Journal of the Royal Statistical Society Series B: Statistical Methodology, 2024, **86**, 302–339

Journal of the Statistics Society
Series B
Statistical Methodology

Discussion Paper Contribution

Proposer of the vote of thanks and contribution to the Discussion of 'the Discussion Meeting on Probabilistic and statistical aspects of machine learning'

Darren Wilkinson

Department of Mathematical Sciences, Durham University, Durham, UK

Address for correspondence: Darren Wilkinson, Department of Mathematical Sciences, Durham University, South Road, Durham DH1 3LE, UK. Email: darren.j.wilkinson@durham.ac.uk

Computational statistics and machine learning (ML) are closely related, and there are many opportunities for cross-fertilization of ideas between the two fields. Both can benefit from greater interaction, and the two papers being discussed here highlight some ways that this can happen.

1 Automatic change-point detection in time series via deep learning

The main focus of this paper is offline detection of a single change-point using labelled training data. Interest is in the automatic generation of new offline detection methods using neural networks whilst providing statistical guarantees of method performance. Theory is developed for a class of multi-layer perceptrons (MLPs) that directly generalize existing cumulative sum-based methods.

The theory developed in the paper applies to a MLP with ReLU activation, and this basic model is amenable to analysis. However, the theory only requires a single layer, and the examples all use MLPs of constant layer width, which is rarely seen in practice. Can the authors provide practical advice on choosing network depth and layer widths sensibly and safely? In particular, are there practical issues relating to the width condition, $m_r m_{r+1} = \mathcal{O}(n \log n)$? The theoretical bounds suggest the need for a lot of training data, but empirically it seems that these may be overly conservative. Do the authors have any insight into this apparent mismatch? Neural networks often work better with scaled data, and min–max scaling is used for the examples in the paper, but is this safe in the presence of heavy-tailed noise?

For the application based on activity data, a more sophisticated neural network architecture is adopted for which the theoretical results provided do not directly apply. What hope is there of extending the theory to such models, and in the absence of this, what practical advice can be given? It is mentioned in the paper that in the absence of labelled data, but in the presence of a full data generating process, a simulator can be used to train the network. This is *simulation-based inference* (SBI) (Cranmer et al., 2020), and it is worth establishing the connection with this literature, where the use of neural networks has become a standard practice in recent years.

2 From denoising diffusions to denoising Markov models

Denoising Markov models (DMMs) are deep generative models for simulating (conditional) samples from a data distribution. Huge training data sets are required, but these can be replaced by a

Received: September 13, 2023. **Accepted:** September 20, 2023

data generating process in the context of SBI. Typical applications are to very high dimensional data such as images, but the methods can also be used for sampling Bayesian posterior distributions. These models are very expensive both to train and sample (even relative to other deep generative models), but are considered state-of-the-art for certain problems.

The paper provides a unifying framework for a broad class of models of the denoising diffusion form with fairly arbitrary state spaces. The emphasis is on continuous time, but the connection with discrete time formulations is clearly articulated. Conditional simulation is also covered and briefly discussed. The approach is to work with continuous time Markov processes on a general state space, and to formulate the (de)noising process in terms of the generator of the Markov process. The resulting optimization targets are shown to generalize several different special cases that have appeared in the literature for particular state spaces.

Details of how to generate samples are missing from the main paper, but are very important in practice. The examples described in the online supplementary material seem to use approximate first-order methods based on a regular time grid, but this is probably not optimal. A benefit of formulating the models in continuous time is the possibility of using higher-order methods with adaptive time steps. At least one of the examples used a time-rescaling—might adaptive time-stepping reduce the need for this, or is that a separate issue? It is sometimes convenient to have a deterministic generation mechanism using a probability flow differential equation (Song et al., 2021). Is such an approach covered by the general DMM framework presented here? What about Schrödinger bridge (Shi et al., 2022) approaches?

Everything depends on using a 'good' neural network architecture for the denoising process, but can anything general be said about how to choose the architecture for a given problem? Do we understand the kinds of problems for which DMMs work well? Why are not these models more widely used for SBI? Given the magnitude of the computational machinery dedicated to the problem, the *g*-and-*k* example was not especially compelling (see, e.g. Figure 8 in the online supplementary material), despite being a fairly standard low-dimensional Bayesian inference problem. Could issues be diagnosed in the absence of ground truth, and could the model be tuned to improve performance if desired? Are there examples in the literature of DMMs being used for problems with a mixed discrete and continuous state space?

3 Summary

These two papers illustrate different aspects of the interaction between statistics and ML. From the perspective of academic statistics, we are likely to see increasing use of modern ML methods in statistical methodology. It is likely to become difficult to draw a clear line between computational statistics and ML, but this comes with challenges, since the language and culture of the two communities remain quite distinct. Programming languages also illustrate potential issues: Python is the language typically used for ML, with tensor frameworks such as TensorFlow (used for Paper 1), JAX (used for Paper 2), and Torch, but most academic statisticians currently use R by default.

The opportunities for sharing ideas between statistics and ML are great and growing. The two papers presented here are important contributions in their own right, and also serve to highlight the potential benefits of narrowing the gap between the two communities. *It therefore gives me great pleasure to propose the vote of thanks.*

Conflict of interest: None declared.

References

Cranmer K., Brehmer J., & Louppe G. (2020). The frontier of simulation-based inference. *PNAS*, *117*(48), 30055–30062. https://doi.org/10.1073/pnas.1912789117

Shi Y., De Bortoli V., Deligiannidis G., & Doucet A. (2022). Conditional simulation using diffusion Schrödinger bridges. In J. Cussens and K. Zhang (Eds.), *Proceedings of the thirty-eighth conference on uncertainty in artificial intelligence: Vol. 180. Proceedings of machine learning research* (pp. 1792–1802). PMLR.

Song Y., Sohl-Dickstein J., Kingma D. P., Kumar A., Ermon S., & Poole B. (2021). 'Score-based generative modeling through stochastic differential equations', *International Conference on Learning Representations*. https://openreview.net/forum?id=PxTIG12RRHS

https://doi.org/10.1093/jrsssb/qkad160
Advance access publication 21 December 2023