Simulation-Based Inference with Quantile Regression

He Jia (贾赫) [1]

Abstract

We present Neural Quantile Estimation (NQE), a novel Simulation-Based Inference (SBI) method based on conditional quantile regression. NQE autoregressively learns individual one dimensional quantiles for each posterior dimension, conditioned on the data and previous posterior dimensions. Posterior samples are obtained by interpolating the predicted quantiles using monotonic cubic Hermite spline, with specific treatment for the tail behavior and multi-modal distributions. We introduce an alternative definition for the Bayesian credible region using the local Cumulative Density Function (CDF), offering substantially faster evaluation than the traditional Highest Posterior Density Region (HPDR). In case of limited simulation budget and/or known model misspecification, a post-processing calibration step can be integrated into NQE to ensure the unbiasedness of the posterior estimation with negligible additional computational cost. We demonstrate that NQE achieves state-of-the-art performance on a variety of benchmark problems.

1. Introduction

Given the likelihood $p(\mathbf{x}|\boldsymbol{\theta})$ of a stochastic forward model and observation data $\mathbf{x}$, Bayes' theorem postulates that the underlying model parameters $\boldsymbol{\theta}$ follow the posterior distribution $p(\boldsymbol{\theta}|\mathbf{x}) \propto p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$, where $p(\boldsymbol{\theta})$ represents the prior. In many applications, however, we are restricted to simulating the data $\mathbf{x} \sim p(\mathbf{x}|\boldsymbol{\theta})$, while the precise closed form of $p(\mathbf{x}|\boldsymbol{\theta})$ remains unavailable. Simulation-Based Inference (SBI), also known as Likelihood-Free Inference (LFI) or Implicit Likelihood Inference (ILI), conducts Bayesian inference directly from these simulations, circumventing the need to explicitly formulate a tractable likelihood function. Early research in this field primarily consists of Approximate Bayesian Computation (ABC) variants, which employ a distance metric in the data space and approximate true posterior samples using realizations whose simulated data are "close enough" to the observation (e.g. Tavaré et al., 1997; Pritchard et al., 1999; Beaumont et al., 2002; 2009). However, these methods are prone to the curse of dimensionality and prove inadequate for higher-dimensional applications.

In recent years, a series of neural-network-based SBI methods have been proposed, which can be broadly categorized into three groups. Neural Likelihood Estimation (NLE, Papamakarios et al., 2019b; Lueckmann et al., 2019) fits the likelihood using a neural density estimator, typically based on Normalizing Flows. The posterior is then evaluated by multiplying the likelihood with the prior, and posterior samples can be drawn using Markov Chain Monte Carlo (MCMC). Neural Posterior Estimation (NPE, Papamakarios & Murray, 2016; Lueckmann et al., 2017; Greenberg et al., 2019) uses neural density estimators to approximate the posterior, thereby enabling direct posterior sample draws without running MCMC. Neural Ratio Estimation (NRE, Hermans et al., 2020) employs classifiers to estimate density ratios, commonly selected as the likelihood-to-evidence ratio. Indeed, Durkan et al. (2020) demonstrates that NRE can be unified with specific types of NPE under a general contrastive learning framework. Each method has its sequential counterpart, namely SNLE, SNPE, and SNRE, respectively. Whereas standard NLE, NPE, and NRE allocate new simulations based on the prior, allowing them to be applied to any observation data (i.e., they are *amortized*), their sequential counterparts allocate new simulations based on the inference results from previous iterations and must be trained specifically for each observation. These neural-network-based methods typically surpass traditional ABC methods in terms of inference accuracy under given simulation budgets. See Cranmer et al. (2020) for a review and Lueckmann et al. (2021) for a comprehensive benchmark of prevalent SBI methods.

Quantile Regression (QR), as introduced by Koenker & Bassett Jr (1978), estimates the conditional quantiles of the response variable over varying predictor variables. Many Machine Learning (ML) algorithms can be extended to quantile regression by simply transitioning to a weighted L_1 loss (e.g. Meinshausen & Ridgeway, 2006; Rodrigues & Pereira, 2020; Tang et al., 2022). In this paper, we introduce Neural

[1]Department of Astrophysical Sciences, Princeton University, USA. Correspondence to: He Jia <hejia@princeton.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

Quantile Estimation (NQE), a new family of SBI methods supplementing the existing NPE, NRE and NLE approaches. NQE successively estimates the one dimensional quantiles of each dimension of $\boldsymbol{\theta}$, conditioned on the data $\mathbf{x}$ and previous $\boldsymbol{\theta}$ dimensions. We interpolate the discrete quantiles with monotonic cubic Hermite splines, adopting specific treatments to account for the tail behavior and potential multimodality of the distribution. Posterior samples can then be drawn by successively applying inverse transform sampling for each dimension of $\boldsymbol{\theta}$. We also develop a post-processing calibration strategy, leading to **guaranteed unbiased posterior estimation** as long as one provides enough ($\lesssim 10^3$) simulations to accurately calculate the empirical coverage. To the best of our knowledge, this constitutes the first demonstration that QR-based SBI methods can attain state-of-the-art performance, matching or surpassing the benchmarks set by existing methods.

The structure of this paper is as follows: In Section 2, we introduce the methodology of NQE, along with a alternative definition for Bayesian credible regions and a post-processing calibration scheme to ensure the unbiasedness of the inference results. In Section 3, we demonstrate that NQE attains state-of-the-art performance across a variety of benchmark problems, together with a realistic application to high dimensional cosmology data. Subsequently, in Section 4, we discuss related works in the literature and potential avenues for future research. The results in this paper can be reproduced with the publicly available `NQE` package [1] based on `pytorch` (Paszke et al., 2019).

2. Methodology

2.1. Quantile Estimation And Interpolation

The cornerstone of most contemporary SBI methods is some form of conditional density estimator, which is used to approximate the likelihood, the posterior, or the likelihood-to-evidence ratio. Essentially, every generative model can function as a density estimator. While Generative Adversarial Networks (Goodfellow et al., 2020) and more recently Diffusion Models (Dhariwal & Nichol, 2021) have shown remarkable success in generating high-quality images and videos, the SBI realm is primarily governed by Normalizing Flows (NF, e.g. Rezende & Mohamed, 2015; Papamakarios et al., 2019a), which offer superior inductive bias for the probabilistic distributions with up to dozens of dimensions frequently encountered in SBI tasks. Our proposed NQE method can also be viewed as a density estimator, as it reconstructs the posterior distribution autoregressively from its 1-dim conditional quantiles.

In a typical SBI setup, one first samples the model parameters $\boldsymbol{\theta}$ from the prior $p(\boldsymbol{\theta})$, and then runs the forward simulations to generate the corresponding observations $\mathbf{x}$. For simplicity, let us start with the scenario of 1-dim θ. Given a dataset $\{\theta, \mathbf{x}\}$ and a neural network $F_\phi(\mathbf{x})$ parameterized by ϕ, one can estimate the median (mean) of θ conditioned on $\mathbf{x}$ by minimizing the L_1 (L_2) loss [2] between θ and $F_\phi(\mathbf{x})$. As a straightforward generalization, one can estimate the τ-th quantile of θ conditioned on $\mathbf{x}$ by minimizing the following weighted L_1 loss,

$$\mathcal{L}_\tau[\theta,\, F_\phi(\mathbf{x})] \equiv (\tau-1) \sum_{\theta<F_\phi(\mathbf{x})} w(\mathbf{x})\, [\theta - F_\phi(\mathbf{x})] + \tau \sum_{\theta\geq F_\phi(\mathbf{x})} w(\mathbf{x})\, [\theta - F_\phi(\mathbf{x})]\,. \quad (1)$$

Here one can introduce an additional $\mathbf{x}$-dependent weight $w(\mathbf{x})$, similar to the fact that one can use simulations allocated from an arbitrary prior to train SNLE. A discussion regarding the choice of $w(\mathbf{x})$ can be found in Appendix B. To reconstruct the full posterior, we require the quantiles at multiple τ's, for which we aggregate the individual loss functions,

$$\mathcal{L}_0[\theta,\, F_\phi(\mathbf{x})] \equiv \sum_\tau \mathcal{L}_\tau[\theta,\, F_\phi(\mathbf{x})]\,. \quad (2)$$

Without loss of generality, we assume the prior of θ is zero outside some interval $[a, b]$. If the prior is positive everywhere on $\mathbb{R}$, one can choose $[a, b]$ such that the prior mass

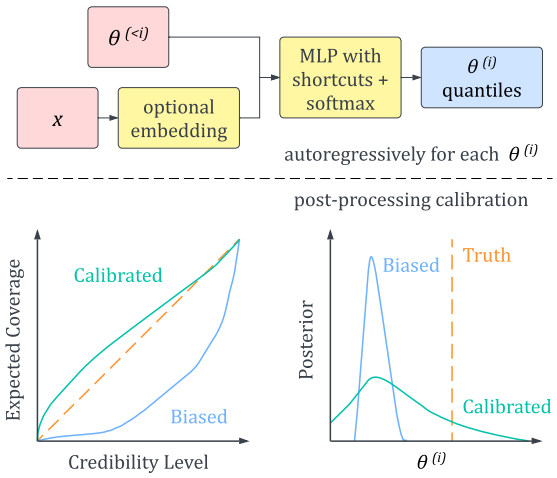

Figure 1. (Top) Network architecture of our NQE method, which autoregressively learns 1-dim conditional quantiles for each posterior dimension. The estimated quantiles are then interpolated to reconstruct the full distribution. (Bottom) A post-processing calibration step can be employed to ensure the unbiasedness of NQE inference results.

[1] https://github.com/h3jia/nqe.

[2] Not to be confused with $\mathcal{L}_0$ and $\mathcal{L}_1$ defined below.

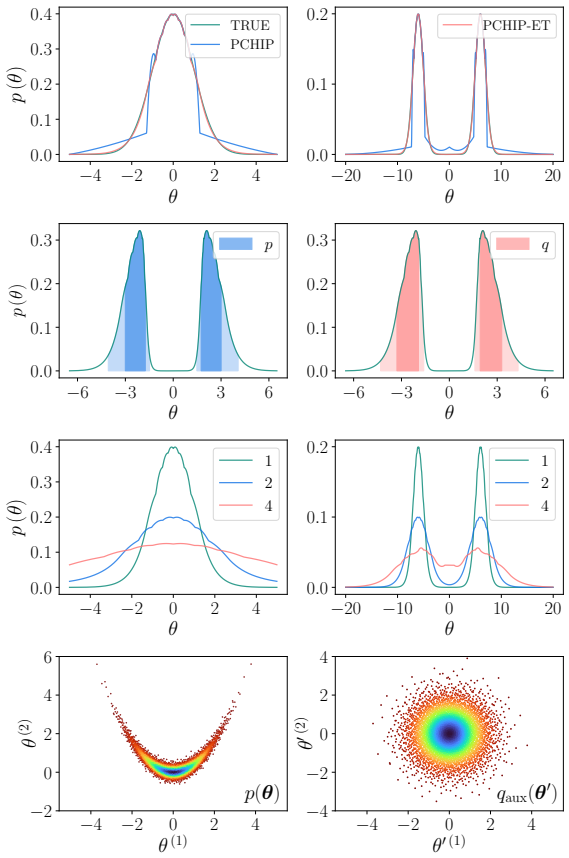

Figure 2. (1st row) Interpolation of Gaussian and Gaussian Mixture distributions. While the original PCHIP algorithm shows significant interpolation artifacts, our modified PCHIP-ET scheme decently reconstructs the distributions with only ~ 15 quantiles. (2nd row) Comparison of the 68.3% and 95.4% credible regions for a mixture of two asymmetric modes, evaluated with HPDR (p-coverage) and QMCR (q-coverage). (3rd row) Broadening of the interpolated posterior, with the broadening factors indicated in the legend. (4th row) The bijective mapping f_{qm} establishes a one-to-one correspondence between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ with the same 1-dim conditional CDF across all the $\theta^{(i)}$ dimensions. The $p-$coverage and $q-$coverage are based on the ranking of $p(\boldsymbol{\theta})$ and $q_{\text{aux}}(\boldsymbol{\theta}')$, respectively.

