# Pseudo-Likelihood Inference for Bayesian System Identification

**Anonymous Authors**[1]

## Abstract

Simulation-Based Inference (SBI) is a common name for an emerging family of approaches that infer the model parameters when the likelihood is intractable. Existing SBI methods either approximate the likelihood, such as Approximate Bayesian Computation (ABC), or directly model the posterior, such as Sequential Neural Posterior Estimation (SNPE). While ABC is efficient on low-dimensional problems, on higher-dimensional tasks, it is generally outperformed by SNPE which leverages function approximation. In this paper, we propose Pseudo-Likelihood Inference (PLI), a new method that brings neural approximation into ABC, making it competitive on challenging Bayesian system identification tasks. By utilizing integral probability metrics, we introduce a smooth likelihood kernel with an adaptive bandwidth that is updated based on information-theoretic trust regions. Thanks to this formulation, our method (i) allows for optimizing neural posteriors via gradient descent, (ii) does not rely on summary statistics, and (iii) enables multiple observations as input. In comparison to SNPE, it leads to improved performance when more data is available. The effectiveness of PLI is evaluated on two classical SBI benchmark tasks and on a highly dynamic physical system, showing particular advantages on stochastic simulations and multi-modal posterior landscapes.

## 1. Introduction

Parametric stochastic simulators are a well-established tool for predicting the behavior of real-world phenomena. These statistical models find widespread use in various scientific fields such as physics, economics, biology, ecology, computer science, and robotics, where they help to gain knowledge about the underlying stochastic processes (Hartig et al., 2011; McGoff et al., 2015) or generate additional data for subsequent downstream tasks (Muratore et al., 2022). In both cases, the practitioner seeks to explain the observations as accurately as possible while incorporating all available information. The output of such a simulator is largely determined by its parameters and their values. When estimating these parameters using Bayesian inference given observations from a physical system, which are inevitably subject to measurement noise, we obtain a distribution over values instead of a point estimate. Additionally, there might be several parameter configurations yielding the same observation, hence rendering the resulting distribution to be multi-modal. Moreover, for many practical use cases, the likelihood function is unknown or too expensive to evaluate. The combination of these difficulties makes obtaining a posterior distribution over simulator parameters challenging for state-of-the-art inference methods, both regarding effectiveness as well as efficiency.

SBI approaches address the issue of intractable likelihoods by using (stochastic) simulators as forward models to generate observations from proposal distributions over parameters. The approaches are also often called *likelihood-free*, which can be easily misunderstood since some of them directly approximate the likelihood (Cranmer et al., 2020). ABC is a family of SBI methods that approximate the posterior with a set of weighted particles which are obtained from Monte Carlo simulations and updated based on an empirical estimation of the intractable likelihood (Sisson et al., 2018). For an ABC approach to work well, three criteria have to be fulfilled: (i) the likelihood kernel is capable of measuring the similarity of observations meaningfully, (ii) the proposal distribution samples close to the posterior, (iii) the decision-making rule balances between accepting a sufficient amount of samples from the proposal and steering inference towards the posterior distribution. Constructing a suitable likelihood kernel often means tailoring summary statistics to the problem at hand. However, recent advances promise to replace this heuristic-based process by employing Integral Probability Metrics (IPMs) to measure statistical distances in observation space (Bernton et al., 2019; Dellaporta et al., 2022; Drovandi & Frazier, 2022). While these methods significantly increase the required number of simulations, approximations of the statistical distances (Gretton et al., 2012;

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Cuturi, 2013) can be computed in parallel, hence facilitating the parallelization of the whole inference pipeline. Following up on the shortcomings of ABC, the family of SNPE approaches provide Bayesian inference methods that leverage conditional neural density estimators to approximate the posterior (Papamakarios & Murray, 2016; Greenberg et al., 2019; Papamakarios et al., 2019; Durkan et al., 2020). The benchmarking study of Lueckmann et al. (2021) concludes that, generally, SNPE approaches are to be preferred over ABC as they are superior in terms of expressibility and accuracy across a wide range of benchmarking tasks. However, it is important to point out that the analysis of the posterior inference has solely been reported for single observations. These single-sample scenarios favor SNPE in high dimensions since ABC relies on summary statistics to evaluate the likelihood. Therefore, it remains an open question whether SNPE methods can transfer their benefits to settings where the (approximated) posterior is conditioned on multiple observations at once.

**Contributions.** By deriving the ABC posterior from a constrained variational inference objective, we introduce a novel SBI method called Pseudo-Likelihood Inference (PLI). Our formulation allows for approximating the posterior with neural density estimators. PLI updates this posterior from pseudo-likelihoods which are exponentially transformed statistical distances computed using IPMs. To further remove heuristics from the inference process, we derive an adaptive bandwidth update of PLI's likelihood kernel that bounds the loss of information based on information-geometric trust-region principles. This way, PLI can update its neural posterior solely given observations from a (stochastic) black-box simulator. Moreover, the usage of IPMs enables PLI to simultaneously condition on a variable number of observations, while SNPE methods need to concatenate them and, therefore, degrade when the number of observations increases. We compare PLI against ABC and one SNPE method on two SBI benchmarking tasks as well as a highly dynamical double pendulum task. For both ABC and PLI, we investigate two IPMs: the maximum mean discrepancy and the Wasserstein distance. Motivated by recent benchmarking results, we chose Automatic Posterior Transformation (APT) (Greenberg et al., 2019) to represent SNPE approaches. Our experiments investigate the dependency of the trained density estimator's performance on the number of observations, where performance is measured in observation as well as parameter space. We show the merits and disadvantages of all methods and conclude with concrete recommendations.

## 2. Bayesian Inference with Intractable Likelihoods

The objective of Bayesian inference is to estimate the posterior parameter distribution $p(\boldsymbol{\xi}|\mathcal{X}^\star)$ given a set of reference data points $\mathcal{X}^\star = \{\mathbf{x}_1^\star, \dots, \mathbf{x}_M^\star\}$ which are assumed to be drawn from the likelihood model $p(\mathbf{x}|\boldsymbol{\xi})$. Given a prior belief over the parameters $p(\boldsymbol{\xi})$, the posterior can be expressed via Bayes' rule

$$p(\boldsymbol{\xi}|\mathcal{X}^\star) \propto p(\mathcal{X}^\star|\boldsymbol{\xi})\, p(\boldsymbol{\xi}). \tag{1}$$

In the following, we describe SBI methodologies that carry out approximate inference of (1) for the case in which the simulator represents an intractable likelihood from which only sampling is possible $\mathcal{X} = \{\mathbf{x}_i \sim p(\mathbf{x}_i|\boldsymbol{\xi}) | i = 1:M\}$, but evaluating the likelihood is infeasible.

### 2.1. Approximate Bayesian Computation (ABC)

ABC methods perform Bayesian inference without explicitly computing the likelihood function $p(\mathcal{X}^\star|\boldsymbol{\xi}) = \int p(\mathcal{X}^\star|\mathcal{X}, \boldsymbol{\xi}) p(\mathcal{X}|\boldsymbol{\xi})\, \mathrm{d}\mathcal{X}$ (Sisson et al., 2018). To approximate the data likelihood, ABC uses Monte Carlo samples $\mathcal{X}$ from the simulator as the reference points and smoothes them with a kernel $K_\beta(D(\mathcal{X}^\star, \mathcal{X}))$ (Karabatsos & Leisen, 2018)

$$p_\beta(\mathcal{X}^\star|\boldsymbol{\xi}) \propto \int K_\beta(D(\mathcal{X}^\star, \mathcal{X})) p(\mathcal{X}|\boldsymbol{\xi}) \mathrm{d}\mathcal{X}. \tag{2}$$

The kernel assesses the similarity of the reference data $\mathcal{X}^\star$ and the simulated data $\mathcal{X}$ based on a distance measure $D$, the kernel type $K$, and the kernel bandwidth $\beta$. The uniform kernel $\mathbb{1}_{\{D(\mathcal{X}^\star, \mathcal{X}) \leq \beta\}}$ (see Table 1) has emerged as the default kernel choice of many ABC methods (Del Moral et al., 2012; Lee, 2012; Lenormand et al., 2013). In this case, the bandwidth represents a rejection threshold that assigns zero probability to all parameters whose simulations lie outside of the $\beta$-ball in terms of the distance $D$. The uniform kernel exhibits the favorable characteristic of converging to the likelihood in the limit

$$\lim_{\beta \to 0} p_\beta(\mathcal{X}^\star|\boldsymbol{\xi}) = \int \mathbb{1}_{\{\mathcal{X}^\star\}}(\mathcal{X}) p(\mathcal{X}|\boldsymbol{\xi}) \mathrm{d}\mathcal{X} = p(\mathcal{X}^\star|\boldsymbol{\xi}).$$

Once the likelihood approximation (2) is obtained, ABC draws samples from the approximate posterior

$$p_\beta(\boldsymbol{\xi}|\mathcal{X}^\star) \propto p_\beta(\mathcal{X}^\star|\boldsymbol{\xi})\, p(\boldsymbol{\xi}). \tag{3}$$

There are multiple ways of implementing the sampling procedure. In rejection ABC (Tavaré et al., 1997), the simplest form, proposal parameters are drawn from the prior distribution $p(\boldsymbol{\xi})$ and are accepted if the simulated data falls close to the true data, as measured by the kernel function.

While rejection ABC yields a simple algorithm with desirable convergence properties, finding posterior samples for small bandwidths $\beta$ in high dimensions often becomes computationally infeasible (Marjoram et al., 2003). Therefore, the research on ABC focuses on three directions of improvement: (i) replacing the prior with a sequentially updated proposal distribution $\tilde{p}_t(\boldsymbol{\xi})$ to reduce the search space

during sampling, (ii) adapting the bandwidth $\beta$ to draw samples with an appropriate acceptance rate, and (iii) finding sufficient statistics to represent the simulated output in low dimensions (Sisson et al., 2018). MCMC-ABC (Marjoram et al., 2003) and SMC-ABC (Sisson et al., 2007; Toni et al., 2009; Del Moral et al., 2012; Lenormand et al., 2013) build upon sampling strategies based on Markov Chain Monte Carlo (MCMC) and Sequential Monte Carlo (SMC) to sequentially update the proposal distribution. MCMC-ABC does not allow for an adaptive bandwidth, and thus, SMC sampling strategies have evolved as the leading ABC methods for these cases Del Moral et al. (2012).

### 2.2. Sequential Monte Carlo ABC

SMC–ABC builds on SMC samplers introduced by Del Moral et al. (2006). Fundamentally, SMC-ABC approximates the posterior distribution through a sequence of intermediate *target posterior* distributions (3) that are characterized by an adaptable bandwidth parameter $\beta_t$. Furthermore, SMC-ABC uses importance sampling from a sequentially updated particle-based proposal distribution $\tilde{p}_t(\boldsymbol{\xi})$ to improve the sample efficiency.

The proposal distribution is represented by an empircal distribution $\tilde{p}_t(\boldsymbol{\xi}) = \sum_{i=0}^{M} \delta_{\boldsymbol{\xi}_t^{(i)}}(\boldsymbol{\xi})$ that is defined by a set of particles $\{\boldsymbol{\xi}_t^{(i)}\}$. Importance sampling then enables the approximation of the target posterior $p_{\beta_t}(\boldsymbol{\xi}|\mathcal{X}^\star)$ from the proposal distribution

$$q_t(\boldsymbol{\xi}) = \sum_{i=1}^{M} W_t^{(i)} \delta_{\boldsymbol{\xi}_t^{(i)}}(\boldsymbol{\xi}); \quad W_t^{(i)} = \frac{p_{\beta_t}(\boldsymbol{\xi}_t^{(i)}|\mathcal{X}^\star)}{\tilde{p}_t(\boldsymbol{\xi}^{(i)})}. \quad (4)$$

Assume now that at inference time $t$, an approximation of the target posterior $p_{\beta_t}(\boldsymbol{\xi}|\mathcal{X}^\star)$ is available by (4). Then SMC-ABC methods follow three steps to carry out inference for the next target posterior $p_{\beta_{t+1}}(\boldsymbol{\xi}|\mathcal{X}^\star)$: (i) A new bandwidth $\beta_{t+1}$ of the target posterior $p_{\beta_{t+1}}(\mathcal{X}^\star|\boldsymbol{\xi})$ is estimated. Typically, the update is based on heuristics, such as the Effective Sample Size (ESS) (Del Moral et al., 2012) to ensure that the particle variance does not degrade. (ii) New proposal particles $\boldsymbol{\xi}_t^{(i)}$ are sampled from a forward Markov kernel $\boldsymbol{\xi}_{t+1}^{(i)} \sim K_t(\boldsymbol{\xi}_{t-1}^{(i)}, \boldsymbol{\xi}_t^{(i)})$ in order to stay close to the target posterior of the next iteration $p_{\beta_{t+1}}(\boldsymbol{\xi}|\mathcal{X}^\star)$. (iii) The weights of the particles are adjusted based on approximations of (4). As the weight update is typically numerically intractable (Del Moral et al., 2006), different SMC-ABC methods (Del Moral et al., 2012; Lee, 2012; Lenormand et al., 2013) have been introduced which propose approximations to the optimal weight update. We refer to Appendix B for a more detailed explanation of SMC-ABC and its different approaches.

## 3. Pseudo-Likelihood Inference

The proposed PLI methodology, summarized in Figure 1, generalizes the ABC approaches by introducing exponential likelihood kernels with adaptive bandwidth updates, which are motivated from a Variational Inference (VI) perspective.

