Bayesian Deep Learning

Maurizio Filippone
maurizio.filippone@kaust.edu.sa

Statistics Program
King Abdullah University of Science and Technology

February 19, 2025

Decision-Making

Decision-making is a critical step in several domains [**Norvig and Russell**, **1995**]:

- Policy-making for the environment
- Healthcare
- Society
- ...

Decision-Making

Decision-making is a critical step in several domains [**Norvig and Russell**, **1995**]:

- Policy-making for the environment
- Healthcare
- Society
- ...

$$\text{Decision Theory} = \text{Probabilistic reasoning} + \text{Utility theory}$$

Decision-Making

Decision-making is a critical step in several domains [**Norvig and Russell**, **1995**]:

- Policy-making for the environment
- Healthcare
- Society
- ...

Decision Theory = Probabilistic reasoning + Utility theory

Is Deep Learning effective in assisting decision-making?

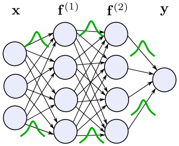
Image prediction: ping-pong ball
Confidence: 99.99%

Illustration: Dianna "Mick" McDougall, Photo: ResNeXtGuesser

Image prediction: pineapple
Confidence: 99.3%

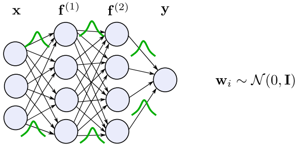
Illustration: Dianna “Mick” McDougall, Photo: ResNeXtGuesser

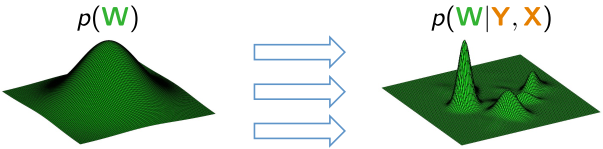
- Inputs : $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$
- Labels : $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$
- Weights : $\mathbf{W} = \{\mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(L)}\}$



$\mathbf{w}_i \sim \mathcal{N}(0, \mathbf{I})$

- Inputs : $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$
- Labels : $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$
- Weights : $\mathbf{W} = \{\mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(L)}\}$

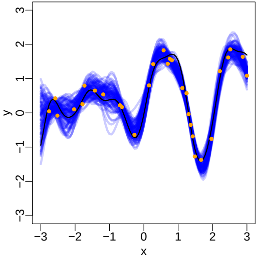


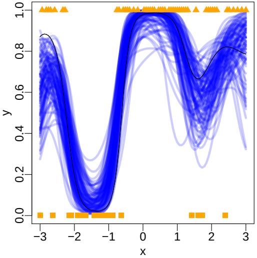


$$p(\mathbf{W}|\mathbf{Y}, \mathbf{X}) = \frac{p(\mathbf{Y}|\mathbf{X}, \mathbf{W})p(\mathbf{W})}{\int p(\mathbf{Y}|\mathbf{X}, \mathbf{W})p(\mathbf{W})d\mathbf{W}}$$

- Predictions consider an infinite number of parameter configurations (e.g., [**Bishop**, **2006**])

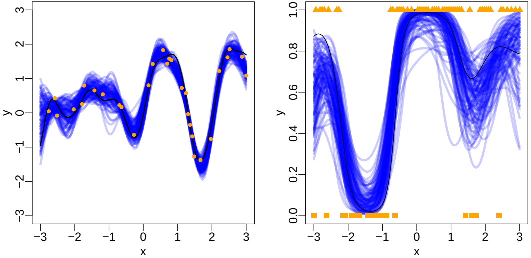
$$p(\mathbf{y}^*|\mathbf{x}^*, \mathbf{Y}, \mathbf{X}) = \int p(\mathbf{y}^*|\mathbf{x}^*, \mathbf{W})p(\mathbf{W}|\mathbf{Y}, \mathbf{X})d\mathbf{W}$$





- Predictions consider an infinite number of parameter configurations (e.g., [**Bishop**, **2006**])

$$p(\mathbf{y}^*|\mathbf{x}^*, \mathbf{Y}, \mathbf{X}) = \int p(\mathbf{y}^*|\mathbf{x}^*, \mathbf{W})p(\mathbf{W}|\mathbf{Y}, \mathbf{X})d\mathbf{W}$$



Combining the flexibility of Deep Learning with sound uncertainty quantification!