outside it is negligible. For example, one can set $[a, b]$ to $[-5, 5]$ for a standard Gaussian prior; in case of heavy-tailed priors, one can also use the (inverse) prior CDF to map the prior support to $[0, 1]$. We then equally divide the interval $[0, 1]$ into n_{bin} bins, and estimate the corresponding $n_{\text{bin}} - 1$ quantiles with $F_\phi(\mathbf{x})$. In this work, we choose $F_\phi(\mathbf{x})$ to be a Multi-Layer Perceptron (MLP) with n_{bin} outputs z_i followed by a softmax layer, such that the i/n_{bin}-th quantile of θ is parameterized as $a + (b - a) \times \sum_{j \leq i} \text{softmax}(z_j)$, and we add shortcut connections (the input layer of MLP is concatenated to every hidden layer) to facilitate more

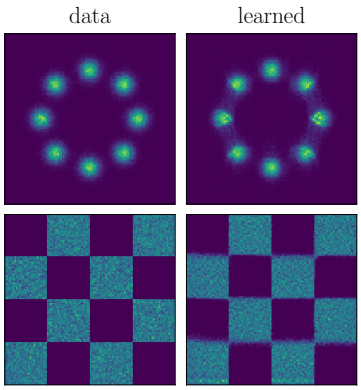

Figure 3. Probability density estimation for two toy examples from Grathwohl et al. (2018). Despite the intricate multimodal structures, NQE is able to faithfully reconstruct both distributions.

efficient information propagation throughout the network. Moreover, an optional embedding network (e.g. Jiang et al., 2017; Radev et al., 2020) can be added before the MLP to more efficiently handle high dimensional data (e.g. the cosmology example in Section 3.3).

For multidimensional $\boldsymbol{\theta}$, we successively apply the aforementioned method to each dimension $\theta^{(i)}$, conditioned on not only the data $\mathbf{x}$ but also all the previous dimensions $\boldsymbol{\theta}^{(j<i)}$. In other words, $F_\phi(\mathbf{x})$ in Equations (1) and (2) is replaced by $F_\phi(\mathbf{x}, \boldsymbol{\theta}^{(j<i)})$, since $\boldsymbol{\theta}^{(j<i)}$ is effectively treated as observation data for the inference of $\theta^{(i)}$. An illustration of the NQE architecture can be found in the top panel of Figure 1. Similar to Flow Matching Posterior Estimation (FMPE, Dax et al., 2023), NQE has an *unconstrained* architecture which does not require specialized NFs.

The estimated conditional quantiles must be interpolated to enable sampling from them. We achieve this by interpolating the Cumulative Distribution Function (CDF) using Piecewise Cubic Hermite Interpolating Polynomial with Exponential Tails (PCHIP-ET), a modified version of the PCHIP scheme (Fritsch & Carlson, 1980), which preserves monotonicity of input data and continuity of first derivatives, ensuring a well-defined Probability Distribution Function (PDF). As depicted in the 1st row of Figure 2, the original PCHIP algorithm presents discernible interpolation artifacts, primarily because polynomials cannot decay rapidly enough to align with the true PDF in the tail regime. To address this issue, we substitute the polynomials with Gaussians within bins identified as tails. A more detailed description of our PCHIP-ET scheme is available in Appendix A. We observe that a satisfactory reconstruction of unimodal distributions can be achieved with $\lesssim 15$ quantiles, while incorporating additional bins may facilitate better convergence in multimodal cases. Samples can then be drawn using inverse

transform sampling with the interpolated CDF.

NQE requires one neural network for each posterior dimension, which can be trained independently on multiple devices to reduce the training wall time. In principle, one can also train NQE by maximizing the joint PDF, similar to the training of NPE. However, such approach will be less efficient than minimizing $\mathcal{L}_0$ in Equation (2), since one needs to compute the PCHIP-ET interpolation for the PDF, while $\mathcal{L}_0$ only depends on the individual quantiles. NQE can also be used to estimate $p(\boldsymbol{\theta})$ distributions with no observation $\mathbf{x}$ to condition on. In this case, we do not need neural networks for the first dimension $\theta^{(1)}$, which can be directly interpolated from the empirical quantiles. In Figure 3, we demonstrate that NQE can successfully model two complicated distributions from Grathwohl et al. (2018).

2.2. Regularization

Numerical derivatives are inherently noisier than integrals, and similarly for the PDF compared with the CDF. To mitigate this issue, we propose the following regularization scheme to improve the smoothness of NQE PDF predictions. Intuitively, a "smooth distribution" means the averaged PDF within every 1-dim bin for quantile prediction, $\langle p \rangle_{\rm bin}$, should be close to the interpolated value between its neighboring bins,

$$\langle p \rangle_{\rm interp} \equiv \max\Bigg[f_1 \times \left(\langle p \rangle_{\rm left} + \langle p \rangle_{\rm right} \right) / 2, \\ f_2 \times \max\left(\langle p \rangle_{\rm left}, \langle p \rangle_{\rm right} \right) \Bigg], \quad (3)$$

with $f_1 = 1.1$ and $f_2 = 0.8$, which leads to the following loss for regularization,

$$\mathcal{L}_1 \equiv \sum_{\rm bins} H\left(\langle p \rangle_{\rm bin} - \langle p \rangle_{\rm interp} \right) \times \\ \left(\log \langle p \rangle_{\rm bin} - \log \langle p \rangle_{\rm interp} \right)^2, \quad (4)$$

where $H(\cdot)$ is the Heaviside function. With Equation (4), we only penalize cases where $\langle p \rangle_{\rm bin} > \langle p \rangle_{\rm interp}$, since we will have $\langle p \rangle_{\rm bin} < \langle p \rangle_{\rm interp}$ between the peaks in multimodal problems, which is therefore a possible feature in the ground truth solution that should not be penalized. For similar reasons, $\langle p \rangle_{\rm interp}$ in Equation (3) is set to be larger than the naive average of $\langle p \rangle_{\rm left}$ and $\langle p \rangle_{\rm right}$, so that the regularization is only activated when necessary. The total loss is then defined as

$$\mathcal{L} \equiv \mathcal{L}_0 \left(1 + \lambda_{\rm reg} \mathcal{L}_1 \right). \quad (5)$$

Note that a linear rescaling of $\boldsymbol{\theta}$ changes $\mathcal{L}_0$ while $\mathcal{L}_1$ remains invariant, which motivates our choice of $\mathcal{L}$ above. We find 0.1 to be a generally reasonable choice for $\lambda_{\rm reg}$, although one may reduce $\lambda_{\rm reg}$ for examples with e.g. sharp spikes or edges in the posterior distribution, if one has such prior knowledge of the typical shape of the posterior.

2.3. Empirical Coverage

Analogous to frequentist confidence regions, Bayesian statistics utilizes credible regions to define the reasonable space for model parameters $\boldsymbol{\theta}$ given $\mathbf{x}$. The most popular choice of Bayesian credible region, namely the highest posterior density region (HPDR, e.g. McElreath, 2020), encloses the $\alpha\%$ samples with the highest PDF for the $\alpha\%$ credible region, achieving the smallest $\boldsymbol{\theta}$ volume for any given credibility level. To test whether a posterior estimator is biased, one checks the empirical coverage, namely the probability of the true model parameters to fall into the $\alpha\%$ credible region over the simulation data. If such probability is larger (smaller) than $\alpha\%$, the posterior estimator is over-conservative (biased) [3]. To compute the empirical coverage in practice, one needs to pick N_o pairs of $(\boldsymbol{\theta}, \mathbf{x})$ from the simulation data, and generate N_r samples for each of them to get the rank of PDF, leading to $\mathcal{O}(N_o N_r)$ neural network calls for NPE and NQE [4]. For NLE and NRE, such cost is further multiplied by N_m, the number of posterior evaluations per effective MCMC sample [5]. Typically one needs to set both N_o and N_r to $\sim 10^2 - 10^3$ so as to get a reliable estimate of the empirical coverage, leading to a moderate computational cost especially for NLE and NRE methods.

A unique characteristics of NQE is that it predicts the distribution quantiles, which explicitly contains the information regarding the *global* properties of the posterior and enables us to propose the following quantile mapping credible region (QMCR) [6], a generalization of the 1-dim equal-tailed credible interval (e.g. McElreath, 2020) for multidimensional distributions. Talts et al. (2018) shows the rank of any 1-dim statistic can be used to define the Bayesian credible region, with HPDR a special case that chooses such statistic as the posterior PDF. With the conditional quantiles predicted by NQE, we introduce an auxiliary distribution $q_{\rm aux}(\boldsymbol{\theta}')$, which we typically set to a multivariate standard Gaussian. We then define a bijective mapping $f_{\rm qm}: \boldsymbol{\theta} \to \boldsymbol{\theta}'$ that establishes a one-to-one correspondence between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ with the same 1-dim conditional CDF,

[3] Note that being well calibrated is a necessary yet not sufficient condition for an estimator to predict the Bayesian optimal posterior, as exemplified by the extreme case where the posterior estimator always outputs the prior.

[4] We ignore the factor $\dim \boldsymbol{\theta}$ for NQE as we define one network call as one evaluation of the whole estimator.

[5] For $\dim \boldsymbol{\theta} \lesssim 5$ one may circumvent MCMC using Importance Sampling, which however becomes inefficient as the dimensionality of $\boldsymbol{\theta}$ grows.

[6] Not to be confused with the quantile mapping technique used to e.g. correct the bias for simulated climate data (Maraun, 2013).

$\int p(\theta^{(i)} \,|\, \mathbf{x},\, \boldsymbol{\theta}^{(j<i)})\, \mathrm{d}\theta^{(i)}$ and $\int q_{\rm aux}(\theta'^{(i)} \,|\, \boldsymbol{\theta}'^{(j<i)})\, \mathrm{d}\theta'^{(i)}$, across all the $\theta^{(i)}$ dimensions [7]. The defining statistic of the credible region is chosen as $\log q_{\rm aux}(\boldsymbol{\theta}')$ with $\boldsymbol{\theta}' = f_{\rm qm}(\boldsymbol{\theta})$, whose rank can be computed analytically using the χ^2 distribution since $q_{\rm aux}(\boldsymbol{\theta}')$ is Gaussian. If the interpolation indicates that $p(\theta^{(i)} \,|\, \mathbf{x},\, \boldsymbol{\theta}^{(j<i)})$ includes multiple modes, we use the *local* CDF within the mode containing θ to define the mapping $f_{\rm qm}$, such that the low PDF regions between the modes are excluded from the credible regions.

A comparison of HPDR and QMCR for a toy distribution can be found in the 2nd row of Figure 2, together with the $f_{\rm qm}$ mapping illustrated in the 4th row. Heuristically, the $\alpha \to 0$ limit of QMCR encloses the (conditional) median across all the dimensions for unimodal distributions, as opposed to the global maximum of the PDF for HPDR. Therefore, unlike HPDR, QMCR is invariant under any 1-dim monotonic transforms of $\boldsymbol{\theta}$, as long as such reparameterization does not give rise to a different identification of multimodality during the CDF interpolation. As shown with the examples below, QMCR typically leads to similar conclusions regarding the (un)biasedness of the posterior estimators as HPDR, but only requires $\mathcal{O}(N_o)$ network calls to evaluate as one no longer needs to generate N_r samples for each observation. Such speed-up allows us to perform posterior calibration in the next subsection with negligible computational cost. For simplicity, in the rest of this paper we will use the term $\boldsymbol{p}-$**coverage (**$\boldsymbol{q}-$**coverage)** for empirical coverage computed with HPDR (QMCR). In addition, we note that due to its autoregressive structure, one can compute the coverage of NQE for the leading $\boldsymbol{\theta}$ dimensions without additional training, which is useful if the unbiasedness of certain $\boldsymbol{\theta}$ dimensions takes precedence over others.

