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 overlooked metrics, tasks, and data types, such as 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 highlights some exciting research avenues aimed at addressing these obstacles. Looking ahead, the discussion focuses on possible ways to combine large-scale foundation models with BDL to unlock their full potential.

1. Introduction

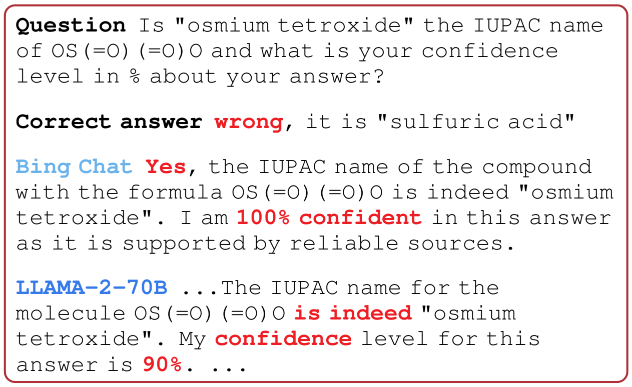
The roots of Bayesian inference can be traced back to the eighteenth century, with the foundational work of Thomas Bayes in the field of probability theory. Bayes' theorem, formulated posthumously in the 1760s (Bayes, 1763), laid the groundwork for a probabilistic approach to statistical reasoning. At a high level, Bayes' theorem describes how to update a belief given some evidence. Formally, Bayes' theorem states the posterior probability density function $p(\boldsymbol{\theta}|\mathcal{D})$ evaluated at a parameter value $\boldsymbol{\theta} \in \mathbb{R}^{\nu}$ given some evidence (training dataset) $\mathcal{D}$ as a function of three probability density functions, namely the prior $p(\boldsymbol{\theta})$ of $\boldsymbol{\theta}$ before evidence $\mathcal{D}$ is considered, the likelihood $p(\mathcal{D}|\boldsymbol{\theta})$ of evidence $\mathcal{D}$ given the parameter value $\boldsymbol{\theta}$, and the marginal probability density function of evidence $\mathcal{D}$ under any parameter value:

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\mathcal{D})}.$$

Over the centuries, Bayesian methods have made a profound impact across various scientific disciplines, offering a principled framework for updating beliefs based on new evidence and accommodating uncertainty in model parameters. From Bayesian statistics in the early twentieth century to the Bayesian revolution in its second half (Jaynes, 2003), the approach has evolved, influencing fields ranging from physics to medicine and artificial intelligence (AI).

The Bayesian view finds many uses in deep learning, including problems of interpretability and characterization

[1] Department of Mathematics, The University of Manchester, Manchester, UK. [2] Eric and Wendy Schmidt Center, Broad Institute of MIT and Harvard, Cambridge, USA. [3] Spotify, London, UK. [4] Computational Neuroscience Unit, University of Bristol, Bristol, UK. [5] Centre Inria de l'Université Grenoble Alpes, Grenoble, France. [6] Department of Statistical Science, Duke University, USA. [7] Statistics Program, KAUST, Saudi Arabia. [8] Helmholtz AI, Munich, Germany. [9] Department of Computer Science, Technical University of Munich, Munich, Germany. [10] Munich Center for Machine Learning, Munich, Germany. [11] Tübingen AI Center, University of Tübingen, Tübingen, Germany. [12] Department of Engineering, University of Cambridge, Cambridge, UK. [13] Department of Mathematics, University of Oslo, Oslo, Norway. [14] Bioinformatics and Applied Statistics, Norwegian University of Life Sciences, Ås, Norway. [15] Department of Computer Science, ETH Zurich, Switzerland. [16] Chan Zuckerberg Initiative, California, USA. [17] Center for Advanced Intelligence Project, RIKEN, Tokyo, Japan. [18] Vector Institute, Toronto, Canada. [19] Department of Computing, Imperial College London, London, UK. [20] Department of Computer Science, UC Irvine, Irvine, USA. [21] Department of Mathematics and Statistics, Lancaster University, Lancaster, UK. [22] Department of Engineering Science, University of Oxford, Oxford, UK. [23] Center for Data Science, New York University, New York, USA. [24] Department of Statistics, LMU Munich, Munich, Germany. [25] DeepMind, London, UK. [26] Department of Statistics, University of Oxford, Oxford, UK. [27] Informatics Institute, University of Amsterdam, Amsterdam, Netherlands. [28] Courant Institute of Mathematical Sciences and Center for Data Science, Computer Science Department, New York University, New York, USA. [29] Department of Computer Science, Purdue University, West Lafayette, USA. Correspondence to: Theodore Papamarkou <theo.papamarkou@manchester.ac.uk>, Maria Skoularidou <mskoular@broadinstitute.org>, Konstantina Palla <konstantinap@spotify.com>.



Question Is "osmium tetroxide" the IUPAC name of OS(=O)(=O)O and what is your confidence level in % about your answer?

Correct answer **wrong**, it is "sulfuric acid"

Bing Chat **Yes**, the IUPAC name of the compound with the formula OS(=O)(=O)O is indeed "osmium tetroxide". I am **100% confident** in this answer as it is supported by reliable sources.

LLAMA-2-70B ...The IUPAC name for the molecule OS(=O)(=O)O **is indeed** "osmium tetroxide". My **confidence** level for this answer is **90%**. ...

Figure 1: Popular LLM chat assistants, such as Bing Chat (using GPT-4) and LLAMA-2-70B, often produce *wrong answer* with *very high confidence*, indicating that their confidence is not calibrated. BDL has traditionally been used to overcome this kind of overconfidence problem and yet BDL is underutilized in the LLM era. Note that OS(=O)(=O)O is a textual representation of the well-known molecule H_2SO_4 and can easily be looked up on Wikipedia. Emphasis and ellipsis ours. Accessed on 2024-01-23.

of predictive uncertainty. Applications of Bayes' theorem estimate the posterior distribution of neural network (NN) parameters, thus providing a probabilistic understanding and interpretation of the parameters. Furthermore, Bayes' theorem underpins posterior predictive distribution estimation, making it possible to quantify the uncertainty of NN predictions. Interpreting the role of NN parameters and quantifying uncertainty in predictions facilitates risk assessment and improves safety in decision-making.

In the last two decades, the Bayesian deep learning (BDL) framework, which combines Bayesian principles with deep learning, has garnered significant attention. Despite its potential to provide uncertainty estimates and improve model interpretability, generalization, and robustness, mainstream adoption of BDL has been sluggish on both the research and application fronts. A primary concern that is often voiced is the lack of scalability of BDL. However, in an era marked by the widespread and rapid adoption of extensively parameterized deep learning models, this paper posits that BDL has untapped potential and can significantly contribute to the current AI landscape. Recognizing the need to revisit the applicability of BDL, especially in the context of largely parameterized deep learning models, this paper aims to critically analyze the existing challenges that hinder the broader acceptance of BDL. By delving into these challenges and proposing avenues for future research, the paper seeks to unlock the full potential of BDL.

The reason Bayesian concepts are not mainstream in deep learning is not that deep learning makes uncertainty obsolete. In fact, reliable epistemic uncertainty is more relevant than ever in a world of massively overparameterized models. For example, out-of-distribution prompts demonstrate that large language models (LLMs) urgently need reliable uncertainty quantification (UQ); see Figure 1. The problem is that exact Bayesian inference is typically too computationally expensive.

Position. This position paper argues that the advancement of BDL can overcome many of the challenges that deep learning faces nowadays. Notably, BDL methods can prove instrumental in meeting the needs of the 21st century for more mature AI systems and safety-critical decision-making algorithms that can reliably assess uncertainties and incorporate existing knowledge. For example, BDL methods can mitigate risks arising from overly confident yet incorrect predictions made by LLMs (see Figure 1). The major impediment to the development of broadly adoptable BDL methods is scalability, yet this paper proposes research directions that promise to make BDL more amenable to contemporary deep learning.

Bayesian approaches to deep learning provide several advantages over frequentist alternatives. First, BDL reduces the importance of hyper-parameter tuning through incorporating relevant hyper-priors (Lampinen & Vehtari, 2001). Second, in contrast to post-hoc regularization techniques for training on small datasets, BDL enables the use of domain knowledge priors (Sam et al., 2024). Third, BDL approaches to decision-making are more advantageous than frequentist approaches in terms of mitigating the asymmetric costs of errors (Tump et al., 2022). Although there exist non-Bayesian approaches promoting the concept of decision calibration in classification problems, which deal with such asymmetric errors and are suitable for decision-making applications (Zhao et al., 2021), BDL has the added advantage of providing uncertainties over predictions, which can enrich decision-making, for example, by deferring a decision to a later stage when more data is gathered and uncertainty is lower. Fourth, in contrast to conformal prediction, BDL does not require the exchangeability assumption and enables dependence between data across spatiotemporal dimensions through appropriate latent variables (Tran et al., 2020).

Paper structure. Section 2 explains why BDL matters by highlighting the strengths of BDL. Section 3 critically reflects on the challenges that current BDL methods face. Section 4 identifies research directions for the development of scalable BDL methods that can overcome these challenges and become as computationally efficient as established deep learning solutions. The paper concludes with final remarks on the future of BDL (Section 5). Appendix A is a self-contained introductory tutorial on the basics of Bayesian methodology and BDL, providing background knowledge on several Bayesian methods discussed in this paper.

2. Why Bayesian Deep Learning Matters

BDL is a computational framework that combines Bayesian inference principles with deep learning models. Unlike traditional deep learning methods that often provide point estimates, BDL provides a full probability distribution over the parameters, allowing for a principled handling of uncertainty. This intrinsic *uncertainty quantification* is particularly valuable in real-world scenarios where data are limited or noisy. Moreover, BDL accommodates the incorporation of prior information, encapsulated in the choice of a prior distribution. This *integration of prior beliefs* serves as an inductive bias, enabling the model to leverage existing knowledge and providing a principled way to incorporate domain expertise. Based on Bayesian principles, BDL allows *updating beliefs* about uncertain parameters *in light of new evidence*, combining prior knowledge with observed data through Bayes' theorem (Bayes, 1763). Several works aim to improve the understanding of BDL (Wilson & Izmailov, 2020; Izmailov et al., 2021b;a; Kristiadi et al., 2022; Papamarkou et al., 2022; Kapoor et al., 2022; Khan & Rue, 2023; Papamarkou, 2023; Qiu et al., 2023).

BDL has shown substantial potential in a range of critical application domains, such as healthcare (Peng et al., 2019; Abdar et al., 2021; Abdullah et al., 2022; Lopez et al., 2023; Band et al., 2021), single-cell biology (Way & Greene, 2018), drug discovery (Gruver et al., 2021; Stanton et al., 2022; Gruver et al., 2023b; Klarner et al., 2023), agriculture (Hernández & López, 2020), astrophysics (Soboczenski et al., 2018; Ferreira et al., 2020), nanotechnology (Leitherer et al., 2021), physics (Cranmer et al., 2021), climate science (Vandal et al., 2018; Luo et al., 2022), smart electricity grids (Yang et al., 2019), wearables (Manogaran et al., 2019; Zhou et al., 2020), robotics (Shi et al., 2021; Mur-Labadia et al., 2023), and autonomous driving (McAllister et al., 2017). This section outlines the strengths of BDL to motivate the advancement of BDL in the era of large-scale AI.

2.1. Uncertainty Quantification

UQ in BDL improves the reliability of the decision-making process and is valuable when the model encounters ambiguous or out-of-distribution inputs (Tran et al., 2022b). In such instances, the model can signal its lack of confidence in the predictions through the associated probability instead of providing underperforming point estimates. The importance of predictive UQ is especially emphasized in the context of AI-informed decision-making, such as in healthcare (Band et al., 2021; Lopez et al., 2023). In safety-critical domains, *reliable UQ* can be used to deploy models more safely by deferring to a human expert whenever an AI system has high uncertainty about its prediction (Tran et al., 2022b; Rudner et al., 2022a; 2023). This capability is also pertinent to address current challenges in language models, where uncertainty quantification can be used to mitigate risks associated with overly confident but incorrect model predictions (Kadavath et al., 2022); see Figure 1 for an example. Similarly, BDL can be useful for modern challenges, such as hallucinations (Ji et al., 2023) and adversarial attacks (Andriushchenko, 2023) in LLMs, or jailbreaking in text-to-image models (Yang et al., 2023b).

In scientific domains, including but not limited to chemistry and material sciences, where experimental data collection is resource-intensive or constrained, parameter spaces are high-dimensional, and models are inherently complex, BDL excels by providing robust estimates of uncertainty. This attribute is particularly crucial for guiding decisions in inverse design problems, optimizing resource utilization through Bayesian experimental design, optimization, and model selection (Stanton et al., 2022; Gruver et al., 2023b; Li et al., 2023; Rainforth et al., 2024; Bamler et al., 2020; Lotfi et al., 2022; Immer et al., 2021a; 2023).

2.2. Data Efficiency

BDL has manifested data efficiency in various contexts. Notably, BDL methods have been developed for few-shot learning scenarios (Yoon et al., 2018; Patacchiola et al., 2020) and for federated learning under limited data (Zhang et al., 2022b).

Unlike many machine learning approaches that may require large datasets to generalize effectively, BDL leverages prior knowledge and updates beliefs as new data become available. This allows BDL to extract meaningful information from *small datasets*, making it more efficient in scenarios where collecting large amounts of data is challenging or costly (Finzi et al., 2021; Immer et al., 2022b; Shwartz-Ziv et al., 2022; Schwöbel et al., 2022; van der Ouderaa et al., 2023). In addition, the *regularization* effect introduced by the probabilistic nature of its Bayesian approach is beneficial in *preventing overfitting* and contributing to better *generalization* from fewer samples (Rothfuss et al., 2022; Sharma et al., 2023). BDL's uncertainty modeling helps resist the influence of outliers, making it well-suited for real-world scenarios with noisy or out-of-distribution data. This also makes it attractive for foundation model fine-tuning, where data are commonly small and sparse, and uncertainty is important.

Furthermore, the UQ capabilities of BDL allow for an informed selection of data points for labeling. Using prior knowledge and continually updating beliefs as new information arrives, BDL optimizes the iterative process of *active learning*, strategically choosing the most informative instances for labeling to enhance model performance (Gal et al., 2017). This capability may be particularly advantageous for addressing the current challenge of efficiently selecting demonstrations in in-context learning scenarios (Margatina et al., 2023) or fine-tuning with human feedback (Casper et al., 2023).

2.3. Adaptability to New and Evolving Domains

By dynamically updating prior beliefs in response to new evidence, BDL allows selective retention of valuable information from previous tasks while adapting to new ones, thus improving *knowledge transfer* across diverse domains and tasks (Rothfuss et al., 2021; 2022; Rudner et al., 2024a). This is crucial for developing AI systems that can adapt to new situations or temporally evolving domains (Nguyen et al., 2018; Rudner et al., 2022b), as in the case of continual or lifelong learning. The contrast with traditional approaches in large-scale machine learning becomes apparent, as these static models assume that the underlying patterns in the data remain constant over time and struggle with the constant influx of new data and changes in underlying patterns.

2.4. Model Misspecification and Interpretability

Bayesian model averaging (BMA) acknowledges and quantifies uncertainty in the choice of model structure. Instead of relying on a single fixed model, BMA considers a distribution of possible models (Hoeting et al., 1998; 1999; Wasserman, 2000). By incorporating model priors and inferring model posteriors, BDL allows BMA to calibrate uncertainty over network architectures (Hubin & Storvik, 2019; Skaaret-Lund et al., 2023). By averaging predictions over different model possibilities, BMA *attenuates* the impact of *model misspecification*, offering a robust framework that accounts for uncertainty in both parameter values and model structures, ultimately leading to more reliable and interpretable predictions (Hubin et al., 2021; Wang et al., 2023a; Bouchiat et al., 2023).

The interpretation of parameters and structures may seem less crucial in BDL, where overparameterized neural networks serve as functional approximations to unknown data-generating processes. However, research is required to establish reproducible and interpretable Bayesian inferences from deep neural networks (DNNs), especially in applications where black-box prediction is not the primary objective, particularly in scientific contexts (Rügamer, 2023; Wang et al., 2023a; Dold et al., 2024). BMA-centric research in BDL can be valuable in these directions.

3. Current Challenges

One of the challenges in BDL is the computational cost incurred (Izmailov et al., 2021b). Despite the BDL advantages outlined in Section 2, within the realm of Bayesian approaches, Gaussian Processes (GPs) remain the preferred choice in computationally demanding scenarios such as scientific discovery (Tom et al., 2023; Griffiths et al., 2023; Strieth-Kalthoff et al., 2023). Showing that BDL works cheaply, or at least with practical efficiency under modern settings in the real world, is one of the most important problems that remains to be addressed. This section aims to explore the complexities of BDL, highlighting two main challenges that contribute to its difficulties in deployment: posterior computation (Figure 2) and prior specification. It is also explored how scalability arises as a main challenge in BDL. The section concludes with difficulties in the adoption of BDL in foundation models. Challenges related to the lack of convergence and performance metrics and benchmarks for BDL are discussed in Appendix B.

