsymbol{x})$, because samples from the product of the marginals are attributed with higher posterior density in expectation (after normalization of the posterior approximation). Therefore, the reverse should actually hold.

To enforce these conditions, we propose **the balancing condition** as an additional term to classical loss functionals for classification such as the binary cross entropy.

Definition 6. *The **balancing condition** involving two probability densities $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ is defined as $\mathbb{E}_{p(\boldsymbol{x})}[\hat{d}(\boldsymbol{x})] + \mathbb{E}_{q(\boldsymbol{x})}[\hat{d}(\boldsymbol{x})] = 1$ or $\mathbb{E}_{p(\boldsymbol{x})}[\hat{d}(\boldsymbol{x})] = \mathbb{E}_{q(\boldsymbol{x})}[1 - \hat{d}(\boldsymbol{x})]$ for any binary classifier or discriminator $\hat{d}(\boldsymbol{x})$.*

Definition 7. *A binary discriminator $\hat{d}$ is said to be **balanced** whenever it satisfies the balancing condition, i.e., whenever*

$$\mathbb{E}_{p(\boldsymbol{x})}[\hat{d}(\boldsymbol{x})] = \mathbb{E}_{q(\boldsymbol{x})}[1 - \hat{d}(\boldsymbol{x})]. \tag{6.27}$$

For the binary cross entropy

$$\mathcal{L}\left[\hat{d}(\boldsymbol{x})\right] \triangleq -\mathbb{E}_{p(\boldsymbol{x})}\left[\log \hat{d}(\boldsymbol{x})\right] - \mathbb{E}_{q(\boldsymbol{x})}\left[\log(1 - \hat{d}(\boldsymbol{x}))\right], \tag{6.28}$$

and given the balancing condition only depends on samples from $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$, the full loss functional with the balancing condition becomes

$$\mathcal{L}_b\left[\hat{d}(\boldsymbol{x})\right] \triangleq \mathcal{L}\left[\hat{d}(\boldsymbol{x})\right] + \lambda\left[\mathbb{E}_{p(\boldsymbol{x})}\left[\hat{d}(\boldsymbol{x})\right] + \mathbb{E}_{q(\boldsymbol{x})}\left[\hat{d}(\boldsymbol{x})\right] - 1\right]^2, \tag{6.29}$$

where λ is a hyperparameter (scalar) controlling the strength of the balancing condition's contribution. The balancing condition could thus be viewed through the lens of regularization. It should be noted however that the balancing condition differs from traditional regularization schemes in the sense that the balancing condition **needs** to be 0 for a discriminator to be balanced. This means that the hyperparameter controlling the strength of the balancing condition (here λ) can be set arbitrarily large, at least in principle. We found $\lambda = 100.0$ to work well across a wide range of problem domains.

Theorem 1. *The Bayes optimal discriminator $d(\boldsymbol{x})$ is balanced.*

Proof. Using the Bayes optimal discriminator

$$d(\boldsymbol{x}) \triangleq \frac{p(\boldsymbol{x})}{p(\boldsymbol{x}) + q(\boldsymbol{x})}, \tag{6.30}$$

and expressing the balancing condition in its integral form,

$$\Leftrightarrow \int d(\boldsymbol{x}) \cdot (p(\boldsymbol{x}) + q(\boldsymbol{x}))\, \mathrm{d}\,\boldsymbol{x}, \tag{6.31}$$

$$\Leftrightarrow \int \frac{p(\boldsymbol{x}) \cdot (p(\boldsymbol{x}) + q(\boldsymbol{x}))}{p(\boldsymbol{x}) + q(\boldsymbol{x})}\, \mathrm{d}\,\boldsymbol{x}, \tag{6.32}$$

$$\Leftrightarrow \int p(\boldsymbol{x})\, \mathrm{d}\,\boldsymbol{x} = 1, \tag{6.33}$$

because $p(\boldsymbol{x})$ is a proper probability density that integrates to 1. □

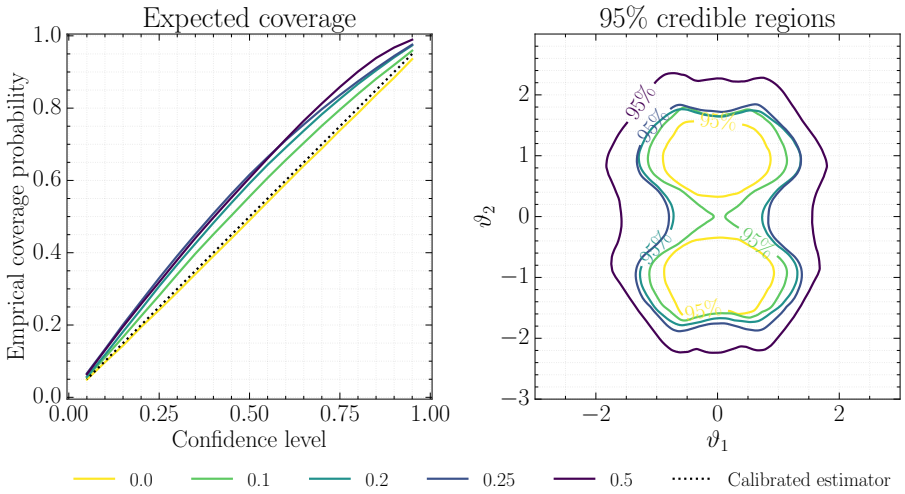

Figure 6.3: Weight-decay directly affects the conservativeness of the posterior estimator on a toy example. Larger values of weight-decay are associated with more conservative posteriors and credible regions. Although classical regularization techniques might prove successful to learn conservative posteriors, there is a lot of tuning involved. Moreover, there are no guarantees that regularizers such as weight decay share the same global optimum. In the case of weight decay, this is because the regularizer implicitly specifies a prior over the weights of the binary classifier.

This result implies that any *balanced* discriminator shares the same global minimum as the Bayes optimal discriminator. It should be noted this is not necessarily guaranteed with other regularization techniques such as weight decay, although they might positively affect the reliability of the binary classifier as well, as shown in Figure 6.3.

Theorem 2. *A balanced discriminator $\hat{d}(\boldsymbol{x})$ satisfies*

$$\mathbb{E}_{p(\boldsymbol{x})}\left[\log \frac{d(\boldsymbol{x})}{\hat{d}(\boldsymbol{x})}\right] \geq 0. \tag{6.34}$$

Proof. From the balancing condition, we have that

$$\int \mathrm{d}\,\boldsymbol{x}\,(p(\boldsymbol{x}) + q(\boldsymbol{x}))\,\hat{d}(\boldsymbol{x}), \tag{6.35}$$

$$= \mathbb{E}_{p(\boldsymbol{x})}\left[\hat{d}(\boldsymbol{x})\right] + \mathbb{E}_{q(\boldsymbol{x})}\left[\hat{d}(\boldsymbol{x})\right], \tag{6.36}$$

$$= 1. \tag{6.37}$$

$(p(\boldsymbol{x}) + q(\boldsymbol{x}))\hat{d}(\boldsymbol{x})$ is therefore a density that integrates to 1 and therefore the KL is defined such that

$$\text{KL}\left[p(\boldsymbol{x}) \;||\; (p(\boldsymbol{x}) + q(\boldsymbol{x}))\hat{d}(\boldsymbol{x})\right] \geq 0, \tag{6.38}$$

$$\Leftrightarrow \mathbb{E}_{p(\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x})}{(p(\boldsymbol{x}) + q(\boldsymbol{x}))\hat{d}(\boldsymbol{x})}\right] \geq 0, \tag{6.39}$$

$$\Leftrightarrow \mathbb{E}_{p(\boldsymbol{x})}\left[\log \frac{d(\boldsymbol{x})}{\hat{d}(\boldsymbol{x})}\right] \geq 0. \tag{6.40}$$

After applying Jensen's inequality, we obtain

$$\log \mathbb{E}_{p(\boldsymbol{x})}\left[\frac{d(\boldsymbol{x})}{\hat{d}(\boldsymbol{x})}\right] \geq \mathbb{E}_{p(\boldsymbol{x})}\left[\log \frac{d(\boldsymbol{x})}{\hat{d}(\boldsymbol{x})}\right] \geq \log 1, \tag{6.41}$$

$$\log \mathbb{E}_{p(\boldsymbol{x})}\left[\log \frac{d(\boldsymbol{x})}{\hat{d}(\boldsymbol{x})}\right] \geq \log 1, \tag{6.42}$$

and finally after exponentiating both sides

$$\mathbb{E}_{p(\boldsymbol{x})}\left[\frac{d(\boldsymbol{x})}{\hat{d}(\boldsymbol{x})}\right] \geq 1. \tag{6.43}$$

However, note that the reverse

$$\mathbb{E}_{p(\boldsymbol{x})}\left[\frac{\hat{d}(\boldsymbol{x})}{d(\boldsymbol{x})}\right] = 1 \tag{6.44}$$

holds as a direct result of the balancing condition, because

$$p(\boldsymbol{x}) = d(\boldsymbol{x})(p(\boldsymbol{x}) + q(\boldsymbol{x})). \tag{6.45}$$

□

Theorem 3. *A balanced discriminator $\hat{d}(\boldsymbol{x})$ satisfies*

$$\mathbb{E}_{q(\boldsymbol{x})}\left[\log \frac{1 - d(\boldsymbol{x})}{1 - \hat{d}(\boldsymbol{x})}\right] \geq 0. \tag{6.46}$$

Proof. From the balancing condition, we have that

$$\int \mathrm{d}\,\boldsymbol{x}\,(p(\boldsymbol{x}) + q(\boldsymbol{x}))\,(1 - \hat{d}(\boldsymbol{x})), \tag{6.47}$$

$$= \mathbb{E}_{p(\boldsymbol{x})}\left[1 - \hat{d}(\boldsymbol{x})\right] + \mathbb{E}_{q(\boldsymbol{x})}\left[1 - \hat{d}(\boldsymbol{x})\right], \tag{6.48}$$

$$= 2 - \int \mathrm{d}\,\boldsymbol{x}\;\hat{d}(\boldsymbol{x})(p(\boldsymbol{x}) + q(\boldsymbol{x})), \tag{6.49}$$

$$= 1. \tag{6.50}$$

$(p(\boldsymbol{x}) + q(\boldsymbol{x}))(1 - \hat{d}(\boldsymbol{x}))$ is therefore a density that integrates to 1 and therefore the KL is defined such that

$$\text{KL}\left[q(\boldsymbol{x}) \;||\; (p(\boldsymbol{x}) + q(\boldsymbol{x}))(1 - \hat{d})(\boldsymbol{x})\right] \geq 0, \tag{6.51}$$

$$\Leftrightarrow \mathbb{E}_{q(\boldsymbol{x})}\left[\log \frac{q(\boldsymbol{x})}{(p(\boldsymbol{x}) + q(\boldsymbol{x}))(1 - \hat{d}(\boldsymbol{x}))}\right] \geq 0, \tag{6.52}$$

$$\Leftrightarrow \mathbb{E}_{q(\boldsymbol{x})}\left[\log \frac{1 - d(\boldsymbol{x})}{1 - \hat{d}(\boldsymbol{x})}\right] \geq 0. \tag{6.53}$$

After applying Jensen's inequality, we obtain

$$\log \mathbb{E}_{1(\boldsymbol{x})}\left[\frac{1-d(\boldsymbol{x})}{1-\hat{d}(\boldsymbol{x})}\right] \geq \mathbb{E}_{p(\boldsymbol{x})}\left[\log \frac{1-d(\boldsymbol{x})}{1-\hat{d}(\boldsymbol{x})}\right] \geq \log 1, \tag{6.54}$$

$$\log \mathbb{E}_{1(\boldsymbol{x})}\left[\log \frac{1-d(\boldsymbol{x})}{1-\hat{d}(\boldsymbol{x})}\right] \geq \log 1, \tag{6.55}$$

and finally after exponentiating both sides

$$\mathbb{E}_{q(\boldsymbol{x})}\left[\frac{1-d(\boldsymbol{x})}{1-\hat{d}(\boldsymbol{x})}\right] \geq 1. \tag{6.56}$$

However, note that the reverse

$$\mathbb{E}_{q(\boldsymbol{x})}\left[\frac{1-\hat{d}(\boldsymbol{x})}{1-d(\boldsymbol{x})}\right] = 1 \tag{6.57}$$

holds as a direct result of the balancing condition, because

$$q(\boldsymbol{x}) = (1-d(\boldsymbol{x}))(p(\boldsymbol{x})+q(\boldsymbol{x})). \tag{6.58}$$

□

Theorem 4. *For a balanced discriminator $\hat{d}(\boldsymbol{x})$, the following relation between the loss terms must hold*

$$\mathcal{L}\left[\hat{d}(\boldsymbol{x})\right] \geq \mathcal{L}\left[d(\boldsymbol{x})\right]. \tag{6.59}$$

Proof. Assume the negation, i.e.,

$$\mathcal{L}\left[\hat{d}(\boldsymbol{x})\right] < \mathcal{L}\left[d(\boldsymbol{x})\right]. \tag{6.60}$$

We obtain the following by expanding the loss-terms into the binary cross-entropy

$$-\mathbb{E}_{p(\boldsymbol{x})}\left[\log \hat{d}(\boldsymbol{x})\right] - \mathbb{E}_{q(\boldsymbol{x})}\left[\log(1-\hat{d}(\boldsymbol{x}))\right] < -\mathbb{E}_{p(\boldsymbol{x})}\left[\log d(\boldsymbol{x})\right] - \mathbb{E}_{q(\boldsymbol{x})}\left[\log(1-d(\boldsymbol{x}))\right], \tag{6.61}$$

$$\Leftrightarrow 0 < \mathbb{E}_{p(\boldsymbol{x})}\left[\log \frac{\hat{d}(\boldsymbol{x})}{d(\boldsymbol{x})}\right] + \mathbb{E}_{q(\boldsymbol{x})}\left[\log \frac{1-\hat{d}(\boldsymbol{x})}{1-d(\boldsymbol{x})}\right]. \tag{6.62}$$

We have reached a contradiction because we have previously shown both terms to be negative whenever the approximate discriminator $\hat{d}(\boldsymbol{x})$ is balanced.

