Mutual information maximization for amortized likelihood inference from sampled trajectories: MINIMALIST

Giulio Isacchini ®,[1,2,*] Natanael Spisak,[1,*] Armita Nourmohammad ®,[2,3,4,†] Thierry Mora ®,[1,†] and Aleksandra M. Walczak[1,†]

[1]*Laboratoire de Physique de l'École Normale Supérieure, CNRS, PSL University, Sorbonne Université,*
and Université Paris Cité, 75005 Paris, France
[2]*Max Planck Institute for Dynamics and Self-organization, Am Faßberg 17, 37077 Göttingen, Germany*
[3]*Department of Physics, University of Washington, 3910 15th Avenue Northeast, Seattle, Washington 98195, USA*
[4]*Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N, Seattle, Washington 98109, USA*

Simulation-based inference enables learning the parameters of a model even when its likelihood cannot be computed in practice. One class of methods uses data simulated with different parameters to infer models of the likelihood-to-evidence ratio, or equivalently the posterior function. Here we frame the inference task as an estimation of an energy function parametrized with an artificial neural network. We present an intuitive approach, named MINIMALIST, in which the optimal model of the likelihood-to-evidence ratio is found by maximizing the likelihood of simulated data. Within this framework, the connection between the task of simulation-based inference and mutual information maximization is clear, and we show how several known methods of posterior estimation relate to alternative lower bounds to mutual information. These distinct objective functions aim at the same optimal energy form and therefore can be directly benchmarked. We compare their accuracy in the inference of model parameters, focusing on four dynamical systems that encompass common challenges in time series analysis: dynamics driven by multiplicative noise, nonlinear interactions, chaotic behavior, and high-dimensional parameter space.

I. INTRODUCTION

Model-based Bayesian inference relies on knowing the probabilistic description of a process. Traditional methods rely on computing the likelihood of the observed data given the model parameters in order to maximize or sample from the posterior. For many models, in particular with multiple interacting degrees of freedom or hidden variables, the likelihood function may be impractical to evaluate. In cases in which drawing data from the generative process is possible, simulation-based inference techniques can be used as a powerful alternative approach for characterizing the underlying model.

Population genetics provides many examples of such problems. The observed quantities in this context are often based on sequencing data and are "far" from the quantities described by population dynamics models: it is possible in principle to write down likelihood functions, but they typically depend on a number of hidden variables that need to be marginalized out, making their evaluation impractical. Approximate Bayesian computation (ABC) was first used for posterior inference in the context of population genetics [1], and since then numerous new approaches to simulation-based inference have been developed to answer particular questions of phylodynamics and sequencing data analysis.

More broadly, methods for simulation-based inference can be organized in two classes [2,3]. In the first class, observations and simulated data are compared within the inference process, as in the original ABC approach [4,5]. Methods belonging to the second class proceed in two stages. First, they use a large number of simulations to learn an approximate model for the likelihood function [6,7], or alternatively the posterior function [8], or the likelihood-to-evidence ratio [9–11] that is amortized over the simulation examples. The amortized model is then used to evaluate the likelihood of observations and to evaluate the posterior for model parameters. With theoretical developments in machine learning and improvements in computing power, this class of algorithms has seen a renewed interest in recent years.

Here, we model the likelihood-to-evidence ratio as a Boltzmann factor and use it to infer the corresponding energy function. We show that the maximum-likelihood estimation of this factor is equivalent to the maximization of a lower bound to mutual information between parameters of the simulation and the simulated data. We exploit this equivalence by testing different lower bounds to mutual information as objective functions for simulation-based inference and optimize the parameters of artificial neural networks to approximate the posterior. We compare this approach to a recently developed method [10,11], where a model of the likelihood-to-evidence
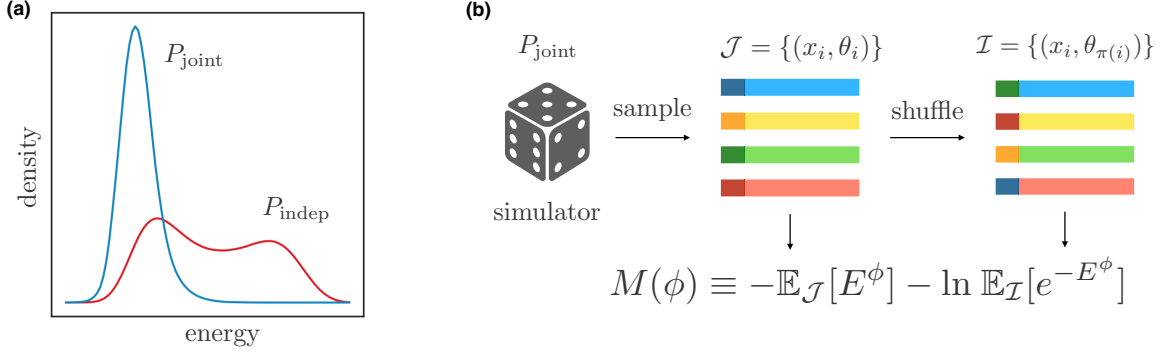
FIG. 1. (a) Distribution of energies of independent and joint pairs (x, θ). Pairs from the joint distribution have lower energy E, while pairs from the independent distribution have higher energy, as the majority of these independent samples are relatively unlikely under a joint model. (b) Schematic of the method. We first sample parameters θ from the prior $P(\theta)$ and then sample observations x from the simulator $P(x|\theta)$ to obtain pairs (x_i, θ_i). To generate pairs (x_i, θ_j) from the independent distribution, we shuffle the two initial vectors. Both sets are used to infer a model of energy E^ϕ by maximizing a log-likelihood in (5) with ϕ parameters of the artificial neural network used for the inference.

ratio is learned through the optimization of a binary classifier operating on simulated data and which has been shown to provide state-of-the-art performance on a variety of tasks [3]. We assess the accuracy of the methods to infer the parameters of four dynamical systems: the Ornstein-Uhlenbeck process, which is a multidimensional stochastic process with analytically tractable posterior; the birth-death process with multiplicative noise-driven dynamics; the susceptible-infected-recovered (SIR) model of epidemiology, which is a simple system with elementary nonlinear interactions; and the Lorenz attractor, which is a dynamical system that exhibits chaotic behavior. The task of finding the parameters of stochastic processes by maximizing the likelihood of observed discrete trajectories is generically difficult [12]. Our experiments encompass the most common challenges in the analysis of time-series data and together show that simulation-based inference offers a viable alternative to analytical methods.

II. METHODS

We aim to estimate parameters θ of a model given a set of data x obtained from stochastic simulation of that model, $P(x|\theta)$, with a prior $P(\theta)$. First, we reinterpret the likelihood-to-evidence ratio in terms of a Boltzmann factor,

$$\frac{P(x|\theta)}{P(x)} = \frac{P(x, \theta)}{P(x)P(\theta)} \equiv \frac{P_{\text{joint}}(x, \theta)}{P_{\text{indep}}(x, \theta)} = \frac{1}{Z} e^{-E(x,\theta)}, \quad (1)$$

where the energy function $E(x, \theta)$ captures how the joint distribution of data and parameters $P_{\text{joint}}(x, \theta) = P(x|\theta)P(\theta)$ deviates from the independent distribution $P_{\text{indep}}(x, \theta) = P(x)P(\theta)$, as shown schematically in Fig. 1(a), in which the pairs (x, θ) sampled from P_{joint} have lower energies than the samples from P_{indep}.

The energy $E(x, \theta)$ is generally a nonlinear function describing the dependence between data and parameters. Z is the partition function, which ensures that the probability density $P(x, \theta)$ is normalized. E and Z are each defined up to constants (additive for E, multiplicative for Z). Given the energy function $E(x, \theta)$, we recover the posterior probability density $P(\theta|x) = \frac{1}{Z} e^{-E(x,\theta)} P(\theta)$.

We will now describe an inference scheme to learn a model of the energy function from simulated data, relying on the flexibility of artificial neural networks. Specifically, we approximate the energy E by a multilayered network E^ϕ characterized by a set of parameters ϕ. Under a given model E^ϕ, the joint distribution is approximated as

$$P_{\text{joint}}^\phi(x, \theta) = \frac{1}{Z^\phi} e^{-E^\phi(x,\theta)} P_{\text{indep}}(x, \theta). \quad (2)$$

We simulate samples from the joint distribution, denoted $\mathcal{J} = \{(x_i, \theta_i)\}_{i=1}^N$, by drawing a model parameter from a prior distribution, $\theta_i \sim P(\theta)$, and simulating $x_i \sim P(x|\theta_i)$.

