Amortised Likelihood-free Inference for Expensive Time-series Simulators with Signatured Ratio Estimation

Joel Dyer[1], Patrick Cannon[2], Sebastian M. Schmon[2,3]

joel.dyer@maths.ox.ac.uk

University of Oxford[1], Improbable[2], Durham University[3]

UNIVERSITY OF OXFORD

InFoMM Industrially Focused Mathematical Modelling

Stochastic simulation models

Stochastic mechanistic models are used widely in scientific and commercial domains, ranging from biology to economics to cybersecurity. Such models often take the form of a simulation model: a computer program that consumes parameters $\boldsymbol{\theta} \in \mathbb{R}^d$ and generates possible instances of datasets $\mathbf{x}$, which can be viewed as a draw from the simulator's likelihood function, i.e. as $\mathbf{x} \sim p(\mathbf{x} \mid \boldsymbol{\theta})$.

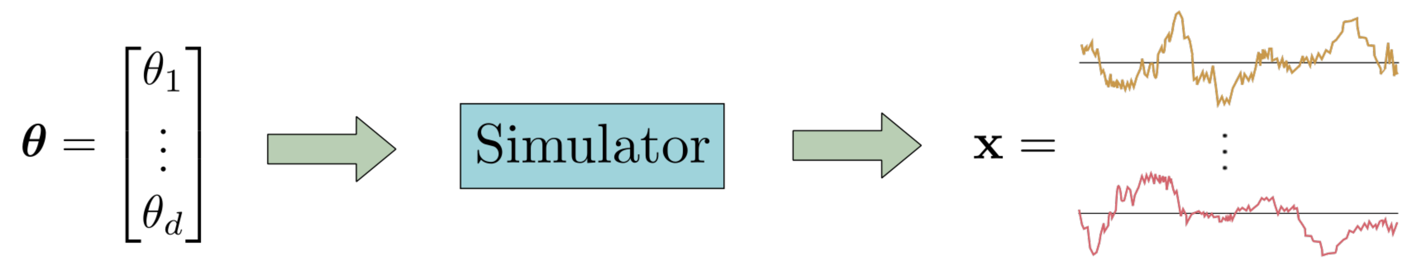

Figure 1 – A stochastic simulator consumes Euclidean parameters $\boldsymbol{\theta}$ and generates are random output $\mathbf{x}$.

Stochastic simulation models are helpful for flexibly capturing causal mechanisms.

Approximate inference for stochastic simulation models

Once a stochastic simulator is constructed, it is often the case that its behaviour varies with changes in the parameter values. Furthermore, one is in many cases interested in finding parameter values that result in simulations that match reality closely. This is an inference problem that can be approached using classical statistical methods, e.g. Bayesian inference:

$$\underbrace{p(\boldsymbol{\theta} \mid \mathbf{y})}_{\text{Updated belief distribution}} = \overbrace{\frac{p(\mathbf{y} \mid \boldsymbol{\theta})}{p(\mathbf{y})}}^{\text{Likelihood-to-evidence ratio}} \underbrace{p(\boldsymbol{\theta})}_{\text{Initial belief distribution}}$$

Figure 2 – Bayesian inference provides a means to updating degrees of belief regarding appropriate parameter values by incorporating both prior knowledge and new evidence.

However, likelihood-based inference techniques often cannot be applied immediately: evaluating the likelihood function for arbitrary simulators is generally an intractable computational task.

Approximate inference via density ratio estimation

Approximate, simulation-based inference techniques that mimic likelihood-based procedures are often used when the likelihood function is unavailable. One option is **density ratio estimation**.

Density ratio estimation can be used to learn the likelihood-to-evidence ratio appearing in Bayes's theorem [1]. This can be achieved by training a probabilistic classifier to distinguish between instances drawn from two collections of data-parameter pairs:

$$\begin{array}{cc} z = 1 & z = 0 \\ \left(\mathbf{x}^{(1)}, \boldsymbol{\theta}^{(1)}\right)\left(\mathbf{x}^{(4)}, \boldsymbol{\theta}^{(4)}\right) & \left(\mathbf{x}^{(1)}, \boldsymbol{\theta}^{(2)}\right)\left(\mathbf{x}^{(4)}, \boldsymbol{\theta}^{(3)}\right) \\ \left(\mathbf{x}^{(2)}, \boldsymbol{\theta}^{(2)}\right)\left(\mathbf{x}^{(5)}, \boldsymbol{\theta}^{(5)}\right) & \left(\mathbf{x}^{(2)}, \boldsymbol{\theta}^{(5)}\right)\left(\mathbf{x}^{(5)}, \boldsymbol{\theta}^{(4)}\right) \\ \left(\mathbf{x}^{(3)}, \boldsymbol{\theta}^{(3)}\right)\left(\mathbf{x}^{(6)}, \boldsymbol{\theta}^{(6)}\right) & \left(\mathbf{x}^{(3)}, \boldsymbol{\theta}^{(6)}\right)\left(\mathbf{x}^{(6)}, \boldsymbol{\theta}^{(1)}\right) \\ \sim p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}) & \sim p(\mathbf{x})p(\boldsymbol{\theta}) \end{array}$$