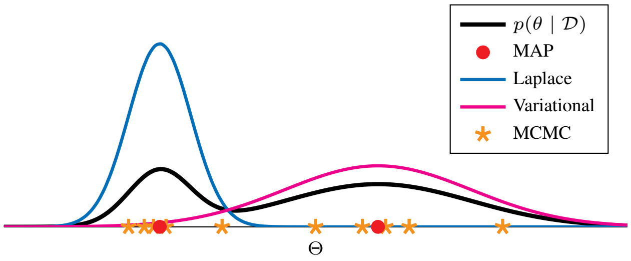


Figure 2: Different BDL methods for approximating a posterior $p(\theta \mid \mathcal{D})$ on a parameter space Θ. While Laplace and Gaussian-based variational approaches yield Gaussian approximations, they generally capture different local modes of the posterior. Ensemble methods use maximum a posteriori estimates as their samples.

3.1. Laplace and Variational Approximations

Laplace and variational approximations use geometric or differential information about the empirical loss to construct closed-form (usually Gaussian) probability measures to approximate the posterior. Despite their simple nature and long history (MacKay, 1992), they often show competitive predictive performance (Daxberger et al., 2021b; Rudner et al., 2022a; Antoran et al., 2023; Rudner et al., 2023). More importantly, their closed-form nature, leveraging automatically computed differential quantities and the foundations of numerical linear algebra, allows theoretical analysis (Kristiadi et al., 2020) and analytical functionality, such as calibration (Kristiadi et al., 2021b;a) and marginalization (Khan et al., 2019; Immer et al., 2021a;b), which are less elegant with stochastic approaches. Laplace-approximated neural networks (Ritter et al., 2018) are particularly tempting because they add no computational cost during training, and require limited computational overhead (comparable to a few epochs) for post-hoc UQ. Moreover, recent variational objectives (Alemi & Poole, 2023) provide alternative means of prediction that avoid internal marginalization.

Alternatively, SWAG (Maddox et al., 2019) is another scalable approximation that creates a Gaussian approximate posterior from stochastic gradient descent (SGD) iterations (Mandt et al., 2017) with a modified learning rate schedule. Similarly to the Laplace approximation, it does not cost much more than standard training. However, SWAG estimates curvature from the trajectory of SGD, rather than the Hessian at a single point. By producing a deterministic probability measure from stochastic gradients, it bridges the gap between deterministic and stochastic procedures.

Despite their analytic strengths, these approximations remain fundamentally local, capturing only a single mode of the multimodal Bayesian neural network (BNN) posterior. Arguably, their most fundamental flaw is that their posterior is dependent on the parametrization of the BNN (MacKay, 1998) and thus inconsistent with some of the most basic properties of probability measures (Kristiadi et al., 2023). Furthermore, the local posterior geometry may be poorly approximated by a Gaussian distribution, which can lead to underconfidence when sampling from the Laplace approximation (Lawrence, 2001), a problem that can be mitigated by linearization (Immer et al., 2021b).

3.2. Ensembles

Deep ensembling involves the retraining of an NN with various initializations, followed by averaging the resulting models. It is effective in approximating the posterior predictive distribution (Wilson & Izmailov, 2020). Recent theoretical advances have established precise connections between ensembles and Bayesian methods (Ciosek et al., 2020; He et al., 2020; Wild et al., 2023).

An open question in BDL is whether one can develop scalable Bayesian inference methods that outperform deep ensembles. Izmailov et al. (2021b) have shown that Hamiltonian Monte Carlo (HMC) often outperforms deep ensembles, but with significant additional computational overhead.

When dealing with large and computationally expensive deep learning models, such as LLMs, the use of deep ensembles may encounter significant challenges due to the associated training and execution costs. Therefore, these large models may motivate research into more efficient architectures and inference paradigms, such as posterior distillation or repulsive ensembles (D'Angelo & Fortuin, 2021), to improve uncertainty calibration and sparser model use.

3.3. Posterior Sampling Algorithms

Within the realm of Markov chain Monte Carlo (MCMC; Brooks et al., 2011) for BDL, stochastic gradient MCMC (SG-MCMC; Nemeth & Fearnhead, 2021) algorithms, such as stochastic gradient Langevin dynamics (SG-LD; Welling & Teh, 2011) and stochastic gradient HMC (SG-HMC; Chen et al., 2014), have emerged as widely adopted tools. Despite offering improved posterior approximations, SG-MCMC algorithms exhibit slower convergence compared to SGD (Robbins, 1951). This deceleration results from the increased iterations required by SG-MCMC to thoroughly explore the posterior distribution beyond locating the mode.

Furthermore, SG-MCMC is still considered expensive for deep learning applications. A step forward in this regard would be to learn from the machine learning and systems community how to make Monte Carlo faster using contemporary hardware (Zhang et al., 2022a; Wang et al., 2023b). Algorithms such as Stein variational gradient descent (SVGD; Liu & Wang, 2016) occupy a middle ground between optimization and sampling, by employing optimization-type updates but with a set of interacting particles. While recent advances show promising results in BNN settings (D'Angelo et al., 2021; D'Angelo & Fortuin, 2021; Pielok et al., 2022), these methods often perform poorly in high-dimensional problems. Alternatively, convergence rates and posterior exploration can be improved with cyclical step-size schedules (Zhang et al., 2020b).

However, despite these advances, the persistent challenges posed by the highly multimodal and high-dimensional nature of BDL posteriors continue to impede the accurate characterization of the full posterior distribution via sampling. There is a need for SG-MCMC algorithms that not only match the speed of SGD, as deployed for optimization in typical deep learning settings, but also deliver high-quality approximations of the posterior to ensure practical utility.

3.4. Prior Specification

The prior over parameters induces a prior over functions, and it is the prior over functions that matters for generalization (Wilson & Izmailov, 2020). Fortunately, the structure in neural network architectures already endows this prior over functions with many desirable properties, such as translation equivariance if a CNN architecture is used. At the same time, defining priors over the parameters is hindered by the complexity and unintelligibility of high-dimensional spaces in BDL. Thus, one aim is to construct informative proper priors on neural network weights that are computationally efficient and favor solutions with desirable model properties (Vladimirova et al., 2019; 2021; Fortuin et al., 2022; Rudner et al., 2023), such as priors that favor models with reliable uncertainty estimates (Rudner et al., 2024a), a high degree of fairness (Rudner et al., 2024b), generalization under covariate shifts (Klarner et al., 2023), equivariance (Finzi et al., 2021), or a high level of sparsity (Ghosh et al., 2018; Polson & Ročková, 2018; Hubin & Storvik, 2019). Weight priors can be cast as neural fields using low-dimensional unit latent variables (Karaletsos et al., 2018; Karaletsos & Bui, 2020) paired with hypernetworks or GPs to express prior knowledge about the field, thus omitting direct parameterizations of beliefs over weights in favor of geometric or other properties of units.

Recent research has developed priors in function space rather than in weight space (Tran et al., 2022a; Rudner et al., 2022b; Qiu et al., 2023). Function-space priors also raise some issues, such as ill-defined variational objectives (Burt et al., 2020; Rudner et al., 2022a) or, in some cases, the need to perform computationally costly GP approximations. There are alternative ways to specify function-space priors beyond GPs. For example, informative function-space priors may be constructed through self-supervising learning (Shwartz-Ziv et al., 2022; Sharma et al., 2023).

3.5. Scalability

The presence of symmetries in the parameter space of NNs yields computational redundancies (Wiese et al., 2023). Addressing the complexity and identifiability issues arising from these symmetries in the context of BDL can significantly impact scalability. Proposed solutions involve the incorporation of symmetry-based constraints in BDL inference methods (Sen et al., 2024) or the design of symmetry-aware priors (Atzeni et al., 2023). However, removing symmetries may not be an optimal strategy, since part of the success of deep learning can be attributed to the overparameterization of NNs, allowing rapid exploration of numerous hypotheses during training or having other positive ‘side effects’ such as induced sparsity (Kolb et al., 2023).

Contrary to the misconception that BNNs inherently suffer from limitations in speed and memory efficiency compared to deterministic NNs, recent advances challenge this notion. For instance, research by Ritter et al. (2021) shows that BNNs can achieve up to four times greater memory efficiency than their deterministic counterparts in terms of the number of parameters. Furthermore, strategies such as recycling the standard training trajectory to construct approximate posteriors, as proposed by Maddox et al. (2019), incur negligible additional computation costs. Hybrid models that combine NNs with GPs, such as deep kernel learning (DKL; Wilson et al., 2016), are also only marginally slower or more memory-consuming than deterministic NNs.

Although UQ is important across various domains, it should not come at the cost of reduced predictive performance. BDL must strike a balance by ensuring that the computational cost of UQ matches that of point estimation. Otherwise, investing computational resources to improve the predictive performance of deep learning models might be a more prudent option. Some may contend that ensembles are less affected by this concern due to their embarrassingly parallel nature. However, in an era where even industry leaders encounter limitations in graphics processing unit (GPU) resources required to train a single large deep learning model, relying solely on parallelism becomes inadequate. Simultaneously achieving time efficiency, memory efficiency, and high model utility (in terms of predictive performance and uncertainty calibration) remains the grand challenge; this is the holy grail of approximate Bayesian inference.

3.6. Foundation Models

Deep learning is in the midst of a paradigm shift into the ‘foundation model’ era, characterized by models with billions, rather than millions, of parameters, with a predominant focus on language rather than vision. BDL approaches to LLMs are relatively unexplored, both in terms of methods and applications. While state-of-the-art approximate inference algorithms can effectively handle models with millions of parameters, only a limited number of works have considered Bayesian approaches to LLMs (Xie et al., 2021; Cohen, 2022; Margatina et al., 2022). In particular, some BDL methods for LLMs have been developed by using Bayesian low-rank adaptation (LoRA; Yang et al., 2024b; Onal et al., 2024), Bayesian optimization (Kristiadi et al., 2024), and Bayesian reward modeling (Yang et al., 2024a).

As discussed in Section 2, BDL emerges as a solution to address limitations in foundation models, particularly in scenarios where data availability is limited. In contexts involving personalized data (Moor et al., 2023) or causal inference applications (Zhang et al., 2023), such as individual treatment effect estimation, where small datasets prevail, the capacity of BDL for uncertainty estimation aligns seamlessly. The fine-tuning settings of foundation models in small-data scenarios is another example. While foundation models are few-shot learners (Brown et al., 2020), BDL offers interpretable uncertainty quantification, which is particularly important in data-limited settings. Moreover, BDL facilitates predictive uncertainty estimation and robust decision-making under uncertainty.

Foundation models represent a valuable frontier for BDL research, particularly around evaluation and applications. What applications of LLMs or transformers are going to benefit from Bayesian inference tools, such as marginalization and priors? More generally, more meaningful applications are needed to convincingly demonstrate that BDL principles go beyond proof-of-concept. The representation of epistemic uncertainty will possibly be most valuable when LLMs or other large-scale NNs are deployed in settings outside of the realm of their training data. For example, Bayesian approaches can be developed and tested in the time series context of applying LLMs in downstream forecasting tasks (Gruver et al., 2023a).

4. Proposed Future Directions

This section, driven by the challenges described in Section 3, presents ongoing research initiatives dedicated to addressing these challenges, particularly focusing on scalability. Subsection 4.7 presents more recent or less widely studied Bayesian research approaches to deep learning. Some topical developments in BDL are discussed in Appendix D.

4.1. Posterior Sampling Algorithms

There is a need for new classes of posterior sampling algorithms that perform better on deep neural networks (DNNs). These algorithms should aim to enhance efficiency, reduce computational overhead, and enable more effective exploration of high-dimensional parameter spaces.

SG-MCMC with tempered posteriors may potentially overcome the issue of sampling from multiple modes. This could be achieved by developing new sampling approaches that can be based on ideas from optimal transport theory (Villani, 2021), score-based diffusion models (Song et al., 2020), and ordinary differential equation (ODE) approaches such as flow matching (Lipman et al., 2022), which use NNs to learn a mapping from a simpler (usually Gaussian) distribution to a complex data distribution (for example, a distribution of images). So, one could plausibly use an NN either to learn a mapping between the BDL posterior and a Gaussian distribution or to use an NN in an MCMC proposal mechanism.

Generally, instead of just focusing on local information about the posterior, there is a need for SG-MCMC algorithms that are able to move rapidly across isolated modes, for instance, using normalizing flows. Since one may not expect to accurately approximate a high-dimensional posterior with respect to all the BNN parameters, novel performance metrics may target lower-dimensional functionals of interest, including UQ as a key piece.

One approach is to incorporate appropriate constraints to attain identifiability, for instance, by making inference on the latent BNN structure (Gu & Dunson, 2023). Instead, one can focus on identifiable functionals for canonical classes of NNs, targeting posterior approximation algorithms for these functionals. Further, one may consider decoupling approaches, which use the BNN as a black box to fit the data-generating model and then choose appropriate loss functions to conduct inference in a second stage.

Another promising approach is running SG-MCMC algorithms in subspaces of the parameter space, for example, linear or sparse subspaces (Izmailov et al., 2020; Li et al., 2024), further enabling the formulation of uncertainty statements for targeted subnetworks (Dold et al., 2024). In the future, SG-MCMC operating on QLoRA (Dettmers et al., 2023) or non-linear subspaces may be constructed. Besides treating subspaces deterministically, posterior dependencies between subspaces can be broken systematically, leading to novel hybrid samplers that combine structured variational inference with MCMC (Alexos et al., 2022) to achieve compute-accuracy trade-offs. Subsampling for BDL can be combined with reasoning about transfer learning (Kirichenko et al., 2023).

4.2. Hybrid Bayesian Approaches

In the future, practical BDL approaches may capture uncertainty over a limited part of the model, while other parts may be estimated efficiently using point estimation. So, one may consider hybrid approaches that combine Bayesian methods with the efficiency of deterministic deep learning.

This could involve developing methods that selectively apply Bayesian approaches in critical areas of the model where capturing uncertainty will be more useful and cheaper, while maintaining a deterministic approach for other parts of the model (Daxberger et al., 2021b). The last-layer Laplace approximation is an example of this (Daxberger et al., 2021a). Such hybrid approaches are a promising area for future research.

Combinations of deep learning methods and GPs have traditionally been limited by the lack of scalability of GPs. However, recent advances in scaling up GP inference are promising for making these hybrid models more widespread. DKL (Wilson et al., 2016) is one example of such a hybrid model. The DKL scalability frontier may be further pushed by exploiting advances in GP scalability.

There exists a prolific literature on connecting BDL and deep Gaussian processes (DGPs; Wilson et al., 2012; Damianou & Lawrence, 2013; Agrawal et al., 2020). This line of work involves neural network GPs (Neal, 1996; de G. Matthews et al., 2018), which are GPs that arise as infinite-width limits of NNs. Theoretical insights into BDL may come from the connection between NNs and GPs.

4.3. Deep Kernel Processes and Machines

Deep kernel processes (DKPs) constitute a family of deep non-parametric approaches to BDL (Aitchison et al., 2021; Ober & Aitchison, 2021a; Ober et al., 2023). A DKP is a DGP, in which one treats the kernels, rather than the features, as random variables. It is possible to derive the prior and perform inference for kernels, without needing DGP features or BNN weights (Aitchison et al., 2021). Thus, DKPs avoid the highly multimodal posteriors caused by permutation symmetries in BDL. It is challenging to accurately approximate these multimodal posteriors with simplified parametric families, for instance, as used in Laplace or variational inference. In contrast, the DKP posterior in practice tends to be unimodal (Yang et al., 2023a). DKPs are a generalization of kernel inverse Wishart processes (Shah et al., 2014), but with non-linear transformations of the kernel, which are useful in representation learning.

Deep kernel machines (DKMs; Milsom et al., 2023; Yang et al., 2023a) go further, by taking the infinite-width limit of a DKP. Usually such an infinite-width limit would eliminate representation learning. However, DKMs carefully temper the likelihood in order to retain representation learning, and are thereby able to attain state-of-the-art predictive performance (Milsom et al., 2023), while their theoretical implications are profound for BDL. DKMs offer key insights into what 'inference in function space' really means and how it relates to representation learning. Specifically, the kernels learned at every layer in a DKM define a 'function space' at every layer. In fact, in a DKM, the true posterior over features is multivariate Gaussian with covariance given by the learned kernel (Aitchison et al., 2021). Representation learning occurs as these function spaces at every layer are modulated by training to focus on the features that matter for predictive performance.

4.4. Semi-Supervised and Self-Supervised Learning

From a Bayesian perspective, one of the surprises in modern deep learning has been the success of semi-supervised learning, where the objective is seemingly arbitrary (or at least, it does not obviously correspond to a likelihood in a known model). Additionally, in Bayesian inference, there are phenomena such as the 'cold posterior effect' (Aitchison, 2021; Wenzel et al., 2020), in which BDL appears to attain more competitive predictive performance by taking the posterior to a power greater than one, thereby shrinking the posterior. In particular, the patterns exploited by semi-supervised learning arise from data curation (Ganev & Aitchison, 2023). If semi-supervised learning is performed on uncurated data, any improvements disappear. This casts doubt on the applicability of semi-supervised learning on real-world uncurated datasets. The cold posterior results can also be explained by underconfident aleatoric uncertainty representation (Kapoor et al., 2022).

