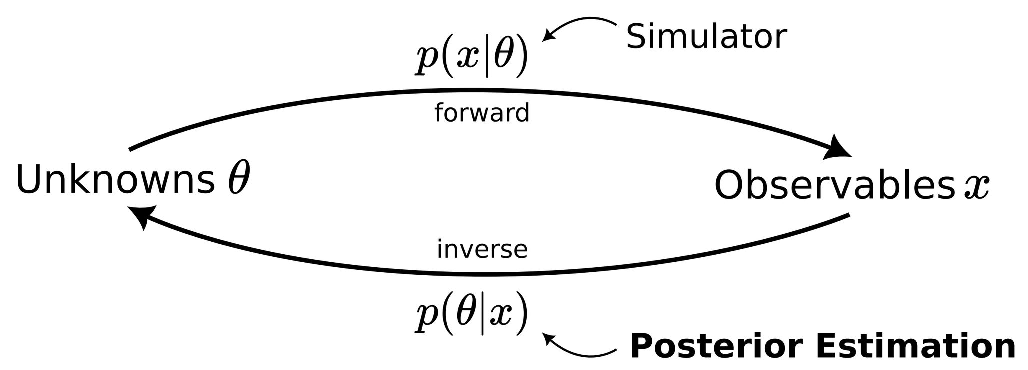
Consistency Models for Scalable and Fast Simulation-Based Inference

Marvin Schmitt
University of Stuttgart, Germany
Valentin Pratz
Heidelberg University & Zuse School ELIZA, Germany
Ullrich Köthe
Heidelberg University, Germany
Paul-Christian Bürkner
TU Dortmund University, Germany
Stefan T. Radev
Rensselaer Polytechnic Institute, United States

Introduction



- **Simulation-based inference (SBI)** allow us to infer the hidden parameters of complex systems by means of simulation.
- The goal of amortized **posterior estimation** is to efficiently approximate the full posterior distribution $p(\theta|x)$ over parameters θ for *any* observable x.

Motivation

- Multi-step models (e.g., diffusion models, flow matching) are flexible, but slow.
- One-step models (e.g., normalizing flows) are constrained by invertible architectures, but fast.
- Consistency models are both **unconstrained** *and* **fast**.

Method

- Explore Consistency Training to train neural posterior estimators from scratch
- Compare four methods: Affine Coupling Flows (**ACF**), Neural Spline Flows (**NSF**), Flow Matching Posterior Estimation (**FMPE**), and Consistency Model Posterior Estimation (**CMPE**; Ours)