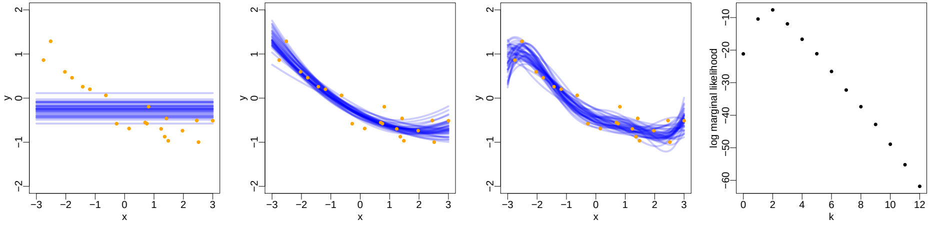
- The normalization term in Bayes theorem can be used for model selection (e.g., [**Bishop**, **2006**; **Gelman et al.**, **2013**])

$$p(\mathbf{Y}|\mathbf{X}, \mathcal{H}) = \int p(\mathbf{Y}|\mathbf{X}, \mathbf{W}, \mathcal{H}) p(\mathbf{W}|\mathcal{H}) d\mathbf{W}$$

- The normalization term in Bayes theorem can be used for model selection (e.g., [**Bishop**, **2006**; **Gelman et al.**, **2013**])

$$p(\mathbf{Y}|\mathbf{X},\mathcal{H}) = \int p(\mathbf{Y}|\mathbf{X},\mathbf{W},\mathcal{H})p(\mathbf{W}|\mathcal{H})d\mathbf{W}$$

- Illustration of different hypotheses $\mathcal{H}$ for a simple linear model:

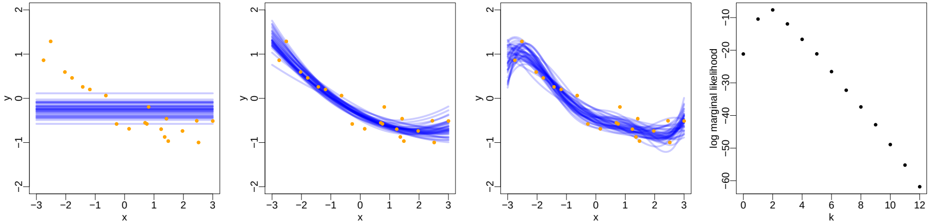


جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

- The normalization term in Bayes theorem can be used for model selection (e.g., [**Bishop**, **2006**; **Gelman et al.**, **2013**])

$$p(\mathbf{Y}|\mathbf{X}, \mathcal{H}) = \int p(\mathbf{Y}|\mathbf{X}, \mathbf{W}, \mathcal{H}) p(\mathbf{W}|\mathcal{H}) d\mathbf{W}$$

- Illustration of different hypotheses $\mathcal{H}$ for a simple linear model:



Unfortunately, for Deep Learning models this is intractable!

- Leverage statistical connections [**Akaike**, **1973**] between the model selection problem:

$$\arg\max_{\mathcal{H}} \{\log [p(\mathbf{Y}|\mathbf{X}, \mathcal{H})]\}$$

and the minimization of the following Kullback-Leibler divergence [**Tran et al.**, **NeurIPS 2021**]:

$$\arg\min_{\mathcal{H}} \{\mathcal{D}_{\mathrm{KL}}[\pi_{\mathrm{TRUE}}(\mathbf{Y}|\mathbf{X}) \,\|\, p(\mathbf{Y}|\mathbf{X}, \mathcal{H})]\}$$

to obtain simple and tractable objectives for model selection.

Research Project Idea

- Leverage statistical connections [**Akaike**, **1973**] between the model selection problem:

$$\arg\max_{\mathcal{H}} \left\{ \log \left[p(\mathbf{Y}|\mathbf{X}, \mathcal{H}) \right] \right\}$$

and the minimization of the following Kullback-Leibler divergence [**Tran et al.**, **NeurIPS 2021**]:

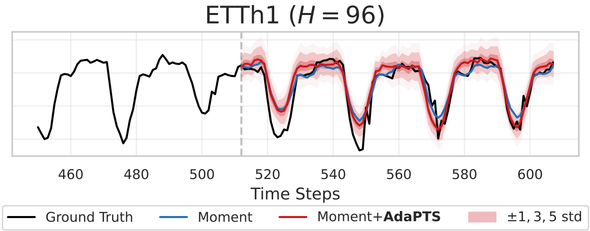
$$\arg\min_{\mathcal{H}} \left\{ \mathcal{D}_{\mathrm{KL}} \left[\pi_{\mathrm{TRUE}}(\mathbf{Y}|\mathbf{X}) \,\|\, p(\mathbf{Y}|\mathbf{X}, \mathcal{H}) \right] \right\}$$

to obtain simple and tractable objectives for model selection.