### 3.1. Exponential Likelihood Kernels

PLI adopts the view of SMC-ABC on approximating a smoothed target posterior $\tilde{p}_t(\boldsymbol{\xi})$ by formulating the following constrained VI problem for each inference step $t$

$$\begin{aligned} \tilde{p}_t(\boldsymbol{\xi}) = \arg\min_{p_t(\boldsymbol{\xi})} & D_{\mathrm{KL}}\left(p_t(\mathcal{X},\boldsymbol{\xi}) \,||\, p(\mathcal{X},\boldsymbol{\xi}|\mathcal{X}^\star)\right), \\ \text{s.t. } & D_{\mathrm{KL}}\left(p_t(\mathcal{X},\boldsymbol{\xi}) \,||\, \tilde{p}_{t-1}(\mathcal{X},\boldsymbol{\xi})\right) \leq \varepsilon, \quad (5) \\ & \int p_t(\mathcal{X},\boldsymbol{\xi}) \mathrm{d}\mathcal{X}\, \mathrm{d}\boldsymbol{\xi} = 1. \end{aligned}$$

This problem is formulated to optimize the target posterior $\tilde{p}_t(\boldsymbol{\xi})$. Since we assume the likelihood $p(\mathcal{X}|\boldsymbol{\xi})$ to be fixed, the joint distribution can be decomposed as $p_t(\mathcal{X},\boldsymbol{\xi}) = p(\mathcal{X}|\boldsymbol{\xi})p_t(\boldsymbol{\xi})$. We further refer to $\tilde{p}_{t-1}(\mathcal{X},\boldsymbol{\xi}) = p(\mathcal{X}|\boldsymbol{\xi})\tilde{p}_{t-1}(\boldsymbol{\xi})$ as the joint representation of the target posterior $\tilde{p}_{t-1}(\boldsymbol{\xi})$ of the previous iteration. The trust-region threshold $\varepsilon$ limits the maximum step size that the target posterior can move away from the previous distribution in the probability manifold. The constrained objective balances between fitting the joint posterior distribution $p(\boldsymbol{\xi},\mathcal{X}|\mathcal{X}^\star)$ and bounding the information loss between two inference steps. The optimal target posterior $\tilde{p}_t(\boldsymbol{\xi})$ at inference time $t$ for the above optimization problem (5) is given by

$$\tilde{p}_t(\boldsymbol{\xi}) \propto \left(\frac{p(\boldsymbol{\xi})}{\tilde{p}_{t-1}(\boldsymbol{\xi})}\right)^{\frac{\nu}{\nu+\eta_t}} \underbrace{\exp\left(\frac{\mathbb{E}_{p(\mathcal{X}|\boldsymbol{\xi})} \log p(\mathcal{X}^\star|\mathcal{X},\boldsymbol{\xi})}{\nu+\eta_t}\right)}_{\hat{p}_t(\mathcal{X}^\star|\boldsymbol{\xi})} \tilde{p}_{t-1}(\boldsymbol{\xi}) \quad (6)$$

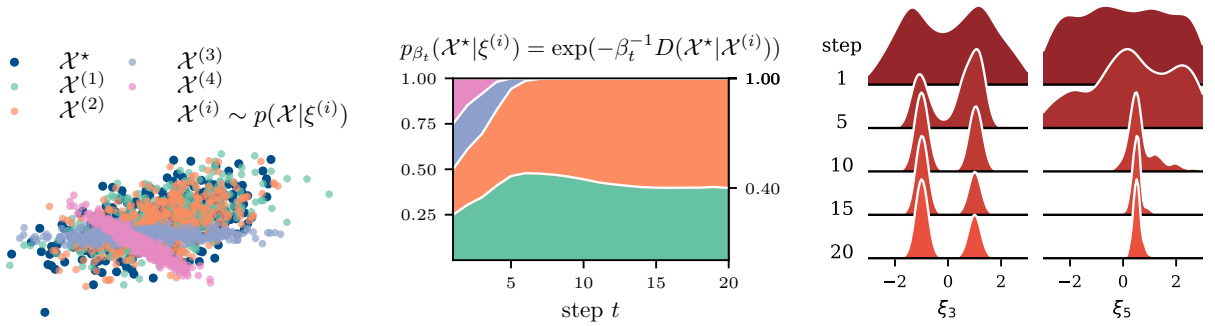
where $\eta_t$ and $\nu$ are the Lagrangian multipliers corresponding to the constraints in (5). Please see Appendix A.1 for more details. Assuming a monotonically decreasing adaptive bandwidth schedule $\eta_t \to 0$, we further show in Appendix A.2 that the series of target posteriors converges to the Bayes posterior (1).

Similar to ABC, we propose to approximate the intractable distribution $p(\mathcal{X}^\star|\mathcal{X},\boldsymbol{\xi})$ with a kernel similarity measure. In this case, we explicitly choose a Gibbs distribution

$$p_\beta(\mathcal{X}^\star|\mathcal{X},\boldsymbol{\xi}) \propto \exp\left(-\frac{1}{2\beta} D(\mathcal{X}^\star,\mathcal{X})\right) \quad (7)$$

with a fixed temperature parameter $\beta_t$ which yields a Gibbs kernel approximation to $\hat{p}_t(\mathcal{X}^\star|\boldsymbol{\xi})$

$$\hat{p}_{\beta_t}(\mathcal{X}^\star|\boldsymbol{\xi}) \propto \exp\left(-\frac{\mathbb{E}_{p(\mathcal{X}|\boldsymbol{\xi})}\left[D(\mathcal{X}^\star,\mathcal{X})\right]}{2\beta(\nu+\eta_t)}\right). \quad (8)$$




(a) Observations (b) Pseudo-likelihood evolution (c) Marginal posterior evolution

Figure 1: Schematic overview of the introduced Pseudo-Likelihood Inference (PLI) approach based on the SLCP task, described in Section 4.2: (a) observations generated from the stochastic simulator for four different parameter configurations $\boldsymbol{\xi}^{(1:4)}$ and the reference observation $\mathcal{X}^\star$, (b) approximation of the likelihood leveraging the pseudo-likelihood with adaptive bandwidth $\beta_t$ as defined in Section 3, (c) the estimated posterior is sequentially updated based on the refined pseudo-likelihood evaluations. We show the marginal evolution of two exemplary system parameters $\xi_3$ and $\xi_5$.

The bandwidth $\beta_t = \beta(\nu+\eta_t)$ is a time-varying temperature parameter that emulates the role of the rejection parameter in the uniform kernel.

**Bandwidth Adaptation from Trust-Region Principles.** The dual formulation of the stochastic search problem (5) leads to a tractable solution for the optimal bandwidth parameter $\beta_t$. While the optimization of the Lagrangian directly leads to the optimal posterior at time $t$, the dual problem involes a maximization of the dual objective w.r.t. $\eta_t$,

$$g(\eta_t) = -\eta_t \epsilon - (\nu + \eta_t) \log \mathop{\mathbb{E}}_{\tilde{p}_{t-1}(\boldsymbol{\xi})} \left[\tilde{p}_{t-1}(\mathcal{X}^\star|\boldsymbol{\xi})\right]. \quad (9)$$

The bandwidth steers the loss of information between successive steps. In early inference stages, the proposal prior $p_{t-1}(\boldsymbol{\xi})$ is typically uninformative and thus the information loss is moderate even if $p_t(\boldsymbol{\xi})$ moves far away from $p_{t-1}(\boldsymbol{\xi})$. In later inference steps, the proposal distribution is typically pronounced and small deviations may lead to significant information loss. Therefore, the bandwidth is expected to decay over time. The optimal bandwidth parameter $\beta_t$ that is obtained by maximizing (9) for $\eta_t$ can be seen as an information-bounded trust region update to move the pseudo-likelihood towards the likelihood.

### 3.2. Bayesian Inference with Pseudo-Likelihoods

Pseudo-Likelihood Inference (PLI) is a sequential SBI methodology which bases on the previously derived target posterior. The method is closely tied to SMC-ABC by (i) sequentially approximating *target posteriors* that converge to the posterior distribution, (ii) adapting the bandwidth sequentially, and (iii) updating the proposal distribution for higher sample efficiency. Instead of representing the posterior through a set of weighted particles, the PLI formulation allows for a variety of powerful parameterized density estimators.

A parameterized density estimator $q_{\boldsymbol{\phi}_t}(\boldsymbol{\xi})$ is trained to approximate the PLI posterior (6) in terms of the M-projection of the Kullback-Leibler (KL) divergence

$$\min_{\boldsymbol{\phi}_t} \; D_{\mathrm{KL}}\left(\tilde{p}_t \,||\, q_{\boldsymbol{\phi}_t}\right) = \max_{\boldsymbol{\phi}} \mathop{\mathbb{E}}_{\tilde{p}_{t-1}(\boldsymbol{\xi})} \left[w \log q_{\boldsymbol{\phi}}(\boldsymbol{\xi})\right],$$
$$w = \left(\frac{p(\boldsymbol{\xi})}{\tilde{p}_{t-1}(\boldsymbol{\xi})}\right)^{\frac{\nu}{\nu+\eta_t}} p_{\beta_t}(\mathcal{X}^\star|\boldsymbol{\xi}). \quad (10)$$

We deduce the PLI methodology by realizing that the M-projection is formulated as an expectation of the proposal distribution (see Appendix A.3 for further details). Thus, the density estimator can be optimized based on data drawn from the proposal distribution. The PLI methodology (see Algorithm 1) can be split into four distinct parts. The first step (lines 4–8) consists of drawing pairs of training samples from the proposal distribution $\boldsymbol{\xi}^{(i)} \sim p_{t-1}(\boldsymbol{\xi}^{(i)})$ and the simulator $\mathcal{X}^{(i)} \sim p(\mathcal{X}|\boldsymbol{\xi}^{(i)})$, and subsequently calculating the distance measure $D(\mathcal{X}^\star, \mathcal{X})$. Then in line 9, the bandwidth $\beta_t$ is optimized by maximizing the dual formulation (9). Afterward, in line 11, the parameterized density estimator is trained to approximate the target posterior $\tilde{p}_t(\boldsymbol{\xi})$ by minimizing (10). Note that the expectation on the proposal distribution enables gradient descent of the neural density estimator without requiring a differentiable simulator. In the final step (line 12), we set the current posterior approximation as the proposal of the next inference step, thus leveraging bootstrapping of the density estimator. While we restrict the analysis in this paper to the M-projection, we emphasize that the I-projection can be leveraged in the same way as shown in Appendix A.4.

**Algorithm 1** Pseudo-Likelihood Inference (PLI)

**input:** reference data $\mathcal{X}^\star$, prior $p(\boldsymbol{\xi})$, stochastic simulator $p(\mathbf{x}|\boldsymbol{\xi})$, IPM $D(\cdot,\cdot)$, posterior approximator $q_{\boldsymbol{\phi}}(\boldsymbol{\xi})$, max. iteration count $T$
initialize the proposal as the prior $\tilde{p}_0(\boldsymbol{\xi}) = p(\boldsymbol{\xi})$
**for** $t$ in $1:T$ **do**
  sample $K$ parameters $\boldsymbol{\xi}^{(k)} \sim \tilde{p}_{t-1}(\cdot)$ with $k = 1:K$
  **for** each $\boldsymbol{\xi}^{(k)}$ **do**
    simulate $\mathcal{X}^{(k)} = \{\mathbf{x}_n^{(k)} \sim p(\mathbf{x}|\boldsymbol{\xi}^{(k)})\}$
    compute the IPM $s^{(k)} = D(\mathcal{X}^{(k)}, \mathcal{X}^\star)$
  **end for**
  update $\eta_t$ by maximizing the dual (9) with non-linear solvers (e.g., Newton-CG)
  calculate the pseudo-likelihood $\hat{p}_{\beta_t}(\mathcal{X}^\star|\boldsymbol{\xi}^{(k)})$ see (8)
  fit the parameter distribution by weighted maximum likelihood (10)

$$\boldsymbol{\phi}_t = \arg\max_{\boldsymbol{\phi}} \sum_{k=1}^{K} w^{(k)} \log q_{\boldsymbol{\phi}}(\boldsymbol{\xi}^{(k)})$$

  set the new proposal prior $\tilde{p}_t(\boldsymbol{\xi}) = q_{\boldsymbol{\phi}_t}(\boldsymbol{\xi})$
**end for**
**output:** approximate posterior distribution $q_{\boldsymbol{\phi}_T}(\boldsymbol{\xi})$

# 4. Experiments

We compare the PLI framework against Wasserstein-ABC (W-ABC) (Bernton et al., 2019) and APT (Greenberg et al., 2019) on three diverse tasks. Our implementation of W-ABC is based on the SMC sampler approach (Del Moral et al., 2006), following Population Monte Carlo (PMC) (Table 1). While the original implementation of W-ABC leverages the r-hit kernel (Lee, 2012), our PMC-ABC implementation is based on Lenormand et al. (2013) because we observed better performance with that approach in preliminary studies. APT was chosen as the representative for the class of SNPE algorithms. We leverage Neural Spline Flows (NSFs) (Durkan et al., 2020) as density estimators for both PLI and APT. Both neural flow configurations share the same base network architecture, but for APT the conditional flow is augmented with an embedding network (Appendix C). All experiments are implemented in JAX (Bradbury et al., 2018).[1] To make the experiments comparable, simulation budgets of PLI and APT were fixed to 5000 samples per inference step over 20 episodes, while ABC moved 1000 particles over 200 episodes.

## 4.1. Evaluation Metrics

When available, the posteriors are compared against the reference posterior samples $\boldsymbol{\xi}^\star$. We also quantify the methods' performances based on their realizations by computing Wasserstein distance and the MMD. The comparison is carried out on 10 000 samples each. Furthermore, we use Posterior Predictive Checks (PPCs) to evaluate the predictive capabilities of the posterior models in the observation space $\int p(\mathcal{X}|\boldsymbol{\xi}) q(\boldsymbol{\xi}|\mathcal{X}^\star)\, \mathrm{d}\boldsymbol{\xi} \approx D(\mathcal{X}^\star, \mathcal{X})$. Due to the computational limits, the PPCs are carried out on 1000 simulations against the reference data. Lueckmann et al. (2021) also reports results with classifier-based tests and the kernelized Stein-discrepancy. However, since the focus of this paper is not on benchmarking, we restrict the analysis to comparing with the Wasserstein and MMD.

## 4.2. Tasks

We evaluate PLI on three tasks: Gaussian Location, Simple-Likelihood Complex-Posterior, and the Furuta pendulum. The first two are common tasks within the SBI community (Lueckmann et al., 2021), while the latter covers system identification of a highly dynamic continuous control system. The tasks' specifications are listed in Appendix C. For each task, we conduct experiments for different numbers of available reference observations $N = \{1, 2, 5, 10, 20, 50, 100\}$. The reference observations are simulated based on a pre-defined ground-truth parameter $\boldsymbol{\xi}^{\mathrm{gt}}$. Despite the fact that PLI and ABC can cope with varying numbers of samples $N$ and numbers of simulations per parameter $M$, we choose $N = M$ for all experiments as it is required by APT. In the following paragraphs, we discuss the results for each task separately, while Figure 2 summarizes all results.