Figure 3 – The data-parameter pairs on the left are drawn from the joint density, such that the parameter "generated" the data to which it is paired. In contrast, the pairs on the right are drawn from the product of the marginals, such that there exists no relationship between each data instance and its parameter partner.

The likelihood-to-evidence ratio can be recovered from the logit of the following class probability estimate learned by the classifier:

$$\text{Classifier learns } d(\mathbf{x}, \boldsymbol{\theta}) := p(z = 1 \mid \mathbf{x}, \boldsymbol{\theta}) = \frac{p(\mathbf{x} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{x} \mid \boldsymbol{\theta})p(\boldsymbol{\theta}) + p(\mathbf{x})p(\boldsymbol{\theta})}$$

$$\Rightarrow p(\boldsymbol{\theta} \mid \mathbf{x}) = \exp\left[\operatorname{logit}(d(\mathbf{x}, \boldsymbol{\theta}))\right] p(\boldsymbol{\theta})$$

Figure 4 – The classifier learns an estimate of the class-1 probability, which can be used to estimate the posterior.

Summary statistics in density ratio estimation

To facilitate the learning task when the data $\mathbf{y}$ is a time-series, it is often practically necessary to find low-dimensional summary statistics for the data.

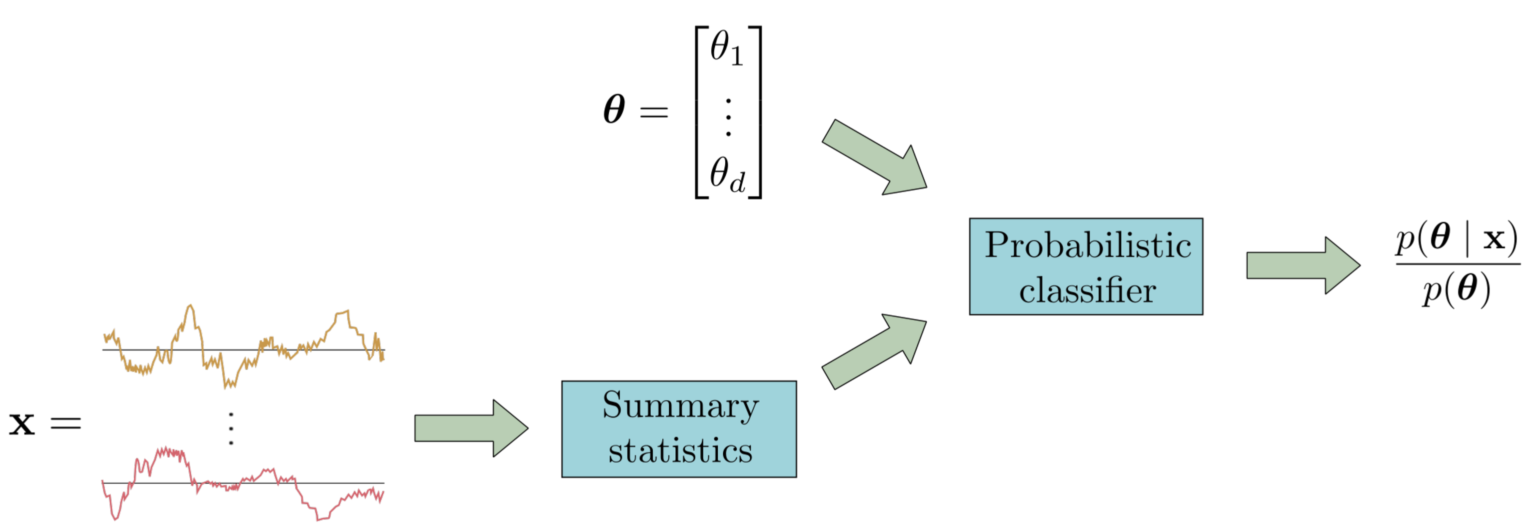

Figure 5 – The high-dimensional time-series data must often be summarised to facilitate learning. Good summary statistics are often unknown and must be learned from data.