The minimization of the binary cross-entropy with a balanced discriminator is thus equivalent to the strict minimization of both $\mathrm{KL}\left[p(\boldsymbol{x}) \mid\mid (p(\boldsymbol{x})+q(\boldsymbol{x}))\hat{d}(\boldsymbol{x})\right]$ and $\mathrm{KL}\left[q(\boldsymbol{x}) \mid\mid (p(\boldsymbol{x})+q(\boldsymbol{x}))(1-\hat{d}(\boldsymbol{x}))\right]$.

□

The above establishes that any approximate balanced discriminator will never – in expectation – be overconfident with respect to its Bayes optimal version for samples of $p(\boldsymbol{x})$ (class label 1) and $q(\boldsymbol{x})$ (class label 0).

Algorithm 7 shows how the balancing condition can be used in simulation-based inference, we refer to this algorithm henceforth as CNRE (conservative NRE). In particular, as an improvement for the previously introduced NRE protocol.

Algorithm 7 Optimization procedure of Conservative Neural Ratio Estimation (CNRE).

Inputs:	Implicit generative model $p(\boldsymbol{x} \mid \boldsymbol{\vartheta})$ and prior $p(\boldsymbol{\vartheta})$
Outputs:	Approximate discriminator $\hat{d}_\psi(\boldsymbol{\vartheta}, \boldsymbol{x})$
Hyperparameters:	Balancing condition strength λ (default = 100.0)

repeat
  $\mathcal{L}[\hat{d}_\psi(\boldsymbol{\vartheta}, \boldsymbol{x})] = -\mathbb{E}_{p(\boldsymbol{\vartheta}, \boldsymbol{x})}[\log \hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x})] - \mathbb{E}_{p(\boldsymbol{\vartheta})p(\boldsymbol{x})}[\log(1 - \hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x}))]$
  $\mathcal{B}[\hat{d}_\psi(\boldsymbol{\vartheta}, \boldsymbol{x})] = \mathbb{E}_{p(\boldsymbol{\vartheta}, \boldsymbol{x})}[\hat{d}_\psi(\boldsymbol{\vartheta}, \boldsymbol{x})] + \mathbb{E}_{p(\boldsymbol{\vartheta})p(\boldsymbol{x})}[\hat{d}_\psi(\boldsymbol{\vartheta}, \boldsymbol{x})]$.
  $\psi = \text{MINIMIZER}\left(\psi,\ \mathcal{L}[\hat{d}_\psi(\boldsymbol{\vartheta}, \boldsymbol{x})] + \lambda(\mathcal{B}[\hat{d}_\psi(\boldsymbol{\vartheta}, \boldsymbol{x})] - 1)^2\right)$
until not converged
return $\hat{d}_\psi(\boldsymbol{\vartheta}, \boldsymbol{x})$.

6.3 EXPERIMENTS

We explore the effect of the balancing condition on benchmarks from various fields ranging from toy problems to real applications. Conservativeness is evaluated through the lens of expected coverage with the intent of demonstrating that the balancing condition leads to more conservative posteriors and credible regions.

6.3.1 *Setup*

Our evaluations consider simulation budgets ranging from 2^{10} up to 2^{17} samples and confidence levels from 0.05 up to 0.95. We train, for every simulation budget, 5 posterior estimators for 100 epochs. The empirical expected coverage probability is evaluated on at least 5,000 unseen samples from the joint $p(\boldsymbol{\vartheta}, \boldsymbol{x})$ and for all confidence levels under consideration. In addition, we repeat the expected coverage evaluation for ensembles of 5 estimators as well. We evaluate the proposed balancing condition on the same problem domains as in Chapter 5 and compare against NRE and NPE.

ENSEMBLE The CNRE ensemble is defined as

$$\hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x}) \triangleq \frac{1}{n} \sum_{i=1}^{n} \hat{d}_i(\boldsymbol{\vartheta}, \boldsymbol{x}), \tag{6.63}$$

to ensure that an ensemble of balanced discriminators is balanced.

6.3.2 *Results*

Across all problem domains we observe that CNRE is conservative. The proposed technique passes the expected coverage diagnostic for all settings. We observe the balancing condition enforces the desired behavior with small simulation budgets, where the uncertainty of the target parameter $\boldsymbol{\vartheta}$ is be expected to be large. A compressed summary of this result in presented in Figure 6.4 through the notion of *area under expected coverage*. The metric quantifies the area under the expected coverage curve for all credible levels, akin to the classical area under curve that quantifies the performance of a binary classifier. From this perspective, the effect of the balancing condition (CNRE) on expected coverage is directly apparent. Figure 6.5 shows all expected coverage curves.

In addition, we evaluate the value balancing condition

$$\mathbb{E}_{p(\boldsymbol{\vartheta}, \boldsymbol{x})} \left[\hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x}) \right] + \mathbb{E}_{p(\boldsymbol{\vartheta}) p(\boldsymbol{x})} \left[\hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x}) \right] \tag{6.64}$$

for both NRE and CNRE. All CNRE-based estimators are balanced after training. We observed that for NRE in particular there is a lot of variability present for small simulation budgets, suggesting that NRE could be overconfident in these regimes. This suspicion is reflected in Figure 6.4 for both NRE and NPE. However, NPE and NRE tend towards being balanced as the simulation budget increases.

Finally, as our ultimate goal targets conservative credible regions, we measure the effect of applying the balancing condition on the size of the estimated credible regions for various credible levels. The results are summarized in Figure 6.6. We observe the desired effect with CNRE: an increase in uncertainty while at the same time tending towards the NRE approximation for large simulation budgets.

It should be noted however that some efficiency in terms of credible region size is lost, i.e., the constraints are not as strong as with NRE. This is to be expected, although both NRE and CNRE share the same global minimum. The question remains whether this relates to the training procedure, or if the balancing condition is effectively enforcing more conservative approximations for larger simulation budgets. Suggesting that either the simulation budget is not sufficiently large to capture all information, or the trained model is not expressive enough.

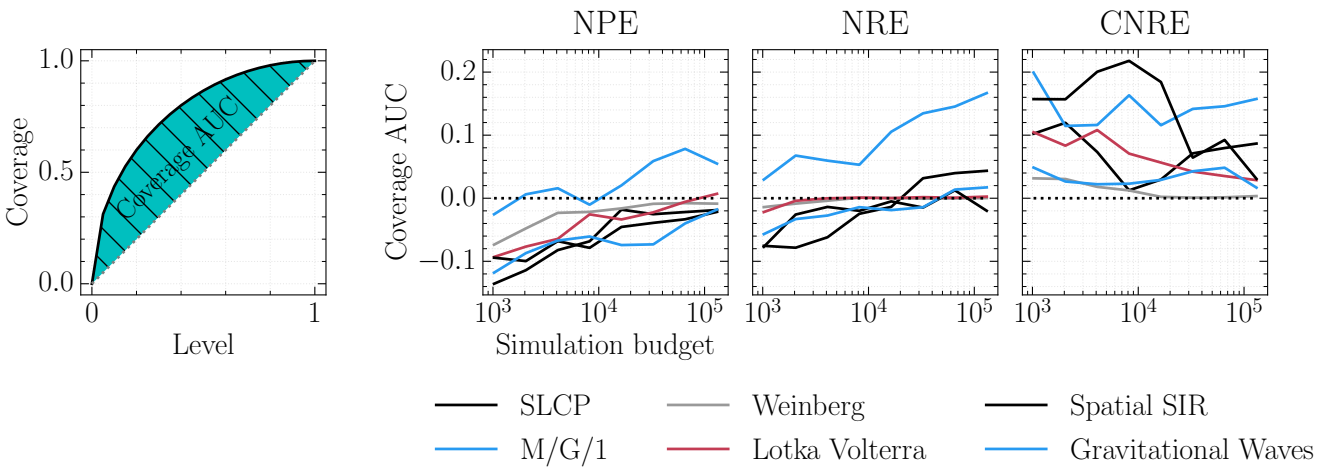

Figure 6.4: Direct comparison between NPE, NRE and CNRE using the area under expected coverage as a metric. The metric is analogous to the classical area under curve, as shown on the left. However, in this case a positive area under curve implies that the learned estimator is conservative in expectation. Whereas a negative value implies that the estimator is overconfident. From this figure, we observe a clear trend between NPE and NRE. They are especially overconfident for small simulation budgets – associated with a lot of uncertainty –, while CNRE is not. On these benchmarks, CNRE is always conservative and enforces a significant amount of uncertainty to smaller simulation budgets while tending towards being a calibrated estimator for large simulation budgets. This particular behavior is illustrated in Figure 6.6.

Although unlikely, the issue relating to the training procedure might still hold some merit as enforcing the balancing condition, especially with large λ, increases the difficulty of the optimization procedure. More training epochs could thus have been required to achieve the same efficiency.

6.4 SUMMARY

In conclusion, the balancing condition is an effective and easily applicable technique that can be used in any binary classification problem. Experiments demonstrate that the technique does in fact increases the reliability of the trained discriminators in practice and their approximated posteriors. Note however that these experiment do not consider the effect of distribution shift. **It might still be possible that cnre fails the expected coverage diagnostic due to a distribution shift with respect to true posterior**, even though its credible regions might be larger than they should be. However, research surrounding

the balancing condition is not complete. There are still several open questions that relate to posterior approximations and their credible regions, i.e., can we connect the balancing condition to differential posterior entropy?