**Gaussian Location.** The task of the Gaussian location model is to infer the mean of a 10-dimensional Gaussian distribution $\mathcal{N}(\mathbf{x}|\mu = \boldsymbol{\xi}, \Sigma = 0.1I)$. Thus, the posterior for this particular problem is available analytically. In Figure 2, we can see the performance with varying numbers of observations $N$. For fewer observations ($N \lessapprox 20$), APT matches the reference posterior better than the other approaches, whereas ABC and PLI outperform APT for larger $N$. W-PLI exhibits a degrading performance with increasing $N$. This disadvantage can be attributed to the Wasserstein distance not scaling well to high dimensions, which has been reported in several recent studies (Drovandi & Frazier, 2022; Dellaporta et al., 2022). The posterior plots in Figure 4 in the appendix give qualitative insights into the prediction performance of the learned posteriors which align well with the quantitative results mentioned above.

**Simple-Likelihood Complex-Posterior.** The SLCP task (Papamakarios et al., 2019) has five parameters and is designed to transform a simple likelihood into a multi-modal posterior distribution. The 8-dimensional observations represent four i.i.d samples from a Gaussian parameterized by five parameters. For further details, please see Appendix C.2. To evaluate the posterior approximations,

[1] Our code will be availible open-source upon publicaiton.

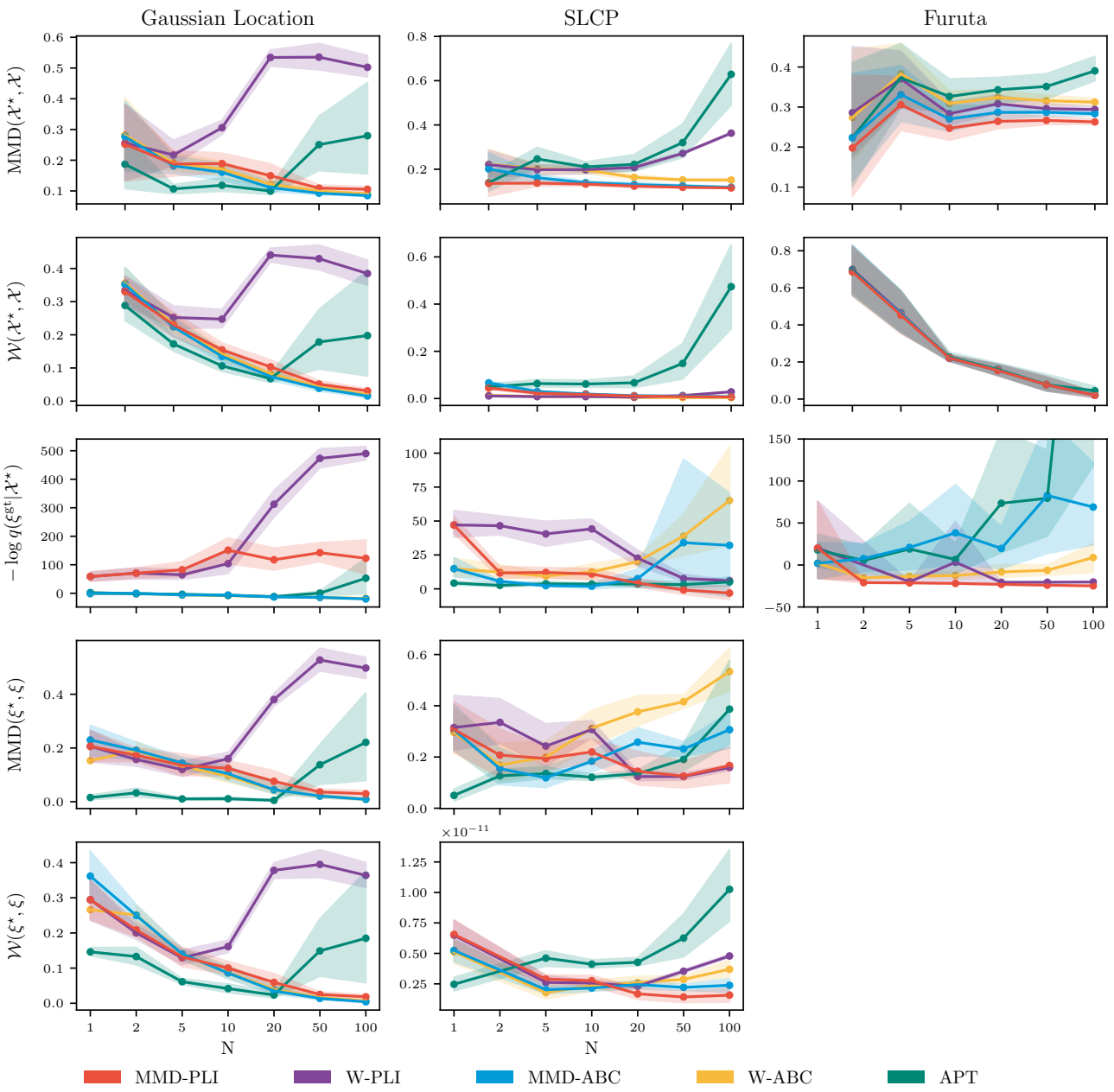



Figure 2: Evaluation of the posterior performance on three different tasks (displayed along the columns). A node represents the mean and standard deviation of 10 experiments with different random seeds, each carried out using $N$ data points for conditioning. The samples from the approximate posterior $\boldsymbol{\xi} \sim q(\boldsymbol{\xi}|\mathcal{X}^\star)$ are compared against the reference posterior samples with the Wasserstein distance and the Maximum Mean Discrepancy (MMD) when available. Additionally, the log probability of the ground-truth parameter $\boldsymbol{\xi}^{gt}$ is evaluated and posterior predictive checks are carried out on all tasks. The ground-truth parameters are described in Appendix C. Lower values are better for all metrics (displayed along the rows).

i.e., to obtain the posterior reference samples, we follow the improved MCMC sampling routine of Lueckmann et al. (2021). As displayed in Figure 2, PLI performs best where a sufficient amount of observations are present ($N \gtrapprox 20$) to evaluate the pseudo-likelihood. Again, the contrary can be seen for APT. While it performs strongly in low data regimes, APT deteriorates with an increasing number of observations. The main difference to the Gaussian location task is that here both ABC variants struggle to capture the multi-modality of the posterior distribution. This effect is

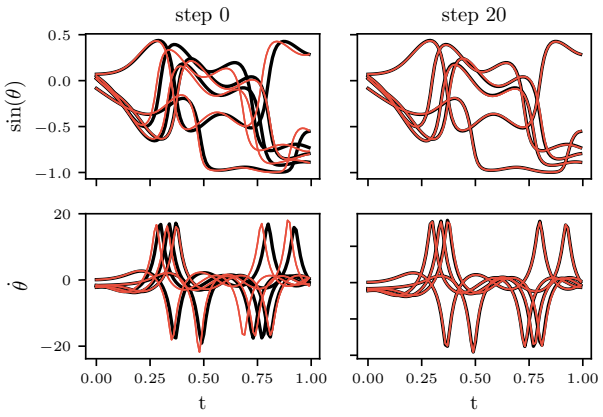



Figure 3: Predictive performance of the posterior learned with MMD-PLI on the Furuta pendulum on $N = 100$. The stochasticity of the environment is removed by synchronizing the initial state between the reference and predicted simulations. Thus, the only discrepancies between trajectories are due to the model not capturing the dynamics parameters of the system. As expected, parameter samples from the prior (left column) show a significant deviation over time. After the inference has been completed (right column) the predictive simulator (MMD-PLI) can completely recover the ground truth dynamics (Reference).

displayed in Figures 5 and 6 which allow for a qualitative assessment of the posterior approximations.

**Furuta Pendulum.** The Furuta pendulum is an inverted double pendulum set-up (Furuta et al., 1992). While the system's dynamics are inherently deterministic, small perturbations of the initial state around its unstable equilibrium point lead to highly diverse trajectories. The observation space is $t_{\text{end}} \times 6$ dimensional, where $t_{\text{end}}$ is the number of time steps per trajectory. We set the sampling frequency of the simulation to 100Hz and the duration to 1 sec, resulting in 600-dimensional observations. For this task, no reference posterior is available, and thus, the analysis is restricted to quantifying the observed data. Given the similarity of the Wasserstein distance and the MMD in parameter and observation space for the first two tasks, we argue that a comparison based on PPCs, i.e. $\mathcal{W}(\mathcal{X}^\star, \mathcal{X})$ and $\text{MMD}(\mathcal{X}^\star, \mathcal{X})$, is sufficient. The metrics reported in Figure 2 show the superior performance of both PLI variants as well as W-ABC. Additionally, Figure 3 exemplifies the predictive performance of the learned MMD–PLI model. The appended posterior plots in Figures 7 and 8 reveal that for $N = 2$, all methods are widely spread over the prior region, yet converge to the ground truth. However, APT cannot recover the ground truth for $N = 100$, whereas MMD-PLI and MMD-ABC center around the ground truth and depicting correlations between the individual parameters.

## 5. Related Work

In the previous sections, we have seen that PLI is algorithmically similar to ABC methods with SMC samplers. Therefore, approximating the likelihood by the empirical pseudo-likelihood (7) enables drawing from the rich toolbox of existing approximate inference algorithms. Here, we show related research fields and how PLI fits among them.

**Sequential Neural Density Estimation.** With the enriched class of neural density estimators, amortized SBI methods have received increasing interest in recent years. Similar to ABC, synthetic samples from the simulator are used to approximate the posterior. Sequential neural density estimation methods can be further classified into methods that directly train a posterior estimator (Papamakarios & Murray, 2016; Greenberg et al., 2019), a neural likelihood (Papamakarios et al., 2019; Glaser et al., 2022), or a neural ratio estimator (Durkan et al., 2020; Miller et al., 2022). All methods have in common that they do not rely on an approximation of the posterior model, but are optimized solely on pairs of parameter samples from a proposal distribution $\boldsymbol{\xi}^{(k)} \sim p_t(\boldsymbol{\xi})$ and its corresponding simulation $\mathbf{x}^{(k)} \sim p(\mathbf{x}|\boldsymbol{\xi}^{(k)})$. We note that the original papers have only reported posteriors conditioned on a single observation $\mathbf{x}$. While technically these methods can incorporate multiple data points, this requires either stacking multiple observations or falling back to summary statistics. As noted in (Tejero-Cantero et al., 2020),[2] neural likelihood estimators can sidestep these requirements by evaluating the log-likelihood of single observations and carrying out MCMC sampling on the joint log-likelihood. Yet, leveraging the neural likelihood restricts the evaluation of the posterior.

**Summary Statistics.** ABC has commonly relied on reducing the dimensionality of the raw observations with summary statistics (Sisson et al., 2018). These summaries must be carefully chosen and are often task-specific, restricting the general applicability of ABC. Recent additions to ABC methods report on replacing summary statistics with the use of statistical distances (Drovandi & Frazier, 2022). While direct comparison of the raw data suffers from the curse of dimensionality, comparing the observations by means of empirical measures sidesteps this issue (Drovandi & Frazier, 2022). Bernton et al. (2019) report on augmenting the likelihood kernel with the Wasserstein distance, while Park et al. (2016) leverages the kernelized approximation of the MMD (Gretton et al., 2012). Other contributions include the Cramér-von-Mises distance (Frazier, 2020) and the energy distance (Nguyen et al., 2020). While statistical distances are appealing due to their general applicability, Drovandi & Frazier (2022) conclude that they are limited by their high computational requirements. Several other

[2] https://www.mackelab.org/sbi/tutorial/14_multi-trial-data-and-mixed-data-types

approaches have been proposed to design summaries automatically. ABC with indirect inference (Gleim & Pigorsch, 2013; Drovandi et al., 2015) leverage an auxiliary model to assess the summaries of the data.

**Particle Mirror Descent.** A similar posterior updating rule to our Equation (6) has been derived in Particle Mirror Descent (PMD) (Dai et al., 2016). PMD tackles particle depletion by incorporating the proposal distribution of the previous round into the optimization process. Furthermore, the authors show that the proposed method converges to the posterior given $m$ posterior samples by $\mathcal{O}(1/\sqrt{m})$. Our version can be seen as extending their approach to the case of intractable likelihoods. We extend PMD to neural density estimators by leveraging the samples from the proposal posterior (6) as a training set.

**Geometric Path and Likelihood Tempering.** Rewriting the optimal posterior (6) reveals a close relation of the optimal PLI posterior (3) and the geometric path formulation (Chopin et al., 2020, p. 335)

$$q^\star \propto p_t^{1-\lambda}(\boldsymbol{\xi})\, p(\mathcal{X}^\star, \boldsymbol{\xi})^\lambda. \tag{11}$$

The optimal posterior moves from the proposal distribution $p_t$ at time $t$ to the desired posterior $p(\boldsymbol{\xi}|\mathcal{X}^\star) \propto p(\mathcal{X}^\star, \boldsymbol{\xi})$ along the geometric path that is parameterized by $\lambda$. The formulation differentiates from likelihood tempering in SMC samplers (Chopin et al., 2020) by leveraging the proposal instead of the prior distribution. Note, however, that for $t = 0$, the proposal mimics the prior, and thus the PLI geometric path has the same boundary values as in classical likelihood tempering. While the above formulation (11) cannot be applied to SMC samplers due to its dependence on the proposal distribution, the geometric path formulation based on the prior distribution gives rise to sequential annealing ABC (Albert et al., 2015).

**Pseudo-Likelihood Methods.** Our inference algorithm derives its name from prior literature on pseudo-likelihood methods in statistics (Gong & Samaniego, 1981; Hall, 1990; Guolo, 2011). The basic idea behind pseudo-likelihood methods is to replace the true likelihood function with a simpler function that can be easily evaluated. In prior works, the emphasis was on analytical tractability. Therefore, the pseudo-likelihood was constructed by assuming that certain parts of the model are independent given the other parts and then approximating the true likelihood function by the product of the individual likelihoods of these parts (Guolo, 2011). We expand the scope of PLI by allowing more general neural function approximations of the likelihood function.