2.4. Posterior Calibration

Hermans et al. (2021) demonstrates that all existing SBI methods may produce biased results when the simulation budget is limited. Intuitively, a biased posterior is too narrow to enclose the true model parameters, so we propose the following calibration strategy as illustrated in the bottom panel of Figure 1. To make a distribution broader, we fix the medians of all 1-dim conditional posteriors and increase the distance between the medians and all other quantiles by a global *broadening factor*. Similar to the $q-$coverage evaluation, we utilize the local quantiles within modes for multimodal distributions. We remove the quantiles that escape from the boundary of the prior and/or the boundary between different modes, and redistribute the corresponding posterior mass to the bins still within the boundary based on the bin mass, so that the local posterior shape is preserved.

[7] If $q_{\rm aux}(\boldsymbol{\theta}')$ is set to a multivariate standard Gaussian, there is no correlation between the different dimensions so we indeed have $q_{\rm aux}(\theta'^{(i)}|\boldsymbol{\theta}'^{(j<i)}) = q_{\rm aux}(\theta'^{(i)})$.

Table 1. Computational cost of the broadening calibration, with NQE being significantly faster than other methods. N_i: number of iterations to solve for the desired coverage. N_o: number of simulated observations for coverage computation. N_r: number of samples per observation for the rank of PDF. N_m: number of posterior evaluations per effective MCMC sample. We assume there is no broadening technique for NPE that does not necessitate MCMC sampling.

	coverage	simulations	network calls
NQE	$\boldsymbol{q}$	$\boldsymbol{N_o}$	$\boldsymbol{\mathcal{O}(N_o)}$
NQE	p	N_o	$\mathcal{O}(N_i N_o N_r)$
NLE	p	N_o	$\mathcal{O}(N_i N_o N_r N_m)$
NPE	p	N_o	$\mathcal{O}(N_i N_o N_r N_m)$
NRE	p	N_o	$\mathcal{O}(N_i N_o N_r N_m)$

The effect of such broadening transform is shown in the 3rd row of Figure 2. We then solve for the *minimum broadening factor* such that the calibrated posterior is unbiased across a series of credibility levels, which we set to $\{0.1,\, 0.5,\, 0.9\}$ throughout this paper. Note that ideally, a good estimator should have empirical coverage that matches the credibility level. However, if this is not possible due to limited training data, over-conservative inference should be preferred over biased results. The broadening factor can also be smaller than 1, in case the original posterior is already too conservative. While one has the freedom to choose the definition of the coverage for the calibration process, the broadened posterior is only guaranteed to be unbiased at the calibrated credibility levels under the same coverage definition.

While similar calibration tricks may also be developed for other SBI methods, it will likely be considerably more expensive than NQE in practice, for the following reasons. Firstly, the evaluation of $q-$coverage is exclusive to NQE, which is faster by at least a factor of N_r than traditional $p-$coverage (with an additional factor of N_m if MCMC is required for sampling). More importantly, we have developed a broadening strategy for NQE that preserves not only the local correlation structure of the posterior but also the ability of fast sampling without MCMC. We are not aware of any similar techniques for existing SBI methods, which estimate the local PDF with no explicit global information of the distribution. For example, while one can broaden an NF-based probability distribution by lowering its *temperature*, i.e. replacing $\log p(\boldsymbol{\theta}|\mathbf{x})$ with $\beta \log p(\boldsymbol{\theta}|\mathbf{x})$, $\beta < 1$, this will necessitate MCMC sampling for NPE (NLE and NRE need MCMC even without broadening). In addition, with the analytical rank evaluation of $q-$coverage, the NQE network outputs can also be reused between different iterations, thus reducing the total network calls by another factor of N_i. We compare the computational cost of broadening calibration for different methods in Table 1.

Algorithm 1 Neural Quantile Estimation (NQE)

1. Training NQE ▶ Sections 2.1 and 2.2
  for $i = 1$ to $\dim \theta$ **do** ▶ parallelizable
    train $F_\phi(\mathbf{x}, \boldsymbol{\theta}^{(j<i)})$ by minimizing $\mathcal{L}$ ▶ Equation (5)
2. Calibrating NQE (Optional) ▶ Sections 2.3 and 2.4
  solve the calibration for unbiased posterior
3. Sampling NQE ▶ Section 2.1
  for $i = 1$ to $\dim \theta$ **do** ▶ sequential
    sample $\theta^{(i)}$ from interpolated $F_\phi(\mathbf{x}, \boldsymbol{\theta}^{(j<i)})$

Such post-processing calibration relies on a reliable calculation of the coverage. The (pointwise) error of empirical coverage due to stochastic sampling can be estimated using binomial distribution (Säilynoja et al., 2022); with $N_o = 10^3$, the maximum error is smaller than 1.6%, regardless of the dimensionality of $\mathbf{x}$ and $\boldsymbol{\theta}$ [8]. In other words, **for any inference task, with the broadening calibration, one only needs $\lesssim 10^3$ simulations in the validation dataset to ensure the unbiasedness of the posterior, if there is no model misspecification**. Nevertheless, the number of network calls required for broadening is different across the various algorithms as compared in Table 1. Using NQE and $q-$coverage, one only needs $\mathcal{O}(N_o)$ calls of the NQE network for the broadening, which is typically negligible compared with the cost for running the simulations and training the neural estimators. In addition, **similar calibration tricks can be used to mitigate partially known model misspecification, as exemplified in Section 3.3 below.** Note that we use the same validation dataset during the training and broadening calibration of NQE, as the one-parameter broadening transform is unlikely to overfit. We summarize the proposed NQE method in Algorithm 1.

In this paper, we focus on the simple broadening calibration, which is guaranteed to converge with $\lesssim 10^3$ validation simulations, regardless of $\dim \mathbf{x}$ and $\dim \boldsymbol{\theta}$. With more simulations, it may be beneficial to employ a more sophisticated calibration scheme to remove the bias without overbroadening the predicted posterior. We plan to conduct a comprehensive survey of such calibration schemes in a follow-up paper. One example is the *quantile shifting* calibration demonstrated with the cosmology example in Section 3.3: for each τ quantile of $p(\theta^{(i)}|\mathbf{x}, \boldsymbol{\theta}^{(<i)})$ predicted by NN, we check if we indeed have τ probability that the true $\theta^{(i)}$ is smaller than the predicted quantile (on the validation dataset) [9]. If not, we calculate the shift required for the τ quantile such that this statement is true. Note that we apply a shift of $\theta^{(i)}$ quantile that is different for each τ and i, but the same for all $\mathbf{x}$ and $\boldsymbol{\theta}^{(<i)}$. In other words, we effectively calculate the bias averaged over the prior, and shift the predicted quantiles accordingly to remove the bias. Strictly speaking, such quantile shifting scheme calibrates the $q-$coverage of all the individual 1-dim conditional posteriors, but not necessarily the $q-$coverage of the multi-dimensional joint posterior. In addition, the number of simulations required for this scheme depends on the dimensionality of $\boldsymbol{\theta}$, in contrast to the global broadening scheme which always converges with $\lesssim 10^3$ validation simulations. We leave a more detailed investigation of such methods for future research; nevertheless, for the cosmology example in Section 3.3, the posterior calibrated with quantile shifting has an almost diagonal empirical coverage and is much narrower than the posterior calibrated with simple global broadening, when there is a significant bias in the uncalibrated posterior due to model misspecification.

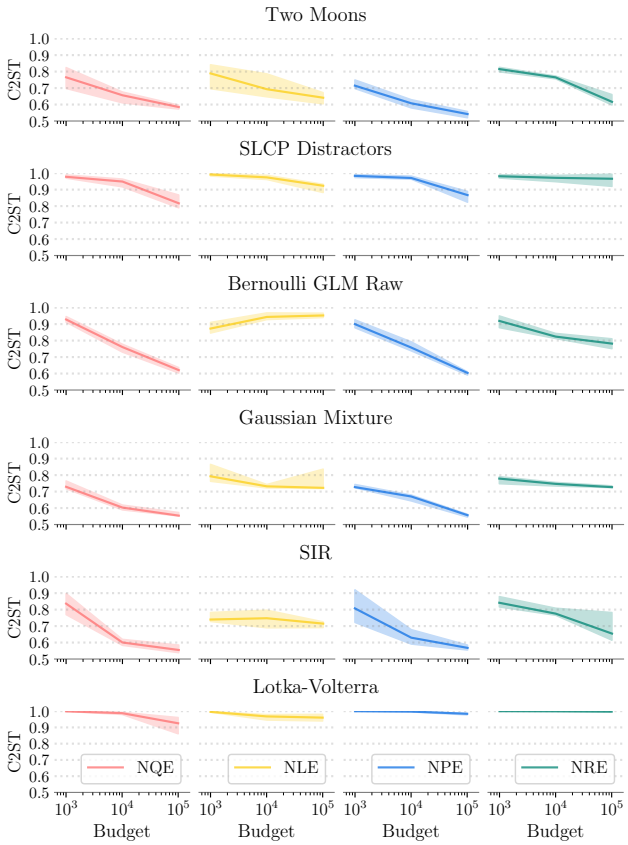

Figure 4. Comparison of C2ST as a function of simulation budget for the six benchmark problems, with lower C2ST values representing better performance of the algorithm. The error bars are estimated using the 25%, 50% and 75% quantiles of C2ST over ten realizations for each problem. (Uncalibrated) NQE achieves state-of-the-art performance across all the examples.

[8] See Appendix E for more discussion on this.

[9] For multi-modal distributions, we use the local quantile within the mode that contains the true $\theta^{(i)}$, similar to the definition of the $q-$coverage in Section 2.3.

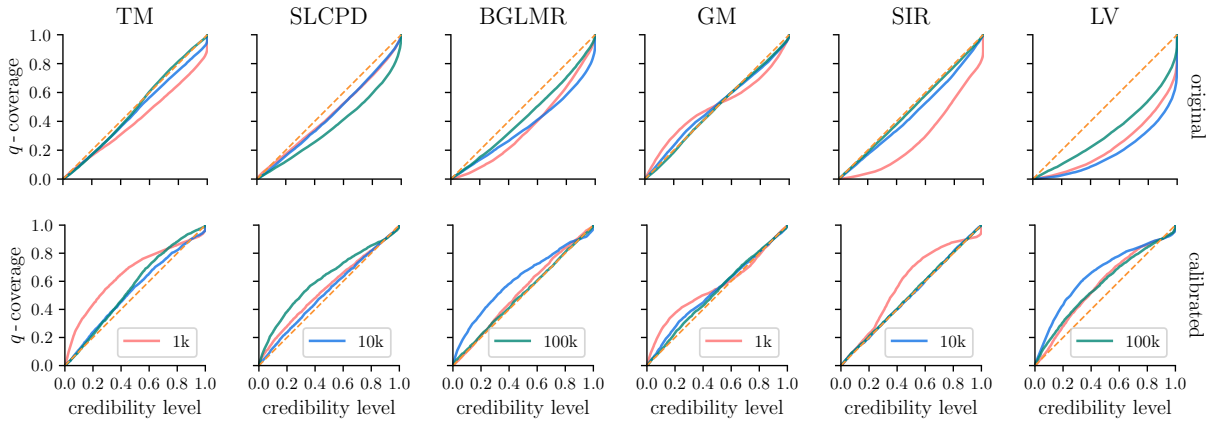

Figure 5. (Top) NQE $q-$coverage for the benchmark problems. Like other SBI methods, with limited simulation budgets, NQE may predict biased posteriors. (Bottom) Calibrated NQE predicts unbiased posteriors for all the problems. Errorbars are small and thus not plotted. See Appendix E for a convergence test and Figure 14 for a similar plot with $p-$coverage.