To learn the neural network parameters ϕ, we need to maximize the log-likelihood of the simulated sample $\mathcal{J}$ under a given model E^ϕ:

$$\mathcal{L}(\phi; \mathcal{J}) = N \mathbb{E}_{\mathcal{J}} \left[\ln P_{\text{joint}}^\phi \right]$$
$$= N(-\mathbb{E}_{\mathcal{J}}[E^\phi] - \ln Z^\phi + \mathbb{E}_{\mathcal{J}}[\ln P_{\text{indep}}]), \quad (3)$$

where $\mathbb{E}_{\mathcal{J}}[\cdot]$ denotes the empirical average over samples $\mathcal{J}$. The partition function is approximated using importance sampling on samples drawn from P_{indep}:

$$Z^\phi = \int e^{-E^\phi(x,\theta)} P_{\text{indep}}(x, \theta) dx \, d\theta \approx \mathbb{E}_{\mathcal{I}}[e^{-E^\phi}], \quad (4)$$

where $\mathbb{E}_{\mathcal{I}}[\cdot]$ is the counting measure over a large set $\mathcal{I}$ of independently drawn parameter and data pairs $(x, \theta) \sim P_{\text{indep}}(x, \theta) = P(\theta)P(x)$. In practice, $\mathcal{I}$ may be obtained by shuffling the indices of $\mathcal{J}$ [11], $\mathcal{I} = \{(x_i, \theta_{\pi(i)})\}$, where π is a random permutation of N elements, possibly multiple times. Counting all possible combinations, the set $\mathcal{I}$ can have maximal size $\max(N_{\mathcal{I}}) = N^2 - N$ under a fixed simulation budget. We denote the relative size of the two sets by $k = N_{\mathcal{I}}/N$.

With this estimate of Z^ϕ, and noting that the last term of Eq. (3) does not depend on ϕ, the problem is reduced to maximizing

$$M(\phi; \mathcal{I}, \mathcal{J}) \equiv -\mathbb{E}_{\mathcal{J}}[E^\phi] - \ln \mathbb{E}_{\mathcal{I}}[e^{-E^\phi}], \quad (5)$$

which in the infinite data limit constitutes a lower bound (the Donsker-Varadhan representation) to mutual information

$I(X; \Theta)$ between simulated data and simulation parameters,

$$M(\phi; \mathcal{I}, \mathcal{J}) \overset{N \to \infty}{=} \int P_{\text{joint}}(x, \theta) \ln \frac{P^\phi_{\text{joint}}(x, \theta)}{P_{\text{indep}}(x, \theta)} \, dx \, d\theta$$
$$= I(X; \Theta) - D_{\text{KL}}\left(P_{\text{joint}} \| P^\phi_{\text{joint}}\right)$$
$$\leqslant I(X; \Theta), \tag{6}$$

where $D_{\text{KL}}(P_{\text{joint}} \| P^\phi_{\text{joint}})$ is the Kullback-Leibler divergence between the true joint distribution and its model (2). This bound was extensively studied in [13] as an estimate of the mutual information from discrete samples drawn from joint distributions. Here, we will use this representation to learn the energy function E^ϕ, which approximates the likelihood-to-evidence ratio. We refer to this method as mutual information neural estimation (MINE). Its rationale is presented schematically in Fig. 1(b).

The connection to mutual information estimation established above opens the possibility of employing other empirical mutual information estimates to perform simulation-based inference. An alternative lower bound to $I(X; \Theta)$, first introduced in [14], is the so-called f-divergence representation (FDIV),

$$L_f(\phi; \mathcal{I}, \mathcal{J}) \equiv -\mathbb{E}_{\mathcal{J}}[E^\phi] - \mathbb{E}_{\mathcal{I}}[e^{-E^\phi - 1}]. \tag{7}$$

This estimator defines an alternative objective function to (5) that can be used to infer an optimal energy model E^ϕ. Note that in the limit of infinite data $N \to \infty$, and when the class of models $\{E^\phi\}_\phi$ can represent the true energy exactly, the maxima of (5) and (7) are both reached at the true value of $E(x, \theta)$ where they give the true value of $I(X; \Theta)$ (see Appendix A). Outside of this limit, using one of these objective functions may prove more beneficial. In particular, the second term of $L_f(\phi; \mathcal{I}, \mathcal{J})$ and its gradients may be reliably estimated by averaging over small batches, unlike the second term of $M(\phi; \mathcal{I}, \mathcal{J})$ because of the logarithm, giving FDIV an advantage for stochastic gradient descent algorithms. While the Donsker-Varadhan bound on the mutual information is tighter, i.e., $L_f(\phi; \mathcal{I}, \mathcal{J}) \leqslant M(\phi; \mathcal{I}, \mathcal{J})$ holds for $N \to \infty$ [13], it is unclear whether it might produce a more reliable estimate of $E(x, \theta)$.

A third alternative is to use the original approach for the likelihood-to-evidence ratio estimation proposed in [10] and [11]. The energy $E(x, \theta)$ may be rewritten in terms of a classifier between the two hypotheses of (x, θ) originating from the joint or independent distribution in a mixture $P_{\text{mix}} = \frac{1}{k+1} P_{\text{joint}} + \frac{k}{k+1} P_{\text{indep}}$ (in [11] $k = 1$). We define

$$d(x, \theta) \equiv P(\text{joint}|x, \theta) = \frac{P_{\text{joint}}(x, \theta)}{P_{\text{joint}}(x, \theta) + k P_{\text{indep}}(x, \theta)}$$
$$= \frac{1}{1 + kZe^{E(x,\theta)}}. \tag{8}$$

The classifier is parametrized by a neural network, $d = d^\phi$, and is trained by minimizing the binary cross-entropy,

$$S(\phi; \mathcal{I}, \mathcal{J}) = -\mathbb{E}_{\mathcal{J}}[\ln d^\phi] - k\mathbb{E}_{\mathcal{I}}[\ln(1 - d^\phi)]. \tag{9}$$

Similarly to objectives (5) and (7), in the $N \to \infty$ limit and when the class of $\{d^\phi\}_\phi$ models contains the true d, this cross-entropy is minimized at the true value $d(x, \theta)$. In [15] it

was shown that it can also be used as an estimator of mutual information by computing the mean logarithm of the predicted likelihood-to-evidence ratio, $\mathbb{E}_{\mathcal{J}}[\ln(kd^\phi/(1 - d^\phi))]$. Indeed, in the infinite data limit, all three objective functions, (5), (7), and (9), share the same optimum; see Appendix A. For high-dimensional random variables, binary cross-entropy (9) sets a tighter lower bound and a more accurate estimate of the mutual information than the f-divergence estimator (7) [15]. However, its accuracy was not directly compared to the proposed estimator in Eq. (5). We will refer to the inference approach based on minimizing the binary cross-entropy loss in Eq. (9) as BCE.

Finally, once an energy model E^ϕ has been trained by optimizing $M(\phi; \mathcal{I}, \mathcal{J})$, $L_f(\phi; \mathcal{I}, \mathcal{J})$, or $S(\phi; \mathcal{I}, \mathcal{J})$ over ϕ, the posterior of parameters given an observation x may be calculated as $P(\theta|x) = (1/Z^\phi)e^{-E^\phi(x,\theta)}P(\theta)$. We note that if the prior is changed, the energy function needs to be re-inferred. When θ is of high dimension, scanning the posterior for all possible values of θ may be impractical. In that case, we generate samples of θ from the posterior using a Markov-chain Monte Carlo method with Metropolis-Hasting acceptance probability:

$$\rho(\theta \to \theta') = \min\left(1, \frac{q(\theta'|\theta)P(\theta')}{q(\theta|\theta')P(\theta)} e^{-[E^\phi(x,\theta') - E^\phi(x,\theta)]}\right), \tag{10}$$

with q an ergodic Markov transition probability in the parameter space. This procedure generalizes in a straightforward way to the case of multiple observations drawn with the same set of parameters.

III. RELATIONS TO OTHER WORK

The presented methods are related to several recent approaches. For completeness, we discuss the similarities and differences between the presented and other methods.

A. Posterior inference methods

An alternative to likelihood-to-evidence ratio estimation (RE) is the framework of neural posterior estimation (NPE), which consists in fitting a conditional density estimator directly to the posterior. Several recently developed NPE algorithms [16–18] have been compared with the BCE method in [3] as part of a public benchmark of ABC, NPE, and RE methods across several simulation-based inference tasks. The comparison reveals that there is no single best algorithm, i.e., NPE and RE approaches yield similar performance and consistently outperform ABC. The benchmark we present below is limited to different RE methods and compares alternative loss functions.

