

# Hands-on Bayesian Neural Networks - a Tutorial for Deep Learning Users

LAURENT VALENTIN JOSPIN, University of Western Australia

WRAY BUNTINE, Monash University

FARID BOUSSAID, University of Western Australia

HAMID LAGA, Murdoch university

MOHAMMED BENNAMOUN, University of Western Australia

Modern deep learning methods have equipped researchers and engineers with incredibly powerful tools to tackle problems that previously seemed impossible. However, since deep learning methods operate as black boxes, the uncertainty associated with their predictions is often challenging to quantify. Bayesian statistics offer a formalism to understand and quantify the uncertainty associated with deep neural networks predictions. This paper provides a tutorial for researchers and scientists who are using machine learning, especially deep learning, with an overview of the relevant literature and a complete toolset to design, implement, train, use and evaluate Bayesian neural networks.

CCS Concepts: • **Mathematics of computing** → **Probability and statistics**; • **Computing methodologies** → **Neural networks**; **Bayesian network models**; *Ensemble methods*; *Regularization*.

Additional Key Words and Phrases: Bayesian methods, Bayesian Deep Learning, Approximate Bayesian methods

## ACM Reference Format:

Laurent Valentin Jospin, Wray Buntine, Farid Boussaid, Hamid Laga, and Mohammed Bennamoun. 2020. Hands-on Bayesian Neural Networks - a Tutorial for Deep Learning Users. *ACM Comput. Surv.* 1, 1 (July 2020), 35 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

## 1 INTRODUCTION

Deep learning has led to a revolution in machine learning, providing solutions to tackle more and more complex and challenging real-life problems. However, deep learning models are prone to **overfitting**, which adversely affects their generalization capabilities. Deep learning models also **tend to be overconfident** about their predictions (when they do provide a confidence interval). All of this is problematic for applications such as self driving cars [74], medical diagnostics [38] or trading and finance [11], where silent failure can lead to dramatic outcomes. Consequently, many approaches have been proposed to mitigate this risk, especially via the use of stochastic neural networks to estimate the uncertainty in the model prediction. The Bayesian paradigm provides a

---

Authors' addresses: Laurent Valentin Jospin, [laurent.jospin@research.uwa.edu.au](mailto:laurent.jospin@research.uwa.edu.au), University of Western Australia, 35 Stirling Hwy, Crawley, Western Australia, 6009; Wray Buntine, [wray.buntine@monash.edu](mailto:wray.buntine@monash.edu), Monash University, Wellington Rd, Monash, Victoria, 3800; Farid Boussaid, [farid.boussaid@uwa.edu.au](mailto:farid.boussaid@uwa.edu.au), University of Western Australia, 35 Stirling Hwy, Crawley, Western Australia, 6009; Hamid Laga, [h.laga@murdoch.edu.au](mailto:h.laga@murdoch.edu.au), Murdoch university, 90 South St, Murdoch, Western Australia, 6150; Mohammed Bennamoun, [mohammed.bennamoun@uwa.edu.au](mailto:mohammed.bennamoun@uwa.edu.au), University of Western Australia, 35 Stirling Hwy, Crawley, Western Australia, 6009.

---

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

© 2020 Association for Computing Machinery.

0360-0300/2020/7-ART \$15.00

<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

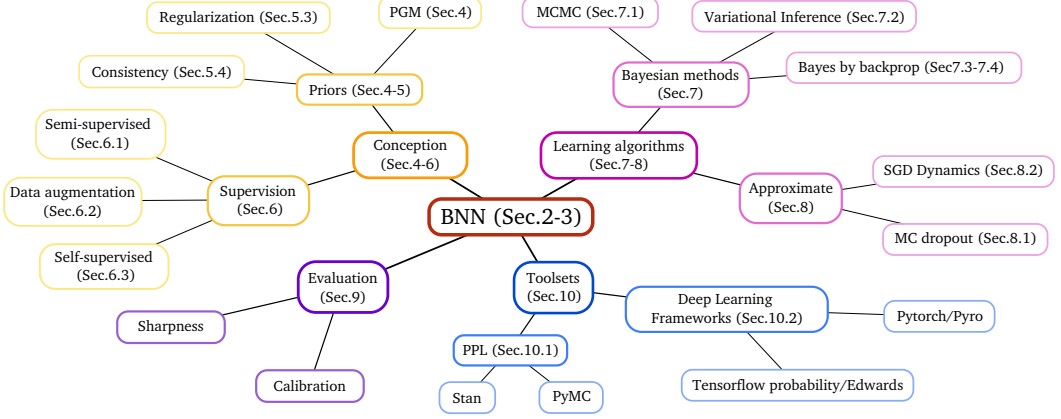


Fig. 1. A mind map of the topics covered in this article. These can broadly be divided into the conception of Bayesian deep neural networks, the different (strictly or approximately Bayesian) learning approaches, the evaluation methods, and the tool sets available to researchers for implementation.

rigorous framework to analyse and train such stochastic neural networks, and more generally to support the development of learning algorithms.

The Bayesian paradigm in statistics is often opposed to the pure frequentist paradigm, a major area of distinction being in hypothesis testing [15]. The Bayesian paradigm is based on two simple ideas. The first is that probability is a measure of belief in the occurrence of events, rather than just some limit in the frequency of occurrence when the number of samples goes towards infinity. The second is that prior beliefs influence posterior beliefs. All of this is summarized by Bayes' theorem, a very simple formula to invert conditional probabilities, and its interpretation in Bayesian statistics.

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)} = \frac{P(D|H)P(H)}{\int_H P(D|H')P(H')dH'} = \frac{P(D, H)}{\int_H P(D, H')dH'}. \quad (1)$$

In the classical interpretation,  $H$  and  $D$  are simply considered as sets of outcomes, while the Bayesian interpretation explicitly considers  $H$  to be a hypothesis, such as deep neural network parameters, and  $D$  to be some data, while in the classical interpretation one cannot define a probability law for an hypothesis.  $P(D|H)$  is called the likelihood,  $P(H)$  the prior,  $P(D)$  the evidence, and  $P(H|D)$  the posterior. We designate  $P(D|H)P(H) = P(D, H)$  as the joint probability of  $D$  and  $H$ .

This interpretation, which can be understood as learning from the data  $D$ , means that the Bayesian paradigm does not just offer a solid approach for the quantification of uncertainty in deep learning models. It also gives a mathematical framework to understand many regularization techniques and learning strategies that are already used in classic deep learning [69].

There is a rich literature in the field of Bayesian (deep) learning, including reviews [50, 85, 89], but none of which explores, in a specific and exhaustive way, the general theory of Bayesian neural networks. However, the field of Bayesian learning is much larger than just stochastic neural networks trained using a Bayesian approach, i.e., Bayesian neural networks. It makes it hard to navigate this literature without prior knowledge of Bayesian methods and advanced statistics, meaning there is an additional layer of complexity for deep learning practitioners willing to understand how to build and use Bayesian neural networks. This is one of the reason which explains why the number of theoretical contributions in the field is large, while the practical applications of Bayesian methods in deep learning are scarce. The other main reason probably

relates to the lack of efficient algorithms to overcome the computational challenges to combine Bayesian methods and big data, or the lack of knowledge about recent contributions to address those challenges.

This paper is meant to fill in this gap. It is conceived as a tutorial for scientists and postgraduate students who are already familiar with standard deep learning approaches, and who are interested in using Bayesian methods. It covers all of the basic principles needed to design, implement, train and evaluate a Bayesian neural network (Fig. 1). It also offers an extensive overview of the relevant literature about Bayesian neural networks, from the early seminal work dating back to the end of the 20th century [54] to the most recent contributions that have not been covered in any of the previously cited reviews. This tutorial also puts a strong focus on the practical aspects. A large number of approaches have been developed to build Bayesian neural networks, sometimes quite different in their intrinsic approach, and a good knowledge of those different methods is a prerequisites for an efficient use of Bayesian neural networks. To the best of our knowledge, there is no prior work in the literature which provides a systematic review of all those different approaches.

We start by defining, in Section 2, the concept of a Bayesian neural network. In Section 3, provide some of the motivations behind the use of deep Bayesian networks and why they are useful. We then present, in Section 4, some important concepts in statistics that are used to conceive and analyse Bayesian neural networks. Then, in Section 5, we present how prior knowledge, an important part of Bayesian statistics, is accounted for in Bayesian deep learning. We also consider the relationships between priors for Bayesian neural networks and regularization of traditional neural networks. In Section 6, we explain how Bayesian design tools are used to adjust the degree of supervision and to define a learning strategy. In Section 7, we explore some of the most important algorithms that are used for Bayesian inference. We review in Section 8 how Bayesian methods were specifically adapted to deep learning, to reduce the computational complexity or memory footprint. In Section 9, we present the methods used to evaluate the performance of a Bayesian neural network. In Section 10, we review the different frameworks that can be used for Bayesian deep learning. Finally, we conclude in Section 11.

## 2 WHAT IS A BAYESIAN NEURAL NETWORK?

A Bayesian neural network is defined slightly differently across the literature, but a common definition is that a Bayesian neural network is a stochastic artificial neural network trained using Bayesian inference (Fig. 2).

The goal of Artificial neural networks (ANNs) is to represent an arbitrary function  $\mathbf{y} = NN(\mathbf{x})$ . Traditional ANNs (e.g., feedforward networks, recurrent networks, branched networks, ...) are built using a succession of one input layer, a number of hidden layers and one output layer. We designate the input variables as  $\mathbf{x}$ , and output variables (predictions) as  $\mathbf{y}$ . In feedforward networks, the simplest architecture, each layer  $l$  is represented as a linear transformation of the previous one, followed by a non linear operation  $nl$  (a.k.a activation function):

$$\begin{aligned} \mathbf{l}_0 &= \mathbf{x}, \\ \mathbf{l}_i &= nl_i(\mathbf{W}_i \mathbf{l}_{i-1} + \mathbf{b}_i) \quad \forall i \in [1, n], \\ \mathbf{y} &= \mathbf{l}_n. \end{aligned} \tag{2}$$

More complex architectures also exist (e.g., networks with multiple inputs, outputs, exotic activation functions, recurrent architectures ...). This means that a given ANN architecture represents a set of functions isomorphic to the set of possible coefficients  $\theta$ , which represent all the weights  $\mathbf{W}$  and biases  $\mathbf{b}$  of the network. Deep learning is the process of regressing the parameters  $\theta$  on some training data  $D$ , usually a series of inputs  $\mathbf{x}$  and their corresponding labels  $\mathbf{y}$ . The standard approach is to approximate a minimal cost point estimate  $\hat{\theta}$  (Fig. 3a) using the back-propagation algorithm,

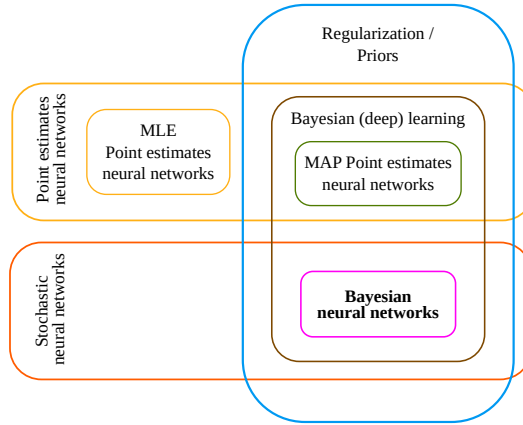


Fig. 2. A classification of neural networks from a statistical point of view. We distinguish point estimate neural networks, where a single instance of parameters is learned, and stochastic neural networks, where a distribution over the parameters is learned. Point estimate models without regularization, which imply implicit uniform prior, are learned using a maximum-likelihood estimator, while point estimate models with regularization are learned with a maximum a-posteriori estimator. Bayesian neural networks are stochastic neural networks with priors.

with all other possible parametrizations discarded. The cost function is often defined as the log likelihood of the training set, sometimes with a regularization term to penalize parametrizations. From a statistician point of view, this can be considered to be a Maximum Likelihood Estimation (MLE), respectively a Maximum A Posteriori (MAP) estimation when regularization is used (Fig. 2).

The point estimate approach is relatively easy (with modern algorithms and software packages), but tends to lack explainability and might generalize in unforeseen and overconfident ways on out-of-training-distribution data points [27, 63]. This property, and inability of ANNs to answer “I don’t know” is problematic in fields where their predictions have critical implications, such as trading, autonomous driving or medical applications. Techniques exist to mitigate this risk [28] based either on a threshold for the softmax-predicted class logit or an additional module to classify out-of-distribution samples. Another method for out-of-distribution detection is the use of Deep Generative Models, a class of ANNs (e.g., Generative Adversarial Networks) meant to encode complex data distributions [79]. Concerns have, however, been raised about these different approaches, either because they are too simple, like a threshold on the softmax logits, or because the additional module or deep generative models used for out-of-distribution detection might themselves suffer from the same overconfidence problems they were supposed to fix in the first place [60]. The flaws of these different approaches is one of the main motivations for the introduction of stochastic neural networks.

**Stochastic neural networks** are a type of ANNs built by introducing stochastic components into the network (by giving the network stochastic activation: Fig. 3b or stochastic weights: Fig. 3c) to simulate multiple possible models  $\theta$  with their associated probability distribution  $p(\theta)$ . They can, therefore, be considered as a special case of **ensemble learning** [99], where instead of training one single model, a set of models is trained and their predictions are aggregated.

The main motivation behind ensemble learning comes from the observation that aggregating the predictions of a large set of average performing but independent predictors can lead to much better predictions than one output by a single well-performing expert predictor [20]. Stochastic neural networks can be used in a similar fashion. It has been observed that their predictions

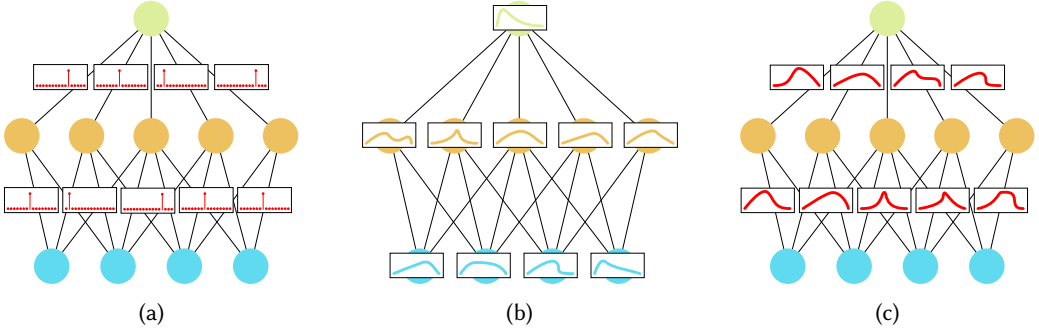


Fig. 3. Point estimate neural networks (a), where only a set of weights is learned, stochastic activation neural networks (b), where only a set of weights along with a probability distribution for the activation is learned, and stochastic coefficients neural networks (c), where a probability distribution over the weights is learned.

can improve over their point-estimate counterparts, even if there is no evidence that the use of stochastic neural networks is the best way to improve accuracy. Instead, the main goal of using a stochastic neural network architecture is to get a better idea of the uncertainty associated with the underlying processes. This is accomplished by comparing the predictions of multiple sampled model parametrization  $\theta$ . If the different models agree the uncertainty is low. If they disagree, then it is high. This process can be summarized as follow:

$$\begin{aligned} \theta &\sim p(\theta), \\ \mathbf{y} &= NN_{\theta}(\mathbf{x}) + \epsilon, \end{aligned} \quad (3)$$

where  $\epsilon$  represents random noise to account for the fact that the function  $NN$  is just an approximation.