Learning good low-dimensional representations of high-dimensional time-series can be difficult when the simulator is expensive and the simulation budget is severely constrained.

Acknowledgements

EPSRC Engineering and Physical Sciences Research Council

IMPROBABLE

Signatured density ratio estimation: SignatuRE

To bypass the task of learning summary statistics for time-series data – which can be difficult in low-simulation budget settings – we propose to use the path signature and corresponding signature kernel to construct probabilistic classifiers for time-series simulators [2]. Kernel methods are known to offer benefits in data-sparse scenarios, where deep learning-based approaches cannot be so readily deployed; applied in conjunction with signature features, this provides us with an alternative to neural probabilistic classifiers that is designed with low-simulation budget environments in mind.

The key idea is to use kernel logistic regression, along with the signature kernel k, on data-parameter pairs to perform the probabilistic binary classification task. Choosing kernel

$$m\left((\mathbf{x}, \boldsymbol{\theta}), (\mathbf{x}', \boldsymbol{\theta}')\right) = k(\mathbf{x}, \mathbf{x}')l(\boldsymbol{\theta}, \boldsymbol{\theta}'), \tag{1}$$

where l is some universal kernel on parameters, we are equipped with a universal kernel that can approximate the desired logit arbitrarily well over compact sets of data-parameter pairs.

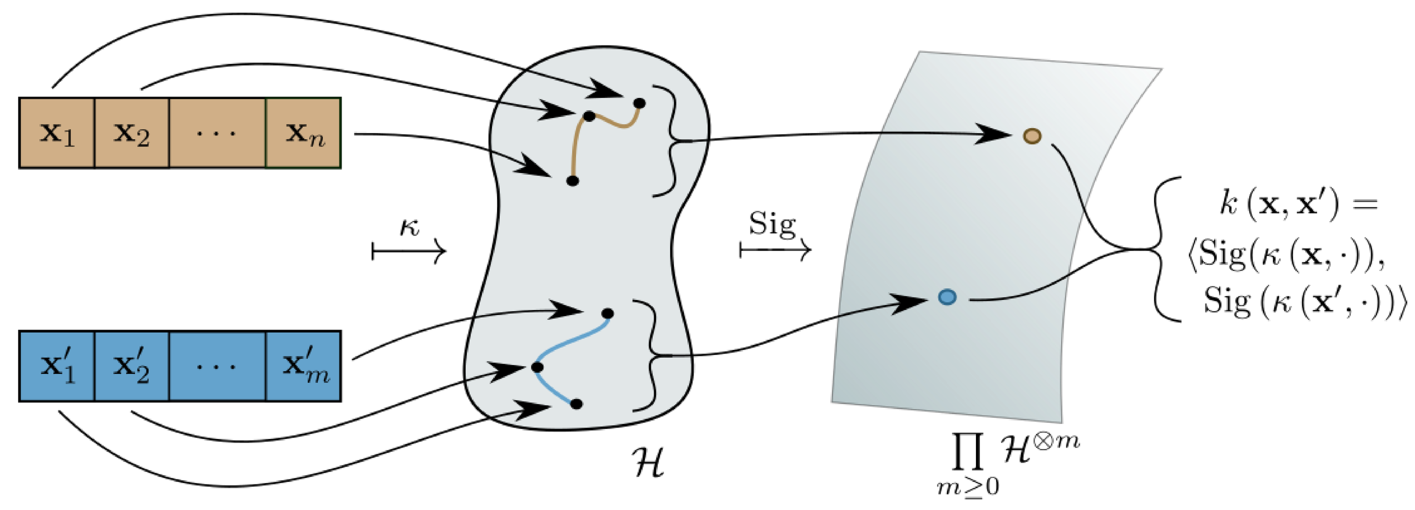

Figure 6 – A schematic showing how the signature kernel embeds the time-series.

Experimental results

To investigate the usefulness of our proposed solution, we compared its ability to recover ground-truth posteriors for a variety of simulators against three baseline alternatives:

a) GRU-ResNet – a neural classifier in which a GRU is used to embed the time-series before feeding the time-series embedding, in addition to the parameters, into a residual network
b) Bespoke ResNet – identical to the above, except that the time-series is cast into hand-crafted summary statistics that are known to be informative about the parameters being inferred. This is an idealised, gold-standard approach that is in general not possible for arbitrary simulators
c) K2-RE – an alternate kernel-based classifier in which the signature kernel k is replaced with the double MMD-based kernel described in Park et al. [3]. This acts as an ablation study, helping to isolate the effect of the signature kernel

We compare these methods on three inference tasks for three simulators: an Ornstein-Uhlenbeck process (results in Figure 7); a moving average model of order 2 (results in Figure 8); and a partially observed stochastic epidemic model generating multivariate time-series (results shown in Table 2).