B. Noise contrastive estimation

In [19], the NPE and the RE methods have been presented as two instances of a more general scheme, which the authors termed as *noise contrastive estimation*. The learning algorithm introduced in [19] is based on a multisample loss function that allows interpolating between the two approaches. Importantly, Ref. [19] also suggests a link between the field of mutual
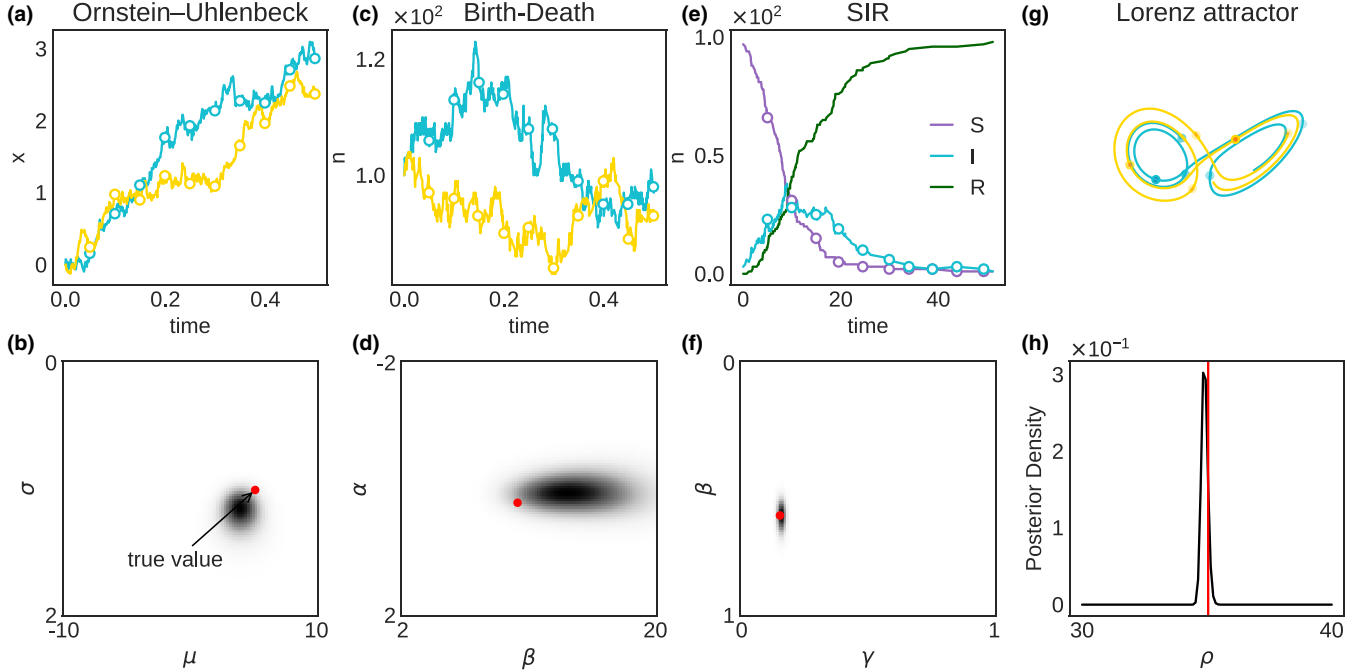
FIG. 2. Example trajectories (a,c,e,g) and posterior inference (b,d,f,h) for Ornstein-Uhlenbeck (a,b), birth-death (c,d), SIR (e,f), and Lorenz attractor (g,h). Trajectories are simulated with the parameter marked in red on the posterior plots. Circles indicate the discrete observations used for inference. Posteriors were estimated over 10 trajectories.

information estimation and simulation-based inference since the multisample loss function is also a lower bound to $I(X; \Theta)$. Here we formalize this link and test its applicability on a variety of examples by proposing two new methods (MINE and FDIV) for likelihood-to-evidence ratio estimation. We detail how our approach fits within the noise contrastive estimation scheme in Appendix B.

C. Learning optimal experimental designs and sufficient statistics

An important direction of research in simulation-based inference aims at finding experimental designs that are most informative about the model parameters [20,21]. In [22], an optimal design is found by maximizing mutual information between design choice and target variable of interest in the experiment. This task extends the original RE framework to optimize over an additional hidden variable, namely the design, that specifies the simulation setup.

In [23] mutual information is maximized to build sufficient summary statistics of the simulated data. This technique allows for automatically learning an optimal representation of the data without expert knowledge of process-specific observables. The statistics obtained this way can then generically be parsed as input to simulation-based inference methods. While the procedure to find summary statistics is similar to the MINE and BCE methods described here, the algorithm of [23] does not exploit the energy model for posterior inference. Instead, the model is discarded and the statistics are independently utilized with another simulation-based inference method.

Concurrently to the first version of this paper, a new development in this line of research [24] has been proposed. The

authors compare different mutual information lower bounds for Bayesian experimental design and observe a consistent trend with our findings, namely that the BCE method performs overall better than other estimators. It is an independent validation of our results using a closely related task.

IV. EXPERIMENTS

We set out to examine the three presented methods for estimating the likelihood-to-evidence ratio (MINE, FDIV, and BCE) to infer the parameters of simple dynamical models from discrete samples of their trajectories. We chose four contexts that together encompass the range of difficulties in the inference of model parameters: (i) the stochastic birth-death process, (ii) the epidemiological susceptible-infected-recovered (SIR) process, (iii) the multidimensional Ornstein-Uhlenbeck process, and (iv) the chaotic system of Lorenz attractor. Example trajectories of each model are shown in Fig. 2.

A. Ornstein-Uhlenbeck process

The Ornstein-Uhlenbeck (OU) process is a multidimensional Markov process driven by additive Gaussian white noise. It is applied in many branches of science, notably to describe the velocity of a Brownian particle [25], the fluctuations of interest rates [26], or the evolution of continuous phenotypic traits [27,28]. The trajectories are a solution of a stochastic differential equation:

$$dx = -\gamma(x - \mu)dt + \sqrt{2}\sigma \, dW, \qquad (11)$$

where μ is the stationary mean and γ is the damping matrix, assumed to be symmetric. W stands for the multidimensional

Wiener process, and σ is the noise amplitude. We use the Euler-Maruyama integration scheme to obtain the numerical solutions of this equation [29]. The corresponding Fokker-Planck equation for this process can be solved to obtain the true posterior (see Appendix C).

In one dimension, we infer the mean μ and the noise strength σ, setting $\gamma = 1$ and using uniform priors $P(\mu) = \mathcal{U}(-10, 10)$ and $P(\sigma) = \mathcal{U}(0, 2)$.

To study how the performance of each method scales with dimension, for $1 \leqslant d \leqslant 5$, we fix $\mu = 0, \sigma = \mathbb{I}$, where $\mathbb{I}$ is the identity matrix, and we infer the damping matrix parametrized as $\gamma = \mathbb{I} + \epsilon g$, where g is a Gaussian orthogonal matrix and $\epsilon < 1$, which ensures that the damping matrix is positive definite (see Appendix C for more details). Since γ is symmetric, in the d-dimensional case we have $\binom{d}{2}$ parameters to infer. The prior is given by the Gaussian orthogonal ensemble distribution density, $P(g) \propto \exp(-d \operatorname{Tr}(g^2)/4)$.

B. Birth-death process

The birth-death process is a discrete one-dimensional Markov process with multiplicative demographic noise. The number of individuals n is subject to variation due to stochastic birth and death events occurring at rates $n\lambda$ and $n\delta$, respectively,

$$n \xrightarrow{n\lambda} n + 1, \qquad n \xrightarrow{n\delta} n - 1. \qquad (12)$$

We use the Gillespie algorithm to sample trajectories from this process [30]. We parametrize the process with the average exponential drift $\alpha = \lambda - \delta$ and the noise timescale $\beta = \lambda + \delta$. We use uniform priors for both of these variables: $P(\alpha) = \mathcal{U}(-2, 2)$ and $P(\beta) = \mathcal{U}(2, 20)$.

C. SIR model

The SIR model is a staple of epidemiological modeling. Any member of the susceptible population S can be infected at a rate β upon contact with one of I infected individuals. The infected individuals can become resistant R at a rate γ:

$$S + I \xrightarrow{\beta} 2I, \qquad I \xrightarrow{\gamma} R. \qquad (13)$$
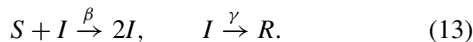
We simulate the trajectories of the SIR model using the Gillespie algorithm [30]. We infer the rates β and γ under uniform priors $P(\beta) = P(\gamma) = \mathcal{U}(0, 1)$ given samples from the (S, I) trajectories.

D. Lorenz attractor

The Lorenz system is a three-dimensional chaotic system governed by the equations

$$\dot{x} = \sigma(y - x), \quad \dot{y} = x(\rho - z) - y, \quad \dot{z} = xy - \beta z. \quad (14)$$

We simulate this deterministic process starting from a random position $(x_0 + \eta, y_0, z_0)$, where η is the noise in the initial position drawn from a uniform distribution, $\eta \sim \mathcal{U}(-0.1, 0.1)$. We fix the parameters $\sigma = 10$ and $\beta = 8/3$ and set out to infer ρ. The ensemble of trajectories starting in the vicinity of x_0 diverge with the characteristic time set by the inverse of the largest Lyapunov exponent of the system $\lambda = \lambda(\rho)$. We start sampling from the trajectories at a random initial time

drawn from a Gamma distribution, $t_0 \sim \Gamma(k = 5, \theta = 2)$. We then take five samples from each trajectory at time windows that are larger than the characteristic time for chaotic divergence, $\Delta t = 2\lambda^{-1}$. We set $\lambda \simeq 0.905$, which is the Lyapunov exponent for $\rho = 28$, a transition point where some but not all the solutions of the Lorenz system are chaotic. We infer the parameter ρ in a chaotic regime using a uniform prior $P(\rho) = \mathcal{U}(30, 40)$.

The artificial neural networks used for all three methods were multilayer perceptrons [31] with two hidden layers and a hyperbolic tangent activation function. This architecture choice was found to be expressive enough across tasks, and, thanks to its simplicity, we could perform a well-grounded comparison of the three methods without advanced regularization techniques (see Appendix D for details on hyperparameters choices). The methods were implemented using TENSORFLOW [32,33] with extensive use of Numpy [34] and Scipy [35] libraries.

V. RESULTS

Given enough data and a powerful enough neural network, we expect the optima of the objective functions I, L_f, and S to converge, and the estimated energy function should approach the true value.