Self-supervised learning is an alternative to semi-supervised learning. Self-supervised learning is based on objectives such as mutual information between latent representations of two augmentations of the same underlying image. From a Bayesian perspective, these objectives appear to be ad hoc, as they do not correspond to any likelihood. However, it is possible to formulate a rigorous likelihood in the form of a recognition-parameterized model (Aitchison & Ganev, 2023). This provides insight into the workings of self-supervised learning and how to generalize it to new settings, such as viewing it as a way to learn Bayesian priors (Shwartz-Ziv et al., 2022; Sharma et al., 2023).

4.5. Mixed Precision and Tensor Computations

The success of deep learning is closely tied to its coupling with modern computing and specialized hardware, leveraging technologies like GPUs. Recent investigations within deep learning on the impact of mixed precision point to a role for Bayes, particularly probabilistic numerics (Oates & Sullivan, 2019), in making more efficient use of computation. Mixed precision introduces uncertainty into the internal computations of a model, which Bayes can effectively propagate to downstream predictions. Furthermore, mixed precision requires making decisions about which precision to use, where Bayes can ensure that these decisions are optimal and sensitive to the relations between numerical tasks. Drawing inspiration from specialized hardware, such as tensor processing units, there is potential for a similar trajectory in BDL to address scalability concerns (Mansinghka, 2009). This suggests that the creation of dedicated hardware for BDL has the potential to spark a reevaluation of inference strategies.

In a parallel vein, accelerating software development is crucial to encouraging deep learning practitioners to adopt Bayesian methods. There is a demand for user-friendly software that facilitates the integration of BDL into various projects. The goal is to make BDL usage competitive in terms of human effort compared to standard deep learning practices. For details on BDL software efforts, see Appendix C.

4.6. Compression Strategies

To decrease the computational cost of BDL models, for both memory efficiency and computational speed, compression strategies are being explored. An approach involves using sparsity-inducing priors to prune large parts of BNNs (Louizos et al., 2017). Alternatively, the prior can serve as an entropy model, enabling the compression of BNN weights (Yang et al., 2023c). Methods such as relative entropy coding and variational Bayesian quantization, where the quantization grid is dynamically refined, provide efficient BNN compression (Yang et al., 2020). These novel tools could also be used to dynamically decode a Bayesian ensemble at test time to various levels of precision or ensemble size, resulting in precision-compute trade-offs.

Furthermore, in the context of compressing NN weights, a viable approach involves obtaining the posterior distribution based on observed data and encoding a sample into a bit sequence to send to a receiver (Havasi et al., 2019). The receiver can then extract the posterior sample and use the corresponding weights to make predictions. In practice, approximations are needed to obtain the posterior, encode the sample, and use the corresponding weights to make predictions. Despite the need for approximations in the process, this method yields commendable trade-offs between compression cost and predictive quality compared to alternatives centered on deterministic weight compression.

4.7. Other Future Directions

Bayesian transfer and continual learning. The transfer learning paradigm is quickly becoming a standard way to deploy deep learning models. As noted in Subsection 2.3, BDL is optimized for transfer learning. The focus is not solely on transferring an initialization as in traditional deep learning; instead, knowledge of the source task may inform the shapes and locations of optima on downstream tasks (Shwartz-Ziv et al., 2022; Rudner et al., 2022b; 2023). Self-supervised learning can also be used to create informative self-supervised priors for transfer learning (Shwartz-Ziv et al., 2022; Sharma et al., 2023). Leveraging its efficiency in learning under temporally-changing data distributions through posterior updates, current efforts in the continual learning context explore approaches that integrate new information either assuming a continuous rate of change (Nguyen et al., 2018; Chang et al., 2022) or incorporating priors for changepoint detection (Li et al., 2021).

Probabilistic numerics. Probabilistic numerics (Hennig et al., 2022) is the study of numerical algorithms as Bayesian decision-makers. As numerical algorithms, such as optimization and linear algebra, are clearly central to deep learning, probabilistic numerics offers interesting prospects for making deep learning both more powerful and Bayesian. As one example, since deep training is now regularly I/O-bound for large models, active management of data loading, during training and UQ, is of increasing interest. Methods that quantify and control the information provided by individual computations, based on their effect on the BDL posterior, are showing promise as a formalism for algorithmic data processing in deep training (Tatzel et al., 2023), using probabilistic numerical linear algebra (Wenger et al., 2022) to select sparse informative ‘views’ on the data.

Singular learning theory. Singular learning theory (SLT; Watanabe, 2009) investigates the relation between Bayesian losses, such as approximations of the marginal log-likelihood, and neural network loss functions, using principles from non-equilibrium statistical mechanics. Recent research has drawn connections between Bayesian methods and SLT (Wei & Lau, 2023).

Conformal prediction. For UQ, alternatives such as conformal prediction have emerged as competitors to Bayesian methods and result in well-calibrated uncertainties (Vovk et al., 2005). Deep learning models can be used to develop conformal prediction algorithms (Meister et al., 2023) and, conversely, conformal prediction methods can be used to quantify or calibrate uncertainty in deep learning models. A Bayesian approach to conformal prediction has started to emerge (Hobbhahn et al., 2022; Murphy, 2023), promising a synergistic approach that combines the strengths of Bayesian reasoning with the well-calibrated UQ offered by conformal prediction.

LLMs as distributions. LLMs may be used flexibly as distribution objects in arbitrarily complex programs and workflows. By taking a Bayesian stance, several questions emerge for exploration. When multiple LLMs interact, how does one perform joint inference? What is an effective approach to marginalize over latent variables generated by LLMs, facilitating joint learning over such latent spaces? Is it possible to adopt tools from computational statistics or approximate inference to perform various forms of reasoning with LLMs? And are there innovative ways to synergize small and large LLMs to amortize inferences just in time?

Meta-models. An intriguing prospect arises when contemplating whether BDL will parallel the trajectory of language models. Could one envision the development of a Bayesian meta-model within the BDL framework (Krueger et al., 2017)? This meta-model, akin to language models, may be fine-tuned to multiple tasks, demonstrating competitive predictive performance across them, thus generalizing approaches in amortized inference (Garnelo et al., 2018; Gordon et al., 2019; Müller et al., 2021).

Sequential decision benchmarks. Standard image-based benchmarks focus exclusively on state-of-the-art predictive performance, where non-Bayesian deep learning algorithms typically have an advantage over BDL. To quantify predictive uncertainty, it is encouraged to shift attention to more thorough simulation studies or scientific applications focused on sequential learning and decision-making, such as experimental design, Bayesian optimization, active learning, or bandits. By prioritizing sequential problems in such contexts, researchers and practitioners can gain insights into how well a model generalizes to new and unseen data, how robust it is under uncertain conditions, and how effectively its uncertainty estimates can be utilized by decision makers in real-world scenarios.

5. Final Remarks

This paper has shown that modern deep learning faces a variety of persistent ethical, privacy, and safety issues, particularly when viewed in the context of different types of data, tasks, and performance metrics. However, many of these issues can be overcome within the framework of Bayesian deep learning, building on foundational principles that have survived two and a half centuries of scientific and machine learning evolution. While a number of technical challenges remain, there is a clear path forward that combines creativity and pragmatism to develop BDL approaches that match the data, hardware, and numerical advances of the twenty-first century, especially in the context of large-scale foundation models. In a future where deep learning models seamlessly integrate into decision-making systems, BDL thus emerges as a crucial building block for more mature AI, adding an extra layer of reliability, safety, and trust.

Acknowledgements

This work was supported in part by funding from the Eric and Wendy Schmidt Center at the Broad Institute of MIT and Harvard (MS). JA is supported by ANR-21-JSTM-0001 grant. VF was supported by a Branco Weiss Fellowship. PH gratefully acknowledges financial support by the DFG Cluster of Excellence 'Machine Learning - New Perspectives for Science', EXC 2064/1, project number 390727645; the German Federal Ministry of Education and Research (BMBF) through the Tübingen AI Center (FKZ: 01IS18039A); and funds from the Ministry of Science, Research and Arts of the State of Baden-Württemberg. JMHL acknowledges support from a Turing AI Fellowship under grant EP/V023756/1. SM acknowledges support from the National Science Foundation (NSF) under the NSF CAREER Award 2047418; NSF Grants 2003237 and 2007719, the Department of Energy, Office of Science under grant DE-SC0022331, as well as gifts from Disney and Qualcomm. CN kindly acknowledges the support of EPSRC grants EP/V022636/1 and EP/Y028783/1. AGW is supported by NSF CAREER IIS-2145492, NSF I-DISRE 193471, NSF IIS-1910266, BigHat Biosciences and Capital One.

The authors thank the ICML reviewers for their reviews and feedback.

References

Abdar, M., Samami, M., Mahmoodabad, S. D., Doan, T., Mazoure, B., Hashemifesharaki, R., Liu, L., Khosravi, A., Acharya, U. R., Makarenkov, V., et al. Uncertainty quantification in skin cancer classification using three-way decision-based Bayesian deep learning. *Computers in Biology and Medicine*, 135:104418, 2021.

Abdullah, A. A., Hassan, M. M., and Mustafa, Y. T. A review on Bayesian deep learning in healthcare: Applications and challenges. *IEEE Access*, 10:36538–36562, 2022.

Agrawal, D., Papamarkou, T., and Hinkle, J. Wide neural networks with bottlenecks are deep Gaussian processes. *Journal of Machine Learning Research*, 21(175):1–66, 2020.

Aitchison, L. A statistical theory of cold posteriors in deep neural networks. *International Conference on Learning Representations*, 2021.

Aitchison, L. and Ganev, S. InfoNCE is variational inference in a recognition parameterised model. *arXiv preprint arXiv:2107.02495*, 2023.

Aitchison, L., Yang, A., and Ober, S. W. Deep kernel processes. In *International Conference on Machine Learning*, 2021.

Alemi, A. A. and Poole, B. Variational prediction. In *Fifth Symposium on Advances in Approximate Bayesian Inference*, 2023.

Alexos, A., Boyd, A. J., and Mandt, S. Structured stochastic gradient MCMC. In *International Conference on Machine Learning*, 2022.

Andriushchenko, M. Adversarial attacks on GPT-4 via simple random search. *Preprint*, 2023.

Antoran, J., Bhatt, U., Adel, T., Weller, A., and Hernández-Lobato, J. M. Getting a CLUE: A method for explaining uncertainty estimates. In *International Conference on Learning Representations*, 2021.

Antorán, J., Allingham, J. U., Janz, D., Daxberger, E., Nalisnick, E., and Hernández-Lobato, J. M. Linearised Laplace inference in networks with normalisation layers and the neural g-prior. In *Fourth Symposium on Advances in Approximate Bayesian Inference*, 2022.

Antoran, J., Padhy, S., Barbano, R., Nalisnick, E., Janz, D., and Hernández-Lobato, J. M. Sampling-based inference for large linear models, with application to linearised Laplace. In *International Conference on Learning Representations*, 2023.

Arbel, J., Pitas, K., Vladimirova, M., and Fortuin, V. A primer on Bayesian neural networks: review and debates. *arXiv preprint arXiv:2309.16314*, 2023.

Atzeni, M., Sachan, M., and Loukas, A. Infusing lattice symmetry priors in attention mechanisms for sample-efficient abstract geometric reasoning. *arXiv preprint arXiv:2306.03175*, 2023.

Bamler, R., Salehi, F., and Mandt, S. Augmenting and tuning knowledge graph embeddings. In *Conference on Uncertainty in Artificial Intelligence*, 2020.

Band, N., Rudner, T. G. J., Feng, Q., Filos, A., Nado, Z., Dusenberry, M. W., Jerfel, G., Tran, D., and Gal, Y. Benchmarking Bayesian deep learning on diabetic retinopathy detection tasks. In *Advances in Neural Information Processing Systems*, 2021.

Bayes, T. An essay towards solving a problem in the doctrine of chances. *Philosophical transactions of the Royal Society of London*, 53:370–418, 1763. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFRS.

Bhatt, U., Antorán, J., Zhang, Y., Liao, Q. V., Sattigeri, P., Fogliato, R., Melançon, G., Krishnan, R., Stanley, J., Tickoo, O., Nachman, L., Chunara, R., Srikumar, M., Weller, A., and Xiang, A. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *Conference on AI, Ethics, and Society*, 2021.

Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P., Horsfall, P., and Goodman, N. D. Pyro: Deep universal probabilistic programming. *Journal of Machine Learning Research*, 20(28):1–6, 2019.

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. Weight uncertainty in neural networks. In *International Conference on Machine Learning*, 2015a.

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. Weight uncertainty in neural networks. In *International Conference on Machine Learning*, 2015b.

Bouchiat, K., Immer, A., Yèche, H., Rätsch, G., and Fortuin, V. Laplace-approximated neural additive models: Improving interpretability with Bayesian inference. *arXiv preprint arXiv:2305.16905*, 2023.

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. *Handbook of Markov chain Monte Carlo*. CRC Press, 2011.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020.

Burt, D. R., Ober, S. W., Garriga-Alonso, A., and van der Wilk, M. Understanding variational inference in function-space. In *Third Symposium on Advances in Approximate Bayesian Inference*, 2020.

Carvalho, C. M., Polson, N. G., and Scott, J. G. Handling sparsity via the horseshoe. In *Artificial Intelligence and Statistics*, 2009.

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*, 2023.

Chang, P. G., Murphy, K. P., and Jones, M. On diagonal approximations to the extended Kalman filter for online training of Bayesian neural networks. In *Continual Lifelong Learning Workshop at ACML 2022*, 2022.

Chen, T., Fox, E., and Guestrin, C. Stochastic gradient Hamiltonian Monte Carlo. In *International Conference on Machine Learning*, 2014.

Ciosek, K., Fortuin, V., Tomioka, R., Hofmann, K., and Turner, R. Conservative uncertainty estimation by fitting prior networks. In *International Conference on Learning Representations*, 2020.

Cohen, S. *Bayesian analysis in natural language processing*. Springer Nature, 2022.

Corani, G. and Mignatti, A. Credal model averaging for classification: representing prior ignorance and expert opinions. *International Journal of Approximate Reasoning*, 56:264–277, 2015.

Cranmer, M., Tamayo, D., Rein, H., Battaglia, P., Hadden, S., Armitage, P. J., Ho, S., and Spergel, D. N. A Bayesian neural network predicts the dissolution of compact planetary systems. *Proceedings of the National Academy of Sciences*, 118(40):e2026053118, 2021.

Damianou, A. and Lawrence, N. D. Deep Gaussian processes. In *International Conference on Artificial Intelligence and Statistics*, 2013.

D'Angelo, F. and Fortuin, V. Repulsive deep ensembles are Bayesian. *Advances in Neural Information Processing Systems*, 2021.

D'Angelo, F., Fortuin, V., and Wenzel, F. On Stein variational neural network ensembles. *arXiv preprint arXiv:2106.10760*, 2021.

Daxberger, E., Kristiadi, A., Immer, A., Eschenhagen, R., Bauer, M., and Hennig, P. Laplace redux - effortless Bayesian deep learning. In *Advances in Neural Information Processing Systems*, 2021a.

Daxberger, E., Nalisnick, E., Allingham, J. U., Antoran, J., and Hernández-Lobato, J. M. Bayesian deep learning via subnetwork inference. In *International Conference on Machine Learning*, 2021b.

de G. Matthews, A. G., Rowland, M., Hron, J., Turner, R. E., and Ghahramani, Z. Gaussian process behaviour in wide deep neural networks. *International Conference on Learning Representations*, 2018.

Detommaso, G., Gasparin, A., Donini, M., Seeger, M., Wilson, A. G., and Archambeau, C. Fortuna: A library for uncertainty quantification in deep learning. *arXiv preprint arXiv:2302.04019*, 2023.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, 2023.

Dhahri, R., Immer, A., Charpentier, B., Günnemann, S., and Fortuin, V. Shaving weights with Occam's razor: Bayesian sparsification for neural networks using the marginal likelihood. *arXiv preprint arXiv:2402.15978*, 2024.

Dold, D., Rügamer, D., Sick, B., and Dürr, O. Semi-structured subspace inference. In *International Conference on Artificial Intelligence and Statistics*, 2024.

Ferreira, L., Conselice, C. J., Duncan, K., Cheng, T.-Y., Griffiths, A., and Whitney, A. Galaxy merger rates up to z $\sim$ 3 using a Bayesian deep learning model: A major-merger classifier using illustrisTNG simulation data. *The Astrophysical Journal*, 895(2):115, 2020.

Finzi, M., Benton, G., and Wilson, A. G. Residual pathway priors for soft equivariance constraints. In *Advances in Neural Information Processing Systems*, 2021.

Fortuin, V. Priors in Bayesian deep learning: a review. *International Statistical Review*, 90(3):563–591, 2022.

Fortuin, V., Garriga-Alonso, A., van der Wilk, M., and Aitchison, L. BNNpriors: A library for Bayesian neural network inference with different prior distributions. *Software Impacts*, 9:100079, 2021.