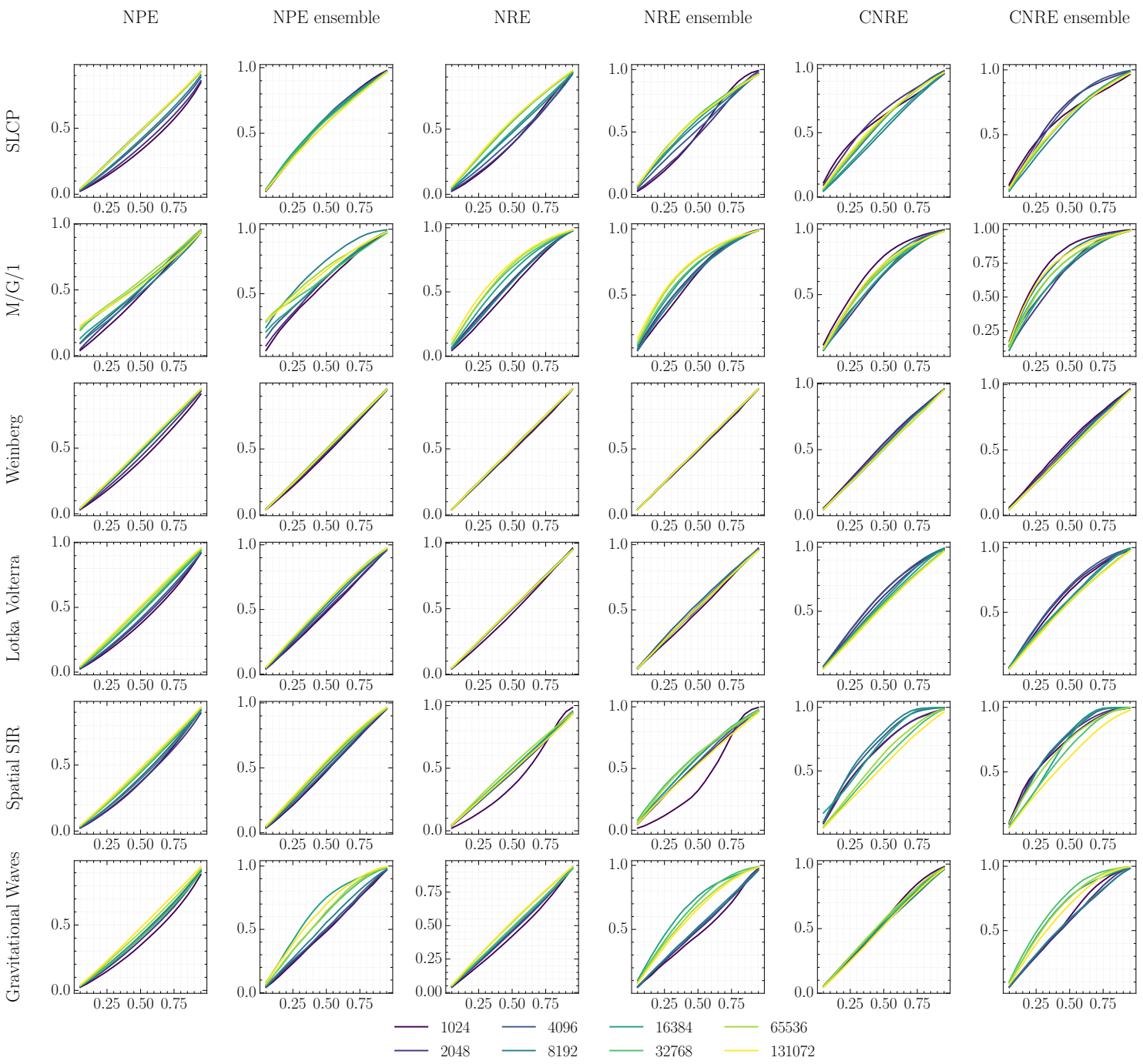

Figure 6.5: Full results of the expected coverage diagnostic for several problem domains and simulation budgets. The results for NPE and NRE are in agreement with the results from Chapter 5: these methods can in fact produce non-conservative approximations. The balancing condition shows a lot of promise. In fact, CNRE employs the same setup as NRE (model architecture, hyperparameters) with the exception of the balancing condition. Interestingly, on average, CNRE seems to be most conservative for small simulation budgets (darker lines), while the reverse is mostly true for NPE and NRE.

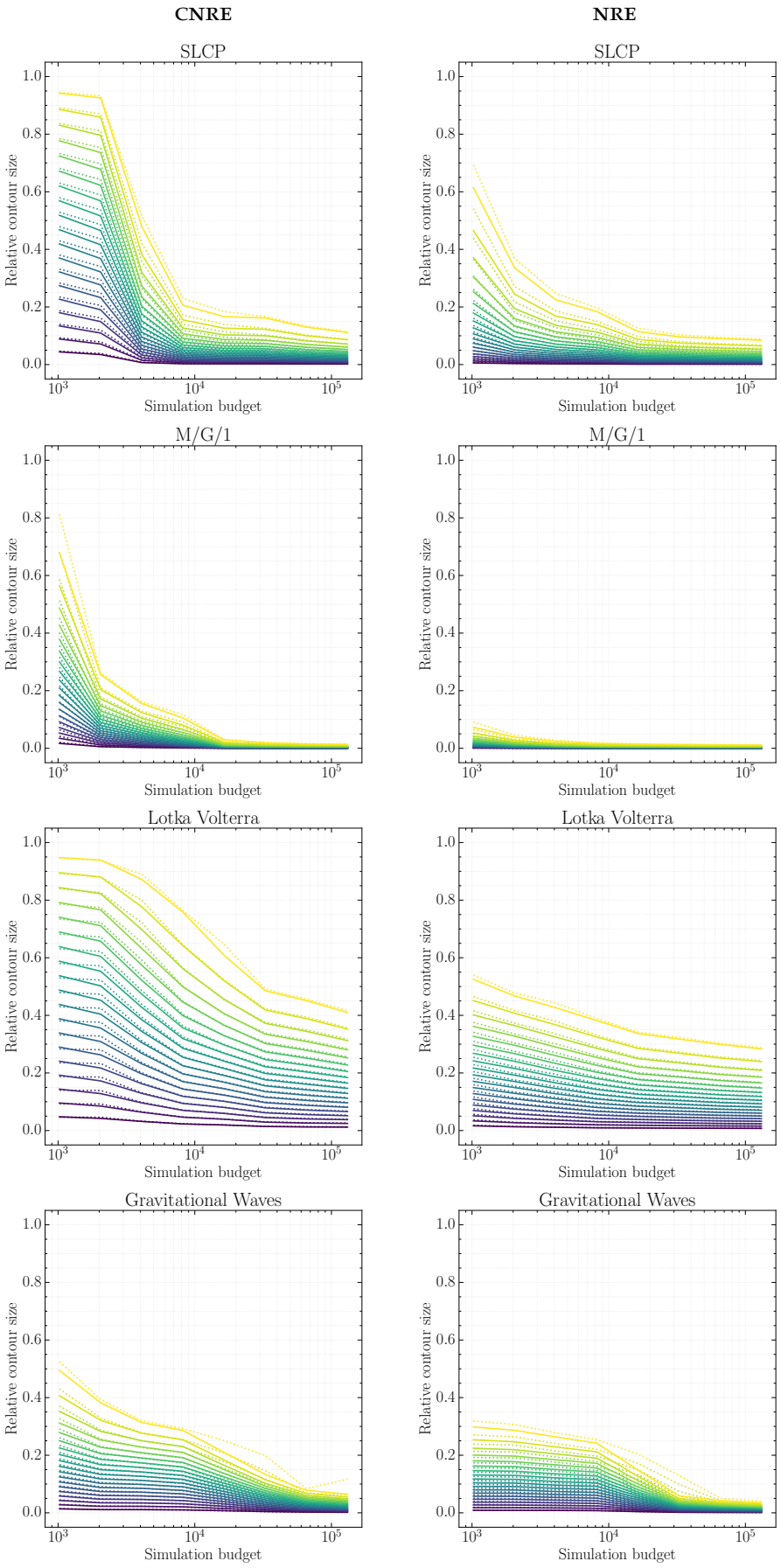

Figure 6.6: Shows the relative size for the credible regions with respect to the prior for various benchmarks. Meaning, a value of 0.95 corresponds to a spanned area of 95% within the support of the prior $p(\boldsymbol{\vartheta})$. Figures on the left show CNRE, while NRE is shown on the right. CNRE shows larger credible regions compared to NRE, especially in small simulation budget regimes, but is eventually on par with NRE.

Part III

CONCLUSION AND PROSPECTS

7

Conclusions

TL;DR

- Current techniques are not suitable for problem domains with a large dimensionality of the target parameter space ϑ. Advances in this area could help specific applications. For instance, posteriors over high-resolution depth of field maps, mass sheets (for gravitational lensing), climate models, and so on. The techniques presented in this dissertation, at least in their current form, are not be able to handle these dimensionalities. Further work is required in this area.
- Density-ratio estimation through binary classification is a powerful technique that enables us to approximate statistical quantities that are otherwise not tractable at training time.
- All vanilla simulation-based inference protocols, despite theoretical guarantees at a global optimum, can produce unreliable (non-conservative) constraints. Diagnostics should be applied to determine the reliability of the estimator before making any scientific conclusion.
- Currently, inference protocols that are amortized and thus do not require retraining or new simulations should be favoured over their non-amortized counterparts because the reliability of amortized techniques can be practically determined without the need of a ground truth (which is never available in practice anyway).
- We introduce a new criterion, the balancing condition, that enforces a binary classifier to be conservative with respect to the Bayes optimal classifier. The criterion comes with theoretical guarantees. Experiments show this reliability translates to the estimated credible regions (Bayesian constraints) in simulation-based inference.

- The balancing condition is applicable to all binary classification problems and could improve the reliability of other classification problems outside the field of simulation-based inference.

7.1 SUMMARY AND TAKE-AWAY MESSAGES

The contributions of this thesis are two-fold. They can be characterized as novel set of techniques and guidelines to solve intractable statistical inference problems. Part i introduces novel simulation-based inference protocols and provides guidelines towards effectively applying these techniques to scientific problems. We introduced two inference protocols; AVO in Chapter 2 and AALR-MCMC (NRE with MCMC) in Chapter 3.

AVO yields a proposal distribution $q_\psi(\boldsymbol{\vartheta})$ across target parameters $\boldsymbol{\vartheta}$ such that the Jensen-Shannon divergence between the marginal $q_{\boldsymbol{\vartheta}}(\boldsymbol{x})$ and an empirical dataset of observables $p_r(\boldsymbol{x})$ is minimized. The technique is recommended whenever a relatively large empirical dataset of observables is available. However, AVO is not suitable for scenarios where this dataset is small. The reason being that a small dataset would quickly saturate the learning signal of the discriminator attempting to optimize adversarial objective. In turn, a saturated learning signal would affect the gradient used in the optimization procedure of the proposal distribution. Another problematic aspect of AVO is its constant and costly reliance on the simulation model to evaluate the gradient with respect to the proposal parameters. Despite these aspects, the ability to deal with a large empirical dataset of observables is rather unique. For this reason, AVO lends itself easily to Empirical Bayes. For instance, by enabling population studies to determine a prior for subsequent deeper analyses.

The major advantage that AALR-MCMC (or NRE without MCMC) brings lies in its ability to approximate posteriors for all observables $\boldsymbol{x}$ supported by the marginal model $p(\boldsymbol{x})$. The technique achieves this by scanning a discretized grid of target parameters, or numerically with MCMC. Contrary to many existing inference protocols, the technique does *not* require new simulations or retraining to achieve this, while at the same time retaining its data efficiency and ability to probe the estimator's statistical quality.

Part ii explores the topic of *reliability* in simulation-based inference and more importantly, why reliability is critical to the sciences. We argue that we should not strive to learn exact approximations, something which is rarely achieved in practice anyway, but rather seek to fit *conservative* approximations. We stress this point by demonstrating that common (Bayesian) simulation-inference protocols *can* produce

non-conservative approximations, despite exactness guarantees these algorithms carry whenever their fitting criteria are globally optimized. Interestingly, our experiments showed various other aspects *(i)* amortization seems to be more conservative compared to non-amortized inference protocols, *(ii)* it is computationally inefficient and impractical to determine the reliability of non-amortized algorithms, limiting their applicability, and finally *(iii)* ensembling posterior estimators increases their joint reliability significantly and should be seen as an easy and effective method to construct more reliability statistical quantities.

Jointly, these results motivated the development of the *balancing condition*; our proposal to increase the *theoretical* and *practical* reliability of simulation-based inference. The technique *(i)* establishes a premise for conservative approximations *(ii)* comes with theoretical guarantees and *(iii)* is directly applicable, independent of the problem setting or hyperparameters such as simulation budget. Jointly, these properties provide an answer to the question: how can we learn reliable approximations whenever fitting criteria are *not* globally optimized? More importantly, the balancing condition is not limited to simulation-based inference. In fact, the technique and associated theory generalizes to all binary classification problems. The balancing condition is subject of further investigation to explore its full potential and fully understand its properties.

7.2 MOVING FORWARD

The scientific method is a constant recurrence over the loop "*theory*, *experiment*, *conclusion*". By building upon the assumptions made in the modelling (theory) and experiment step, this thesis contributes to the automation of the last step: the conclusion. However, our ultimate goal is to mechanize the complete scientific loop. In the following sections we give a few promising research directions that could enable the full automation of science, including the apparent creative process of modelling.

7.2.1 *Optimal Bayesian Experimental Design*

The field of Optimal Bayesian Experimental Design in its most basic setting attempts to answer the question: "What experimental configuration ψ has the potential to maximally reduce the uncertainty surrounding the target parameter $\boldsymbol{\vartheta}$?". Formally, it does so by assigning a utility score to an experimental configuration ψ in terms of the expected information gain

$$U(\psi) \triangleq \mathbb{E}_{p(\boldsymbol{x})} \left[\mathbb{H} \left[p(\boldsymbol{\vartheta}) \right] - \mathbb{H} \left[p(\boldsymbol{\vartheta} \mid \boldsymbol{x}, \psi) \right] \right], \tag{7.1}$$

which can be re-expressed as

$$U(\psi) \triangleq \mathbb{E}_{p(\boldsymbol{\vartheta},\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{\vartheta} \mid \boldsymbol{x}, \psi)}{p(\boldsymbol{\vartheta})}\right]. \tag{7.2}$$

Given this formulation, the objective of Optimal Bayesian Experimental Design is to find

$$\psi^* = \arg\max_{\psi} \; U(\psi). \tag{7.3}$$

The issue in Optimal Bayesian Experimental Design is the intractable utility function $U(\psi)$, akin to the problems we have been dealing with throughout this dissertation. To solve this in a naive fashion, we could apply NRE to approximate this quantity for a given experimental configuration ψ through

$$U(\psi) \approx \mathbb{E}_{p(\boldsymbol{\vartheta},\boldsymbol{x})}\left[\log \hat{r}(\boldsymbol{x} \mid \boldsymbol{\vartheta})\right], \tag{7.4}$$

where $\hat{r}(\boldsymbol{x} \mid \boldsymbol{\vartheta})$ is a likelihood-to-evidence ratio estimator trained to differentiate between samples from the conditional joint $p(\boldsymbol{\vartheta}, \boldsymbol{x} \mid \psi)$ and $p(\boldsymbol{\vartheta})p(\boldsymbol{x} \mid \psi)$. However, to maximize $U(\psi)$, we would need to be able to evaluate the utility for different values of ψ. Implying that we would need to re-apply this procedure and consequently retrain and resimulate a dataset of observables for every distinct evaluation of the utility function. Clearly, this greedy approach does not scale.