To confirm the validity of the proposed methods, in the first set of experiments we use a large number of simulations to study the convergence of the posterior functions. To this end, we choose a hypothesis θ^* and simulate M trajectories, $x_{1:M} = \{x_m\}_{m=1}^M$, $x_m \sim P(x|\theta^*)$ with $M = 2$ for SIR, and $M = 5$ for the other tasks. We evaluate the posteriors

$$\hat{P}_l(\theta|x_{1:M}) = \frac{1}{(Z_l^\phi)^M} \exp\left[-\sum_{m=1}^M E_l^\phi(x_m, \theta)\right] P(\theta), \quad (15)$$

with l indexing one of the three methods (MINE, FDIV, or BCE) and ϕ is the optimal one for each method (we dropped the explicit dependency on ϕ in $\hat{P}_l$ for ease of notation). The posteriors converge when the amortized inference is done on a training set with at least $N = 10^7$ samples (and 10^7 samples for validation); see Fig. 3. We define a reference posterior $P^\infty(\theta|x_{1:M}) \equiv \langle \hat{P}_l(\theta|x_{1:M}) \rangle_l$, obtained with $N = 10^7$ as the average over three estimators. In the case of the Ornstein-Uhlenbeck process, where the true posterior $P(\theta|x_{1:M})$ can be calculated analytically, $P^\infty(\theta|x_{1:M})$ agrees with the analytical prediction. This first result confirms the validity of our approach.

With reducing sample size N, the amortized posteriors differ. To study the performance of the three methods under different simulation budgets N for each task, we simulate $N_{\text{tot}} = 2 \times 10^7$ samples $\mathcal{J} = \{(x_i, \theta_i)\}$. We perform the inference of the amortized likelihood-to-evidence ratio with varying simulation budgets, where both the training and the validation data are equal-sized subsamples of $\mathcal{J}$ with $N \in \{10^4, 10^5, 10^6\}$. Inference and comparison are performed 10 times on independent subsamples of $\mathcal{J}$. To obtain samples from the independent set $\mathcal{I}$, we shuffle the joint samples $k = 5$ times, and so $N_{\mathcal{I}} = 5N$. Larger shuffled data $N_{\mathcal{I}}$ can improve the inference but at the cost of computing power, which sets a tradeoff between performance and training time.
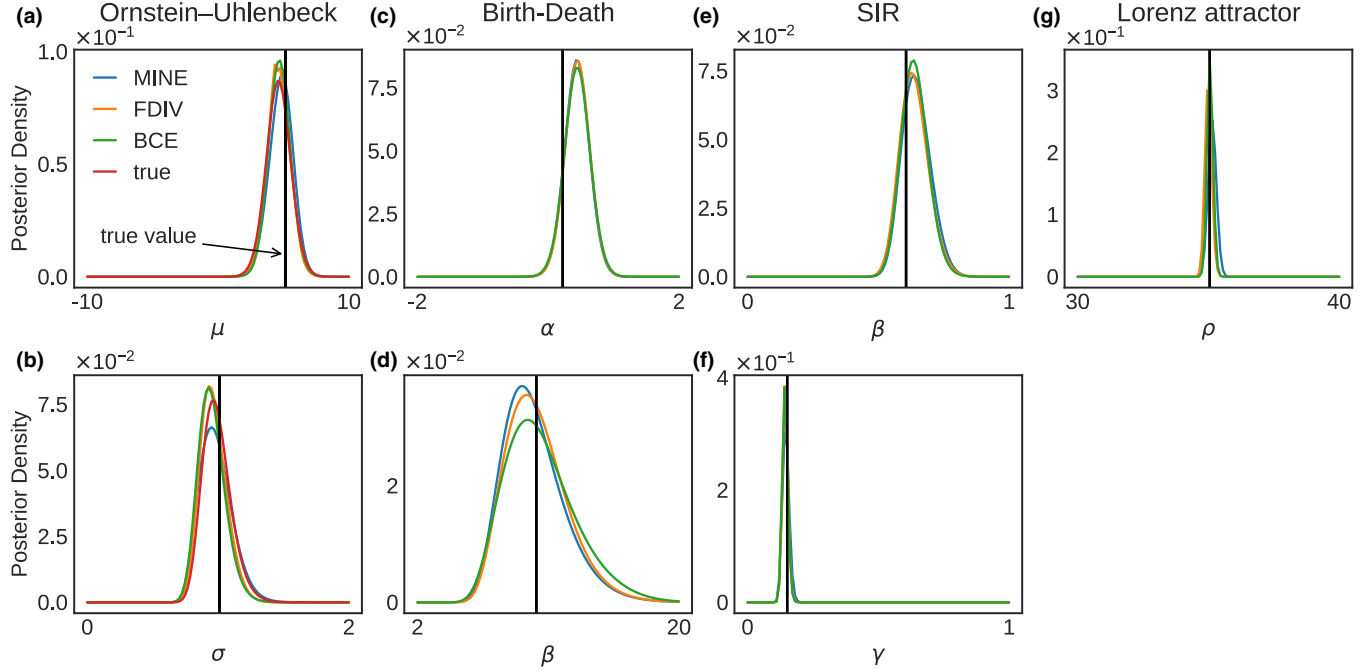
FIG. 3. Convergence of the posteriors for Ornstein-Uhlenbeck (a,b), birth-death (c,d), SIR (e,f), and Lorenz attractor (g). To evaluate the posteriors, we first choose a reference hypothesis θ^*: for Ornstein-Uhlenbeck $\mu^* = 5$ and $\sigma^* = 1$, for birth-death $\beta^* = 10$ and $\alpha^* = 0.2$, for SIR $\beta^* = 0.6$ and $\gamma^* = 0.2$, and for Lorenz attractor $\rho^* = 35$. We show posteriors $P^\infty(\theta|x_{1:M})$ calculated using models trained with $N = 10^7$ trajectories. For the Ornstein-Uhlenbeck process, the exact posterior density $P(\theta|x_{1:M})$ is also shown.

We compare the accuracy of the three inference methods (MINE, FDIV, BCE) for the four tasks (OU, birth-death, SIR, Lorenz) based on the three following metrics.

A. Global comparison

The first metric used for the benchmark is the mutual information given a density estimator, computed with Eq. (5). For each N, it is evaluated on test data composed of the remaining $N_{\text{tot}} - N$ samples. Unlike the other two comparisons (see below), it is a global metric that tests the approximation of the likelihood-to-evidence ratio over all θ and x values. For this reason, we use it to perform hyperparameter tuning for each task and each objective with $N = 10^5$; see Appendix D.

For all four tasks, the value of the estimated mutual information grows with the simulation budget and yields comparable performances for the three methods (Fig. 4, a, d, g, and j). For the Ornstein-Uhlenbeck process, the BCE method reaches consistently higher values of mutual information. For the Lorenz attractor, the MINE method is significantly outperformed by the other methods in the low data limit ($N = 10^4$).

B. Posterior comparison

The objective of simulation-based inference is to find the posterior distribution over model parameters. To characterize the inference accuracy as a function of the simulation budget N and the method l, we evaluate the Jensen-Shannon divergence between the inferred and the reference posterior $D_{\text{JS}}(P^\infty(\theta|x_{1:M}), \hat{P}_l^N(\theta|x_{1:M}))$ [where $D_{\text{JS}}(p, q) = (1/2) \int [p(x) \ln (p(x)/m(x)) + q(x) \ln (q(x)/m(x))] dx$ with $m(x) = [p(x) + q(x)]/2$], by scanning through the parameter

space with the prior $P(\theta)$. A larger Jensen-Shannon divergence indicates a larger deviation between the inferred posterior and the reference (i.e., a lower performance).

All methods show comparable performances, and the Jensen-Shannon divergence decays as a function of the simulation budget N (Fig. 4 b, e, h, and k, and Fig. 5 for the Ornstein-Uhlenbeck process with $d \geqslant 1$). At $N = 10^4$, the accuracy of the posterior inference is decreased for all objective functions, as reflected by the large variance of the D_{JS}. In the case of the Lorenz attractor, this simulation budget is also insufficient for the MINE method, which performs significantly worse than the classifier-based (BCE) and f-divergence (FDIV) estimators.

Figure 5 presents a comparison of the objective functions for the inference of the damping matrix elements in the high-dimensional Ornstein-Uhlenbeck process. We compare the performance in terms of the estimated posterior's divergence from the analytical prediction. We compute the Jensen-Shannon divergence independently for the marginal posterior distributions over each element of the damping matrix [a total of $\binom{d}{2}$ unique elements in dimension d]. We present the results independently for diagonal and off-diagonal elements of the matrix γ [in dimension d there are d diagonal and $d(d - 1)/2$ off-diagonal elements].