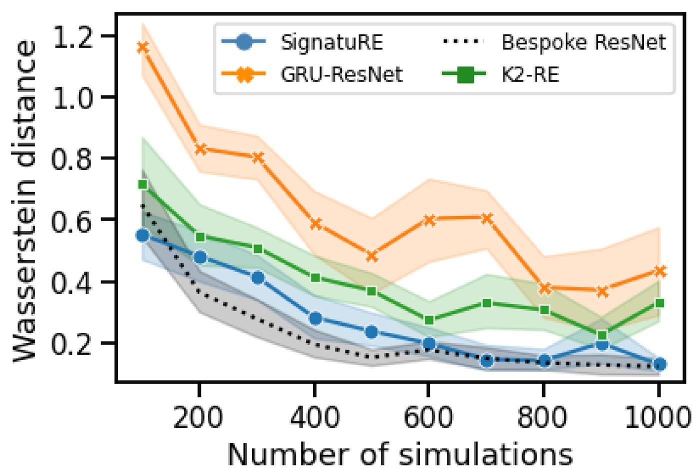

Figure 7 – Experimental results for the Ornstein-Uhlenbeck process.

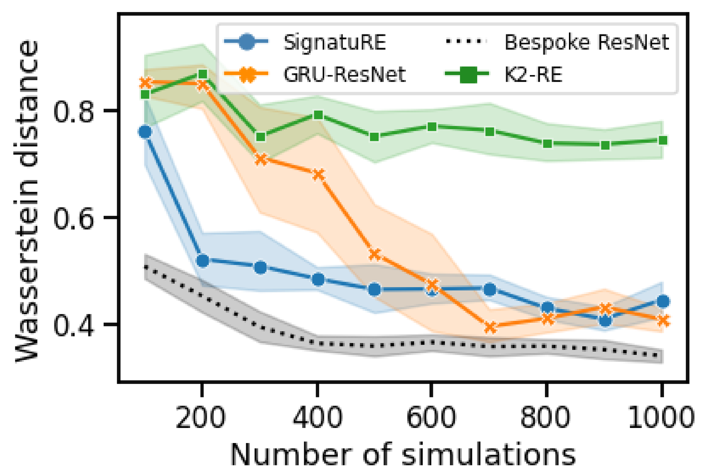

Figure 8 – Experimental results for the moving average model of order 2.

Method	Simulation budget				
	50	100	200	500	1000
GRU-ResNet	0.434	0.425	0.355	0.273	*0.090*
K2-RE	*0.417*	0.432	0.407	0.454	0.431
K2-RE-5	0.440	0.427	0.374	*0.206*	0.255
SignatuRE	0.430	*0.411*	*0.351*	0.513	0.321
SignatuRE-5	**0.241**	**0.333**	**0.176**	**0.133**	**0.083**
Bespoke ResNet	0.379	0.222	0.146	0.104	0.092

Table 2 – Median Wasserstein distance from SMC-ABC [4] posterior for the partially-observed epidemic model (from 10 seeds). Smaller values are better. **Bold** and *italics* indicate best and second-best, respectively, of the methods that do not use pre-defined summary statistics.

From these experiments, we see that SignatuRE can outperform even sophisticated neural architectures when the simulation budget is extremely constrained. Such an approach may therefore be useful for very expensive simulators.

References

[1] Joeri Hermans, Volodimir Begy, and Gilles Louppe. Likelihood-free mcmc with amortized approximate ratio estimators. In *International Conference on Machine Learning*, pages 4239–4248. PMLR, 2020.

[2] Joel Dyer, Patrick W. Cannon, and Sebastian M. Schmon. Amortised Likelihood-free Inference for Expensive Time-series Simulators with Signatured Ratio Estimation . In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pages 11131–11144, 2022.

[3] Mijung Park, Wittawat Jitkrittum, and Dino Sejdinovic. K2-ABC: Approximate Bayesian computation with kernel embeddings. *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS 2016*, 41:398–407, 2016.

[4] Mark A. Beaumont, Jean-Marie Cornuet, Jean-Michel Marin, and Christian P. Robert. Adaptive approximate bayesian computation. *Biometrika*, 96(4):983–990, 2009. ISSN 00063444, 14643510. URL `http://www.jstor.org/stable/27798882`.