However, one could in fact amortize the estimation of $U(\psi)$ in the same way we amortized posterior estimation. To accomplish this, we only require a single presimulated dataset containing samples from the joint $p(\boldsymbol{\vartheta}, \boldsymbol{x}, \psi)$ and two ratio estimators that are trained in specific ways. In contrast to the previously outlines approach, both can now rely on the same presimulated dataset during training. Under the reasonable assumption that $\boldsymbol{\vartheta}$ and ψ are independent, the first ratio estimator type is trained to approximate

$$r(\boldsymbol{x} \mid \boldsymbol{\vartheta}, \psi) = \frac{p(\boldsymbol{\vartheta}, \boldsymbol{x}, \psi)}{p(\boldsymbol{\vartheta})p(\boldsymbol{x}, \psi)} = \frac{p(\boldsymbol{\vartheta} \mid \boldsymbol{x}, \psi)}{p(\boldsymbol{\vartheta})}. \tag{7.5}$$

The second ratio estimator type approximates

$$r(\boldsymbol{\vartheta}, \boldsymbol{x} \mid \psi) = \frac{p(\boldsymbol{\vartheta}, \boldsymbol{x}, \psi)}{p(\boldsymbol{\vartheta}, \boldsymbol{x})p(\psi)} = \frac{p(\boldsymbol{\vartheta}, \boldsymbol{x} \mid \psi)}{p(\boldsymbol{\vartheta}, \boldsymbol{x})}. \tag{7.6}$$

Jointly, both ratio estimators can be combined to compute

$$\mathbb{E}_{p(\boldsymbol{\vartheta},\boldsymbol{x})}\left[r(\boldsymbol{\vartheta}, \boldsymbol{x} \mid \psi) \log r(\boldsymbol{x} \mid \boldsymbol{\vartheta}, \psi)\right] \tag{7.7}$$

for a given experimental configuration ψ. This is equivalent to computing the expected information gain of an experimental configuration because

$$= \mathbb{E}_{p(\boldsymbol{\vartheta},\boldsymbol{x})}\left[r(\boldsymbol{\vartheta},\boldsymbol{x}\,|\psi)\log r(\boldsymbol{x}\,|\,\boldsymbol{\vartheta},\psi)\right], \tag{7.8}$$

$$= \int\!\!\int \mathrm{d}\,\boldsymbol{\vartheta}\,\mathrm{d}\,\boldsymbol{x}\;\; p(\boldsymbol{\vartheta},\boldsymbol{x})\frac{p(\boldsymbol{\vartheta},\boldsymbol{x}\,|\psi)}{p(\boldsymbol{\vartheta},\boldsymbol{x})}\log\frac{p(\boldsymbol{\vartheta}\,|\,\boldsymbol{x},\psi)}{p(\boldsymbol{\vartheta})}, \tag{7.9}$$

$$= \int\!\!\int \mathrm{d}\,\boldsymbol{\vartheta}\,\mathrm{d}\,\boldsymbol{x}\;\; p(\boldsymbol{\vartheta},\boldsymbol{x}\,|\psi)\log\frac{p(\boldsymbol{\vartheta}\,|\,\boldsymbol{x},\psi)}{p(\boldsymbol{\vartheta})}, \tag{7.10}$$

$$= U(\psi). \tag{7.11}$$

The above implies that we can reuse previously simulated points and essentially reweigh them to estimate the expectation for all experimental configurations supported by $p(\psi)$.

Figure 7.1 demonstrates this approach on a problem concerned a simulation of high energy particle collisions $e^+e^- \to \mu^+\mu^-$, where the experimental configuration ψ is the beam energy. The simulator outputs the Weinberg angle x in the standard model of particle physics. From the scattering angle, one can directly derive Fermi's constant $\boldsymbol{\vartheta}$. Because of this, we analytically know that inference procedures should be insensitive around ψ =45 GeV, which is reflected in the estimated information gain.

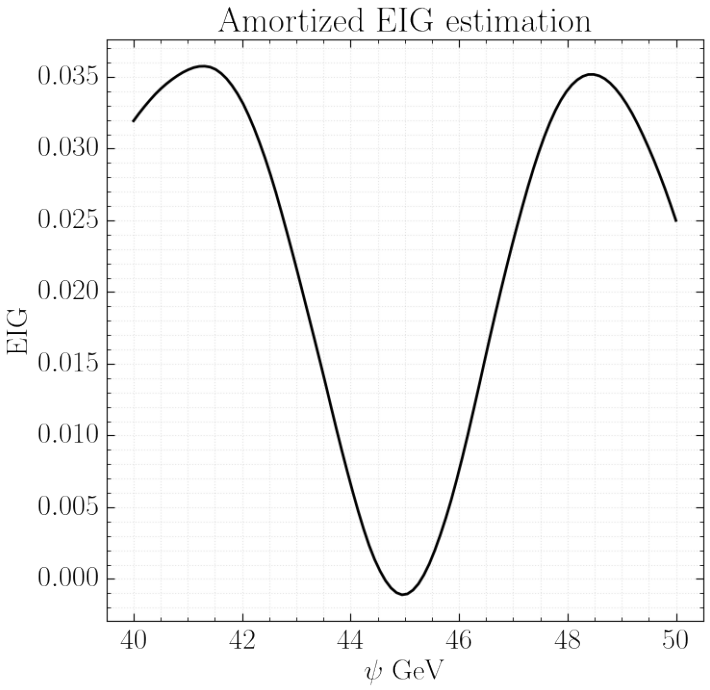

Figure 7.1: Amortization of the estimated information gain of the Weinberg problem used through this thesis. The expected information gain is estimated using the formulation put forward in Equation 7.7.

While in low-dimensional settings scanning might be a effective approach to maximize $U(\psi)$, it does not scale to larger dimensions. However, due to the amortization and the fully differentiable ratio estimators, we are able to directly compute gradients with respect to ψ through automatic differentiation. The availability of these gradients allow us to apply gradient ascent to directly maximize the expected information gain.

There are of course many nuances to the setup of Optimal Bayesian Experimental Design. While the above mainly focusses on maximizing the expected information gain under the assumption that one facilitates a single experiment, it does not necessarily mean that this is the optimal strategy (policy) across a certain budget or number of experiments. Amortizing the expected information gain in these settings could provide tremendous yields in computational performance for the estimation of the expected information gain across a horizon of experiments (utility) and open the door for the effective symbiosis between experimental design and the reinforcement learning field.

7.2.2 *Hypothesis synthesis*

The next major milestones in this line of research — of interest to many domain scientists — is the exploration of the hypothesis space to search for new physics and rediscover well-established facts. Although this task is inherently intractable, there are connections to be made with recent advances in AutoML, symbolic regression, neural program induction, (invertible) generative models, and the research presented in this dissertation. In particular, potential future work could investigate how a posterior over (sub)programs should be computed within existing simulation code. By sampling programs from such a hypothetical posterior, domain scientists would be able to efficiently probe the hypothesis space because inadequate programs are not evaluated due to their small or zero posterior probability. In addition to the computational advantages that stem from a constrained search space, this particular formulation lends itself to a well-defined and natural optimization criterion to search for and score probabilistic (sub)programs; the maximization of the likelihood-ratio, or Bayes factor between hypotheses. Moreover, this formulation could also be used to probe the validity of the *key assumption in simulation-based inference*, i.e., the assumption that simulation model conforms with nature. Testing for model misspecification under this formulation would thus correspond to testing whether the sampled residual (sub)programs are empty.

The above sketches a conceptual picture of a potential hypothesis search, where the representation of a theory is encoded as a com-

puter program. Ignoring research avenues such as fast conditional approximate simulation, program representation and sampling, this conceptual research avenue actually directly integrates within existing simulation-based inference workflows. With the added problem that we now consider various distinct versions of a computer simulator. Although, parameterized by these subprograms and essentially encoding the likelihood model as $p(\boldsymbol{x} \mid z)$, where z is a (potentially latent) program description. While it might seem that this formulation brings additional problems such as the model parameter space not being static, it does in fact generalize to the general Bayesian simulation-based inference setting. The only difference being that the marginal model $p(x)$ is now defined as

$$p(x) = \int \mathrm{d}\, z \; p(z) p(\boldsymbol{x} \,|z), \tag{7.12}$$

where $p(z)$ is a prior across all potential subprograms. Note that, much like the likelihood model in simulators, this prior does not necessarily have to be defined *explicitly*. Rather, it's definition can be done implicitly by means of some sampling or program synthesis procedure. For instance, syntactic constraints of the programming language.

Bibliography

[1] Donald E. Knuth. "Computer Programming as an Art." In: 17.12 (1974). ISSN: 0001-0782. DOI: 10.1145/361604.361612. URL: https://doi.org/10.1145/361604.361612.

[2] Adam Paszke et al. "Automatic differentiation in pytorch." In: (2017).

[3] J. D. Hunter. "Matplotlib: A 2D graphics environment." In: *Computing in Science & Engineering* 9.3 (2007), pp. 90–95. DOI: 10.1109/MCSE.2007.55.

[4] Stefan Van Der Walt, S Chris Colbert, and Gael Varoquaux. "The NumPy array: a structure for efficient numerical computation." In: *Computing in Science & Engineering* 13.2 (2011), p. 22.

[5] Thomas Kluyver et al. "Jupyter Notebooks – a publishing format for reproducible computational workflows." In: *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. Ed. by F. Loizides and B. Schmidt. IOS Press. 2016, pp. 87 –90.

[6] Benoit B Mandelbrot. *The fractal geometry of nature*. San Francisco, CA: Freeman, 1982. URL: https://cds.cern.ch/record/98509.

[7] Gardner Martin. "The fantastic combinations of John Conway's new solitaire game "life" by Martin Gardner." In: *Scientific American* 223 (1970), pp. 120–123.

[8] Stephen Wolfram. "Statistical mechanics of cellular automata." In: *Reviews of Modern Physics* 55.3 (July 1983), pp. 601–644. DOI: 10.1103/RevModPhys.55.601.

[9] Gilles Louppe, Joeri Hermans, and Kyle Cranmer. "Adversarial Variational Optimization of Non-Differentiable Simulators." In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Ed. by Kamalika Chaudhuri and Masashi Sugiyama. Vol. 89. Proceedings of Machine Learning Research. PMLR, 2019, pp. 1438–1447. URL: https://proceedings.mlr.press/v89/louppe19a.html.

[10] Joeri Hermans, Volodimir Begy, and Gilles Louppe. "Likelihood-free MCMC with Amortized Approximate Ratio Estimators." In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 4239–4248. URL: https://proceedings.mlr.press/v119/hermans20a.html.

[11] Joeri Hermans et al. "Towards constraining warm dark matter with stellar streams through neural simulation-based inference." In: *Monthly Notices of the Royal Astronomical Society* 507.2 (Aug. 2021), pp. 1999–2011. ISSN: 0035-8711. DOI: 10.1093/mnras/stab2181. eprint: https://academic.oup.com/mnras/article-pdf/507/2/1999/40078147/stab2181.pdf. URL: https://doi.org/10.1093/mnras/stab2181.

[12] Joeri Hermans et al. "Averting A Crisis In Simulation-Based Inference." In: *arXiv e-prints*, arXiv:2110.06581 (Oct. 2021). arXiv: 2110.06581 [stat.ML].

[13] Joeri Hermans and Gilles Louppe. "Gradient Energy Matching for Distributed Asynchronous Gradient Descent." In: *arXiv e-prints*, arXiv:1805.08469 (May 2018). arXiv: 1805.08469 [cs.LG].

[14] Volodimir Begy et al. "Simulating Data Access Profiles of Computational Jobs in Data Grids." In: *2019 15th International Conference on eScience (eScience)*. 2019, pp. 394–402. DOI: 10.1109/eScience.2019.00051.

[15] Johann Brehmer et al. "Mining for Dark Matter Substructure: Inferring Subhalo Population Properties from Strong Lenses with Machine Learning." In: *The Astrophysical Journal* 886.1 (2019), p. 49. DOI: 10.3847/1538-4357/ab4c41. URL: https://doi.org/10.3847/1538-4357/ab4c41.

[16] Diederik P Kingma and Max Welling. "Auto-Encoding Variational Bayes." In: *arXiv e-prints*, arXiv:1312.6114 (Dec. 2013), arXiv:1312.6114. arXiv: 1312.6114 [stat.ML].