**General Variational Inference.** Knoblauch et al. (2022) proposed the Rule of Three (RoT), which decomposes a wide variety of Bayesian inference methods into their representation of the objective/loss, the discrepancy measure between the posterior and the prior, and the search space of probability distributions that the optimization routine has access to. As has been shown in the previous section, ABC, including the PLI methodology, can be integrated into the RoT paradigm as $P(K_\beta(\mathcal{X}^\star, \mathcal{X}), \mathrm{KL}, \mathcal{P}(\Theta))$ (following the notation of Knoblauch et al. (2022)). Most crucially, ABC methods are characterized through their loss based on the likelihood kernel. Knoblauch et al. (2022) states that specifying a loss function instead of choosing the log-likelihood as a loss addresses model misspecification. Probably Approximately Correct (PAC)-Bayes (Guedj, 2019) can be seen as a generalization of ABC as it covers the whole space of possible loss functions $l(\mathcal{X}^\star, \boldsymbol{\xi})$. The PAC-Bayesian theory provides a broad array of risk bounds for generalized Bayesian learning methods.

## 6. Conclusion

This paper proposes Pseudo-Likelihood Inference (PLI), a new addition to the toolbox of SBI methods that is targeted for Bayesian inference tasks in which the posterior is conditioned on multiple observations. For that, we derive a softened ABC posterior from a constrained variational inference problem and leverage IPMs between the empirical observations to assess the intractable likelihood. The derived posterior formulation enables the learning of flexible neural density estimators from black-box simulators, thereby extending the range of applicability for ABC methods.

The goal of our experiments is to assess how well PLI, ABC, and SNPE perform in terms of prediction when given varying amounts of data to condition on. In situations where there is limited data and uncertainty, SNPE-based methods perform better than ABC and PLI, which rely on statistical distances. However, when there is more data available, ABC and PLI methods perform better. When the posterior distribution is straightforward, MMD-ABC is efficient at reproducing it using fast particle updates. However, for complex posterior distributions, MMD-PLI is a better option because it has a more adaptable neural density estimator. Additionally, PLI can evaluate the posterior probability, which can be useful for later tasks that require quantifying uncertainty. Throughout our experiments, we discovered that PLI and ABC with the Wasserstein distance performed significantly worse than MMD, as supported by previous research. As a result, we suggest using MMD over the Wasserstein distance in general for SBI. While our results were obtained on SBI benchmark tasks as well as time series data, the presented method can be easily extended to different types of data such as images.

Encouraged by the findings on the Furuta pendulum in simulations, we intend to expand the study to real-world data to assess the usefulness of SBI methods further. We are particularly interested in how well SBI methods handle model mismatch, as this is a common issue when working with real-world data.

## References


Albert, C., Künsch, H. R., and Scheidegger, A. A simulated annealing approach to approximate bayes computations. *Statistics and Computing*, 25(6):1217–1232, 2015.

Beaumont, M. A. Approximate bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution, and Systematics*, 41(1):379–406, 2010.

Bernton, E., Jacob, P. E., Gerber, M., and Robert, C. P. Approximate bayesian computation with the wasserstein distance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(2):235–269, 2019.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., and Zhang, Q. JAX: composable transformations of Python+NumPy programs, 2018.

Chopin, N., Papaspiliopoulos, O., et al. *An introduction to sequential Monte Carlo*, volume 4. Springer, 2020.

Cranmer, K., Brehmer, J., and Louppe, G. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, 2013.

Dai, B., He, N., Dai, H., and Song, L. Provable bayesian inference via particle mirror descent. In *Artificial Intelligence and Statistics*, 2016.

Daniel, C., Neumann, G., Kroemer, O., and Peters, J. Hierarchical relative entropy policy search. *Journal of Machine Learning Research*, 17:93:1–93:50, 2016.

Deisenroth, M. P., Neumann, G., Peters, J., et al. A survey on policy search for robotics. *Foundations and Trends in Robotics*, 2(1–2):1–142, 2013.

Del Moral, P., Doucet, A., and Jasra, A. Sequential monte carlo samplers. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 68(3):411–436, 2006.

Del Moral, P., Doucet, A., and Jasra, A. An adaptive sequential monte carlo method for approximate bayesian computation. *Statistics and Computing*, 22(5):1009–1020, 2012.

Dellaporta, C., Knoblauch, J., Damoulas, T., and Briol, F. Robust bayesian inference for simulator-based models via the MMD posterior bootstrap. In *Artificial Intelligence and Statistics*, 2022.

Drovandi, C. C. and Frazier, D. T. A comparison of likelihood-free methods with and without summary statistics. *Statistics and Computing*, 32(3):42, 2022.

Drovandi, C. C., Pettitt, A. N., and Lee, A. Bayesian indirect inference using a parametric auxiliary model. *Statistical Science*, 30(1):72–95, 2015.

Durkan, C., Murray, I., and Papamakarios, G. On contrastive learning for likelihood-free inference. In *International Conference on Machine Learning*, 2020.

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

Furuta, K., Yamakita, M., and Kobayashi, S. Swing-up control of inverted pendulum using pseudo-state feedback. *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*, 206 (4):263–269, 1992.

Glaser, P., Arbel, M., Doucet, A., and Gretton, A. Maximum likelihood learning of energy-based models for simulation-based inference. *arXiv preprint arXiv:2210.14756*, 2022.

Gleim, A. and Pigorsch, C. Approximate bayesian computation with indirect summary statistics. *Draft paper: http://ect-pigorsch. mee. uni-bonn. de/data/research/papers*, 2013.

Gong, G. and Samaniego, F. J. Pseudo maximum likelihood estimation: theory and applications. *The Annals of Statistics*, pp. 861–869, 1981.

Greenberg, D. S., Nonnenmacher, M., and Macke, J. H. Automatic posterior transformation for likelihood-free inference. In *International Conference on Machine Learning*, 2019.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *Journal of Machine Learning Research*, 13(1):723–773, 2012.

Guedj, B. A primer on pac-bayesian learning. *arXiv preprint arXiv:1901.05353*, 2019.

Guolo, A. Pseudo-likelihood inference for regression models with misclassified and mismeasured variables. *Statistica Sinica*, pp. 1639–1663, 2011.

Hall, P. Pseudo-likelihood theory for empirical likelihood. *The Annals of Statistics*, pp. 121–140, 1990.

Hartig, F., Calabrese, J. M., Reineking, B., Wiegand, T., and Huth, A. Statistical inference for stochastic simulation models–theory and application. *Ecology letters*, 14(8): 816–827, 2011.

Karabatsos, G. and Leisen, F. An approximate likelihood perspective on ABC methods. *Statistics Surveys*, 12:66 – 104, 2018.

Knoblauch, J., Jewson, J., and Damoulas, T. An optimization-centric view on bayes' rule: Reviewing and generalizing variational inference. *Journal of Machine Learning Research*, 23(132):1–109, 2022.

Lee, A. On the choice of MCMC kernels for approximate bayesian computation with SMC samplers. In *Winter Simulation Conference*, pp. 27:1–27:12, 2012.

Lenormand, M., Jabot, F., and Deffuant, G. Adaptive approximate bayesian computation for complex models. *Computational Statistics*, 28(6):2777–2796, 2013.

Lueckmann, J., Boelts, J., Greenberg, D. S., Gonçalves, P. J., and Macke, J. H. Benchmarking simulation-based inference. In *Artificial Intelligence and Statistics*, 2021.

Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. Markov chain monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328, 2003.

McGoff, K., Mukherjee, S., and Pillai, N. Statistical inference for dynamical systems: A review. *Statistics Surveys*, 9:209–252, 2015.

Miller, B. K., Weniger, C., and Forré, P. Contrastive neural ratio estimation. *arXiv preprint arXiv:2210.06170*, 2022.

Muratore, F., Gruner, T., Wiese, F., Belousov, B., Gienger, M., and Peters, J. Neural posterior domain randomization. In *Conference on Robot Learning*, 2021.

Muratore, F., Ramos, F., Turk, G., Yu, W., Gienger, M., and Peters, J. Robot learning from randomized simulations: A review. *Frontiers in Robotics and AI*, 9, 2022.

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

Papamakarios, G. and Murray, I. Fast $\epsilon$-free inference of simulation models with bayesian conditional density estimation. In *Advances in Neural Information Processing Systems*, 2016.

Papamakarios, G., Sterratt, D. C., and Murray, I. Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. In *Artificial Intelligence and Statistics*, 2019.

Park, M., Jitkrittum, W., and Sejdinovic, D. K2-ABC: approximate bayesian computation with kernel embeddings. In *Artificial Intelligence and Statistics*, 2016.

Sisson, S. A., Fan, Y., and Tanaka, M. M. Sequential monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 104(6):1760–1765, 2007.

Sisson, S. A., Fan, Y., and Beaumont, M. *Handbook of approximate Bayesian computation (1st ed.)*. CRC Press, 2018.

Tavaré, S., Balding, D. J., Griffiths, R. C., and Donnelly, P. Inferring Coalescence Times From DNA Sequence Data. *Genetics*, 145(2):505–518, 1997.

Tejero-Cantero, A., Boelts, J., Deistler, M., Lueckmann, J.-M., Durkan, C., Gonçalves, P. J., Greenberg, D. S., and Macke, J. H. sbi: A toolkit for simulation-based inference. *Journal of Open Source Software*, 5(52):2505, 2020.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, M. P. Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of The Royal Society Interface*, 6(31): 187–202, 2009.

# A. Algorithmic Details: Pseudo-Likelihood Inference (PLI)

## A.1. Deriving the Optimal PLI Parameter Distribution

The stochastic search problem (5) is restated here for readability

$$\begin{aligned}\tilde{p}_t = \min_{p_t} \quad & D_{\mathrm{KL}}\left(p_t(\mathcal{X},\boldsymbol{\xi}) \,||\, p(\mathcal{X},\boldsymbol{\xi}|\mathcal{X}^\star)\right) = \min_{p_t} \quad \mathop{\mathbb{E}}_{p_t(\mathcal{X},\boldsymbol{\xi})}\left[-\log p(\mathcal{X}^\star|\mathcal{X},\boldsymbol{\xi})\right] + D_{\mathrm{KL}}\left(p_t(\mathcal{X},\boldsymbol{\xi}) \,||\, p(\mathcal{X},\boldsymbol{\xi})\right),\\ \text{s.t.} \quad & D_{\mathrm{KL}}\left(p_t(\mathcal{X},\boldsymbol{\xi}) \,||\, \tilde{p}_{t-1}(\mathcal{X},\boldsymbol{\xi})\right) \leq \varepsilon,\\ & \iint p_t(\mathcal{X},\boldsymbol{\xi})\,\mathrm{d}\mathcal{X}\,\mathrm{d}\boldsymbol{\xi} = 1.\end{aligned}$$

For simplicity, we will write $p_t(\mathcal{X},\boldsymbol{\xi}) := p_t(\mathcal{X},\boldsymbol{\xi}|\mathcal{X}^\star)$ for the joint distribution without explicitly stating its dependence on the observed data $\mathcal{X}^\star$. Instead of optimizing the VI objective directly, we add a tempering constant $\nu$ to the KL between the joint distributions $p_t(\mathcal{X},\boldsymbol{\xi})$ and $p(\mathcal{X},\boldsymbol{\xi})$

$$\min_{p_t} \quad \mathop{\mathbb{E}}_{p_t(\mathcal{X},\boldsymbol{\xi})}\left[-\log p(\mathcal{X}^\star|\mathcal{X},\boldsymbol{\xi})\right] + \nu D_{\mathrm{KL}}\left(p_t(\mathcal{X},\boldsymbol{\xi}) \,||\, p(\mathcal{X},\boldsymbol{\xi})\right).$$

For $\nu = 1$, the VI objective is recovered, while $\nu = 0$ resorts to minimizing the distance between simulated and reference data. The constrained optimization problem (5) can be reformulated with Lagrange multipliers as

$$\begin{aligned}\mathcal{L} = & \iint -p_t(\mathcal{X},\boldsymbol{\xi}) \log p(\mathcal{X}^\star|\mathcal{X},\boldsymbol{\xi})\,\mathrm{d}\mathcal{X}\mathrm{d}\boldsymbol{\xi}\\ & + \nu \iint p_t(\mathcal{X},\boldsymbol{\xi}) \log \frac{p_t(\mathcal{X},\boldsymbol{\xi})}{p(\mathcal{X},\boldsymbol{\xi})}\,\mathrm{d}\mathcal{X}\mathrm{d}\boldsymbol{\xi}\\ & + \eta\left(\iint p_t(\mathcal{X},\boldsymbol{\xi}) \log \frac{p_t(\mathcal{X},\boldsymbol{\xi})}{\tilde{p}_{t-1}(\mathcal{X},\boldsymbol{\xi})}\,\mathrm{d}\mathcal{X}\mathrm{d}\boldsymbol{\xi} - \varepsilon\right)\\ & + \lambda\left(\iint p_t(\mathcal{X},\boldsymbol{\xi})\,\mathrm{d}\mathcal{X}\mathrm{d}\boldsymbol{\xi} - 1\right)\\ = & \iint p_t(\mathcal{X},\boldsymbol{\xi})\left[-\log p(\mathcal{X}^\star|\mathcal{X},\boldsymbol{\xi}) + \nu \log \frac{p_t(\mathcal{X},\boldsymbol{\xi})}{p(\mathcal{X},\boldsymbol{\xi})} + \eta \log \frac{p_t(\mathcal{X},\boldsymbol{\xi})}{\tilde{p}_{t-1}(\mathcal{X},\boldsymbol{\xi})} + \lambda\right]\,\mathrm{d}\mathcal{X}\mathrm{d}\boldsymbol{\xi} - \eta\varepsilon - \lambda\\ = & \int p_t(\boldsymbol{\xi})\left[-\int p(\mathcal{X}|\boldsymbol{\xi}) \log p(\mathcal{X}^\star|\mathcal{X},\boldsymbol{\xi})\mathrm{d}\mathcal{X} + \nu \log \frac{p_t(\boldsymbol{\xi})}{p(\boldsymbol{\xi})} + \eta \log \frac{p_t(\boldsymbol{\xi})}{\tilde{p}_{t-1}(\boldsymbol{\xi})} + \lambda\right]\mathrm{d}\boldsymbol{\xi} - \eta\varepsilon - \lambda \qquad (12)\end{aligned}$$