This can be leveraged for architecture search!

Research Project Idea - Why does it matter?

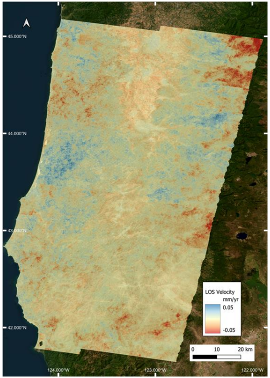
- Occam's razor [**Solomonoff**, **1964**] for Deep Learning
- Efficient (compute and size) Deep Learning models
- Parsimonious probabilistic foundation models

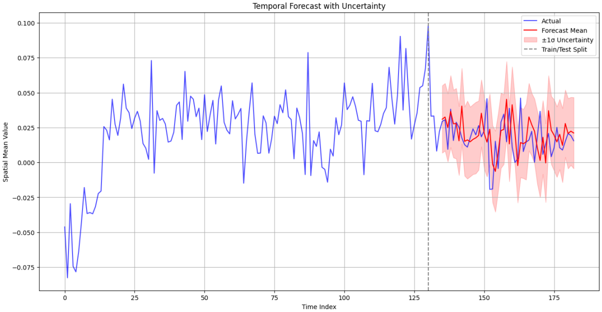


Research Project Idea - Why KAUST?

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

- New foundation models for applications in Statistics





جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

- Key collaborations in Deep Learning and Statistics worldwide

Position: Bayesian Deep Learning is Needed in the Age of Large-Scale AI

Theodore Papamarkou [1] Maria Skoularidou [2] Konstantina Palla [3] Laurence Aitchison [4] Julyan Arbel [5] David Dunson [6] Maurizio Filippone [7] Vincent Fortuin [8 9 10] Philipp Hennig [11] José Miguel Hernández-Lobato [12] Aliaksandr Hubin [13 14] Alexander Immer [15] Theofanis Karaletsos [16] Mohammad Emtiyaz Khan [17] Agustinus Kristiadi [18] Yingzhen Li [19] Stephan Mandt [20] Christopher Nemeth [21] Michael A. Osborne [22] Tim G. J. Rudner [23] David Rügamer [10 24] Yee Whye Teh [25 26] Max Welling [27] Andrew Gordon Wilson [28] Ruqi Zhang [29]

Abstract

In the current landscape of deep learning research, there is a predominant emphasis on achieving high predictive accuracy in supervised tasks involving large image and language datasets. However, a broader perspective reveals a multitude of over-uncertainty, active and continual learning, and scientific data, that demand attention. Bayesian deep learning (BDL) constitutes a promising avenue, offering advantages across these diverse settings. This paper posits that BDL can elevate the capabilities of deep learning. It revisits the strengths of BDL, acknowledges existing challenges, and

Spatial Statistics 60 (2024) 100825

Contents lists available at ScienceDirect

Spatial Statistics

journal homepage: www.elsevier.com/locate/spasta

ELSEVIER

SPATIAL STATISTICS

Check for updates

Spatial Bayesian neural networks

Andrew Zammit-Mangion [a,*], Michael D. Kaminski [a], Ba-Hien Tran [b], Maurizio Filippone [c], Noel Cressie [a]

[a] School of Mathematics and Applied Statistics, University of Wollongong, Australia
[b] Paris Research Centre, Huawei Technologies, France
[c] Statistics Program, King Abdullah University of Science and Technology, Saudi Arabia

ARTICLE INFO

Keywords:
Gaussian process
Hamiltonian Monte Carlo

ABSTRACT

Statistical models for spatial processes play a central role in analyses of spatial data. Yet, it is the simple, interpretable, and well understood models that are routinely employed even though,

- Unique World-class AI and Statistics research environment!
- Vision 2030 drives methodological questions!

Thank you!

Questions?