[17] Ian J. Goodfellow et al. "Generative Adversarial Networks." In: *arXiv e-prints*, arXiv:1406.2661 (June 2014), arXiv:1406.2661. arXiv: 1406.2661 [stat.ML].

[18] Kyle Cranmer, Juan Pavez, and Gilles Louppe. "Approximating Likelihood Ratios with Calibrated Discriminative Classifiers." In: *arXiv e-prints*, arXiv:1506.02169 (June 2015), arXiv:1506.02169. arXiv: 1506.02169 [stat.AP].

[19] S. Mohamed and B. Lakshminarayanan. "Learning in Implicit Generative Models." In: *ArXiv e-prints* (Oct. 2016). arXiv: 1610.03483 [stat.ML].

[20] Johann Brehmer et al. "MadMiner: Machine learning-based inference for particle physics." In: *arXiv e-prints*, arXiv:1907.10621 (July 2019), arXiv:1907.10621. arXiv: 1907.10621 [hep-ph].

[21] Johann Brehmer. "Simulation-based inference in particle physics." In: *Nature Reviews Physics* 3.5 (Jan. 2021), pp. 305–305. DOI: 10.1038/s42254-021-00305-6. arXiv: 2010.06439 [hep-ph].

[22] Johann Brehmer, Kyle Cranmer, and F. Kling. "Improving inference with matrix elements and machine learning." In: *International Journal of Modern Physics A* 35, 2041008 (June 2020), p. 2041008. DOI: 10.1142/S0217751X20410080. arXiv: 1906.01578 [hep-ph].

[23] Johann Brehmer et al. "Constraining effective field theories with machine learning." In: *European Physical Journal Web of Conferences*. Vol. 245. European Physical Journal Web of Conferences. Nov. 2020, 06026, p. 06026. DOI: 10.1051/epjconf/202024506026.

[24] Kyle Cranmer et al. "Reframing Jet Physics with New Computational Methods." In: *arXiv e-prints*, arXiv:2105.10512 (May 2021), arXiv:2105.10512. arXiv: 2105.10512 [hep-ph].

[25] Tim Salimans et al. "Improved Techniques for Training GANs." In: *arXiv e-prints*, arXiv:1606.03498 (June 2016), arXiv:1606.03498. arXiv: 1606.03498 [cs.LG].

[26] Luke Metz et al. "Unrolled Generative Adversarial Networks." In: *arXiv e-prints*, arXiv:1611.02163 (Nov. 2016), arXiv:1611.02163. arXiv: 1611.02163 [cs.LG].

[27] Martin Arjovsky, Soumith Chintala, and Léon Bottou. "Wasserstein GAN." In: *arXiv e-prints*, arXiv:1701.07875 (Jan. 2017), arXiv:1701.07875. arXiv: 1701.07875 [stat.ML].

[28] Ishaan Gulrajani et al. "Improved Training of Wasserstein GANs." In: *arXiv e-prints*, arXiv:1704.00028 (Mar. 2017), arXiv:1704.00028. arXiv: 1704.00028 [cs.LG].

[29] Kevin Roth et al. "Stabilizing Training of Generative Adversarial Networks through Regularization." In: *arXiv e-prints*, arXiv:1705.09367 (May 2017), arXiv:1705.09367. arXiv: 1705.09367 [cs.LG].

[30] Martin Arjovsky and Léon Bottou. "Towards Principled Methods for Training Generative Adversarial Networks." In: *arXiv e-prints*, arXiv:1701.04862 (Jan. 2017), arXiv:1701.04862. arXiv: 1701.04862 [stat.ML].

[31] Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. "The numerics of gans." In: *arXiv preprint arXiv:1705.10461* (2017).

[32] Joe Staines and David Barber. "Variational Optimization." In: *arXiv e-prints*, arXiv:1212.4507 (Dec. 2012), arXiv:1212.4507. arXiv: 1212.4507 [stat.ML].

[33] J Staines and D Barber. "Optimization by variational bounding." In: *ESANN 2013 proceedings, 21st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. 2013, pp. 473–478.

[34] D. Wierstra et al. "Natural Evolution Strategies." In: *ArXiv e-prints* (June 2011). arXiv: 1106.4487 [stat.ML].

[35] Ronald J Williams. "Simple statistical gradient-following algorithms for connectionist reinforcement learning." In: *Machine learning* 8.3-4 (1992), pp. 229–256.

[36] Markus Giftthaler et al. "Automatic differentiation of rigid body dynamics for optimal control and estimation." In: *Advanced Robotics* 31.22 (2017), pp. 1225–1237.

[37] Yuanming Hu et al. "Difftaichi: Differentiable programming for physical simulation." In: *arXiv preprint arXiv:1910.00935* (2019).

[38] L Heinrich et al. "Differentiable Simulators for HEP." In: ().

[39] Eric Heiden et al. "NeuralSim: Augmenting differentiable simulators with neural networks." In: *arXiv preprint arXiv:2011.04217* (2020).

[40] Atılım Günes Baydin et al. "Differentiable Programming in High-Energy Physics." In: *Submitted as a Snowmass LOI* (2020).

[41] Marco Chianese et al. "Differentiable strong lensing: uniting gravity and neural nets through differentiable probabilistic programming." In: *Monthly Notices of the Royal Astronomical Society* 496.1 (2020), pp. 381–393.

[42] Filipe de Avila Belbute-Peres et al. "End-to-end differentiable physics for learning and control." In: *Advances in neural information processing systems* 31 (2018), pp. 7178–7189.

[43] Yi-Ling Qiao et al. "Scalable differentiable physics for learning and control." In: *arXiv preprint arXiv:2007.02168* (2020).

[44] Matthew M Loper and Michael J Black. "OpenDR: An approximate differentiable renderer." In: *European Conference on Computer Vision*. Springer. 2014, pp. 154–169.

[45] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. "Neural 3D Mesh Renderer. CoRR abs/1711.07566 (2017)." In: *arXiv preprint arXiv:1711.07566* (2017).

[46] Tzu-Mao Li et al. "Differentiable Monte Carlo Ray Tracing through Edge Sampling." In: *ACM Trans. Graph. (Proc. SIGGRAPH Asia)* 37.6 (2018), 222:1–222:11.

[47] Merlin Nimier-David et al. "Mitsuba 2: A Retargetable Forward and Inverse Renderer." In: *Transactions on Graphics (Proceedings of SIGGRAPH Asia)* 38.6 (Dec. 2019). DOI: 10.1145/3355089.3356498.

[48] Cheng Zhang et al. "A Differential Theory of Radiative Transfer." In: *ACM Trans. Graph.* 38.6 (2019), 227:1–227:16.

[49] Rajesh Ranganath et al. "Operator Variational Inference." In: *arXiv e-prints*, arXiv:1610.09033 (Oct. 2016), arXiv:1610.09033. arXiv: 1610.09033 [stat.ML].

[50] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. "Learning representations by back-propagating errors." In: *nature* 323.6088 (1986), pp. 533–536.

[51] Ning Qian. "On the momentum term in gradient descent learning algorithms." In: *Neural networks* 12.1 (1999), pp. 145–151.

[52] Torbjörn Sjöstrand, Stephen Mrenna, and Peter Skands. "A brief introduction to PYTHIA 8.1." In: *Computer Physics Communications* 178.11 (June 2008), pp. 852–867. DOI: 10.1016/j.cpc.2008.01.036. arXiv: 0710.3820 [hep-ph].

[53] O. Crosby. *PYTHIA, pythia*. Apr. 2020. URL: https://doi.org/10.1002/9783527809080.cataz13886.

[54] Michael U Gutmann and Aapo Hyvärinen. "Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics." In: *Journal of Machine Learning Research* 13.Feb (2012), pp. 307–361.

[55] Kyle Cranmer, Juan Pavez, and Gilles Louppe. "Approximating likelihood ratios with calibrated discriminative classifiers." In: *arXiv preprint arXiv:1506.02169* (2015).

[56] K Cranmer et al. "Experiments using machine learning to approximate likelihood ratios for mixture models." In: *Journal of Physics: Conference Series*. Vol. 762. 1. IOP Publishing. 2016, p. 012034.

[57] Ritabrata Dutta et al. "Likelihood-free inference by ratio estimation." In: *arXiv preprint arXiv:1611.10242* (2016).

[58] Michael U Gutmann et al. "Likelihood-free inference via classification." In: *Statistics and Computing* 28.2 (2018), pp. 411–425.

[59] Mihaela Rosca et al. "Variational approaches for auto-encoding generative adversarial networks." In: *arXiv preprint arXiv:1706.04987* (2017).

[60] Yaroslav Ganin et al. "Synthesizing programs for images using reinforced adversarial learning." In: *International Conference on Machine Learning*. PMLR. 2018, pp. 1666–1675.

[61] Vincent Dumoulin et al. "Adversarially learned inference." In: *arXiv preprint arXiv:1606.00704* (2016).

[62] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. "Adversarial feature learning." In: *arXiv preprint arXiv:1605.09782* (2016).

[63] Sebastian Lutz, Konstantinos Amplianitis, and Aljosa Smolic. "Alphagan: Generative adversarial networks for natural image matting." In: *arXiv preprint arXiv:1807.10088* (2018).

[64] Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. "Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks." In: *arXiv preprint arXiv:1701.04722* (2017).

[65] Ferenc Huszár. "Variational inference using implicit distributions." In: *arXiv preprint arXiv:1702.08235* (2017).

[66] Atilim Güneş Baydin et al. "Etalumis: Bringing probabilistic programming to scientific simulators at scale." In: *Proceedings of the international conference for high performance computing, networking, storage and analysis*. 2019, pp. 1–24.

[67] Minh-Ngoc Tran, David J Nott, and Robert Kohn. "Variational Bayes with intractable likelihood." In: *Journal of Computational and Graphical Statistics* 26.4 (2017), pp. 873–882.

[68] Dustin Tran, Rajesh Ranganath, and David Blei. "Hierarchical implicit models and likelihood-free variational inference." In: *Advances in Neural Information Processing Systems*. 2017, pp. 5523–5533.

[69] Adam McCarthy, Blanca Rodriguez, and Ana Minchole. "Variational inference over non-differentiable cardiac simulators using bayesian optimization." In: *arXiv preprint arXiv:1712.03353* (2017).

[70] Nicholas Metropolis et al. "Equation of State Calculations by Fast Computing Machines." In: *The Journal of Chemical Physics* 21.6 (June 1953), pp. 1087–1092. ISSN: 0021-9606, 1089-7690. URL: https://doi.org/10.1063/1.1699114.

[71] W.K. Hastings. "Monte Carlo sampling methods using Markov chains and their applications." In: *Biometrika* 57.1 (Apr. 1970), pp. 97–109. ISSN: 1464-3510, 0006-3444. URL: https://doi.org/10.1093/biomet/57.1.97.

[72] Simon Duane et al. "Hybrid Monte Carlo." In: *Phys. Lett. B* 195.2 (Sept. 1987), pp. 216–222. ISSN: 0370-2693. URL: https://doi.org/10.1016/0370-2693(87)91197-x.

[73] Radford M Neal. "MCMC using Hamiltonian dynamics." In: *Handbook of Markov Chain Monte Carlo* 2.11 (2011), p. 2. URL: https://arxiv.org/abs/1206.1901.

[74] David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[75] Daniel Foreman-Mackey et al. "emcee: The MCMC Hammer." In: *Publ. Astron. Soc. Pac.* 125.925 (Mar. 2013), pp. 306–312. ISSN: 0004-6280, 1538-3873. URL: https://doi.org/10.1086/670067.

[76] Michael Betancourt. "A conceptual introduction to Hamiltonian Monte Carlo." In: *arXiv preprint arXiv:1701.02434* (2017). URL: https://arxiv.org/abs/1701.02434.

[77] Matthew D Hoffman, Andrew Gelman, et al. "The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo." In: *J. Mach. Learn. Res.* 15.1 (2014), pp. 1593–1623.

[78] E.S. Pearson J. Neyman. "IX. On the problem of the most efficient tests of statistical hypotheses." In: *Phil. Trans. R. Soc. Lond. A* 231.694-706 (Feb. 1933), pp. 289–337. ISSN: 0264-3952, 2053-9258. URL: https://doi.org/10.1098/rsta.1933.0009.

[79] Michael U. Gutmann et al. "Likelihood-free inference via classification." In: *Stat Comput* 28.2 (Mar. 2017), pp. 411–425. ISSN: 0960-3174, 1573-1375. URL: https://doi.org/10.1007/s11222-017-9738-6.

[80] Johann Brehmer et al. "Mining gold from implicit models to improve likelihood-free inference." In: *Proc Natl Acad Sci USA* 117.10 (Feb. 2020), pp. 5242–5249. ISSN: 0027-8424, 1091-6490. URL: https://doi.org/10.1073/pnas.1915980117.

[81] Ian Goodfellow et al. "Generative adversarial nets." In: *Advances in neural information processing systems*. 2014, pp. 2672–2680.