Here we leveraged the assumption that the likelihood $p(\mathcal{X}|\boldsymbol{\xi})$ is fixed for all joint distributions, and thus the joint distributions can be split into the likelihood $p(\mathcal{X}^\star|\boldsymbol{\xi})$ and their associated prior/proposal distributions. The gradient of the Lagrangian vanishes for the optimal parameter distribution

$$\left.\frac{\partial \mathcal{L}}{\partial p_t}\right|_{p_t=\tilde{p}_t} = -\int p(\mathcal{X}|\boldsymbol{\xi}) \log p(\mathcal{X}^\star|\mathcal{X},\boldsymbol{\xi})\mathrm{d}\mathcal{X} + \nu\left[\log \frac{\tilde{p}_t(\boldsymbol{\xi})}{p(\boldsymbol{\xi})} + 1\right] + \eta\left[\log \frac{\tilde{p}_t(\boldsymbol{\xi})}{\tilde{p}_{t-1}(\boldsymbol{\xi})} + 1\right] + \lambda = 0$$

Reformulation yields

$$\begin{aligned}\tilde{p}_t(\boldsymbol{\xi}) &= p^{\frac{\nu}{\nu+\eta}}(\boldsymbol{\xi})\,\tilde{p}_{t-1}^{\frac{\eta}{\nu+\eta}}(\boldsymbol{\xi})\,\exp\left(\frac{1}{\nu+\eta}\int p(\mathcal{X}|\boldsymbol{\xi}) \log p(\mathcal{X}^\star|\mathcal{X},\boldsymbol{\xi})\mathrm{d}\mathcal{X} - \frac{\nu+\eta+\lambda}{\nu+\eta}\right)\\ &\propto \left(\frac{p(\boldsymbol{\xi})}{\tilde{p}_{t-1}(\boldsymbol{\xi})}\right)^{\frac{\nu}{\nu+\eta}} \exp\left(\frac{1}{\nu+\eta}\int p(\mathcal{X}|\boldsymbol{\xi}) \log p(\mathcal{X}^\star|\mathcal{X},\boldsymbol{\xi})\mathrm{d}\mathcal{X}\right)\tilde{p}_{t-1}(\boldsymbol{\xi}) \qquad (13)\end{aligned}$$

We denote the proposal likelihood associated with the proposal prior $\tilde{p}_{t-1}(\boldsymbol{\xi})$ by $\tilde{p}_{t-1}(\mathcal{X}^\star|\boldsymbol{\xi})$. The normalization constant $Z = \exp((\nu+\eta+\lambda)/(\nu+\eta))$ can then be written as

$$\psi(\eta) := \frac{\nu+\eta+\lambda}{\nu+\eta} = \log \mathop{\mathbb{E}}_{\tilde{p}_{t-1}(\boldsymbol{\xi})}\left[\tilde{p}_{t-1}(\mathcal{X}^\star|\boldsymbol{\xi})\right].$$

The dual associated with the Lagrangian can be obtained by inserting (13) into the Lagrangian (12)

$$g(\eta) = -\eta\varepsilon - (\nu+\eta)\psi(\eta).$$

### A.2. Convergence of the PLI Posterior

Here we show the convergence of the PLI posterior (6) to the true posterior in two steps. First, we show that the joint optimization problem is a lower bound to the marginal posterior optimization problem in terms of the KL divergence. Then we employ the fact that the posterior derived from the marginal posterior optimizatoin problem approximates the true posterior in the limit $\eta_t \to 0$. To demonstrate the convergence to the PLI posterior, we derive the following lower bound on the constrained optimization problem (5),

$$\begin{aligned}
D_{\mathrm{KL}}\left(p_t(\mathcal{X}, \boldsymbol{\xi}) \,||\, p(\mathcal{X}, \boldsymbol{\xi}|\mathcal{X}^\star)\right) &= -\mathop{\mathbb{E}}_{p_t(\mathcal{X}, \boldsymbol{\xi})}\left[\log p(\mathcal{X}^\star|\mathcal{X}, \boldsymbol{\xi})\right] + D_{\mathrm{KL}}\left(p_t(\mathcal{X}, \boldsymbol{\xi}) \,||\, p(\mathcal{X}, \boldsymbol{\xi})\right) \\
&= -\mathop{\mathbb{E}}_{p_t(\boldsymbol{\xi})}\mathop{\mathbb{E}}_{p(\mathcal{X}|\boldsymbol{\xi})}\left[\log p(\mathcal{X}^\star|\mathcal{X}, \boldsymbol{\xi})\right] + D_{\mathrm{KL}}\left(p_t(\boldsymbol{\xi}) \,||\, p(\boldsymbol{\xi})\right) \\
&\geq -\mathop{\mathbb{E}}_{p_t(\boldsymbol{\xi})}\left[\log \mathop{\mathbb{E}}_{p(\mathcal{X}|\boldsymbol{\xi})}\left[p(\mathcal{X}^\star|\mathcal{X}, \boldsymbol{\xi})\right]\right] + D_{\mathrm{KL}}\left(p_t(\boldsymbol{\xi}) \,||\, p(\boldsymbol{\xi})\right) \\
&= -\mathop{\mathbb{E}}_{p_t(\boldsymbol{\xi})}\left[\log p(\mathcal{X}^\star|\boldsymbol{\xi})\right] + D_{\mathrm{KL}}\left(p_t(\boldsymbol{\xi}) \,||\, p(\boldsymbol{\xi})\right) = D_{\mathrm{KL}}\left(p_t(\boldsymbol{\xi}) \,||\, p(\boldsymbol{\xi}|\mathcal{X}^\star)\right).
\end{aligned}$$

In the third line, we use Jensen's inequality to pull the logarithm out of the inner expectation. The inner expectation represents a marginalization over $\mathcal{X}$, thus rendering that term as the likelihood of the observed data $p(\mathcal{X}^\star|\boldsymbol{\xi})$. The above formulation means that our optimization problem is an upper bound on the original inference task, i.e., minimizing the KL divergence from the posterior distribution. The optimal posterior reads as follows:

$$\tilde{p}_t(\boldsymbol{\xi}) \propto \left(\frac{p(\boldsymbol{\xi})}{\tilde{p}_{t-1}(\boldsymbol{\xi})}\right)^{\frac{1}{1+\eta_t}} \exp\left(\frac{\log p(\mathcal{X}^\star|\boldsymbol{\xi})}{1+\eta_t}\right) \tilde{p}_{t-1}(\boldsymbol{\xi}).$$

When we assume that $\eta_t(t)$ is a monotonically decreasing function that goes to 0 for $t \to \infty$, we directly see that the target posterior of the lower bound converges to the true posterior distribution,

$$\begin{aligned}
\lim_{t\to\infty} \tilde{p}_t(\boldsymbol{\xi}) &\propto \lim_{\eta_t \to 0} \left(\frac{p(\boldsymbol{\xi})}{\tilde{p}_{t-1}(\boldsymbol{\xi})}\right)^{\frac{1}{1+\eta_t}} \exp\left(\frac{\log p(\mathcal{X}^\star|\boldsymbol{\xi})}{1+\eta_t}\right) \tilde{p}_{t-1}(\boldsymbol{\xi}) \\
&= \exp\left(\log p(\mathcal{X}^\star|\boldsymbol{\xi})\right) p(\boldsymbol{\xi}) \\
&= p(\mathcal{X}^\star|\boldsymbol{\xi}) p(\boldsymbol{\xi}).
\end{aligned}$$

From here it follows that the minimization of the KL between the joint distributions leads to the minimization of the KL between the marginals, and thus, the posterior is correctly approximated when the KL of the joint problem approaches 0.

### A.3. Weighted Maximum Likelihood Optimization (M-projection)

The M-projection of the optimal posterior onto the approximation family results in a weighted maximum likelihood formulation

$$\begin{aligned}
\min_{\boldsymbol{\phi}} \; D_{\mathrm{KL}}\left(\tilde{p}_t(\boldsymbol{\xi}) \,||\, q_{\boldsymbol{\phi}}(\boldsymbol{\xi})\right) &= \max_{\boldsymbol{\phi}} \mathop{\mathbb{E}}_{\tilde{p}_t(\boldsymbol{\xi})}\left[\log q_{\boldsymbol{\phi}}(\boldsymbol{\xi})\right] \\
&= \max_{\boldsymbol{\phi}} \int \log q_{\boldsymbol{\phi}}(\boldsymbol{\xi}) \, \tilde{p}_t(\boldsymbol{\xi}) \, \mathrm{d}\boldsymbol{\xi} \\
&= \max_{\boldsymbol{\phi}} \int \frac{1}{Z} \left(\frac{p(\boldsymbol{\xi})}{\tilde{p}_{t-1}(\boldsymbol{\xi})}\right)^{\nu/(\nu+\eta)} p_{\beta_t}(\mathcal{X}^\star|\boldsymbol{\xi}) \, \tilde{p}_{t-1}(\boldsymbol{\xi}) \log q_{\boldsymbol{\phi}}(\boldsymbol{\xi}) \, \mathrm{d}\boldsymbol{\xi} \\
&= \max_{\boldsymbol{\phi}} \mathop{\mathbb{E}}_{\tilde{p}_{t-1}(\boldsymbol{\xi})} \left[\underbrace{\left(\frac{p(\boldsymbol{\xi})}{p_{t-1}(\boldsymbol{\xi})}\right)^{\nu/(\nu+\eta)} p_{\beta_t}(\mathcal{X}^\star|\boldsymbol{\xi})}_{w} \log q_{\boldsymbol{\phi}}(\boldsymbol{\xi})\right].
\end{aligned}$$

The weighting term $w$ is independent of $\boldsymbol{\phi}$, and as such, the formulation facilitates optimizing neural density estimators with gradient descent. This optimization resolves to weighted maximum likelihood where the weights are obtained from the pseudo-likelihood (7). The weighted maximum likelihood formulation can be optimized in closed form for linear Gaussian models (Deisenroth et al., 2013), with Expectation-Maximization (EM) using Gaussian Mixture Models (GMMs) (Daniel et al., 2016) or with gradient descent as we do it in this paper.

### A.4. Optimizing with the I-projection

We can reformulate the minimization problem for the I-projection in the following way

$$\min_{\boldsymbol{\phi}} D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\boldsymbol{\xi}) \,||\, \tilde{p}_t(\boldsymbol{\xi})\right) = \min_{\boldsymbol{\phi}} \mathop{\mathbb{E}}_{q_{\boldsymbol{\phi}}(\boldsymbol{\xi})} \left[ \log \frac{q_{\boldsymbol{\phi}}(\boldsymbol{\xi})}{Z_{\boldsymbol{\phi}}^{-1} \left(\frac{p(\boldsymbol{\xi})}{p_{t-1}(\boldsymbol{\xi})}\right)^{\nu/(\nu+\eta)} p_{\beta_t}(\mathcal{X}^\star|\boldsymbol{\xi})\, \tilde{p}_{t-1}(\boldsymbol{\xi})} \right] \tag{14}$$

$$= \min_{\boldsymbol{\phi}} D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\boldsymbol{\xi}) \,||\, \tilde{p}_{t-1}(\boldsymbol{\xi})\right) - \mathop{\mathbb{E}}_{\tilde{p}_{t-1}(\boldsymbol{\xi})} \left[ \frac{q_{\boldsymbol{\phi}}(\boldsymbol{\xi})}{\tilde{p}_{t-1}(\boldsymbol{\xi})} \log \left( \frac{\left(\frac{p(\boldsymbol{\xi})}{p_{t-1}(\boldsymbol{\xi})}\right)^{\nu/(\nu+\eta)} p_{\beta_t}(\mathcal{X}^\star|\boldsymbol{\xi})}{Z_{\boldsymbol{\phi}}} \right) \right].$$

The equation above alleviates the issue of back-propagating through the simulator by using importance sampling. The optimization problem can be understood as fitting the posterior estimator $q$ to the proposal $p(\boldsymbol{\xi})$ while having a regularization term that forces the distribution to fit the reference data. The temperature parameter $\eta$ can thus be interpreted as weighting the regularization term. Small values of $\eta$ put more emphasis on the regularization term while large values concentrate on containing the information of the proposal.

For linear Gaussian models, a closed form solution of the above KL exists, thus (14) can be optimized with non-linear optimizers. The optimal temperature parameter $\eta^\star$ can be obtained by maximizing the dual using a non-linear optimization solver.