A Bayesian Neural Network (BNN) can then be defined as any stochastic artificial neural network trained using Bayesian inference [54]. To design a BNN, the first step is the choice of a deep neural network architecture, i.e., of a functional model. Then, one has to choose a stochastic model, i.e., a prior distribution over the possible model parametrization  $p(\theta)$  and a prior confidence in the predictive power of the model  $p(\mathbf{y}|\mathbf{x}, \theta)$  (Fig. 4a). The model parametrization can be considered to be the hypothesis  $H$  and the training set is the data  $D$ . In the rest of this paper we will designate the model parameter as  $\theta$ , and use  $D$  to designate the training set,  $D_{\mathbf{x}}$  to designate the training features and  $D_{\mathbf{y}}$  to designate the training labels. This is to distinguish between the training data and any input/output pair  $(\mathbf{x}, \mathbf{y})$ . Applying Bayes theorem, and enforcing independence between the model parameters and the inputs, the Bayesian posterior can then be written as:

$$p(\theta|D) = \frac{p(D_{\mathbf{y}}|D_{\mathbf{x}}, \theta)p(\theta)}{\int_{\theta} p(D_{\mathbf{y}}|D_{\mathbf{x}}, \theta')p(\theta')d\theta'} \propto p(D_{\mathbf{y}}|D_{\mathbf{x}}, \theta)p(\theta). \quad (4)$$

Computing this distribution, and moreover sampling from it using standards methods, is usually an intractable problem, especially since computing the evidence  $\int_{\theta} p(D_{\mathbf{y}}|D_{\mathbf{x}}, \theta')p(\theta')d\theta'$  is hard. To address this, two approaches are possible. The first one is to use a Markov Chain Monte Carlo algorithm, which allows to sample the posterior directly, but needs to cache a collection of samples  $\Theta$ . The second one is to use a variational inference approach, which learns a variational distribution  $q_{\phi}(\theta)$  to approximate the exact posterior (Fig. 4b). Both of these methods bypass the computation of the evidence (denominator of Eq. 4), which explains why the posterior is often given up to a scaling constant, as we do in the rest of this tutorial for simplicity. These algorithms are presented

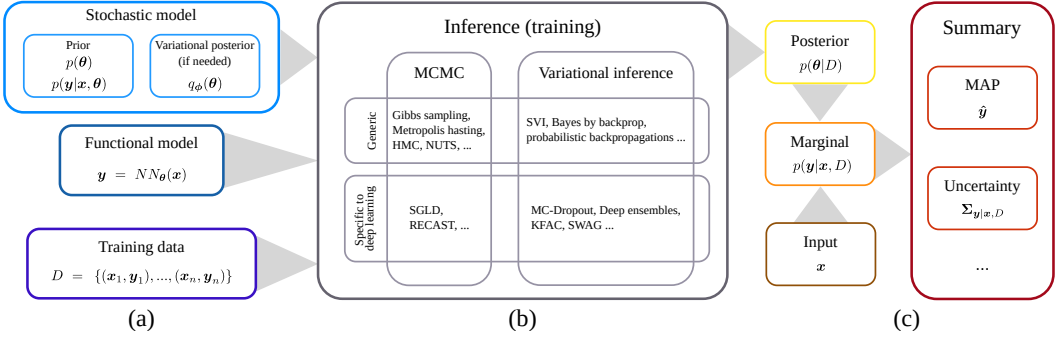


Fig. 4. Conventional workflow to design, train and use a Bayesian neural network. Design (a), includes the choice of artificial neural network architecture, but also the Stochastic model, including the prior and possibly a variational posterior family. Training (b), usually referred to as inference in statistics, implies different techniques described in Section 7 for the generic methods, and Section 8 for the methods specific to deep learning. Since inference outputs a posterior distribution over the model coefficients, one has to marginalize it to get the distribution of possible predictions, and summarize this marginal distribution to output the final predictions (c).

in more details in Section 7 for the generic methods and Section 8 for the methods specific to deep learning.

Given the Bayesian posterior, or its variational approximation, it becomes possible to compute a marginal probability distribution [93] of the output, given a certain input, which quantifies exactly the model's uncertainty:

$$p(y|x, D) = \int_{\Theta} p(y|x, \theta') p(\theta'|D) d\theta'. \quad (5)$$

In practice, the distribution  $p(y|x, D)$  is sampled indirectly using Equation 3.  $\theta$  is sampled from the variational distribution  $q_{\phi}(\theta)$  or uniformly in  $\Theta$ . The final prediction is summarized by a few statistics computed using a Monte-carlo approach (Fig. 4c).

To summarize the predictions of a BNN used to perform **regression**, the usual procedure is to perform model averaging [17]:

$$\hat{y} = \frac{1}{|\Theta|} \sum_{\theta_i \in \Theta} NN_{\theta_i}(x). \quad (6)$$

This approach is so common in ensemble learning that it is sometimes called ensembling. To quantify uncertainty, the covariance matrix can be computed as follows:

$$\Sigma_{y|x, D} = \frac{1}{|\Theta| - 1} \sum_{\theta_i \in \Theta} (NN_{\theta_i}(x) - \hat{y}) (NN_{\theta_i}(x) - \hat{y})^T. \quad (7)$$

When performing **classification**, the average model prediction will give the relative probability of each class, which can be considered as a measure of uncertainty in this case:

$$\hat{p} = \frac{1}{|\Theta|} \sum_{\theta_i \in \Theta} NN_{\theta_i}(x). \quad (8)$$

The final prediction (when the costs of giving a false positive are equal across all classes) is taken as the most likely class:

$$\hat{y} = \arg \max_i p_i \in \hat{p}. \quad (9)$$



If the cost of a false positive varies across different classes, it should be used to compute the risk and choose the minimal risk prediction.

### 3 MOTIVATION FOR BAYESIAN METHODS IN DEEP LEARNING

Defining a prior belief  $p(\theta)$  on the model parametrization (Section 2) is regarded by some users to be hard if not impossible. Defining a prior for a simple functional model is considered intuitive, e.g., explicitly adding a regularization term to favor a lower degree polynomial function or a smoother function [54]. However, defining priors is harder for the multi-layer models used in deep learning.

So, why do we bother to use Bayesian methods for deep learning given that it is hard to clearly comprehend deep neural networks behavior when defining the priors? The functional relationship encoded by an artificial neural network implicitly represents the conditional probability  $p(y|x, \theta)$ , and Bayes formula is an appropriate tool to use to invert conditional probabilities, even if one has a priori little insight about  $p(\theta)$ . While there are very strong theoretical principles and schema on which this Bayes formula can be based [76], we focus in this section on some practical benefits of using Bayesian Deep networks.

First, Bayesian methods provide a natural approach to **quantify uncertainty** in deep learning. Bayesian neural networks often have better calibration than classical neural networks [46, 58, 66], i.e., their predicted uncertainty is more consistent with the observed errors. In other words, they are neither overconfident nor underconfident compared to their non-Bayesian counterpart.

Working with a Bayesian neural network allows to distinguish between **epistemic uncertainty**, i.e., the uncertainty due to a lack of knowledge, measured by  $p(\theta|D)$ , which can be reduced with more data, and **aleatoric uncertainty**, i.e., the uncertainty due to the (partially) aleatoric nature of the data and measured by  $p(y|x, \theta)$  [14, 44]. This makes BNNs very data efficient, as they can learn from a small dataset without overfitting. At prediction time, out-of-training distribution points will just lead to high epistemic uncertainty. It also makes BNNs an interesting tool for active learning [19, 88], as one can interpret the model predictions and see if, for a given input, different probable parametrizations lead to different predictions. In this latter case, labelling this specific input will effectively reduce the epistemic uncertainty.

Furthermore, the No-free-lunch theorem for machine learning [94] can be interpreted as saying that any supervised learning algorithm includes some kind of implicit prior (while this interpretation is more philosophical than mathematical, and thus subject to discussion). Bayesian methods, when used correctly, will at least make the prior explicit. Now, if **integrating prior knowledge** seems hard with tools that are basically black boxes, it is not impossible. In Bayesian deep learning, priors are often considered as soft constraints, like regularization. Most regularization methods already used for point estimate neural networks can be understood from a Bayesian perspective as setting a prior, as demonstrated in Section 5.3. Moreover, previously learned posterior can be recycled as prior when new data becomes available. This makes Bayesian neural networks a valuable tool for online learning [64].

Last but not least, **the Bayesian paradigm enables the analysis of learning methods and draws links between them**. Some methods initially not presented as Bayesian can be **implicitly understood** as being approximate Bayesian, like regularization (Sec.5.3) or ensembling (Sec.8.2.2). This, in turn, supports the understanding of why certain methods that are easier to use than a strict application of the Bayesian algorithms can still give meaningful results from a Bayesian perspective. In fact, most Bayesian neural network architectures used in practice rely on methods that are approximately or implicitly Bayesian (Sec.8), because the exact algorithms are often too expensive. The Bayesian paradigm also provides a systematic framework to design new learning and regularization strategies, even for point-estimate models.

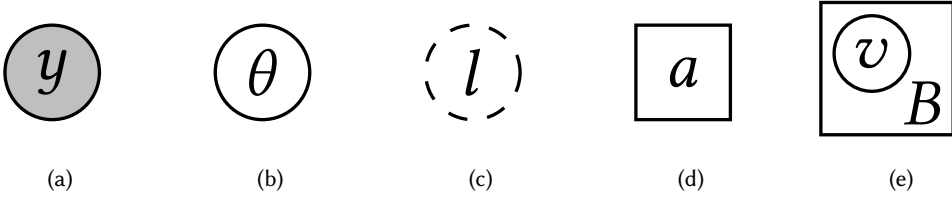


Fig. 5. PGM symbols, observed variables are in light gray circles (a), unobserved variables are in white circles (b), deterministic functions of other variables are in dashed circles (c) and parameters are in rectangles (d). Plates, represented as a rectangle around a part of the graph, indicate multiple independent instances of the framed subgraph as a batch  $B$  (e).

## 4 STOCHASTIC MODELS FOR BAYESIAN DEEP LEARNING

When designing a Bayesian neural network, one has to choose a deep network architecture, i.e., a functional model, but also a stochastic model, i.e., which variables are treated as stochastic variables and their a priori distribution. We will not cover the design of the functional model in this tutorial, as almost any model used for point estimate networks can be used in Bayesian deep learning, and a rich literature on the subject exists already [71]. Instead, in this section, we will focus on how to design the stochastic model.

We will present probabilistic graphical models (PGMs), a tool used to represent stochastic variables and their conditional relations, and specifically Bayesian belief networks, a class of PGMs used in Bayesian statistics. We will then show how to implement the stochastic model of a BNN from its PGM.

### 4.1 Probabilistic graphical models

Probabilistic graphical models (PGMs) are a tool that statisticians use to represent interdependence of multivariate stochastic variables and decompose their probability distributions accordingly using graphs. PGMs cover a large variety of models. In this tutorial, we will cover only Bayesian Networks, sometimes also called belief networks or Bayesian belief networks (BBN), which are PGMs represented using an acyclic directed graph. For a detailed review of the uses of PGMs as a representation of learning algorithms, the reader should refer to [9].

Even though both are acyclic directed graphs, one should not confuse BBNs with BNNs. BNNs represent a set of functional relations, such as the one shown in Equation (2), with an a priori distribution, while BBNs represent the inner structure of the joint probability distribution of the variables considered in the model. When conceiving a BNN, the corresponding BBN represents the base structure of the prior and, eventually, the variational posterior when using variational inference (see Section 7.2).

In a PGM, variables  $\mathbf{v}_i$  are the nodes in the graph, designated with different symbols to distinguish the nature of the considered variables (Fig. 5). A directed link, the only kind of link allowed in a BBN, means that the target variable probability distribution is defined conditioned on the source variable (but the converse is false, as the source variable probability distribution is not defined conditioned on the target variable). This allows the computation of the corresponding joint probability distribution of all  $\mathbf{v}_i$  in the graph:

$$p(\mathbf{v}_1, \dots, \mathbf{v}_n) = \prod_{i=1}^n p(\mathbf{v}_i | \text{parents}(\mathbf{v}_i)). \quad (10)$$

To complete the BBN, one has to define all the probability distributions  $p(\mathbf{v}_i | \text{parents}(\mathbf{v}_i))$ . The type of distribution used will depend on the context. Once the  $p(\mathbf{v}_i | \text{parents}(\mathbf{v}_i))$  are defined, the



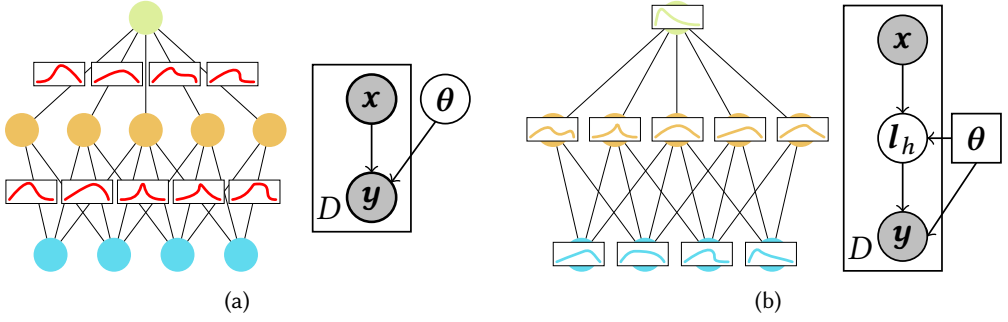


Fig. 6. The Bayesian belief networks corresponding to (a) coefficients as stochastic variables, which correspond to a conventional Bayesian regression and (b) activations as stochastic variables, which require accounting for the chained dependency.

BBN describes a data generation process. Parents are sampled before their children, which is always possible, as the graph is acyclic, and all the variables together represent a sample from the joint probability distribution  $p(\mathbf{v}_1, \dots, \mathbf{v}_n)$ .

Models usually learn from multiple examples sampled from the same distribution. To highlight this fact, the plate notation (Fig. 5e) has been introduced. A plate indicates that the variables  $(\mathbf{v}_1, \dots, \mathbf{v}_n)$  in the subgraph encapsulated by the plate are copied along a given batch dimension. A plate implies independence between all the duplicated nodes. This fact can be exploited to compute the joint probability of a batch  $B = \{(\mathbf{v}_1, \dots, \mathbf{v}_n)_b : b = 1, \dots, |B|\}$  as:

$$p(B) = \prod_{(\mathbf{v}_1, \dots, \mathbf{v}_n) \in B} p(\mathbf{v}_1, \dots, \mathbf{v}_n). \quad (11)$$

In a PGM, one distinguishes between observed variables, depicted in gray circles (Fig. 5a), which are treated as the data, and unobserved, also called latent, variables in white circle (Fig. 5b), which are treated as the hypothesis. From the joint probability derived from the PGM, defining the posterior for the latent variables given the observed variables is straightforward using Bayes formula:

$$p(\mathbf{v}_{latent} | \mathbf{v}_{obs}) = \frac{p(\mathbf{v}_{obs}, \mathbf{v}_{latent})}{\int_{\mathbf{v}_{latent}} p(\mathbf{v}_{obs}, \mathbf{v}_{latent}) d\mathbf{v}_{latent}} \propto p(\mathbf{v}_{obs}, \mathbf{v}_{latent}). \quad (12)$$

#### 4.2 Defining the stochastic model of a BNN from a PGM

Consider the two models presented in Fig. 6, with both the BNN and the corresponding BBN depicted. The BBN for the stochastic weights case (Fig. 6a) could represent the following data generation process, assuming the neural network is meant to do regression:

$$\begin{aligned} \theta &\sim p(\theta) = \mathcal{N}(\mu, \Sigma), \\ \mathbf{y} &\sim p(\mathbf{y} | \mathbf{x}, \theta) = \mathcal{N}(\text{NN}_\theta(\mathbf{x}), \Sigma). \end{aligned} \quad (13)$$

The choice of normal laws  $\mathcal{N}(\mu, \Sigma)$  is purely arbitrary, but common in practice.