[82] Masatoshi Uehara et al. "Generative adversarial nets from a density ratio estimation perspective." In: *arXiv preprint arXiv:1610.02920* (2016).

[83] Ryan Turner et al. "Metropolis-Hastings Generative Adversarial Networks." In: *arXiv preprint arXiv:1811.11357* (2018).

[84] Samaneh Azadi et al. "Discriminator rejection sampling." In: *arXiv preprint arXiv:1810.06758* (2018).

[85] Dustin Tran, Rajesh Ranganath, and David M Blei. "Deep and hierarchical implicit models." In: *arXiv preprint arXiv:1702.08896* 7 (2017).

[86] Pierre Baldi et al. "Parameterized neural networks for high-energy physics." In: *Eur. Phys. J. C* 76.5 (Apr. 2016), p. 235. ISSN: 1434-6044, 1434-6052. arXiv: 1601.07913 [hep-ex]. URL: https://doi.org/10.1140/epjc/s10052-016-4099-4.

[87] Hongyu Ren et al. "Adversarial Constraint Learning for Structured Prediction." In: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, July 2018, pp. 5140–5150. URL: https://doi.org/10.24963/ijcai.2018/366.

[88] Simon Tavaré et al. "Inferring coalescence times from DNA sequence data." In: *Genetics* 145.2 (1997), pp. 505–518. URL: http://www.genetics.org/content/genetics/145/2/505.full.pdf.

[89] J.K. Pritchard et al. "Population growth of human Y chromosomes: A study of Y chromosome microsatellites." In: *Mol. Biol. Evol.* 16.12 (Dec. 1999), pp. 1791–1798. ISSN: 0737-4038, 1537-1719. URL: https://doi.org/10.1093/oxfordjournals.molbev.a026091.

[90] Mark A Beaumont, Wenyang Zhang, and David J Balding. "Approximate Bayesian computation in population genetics." In: *Genetics* 162.4 (2002), pp. 2025–2035. URL: http://www.genetics.org/content/162/4/2025.

[91] Jean-Michel Marin et al. "Approximate Bayesian computational methods." In: *Stat Comput* 22.6 (Oct. 2011), pp. 1167–1180. ISSN: 0960-3174, 1573-1375. URL: https://doi.org/10.1007/s11222-011-9288-2.

[92] Tina Toni et al. "Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems." In: *J. R. Soc. Interface.* 6.31 (July 2008), pp. 187–202. ISSN: 1742-5689, 1742-5662. URL: https://doi.org/10.1098/rsif.2008.0172.

[93] P. Marjoram et al. "Markov chain Monte Carlo without likelihoods." In: *Proceedings of the National Academy of Sciences* 100.26 (Dec. 2003), pp. 15324–15328. ISSN: 0027-8424, 1091-6490. eprint: https://www.pnas.org/content/100/26/15324.full.pdf. URL: https://doi.org/10.1073/pnas.0306899100.

[94] Daniel Wegmann, Christoph Leuenberger, and Laurent Excoffier. "Efficient Approximate Bayesian Computation Coupled With Markov Chain Monte Carlo Without Likelihood." In: *Genetics* 182.4 (June 2009), pp. 1207–1218. ISSN: 0016-6731, 1943-2631. URL: https://doi.org/10.1534/genetics.109.102509.

[95] Paul Fearnhead and Dennis Prangle. "Constructing summary statistics for approximate Bayesian computation: Semi-automatic approximate Bayesian computation. Semi-automatic Approximate Bayesian Computation." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74.3 (May 2012), pp. 419–474. ISSN: 1369-7412. URL: https://doi.org/10.1111/j.1467-9868.2011.01010.x.

[96] Traiko Dinev and Michael U Gutmann. "Dynamic likelihood-free inference via ratio estimation (dire)." In: *arXiv preprint arXiv:1810.09899* (2018).

[97] Wing Wong et al. "Learning Summary Statistic for Approximate Bayesian Computation via Deep Neural Network." In: *STAT SINICA* (2018), pp. 1595–1618. ISSN: 1017-0405. URL: https://doi.org/10.5705/ss.202015.0340.

[98] Radford M. Neal and Geoffrey E. Hinton. "A View of the Em Algorithm that Justifies Incremental, Sparse, and other Variants." In: *Learning in Graphical Models*. Springer Netherlands, 1998, pp. 355–368. URL: https://doi.org/10.1007/978-94-011-5014-9_12.

[99] Matthew D Hoffman et al. "Stochastic variational inference." In: *The Journal of Machine Learning Research* 14.1 (2013), pp. 1303–1347.

[100] Tim Salimans, Diederik Kingma, and Max Welling. "Markov chain monte carlo and variational inference: Bridging the gap." In: *International Conference on Machine Learning*. 2015, pp. 1218–1226.

[101] Samuel Gershman and Noah Goodman. "Amortized inference in probabilistic reasoning." In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 36. 36. 2014.

[102] Daniel Ritchie, Paul Horsfall, and Noah D Goodman. "Deep amortized inference for probabilistic programs." In: *arXiv preprint arXiv:1610.05735* (2016).

[103] Ronald J. Williams. "Simple statistical gradient-following algorithms for connectionist reinforcement learning." In: *Mach Learn* 8.3-4 (May 1992), pp. 229–256. ISSN: 0885-6125, 1573-0565. URL: https://doi.org/10.1007/bf00992696.

[104] Richard S Sutton et al. "Policy gradient methods for reinforcement learning with function approximation." In: *Advances in neural information processing systems*. 2000, pp. 1057–1063.

[105] Arthur Pesah, Antoine Wehenkel, and Gilles Louppe. "Recurrent machines for likelihood-free inference." In: *arXiv preprint arXiv:1811.12932* (2018).

[106] George Papamakarios and Iain Murray. "Fast ε-free inference of simulation models with Bayesian conditional density estimation." In: *Advances in Neural Information Processing Systems*. 2016, pp. 1028–1036.

[107] Jan Boelts et al. "Comparing neural simulations by neural density estimation." In: *2019 Conference on Cognitive Computational Neuroscience*. Cognitive Computational Neuroscience, 2019, pp. 1289–1299. URL: https://doi.org/10.32470/ccn.2019.1291-0.

[108] David Greenberg, Marcel Nonnenmacher, and Jakob Macke. "Automatic Posterior Transformation for Likelihood-Free Inference." In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. Long Beach, California, USA: PMLR, 2019, pp. 2404–2414. URL: http://proceedings.mlr.press/v97/greenberg19a.html.

[109] George Papamakarios and Iain Murray. "Sequential Neural Likelihood: Fast Likelihood-free Inference with Autoregressive Flows." In: *arXiv preprint arXiv:1805.07226* (2018). URL: https://arxiv.org/abs/1805.07226.

[110] Kyle Cranmer, Juan Pavez, and Gilles Louppe. "Approximating likelihood ratios with calibrated discriminative classifiers." In: *arXiv preprint arXiv:1506.02169* (2015). URL: https://arxiv.org/abs/1506.02169.

[111] Michael U Gutmann and Jukka Corander. "Bayesian optimization for likelihood-free inference of simulator-based statistical models." In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 4256–4302. URL: http://jmlr.org/papers/v17/15-017.html.

[112] Victor M.H. Ong et al. "Variational Bayes with synthetic likelihood." In: *Stat Comput* 28.4 (Aug. 2017), pp. 971–988. ISSN: 0960-3174, 1573-1375. URL: https://doi.org/10.1007/s11222-017-9773-3.

[113] Edward Meeds and Max Welling. "GPS-ABC: Gaussian Process Surrogate Approximate Bayesian Computation." In: *arXiv preprint arXiv:1401.2838* (2014). URL: https://arxiv.org/abs/1401.2838.

[114] Jan-Matthis Lueckmann et al. "Likelihood-free inference with emulator networks." In: *arXiv preprint arXiv:1805.09294* (2018). URL: https://arxiv.org/abs/1805.09294.

[115] Johann Brehmer et al. "A guide to constraining effective field theories with machine learning." In: *Phys. Rev. D* 98.5 (Sept. 2018), p. 052004. ISSN: 2470-0010, 2470-0029. URL: https://doi.org/10.1103/physrevd.98.052004.

[116] David Ha and Jürgen Schmidhuber. "World Models." In: *arXiv preprint arXiv:1803.10122* (2018).

[117] Donald B Rubin. "Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician." In: *The Annals of Statistics* (1984), pp. 1151–1172.

[118] P. Skands, S. Carrazza, and J. Rojo. "Tuning PYTHIA 8.1: The Monash 2013 tune." In: *Eur. Phys. J. C* 74.8 (Aug. 2014), p. 3024. ISSN: 1434-6044, 1434-6052. URL: https://doi.org/10.1140/epjc/s10052-014-3024-y.

[119] Alfred J. Lotka. "Analytical Note on Certain Rhythmic Relations in Organic Systems." In: *Proc Natl Acad Sci USA* 6.7 (June 1920), pp. 410–415. ISSN: 0027-8424, 1091-6490. URL: https://doi.org/10.1073/pnas.6.7.410.

[120] Arthur Gretton et al. "A kernel two-sample test." In: *Journal of Machine Learning Research* 13.Mar (2012), pp. 723–773.

[121] J W Nightingale, S Dye, and Richard J Massey. "AutoLens: Automated modeling of a strong lens's light, mass, and source." In: *Mon. Not. R. Astron. Soc.* 478.4 (May 2018), pp. 4738–4784. ISSN: 0035-8711, 1365-2966. URL: https://doi.org/10.1093/mnras/sty1264.

[122] Kaiming He et al. "Deep Residual Learning for Image Recognition." In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2016, pp. 770–778. URL: https://doi.org/10.1109/cvpr.2016.90.

[123] Robert Kormann, Peter Schneider, and Matthias Bartelmann. "Isothermal elliptical gravitational lens models." In: *Astron. Astrophys.* 284 (1994), pp. 285–299.

[124] Conor Durkan, Iain Murray, and George Papamakarios. "On Contrastive Learning for Likelihood-free Inference." In: *arXiv e-prints*, arXiv:2002.03712 (Feb. 2020), arXiv:2002.03712. arXiv: 2002.03712 [stat.ML].

[125] P. J. E. Peebles. "Large-scale background temperature and mass fluctuations due to scale-invariant primeval perturbations." In: *ApJ* 263 (Dec. 1982), pp. L1–L5. DOI: 10.1086/183911.

[126] G. R. Blumenthal et al. "Formation of galaxies and large-scale structure with cold dark matter." In: *Nature* 311 (Oct. 1984), pp. 517–525. DOI: 10.1038/311517a0.

[127] Ben Moore et al. "Dark matter substructure within galactic halos." In: *ApJ* 524.1 (Dec. 1999), pp. 9–20. DOI: 10.1086/323695. arXiv: astro-ph/0108224 [astro-ph]. URL: https://ui.adsabs.harvard.edu/abs/2001ApJ...563....9M.

[128] V. Avila-Reese, C. Firmani, and X. Hernández. "On Formation and Evolution of Disk Galaxies: Cosmological Initial Conditions and the Gravitational Collapse." In: *Astrophys.J.505:37,1998* 505.1 (Oct. 19, 1997), pp. 37–49. DOI: 10.1086/306136. arXiv: astro-ph/9710201 [astro-ph]. URL: https://ui.adsabs.harvard.edu/abs/1998ApJ...505...37A.

[129] DH Zhao et al. "The growth and structure of dark matter haloes." In: *MNRAS* 339.1 (Feb. 2003), pp. 12–24. DOI: 10.1046/j.1365-8711.2003.06135.x. arXiv: astro-ph/0204108 [astro-ph]. URL: https://ui.adsabs.harvard.edu/abs/2003MNRAS.339...12Z.

[130] Stefan Hofmann, Dominik J Schwarz, and Horst Stoecker. "Damping scales of neutralino cold dark matter." In: *Phys. Rev. D* 64.8 (Oct. 2001), p. 083507. DOI: 10.1103/PhysRevD.64.083507. arXiv: astro-ph/0104173 [astro-ph]. URL: https://ui.adsabs.harvard.edu/abs/2001PhRvD..64h3507H.

[131] Aurel Schneider, Robert E Smith, and Darren Reed. "Halo mass function and the free streaming scale." In: *MNRAS* 433.2 (Aug. 2013), pp. 1573–1587. DOI: 10.1093/mnras/stt829. arXiv: 1303.0839 [astro-ph.CO]. URL: https://ui.adsabs.harvard.edu/abs/2013MNRAS.433.1573S.

[132] Edmund Bertschinger. "Effects of cold dark matter decoupling and pair annihilation on cosmological perturbations." In: *Phys. Rev. D* 74.6 (2006), p. 063509. DOI: 10.1103/PhysRevD.74.063509.

[133] J. R. Bond and A. S. Szalay. "The collisionless damping of density fluctuations in an expanding universe." In: *ApJ* 274 (Nov. 1983), pp. 443–468. DOI: 10.1086/161460.