The efficacy of the three methods is comparable and the average performance does not significantly decrease up to $d = 5$, for which we estimate 15 elements of matrix g. At the same time, the variation in the Jensen-Shannon divergence increases with dimension as the inference task becomes more difficult. This is particularly pronounced for the MINE method in dimensions 4 and 5 at low ($N = 10^4$) as well as intermediate ($N = 10^5$) simulation budgets. In these instances,
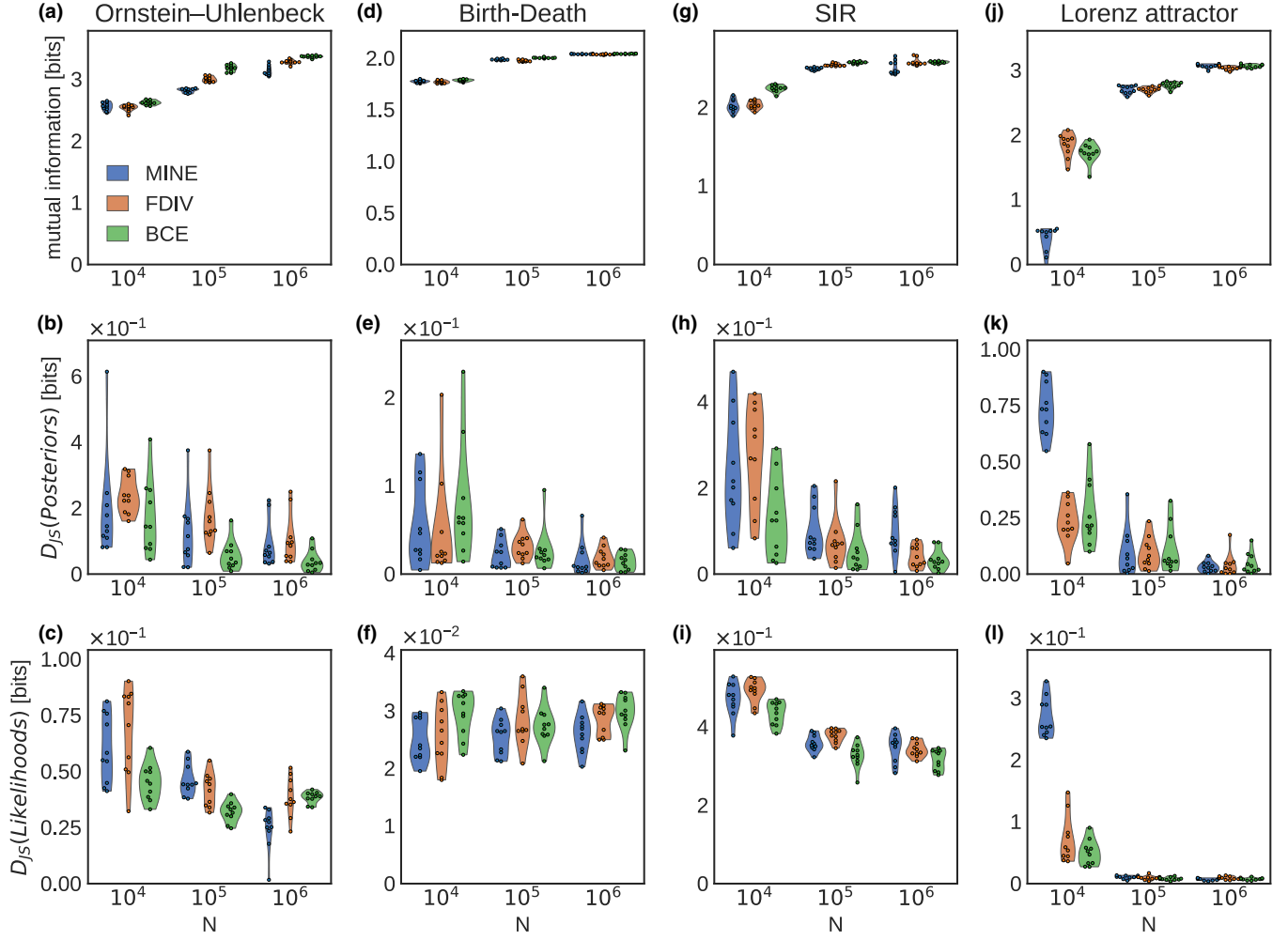
FIG. 4. Benchmarking of the methods. We compare the three objectives M, Eq. (5) (MINE), L_f, Eq. (7) (FDIV), and S, Eq. (8) (BCE) for three different metrics. We perform 10 replicates of the inference and comparison for simulation budgets $N \in \{10^4, 10^5, 10^6\}$ for four systems: Ornstein-Uhlenbeck (a,b,c), birth-death (d,e,f), SIR (g,h,i), and Lorenz attractor (j,k,l). In the first row, we compare the mutual information on held out test data using Eq. (5) with the estimated E^ϕ. For the following two metrics, we need to instead choose a hypothesis θ^*; see Fig. 3 for the exact values. In the second row, we compare the Jensen-Shannon divergence $D_{JS}(P^\infty(\theta|x_{1:M}), \hat{P}_l^N(\theta|x_{1:M}))$ between the reference and inferred posteriors. In the last row, we compare the Jensen-Shannon divergence $D_{JS}(P(x|\theta^*), \hat{P}_l^N(x|\theta^*))$ using sampled trajectories from the simulator $P(x|\theta^*)$ and the inferred distribution $\hat{P}_l(x|\theta^*)$.

using the f-divergence objective function yields the best performance.

C. Likelihood comparison

The third metric is the Jensen-Shannon divergence $D_{JS}(P(x|\theta^*), \hat{P}_l^N(x|\theta^*))$ between the true and approximated likelihood for a given model θ. This D_{JS} cannot be directly evaluated by summing over x, because it is typically of high dimension. We thus rely on samples from these two distributions and infer an additional classifier to estimate D_{JS}; see Appendix E.

The performance of the three methods is comparable (Fig. 4, c, f, i, and l). For the Ornstein-Uhlenbeck process, the BCE infers more accurate likelihood functions at $N = 10^4$ and 10^5 but it is outperformed by MINE at higher simulation budgets.

The results of the benchmark shown in Fig. 4 suggest that all estimators show reliable performances across different tasks and simulation budgets. While the first metric is global and the two other metrics are local, they draw a consistent picture. A higher simulation budget enhances the performance of all methods. The BCE method tends to perform better at the lowest simulation budget. All three methods perform similarly in the middle and high data regimes.

VI. CONCLUSION

We analyzed the problem of inferring an amortized estimator for the likelihood-to-evidence ratio over model parameters, using simulated data. We showed that this inference can be performed by maximization of the mutual information between simulated data and parameters of the model. This formulation captures an intuition that inference can be performed when we can extract the dependence between parameters and
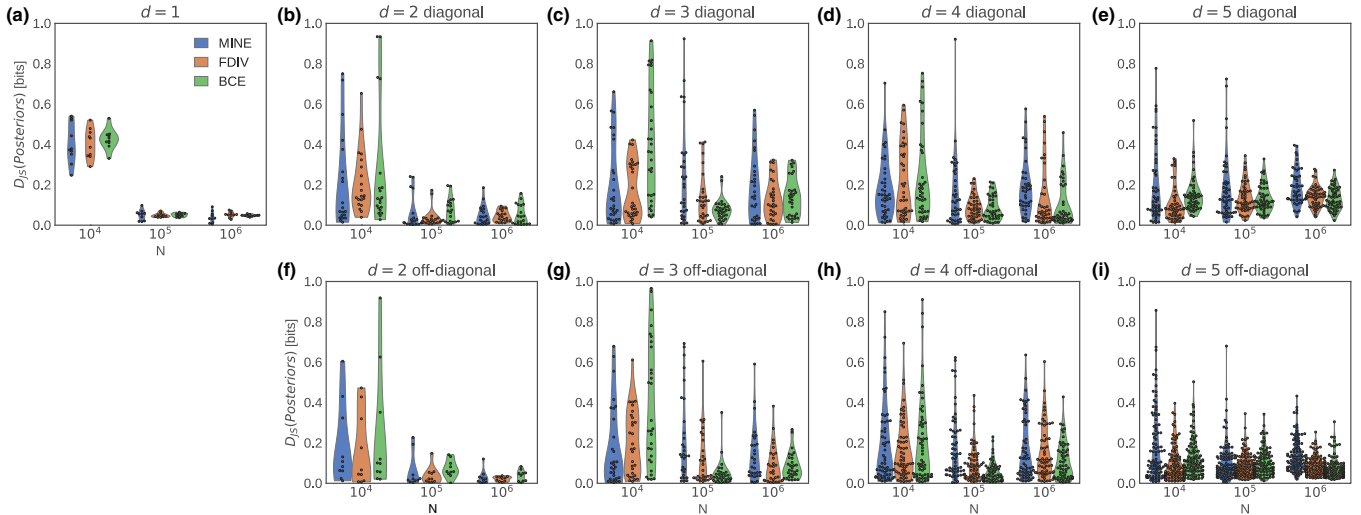
FIG. 5. The Ornstein-Uhlenbeck process in $d \geqslant 1$. We compare the three objective functions M (MINE), L_f (FDIV), and S (BCE) for dimension $d = 1, 2, 3, 4, 5$. We perform 10 replicates of the inference with simulation budgets $N = 10^4, 10^5, 10^6$ for changing dimension $d = 1, 2, 3, 4, 5$. We compare the posterior with the analytical prediction for a hypothesis $(g^*)_{ij} = -1$. We compute the Jensen-Shannon divergence $D_{\mathrm{JS}}(P(\theta|x_{1:M}), \hat{P}_l^N(\theta|x_{1:M}))$ between the true and inferred posteriors for each element of the damping matrix γ independently. We show the divergence for diagonal (a–e) and off-diagonal terms (f–i).

observed data, as measured by the mutual information. Our formalism opens up possibilities for using algorithms and techniques developed in the context of mutual information estimation [36] for inverse problems.

The likelihood function we propose is equivalent to the mutual information bound analyzed in [13]. However, while in [13] the focus is on the estimation of the absolute value of this quantity, we are interested in the inferred energy function that can be used to evaluate the posterior distribution for model parameters. Previous work that used classifiers for simulation-based inference [10,11] also fits naturally within our framework since logistic regression is linked to mutual information estimation [15]. The methods we studied rely on two lower bound estimators of mutual information, which are based on (i) the Donsker-Varadhan [13], and (ii) f-divergence representations of the Kullback-Leibler divergence [14]. It would be interesting to explore other known mutual information estimators for simulation-based inference [36].