If the neural network is meant to do classification then the model would have a categorical law  $\text{Cat}(p_i)$  to sample the prediction instead of a normal distribution:

$$\begin{aligned} \theta &\sim p(\theta) = \mathcal{N}(\mu, \Sigma), \\ \mathbf{y} &\sim p(\mathbf{y} | \mathbf{x}, \theta) = \text{Cat}(\text{NN}_\theta(\mathbf{x})). \end{aligned} \quad (14)$$

Then, one can use the fact that multiple data points from the training set are independent, which is indicated by the plate notation in Fig. 6, to write the probability of the training set as:

$$p(D_{\mathbf{y}}|D_{\mathbf{x}}, \theta) = \prod_{(\mathbf{x}, \mathbf{y}) \in D} p(\mathbf{y}|\mathbf{x}, \theta). \quad (15)$$

In the case of stochastic activations (Fig. 6b), the data generation process might become:

$$\begin{aligned} \mathbf{l}_0 &= \mathbf{x}, \\ \mathbf{l}_i &\sim p(\mathbf{l}_i|\mathbf{l}_{i-1}) = nl_i(\mathcal{N}(\mathbf{W}_i \mathbf{l}_{i-1} + \mathbf{b}_i, \Sigma)) \quad \forall i \in [1, n], \\ \mathbf{y} &= \mathbf{l}_n. \end{aligned} \quad (16)$$

The formulation of the joint probability for Bayes formula is slightly more complex, as we have to account for the chained dependency spanned by the BBN over the multiple latent variables  $\mathbf{l}_{[1, n-1]}$ :

$$p(D_{\mathbf{y}}, \mathbf{l}_{[1, n-1]}|D_{\mathbf{x}}) = \prod_{(\mathbf{l}_0, \mathbf{l}_n) \in D} \left( \prod_{i=1}^n p(\mathbf{l}_i|\mathbf{l}_{i-1}) \right). \quad (17)$$

It is sometimes possible, and often desirable, to define  $p(\mathbf{l}_i|\mathbf{l}_{i-1})$  such that the BBN described in Fig. 6a and the one described in Fig. 6b can be considered equivalent, e.g., sampling  $\mathbf{l}$  as:

$$\begin{aligned} \mathbf{W} &\sim \mathcal{N}(\mu_{\mathbf{W}}, \Sigma_{\mathbf{W}}), \\ \mathbf{b} &\sim \mathcal{N}(\mu_{\mathbf{b}}, \Sigma_{\mathbf{b}}), \\ \mathbf{l} &= nl(\mathbf{W} \mathbf{l}_{-1} + \mathbf{b}) \end{aligned} \quad (18)$$

is equivalent to sampling  $\mathbf{l}$  as:

$$\mathbf{l} \sim nl(\mathcal{N}(\mu_{\mathbf{W}} \mathbf{l}_{-1} + \mu_{\mathbf{b}}, (\mathbf{I} \otimes \mathbf{l}_{-1})^T \Sigma_{\mathbf{W}} (\mathbf{I} \otimes \mathbf{l}_{-1}) + \Sigma_{\mathbf{b}})), \quad (19)$$

where  $\otimes$  denotes a Kronecker product.

The basic Bayesian regression architecture depicted in Fig. 6a is more common in practice. The alternative formulation, depicted in Fig. 6b, is sometimes used as it allows to compress the number of optimisation parameters when using variational inference [92]. This provides different options when defining the prior.

## 5 SETTING THE PRIORS

**Setting the prior of small, causal, probabilistic models is very intuitive.** The same cannot be said about Deep Neural Networks, for which setting a good prior is often a tedious and unintuitive task. The main problem is that it is not really explicit how models with a very large number of parameters and a nontrivial architecture like ANNs will generalize for a given parametrization [98].

In this section, we present the common practice, and associated issues, related to the statistical unidentifiability of ANNs. Then, in Section 5.3, **we present how the prior in Bayesian Deep Learning can be related to regularization for the point estimate algorithms.** From there, we show how conventional regularization methods can give us some insight on how to choose better priors and, conversely, how a Bayesian analysis can help design new objective functions.

### 5.1 A good default

**For basic architectures, such as Bayesian Regression with an ANN (Fig. 6a), a standard procedure is to use a normal prior with mean 0 and diagonal covariance  $\sigma \mathbf{I}$  on the coefficients of the network:**

$$p(\theta) = \mathcal{N}(\mathbf{0}, \sigma \mathbf{I}). \quad (20)$$

This approach is equivalent to a weighted  $\ell_2$  regularization with weights  $1/\sigma$  when training a point estimate network, as we demonstrate in Section 5.3. The documentation of the probabilistic

programming language Stan [10] provided some examples [21] on how to choose  $\sigma$ , knowing the expected scale of the considered parameters.

Yet if this approach is often used in practice, there is no theoretical argument that makes it better than any other formulation [80]. The normal law is favored for its mathematical properties and the simple formulation of its log, as the log of the probability distribution is used in most learning algorithms.

## 5.2 Addressing unidentifiability in Bayesian neural networks

One of the main problems with Bayesian deep learning is that deep neural networks are over-parametrized models (i.e., with many equivalent parametrizations) [59]. This is referred to as statistical unidentifiability (i.e., inference does not lead to a unique answer). This can lead to complex multimodal posteriors that are hard to sample and approximate when training a BNN. There are two solutions to deal with such problems, changing the functional model parametrization or constrain the support of the prior to remove unidentifiability.

The two most common classes of non-uniqueness in ANNs that one might want to address are weight-space symmetry and scaling symmetry. Both are not a concern for point estimate neural networks but might be for BNNs.

**Weight-space symmetry** implies that one can build an equivalent parametrization of an ANN with at least one hidden layer by permuting two rows in the weights  $\mathbf{W}_i$  (respectively the bias  $\mathbf{b}_i$ ) of one of the hidden layers and the corresponding columns in the following layer weights matrix  $\mathbf{W}_{i+1}$ . This means that as the number of hidden layers and units in the hidden layers grows, the number of equivalent representations, which would roughly correspond to modes in the posterior distribution, grows factorially. A mitigation strategy is to enforce the bias vector in each layer to be sorted in ascending or descending order. However the practical effects of doing this are unknown, and it is possible that weight-space symmetry may implicitly support the exploration of the parameter space during the early stages of the optimisation.

**Scaling symmetry** is an unidentifiability problem arising when using non-linearities with the property  $nl(ax) = anl(x)$ , which is the case of RELU and Leaky-RELU, two popular non-linearities in modern machine learning. In this case assigning layers  $l$  and  $l+1$  weights  $\mathbf{W}_l, \mathbf{W}_{l+1}$  becomes strictly equivalent to assigning  $\alpha\mathbf{W}_l, 1/\alpha\mathbf{W}_{l+1}$ . This can reduce convergence speed for point estimate neural networks, a problem that is addressed in practice with various activation normalization techniques [1]. For BNN, this is slightly more complex, as the scaling symmetry influences the posterior, making it harder to approximate its shape. Some authors have proposed to use Givens transformations (sometimes also called Givens rotations elsewhere in the literature) to constrain the norm of the hidden layers [70] to address the scaling symmetry issue. In practice, using a Gaussian prior already reduces the scaling symmetry problem, as it will favor weights with the same Frobenius norm on each layer. A soft version of activation normalization can also be implemented by using a consistency condition, as explained in Section 5.4. The additional complexity of sampling the network parameters in a constrained space is not worth it from a computational complexity point of view.

## 5.3 The link between regularization and priors

The usual learning procedure for a point-estimate neural network is to find the set of parameters  $\theta$  which minimize some loss function, built using the data in the training set:

$$\hat{\theta} = \arg \min_{\theta} \text{loss}_{D_x, D_y}(\theta). \quad (21)$$

Assuming that the loss is minus the log-likelihood function (which is always the case, up to an additive constant), the problem can be rewritten as:

$$\hat{\theta} = \arg \max_{\theta} p(D_{\mathbf{y}}|D_{\mathbf{x}}, \theta), \quad (22)$$

which would be the first half of the model according to the Bayesian paradigm. Now assume we also have a prior for  $\theta$ , and we want to find the most likely point estimate from the posterior. The problem then become:

$$\hat{\theta} = \arg \max_{\theta} p(D_{\mathbf{y}}|D_{\mathbf{x}}, \theta)p(\theta). \quad (23)$$

Next, as it is easier to optimize, one would go back to a log-likelihood formulation:

$$\hat{\theta} = \arg \min_{\theta} \text{loss}_{D_{\mathbf{x}}, D_{\mathbf{y}}}(\theta) + \text{reg}(\theta). \quad (24)$$

If this formulation seems familiar, it is not a surprise. This is usually how regularization is applied in machine learning and in many other fields. The implication here is that, as prior, we have:

$$p(\theta) \propto e^{-\text{reg}(\theta) + \text{cst}}. \quad (25)$$

For some of the regularisers that are used in practice, this might be an ill-posed distribution, but the general idea is there. Another argument, less formal, is that regularization acts as a soft constraint on the search space, the same as what a prior does for a posterior.

#### 5.4 Prior with a consistency condition

Using the formulation in Equation 25, it becomes possible in certain cases to use the expected behavior of the functional model to extend the prior. To do so, one usually defines a consistency condition  $C(\theta, \mathbf{x})$  to evaluate the relative log-likelihood of a prediction given the input  $\mathbf{x}$  and parameter set  $\theta$ . For example,  $C$  can be set to favor sparse or regular predictions, to encourage monotonicity of predictions with respect to some input variables (e.g., probability of getting the flu increases with age), or to favor decision boundaries in low density regions when doing semi-supervised learning (Sec.6.1).  $C$  should be averaged over all possible inputs:

$$C(\theta) = \int_{\mathbf{x}} C(\theta, \mathbf{x})p(\mathbf{x})d\mathbf{x}. \quad (26)$$

In practice, as  $p(\mathbf{x})$  is unknown,  $C(\theta)$  is approximated from the features in the training set:

$$C(\theta) \approx \frac{1}{|D_{\mathbf{x}}|} \sum_{\mathbf{x} \in |D_{\mathbf{x}}|} C(\theta, \mathbf{x}). \quad (27)$$

We can now write a function proportional to the prior with the consistency condition included:

$$p(\theta|D_{\mathbf{x}}) \propto p(\theta) \exp\left(-\frac{1}{|D_{\mathbf{x}}|} \sum_{\mathbf{x} \in |D_{\mathbf{x}}|} C(\theta, \mathbf{x})\right), \quad (28)$$

where  $p(\theta)$  is the prior without the consistency condition.

## 6 DEGREE OF SUPERVISION AND ALTERNATIVE FORMS OF PRIOR KNOWLEDGE

The architecture presented so far focused mainly on the use of Bayesian neural networks in a supervised learning setting. However in real world applications, getting ground truth labels can be expensive, and new learning strategies should be adopted [72]. We now present how to adapt Bayesian neural networks for different degrees of supervision. While doing so, we also demonstrate



Fig. 7. Bayesian belief networks corresponding to (a) learning with noisy labels and (b) semi-supervised learning.

how PGMs in general and Bayesian Belief networks in particular are useful to design or interpret learning strategies.

In particular, the formulation of the Bayesian posterior, which is derived from the different PGMs presented below (Figs. 8,9, and 10), can be used (Sec.5.3) to obtain a suitable loss function for a maximum a posteriori estimator in the case where a point estimate neural network is sufficient for the considered use cases.

### 6.1 Noisy labels and semi-supervised learning

The inputs in the training sets can be uncertain, either because the labels  $D_y$  are corrupted by noise [61], or because a number of points are unlabelled, i.e., the setting for a semi-supervised learning approach.

In the case of noisy labels, one should extend the BBN to add a new variable for the noisy labels  $\tilde{y}$  conditioned on  $y$  (Fig. 7a). It is common, as the noise level itself is often unknown, to add a variable  $\sigma$  to characterize the noise. Frenay and Verleysen [16] proposed a taxonomy of the different approaches to integrate  $\sigma$  in a PGM (Fig. 8), where they distinguish three cases: the noise completely at random (NCAR); noise at random (NAR); and noise not at random (NNAR) models. In the NCAR model,  $\sigma$  is independent of any other variables, by definition the noise is then homoscedastic. In the NAR model,  $\sigma$  is dependent on the true label  $y$  but still independent of the features, while the NNAC models also account for the influence of the features  $x$ , e.g., if the level of noise in an image increases the chances that the image has been mislabeled. Both NAR and NNAC models represent heteroscedastic (i.e., the antonym of homoscedastic) noise.

Those models are slightly more complex than a pure supervised BBN as presented in Section 4 but can be treated in a similar fashion, by deriving the formula for the posterior from the PGM (Eq.12) and applying the chosen inference algorithm. We present here the procedure for a NNAR model, the most general one. The posterior becomes:

$$p(\mathbf{y}, \sigma, \theta | D) \propto p(D_{\tilde{y}} | \mathbf{y}, \sigma) p(\sigma | D_x, \mathbf{y}) p(\mathbf{y} | D_x, \theta) p(\theta). \quad (29)$$

During the prediction phase, for each tuple  $(\mathbf{y}, \sigma, \theta)$  sampled from the posterior,  $\mathbf{y}$  and  $\sigma$  can just be disregarded.

In the case of partially labelled data, also known as semi-supervised learning, (Fig. 7b), the dataset  $D$  is split into labelled  $L$  and unlabelled  $U$  examples. In theory, this PGM can be considered equivalent to the one used in the supervised learning case depicted in Fig. 6a, but in this case the unobserved data  $U$  would bring no information. The additional information of unlabelled data comes from the prior and only the prior. In traditional machine learning, the most common approaches to

implement semi-supervised learning are either to use some kind of data-driven regularization [86] or to rely on pseudo labels [82], Bayesian learning is no different.

**Data-driven regularization** implies modifying the prior assumptions, thus the stochastic model, to be able to extract meaningful information from the unlabeled dataset  $U$ . Two common ways to approach this process exist.

The first way is to condition the prior distribution of the model parameters on the unlabeled examples to favor certain properties of the model, such as a decision boundary in a low density region, i.e., using a distribution  $p(\theta|U)$  instead of  $p(\theta)$ . This implies writing the stochastic model as:

$$p(\theta|D) \propto p(L_y|L_x, \theta)p(\theta|U), \quad (30)$$

where  $p(\theta|U)$  is a prior with a consistency condition, as defined in Equation 28.