[134] Scott Dodelson and Lawrence M Widrow. "Sterile neutrinos as dark matter." In: *Phys. Rev. Lett.* 72.1 (Jan. 1994), pp. 17–20. DOI: 10.1103/PhysRevLett.72.17. arXiv: hep-ph/9303287

[hep-ph]. URL: https://ui.adsabs.harvard.edu/abs/1994PhRvL..72...17D.

[135] Paul Bode, Jeremiah P. Ostriker, and Neil Turok. "Halo Formation in Warm Dark Matter Models." In: *ApJ* 556.1 (July 2001), pp. 93–107. DOI: 10.1086/321541. arXiv: astro-ph/0010389 [astro-ph].

[136] Robert E Smith and Katarina Markovic. "Testing the Warm Dark Matter paradigm with large-scale structures." In: *Phys. Rev. D* 84.6, 063507 (Sept. 2011), p. 063507. DOI: 10.1103/PhysRevD.84.063507. arXiv: 1103.2134 [astro-ph.CO]. URL: https://ui.adsabs.harvard.edu/abs/2011PhRvD..84f3507S.

[137] RA Ibata et al. "Uncovering cold dark matter halo substructure with tidal streams." In: *MNRAS* 332.4 (June 2002), pp. 915–920. DOI: 10.1046/j.1365-8711.2002.05358.x. arXiv: astro-ph/0110690 [astro-ph]. URL: https://ui.adsabs.harvard.edu/abs/2002MNRAS.332..915I.

[138] Kathryn V Johnston, David N Spergel, and Christian Haydn. "How lumpy is the Milky Way's dark matter halo?" In: *ApJ* 570.2 (2002), pp. 656–664. ISSN: 0004-637X. DOI: 10.1086/339791.

[139] Joo Heon Yoon, Kathryn V Johnston, and David W Hogg. "Clumpy streams from clumpy halos: detecting missing satellites with cold stellar structures." In: *ApJ* 731.1, 58 (Apr. 2011), p. 58. DOI: 10.1088/0004-637X/731/1/58. arXiv: 1012.2884 [astro-ph.GA]. URL: https://ui.adsabs.harvard.edu/abs/2011ApJ...731...58Y.

[140] Raymond G Carlberg. "Dark matter sub-halo counts via star stream crossings." In: *ApJ* 748.1, 20 (Mar. 2012), p. 20. DOI: 10.1088/0004-637X/748/1/20. arXiv: 1109.6022 [astro-ph.CO]. URL: https://ui.adsabs.harvard.edu/abs/2012ApJ...748...20C.

[141] Denis Erkal and Vasily Belokurov. "Forensics of subhalo–stream encounters: the three phases of gap growth." In: *MNRAS* 450.1 (2015), pp. 1136–1149. ISSN: 0035-8711. DOI: 10.1093/mnras/stv655.

[142] Denis Erkal and Vasily Belokurov. "Properties of dark subhaloes from gaps in tidal streams." In: *MNRAS* 454.4 (2015), pp. 3542–3558. ISSN: 0035-8711. DOI: 10.1093/mnras/stv2122.

[143] Jo Bovy, Denis Erkal, and Jason L. Sanders. "Linear perturbation theory for tidal streams and the small-scale CDM power spectrum." In: *MNRAS* 466.1 (2017), pp. 628–668. ISSN: 0035-8711. DOI: 10.1093/mnras/stw3067. arXiv: 1606.03470 [astro-ph.GA].

[144] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. "The frontier of simulation-based inference." In: *Proceedings of the National Academy of Sciences* (2020).

[145] Nilanjan Banik et al. "Probing the nature of dark matter particles with stellar streams." In: *JCAP07(2018)061* 7.07, 061 (Apr. 12, 2018), p. 061. DOI: 10.1088/1475-7516/2018/07/061. arXiv: http://arxiv.org/abs/1804.04384v2 [astro-ph.CO]. URL: http://adsabs.harvard.edu/abs/2018JCAP...07..061B.

[146] Nilanjan Banik et al. "Novel constraints on the particle nature of dark matter from stellar streams." In: *arXiv e-prints*, arXiv:1911.02663 (Nov. 2019), arXiv:1911.02663. arXiv: 1911.02663 [astro-ph.GA].

[147] In: ().

[148] J. Bovy. "galpy: A python Library for Galactic Dynamics." In: *The Astrophysical Journal Supplement Series* 216.2, 29 (Feb. 2015), p. 29. DOI: 10.1088/0067-0049/216/2/29. arXiv: 1412.3451. URL: http://adsabs.harvard.edu/abs/2015ApJS..216...29B.

[149] Nicola C. Amorisco et al. "Gaps in globular cluster streams: giant molecular clouds can cause them too." In: *MNRAS* 463.1 (June 8, 2016), pp. L17–L21. ISSN: 1745-3925. DOI: 10.1093/mnrasl/slw148. arXiv: 1606.02715v2 [astro-ph.GA].

[150] Denis Erkal, Sergey E. Koposov, and Vasily Belokurov. "A sharper view of Pal 5's tails: Discovery of stream perturbations with a novel non-parametric technique." In: *MNRAS* 470.1 (Sept. 5, 2016), pp. 60–84. DOI: 10.1093/mnras/stx1208. arXiv: 1609.01282v2 [astro-ph.GA].

[151] S. Pearson, A. M. Price-Whelan, and K. V. Johnston. "Gaps and length asymmetry in the stellar stream Palomar 5 as effects of Galactic bar rotation." In: *Nature Astronomy* 1 (Mar. 14, 2017), pp. 633–639. DOI: 10.1038/s41550-017-0220-3. arXiv: 1703.04627v2 [astro-ph.GA]. URL: http://adsabs.harvard.edu/abs/2017NatAs...1..633P.

[152] Nilanjan Banik and Jo Bovy. "Effects of baryonic and dark matter substructure on the Pal 5 stream." In: *MNRAS* ().

[153] C. J. Grillmair and O. Dionatos. "Detection of a 63 deg Cold Stellar Stream in the Sloan Digital Sky Survey." In: *ApJ* 643.1 (2006), pp. L17–L20. DOI: 10.1086/505111. eprint: astro-ph/0604332. URL: http://adsabs.harvard.edu/abs/2006ApJ...643L..17G.

[154] Jeremy J Webb and Jo Bovy. "Searching for the GD-1 stream progenitor inGaiaDR2 with directN-body simulations." In: *MNRAS* 485.4 (2019), 5929–5938. ISSN: 1365-2966. DOI: 10.1093/mnras/stz867. URL: http://dx.doi.org/10.1093/mnras/stz867.

[155] Nilanjan Banik et al. "Evidence of a population of dark subhalos from Gaia and Pan-STARRS observations of the GD-1 stream." In: *arXiv e-prints*, arXiv:1911.02662 (Nov. 2019), arXiv:1911.02662. arXiv: 1911.02662 [astro-ph.GA].

[156] Mark R. Lovell et al. "The properties of warm dark matter haloes." In: *MNRAS* 439.1 (Aug. 6, 2013), pp. 300–317. DOI: 10.1093/mnras/stt2431. arXiv: 1308.1399v2 [astro-ph.CO].

[157] Volker Springel et al. "The Aquarius Project: the subhaloes of galactic haloes." In: *MNRAS* 391.4 (Dec. 2008), pp. 1685–1711. DOI: 10.1111/j.1365-2966.2008.14066.x. arXiv: 0809.0898. URL: http://adsabs.harvard.edu/abs/2008MNRAS.391.1685S.

[158] Elena D'Onghia et al. "Substructure depletion in the Milky Way halo by the disk." In: *ApJ* 709.2 (2010), p. 1138. ISSN: 0004-637X. DOI: 10.1088/0004-637x/709/2/1138.

[159] Till Sawala et al. "Shaken and Stirred: The Milky Way's Dark Substructures." In: *MNRAS* 467.4 (Sept. 6, 2016), pp. 4383–4400. ISSN: 0035-8711. DOI: 10.1093/mnras/stx360. arXiv: 1609.01718v1 [astro-ph.GA].

[160] Shea Garrison-Kimmel et al. "Not so lumpy after all: modelling the depletion of dark matter subhaloes by Milky Way-like galaxies." In: *MNRAS* 471.2 (2017), pp. 1709–1727. DOI: 10.1093/mnras/stx1710. arXiv: 1701.03792 [astro-ph.GA].

[161] Tyler Kelley et al. "Phat ELVIS: The inevitable effect of the Milky Way's disc on its dark matter subhaloes." In: *MNRAS* 487.3 (Aug. 2019), pp. 4409–4423. DOI: 10.1093/mnras/stz1553. arXiv: 1811.12413 [astro-ph.GA].

[162] Jeremy J. Webb and Jo Bovy. "And In The Darkness Unbind Them: High-Resolution Simulations of Dark Matter Subhalo Disruption in a Milky Way-like Tidal Field." In: *arXiv e-prints*, arXiv:2006.06695 (June 2020), arXiv:2006.06695. arXiv: 2006.06695 [astro-ph.GA].

[163] T. J. L. de Boer, D. Erkal, and M. Gieles. "A closer look at the spur, blob, wiggle, and gaps in GD-1." In: *MNRAS* 494.4 (Nov. 13, 2019), pp. 5315–5332. DOI: 10.1093/mnras/staa917. arXiv: 1911.05745v1 [astro-ph.GA]. URL: https://ui.adsabs.harvard.edu/abs/2020MNRAS.494.5315D.

[164] J. Neyman and Elizabeth L. Scott. "Consistent Estimates Based on Partially Consistent Observations." In: *Econometrica* 16.1 (1948), pp. 1–32. ISSN: 00129682, 14680262. URL: http://www.jstor.org/stable/1914288.

[165] James O Berger, Brunero Liseo, Robert L Wolpert, et al. "Integrated likelihood methods for eliminating nuisance parameters." In: *Statistical science* 14.1 (1999), pp. 1–28.

[166] Jo Bovy. "Constraining the Small-Scale Clustering of Dark Matter with Stellar Streams." In: *Illuminating Dark Matter*. Ed. by Rouven Essig, Jonathan Feng, and Kathryn Zurek. Vol. 56. Jan. 2019, pp. 9–18.

[167] Daniel Gilman et al. "Warm dark matter chills out: constraints on the halo mass function and the free-streaming length of dark matter with eight quadruple-image strong gravitational lenses." In: *MNRAS* 491.4 (Feb. 2020), pp. 6077–6101. DOI: 10.1093/mnras/stz3480. arXiv: 1908.06983 [astro-ph.CO]. URL: https://ui.adsabs.harvard.edu/abs/2020MNRAS.491.6077G.

[168] Johann Brehmer et al. "Constraining effective field theories with machine learning." In: *Phys. Rev. Lett.* 121.11, 111801 (Sept. 2018), p. 111801. DOI: 10.1103/PhysRevLett.121.111801. arXiv: 1805.00013 [hep-ph]. URL: https://ui.adsabs.harvard.edu/abs/2018PhRvL.121k1801B.

[169] Johann Brehmer et al. "Mining gold from implicit models to improve likelihood-free inference." In: *Proceedings of the National Academy of Sciences* 117.10 (2020), pp. 5242–5249.

[170] Glen Cowan et al. "Asymptotic formulae for likelihood-based tests of new physics." In: *Eur.Phys.J.C71:1554,2011* 71.2 (July 10, 2010). DOI: 10.1140/epjc/s10052-011-1554-0. arXiv: 1007.1727 [physics.data-an].

[171] Samuel S Wilks. "The large-sample distribution of the likelihood ratio for testing composite hypotheses." In: *The annals of mathematical statistics* 9.1 (1938), pp. 60–62.

[172] J. Neyman. "Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability." In: *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences* 236.767 (1937), pp. 333–380. ISSN: 00804614. URL: http://www.jstor.org/stable/91337.

[173] Robert Schall. "The empirical coverage of confidence intervals: Point estimates and confidence intervals for confidence levels." In: *Biometrical journal* 54.4 (2012), pp. 537–551.

[174] Charlotte Strege et al. "Fundamental statistical limitations of future dark matter direct detection experiments." In: *Phys. Rev. D* 86.2, 023507 (July 2012), p. 023507. DOI: 10.1103/PhysRevD.86.023507. arXiv: 1201.3631 [hep-ph]. URL: https://ui.adsabs.harvard.edu/abs/2012PhRvD..86b3507S.

[175] D. Prangle et al. "Diagnostic tools of approximate Bayesian computation using the coverage property." In: *arXiv e-prints*, arXiv:1301.3166 (Jan. 2013), arXiv:1301.3166. arXiv: 1301.3166 [stat.ME].

[176] Sean Talts et al. "Validating Bayesian inference algorithms with simulation-based calibration." In: *arXiv preprint arXiv:1804.06788* (2018).

[177] Niccolo Dalmasso et al. "Validation of approximate likelihood and emulator models for computationally intensive simulations." In: *International Conference on Artificial Intelligence and Statistics*. May 2020, arXiv:1905.11505, pp. 3349–3361. arXiv: 1905.11505 [stat.ME].