We showed that the mutual information-based methods (MINE and FDIV), implemented in flexible neural networks, can reliably infer the posterior of the parameters and give consistent results with the previously proposed classifier-based technique (BCE) [10,11] when the simulation budget is sufficient. We benchmarked the three approaches and found that their performances are comparable in the intermediate data regime, while in the low data regime the classifier-based method performs consistently better. The main limitation of the two proposed objective functions M and L_f is that they require large simulation budgets for accurate inference.

Our choice to implement the neural network as a multi-layer perceptron with two hidden layers was motivated by having a simple and reliable architecture to better focus on the relative performance of the different objective functions. For the specific task of inference of model parameters from discrete samples of trajectories, absolute performance could be increased by choosing network architectures adapted to the data structure such as convolutional and recurrent layers.

Existing approaches to simulation-based inference, such as ABC, suffer from the need to define ad-hoc summary statistics to be matched between data and model. An important property of mutual information is its invariance upon the reparametrization of its variables. This enables inference and comparison of different parametrizations of the observed data, as different choices can be evaluated using the absolute value of the mutual information. A specific application that could be interesting to explore is inference for population genetics models, where the choice of summary statistics to use for ABC analysis has always been critical, and the ability to flexibly compare different parametrization choices greatly improves performance, as shown in [23]. Another possibility would be to explore more principled regularization techniques such as the information bottleneck method [37]. This approach could be used to infer summary statistics of the data that are maximally informative of the parameters of the model. Then the summary statistics could be added as additional variables for the observations of related tasks, such as model extensions, in a transfer learning fashion.

In conclusion, our work helps to clarify the link between mutual information estimation and simulation-based inference. We believe that this connection can be a fruitful source of improved methods for amortized inference.

The code for the algorithms presented in this paper is available in Ref. [38].

APPENDIX A: DIFFERENT OBJECTIVE FUNCTIONS SHARE AN OPTIMUM

We study the following objective functions:

$$M(\phi; \mathcal{I}, \mathcal{J}) = -\mathbb{E}_{\mathcal{J}}[E^\phi] - \ln \mathbb{E}_{\mathcal{I}}[e^{-E^\phi}], \tag{A1}$$

$$L_f(\phi; \mathcal{I}, \mathcal{J}) = -\mathbb{E}_{\mathcal{J}}[E^\phi] - \mathbb{E}_{\mathcal{I}}[e^{-E^\phi - 1}], \tag{A2}$$

$$S(\phi; \mathcal{I}, \mathcal{J}) = -\mathbb{E}_{\mathcal{J}}[\ln d^\phi] - k\mathbb{E}_{\mathcal{I}}[\ln\left(1 - d^\phi\right)]$$

$$= -\mathbb{E}_{\mathcal{J}}\left[\ln \frac{1}{1 + kZe^{E^\phi}}\right] - k\mathbb{E}_{\mathcal{I}}\left[\ln \frac{k}{k + Z^{-1}e^{-E^\phi}}\right]. \tag{A3}$$

In the infinite data limit, the empirical averages converge and we can rewrite all objectives as functional of the energy model:

$$M(E^\phi) = -\int E^\phi(x, \theta) P_{\text{joint}}(x, \theta)\, dx\, d\theta - \ln \int e^{-E^\phi(x,\theta)} P_{\text{indep}}(x, \theta)\, dx\, d\theta, \tag{A4}$$

$$L_f(E^\phi) = -\int \left(E^\phi(x, \theta) P_{\text{joint}}(x, \theta) + e^{-E^\phi(x,\theta)-1} P_{\text{indep}}(x, \theta)\right) dx\, d\theta, \tag{A5}$$

$$S(E^\phi) = -\int \left(\ln \frac{1}{1 + kZe^{E^\phi(x,\theta)}} P_{\text{joint}}(x, \theta) + k \ln \frac{k}{k + Z^{-1}e^{-E^\phi(x,\theta)}} P_{\text{indep}}(x, \theta)\right) dx\, d\theta. \tag{A6}$$

In this limit, the three optima of the objective functions are equivalent and recover the likelihood-to-evidence ratio. To see this, we take the functional derivative with respect to the energy model E^ϕ,

$$\frac{\delta M}{\delta E^\phi(x, \theta)} = -P_{\text{joint}}(x, \theta) + \frac{1}{\int e^{-E^\phi(x,\theta)} P_{\text{indep}}(x, \theta)\, dx\, d\theta} e^{-E^\phi(x,\theta)} P_{\text{indep}}(x, \theta), \tag{A7}$$

$$\frac{\delta L_f}{\delta E^\phi(x, \theta)} = -P_{\text{joint}}(x, \theta) + e^{-E^\phi(x,\theta)-1} P_{\text{indep}}(x, \theta), \tag{A8}$$

$$\frac{\delta S}{\delta E^\phi(x, \theta)} = k\left(\frac{Ze^{E^\phi(x,\theta)} P_{\text{joint}}(x, \theta) - P_{\text{indep}}(x, \theta)}{1 + kZe^{E^\phi(x,\theta)}}\right)\left(1 + \frac{\delta \ln Z}{\delta E^\phi(x, \theta)}\right), \tag{A9}$$

and we find that they vanish at energies E_M, E_f, and E_S, respectively:

$$E_M = -\ln \frac{P_{\text{joint}}(x, \theta)}{P_{\text{indep}}(x, \theta)} - \ln Z, \tag{A10}$$

$$E_f = -\ln \frac{P_{\text{joint}}(x, \theta)}{P_{\text{indep}}(x, \theta)} - 1, \tag{A11}$$

$$E_S = -\ln \frac{P_{\text{joint}}(x, \theta)}{P_{\text{indep}}(x, \theta)} - \ln Z. \tag{A12}$$

All three are equal to the logarithm of the likelihood-to-evidence ratio up to constant factors. We note that the second derivatives are different in the three cases, and therefore convergence to the optima E_M, E_f, and E_S will in general be different.

APPENDIX B: NOISE CONTRASTIVE ESTIMATION AND MUTUAL INFORMATION

In this Appendix, we show how our work fits within the framework of noise contrastive estimation (NCE) and how it relates to the existing contrastive learning approaches to simulation-based inference. The NCE methods estimate a probability density $p(y)$ by comparison to a reference noise distribution $q(y)$ [39,40]:

$$p(y) = \frac{1}{Z} e^{-E(y)} q(y), \tag{B1}$$

which reduces the problem to approximating the density ratio. The original method [39] consists of the inference of the density ratio model using logistic regression (minimizing binary cross entropy) on samples from both distributions, p, q. This framework encompasses the likelihood-to-evidence ratio inference problem where $p(y) = P_{\text{joint}}(x, \theta)$ and $q(y) = P_{\text{indep}}(x, \theta)$ and one minimizes $S(\phi; \mathcal{I}, \mathcal{J})$ to find E^ϕ.

An alternative approach proposed in the noise contrastive estimation literature [41,42] focuses on the estimation of conditional probability functions,

$$p(y|z) = \frac{1}{Z(z)} e^{-E(z,y)} q(y), \tag{B2}$$

where now the partition function explicitly depends on the conditioned variable z. The new density ratio can be inferred

by optimizing the so-called ranking objective [41]. This objective function is typically used to rank a positive sample from the target distribution $p(y|z)$ above k samples from the reference noise $q(y)$ for the input z [42].

In simulation-based inference, this family of methods has been used for posterior estimation, where $p(y|z) = P(\theta|x)$ is the unknown posterior and $q(y) = P(\theta)$ is the prior. In our notations, the ranking objective function reads

$$L_r(\phi; \mathcal{J}) = \mathbb{E}_{\mathcal{J}}\left[\int \ln\left(\frac{e^{-E^\phi(x,\theta)}}{e^{-E^\phi(x,\theta)} + \sum_{j=1}^{k} e^{-E^\phi(x,\theta_j)}} \right) \prod_{i=1}^{k} P(\theta_i) d\theta_i \right]. \quad \text{(B3)}$$

This method is known as the sequential neural posterior estimation (SNPE) proposed in [18], building on the work in Refs. [16,17]. It is useful to note that the ranking loss $L_r(\phi; \mathcal{J})$ has also been used to construct a high-bias and low-variance estimator of mutual information [19,43].

Reference [19] proposes that also the binary classification approach of [11] is a special case of the above inference for $k = 1$. However, the ranking objective $L_r(\phi; \mathcal{J})$ with $k = 1$ is distinct from $S(\phi; \mathcal{I}, \mathcal{J})$ and the two methods cannot be identified as one. In Ref. [42], the cross-entropy has been compared to the ranking loss and shown to generically outperform it in the context of neural language processing.