The second way is to assume some kind of dependency across the observed and unobserved labels in the dataset. This type of Bayesian semi-supervised learning relies on undirected PGM [96] to build the prior, or a least not assuming independence [48] between different training pairs  $(\mathbf{x}, \mathbf{y})$ . To keep things simple, we represent this fact by dropping the plate around  $\mathbf{y}$  in Fig. 7b. The posterior is thus written in the usual way (Eq.4), the main difference is that now  $p(D_y|D_x, \theta)$  is chosen to enforce some kind of consistency across the dataset. For example, it could be defined by assuming that two points close together (according to a certain notion of closeness which depends on the input space) are likely to have similar labels  $\mathbf{y}$  with a level of uncertainty that increases with the distance.

Both approaches have a similar effect and the choice of one over the other will depend on the mathematical formulation one favors to build the model.

The semi-supervised learning strategy can also be reformulated as having a weak predictor that is able to give some **pseudo labels**  $\tilde{\mathbf{y}}$ , sometimes with some confidence level. Many of the algorithms that are used for semi-supervised learning use an initial version of the model, trained with the labeled examples [51], to generate the pseudo labels  $\tilde{\mathbf{y}}$  and train the final model with those labels. This is problematic for Bayesian neural networks, as if the prediction uncertainty is accounted for, then, it becomes impossible to reduce the uncertainty associated with the unlabelled data, at least not without an additional hypothesis in the prior. Using a simpler model [53] to get the pseudo labels, even if less current in practice, can help mitigate that problem.

## 6.2 Data augmentation

Data augmentation is a strategy to significantly increase the diversity of data available to train deep models, without actually collecting new data. It relies on transformations which act on the input

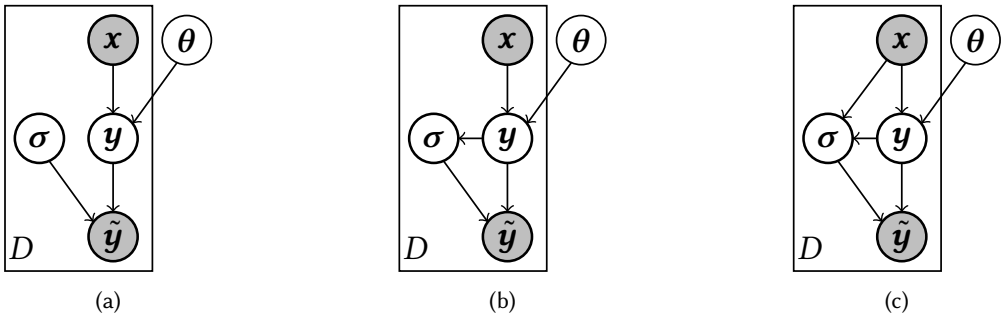


Fig. 8. Bayesian belief networks corresponding to (a) the noise completely at random (NCAR), (b) noise at random (NAR) and (c) noise not at random (NNAR) models.



without changing the label to generate an augmented dataset  $A(D)$ , e.g., applying rotations, flipping or adding noise in the case of images. Data augmentation is now at the forefront of state-of-the-art techniques in image processing [82] and increasingly in natural language processing [3].

From a Bayesian perspective, the additional information is brought by the knowledge of the augmentation process, rather than actual additional data.  $A(D)$  could contain an infinite set of possible variants of the initial dataset, e.g., when using continuous transforms such as rotations or additional noise. In practice,  $A(D)$  is sampled on the fly during training, rather than caching all possible augmentations in the training set in advance. This process is straightforward when training point estimate neural networks, but there are some subtleties when applying it with Bayesian statistics. The main concern is that the posterior of interest is  $p(\phi|D, Aug)$ , where  $Aug$  represents some knowledge about augmentation, not  $p(\phi|A(D), D)$ . This means, from a Bayesian perspective, that data augmentation should not, or at least not only, be treated as additional data but instead as an implicit transformation of the model, stated otherwise, as a prior. Moreover, the notion of data augmentation in the traditional statistical context was instead concerned with modelling missing values [84], a related but different problem.

The effect of adding more data points or counting similar data points multiple times will be to concentrate the posterior more around the MAP point-estimate, i.e., assuming a model with a given likelihood  $p(y|x, \theta)$ , counting the same data  $n$  times is equivalent to changing the likelihood to:

$$p^{aug}(y|x, \theta) \propto p(y|x, \theta)^n, \quad (31)$$

which is not incorrect, from a Bayesian point of view, as it can be done by modifying the likelihood in the original stochastic model. However, this poses a problem to account for epistemic uncertainty correctly when the size of the augmented dataset becomes infinite.

It can be argued instead that data augmentation should be implemented in the stochastic model. The idea is that if one is given data  $D$ , then one could also have been given data  $D'$ , where each element in  $D$  is replaced by an augmentation. Then  $D'$  is a different perspective of the data  $D$ . To model this, we have an augmentation distribution  $p(x'|x, Aug)$  that augments the observed data  $x$  using the augmentation model  $Aug$  to generate (probabilistically)  $x'$ . This  $x'$  is data in the vicinity of  $x$  (Fig. 9).  $x'$  can then be marginalized to simplify the stochastic model. The posterior is then given by:

$$p(\theta|D, Aug) = \int_{x'} p(\theta, x'|D, Aug) dx' \propto \left( \int_{x'} p(y|x', \theta) p(x'|x, Aug) dx' \right) p(\theta). \quad (32)$$

By setting:

$$p(y|x, \theta, Aug) = \int_{x'} p(y|x', \theta) p(x'|x, Aug) dx' = E_{x' \sim p(x'|x, Aug)} [p(y|x', \theta)], \quad (33)$$

we can define the augmented Bayesian posterior as:

$$p(\theta|D, Aug) \propto p(D_y|D_x, \theta, Aug) p(\theta), \quad (34)$$

This is a probabilistic counterpart to vicinal risk [12]. Using this form of the augmented likelihood prevents the multiple counting which would occur with the naive approach above.

In practice, this means we can perform the integral in Equation (33) using Monte Carlo, so we sample a small set of augmentations  $A_x$  according to  $p(x'|x, Aug)$  and average:

$$p(y|x, \theta, Aug) \approx \frac{1}{|A_x|} \sum_{x' \in A_x} p(y|x', \theta). \quad (35)$$

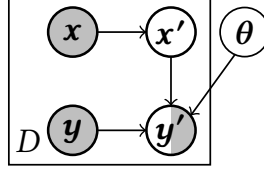


Fig. 9. Bayesian belief network corresponding to a data augmentation pipeline by sampling intermediate data  $x'$ .

The relevant cost function to use in training then becomes:

$$-\log p(D_y | D_x, \theta, Aug) \approx - \sum_{(x,y) \in D} \log \left( \frac{1}{|A_x|} \sum_{x' \in A_x} p(y|x', \theta) \right). \quad (36)$$

$A_x$  can contain as few as a single element, as long as it is re-sampled for each optimization iteration. This greatly simplifies Equation (36), especially when distributions from the exponential family are used to build the stochastic model.

An extension of this approach works in the context of semi-supervised learning where one adds a training cost to encourage consistency of predictions under augmentation [82, 95], where unlabeled data is used to build the samples for the consistency term. Note that this does not add labelling to the unlabeled examples. It adds a term to encourage consistency between the labels for an unlabeled data and its augmentation.

### 6.3 Meta learning, transfer learning, and self-supervised learning

**Meta learning** [33], in the broadest sense, is the use of machine learning algorithms **to assist in the training and optimization of other machine learning models**. The meta-knowledge acquired by meta learning can be distinguished from standard knowledge in the sense that it is applicable over a set of related tasks  $T$  rather than a single task.

**Transfer learning** designates methods where some **intermediate knowledge acquired** on a given problem **is reused to address a different problem**. In deep learning, it is used mostly for **domain adaptation**, when labelled data are available, abundant in a domain in some way similar to the domain of interest and scarce in the domain of interest [67]. Alternatively, some pre-trained models [73] are also a solution used to study architectures so large that training from scratch multiple times becomes inconvenient in practice.

**Self-supervised learning** is a learning strategy where the **data itself provides the labels** [36]. Since the labels obtainable by the data directly do not match the task of interest, the problem is approached as meta learning with a pretext (or proxy) task in addition to the task of interest. The use of the products of self-supervision is now generally regarded as an essential step in some areas. For instance, in natural language processing, most state of the art methods use these pre-trained models [73].

**In our Bayesian understanding of meta learning, derived from the broad definition above, we consider both transfer learning and self-supervised learning to be special cases of meta learning.**

A common approach for meta learning in Bayesian statistics is to recast the problem as hierarchical Bayes [25], where the prior for each task  $p(\theta_t | \xi)$  is conditioned on a new global variable  $\xi$  (Fig. 10a).  $\xi$  can represent some continuous meta-parameters (the focus of this tutorial) or discrete information about the structure of the BNN (the case of learning probable functional models) or the underlying subgraph of the PGM (the case of learning probable stochastic models). Multiple levels can be added to organise the tasks in a more complex hierarchy if need be, but we will present only



Fig. 10. Bayesian belief networks corresponding (a) general supervised hierarchical Bayes and (b) the joint self-supervised learning strategy.  $l$  is an intermediate activation layer in the network defined by design,  $\theta_s$  represents the shared parameters,  $\theta_p$  the parameters specific to the pretext task,  $\theta_t$  the parameters specific to the main task,  $y_p$  the labels obtained from the data and  $y_t$  the labels for the main task (some of which might be unobserved [4])

the case with one level as generalizing is straightforward. The general posterior becomes:

$$p(\theta, \xi | D) \propto \left( \prod_{t \in T} p(D_y^t | D_x^t, \theta_t) p(\theta_t | \xi) \right) p(\xi). \quad (37)$$

In practice thus, the problem is often approached with empirical Bayes (Sec.7.4), and only a point estimate  $\hat{\xi}$  is considered for the global variable, ideally the MAP estimate obtained by marginalizing  $p(\theta, \xi | D)$  and selecting the most likely point, but this is not always the case.

In transfer learning, the usual approach would be to set  $\hat{\xi} = \theta_m$ , with  $\theta_m$  the coefficients of the main task. The new prior can then be obtained from  $\hat{\xi}$ , for example:

$$p(\theta | \xi) = \mathcal{N}(\text{sel}(\xi), 0, \sigma I), \quad (38)$$

where  $\text{sel}$  is a selection function for the parameters to transfer and  $\sigma$  is a parameter to tune manually. Unselected parameters are assigned a mean of 0 by convention but other methods to design those parts of the priors can be used. If a BNN has been trained for the main task  $\sigma$  can be estimated on the previous posterior but it will still be required to slightly scale it up to account for the additional uncertainty.

Self-supervised learning can be implemented in two steps, first learning the pretext task and then use transfer learning. This can be considered overly complex, but might be required if the pretext task has a high computational complexity (e.g., BERT models in natural language processing [73]). Recent contributions have shown that jointly learning the pretext task and the final task (Fig. 10b) can improve the results obtained in self-supervised learning [4]. This approach, which is closer to hierarchical Bayes, also allows setting the prior a single time while still retaining the benefits of self-supervised learning.

## 7 BAYESIAN INFERENCE ALGORITHMS

A priori, one does not have to undergo a learning phase when using a BNN, just sample the posterior and do model averaging (see Equations (4) and (6)). But sampling the posterior is not easy in the general case. If the conditional probability of the data  $P(D|H)$  and the probability of the model  $P(H)$  are given by our prior and our model, the integral for the evidence term  $\int_H P(D|H')P(H')dH'$  might be excessively difficult to compute. For non-trivial models, and even if the evidence has been computed, it is really hard to sample the posterior directly due to the high dimensionality

of the sampling space and the non-trivial transformation from the samples of a uniform random variable. Instead of using traditional methods to sample the posterior, such as inversion sampling or rejection sampling, dedicated algorithms are used. The most popular ones are Markov Chain Monte Carlo methods, a family of algorithms to exactly sample the posterior, or variational inference, a method for learning an approximation of the posterior (Fig. 4). This section reviews these methods.

### 7.1 Markov Chain Monte Carlo

The idea of Markov Chain Monte Carlo methods is to construct a Markov chain, a sequence of random samples  $S_i$ , which probabilistically depend only on the previous sample  $S_{i-1}$ , such that the elements of the sequence eventually are distributed following a desired distribution. Unlike standard, simple, low-dimensional sampling methods, like rejection or inversion sampling, most MCMC algorithms require some initial burn-in time before the underlying Markov chain converges to the desired distribution. Also, the successive  $S_i$  might be autocorrelated. This means that a large set of samples  $\Theta$  has to be generated and subsampled to get approximately independent samples from the underlying distribution. Moreover a certain number of initial samples has to be discarded at the beginning of the sequence, and the exact amount is not always easy to define. Last but not least, the final collection of samples  $\Theta$  has to be cached after training, which can be very expensive even for average size deep learning models.

Despite their inherent drawbacks, MCMC methods can be considered among the best available and the most popular solutions for sampling from exact posterior distributions in Bayesian statistics [2]. However not all MCMC algorithms are relevant for Bayesian deep learning. Gibbs sampling [22], for example, is very popular in general statistics and unsupervised machine learning, but is rarely used for BNNs. The class of MCMC methods which are the most relevant for Bayesian neural networks are the Metropolis-Hastings algorithms [13]. The property which makes Metropolis-Hastings algorithms popular is that they do not require knowledge about the exact probability distribution  $P(\mathbf{x})$  to sample from. Instead, a function  $f(\mathbf{x})$  that is proportional to that distribution is sufficient. This is the case of a Bayesian posterior distribution, which is usually quite easy to compute except for the evidence term.

The basic idea of the Metropolis-Hastings algorithm is to start with a random initial guess,  $\mathbf{x}_0$ , and then sample a new candidate point “around” the previous one. If this candidate is more likely than the previous one (according to the distribution we want to sample from), then it is accepted. If it is less likely, then it is accepted with a certain probability, rejected otherwise.

More formally, the algorithm, see Algorithm 1, is constructed with a proposal distribution  $Q(\mathbf{x}'|\mathbf{x})$ , which tells us how to sample “around” the previous sample.

---

#### Algorithm 1 Metropolis-Hastings

---

```

Draw  $\mathbf{x}_0 \sim \text{Initial}$ 
while  $n = 0$  to  $N$  do
  Draw  $\mathbf{x}' \sim Q(\mathbf{x}|\mathbf{x}_n)$ 
   $p = \min\left(1, \frac{Q(\mathbf{x}'|\mathbf{x}_n) f(\mathbf{x}')}{Q(\mathbf{x}_n|\mathbf{x}') f(\mathbf{x}_n)}\right)$ 
  Draw  $k \sim \text{Bernoulli}(p)$ 
  if  $k$  then
     $\mathbf{x}_{n+1} = \mathbf{x}'$ 
     $n = n + 1$ 
  end if
end while

```

---

The acceptance probability  $p$  of Algorithm 1 can be shown to yield the most accepting of all reversible Markov chains. Furthermore, the probability  $p$  can be simplified if  $Q$  is chosen to be symmetric, i.e.,  $Q(\mathbf{x}'|\mathbf{x}_n) = Q(\mathbf{x}_n|\mathbf{x}')$ , and the formula for the acceptance rate becomes:

$$p = \min \left( 1, \frac{f(\mathbf{x}')}{f(\mathbf{x}_n)} \right). \quad (39)$$

In this situation, the algorithm is simply called the Metropolis method. Common choices for  $Q$  can be a normal distribution  $Q(\mathbf{x}'|\mathbf{x}_n) = \mathcal{N}(\mathbf{x}_n, \sigma^2)$ , or a uniform distribution  $Q(\mathbf{x}'|\mathbf{x}_n) = \mathcal{U}(\mathbf{x}_n - \varepsilon, \mathbf{x}_n + \varepsilon)$ , centered around the previous sample. But sometimes, one has to deal with non-symmetric proposal distribution, e.g., to accommodate a constraint in the model, like if the domain of the distribution under consideration is bounded. In that case, one has to take into account the correction term imposed by the full Metropolis-Hasting algorithm.