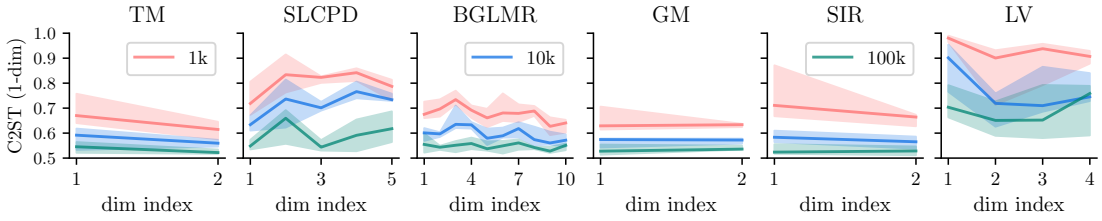

Figure 6. The C2ST values for the 1-dim *uncalibrated* marginal posteriors. We do not observe a clear trend of increasing C2ST with respect to the ordering of the dimensions.

3. Numerical Experiments

3.1. SBI Benchmark Problems

We assess the performance of NQE on six benchmark problems, with detailed specifications provided in Appendix C. All results for methods other than NQE are adopted from Lueckmann et al. (2021). As discussed in Appendix F, we conduct a mild search of hyperparameters for NQE, but in the end use the same set of hyperparameters across all the benchmark problems, although it is possible to further improve the performance by tuning the hyperparameters based on specific posterior structures. For example, increasing the number of predicted quantiles will be beneficial for multimodal problems with large simulation budgets. To evaluate the performance of SBI algorithms, we employ Classifier-based 2-Sample Testing (C2ST) as implemented in the `sbibm` package (Lopez-Paz & Oquab, 2016; Lueckmann et al., 2021). Lower C2ST values denote superior results, with 0.5 signifying a perfect posterior and 1.0 indicating complete failure.

We plot the C2ST results for the benchmark problems in Figure 4, showing that (uncalibrated) NQE achieves state-of-the-art performance across all the examples. In Figure 5, we compare the NQE $q-$coverage before and after broadening: with the broadening calibration, NQE consistently predicts unbiased posterior for all the problems. While Figure 5 utilizes 10^4 simulations to enhance the smoothness of the coverage curves, a convergence test in Appendix E shows that $\lesssim 10^3$ simulations are sufficient for most cases. The exact values of the broadening factor can be found in Figure 15. In Figure 16, we find that the C2ST is generally similar or slightly worse after the global broadening calibration: this is likely due to the nature of the C2ST metric, since a conservative posterior will be similarly penalized as a biased posterior, although the former should be preferred over the latter for most scientific applications (e.g. Hermans et al., 2021; Delaunoy et al., 2022).

3.2. Order of Model Parameters

Due to its autoregressive structure, NQE's performance may be affected by the order of $\boldsymbol{\theta}$ dimensions. While each 1-dim conditional distribution $p(\theta^{(i)} \,|\, \mathbf{x}, \boldsymbol{\theta}^{(j<i)})$ is estimated in-

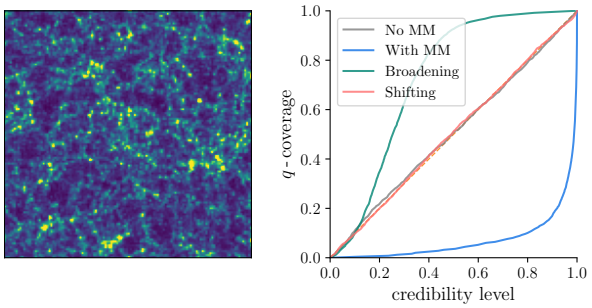

Figure 7. (Left) Sample image of the simulated data. The task is to infer two parameters of our Universe, Ω_m and σ_8, from such images. (Right) The $q-$coverage for uncalibrated NQE without model misspecification (No MM), uncalibrated NQE with model misspecification (With MM), and NQE with model misspecification but calibrated using a broadening factor of 4.2 (Broadening) and using the quantile shifting method (Shifting). Both calibration methods eliminate the bias due to *known* model misspecification, with quantile shifting achieving almost exact empirical coverage whereas global broadening being over-conservative.

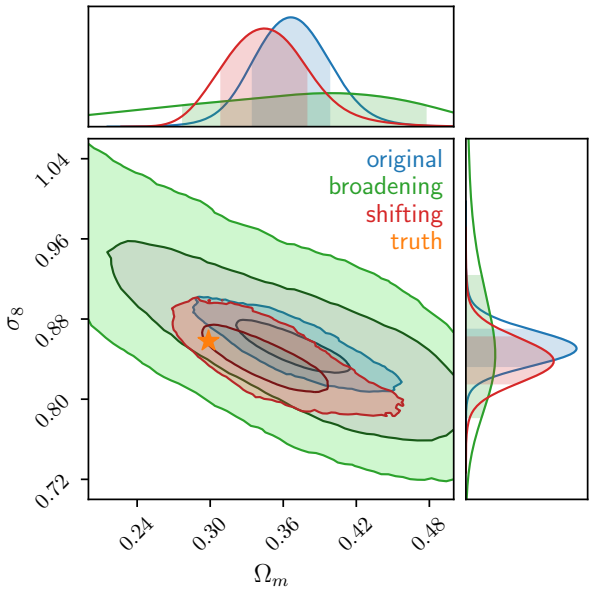

Figure 8. Comparison of the uncalibrated posterior and the posteriors calibrated with two different schemes. The quantile shifting scheme removes the bias without over-broadening the posterior.

dependently, the 1-dim marginal posterior $p(\theta^{(i)}|\mathbf{x})$ does depend on the estimation for all the previous $\boldsymbol{\theta}^{(j<i)}$ that are correlated with $\theta^{(i)}$, therefore one may expect the marginals for the latter dimensions to be less accurate than the former dimensions as the error will accumulate. To study this effect, we compute all the 1-dim marginal C2ST's for the benchmark problems and plot them with respect to the dimension indices in Figure 6. Contrary to the conjecture above, we find no clear dependence between the marginal C2ST and the dimension index. Nevertheless, this may be due to the relative low posterior dimensionality of the benchmark problems, such that the accumulation of per-dimension error has not become the dominant contribution. We still recommend ordering the $\boldsymbol{\theta}$ dimensions based on the relative importance of the parameters, especially for applications to higher ($\gtrsim 10$) dimensional posteriors. We note that similar to the TMNRE approach (Miller et al., 2021), one may estimate the individual marginal posteriors with NQE, if the high dimensionality makes it impractical to accurately model the joint posterior.

3.3. Application to Cosmology

The cosmological large scale structures contain ample information regarding the origin and future of our universe, which can be inferred from the locations and/or shapes of the galaxies (e.g. Dodelson & Schmidt, 2020), however the optimal strategy to extract the information remains an unsolved problem. While at larger scales the power spectra carry most of the information and can be well modeled with a Gaussian likelihood, at smaller scales the highly nonlinear evolution render SBI methods necessary for the optimal inference.

Unfortunately, the small-scale baryonic physics is still poorly understood, leading to potential model misspecification which can bias the SBI inference (e.g. Modi et al., 2023). As we do not know the exact forward model for our Universe, the best we can do is to make sure our SBI estimator is unbiased on all the well-motivated baryonic physics models, which requires a massive amount of expensive cosmological hydrodynamic simulations (e.g. Villaescusa-Navarro et al., 2021). However, with NQE one can first train it using cheap (therefore less realistic) simulations and then calibrate it using all available high fidelity (therefore much more expensive) simulations to make sure the uncertainties of baryonic physics have been properly accounted for [10]. Note that one only needs $\lesssim 10^3$ simulations for each baryonic model to calibrate NQE, which is far fewer than the amount required to directly train field-level SBI with them. Such approach is demonstrated in Figures 7 and 8, where we show that the bias due to model misspecification can be mitigated by the calibration of NQE. As the model misspecification introduces a large systematic bias, we find that

[10] Here we assume the model misspecification is at least *partially known*, in the sense that our selection of baryonic physics models "includes" the correct model for our Universe. The post-processing calibration cannot mitigate *completely unknown* model misspecification.

the global broadening calibration makes the posterior over-conservative while the quantile shifting scheme eliminates the bias without over-broadening the posterior, highlighting the benefits of such more advanced calibration methods that will be examined more thoroughly in a follow-up paper. More details regarding this example can be found in Appendix D.

4. Discussion

The main contribution of this work is to introduce Neural Quantile Estimation (NQE), a novel class of SBI methods that incorporate the concept of quantile regression, with competitive performance across various examples. Strictly speaking, our paper presents Neural Quantile *Posterior* Estimation, a method that can be extended to Neural Quantile *Likelihood* Estimation, which fits the likelihood $p(\mathbf{x}|\boldsymbol{\theta})$ with conditional quantiles. We note that the idea of interpolating predicted quantiles has been explored for e.g. time series forecasting (Gasthaus et al., 2019; Sun et al., 2023). Nonetheless, to our knowledge our paper is the first work that implements this idea in the SBI framework, with a dedicated interpolation scheme that minimizes the potential artifacts. In addition, Jeffrey & Wandelt (2020) uses a similar architecture to predict the moments of the posterior. Montel et al. (2023) proposes to autoregressively apply marginal NRE estimators to obtain the joint distribution, which outperforms standard NRE in their benchmarks.

As shown in Hermans et al. (2021), all existing SBI methods may predict biased results in practice: while the Bayesian optimal posterior has perfect calibration, there is no guarantee regarding the unbiasedness of SBI algorithms trained with insufficient number of simulations. However, with the post-processing calibration step, **NQE is guaranteed to be unbiased should there be no unknown model misspecification**, in the sense that the credible regions of the posterior will enclose no fewer samples samples than their corresponding credibility levels, as long as one has $\lesssim 10^3$ validation data to reliably compute the empirical coverage for the broadening calibration. While Balanced Neural Ratio Estimation (BNRE, Delaunoy et al., 2022) pursues similar goals of robust SBI inference, the unbiasedness of BNRE depends on the choice of their regularization parameter, so in principle they need to tune this parameter for each task to obtain best results. Unfortunately, the coverage evaluation is considerably more expensive for NRE methods which relies on MCMC sampling, making the coverage-based tuning of BNRE computationally prohibitive for higher dimensional applications. On the other hand, the broadening calibration of NQE can be applied with negligible computational cost, with the calibrated NQE manifestly unbiased as the empirical coverage has been explictly corrected during the broadening process. In addition, one can also mitigate the bias due to *partially known* model misspecification by calibrating the NQE posterior.

Before concluding this paper, we enumerate several promising directions for future study. First of all, NQE can be straightforwardly generalized to Sequential NQE (SNQE), which will be presented in a separate paper. Second, while our PCHIP-ET scheme shows competitive performance across various problems, it does not have continuous PDF derivatives, which may be improved by a higher order interpolation scheme. Moreover, in this work we mostly restrict to a global broadening transform for the calibration of NQE, which eliminates the bias at the cost of being possibly too conservative for certain credibility levels. As shown in Section 3.3, a more advanced calibration strategy would be useful, in particular for problems with a large systematic bias, so that one can calibrate biased posteriors without losing too much constraining power.