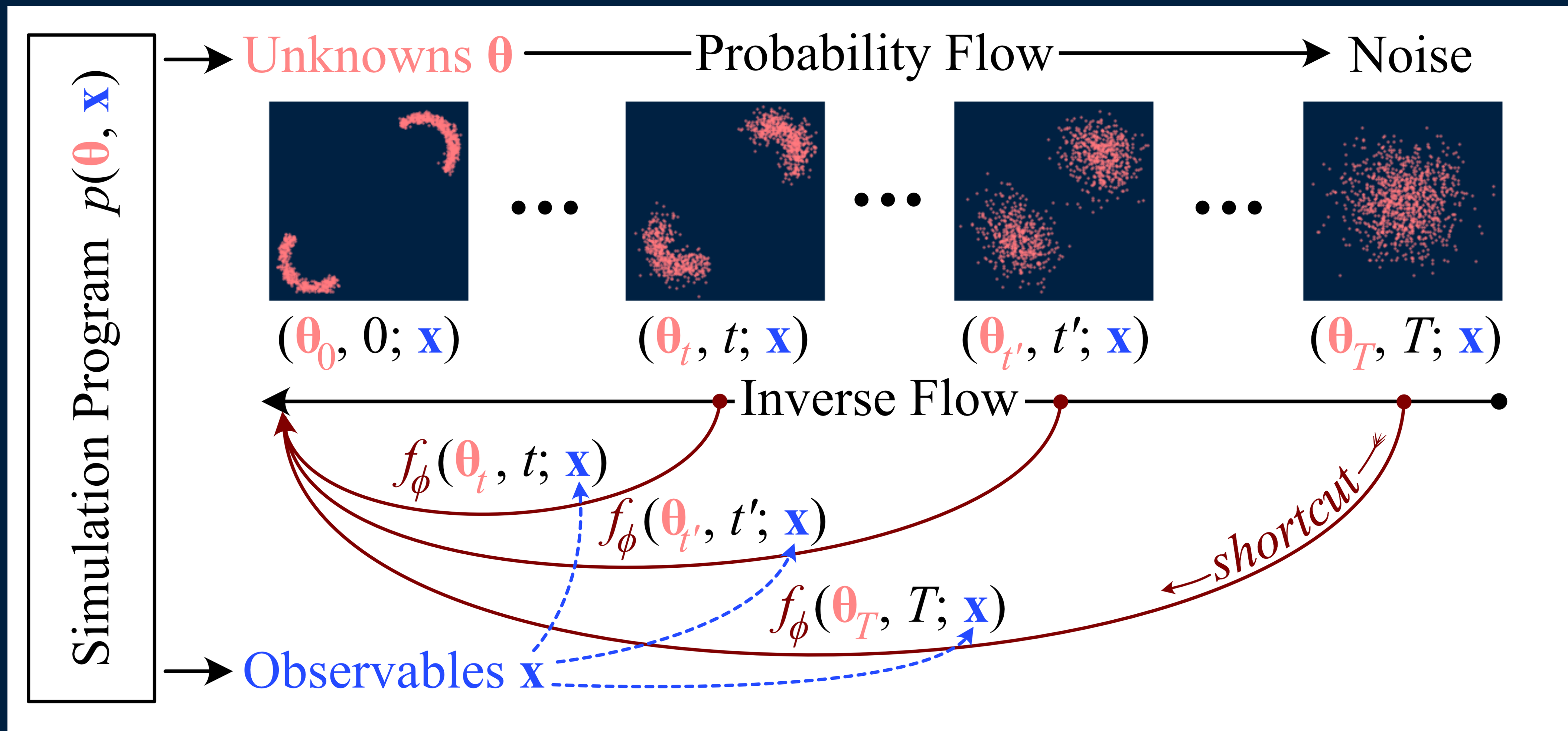
Consistency Function: Ensure that for $t = 0$, function is the identity: $f_{\boldsymbol{\phi}}(\boldsymbol{\theta}, t; \mathbf{x}) = c_{\text{skip}}(t)\boldsymbol{\theta} + c_{\text{out}}(t)F_{\boldsymbol{\phi}}(\boldsymbol{\theta}, t; \mathbf{x})$