The spread of the proposal distribution has to be tweaked. If it is too large, the rejection rate will be too high. If it is too small the samples will be more auto-correlated. There is no general method to tweak those parameters. Yet a clever strategy to get the new proposed sample  $\mathbf{x}'$  can reduce the impact of such parameters. This is why the Hamiltonian Monte-Carlo method has been proposed.

The Hamiltonian Monte-Carlo algorithm [62] is another example of Metropolis-Hasting algorithm for continuous distributions designed with a clever scheme to draw a new proposal  $\mathbf{x}'$  to ensure that as few samples as possible are rejected and there is as few correlation as possible between samples. Moreover, the burn in time is extremely short. The process to generate a new jump is based on Hamiltonian mechanics. First, we assume that from the actual position  $\mathbf{x}_n$ , we will move with an initial random velocity  $\mathbf{v}$  from a proposal distribution  $Q(\mathbf{v})$ . Then, we define the Hamiltonian of the system as:

$$H(\mathbf{x}, \mathbf{v}) = \log(P(\mathbf{x})) + \log(Q(\mathbf{v})) = \log(f(\mathbf{x})) + cst + \log(Q(\mathbf{v})), \quad (40)$$

with  $\log(P(\mathbf{x}))$  being considered as the potential energy and  $\log(Q(\mathbf{v}))$  as the kinetic energy. We then let the system move for a given time  $T$ . The corresponding dynamical system is parametrized by the following PDE:

$$\begin{cases} \frac{\partial \mathbf{x}}{\partial t} = \frac{\partial H}{\partial \mathbf{v}} = \frac{\partial \log(Q(\mathbf{v}))}{\partial \mathbf{v}} \\ \frac{\partial \mathbf{v}}{\partial t} = -\frac{\partial H}{\partial \mathbf{x}} = -\frac{\partial \log(f(\mathbf{x}))}{\partial \mathbf{x}} \end{cases} \quad (41)$$

with initial conditions  $\mathbf{x}'_0 = \mathbf{x}_n$  and  $\mathbf{v}_0 = \mathbf{v}'$ . The advantage is that we still need to know the distribution  $P$  only up to a scaling factor. We accept the proposed sample  $\mathbf{x}'_T$  with probability  $p$  computed as:

$$p = \min \left( 1, \frac{\exp[H(\mathbf{x}'_T, \mathbf{v}_T)]}{\exp[H(\mathbf{x}'_0, \mathbf{v}_0)]} \right). \quad (42)$$

Now, since the Hamiltonian is preserved over time,  $p$  is supposed to be equal to 1, and the new sample is never rejected. The problem is that getting an exact solution is often hard, if not impossible, so one has to rely on numerical integration instead. To do so, it is required to use a symplectic integrator. This is an integrator that represents a discrete Hamiltonian system  $H'(\mathbf{x}, \mathbf{v})$ , ideally only slightly perturbed compared to the original continuous one  $H(\mathbf{x}, \mathbf{v})$ . This preserves important properties of the underlying Markov chain of the MCMC algorithm, the main one being that if the considered Hamiltonian path loops back on the starting point  $\mathbf{x}_0$ , the numerical integrator will also loop back on  $\mathbf{x}_0$ . Note that most numerical integrators, including the popular Runge-Kutta scheme,

are not symplectic. A good choice of symplectic integrator is the leapfrog, with timestep  $\Delta t$ :

$$\begin{cases} \mathbf{v}_{t+\Delta t/2} = \mathbf{v}_t - \frac{\Delta t}{2} \frac{\partial \log(f(\mathbf{x}'))}{\partial \mathbf{x}'} \\ \mathbf{x}'_{t+\Delta t} = \mathbf{x}'_t + \Delta t \frac{\partial \log(Q(\mathbf{v}_{t+\Delta t/2}))}{\partial \mathbf{v}} \\ \mathbf{v}_{t+\Delta t} = \mathbf{v}_{t+\Delta t/2} - \frac{\Delta t}{2} \frac{\partial \log(f(\mathbf{x}'_{t+\Delta t}))}{\partial \mathbf{x}'} \end{cases} \quad (43)$$

With this symplectic integrator, we have the following relation between the original Hamiltonian  $H(\mathbf{x}, \mathbf{v})$  and the modified Hamiltonian  $H'(\mathbf{x}, \mathbf{v})$ :

$$H(\mathbf{x}, \mathbf{v}) = H'(\mathbf{x}, \mathbf{v}) + O(\Delta t^2). \quad (44)$$

This means that it is easy to adaptively adjust the integration step  $\Delta t$  to tweak the acceptance probability  $p$ .

Now, the problem is to choose the proposal distribution  $Q(\mathbf{v})$ , and the integration time  $T$ .  $Q(\mathbf{v})$  is usually chosen to be a normal distribution  $\mathcal{N}(\mathbf{0}, \Sigma)$ . The most obvious choice for  $\Sigma$  is  $\sigma^2 \mathbf{I}$ .  $\sigma^2$  can be increased to augment the odds of sampling a new point far away from the previous samples. The choice of the integration time  $T$  is critical. If it is too short, then successive samples are at risk of being autocorrelated. If it is too large, then the Hamiltonian path might loop and a lot of time will be lost integrating the same things over and over again.

An improvement over the classic HMC algorithm to automatically adapt the integration time  $T$  has been proposed [31], called No-U-Turn sampler (NUTS for short), which most software packages for Bayesian statistics implement.

## 7.2 Variational inference

MCMC algorithms are the goto tools to sample from the exact posterior in Bayesian statistics. However their lack of scalability, even if somehow mitigable, has made them less popular for Bayesian Deep Learning, due to the huge sizes of models usually considered in deep learning. Variational inference [6], which scales better than MCMC algorithms, gained a lot of popularity.

Variational inference is not an exact method. Rather than allowing sampling from the exact posterior, the idea is to have a distribution  $q_\phi(H)$ , parametrized by a set of parameters  $\phi$ , called the variational distribution. The values of the parameters  $\phi$  is then inferred, or learned (to use a word closer to the machine learning jargon), such that the variational distribution  $q_\phi(H)$  is as close as possible to the exact posterior  $P(H|D)$ . The measure of closeness most readily used is the KL-divergence, which is a measure of closeness between probability distribution functions, and is a function of  $\phi$ :

$$D_{KL}(q_\phi||P) = \int_H q_\phi(H') \log \left( \frac{q_\phi(H')}{P(H'|D)} \right) dH'. \quad (45)$$

There is an apparent problem here, which is that to compute  $D_{KL}(q_\phi||P)$ , it looks like we need to compute  $P(H|D)$  anyway. Fortunately, this is not the case, and a different, easily derived formula called the evidence lower bound, or ELBO, is used instead:

$$ELBO = \int_H q_\phi(H') \log \left( \frac{P(H', D)}{q_\phi(H')} \right) dH' = \log(P(D)) - D_{KL}(q_\phi||P). \quad (46)$$

Since  $\log(P(D))$  only depends on the prior, minimizing  $D_{KL}(q_\phi||P)$  is equivalent to maximizing the ELBO.

Almost any large scale optimization method that is used in classic machine learning should be applicable to optimize the ELBO. The most popular one in neural networks is Stochastic variational inference [32], which is in fact stochastic gradient descent applied to variational inference. This



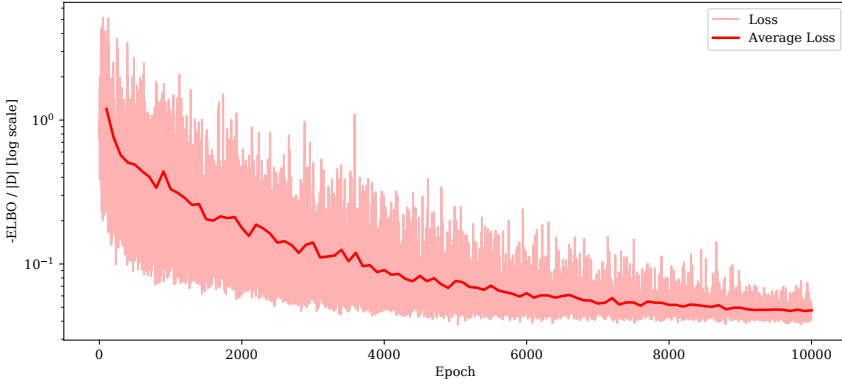


Fig. 11. Typical training curve for Bayes by backprop. Each sample looks random but the curve still has a downwards trend. The ELBO has been normalized in the graph for comparison purposes.

allows the algorithm to scale to the large datasets that are encountered in modern machine learning, since the ELBO can be computed on a single mini-batch at each iteration.

Note that convergence, when learning the posterior with variational inference, will be relatively slow compared to the usual backpropagation. Moreover, most implementations use a minimal number of samples to evaluate the ELBO, often just one, before taking a gradient step. This means that the ELBO as a function of the iteration will be quite noisy.

In traditional machine learning and statistics,  $q_\phi(H)$  is mostly constructed from distributions in the exponential family, e.g., multivariate normals [26], gammas and Dirichlets. The ELBO can then be dramatically simplified into components [23] leading to a generalization of the well known expectation-maximization algorithm. To account for correlations between the large number of parameters, certain approximations are made, for instance block diagonal covariance matrices can be used [75], or low rank plus diagonal [55].

### 7.3 Bayes by backpropagation

Variational inference offers a good mathematical tool for Bayesian inference, but it needs to be adapted to deep learning. The main problem is that stochasticity stops backpropagation from functioning for the internal nodes in a network [9]. Different solutions have been proposed to mitigate this problem, like probabilistic backpropagation [29] or Bayes-by-backprop [7]. Both methods serve the same goal, but as Bayes-by-backprop may look more familiar to deep learning practitioners, we will focus on it. Bayes-by-backprop is indeed a practical implementation of variational inference combined with a reparametrization trick [42] to ensure backpropagation works as usual.

The idea is to use some random variable  $\varepsilon \sim q(\varepsilon)$ , as a non-variational source of noise.  $\theta$  is not sampled directly but obtained via a deterministic transformation  $t(\varepsilon, \phi)$  such that  $\theta = t(\varepsilon, \phi)$  follows  $q_\phi(\theta)$ .  $\varepsilon$  is sampled and thus changes at each iteration but can still be considered as a constant with regard to other variables. All other transformations being non-stochastic, backpropagation works as usual for the variational parameters  $\phi$ . The general formula for the ELBO becomes:

$$ELBO = \int_{\varepsilon} q_{\phi}(t(\varepsilon, \phi)) \log \left( \frac{P(t(\varepsilon, \phi), D)}{q_{\phi}(t(\varepsilon, \phi))} \right) |Det(\nabla_{\varepsilon} t(\varepsilon, \phi))| d\varepsilon. \quad (47)$$

This is tedious to work with. Instead, to estimate the gradient of the ELBO, it has been proposed [7] to use the fact that if  $q_\phi(\theta)d\theta = q(\varepsilon)d\varepsilon$ , then, for a differentiable function  $f(\theta, \phi)$  we have:

$$\frac{\partial}{\partial \phi} \int_{\phi} q_\phi(\theta') f(\theta', \phi) d\theta' = \int_{\varepsilon} q(\varepsilon) \left( \frac{\partial f(\theta, \phi)}{\partial \theta} \frac{\partial \theta}{\partial \phi} + \frac{\partial f(\theta, \phi)}{\partial \phi} \right) d\varepsilon. \quad (48)$$

A proof is provided in [7]. We also provide an alternative proof to give more details on when we can assume  $q_\phi(\theta)d\theta = q(\varepsilon)d\varepsilon$ , in appendix A. A sufficient condition is for  $t(\varepsilon, \phi)$  to be invertible with respect to  $\varepsilon$  and the distributions  $q(\varepsilon)$  and  $q_\phi(\theta)$  to not be degenerated probability distributions.

For the case where the weights are treated as stochastic variables, and thus hypothesis  $H$ , which is the case considered in the original Bayes-by-backprop paper, the training loop can then be implemented as described in Algorithm 2.

---

**Algorithm 2** Bayes-by-backprop
 

---

```

 $\phi = \phi_0$ 
for  $i = 0$  to  $N$  do
  Draw  $\varepsilon \sim q(\varepsilon)$ 
   $\theta = t(\varepsilon, \phi)$ 
   $f(\theta, \phi) = \log(q_\phi(\theta)) - \log(p(D_y | D_x, \theta)p(\theta))$ 
   $\Delta_\phi f = \text{backprop}_{\phi}(f)$ 
   $\phi = \phi - \alpha \Delta_\phi f$ 
end for

```

---

The objective function  $f$  corresponds to an estimate of the ELBO from a single sample. This means that the gradient estimate will be noisy. The convergence graph will also be noisy, much more than in the case of classic backpropagation (Fig. 11). To get a better estimate of the convergence, one can average the loss over multiple epochs.

Algorithm 2 is very similar to the classic training loop for point estimate deep learning. So much in fact that many recent contributions to optimization for deep learning are straightforward to use for Bayes-by-backprop, e.g., to use the Adam optimizer [41] instead of stochastic gradient descent. Natural gradient descent has also been advocated as an optimization tool for variational inference for BNN [65], as the distributions from the exponential family lead to, if parametrized in natural space, a very simple formulation of the natural gradient [39].

Note also that, even if the original algorithm was presented for BNN with stochastic weights, it is straightforward to adapt it for BNN with stochastic activations. In that case, the activations  $\mathbf{l}$  represent the hypothesis  $H$  and the weights  $\theta$  are part of the variational parameters  $\phi$ .

#### 7.4 Learning the prior

Despite this being counter-intuitive, it is possible to learn the prior even if one learns the posterior afterwards. This is meaningful if there is some aspect of the prior that are set using prior knowledge, allowing to learn the remaining free parameters before getting the posterior. In standard Bayesian statistics, this is known as **empirical Bayes** and is usually a valid approximation when the dimensions of the prior parameters being learned are far less than the dimensions of the model parameters.

Given a parametrized prior distribution  $p_\xi(H)$ , maximizing the likelihood of the data is a good method to learn the parameters  $\xi$ :

$$\hat{\xi} = \arg \max_{\xi} P(D|\xi) = \arg \max_{\xi} \int_H p_\xi(D|H') p_\xi(H') dH'. \quad (49)$$

Finding  $\hat{\xi}$  directly in general is an intractable problem, but when using variational inference, there is a specific property of the ELBO that one could exploit [26]. Namely, that the ELBO is the log-likelihood of the data minus the KL-divergence of the variational posterior and prior (see Equation 46):

$$\log(P(D|\xi)) = ELBO + D_{KL}(q_\phi||P). \quad (50)$$

This means that maximizing the ELBO, now a function of both  $\xi$  and  $\phi$ , is equivalent to maximising a lower bound on the log-likelihood of the data, a bound which is tighter when  $q_\phi$  is from a general family so able to fit  $P$  well. The Bayes-by-backprop algorithm presented in Section 7.3 needs only to be slightly modified, see Algorithm 3.