## B. Algorithmic Details: Sequential Monte Carlo ABC

The foundations of SMC-ABC have been laid by Del Moral et al. (2006), who introduced SMC samplers. SMC samplers describe an approximate inference routine in which the posterior is approximated through a sequence of intermediate target posteriors. In the context of ABC, the sequence of intermediate posteriors is defined by an adaptive bandwidth $\beta_t$ of the approximate posterior $p_{\beta_t}(\boldsymbol{\xi}|\mathcal{X}^\star)$ (3). Additionally, the sample efficiency of ABC is improved by replacing the prior as the sampling distribution with a proposal distribution $\tilde{p}_t(\boldsymbol{\xi})$. The proposal distribution is represented by a set of particles $\tilde{p}_t(\boldsymbol{\xi}) = \sum_{i=1}^M \delta_{\boldsymbol{\xi}_t^{(i)}}(\boldsymbol{\xi})$ and through importance sampling an approximation of the target posterior $p_{\beta_t}(\boldsymbol{\xi}|\mathcal{X}^\star)$ can be obtained

$$q_t(\boldsymbol{\xi}) = \sum_{i=1}^M W_t^{(i)} \delta_{\boldsymbol{\xi}_t^{(i)}}(\boldsymbol{\xi}); \quad W_t^{(i)} = \frac{p_{\beta_t}(\boldsymbol{\xi}_t^{(i)}|\mathcal{X}^\star)}{\tilde{p}_t(\boldsymbol{\xi}^{(i)})}, \tag{15}$$

where $W_t^{(i)}$ denote the importance weights. The proposal distribution $\tilde{p}_t(\boldsymbol{\xi})$ should ideally stay close to the target posterior $p_{\beta_t}(\boldsymbol{\xi}|\mathcal{X}^\star)$ to improve the sample efficiency. Therefore, the proposal distribution is updated based on a Markov kernel $K_{t+1}(\boldsymbol{\xi}_t, \boldsymbol{\xi}_{t+1})$ which is the transition probability from $\boldsymbol{\xi}_t$ to $\boldsymbol{\xi}_{t+1}$. The update of the proposal distribution is typically numerically intractable as it requires marginalization, i.e., integration over $\boldsymbol{\xi}_t$ for each inference step $0:t$

$$\tilde{p}_{t+1}(\boldsymbol{\xi}_{t+1}) = \int K_{t+1}(\boldsymbol{\xi}_t, \boldsymbol{\xi}_{t+1}) \tilde{p}_t(\boldsymbol{\xi}_t) \mathrm{d}\boldsymbol{\xi}_t.$$

Table 1: Kernels used for the ABC and PLI variants during the experiments in Section 4. $D$ denotes an IPM which can been reagarded as a distance measure between the reference data $\mathcal{X}^\star$ and samples $\mathcal{X}$. $\beta$ is the kernel bandwith parameter introduced in Section 3.1. The Effective Sample Size (ESS) is defined as the inverse of the weights' variance.

| $K_\beta(D(\mathcal{X}, \mathcal{X}^\star))$ | Algorithm | $\beta_t$ estimation | Update |
|---|---|---|---|
| $\mathbb{1}_{\{D(\mathcal{X},\mathcal{X}^\star)\leq\beta\}}$ | SMC ABC | ESS | $\beta_t^\star = \arg\min_{\beta_t} \mathrm{ESS}(w_t, \beta_t) - \alpha \mathrm{ESS}(w_{t-1}, \beta_{t-1})$ |
| | Adaptive PMC ABC | $\alpha$-Quantile | $\beta_t^\star = Q_{D(\mathcal{X}_t,\mathcal{X}^\star)}(\alpha)$ |
| $\exp\left(-\frac{D(\mathcal{X},\mathcal{X}^\star)}{2\beta}\right)$ | PLI | Trust-region | $\beta_t^\star = \arg\min_\beta 2\beta\left(\varepsilon + \log \mathbb{E}_{\tilde{p}_t(\boldsymbol{\xi})}\left[p_\beta(\mathcal{X}^\star|\boldsymbol{\xi})\right]\right)$ |

To alleviate the computational burden, Del Moral et al. (2006) show that the joint representation of the proposal distribution $p_t(\boldsymbol{\xi}_{0:t})$ can be efficiently calculated as it only requires solving the product over $t$ transitions

$$\tilde{p}_t(\boldsymbol{\xi}_{0:t}) = \tilde{p}_0(\boldsymbol{\xi}_0) \prod_{\tau=0}^{t} K_{t+1}(\boldsymbol{\xi}_t, \boldsymbol{\xi}_{t+1}).$$

We define the joint proposal distribution as the empirical distribution $p_t(\boldsymbol{\xi}_{0:t}) = \sum_{i=1}^{M} \delta_{\boldsymbol{\xi}_{0:t}^{(i)}}(\boldsymbol{\xi}_{0:t})$ defined by a set of joint particles $\boldsymbol{\xi}_{0:t}^{(i)}$. Thus, the joint posterior approximation of $p_{\beta_t}(\boldsymbol{\xi}_{0:t}|\mathcal{X}^\star)$ based on the importance weights reads as

$$\begin{aligned} q_t(\boldsymbol{\xi}_{0:t}) &= \sum_{i=1}^{M} w_t^{(i)} \delta_{\boldsymbol{\xi}_{0:t}^{(i)}}(\boldsymbol{\xi}_{0:t}); \quad w_t^{(i)} = \frac{p_{\beta_t}(\boldsymbol{\xi}_{0:t}^{(i)}|\mathcal{X}^\star)}{\tilde{p}_t(\boldsymbol{\xi}_{0:t}^{(i)})} \\ \Rightarrow \quad q_t(\boldsymbol{\xi}_t) &= \sum_{i=1}^{M} w_t^{(i)} \delta_{\boldsymbol{\xi}_t^{(i)}}(\boldsymbol{\xi}_t); \quad w_t^{(i)} = \frac{p_{\beta_t}(\boldsymbol{\xi}_{0:t}^{(i)}|\mathcal{X}^\star)}{\tilde{p}_t(\boldsymbol{\xi}_{0:t}^{(i)})}. \end{aligned}$$

The marginal target posterior approximation $q_t(\boldsymbol{\xi}_t)$ can be directly recovered from the joint approximation $q_t(\boldsymbol{\xi}_{0:t})$. Furthermore, both distributions share their weights which means that it is only required to estimate the weights $w_t$ in order to approximate the target posteriors $p_{\beta_t}(\boldsymbol{\xi}_t|\mathcal{X}^\star)$. In general, the probability of the target joint posterior $p_{\beta_t}(\boldsymbol{\xi}_{0:t}^{(i)}|\boldsymbol{\xi})$ is intractable. Therefore, the authors introduce an auxiliary backward Markov kernel $L_t(\boldsymbol{\xi}_{t+1}, \boldsymbol{\xi}_t)$ to simplify the computation

$$p_{\beta_t}(\boldsymbol{\xi}_{0:t}|\mathcal{X}^\star) = p_{\beta_t}(\boldsymbol{\xi}_t|\mathcal{X}^\star) \prod_{\tau=1}^{t} L_\tau(\boldsymbol{\xi}_{\tau+1}, \boldsymbol{\xi}_\tau).$$

Assume now that a posterior approximation of the target posterior $p_{\beta_t}(\boldsymbol{\xi}_t|\mathcal{X}^\star)$ is available through the set of weighted particles $\{(w_t^{(i)}, \boldsymbol{\xi}_t^{(i)})\}$ and the particles of the proposal distribution $\tilde{p}_t(\boldsymbol{\xi}) = \sum_{i=1}^{M} \delta_{\boldsymbol{\xi}_t^{(i)}}(\boldsymbol{\xi})$ are updated based on a kernel transition $\boldsymbol{\xi}_{t+1}^{(i)} \sim K_{t+1}(\boldsymbol{\xi}_t^{(i)}, \boldsymbol{\xi}_{t+1}^{(i)})$, then the importance weights $w_{t+1}$ are updated based on the following recursion

$$\begin{aligned} w_{t+1} &= \frac{p_{\beta_{t+1}}(\boldsymbol{\xi}_{0:t+1})}{\tilde{p}_{t+1}(\boldsymbol{\xi}_{0:t+1})} = \frac{p_{\beta_{t+1}}(\boldsymbol{\xi}_{t+1}) \prod_{\tau=1}^{t} L_t(\boldsymbol{\xi}_{\tau+1}, \boldsymbol{\xi}_\tau)}{\tilde{p}_0(\boldsymbol{\xi}_0) \prod_{\tau=0}^{t} K_t(\boldsymbol{\xi}_\tau, \boldsymbol{\xi}_{\tau+1})} \\ &= \underbrace{\frac{p_{\beta_{t+1}}(\boldsymbol{\xi}_{t+1}) L_t(\boldsymbol{\xi}_{t+1}, \boldsymbol{\xi}_t)}{p_{\beta_t}(\boldsymbol{\xi}_t) K_{t+1}(\boldsymbol{\xi}_t, \boldsymbol{\xi}_{t+1})}}_{\hat{w}_{t+1}} \underbrace{\frac{p_{\beta_t}(\boldsymbol{\xi}_t) \prod_{\tau=1}^{t-1} L_t(\boldsymbol{\xi}_{\tau+1}, \boldsymbol{\xi}_\tau)}{\tilde{p}_0(\boldsymbol{\xi}_0) \prod_{\tau=0}^{t-1} K_t(\boldsymbol{\xi}_\tau, \boldsymbol{\xi}_{\tau+1})}}_{w_t}. \end{aligned}$$

Thus, the sequential update is performed by updating the current weights with the marginal weights $\hat{w}_{t+1}$. Up until now, the choice of the backward kernel has been neglected. As it is an auxiliary quantity, several approximations can be made to model the backward kernel. Del Moral et al. (2006) refer to the optimal backward kernel as the Markov kernel that minimizes the variance of the particles

$$L_t^{\text{opt}}(\boldsymbol{\xi}_{t+1}, \boldsymbol{\xi}_t) = \frac{\tilde{p}(\boldsymbol{\xi}_t) K_{t+1}(\boldsymbol{\xi}_t, \boldsymbol{\xi}_{t+1})}{\tilde{p}_{t+1}(\boldsymbol{\xi}_{t+1})}.$$

They further show that the optimal backward kernel recovers the marginal weights from (15)

$$w_{t+1}^{\text{opt}} = \frac{p_{\beta_{t+1}}(\boldsymbol{\xi}|\mathcal{X}^\star)}{\tilde{p}_{t+1}(\boldsymbol{\xi})}.$$

In general, the optimal backward kernel is numerically intractable and has lead to several other approximations which are summarized in Table 2. Depending on choice of approximation, a number of different SMC-ABC methods have evolved, namely the classical SMC-ABC approach by Del Moral et al. (2012), PMC-ABC (Toni et al., 2009; Beaumont, 2010; Lenormand et al., 2013), and the Metropolis-Hastings (MH)-ABC (Lee, 2012). Please refer to those references and Algorithm 2 and Algorithm 3 for implementation details of the approaches.

**Algorithm 2** Sequential Monte Carlo-ABC (Del Moral et al., 2012)

**input:** reference data $\mathcal{X}^\star$, prior $p(\boldsymbol{\xi})$, stochastic simulator $p(\mathbf{x}|\boldsymbol{\xi})$, IPM $D(\cdot,\cdot)$, max. iteration count $T$, resampling threshold $V$, forward kernel $K(\boldsymbol{\xi}_t, \boldsymbol{\xi}_{t+1})$
initialize particles $\boldsymbol{\xi}_0^{(k)} \sim p(\cdot)$, initialize particle weights $w_0^{(k)} = 1/K$
**for** $t$ in $1\!:\!T$ **do**
  **for** each $\boldsymbol{\xi}^{(k)}$ **do**
    simulate $\mathcal{X}_{t-1}^{(k)} = \{\mathbf{x}_n^{(k)} \sim p(\mathbf{x}|\boldsymbol{\xi}_{t-1}^{(k)})\}$
    compute the IPM $s_{t-1}^{(k)} = D(\mathcal{X}^\star, \mathcal{X}_{t-1}^{(k)})$
  **end for**
  update the bandwidth $\beta_t$ by solving

$$\mathrm{ESS}(\{w_t^{(k)}\}, \beta_t) = \alpha \mathrm{ESS}(\{w_{t-1}^{(k)}\}, \beta_{t-1})$$

$$\text{with} \qquad w_t^{(k)} \propto w_{t-1}^{(k)} \frac{\mathbb{1}_{\{s_{t-1}^{(k)} \leq \beta_t\}}}{\mathbb{1}_{\{s_{t-1}^{(k)} \leq \beta_{t-1}\}}}$$

  **if** $\mathrm{ESS}(\{w_t^{(k)}\}, \beta_t) < V$ **then**
    resample K new particles $\boldsymbol{\xi}_{t-1}^{(k)}$ from the set of particles $\{w_{t-1}^{(k)}, \boldsymbol{\xi}_{t-1}^{(k)}\}$
    set weights $w_t^{(k)} = 1/K$
  **end if**
  sample $K$ new particles $\boldsymbol{\xi}_t^{(k)} \sim K(\boldsymbol{\xi}_{t-1}^{(k)}, \boldsymbol{\xi}_t^{(k)})$ with $k = 1\!:\!K$
**end for**
**output:** posterior particles $\boldsymbol{\xi}_T^{(k)}$

Table 2: Approximations of the optimal backward kernel $L_t^{\mathrm{opt}}(\boldsymbol{\xi}_{t+1}, \boldsymbol{\xi}_t)$ lead to different SMC-ABC approaches.

| Algorithm | Assumption | $\tilde{L}_t$ | $\hat{w}_{t+1}$ |
|---|---|---|---|
| Optimal | - | $\frac{\tilde{p}_t(\boldsymbol{\xi}_t) K_{t+1}(\boldsymbol{\xi}_t, \boldsymbol{\xi}_{t+1})}{\tilde{p}_{t+1}(\boldsymbol{\xi}_{t+1})}$ | - |
| PMC-ABC | $p_t \approx \tilde{p}_{\beta_t}$ | $\frac{p_{\beta_t}(\boldsymbol{\xi}_t \vert \mathcal{X}^\star) K_{t+1}(\boldsymbol{\xi}_t, \boldsymbol{\xi}_{t+1})}{\int p_{\beta_t}(\boldsymbol{\xi}_t \vert \mathcal{X}^\star) K_{t+1}(\boldsymbol{\xi}_t, \boldsymbol{\xi}_{t+1}) \mathrm{d}\boldsymbol{\xi}_t}$ | $\frac{p_{\beta_{t+1}}(\boldsymbol{\xi}_{t+1} \vert \mathcal{X}^\star)}{\int p_{\beta_t}(\boldsymbol{\xi}_t \vert \mathcal{X}^\star) K_{t+1}(\boldsymbol{\xi}_t, \boldsymbol{\xi}_{t+1}) \mathrm{d}\boldsymbol{\xi}_t}$ |
| SMC-ABC | $p_{\beta_{t+1}} \approx p_{\beta_t}$ | $\frac{p_{\beta_t}(\boldsymbol{\xi}_t \vert \mathcal{X}^\star) K_{t+1}(\boldsymbol{\xi}_t, \boldsymbol{\xi}_{t+1})}{p_{\beta_t}(\boldsymbol{\xi}_{t+1} \vert \mathcal{X}^\star)}$ | $\frac{p_{\beta_{t+1}}(\boldsymbol{\xi}_t \vert \mathcal{X}^\star)}{p_{\beta_t}(\boldsymbol{\xi}_t \vert \mathcal{X}^\star)}$ |
| MH-ABC | $L_t(\boldsymbol{\xi}_{t+1}, \boldsymbol{\xi}_t) = K_{t+1}(\boldsymbol{\xi}_{t+1}, \boldsymbol{\xi}_t)$ | | $\frac{p_{\beta_{t+1}}(\boldsymbol{\xi}_{t+1} \vert \mathcal{X}^\star) K_{t+1}(\boldsymbol{\xi}_{t+1}, \boldsymbol{\xi}_t)}{p_{\beta_t}(\boldsymbol{\xi}_t \vert \mathcal{X}^\star) K_{t+1}(\boldsymbol{\xi}_t, \boldsymbol{\xi}_{t+1})}$ |