Acknowledgements

We thank Sihao Cheng, Biwei Dai, ChangHoon Hahn, Francois Lanusse, Jiaxuan Li, Yin Li, Peter Melchior, Chirag Modi, Nikhil Padmanabhan, Oliver Philcox and David Spergel for helpful discussions. The work presented in this article was performed on computational resources managed and supported by Princeton Research Computing, a consortium of groups including the Princeton Institute for Computational Science and Engineering (PICSciE) and the Office of Information Technology's High Performance Computing Center and Visualization Laboratory at Princeton University.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Beaumont, M. A., Zhang, W., and Balding, D. J. Approximate bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

Beaumont, M. A., Cornuet, J.-M., Marin, J.-M., and Robert, C. P. Adaptive approximate bayesian computation. *Biometrika*, 96(4):983–990, 2009.

Behnel, S., Bradshaw, R., Citro, C., Dalcin, L., Seljebotn, D. S., and Smith, K. Cython: The best of both worlds. *Computing in Science & Engineering*, 13(2):31–39, 2010.

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.

Dax, M., Wildberger, J., Buchholz, S., Green, S. R., Macke, J. H., and Schölkopf, B. Flow matching for scalable simulation-based inference. *arXiv preprint arXiv:2305.17161*, 2023.

Delaunoy, A., Hermans, J., Rozet, F., Wehenkel, A., and Louppe, G. Towards reliable simulation-based inference with balanced neural ratio estimation. *arXiv preprint arXiv:2208.13624*, 2022.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

Dodelson, S. and Schmidt, F. *Modern cosmology*. Academic press, 2020.

Durkan, C., Murray, I., and Papamakarios, G. On contrastive learning for likelihood-free inference. In *International Conference on Machine Learning*, pp. 2771–2781. PMLR, 2020.

Fritsch, F. N. and Carlson, R. E. Monotone piecewise cubic interpolation. *SIAM Journal on Numerical Analysis*, 17 (2):238–246, 1980.

Gasthaus, J., Benidis, K., Wang, Y., Rangapuram, S. S., Salinas, D., Flunkert, V., and Januschowski, T. Probabilistic forecasting with spline quantile function rnns. In *The 22nd international conference on artificial intelligence and statistics*, pp. 1901–1910. PMLR, 2019.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Grathwohl, W., Chen, R. T., Bettencourt, J., Sutskever, I., and Duvenaud, D. Ffjord: Free-form continuous dynamics for scalable reversible generative models. *arXiv preprint arXiv:1810.01367*, 2018.

Greenberg, D., Nonnenmacher, M., and Macke, J. Automatic posterior transformation for likelihood-free inference. In *International Conference on Machine Learning*, pp. 2404–2414. PMLR, 2019.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hermans, J., Begy, V., and Louppe, G. Likelihood-free mcmc with amortized approximate ratio estimators. In *International Conference on Machine Learning*, pp. 4239–4248. PMLR, 2020.

Hermans, J., Delaunoy, A., Rozet, F., Wehenkel, A., Begy, V., and Louppe, G. A trust crisis in simulation-based inference? your posterior approximations can be unfaithful. *arXiv preprint arXiv:2110.06581*, 2021.

Jeffrey, N. and Wandelt, B. D. Solving high-dimensional parameter inference: marginal posterior densities & moment networks. *arXiv preprint arXiv:2011.05991*, 2020.

Jiang, B., Wu, T.-y., Zheng, C., and Wong, W. H. Learning summary statistic for approximate bayesian computation via deep neural network. *Statistica Sinica*, pp. 1595–1618, 2017.

Kermack, W. O. and McKendrick, A. G. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721, 1927.

Koenker, R. and Bassett Jr, G. Regression quantiles. *Econometrica: journal of the Econometric Society*, pp. 33–50, 1978.

Li, Y., Lu, L., Modi, C., Jamieson, D., Zhang, Y., Feng, Y., Zhou, W., Kwan, N. P., Lanusse, F., and Greengard, L. pmwd: A differentiable cosmological particle-mesh n-body library. *arXiv preprint arXiv:2211.09958*, 2022a.

Li, Y., Modi, C., Jamieson, D., Zhang, Y., Lu, L., Feng, Y., Lanusse, F., and Greengard, L. Differentiable cosmological simulation with adjoint method. *arXiv preprint arXiv:2211.09815*, 2022b.

Lopez-Paz, D. and Oquab, M. Revisiting classifier two-sample tests. *arXiv preprint arXiv:1610.06545*, 2016.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Lotka, A. J. Analytical note on certain rhythmic relations in organic systems. *Proceedings of the National Academy of Sciences*, 6(7):410–415, 1920.

Lueckmann, J.-M., Goncalves, P. J., Bassetto, G., Öcal, K., Nonnenmacher, M., and Macke, J. H. Flexible statistical inference for mechanistic models of neural dynamics. *Advances in neural information processing systems*, 30, 2017.

Lueckmann, J.-M., Bassetto, G., Karaletsos, T., and Macke, J. H. Likelihood-free inference with emulator networks. In *Symposium on Advances in Approximate Bayesian Inference*, pp. 32–53. PMLR, 2019.

Lueckmann, J.-M., Boelts, J., Greenberg, D., Goncalves, P., and Macke, J. Benchmarking simulation-based inference. In *International Conference on Artificial Intelligence and Statistics*, pp. 343–351. PMLR, 2021.

Maraun, D. Bias correction, quantile mapping, and downscaling: Revisiting the inflation issue. *Journal of Climate*, 26(6):2137–2143, 2013.

McElreath, R. *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC press, 2020.

Meinshausen, N. and Ridgeway, G. Quantile regression forests. *Journal of machine learning research*, 7(6), 2006.

Miller, B. K., Cole, A., Forré, P., Louppe, G., and Weniger, C. Truncated marginal neural ratio estimation. *Advances in Neural Information Processing Systems*, 34:129–143, 2021.

Modi, C., Pandey, S., Ho, M., Hahn, C., Blancard, B. R.-S., and Wandelt, B. Sensitivity analysis of simulation-based inference for galaxy clustering, 2023.

Moler, C. B. *Numerical computing with MATLAB*. SIAM, 2004.

Montel, N. A., Alvey, J., and Weniger, C. Scalable inference with autoregressive neural ratio estimation. *arXiv preprint arXiv:2308.08597*, 2023.

Papamakarios, G. and Murray, I. Fast ε-free inference of simulation models with bayesian conditional density estimation. *Advances in neural information processing systems*, 29, 2016.

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. Normalizing flows for probabilistic modeling and inference, arxiv e-prints. *arXiv preprint arXiv:1912.02762*, 2019a.

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

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., and Feldman, M. W. Population growth of human y chromosomes: a study of y chromosome microsatellites. *Molecular biology and evolution*, 16(12):1791–1798, 1999.

Radev, S. T., Mertens, U. K., Voss, A., Ardizzone, L., and Köthe, U. Bayesflow: Learning complex stochastic models with invertible neural networks. *IEEE transactions on neural networks and learning systems*, 33(4):1452–1466, 2020.

Rezende, D. and Mohamed, S. Variational inference with normalizing flows. In *International conference on machine learning*, pp. 1530–1538. PMLR, 2015.

Rodrigues, F. and Pereira, F. C. Beyond expectation: Deep joint mean and quantile regression for spatiotemporal problems. *IEEE transactions on neural networks and learning systems*, 31(12):5377–5389, 2020.

Säilynoja, T., Bürkner, P.-C., and Vehtari, A. Graphical test for discrete uniformity and its applications in goodness-of-fit evaluation and multiple sample comparison. *Statistics and Computing*, 32(2):32, 2022.

Simola, U., Cisewski-Kehe, J., Gutmann, M. U., and Corander, J. Adaptive approximate bayesian computation tolerance selection. *Bayesian analysis*, 16(2):397–423, 2021.

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.

Sun, R., Li, C.-L., Arik, S. Ö., Dusenberry, M. W., Lee, C.-Y., and Pfister, T. Neural spline search for quantile probabilistic modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 9927–9934, 2023.

Talts, S., Betancourt, M., Simpson, D., Vehtari, A., and Gelman, A. Validating bayesian inference algorithms with simulation-based calibration. *arXiv preprint arXiv:1804.06788*, 2018.

Tang, W., Shen, G., Lin, Y., and Huang, J. Nonparametric quantile regression: Non-crossing constraints and conformal prediction. *arXiv preprint arXiv:2210.10161*, 2022.

Tavaré, S., Balding, D. J., Griffiths, R. C., and Donnelly, P. Inferring coalescence times from dna sequence data. *Genetics*, 145(2):505–518, 1997.

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.

Villaescusa-Navarro, F., Anglés-Alcázar, D., Genel, S., Spergel, D. N., Somerville, R. S., Dave, R., Pillepich, A., Hernquist, L., Nelson, D., Torrey, P., et al. The camels project: Cosmology and astrophysics with machine-learning simulations. *The Astrophysical Journal*, 915 (1):71, 2021.

A. Piecewise Cubic Hermite Interpolating Polynomial with Exponential Tails

We interpolate the CDF of the conditional 1-dim distributions using the quantiles predicted by NQE. Our interpolation scheme is based on Piecewise Cubic Hermite Interpolating Polynomial (PCHIP, Fritsch & Carlson, 1980; Moler, 2004), which preserves the monotonicity of the input data and has continuous first order derivatives. The values of the interpolated function at the $k-$th and $(k+1)-$th nodes, y_k and y_{k+1}, match the values of the target function, while the derivatives, y'_k and y'_{k+1}, are given by the two-side scheme for non-boundary points,

$$\begin{aligned} &h_k \equiv x_{k+1} - x_k, \quad d_k \equiv (y_{k+1} - y_k)/h_k, \\ &y'_k = \begin{cases} 0, & d_k d_{k+1} \le 0 \\ \dfrac{w_1 + w_2}{w_1/d_{k-1} + w_2/d_k}, & d_k d_{k+1} > 0, \end{cases} \quad \text{where } w_1 = 2h_k + h_{k-1},\ w_2 = h_k + 2h_{k-1}. \end{aligned} \tag{6}$$

For boundary points, we use the following one-side scheme for the left end (similarly for the right end),

$$y'_1 = \frac{(2h_1 + h_2)d_1 - h_1 d_2}{h_1 + h_2}, \tag{7}$$

which however is clipped to $[p_{\rm tl} d_1,\ 3d_1]$ for $d_1 \ge 0$ and $[3d_1,\ p_{\rm tl} d_1]$ for $d_1 < 0$, with $p_{\rm tl}$ a hyperparameter typically set to 0.6. Note that for well-defined CDF data, one always has $d_k > 0$ in Equations (6) and (7). Fritsch & Carlson (1980) shows a sufficient condition for the interpolation to preserve monotonicity is $0 \le y'_k/d_k \le 3$ and $0 \le y'_{k+1}/d_k \le 3$ [11], which is satisfied by Equations (6) and (7).

With y_k, y_{k+1}, y'_k and y'_{k+1}, the interpolation gives

$$\begin{aligned} y_{\rm interp}(x) &= h_{00}(t) \times y_k + h_{10}(t) \times h_k y'_k + h_{01}(t) \times y_{k+1} + h_{11}(t) \times h_k y'_{k+1}, \quad \text{where} \\ h_{00}(t) &= 2t^3 - 3t^2 + 1, \\ h_{10}(t) &= t^3 - 2t^2 + t, \\ h_{01}(t) &= -2t^3 + 3t^2, \\ h_{11}(t) &= t^3 - t^2, \\ t &\equiv (x - x_k)/(x_{k+1} - x_k). \end{aligned} \tag{8}$$

As shown in Figure 2, this interpolation scheme generates notable artifacts in the PDF, due to the challenge posed by fitting polynomials to the exponentially declining tail of the probability density.