Fortuin, V., Garriga-Alonso, A., Ober, S. W., Wenzel, F., Rätsch, G., Turner, R. E., van der Wilk, M., and Aitchison, L. Bayesian neural network priors revisited. In *International Conference on Learning Representations*, 2022.

Gal, Y. and Ghahramani, Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, 2016.

Gal, Y., Islam, R., and Ghahramani, Z. Deep Bayesian active learning with image data. In *International Conference on Machine Learning*, 2017.

Ganev, S. K. and Aitchison, L. Semi-supervised learning with a principled likelihood from a generative model of data curation. In *International Conference on Learning Representations*, 2023.

Garnelo, M., Rosenbaum, D., Maddison, C., Ramalho, T., Saxton, D., Shanahan, M., Teh, Y. W., Rezende, D., and Eslami, S. A. Conditional neural processes. In *International Conference on Machine Learning*, 2018.

Garriga-Alonso, A. and Fortuin, V. Exact Langevin dynamics with stochastic gradients. In *Third Symposium on Advances in Approximate Bayesian Inference*, 2021.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. *Bayesian data analysis*. Chapman and Hall/CRC, 3rd edition, 2013.

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., and Modrák, M. Bayesian workflow. *arXiv preprint arXiv:2011.01808*, 2020.

George, E. I. Dilution priors: compensating for model space redundancy. In *Borrowing Strength: Theory Powering Applications–A Festschrift for Lawrence D. Brown*, volume 6, pp. 158–166. Institute of Mathematical Statistics, 2010.

Ghosh, S., Yao, J., and Doshi-Velez, F. Structured variational learning of bayesian neural networks with horseshoe priors. In *International Conference on Machine Learning*, 2018.

Goli, L., Reading, C., Sellán, S., Jacobson, A., and Tagliasacchi, A. Bayes' rays: uncertainty quantification in neural radiance fields. *Conference on Computer Vision and Pattern Recognition*, 2024.

Gordon, J., Bruinsma, W. P., Foong, A. Y., Requeima, J., Dubois, Y., and Turner, R. E. Convolutional conditional neural processes. In *International Conference on Learning Representations*, 2019.

Graves, A. Practical variational inference for neural networks. In *Advances in Neural Information Processing Systems*, 2011.

Griffiths, R.-R., Klarner, L., Moss, H. B., Ravuri, A., Truong, S., Stanton, S., Tom, G., Rankovic, B., Du, Y., Jamasb, A., et al. GAUCHE: A library for Gaussian processes in chemistry. In *Advances in Neural Information Processing Systems*, 2023.

Grün, B. and Hofmarcher, P. Identifying groups of determinants in Bayesian model averaging using Dirichlet process clustering. *Scandinavian Journal of Statistics*, 48 (3):1018–1045, 2021.

Grünwald, P. and Van Ommen, T. Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12(4):1069–1103, 2017.

Gruver, N., Stanton, S., Kirichenko, P., Finzi, M., Maffettone, P., Myers, V., Delaney, E., Greenside, P., and Wilson, A. G. Effective surrogate models for protein design with Bayesian optimization. In *ICML Workshop on Computational Biology*, 2021.

Gruver, N., Finzi, M., Qiu, S., and Wilson, A. G. Large language models are zero-shot time series forecasters. *arXiv preprint arXiv:2310.07820*, 2023a.

Gruver, N., Stanton, S., Frey, N., Rudner, T. G., Hotzel, I., Lafrance-Vanasse, J., Rajpal, A., Cho, K., and Wilson, A. G. Protein design with guided discrete diffusion. *Advances in Neural Information Processing Systems*, 2023b.

Gu, Y. and Dunson, D. B. Bayesian pyramids: identifiable multilayer discrete latent structure models for discrete data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(2):399–426, 2023.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International Conference on Machine Learning*, 2017.

Gustafsson, F. K., Danelljan, M., and Schon, T. B. Evaluating scalable Bayesian deep learning methods for robust computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

Havasi, M., Peharz, R., and Hernández-Lobato, J. M. Minimal random code learning: Getting bits back from compressed model parameters. In *International Conference on Learning Representations*, 2019.

He, B., Lakshminarayanan, B., and Teh, Y. W. Bayesian deep ensembles via the neural tangent kernel. In *Advances in Neural Information Processing Systems*, 2020.

Hennig, P., Osborne, M. A., and Kersting, H. P. *Probabilistic Numerics: Computation as Machine Learning*. Cambridge University Press, 2022.

Hernández, S. and López, J. L. Uncertainty quantification for plant disease detection using Bayesian deep learning. *Applied Soft Computing*, 96:106597, 2020.

Hobbhahn, M., Kristiadi, A., and Hennig, P. Fast predictive uncertainty for classification with Bayesian deep networks. In *Conference on Uncertainty in Artificial Intelligence*, 2022.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. Bayesian model averaging. In *AAAI Workshop on Integrating Multiple Learned Models*, 1998.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. Bayesian model averaging: a tutorial. *Statistical Science*, 14(4):382–417, 1999.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. Stochastic variational inference. *Journal of Machine Learning Research*, 14(1):1303–1347, May 2013.

Hubin, A. and Storvik, G. Combining model and parameter uncertainty in Bayesian neural networks. *arXiv preprint arXiv:1903.07594*, 2019.

Hubin, A. and Storvik, G. Sparse Bayesian neural networks: bridging model and parameter uncertainty through scalable variational inference. *Mathematics*, 12(6):788, 2024.

Hubin, A., Storvik, G., and Frommlet, F. Flexible Bayesian nonlinear model configuration. *Journal of Artificial Intelligence Research*, 72:901–942, 2021.

Ibrahim, J. G. and Laud, P. W. On Bayesian analysis of generalized linear models using Jeffreys's prior. *Journal of the American Statistical Association*, 86(416):981–986, 1991.

Immer, A., Bauer, M., Fortuin, V., Rätsch, G., and Khan, M. E. Scalable marginal likelihood estimation for model selection in deep learning. In *International Conference on Machine Learning*, 2021a.

Immer, A., Korzepa, M., and Bauer, M. Improving predictions of Bayesian neural nets via local linearization. In *International Conference on Artificial Intelligence and Statistics*, 2021b.

Immer, A., Hennigen, L. T., Fortuin, V., and Cotterell, R. Probing as quantifying inductive bias. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 1839–1851, 2022a.

Immer, A., van der Ouderaa, T., Rätsch, G., Fortuin, V., and van der Wilk, M. Invariance learning in deep neural networks with differentiable Laplace approximations. *Advances in Neural Information Processing Systems*, 2022b.

Immer, A., Van Der Ouderaa, T. F., Van Der Wilk, M., Ratsch, G., and Schölkopf, B. Stochastic marginal likelihood gradients using neural tangent kernels. In *International Conference on Machine Learning*, 2023.

Izmailov, P., Maddox, W. J., Kirichenko, P., Garipov, T., Vetrov, D., and Wilson, A. G. Subspace inference for Bayesian deep learning. In *Conference on Uncertainty in Artificial Intelligence*, 2020.

Izmailov, P., Nicholson, P., Lotfi, S., and Wilson, A. G. Dangers of Bayesian model averaging under covariate shift. *Advances in Neural Information Processing Systems*, 2021a.

Izmailov, P., Vikram, S., Hoffman, M. D., and Wilson, A. G. G. What are Bayesian neural network posteriors really like? In *International Conference on Machine Learning*, 2021b.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, 2018.

Janz, D., Hron, J., Mazur, P., Hofmann, K., Hernández-Lobato, J. M., and Tschiatschek, S. Successor uncertainties: exploration and uncertainty in temporal difference learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Jaynes, E. T. *Probability theory: The logic of science*. Cambridge University Press, 2003.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 2023.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., Ganguli, D., Hernandez, D., Jacobson, J., Kernion, J., Kravec, S., Lovitt, L., Ndousse, K., Olsson, C., Ringer, S., Amodei, D., Brown, T., Clark, J., Joseph, N., Mann, B., McCandlish, S., Olah, C., and Kaplan, J. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Kapoor, S., Maddox, W. J., Izmailov, P., and Wilson, A. G. On uncertainty, tempering, and data augmentation in Bayesian classification. *Advances in Neural Information Processing Systems*, 2022.

Karaletsos, T. and Bui, T. D. Hierarchical Gaussian process priors for Bayesian neural network weights. In *Advances in Neural Information Processing Systems*, 2020.

Karaletsos, T., Dayan, P., and Ghahramani, Z. Probabilistic meta-representations of neural networks. *arXiv preprint arXiv:1810.00555*, 2018.

Khan, M. E. and Rue, H. The Bayesian learning rule. *Journal of Machine Learning Research*, 24(281):1–46, 2023.

Khan, M. E., Immer, A., Abedi, E., and Korzepa, M. Approximate inference turns deep networks into Gaussian processes. In *Advances in Neural Information Processing Systems*, 2019.

Kirichenko, P., Izmailov, P., and Wilson, A. G. Last layer re-training is sufficient for robustness to spurious correlations. In *International Conference on Learning Representations*, 2023.

Klarner, L., Rudner, T. G. J., Reutlinger, M., Schindler, T., Morris, G. M., Deane, C., and Teh, Y. W. Drug Discovery under Covariate Shift with Domain-Informed Prior Distributions over Functions. In *International Conference on Machine Learning*, 2023.

Kolb, C., Müller, C. L., Bischl, B., and Rügamer, D. Smoothing the edges: A general framework for smooth optimization in sparse regularization using Hadamard overparametrization. *arXiv preprint arXiv:2307.03571*, 2023.

Kou, S., Gan, L., Wang, D., Li, C., and Deng, Z. BayesDiff: estimating pixel-wise uncertainty in diffusion via Bayesian inference. In *International Conference on Learning Representations*, 2024.

Kristiadi, A., Hein, M., and Hennig, P. Being Bayesian, even just a bit, fixes overconfidence in ReLU networks. In *International Conference on Machine Learning*, 2020.

Kristiadi, A., Hein, M., and Hennig, P. An infinite-feature extension for Bayesian ReLU nets that fixes their asymptotic overconfidence. In *Advances in Neural Information Processing Systems*, 2021a.

Kristiadi, A., Hein, M., and Hennig, P. Learnable uncertainty under Laplace approximations. In *Conference on Uncertainty in Artificial Intelligence*, 2021b.

Kristiadi, A., Hein, M., and Hennig, P. Being a bit frequentist improves Bayesian neural networks. In *International Conference on Artificial Intelligence and Statistics*, 2022.

Kristiadi, A., Dangel, F., and Hennig, P. The geometry of neural nets' parameter spaces under reparametrization. In *Advances in Neural Information Processing Systems*, 2023.

Kristiadi, A., Strieth-Kalthoff, F., Skreta, M., Poupart, P., Aspuru-Guzik, A., and Pleiss, G. A sober look at LLMs for material discovery: are they actually good for Bayesian optimization over molecules? *arXiv preprint arXiv:2402.05015*, 2024.

Krueger, D., Huang, C.-W., Islam, R., Turner, R., Lacoste, A., and Courville, A. Bayesian hypernetworks. *arXiv preprint arXiv:1710.04759*, 2017.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 2017.

Lampinen, J. and Vehtari, A. Bayesian approach for neural networks—review and case studies. *Neural networks*, 14 (3):257–274, 2001.

Langford, J. and Shawe-Taylor, J. PAC-Bayes & margins. In *Advances in Neural Information Processing Systems*, 2002.

Lawrence, N. D. *Variational inference in probabilistic models*. PhD thesis, University of Cambridge, 2001.

Lee, J., Bahri, Y., Novak, R., Schoenholz, S. S., Pennington, J., and Sohl-Dickstein, J. Deep neural networks as Gaussian processes. In *International Conference on Learning Representations*, 2017.

Leitherer, A., Ziletti, A., and Ghiringhelli, L. M. Robust recognition and exploratory analysis of crystal structures via Bayesian deep learning. *Nature Communications*, 12 (1):6234, 2021.

Li, A., Boyd, A., Smyth, P., and Mandt, S. Detecting and adapting to irregular distribution shifts in Bayesian online learning. *Advances in Neural Information Processing Systems*, 2021.

Li, J., Miao, Z., Qiu, Q., and Zhang, R. Training Bayesian neural networks with sparse subspace variational inference. In *International Conference on Learning Representations*, 2024.

Li, Y. and Clyde, M. A. Mixtures of g-priors in generalized linear models. *Journal of the American Statistical Association*, 113(524):1828–1845, 2018.

Li, Y. L., Rudner, T. G., and Wilson, A. G. A study of Bayesian neural network surrogates for Bayesian optimization. *arXiv preprint arXiv:2305.20028*, 2023.

Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2022.

Liu, Q. and Wang, D. Stein variational gradient descent: A general purpose Bayesian inference algorithm. *Advances in Neural Information Processing Systems*, 2016.

Lopez, J. L., Rudner, T. G. J., and Shamout, F. Informative priors improve the reliability of multimodal clinical data classification. In *Machine Learning for Health Symposium Findings*, 2023.

Lotfi, S., Izmailov, P., Benton, G., Goldblum, M., and Wilson, A. G. Bayesian model selection, the marginal likelihood, and generalization. In *International Conference on Machine Learning*, pp. 14223–14247, 2022.

Louizos, C., Ullrich, K., and Welling, M. Bayesian compression for deep learning. *Advances in Neural Information Processing Systems*, 2017.

Luo, X., Nadiga, B. T., Park, J. H., Ren, Y., Xu, W., and Yoo, S. A Bayesian deep learning approach to near-term climate prediction. *Journal of Advances in Modeling Earth Systems*, 14(10):e2022MS003058, 2022.

MacKay, D. J. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992.

MacKay, D. J. Bayesian neural networks and density networks. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 354(1):73–80, 1995.

MacKay, D. J. Choice of basis for Laplace approximation. *Machine Learning*, 33:77–86, 1998.

Maddox, W. J., Izmailov, P., Garipov, T., Vetrov, D. P., and Wilson, A. G. A simple baseline for Bayesian uncertainty in deep learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Mandt, S., Hoffman, M. D., and Blei, D. M. Stochastic gradient descent as approximate Bayesian inference. *Journal of Machine Learning Research*, 18(134):1–35, 2017.

Manogaran, G., Shakeel, P. M., Fouad, H., Nam, Y., Baskar, S., Chilamkurti, N., and Sundarasekar, R. Wearable IoT smart-log patch: An edge computing-based Bayesian deep learning network system for multi access physical monitoring system. *Sensors*, 19(13):3030, 2019.

Mansinghka, V. K. *Natively probabilistic computation*. PhD thesis, Massachusetts Institute of Technology, Department of Brain and Cognitive Sciences, 2009.

Margatina, K., Barrault, L., and Aletras, N. On the importance of effectively adapting pretrained language models for active learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022.

Margatina, K., Schick, T., Aletras, N., and Dwivedi-Yu, J. Active learning principles for in-context learning with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023.

Martens, J. Deep learning via Hessian-free optimization. In *International Conference on Machine Learning*, 2010.

McAllister, R., Gal, Y., Kendall, A., Van Der Wilk, M., Shah, A., Cipolla, R., and Weller, A. Concrete problems for autonomous vehicle safety: Advantages of Bayesian deep learning. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 2017.

Meister, J. A., Nguyen, K. A., Kapetanakis, S., and Luo, Z. A novel deep learning approach for one-step conformal prediction approximation. *Annals of Mathematics and Artificial Intelligence*, pp. 1–28, 2023.

Milsom, E., Anson, B., and Aitchison, L. Convolutional deep kernel machines. *arXiv preprint arXiv:2309.09814*, 2023.

Molchanov, D., Ashukha, A., and Vetrov, D. Variational dropout sparsifies deep neural networks. In *International Conference on Machine Learning*, 2017.

Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., and Rajpurkar, P. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

Müller, S., Hollmann, N., Arango, S. P., Grabocka, J., and Hutter, F. Transformers can do Bayesian inference. In *International Conference on Learning Representations*, 2021.

Mur-Labadia, L., Martinez-Cantin, R., and Guerrero, J. J. Bayesian deep learning for affordance segmentation in images. *arXiv preprint arXiv:2303.00871*, 2023.

Murphy, K. P. *Probabilistic machine learning: Advanced topics*. MIT Press, 2023.

Nabarro, S., Ganev, S., Garriga-Alonso, A., Fortuin, V., van der Wilk, M., and Aitchison, L. Data augmentation in Bayesian neural networks and the cold posterior effect. In *Uncertainty in Artificial Intelligence*, 2022.

Neal, R. M. Priors for infinite networks. *Bayesian learning for neural networks*, pp. 29–53, 1996.

Neklyudov, K., Molchanov, D., Ashukha, A., and Vetrov, D. P. Structured Bayesian pruning via log-normal multiplicative noise. In *Advances in Neural Information Processing Systems*, 2017.

Neklyudov, K., Molchanov, D., Ashukha, A., and Vetrov, D. Variance networks: when expectation does not meet your expectations. In *International Conference on Learning Representations*, 2018.

Nemeth, C. and Fearnhead, P. Stochastic gradient Markov chain Monte Carlo. *Journal of the American Statistical Association*, 116(533):433–450, 2021.

Nguyen, C. V., Li, Y., Bui, T. D., and Turner, R. E. Variational continual learning. In *International Conference on Learning Representations*, 2018.