**Algorithm 3** Adaptive Population Monte Carlo ABC (Lenormand et al., 2013)

**input:** reference data $\mathcal{X}^\star$, prior $p(\boldsymbol{\xi})$, stochastic simulator $p(\mathbf{x}|\boldsymbol{\xi})$, IPM $D(\cdot,\cdot)$, max. iteration count $T$, resampling threshold $V$, forward kernel $K_{t+1}(\boldsymbol{\xi}_t, \boldsymbol{\xi}_{t+1})$, $\alpha$-Quantile $\alpha$
initialize particles $\boldsymbol{\xi}_0^{(k)} \sim p(\cdot)$, initialize particle weights $w_0^{(k)} = 1/K$, $K_\alpha = \alpha K$

**for** $t$ in $1{:}T$ **do**
  choose the $K_\alpha$ best particles $\hat{\boldsymbol{\xi}}_{t-1}^{(k)}$ with $k = 1 : K_\alpha$
  sample $K - K_\alpha$ new proposal particles $\tilde{\boldsymbol{\xi}}_t^{(l)} \sim K_t(\hat{\boldsymbol{\xi}}_{t-1}^{(k)}, \tilde{\boldsymbol{\xi}}_t^{(l)})$ with $l = 1 : K - K_\alpha$
  **for** each $\tilde{\boldsymbol{\xi}}_t^{(l)}$ **do**
    simulate $\mathcal{X}_t^{(l)} = \{\mathbf{x}_n^{(l)} \sim p(\mathbf{x}|\tilde{\boldsymbol{\xi}}_t^{(l)})\}$
    compute the IPM $s_t^{(l)} = D(\mathcal{X}^\star, \mathcal{X}_t^{(l)})$
  **end for**
  update the bandwidth $\beta_t$ based on the $\alpha$-Quantile

$$\beta_t = \mathcal{Q}_{\{s_{t-1}^{(k)}, s_t^{(l)}\}}(\alpha)$$

  update weights

$$w_t^{(k)} = \frac{p(\tilde{\boldsymbol{\xi}}_t^{(k)})}{\sum_{i=1}^{K_\alpha} \frac{w_{t-1}^{(i)}}{\sum_{j=1}^{K_\alpha} w_{t-1}^{(j)}} K_t(\boldsymbol{\xi}_{t-1}^{(i)}, \tilde{\boldsymbol{\xi}}_t^{(k)})}$$

  set the $K$ new particles $\boldsymbol{\xi}_t^k = \{\hat{\boldsymbol{\xi}}_{t-1}^{(k)}, \tilde{\boldsymbol{\xi}}_t^{(k)}\}$ and sort them in increasing order based on their IPM $s_t^{(k)}$
  update forward kernel $K_{t+1}(\boldsymbol{\xi}_t, \boldsymbol{\xi}_{t+1})$ with particles $\boldsymbol{\xi}_t^{(k)}$
**end for**
**output:** posterior particles $\boldsymbol{\xi}_T^{(k)}$

## C. Experimental Details

Here we detail the experimental configurations to reproduce the results covered in Figure 2.

Table 3: Hyper-parameter settings of the SBI methods as used for the experiments in Section 4. Forward slashes symbolize layers of a neural network.

| Parameter | Value |
|---|---|
| PLI (Ours) | |
| Likelihood kernel | Exponential Kernel |
| Model | Neural Spline Flow (NSF) |
| Bijector | Rational Quadratic Spline with param size $D$ |
| # Bins | 10 |
| Conditioning MLP | input dim / 50 / 50 / 50 / $D$ |
| # Bijectors / Transforms | 5 |
| Base distribution | $\mathcal{N}(\mathbf{0}, \mathbf{1})$ |
| Learning rate | $1 \times 10^{-5}$ |
| Epochs | 20 |
| Train samples per iteration | 5000 |
| Batch size | 125 |
| PMC-ABC (Lenormand et al., 2013) | |
| Likelihood kernel | Uniform Kernel |
| Likelihood update | $\alpha$-Quantile, $\alpha = 0.1$ (see Table 1) |
| $\alpha$ | 0.1 |
| Reverse transition kernel $\tilde{L}_t$ | reverse PMC kernel (see Table 2) |
| Particles | 1000 |
| Epochs | 200 |
| Perturbation kernel | GMM with 5 components |
| APT (Greenberg et al., 2019) | |
| Model | Conditional Neural Spline Flow (NSF) |
| Bijector | Rational Quadratic Spline with param size $D$ |
| # Bins | 10 |
| # Bijectors / Transforms | 5 |
| Conditioning MLP | input dim / 32 / 32 / 32 / $D$ |
| Base distribution | $\mathcal{N}(\mathbf{0}, \mathbf{1})$ |
| # Atoms | 10 |
| Learning rate | $1 \times 10^{-5}$ |
| Epochs | 20 |
| Train samples per epoch | 5000 |
| Batch size | 500 |

### C.1. Gaussian Location

The Gaussian location model is a 10-dimensional Gaussian model. The ten dimensional parameters $\boldsymbol{\xi} \in [-1, 1]^{10}$ define the means of the model $\mathcal{N}(\mathbf{x}|\mu = \boldsymbol{\xi}, \Sigma = 0.1I)$. We choose a Gaussian prior $p(\boldsymbol{\xi}) = \mathcal{N}(\boldsymbol{\xi}|\mathbf{0}, 0.1I)$ for which the posterior can be recovered in closed form. The ground-truth parameter is sampled uniformly within the posterior support $\boldsymbol{\xi}^{\text{gt}} \sim \mathcal{U}(-\mathbf{1}, \mathbf{1})$.

### C.2. Simple-Likelihood Complex-Posterior

The Simple-Likelihood Complex-Posterior (SLCP) task consists of a 5-dimensional parameter space $\boldsymbol{\Xi} \in [-3, 3]^5$ with a uniform prior $\mathcal{U}(-\mathbf{3}, \mathbf{3})$. The ground-truth parameter, from which the reference observations are generated, is set to $\boldsymbol{\xi}^{(\text{gt})} = (0.7, 1.5, -1.0, -0.9, 0.6)^T$. The observations represent four samples from a 2-dimensional Gaussian distribution

$$\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, ]^T, \quad \mathbf{x}_i = \mathcal{N}(\mathbf{x}_i, \boldsymbol{\mu}(\boldsymbol{\xi}), \Sigma(\boldsymbol{\xi})).$$

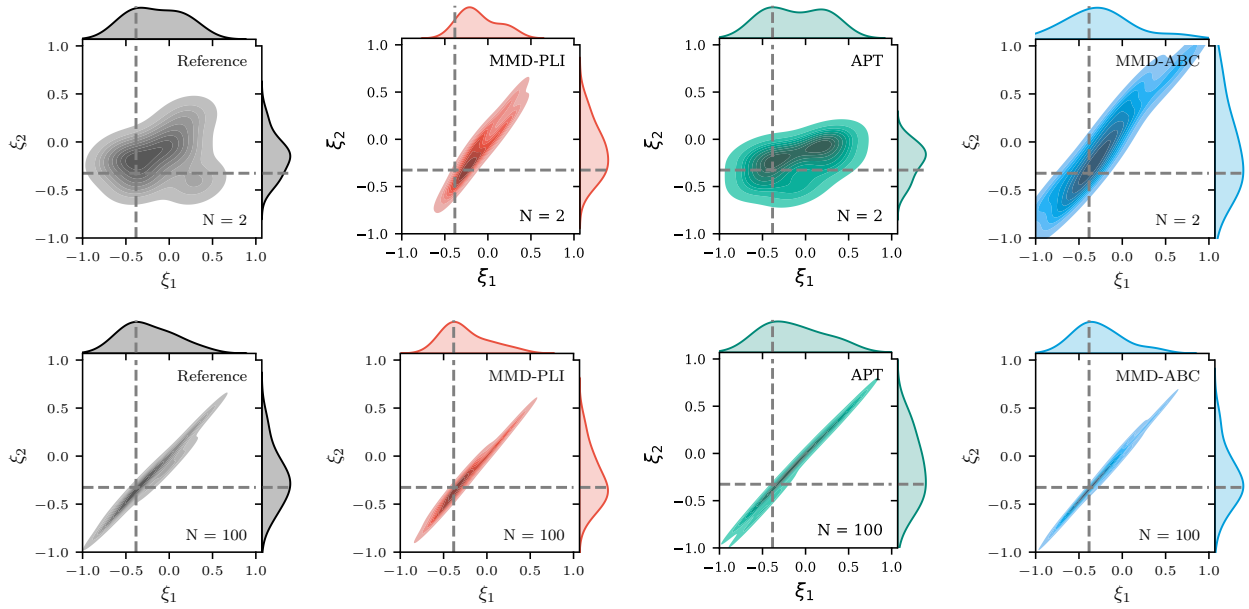



Figure 4: Slice of the posterior through the $\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2$ plane. The upper row shows experiments conducted on $N = 2$ reference observations. The lower row shows the approximate posteriors for $N = 100$ reference observations. The dotted line represents the ground truth parameter $\boldsymbol{\xi}^{(\mathrm{gt})}$ that was used to generate $\mathcal{X}^\star$. All approaches show that the posterior becomes denser when conditioned on more data.

For further information, we refer to the SBI benchmarking paper from Lueckmann et al. (2021).

### C.3. The Furuta Pendulum

This inverted double pendulum can be described by the angular deflection $[\theta_r, \theta_p]$ of the rods *w.r.t.* their equilibrium position $[0, 0]$. The equations of motion can be derived by formulating the Euler-Lagrange equation (Muratore et al., 2021):

$$\begin{bmatrix} -d_r\dot{\theta}_r + \tau \\ -d_p\dot{\theta}_p \end{bmatrix} = \begin{bmatrix} \frac{1}{12}m_r l_r^2 + m_p l_r^2 + \frac{1}{4}m_p l_p^2 \sin^2\theta_p & \frac{1}{2}m_p l_p l_r \cos\theta_p \\ \frac{1}{2}m_p l_p l_r \cos\theta_p & \frac{1}{3}m_p l_p^2 \end{bmatrix} \begin{bmatrix} \ddot{\theta}_r \\ \ddot{\theta}_p \end{bmatrix} + \begin{bmatrix} \frac{1}{4}m_p l_p^2 \sin 2\theta_p \, \dot{\theta}_r\dot{\theta}_p - \frac{1}{2}m_p l_p l_r \sin\theta_p \, \dot{\theta}_p^2 \\ -\frac{1}{8}m_p l_p^2 \sin 2\theta_p \, \dot{\theta}_r^2 + \frac{1}{2}m_p l_p g \sin\theta_p \end{bmatrix}.$$

Here, the mild assumption is made that the pole length is significantly greater than its diameter for which the moments of inertia of the poles around their pivot are $J_i = 1/3 \, m_i l_i^2, \; i \in \{r, p\}$. The mass matrix contains entries from the translatory and rotational movement of the two poles. As the reference coordinate systems are constantly rotating *w.r.t.* the basis coordinate system, Coriolis forces occur. They are complemented by gravitation which works on the rotational pole. The left-hand side considers damping in the joints, represented by the damping coefficients $d_r$ and $d_p$, and the torque $\tau$ which is applied from a servo motor. For this paper, we omit external forces, i.e., $\tau = 0\,\mathrm{Nm}$. The Furuta pendulum is set into motion by perturbing the initial state around its unstable equilibrium.

For the system identification tasks we select the five system parameters

$$\boldsymbol{\xi} = \begin{bmatrix} g \\ l_r \\ m_r \\ l_p \\ m_p \end{bmatrix} \in \begin{bmatrix} [9, 11] \\ [0.08, 0.09] \\ [0.08, 0.1] \\ [0.12, 0.135] \\ [0.02, 0.03] \end{bmatrix}; \qquad \boldsymbol{\xi}^{\mathrm{gt}} = \begin{bmatrix} 9.81 \\ 0.085 \\ 0.095 \\ 0.129 \\ 0.024 \end{bmatrix}$$

with a uniform prior on the predefined ranges.

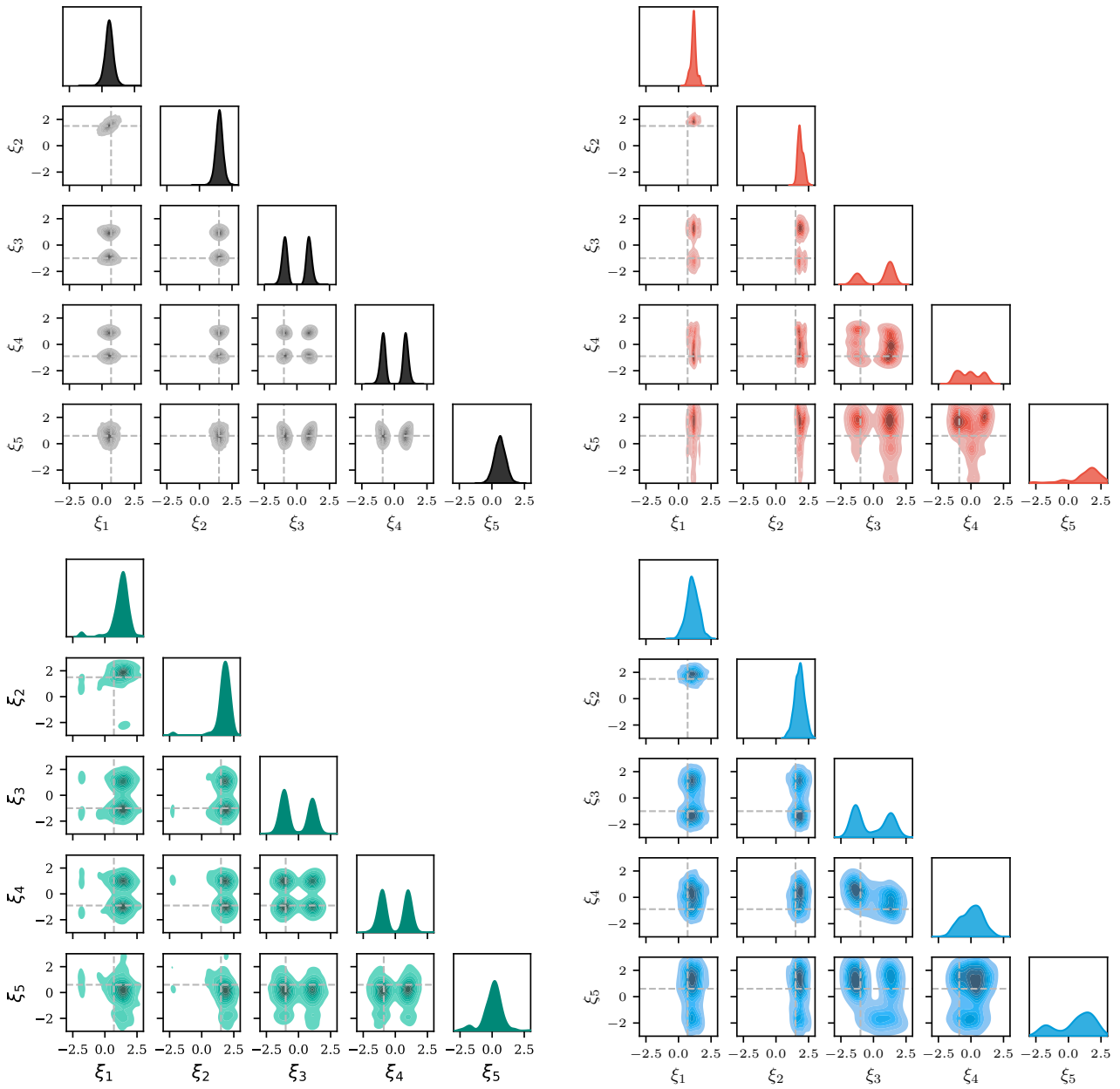



Figure 5: Results of posterior inference on the SLCP task with $N = 2$ reference observations. The unimodal distribution of the parameters $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$) are depicted well by all approaches. On the contrary, the multi-modality is only represented properly by the APT posterior (— Reference, — MMD-PLI, — APT, — MMD-ABC).