In response to this challenge, we fit the local distribution with Gaussian tails whenever necessary. In this regime, the fitting PDF is given by

$$p(x) = p_0\, e^{a(x-x_0)^2 + \frac{p'_0}{p_0}(x - x_0)}, \tag{9}$$

with $p(x_0) = p_0$ and its first derivative $p'(x_0) = p'_0$ continuous at the end point of the bin. We then solve the a parameter by requiring that the PDF has correct normalization within the bin, which can be computed via the following indefinite integrals. For $p'_0 \ne 0$, we have

$$\int p\, {\rm d}x = \begin{cases} \dfrac{\sqrt{\pi}}{2\sqrt{|a|}}\, p_0\, e^{-\frac{p'^2_0}{4|a|p_0^2}} \times {\rm erfi}\left[\dfrac{2|a|p_0(x - x_0) + p'_0}{2\sqrt{|a|}p_0}\right] + C, & a > 0 \\ \dfrac{p_0^2\, e^{\frac{p'_0}{p_0}(x - x_0)}}{p'_0} + C, & a = 0 \\ \dfrac{\sqrt{\pi}}{2\sqrt{|a|}}\, p_0\, e^{\frac{p'^2_0}{4|a|p_0^2}} \times {\rm erf}\left[\dfrac{2|a|p_0(x - x_0) - p'_0}{2\sqrt{|a|}p_0}\right] + C, & a < 0 \end{cases} \tag{10}$$

[11] Indeed 3 is the largest number for the criterion of this form.

while for $p'_0 = 0$,

$$\int p\,\mathrm{d}x = \begin{cases} \dfrac{\sqrt{\pi}}{2\sqrt{|a|}}\, p_0 \times \mathrm{erfi}\left[\sqrt{|a|}(x - x_0)\right] + C, & a > 0 \\ p_0(x - x_0) + C, & a = 0 \\ \dfrac{\sqrt{\pi}}{2\sqrt{|a|}}\, p_0 \times \mathrm{erf}\left[\sqrt{|a|}(x - x_0)\right] + C, & a < 0 \end{cases} \tag{11}$$

where $\mathrm{erf}(\cdot)$ and $\mathrm{erfi}(\cdot)$ are the error function and imaginary error function, respectively. For $a \neq 0$ and $p'_0 \neq 0$, we use the following expressions which are analytically equivalent but numerically more stable,

$$\int_0^h p\,\mathrm{d}x = \begin{cases} \dfrac{p_0}{\sqrt{|a|}} \left\{ e^{|a|h^2 + p'_0 h/p_0} D\left[\dfrac{2|a|p_0 h + p'_0}{2\sqrt{|a|}p_0}\right] - D\left[\dfrac{p'_0}{2\sqrt{|a|}p_0}\right] \right\}, & a > 0 \\ \dfrac{\sqrt{\pi}\, p_0}{2\sqrt{|a|}} \left\{ e^{-|a|h^2 + p'_0 h/p_0} \mathrm{erfcx}\left[\dfrac{p'_0 - 2|a|p_0 h}{2\sqrt{|a|}p_0}\right] - \mathrm{erfcx}\left[\dfrac{p'_0}{2\sqrt{|a|}p_0}\right] \right\}, & a < 0 \end{cases} \tag{12}$$

where $D(\cdot)$ is Dawson's integral and $\mathrm{erfcx}(\cdot)$ is the scaled complementary error function. Nonetheless, in rare cases where $a < 0$ we set $a = 0$ and give up the continuity condition for the first derivative of PDF, and instead solve p'_0 for the correct normalization within the bin.

Our criterion for deciding whether a bin should be fitted with exponential tails is as follows. First of all, the leftmost and rightmost bins have one-sided exponential tails as long as their averaged PDF is smaller than 0.6 times the averaged PDF in the bin next to them, otherwise the edge bins likely have a hard truncation by the prior and are therefore fitted with polynomials. In addition, we also allow other bins to have double, i.e. $p_{\mathrm{exp}}^{(\mathrm{left})}$ from left endpoint x_k towards right and $p_{\mathrm{exp}}^{(\mathrm{right})}$ from right endpoint x_{k+1} towards left, exponential tails to account for potential multimodality. For each bin $[x_k, x_{k+1}]$, we attempt to fit the distribution with double exponential tails, and compute

$$f_{\mathrm{split}} \equiv \max[\, p_{\mathrm{exp}}^{(\mathrm{left})}(x_{k+1}) \,/\, p_0(x_{k+1}),\, p_{\mathrm{exp}}^{(\mathrm{right})}(x_k) \,/\, p_0(x_k)\,]. \tag{13}$$

Note that the PDF is no longer strictly continuous at the bin endpoints when fitted with double exponential tails, and f_{split} quantifies such discontinuity. We then switch to double exponentials only for bins with local minimum $f_{\mathrm{split}} < 0.01$, and stick with the PCHIP polynomials for the remaining bins. The rationale behind this is intuitive: a smaller f_{split} implies a likely gap between two isolated peaks of the PDF (see, for instance, the top right panel of Figure 2), which can be better fitted with two exponential tails extending from both sides. Our PCHIP-ET scheme incorporates the inductive bias that for most SBI problems the tails of probabilistic distributions can be well modeled by Gaussians; if this is not the case, one may replace the Gaussian with e.g. student's t or Cauchy for long-tailed distributions.

B. Weights in $\mathcal{L}_0$

In this work, we use NQE to predict the quantiles equally spaced between $[0,\, 1]$, which tends to put more emphasis on the regions with larger PDF where the neighboring quantiles are closer to each other, leading to potential instability in the tail regions. Instead of directly weighting the different terms in $\mathcal{L}_0$, we adopt the following dropout strategy: for each training batch, we only keep $0 < p_0 \leq 1$ of the terms in $\mathcal{L}_0$ using a no-replacement multinomial sampling with weights proportional to $\langle p \rangle_{\mathrm{avg}}^{-f_0}$, $\langle p \rangle_{\mathrm{avg}} \equiv (\langle p \rangle_{\mathrm{left}} + \langle p \rangle_{\mathrm{right}})/2$, with $p_0 = 0.5$ and $f_0 = 1$ by default. This will effectively upweight the quantiles where the PDF is small, while the no-replacement sampling prevents specific terms from having too large weights that dominate the whole loss function.

C. Benchmark Problems

We use the following problems from Lueckmann et al. (2021) to benchmark the performance of the SBI methods. The "ground truth" posterior samples are available for all the problems.

C.1. Two Moons (TM)

A toy problem with complicated global (bimodality) as well as local (crescent shape) structures.

Prior $\mathcal{U}(-\mathbf{1}, \mathbf{1})$

Simulator $\boldsymbol{x}|\boldsymbol{\theta} = \begin{bmatrix} r\cos(\alpha) + 0.25 \\ r\sin(\alpha) \end{bmatrix} + \begin{bmatrix} -|\theta_1 + \theta_2|/\sqrt{2} \\ (-\theta_1 + \theta_2)/\sqrt{2} \end{bmatrix}$, where $\alpha \sim \mathcal{U}(-\pi/2, \pi/2)$, $r \sim \mathcal{N}(0.1, 0.01^2)$

Dimensionality $\boldsymbol{\theta} \in \mathbb{R}^2, \mathbf{x} \in \mathbb{R}^2$

References Greenberg et al. (2019)

C.2. SLCP with Distractors (SLCPD)

A challenging problem designed to have a simple likelihood and a complex posterior, with uninformative dimensions (distractors) added to the observation.

Prior $\mathcal{U}(-\mathbf{3}, \mathbf{3})$

Simulator $\mathbf{x}|\boldsymbol{\theta} = (\mathbf{x}_1, \ldots, \mathbf{x}_{100})$, $\mathbf{x} = p(\mathbf{y})$,

where p re-orders the dimensions of $\mathbf{y}$ with a fixed random permutation;

$\mathbf{y}_{[1:8]} \sim \mathcal{N}(\boldsymbol{m}_{\boldsymbol{\theta}}, \boldsymbol{S}_{\boldsymbol{\theta}})$, $\mathbf{y}_{[9:100]} \sim \frac{1}{20}\sum_{i=1}^{20} t_2(\boldsymbol{\mu}^i, \boldsymbol{\Sigma}^i)$,

where $\boldsymbol{m}_{\boldsymbol{\theta}} = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}$, $\boldsymbol{S}_{\boldsymbol{\theta}} = \begin{bmatrix} s_1^2 & \rho s_1 s_2 \\ \rho s_1 s_2 & s_2^2 \end{bmatrix}$, $s_1 = \theta_3^2, s_2 = \theta_4^2, \rho = \tanh\theta_5$, $\boldsymbol{\mu}^i \sim \mathcal{N}(0, 15^2\boldsymbol{I})$,

$\boldsymbol{\Sigma}^i_{j,k} \sim \mathcal{N}(0, 9)$ for $j > k$, $\boldsymbol{\Sigma}^i_{j,j} = 3e^a$ with $a \sim \mathcal{N}(0, 1)$, $\boldsymbol{\Sigma}^i_{j,k} = 0$ otherwise

Dimensionality $\boldsymbol{\theta} \in \mathbb{R}^5, \mathbf{x} \in \mathbb{R}^{100}$

References Greenberg et al. (2019)

C.3. Bernoulli GLM Raw (BGLMR)

Inference of a 10-parameter Generalized Linear Model (GLM) with raw Bernoulli observations.

Prior $\beta \sim \mathcal{N}(0, 2)$, $\mathbf{f} \sim \mathcal{N}(\mathbf{0}, (\boldsymbol{F}^\top\boldsymbol{F})^{-1})$,

$\boldsymbol{F}_{i,i-2} = 1$, $\boldsymbol{F}_{i,i-1} = -2$, $\boldsymbol{F}_{i,i} = 1 + \sqrt{\frac{i-1}{9}}$, $\boldsymbol{F}_{i,j} = 0$ otherwise, $1 \le i, j \le 9$

Simulator $\mathbf{x}|\boldsymbol{\theta} = (\mathbf{x}_1, \ldots, \mathbf{x}_{100})$, $x_i \sim \text{Bern}(\eta(\mathbf{v}_i^\top\mathbf{f} + \beta))$, $\eta(\cdot) = \exp(\cdot)/(1 + \exp(\cdot))$

frozen input between time bins $i - 8$ and i: $\boldsymbol{v}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$

Dimensionality $\boldsymbol{\theta} \in \mathbb{R}^{10}, \mathbf{x} \in \mathbb{R}^{100}$

Fixed Parameters Duration of task $T = 100$

C.4. Gaussian Mixture (GM)

Inferring the common mean of a mixture of two Gaussians, one with much broader covariance than the other.

Prior $\mathcal{U}(-\mathbf{10}, \mathbf{10})$

Simulator $\mathbf{x}|\boldsymbol{\theta} \sim 0.5\,\mathcal{N}(\mathbf{x}|\boldsymbol{m}_{\boldsymbol{\theta}} = \boldsymbol{\theta}, \boldsymbol{S} = \boldsymbol{I}) + 0.5\,\mathcal{N}(\mathbf{x}|\boldsymbol{m}_{\boldsymbol{\theta}} = \boldsymbol{\theta}, \boldsymbol{S} = 0.01 \odot \boldsymbol{I})$

Dimensionality $\boldsymbol{\theta} \in \mathbb{R}^2, \mathbf{x} \in \mathbb{R}^2$

References Sisson et al. (2007); Beaumont et al. (2009); Toni et al. (2009); Simola et al. (2021)

C.5. SIR

An epidemiological model describing the numbers of individuals in three possible states: susceptible S, infectious I, and recovered or deceased R.

Prior	$\beta \sim \text{LogNormal}(\log(0.4), 0.5),\ \gamma \sim \text{LogNormal}(\log(1/8), 0.2)$
Simulator	$\mathbf{x}\|\boldsymbol{\theta} = (x_1, \dots, x_{10}),\ x_i \sim \mathcal{B}(1000, \frac{I}{N}),$ where I is simulated from $\frac{dS}{dt} = -\beta\frac{SI}{N}$ $\frac{dI}{dt} = \beta\frac{SI}{N} - \gamma I$ $\frac{dR}{dt} = \gamma I$
Dimensionality	$\boldsymbol{\theta} \in \mathbb{R}^2, \mathbf{x} \in \mathbb{R}^{10}$
Fixed Parameters	Population size $N = 1000000$ and duration of task $T = 160$ Initial conditions: $(S(0), I(0), R(0)) = (N-1, 1, 0)$
References	Kermack & McKendrick (1927)

C.6. Lotka-Volterra (LV)

An influential ecology model describing the dynamics of two interacting species.

Prior	$\alpha \sim \text{LogNormal}(-0.125, 0.5),\ \beta \sim \text{LogNormal}(-3, 0.5),$ $\gamma \sim \text{LogNormal}(-0.125, 0.5),\ \delta \sim \text{LogNormal}(-3, 0.5)$
Simulator	$\mathbf{x}\|\boldsymbol{\theta} = (\mathbf{x}_1, \dots, \mathbf{x}_{10}),\ \mathbf{x}_{1,i} \sim \text{LogNormal}(\log(X), 0.1),\ \mathbf{x}_{2,i} \sim \text{LogNormal}(\log(Y), 0.1),$ X and Y are simulated from $\frac{dX}{dt} = \alpha X - \beta XY$ $\frac{dY}{dt} = -\gamma Y + \delta XY$
Dimensionality	$\boldsymbol{\theta} \in \mathbb{R}^4, \mathbf{x} \in \mathbb{R}^{20}$
Fixed parameters	Duration of task $T = 20$ Initial conditions: $(X(0), Y(0)) = (30, 1)$
References	Lotka (1920)

D. Details of the Cosmology Application

We run 10^4 dark-matter-only Particle Mesh (PM) simulations with 128^3 particles in 256^3 Mpc/h^3 boxes using the `pmwd` code (Li et al., 2022a;b), and generate two 128^2 projected overdensity fields $\delta(\boldsymbol{x}) \equiv \rho(\boldsymbol{x})/\bar{\rho}$ from each simulation by dividing the box into two halves along the z axis as the observation data. We use 80% simulations for training, 10% for validation, and 10% for test. We evaluate the calibration of NQE with the validation data, and plot Figures 7 and 8 with the test data. The model parameters are Ω_m, the total matter density today, and σ_8, the RMS matter fluctuation today in linear theory, with uniform priors $0.1 \leq \Omega_m \leq 0.5$ and $0.5 \leq \sigma_8 \leq 1.1$.

As a proof-of-concept example, we substitute the expensive cosmological hydrodynamic simulations with a post-processing scale-independent bias [12] model over the density fields from the dark-matter-only simulations, i.e. $\delta(\boldsymbol{x}) \to b\,\delta(\boldsymbol{x})$ with $b = 1.02$ [13]. In other words, we train NQE with $b = 1$ simulations but requires the inference to be unbiased for $b = 1.02$, which is achieved via the calibration of NQE. A ResNet (He et al., 2016) with 10 convolutional layers is utilized as the embedding network for a more efficient inference with the high dimensional data.

[12] Here the *bias* means any deviation of the actual observed field with respect to the dark-matter-only density field.
[13] But we still require that $\delta(\boldsymbol{x}) > 0$.

E. Convergence Test of Coverage Evaluation

We check the convergence of the $q-$coverage evaluation in Figures 9 to 11. While Figure 5 in the main paper uses 10^4 simulations to enhance the smoothness of the coverage curves, in most cases $\lesssim 10^3$ simulations should be sufficient for the evaluation of $q-$coverage. Actually, the (pointwise) standard error of empirical coverage $p_{\rm ec}$ can be estimated using the properties of binomial distribution as $\Delta p_{\rm ec} = \sqrt{p_{\rm ec}\left(1-p_{\rm ec}\right)/N_o}$, where N_o is the number of simulations for the coverage evaluation (Säilynoja et al., 2022). Therefore, with $N_o = 10^3$, one has $\Delta p_{\rm ec} < 1.6\%$ for all $p_{\rm ec} \in [0,\, 1]$.

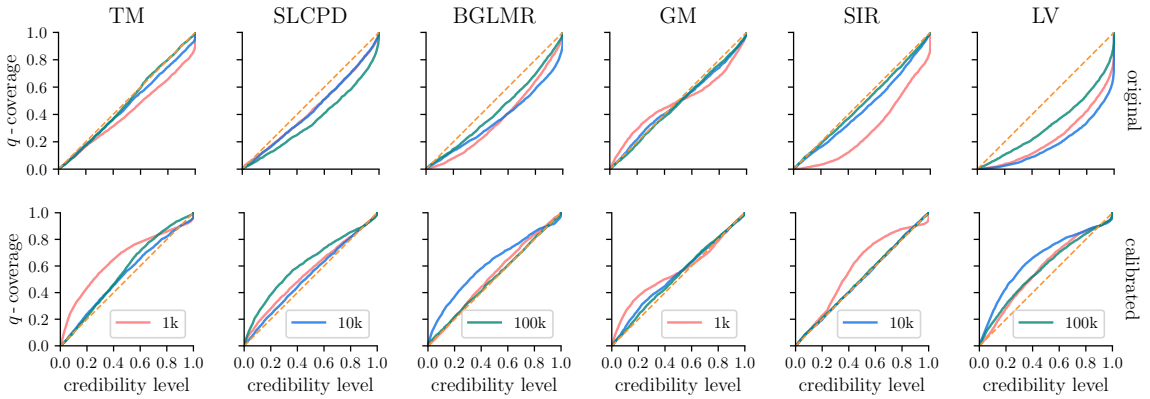

Figure 9. Similar to Figure 5, but using 2,000 simulations for the evaluation of $q-$coverage.

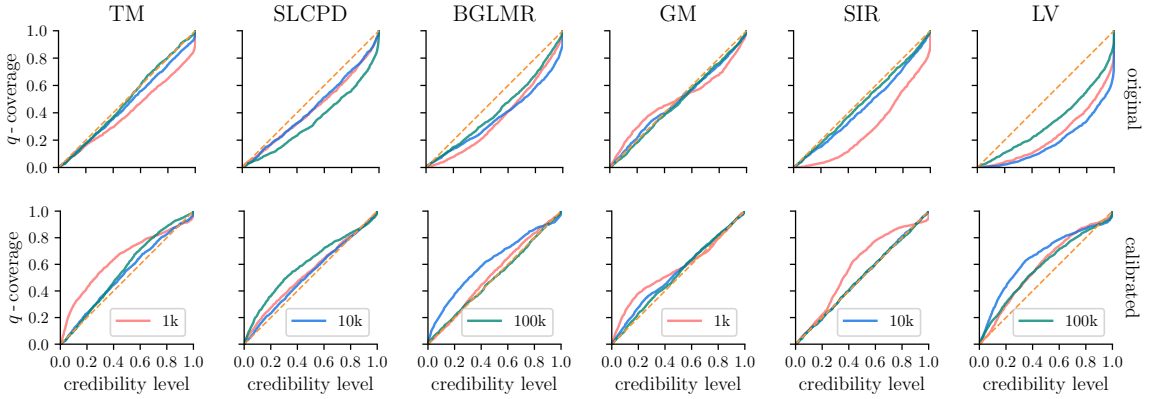

Figure 10. Similar to Figure 5, but using 1,000 simulations for the evaluation of $q-$coverage.

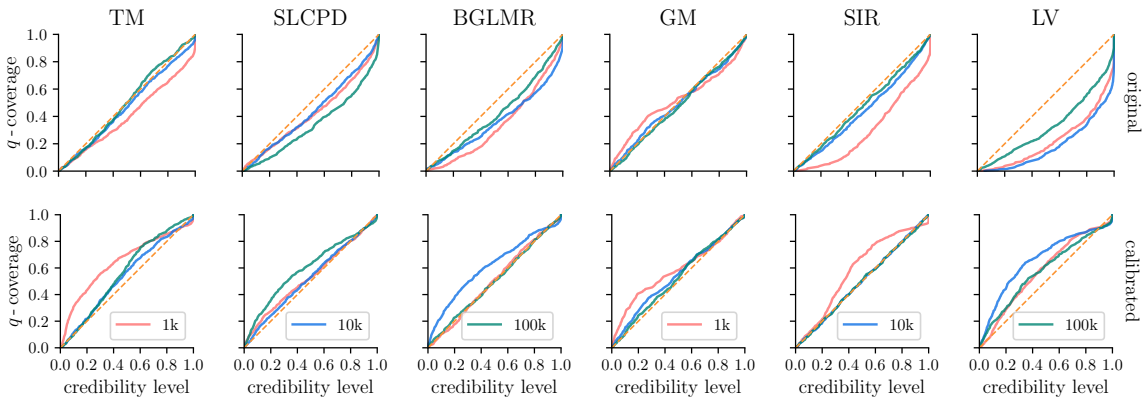

Figure 11. Similar to Figure 5, but using 500 simulations for the evaluation of $q-$coverage.

F. Hyperparameter Choices

Table 2. Our baseline choice of NQE hyperparameters.

hyperparameter	value
$p_{\rm tl}$	0.6
p_0	0.5
f_0	1.
f_1	1.1
f_2	0.8
$\lambda_{\rm reg}$	0.1
# of MLP hidden layers	10
# of MLP hidden neurons per layer	512
$n_{\rm bin}$	16

We train all the models on NVIDIA A100 MIG GPUs using the AdamW optimizer (Loshchilov & Hutter, 2017), and find the wall time of NQE training to be comparable to existing methods like NPE. Our PCHIP-ET scheme has been implemented with Cython (Behnel et al., 2010), so that its evaluation is much faster than the quantile regression network calls for typical real-world examples. We conduct a mild search for $\{f_0, \lambda_{\rm reg}, n_{\rm bin}\}$ in Figures 12 and 13, which leads to our baseline choice in Table 2. We reduce the stepsize by 10% after every 5 epochs, and terminate the training if the loss does not improve after 30 epochs or when the training reaches 300 epochs.

We find that some tasks require a different stepsize while some tasks exhibit significant overfitting, so we train 9 realizations for each network with $\{$initial step size = 5e-4, 1e-4, 2e-5$\} \times \{$AdamW weight decay = 0, 1, 10$\}$, and choose the realization with the smallest loss function. Nevertheless, most problems have a clear preference regarding these two parameters so it should be straightforward to tune them for specific problems in practice.

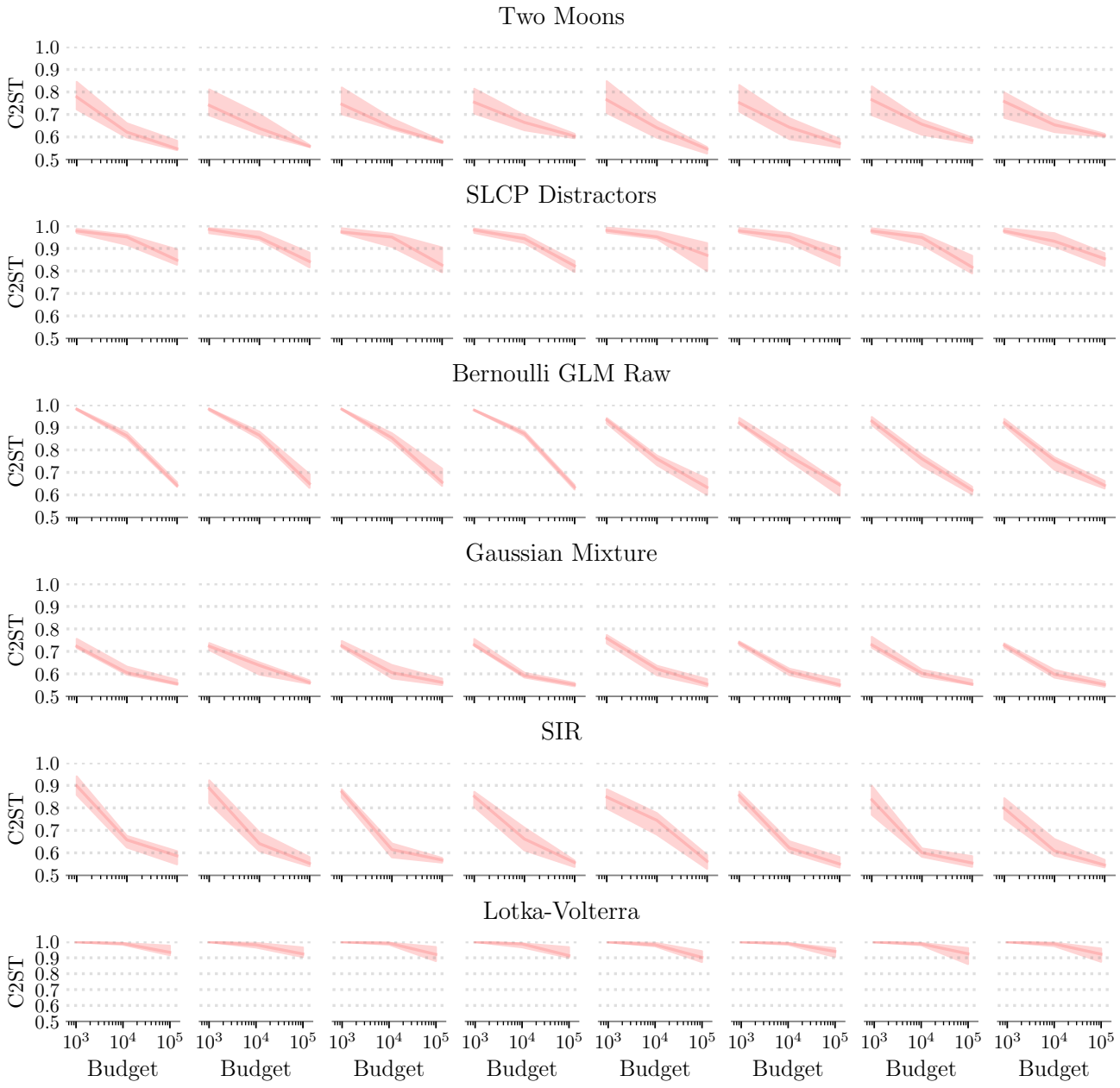

Figure 12. A survey of NQE performance across different choices of hyperparameters. From left to right, we set f_0 as (0, 0, 0, 0, 1, 1, 1, 1), and set λ_{reg} as (0, 0.01, 0.1, 1, 0, 0.01, 0.1, 1). All other parameters are the same as Table 2.

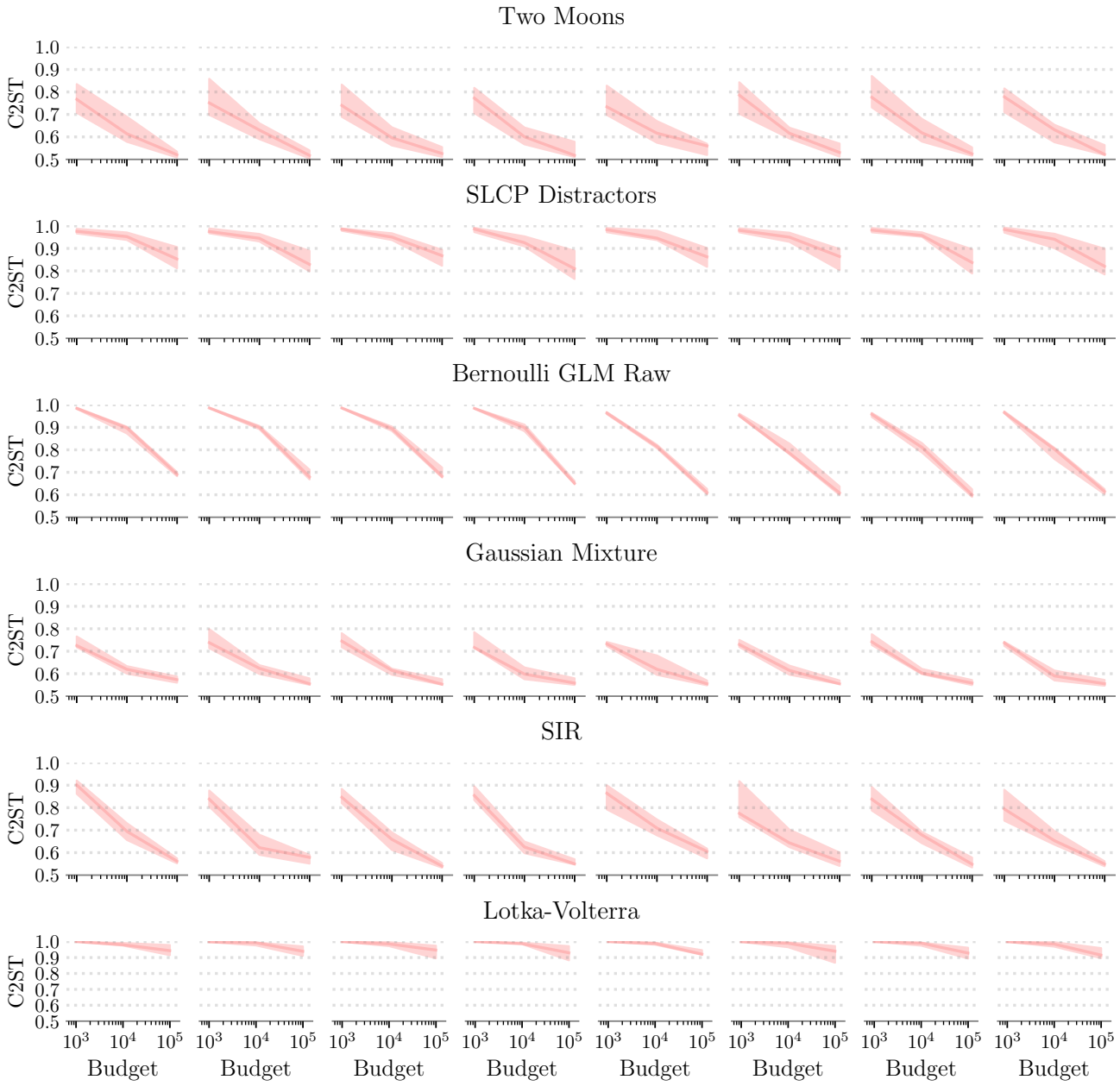

Figure 13. Same as Figure 12, but using 25 quantile bins. Increasing the number of bins is helpful for multimodal problems (e.g. TM) with large simulation budgets.

G. Additional Plots

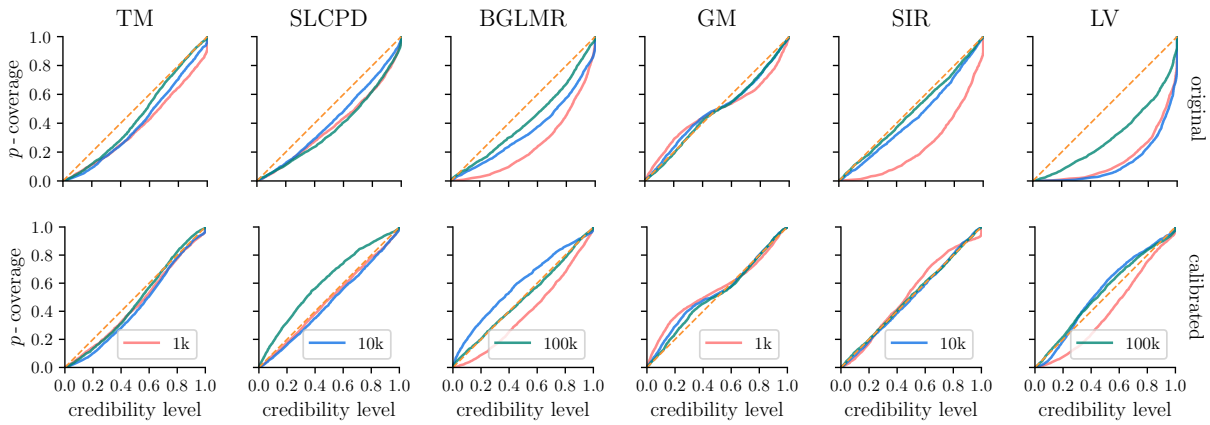

Figure 14. Empirical coverage results using $p-$coverage, while the calibration is still evaluated using $q-$coverage. We find that the $p-$coverage results are qualitatively similar to the $q-$coverage in most cases, and the broadening calibration with $q-$coverage in the main text also mitigates the bias for the $p-$coverage. Nevertheless, one can always solve the broadening factor directly with $p-$coverage if one wishes the $p-$coverage to be strictly unbiased, at the cost of more network calls required than using $q-$coverage.

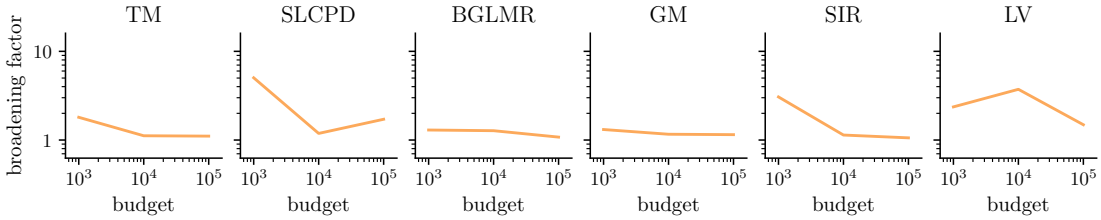

Figure 15. The actual broadening factor applied to remove the bias for the benchmark problems.

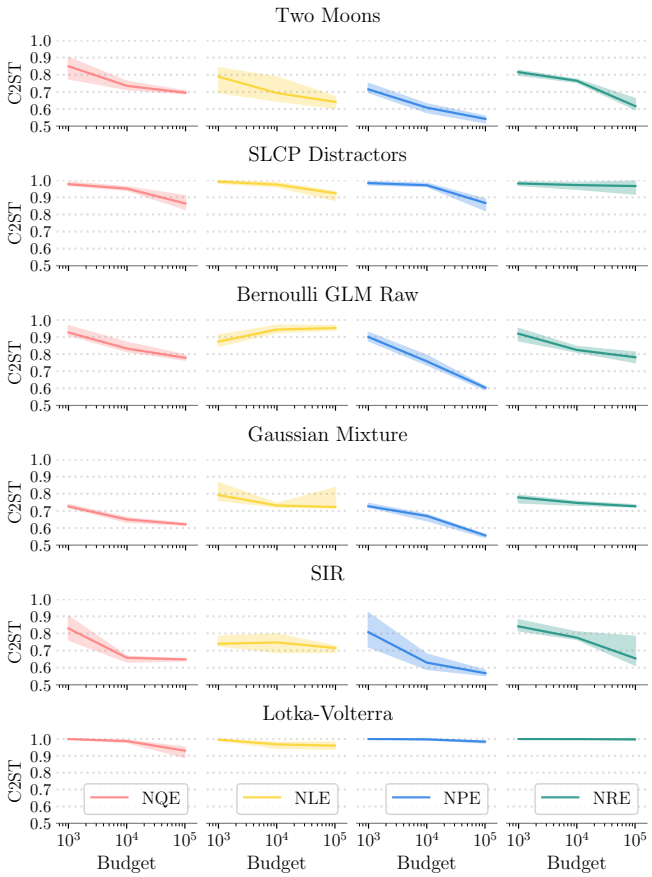

Figure 16. Similar to Figure 4, but for NQE calibrated with the global broadening scheme. The C2ST of calibrated NQE is generally similar to or slightly worse than uncalibrated NQE in Figure 4.