[178] Gaia Collaboration and Brown. "Gaia Data Release 2. Summary of the contents and survey properties." In: *Astronomy and Astrophysics* 616, A1 (Aug. 2018), A1. DOI: 10.1051/0004-6361/201833051. arXiv: 1804.09365 [astro-ph.GA].

[179] Gaia Collaboration and Prusti. "The Gaia mission." In: *A&A* 595, A1 (Nov. 2016), A1. DOI: 10.1051/0004-6361/201629272. arXiv: 1609.04153 [astro-ph.IM].

[180] Kenneth C. Chambers and Pan-STARRS Team. "The Pan-STARRS1 Survey Data Release." In: *American Astronomical Society Meeting Abstracts 229*. Vol. 229. American Astronomical Society Meeting Abstracts. Jan. 2017, 223.03, p. 223.03.

[181] Sergey Ioffe and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." In: *arXiv preprint arXiv:1502.03167* (2015).

[182] Günter Klambauer et al. "Self-normalizing neural networks." In: *Advances in Neural Information Processing Systems*. 2017, pp. 971–980.

[183] Ilya Loshchilov and Frank Hutter. "Decoupled weight decay regularization." In: *arXiv preprint arXiv:1711.05101* (2017).

[184] Brandon Yang et al. "Condconv: Conditionally parameterized convolutions for efficient inference." In: *Advances in Neural Information Processing Systems*. 2019, pp. 1307–1318.

[185] David Ha, Andrew Dai, and Quoc V Le. "Hypernetworks." In: *arXiv preprint arXiv:1609.09106* (2016).

[186] Geoffrey E Hinton et al. "Improving neural networks by preventing co-adaptation of feature detectors." In: *arXiv preprint arXiv:1207.0580* (2012).

[187] Jarno Lintusaari et al. "Fundamentals and Recent Developments in Approximate Bayesian Computation." en. In: *Syst. Biol.* 66.1 (Jan. 2017), e66–e82. ISSN: 1063-5157, 1076-836X. DOI: 10.1093/sysbio/syw077. URL: http://dx.doi.org/10.1093/sysbio/syw077.

[188] Dennis Prangle. "Adapting the ABC Distance Function." en. In: *Bayesian Anal.* 12.1 (Mar. 2017), pp. 289–309. ISSN: 1936-0975, 1931-6690. DOI: 10.1214/16-BA1002. URL: https://projecteuclid.org/euclid.ba/1460641065.

[189] Xiangdong Shi and George M Fuller. "New dark matter candidate: nonthermal sterile neutrinos." In: *Phys. Rev. Lett.* 82.14 (Apr. 1999), p. 2832. DOI: 10.1103/PhysRevLett.82.2832. arXiv: astro-ph/9810076 [astro-ph]. URL: https://ui.adsabs.harvard.edu/abs/1999PhRvL..82.2832S.

[190] Kevork Abazajian, George M. Fuller, and Mitesh Patel. "Sterile neutrino hot, warm, and cold dark matter." In: *Phys. Rev. D* D64.2 (2001), p. 023501. DOI: 10.1103/physrevd.64.023501. arXiv: astro-ph/0101524 [astro-ph].

[191] Takehiko Asaka and Mikhail Shaposhnikov. "The νMSM, dark matter and baryon asymmetry of the universe." In: *Phys. Rev. B* 620.1-2 (2005), pp. 17–26. ISSN: 0370-2693. DOI: 10.1016/j.physletb.2005.06.020.

[192] Alexey Boyarsky, Oleg Ruchayskiy, and Mikhail Shaposhnikov. "The role of sterile neutrinos in cosmology and astrophysics." In: *Annual Review of Nuclear and Particle Science* 59.1 (Nov. 2009), pp. 191–214. DOI: 10.1146/annurev.nucl.010909.083654. arXiv: 0901.0011 [hep-ph]. URL: https://ui.adsabs.harvard.edu/abs/2009ARNPS..59..191B.

[193] Jan-Matthis Lueckmann et al. "Benchmarking Simulation-Based Inference." In: *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*. Ed. by Arindam Banerjee and Kenji Fukumizu. Vol. 130. Proceedings of Machine Learning Research. PMLR, 2021, pp. 343–351.

[194] Johann Brehmer et al. "A Guide to Constraining Effective Field Theories with Machine Learning." In: *Phys. Rev. D* 98.5 (2018), p. 052004. DOI: 10.1103/PhysRevD.98.052004. arXiv: 1805.00020 [hep-ph].

[195] Johann Brehmer et al. "Mining for Dark Matter Substructure: Inferring subhalo population properties from strong lenses with machine learning." In: *The Astrophysical Journal* 886.1 (2019), p. 49.

[196] Joeri Hermans et al. "Towards constraining warm dark matter with stellar streams through neural simulation-based inference." In: (Nov. 2020). arXiv: 2011.14923 [astro-ph.GA].

[197] Niccolò Dalmasso, Rafael Izbicki, and Ann B Lee. "Confidence Sets and Hypothesis Testing in a Likelihood-Free Inference Setting." In: *arXiv preprint arXiv:2002.10399* (2020).

[198] Georges Aad et al. "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC." In: *Physics Letters B* 716.1 (2012), pp. 1–29.

[199] Daniel Gilman et al. "Probing the nature of dark matter by forward modelling flux ratios in strong gravitational lenses." In: *Monthly Notices of the Royal Astronomical Society* 481.1 (2018), pp. 819–834.

[200] N. Aghanim et al. "Planck 2018 results. VI. Cosmological parameters." In: *Astron. Astrophys.* 641 (2020). [Erratum: Astron.Astrophys. 652, C4 (2021)], A6. DOI: 10.1051/0004-6361/201833910. arXiv: 1807.06209 [astro-ph.CO].

[201] Benjamin P Abbott et al. "GW151226: observation of gravitational waves from a 22-solar-mass binary black hole coalescence." In: *Physical review letters* 116.24 (2016), p. 241103.

[202] George EP Box and George C Tiao. *Bayesian inference in statistical analysis*. Vol. 40. John Wiley & Sons, 1973.

[203] Rob J Hyndman. "Computing and graphing highest density regions." In: *The American Statistician* 50.2 (1996), pp. 120–126.

[204] Erich L Lehmann and Joseph P Romano. *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

[205] David Lopez-Paz and Maxime Oquab. "Revisiting classifier two-sample tests." In: *arXiv preprint arXiv:1610.06545* (2016).

[206] Arthur Gretton et al. "A kernel two-sample test." In: *The Journal of Machine Learning Research* 13.1 (2012), pp. 723–773.

[207] Yoshua Bengio, Li Yao, and Kyunghyun Cho. "Bounding the test log-likelihood of generative models." In: *arXiv preprint arXiv:1311.6184* (2013).

[208] Gintare Karolina Dziugaite, Daniel M Roy, and Zoubin Ghahramani. "Training generative neural networks via maximum mean discrepancy optimization." In: *arXiv preprint arXiv:1505.03906* (2015).

[209] Laurent Dinh, David Krueger, and Yoshua Bengio. "Nice: Nonlinear independent components estimation." In: *arXiv preprint arXiv:1410.8516* (2014).

[210] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. "Density estimation using Real NVP." In: *arXiv preprint arXiv:1605.08803* (2016). URL: https://arxiv.org/abs/1605.08803.

[211] Owen Thomas et al. "Likelihood-free inference by ratio estimation." In: *Bayesian Analysis* (2016).

[212] Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. "Density-ratio matching under the Bregman divergence: a unified framework of density-ratio estimation." In: *Annals of the Institute of Statistical Mathematics* 64.5 (2012), pp. 1009–1044.

[213] Donald B. Rubin. "Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician." In: *The Annals of Statistics* 12.4 (1984), pp. 1151 –1172. DOI: 10.1214/aos/1176346785. URL: https://doi.org/10.1214/aos/1176346785.

[214] Jonathan K Pritchard et al. "Population growth of human Y chromosomes: a study of Y chromosome microsatellites." In: *Molecular biology and evolution* 16.12 (1999), pp. 1791–1798.

[215] Scott A Sisson, Yanan Fan, and Mark A Beaumont. "Overview of ABC." In: *Handbook of approximate Bayesian computation*. Chapman and Hall/CRC, 2018, pp. 3–54.

[216] Tina Toni and Michael P. H. Stumpf. "Simulation-based model selection for dynamical systems in systems and population biology." In: *Bioinformatics* 26.1 (Oct. 2009), pp. 104–110. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp619. eprint: https://academic.oup.com/bioinformatics/article-pdf/26/1/104/16893632/btp619.pdf. URL: https://doi.org/10.1093/bioinformatics/btp619.

[217] S.A. Sisson, Y. Fan, and M.M. Tanaka. "Sequential Monte Carlo without likelihoods." In: *Proceedings of the National Academy of Sciences* 104.6 (Jan. 2007), pp. 1760–1765. ISSN: 0027-8424, 1091-6490. URL: https://doi.org/10.1073/pnas.0607208104.

[218] Mark A Beaumont et al. "Adaptive approximate Bayesian computation." In: *Biometrika* 96.4 (2009), pp. 983–990.

[219] George Papamakarios and Iain Murray. "Fast ε-free inference of simulation models with bayesian conditional density estimation." In: *Advances in neural information processing systems*. 2016, pp. 1028–1036.

[220] Jan-Matthis Lueckmann et al. "Flexible statistical inference for mechanistic models of neural dynamics." In: *arXiv preprint arXiv:1711.01861* (2017).

[221] David Greenberg, Marcel Nonnenmacher, and Jakob Macke. "Automatic posterior transformation for likelihood-free inference." In: *International Conference on Machine Learning*. PMLR. 2019, pp. 2404–2414.

[222] George Papamakarios, David Sterratt, and Iain Murray. "Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows." In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR. 2019, pp. 837–848.

[223] Conor Durkan, Iain Murray, and George Papamakarios. "On contrastive learning for likelihood-free inference." In: *International Conference on Machine Learning*. PMLR. 2020, pp. 2771–2781.

[224] Kyle Cranmer et al. *"Active Sciencing" with Reusable Workflows*. https://github.com/cranmer/active_sciencing. 2017.

[225] Vito Volterra. "Fluctuations in the abundance of a species considered mathematically." In: *Nature* 118.2972 (1926), pp. 558–560.

[226] LIGO Scientific Collaboration. *LIGO Algorithm Library - LALSuite*. free software (GPL). 2018. DOI: 10.7935/GT1W-FZ16.

[227] C. M. Biwer et al. "PyCBC Inference: A Python-based parameter estimation toolkit for compact binary coalescence signals." In: *Publ. Astron. Soc. Pac.* 131.996 (2019), p. 024503. DOI: 10.1088/1538-3873/aaef0b. arXiv: 1807.10312 [astro-ph.IM].

[228] Alvaro Tejero-Cantero et al. "SBI–A toolkit for simulation-based inference." In: *arXiv preprint arXiv:2007.09114* (2020).

[229] Jing Lin. "An integrated procedure for bayesian reliability inference using MCMC." In: *Journal of Quality and Reliability Engineering* 2014 (2014).

[230] David W Hogg and Daniel Foreman-Mackey. "Data analysis recipes: Using markov chain monte carlo." In: *The Astrophysical Journal Supplement Series* 236.1 (2018), p. 11.

[231] John F Geweke et al. *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*. Tech. rep. Federal Reserve Bank of Minneapolis, 1991.

[232] Andrew Gelman and Donald B Rubin. "Inference from iterative simulation using multiple sequences." In: *Statistical science* 7.4 (1992), pp. 457–472.

[233] Adrian E Raftery and Steven Lewis. *How many iterations in the Gibbs sampler?* Tech. rep. WASHINGTON UNIV SEATTLE DEPT OF STATISTICS, 1991.

[234] Anand Dixit and Vivekananda Roy. "MCMC diagnostics for higher dimensions using Kullback Leibler divergence." In: *Journal of Statistical Computation and Simulation* 87.13 (2017), pp. 2622–2638.

[235] Anand Ulhas Dixit. "Developments in MCMC diagnostics and sparse Bayesian learning models." PhD thesis. Iowa State University, 2018.

[236] Vivekananda Roy. "Convergence diagnostics for markov chain monte carlo." In: *Annual Review of Statistics and Its Application* 7 (2020), pp. 387–412.

[237] François Rozet and Gilles Louppe. "Arbitrary Marginal Neural Ratio Estimation for Likelihood-free Inference." MA thesis. University of Liège, Belgium, 2021. URL: https://hdl.handle.net/2268.2/12993.

[238] Niccolò Dalmasso et al. "Likelihood-Free Frequentist Inference: Bridging Classical Statistics and Machine Learning in Simulation and Uncertainty Quantification." In: *arXiv preprint arXiv:2107.03920* (2021).

[239] Mohamed Ishmael Belghazi et al. "Mine: mutual information neural estimation." In: *arXiv preprint arXiv:1801.04062* (2018).

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this or any other University.

Joeri Hermans