---

**Algorithm 3** Bayes by backprop with parametric prior
 

---

```

 $\xi = \xi_0$ 
 $\phi = \phi_0$ 
for  $i = 0$  to  $N$  do
  Draw  $\varepsilon \sim q(\varepsilon)$ 
   $\theta = t(\varepsilon, \phi)$ 
   $f(\theta, \phi, \xi) = \log(q_\phi(\theta)) - \log(p_\xi(D_y|D_x, \theta)p_\xi(\theta))$ 
   $\Delta_\xi f = \text{backprop}_\xi(f)$ 
   $\Delta_\phi f = \text{backprop}_\phi(f)$ 
   $\xi = \xi - \alpha_\xi \Delta_\xi f$ 
   $\phi = \phi - \alpha_\phi \Delta_\phi f$ 
end for
  
```

---

## 8 ADAPTING BAYESIAN METHODS FOR DEEP LEARNING

We presented so far the fundamental theory to design and train Bayesian neural networks. However, the aforementioned methods are still not easily applicable to most large scale architectures that are used today in deep learning. Recent research has also shown that being only approximately Bayesian is enough to get a decent, correctly calibrated model with uncertainty estimates [46]. In this section, we present how Bayesian methods can be adapted specifically or approximated for deep learning, resulting in new, more efficient dedicated algorithms. We classified different approaches to speed up either training or predictions into two broad categories (Fig. 12): the inference algorithms and the other simplification and compression methods. Specific inference algorithms can then be further classified based on the source of stochasticity they rely on and whether they work like an MCMC algorithm (they generate a sequence of samples from the posterior) or can be considered as a form of variational inference (they learn the parameters of an intermediate distribution to approximate the posterior). The two main sources of stochasticity that can be exploited to drive those methods are noise layers, the most renown type being dropout, or the noise induced by stochastic gradient descent.

### 8.1 Bayes via Dropout

Dropout has initially been proposed as a regularization procedure used during training [83]. The procedure is to apply multiplicative noise to the previous layer. By far, the most used type of noise is Bernoulli noise, but other types of noise are sometimes used instead, like Gaussian Noise, in which case the procedure is called Gaussian Dropout [83].

Dropout can also be used during the evaluation phase as a form of ensemble learning. This offers the ability to obtain a distribution for the output predictions [18, 52]. This procedure, called

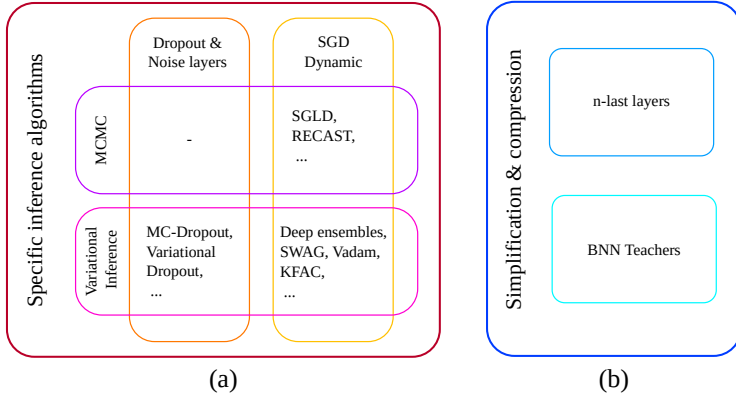


Fig. 12. Simplification algorithm for Bayesian neural networks can be broadly classified into (a) dedicated inference algorithms and (b) other simplifications and compression strategies. Dedicated inference algorithms can be further distinguished by, first if they work like a MCMC algorithm or variational inference and second by the source of stochasticity they rely on, i.e., either Dropout & other noise layer or Stochastic gradient descent dynamic.

**Monte Carlo Dropout**, can be considered as being variational inference, with a variational posterior distribution defined for each weight matrix as:

$$\begin{aligned} z_{i,j} &\sim \text{Bernoulli}(p_i) \\ \mathbf{W}_i &= \mathbf{M}_i \cdot \text{diag}(\mathbf{z}_i), \end{aligned} \quad (51)$$

with  $\mathbf{z}_i$  the random activation or inactivations coefficients and  $\mathbf{M}_i$  the weights matrix before dropout is applied.  $p_i$  is the activation probability for layer  $i$  and can be learned or set manually.

The equivalence between a standard objective function used for training with dropout and additional  $\ell^2$  weight regularization:

$$\mathcal{L}_{\text{dropout}} = \frac{1}{N} \sum_D f(\mathbf{y}, \hat{\mathbf{y}}) + \lambda \sum_{\theta} \theta_i^2, \quad (52)$$

and the ELBO for variational inference assuming a normal prior on the weights and the distribution presented in Equation 51 as variational posterior, has been demonstrated [18] using an argument similar to the one presented in Section 5.3.

**MC-Dropout is a very convenient technique to do Bayesian Deep learning, as it is straightforward to implement and requires little additional knowledge or modelling effort compared to traditional methods. Moreover, it often leads to a faster training phase compared to other variational inference approaches.**

On the other hand, MC-Dropout might lack some expressiveness and thus does not fully capture the uncertainty associated with the model predictions. It also lacks some flexibility compared to other methods when performing Bayesian online or active learning.

**A variant of MC-Dropout is Variational Dropout**, which attempts to apply Gaussian Dropout at evaluation time [43]. Variational Dropout has been criticized as not being Bayesian [34], mainly due to the fact that the authors have chosen an objective function where Gaussian Dropout is the only form of regularization used. In that case, the implied prior, i.e., log-uniform, corresponds to a degenerate probability distribution, which in turn leads to a degenerate posterior. It is important to stress that when using dropout as a Bayesian approximation, one has to account for the prior regularization on top of dropout during training.

## 8.2 Bayes via stochastic gradient descent

Stochastic Gradient Descent (SGD) and related algorithms are at the core of modern machine learning. The initial goal of SGD is to provide an algorithm to converge to a point estimate corresponding to an optimal solution while having only noisy estimates of the gradient of the objective function, especially when the training data has to be split into mini-batches. The parameter update rule at time  $t$  can be written as:

$$\Delta\theta_t = \frac{\epsilon_t}{2} \left( \frac{N}{n} \nabla \log(p(D_t, \theta_t)) + \nabla \log(p(\theta_t)) \right), \quad (53)$$

where  $D_t$  is a minibatch subsampled at time  $t$  from the complete dataset  $D$ ,  $\epsilon_t$  is the learning rate at time  $t$ ,  $N$  is the size of the whole dataset and  $n$  the size of the minibatch.

SGD, or related optimization algorithms like ADAM [41], can be reinterpreted as a Markov-chain algorithm [56], with the sampling of the minibatch acting as a source of stochasticity. Usually the hyperparameters of the algorithm are tweaked to ensure the chain converges to a Dirac distribution, the final point estimate indicator distribution, especially by reducing  $\epsilon_t$  towards 0 while ensuring  $\sum_{t=0}^{\infty} \epsilon_t = \infty$ . However, this does not have to be the case. If the learning rate is reduced toward a constant, the underlying Markov-chain will converge to a stationary distribution. If a Bayesian prior is accounted for in the objective function, then this stationary distribution can be an approximation of the corresponding posterior.

**8.2.1 MCMC algorithms based on SGD dynamic.** Based on this observation, a specific MCMC method has been developed from the SGD algorithm, called Stochastic Gradient Langevin Dynamic (SGLD) [91]. The idea to couple SGD with Langevin Dynamic leads to a slightly modified update step:

$$\begin{aligned} \Delta\theta_t &= \frac{\epsilon_t}{2} \left( \frac{N}{n} \nabla \log(p(D_t, \theta_t)) + \nabla \log(p(\theta_t)) \right) + \eta_t \\ \eta_t &\sim \mathcal{N}(0, \epsilon_t). \end{aligned} \quad (54)$$

It has been shown that this method leads to a Markov chain sampling the exact posterior if  $\epsilon_t$  goes toward 0 [91]. The problem is, if  $\epsilon_t$  goes toward 0, then the successive samples become more and more auto-correlated. To leverage this problem, the authors proposed to stop reducing  $\epsilon_t$  at some point, thus making the samples only an approximation of the exact posterior. Still, SGLD offers better theoretical guarantee if the dataset has to be split into minibatches compared to other MCMC methods, which makes the algorithm very useful in Bayesian deep learning.

To favor exploration of the posterior over exact sampling, SGD can also be used in a slightly different manner, that coarsely approximates the Bayesian posterior while being easy to implement in practice. The idea is to use warm restarts, i.e., restarting the gradient descent with a large learning rate  $\epsilon_0$  and possibly new random weights  $\theta_0$ .

This method is clearly a coarse approximation of a true Bayesian approach, but it offers multiple advantages. The main one is to avoid the mode collapse problem [49]. In the case of BNN, the true Bayesian posterior is usually a complex multimodal distribution, as multiple and sometimes not equivalent parametrizations  $\theta$  of the network can fit the training dataset. Favoring exploration over precise reconstruction can help to get a better picture of those different modes. Then, as  $\theta$  sampled from the same mode are more likely to make the model generalize in a similar manner, using warm restarts will enable a much better estimate of the model uncertainty when processing unseen data, even if this approach is not perfectly Bayesian.

Algorithms exploiting this idea have been presented in the literature, e.g., RECAST [78], a modification of SGLD with warm restarts.

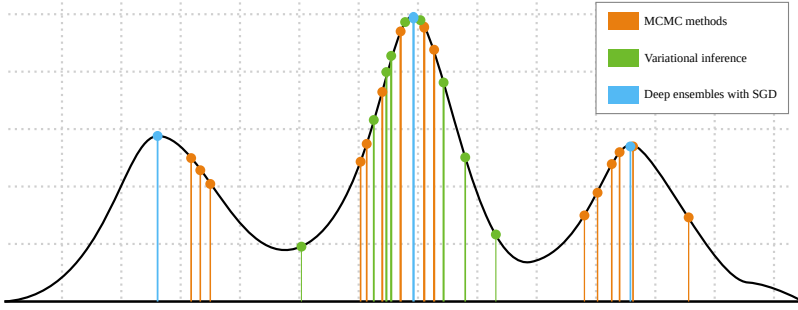


Fig. 13. Different techniques to sample the posterior. MCMC algorithms sample everywhere but successive samples might be correlated and accumulate in a specific region of the posterior, variational inference use a parametric distribution but can suffer from mode collapse while deep ensembles with SGD focuses on the different modes of the distribution.

These methods still suffer from the main drawbacks of MCMC methods, being their huge memory footprint. This is why some authors have proposed methods that function more like traditional variational inference and less like a MCMC algorithm.

**8.2.2 Variational inference based on SGD dynamic.** Instead of deriving a MCMC algorithm from the SGD dynamic, it is also possible to use this to fit a variational distribution, an approach that can be considered to be variational inference.

Many approaches are based on Laplace approximation, which allows to fit a Gaussian posterior by just using the maximum a posteriori estimate as mean and the Hessian of the loss, assuming the loss represents a log likelihood, as an inverse covariance matrix. Computing the Hessian matrix, and thus the curvature of the loss, is usually intractable for modern large size neural networks architecture, so approximations are required, like KFAC [75], a scalable method using kronecker factorization to fit a Gaussian with a block diagonal covariance matrix as the posterior. Another way to get an estimation of second order derivatives is to estimate gradient variance over multiple SGD iterations. SWAG [55] exploits this idea to fit a Gaussian posterior with a low rank plus diagonal covariance matrix. A similar method, named Vadam [40], has been proposed for the Adam algorithm. However, if those methods are able to capture the fine shape of one mode of the posterior, they cannot fit multiple modes.

In [49] the authors propose to use warm restarts to get different point estimate networks instead of fitting a parametric distribution. This method, called deep ensembles (see Fig. 13), has been used in the past to perform model averaging, but the main contribution of [49] is to show that it enables well calibrated error estimates. While the authors of [49] claim that their method is non-Bayesian, it has been shown that their approach can be understood from a Bayesian point of view [68, 93], especially if regularization is used (i.e., a prior is implied), the different points estimates should correspond to modes of a Bayesian posterior. To be truly Bayesian the relative posterior probability of those different modes should be accounted for during model averaging. This can be interpreted as approximating the posterior with a distribution parametrized as multiple dirac deltas, which can be considered to be some sort of variational inference, even if for such variational distribution it is impossible to compute the ELBO in a sense that is meaningful for traditional optimization:

$$q_{\phi}(\theta) = \sum_{\theta_i \in \phi} \alpha_{\theta_i} \delta_{\theta_i}(\theta), \quad (55)$$

with the  $\alpha_{\theta_i}$  as positive constants that sum to 1.



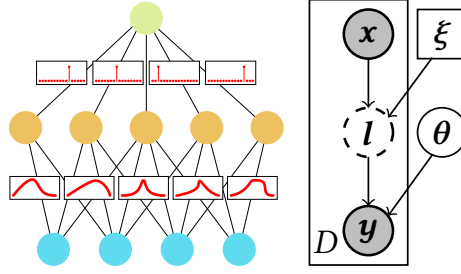


Fig. 14. Bayesian belief networks and Bayesian Neural Network architecture corresponding to a last-layer architecture.

### 8.3 Bayesian inference on the (n-)last layer(s) only

The main concern about being truly Bayesian for a whole ANN is that the architecture of deep neural networks makes it quite redundant to account for uncertainty for a large number of successive layers. Instead, recent research tried to use only a few stochastic layers, usually positioned at the end of the networks [8, 97].

Training only a few stochastic layers can drastically simplify the learning procedure. It removes many conception and training problems while still being able to give meaningful results from a Bayesian perspective. It can be interpreted as learning jointly a point-estimate transformation followed by a shallow BNN.

Training a (n)last-layer BNN seems non-trivial at first, but it is in fact very similar to learning the parameters for the prior, as presented in Section 7.4. The weights of the non-bayesian layers should be considered as both prior and variational-posterior parameters.

### 8.4 Bayesian teachers

Even with methods to speedup the training, evaluation using Monte Carlo samples is still problematic in applications where real time performances are important, e.g., self-driving cars. Storing a large set of parametrization  $\Theta$  obtained via MCMC is problematic in settings with limited memory, like smartphones and embedded devices.

Some authors proposed a solution to speed up evaluation time when using a BNN, derived from an approach already used in Bayesian modelling [81]. The approach is to train a non-stochastic ANN to predict the marginal probability  $p(y|x, D)$ , using a BNN as a teacher [45]. This is related to the idea of knowledge distillation [30, 57] where possibly several pre-trained knowledge sources can be used to train a more functional system.