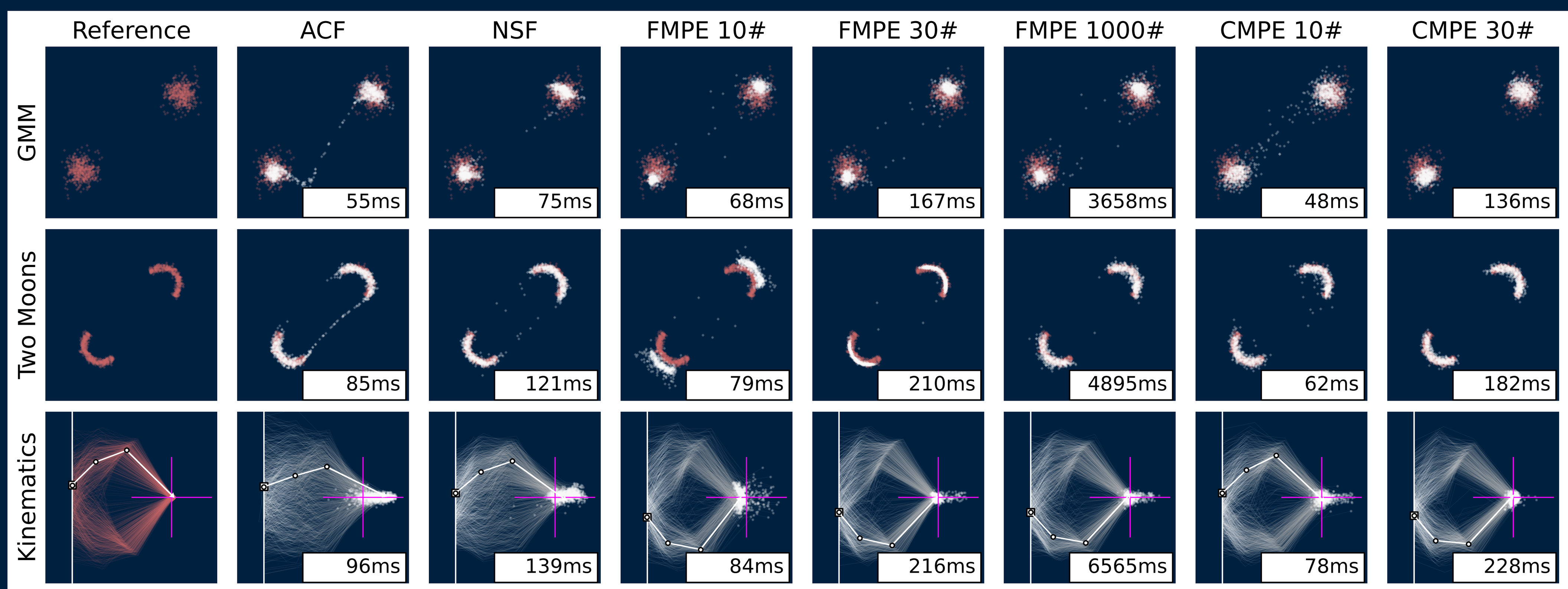
Optimization Objective:

$$\mathbb{E}\left[\lambda(t_i)\left\| f_{\boldsymbol{\phi}}(\boldsymbol{\theta} + t_{i+1}\mathbf{z}, t_{i+1}; \mathbf{x}) - \underbrace{f_{\boldsymbol{\phi}^-}(\boldsymbol{\theta} + t_i\mathbf{z}, t_i; \mathbf{x})}_{\texttt{stop\_gradient}}\right\|\right],$$

where $\lambda(t)$ is a weighting function and $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

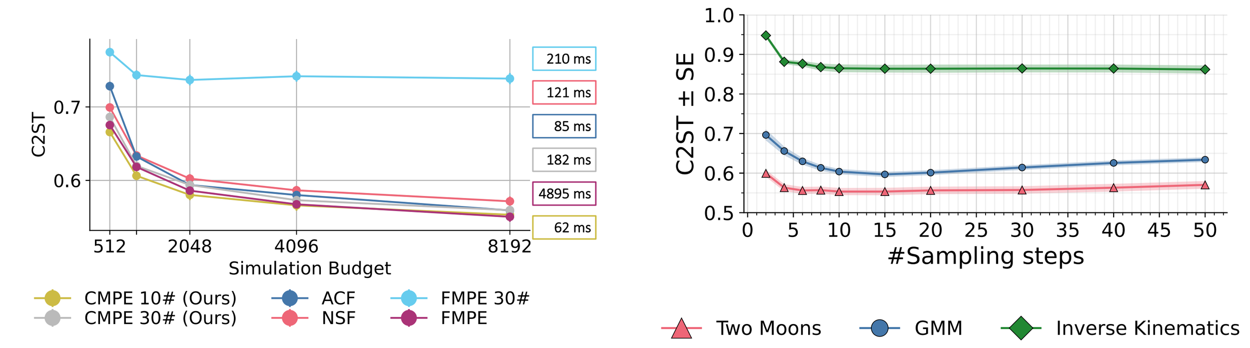
Consistency Models Enable **Fast** Posterior Approximation with **Unconstrained** Architectures.



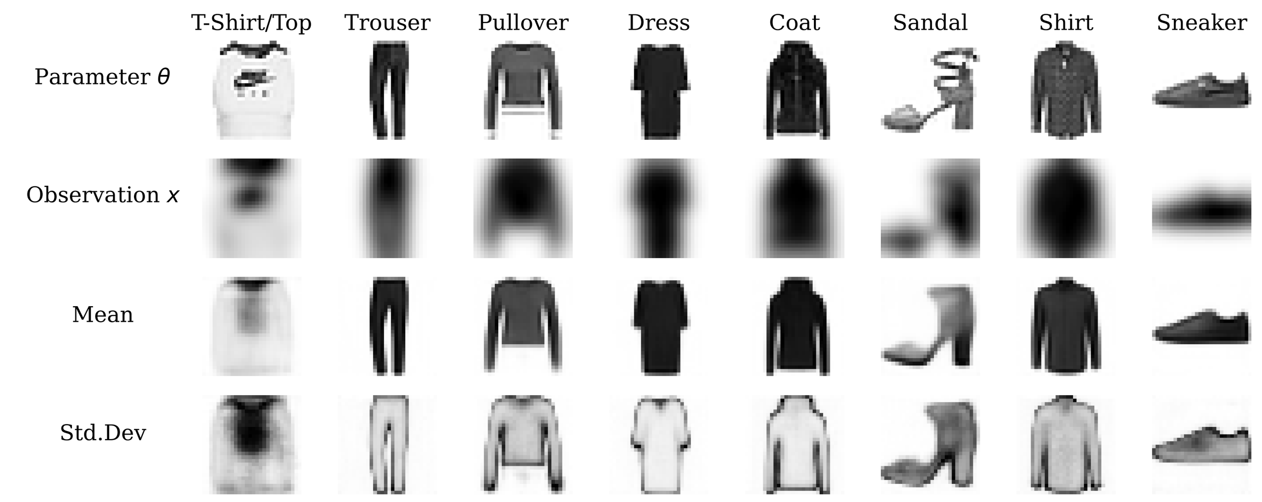


Experiments

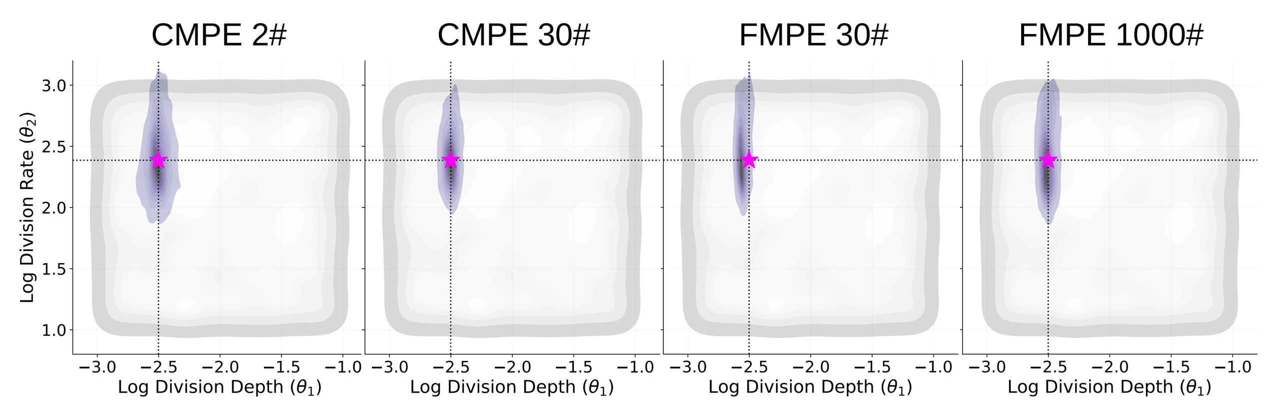
Three low-dimensional benchmarks: The tasks feature multi-modal distributions. See the center figure (bottom) for examples with sampling durations. Below on the left, we provide performance on the two moons benchmark as a function of the training budget for different methods. $K\#$ indicates K sampling steps. For C2ST, lower is better.



Bayesian denoising: Denoising Fashion MNIST shows that CMPE is able to handle higher-dimensional problems as well. For the figure below, we used a U-Net architecture and 60 000 training images.



Tumor spheroid growth: A multi-scale hybrid discrete-continuum model describing the growth of a 2D tumor spheroid. The plot below shows bivariate posteriors for two parameters.



Limitations

- No closed-form likelihood computation possible.
- Non-monotonic relationship between compute and sample quality.
- Slightly increased training time (25%).