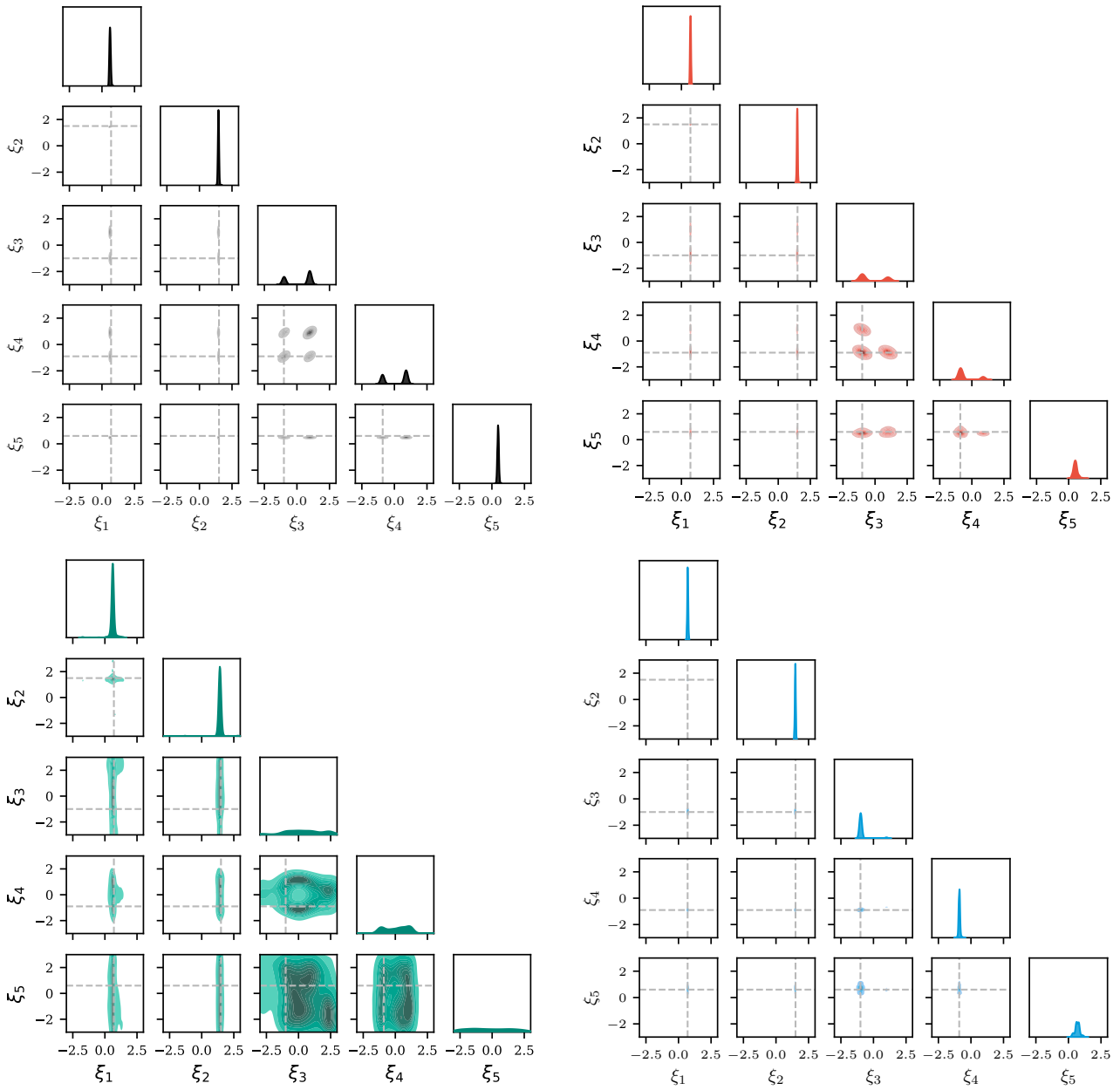



Figure 6: Results of posterior inference on the SLCP task with $N = 100$ reference observations. Compared to the posterior given $N = 2$ observations (Figure 5), the posterior (━ Reference) is distributed tightly around distinct points. Here, ━ MMD-PLI captures all modes of the posterior, ━ MMD-ABC centers around a uni-mode, while ━ APT cannot represent the multi-modality.

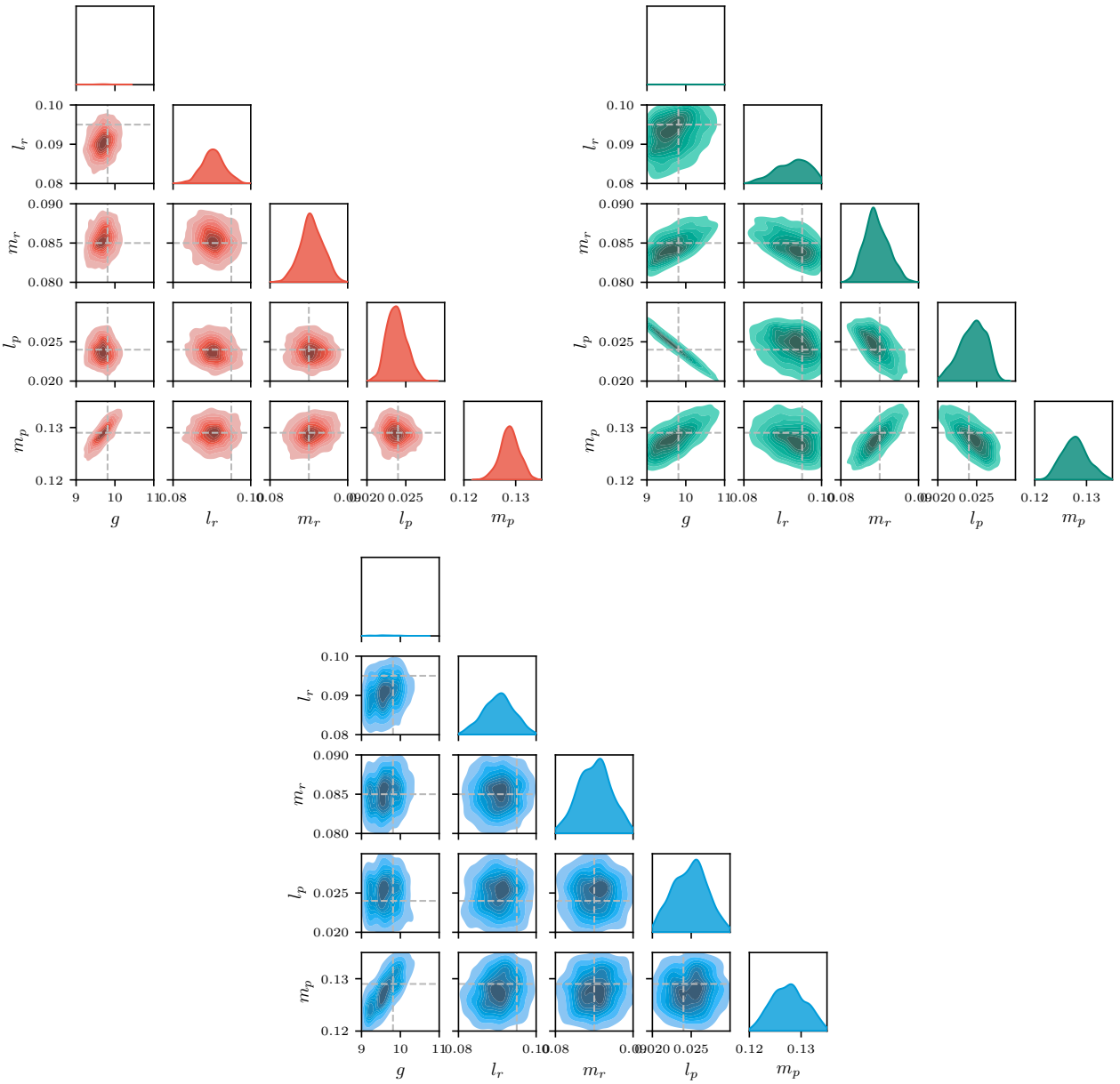



Figure 7: Results of posterior inference on the Furuta pendulum with $N = 2$ reference observations. All methods center around the ground truth parameter. — APT finds the expected correlations among the parameters while — MMD-PLI and — MMD-ABC remain more widespread.

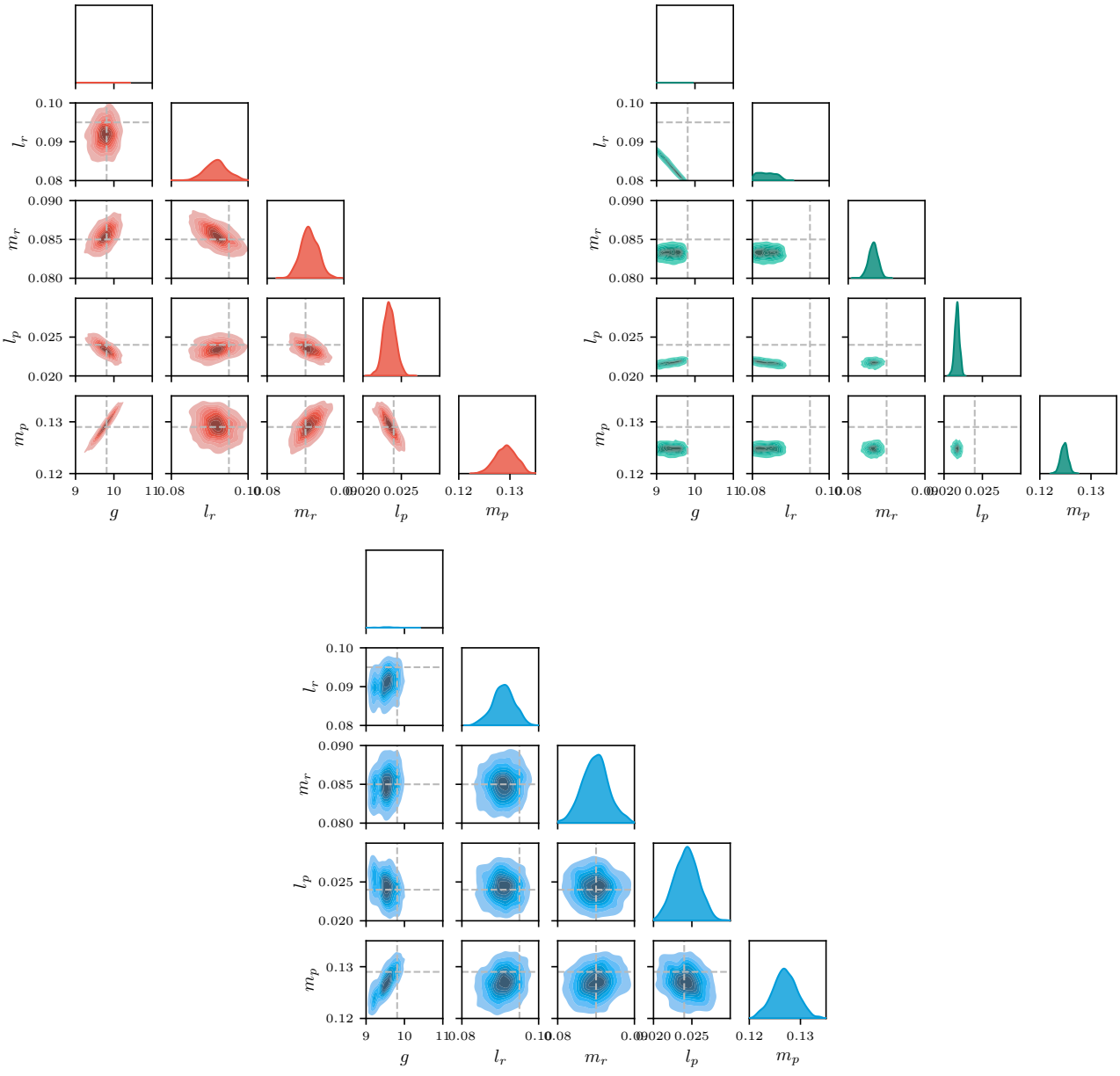



Figure 8: Results of posterior inference on the Furuta pendulum with $N = 100$ reference observations. All models capture the ground-truth parameter well. In contrast to the $N = 2$ setting (Figure 7) MMD-PLI reveals pairwise correlations between the domain parameters, and MMD-ABC is less densely distributed. Note, that APT clusters outside of the ground-truth parameter.