To do so, the KL-divergence between the parametric distribution  $q_{\omega}(y|x)$ , where  $\omega$  are the coefficients of the student network, and the marginal distribution  $p(y|x, D)$  is minimized:

$$\hat{\omega} = \arg \min_{\omega} D_{KL}(p(y|x, D) || q_{\omega}(y|x)), \quad (56)$$

which is intractable. Korattikara et al. [45] propose a Monte Carlo approximation:

$$\hat{\omega} = \arg \min_{\omega} -\frac{1}{|\Theta|} \sum_{\theta_i \in \Theta} E_{p(y|x, \theta_i)} [\log(q_{\omega}(y|x))]. \quad (57)$$

This quantity can be estimated on a training dataset  $D'$ . The big advantage of this method is that it only requires the features  $x$  during training, as the probability of labels  $p(y|x, \theta)$  is defined by the BNN stochastic model (see Section 4.2). Without the need of labeling, one of the most expensive part of a machine learning workflow,  $D'$  can be much larger than the original set  $D$  used to train the

teacher BNN, so that the student network is far less likely to suffer from unexpected generalization flaws.

It has also been observed that, for classification problems, simply using the class probabilities outputted by a Bayesian teacher rather than one-hot labels can help a student to learn while retaining calibration and uncertainty from the original Bayesian neural networks [57].

Now for the case where one wants to compress a large set of samples generated using a MCMC algorithm, a Bayesian teacher can also be an interesting solution [90]. Instead of caching the collection of samples  $\Theta$ , a Generative model  $G$ , a GAN in [90], is trained against the MCMC samples to generate the coefficients  $\theta_i$  at evaluation time. This approach is similar somehow to variational inference, as the generative model represents in fact a parametric distribution, but the proposed algorithm allows to train a much more complex model than the distributions usually considered for variational inference.

## 9 EVALUATING BAYESIAN NEURAL NETWORKS PERFORMANCES

One big challenge with Bayesian neural networks is to evaluate their performance, as they do not output directly a point estimate prediction  $\hat{\mathbf{y}}$  but a conditional probability distribution  $p(\mathbf{y}|\mathbf{x}, D)$ , from which an optimal estimate  $\hat{\mathbf{y}}$  can later be extracted.

The predictive performance, sometimes called sharpness in statistics, of the network can be assessed by treating the estimator  $\hat{\mathbf{y}}$  as the prediction. This procedure often depends on the type of data the network is meant to treat, and many different metrics, e.g., MSE,  $\ell_n$  distances, cross-entropy, etc., are used in practice. Covering these metrics is out of the scope of this tutorial, for more details the reader can refer to [35].

Checking that the predicted posterior  $p(\mathbf{y}|\mathbf{x}, D)$  is well behaved, i.e., that the network is neither overconfident nor underconfident about its prediction, is also important. The standard method to do this is using a calibration curve, also called a reliability diagram [37, 47].

In practice, the calibration curve is a function  $\check{p} : [0, 1] \rightarrow [0, 1]$  representing the observed probability  $\check{p}$ , or empirical frequency, as a function of the predicted probability  $\hat{p}$  (Fig. 15). If  $\check{p} < \hat{p}$ , then the model is overconfident, otherwise, it is underconfident. A well calibrated model should have  $\check{p} \cong \hat{p}$ . This approach is really useful in practice. It requires choosing a set of events  $\mathcal{E}$  with different predicted probabilities and measuring the empirical frequency of each event using a test set  $T$ .

For a binary classifier, the set of test events can be chosen as the set of all sets of datapoints with predicted probabilities of acceptance in interval  $[p - \delta, p + \delta]$  for a chosen  $\delta$ , or alternatively  $[0, p]$  or  $[1 - p, 1]$  for small datasets. The empirical frequency is given by:

$$\check{p} = \frac{\sum_{\hat{\mathbf{y}} \in T_y} \check{\mathbf{y}} \cdot \mathbb{I}_{[\hat{p}-\delta, \hat{p}+\delta]}(\hat{\mathbf{y}})}{\sum_{\hat{\mathbf{y}} \in T_y} \mathbb{I}_{[\hat{p}-\delta, \hat{p}+\delta]}(\hat{\mathbf{y}})}. \quad (58)$$

For multi-class classifiers, the calibration curve can be checked for each class against all the other classes independently. In this case, the problem is reduced to a binary classifier.

Regression problems are slightly more complex, as the network will not output a confidence level, like a classifier, but a distribution of possible outputs. The solution is to use an intermediate statistic with a known probability distribution.

Assuming Independence between the  $\hat{\mathbf{y}}$  for a large enough set of different, randomly selected, inputs  $\mathbf{x}$ , one can assume that the sum of normalized squared errors follows a chi-square law:

$$(\hat{\mathbf{y}} - \check{\mathbf{y}})^T \Sigma_{\hat{\mathbf{y}}}^{-1} (\hat{\mathbf{y}} - \check{\mathbf{y}}) \sim \chi_{Dim(\mathbf{y})}^2. \quad (59)$$

This allows attributing to each data point in the test set  $T$  a predicted probability as the probability of observing a variance-normalized distance between the prediction and the true value equal or

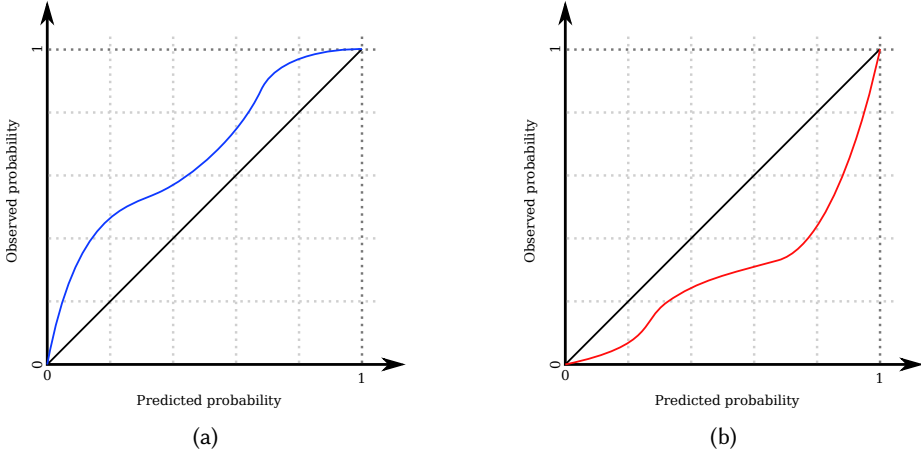


Fig. 15. Examples of calibration curve for underconfident (a) and overconfident (b) models.

lower than the actual distance. Formally, the predicted probability is computed as:

$$\hat{p}_i = X_{Dim(\mathbf{y})}^2 \left( (E_{p(\mathbf{y}|\mathbf{x}_i, D)}[\mathbf{y}] - \mathbf{y}_i)^T \Sigma_{\mathbf{y}|\mathbf{x}_i, D}^{-1} (E_{p(\mathbf{y}|\mathbf{x}_i, D)}[\mathbf{y}] - \mathbf{y}_i) \right) \quad \forall (\mathbf{y}_i, \mathbf{x}_i) \in T, \quad (60)$$

where  $X_{Dim(\mathbf{y})}^2$  is the cumulative distribution of the chi-square law with  $Dim(\mathbf{y})$  degrees of freedom. The observed probability can be computed as:

$$\check{p}_i = \frac{1}{|T|} \sum_{j=1}^{|T|} \mathbb{I}_{[0, \infty)}(\hat{p}_j - \hat{p}_i). \quad (61)$$

Now, giving the whole calibration curve for a given stochastic model is often a good idea, as it allows to see where the model is likely to be overconfident or underconfident. It also allows, to a certain extent, to recalibrate the model [47]. However, it might also be necessary to provide a summary statistic, to ease comparison or interpretation. The Area Under Curve (AUC) is a standard metric of the form:

$$AUC = \int_0^1 \check{p} d\hat{p}. \quad (62)$$

An AUC of 0.5 indicates that the model is, on average and only on average, well calibrated.

Less used in practice, the distance from the ideal calibration curve (defined in the functions Hilbert space) is also a good indicator for the calibration of a model:

$$d(\check{p}, \hat{p}) = \sqrt{\int_0^1 (\check{p} - \hat{p})^2 d\hat{p}}. \quad (63)$$

When  $d(\check{p}, \hat{p}) = 0$ , the model is perfectly calibrated.

Other measures have been proposed. Examples include the Expected Calibration Error, and some variants [63], which are discretized variants of the distance from the ideal calibration curve.

Those measures are good to summarize the calibration but not so much to properly learn calibrated models. The tools used in statistics to optimize calibration are scoring rules [24]. A scoring rule is a function  $S(P, x)$  where  $P$  is a predictive probability distribution, like the marginal given by a BNN, and  $x$  is an event. Given a target distribution  $Q$ , we write  $\bar{S}(P, Q) = E_{x \sim Q} [S(P, x)]$ . Scoring rules are, by convention, positively oriented such that the optimization objective becomes

maximizing  $\bar{S}(P, Q)$ . An important property of scoring rules is that their maxima favor a well calibrated model, that is:

$$\bar{S}(Q, Q) \geq \bar{S}(P, Q) \quad \forall P, \quad (64)$$

in which case the scoring rule is called a proper scoring rule. If we also have  $\bar{S}(Q, Q) = \bar{S}(P, Q) \Rightarrow P = Q$  then the scoring rule is said to be strictly proper. In theory, to get a proper calibrated model it is sufficient to ensure that the loss used for training corresponds to a (strictly) proper scoring rule. In practice this is more complex, as the expected value of the score can be evaluated accurately only for large datasets. This can cause miscalibration errors when the trained model is presented with data that is slightly shifted from the original training set [66]. This is where Bayesian models can really shine, as they follow a proper scoring rule by design and accounting for epistemic uncertainty along with aleatoric uncertainty makes them more resilient, even if not immune, to such changes. Yet, it might happen that a trained BNN is miscalibrated. In that case, the prior might need to be changed, as the assumptions made are too specific and probably not reasonable. Alternatively the training strategies, especially if the variational distribution used during training is not able to correctly approximate the posterior, might need to be reconsidered.

## 10 BAYESIAN DEEP LEARNING PROGRAMMING FRAMEWORKS

Before we conclude this tutorial, we quickly present in this section some frameworks that can be used for Bayesian Deep Learning.

### 10.1 Probabilistic Programming Languages

Probabilistic Programming Languages (PPLs) are meant to describe probability distributions while accounting for relations between variables in a way similar to PGMs to speed up inference. Most of them include many different optimizations algorithms, e.g., MCMC, variational inference and maximum likelihood estimators, while some only focus on specific methods.

Stan [10] and PyMC3 [77] are popular PPL for general use. Stan is written in C++ and uses its own language to define models. APIs are provided for all popular scientific software packages, including Python, R, Matlab, Julia and a command line interface. PyMC3, on the other hand, is written in Python and is based on Theano for automatic differentiation.

The problem when using pure PPL to design BNN is that they are built with complex probabilistic models in mind, but not specifically deep learning. It is definitely possible to write a BNN using a PPL, but most of the work will have to be done almost from scratch for the deep learning part.

### 10.2 Probabilistic Programming in Deep Learning Framework

PPL and deep learning frameworks have more or less the same requirements. The main ones are automatic differentiation and large scale optimization routines. From that observation, recent developments have focused on integrating PPL into deep learning frameworks. The results were Pyro [5], constructed on top of PyTorch, and Edward [87], constructed on top of TensorFlow. Edward has later on been extended and integrated into a larger sub-module of Tensorflow, Tensorflow probability.

The seamless integration of these PPLs with their respective deep learning frameworks makes them ideal tools for designing, training and using BNN. They also allow to easily re-use components that were designed for point estimate neural networks.

Pyro has been developed with a strong focus on variational inference. It is based on a seemingly functional programming paradigm, coupled with a context manager to easily modify stochastic functions on the fly. It includes a wrapper class to transform PyTorch layers into probabilistic layers, allowing to replace their parameters with sample sites on the fly.

Edward and tensorflow probability include some higher level structures, especially probabilistic layers usable on their own, and some lower level structure to build probabilistic networks.

Without doing a detailed comparison, we would recommend the use of the one integrated to your deep learning framework of choice, as both frameworks have similar potential. Pyro is probably better suited for dynamic PGM, as pytorch uses a dynamic computation graph, while Edward2 with TensorFlow probability will suit you better if you would like to use already existing Bayesian layers in a fixed architecture.

## 11 CONCLUSION

This tutorial covered the design, implementation, training, usage and evaluation of Bayesian Neural Networks. While the basic idea is simple, i.e., just training an artificial neural network with some probability distribution attached to its weights, the challenges implied, i.e., designing efficient algorithms to train or use a BNN, are still hard to address. This explains why BNNs are still scarce in practical applications. However, the potential applications of BNN are huge, and this paradigm offers a promising avenue to apply deep learning in areas where a system is not allowed to fail to generalize without emitting a warning. Moreover, there is more links than one could imagine a priori between traditional deep learning algorithms and the Bayesian paradigm, meaning that Bayesian methods, even if not applied directly, can help design new learning and regularization strategies.

## ACKNOWLEDGMENTS

This material is partially based on research sponsored by the Australian Research Council (Grants DP150100294 and DP150104251), and Air Force Research Laboratory and DARPA under agreement number FA8750-19-2-0501.

## REFERENCES

- [1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *CoRR*, arXiv:1607.06450, 2016. In NIPS 2016 Deep Learning Symposium.
- [2] Rémi Bardenet, Arnaud Doucet, and Chris Holmes. On Markov Chain Monte Carlo methods for tall data. *J. Mach. Learn. Res.*, 18(1):1515–1557, January 2017. ISSN 1532–4435.
- [3] M Saiful Bari, Muhammad Tasnim Mohiuddin, and Shafiq Joty. MultiMix: A robust data augmentation strategy for cross-lingual nlp. In *ICML*, 2020.
- [4] L. Beyer, X. Zhai, A. Oliver, and A. Kolesnikov. S4L: Self-supervised semi-supervised learning. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1476–1485, 2019.
- [5] Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D. Goodman. Pyro: Deep universal probabilistic programming. *J. Mach. Learn. Res.*, 20(1):973â–978, January 2019. ISSN 1532–4435.
- [6] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.
- [7] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1613–1622, 2015.
- [8] Nicolas Brosse, Carlos Riquelme, Alice Martin, Sylvain Gelly, and ÃLric Moulines. On last-layer algorithms for classification: Decoupling representation from uncertainty estimation. *CoRR*, abs/2001.08049, 2020. URL <http://arxiv.org/abs/2001.08049>.
- [9] W. L. Buntine. Operations for learning with graphical models. *Journal of Artificial Intelligence Research*, 2:159–225, Dec 1994. ISSN 1076–9757.
- [10] Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1), 2017.
- [11] Rodolfo C. Cavalcante, Rodrigo C. Brasileiro, Victor L.F. Souza, Jarley P. Nobrega, and Adriano L.I. Oliveira. Computational intelligence and financial markets: A survey and future directions. *Expert Systems with Applications*, 55:194–211,

