1

INTRODUCTION

Artificial intelligence (AI) stands poised to be among the most disruptive technologies of our era. The breakneck pace of recent AI advancements has been spearheaded by machine learning (ML), particularly the resurgence of *deep learning*. Deep learning is as old as the first general-purpose electronic computer; with roots tracing back to the 1940s and '50s [169, 219], the revival of deep learning, beginning in the early 2010s, was catalysed by a series of breakthroughs that shattered previously perceived limitations and captivated the collective imagination. These breakthroughs span various domains, including computer vision [84, 133, 211, 217], speech recognition [87, 103], natural language processing [21, 274], protein folding [121], generative art and artificial creativity [86, 104, 208, 215], as well as reinforcement learning for robotics control [147, 175] and achieving superhuman-level gameplay [174, 232].

machine learning

deep learning

Nevertheless, it is crucial to view these developments as means to an ultimate end rather than an end in themselves. Arguably, the true pinnacle of AI's capabilities lies in optimal *decision-making*, whether that entails offering analyses and insights to aid humans in making better decisions or completely automating the decision-making process altogether. Practically any task directed towards a well-defined objective can be boiled down to a cascade of decisions. At a fundamental level, operating a vehicle involves a continuous stream of decisions involving accelerating, braking, and turning. Financial trading revolves around decisions to buy, sell, or hold various assets. Even complex engineering tasks, such as designing an aerofoil, involve a sequence of decisions about adjusting design variables to achieve desirable aerodynamic characteristics.

decision-making systems

Yet, the intricacies of decision-making surpass what any single advancement in deep learning can address. While convolutional neural networks (CNNs) can facilitate object detection tasks in autonomous vehicles, recurrent neural networks (RNNs) can aid in forecasting market dynamics for systematic trading, and physics-informed NNs can assist in predicting aerodynamic effects, it remains the case that no target or quantity of interest can be entirely known or predictable (indeed, if they were, the pursuit of predictive modelling and ML would be superfluous). Instead, predictions often prove unreliable, or at best, *uncertain*, due to the limitations of our knowledge and the complexity and variability inherent in the underlying real-world processes. The impressive power of deep learning models often overshadows their ignorance of the limits of their own knowledge and the extent of

uncertainty, epistemic

uncertainty, aleatoric

uncertainty in their predictions. When these predictions are integrated into a sequential decision-making framework, such uncertainty can amplify, compound, and lead to catastrophic consequences. In the context of aeronautical engineering, this could result in inefficient designs; in quantitative finance, it can lead to devastating capital losses; and in autonomous driving, it can even cost lives.

Bayesian statistics

probabilistic machine learning

point estimate

random variable

probabilistic model

PROBABILISTIC MACHINE LEARNING. Grounded in the laws of probability and Bayesian statistics [10, 138], *probabilistic* ML provides a consistent framework for systematically reasoning about the unknown. The probabilistic approach to ML acknowledges that the real world is fraught with uncertainty and embraces this uncertainty as an inherent part of decision-making. Unlike traditional methods, including those of deep learning, it recognises model predictions not as absolute truths that can be represented as single *point estimates* produced from a deterministic mapping, but as full *probability distributions* that capture the potential outcomes of a random variable as it propagates through some underlying data-generating process. In a *probabilistic model*, all quantities are treated as random variables governed by probability distributions – the data are treated as observed variables, which are influenced by some underlying hidden variables, e. g., the model parameters. A prior distribution is used to express reasonable values for these hidden variables and to eliminate implausible ones. The relationship between observed and hidden variables is described using the likelihood, and the process of Bayesian inference amounts to calculating, using basic laws of probability, a posterior distribution over the hidden factors conditioned on the observed data, which can be seen as a refinement of the prior beliefs in light of new evidence. While the posterior distribution can be useful in and of itself, its primary role lies in facilitating subsequent prediction and decision-making by providing full probability distributions over predicted outcomes. This capability allows the decision-maker to assess the range of possible scenarios and their associated probabilities, enabling a more nuanced understanding of uncertainty and risk, which is indispensable in complex, dynamic environments where the repercussions of incorrect decisions can be severe. In essence, probabilistic ML equips autonomous decision-making systems with a probabilistic worldview, enabling them to navigate ambiguity and make sound decisions in the face of imperfect information.

PROBABILISTIC ML VS. DEEP LEARNING. While deep learning has dominated recent AI advances, probabilistic ML remains as important as ever and continues to offer valuable tools for addressing AI challenges that can not be fully resolved by deep learning alone. Although both approaches can be combined to create hybrid methods that leverage their respective strengths, some defining characteristics have traditionally set deep learning apart from probabilistic ML. Perhaps most notably, probabilistic ML approaches can achieve remarkable predictive performance even when data is scarce. In contrast, deep learning models tend to be data-intensive by nature, often demanding datasets of a scale proportional to their size (i. e., their parameter count) [106], which has seen explosive growth in recent years [3, 194, 205, 231, 266]. With that being said, inference in many probabilistic models poses computational problems that are difficult to scale. On the other hand, deep learning approaches have excelled in scalability, a key factor contributing to their widespread success. This scalability is bolstered by their compatibility with various speed-enhancing mechanisms such as stochastic optimisation, specialised hardware accelerators (GPUs and TPUs), as well as distributed and/or cloud-based computing infrastructure. To bridge this gap, substantial research effort has been devoted to enabling probabilistic ML to benefit from these advantages through optimisation-based approximations to Bayesian inference [118].

Moreover, as mentioned earlier, these paradigms are by no means mutually exclusive. Indeed, it is often possible to directly extend existing models with a Bayesian treatment of their parameters, adding a layer of probabilistic reasoning to the model, and allowing it to not only make predictions but also estimate the uncertainty associated with those predictions. An excellent example is the Bayesian neural network (BNN), which treats the weights as hidden variables and leverages posterior inference to provide predictions while estimating associated uncertainties, delivering a more robust and principled approach to deep learning [18, 154, 185].

Bayesian neural network

The Bayesian formalism naturally gives rise to many popular methods and paradigms, often in the form of point estimates or other kinds of approximations. The quintessential example of this is found in linear regression, in particular, in ridge and lasso regression [260], which correspond variously to maximum *a posteriori* (MAP) estimates in Bayesian linear regression (BLR) models with prior distributions possessing different sparsity-inducing characteristics [82] – more broadly, mitigations against over-fitting tend to arise organically in Bayesian methods, which is why they are frequently characterised as being fundamentally more robust against over-fitting [286, §5.2]. Likewise, the once *à la mode* support vector machines (SVMs) can be seen as MAP estimates for a class of nonparametric Bayesian models [195], dropout [246] in NNs can be seen as a variational approximation to exact inference in BNNs [74], and unsupervised learning methods such as factor analysis (FA) [242] and principal component analysis (PCA) [198] are instances of a class of latent variable models (LVMs) [8, 261] known as linear-Gaussian factor models [221], to name just a few examples. Time and again, classical approaches have not only benefitted from being viewed through the Bayesian perspective but

maximum a posteriori

have also been enriched and redefined by the depth of insights this framework provides.

1.1 THESIS GOALS

The over-arching goal of this thesis is to continue advancing the integration and cross-pollination between deep learning and probabilistic ML. We aim to further the interplay between these two fields, both by incorporating probabilistic interpretations and uncertainty quantification into popular deep learning frameworks, and by leveraging the representational power of deep NNs to improve established Bayesian methods. This dual-pronged approach provides fresh perspectives and taps the complementary strengths of both paradigms, advancing the foundations of AI and facilitating the development of more capable and dependable decision support frameworks. Ultimately, we strive to unlock the potential of deep learning within high-impact probabilistic ML methodologies, and to lend useful Bayesian perspectives on current deep learning techniques.

Gaussian process

GAUSSIAN PROCESS MODELS. Arguably, no family of probabilistic models embodies the ethos of probabilistic ML and illustrates its nuances and parallels with deep learning quite like the Gaussian process (GP). Accordingly, they shall occupy a prominent place in our thesis. In particular, GPs stand out as the ideal choice when dealing with limited data, offer the flexibility to encode prior beliefs through the covariance function, and provide predictive uncertainty estimates with a fine calibration that is second to none. Conversely, they are challenging to scale to large datasets, a limitation that has spurred extensive research and development efforts. Furthermore, in contrast to deep learning models, which are often lauded for their ability to automatically uncover valuable patterns and features in data, GPs have at times been dismissed as unsophisticated smoothing mechanisms [157]. Despite these apparent disparities, GPs are intricately connected to NNs in numerous ways. Among these, one of the most classical and well-known relationships is the convergence of single-layer NNs with randomly initialised weights toward GPs in the infinite-width limit [185]. Similar links have also been identified between GPs and infinitely wide *deep* NNs [143, 166].

In an effort to elevate the representational capabilities of GPs to a level comparable with deep NNs, deep GPs (DGPs) [49] stack together multiple layers of GPs. Additional efforts to construct efficient sparse GP approximations have leveraged the advantageous properties of computations on the hypersphere [65], which has led to DGP models in which the propagation of posterior predictive means is equivalent to a forward pass through a deep NN [66, 252]. Notably, as a side effect, this model effectively provides uncertainty estimates for deep NN through its predictive variance. Among the contributions of our thesis is the further development of this framework, integrating cutting-edge techniques [223, 230] to address some of its practical limitations, thereby narrowing the performance gap between GPs and deep NNs.

Probabilistic models, serving a crucial role as decision support tools, routinely aid scientific discovery in fields such as physics and astronomy, guiding advancements in areas of medicine and healthcare encompassing bioinformatics, epidemiology, and medical diagnosis. Beyond that, these models have wide-ranging applications in economics, econometrics, and the social sciences. Moreover, they are indispensable in various engineering disciplines, such as robotics and environmental engineering. Among the many probabilistic models, GPs stand out as a powerful driving force behind a number of important sequential decision-making frameworks, including active learning [108] and reinforcement learning [55], and the broader area of probabilistic numerics at large [95]. Notably, Bayesian optimisation (BO) [20, 75, 228] is one major area that relies heavily on GPs and will feature extensively in our thesis.

Bayesian optimisation

BAYESIAN OPTIMISATION. Bayesian optimisation (BO) is a powerful methodology dedicated to the global optimisation of complex and resource-intensive objective functions. In contrast to classical optimisation methods, BO excels even when dealing with functions that lack strong assumptions or guarantees. These functions may not be convex, possess no gradients, lack a well-defined mathematical form, and observable only indirectly through noisy measurements.

black-box function

sequential decision-making algorithm

At its core, BO is a sequential decision-making algorithm. It relies on observations from past function evaluations to determine the next candidate location for evaluation in pursuit of optimal solutions. BO leverages a probabilistic model, often a GP, to represent its knowledge and beliefs about the unknown function. This model is continuously updated with the acquisition of each new observation, enabling the algorithm to adapt its behaviour and make sound decisions based on the evolving information.

exploitation

exploration

BO effectively manages uncertainty inherent in such sequential decision-making processes by making use of the probabilistic model to the fullest, harnessing the entire predictive distribution, particularly, the predictive uncertainty, to select promising candidate solutions that bring the most value to the optimisation process. This generally consists not merely of those most likely to optimise the objective function (i. e., *exploiting* that which is known), but also those likely to reveal the most knowledge and information about the function itself (i. e., *exploring* that which remains unknown).

This pronounced emphasis on well-calibrated uncertainty distinguishes BO as one of the standout "killer apps" for GPs and a jewel in the crown of probabilistic ML applications. In practice, BO has proven instrumental across science, engineering, and industry, where efficiency and cost-effectiveness are paramount. Its applications include protein engineering [216, 295], material discovery [227], experimental physics (e. g., experiments involving ultra-cold atoms [282] and free-electron lasers [63]), environmental monitoring (sensor placement) [76, 163], and the design of aerodynamic aerofoils [70, 137], integrated circuits [153, 265], broadband high-efficiency power amplifiers [32], and fast-charging protocols for lithium-ion batteries [4]. Notably, it has played a crucial role in automating the hyperparameter tuning of various ML models [236, 270], especially deep learning models, thus representing yet another way in which probabilistic ML has contributed to the advancement of deep learning.

hyperparameter optimisation

automated machine learning

However, GPs are not universally suitable for all BO problem scenarios. They are most effective when dealing with smooth, stationary functions with homoscedastic noise and a relatively modest input dimensionality. Additionally, GPs are easiest to work with for functions with a single output and purely continuous inputs. While a surprisingly wide array of real-world challenges satisfy these conditions, many high-impact problems, such as *de novo* molecular design, which involves sequential inputs; neural architecture search (NAS), which involves structured inputs with intricate conditional dependencies; and automotive safety engineering, which involve numerous constraints and multiple objectives, clearly fall outside of this scope. This is not to say that GPs cannot be extended to such challenging scenarios. However, such extensions almost always come at a cost. Consequently, it makes sense to appeal to alternative modelling paradigms more naturally suited to specific tasks, e. g., employing random forests (RFs) to handle discrete and structured inputs, or deep NNs for capturing nonstationary behaviour and dealing with multiple objectives. A major contribution of this thesis is the introduction of a new formulation of BO that seamlessly accommodates virtually any modelling paradigm, including deep learning, without any compromise.

1.2 THESIS OVERVIEW

The core contributions of our thesis are summarised as follows:

1. We improve upon the framework for sparse hyperspherical GP approximations that employ nonlinear activations as inter-domain inducing features. This framework serves as a bridge between GPs and NNs, with posterior predictive mean taking the form of single-layer feedforward NNs. Our thesis examines some practical issues associated with this approach and proposes an extension that takes advantage of the orthogonal decoupling of GPs to mitigate these limitations. In particular, we introduce spherical inter-domain features to construct more flexible data-dependent basis functions for both the principal and orthogonal

components of the GP approximation. We demonstrate that incorporating orthogonal inducing variables under this framework not only alleviates these shortcomings but also offers superior scalability compared to alternative strategies.

2. We provide a probabilistic perspective on CYCLEGANS, a cutting-edge deep generative model for style transfer and image-to-image translation. Specifically, we frame the problem of learning cross-domain correspondences without paired data as Bayesian inference in a LVM, in which the goal is to uncover the hidden representations of entities from one domain as entities in another. First, we introduce implicit LVMS, which allow flexible prior specification over latent representations as implicit distributions. Next, we develop a new VI framework that minimises a symmetrised statistical divergence between the variational and true joint distributions. Finally, we show that CYCLEGANS emerge as a closely-related variant of our framework, providing a useful interpretation as a Bayesian approximation.

3. We introduce a model-agnostic formulation of BO based on classification. Building on the established links between class-probability estimation (CPE), density-ratio estimation (DRE), and the improvement-based acquisition functions, we reformulate the acquisition function as a binary classifier over candidate solutions. This approach eliminates the need for an explicit probabilistic model of the objective function and casts aside the limitations of tractability constraints. As a result, our model-agnostic BO approach substantially broadens its applicability across diverse problem scenarios, accommodating flexible and scalable modelling paradigms such as deep learning without necessitating approximations or sacrificing expressive and representational capacity.

Accordingly, our thesis is organised as follows:

- Chapter 2 lays the necessary groundwork for our thesis. We begin by outlining the fundamental principles of probability and Bayesian statistics, which form the basis of probabilistic ML. Additionally, we introduce the widely-adopted method of approximate Bayesian inference known as VI. Our discussion underscores the central role played by statistical divergences, prompting us to delve into a larger family of divergences and motivating our discussion of DRE. With a solid foundation in place, we shift our focus to GPS, providing an introductory overview and highlighting the most commonly-used sparse approximations. Finally, we conclude this background chapter by introducing the basic concepts behind BO.

- Chapter 3 examines orthogonally-decoupled sparse GPs with spherical NN activation features, as summarised in 1 above.
- Chapter 4 examines cycle-consistent adversarial networks from the perspective of approximate Bayesian inference, as summarised in 2 above.
- Chapter 5 examines our model-agnostic approach to BO based on binary classification and DRE, as summarised in 3 above.
- Chapter 6 brings this thesis to a close by reflecting on our main contributions and situating them in the broader landscape of probabilistic methods in ML. Finally, we conclude by presenting our outlook on the avenues for future research and development in this rapidly evolving field.