APPENDIX C: THE ORNSTEIN-UHLENBECK PROCESS IN DIMENSION d

The trajectories $x(t)$ are solutions to a stochastic differential equation,

$$dx = -\gamma(x - \mu)dt + \sqrt{2}\sigma \, dW, \quad \text{(C1)}$$

where x is a d-dimensional coordinate, μ is its long-term average, γ is a $d \times d$ damping matrix, σ is the noise strength, and W is a d-dimensional Wiener process. From a trajectory $x(t)$ we sample n values every Δt so that $x = \{x_i = x(i\Delta t)\}$. To find the analytical expression for the likelihood of these observations, we write the corresponding Fokker-Planck equation for the density $P = P(x_i, t + \Delta t|x_{i-1}, t)$,

$$\frac{dP}{dt} = -\nabla[\gamma(x - \mu)P] + \sigma\sigma^T \nabla^2 P, \quad \text{(C2)}$$

solved with a multivariate Gaussian distribution density

$$P(x_i, t + \Delta t|x_{i-1}, t) = \frac{1}{\sqrt{(2\pi)^d \det \Sigma}} e^{-\frac{1}{2}(x_i - \langle x_i \rangle)^T \Sigma^{-1}(x_i - \langle x_i \rangle)}, \quad \text{(C3)}$$

with mean

$$\langle x_i \rangle = e^{-\gamma \Delta t} x_{i-1} + (1 - e^{-\gamma \Delta t})\mu, \quad \text{(C4)}$$

and a covariance matrix given by

$$\Sigma = 2 \int_0^{\Delta t} ds \, e^{\gamma(s - \Delta t)} \sigma\sigma^T e^{\gamma^T(s - \Delta t)}. \quad \text{(C5)}$$

Both expressions simplify when we set $\sigma = \mathbb{I}$ and $\mu = 0$. For symmetric γ ($\gamma = \gamma^T$) we can find an orthogonal eigenbasis $r(\gamma)$ in which the damping matrix is diagonal,

$$\gamma = r(\gamma) \Gamma r(\gamma)^T, \quad \text{(C6)}$$

where Γ is a diagonal matrix and $r(\gamma)r(\gamma)^T = \mathbb{I}$. The covariance matrix is also diagonal in this basis, which allows us to

compute the integral when $\sigma = \mathbb{I}$ so that

$$\Sigma(\Delta t) = r(\gamma) \Gamma^{-1}(1 - e^{2\Gamma \Delta t})r(\gamma)^T. \quad \text{(C7)}$$

To ensure that γ is symmetric and positive definite (which is required so that the trajectories do not diverge and a steady state exists), we choose the following parametrization:

$$\gamma = \mathbb{I} + \epsilon(d)g, \quad \text{(C8)}$$

where g is a random matrix from the Gaussian orthogonal ensemble with density

$$P(g) \propto e^{-\frac{d}{4}\text{Tr}(g^2)}. \quad \text{(C9)}$$

The eigenvalues of g can be both positive and negative; in particular, the lowest eigenvalue is distributed according to the Tracy-Widom law with mean $\mu_g = \sqrt{2d}$ and standard deviation of $\sigma_g = \sqrt{2}d^{1/6}$. Choosing $\epsilon(d) = (\mu_g + 2\sigma_g)^{-1}$ ensures that the eigenvalues of γ are all positive with good confidence.

APPENDIX D: NEURAL NETWORK ARCHITECTURE AND LEARNING HYPERPARAMETERS

The $M(\phi; \mathcal{I}, \mathcal{J})$ objective function is invariant with respect to a global shift in energy, $M(E^\phi + E_0) = M(E^\phi)$, since any shift E_0 can be incorporated in the partition function Z to obtain the same likelihood-to-evidence ratio. We choose an energy gauge in which the "free energy" vanishes, $-\ln Z = 0$. As suggested in [44], we do so by adding a regularization term of the form $-\lambda_Z(\ln Z)^2$ to the likelihood function. Since the constraint $Z = 1$ may be satisfied by adding the right constant E_0 to the energy function, this regularization does not affect the result of the optimization. We fixed the strength of this term to $\lambda_Z = 10^{-3}$.

To perform the benchmark of the methods, we used the same neural network architecture for all three objective functions: a multilayer perceptron [31] with two hidden layers of 50 nodes each. Each node processes a linear combination of the inputs and adds a constant term (bias). A hyperbolic tangent activation function is then applied to the result of this linear map. Between the second hidden layer and the output of the network, we do not apply the activation function. We implemented L_2 regularization on network weights, with regularization strength λ_2. We optimized the network weights using stochastic gradient descent and the RMSprop [45] optimization algorithm with learning rate l_r and a size of mini batches b.

We tuned the hyperparameter by inference of five replicate models on $N = 10^5$ training data for each objective function and combination of hyperparameters, $\lambda_2 \in \{10^{-4}, 10^{-5}, 10^{-6}\}$, $l_r \in \{10^{-2}, 10^{-3}, 10^{-4}\}$, and $b \in \{10^3, 10^4\}$.
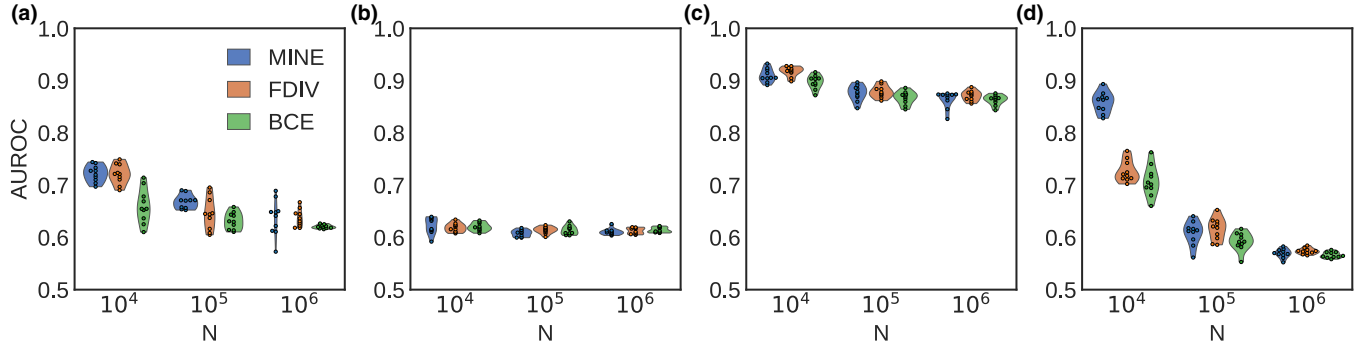
FIG. 6. We compare the three objective functions M (MINE), L_f (FDIV), and S (BCE) using the area under the receiver operating characteristic (AUROC) of a classifier trained to distinguish samples from the simulator $P(x|\theta^*)$ and samples from the inferred estimators $\hat{P}(x|\theta^*)$ for a specific hypothesis θ^*. AUROCs closer to $1/2$ mean that the model is performing well, as its samples are indistinguishable from samples from the true distribution. This metric gives a consistent picture with respect to the D_{JS} metric presented in the main text (Fig. 4, c, f, i, and l).

We evaluated the mutual information estimate $M(\phi; \mathcal{I}, \mathcal{J})$ on $N = 10^5$ independent samples (test set). For each of the three methods, we chose hyperparameters for which the mutual information was highest.

APPENDIX E: METHODS FOR LIKELIHOOD COMPARISON

We outline here the method for calculating the Jensen-Shannon divergence between two distributions for which an analytical density is not known, but instead we can sample from the two distributions. This will be the case for the likelihood comparison where we will compare the true likelihood and an inferred model of the likelihood.

We generate $M = 5 \times 10^4$ samples $\{x_m^*\}_{m=1}^M \sim P(x|\theta^*)$ from the true simulator. To generate samples for the inferred estimators $\{\hat{x}_m\}_{m=1}^M \sim \hat{P}_l^N(x|\theta^*)$, we perform rejection sampling on samples from the marginal probability $P(x)$. To produce samples from $P(x)$, we discard the parameters θ from the samples $\{(x_i, \theta_i)\}_{i=1}^{N_{\mathrm{tot}}}$. Rejection sampling is based on the identity $P(x|\theta) = P(x)Z^{-1}e^{-E(x,\theta)}$, where the likelihood-to-evidence ratio is approximated by an estimator. For each simulation budget N and method l we generate $\{\hat{x}_m\}_{m=1}^M$ samples by rejection sampling with acceptance probability $e^{-E_l^\phi}/Z_l^\phi$. In the last row of Fig. 4, the methods are compared using this metric.

By mixing the samples from $P(x|\theta^*)$ and $\hat{P}_l^N(x|\theta^*)$ in equal proportion, we construct an ensemble of samples from $P_{\mathrm{mix}}(x|\theta^*) = \frac{1}{2}[P(x|\theta^*) + \hat{P}_l^N(x|\theta^*)]$. We then train two clas-

sifiers, one between samples from $P(x|\theta^*)$ and $P_{\mathrm{mix}}(x|\theta^*)$, and the second between samples from $P_{\mathrm{mix}}(x|\theta^*)$ and $\hat{P}_l^N(x|\theta^*)$. We again exploit the fact that an optimal classifier is the ratio of the two likelihoods, and we can read off the two corresponding Kullback-Leibler divergences, $D_{\mathrm{KL}}(P||P_{\mathrm{mix}})$ and $D_{\mathrm{KL}}(\hat{P}_l^N||P_{\mathrm{mix}})$, from its estimate (5). The value of the Jensen-Shannon divergence is then the average

$$D_{\mathrm{JS}}(P||\hat{P}_l^N) = \frac{1}{2}\big(D_{\mathrm{KL}}(P||P_{\mathrm{mix}}) + D_{\mathrm{KL}}(\hat{P}_l^N||P_{\mathrm{mix}})\big). \quad \text{(E1)}$$

An alternative measure of performance of the model of the likelihood is the AUROC characteristic of the optimal classifier between samples from the true and an inferred likelihood, which we will use below. The best models should result in indistinguishable sets of samples for which AUROC $= 1/2$. A failure to capture the mutual information between parameters θ and simulated data x makes the two sets distinguishable and AUROC $= 1$.