2016. ISSN 0957-4174.
- [12] Olivier Chapelle, Jason Weston, Léon Bottou, and Vladimir Vapnik. Vicinal risk minimization. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 416–422. MIT Press, 2001.
  - [13] Siddhartha Chib and Edward Greenberg. Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49(4):327–335, 1995.
  - [14] Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1184–1193, 2018.
  - [15] Alexander Etz, Quentin F. Gronau, Fabian Dablander, Peter A. Edelsbrunner, and Beth Baribault. How to become a Bayesian in eight easy steps: An annotated reading list. *Psychonomic Bulletin & Review*, 25:219–234, 2018.
  - [16] B. Frenay and M. Verleysen. Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869, 2014.
  - [17] Yarin Gal and Zoubin Ghahramani. Bayesian convolutional neural networks with Bernoulli approximate variational inference. *CoRR*, abs/1506.02158, 2015. URL <http://arxiv.org/abs/1506.02158>.
  - [18] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning - Volume 48*, ICML’16, page 1050–1059, 2016.
  - [19] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep Bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML’17, page 1183–1192, 2017.
  - [20] Francis Galton. Vox Populi. *Nature*, 75(1949):450–451, Mar 1907. ISSN 1476-4687.
  - [21] Andrew Gelman and other Stan developers. Prior choice recommendations, 2020. Retrieved from <https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations> [last seen 13.07.2020].
  - [22] Edward I. George, George Casella, and Edward I. George. Explaining the Gibbs sampler. *The American Statistician*, 1992.
  - [23] Zoubin Ghahramani and Matthew J. Beal. Propagation algorithms for variational Bayesian learning. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 507–513. MIT Press, 2001.
  - [24] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.
  - [25] Erin Grant, Chelsea Finn, Sergey Levine, Trevor Darrell, and Thomas L. Griffiths. Recasting gradient-based meta-learning as hierarchical Bayes. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.
  - [26] Alex Graves. Practical variational inference for neural networks. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2348–2356. Curran Associates, Inc., 2011.
  - [27] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural network. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML’17, pages 1321–1330, 2017.
  - [28] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations, ICLR 2017, Conference Track Proceedings*, 2017.
  - [29] José Miguel Hernández-Lobato and Ryan P. Adams. Probabilistic backpropagation for scalable learning of Bayesian neural networks. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML’15, page 1861–1869, 2015.
  - [30] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. In NIPS 2014 Deep Learning Workshop.
  - [31] Matthew D Hoffman and Andrew Gelman. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.
  - [32] Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *J. Mach. Learn. Res.*, 14(1):1303–1347, May 2013. ISSN 1532-4435.
  - [33] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *CoRR*, abs/2004.05439, 2020. URL <http://arxiv.org/abs/2004.05439>.
  - [34] Jiri Hron, Alex Matthews, and Zoubin Ghahramani. Variational Bayesian dropout: pitfalls and fixes. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2019–2028, 2018.
  - [35] Katarzyna Janocha and Wojciech Marian Czarnecki. On loss functions for deep neural networks in classification. *Schedae Informaticae*, 1/2016, 2017. ISSN 2083-8476. doi: 10.4467/20838476si.16.004.6185. URL <http://dx.doi.org/10.4467/20838476si.16.004.6185>.
  - [36] L. Jing and Y. Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020.



- [37] Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17*, page 5580–5590, 2017. ISBN 9781510860964.
- [38] J. Ker, L. Wang, J. Rao, and T. Lim. Deep learning applications in medical image analysis. *IEEE Access*, 6:9375–9389, 2018.
- [39] M. E. Khan and D. Nielsen. Fast yet simple natural-gradient descent for variational inference in complex models. In *2018 International Symposium on Information Theory and Its Applications (ISITA)*, pages 31–35, 2018.
- [40] Mohammad Khan, Didrik Nielsen, Voot Tangkaratt, Wu Lin, Yarin Gal, and Akash Srivastava. Fast and scalable Bayesian deep learning by weight-perturbation in Adam. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2611–2620, 2018.
- [41] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014.
- [42] Diederik P Kingma, Max Welling, et al. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.
- [43] Durk P Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2575–2583. Curran Associates, Inc., 2015.
- [44] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural Safety*, 31(2):105–112, 2009. ISSN 0167–4730. Risk Acceptance and Risk Communication.
- [45] Anoop Korattikara, Vivek Rathod, Kevin Murphy, and Max Welling. Bayesian dark knowledge. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2, NIPS'15*, page 3438–3446, 2015.
- [46] Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Being Bayesian, even just a bit, fixes overconfidence in ReLU networks. *CoRR*, abs/2002.10118, 2020. URL <http://arxiv.org/abs/2002.10118>.
- [47] Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2796–2804, 2018.
- [48] R. Kunwar, U. Pal, and M. Blumenstein. Semi-supervised online Bayesian network learner for handwritten characters recognition. In *2014 22nd International Conference on Pattern Recognition*, pages 3104–3109, 2014.
- [49] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6402–6413. Curran Associates, Inc., 2017.
- [50] Jouko Lampinen and Aki Vehtari. Bayesian approach for neural networks—review and case studies. *Neural Networks*, 14(3):257 – 274, 2001. ISSN 0893-6080.
- [51] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, 2013.
- [52] Yingzhen Li and Yarin Gal. Dropout inference in Bayesian neural networks with alpha-divergences. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17*, pages 2052–2061, 2017.
- [53] Zhun Li, ByungSoo Ko, and Ho-Jin Choi. Naive semi-supervised deep learning using pseudo-label. *Peer-to-Peer Networking and Applications*, 12(5):1358–1368, 2019. ISSN 1936-6450.
- [54] David J. C. MacKay. A practical Bayesian framework for backpropagation networks. *Neural Computation*, 4(3):448–472, 1992.
- [55] Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Petrov, and Andrew Gordon Wilson. A simple baseline for Bayesian uncertainty in deep learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 13153–13164. Curran Associates, Inc., 2019.
- [56] Stephan Mandt, Matthew D Hoffman, and David M Blei. Stochastic gradient descent as approximate Bayesian inference. *The Journal of Machine Learning Research*, 18(1):4873–4907, 2017.
- [57] Aditya Krishna Menon, Ankit Singh Rawat, Sashank J. Reddi, Seungyeon Kim, and Sanjiv Kumar. Why distillation helps: a statistical perspective. *CoRR*, abs/2005.10419, 2020. URL <https://arxiv.org/abs/2005.10419>.
- [58] John Mitros and Brian Mac Namee. On the validity of Bayesian neural networks for uncertainty estimation. In *AICS*, 2019.
- [59] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012. ISBN 0262018020.
- [60] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *International Conference on Learning Representations*, 2019.

- [61] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 1196–1204. Curran Associates, Inc., 2013.
- [62] Radford M Neal et al. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11):2, 2011.
- [63] Jeremy Nixon, Michael W. Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.
- [64] Manfred Opper and Ole Winther. A Bayesian approach to on-line learning. *On-line learning in neural networks*, pages 363–378, 1998.
- [65] Kazuki Osawa, Siddharth Swaroop, Mohammad Emtiyaz E Khan, Anirudh Jain, Runa Eschenhagen, Richard E Turner, and Rio Yokota. Practical deep learning with Bayesian principles. In *Advances in Neural Information Processing Systems 32*, pages 4287–4299. Curran Associates, Inc., 2019.
- [66] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 13991–14002. Curran Associates, Inc., 2019.
- [67] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.
- [68] Tim Pearce, Felix Leibfried, Alexandra Brintrup, Mohamed Zaki, and Andy Neely. Uncertainty in neural networks: Approximately Bayesian ensembling. In *AISTATS 2020*, 2020.
- [69] Nicholas G Polson, Vadim Sokolov, et al. Deep learning: a Bayesian perspective. *Bayesian Analysis*, 12(4):1275–1304, 2017.
- [70] Arya A Pourzanjani, Richard M Jiang, Brian Mitchell, Paul J Atzberger, and Linda R Petzold. Bayesian inference over the Stiefel manifold via the Givens representation. *CoRR*, abs/1710.09443, 2017. URL <http://arxiv.org/abs/1710.09443>.
- [71] Samira Pouyanfar, Saad Sadiq, Yilin Yan, Haiman Tian, Yudong Tao, Maria Presa Reyes, Mei-Ling Shyu, Shu-Ching Chen, and S. S. Iyengar. A survey on deep learning: Algorithms, techniques, and applications. *ACM Comput. Surv.*, 51(5), September 2018. ISSN 0360–0300.
- [72] Guo-Jun Qi and Jiebo Luo. Small data challenges in big data era: A survey of recent progress on unsupervised and semi-supervised methods. *CoRR*, abs/1903.11260, 2019. URL <http://arxiv.org/abs/1903.11260>.
- [73] Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. Pre-trained models for natural language processing: A survey. *CoRR*, abs/2003.08271, 2020.
- [74] Qing Rao and Jelena Rftunikj. Deep learning for self-driving cars: Chances and challenges. In *Proceedings of the 1st International Workshop on Software Engineering for AI in Autonomous Systems*, SEFAIS '18, pages 35–38, 2018.
- [75] Hippolyt Ritter, Aleksandar Botev, and David Barber. A scalable laplace approximation for neural networks. In *International Conference on Learning Representations*, 2018.
- [76] Christian Robert. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media, 2007.
- [77] John Salvatier, Thomas V Wiecki, and Christopher Fonnesbeck. PyMC3: Python probabilistic programming framework. *PeerJ Computer Science*, 2:e55, 2016. <https://doi.org/10.7717/peerj-cs.55>.
- [78] Nabeel Seedat and Christopher Kanan. Towards calibrated and scalable uncertainty representations for neural networks. *CoRR*, abs/1911.00104, 2019. URL <http://arxiv.org/abs/1911.00104>.
- [79] Joan Serrà, David Álvarez, Vicenç Gámez, Olga Slizovskaia, Josá F. Nájiz, and Jordi Luque. Input complexity and out-of-distribution detection with likelihood-based generative models. *CoRR*, abs/1909.11480, 2020. URL <http://arxiv.org/abs/1909.11480>.
- [80] Daniele Silvestro and Tobias Andermann. Prior choice affects ability of Bayesian neural networks to identify unknowns. *CoRR*, abs/2005.04987, 2020. URL <http://arxiv.org/abs/2005.04987>.
- [81] Edward Snelson and Zoubin Ghahramani. Compact approximations to Bayesian predictive distributions. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, page 840, 2005.
- [82] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. FixMatch: Simplifying semi-supervised learning with consistency and confidence. *CoRR*, abs/2001.07685, 2020. URL <https://arxiv.org/abs/2001.07685>.
- [83] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.
- [84] Martin A. Tanner and Wing Hung Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540, 1987.
- [85] D. M. Titterton. Bayesian methods for neural networks and related models. *Statist. Sci.*, 19(1):128–139, 02 2004.
- [86] Adrian Corduneanu Tommi and Tommi Jaakkola. On information regularization. In *Proceedings of the 19th UAI*, 2003.

- [87] Dustin Tran, Matthew D. Hoffman, Rif A. Saurous, Eugene Brevdo, Kevin Murphy, and David M. Blei. Deep probabilistic programming. *CoRR*, abs/1701.03757, 2017. URL <http://arxiv.org/abs/1701.03757>.
- [88] Toan Tran, Thanh-Toan Do, Ian Reid, and Gustavo Carneiro. Bayesian generative active deep learning. *CoRR*, abs/1904.11643, 2019. URL <http://arxiv.org/abs/1904.11643>.
- [89] Hao Wang and Dit-Yan Yeung. Towards Bayesian deep learning: A framework and some existing methods. *IEEE Trans. on Knowl. and Data Eng.*, 28(12):3395–3408, December 2016. ISSN 1041–4347.
- [90] Kuan-Chieh Wang, Paul Vicol, James Lucas, Li Gu, Roger Grosse, and Richard Zemel. Adversarial distillation of Bayesian neural network posteriors. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5190–5199, 2018.
- [91] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th international conference on machine learning*, ICML ’11, pages 681–688, 2011.
- [92] Yeming Wen, Paul Vicol, Jimmy Ba, Dustin Tran, and Roger Grosse. Flipout: Efficient pseudo-independent weight perturbations on mini-batches. In *International Conference on Learning Representations*, 2018.
- [93] Andrew Gordon Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. *CoRR*, abs/2002.08791, 2020. URL <http://arxiv.org/abs/2002.08791>.
- [94] D. H. Wolpert. The lack of a priori distinctions between learning algorithms. *Neural Computation*, 8(7):1341–1390, 1996.
- [95] Qizhe Xie, Zihang Dai, Eduard H. Hovy, Minh-Thang Luong, and Quoc V. Le. Unsupervised data augmentation. *CoRR*, abs/1904.12848, 2019. URL <http://arxiv.org/abs/1904.12848>.
- [96] Shipeng Yu, Balaji Krishnapuram, R  mer Rosales, and R. Bharat Rao. Bayesian co-training. *Journal of Machine Learning Research*, 12(80):2649–2680, 2011.
- [97] Jiaming Zeng, Adam Lesnikowski, and Jose M. Alvarez. The relevance of Bayesian layer positioning to model uncertainty in deep Bayesian active learning. *CoRR*, abs/1811.12535, 2018. URL <http://arxiv.org/abs/1811.12535>.
- [98] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR*, 2017.
- [99] Zhi-Hua Zhou. *Ensemble Methods: Foundations and Algorithms*. Chapman and Hall/CRC, 1st edition, 2012.

## A A PROOF OF EQUATION 48

Given a probability space  $(\Omega, \mathcal{F}, P)$ , where  $\Omega$  is a set of outcomes,  $\mathcal{F}$  is a  $\sigma$ -algebra of  $\Omega$  representing possible events and  $P$  is a measure defined on  $\mathcal{F}$  and assigning value 1 to  $\Omega$ , representing the probability of an event, let's assume that we have a probability distribution  $q_\phi(\theta)$  for a given random variable  $\theta$ , a probability distribution  $q(\varepsilon)$  for a given random variable  $\varepsilon$  and a functional relation  $t(\varepsilon, \phi)$  such that  $t(\varepsilon, \phi)$  is distributed like  $q_\phi(\theta)$  and  $t(\varepsilon, \phi)$  is a bijection with respect to  $\varepsilon$ .

Now we have:

$$P(\theta^{-1}(t(E, \phi))) = P(\varepsilon^{-1}(E)) \quad \forall E \in \mathcal{E}(\mathcal{F}), \quad (65)$$

with  $\mathcal{E}(\mathcal{F}) = \{\{\varepsilon(\omega) : \omega \in e\} : e \in \mathcal{F}\}$ ,  $t(E, \phi) = \{t(\varepsilon, \phi) : \varepsilon \in E\}$ ,  $\varepsilon^{-1}(E) = \bigcup_{e \in \mathcal{F} \wedge \varepsilon(e) \subseteq E} e$  and  $\theta^{-1}(t(E, \phi)) = \bigcup_{e \in \mathcal{F} \wedge \theta(e) \subseteq t(E, \phi)} e$ . Since  $t(\varepsilon, \phi)$  is a bijection with respect to  $\varepsilon$  we have  $\varepsilon^{-1}(E) = \theta^{-1}(t(E, \phi))$ . This implies:

$$\int_{\theta \in t(E, \phi)} q_\phi(\theta) d\theta = \int_{\varepsilon \in E} q(\varepsilon) d\varepsilon \quad \forall E \in \mathcal{E}(\mathcal{F}). \quad (66)$$

which in turn implies:

$$q_\phi(\theta) d\theta = q(\varepsilon) d\varepsilon \quad (67)$$

for non-degenerated probability distributions  $q_\phi(\theta)$  and  $q(\varepsilon)$ .

Now, given a differentiable function  $f(\phi, \theta)$ , we have:

$$\int_{\theta \in t(\varepsilon(\Omega), \phi)} f(\phi, \theta) q_\phi(\theta) d\theta = \int_{\varepsilon \in \varepsilon(\Omega)} f(\phi, t(\varepsilon, \phi)) q(\varepsilon) d\varepsilon \quad (68)$$

which implies Equations (48).