Oates, C. J. and Sullivan, T. J. A modern retrospective on probabilistic numerics. *Statistics and Computing*, 29(6): 1335–1351, 2019.

Ober, S. and Aitchison, L. A variational approximate posterior for the deep Wishart process. *Advances in Neural Information Processing Systems*, 2021a.

Ober, S. W. and Aitchison, L. Global inducing point variational posteriors for Bayesian neural networks and deep Gaussian processes. In *International Conference on Machine Learning*, 2021b.

Ober, S. W., Anson, B., Milsom, E., and Aitchison, L. An improved variational approximate posterior for the deep Wishart process. In *Conference on Uncertainty in Artificial Intelligence*, 2023.

Onal, E., Flöge, K., Caldwell, E., Sheverdin, A., and Fortuin, V. Gaussian stochastic weight averaging for Bayesian low-rank adaptation of large language models. *arXiv preprint arXiv:2405.03425*, 2024.

Osband, I., Wen, Z., Asghari, S. M., Dwaracherla, V., Lu, X., Ibrahimi, M., Lawson, D., Hao, B., O'Donoghue, B., and Van Roy, B. The neural testbed: Evaluating joint predictions. *Advances in Neural Information Processing Systems*, 2022.

Osband, I., Wen, Z., Asghari, S. M., Dwaracherla, V., Ibrahimi, M., Lu, X., and Van Roy, B. Approximate Thompson sampling via epistemic neural networks. In *Conference on Uncertainty in Artificial Intelligence*, 2023.

Papadopoulos, H., Vovk, V., and Gammerman, A. Conformal prediction with neural networks. In *International Conference on Tools with Artificial Intelligence*, 2007.

Papamarkou, T. Approximate blocked Gibbs sampling for Bayesian neural networks. *Statistics and Computing*, 33, 2023.

Papamarkou, T., Hinkle, J., Young, M. T., and Womble, D. Challenges in Markov chain Monte Carlo for Bayesian neural networks. *Statistical Science*, 37(3):425–442, 2022.

Parrado-Hernández, E., Ambroladze, A., Shawe-Taylor, J., and Sun, S. PAC-Bayes bounds with data dependent priors. *Journal of Machine Learning Research*, 13(1): 3507–3531, 2012.

Patacchiola, M., Turner, J., Crowley, E. J., O' Boyle, M., and Storkey, A. J. Bayesian meta-learning for the few-shot setting via deep kernels. In *Advances in Neural Information Processing Systems*, 2020.

Pearce, T., Leibfried, F., and Brintrup, A. Uncertainty in neural networks: Approximately Bayesian ensembling. In *International Conference on Artificial Intelligence and Statistics*, pp. 234–244. PMLR, 2020.

Peng, W., Ye, Z.-S., and Chen, N. Bayesian deep-learning-based health prognostics toward prognostics uncertainty. *IEEE Transactions on Industrial Electronics*, 67(3):2283–2293, 2019.

Pielok, T., Bischl, B., and Rügamer, D. Approximate Bayesian inference with Stein functional variational gradient descent. In *International Conference on Learning Representations*, 2022.

Polson, N. G. and Ročková, V. Posterior concentration for sparse deep learning. *Advances in Neural Information Processing Systems*, 2018.

Qiu, S., Rudner, T. G. J., Kapoor, S., and Wilson, A. G. Should we learn most likely functions or parameters? In *Advances in Neural Information Processing Systems*, 2023.

Rainforth, T., Foster, A., Ivanova, D. R., and Bickford Smith, F. Modern bayesian experimental design. *Statistical Science*, 39(1):100–114, 2024.

Ritter, H. and Karaletsos, T. TyXe: Pyro-based Bayesian neural nets for Pytorch. In *Proceedings of Machine Learning and Systems*, 2022.

Ritter, H., Botev, A., and Barber, D. A scalable Laplace approximation for neural networks. In *International Conference on Learning Representations*, 2018.

Ritter, H., Kukla, M., Zhang, C., and Li, Y. Sparse uncertainty representation in deep learning with inducing weights. In *Advances in Neural Information Processing Systems*, 2021.

Robbins, H. E. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

Rothfuss, J., Fortuin, V., Josifoski, M., and Krause, A. PACOH: Bayes-optimal meta-learning with PAC-guarantees. In *International Conference on Machine Learning*, 2021.

Rothfuss, J., Josifoski, M., Fortuin, V., and Krause, A. PAC-Bayesian meta-learning: From theory to practice. *arXiv preprint arXiv:2211.07206*, 2022.

Rudner, T. G. J., Chen, Z., Teh, Y. W., and Gal, Y. Tractable function-space variational inference in Bayesian neural networks. In *Advances in Neural Information Processing Systems*, 2022a.

Rudner, T. G. J., Smith, F. B., Feng, Q., Teh, Y. W., and Gal, Y. Continual Learning via Sequential Function-Space Variational Inference. In *International Conference on Machine Learning*, 2022b.

Rudner, T. G. J., Kapoor, S., Qiu, S., and Wilson, A. G. Function-Space Regularization in Neural Networks: A Probabilistic Perspective. In *International Conference on Machine Learning*, 2023.

Rudner, T. G. J., Pan, X., Li, Y. L., Shwartz-Ziv, R., and Wilson, A. G. Uncertainty-aware priors for finetuning pretrained models. In *Preprint*, 2024a.

Rudner, T. G. J., Zhang, Y. S., Wilson, A. G., and Kempe, J. Mind the GAP: Improving robustness to subpopulation shifts with group-aware priors. In *International Conference on Artificial Intelligence and Statistics*, 2024b.

Rügamer, D. A new PHO-rmula for improved performance of semi-structured networks. In *International Conference on Machine Learning*, 2023.

Russo, D. J., Van Roy, B., Kazerouni, A., Osband, I., and Wen, Z. A tutorial on Thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018.

Sam, D., Pukdee, R., Jeong, D. P., Byun, Y., and Kolter, J. Z. Bayesian neural networks with domain knowledge priors. *arXiv preprint arXiv:2402.13410*, 2024.

Schraudolph, N. N., Yu, J., and Günter, S. A stochastic quasi-Newton method for online convex optimization. In *International Conference on Artificial Intelligence and Statistics*, 2007.

Schwöbel, P., Jørgensen, M., Ober, S. W., and Van Der Wilk, M. Last layer marginal likelihood for invariance learning. In *International Conference on Artificial Intelligence and Statistics*, 2022.

Sen, D., Papamarkou, T., and Dunson, D. Bayesian neural networks and dimensionality reduction. In *Handbook of Bayesian, fiducial, and frequentist inference*. Chapmann and Hall/CRC Press, 2024.

Shah, A., Wilson, A., and Ghahramani, Z. Student-t processes as alternatives to Gaussian processes. In *International Conference on Artificial Intelligence and Statistics*, 2014.

Sharma, M., Rainforth, T., Teh, Y. W., and Fortuin, V. Incorporating unlabelled data into Bayesian neural networks. *arXiv preprint arXiv:2304.01762*, 2023.

Shen, Y., Daheim, N., Cong, B., Nickl, P., Marconi, G. M., Bazan, C., Yokota, R., Gurevych, I., Cremers, D., and Khan, M. E. Variational learning is effective for large deep networks. In *International Conference on Machine Learning*, 2024.

Shi, L., Copot, C., and Vanlanduit, S. A Bayesian deep neural network for safe visual servoing in human–robot interaction. *Frontiers in Robotics and AI*, 8:687031, 2021.

Shwartz-Ziv, R., Goldblum, M., Souri, H., Kapoor, S., Zhu, C., LeCun, Y., and Wilson, A. G. Pre-train your loss: Easy Bayesian transfer learning with informative priors. In *Advances in Neural Information Processing Systems*, 2022.

Skaaret-Lund, L., Storvik, G., and Hubin, A. Sparsifying Bayesian neural networks with latent binary variables and normalizing flows. *arXiv preprint arXiv:2305.03395*, 2023.

Soboczenski, F., Himes, M. D., O'Beirne, M. D., Zorzan, S., Baydin, A. G., Cobb, A. D., Gal, Y., Angerhausen, D., Mascaro, M., Arney, G. N., et al. Bayesian deep learning for exoplanet atmospheric retrieval. *arXiv preprint arXiv:1811.03390*, 2018.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020.

Stanton, S., Maddox, W., Gruver, N., Maffettone, P., Delaney, E., Greenside, P., and Wilson, A. G. Accelerating

Bayesian optimization for biological sequence design with denoising autoencoders. In *International Conference on Machine Learning*, 2022.

Strieth-Kalthoff, F., Hao, H., Rathore, V., Derasp, J., Gaudin, T., Angello, N. H., Seifrid, M., Trushina, E., Guy, M., Liu, J., and et al. Delocalized, asynchronous, closed-loop discovery of organic laser emitters. *ChemRxiv*, 2023.

Sun, S., Zhang, G., Shi, J., and Grosse, R. Functional variational Bayesian neural networks. In *International Conference on Learning Representations*, 2019.

Sunde, L.-M. S. Spherical priors for Bayesian deep learning. Master's thesis, University of Oslo, 2023.

Tatzel, L., Wenger, J., Schneider, F., and Hennig, P. Accelerating generalized linear models by trading off computation for uncertainty. *arXiv preprint arXiv:2310.20285*, 2023.

Tom, G., Hickman, R. J., Zinzuwadia, A., Mohajeri, A., Sanchez-Lengeling, B., and Aspuru-Guzik, A. Calibration and generalizability of probabilistic models on low-data chemical datasets with DIONYSUS. *Digital Discovery*, 2023.

Tran, B.-H., Rossi, S., Milios, D., and Filippone, M. All you need is a good functional prior for Bayesian deep learning. *Journal of Machine Learning Research*, 23(1), 2022a.

Tran, D., Liu, J., Dusenberry, M. W., Phan, D., Collier, M., Ren, J., Han, K., Wang, Z., Mariet, Z., Hu, H., Band, N., Rudner, T. G. J., Singhal, K., Nado, Z., van Amersfoort, J., Kirsch, A., Jenatton, R., Thain, N., Yuan, H., Buchanan, K., Murphy, K., Sculley, D., Gal, Y., Ghahramani, Z., Snoek, J., and Lakshminarayanan, B. Plex: Towards Reliability Using Pretrained Large Model Extensions. In *ICML 2022 Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward*, 2022b.

Tran, M.-N., Nguyen, N., Nott, D., and Kohn, R. Bayesian deep net GLM and GLMM. *Journal of Computational and Graphical Statistics*, 29(1):97–113, 2020.

Tump, A. N., Wolf, M., Romanczuk, P., and Kurvers, R. H. J. M. Avoiding costly mistakes in groups: the evolution of error management in collective decision making. *PLOS Computational Biology*, 18(8):1–21, 2022.

van der Ouderaa, T. F., Immer, A., and van der Wilk, M. Learning layer-wise equivariances automatically using gradients. In *Advances in Neural Information Processing Systems*, 2023.

Vandal, T., Kodra, E., Dy, J., Ganguly, S., Nemani, R., and Ganguly, A. R. Quantifying uncertainty in discrete-continuous and skewed data with Bayesian deep learning. In *International Conference on Knowledge Discovery & Data Mining*, 2018.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.-C. Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC. *Bayesian Analysis*, 16(2):667–718, 2021.

Villani, C. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2021.

Vladimirova, M., Verbeek, J., Mesejo, P., and Arbel, J. Understanding priors in Bayesian neural networks at the unit level. In *International Conference on Machine Learning*, 2019.

Vladimirova, M., Arbel, J., and Girard, S. Bayesian neural network unit priors and generalized Weibull-tail property. In *Asian Conference on Machine Learning*, 2021.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Wainwright, M. J. and Jordan, M. I. *Graphical Models, Exponential Families, and Variational Inference*. Now Publishers Inc., Hanover, MA, USA, 2008.

Wang, Y., Rudner, T. G. J., and Wilson, A. G. Visual explanations of image-text representations via multi-modal information bottleneck attribution. In *Advances in Neural Information Processing Systems*, 2023a.

Wang, Z., Chen, Y., Song, Q., and Zhang, R. Enhancing low-precision sampling via stochastic gradient Hamiltonian Monte Carlo. *arXiv preprint arXiv:2310.16320*, 2023b.

Wasserman, L. Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, 44(1):92–107, 2000.

Watanabe, S. *Algebraic Geometry and Statistical Learning Theory*. Cambridge University Press, USA, 2009.

Way, G. P. and Greene, C. S. Bayesian deep learning for single-cell analysis. *Nature Methods*, 15(12):1009–1010, 2018.

Wei, S. and Lau, E. Variational Bayesian neural networks via resolution of singularities. *arXiv preprint arXiv:2302.06035*, 2023.

Welling, M. and Teh, Y. W. Bayesian learning via stochastic gradient Langevin dynamics. In *International Conference on Machine Learning*, 2011.

Wen, Z., Osband, I., Qin, C., Lu, X., Ibrahimi, M., Dwaracherla, V., Asghari, M., and Van Roy, B. From predictions to decisions: The importance of joint predictive distributions. *arXiv preprint arXiv:2107.09224*, 2021.

Wenger, J., Pleiss, G., Pförtner, M., Hennig, P., and Cunningham, J. P. Posterior and computational uncertainty in Gaussian processes. In *Advances in Neural Information Processing Systems*, 2022.

Wenzel, F., Roth, K., Veeling, B., Swiatkowski, J., Tran, L., Mandt, S., Snoek, J., Salimans, T., Jenatton, R., and Nowozin, S. How good is the Bayes posterior in deep neural networks really? In *International Conference on Machine Learning*, 2020.

Wiese, J. G., Wimmer, L., Papamarkou, T., Bischl, B., Günnemann, S., and Rügamer, D. Towards efficient MCMC sampling in Bayesian neural networks by exploiting symmetry. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2023.

Wild, V. D., Ghalebikesabi, S., Sejdinovic, D., and Knoblauch, J. A rigorous link between deep ensembles and (variational) Bayesian methods. In *Conference on Neural Information Processing Systems*, 2023.

Williams, P. M. Bayesian regularization and pruning using a Laplace prior. *Neural Computation*, 7(1):117–143, 1995.

Wilson, A. G. and Izmailov, P. Bayesian deep learning and a probabilistic perspective of generalization. *Advances in Neural Information Processing Systems*, 2020.

Wilson, A. G., Knowles, D. A., and Ghahramani, Z. Gaussian process regression networks. In *International Conference on Machine Learning*, 2012.

Wilson, A. G., Hu, Z., Salakhutdinov, R., and Xing, E. P. Deep kernel learning. In *International Conference on Artificial Intelligence and Statistics*, 2016.

Wilson, A. G., Izmailov, P., Hoffman, M. D., Gal, Y., Li, Y., Pradier, M. F., Vikram, S., Foong, A., Lotfi, S., and Farquhar, S. Evaluating approximate inference in bayesian deep learning. In *NeurIPS 2021 Competitions and Demonstrations Track*, 2022.

Xie, S. M., Raghunathan, A., Liang, P., and Ma, T. An explanation of in-context learning as implicit Bayesian inference. In *International Conference on Learning Representations*, 2021.

Yang, A. X., Robeyns, M., Milsom, E., Anson, B., Schoots, N., and Aitchison, L. A theory of representation learning gives a deep generalisation of kernel methods. In *International Conference on Machine Learning*, 2023a.

Yang, A. X., Robeyns, M., Coste, T., Wang, J., Bou-Ammar, H., and Aitchison, L. Bayesian reward models for LLM alignment. *arXiv preprint arXiv:2402.13210*, 2024a.

Yang, A. X., Robeyns, M., Wang, X., and Aitchison, L. Bayesian low-rank adaptation for large language models. *International Conference on Learning Representations*, 2024b.

Yang, Y., Li, W., Gulliver, T. A., and Li, S. Bayesian deep learning-based probabilistic load forecasting in smart grids. *IEEE Transactions on Industrial Informatics*, 16 (7):4703–4713, 2019.

Yang, Y., Bamler, R., and Mandt, S. Variational Bayesian quantization. In *International Conference on Machine Learning*, 2020.

Yang, Y., Hui, B., Yuan, H., Gong, N., and Cao, Y. SneakyPrompt: Evaluating robustness of text-to-image generative models' safety filters. In *IEEE Symposium on Security and Privacy*, 2023b.

Yang, Y., Mandt, S., and Theis, L. An introduction to neural data compression. *Foundations and Trends® in Computer Graphics and Vision*, 15(2):113–200, 2023c.

Yoon, J., Kim, T., Dia, O., Kim, S., Bengio, Y., and Ahn, S. Bayesian model-agnostic meta-learning. In *Advances in Neural Information Processing Systems*, 2018.

Zellner, A. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian Inference and Decision Techniques*, 1986.

Zhang, J., Jennings, J., Zhang, C., and Ma, C. Towards causal foundation model: on duality between causal inference and attention. *arXiv preprint arXiv:2310.00809*, 2023.

Zhang, R., Cooper, A. F., and De Sa, C. AMAGOLD: amortized Metropolis adjustment for efficient stochastic gradient MCMC. In *International Conference on Artificial Intelligence and Statistics*, 2020a.

Zhang, R., Li, C., Zhang, J., Chen, C., and Wilson, A. G. Cyclical stochastic gradient MCMC for Bayesian deep learning. In *International Conference on Learning Representations*, 2020b.

Zhang, R., Wilson, A. G., and De Sa, C. Low-precision stochastic gradient Langevin dynamics. In *International Conference on Machine Learning*, 2022a.

Zhang, X., Li, Y., Li, W., Guo, K., and Shao, Y. Personalized federated learning via variational Bayesian inference. In *International Conference on Machine Learning*, 2022b.

Zhao, S., Kim, M. P., Sahoo, R., Ma, T., and Ermon, S. Calibrating predictions to decisions: a novel approach to multi-class calibration. In *Advances in Neural Information Processing Systems*, 2021.

Zhou, Z., Yu, H., and Shi, H. Human activity recognition based on improved Bayesian convolution network to analyze health care data using wearable iot device. *IEEE Access*, 8:86411–86418, 2020.

A. Background

This appendix provides background knowledge on several Bayesian methods that underpin Bayesian deep learning (BDL). It can be used as a self-contained introductory tutorial on the basics of Bayesian methodology and BDL. For a more detailed coverage, the reader is referred to the references provided herein.

A.1. Laplace Approximations

Laplace approximations constitute a method for constructing a Gaussian process (GP) posterior on the output of a neural network, leveraging automatic differentiation and numerical linear algebra. Consider a neural network $\mathbf{f}(\mathbf{x}, \boldsymbol{\theta})$ that maps input $\mathbf{x}$ and parameters $\boldsymbol{\theta} \in \mathbb{R}^{\nu}$ (representing, for example, network weights and biases) to an output $\mathbf{y}$. The neural network is trained to find the parameters $\tilde{\boldsymbol{\theta}}$ that minimize a regularized empirical risk function $\mathcal{L}(\boldsymbol{\theta})$ on supervised training data $\mathcal{D} = (\mathbf{x}_i, \mathbf{y}_i)_{i=1,\ldots,n}$.

$$\tilde{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^{\nu}} \mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ell(\mathbf{y}_i, \mathbf{f}(\mathbf{x}_i, \boldsymbol{\theta})) + r(\boldsymbol{\theta}),$$

where ℓ and r are a training loss and regularizer, respectively. The parameter value $\tilde{\boldsymbol{\theta}}$ is found using the same approach as in non-Bayesian deep learning, employing stochastic optimization. It is possible to interpret the value $\tilde{\boldsymbol{\theta}}$ obtained by training the neural network. In particular, minimizing $\mathcal{L}$ is equivalent to maximizing the exponential of negative $\mathcal{L}$, since the exponential function is strictly increasing:

$$\begin{aligned}
\tilde{\boldsymbol{\theta}} &= \operatorname*{argmax}_{\boldsymbol{\theta} \in \mathbb{R}^{\nu}} \exp(-\mathcal{L}(\boldsymbol{\theta})) \\
&= \operatorname*{argmax}_{\boldsymbol{\theta} \in \mathbb{R}^{\nu}} \left(\prod_{i=1}^{n} \exp(-\ell(\mathbf{y}_i, \mathbf{f}(\mathbf{x}_i, \boldsymbol{\theta}))) \exp(-r(\boldsymbol{\theta})) \right) \\
&= \operatorname*{argmax}_{\boldsymbol{\theta} \in \mathbb{R}^{\nu}} \prod_{i=1}^{n} p(\mathbf{y}_i \mid \mathbf{f}(\mathbf{x}_i, \boldsymbol{\theta})) p(\boldsymbol{\theta}) \\
&= \operatorname*{argmax}_{\boldsymbol{\theta} \in \mathbb{R}^{\nu}} p(\boldsymbol{\theta} \mid \mathcal{D}),
\end{aligned}$$

where ℓ is re-interpreted as a negative log-likelihood, and r as a negative log-prior. This interpretation is valid for commonly used choices of these quantities in deep learning. The log-likelihood ℓ is commonly the logarithm of a distribution from the exponential family. Typical choices of r are variants of the l_2 loss, such as the logarithm of a Gaussian prior on the parameters.

Under this interpretation, automatic differentiation can be used to compute a second-order Taylor approximation of $\mathcal{L}$ around $\tilde{\boldsymbol{\theta}}$, and thus a Gaussian approximation for $p(\boldsymbol{\theta} \mid \mathcal{D})$ can be acquired:

$$\log p(\boldsymbol{\theta} \mid \mathcal{D}) \approx \mathcal{L}(\tilde{\boldsymbol{\theta}}) + \frac{1}{2}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^T \boldsymbol{\Psi} (\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}) = \log \mathcal{N}(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}, -\boldsymbol{\Psi}^{-1}). \tag{1}$$

$\boldsymbol{\Psi}$ is nominally the Hessian of $\mathcal{L}$. Due to its quadratic dependence on ν, approximations are typically used. Of particular interest is the generalized Gauss-Newton (GGN) matrix $\mathbf{G}$ (Schraudolph et al., 2007; Martens, 2010),

$$\boldsymbol{\Psi} \approx \mathbf{G} = \sum_{i=1}^{n} \mathbf{J}_{\tilde{\boldsymbol{\theta}}}(\mathbf{x}_i) \left(\nabla_{\mathbf{f}} \nabla_{\mathbf{f}}^T \ell(\mathbf{y}_i, \mathbf{f}(\mathbf{x}_i, \tilde{\boldsymbol{\theta}})) \right) \mathbf{J}_{\tilde{\boldsymbol{\theta}}}^T(\mathbf{x}_i) + \nabla \nabla^T r(\tilde{\boldsymbol{\theta}}), \tag{2}$$

which can be evaluated using the closed-form Hessian of the loss with respect to the logit inputs, and the Jacobian

$$[\mathbf{J}_{\tilde{\boldsymbol{\theta}}}(\mathbf{x}_i)]_{a,b} = \left. \frac{\partial f_b(\mathbf{x}_i, \boldsymbol{\theta})}{\partial \theta_a} \right|_{\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}}$$

of the neural network $\mathbf{f}$. This matrix has a low rank that allows efficient manipulation, such as computing the inverse required in Equation (2). To propagate this approximate belief on $\boldsymbol{\theta}$ to the output of $\mathbf{f}$, it is common to linearize the network with respect to $\boldsymbol{\theta}$ around $\tilde{\boldsymbol{\theta}}$:

$$\mathbf{f}(\mathbf{x}, \boldsymbol{\theta}) \approx \mathbf{f}(\mathbf{x}, \tilde{\boldsymbol{\theta}}) + (\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^T \mathbf{J}_{\tilde{\boldsymbol{\theta}}}(\mathbf{x}).$$

Note that this approximation is made with respect to $\boldsymbol{\theta}$; the neural network remains a non-linear function of its input $\mathbf{x}$.

Under this linearization, the posterior on $\mathbf{f}(\mathbf{x})$ associated with the Gaussian posterior on $\boldsymbol{\theta}$ of Equation (1) is a GP:

$$p(\mathbf{f}(\mathbf{x}) \mid \mathcal{D}) \approx \mathcal{GP}\left(\mathbf{f}(\cdot), \mathbf{f}(\mathbf{x}, \tilde{\boldsymbol{\theta}}), -\mathbf{J}_{\tilde{\boldsymbol{\theta}}}(\cdot)\boldsymbol{\Psi}^{-1}\mathbf{J}_{\tilde{\boldsymbol{\theta}}}(\cdot)\right).$$

The mean function of the GP corresponds to the trained neural network $\mathbf{f}(\cdot, \tilde{\boldsymbol{\theta}})$ used in non-Bayesian deep learning. The GP kernel is the posterior version of the neural tangent kernel (Jacot et al., 2018). This concrete practical connection enables Laplace approximations to be used as a drop-in method in deep learning; the neural network is trained or a pre-trained one is used. Subsequently, the GGN matrix and Jacobian are computed. The trained neural network is then kept as a point estimate, now serving as the posterior mean of the GP, augmented with structured GP uncertainty. The computational overhead at training time is limited to the numerical linear algebra of cost that is linear in the training set size n and in the parameter space dimension ν. At test time, inference for a given input $\mathbf{x}'$ requires one backward pass to compute the Jacobian $\mathbf{J}_{\tilde{\boldsymbol{\theta}}}(\mathbf{x}')$, resulting in a constant overhead compared to the forward pass needed to compute $\mathbf{f}(\mathbf{x}', \tilde{\boldsymbol{\theta}})$.

An advantage of the Laplace approximation is that it enables to compute the marginal likelihood of the approximate posterior in closed form (Immer et al., 2021a), which can be used for Bayesian model selection in neural networks, for instance, for invariance learning (Immer et al., 2022b), linguistic probing of language models (Immer et al., 2022a), or neural network pruning (Dhahri et al., 2024). Thus, the Laplace approximation makes BDL more computationally feasible.

A.2. Variational Inference

Variational inference is an approach to approximate inference that seeks to avoid the intractability of exact inference by framing posterior inference as a variational optimization problem. Consider some stochastic parameters $\boldsymbol{\Theta}$, data $\mathcal{D}$, a likelihood function $p(\mathcal{D} \mid \boldsymbol{\theta})$, a prior $p(\boldsymbol{\theta})$, and the posterior $p(\boldsymbol{\theta} \mid \mathcal{D})$ given by

$$p(\boldsymbol{\theta} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\mathcal{D})}.$$

Variational inference approximates $p(\boldsymbol{\theta} \mid \mathcal{D})$ by solving the variational problem

$$\min_{q_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) \in \mathcal{Q}_{\boldsymbol{\Theta}}} \mathbb{D}_{\mathrm{KL}}(q_{\boldsymbol{\Theta}} \mid\mid p_{\boldsymbol{\theta}|\mathcal{D}}) \tag{3}$$

with respect to a variational distribution $q(\boldsymbol{\theta})$ within some variational family of distributions $\mathcal{Q}_{\boldsymbol{\Theta}}$ (Wainwright & Jordan, 2008). In expression (3), $\mathbb{D}_{\mathrm{KL}}$ denotes the Kullback-Leibler (KL) divergence. Since the posterior $p(\boldsymbol{\theta} \mid \mathcal{D})$ is the distribution to be approximated and as such is not accessible, the variational problem described by expression (3) cannot be solved directly. However, it can be shown that solving this variational problem is mathematically equivalent to maximizing the variational objective

$$\mathcal{F}(q(\boldsymbol{\theta})) = \mathbb{E}_{q(\boldsymbol{\theta})}[\log p(\mathcal{D} \mid \boldsymbol{\theta})] - \mathbb{D}_{\mathrm{KL}}(q_{\boldsymbol{\Theta}} \mid\mid p_{\boldsymbol{\Theta}})$$

with respect to a variational distribution $q_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) \in \mathcal{Q}_{\boldsymbol{\Theta}}$. Put another way,

$$\min_{q_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) \in \mathcal{Q}_{\boldsymbol{\Theta}}} \mathbb{D}_{\mathrm{KL}}(q_{\boldsymbol{\Theta}} \mid\mid p_{\boldsymbol{\theta}|\mathcal{D}}) \Longleftrightarrow \max_{q_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) \in \mathcal{Q}_{\boldsymbol{\Theta}}} \mathcal{F}(q(\boldsymbol{\theta})).$$

The variational objective $\mathcal{F}(q(\boldsymbol{\theta}))$ is commonly referred to as the evidence lower bound (ELBO), since it can be shown that

$$\log p(\mathcal{D}) = \mathcal{F}(q(\boldsymbol{\theta})) + \mathbb{D}_{\mathrm{KL}}(q_{\boldsymbol{\Theta}} \mid\mid p_{\boldsymbol{\theta}|\mathcal{D}}),$$

which, by non-negativity of the KL divergence, implies that $\log p(\mathcal{D}) \geq \mathcal{F}(q(\boldsymbol{\theta}))$. So, the variational objective is a lower bound on the evidence, that is, on the log-marginal likelihood $\log p(\mathcal{D})$. Finally, it is noted that $\log p(\mathcal{D}) = \mathcal{F}(q(\boldsymbol{\theta}))$ if and only if $q(\boldsymbol{\theta}) = p(\boldsymbol{\theta} \mid \mathcal{D})$, which means that the ELBO is perfectly tight if and only if the variational distribution is equal to the posterior.

In general, variational inference is not guaranteed to converge to the posterior $p(\boldsymbol{\theta} \mid \mathcal{D})$ unless the variational objective is convex in the variational parameters and the posterior is a member of the variational family, that is, $p(\boldsymbol{\theta} \mid \mathcal{D}) \in \mathcal{Q}_{\boldsymbol{\Theta}}$. Various approximate inference methods have been developed to solve the variational problem described by expression (3).

These methods make different assumptions about the variational family $\mathcal{Q}_{\mathbf{\Theta}}$, and therefore result in different posterior approximations.

For variational inference with neural networks, two well-established methods are Monte Carlo dropout (Gal & Ghahramani, 2016) and Gaussian mean-field variational inference (also referred to as Bayes-by-backprop; Blundell et al., 2015b; Graves, 2011). These methods are suited for stochastic mini-batch-based variational inference and can be scaled to large neural networks (Hoffman et al., 2013). Recent work on function-space variational inference (FSVI; Sun et al., 2019; Rudner et al., 2022a;b) in Bayesian neural networks (BNNs) frames variational inference as optimization over induced functions, that is,

$$\min_{q_{\mathbf{F}}(\mathbf{f}) \in \mathcal{Q}_{\mathbf{F}}} \mathbb{D}_{\mathrm{KL}}(q_{\mathbf{F}} \,||\, p_{\mathbf{F}|\mathcal{D}})$$

for

$$p(\mathbf{f} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \mathbf{f})\, p(\mathbf{f})}{p(\mathcal{D})}$$

with a suitably defined prior distribution $p(\mathbf{f})$ over functions. FSVI has been shown to result in state-of-the-art predictive uncertainty estimates in computer vision tasks (Rudner et al., 2022a).

A.3. Ensembles

Deep ensembling refers to a procedure where a neural network architecture is re-trained multiple times with different initializations to find different parameter settings, and then the resulting predictive distributions at those parameter settings are averaged at test time (Lakshminarayanan et al., 2017). In practice, deep ensembles provide a simple approach to representing epistemic uncertainty by capturing the variability in model predictions. This approach contrasts with more conventional Bayesian methods that involve sampling from the posterior distribution. Although unorthodox as an approximate inference procedure, deep ensembles often provide a closer approximation to the true posterior predictive distribution than many conventional approximate inference methods in deep learning (Wilson & Izmailov, 2020; Izmailov et al., 2021b), such as variational inference with a Gaussian approximate posterior.

In particular, one minimizes the standard loss for different initializations, which is often equivalent to minimizing a negative log-posterior $\log p(\boldsymbol{\theta} \mid \mathcal{D})$ to obtain

$$\tilde{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^{\nu}} \mathcal{L}(\boldsymbol{\theta}) = \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^{\nu}} (-\log p(\boldsymbol{\theta} \mid \mathcal{D})) = \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^{\nu}} (-\log p(\mathcal{D} \mid \boldsymbol{\theta}) - \log p(\boldsymbol{\theta}))\,,$$

where $\boldsymbol{\theta} \in \mathbb{R}^{\nu}$ are the neural network parameters, and $\mathcal{D}$ represents the training dataset. The negative log-likelihood $-\log p(\mathcal{D} \mid \boldsymbol{\theta})$ may correspond to cross-entropy loss, and a Gaussian prior $-\log p(\theta)$ corresponds to standard ℓ_2 regularization or weight decay. After finding different local solutions $\tilde{\boldsymbol{\theta}}_1, \ldots, \tilde{\boldsymbol{\theta}}_s$ starting from different initializations, one averages the predictive distributions to make predictions given a test input x':

$$p(\mathbf{y} \mid \mathbf{x}', \mathcal{D}) = \frac{1}{s} \sum_{i=1}^{s} p(\mathbf{y} \mid \mathbf{x}', \tilde{\boldsymbol{\theta}}_i). \tag{4}$$

Initially, the procedure of neural network ensembling at test time was not framed in probabilistic terms and was frequently described as a 'non-Bayesian' alternative to standard approximate inference methods such as the Laplace approximation. However, Equation (4) can be seen as approximating the true posterior predictive distribution

$$p(\mathbf{y} \mid \mathbf{x}', \mathcal{D}) = \int p(\mathbf{y} \mid \mathbf{x}', \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathcal{D})\, d\boldsymbol{\theta}. \tag{5}$$

There are different ways to interpret this predictive distribution. Several works have explored the connections between Bayesian inference and deep ensembles (Ciosek et al., 2020; Gustafsson et al., 2020; He et al., 2020; Pearce et al., 2020; Wilson & Izmailov, 2020; Izmailov et al., 2021b; D'Angelo & Fortuin, 2021; D'Angelo et al., 2021). One interpretation views Equation (4) as a Monte Carlo approximation of Equation (5), where the posterior of the parameters is represented as a set of point masses centered at different modes, which may be viewed as approximate posterior samples. However, this interpretation is not the most insightful.

It is more enlightening to view approximate inference as the task of accurately approximating the integral in Equation (5). From this perspective, the focus is not on collecting posterior samples. For a fixed computational budget, a Monte Carlo average of predictive distribution values based on exact posterior samples can provide a poor approximation of the integral relative to alternatives. A more compelling approach to numerical integration is to choose parameter values that represent typical points in the posterior, indicative of regions with significant posterior probability mass, and that yield diverse predictions on the test set. A heuristic to achieve this goal is to choose points corresponding to different posterior modes, as achieved by deep ensembles (Wilson & Izmailov, 2020). In practice, there is more functional variability across different posterior modes compared to samples in the vicinity of a single mode, such as the ones found from a variational Gaussian approximation of the posterior.

These observations are corroborated in practice by experiments. Deep ensembles tend to provide a closer approximation to the posterior predictive distribution, represented by exhaustive Hamiltonian Monte Carlo (HMC) sampling, than conventional unimodal posterior approximations (Izmailov et al., 2021b). The success of deep ensembles suggests that achieving a closer approximation to the posterior predictive distribution can lead to better predictive performance, highlighting the potential for further research. There are many natural ways to approximate the posterior predictive distribution. An obvious approach is to use a mixture of Gaussians centered at posterior modes, rather than a mixture of point masses. This approach has been found to approximate the posterior predictive distribution more closely and achieve better predictive performance in the NeurIPS 2021 approximate Bayesian inference competition (Wilson & Izmailov, 2020; Wilson et al., 2022; Shen et al., 2024).

A more general lesson to be extracted from these findings is that it is often not reasonable to consider whether a method is 'Bayesian' as a binary; different approximate inference procedures fall onto a spectrum representing how closely they approximate the true posterior predictive distribution. Different methods provide better or worse approximations, depending on the model and the data. In the case where the parameter posterior is unimodal, deep ensembles are less useful as an inference procedure. On the other hand, if many modes are available and the modes correspond to functions that make different predictions, then deep ensembles are sensible as an approximate Bayesian inference procedure, especially under computational constraints when it is not feasible to represent many different parameter settings.[1]

A.4. Posterior Sampling Algorithms

Sampling algorithms, particularly Markov chain Monte Carlo (MCMC) methods, are widely used for Bayesian posterior inference. These algorithms work by constructing a Markov chain whose equilibrium distribution matches the desired (target) distribution. Updating the parameters by realizing a Markov chain yields samples from the target distribution, provided a sufficient number of updates are performed. Given a dataset $\mathcal{D}$, a model with parameters $\boldsymbol{\theta} \in \mathbb{R}^{\nu}$, and a prior $p(\boldsymbol{\theta})$, the aim is to sample from the target posterior $p(\boldsymbol{\theta} \mid \mathcal{D}) \propto \exp(-U(\boldsymbol{\theta}))$, where the energy function is

$$U(\boldsymbol{\theta}) = -\sum_{\mathbf{x} \in \mathcal{D}} \log p(\mathbf{x} \mid \boldsymbol{\theta}) - \log p(\boldsymbol{\theta}).$$

Simulating the following continuous-time stochastic differential equation (SDE) produces samples with $p(\boldsymbol{\theta} \mid \mathcal{D})$ as its stationary distribution:

$$\mathrm{d}\boldsymbol{\theta} = -\nabla U(\boldsymbol{\theta}_t)\mathrm{d}t + 2\mathrm{d}B_t. \tag{6}$$

$\nabla U(\boldsymbol{\theta})$ is the drift term of the SDE that guides the generated samples towards the posterior distribution, and B_t is Brownian motion which introduces randomness into the process. The SDE in Equation (6) is also known as the Langevin diffusion equation and is used as the basis of many Monte Carlo sampling algorithms (Nemeth & Fearnhead, 2021). If the Langevin diffusion equation is considered over a small time interval $\alpha > 0$, then a discrete-time version of it can be derived via the Euler-Maruyama approximation as

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \alpha \nabla U(\boldsymbol{\theta}_k) + \sqrt{2\alpha}\boldsymbol{\xi}_{k+1}, \tag{7}$$

where $\alpha > 0$ is the step size parameter and $\boldsymbol{\xi}$ is standard Gaussian noise. This discrete-time algorithm is known as the unadjusted Langevin algorithm, or the Langevin Monte Carlo algorithm. However, unlike the continuous-time Langevin diffusion equation, the discrete-time unadjusted Langevin algorithm does not simulate samples with $p(\boldsymbol{\theta} \mid \mathcal{D})$ as its stationary distribution, but instead produces samples that are only approximately drawn from $p(\boldsymbol{\theta} \mid \mathcal{D})$. The discretization of the SDE leads to a bias in the posterior samples, which can be reduced by decreasing the step size parameter α.

[1] For more information on how deep ensembles facilitate approximate Bayesian inference, see the webpage https://cims.nyu.edu/~andrewgw/deepensembles/.

For large datasets, the unadjusted Langevin algorithm (7) can be computationally expensive due to the need to sum over the entire dataset when evaluating $\nabla U(\boldsymbol{\theta})$. Stochastic gradient Langevin dynamics (SGLD; Welling & Teh, 2011) reduces the computational cost by using a stochastic gradient estimator $\nabla \tilde{U}$, an unbiased estimator of ∇U based on a subset of the dataset $\mathcal{D}$. SGLD has initiated a line of research on stochastic gradient MCMC (SG-MCMC) algorithms. It updates the vector of parameters $\boldsymbol{\theta}$ at the $(k+1)$-th step according to

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \alpha \nabla \tilde{U}(\boldsymbol{\theta}_k) + \sqrt{2\alpha}\boldsymbol{\xi}_{k+1}.$$

The key difference between SGLD and stochastic gradient descent (SGD) is the additional Gaussian noise in each step of SGLD, which allows it to characterize the full parameter posterior distribution rather than converging to a single point.

Other notable variants of SG-MCMC include stochastic gradient HMC (SG-HMC; Chen et al., 2014), which accelerates convergence using auxiliary momentum variables, and cyclical SG-MCMC (Zhang et al., 2020b), which employs a cyclical step size schedule to efficiently explore multiple modes of the parameter posterior distribution. There have also been efforts to mitigate the bias in SG-MCMC methods by using Metropolis adjustments (Zhang et al., 2020a; Garriga-Alonso & Fortuin, 2021).

A.5. Prior Specification

The specification of prior $p(\boldsymbol{\theta} \mid \mathcal{M})$ on a vector of parameters $\boldsymbol{\theta} \in \mathbb{R}^\nu$ of a statistical model $\mathcal{M}$ has been a central part of Bayesian analysis, allowing to incorporate existing domain knowledge or expert opinion into statistical inference. A prior is called informative if it reflects such knowledge. Specific edge cases of informative priors include strongly informative priors, which dominate over the information coming from the observed data (likelihood), and weakly informative priors, which align with existing knowledge in a vague way so that the posterior is regularized to be data-informed and to be based on prior knowledge. In some cases, prior knowledge does not exist or a modeler does not want to rely on subjective knowledge. In such cases, uninformative or objective priors are used, where a common choice is a near-flat or even a uniform prior over the parameters. Another choice of objective prior worth mentioning is the reference prior, which is constructed to maximize some distance or divergence between the posterior and the chosen prior. Finally, in modern applications, priors are often selected to incorporate some desired properties into the model, such as regularization or sparsity. The model $\mathcal{M}$ is considered herein to be a BNN. However, the priors discussed below are most commonly used in other statistical models, from which they have been typically adopted for BDL.

A common approach is to specify independent and identically distributed (i.i.d.) priors for the BNN parameters $\boldsymbol{\theta}$. More specifically, a common default choice is to use a zero-centered isotropic Gaussian prior

$$p(\boldsymbol{\theta} \mid \mathcal{M}) = \prod_{i=1}^{\nu} \mathcal{N}(\theta_i; 0, \sigma^2),$$

which corresponds to l_2 regularization in the sense of maximum a posteriori (MAP) solutions with $\lambda = 1/(2\sigma^2)$. Thus, the larger the prior variance, the less regularization is incorporated, and vice versa. Combined with specific activation functions, such as the logistic function, which is close to linear around 0, choosing a small σ^2 results in more linear behavior of the neurons and their compositions, while large σ^2 allows for more non-linear behavior. Thus, the popular approach of choosing standard Gaussian priors is not satisfactory in most cases and may lead to misspecified models. This, in turn, can cause the cold posterior effect that has been known to be the case for linear models (Grünwald & Van Ommen, 2017), but is also observed for BNNs (Wenzel et al., 2020; Fortuin et al., 2022; Nabarro et al., 2022). For a specific problem, σ^2 can be chosen via hyper-parameter tuning or empirical Bayes. Moreover, a direct translation of the tuned σ^2 for some architecture (or, equivalently, λ for frequentist neural networks) is possible. Another approach is to impose an inverse-Gamma hyper-prior on σ^2, for example, $\sigma^2 \sim \Gamma^{-1}(\alpha, \beta)$, see Lampinen & Vehtari (2001). To incorporate prior dependencies between the parameters, i.i.d. Gaussian priors can be extended to multivariate normals with a zero mean vector and a covariance matrix $\boldsymbol{\Sigma}$, i.e.,

$$p(\boldsymbol{\theta} \mid \mathcal{M}) = \mathcal{N}_\nu(\boldsymbol{\theta}; \mathbf{0}, \boldsymbol{\Sigma}),$$

with the possibility to use an inverse-Wishart hyper-prior on $\boldsymbol{\Sigma}$. Similarly to the i.i.d. Gaussian priors, independent Laplace priors can be used, i.e.,

$$p(\boldsymbol{\theta} \mid \mathcal{M}) = \prod_{i=1}^{\nu} \mathrm{Laplace}(\theta_i; 0, b),$$

which in the MAP sense correspond to the l_1 regularization (Williams, 1995) with $\lambda = b^{-1}$. Choosing the scale parameter b can be done similarly to how σ^2 is chosen for Gaussian priors. Furthermore, Student-t priors have been used in the context of BNNs (Neklyudov et al., 2018). Heavy-tailed priors are possibly more robust towards model misspecification in the sense of the cold posterior effect (Fortuin et al., 2022).

Another desirable property that is often integrated into BNNs is sparsity. Mixtures of Gaussians have been popular in this context, including a scale mixture of Gaussians prior (Blundell et al., 2015a),

$$p(\boldsymbol{\theta} \mid \mathcal{M}) = \prod_{i=1}^{\nu} \left(\pi \mathcal{N}(\theta_i; 0, \sigma_1^2) + (1-\pi)\mathcal{N}(\theta_i; 0, \sigma_2^2) \right),$$

with $\sigma_1^2 > \sigma_2^2$ and $\sigma_2^2 \ll 1$. Similarly, one can use horseshoe priors (Carvalho et al., 2009),

$$p(\boldsymbol{\theta} \mid \mathcal{M}) = \prod_{i=1}^{\nu} \mathcal{N}(\theta_i; 0, \sigma^2 \tau_i^2),$$

where τ_i is the local shrinkage parameter that has a half-Cauchy hyperprior $\tau_i \sim \mathrm{C}^+(0,1)$, and σ is the global shrinkage parameter. Finally, another sparsity-inducing prior is the (improper) log-uniform prior (Molchanov et al., 2017),

$$p(\boldsymbol{\theta} \mid \mathcal{M}) = \prod_{i=1}^{\nu} \mathrm{LogU}_{\infty}(\theta_i) \propto \prod_{i=1}^{\nu} \frac{1}{\theta_i},$$

and its proper counterpart (Neklyudov et al., 2017),

$$p(\boldsymbol{\theta} \mid \mathcal{M}) \propto \prod_{i=1}^{\nu} \mathrm{LogU}_{\infty}(\theta_i) \mathrm{I}_{[a,b]}(\log \theta_i).$$

Some priors based on directional statistics have been explored for BNNs (Sunde, 2023), but have not gained widespread adoption. Similarly, Jeffreys priors (Ibrahim & Laud, 1991) have not been used extensively in this context. Although Zellner's g-priors (Zellner, 1986) and mixtures of g-priors (Li & Clyde, 2018) are highly popular in linear models, their application in BDL has only recently garnered attention (Antorán et al., 2022).

In Bayesian statistics, model uncertainty has been studied extensively for several decades (Hoeting et al., 1998; 1999; Wasserman, 2000). Within this framework, rather than having a single model $\mathcal{M}$, multiple BNN architectures $\{\mathcal{M}_1, \ldots, \mathcal{M}_t\}$ from a model space $\mathbb{M}$ are considered, making use of both $p(\boldsymbol{\theta} \mid \mathcal{M})$ and $p(\mathcal{M})$. Recent research has focused on model uncertainty with respect to a model space $\mathbb{M}$ defined by different patterns of weight inclusion (Hubin & Storvik, 2019; Skaaret-Lund et al., 2023), resulting in 2^ν models in $\mathbb{M}$. This requires additional model priors. If a model $\mathcal{M} = (\gamma_1, \ldots, \gamma_\nu)$ with $\gamma_i \in \{0,1\}, i \in \{1, \ldots, \nu\}$, is assumed, then Hubin & Storvik (2019) and Skaaret-Lund et al. (2023) propose

$$p(\mathcal{M}) = \prod_{i=1}^{\nu} \mathrm{Bernoulli}(\gamma_i; \rho_i),$$

with ρ_i being the prior inclusion probability for a specific weight. Similarly, Hubin & Storvik (2024) propose to use

$$p(\mathcal{M}) \propto \prod_{i=1}^{\nu} \mathrm{BetaBinomial}(\gamma_i; 1, a_i, b_i).$$

These two types of prior are common in the Bayesian model-averaging literature (Hoeting et al., 1999; Corani & Mignatti, 2015). However, more advanced model priors that incorporate dependencies between parameter inclusions through, for example, Dirichlet process hyper-priors (Grün & Hofmarcher, 2021) or dilution priors (George, 2010), have not yet been studied in the context of BDL. It is also noteworthy that, for model priors, inclusion probabilities for specific covariates for the input layer can be adjusted by experts according to prior knowledge, thus allowing the incorporation of domain-specific information into inference for BNNs.

Incorporating prior knowledge directly into parameter priors presents a challenge in general. However, recent advances in probabilistic modeling for neural networks have shown that incorporating prior knowledge is possible. One approach involves leveraging auxiliary objectives to create data-driven priors (Lopez et al., 2023; Rudner et al., 2023; 2024a; Sam et al., 2024). Another approach to specifying meaningful priors for neural networks is to adopt a function-space perspective. In this approach, BNNs generate a distribution $p(\mathbf{f})$ over functions when sampling from parameter priors. A functional prior, such as a Gaussian process $p(\mathbf{f}) = \mathcal{GP}(\boldsymbol{\mu}(\cdot), \mathbf{K}(\cdot, \cdot))$, can then be assumed for the function-space output, allowing the incorporation of expert knowledge about the mean and covariance functions for a specific phenomenon of interest. However, a direct application of this type of functional prior can be problematic due to potential mismatches between the support of the GP and the outputs of the BNN; see Burt et al. (2020); Rudner et al. (2022a). For the same reason, using the KL divergence to pre-train the priors over the parameters to match the chosen GP prior is problematic. Rudner et al. (2022a) resolve this issue by considering a KL divergence between distributions over functions that are absolutely continuous to one another by design. Tran et al. (2022a) also tackle this challenge by using the 1-Wasserstein distance instead of the KL divergence to learn the parameters of the priors in the weight space that match a chosen GP prior.

This section offers only a concise glimpse into the extensive literature on priors for BNNs. For a more comprehensive understanding and a relatively recent review, the reader is referred to Fortuin (2022).

A.6. Deep Kernel Processes

A deep kernel process (DKP; Aitchison et al., 2021) is a Bayesian model that places a prior on a deep sequence of kernel representations. This is a change of perspective compared to other deep Bayesian models such as deep GPs (DGPs; Damianou & Lawrence, 2013) or BNNs (MacKay, 1995), which place priors over intermediate layer features or weights. DKPs are equivalent to DGPs whenever the kernel function is isotropic (such as the radial basis function and Matérn kernel). To illustrate this, consider a DGP with an isotropic kernel $\mathbf{C}$, where each layer is modeled as a multivariate Gaussian conditioned on the preceding layer,

$$\mathbf{F}_0 = \mathbf{X}, \tag{8a}$$

$$g(\mathbf{F}_j \mid \mathbf{F}_{j-1}) = \prod_{i=1}^{r_j} \mathcal{N}(\mathbf{f}_{i,j}; \mathbf{0}, \mathbf{C}(\mathbf{F}_j)), \tag{8b}$$

$$g(\mathbf{Y} \mid \mathbf{F}_{\eta+1}) = \prod_{i=1}^{r_j} \mathcal{N}(\mathbf{y}_i; \mathbf{f}_{i,\eta+1}, \sigma^2 \mathbf{I}). \tag{8c}$$

Here, $\mathbf{F}_j \in \mathbb{R}^{n \times r_j}$ are the feature representations in each intermediate layer $j \in \{1, \ldots, \eta\}$, $\mathbf{X} \in \mathbb{R}^{n \times r_0}$ are the inputs, $\mathbf{Y} \in \mathbb{R}^{n \times \rho_{\eta+1}}$ are the labels, and n are the number of data points. The subscript i denotes individual features, so that $\mathbf{f}_{i,j} \in \mathbb{R}^n$ is the i-th feature at layer j, and $\mathbf{y}_i \in \mathbb{R}^n$ is the i-th output for all data points. r_j is the number of features per data point at layer j, or 'the width of layer' j. To obtain a kernel process, one needs to consider covariance matrices. So, for each layer, the Gram matrix $\mathbf{G}_j = \mathbf{F}_j \mathbf{F}_j^T / r_j \in \mathbb{R}^{n \times n}$ is defined. Since $\mathbf{G}_j$ is the outer product of i.i.d. Gaussian samples with covariance $\mathbf{C}(\mathbf{F}_j)$, it must be Wishart distributed,

$$g(\mathbf{G}_j \mid \mathbf{F}_{j-1}) = \mathcal{W}(\mathbf{G}_j; \mathbf{C}(\mathbf{F}_{j-1})/r_j, r_j).$$

Furthermore, by the isotropic assumption, there is a function $\mathbf{K}(\cdot)$ over Gram matrices such that $\mathbf{K}(\mathbf{G}_j) = \mathbf{C}(\mathbf{F}_j)$; see Aitchison et al. (2021). The ability to define the kernel function in terms of Gram matrices means that it is possible to write the DGP in Equation (8) as a DKP with Wishart priors,

$$g(\mathbf{G}_1 \mid \mathbf{X}) = \mathcal{W}(\mathbf{G}_1; \mathbf{X}\mathbf{X}^T / r_0, r_0), \tag{9a}$$

$$g(\mathbf{G}_j \mid \mathbf{G}_{j-1}) = \mathcal{W}(\mathbf{G}_j; \mathbf{K}(\mathbf{G}_{j-1})/r_j, r_j), \tag{9b}$$

$$g(\mathbf{y}_i \mid \mathbf{G}_\eta) = \mathcal{N}(\mathbf{y}_i; \mathbf{0}, \mathbf{K}(\mathbf{G}_\eta) + \sigma^2 \mathbf{I}). \tag{9c}$$

Since Equation (9) places Wishart priors on the intermediate Gram matrix representations, the resulting process is known as a deep Wishart process (DWP; Aitchison et al., 2021). Deep inverse Wishart processes (DIWPs; Aitchison et al., 2021) are defined using inverse Wishart process priors over kernels (Shah et al., 2014) instead.

Similarly to other deep Bayesian models, closed-form inference of general DKPs is not possible. Aitchison et al. (2021), Ober & Aitchison (2021a) and Ober et al. (2023) have developed approximate posteriors over Gram matrices for DWPs and DIWPs to allow for variational inference. However, despite the use of approximate posteriors, the computational cost of training a DWP or DIWP remains considerable. This is because the number of parameters scales quadratically with the number n of data points, and evaluating the log-probabilities of their approximate posteriors (necessary when evaluating the ELBO) scales cubically with n. To address this scalability challenge, inducing point approximations offer a solution. In particular, global inducing point methods (Ober & Aitchison, 2021b) enable the training of DKPs with linear scaling in the number of data points.

Using inducing point schemes, Ober et al. (2023) have empirically demonstrated that approximate posteriors for DWPs perform better than DGP approximate posteriors. Aitchison et al. (2021) argue that DWPs are expected to perform better due to BNN and DGP priors and posteriors being highly multimodal. In particular, rotation and permutation symmetries in features or weights are not adequately accounted for by common BNN and DGP approximate posteriors. DKPs sidestep this multimodality issue, as Gram matrices inherently avoid these symmetries; an arbitrary rotation or permutation in feature space can be represented by the mapping $\mathbf{F} \mapsto \mathbf{FU}$, where $\mathbf{U}$ is unitary, yet the corresponding Gram matrix is invariant under this transformation since $\mathbf{G} = \mathbf{F}\mathbf{F}^T \mapsto (\mathbf{FU})(\mathbf{FU})^T = \mathbf{G}$.

A.7. Deep Kernel Machines

Deep kernel machines (DKMs; Yang et al., 2023a) are an infinite-width analog of DKPs. They have practical benefits in being easier to implement and cheaper to train, and also theoretical benefits as they can be linked to the existing infinite-width neural network literature. However, DKMs are not strictly Bayesian. Usually, taking an infinite-width limit of a DKP or DGP results in a neural network Gaussian process (NNGP; Lee et al., 2017; Agrawal et al., 2020). The infinite-width limit is taken carefully, in such a way so as to retain flexibility in intermediate Gram representations.

A DKM can be obtained from the DGP of Equation (8) as follows. Consider the following approximate posterior for the features in each intermediate layer $j \in \{1, \ldots, \eta\}$:

$$h(\mathbf{F}_j) = \prod_{i=1}^{r_j} \mathcal{N}(\mathbf{f}_{i,j}; \mathbf{0}, \mathbf{G}_j).$$

Moreover, consider a standard GP approximate posterior for the final layer:

$$h(\mathbf{F}_{\eta+1}) = \prod_{i=1}^{r_{\eta+1}} \mathcal{N}(\mathbf{f}_{i,\eta+1}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}).$$

Here, $\mathbf{G}_1, \ldots, \mathbf{G}_\eta, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{r_{\eta+1}}$, and $\boldsymbol{\Sigma}$ are variational parameters. Although this approximate posterior family may seem restrictive, the intermediate layer part contains the true posterior in the infinite-width limit; see Appendix E of Yang et al. (2023a). A lower bound for the marginal likelihood can be obtained via the ELBO

$$\text{ELBO} = \sum_{i=1}^{r_{\eta+1}} \left\{ \mathbb{E}_{h(\mathbf{F}_{\eta+1})}\left[\log g(\mathbf{y}_i \mid \mathbf{f}_{i,\eta+1})\right] - \mathbb{D}_{\mathrm{KL}}(h(\mathbf{f}_{i,\eta+1}) \mid\mid g(\mathbf{f}_{i,\eta+1} \mid \mathbf{F}_\eta)) \right\} - \sum_{j=1}^{\eta} \beta_j r_j \mathbb{D}_{\mathrm{KL}}(h(\mathbf{f}_j) \mid\mid g(\mathbf{f}_j \mid \mathbf{F}_{j-1})),$$

where tempering is employed using the parameter β_j. As with DWPs, an isotropic kernel function is assumed, which means that $\mathbf{C}(\mathbf{F}_j) = \mathbf{K}(\mathbf{G}_j)$. As the intermediate layers become wider by sending $r \to \infty$ with $r_j = r\rho_j$, the dependency on $\mathbf{F}_j$ disappears. If no tempering is applied (that is, $\beta_j = 1$), then the following objective is recovered:

$$\frac{\text{ELBO}}{r} \to -\sum_{j=1}^{\eta} \rho_j \mathbb{D}_{\mathrm{KL}}(\mathcal{N}(\mathbf{0}, \mathbf{G}_j) \mid\mid \mathcal{N}(\mathbf{0}, \mathbf{K}(\mathbf{G}_{j-1}))). \tag{10}$$

The objective (10) is maximized when $\mathbf{G}_j = \mathbf{K}(\mathbf{G}_{j-1})$, which is the same as the corresponding NNGP (Lee et al., 2017; Agrawal et al., 2020). If tempering is carried out according to the width with $\beta_j = 1/r$, then the following objective is obtained:

$$\begin{aligned} \text{ELBO} \to &\sum_{i=1}^{r_{\eta+1}} \mathbb{E}_{h(\mathbf{F}_{\eta+1})}\left[\log g(\mathbf{y}_i \mid \mathbf{f}_{i,\eta+1})\right] \\ &- \sum_{i=1}^{r_{\eta+1}} \mathbb{D}_{\mathrm{KL}}(\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}) \mid\mid \mathcal{N}(\mathbf{0}, \mathbf{K}(\mathbf{G}_\eta))) \\ &- \sum_{j=1}^{\eta} \rho_j \mathbb{D}_{\mathrm{KL}}(\mathcal{N}(\mathbf{0}, \mathbf{G}_j) \mid\mid \mathcal{N}(\mathbf{0}, \mathbf{K}(\mathbf{G}_{j-1}))) + \text{constant}. \end{aligned} \tag{11}$$

A model that optimizes the objective (11) is called a DKM, and (11) is known as the DKM objective. In the limit, the DKM objective does not depend on intermediate features $\mathbf{F}_j$, which means that learned representations in a DKM are described entirely by deterministic Gram matrices $\mathbf{G}_1, \ldots, \mathbf{G}_\eta$. To interpret the DKM objective, notice that the likelihood term encourages data fitting, and the KL divergences regularize the model toward the NNGP (Lee et al., 2017; Agrawal et al., 2020). The amount of representation learning in the DKM can be controlled by varying the ρ_j parameters. In contrast, the lack of likelihood term in the NNGP objective (10) prevents representation learning from occurring in the NNGP; intermediate Gram matrices are fixed and depend only on the input data.

Similarly to DKP objectives, the DKM objective is computationally infeasible to optimize for large datasets, with cubic scaling in the number of data points. However, Yang et al. (2023a) have shown that the DKM objective can be optimized with linear scaling if global inducing point methods are used. DKMs have been extended to convolutional architectures, achieving performance nearly on par with neural networks on CIFAR-10 (Milsom et al., 2023).

B. Diagnostics, Metrics and Benchmarks

Currently, there is a lack of convergence and performance metrics specifically for the needs of BDL. Developing such tools can help identify the goals in BDL as well as assess their progress. Besides, the choice of evaluation metrics, datasets and benchmarks lack consensus in the BDL community which reflects a difficulty in clearly defining the goals of BDL in a field traditionally viewed through frequentist lens, specifically in terms of performance on test data. Many of the general Bayesian diagnostic and evaluation approaches are proposed through Bayesian workflow (Gelman et al., 2020). This appendix discusses the most relevant approaches for BDL.

Convergence diagnostics in parameter space. The analysis of convergence and sampling efficiency (Gelman et al., 2013; Vehtari et al., 2021) for SG-MCMC sampling becomes a delicate matter, which is currently bypassed by a rather simplistic analysis of these quantities using summary statistics of predictive distributions. More generally, verifying the convergence of inference algorithms in the high-dimensional and multimodal settings of BDL models is not straightforward. Convergence checks designed for BNNs need to be further studied.

Performance metrics in predictive space. BDL and GP literature often focus on the mean of the predictive distribution, overlooking the analysis of variance of the predictive distribution. Some performance metrics are commonly used to assess variance levels, for example, by evaluating the log-likelihood or the entropy of predictions for test data (Rudner et al., 2022a; 2023). However, a systematic way to characterize the predictive uncertainty in BDL inference (apart from binary classification problems where AUROC and AUPRC are widely used) is currently lacking (Arbel et al., 2023). The challenge of setting metrics for the assessment of epistemic and aleatoric uncertainty slows the progress in BDL and could potentially be addressed by establishing widely accepted benchmarks for BDL methods.

Performance metrics in misspecified settings. Addressing challenges related to distribution shift and test data performance requires the development of robust performance metrics. To establish BDL model reliability under distribution shift, tighter generalization bounds, such as those provided by the PAC-Bayes framework (Langford & Shawe-Taylor, 2002; Parrado-Hernández et al., 2012), are crucial to obtain probabilistic guarantees on model performance. Furthermore, in misspecified settings, evaluating calibration becomes paramount. Innovative techniques, such as two-stage calibration (Guo et al., 2017) and conformal prediction (Papadopoulos et al., 2007) or its Bayesian counterpart (Hobbhahn et al., 2022), offer practical solutions by refining predicted probabilities and quantifying predictive uncertainty, respectively. These approaches collectively contribute to a more comprehensive evaluation of model performance in scenarios where the underlying assumptions may not align with the true data distribution.

Probabilistic treatment of datasets. Probabilistic treatments of data as a first-class citizen that can be reasoned about in BNNs seem promising. Such probabilistic approaches may help create more focused and useful datasets to represent the knowledge contained in vast data sources, improving the ability to train and maintain large models.

C. Software Usability

Applying a BDL approach to a real-world problem is still a more complex endeavor than opting for an off-the-shelf standard deep learning solution, which limits the real-world adoption of BDL. Software development is key to encouraging deep learning practitioners to use Bayesian methods. More generally, there is a need for software that would make it easier for practitioners to try BDL in their projects. The use of BDL must become competitive in human effort with standard deep learning.

Some efforts have been made to develop software packages, libraries or probabilistic programming languages (PPLs) on top of deep learning frameworks. `bayesianize` (Ritter et al., 2021), `bnn_priors` (Fortuin et al., 2021), `Laplace` (Daxberger et al., 2021a), `Pyro` (Bingham et al., 2019) and `TyXe` (Ritter & Karaletsos, 2022) are software species built on `PyTorch`, `TensorFlow Probability` is a library built on `TensorFlow`, and `Fortuna` (Detommaso et al., 2023) is a library built on JAX. It would help to make further progress with contributions from the probabilistic programming community.

PPLs, such as `Pyro`, play a role in simplifying the application of probabilistic reasoning to deep learning. In fact, abstractions of the probabilistic treatment of NNs in a PPL, such as those performed in the BDL library `TyXe`, can simplify the application of priors and inference techniques to arbitrary NNs, as demonstrated in a variety of models implemented in `TyXe`. Porting such ideas to modern problem settings involving LLMs and more bespoke probabilistic structures would enable the use of BDL in real-world problems.

Contemporary deep learning pushes the limits of scale in all dimensions: datasets, parameter spaces, and structured function-valued output. For point estimation, the community has been developing array-centric programming paradigms that allow sharding, partial evaluations, currying, and more. BDL should be able to map these ideas to develop analogous software.

D. Topical Developments

This appendix provides topical or specialized areas of BDL for future development. These include BDL for human-AI interaction, lifelong and decentralized learning, Bayesian reinforcement learning (RL), and domain-specific BDL models.

Human-AI interaction and explainable AI. Enabling AI systems to communicate and explain their uncertainty can build trust and improve the interaction between AI systems and humans. While efforts by the community have been made to explain the predictions of DNNs, recent efforts aim to explain the uncertainty of BDL methods (Antoran et al., 2021; Bhatt et al., 2021). Understanding which input patterns are responsible for high predictive uncertainty can build trust in AI systems and can provide insights about input regions where data is sparse. For example, when training a loan default predictor, a data scientist can identify population subgroups (by age, gender, or race) underrepresented in the training data. Collecting more data from these groups can lead to more accurate predictions for a wider range of clients.

Lifelong and decentralized learning. A contemporary research direction is to go beyond the ‘static’ train-test framework and focus on ‘dynamic’ problems where the test set is not known. This includes cases where predictive performance, robustness and safety are important and there are realistic constraints on the infrastructure. Two such problems are lifelong learning and decentralized learning. Focusing on such problems is expected to lead to a new regime in which Bayesian ideas can be useful for deep learning.

Efficient exploration in RL. RL is an area where BDL has shown potential. As an example, Thompson sampling (TS) is known to be a commonly used heuristic for decision making that ‘randomly selects an action, according to the probability that it is optimal’ (Russo et al., 2018). TS balances exploration with exploitation and in its exact form requires sampling from the Bayesian posterior. In practice, approximations are often used, and recent work has shown that the quality of the resulting multivariate joint predictive distribution over multiple test inputs is important for decision-making (Wen et al., 2021; Osband et al., 2023). This is relevant, as typical Bayesian and non-Bayesian methods are commonly evaluated by assessing the quality of marginal predictions over individual test inputs, ignoring potential dependencies (Osband et al., 2022). While deep ensembles are a typical baseline for capturing uncertainty, BDL methods based on the last-layer Laplace approximation can outperform deep ensembles in the quality of joint multivariate predictions (Antoran et al., 2023). Developing methods that achieve trade-offs between computational cost and the quality of their joint multivariate predictions is an area where further research is needed (Osband et al., 2023). Another active area of research at the intersection of RL and BDL aims to produce accurate posterior approximations of value functions (for example, Q functions) given data from interactions with an environment (Janz et al., 2019). This setting is different from typical Bayesian supervised learning as, in this case, the output of value functions is not directly observed, and only rewards are available.

Computer vision. BDL approaches to computer vision tasks have been developed. For instance, Kou et al. (2024) employ BDL in diffusion models to construct a pixel-wise uncertainty estimator for image generation. Goli et al. (2024) use BDL to evaluate uncertainty in pre-trained neural radiance fields in the context of computer graphics. Future research in BDL for computer vision may focus on improving predictive performance and further developing UQ methods. Computer vision, along with natural language processing, constitute applications that may promote the adoption of BDL.

Domain-specific BDL models. There are many opportunities to develop Bayesian methods in combination with deep learning models that are tailored for specific domains, taking into account the characteristics of the data and the tasks involved. This can involve exploring hierarchical models, transfer learning, or meta-learning approaches. An example is molecular property prediction, where many different datasets are available, but each of them has limited available data (Klarner et al., 2023). There is scope to combine deep learning models that learn molecular feature representations with Bayesian methods that receive those representations as inputs. The latter methods can capture uncertainty and make predictions in data-limited settings for each individual task, while the deep learning features are shared across tasks.