APPENDIX F: SUPPLEMENTARY BENCHMARK RESULTS

Figure 6 presents a comparison of the objective functions applied to the four tasks: the Ornstein-Uhlenbeck process in dimension 1 (A), the birth-death process (B), the SIR process (C), and the Lorenz attractor (D), using the AUROC between the true and estimated likelihood as a measure of performance. The results are consistent with the Jensen-Shannon divergence metric presented in the main text (Fig. 4).

[1] S. Tavaré, D. J. Balding, R. C. Griffiths, and P. Donnelly, Inferring coalescence times from DNA sequence data, Genetics **145**, 505 (1997).

[2] K. Cranmer, J. Brehmer, and G. Louppe, The frontier of simulation-based inference, Proc. Natl. Acad. Sci. (USA) **117**, 30055 (2020).

[3] J.-M. Lueckmann, J. Boelts, D. S. Greenberg, P. J. Gonçalves, and J. H. Macke, Benchmarking simulation-based inference, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, edited by A. Banerjee, and K. Fukumizu, Vol. 130 (PMLR, 2021), pp. 343–351.

[4] M. A. Beaumont, W. Zhang, and D. J. Balding, Approximate Bayesian computation in population genetics, Genetics **162**, 2025 (2002).

[5] S. A. Sisson, Y. Fan, and M. A. Beaumont, *Handbook of Approximate Bayesian Computation* (Chapman and Hall/CRC, New York, 2018).

[6] S. N. Wood, Statistical inference for noisy nonlinear ecological dynamic systems, Nature (London) **466**, 1102 (2010).

[7] G. Papamakarios, D. Sterratt, and I. Murray, Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows, in *Proceedings of the Twenty-Second International*

Conference on Artificial Intelligence and Statistics, edited by K. Chaudhuri and M. Sugiyama, Vol. 89 of *Proceedings of Machine Learning Research* (PMLR, 2019), pp. 837–848.

[8] D. Rezende and S. Mohamed, Variational inference with normalizing flows, in *Proceedings of the 32nd International Conference on Machine Learning*, edited by F. Bach and D. Blei, Vol. 37 of *Proceedings of Machine Learning Research* (PMLR, 2015), pp. 1530–1538.

[9] G. L. Kyle Cranmer and J. Pavez, Approximating likelihood ratios with calibrated discriminative classifiers (unpublished).

[10] O. Thomas, R. Dutta, J. Corander, S. Kaski, and M. U. Gutmann, Likelihood-free inference by ratio estimation, *Bayesian Analysis*, Vol. 17 (2021), pp. 1–31.

[11] J. Hermans, V. Begy, and G. Louppe, Likelihood-free MCMC with amortized approximate ratio estimators, 2020, *Proceedings of the 37th International Conference on Machine Learning*, edited by H. Daumé III and A. Singh Vol. 119 of *Proceedings of Machine Learning Research* (PMLR, 2020), pp. 4239–4248.

[12] F. Ferretti, V. Chardès, T. Mora, A. M. Walczak, and I. Giardina, Building General Langevin Models from Discrete Datasets, Phys. Rev. X **10**, 031018 (2020).

[13] M. I. Belghazi, A. Baratin, S. Rajeswar, S. Ozair, Y. Bengio, A. Courville, and R. D. Hjelm, MINE: Mutual information neural estimation, 2018, *Proceedings of the 35th International Conference on Machine Learning*, edited by J. Dy and A. Krause, Vol. 80 of *Proceedings of Machine Learning Research* (PMLR, 2018), pp. 531–540.

[14] X. Nguyen, M. J. Wainwright, and M. I. Jordan, Estimating divergence functionals and the likelihood ratio by convex risk minimization, IEEE Trans. Inf. Theor. **56**, 5847 (2010).

[15] S. Mukherjee, H. Asnani, and S. Kannan, CCMI: Classifier based Conditional Mutual Information Estimation, *Proceedings of the 35th Uncertainty in Artificial Intelligence Conference*, edited by R. Adams and V. Gogate (UAI 2019).

[16] G. Papamakarios and I. Murray, Fast ϵ-free inference of simulation models with bayesian conditional density estimation, in *Advances in Neural Information Processing Systems*, edited by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Curran Associates, New York, 2016), Vol. 29.

[17] J.-M. Lueckmann, P. J. Goncalves, G. Bassetto, K. Öcal, M. Nonnenmacher, and J. H. Macke, Flexible statistical inference for mechanistic models of neural dynamics, in *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, New York, 2017), Vol. 30.

[18] D. Greenberg, M. Nonnenmacher, and J. Macke, Automatic posterior transformation for likelihood-free inference, in *Proceedings of the 36th International Conference on Machine Learning*, edited by K. Chaudhuri and R. Salakhutdinov, Vol. 97 of *Proceedings of Machine Learning Research* (PMLR, 2019), pp. 2404–2414.

[19] C. Durkan, I. Murray, and G. Papamakarios, On contrastive learning for likelihood-free inference, *Proceedings of the 36th International Conference on Machine Learning*, edited by H. Dumé II and A. Singh, Vol. 119 of *Proceedings of Machine Learning Research* (PMLR, 2020), pp. 2771–2781.

[20] S. Kleinegesse and M. U. Gutmann, Efficient Bayesian experimental design for implicit models, *Proceedings of the Twenty-Second International Conference on Artificial Intelli-*

gence and Statistics, edited by K. Chaudhuri and M. Sugiyama, Vol. 89 of *Proceedings of Machine Learning Research* (PMLR, 2019), pp. 476–485.

[21] A. Foster, and M. Jankowiak, E. Bingham, P. Horsfall, Y. W. Teh, T. Rainforth, and N. Goodman (unpublished).

[22] S. Kleinegesse and M. U. Gutmann, Efficient Bayesian experimental design for implicit models, in *Proceedings of Machine Learning Research* (PMLR, 2019), pp. 476–485.

[23] Y. Chen, D. Zhang, M. U. Gutmann, A. Courville, and Z. Zhu, Neural approximate sufficient statistics for implicit models, *Ninth International Conference on Learning Representations*, in *International Conference on Learning Representations (ICLR)* Virtual Conference, 2021.

[24] S. Kleinegesse and M. U. Gutmann, Gradient-based Bayesian experimental design for implicit models using mutual information lower bounds (unpublished).

[25] G. E. Uhlenbeck and L. S. Ornstein, On the theory of the Brownian motion, Phys. Rev. **36**, 823 (1930).

[26] O. Vasicek, An equilibrium characterization of the term structure, J. Financ. Econ. **5**, 177 (1977).

[27] L. L. Cavalli-Sforza and A. W. F. Edwards, Phylogenetic analysis. models and estimation procedures, Am. J. Hum. Genet. **19**, 233 (1967).

[28] J. Felsenstein, Phylogenies and quantitative characters, Ann. Rev. Ecol. Systematics **19**, 445 (1988).

[29] P. E. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*, Stochastic Modelling and Applied Probability (Springer-Verlag, Berlin, 1992).

[30] D. T. Gillespie, Exact stochastic simulation of coupled chemical reactions, J. Phys. Chem. **81**, 2340 (1977).

[31] D. Rumelhart, G. E. Hinton, and R. J. Williams, Learning internal representations by error propagation, *Parallel distributed processing: explorations in the microstructure of cognition* (MIT press, Cambridge, 1986), Vol. 1, pp. 318–362.

[32] M. Abadi *et al.*, TensorFlow: Large-scale machine learning on heterogeneous systems, 2015, Software available from tensorflow.org.

[33] F. Chollet *et al.*, Keras, 2015 Software available from https://github.com/fchollet/keras.

[34] C. R. Harris *et al.*, Array programming with NumPy, Nature (London) **585**, 357 (2020).

[35] P. Virtanen *et al.*, SciPy 1.0: Fundamental algorithms for scientific computing in python, Nat. Methods **17**, 261 (2020).

[36] B. Poole, S. Ozair, A. van den Oord, A. A. Alemi, and G. Tucker, On variational bounds of mutual information, *Proceedings of the 36th International Conference on Machine Learning*, edited by K. Chaudhuri and R. Salakhutdinov, Vol. 97 of *Proceedings of Machine Learning Research* (PMLR,2019).

[37] N. Tishby, F. C. Pereira, and W. Bialek, The information bottleneck method, *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing* (1999), pp. 368–377.

[38] github.com/statbiophys/MINIMALIST.

[39] M. Gutmann and A. Hyvärinen, Noise-contrastive estimation: A new estimation principle for unnormalized statistical models, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, edited by Y. W. Teh and M. Titterington, Vol. 9 of *Proceedings of Machine Learning Research* (PMLR, 2010), pp. 297–304.

[40] M. U. Gutmann and A. Hyvärinen, Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics, J. Mach. Learn. Res. **13**, 307 (2012).

[41] R. Jozefowicz, O. Vinyals, M. Schuster, N. Shazeer, and Y. Wu, Exploring the limits of language modeling, 2016 (unpublished).

[42] Z. Ma and M. Collins, Noise contrastive estimation and negative sampling for conditional models: Consistency and statistical efficiency, in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, 2018), pp. 3698–3707.

[43] A. van den Oord, Y. Li, and O. Vinyals, Representation learning with contrastive predictive coding, 2019 (unpublished).

[44] K. Choi and S. Lee, Regularized mutual information neural estimation, 2020 (unpublished).

[45] Rmsprop, http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf.