Representing and Quantifying Predictive Uncertainty in Machine Learning

Lisa Wimmer

München 2025

Lisa Wimmer

Representing and Quantifying Predictive Uncertainty in Machine Learning

Dissertation an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

Eingereicht von Lisa Wimmer am 02.08.2025

Erster Berichterstatter:	Prof. Dr. Bernd Bischl
Zweiter Berichterstatter:	Prof. Dr. Eyke Hüllermeier
Dritter Berichterstatter:	Prof. Dr. Willem Waegeman

Tag der mündlichen Prüfung: 18.11.2025

Acknowledgments

I would like to express my sincere and heartfelt gratitude to

... Prof. Dr. Bernd Bischl, for taking me on as part of his group, supervising my thesis, and providing guidance throughout. I am deeply grateful for the generous funding and academic freedom that allowed me to follow my heart.

... Prof. Dr. Eyke Hüllermeier and Prof. Dr. Willem Waegeman, for their willingness to dedicate their valuable time as the second and third reviewer of this thesis.

... Prof. Dr. Michael Schomaker, for his availability to be part of my doctoral committee.

... Prof. Dr. David Rügamer, for his availability to be part of my doctoral committee, but mostly, for his time and mentorship, both of which he has offered generously from day one.

... Dr. Ludwig Bothmann, who made research, teaching and all the other little academic pleasures run smoothly, always lending an ear and making me feel appreciated as a person.

... all the lovely people, past and present, from AKA7, 341 and 344. They supplied the fun (and the snacks) in this rollercoaster ride.

... Conni, who deserves special mention for being not only my colleague, but also my best friend, almost-neighbor, indefatigable well of support, and Maus.

... my amazing friends from outside academia. Despite having not the slightest idea what I am doing, they have carried me to this moment with untiring pride and enthusiasm.

... my parents, who have even less clue, which has never deterred them from standing by me with incredible trust, support and patience. Sitting through a child's twenty-five academic semesters is no small feat.

... and lastly, FTN, who has made the final months of my PhD so much more beautiful.

Summary

Machine learning has long been the art of *predictive* modeling: designing powerful, often black-box, algorithms that excel at prediction tasks. With real-world applications came the need for a systematic acknowledgment of potential prediction errors through statements of confidence. In what may be seen as a reconciliation of machine learning with classical statistics, the field of *uncertainty estimation* emerged. Much research has been dedicated to this endeavor, yet reasoning about uncertainty remains difficult. This thesis discusses predictive uncertainty estimation as a compositional problem. It involves [1] finding a *representation* that can be associated with some degree of uncertainty, and [2] *quantifying* this uncertainty by appropriate measures.

1. Distributions over random variables are a natural way to express uncertainty. The Bayesian paradigm provides a canonical framework for distributional representation. Weighted according to their posterior probability, all possible models combine to a predictive density. This bi-level approach pertains to the distinct sources of uncertainty that should enter any meaningful representation: *aleatoric* uncertainty due to imperfect data—the considered features cannot predict the target beyond doubt, even given the correct model, and *epistemic* uncertainty due to imperfect knowledge—the observed data do not suffice to establish which model has generated them. For the neural network models present in many of today's applications, Bayesian inference is severely inhibited by intractable, multi-modal posterior densities. The practical consequence is approximate inference with questionable representation quality. As contributions **[C1]** and **[C2]** in Part II of this thesis lay out, however, complex neural network posteriors are highly structured by symmetric patterns. The first article exploits symmetries to reduce the scope of the inference target and shows the merit of multi-start, sampling-based inference. Building on this, the second work investigates posterior multi-modality in depth and proposes a method for multi-basin sampling from favorable starting locations. Both contribute to facilitating exact inference for faithful uncertainty representation.

2. Given a suitable representation, making uncertainty operational still requires quantification. This typically means aggregating information contained in the predictive density to some summary statistics that can be separately attributed to aleatoric and epistemic uncertainty. The prevailing measures for this suffer from a number of conceptual flaws addressed in contributions **[C3]** and **[C4]**. Specifically, the first article provides a critical discussion of the popular entropy decomposition for classification tasks. In follow-up work, the second article connects entropy-based uncertainty to shortcut learning and shows how the ensuing quantification results point to a more global issue: transferring the above to scenarios where the data-generating process changes from training to deployment. While such non-*i.i.d.* settings are largely beyond the scope of this thesis, their practical relevance cannot be overstated.

Arguably, there is a third component to the uncertainty estimation problem. The absence of an observable ground-truth uncertainty impedes the evaluation of any progress. These challenges notwithstanding, uncertainty estimates have proven useful in downstream applications. Contribution **[C5]** proposes an end-to-end classification pipeline for imaging data where uncertainty serves as a decision criterion in dynamic data acquisition under a constrained labeling budget. Lastly, the remaining article **[C6]** contributes to entirely label-free, self-supervised learning by demonstrating how the injection of uncertainty—by means of a diversity-enhanced ensemble component—encourages better latent representations.

Zusammenfassung

Maschinelles Lernen galt stets als die Kunst *prädiktiver* Modellierung: die Konzeption mächtiger, oft intransparenter, Algorithmen, die exzellente Vorhersagen liefern. Mit großflächiger Anwendung entstand ein Bedarf nach systematischer Anerkennung von potentiellen Vorhersagefehlern durch Konfidenzaussagen. Im Zusammenspiel von maschinellem Lernen und klassischer Statistik bildete sich das Feld der *Unsicherheitsschätzung*. Diesem Ziel ist viel Forschung gewidmet worden; dennoch bleibt der Umgang mit Unsicherheit kompliziert. Die vorliegende Arbeit behandelt Unsicherheitsschätzung als kompositionelles Problem. Dieses umfasst [1] die Entwicklung von *Repräsentationen*, die mit einem Grad an Unsicherheit assoziiert werden können, und [2] die *Quantifizierung* dieser Unsicherheit durch geeignete Maße.

1. Verteilungen über Zufallsvariablen sind eine natürliche Wahl zur Abbildung von Unsicherheit. Das Bayesianische Paradigma ermöglicht eine kanonische Darstellungsform für Verteilungsrepräsentationen. Alle denkbaren Modelle werden, gewichtet nach ihrer *a posteriori*-Wahrscheinlichkeit, in einer Vorhersagedichte kombiniert. Dieser zweistufige Ansatz reflektiert die unterschiedlichen Quellen von Unsicherheit, die Teil jeder sinnvollen Repräsentation sein sollten: *Aleatorische* Unsicherheit durch mangelhafte Daten—die betrachteten Variablen können, auch mit dem korrekten Modell, die Zielvariable nicht zweifelsfrei vorhersagen, und *epistemische* Unsicherheit—die beobachteten Daten reichen nicht aus, um dasjenige Modell zu identifizieren, das sie generiert hat. Für die heutzutage üblichen neuronalen Netze wird Bayesianische Inferenz erheblich durch multimodale Dichten, die sich einer expliziten mathematischen Darstellung entziehen, beeinträchtigt. Infolgedessen sind approximative Inferenzmethoden mit fragwürdiger Repräsentationsgüte weit verbreitet. Wie jedoch in den Beiträgen **[C1]** und **[C2]** in Teil II der vorliegenden Arbeit beschrieben, sind komplexe Posteriori-Dichten in neuronalen Netzen hochgradig durch symmetrische Muster strukturiert. Der erste Artikel nutzt Symmetrien, um den Umfang des Inferenzproblems zu reduzieren, und demonstriert die Nützlichkeit ziehungsbasierter Verfahren mit multiplen Startpunkten. Darauf aufbauend untersucht die zweite Publikation multimodale Dichten im Detail und schlägt eine ziehungsbasierte Methode mit günstig gewählten Startpunkten vor. Beide Arbeiten tragen zur Erleichterung exakter Inferenz für annahmegetreue Unsicherheitsrepräsentation bei.

2. Auch mit einer geeigneten Repräsentation erfordert die Operationalisierung von Unsicherheit eine Form von Quantifizierung. Üblicherweise bedeutet dies, die in der Vorhersagedichte enthaltene Information in einer zusammenfassenden Statistik zu aggregieren, die anschließend der aleatorischen und epistemischen Unsicherheit separat zugeordnet werden kann. Die dafür vorherrschenden Maße weisen konzeptuelle Mängel auf, die in den Beiträgen **[C3]** und **[C4]** adressiert werden. Speziell beinhaltet der erste Artikel eine kritische Diskussion der verbreiteten Entropiedekomposition für Klassifikationsprobleme. Daran anschließend stellt der zweite Artikel einen Zusammenhang zwischen entropiebasierter Unsicherheit und dem Lernen sogenannter *shortcuts* her, und weist hierdurch auf ein größeres Problem hin: die Übertragung der obigen Ausführungen auf Szenarien, in denen sich der datengenerierende Prozess zwischen Training und tatsächlichem Modelleinsatz verändert. Zwar gehen solche Situationen größtenteils über den Umfang dieser Arbeit hinaus, doch sind sie von enormer praktischer Bedeutung.

Das Problem der Unsicherheitsschätzung hat gewissermaßen noch eine dritte Komponente. Das Fehlen beobachtbarer, wahrer Unsicherheitswerte erschwert die Beurteilung jeglichen Fortschritts.

Trotz dieser Herausforderungen haben sich Unsicherheitsschätzungen in nachgelagerten Anwendungen als nützlich erwiesen. Artikel [C5] entwickelt eine vollständig trainierbare Prozedur zur Klassifikation von Bilddaten, in der Unsicherheit als Entscheidungskriterium für dynamische Datenakquise mit begrenztem Annotierungsbudget fungiert. Schließlich trägt Artikel [C6] zu vollständig annotationsfreiem selbstüberwachten Lernen bei, indem aufgezeigt wird, wie die Injektion von Unsicherheit—durch eine zur Diversität animierten Ensemble-Komponente—bessere latente Repräsentationen begünstigt.

Contents

List of Figures

Abbreviations

ABI	approximate Bayesian inference
AI	artificial intelligence
BNN	Bayesian neural network
CPE	cold posterior effect
DGP	data-generating process
DL	deep learning
ERM	empirical risk minimization
HMC	Hamiltonian Monte Carlo
i.i.d.	independent and identically distributed
KL	Kullback-Leibler
LA	Laplace approximation
LLM	large language model
LMC	linear mode connectivity
(L)PPD	(log) posterior predictive density
MAP	maximum *a posteriori*
MCMC	Markov chain Monte Carlo
ML	machine learning
MLP	multi-layer perceptron
NLL	negative log-likelihood
NN	neural network
NUTS	no-U-turn sampler
OOD	out of distribution
RV	random variable
SBI	sampling-based inference
SGD	stochastic gradient descent
SCL	shortcut learning
SLL	self-supervised learning
UE	uncertainty estimation
UQ	uncertainty quantification
UR	uncertainty representation
VI	variational inference
ZDI	zero-mean, diagonal, isotropic

I

Introduction and Background

1. Introduction

The past few years have seen an unprecedented surge in applications for artificial intelligence (AI). Machines can process huge quantities of data, uncover latent patterns, and handle computations in seconds that would take humans years to solve. Many fields have thus realized the potential of predictive machine learning (ML) to assist decisions. Examples include the prediction of molecular properties in drug discovery (Klarner et al., 2023), modeling large-scale events in climate science (Eyring et al., 2024), and characterizing stellar objects in astrophysics (Tamames-Rodero et al., 2025). The 2024 selection of Nobel prize laureates reflects the growing recognition of ML among the most respected sciences[1]. Outside academic circles, the meteoric rise of large language models (LLMs) has brought AI into the everyday lives of millions of users.

Alas, these models' capabilities can be deceptive. They are still abstractions and thus inherently imperfect. Most people would probably concede that, say, predicting job positions from motorists' preferred car brands is simplistic and saddled with considerable inaccuracy. Gauging trustworthiness becomes more difficult with increasing predictive power and opacity of the model. This is all the more true for systems that are not incentivized to advertise their mistakes. LLMs appear especially prone to exuding confidence despite being wrong on a regular basis (Papamarkou et al., 2024; Bo et al., 2025). It is to be hoped that the scientists who develop models recognize their conceptions' fallibility. The same judgmental burden can hardly be placed on regular users. Of course, even a general awareness of potential errors has limited use in assessing the trustworthiness of individual predictions. Tran et al. (2022) posit that, therefore, ***reliable*** models need to "represent their own uncertainty". Precisely what we should mean by ***uncertainty*** is not easily defined in one sentence and will be discussed in Part I of this thesis. Intuitively, predictions reflecting a model's uncertainty might come in the shape of *"I am 98% sure this person holds an upper management position"*. All models that abstract from the real world are indisputably affected by uncertainty—they just do not always say so.

> "As far as the laws of mathematics refer to reality, they are not certain; and as far as they are certain, they do not refer to reality." *Albert Einstein*

Faithful ***uncertainty estimation (UE)*** is the key problem addressed by this thesis. Given that uncertainty is the bread and butter of statisticians, who have been devising models for a long time, it may seem surprising that the ML community is still debating this. ML predictors, however, largely emancipated from the distributional assumptions of statistical modeling, do not admit trustworthy confidence estimates[2] out of the (black) box (Guo et al., 2017; Ovadia et al., 2019; Fisher and Marzouk, 2024). Goodfellow et al. (2015) alerted the community to the risk of confidently wrong predictions by deliberate image manipulation (see Fig. 1.1). The perturbation is invisible to the human eye, making such failure modes hard to anticipate.

[1] `https://www.nobelprize.org/all-nobel-prizes-2024/`

[2] We will use the term *confidence* to mean the opposite of *uncertainty*.

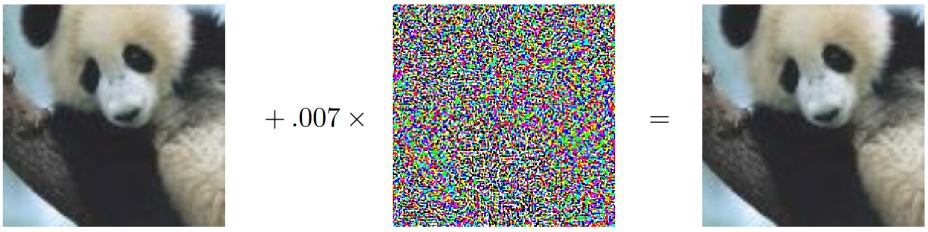

Figure 1.1.: Confidently wrong prediction. Famous example of an image classifier being fooled into declaring an image showing a panda as a "gibbon" with very low uncertainty. Adapted from Goodfellow et al. (2015).

Despite ongoing work to rectify this disposition, **UE** remains an open issue[3] (e.g., Yadkori et al., 2024; Aichberger et al., 2024; Pituk et al., 2025). The potential consequences are much more dire than misclassifying animals or job positions. Recalling one of the above-cited applications, comparable failures in climate modeling will misallocate resources and endanger human lives. Advancing **UE** is paramount in a world increasingly relying on AI.

We follow Hofman et al. (2024) in distinguishing two components to the overall problem.

1. ***Uncertainty representation (UR)***. First, we need a way of representing a multiplicity of options that can be associated with uncertainty. Accounting for the relevant sources and handling intractable representations are key challenges in this regard. Contributions [**C1**] and [**C2**] in Part II of this thesis propose solutions for feasible **UR**.
2. ***Uncertainty quantification[4] (UQ)***. Second, suitable measures must be found that quantify the level of uncertainty embodied by the represented multiplicity. Contributions [**C3**] and [**C4**] critically investigate the current practice for quantifying uncertainty in classification tasks. Arguably, many of the pitfalls in **UQ** arise from the attempt to describe complicated representational objects by single-value summary statistics.

A third aspect makes the problem still more thorny: we lack good evaluation protocols for **UE** because there is no ground truth to observe. We can assert that the driver in the previous example works in upper management, but no such straightforward reality check exists to determine whether 98% is an appropriate amount of confidence. Despite these difficulties, uncertainty-aware methods have been shown to yield good results in many settings. Contributions [**C5**] and [**C6**] describe examples for the successful use of uncertainty estimates in downstream applications.

[3] The *Conference on Uncertainty in Artificial Intelligence* just held its 41st meeting and shows no signs of stopping.
[4] The literature largely refers to the process of producing uncertainty estimates by the term "quantification"; we emphasize the problem's compositional nature by distinguishing between **UR** and **UQ**.

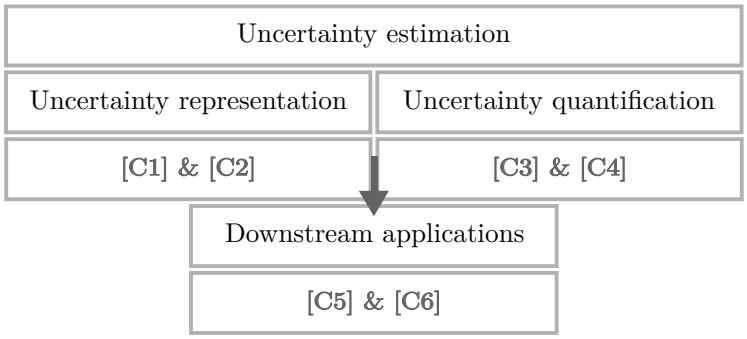

Figure 1.2.: Schematic summary of thesis contributions.

Outline This work is a thesis by publication. Part I provides the methodological background for the contributions in Part II. We begin by setting down our general notation and proceed to outline the central **UE** problem in Sec. 2.1. Subsequently, we address the problem components of **UR** and **UQ**. The focus of the respective sections follows the topics addressed by contributions **[C1]**–**[C4]**, putting special emphasis on representation by Bayesian deep learning and quantification by entropic measures. Sec. 2.4 examines how uncertainty estimates can be evaluated in absence of an observable ground truth. Lastly, we summarize in Sec. 2.5 how downstream applications employ uncertainty estimates as proposed in **[C5]** and **[C6]**. Part II collects the articles constituting this thesis, grouped by area of contribution: *representation*, *quantification*, and *applications* of predictive uncertainty (Fig. 1.2). We conclude with some remarks on future research for **UE**.

2. Methodological Background

2.1. Uncertainty-Aware Machine Learning

This chapter lays down notation, introduces the concept of *uncertainty*, and outlines the central *uncertainty estimation* problem discussed in this thesis.

2.1.1. General Notation

Supervised Learning The notation in this thesis is inspired by Bischl et al. (2023). We focus on ***supervised*** ML, i.e., learning predictive models from data annotated with the quantity of interest. The observed ***features*** (also called ***covariates***) $\boldsymbol{x}$ are a realization of the random variable (RV) $X \sim p_X$ with density function $p(\boldsymbol{x})$, short for $p(X = \boldsymbol{x})$. The feature values occupy a ***feature space*** $\mathcal{X}$ (frequently, $\mathcal{X} \subseteq \mathbb{R}^P$). Similarly, we have ***labels*** (or ***targets***) y as a realization of the RV $Y \sim p_Y$ with density function $p(y)$. The ***label space*** $\mathcal{Y} \ni y$ is typically given by $\mathcal{Y} = \{c_1, \ldots, c_K\}$ for ***classification*** and $\mathcal{Y} = \mathbb{R}$ for ***regression*** tasks. We observe feature-label tuples presumed to be *i.i.d.* (independent and identically distributed) samples from the ***data-generating process (DGP)*** p_{XY}, collected in the ***dataset*** $\mathcal{D} = ((\boldsymbol{x}^{(1)}, y^{(1)}), \ldots, (\boldsymbol{x}^{(N)}, y^{(N)}))$. The DGP is usually inaccessible. We further assume that the relationship between features and labels can be characterized by a learnable function $h : \mathcal{X} \to \mathcal{Y}'$. We call such functions ***hypotheses*** and imply the existence of a true hypothesis h^* creating realizations of p_{XY}. The structure of the co-domain $\mathcal{Y}'$ depends on the ***task*** at hand. Some hypotheses output hard labels ($\mathcal{Y}' = \mathcal{Y}$), others, distributions ($\mathcal{Y}' = \mathbb{P}(\mathcal{Y})$) or sets ($\mathcal{Y}' = \mathcal{S}(\mathcal{Y})$)[1]. Structural assumptions about h are collected in the formal condition $\mathcal{A}$ encoding ***inductive biases***. $\mathcal{A}$ gives rise to the ***hypothesis space***

$$\mathcal{H} = \{h : \mathcal{X} \to \mathcal{Y}' \mid \mathcal{A}\} \tag{1}$$

of all admissible functions[2]. We consider parametric hypotheses, i.e., for given $\mathcal{A}$, each hypothesis is fully determined by a ***parameter*** vector $\boldsymbol{\omega} \in \boldsymbol{\Omega}$ (usually, $\boldsymbol{\Omega} = \mathbb{R}^D$). We sometimes write $h_{\boldsymbol{\omega}}$ to emphasize this relationship, but we will also refer to parameterizations as hypotheses in a *pars pro toto* spirit. For now, we take $\boldsymbol{\omega}$ to be unknown but fixed; later, we will treat it as the realization of a RV $\Omega \sim p_\Omega$. Our definition of the hypothesis space thus extends to

$$\mathcal{H} = \{h_{\boldsymbol{\omega}} : \mathcal{X} \to \mathcal{Y}' \mid \mathcal{A} \wedge \boldsymbol{\omega} \in \boldsymbol{\Omega}\}. \tag{2}$$

[1] We will discuss in Sec. 2.1.2 that X is not always sufficient to fully explain Y, such that h^* may be a non-deterministic function. Still, we *observe* labels as realizations from $\mathcal{Y}$.

[2] For instance, we might restrict the hypothesis space to constant functions via $\mathcal{A}$: $h(\boldsymbol{x}) = c \in \mathbb{R}$.

Each ***prediction*** $h_{\boldsymbol{\omega}}(\boldsymbol{x})$ incurs a point-wise ***loss*** $L : \mathcal{Y} \times \mathcal{Y}' \to \mathbb{R}$ quantifying deviations from the observed label. The ***risk*** $\mathcal{R}$ is the expected loss for an arbitrary sample from p_{XY}. Lacking access to the DGP, we resort in practice to an empirical version computed from the observed data:

$$\mathcal{R}_{\text{emp}}(h_{\boldsymbol{\omega}}) = \sum_{i=1}^{N} L\left(y, h_{\boldsymbol{\omega}}(\boldsymbol{x}^{(i)})\right). \tag{3}$$

Training in supervised learning mainly occurs by ***empirical risk minimization (ERM)***. Given data $\mathcal{D}$, a hypothesis space $\mathcal{H}$, risk, and some optimization procedure, the ***learner*** (endowed with ***hyperparameters*** that are predefined rather than determined during training) produces a risk-optimal hypothesis according to the following optimization problem:

$$\min_{h_{\boldsymbol{\omega}} \in \mathcal{H}} \left\{\mathcal{R}_{\text{emp}}(h_{\boldsymbol{\omega}}) + \lambda \Delta(h_{\boldsymbol{\omega}})\right\}. \tag{4}$$

Δ denotes a ***regularization*** term that can be chosen to discourage certain hypotheses with severity $\lambda \geq 0$. A common choice is $L2$ regularization with $\Delta = \|\boldsymbol{\omega}\|_2^2$, penalizing parameters of large magnitude. Regularization enforces what Wilson (2025) calls *soft* inductive biases, nudging the optimizer toward preferred solutions (for $\lambda < \infty$). This is in contrast to the *hard* inductive biases $\mathcal{A}$ that rule out some hypotheses completely. Regularization usually favors simple, sparse solutions in the hope of avoiding ***overfitting*** to the training observations (e.g., Murphy, 2012). Ultimately, we wish to balance sufficiently expressive hypotheses and regularization to achieve ***generalization*** to unseen ***test*** data.

Deep Learning This thesis focuses on ***deep learning (DL)*** with ***neural networks (NNs)***. We only introduce some additional notation here (for a comprehensive introduction to DL, see, e.g., Bishop and Bishop, 2024). Loosely following Murphy (2022), we cast NNs as feature extractors with a predictor on top. The feature extractor $\phi : \mathcal{X} \to \mathbb{R}^{D-\mathcal{L}}$ learns a ***latent representation*** of the data and is parameterized with $\boldsymbol{\omega}_{-\mathcal{L}}$. NNs realize composite, highly non-linear transformations organized in ***layers***. We refer to any layers between the first (input) and the last (output) layer as ***hidden***. Early NNs were arranged as ***multi-layer perceptrons (MLPs)***, where the feature extractor is given by consecutive layers of fully-connected ***neurons***. Each neuron in layer ℓ computes a linear combination $\boldsymbol{W}^{\ell}\boldsymbol{z}^{\ell-1} + \boldsymbol{b}^{\ell}$ of the outputs $\boldsymbol{z}^{\ell-1}$ of all incoming neurons, using trainable ***weight*** matrices $\boldsymbol{W}$ and ***bias*** vectors $\boldsymbol{b}$, and applies a subsequent (non-linear) ***activation function***. Modern architectures include other layer types (e.g., extracting spatial image features) and more varied ways of connecting neurons. We collect all NN parameters[3] in $\boldsymbol{\omega} = (\boldsymbol{\omega}_{-\mathcal{L}}, \boldsymbol{\omega}_{\mathcal{L}}) \in \mathbb{R}^D$; these can easily number in the billions. The full model computes

$$h_{\boldsymbol{\omega}}(\boldsymbol{x}) = \tau_{\boldsymbol{\omega}_{\mathcal{L}}}(\phi_{\boldsymbol{\omega}_{-\mathcal{L}}}(\boldsymbol{x})), \tag{5}$$

where $\tau_{\boldsymbol{\omega}_{\mathcal{L}}} : \mathbb{R}^{D-\mathcal{L}} \to \mathcal{Y}'$ [4]. Since NNs hardly ever admit analytical solutions, training is based on numerical optimization—frequently, some form of ***stochastic gradient descent (SGD)*** on subsets (***minibatches***) of the training data (Bishop and Bishop, 2024).

[3] We sometimes use the terms *parameters* and *weights* interchangeably, no longer distinguishing between $\boldsymbol{W}$ and $\boldsymbol{b}$.

[4] For instance, in classification with distributional predictions, $\tau_{\boldsymbol{\omega}_{\mathcal{L}}}$ might be the composition of a linear transformation with coefficients $\boldsymbol{\omega}_{\mathcal{L}}$ and the ***softmax*** function mapping to numbers in $[0, 1]$ that sum to 1.

2.1.2. Predictive Uncertainty

Defining Uncertainty Given the elusive nature of uncertainty, it is perhaps not surprising that there is no unifying definition in the context of ML[5]. We will regard uncertainty under a lens of ***multiplicity***. Informally speaking, we mean by multiplicity that there is more than one ***plausible*** option; most notably, in predictions and hypotheses. Senge et al. (2014) define plausibility for hypotheses as *compatible* with and *strongly supported* by the data. Note that such statements require some form of **UR**. For instance, compatibility might translate to non-zero probability of occurrence in a distributional framework. This point will be revisited in Sec. 2.2. For now, we adopt the following informal definition, where ***outcomes*** are elements in some space of interest:

> ***Uncertainty*** State of multiplicity in plausible outcomes, where plausibility must be established through a representational framework.

At this point, one might ask if we could not rid ourselves of all uncertainty by building models powerful enough to refute all but the correct option. In the words of Cuzzolin (2021), is uncertainty not "simply a fig leaf for our ignorance and lack of understanding of nature phenomena"? We believe that eliminating uncertainty altogether is neither possible nor desirable. Even if we could become Laplace's omniscient demon, one appeal of modeling is to achieve a level of abstraction from the underlying process, allowing us to predict something close to the truth at a fraction of the cost it would take to be exact. What is more, Heisenberg's principle implies that uncertainty is not just a human shortcoming but a fundamental part of our physical world.

For this thesis, we are mainly interested in ***predictive uncertainty***, which refers to a single instance of observed features and some hypothesis effecting its prediction:

> ***Predictive uncertainty*** State of multiplicity in plausible values for the target Y, given features $X = \boldsymbol{x}$ and a hypothesis $h \in \mathcal{H}$, on the level of a single observation.

This is a decidedly prediction-centered perspective on uncertainty. In other contexts, like classical statistics, modelers also care about uncertainty surrounding (interpretable) model coefficients.

Sources of Uncertainty Multiplicity in outcomes can arise at different stages of a learning process (e.g., Psaros et al., 2023). Crucially, any statement of uncertainty only captures such multiplicity as is explicitly accounted for and must be treated contingent on all other quantities. The literature has not quite reached a consensus on the appropriate terminology and relevant sources of uncertainty (Gruber et al., 2025). That said, most works have some version of a bi-causal understanding (e.g., Hüllermeier and Waegeman, 2021) which we lay out by means of an example.

Consider a die that is not necessarily fair. The task is to predict the pips after rolling the die with options in $\mathcal{Y} = \{1, \ldots, 6\}$. A featureless predictor will observe a number of tosses and predict according to the empirical frequencies (corresponding to the left graph in Fig. 2.1). No matter how many repetitions are recorded, we could never be sure what the die will show before it is rolled. This type of multiplicity is called ***aleatoric uncertainty (AU)***, the name originating, tellingly,

[5] Other fields have committed to definitions suited to their respective goals. Insurance theory, for instance, distinguishes coverage-invoking events under *risk* (known outcomes and event probabilities) and *uncertainty* (one or both unknown; Dionne and Harrington, 1992). Our purposes require more general notions.

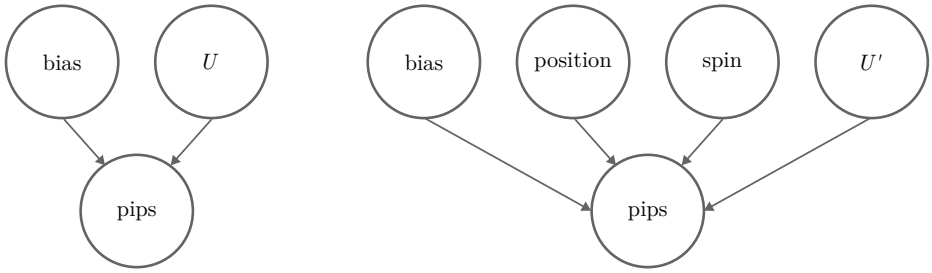

Figure 2.1.: Modeling dicing outcomes. *Left:* Graph corresponding to a featureless predictor, where only the bias of the die, governing the empirical frequencies of each outcome, and a stochastic nuisance variable U capturing randomness are used. *Right:* Graph corresponding to an enhanced predictor consulting position and spin as additional features.

from the Latin *alea* (die, game of dice; Lewis and Short, 1879). We will highlight it by **AU** as one of the key terms in this thesis. Despite some terminological debate, most definitions agree on *irreducibility* being the distinguishing property of the **AU**—even, like in the dicing example, given infinite data (Hüllermeier and Waegeman, 2021).

The overall uncertainty can be reduced with a more elaborate dataset. Suppose we could observe the initial position of the die as well as the spin of the toss (right graph in Fig. 2.1). This new setup encompasses different sources of uncertainty. First, we realize that, in the frequencies-only model, parts of the uncertainty came from omitting informative variables like the position and spin. Second, we can imagine that measuring such quantities can only be realized with finite precision and might well be fraught with measurement error. Gruber et al. (2025) collect these cases in an unobserved RV Z that can take on different roles. In this work, we posit the **AU** as an inherent property of the DGP creating the *observed* data. The full process of tossing a die—including all relevant physical quantities—may be deterministic. The DGP of the observed data, p_{XY}, however, is not generally the DGP of the true phenomenon, p_{XYZ}. Fundamentally, this means that the **AU** depends on the choice of features and targets, i.e., on $\mathcal{X}, \mathcal{Y}$ and p_{XY}.

> ***Aleatoric uncertainty*** State of multiplicity in plausible values for the target Y, given features $X = \boldsymbol{x}$; denoted by **AU** and irreducible with more observations from p_{XY}.

Accepting that some baseline **AU** will generally remain, even with perfect knowledge about p_{XY}, we need to confront the multiplicity in *modeling* this mapping. There are usually several plausible hypotheses compatible with and supported to some degree by the observed data. This type of multiplicity is called ***epistemic uncertainty (EU)*** in a nod to the Greek word *epistēmē* (knowledge, understanding; Steup and Ram, 2024). We will highlight it as **EU**. In the notation of Sec. 2.1.1, **EU** may concern both the structural assumptions $\mathcal{A}$ and the values of parameters $\boldsymbol{\omega}$. For a meaningful notion, we need to fix $\mathcal{H}$ besides $\mathcal{X}, \mathcal{Y}$ and p_{XY}[6].

> ***Epistemic uncertainty*** State of multiplicity in hypotheses h that could have plausibly generated the observed data $(X = \boldsymbol{x}, Y = y)$; denoted by **EU**.

[6]This is because many models rely on *derived* features that differ across model classes, such that modifying $\mathcal{H}$ effectively changes $\mathcal{X}$ (for instance, a simple linear model can only make use of the features as-is, whereas NNs *map* features to highly non-linear representations; cf. Fig. 5 in Hüllermeier and Waegeman, 2021).

Complementing the notion of **AU** being constant for a given setting $(\mathcal{X}, \mathcal{Y}, p_{XY}, \mathcal{H})$, **EU** is usually framed as *reducible* by gathering information (e.g., observing more data; Hüllermeier and Waegeman, 2021). In order to understand which information can reduce **EU** and to what degree, we need to consider its different aspects. ***Structural*** uncertainty (Jiménez et al., 2025, therein called *model* uncertainty) arises from model misspecification (i.e., $\mathcal{H}$ does not contain the true hypothesis) and is notoriously hard to capture, let alone reduce. ***Distributional*** uncertainty (Malinin and Gales, 2018; Jiménez et al., 2025) relates to mismatches between p_{XY} and the data distribution at the time of model deployment. Short of prescience, it is at best partly reducible unless we make strong assumptions on the shifts that can occur. Lastly, ***estimation*** uncertainty can prevent us from learning h^* even with sufficiently expressive $\mathcal{H}$ and congruent distributions. Jiménez et al. (2025) identify two different causes: limited *data* and non-deterministic *procedures*. While the former can be remedied by observing more data, the latter is harder to resolve. Parts of the procedural uncertainty arise *by necessity*, as a consequence of limited resources: time (e.g., finitely many iterations), computing power (e.g., approximate operations), or memory (e.g., data minibatching). Interestingly, procedural uncertainty can also occur *by design* with models that induce a non-bijective mapping between parameters and hypotheses (Huang et al., 2023). In such scenarios, when $\boldsymbol{\omega}, \boldsymbol{\omega}' \in \boldsymbol{\Omega}$ parameterize the exact same function, we cannot know for sure which of the equivalent parameterizations will be chosen. We will discuss this phenomenon in detail for NNs in Sec. 2.2.4. Fig. 2.2 summarizes the sources of uncertainty as listed above.

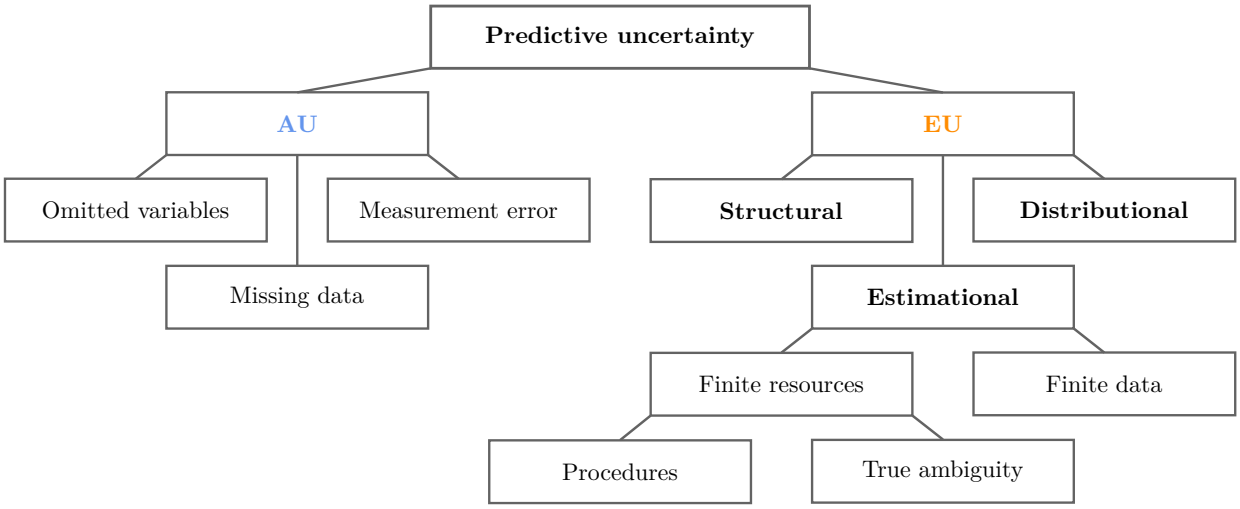

Figure 2.2.: Sources of uncertainty as discussed in this thesis.

It should be clear by now just how complex the picture on predictive uncertainty can become. Still, the above definitions only start in the middle of the modeling process, essentially treating the observed data as fixed[7]. For a fully comprehensive picture, it would further be necessary to address multiplicity in the way the data have been collected and preprocessed (for a thorough discussion, see Hoffmann et al., 2021).

[7] Ignoring sources is a deliberate choice that can be perfectly justified (as stated in the beginning, all uncertainty that is to be estimated also needs to be represented) but is rarely made explicit.

2.1.3. Uncertainty Estimation Problem

Working Assumptions The previous section mentioned numerous potential sources of uncertainty in practical applications. For the remainder of the text, we consider the following setting unless stated otherwise. We assume—in the terminology of Bernardo and Smith (1994)—an $\mathcal{H}$-***closed*** view, meaning that $\mathcal{H}$ contains the true hypothesis and there is *no structural uncertainty*[8]. As Hüllermeier and Waegeman (2021) argue, this is not necessarily a drastic restriction for sufficiently expressive $\mathcal{H}$. Without loss of generality, we take the structural inductive biases $\mathcal{A}$ to be a singleton condition describing just one model class, and confine all **EU** to the parametric sphere[9]. Effectively, the strength of the $\mathcal{H}$-closed assumption then depends on $\mathcal{A}$ (Draper, 1995). Furthermore, we will address specific aspects of *distributional* uncertainty in Sec. 2.1.4 and 2.5.2, but for the most part, the contributions in this thesis concern the *estimation* share of the **EU**.

To summarize, we treat all uncertainty that can be reduced by gathering more information as **EU**, where the context $(\mathcal{X}, \mathcal{Y}, p_{XY}, \mathcal{H})$ remains fixed. **EU** expresses imperfect knowledge about h^* and decreases all the way to zero with infinite data if we assume no distribution shifts, unlimited resources, and uniquely specified hypotheses. Any uncertainty that is not reducible in this sense even with the true mapping h^* is **AU**, and thus solely a property of the DGP p_{XY}. This distinction reflects an important assumption, namely, the treatment of **AU** and **EU** as *independent* sources (the former relating to the data; the latter, to the modeling process). In other words, we suppose that total uncertainty (**TU**) decomposes *additively* (Hüllermeier and Waegeman, 2021):

$$\textbf{TU} = \textbf{AU} + \textbf{EU}. \tag{6}$$

This premise will be revisited in Sec. 2.3.

Uncertainty Estimation Problem Equipped with the above working definition, we consider again the central ***uncertainty estimation*** problem studied in this thesis (cf. Fig. 1.2). The goal of **UE** is to obtain trustworthy estimates of predictive uncertainty for a given test observation. As explained in Sec. 1, **UE** is a two-step procedure. First, we need a way to ***represent*** uncertainty by allowing for multiplicity in appropriate places. This implies a choice of sources to be covered in the spirit of the previous section. Subsequently, we look for summary statistics that ***quantify*** the level of uncertainty in a given representation.

UR	Expressing multiplicity through appropriate representational objects.

UQ	Describing representations of multiplicity by appropriate measures.

The lack of an observable ground truth to emulate makes progress in **UE** hard to judge. We discuss evaluation protocols in Sec. 2.4. For now, suffice it to say that models should *faithfully* estimate uncertainty relative to their predictive performance. This means, in particular, that they are neither *overconfident* (too progressive; reporting lower uncertainty than performance calls for) nor *underconfident* (too conservative; reporting excessive uncertainty).

[8] This understanding is more narrow than what is discussed by Gruber et al. (2025): we care only about whether the DGP p_{XY} of the *observed* data, not the actual latent process p_{XYZ}, is covered by $\mathcal{H}$. This allows us to clearly frame any remaining uncertainty over $Y|X = \boldsymbol{x}$ given the true model h^* as **AU**. Work specifically investigating misspecification includes Cervera et al. (2021); Kato et al. (2022); Masegosa (2020).

[9] We can make this model class arbitrarily complex; for instance, formulating it as a sum over parameterized sub-classes whose coefficients can be selectively set to zero so as to recover only a single sub-class.

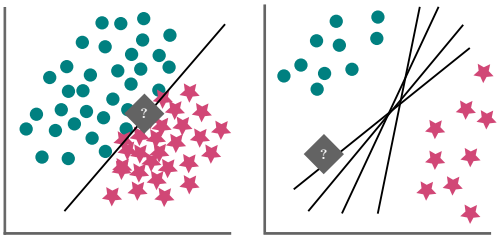

Figure 2.3.: Idealized example cases for sources of uncertainty in binary classification. *Left:* High **AU** but low **EU** for a test point (gray diamond) that falls close to the relatively unambiguous decision boundary between the two classes (teal-colored circles *vs* pink stars). *Right:* High **EU** for the test point due to many hypotheses supported by the data in the given area; the **AU** is fairly low because most hypotheses place the test point into the class of circles. Adapted from Hüllermeier and Waegeman (2021).

UE should account for both sources of multiplicity. For one, omitting uncertainty components carries a risk of models becoming overconfident (Wilson and Izmailov, 2020). On the other hand, the attribution to **AU** and **EU** can help understand learning dynamics, support evaluation, and inform downstream decisions (Sec. 2.4, 2.5; Mucsányi et al., 2023). Fig. 2.3 depicts for a toy example the respective cases of high **AU** and **EU** in a test observation.

2.1.4. Beyond *i.i.d.* Data

In most ML research scenarios, where it is desirable to use all available data for model training, methods are developed and evaluated using some form of train-test split $\mathcal{D} = (\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}})$ such that $\mathcal{D}_{\text{test}} \sim p_{XY}$ (possibly, constructing several pairs by repeated resampling; Bischl et al., 2023). At the same time, it is rarely disputed that most practical applications do not obey an *i.i.d.* process. Rather, we face some form of ***distribution shift*** where $\mathcal{D}_{\text{test}} \nsim p_{XY}$ (e.g., Nickel, 2024). Generalization to unseen data must be expected to worsen when distribution shift is present because key modeling assumptions are violated. **UE** bears an interesting relationship to distribution shifts. We might experience a deterioration in **UE** quality if model properties that hold for in-distribution data do not carry over[10], but we can also put the cart before the horse and use **UE** to *detect* distribution shifts. A test observation from outside the training distribution, the thinking goes, should provoke strong multiplicity in plausible target values (e.g., Ovadia et al., 2019). We discuss how this fact is used for evaluating **UE** in Sec. 2.4.2.

The idea of detecting distribution shifts by uncertainty relies on the model finding the ***out-of-distribution (OOD)*** point unusual. As Li et al. (2025) argue, however, this might be asking too much from models trained on in-distribution data. Relying on **UE** to identify OOD observations becomes especially problematic when models pick up patterns in unforeseen ways. For instance, a model trained on animal images might confidently classify the image of a golf ball (OOD) as "cow" because it has learned to associate grass in the background with cows, which is not unreasonable but brittle. Prediction from such ***spurious*** features is called ***shortcut learning (SCL)***. SCL can be seen as an umbrella term for some previously disconnected, mysterious-seeming phenomena (e.g., adversarial attacks, cf. Fig. 1.1; Steinmann et al., 2024). The spurious patterns may reflect real-world correlations (like cows and grass) or come as an artifact of the modeling process

[10] **UE** under the Bayesian paradigm (Sec. 2.2.2), for instance, is not guaranteed to remain consistent under shifts since a correctly specified likelihood is a key assumption (Izmailov et al., 2021; Knoblauch et al., 2022).

(like models learning to associate medical conditions with image tokens from a specific provider of imaging equipment). They are often subtle and can affect different data modalities (Geirhos et al., 2020). [**C5**] provides evidence that SCL can indeed affect **UQ** (Sec. 2.3).

Despite the ubiquity of distribution shifts, **UE** with non-*i.i.d.* data is underexplored and often treated like an edge case (Zhou and Levine, 2021; Xiao et al., 2021; Rudner et al., 2024). Many voices have been arguing for years that robust generalization requires a *causal* perspective (Xia et al., 2021; Schölkopf et al., 2021; Binkyte et al., 2025). Even without explicit causal frameworks, relatively practical-seeming changes like training on data from multiple environments encoding the shift seem to help (Wald et al., 2021). We will leave the discussion at that because the problems addressed in this thesis are present in the *i.i.d.* setting already. Still, we emphasize that adopting a causal view might prove indispensable for the real-world deployment of ML models.

2.2. Uncertainty Representation

This chapter introduces schemes of uncertainty representation based on the fundamental concept of *multiplicity*. We discuss *Bayesian* learning, as the key representational tool, in detail and put special focus on inference for *Bayesian neural networks* to motivate the work of [**C1**] and [**C2**].

2.2.1. Representation by Multiplicity

Multiplicity

We defined uncertainty in Sec. 2.1.2 as a "state of multiplicity in plausible outcomes, where plausibility must be established through a representational framework". Such multiplicity is important in two different places. First, right at the core of our task, predictive uncertainty requires ambiguity in *predictions*. This provides information about which target values are plausible to which degree given our current state of knowledge. Second, we need to account for the presence of multiple *hypotheses*. Since we cannot be entirely sure which hypothesis has created the observed data, we want to incorporate the ones that cannot be ruled out. Expressing uncertainty over hypotheses, in particular, is different in spirit from standard ERM which refutes all but the loss-optimal one. Clearly, multiplicity in hypotheses exists on a level *above* the multiplicity of predictions in the sense that the former propagates to the latter (see Fig. 2.4). Recalling the components of uncertainty, **TU** corresponds directly to the multiplicity in predictions. **AU** is a property of the DGP and as such independent of the modeling process. As we will see later, it can only be approximated in a conditional notion for a given hypothesis. **EU** surrounds the degree to which multiplicity in hypotheses translates to multiplicity in predictions[11].

There are two straightforward ways of realizing multiplicity in a mathematical sense (Hüllermeier and Waegeman, 2021). The more general is *sets* of outcomes that simply enumerate plausible options. In the dicing example from the previous chapter, this might be the set of predictions $\{1, 3, 5\}$. Intuitively, the *size* of the set expresses the level of uncertainty. We can also cast outcomes as

[11] Note that these descriptions pertain to individual observations. Several hypotheses can have plausibly generated the whole of the observed data, while for a specific instance, the situation is fairly unambiguous (cf. Fig. 2.3).

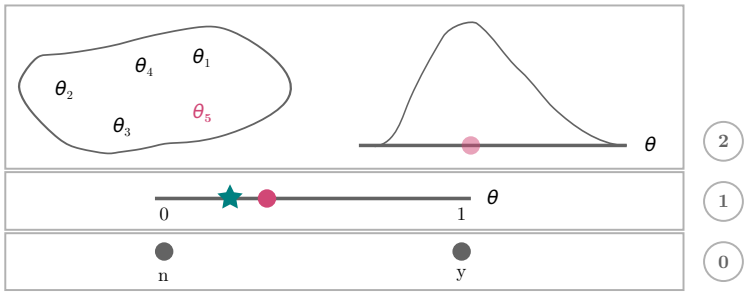

Figure 2.4.: Bi-level uncertainty representation. On level 0, we can only *observe* hard labels ∈ {y, n}. Level 1 represents a multiplicity in *predictions* by the probability θ of the label being "y". The pink dot marks the point-estimate hypothesis, slightly deviating from the ground truth (teal-colored star). Depending on the choice of how to express multiplicity in *hypotheses*, we have a *set* of (*left*) or a *distribution* over (*right*) hypotheses on level 2. Adapted from [C3].

realizations of some RV, thus obtaining a *distribution* assigning probabilities of occurrence to each outcome. In the dicing experiment, we might predict a probability vector $(0.6, 0, 0.2, 0, 0.2, 0)$ for pips 1 through 6. Like the predictive set, it rules out digits $2, 4, 6$ as impossible but also signals that 1 is the most likely outcome. The additional information is bought at the expense of stronger assumptions. Robert (2007) calls this the "probabilization of uncertainty, that is, (...) an axiomatic reduction from the notion of unknown to the notion of random". The level of uncertainty manifests in the *shape* of the distribution: more dispersed (concentrated) distributions imply higher (lower) uncertainty.

Representation Schemes

In theory, we could define representation schemes from arbitrary combinations of sets and distributions on both levels of multiplicity. Fig. 2.5 shows a taxonomy (partly inspired by Hofman et al., 2024) for representations used in practical applications.

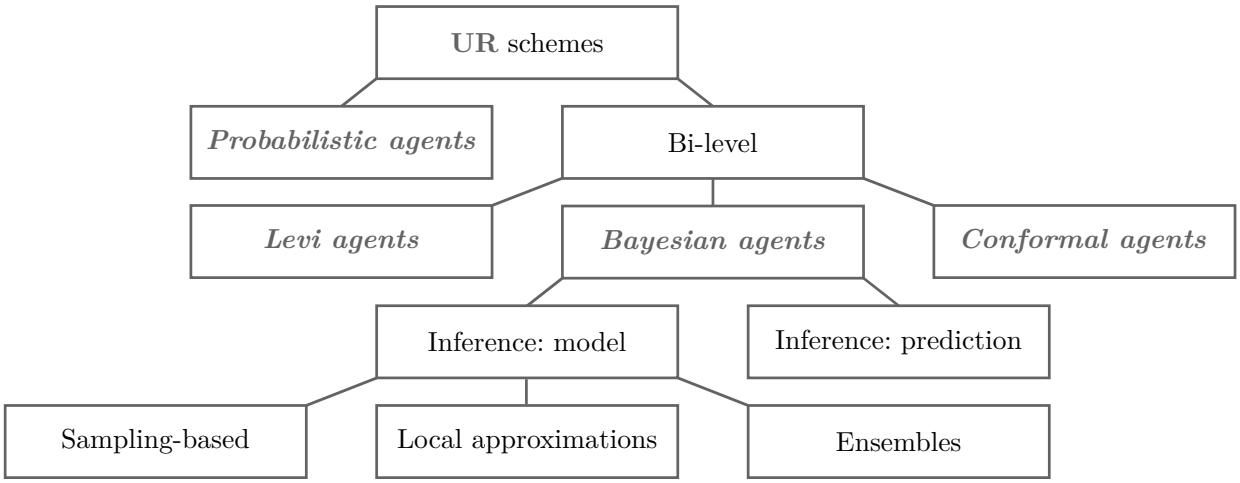

Figure 2.5.: Taxonomy of **UR** schemes as discussed in this thesis. Inspired by Hofman et al. (2024).

Probabilistic agents, in the terms of Hofman et al. (2024), are expressly single-level in that they only consider multiplicity of predictions without accounting for the **EU**. This includes classifiers using a single model to output categorical distributions (e.g., NNs with softmax activation in the final layer). For regression, an example is to have NNs learn mean and variance of a Gaussian predictive distribution (Nix and Weigend, 1994). It was mainly these probabilistic agents that sparked research in **UQ** by proving overconfident (famously pointed out in Guo et al., 2017). Since then, it has become consensus that this failure mode is at least partly due to omitting the **EU** (e.g., Wilson and Izmailov, 2020).

Therefore, we are mainly interested in bi-level representations. The most common types are ***Levi agents***, which express multiplicity in hypotheses by credal sets and multiplicity in predictions by distributions, and ***Bayesian agents*** with distributional representations on both levels. The latter are the focus of this thesis. We can further distinguish how multiplicity in hypotheses is realized by Bayesian agents. Broadly speaking, Bayesian learning involves a distribution over hypotheses which is updated according to the observed data (more details in Sec. 2.2.2). While a special class of Dirichlet-based models (e.g., Charpentier et al., 2020; Kopetzki et al., 2021) directly infers the parameters of the predictive distribution, most methods attach probabilistic beliefs to the parameters of the *model* that produces the predictive distribution. Further down, different approaches exist to compute the involved—usually intractable—distributions in this second category. We discuss ***approximate Bayesian inference (ABI)*** methods in Sec. 2.2.3.

While the previously listed frameworks use distributions to express multiplicity in predictions, a number of approaches that we call ***conformal agents*** (for lack of a better name) lead to set-valued predictions. They are united by the promise of coverage guarantees without explicit distributional assumptions and include conformal prediction as well as methods based on data resampling (e.g., bootstrapping; Barber et al., 2020, 2021; Angelopoulos and Bates, 2023). Yet other approaches (for example, expressing uncertainty via distances in latent representation space; van Amersfoort et al., 2020) exist outside the above taxonomy. For a fairly comprehensive overview see, e.g., Gawlikowski et al. (2023).

2.2.2. Bayesian Learning Paradigm

Bayesian learning is one of the big paradigms in statistics. It professes to be principled and well-rooted in theory, leading to coherent systems and rational decision rules in many frameworks (Gelman et al., 2021). At the core of Bayesian statistics is the interpretation of model parameters[12] as *random*, in contrast to the frequentist view of them as unknown but fixed quantities.

Inference Bayesian learning starts from a ***prior*** belief that encodes *a priori* (i.e., before observing any data) information over the parameters $\Omega \sim p_\Omega$, where realizations $\boldsymbol{\omega}$ have prior density $q(\boldsymbol{\omega})$. While the possibility to encode actual domain knowledge in the prior is an oft-cited appeal of Bayesian learning, the prior can simply be a vehicle to incorporate inductive biases that would be framed as *regularization* in ERM[13]. For instance, a zero-mean Gaussian prior encodes a belief

[12] We assume a *weight-space* view with parametric hypotheses $h_{\boldsymbol{\omega}}$, such that inference over hypotheses is equivalent to inference over parameters. This follows from the assumption of a singleton structural condition $\mathcal{A}$ made in Sec. 2.1.3; we will briefly touch on the complementary, *function-space* view in Sec. 2.2.4.

[13] In particular, the Bayesian view does not claim that parameters are actually products of random processes, nor does it preclude the existence of a true parameter $\boldsymbol{\omega}^*$ that has generated the observed data.

about parameter sparsity. As Robert (2007) puts it, "The choice of a prior distribution does not require any kind of belief in this distribution. (...) [The prior] should rather be considered either a tool that provides a single inferential procedure with acceptable frequentist properties (...), or a way to summarize the available prior information and the uncertainty surrounding this information". ***Inference*** consists of updating the prior to a ***posterior*** density according to Bayes' theorem, with information about the observed data encoded in the ***likelihood*** $p(\mathcal{D}|\boldsymbol{\omega})$:

$$p(\boldsymbol{\omega}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\omega})q(\boldsymbol{\omega})}{\int_{\boldsymbol{\Omega}} p(\mathcal{D}|\boldsymbol{\omega})q(\boldsymbol{\omega})\,\mathrm{d}\boldsymbol{\omega}} = \frac{p(\mathcal{D}|\boldsymbol{\omega})q(\boldsymbol{\omega})}{p(\mathcal{D})}. \tag{7}$$

This updating rule remains consistent in a sequential manner when observing new data, taking the former posterior as the new prior belief (Gelman et al., 2021). If the prior is specified so as to attach positive probability to the true model, the posterior will *contract* on h^* given enough data (see Fig. 2.6), meaning we are guaranteed accurate inference in the infinite limit (Wilson and Izmailov, 2020). There is a well-established link between this updating rule and regularized ERM. Generalizing previous results, Knoblauch et al. (2022) show that Bayesian inference can be viewed as a special case of what they call the "Rule of Three". When optimizing over the space $\mathbb{P}(\boldsymbol{\Omega})$ of all probability measures on the parameters, minimizing the ***negative log-likelihood (NLL)*** loss[14], and penalizing the ***Kullback-Leibler (KL) divergence*** D_{KL} quantifying deviations from the prior q, we recover the Bayes posterior as

$$\begin{aligned} p(\boldsymbol{\omega}|\mathcal{D}) &\in \underset{Q\in\mathbb{P}(\boldsymbol{\Omega})}{\arg\min} \left\{ \mathbb{E}_{Q(\boldsymbol{\omega})}\left[-\log p(\mathcal{D}|\boldsymbol{\omega})\right] + D_{\mathrm{KL}}(Q\|q) \right\} \\ &= \underset{Q\in\mathbb{P}(\boldsymbol{\Omega})}{\arg\min} \left\{ \mathbb{E}_{Q(\boldsymbol{\omega})}\left[-\log p(\mathcal{D}|\boldsymbol{\omega})\right] + \mathbb{E}_{Q(\boldsymbol{\omega})}[-\log q(\boldsymbol{\omega})] + \mathbb{E}_{Q(\boldsymbol{\omega})}[\log Q(\boldsymbol{\omega})] \right\}. \end{aligned} \tag{8}$$

The regularizing nature of the prior becomes evident in the D_{KL} term. Note that replacing the KL divergence by a scalar-valued penalty leads to a (Dirac mass on a) point estimate, amounting to standard regularized ERM (Knoblauch et al., 2022). The full Bayesian formulation can also be translated to a ***maximum a posteriori (MAP)*** point estimate given by the mode of the posterior density (not necessarily coinciding with the ERM result; Murphy, 2012).

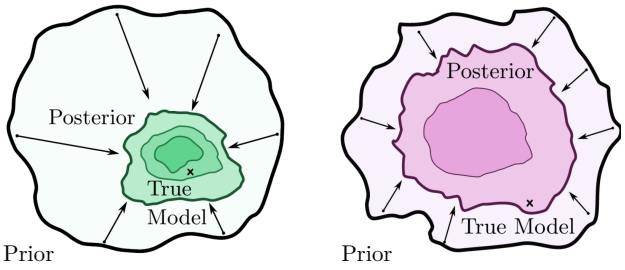

Figure 2.6.: Posterior contraction. Bayesian updating causes the posterior density to concentrate relative to the prior belief. *Left:* efficient contraction, leading to a sharp posterior. *Right:* weak concentration, leading to a vague posterior. In both cases, the model is correctly specified, such that the true model could be identified with sufficient data. Adapted from Wilson and Izmailov (2020).

[14] Since maximizing the log-likelihood is equivalent to minimizing the NLL, the latter can be viewed as a loss function for this classical optimization principle. The standard choices in DL of cross-entropy for classification and $L2$ for regression, if the conditional distribution of the target follows a Gaussian distribution, are NLL-type loss functions (Murphy, 2012).

Predictions Equipped with the posterior density, we obtain a full predictive distribution for any new observation $(\boldsymbol{x}_+, y_+)$. This gives rise to the ***posterior predictive density (PPD)***:

$$p(y_+|\boldsymbol{x}_+, \mathcal{D}) = \int_{\boldsymbol{\Omega}} p(y_+|\boldsymbol{x}_+, \boldsymbol{\omega}) p(\boldsymbol{\omega}|\mathcal{D}) \, \mathrm{d}\boldsymbol{\omega}. \tag{9}$$

This process of taking the expectation over all possible hypotheses, weighted according to their posterior probability, is called ***Bayesian model averaging (BMA)***. From the BMA we can derive predictive uncertainty. In accordance with our previous definitions, the **TU** expresses multiplicity in predictions, so it is equated with the uncertainty surrounding $p(y_+|\boldsymbol{x}_+, \mathcal{D})$. The **AU**, as a property of the latent DGP, can only be estimated. We approximate it as the expected uncertainty of a single hypothesis $p(y_+|\boldsymbol{x}_+, \boldsymbol{\omega})$. Marginalization over individual parameters $\boldsymbol{\omega}$ is governed by our posterior belief about which hypothesis is most likely to have generated the observed data: we assess the uncertainty of $p(y_+|\boldsymbol{x}_+, \boldsymbol{\omega})$ for each $\boldsymbol{\omega}$ (i.e., the multiplicity remaining in the target values, given features $\boldsymbol{x}$, if $\boldsymbol{\omega}$ were actually the true hypothesis), and aggregate the individual uncertainties according to $p(\boldsymbol{\omega}|\mathcal{D})$. The **EU** reflects how much the plausible hypotheses $p(y_+|\boldsymbol{x}_+, \boldsymbol{\omega})$ differ across $\boldsymbol{\omega}$[15]. Unfortunately, the posterior density is rarely available in closed form due to the intractable integral (***evidence***) in the denominator of Eq. (7). The best we can hope for is to collect a finite number S of posterior samples by means of ABI (Sec. 2.2.3). With samples instead of a full distribution, the BMA is approximated by ***Monte Carlo integration***, converging almost surely to Eq. (9) for $S \to \infty$ (Andrieu et al., 2003; Robert, 2007):

$$p(y_+|\boldsymbol{x}_+, \mathcal{D}) \approx \frac{1}{S} \sum_{s=1}^{S} p(y_+|\boldsymbol{x}_+, \boldsymbol{\omega}_s). \tag{10}$$

Criticism Note that there are already two approximations in place: drawing samples from a distribution that might not be the true posterior, and combining finitely many of them to approximate the PPD. Indeed, in what Ghahramani (2008) dubbed the "ABI conundrum", the use of such approximations risks invalidating the coherence arguments of Bayesian statistics (Farquhar, 2022; Knoblauch et al., 2022; Pituk et al., 2025). More recently, there have been attempts to circumvent these problems and estimate the PPD directly (Hollmann et al., 2023; Nagler and Rügamer, 2025). Even precise Bayesian inference is not without criticism (e.g., Gelman and Yao, 2021). In particular, it is unclear how we should represent a situation of total ***ignorance*** (Senge et al., 2014). The least informative prior belief in the Bayesian paradigm is a uniform distribution, which is often justified by Laplace's "principle of insufficient reason"—in the absence of any evidence to the contrary, it is rational to spread the probability mass evenly across all possible outcomes (Seidenfeld, 1986; Gelman et al., 2021). This is not the same as actual ignorance; the uniform distribution might as well reflect perfect knowledge that the outcomes are equiprobable. Dubois et al. (1996) argue that (Bayesian) distributional settings are fundamentally tied to situations of decision-making[16] and that representing ignorance requires more general belief frameworks. We will discuss in Sec. 2.3 how undesired behavior can occur in **UQ** with Bayesian distributional representations. Yet, despite its limitations, the Bayesian paradigm remains one of our best shots at coherent inference from data (Papamarkou et al., 2024). As mentioned before,

[15] We will see in Sec. 2.3 that it is not quite clear which shape of $p(\boldsymbol{\omega}|\mathcal{D})$ should correspond to maximum uncertainty.

[16] Before observing any throws in the dicing example, many people would presumably agree that the probability of throwing a 1 is smaller than throwing a number between 2 and 6. This, however, reflects not genuine ignorance, which would make such a statement impossible, but rather how we should act in a fictitious betting situation where it is irrational to assume any other than a uniform probability distribution.

proponents of the Bayesian idea emphasize the possibility of incorporating domain knowledge, if we have it, into the prior belief (e.g., Cinquin and Bamler, 2025). The sequential nature of the consistent updating rule lends itself to practically relevant settings like continual or federated learning (e.g., Zhang et al., 2022). Bayes' theorem further provides methods for model selection (for example, via the evidence; Piironen and Vehtari, 2017). Most important, Bayesian learning handles predictive uncertainty in a principled manner. We get access to an entire predictive distribution, and both levels of uncertainty are represented in the PPD.

2.2.3. Approximate Bayesian Inference

Sampling-Based Inference

The intractability of Bayes' theorem for non-trivial settings—especially ***Bayesian neural networks (BNNs)***—makes it necessary to approximate the posterior density. Such ABI can be largely divided into two families of approaches, ***sampling-based inference (SBI)*** and ***local approximations*** (cf. Fig. 2.5). As a third alternative, finite ***ensembles*** do not fit squarely into either category but can be viewed as approximately Bayesian (e.g., Wilson and Izmailov, 2020). They have become a staple method of **UE** thanks to their empirical success. SBI is the focus of contributions **[C1]** and **[C2]**, while local approximations serve as baselines in **[C1]**, **[C2]** and **[C3]**, and ensembles play a role in all contributions except **[C5]**. We briefly introduce the respective concepts here and discuss ABI in the light of BNNs in Sec. 2.2.4.

Markov Chain Monte Carlo SBI aims at sampling from the actual posterior density (see Fig. 2.7). Recall from Eq. (7) that computing the posterior is inhibited by the evidence term in the denominator serving as a normalizing constant. ***Markov chain Monte Carlo (MCMC)*** techniques yield—under certain assumptions—samples from the true posterior while requiring access only to an *unnormalized* version of it. Samplers traverse the state space such that the time spent in each state is proportional to the target density $p(\boldsymbol{\omega})$[17]. This is achieved by generating proposals (for instance, sampling from a tractable, auxiliary distribution) and accepting them at a specific rate. One of the simplest algorithms is ***random walk Metropolis-Hastings***, where proposals are drawn from a Gaussian centered at the current state (Betancourt, 2018). Since acceptance probabilities are governed by ratios of densities, the normalization constants conveniently cancel out (Murphy, 2012). MCMC carries *Markov chain* in its name because it produces a series of correlated samples. We will emphasize this by denoting such ***chains*** as $\{\boldsymbol{\omega}_1 \to \cdots \to \boldsymbol{\omega}_S\}$, or $\boldsymbol{\omega}_S^{\rightarrow}$ for short. Each draw depends on the previous one, and the first one, on the initial location (Gelman and Shirley, 2011). The ***stationary distribution*** of $\boldsymbol{\omega}_S^{\rightarrow}$, as $S \to \infty$, is the target density. At the same time, finite-length chains give rise to *Monte Carlo* estimators. As in Eq. (10), we can approximate any expectation for a function $g(\boldsymbol{\omega})$ with arbitrary precision by the sample mean over $g(\boldsymbol{\omega}_s)$, $s \in \{1, \ldots, S\}$, as $S \to \infty$ (Murphy, 2012).

[17] We will omit the conditioning on $\mathcal{D}$ in this section to ease notation and to emphasize that MCMC works for arbitrary distributions over $\boldsymbol{\omega}$ that are known up to normalization.

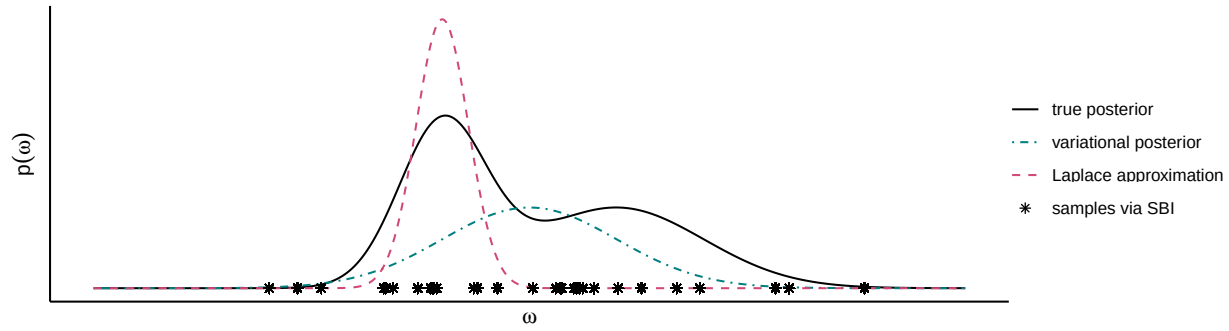

Figure 2.7.: Illustration of ABI methods for $\boldsymbol{\omega} \in \mathbb{R}$. The *variational*, uni-modal approximation (teal, dotted) attempts to cover both modes of the true posterior (black, solid), whereas the *Laplace* approximation (pink, dashed) concentrates at the MAP estimate. *Samples* from the posterior (asterisks) should recover the true density as faithfully as possible. Adapted from Papamarkou et al. (2024).

Convergence The key concern in MCMC is whether $\boldsymbol{\omega}_S^{\rightarrow}$ actually approaches the stationary distribution $p(\boldsymbol{\omega})$. For this to happen, the Markov chain must be ***ergodic***—loosely speaking, this entails the possibility to reach any state with positive probability, no risk of getting trapped in cycles, and the ability to revisit any state in finite return time (Gelman et al., 2021; Asmussen and Glynn, 2011). While guarantees in the limit $S \to \infty$ provide a solid theoretical foundation, the practical concern is, of course, to ensure convergence for $S \ll \infty$. There is no way of knowing precisely when the stationary distribution has been recovered. In practice, all we can do is monitor sampling diagnostics until we cease to detect a *lack of* convergence (Gelman and Shirley, 2011). Two aspects in particular impact finite-time convergence (Martin et al., 2024):

1. *Unbiasedness.* We need to make sure that the sampler reaches the relevant areas of the state space in reasonable time. The probability of a proposal to lie in a given region should be governed by the ***posterior mass*** of that region (the product of density and volume; Speagle, 2020). The collection of high-mass values for which the (log) density is ϵ-close to its expectation[18] is called ***typical set*** for some $\epsilon > 0$ (Murphy, 2022). For computations that involve expectations over the posterior, such as the PPD in Eq. (9), "evaluating the integrand outside of the typical set has negligible effect on expectations and hence is a waste of precious computational resources" (Betancourt, 2018).

2. *Efficiency.* Not only should the sampler be able to find the typical set, but it should do so quickly, and explore it efficiently afterwards. The ***effective sample size (ESS)*** quantifies how many hypothetical, *i.i.d.* samples provide the same amount of information as $\boldsymbol{\omega}_S^{\rightarrow}$. Due to the autocorrelative structure of MCMC, we need more than S draws for an ESS of S. Specifically, samples from the initial ***burn-in phase*** of navigating to the typical set should be discarded (though it can be hard to know when this point is reached; Speagle, 2020).

Obviously, we strive for chains with large ESS, which translates to good proposals with high acceptance probability. Unfortunately, the ***curse of dimensionality*** implies exponential growth in volume with respect to the number of parameters, while the density follows the reverse law. For any given location, the volume increases as a function of distance from this location, creating large

[18] Note that the posterior mass is directly related to expected values. The expectation is an infinite sum of parameter states weighted by their density, and as such dominated by high-mass regions that contain a sufficient number of parameter states (volume) with fairly high density.

gaps between current states and proposals (Speagle, 2020). Naive guess-and-check samplers like Metropolis-Hastings will continue to produce proposals from the vast volume of states outside the typical set and reject them due their low density. Efficient sampling in high dimensions requires more clever strategies that make use of the typical set's geometry (Betancourt, 2018).

Hamiltonian Monte Carlo ***Hamiltonian Monte Carlo*** (***HMC***; Duane et al., 1987) is such a geometry-informed sampler that has emerged as something of a gold standard in BNNs (Štrumbelj et al., 2024; Pituk et al., 2025). It allows for long-distance steps with high acceptance probability while preserving ergodicity (Neal, 2011). The key idea is to exploit directional information in the posterior density, which is readily available in the (first-order) gradient. The gradient vectors, however, point to high-density modes with steep slopes rather than the typical set. Betancourt (2018) uses the analogy of a planet (the mode) surrounded by an orbit (the typical set). For a satellite to remain on the orbit, without being pulled in toward the planet by gravitational forces (the gradient), we need a balancing force. This force is introduced via the ***momentum*** $\boldsymbol{\rho}$ in HMC. We augment the target density $p(\boldsymbol{\omega})$ to the joint, ***canonical density***

$$p(\boldsymbol{\omega}, \boldsymbol{\rho}) = \exp(-\mathbf{H}(\boldsymbol{\omega}, \boldsymbol{\rho})), \tag{11}$$

where $\mathbf{H}$ denotes the ***Hamiltonian*** defined as

$$\mathbf{H}(\boldsymbol{\omega}, \boldsymbol{\rho}) = -\log p(\boldsymbol{\rho}|\boldsymbol{\omega}) - \log p(\boldsymbol{\omega})) = \underbrace{U(\boldsymbol{\rho}, \boldsymbol{\omega})}_{\text{potential energy}} + \underbrace{K(\boldsymbol{\omega})}_{\text{kinetic energy}}. \tag{12}$$

Alg. 1 briefly summarizes the transition from one parameter state to the next (for details, see Betancourt, 2018). The simulation of intermediate states $(\boldsymbol{\omega}_j, \boldsymbol{\rho}_j)$ is governed by a system of

Algorithm 1 HMC iteration

1: **Input:** Current state $\boldsymbol{\omega}$, step size β, number of steps J, mass matrix $\boldsymbol{M}$
2: **Sample** momentum $\boldsymbol{\rho} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{M})$
3: Set $(\boldsymbol{\omega}_0, \boldsymbol{\rho}_0) \leftarrow (\boldsymbol{\omega}, \boldsymbol{\rho})$
4: **for** $j = 1$ to J **do**
5: Simulate $(\boldsymbol{\omega}_j, \boldsymbol{\rho}_j)$ according to Hamiltonian dynamics with step size β
6: **end for**
7: Compute acceptance probability: $\alpha = \min\left(1, \exp\left[\mathbf{H}(\boldsymbol{\omega}_0, \boldsymbol{\rho}_0) - \mathbf{H}(\boldsymbol{\omega}_J, -\boldsymbol{\rho}_J)\right]\right)$
8: With probability α, set $\boldsymbol{\omega} \leftarrow \boldsymbol{\omega}_J$
9: **Return** $\boldsymbol{\omega}$

differential equations that preserve the ***energy***—i.e., they evolve the parameters and momentum while keeping the Hamiltonian constant. Momentum resampling effectively leads the HMC chain to alternate between exploring level sets of constant energy and moving across the level sets in a random walk (Betancourt, 2018). Frequently, the ***leapfrog method*** is used to approximate the Hamiltonian dynamics in line 5 of Alg. 1 (Neal, 2011). It requires two hyperparameters, the step size and number of steps, which greatly influence the performance of HMC. Providing an effective way to adaptively tune both of these levers, Hoffman and Gelman (2014) introduced the ***no-U-turn sampler (NUTS)***, which remains state of the art to this day (Štrumbelj et al., 2024). We use NUTS in both [**C1**] and [**C2**]. Several adaptations have been proposed to make HMC more amenable to BNN inference, including minibatch variants, adaptive tuning strategies, and hybrids with normalizing flows (Cobb and Jalaian, 2021; Alexos et al., 2022; Riou-Durand et al., 2023; Grenioux et al., 2023).

Local Approximations

Local approximations do not aspire to sample from the true posterior. They find a surrogate that locally resembles the target density (cf. Fig. 2.7) and admits efficient sampling. The central framework in local ABI is ***variational inference (VI)***, rooted in statistical physics and later popularized for Bayesian learning (MacKay, 2003; Wainwright and Jordan, 2008; Kucukelbir et al., 2017). More recent adaptions of VI to BNNs include works by Blundell et al. (2015); Gal and Ghahramani (2016); Shen et al. (2024). VI translates the inference problem into an optimization task. The parameters of an approximate distribution from a tractable family are chosen such that its KL divergence to the posterior is minimal. Obviously, the success of local approximations depends on the disparity between the true and the variational density. Gaussian distributions, even as simple as such with diagonal covariance structure, are a standard choice. Variational approximations can be crude when the true posterior is far from normal[19]. In particular, BNN posteriors are notoriously ***multi-modal***, i.e., they attain locally maximal density at several locations (e.g., Gelberg et al., 2024, see Sec. 2.2.4 for an extensive discussion). This may lead VI methods to either mode-*covering* (diffusing the posterior mass widely across modes) or mode-*seeking* (concentrating the mass at a single mode) behavior. For BNNs, the ***Laplace approximation (LA)*** has become popular in recent years (MacKay, 2003; Daxberger et al., 2021; Antorán et al., 2022). It can be seen as a special instance of VI with a Gaussian variational family where the mean and covariance, respectively, are constructed from a MAP estimate and the local curvature around it (Blei et al., 2017). This allows for the LA to be applied post-hoc (after standard loss-based training, where the prior is implicitly specified through a regularization term) at relatively low cost (Daxberger et al., 2021).

Deep Ensembles

We have seen two ways of producing samples $\boldsymbol{\omega}_s$ to compute the approximate PPD in Eq. (10): SBI via MCMC, sampling the true posterior, and local approximations, drawing from a surrogate that is as close as possible to the target density. Both approaches can be interpreted as producing an ***ensemble*** prediction for $p(y_+|\boldsymbol{x}_+, \mathcal{D})$ by averaging S individual opinions with weights governed by the (approximate) posterior. Indeed, the exact BMA in Eq. (9) constitutes an infinite ensemble in this sense (Wilson, 2020). A third alternative to ABI is the construction of *explicit* ensembles as a committee of base learners optimized through loss-based training (Schapire, 1990; Breiman, 1996). Lakshminarayanan et al. (2017) introduced ***deep ensembles (DEs)***, where S NNs that only differ in their random weight initialization[20] (also referred to as a *homogeneous* ensemble; Wang and Wang, 2025) are trained in parallel. Their individual predictions are averaged to a consensus afterwards. As such, DEs are not directly conceived from a Bayesian learning paradigm. Still, they are what Mlodozeniec et al. (2024) call "implicitly Bayesian": they learn a prediction rule that induces data-dependent prior and posterior weight distributions (Loaiza-Ganem et al., 2025). As Arbel et al. (2023) put it, "they can be seen as performing a very rough Monte

[19] The Bernstein-von Mises theorem states that, under a number of regularity conditions, the posterior distribution becomes Gaussian in the limit of infinite data (Van Der Vaart, 1998). However, for BNNs, it cannot generally be assumed that these conditions are met, and the number of observations typically remains small in relation to the number of parameters (e.g., Izmailov et al., 2020).

[20] Other ensemble approaches like bagging (Breiman, 1996) use data subsampling as an additional means of decorrelation, but Lakshminarayanan et al. (2017) find that using all available data works better for DEs.

Carlo estimate of the posterior distribution over weights". The main reason why DEs are handled on par with more rigorous approaches, despite ongoing debates on how exactly they implement Bayesian principles (e.g., D'Angelo and Fortuin, 2021; Wild et al., 2023), is that they frequently outperform BNNs using other ABI strategies (with surprisingly small ensemble sizes for expressive base learners; Lobacheva et al., 2020; Wang and Wang, 2025). We thus follow Wilson and Izmailov (2020) in concluding that DEs "are not a competing approach to Bayesian inference, but can be viewed as a compelling mechanism for Bayesian marginalization"[21].

2.2.4. Bayesian Deep Learning

Peculiarities of Bayesian Neural Networks

The need for approximate inference exists for any model with intractable posterior but is severely aggravated in BNNs due to some idiosyncrasies that set them apart from more traditional models. At first glance, the sole difference seems to be that BNNs have vastly more parameters. It is well known that covering high-dimensional spaces requires an exponentially growing amount of samples because distances may become larger than our intuition from few dimensions suggests (Speagle, 2020, or, as the 1979 horror movie *Alien* put it, "in space, no one can hear you scream"). This makes the scaling of ABI methods challenging[22]. More important, the huge number of parameters creates a non-trivial interplay between inductive biases and the data that has made NNs seem mysterious for years (Wilson, 2025) and has called the validity of Bayesian DL into question (Wenzel et al., 2020; Farquhar, 2022; Pituk et al., 2025). Building on the characterization in Knoblauch et al. (2022), we summarize the peculiarities of BNNs in the following:

1. Inference is prohibitively *expensive*.
2. It is hard to elicit sensible *priors*.
3. The *likelihood* is misspecified because it is chosen for maximum predictive power, rather than to describe the DGP. In particular, it is *non-identifiable* because many parameterizations induce identical functional mappings. As a consequence, we observe *underfitting* when the data are insufficient for posterior contraction, and the posterior density is highly *multi-modal*.

Prior Choice

The original Bayesian idea of representing existing knowledge in the prior distribution is difficult to implement in BNNs featuring high-dimensional parameters without a clear interpretation. We can still encode sensible inductive biases that are known to generalize well. In absence of actual domain knowledge, it seems reasonable to keep the prior relatively vague so the Bayesian updating rule can be dominated by the information in the data. Nalisnick (2018) lists several options for such priors, including Gaussians, more heavy-tailed (for example, student-t) distributions, and minimally-informative alternatives like Jeffreys priors. The choice of ***zero-mean, diagonal, isotropic (ZDI)*** Gaussian distributions $\mathcal{N}(\mathbf{0}, \text{diag}(\sigma))$ with $\sigma > 0$ has become especially popular for BNNs

[21] It should be noted that this argument only holds for ensembles whose members share the same functional form, not arbitrary model combinations (Minka, 2002).

[22] Izmailov et al. (2021) used hundreds of tensor processing units in their seminal paper on the characteristics of BNN posteriors, which exceeds by far the resources most practitioners can claim.

(e.g., Fortuin et al., 2022). This is not entirely unjustified: ZDI Gaussians encourage parameter sparsity and function smoothness while providing unbounded support over the whole parameter space (Nalisnick, 2018). On the other hand, we know that NN parameters are correlated and assembled in hierarchical, layer-like structures (Papamarkou et al., 2022; Arbel et al., 2023). Martin and Mahoney (2021) suggest that weight matrices with strong correlations could be related to NNs performing well, but this is not reflected in a diagonal prior. Furthermore, as we will discuss later, the uni-modality of Gaussians seems hardly appropriate for BNN posteriors. These concerns have led some authors to propose *functional* priors instead (e.g., Tran et al., 2022). In function space, they argue, we have a better grasp of what reasonable inductive biases look like. However, finding *tractable* functions that yield *meaningful* priors is not straightforward either. Regarding the former, most work so far has made use of Gaussian processes (Fortuin, 2022); for the latter, empirical-Bayes-style variants have been proposed that infer the prior from some upstream task (Shwartz-Ziv et al., 2022; Rudner et al., 2023). Also, function-space priors do not combine readily with existing inference methods. SBI has been designed to work in weight space, which is why functional priors remain limited to VI (Sun et al., 2019; Cinquin and Bamler, 2025) or require novel inference routines. While the relevant prior is arguably the one in function space, vague weight priors might still be meaningful when paired with highly complex likelihoods pushing them forward to function space (Wilson and Izmailov, 2020). Since neither perspective is without issues, specifying useful priors remains an open problem.

Likelihood Overparameterization

Overparameterization Concerns of scalability and prior choice arise immediately from the sheer number of parameters. An arguably more consequential property, however, lies in in the joint consideration of weights and data: modern (B)NNs are heavily ***overparameterized***. The term is used widely yet seldom with a strict definition, suggesting something of a surplus of parameters (even if the nominal amount is not necessarily decisive for model behavior; e.g., Efron, 1983). One perspective on overparameterization is that of a veritable redundancy. Redundancy is closely linked to ***compressibility***, i.e., the existence of some latent, lower-dimensional structure in the weight space that allows for smaller or sparser models with the *same* predictive properties (Kwon et al., 2024; Kolb et al., 2025). This indicates ***non-identifiability*** of the original model. We will take a closer look at such redundancies in the following section. By no means, though, does redundancy imply a need to avoid it. The power of modern NNs arises precisely from their vast surplus of parameters. A recent body of work studying this phenomenon has operated with a slightly different notion of overparameterization. Here, the focus is on the complexity induced by the parameters (which is not straightforward to measure; Dherin et al., 2022; Curth et al., 2023; Patil et al., 2024). The key observation, pioneered by Belkin et al. (2019) and challenging previous convictions, is that of two "regimes" in generalization. The ***underparameterized*** regime follows traditional wisdom: while the training error keeps decreasing for growing complexity (in the sense of ***capacity*** of $\mathcal{H}$), the test error assumes a U-shape when the model starts to overfit to the training data. Further increasing complexity, which no one would have thought reasonable before the age of large NNs, pushes the test error toward an ***interpolation threshold***. This threshold marks the capacity required to "memorize" the training data (i.e., fit random datasets of size N). Subsequently, the error enters the ***overparameterized*** regime, decreasing again until well below its previous minimum. The resulting shape has inspired the name ***double descent***.

Implications for Bayesian Inference Overparameterization remains a hot topic in DL (Bubeck and Sellke, 2022; Huh et al., 2023; Peleg and Hein, 2024; Simon et al., 2024; Wilson, 2025). For the Bayesian community, the picture looks less benign. Overparameterization fundamentally challenges the Bayesian update rule turning prior into posterior beliefs (Roy et al., 2024). We have seen that specifying priors for BNNs is difficult. The use of overparameterized models further violates the likelihood principle. Rather than describing the DGP, we look for functions that can fit any size-N dataset sufficiently well to generalize in a useful manner, willing to accept that the parameters characterizing those functions may no longer be identifiable (Knoblauch et al., 2022). As a consequence, it has been observed that BNNs tend to *underfit* (i.e., fail to represent even the training data appropriately; Fisher and Marzouk, 2024; Miani et al., 2025). The balance between prior and data seems to be disturbed for overparameterized models. At second glance, this is not surprising. "Increasing the dimension of the model space for a fixed number of observations amounts to placing more weight on the prior" (Knoblauch et al., 2022), such that "the influence of the prior becomes magnified in the interpolating regime" (Hodgkinson et al., 2023). This perspective helps to explain earlier findings about a ***cold posterior effect (CPE)***, describing a situation where artificial reduction of the posterior dispersion ("tempering") leads to better results (Wenzel et al., 2020). The CPE has been attributed to poor priors and misspecified likelihoods, possibly in conjunction with data augmentation (Adlam et al., 2020; Aitchison, 2021; Kapoor et al., 2022; Noci et al., 2021; Fortuin et al., 2022). More recently, Zhang et al. (2024) established a rigorous link between the CPE and BNNs underfitting. Now, how does overparameterization affect *uncertainty*? The CPE suggests that "cooling" posteriors, which is achieved by overcounting the data at the expense of prior influence and sharpening the posterior to concentrate more strongly, mitigates the underfitting effect. This implies that the untempered posterior remains too vague. As Fisher and Marzouk (2024) put it, "in overparameterized networks, predictive uncertainty likely reflects an inability to completely forget the prior given the training data—that is, an inability to make confident predictions". Obviously, such observations are problematic for the **UE** endeavor. This is not to say that the issue cannot be resolved (for instance, Zhang et al., 2024, propose a modified update rule to eliminate the CPE). Still, there appears to be a gap between state-of-the-art models and the traditional Bayesian machinery.

Posterior Landscape

The delicate prior-likelihood balance bears implications for the posterior ***landscape***. It determines the success of the previously discussed ABI methods, i.e., how well samplers can traverse (or local approximations, cover) the full landscape. This question is at the core of contributions [**C1**] and [**C2**]. As of now, the picture is inconclusive. We do not have analytical expressions for BNN posteriors. Visual inspection usually relies heavily on dimensionality reduction techniques that may discard valuable information (e.g., Garipov et al., 2018; Li et al., 2018). Following the distinction in Freeman and Bruna (2017), two aspects decide how benign or averse the posterior landscape presents itself to the goals of ABI. The first is of *topological* nature and addresses the degree of ***multi-modality*** in the posterior hypersurface. In particular, we wish to know how many modes exist and whether they are isolated or connected. This is, for instance, relevant to the quality of (uni-modal) local approximations to the posterior.

> ***Posterior mode*** Local optimum in the hypersurface of the posterior density, embedded in a locally convex surrounding region.

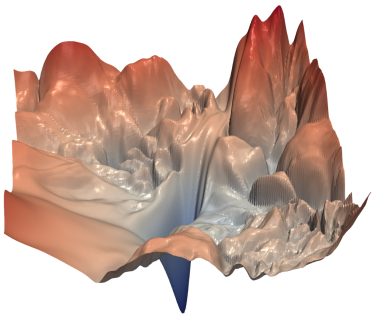

Figure 2.8.: Example for a loss landscape visualized in two dimensions. Taken from Li et al. (2018).

The second question pertains to *geometric* properties and is concerned with ***curvature***. Both aspects together control the presence and representation of **EU**. Unfortunately, the posterior landscape remains underexplored, but there is a rich body of literature on *loss* landscapes of NNs. By virtue of the Bayesian update rule we know that

$$p(\boldsymbol{\omega}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\omega})q(\boldsymbol{\omega}). \tag{13}$$

Taking the negative logarithm of Eq. (13) to obtain

$$-\log p(\boldsymbol{\omega}|\mathcal{D}) = \underbrace{-\log p(\mathcal{D}|\boldsymbol{\omega})}_{\text{NLL loss}} - \log q(\boldsymbol{\omega}) + \text{const} \tag{14}$$

exposes that the (negative log) posterior hypersurface is equivalent to the loss hypersurface modulated by the prior. Research on loss landscapes began in the 1990s and continues to this day with results being challenged at a fast rate. Fig. 2.8 shows an example of how such a landscape may look for architectures with millions of weights. Overall, the loss landscape seems to be governed by two forces engaged in a tug-of-war: unidentifiable parameters and overparameterization.

Parameter Symmetries The pioneering work of the 90s occurred outside the Bayesian paradigm and focused on questions of identifiability. MLPs with bounded activation functions were the predominant architecture of the time. In a progression of increasingly more general results, Hecht-Nielsen (1990); Sussmann (1992); Chen et al. (1993); Albertini and Sontag (1994); Kůrková and Kainen (1994) established that the parameter space of such MLPs contains a number of states that parameterize the *exact same* functional output. In other words, we have a many-to-one mapping from parameter to function space. This property constitutes the following equivalence relation:

$$\boldsymbol{\omega} \sim \boldsymbol{\omega}' \iff h_{\boldsymbol{\omega}}(\boldsymbol{x}) = h_{\boldsymbol{\omega}'}(\boldsymbol{x}) \quad \forall \boldsymbol{x} \in \mathcal{X}, \quad \boldsymbol{\omega}, \boldsymbol{\omega}' \in \boldsymbol{\Omega}.$$

The transformations $\boldsymbol{\Omega} \to \boldsymbol{\Omega}$ inducing such equivalence relations form a ***symmetry***[23] group (Vlačić and Bölcskei, 2021). Fig. 2.9 shows an example case of symmetric patterns. Following Chen et al. (1993); Kůrková and Kainen (1994), we refer to any two parameter vectors related by $\sim$ as ***equioutput*** or ***functionally equivalent***, whereas $\boldsymbol{\omega} \nsim \tilde{\boldsymbol{\omega}}$ are ***functionally diverse***. We call two NNs $h_{\boldsymbol{\omega}}, h_{\boldsymbol{\omega}'}$ with $\boldsymbol{\omega} \sim \boldsymbol{\omega}'$ ***isomorphic*** (e.g., Rolnick and Körding, 2020).

[23] We focus solely on what Villar et al. (2023) call *passive* symmetries, emerging as a consequence of modeling choices. This is in contrast to an entirely different field studying *active* symmetries in terms of invariance or equivariance according to laws of the physical world (e.g., Cohen and Welling, 2016).

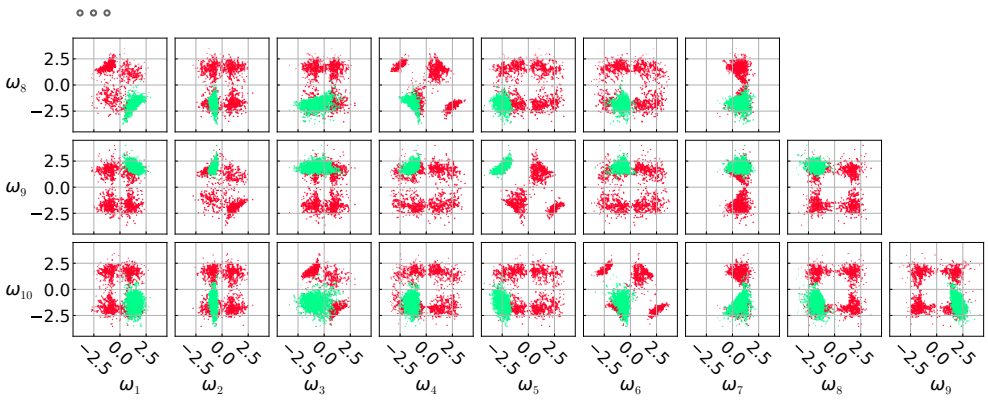

Figure 2.9.: Parameter symmetries. Selected pairwise profiles of samples (red) from the posterior density, as obtained by SBI, of a single-hidden-layer BNN with weights $\boldsymbol{\omega} \in \mathbb{R}^9$. Replicating symmetric patterns (discovered by Alg. 1 in [C1]) are highlighted in green. Adapted from [C1].

Functional equivalence Relation between two or more parameter vectors that parameterize isomorphic models, i.e., lead to identical function results for any input.

MLPs with certain activation functions have at least two built-in mechanisms of generating equioutput parameters. First, we can freely rearrange the neurons of fully-connected layers due to the commutative nature of the sum each neuron computes. This leads to a large[24] (but *countable*) number of ***permutation symmetries***. Ziyin (2024) note that there might be additional permutation patterns when neurons from different layers can be exchanged at functional equivalence. Second, different activation functions admit equioutput states (see, e.g., Kunin et al., 2021, for a taxonomy). The early works focused on odd activation functions[25] popular at the time, like the hyperbolic tangent, that induce a *countable* number of ***sign-flip symmetries*** by changing the signs of all incoming and outgoing weights in a neuron. Frequently, completeness results that prove identifiability up to symmetric parameter transformations exist for such architectures (Zhao et al., 2025). Later work (Neyshabur et al., 2015; Phuong and Lampert, 2020; Bona-Pellissier et al., 2023) focused on the ReLU activation function (Nair and Hinton, 2010) more common in modern architectures. ReLU-type, homogeneous activations effect *uncountably* many equioutput states through ***rescaling symmetries***[26]. To summarize, we have the following common types of functional equivalence[27], all of which can be expressed by means of a projection matrix:

Permutation symmetry Let $\mathbf{P} \in \{0,1\}^{D\times D}$ with $\mathbf{P}\mathbf{e}_i = \mathbf{e}_{\pi(i)} \forall i \in \{1,\dots,D\}$ be a permutation matrix, where $\mathbf{e}_i$ is the i-th basis vector of $\mathbb{R}^D$ and π permutes the entries of $\mathbf{e}_i$. Then, the model $h_{\boldsymbol{\omega}}$ exhibits a *permutation symmetry* if

$$h_{\boldsymbol{\omega}}(\boldsymbol{x}) = h_{\mathbf{P}\boldsymbol{\omega}}(\boldsymbol{x}) \quad \forall \boldsymbol{x} \in \mathcal{X}, \quad \boldsymbol{\omega} \in \boldsymbol{\Omega}.$$

[24] Hecht-Nielsen (1990) showed that there are at least $\Pi_{\ell=1}^{L} N_\ell!$ permutation-induced symmetries for L hidden layers with N_ℓ neurons in the ℓ-th layer (e.g., for $L = 3$ with 5 neurons in each hidden layer, we already have over 1m functionally equivalent permutations).

[25] Odd functions $f : \mathbb{R} \to \mathbb{R}$ obey $f(-z) = -f(z) \quad \forall z \in \mathbb{R}$ (e.g., Albertini and Sontag, 1994).

[26] For $f : \mathbb{R} \to \mathbb{R}_0^+$ the ReLU activation and $c \geq 0$, we have $f(c \cdot z) = c \cdot f(z) \quad \forall z \in \mathbb{R}$ (Neyshabur et al., 2015).

[27] We view negative rescaling as a composition of positive rescaling and a sign-flip symmetry.

Sign-flip symmetry Let $\mathbf{S} = \text{diag}(\boldsymbol{\alpha})$ with $\boldsymbol{\alpha} \in \{-1, 1\}^D$ be a sign-flip matrix. Then, the model $h_{\boldsymbol{\omega}}$ exhibits a *sign-flip symmetry* if

$$h_{\boldsymbol{\omega}}(\boldsymbol{x}) = h_{\mathbf{S}\boldsymbol{\omega}}(\boldsymbol{x}) \quad \forall \boldsymbol{x} \in \mathcal{X}, \quad \boldsymbol{\omega} \in \boldsymbol{\Omega}.$$

(Positive) rescaling symmetry Let $\mathbf{R} = \text{diag}(\alpha_1, \ldots, \alpha_D)$ with $\alpha_i > 0 \; \forall i \in \{1, \ldots, D\}$ be a positive rescaling matrix. Then, the model $h_{\boldsymbol{\omega}}$ exhibits a *positive rescaling symmetry* if

$$h_{\boldsymbol{\omega}}(\boldsymbol{x}) = h_{\mathbf{R}\boldsymbol{\omega}}(\boldsymbol{x}) \quad \forall \boldsymbol{x} \in \mathcal{X}, \quad \boldsymbol{\omega} \in \boldsymbol{\Omega}.$$

Functionally equivalent parameters vastly outnumber functionally diverse ones—for each of the latter, (possibly infinitely) many symmetric replicates exist. The above types of symmetries are not even exhaustive; others (e.g., due to batch normalization) have been noted in the literature (Rolnick and Körding, 2020; Kunin et al., 2021; Armenta and Jodoin, 2021).

Multi-Modality and Mode Connectivity Assuming that a reasonable loss function assigns identical values to identical network outputs, we can view functionally equivalent parameters as being in the same ***level set*** of constant loss (Freeman and Bruna, 2017). Obviously, sets of low loss are of particular interest. The relative location of solutions within low-loss level sets is decisive for how efficiently samplers (or optimizers) can traverse the parameter space: are we facing a vast number of isolated modes, as the existence of symmetries might suggest, or is the reality more benign? This has become known as the question of ***mode connectivity***, a term coined by Garipov et al. (2018). The authors extend previous work by Freeman and Bruna (2017) in finding curved connectors between local optima along which the loss remains near-constant (see Fig. 2.10). Draxler et al. (2018) introduced the idea of "no loss barriers" on the connecting curves, meaning that optimization can proceed unhindered without leaving the area of low loss. The initial findings sparked a line of research refining and expanding the notion of connectivity (Kuditipudi et al., 2019; Benton et al., 2021; Zhao et al., 2023; Adilova et al., 2024), in particular after Frankle et al. (2020) popularized the idea that a situation as simple as ***linear mode connectivity (LMC)*** might characterize robust solutions. Brea et al. (2019) were among the first to note that the search for constant-loss directions bears a natural link to symmetries: if functional equivalence implies identical loss, then symmetric parameters are loss-invariant by design. Subsequent work (Entezari et al., 2022; Jordan et al., 2023; Ainsworth et al., 2023; Ferbach et al., 2023; Lim, 2024) returned to the intriguing concept of LMC. A pile of evidence has since been collected that sufficiently large architectures exhibit LMC[28] *modulo permutation symmetries* ("meaning that there exist permutations (...) that enable any networks trained with the same dataset and SGD procedure to be linearly interpolated with low barriers"; Sharma et al., 2024). Summarizing the past few years of research, the following explanation seems plausible: *continuous* rescaling symmetries produce *connected* level sets (rescaling creates continuous lines of interpolation), while *discrete* permutation symmetries produce *disconnected* replicates (permuting two weights, for example, might map a parameter vector to an entirely different part of the space), such that connectivity can be enforced by controlling for permutations (Draxler et al., 2018; Ainsworth et al., 2023; Ferbach et al., 2023; Theus et al., 2025; Zhao et al., 2025).

[28]The existing literature supports this LMC hypothesis only for pairs of NNs. Recent work conjectures that a stronger claim including many NNs simultaneously might be justifiable (Sharma et al., 2024; Ito et al., 2025).

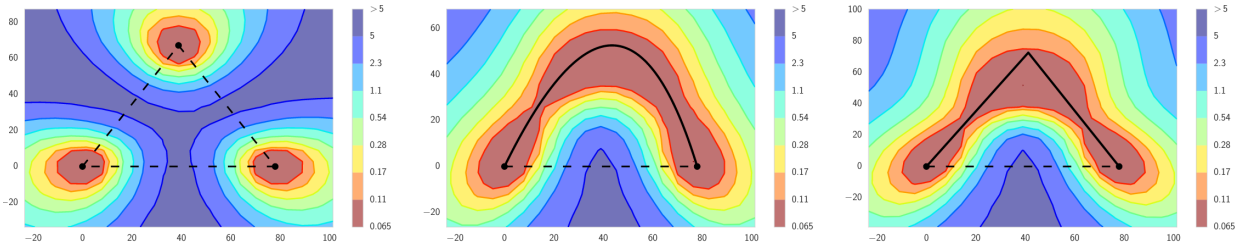

Figure 2.10.: Mode connectivity visualized in two dimensions. *Left:* three disconnected modes. *Middle:* curved path of low loss. *Right:* polygonal path of low loss. Warmer colors (bottom end of scale) indicate lower loss. Taken from Garipov et al. (2018).

Curvature Along the way, a link has been established between the hypothesis of connected optima and overparameterization: "far into or within the overparameterized regime (...), the number of subspaces that make up the global minima manifold is much greater than the number of subspaces that make up the symmetry-induced (nonglobal) critical points. In this sense the global minima manifold is huge" (Simsek et al., 2021). In partial overlap to the topological considerations, a geometric strand of work has found evidence that overparameterization promotes the type of *flat* areas of attraction[29] that had previously been linked to good generalization (Li et al., 2018; Feng and Tu, 2021; Foret et al., 2021; Pittorino et al., 2022; Iyer et al., 2023). In that sense, strongly overparameterized models are expected to have a more benign curvature.

It seems that overparameterization indeed softens much of what symmetries threaten to imply for multi-modality[30] and non-convexity in the loss landscape, possibly leading to what Martin and Mahoney (2021) call a "ruggedly convex" hypersurface. Work on loss landscapes and connectivity is ongoing (e.g., Zhao et al., 2024; Martinelli et al., 2025; Tian et al., 2025). As Ziyin et al. (2025) muse, "[l]astly, it is worth noting that on its own, symmetry is neither good nor bad. (...) Having the right degree of symmetry might thus be crucial for achieving both smooth optimization and good generalization".

Role of the Prior With this intuition about the loss hypersurface, let us turn to the prior as the second influencing factor for the posterior landscape (see Eq. (13)). It appears that posterior multi-modality depends, again, on the interplay between data and prior (Ziyin, 2024; Laurent et al., 2024; Kobialka et al., 2025). To understand this, consider two edge cases.

1. The data signal is so strong that the prior has essentially no influence (severely underparameterized regime). Then, every solution favored by the data represents a mode in the *likelihood* with many symmetric replicates, and the posterior hypersurface is also highly multi-modal.
2. In the other extreme case of perfect prior dominance, which follows from severe overparameterization, the posterior is equal to the prior. Then, assuming a ZDI Gaussian prior, there are no modes other than the one in the origin around which the posterior is located.

So are we caught between a rock and a hard place, facing either an extremely adverse posterior or a perfectly convex one where inference is trivial but the data provide zero information? Of

[29] While the notion of flatness is widely adopted, He et al. (2019) point out that, in high-dimensional spaces, the probability of observing either flat or sharp curvature in all directions from the mode decreases.

[30] It should be noted that not all multi-modality results from symmetries; it is quite conceivable that the observed data do not suffice to choose between functionally diverse hypotheses ([**C1**]; Fisher and Marzouk, 2024).

course, reality will manifest as some kind of middle case. Due to Eq. (13), functionally equivalent parameters translate to posterior symmetries if they are also *a priori* equiprobable. Under the standard choice of ZDI Gaussian priors, this holds for parameters that can be transformed into each other by permutation or sign flipping[31]. Consequently, the effect of discrete symmetries on the loss landscape will propagate directly to the posterior landscape. We also need to take into account the degree of overparameterization which, as we have seen, controls both mode connectivity and curvature. In the earlier work of **[C1]**, we consider small architectures which exhibit strong symmetric posterior patterns, tending more toward the first case (cf. Fig. 2.9). We later find evidence in **[C2]** of posteriors showing signs of more connected, less pronounced modes. Kobialka et al. (2025) reconcile both findings in experimental comparisons, attributing the degree of multi-modality to the degree of overparameterization. For the continuous rescaling symmetries introduced by ReLU-type activations, the picture is even more complicated. While appropriate scaling of weights leaves the functional output unchanged, ZDI Gaussian priors are not generally rescaling-invariant as they favor smaller-norm parameters.

To the best of our knowledge, the posterior landscape is not yet fully understood. The insights on overparameterization in conjunction with underfitting and the CPE suggest that large architectures with ReLU-type activations lean more toward the prior-dominated scenario with the convoluted picture continuous symmetries evoke. We cannot dispel the notion that these findings fundamentally challenge the Bayesian paradigm. Still, for the time being, there are some good practices to be deducted for feasible ABI.

Feasible Inference

Symmetry Removal We have seen that the structure of the posterior landscape is subject to a struggle between symmetries promoting multi-modality and regularization enforcing smoothness. It seems intriguing, therefore, to *remove* symmetries altogether. Several works, including **[C1]**, have experimented with parameter constraints (frequently involving unit-norm weights or bias sorting; Pourzanjani et al., 2017; Pittorino et al., 2022; Stock et al., 2019) to this end. Other proposals introduce invariances into the optimization algorithm (Neyshabur et al., 2015; Meng et al., 2019), activation function (Lim, 2024), or loss function (Ziyin et al., 2024). Their results suggest that symmetry removal can improve performance and soften NN *loss* landscapes. The impact on *Bayesian* inference remains somewhat unclear as most work has been conducted from an ERM angle (Lim, 2024; Ziyin et al., 2025, evaluate their proposals on a BNN task but focus on performance rather than posterior approximation quality).

Weight Subspaces Another conclusion that can be drawn from the insights on loss landscapes is that, if we choose to believe in the existence of a manifold on which most good optima lie, we could transfer inference to this subspace. The early findings on parameter symmetries immediately led to the idea of a representative "search set" that contains all the interesting solutions. For the small MLPs considered at the time, it was possible to give precise geometric descriptions (Hecht-Nielsen, 1990; Chen et al., 1993). More recently, a number of ideas have emerged about how to recover interesting, lower-dimensional subspaces in the hope of facilitating the inference task. The most basic methods restrict inference to the last network layer, treating all other parameters

[31] ZDI Gaussians fully factorize. This immediately implies invariance to permutations. Since all the marginal distributions are symmetrical around zero, they are also invariant to sign flips.

as non-stochastic (Kristiadi et al., 2020; Daxberger et al., 2021). However, the approximation quality of these simple heuristics has been called into question (Sharma et al., 2023). A number of approaches finds the dimensions along which most variability occurs[32]. These include subspaces based on the maximum-variance directions in the trajectories of a preceding SGD optimization (Maddox et al., 2019; Izmailov et al., 2020), joint updates in VI of parameters and subspace dimensions (Li et al., 2024), and separating the parameter space into prior- and likelihood-dominated partitions based on an eigenspectrum analysis of the loss Hessian (Constantine et al., 2016). Yet another strand of work builds on the mode connectivity literature and uses the paths linking pre-trained solutions (Izmailov et al., 2020; Dold et al., 2025). While these curves between pairs of models are usually low-dimensional by construction, recent work by Tian et al. (2025) suggests that global subspaces (i.e., connected volumes spanning the whole of $\boldsymbol{\Omega}$) retain the original dimensionality, limiting the benefits of subspace inference.

Multi-Start Strategies Lastly, in what seems an almost trivial idea, it is beneficial to conduct ABI from multiple starting points (Wilson and Izmailov, 2020; Izmailov et al., 2021; Papamarkou et al., 2024). Wilson et al. (2021) even postulate that "multi-modal approximations to the posterior should become a new standard in Bayesian deep learning, and the multi-modality may even be more important than the quality of approximation within each of the modes". The fact that cloud computing has greatly increased access to distributed machines means that parallel computations can make for efficient use of the available resources. Multi-start strategies echo the idea of ensembling, hedging against the risk of poor initialization and capturing more of the present functional diversity. For instance, a simple improvement over the uni-modality of local Gaussian approximations is to use an ensemble of such (Eschenhagen et al., 2021; Wilson et al., 2021). [**C1**] and [**C2**] show for SBI that running multiple chains in parallel achieves what a single, longer chain can cover. There are some aspects to consider with multi-start approaches. In particular, assessing convergence requires diagnostics that account for cross-chain mixing, typically by penalizing between-chain variability (Vehtari et al., 2021; Margossian et al., 2022). The existence of symmetries challenges this type of convergence monitoring in parameter space. When symmetry-induced multi-modality is present, chains should arguably not be penalized for being attracted to different modi. [**C2**] proposes convergence metrics in function space where any multi-modality pertains to functional diversity. It has further proven effective to warm-start multi-chain SBI in an informed manner; otherwise, posterior landscapes of the more adverse kind can trap samplers so they do not recover in reasonable time. To this end, leveraging the multi-start character of DEs seems to work quite well ([**C2**]; Sommer et al., 2025; Rundel et al., 2025).

Most of the above-cited approaches remain confined to exploring only parts of the parameter space. Recall, however, from Sec. 2.2.3 that what matters is the typical set, which might occupy the space in counterintuitive ways. The relatively recent field of ***weight-space learning*** studies how parameters inform model performance (Navon et al., 2023; Zhang et al., 2023; Zhou et al., 2023; Schürholt, 2024). As symmetries play a crucial role there, we may hope that insights from weight-space learning will help our understanding of posterior landscapes in the future.

[32] The idea of some parameters being more active than others resonates with the benign overfitting exhibited by overparameterized models. Sparse NNs often perform on par with their dense counterparts (e.g., Lotfi et al., 2022). Parameters taking on different roles also has a structural component to it: we find in [**C2**] that first- and last-layer weights are more tightly linked to the input and output, while weights in deeper layers seem to be exchanging roles rather fluently. This intuition is complemented by higher degrees of sparsity (Kolb et al., 2025) and better robustness to perturbations (Zhang et al., 2022) in hidden layers. According to Adilova et al. (2024), "it can be claimed that the loss surface has a pronounced layer-wise structure".

Summary: Representing Uncertainty

We have discussed the problem of **UR** under the working assumptions of Sec. 2.1.3 and with a focus on Bayesian agents. In summary, the interplay between prior and data seems to be different in overparameterized NNs from what traditional wisdom suggests. The picture is not yet conclusive, also owing to the fact that some papers focus on small architectures in the attempt of a deeper understanding while others studying larger models arrive at seemingly contradicting conclusions. From a topological and geometric perspective, the relative dominance of the prior due to overparameterization seems to have a benign effect in smoothing the rugged hypersurface that might be expected from parameter symmetries. There is evidence, however, of the Bayesian principle being violated. This means that the representation of **EU** might not be faithful. Fisher and Marzouk (2024) go so far as to raise "questions of whether overparameterized BNNs can successfully 'forget their priors' to learn from data, and whether a fully Bayesian model of uncertainty is suitable for producing low generalization error". The future will have to show whether the Bayesian paradigm can adapt to the challenges of overparameterized NNs while preserving its theoretical rigor.

2.3. Uncertainty Quantification

This chapter discusses how uncertainty can be *quantified* from a given representation. We focus on widely-used measures derived from Shannon *entropy* and discussed critically by [C3], [C4].

2.3.1. Goals

Summary Statistics for Uncertainty

We turn now from the *representation* of predictive uncertainty to its *quantification*. This is the main focus of contributions [C3] and [C4]. The quest is to get a single number measuring uncertainty on the level of individual observations, possibly disentangled into **AU** and **EU**[33]. It is important to realize that summary statistics come at a cost. Compressing the information contained in an entire distribution—which to obtain we go to great lengths—means that some of the information will be lost. Still, there are reasons for quantifying uncertainty this way.

1. *Model evaluation and interpretation.* In predictive modeling, it is natural to evaluate models by their performance. Here, we are also interested in the *reliability* of predictions: does the model have a good sense of when it might be wrong? This notion of alignment between predictive performance and model confidence serves as a criterion of model quality. Decomposing measures of **TU** further allows to drill down into the components for model interpretation. High levels of **AU** in many observations might signal, for instance, that a decision boundary is not well-suited to separate classes without overlap (cf. Fig. 2.3).

[33] One might be tempted to conclude that, when the ultimate quantity of interest is just total predictive uncertainty, representing the individual sources is futile, and the entire endeavor of bi-level representations is needlessly complicated. But, crucially, the **TU** resorting from representing both the **AU** and **EU** components (under a posterior with positive dispersion) is generally *not* the same as the value we get from ignoring the epistemic part (which would translate to a Dirac delta posterior distribution).

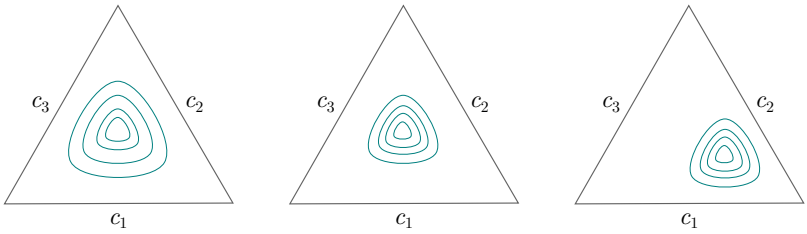

Figure 2.11.: Desiderata in **UQ**. Each point on the simplex $\Delta_K = \{\boldsymbol{\theta} \in [0,1]^3 : \|\boldsymbol{\theta}\|_1 = 1\}$ corresponds to a categorical first-order distribution $\boldsymbol{\theta}^{\boldsymbol{\omega}} = p(y_+|\boldsymbol{x}_+, \boldsymbol{\omega})$, where $\theta_i^{\boldsymbol{\omega}} = p(y_+ = c_i|\boldsymbol{x}_+, \boldsymbol{\omega})$ for $i \in \{1,2,3\}$. The corners equal full certainty for the respective class, and the barycenter is given by $\boldsymbol{\theta}^{\boldsymbol{\omega}} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. Contour lines represent the density of the second-order distribution $Q(\boldsymbol{\omega})$. *Left:* we expect higher **EU** relative to the *middle* case because $Q(\boldsymbol{\omega})$ is more dispersed. *Right:* $Q(\boldsymbol{\omega})$ concentrates further from the barycenter, so the **AU** should be lower than in the *middle* case. Inspired by Hofman et al. (2024).

2. *Taking action.* Downstream decision-making can be linked to a confidence threshold—for example, have a human look at instances for which uncertainty exceeds the threshold. Component-wise **UQ** additionally enables targeted decisions to modify the learning process. In the above situation, we might conclude that increasing model capacity will alleviate the problem.

Desiderata

We have focused on **UR** with Bayesian agents. For the following discussion, the scope is extended to general bi-level distributional representations. We only require some ***second-order distribution*** with density Q, assessing the plausibility of hypotheses $\boldsymbol{\omega}$ that parameterize ***first-order distributions*** $p(y_+|\boldsymbol{x}_+, \boldsymbol{\omega})$. This gives rise to a predictive density that coincides with the Bayesian PPD (Eq. (9)) if Q is a posterior obtained from a Bayesian updating rule:

$$p(y_+|\boldsymbol{x}_+, \mathcal{D}) = \int_{\boldsymbol{\Omega}} p(y_+|\boldsymbol{x}_+, \boldsymbol{\omega}) Q(\boldsymbol{\omega}) \, \mathrm{d}\boldsymbol{\omega}. \tag{15}$$

We can come up with some properties for meaningful uncertainty measures. It seems intuitively reasonable to demand that—for any component—uncertainty has a zero lower bound and can be quantified on a continuous scale. Furthermore, the lower bound should be attained by Dirac delta distributions, i.e., when there is but a single plausible prediction or hypothesis. **[C3]** formalizes these and further desiderata in discrete distributions. Fig. 2.11 illustrates for a three-way classification example a number of situations to which we can attach expected behaviors.

2.3.2. Entropic Measures for Distributional Representations

Entropy

The appropriateness of **UQ** fundamentally depends on the representation of choice (see, e.g., Hoarau et al., 2025, for a taxonomy of available options). Two types of metrics have been widely adopted in practice: ***variance***-based measures for continuous distributions (Kendall and Gal, 2017; Depeweg et al., 2018; Duan et al., 2024), and ***entropy***-based measures for categorical distributions (Mucsányi et al., 2024; Bickford Smith et al., 2024).

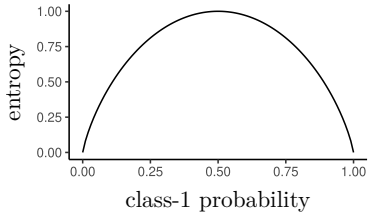

Figure 2.12.: Shannon entropy for a Bernoulli experiment. The entropy is a function of the class-1 probability (for classes $\in \{0, 1\}$) that directly governs the degree of uniformity in this binary task. It assumes its maximum value when both classes are equiprobable, decreasing symmetrically to both extremes (reflecting invariance to re-ordering of classes), and vanishes when any class is attained with probability one.

Given that many of today's applications—for instance, in computer vision or medical diagnosis—are essentially classification tasks, a set of measures derived from the decomposition of entropy (Houlsby et al., 2011) has become quite popular. In physical thermodynamics, entropy describes the *disorder* in a system. Shannon (1948) introduced a related concept to capture the degree of *surprise* contained in RVs. The entropy $\mathbb{H}$ of a discrete RV Z with event space $\mathcal{Z}$ is defined as

$$\mathbb{H}(Z) = -\sum_{z \in \mathcal{Z}} p(z) \log p(z). \tag{16}$$

The base of the logarithm[34] is usually set to two in an information-theoretic interpretation as the minimum number of bits required for efficient encoding. Properties of discrete entropy include boundedness ($\mathbb{H} \in [0, \log K]$, where K is the number of classes), invariance to any re-ordering of categories, and continuity in $p(z)$ (Cover and Thomas, 2006). Intuitively, entropy measures how much of a reduction in uncertainty we can expect by actually observing realizations of the RV. In the extreme case of a Dirac delta variable $Z = \delta_z$, we do not gain any information by observing Z as we know already that $Z = z$ with probability one. The entropy of this RV will be zero. We can expect maximum reduction of uncertainty, on the other hand, for uniform Z: when all outcomes are equally likely, observation is maximally informative. Consequently, $\mathbb{H}$ will assume its highest-possible value. Entropy can thus be viewed as the degree of *uniformity* in RVs (see Fig. 2.12). The ***maximum entropy principle*** states that, from a set of distributions satisfying given constraints, we should pick the distribution with maximum entropy to avoid making any more assumptions than strictly necessary. For a bounded RV and an assumption on the expected value, this leads to the uniform distribution. Then, the maximum entropy law implements Laplace's principle of insufficient reason dictating that, in absence of any evidence to the contrary, probability should be spread equally across all possible outcomes (Seidenfeld, 1986).

Entropy Decomposition

Components Shannon entropy has intrigued researchers in **UQ** not least because it admits a neat disaggregation into components that can be attributed to **AU** and **EU** (Houlsby et al.,

[34] The definition makes use of the convention $\lim_{a \to 0} a \log a = 0$.

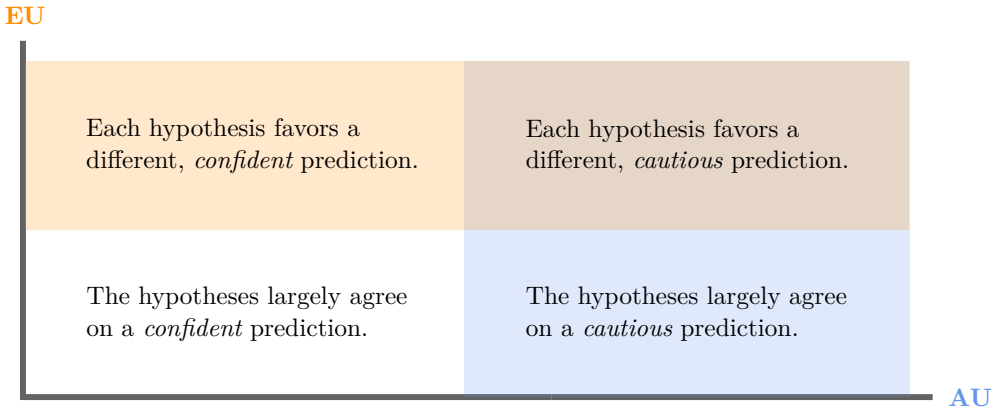

Figure 2.13.: Cases for entropic uncertainty measures arising from high and low levels of both uncertainty components. High **AU** manifests in *cautious* predictions; high **EU**, in *disagreement* with the consensus.

2011; Depeweg et al., 2018). Let Y_+ be the RV of test labels and denote by $\bar{\boldsymbol{\theta}}$ the ***consensus prediction*** resulting from marginalizing over hypotheses according to Eq. (15). Then, we have[35]

$$\begin{aligned} \underbrace{\mathbb{H}(Y_+)}_{\textbf{TU}} &= \underbrace{\mathbb{H}(Y_+|\Omega)}_{\textbf{AU}} + \underbrace{I(Y_+, \Omega)}_{\textbf{EU}} \\ &= \mathbb{E}_{Q(\boldsymbol{\omega})}[\mathbb{H}(Y_+|\boldsymbol{\omega})] + \mathbb{E}_{Q(\boldsymbol{\omega})}[D_{\mathrm{KL}}(p(Y_+|\boldsymbol{\omega})\|\bar{\boldsymbol{\theta}})]. \end{aligned} \tag{17}$$

$\mathbb{H}(Y_+|\Omega)$ is the ***conditional entropy***: we quantify the entropy for each first-order distribution resulting from a given hypothesis $\boldsymbol{\omega}$ (measuring the associated uncertainty if $p(y_+|\boldsymbol{x}_+, \boldsymbol{\omega})$ were the true data-generating distribution), and take the expectation over the second-order distribution on $\boldsymbol{\omega}$. $I(Y_+, \Omega)$ is the ***mutual information*** between Y_+ and Ω, reflecting the reduction in uncertainty about Ω we can expect from observing Y_+. Mutual information is a general measure of statistical dependence and zero if and only if Y_+ and Ω are independent. The more the individual hypotheses $p(Y_+|\boldsymbol{\omega})$ diverge from the consensus prediction $\bar{\boldsymbol{\theta}}$, the higher the **EU**. Fig. 2.13 summarizes different scenarios that arise from varying levels of **AU** and **EU**.

Finite-Sample Approximation When we lack a closed-form expression to compute expectations over $Q(\boldsymbol{\omega})$, as in ABI, we can approximate Eq. (17) with S samples from the second-order distribution. This admits the following finite-sample version of the entropy decomposition:

$$\mathbb{H}\left(\frac{1}{S}\sum_{s=1}^{S} p(y_+|\boldsymbol{x}_+, \boldsymbol{\omega}_s)\right) = \frac{1}{S}\sum_{s=1}^{S} \mathbb{H}(Y_+|\boldsymbol{\omega}_s)) + \frac{1}{S}\sum_{s=1}^{S} D_{\mathrm{KL}}(p(Y_+|\boldsymbol{\omega}_s)\|\bar{\boldsymbol{\theta}}_S). \tag{18}$$

$\bar{\boldsymbol{\theta}}_S$ denotes the consensus prediction from a size-S Monte Carlo estimate of Eq. 15. The **TU** and **AU** terms of Eq. (18) are easily computed from the discrete entropy definition in Eq. (16); the **EU** component results as a residual quantity by virtue of the additive relationship.

[35] We omit the dependence on the features $X = \boldsymbol{x}$ for better readability.

2.3.3. Critique of Entropic Measures

Measuring Ignorance

Despite the simplicity and mathematical appeal of the entropy decomposition, it turns out that the components exhibit counterintuitive behavior in a number of situations. [**C3**] analyzes these inconsistencies in a systematic manner. The criticism can largely be attributed to two conceptual issues: the representation of ***ignorance***, and the assumption of component ***additivity***.

Conflict *vs* Ignorance The Bernstein-von Mises theorem states that the posterior weakly converges to a Gaussian and concentrates to a Dirac delta measure in the limit of infinite data (Van Der Vaart, 1998). Consequently, in a situation of perfect knowledge, there is only a single hypothesis, the divergence term in Eq. (17) and (18) vanishes, and we have **TU** = **AU**. The other extreme case of total *ignorance* is less intuitive. As mentioned in Sec. 2.2.2, the ability of distributional representations to express ignorance has been questioned on a fundamental level. The uniform, as the least-informative distribution according to the Bayesian paradigm, reflects more a situation of indecision than of veritable ignorance. In particular, a truly ignorant modeler would not be able to conclude that all outcomes are equally likely (which is a rather informed statement). The problem is exacerbated by the entropy decomposition. Even if we are willing to accept that the uniform is the best option within a distributional framework, **EU** in terms of mutual information is not maximized by a uniform second-order distribution. Rather, we obtain maximum **EU** when Q is a Dirac mixture of maximally diverging hypotheses ([**C3**]). In this sense, mutual information predominantly captures ***disagreement*** or ***conflict*** (cf. Fig. 2.13; see also Shoja and Soofi, 2017). This behavior is most obvious in the version of the decomposition where the **EU** is expressed by KL divergence, and quite at odds with the Bayesian intuition about uniform beliefs.

Shortcut-Induced Conflict What kind of situation promotes conflicting hypotheses? As explained before, with Bayesian agents, the posterior density ascribes plausibility to functionally diverse hypotheses when the data do not suffice to rule out alternatives. Assuming there is but one process generating the observed data, this indicates that one of the plausible-seeming explanations at most can be the one corresponding to the true DGP. [**C4**] suggests that the presence of ***shortcuts*** (see Sec. 2.1.4) in the data can encourage this behavior. Shortcuts prompt the model to assign credibility to hypotheses that merely reflect spurious patterns in the observed data. That said, SCL does not *imply* disagreement (all hypotheses might settle for the same shortcut), nor does disagreement provide *proof* of SCL (complex or ambiguous data might lead to multiple, functionally diverse hypotheses without shortcuts being present).

Component Additivity

The second fundamental issue with entropy decomposition relates to the additivity of **AU** and **EU**. In theory, this seems a reasonable assumption to make since the components pertain to distinct sources that act independent of each other (for fixed quantities $\mathcal{H}, \mathcal{X}, \mathcal{Y}, p_{XY}$). Uncertainty arising from the data and uncertainty due to a lack of knowledge have different root causes. Unfortunately, this clear separation does not hold for the *estimated* quantities. Whenever **EU**

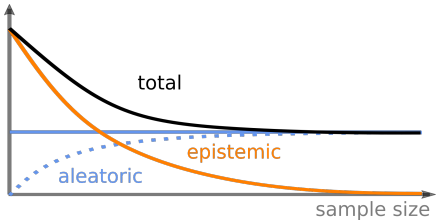

Figure 2.14.: Entropy decomposition for idealized learning scenario with an increasing number of observations. The black line marks **TU**; the solid blue line, AU as a theoretical quantity; the dotted blue line, the estimate of AU; the orange line, EU. Taken from [C3].

is present, the estimation of the AU will be afflicted. This is obvious from the computation of the AU estimate in Eq. (17): lacking knowledge about the true hypothesis $\boldsymbol{\omega}^*$, we resort to the expected entropy according to our *belief* about $Q(\boldsymbol{\omega})$. As Jiménez et al. (2025) caution, "one can only make meaningful statements about aleatoric uncertainty once epistemic uncertainty is *correctly* estimated as sufficiently low". Indeed, empirical evidence suggests that the estimates for both components tend to be correlated (Mucsányi et al., 2024; Hoarau et al., 2025). The decomposition shows inconsistent behavior that is further driven by the bounds of the entropy measure in a K-way classification problem. To see this, consider a hypothetical learning situation where increasingly more data are observed (equated with a gain in knowledge; Fig. 2.14). In the very beginning, it is reasonable to suppose that **TU** and EU assume their respective maximum values as observing more data (from the DGP) can only *reduce* uncertainty in this idealized setting. The AU is a latent, constant property of the DGP. We see that **TU** and EU coincide because, being entropic quantities, they are both upper-bounded by $\log K$ (Cover and Thomas, 2006). In conjunction with additivity, this implies immediately that the AU should be zero, which is not sensible. Demanding that the AU be a positive constant while **TU** = EU breaks additivity (or, *vice versa*, positive AU and additivity lead to **TU** > EU). We see that the entropic measures effectively produce a lower-bound estimate for the true AU (dotted line in Fig. 2.14). One might argue that the problem is only present in the beginning of the learning process and becomes less severe over time. Yet, it is precisely the *low-knowledge* regime—where the largest modeling errors must be expected—in which predictive uncertainty is vital. To make matters worse, the depicted behavior is idealized in the sense that each component measures what it should; in practice, the picture will be more convoluted with EU seeping into the AU estimates. For instance, EU has been shown to assume surprisingly low levels or even exhibit erratic, non-monotonic behavior for increasing sample size (Fellaji and Pennerath, 2024).

Alternative Proposals

In work challenging the current *modus operandi* more radically, Bengs et al. (2022) investigate whether it is possible to *learn* faithful values of (epistemic) uncertainty in a loss-based fashion, but report a negative result. Others have echoed and extended the criticism voiced in [C3]. Notably, Bickford Smith et al. (2024) find that the decomposition lacks conceptual clarity and that all we can ever expect in finite-sample settings is more or less poor estimates, which might be especially brittle in the oft-neglected case of non-*i.i.d.* data. On a similar note, Jiménez et al. (2025) call for explicit consideration of model biases (meaning systematic prediction errors), implying the abolition of the convenient, but possibly optimistic, $\mathcal{H}$-closed view. Schweighofer et al. (2025) find

fault with the component definitions themselves—arguing, for instance, that the **AU** should be a property of the predicting model—and propose a framework comprising a more diverse set of measures. Lastly, Kotelevskii and Panov (2024) suggest to use the Bayes risk (i.e., the expected loss of the true model on a random sample from the DGP, which is obviously a latent quantity) to measure the **AU**, and treat any "excess" risk exhibited by the trained model as **EU**.

Summary: Quantifying Uncertainty

It is not straightforward to disentangle the roles of **UR** and **UQ** in causing inconsistencies of uncertainty estimates. In particular, we cannot expect faithful **UE** when representational frameworks contain holes and require numerous approximations in practical implementations. Depending on the choice of measures, **UQ** adds its own set of issues. The somewhat paradoxical situation of having to estimate uncertainty while being uncertain raises the question of whether component additivity can be maintained in finite-sample situations. Moreover, the fact that uniform second-order distributions do not invoke maximum **EU** challenges a well-justified Bayesian principle. Further work in this important avenue should surely be encouraged.

2.4. Evaluating Uncertainty Estimates

This chapter summarizes how uncertainty estimates can be *evaluated* in the absence of an observable ground truth, distinguishing between *nominal* and *decision-based* criteria.

2.4.1. Nominal Evaluation

Calibration-Based Measures

We have frequently alluded to *faithful* uncertainty estimates, avoiding claims of correctness we cannot make due to the inherent subjectivity of the **EU** and the unobservable nature of the **AU**. In the following, we lay out how uncertainty estimates can be evaluated despite this fundamental opacity. We can largely discern two categories of evaluation protocols. The first is concerned with ***nominal*** levels of uncertainty in the sense that the predicted uncertainty is appropriate for a given observation relative to performance (at least, on average). When the model is too conservative, the uncertainty estimates are so vague as to be useless; too optimistic, and the predictions cannot be trusted. Nominal evaluation often allows for some degree of interpretability. On the downside, it typically remains on the level of **TU** (Sluijterman et al., 2024). The second type cares only about ***decisions*** based on uncertainty estimates, and judges the latter by the quality of the downstream action. This approach allows for a more fine-grained evaluation when decisions are made on the grounds of individual uncertainty components, yet it is rather heuristical.

Calibration For predictive densities, it is quite natural to demand that "[t]he probability that a system outputs for an event should reflect the true frequency of that event" (Kumar et al., 2019). This describes the concept of ***calibration*** (see, e.g., Filho et al., 2023, for an overview). Calibration has been studied mostly in the context of classification, although adaptations for regression exist (e.g., Song et al., 2019). Consider predictions $\boldsymbol{\theta} = p(y_+|\boldsymbol{x}_+, \mathcal{D}) \in \Delta_K$ and let k^* denote the predicted class invoking the highest probability, i.e., $k^* = \arg\max_{k\in\{1,\dots,K\}} \theta_k$. We call a classifier calibrated according to the following notions (which coincide for binary classification; Guo et al., 2017; Vaicenavicius et al., 2019; Widmann et al., 2019):

Weak calibration in classification A classifier is *weakly calibrated* if

$$\mathbb{P}(Y = k^*|\theta_{k^*}) = \theta_{k^*}.$$

Strong calibration in classification A classifier is *strongly calibrated* if

$$\mathbb{P}(Y = k|\boldsymbol{\theta}) = \theta_k \quad \forall k \in \{1, \dots, K\}.$$

Take the example of a diagnostic tool predicting a genetic malfunction in three categories {none, some, full}, and focus on the subset of patients for whom $\boldsymbol{\theta} = (0.9, 0.05, 0.05)$. Weak calibration is satisfied if the real share of patients in this group without malfunction is 90%. It is easy to see that the tool, were the true share of patients much higher or lower, would be of limited use to manage medical resources (and might falsely reassure or needlessly worry patients). Strong calibration demands further that the respective shares of patients with some and no malfunction be 5%. The same considerations hold for other groups with different prediction vectors.

Calibration should be viewed in conjunction with ***sharpness***. For example, a classifier always predicting the majority class with a probability equal to its empirical frequency will be perfectly calibrated (in the weak sense), but obviously useless. Furthermore, calibration is usually interpreted *marginally* across the entire test set, possibly smoothing over interesting sub-population-level effects. Some works have proposed more fine-grained notions based on subgroups (Hansen et al., 2024, for fairness) or domains (Wald et al., 2021, for distribution shifts).

In theory, the use of ***proper scoring rules*** (Gneiting and Raftery, 2007), which are optimized by the true DGP with correct probabilities, encourages model calibration. The standard ***cross-entropy*** NLL loss for classification is even *strictly* proper, i.e., the true DGP is its *unique* minimizer (in the absence of model misspecification and in the limit of infinite data). In practice, over- or underfitting can cause miscalibration (Guo et al., 2017; Mukhoti et al., 2020; Minderer et al., 2021), especially in non-Bayesian models that output *pseudo-probabilities* (e.g., softmax scores that look like probabilities but are not the result of a probabilistic updating rule). Unfortunately, as discussed in Sec. 2.2.4, BNNs are not immune to miscalibration either due to the numerous obstacles impeding ABI (Wilson and Izmailov, 2020).

Recalibration Besides encouraging calibration during training, there are methods to fix miscalibration *ex post* by means of ***recalibration***. This approach follows from the idea that every model has a canonical ***calibration map***, which is the identity for perfectly calibrated models. The calibration map can be used to rectify miscalibration (of course, in practice, it needs to be estimated; Vaicenavicius et al., 2019; Kängsepp et al., 2025). Many popular recalibration techniques are based on *tempering* predictions, which connects back to the CPE in BNNs.

Calibration Errors A number of ***calibration errors*** have been proposed to assess the quality of uncertainty estimates. The most common ones are based on the above *weak* notion of calibration, measuring the discrepancy between model *confidence* (in terms of top-1 probability) and *accuracy* (the share of observations actually belonging to the predicted class). Recall that calibration is a conditional notion grouping data based on having the *exact same* prediction. Obtaining reasonably good estimators from finite data usually requires a ***binning*** scheme to group data with *similar* predictions into $B < N$ bins. Let a_b and c_b be the accuracy and the average posterior probability of the predicted class, respectively, in the b-th bin that contains N_b observations. Then, we have the following measure according to Naeini et al. (2015); Guo et al. (2017), with $\downarrow$ indicating that lower values are better:

Expected classification error (ECE) The *ECE* is defined as

$$\downarrow \varepsilon_{\text{ECE}} = \sum_{b=1}^{B} \frac{N_b}{N} |a_b - c_b|.$$

Analogous to the ECE, we can define a ***maximum calibration error (MCE)*** via the maximum, rather than average, discrepancy between accuracy and confidence (Naeini et al., 2015; Guo et al., 2017). Binning is frequently based on equally-sized intervals of estimated probability. This simple equidistant rule is not always the most appropriate one. Other, adaptive schemes have been suggested (Nixon et al., 2019; Widmann et al., 2019). For instance, the ***adaptive calibration error (ACE)*** is designed so as to produce bins with a roughly constant number of observations in each (Nixon et al., 2019):

Adaptive classification error (ACE) The *ACE* is defined as

$$\downarrow \varepsilon_{\text{ACE}} = \frac{1}{KR} \sum_{k=1}^{K} \sum_{r=1}^{R} |a_{k,r} - c_{k,r}|.$$

The adaptive bins r are determined via quantiles in the ordered data (class-specific and depending on R)[36]. Note that the ACE is formulated as a *multi-class* measure by taking the sum over all classes; the same can be done to adapt the ECE. We use the ECE in [**C3**] and [**C6**] to evaluate **UE** in different ablations, and report ECE and (thresholded) ACE in our experiments for [**C6**].

With calibration errors in finite-data situations, we face a trade-off where too *many* observations per bin can gloss over miscalibration (especially when the effects of under- and overconfidence cancel out) and too *few* produce brittle estimates (Nixon et al., 2019). Overall, the necessity of binning remains a downside of calibration errors, not least because the choice of a binning strategy (number of bins, allocation) adds hyperparameters that must be set (Roelofs et al., 2022).

[36] Nixon et al. (2019) suggest that the ACE should possibly be thresholded if there are many points with near-zero probabilities; otherwise, the adaptive binning scheme can produce very heterogeneous bins by grouping observations with rarely occurring, quite different probability values.

Likelihood-Based Measures

A slightly different evaluation strategy targets the predictive distribution more directly. It is based on the idea that the PPD will attain large values on future observations if it captures the DGP well: then, under the assumed model, the test observations are highly likely. We can formulate this intuition via the ***expected log-PPD (LPPD)***, which measures the predictive density on an arbitrary instance drawn from p_{XY} (Gelman et al., 2014). It is equal to the negative KL divergence from the true distribution (up to a constant; Deshpande et al., 2024). We cannot compute the expected LPPD for lack of access to p_{XY}. If, as the assumption goes, the test observations are drawn *i.i.d.* from the DGP, the ***test LPPD*** is a Monte Carlo estimator for the expected LPPD (↑ signaling that higher values are better; Gelman et al., 2014):

Test LPPD The *test LPPD* is defined as

$$\uparrow \varepsilon_{\text{LPPD}} = \frac{1}{N} \sum_{i=1}^{N} \log p(y_+^{(i)} | \boldsymbol{x}_+^{(i)}, \mathcal{D}).$$

We use the test LPPD to evaluate sampling schemes in **[C1]**, **[C2]** and **[C6]**. Again, strictly proper scoring rules (e.g., cross-entropy or $L2$ loss) incentivize high test LPPD values because they encourage truthful predictions. Recall, however, that they are only guaranteed to recover the true density for correctly specified models and infinite data. The test LPPD can thus be negatively affected by misspecification and finite-size test data, possibly leading the test LPPD to misjudge posterior approximation quality (Deshpande et al., 2024).

Interval-Based Measures

In regression tasks, it is common to derive *intervals* $\mathcal{I}(\cdot)$ from the predictive density (some methods, like conformal prediction, directly output predictive intervals; e.g., Angelopoulos and Bates, 2023). For such interval-valued predictions, we are often interested in the degree of ***coverage***, indicating whether the true label lies within the predicted range. This can be formalized in the following intuitive manner (Sluijterman et al., 2024), where $\mathbb{I}[\cdot]$ denotes the indicator function:

Coverage probability (CP) The prediction interval *coverage probability* is defined as

$$\uparrow \varepsilon_{\text{CP}} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}[y_+^{(i)} \in \mathcal{I}(\boldsymbol{x}_+^{(i)})].$$

Vacuous sets trivially achieve coverage, which is why interval width is usually consulted alongside coverage to judge the faithfulness of uncertainty estimates (Kompa et al., 2021). We use coverage in **[C2]** to evaluate the posterior approximation quality of individual chains in MCMC.

2.4.2. Decision-Based Evaluation

Abstention

Judging uncertainty estimates by downstream decisions is rather less principled. In a way, it can afford to sidestep theoretical shortcomings so long as they do not interfere with decision-making. On the other hand, such frameworks run a risk of convoluting the goodness of uncertainty estimates with other, decision-related factors. It is probably good advice to use a combination of nominal and decision-based evaluation in order to assess both theoretical properties and practical usefulness.

A simple case of decision-based evaluation is prediction under ***abstention***, where the model refrains from predicting when the estimated uncertainty exceeds a threshold. The criterion can derive from total predictive uncertainty or either of its components. Performance is then evaluated only on the data for which predictions are available. Sliding the threshold from high to low uncertainty cut-offs allows to draw what we might call ***performance-rejection curves***. Models with faithful uncertainty estimates should exhibit steep performance increases for higher rejection rates, whereas useless estimates fail to filter the hard-to-predict cases (e.g., Hofman et al., 2024; Sale et al., 2024). Beyond being an evaluation technique, this process allows for intervention in real-world applications. For instance, decision-makers can opt to collect more information on the observation in question, or defer the decision to a human expert entirely.

Downstream Applications

Lastly, we can use the performance of downstream ***proxy tasks***, in which uncertainty estimates play some intermediate role, as a criterion to evaluate **UE**. Methods that make use of uncertainty include ***active learning*** (for data efficiency; Settles, 2010), ***sample prioritization*** (for faster training; Tata et al., 2022), ***pseudo-labeling*** (for partially supervised settings; Lee, 2013), and ***Bayesian optimization*** (for cost-efficient tuning; Jones and Schonlau, 1998). Since these techniques are predominantly interested in parts of the input space about which additional *knowledge* should be gathered, they often focus on the **EU** (Nguyen et al., 2022; Stanton et al., 2023; Rodemann et al., 2023). If **UE** works as desired, the uncertainty-informed acquisition should beat random sampling in efficiency by a solid margin. Another task that draws naturally on uncertainty is the detection of distribution shifts, the idea being that unusual observations invoke cautious predictions (e.g., Ovadia et al., 2019; Mucsányi et al., 2023). In all of these applications, different types of **UR** and **UQ** can appear. In the following, last section of Part I of this thesis, we present a number of applications in some more detail.

Summary: Evaluating Uncertainty

In the absence of an observable ground truth that would admit direct comparison, a plethora of evaluation measures has emerged. All come with their own advantages and shortcomings. Some of these criteria are applicable only to certain types of **UR** (e.g., likelihood-based measures), which impedes broader comparisons. In a recent publication, Manchingal et al. (2025) propose a unified evaluation framework. We believe that such work is vital for future research because evaluation maintains a gate-keeping effect on the progress in other areas of **UE**.

2.5. Applications

This chapter presents selected downstream *applications* for uncertainty estimates, notably, *active learning* as used in [C5] and *self-supervised learning* studied in [C6].

2.5.1. Active Learning

Active Learning Procedure

Most of supervised ML works with fully-labeled datasets. Making use of the entire data is usually the best strategy in these situations. Tasks in semi-supervised or automated ML, on the other hand, frequently entail a phase of sequential data acquisition under a limited budget. Cost-effective learning then requires careful selection of points to evaluate. Similarly, training models in applications where only a few labeled data are available and the rest need to be acquired from a (human) source relies on efficient use of the labeling budget. This is a typical situation in medical contexts depending on experts whose time is both short and valuable.

[C5] proposes an ***active learning (AL)*** pipeline for the classification of images from wildlife camera traps. Such cameras produce a large number of photographs, and having them all labeled by experts is neither practical nor necessary. Starting from a seed dataset with given labels, AL iteratively finds the most informative observations to acquire annotations for, allowing the model to be trained to good performance while keeping the labeling costs as low as possible. Extending the supervised setting, we now have data $\mathcal{D} = (\mathcal{D}_L, \mathcal{D}_U)$, where $\mathcal{D}_L \in (\mathcal{X} \times \mathcal{Y})^{N_L}$ is a set of *labeled* training data, and $\mathcal{D}_U \in \mathcal{X}^{N_U}$ is a pool of *unlabeled* observations. Let further

$$\alpha : \mathcal{H} \times \mathcal{X} \to \mathbb{R}, (h, \boldsymbol{x}) \mapsto \alpha(h, \boldsymbol{x})$$

be an ***acquisition function*** assigning some utility to observations under the current model. In addition, we have an ***oracle*** $o : \mathcal{X} \to \mathcal{Y}, \boldsymbol{x} \mapsto y$ returning labels for query instances. Alg. 2 provides an overview of the AL procedure (cf. Settles, 2010; Budd et al., 2021). It starts by training a model (by ERM or Bayesian updating, depending on the framework) on the initial set of labeled data. In each iteration, the current model is used to compute the utility of the observations in the unlabeled pool according to the acquisition function. The standard AL formulation assumes single-point acquisition in each round, but since this is expensive for models with high training cost, most applications in DL modify the procedure to acquire *batches* of data[37] (expressed by the top M operation in Alg. 2; Pinsler et al., 2019). After obtaining labels from the oracle, learning resumes with training on the newly enriched set of labeled data. The process runs until some stopping criterion, like the exhaustion of a predefined budget or the achievement of a target performance, is reached. While it has been suggested to only fine-tune the model of the previous iteration to the new data, AL implementations frequently accept the cost of training from scratch to avoid long-term dependencies in the optimization trajectory (Gal et al., 2017; Budd et al., 2021). That said, it is common to use pre-trained models in the sense that a part of the weights is inherited from some upstream task and kept fixed during AL. We also do this in [C5].

[37] Acquiring data in batches is not only cheaper but actually reasonable in overparameterized models, where the addition of a single observation to the dataset is not likely to effect much change in the parameters.

Algorithm 2 Active learning

Input: Labeled data $\mathcal{D}_L$, unlabeled data $\mathcal{D}_U$, hypothesis space $\mathcal{H}$, acquisition function α, oracle o, batch size M, stopping criterion $\mathcal{S}$
$j \leftarrow 1$, $\mathcal{D}_L^{[0]} \leftarrow \mathcal{D}_L$, $\mathcal{D}_U^{[0]} \leftarrow \mathcal{D}_U$
while $\mathcal{S}$ not met **do**
  Train $h^{[j]} \in \mathcal{H}$ on $\mathcal{D}_L^{[j-1]}$
  Acquire $\mathcal{B}_{\boldsymbol{x}} = \operatorname{top} M[\alpha(h^{[j]}, \boldsymbol{x}_+)]$ over $\boldsymbol{x}_+ \in \mathcal{D}_U^{[j]}$
  Query $\mathcal{B}_y = \{y_+ \leftarrow o(\boldsymbol{x}_+)\}$ for $\boldsymbol{x}_+ \in \mathcal{B}_{\boldsymbol{x}}$
  Update $\mathcal{D}_U^{[j]} \leftarrow \mathcal{D}_U^{[j-1]} \setminus \mathcal{B}_{\boldsymbol{x}}$, $\mathcal{D}_L^{[j]} \leftarrow \mathcal{D}_L^{[j-1]} \cup (\mathcal{B}_{\boldsymbol{x}}, \mathcal{B}_y)$
  $j \leftarrow j + 1$
end while
return $h^{[j-1]}$

Uncertainty-Based Label Acquisition

Uncertainty in Acquisition It is easy to see how the success of AL depends on the ability to acquire high-utility observations. The acquisition function has the role of a *surrogate*: we try to guess, before knowing its label, how much adding an observation to the training set helps improve the empirical risk (or infer the posterior distribution). This notion of informativeness is operationalized by finding especially *representative* (in the sense of high density under the current training distribution) or *heterogeneous* (in the sense of low predictive confidence with the current knowledge) observations (Aggarwal et al., 2014). We are interested in the heterogeneity-based technique of ***uncertainty sampling***, where the acquisition function provides a score $\mathcal{U}$ depending on the predictive uncertainty assigned to unlabeled observations $\boldsymbol{x}_+$:

$$\alpha(h, \boldsymbol{x}_+) = \mathcal{U}(h, \boldsymbol{x}_+). \tag{19}$$

In principle, this can be any type of uncertainty measure derived by **UQ** from any type of **UR**. The entropy decomposition of Eq. (17) was originally proposed for the purpose of AL, advising to use the mutual information (i.e., **EU**) term as an acquisition criterion (Houlsby et al., 2011). Across a number of experiments, Nguyen et al. (2022) confirm that the **EU** works well for many cases, especially for more expressive learners with weak inductive biases. They note further that **EU**-based acquisition bears similarities to criteria of expected model change[38].

Budget Considerations We have argued in Sec. 2.2 for a bi-level **UR** to avoid model overconfidence. For AL, which is light on the labeling but heavy on the computational budget, the cost of representing the various sources of uncertainty must be carefully weighed against the benefit for data acquisition. For example, Beluch et al. (2018) note that uncertainty estimates obtained from (deep) ensembles achieve superior performance but come at a high computational overhead. We find in [**C5**] that a single-level distributional representation only capturing the **AU** serves to reduce sampling costs effectively. While a Bayesian formulation may have accelerated the sampling procedure (as suggested in Gal et al., 2017), opting for a cheaper compromise reflected the budgetary constraints of the given application.

[38] This seems to recall the interpretation of the mutual information as the expected reduction in uncertainty about the model parameters by observing the true label—observations with high **EU** might be seen, then, as particularly influential for the shape of a model if they were part of the training data.

Sample Selection Bias An important thing to note is that the adaptive acquisition scheme in AL introduces a distributional mismatch between the labeled and unlabeled data. By design, the training dataset is populated with the observations of highest utility in any given iteration. The resulting set consists of points that are not random, but carefully chosen, samples from the DGP. Compromising the *i.i.d.* paradigm this way introduces a ***sample selection bias*** (Dasgupta and Hsu, 2008; Moreno-Torres et al., 2012). Interestingly, Farquhar et al. (2021); Murray et al. (2021) find that sample selection bias can be *helpful* in overparameterized models because it acts as a regularizing force. Since uncertainty sampling leads to a collection of hard-to-fit points, the acquisition process increases the difficulty of the task relative to model complexity. Intuitively, this effect is especially pronounced in early iterations with small training sets, where a strong, optimistic overfitting bias is countered by a pessimistic sample selection bias.

2.5.2. Distribution Shift Detection

Types of Distribution Shift

While ***distribution shifts*** exist in AL by design, they are harder to foresee in other settings. As briefly mentioned in Sec. 2.1.4, uncertainty estimates are being used to *detect* OOD data, following the rationale that unusual observations should elicit high predictive uncertainty (e.g., Ovadia et al., 2019). Moreno-Torres et al. (2012) propose a comprehensive framework that characterizes distribution shifts—understood as happening between the training and the test phase—arising from the relationship between features and targets. They distinguish what they call $X \to Y$ and $Y \to X$ problems. In $X \to Y$ problems, the features are causative for the class label (e.g., in credit scoring, where covariates determine the ability to pay back loans), whereas for $Y \to X$, the class controls the distribution of features (e.g., in medical diagnosis, where conditions determine symptoms). At least three types of distribution shift can occur. ***Covariate shift*** describes the situation where p_X evolves from training to test, relevant only in $X \to Y$ situations (i.e., a change in the instance-labeling process). An analogous shift in p_Y, applicable to $Y \to X$ problems (i.e., a change in the instance-generating process), is called ***label*** or ***prior probability shift***. Both of these cases affect only marginal distributions. Changes of the *conditional* distribution $p_{Y|X}$ ($X \to Y$) or $p_{X|Y}$ ($Y \to X$) are called ***concept shift***. Fig. 2.15 illustrates the different cases in causal graphs, where C denotes a context variable effecting shifts. A related concept that might

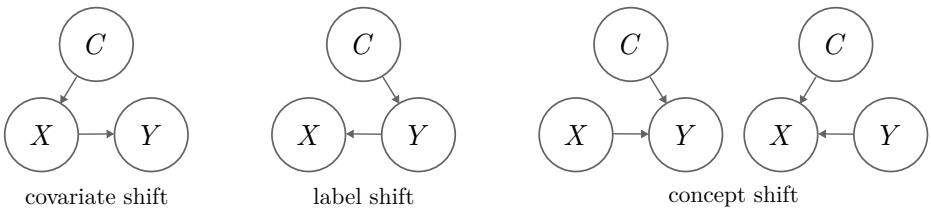

Figure 2.15.: Types of distribution shift, as characterized in Moreno-Torres et al. (2012), depicted by directed acyclic graphs. *Left:* change in marginal distribution p_X. *Middle:* change in marginal distribution p_Y. *Right:* change in conditional distribution $p_{Y|X}$ or $p_{X|Y}$. Adapted from Kull and Flach (2014).

be framed as a special case of label shift is given by ***open set recognition***, exposing models to new classes that were never present during training (Mucsányi et al., 2023).

Uncertainty for Shift Detection

Idea Intuitively, situations of concept shift are most difficult to handle because they affect the very conditional density we try to model in supervised ML (Kull and Flach, 2014). The hope is that such shifts can be detected with faithful **UE**. As the name suggests, distribution shifts are a *gradual* notion. Even instances that seem to originate from a completely different DGP (e.g., images of animals fed to a model trained on images of cars) need not have zero density under the training distribution. With rising OOD-ness, we expect our model to report progressively higher predictive uncertainty (Ovadia et al., 2019). In particular, it seems reasonable to assume that uncertainty on OOD data should manifest in the **EU** component, with the model signaling that no single hypothesis is especially plausible to have generated the unusual observation. This reflects the conflict-driven notion embodied by **EU** measures like the mutual information (D'Angelo and Fortuin, 2021; Tran et al., 2022; Mucsányi et al., 2023). As mentioned in Sec. 2.4, OOD detection can be used in this spirit as a proxy task to *evaluate* uncertainty estimates. We employ the OOD evaluation technique in the experiments of [**C3**] and [**C6**]. A separate line of research has focused on *explaining* distribution shifts (Kulinski and Inouye, 2023), which might bear interesting implications for understanding **UE**.

Criticism We note in [**C4**] that the ability to recognize OOD data as unusual depends fundamentally on the latent representation realized by the model. When the prediction relies on spurious patterns, or shortcuts (recall the example in Sec. 2.1.4 of classifying a golf ball on grass as a cow), there is no way of knowing how the model will react to an OOD observation. In a recent publication, Li et al. (2025) claim that OOD detection asks a question it cannot hope to answer. The authors argue that, rather than investigating "whether an input belongs to the training distribution or some different distribution, [**UE** methods] instead ask if the input leads to atypical model representations or unconfident predictions". Furthermore, the paper makes the interesting observation that the fundamental idea of *reducibility* in **EU** undermines its use in OOD detection: "If measuring epistemic uncertainty were the correct approach to OOD detection, then low epistemic uncertainty implies that OOD points do not exist in this setting [of infinite in-distribution data]". To reconcile the criticism in Li et al. (2025) with the ongoing use of uncertainty to identify distribution shifts, it can be guessed that many applications purporting to do OOD detection actually *wish* to answer the question if the input evokes "unconfident predictions". This is certainly the case for the ablations in [**C3**] and [**C6**]. In that sense, it might be that the proxy task of OOD detection is merely framed in an unfortunate manner.

2.5.3. Self-Supervised Learning

Self-Supervised Paradigm

The last application stems from a somewhat different field that has gained attention with the rise of natural language processing and, more recently, foundation models. ***Self-supervised learning (SSL)*** creates from auxiliary ***pretext tasks*** latent representations of the data to be used in downstream problems. This concept implements the idea that underlying patterns in the input (e.g., text semantics) are helpful for many tasks irrespective of the particular goal (e.g., hate speech detection or machine translation; Hendrycks et al., 2019). Crucially, the pretext task is solved

without any labels. It exploits relations within the data by minimizing special loss functions. A rich body of literature exists now that builds on early work concerned mainly with dimensionality reduction (Hinton and Salakhutdinov, 2006). Many methods are representatives of what Arora et al. (2019) termed ***contrastive learning*** (see, e.g., Jaiswal et al., 2020, for a comprehensive review). Contrastive methods use loss functions that encourage similarity in representations, or ***embeddings***, of semantically similar observations (*positive* pairs). Architectures in contrastive SSL often assume a Siamese-like structure (cf. Fig. 2.16; Chen and He, 2021; Zbontar et al., 2021). Obviously, the notion of semantic similarity needs to be operationalized in some way. Chen

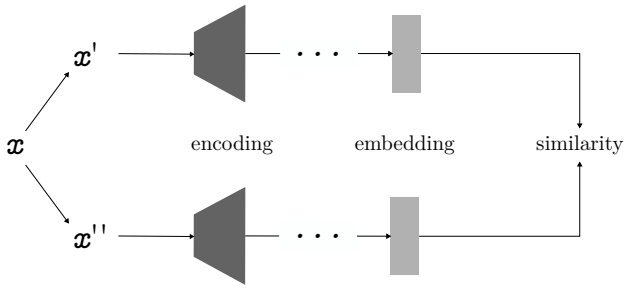

Figure 2.16.: Siamese network structure in SSL. Two augmented versions of input $\boldsymbol{x}$, constituting a positive pair, are embedded into the latent space under optimization of a similarity-based loss. Adapted from [C6].

et al. (2020) proposed the now-popular ***SimCLR*** framework, where positive pairs are created by augmentation of the original input, and the loss function formalizes similarity via an inner product between the pair's embedding vectors. In order to prevent the embeddings of positive pairs from collapsing to the same point in the latent space, it has proven effective to enforce decorrelation Zbontar et al. (2021); Bardes et al. (2022). Huang et al. (2023) find in a formal analysis that contrastive SSL creates well-generalizing embeddings when positive samples are aligned in the same latent class, and the latent classes are well separated.

Uncertainty in Intermediate Representations

Incorporating Multiplicity Most SSL architectures are trained to map each augmented input to a single point in latent space. Vilnis and McCallum (2015) realized early that distributional representations, which they created as Gaussians over word embeddings, may improve over point estimates because they capture uncertainty and produce a regularizing effect. Oh et al. (2019) later extended the idea to arbitrary distributions that admit sampling. [C6] takes a more pragmatic approach, exploiting the conceptual simplicity and empirical success of DEs. Rather than making distributional assumptions, we realize multiple embeddings per instance by employing an ensemble of *sub-networks* (ensembling only parts of the architecture, much like in Havasi et al., 2021). This enhanced embedding block can be used as a plug-in component for any contrastive-type SSL network at relatively low computational overhead. Specifically, [C6] explores the benefit of ensemble embeddings for twin-like architectures as in Fig. 2.16.

Enforcing Diversity Many authors have advocated for ***diversity*** in ensembles, citing it as a decisive factor for DE's success (Stickland and Murray, 2020; Rame and Cord, 2021; Nam et al., 2021; Turkoglu et al., 2022). Usually, diversity is understood as the variability in

predictions for a given observation. Abe et al. (2022) suggest that diversity drives the "Jensen gap", i.e., the expected performance gain of the ensemble over the average member's prediction[39]. In a recent paper, Wang and Wang (2025) challenge this view somewhat. They argue that the average performance of individual members has little practical relevance, and that the diversity of predictions across *different* observations must be taken into consideration. For the purpose of SSL, where the hedging effect of embedding multiplicity is arguably most important, we find that promoting observation-level diversity is beneficial. To this end, **[C6]** adds a term to the SimCLR loss that penalizes ensemble embeddings whose variability fails to meet a threshold value.

Summary: Using Uncertainty

Multiplicity can be helpful to hedge against overconfidence in intermediate representations, justifying the additional cost of producing multiple latent outputs. Uncertainty estimates are also being used more directly as decision criteria for downstream tasks. We have presented examples of dynamic data acquisition and detection of OOD observations in this regard. Overall, the usefulness of **UE** depends critically on the quality of solutions to the **UR** and **UQ** subproblems.

This concludes the introductory notes motivating the contributions in the following Part II of this thesis. We will revisit the current challenges for **UE** in Part III and end with an outlook on promising directions for future research.

[39] For any strictly convex loss L, Jensen's inequality states that $L(y, \bar{h}) \leq \mathbb{E}_{Q(\boldsymbol{\omega})} L(y, h_{\boldsymbol{\omega}}(\boldsymbol{x}))$, where $\bar{h}$ denotes the consensus obtained from averaging over the ensemble members' predictions (Abe et al., 2022).

II

Contributions

3. Uncertainty Representation

3.1. [C1] Exploiting Symmetry in Bayesian Neural Network Inference

Contributing Article

J. Gregor Wiese*, **Lisa Wimmer***, Theodore Papamarkou, Bernd Bischl, Stephan Günnemann, David Rügamer (2023). Towards Efficient MCMC Sampling in Bayesian Neural Networks by Exploiting Symmetry. *In: Koutra, D., Plant, C., Gomez Rodriguez, M., Baralis, E., Bonchi, F. (eds) Machine Learning and Knowledge Discovery in Databases: Research Track. ECML PKDD 2023. Lecture Notes in Computer Science, vol. 14169, pp. 459–474, Springer, Cham.* https://doi.org/10.1007/978-3-031-43412-9_27.

An extended abstract of this article was further published as an invited contribution:

J. Gregor Wiese*, **Lisa Wimmer***, Theodore Papamarkou, Bernd Bischl, Stephan Günnemann, David Rügamer (2024). Towards Efficient MCMC Sampling in Bayesian Neural Networks by Exploiting Symmetry (Extended Abstract). *In: Larson, K. (eds) Proceedings of the 33rd International Joint Conferences on Artificial Intelligence (IJCAI-24). Sister Conferences Best Papers Track, pp. 8466-8470.* https://doi.org/10.24963/ijcai.2024/943.

* Shared first authorship.

Author Contributions

JGW and LW share first authorship. JGW supplied the fundamental ideas for the reformulation of the posterior predictive density and the Markov chain upper bound. LW helped derive the formal characterization of the weight space via the posterior reference set. LW played a key role in shaping the structure of the manuscript and took major responsibility for the text that was drafted jointly by JGW, LW and DR with valuable input from TP. JGW developed the algorithm for symmetry removal and conducted the sampling-based part of the experiments, while LW implemented and conducted experiments for the sampling-free methods. BB and SG provided critical feedback and revisions. DR and TP took on active supervision, offering guidance and support throughout the project.

Note: The project is follow-up work to JGW's master's thesis which he completed at Technical University of Munich, advised by DR and formally supervised by SG.

Awards The article was elected ”Best Paper” in the Research Track of ECML PKDD 2023.

Supplementary Material `https://github.com/jgwiese/mcmc_bnn_symmetry/blob/main/sub_44_supplementary_material.pdf`.

Code Repository `https://github.com/jgwiese/mcmc_bnn_symmetry`.

3.2. [C2] Feasible Sample-Based Inference via Mode-Connectedness

Contributing Article

Emanuel Sommer*, **Lisa Wimmer***, Theodore Papamarkou, Ludwig Bothmann, Bernd Bischl, David Rügamer (2024). Connecting the Dots: Is Mode-Connectedness the Key to Feasible Sample-Based Inference in Bayesian Neural Networks? *In: Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J. Berkenkamp, F. (eds) Proceedings of the 41st International Conference on Machine Learning, pp. 45988–46018, PMLR 235.* `https://proceedings.mlr.press/v235/sommer24a.html`.

* Shared first authorship.

Author Contributions

ES and LW share first authorship. LW supplied large parts of the background and literature review. LW played a key role in shaping the structure of the manuscript and took major responsibility for the text that was drafted jointly by ES, LW and DR with valuable input from TP. ES conceived, implemented and conducted the empirical investigation, including a comprehensive coding framework, convergence diagnostics, and numerous experiments for sampling-based inference. LW implemented and conducted experiments for the sampling-free methods. LB and BB provided critical feedback and revisions. DR took on active supervision, offering guidance and support throughout the project.

Code Repository `https://github.com/EmanuelSommer/bnn_connecting_the_dots`.

4. Uncertainty Quantification

4.1. [C3] Pitfalls of Quantifying Uncertainty with Entropic Measures

Contributing Article

Lisa Wimmer, Yusuf Sale, Paul Hofman, Bernd Bischl, Eyke Hüllermeier (2023). Quantifying Aleatoric and Epistemic Uncertainty in Machine Learning: Are Conditional Entropy and Mutual Information Appropriate Measures? *In: Evans, R. J. and Shpitser, I. (eds) Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence, pp. 2282–2292, PMLR 216.* `https://proceedings.mlr.press/v216/wimmer23a.html`.

Author Contributions

LW, as the first author, shaped the structure of the manuscript and took major responsibility for the text that was drafted mainly by LW and EH. LW conceived, implemented and conducted the experiments with valuable input from PH. YS fleshed out the axiomatic definition of desirable measures based on ideas by EH, streamlined mathematical notation, and gave helpful feedback on the manuscript. BB provided critical feedback and revisions. EH initiated the project and took on active supervision, offering guidance and support throughout the project.

Supplementary Material
`https://proceedings.mlr.press/v216/wimmer23a/wimmer23a-supp.pdf`.

Code Repository `https://github.com/lisa-wm/entropybaseduq`.

4.2. [C4] Predictive Uncertainty in the Presence of Shortcut Learning

Contributing Article

Lisa Wimmer, Bernd Bischl, Ludwig Bothmann (2025). Trust Me, I Know the Way: Predictive Uncertainty in the Presence of Shortcut Learning. *In: Workshop on Spurious Correlation and Shortcut Learning: Foundations and Solutions at the 13th International Conference on Learning Representations (ICLR).* `https://arxiv.org/abs/2502.09137`.

Author Contributions

LW, as the first author, developed the idea and was responsible for the design, implementation, and analysis of the empirical study. LW wrote the manuscript, with LB and BB providing critical feedback and revisions. LB took on active supervision, offering guidance and support throughout the project.

Code Repository `https://github.com/lisa-wm/shortcuts_uncertainty`.

5. Downstream Applications

5.1. [C5] Uncertainty-Informed Active Learning for Wildlife Image Classification

Contributing Article

Ludwig Bothmann, **Lisa Wimmer**, Omid Charrakh, Tobias Weber, Hendrik Edelhoff, Wibke Peters, Hien Nguyen, Caryl Benjamin, Annette Menzel (2023). Automated wildlife image classification: An active learning tool for ecological applications. *Ecological Informatics 77, pp. 102231.* `https://doi.org/10.1016/j.ecoinf.2023.102231`.

Author Contributions

LB, as the first author, developed the idea and was responsible for the conception, design, and methodology. LB wrote the initial code base together with OC. LW provided critical feedback and revisions to the manuscript mainly written by LB. LW and TW developed a comprehensive coding framework, which is available as a package to third-party users, based on code by LB and OC. LW implemented and conducted the empirical investigation for the article. HN helped preprocessing the data. HE, WP, HN, CB, and AM supplied domain knowledge and valuable input to the manuscript.

Code Repository `https://github.com/slds-lmu/wildlife-ml`, `https://github.com/slds-lmu/wildlife-experiments`.

5.2. [C6] Leveraging Sub-Network Ensembles in Self-Supervised Learning

Contributing Article

Amirhossein Vahidi, **Lisa Wimmer**, Hüseyin Anil Gündüz, Bernd Bischl, Eyke Hüllermeier, Mina Rezaei (2024). Diversified Ensemble of Independent Sub-Networks for Robust Self-Supervised Representation Learning. *In: Bifet, A., Davis, J., Krilavičius, T., Kull, M., Ntoutsi, E., Žliobaitė, I. (eds) Machine Learning and Knowledge Discovery in Databases. Research Track. ECML PKDD 2024. Lecture Notes in Computer Science, vol. 14941, pp. 38–55, Springer, Cham.* https://doi.org/10.1007/978-3-031-70341-6_3.

Author Contributions

AV, as the first author, developed the idea together with MR. LW supplied background knowledge on uncertainty quantification and took on responsibility for critical revision of the text that was drafted mainly by AV and MR. LW designed, implemented and conducted the experiments of the NLP task of the empirical evaluation, while HAG did the same for the T6SS identification task. BB and EH provided critical feedback and revisions. MR took on active supervision, offering guidance and support throughout the project.

Note: The project is follow-up work to AV's master's thesis which he completed at LMU Munich, formally supervised by MR.

Supplementary Material https://static-content.springer.com/esm/chp%3A10.1007%2F978-3-031-70341-6_3/MediaObjects/629146_1_En_3_MOESM1_ESM.pdf.

III

Conclusion

Concluding Remarks

Contributions The importance of teaching models to output faithful uncertainty estimates alongside their predictions can hardly be overstated. We have contributed to this endeavor by advancing the ***representation*** and ***quantification*** of predictive uncertainty in several regards. Articles [**C1**] and [**C2**] explore for distributional representations how sampling-based Bayesian inference can be facilitated. To this end, we exploit the special structures induced in the posterior landscape by overparameterization. The almost trivial-seeming task of attaching numeric values to uncertainty representations proves surprisingly delicate. A wide-spread set of measures based on entropic quantities suffers from inconsistencies pointed out by contributions [**C3**] and [**C4**]. Still, there is much to be gained by good solutions to **UE**. Multiplicity is at play in many stages of the learning process—besides the actual predictions, latent representations and decisions along the way benefit from allowing for uncertainty. Articles [**C5**] and [**C6**] provide examples of employing intermediate-stage uncertainty estimates that improve performance in downstream tasks. Overall, there seems to be a unifying rationale of not putting all one's eggs into a single basket when facing finite-resource settings in an increasingly more ambiguous environment.

Limitations Our contributions necessarily leave some aspects unresolved. For one, our work studies specific, often small, models and tasks that are simplified compared to the world outside. This builds intuition but naturally falls short of capturing all complexities. Our focus on Bayesian deep learning and entropic uncertainty measures comes at the expense of other noteworthy approaches to the foundational topics of this thesis. Similarly, the applications in active and self-supervised learning are but small steps in fields deserving of greater attention from the community. More fundamentally, there are sources of multiplicity that the contributions in this thesis do not account for. Uncertainties induced by data collection—from the conception of the task to preprocessing choices—and algorithmic procedures warrant more careful consideration in particular. We trust that these and further challenges will be taken on by future research.

Open Challenges

It is still early days in **UE**. Even ML as a whole is far from a mature field. The past years have seen a constant influx of new contributors, and more than one hype train calling at the station. Much of this is exciting and commendable: entry barriers have come down for young researchers, novel methods are being proposed at a fast rate, and a general spirit of open-access publishing encourages democratization. At the same time, growth and acceleration have caused the field to swell to a size that brings about fragmentation. The community, if we can still call it that, is entirely too large to keep track of all the diverse ideas out there, let alone exchange in dialogue. This carries a risk of overlooking cross-connections that could steer the publication frenzy into fruitful directions. The amount of resources being poured into ML—which is something to praise for sure—quietly enables the type of inward-looking research that is oblivious to a higher purpose. It is to be hoped that upcoming work will invite insights from other disciplines, such as causal inference or the sciences ultimately using ML, and take on a more comprehensive view. We identify a number of key challenges to be confronted by the collective field.

Understand Learning Dynamics The ongoing work on *learning dynamics* of large NNs is urgently needed to improve the representation of uncertainty. Sec. 2.2.4 attempted to reconcile the current state of Bayesian inference with knowledge about DL in the overparameterized regime. Questions of non-identifiability, regularization and generalization concern predictive uncertainty estimates immediately: our ability to capture the structure of the posterior landscape determines how well the **EU** is represented, affecting, in turn, its quantification and that of the **AU**. Still, a conclusive picture is wanting about much that is going on. Million-dimensional spaces preclude any chance of intuitive reasoning. While empirical studies can make informed guesses, it seems that **UE** would benefit from more rigorous mathematical theory. Otherwise, the absence of observable uncertainties might trap the field in a habit of piling up conjectures that are hard to refute until superseded by contradictory evidence. LLMs and their constant interaction with unsuspecting users make this issue all the more pressing. Research into **UE** for foundation models is happening (e.g., Yadkori et al., 2024; Aichberger et al., 2024) but still in its infancy. Solid theoretical frameworks will be necessary to get a firmer grip on these huge black boxes.

Develop Evaluation Protocols We have repeatedly lamented the lack of an observable ground truth for predictive uncertainty. While the **AU** is merely latent, and thus at least accessible in simulation studies, the inherent subjectivity of the **EU** calls the very existence of a true value into question. The community needs to find better *evaluation* protocols for **UE** nonetheless. At the moment, most works are forced to pick their choice of evaluation metrics with notorious deficiencies. There is no such thing as a perfect metric, but more transparency of evaluation protocols could help to create awareness (and perhaps, better metrics). Transparency should also extend to documenting systematically which sources of uncertainty are accounted for and which are disregarded. Overall, a higher degree of standardization seems indispensable. This includes benchmark datasets with known uncertainty and easy-to-use software. The latter is especially important to streamline research efforts by saving time on writing what could be boilerplate code, and to meet a higher standard of quality. Some notable efforts have been made already (e.g., Mucsányi et al., 2024; Lehmann et al., 2025); continued improvement should be encouraged.

Make Uncertainty Affordable Representing multiplicity often comes with an increased computational burden. Several hypotheses and predictions require more budget than one of each. The cost for some approaches, such as MCMC sampling, is downright prohibitive in big models. Efforts on developing new methods thus need to prioritize *cost efficiency* to ensure adoption across the community. This is not merely an economic concern—budgetary considerations matter on a larger scale. In front of a laptop, it is all too easy to forget about the environmental consequences of training, evaluating and prompting models. For the inherently cost-multiplying research on **UE**, however, effects on climate must be seriously factored in. Impact statements in conference proceedings are a laudable impulse, but obviously, more is needed to reduce our share of AI's ecological footprint. From a societal perspective, driving down costs is imperative to create a more inclusive scientific field. It is no secret that ML research is dominated by institutions in the so-called West which tend to have access to financial resources. We cannot afford to confront other, less well-situated actors with a choice between securing a place at the table and sacrificing scientific rigor. Therefore, developing approaches to **UE** that are simultaneously principled and inexpensive poses a major task for future research.

Communicate Uncertainty Lastly, our field faces a challenge beyond technical concerns. We must find ways of *communicating* uncertainty such that it is actually helpful. This reaches deep into human psychology. Even if we manage to overcome all obstacles of representing and quantifying uncertainty in standardized, cost-efficient frameworks, there is no guarantee that the results are put to the intended use. Take the example of a radiologist receiving support from a diagnostic tool. How can we make sure she understands that any estimated probability of a patient developing cancer is contingent on a sheer endless number of modeling choices? The diagnosis might have been different had a slightly sharper prior, or a more aggressive optimizer, or a somewhat modified dataset been used. Decision-makers can be easily overwhelmed by such complexity that bears little relation to their own fields of expertise. This is by no means a criticism of users' capabilities, but simply to acknowledge the immense difficulty of judging products of predictive modeling. While the last point holds for ML in general, there is reason to worry that **UE** could aggravate the danger of misinterpretation. Uncertainty estimates might create a false sense of security by providing a range or probability to contextualize predictions. Paradoxically, a gradual confidence for patients being cancer-free, suggesting careful deliberation of possible scenarios, could inspire greater trust than a blunt yes-or-no diagnosis. Of course, enforcing trustworthy predictions is the very goal of **UE**. However, there is no way of knowing which level of statistical literacy, technological skepticism or risk aversion our estimates will meet in human decision-makers. We cannot hope to make AI tools safe beyond failure. Still, at some point, the community will have to engage in discussions with the public at large. We perceive at least three aspects to tackle.

1. *Education.* In the era of AI, we are surrounded by data and numbers. Enabling people to handle this information is paramount for them to retain agency and assess narratives put out by actors with multitudinous agendas, be they well-meaning or malevolent.

2. *Transparency.* Building on the case for standardized evaluation protocols, we should hold ourselves to high standards regarding the comprehensive and transparent communication of uncertainty. Work on explaining uncertainty estimates is already ongoing (e.g., Van Der Bles et al., 2019; Antorán et al., 2021; Watson and O'Hara, 2023; Spiegelhalter, 2024), and movements of open science have much to offer on the topic of disseminating scientific results.

3. *User centricity.* In what is possibly the hardest challenge, we need to meet people where they are. Teaching initiatives and academic standards alone cannot ensure adequate reception of model outputs. Humans come with diverse cultural, societal and educational backgrounds, and we should make every effort at tailoring uncertainty estimates to their individual perspectives. Here, finally, the LLMs we have criticized for being opaque offer an unprecedented opportunity: they communicate in a customizable, deeply human-sounding fashion that allows for back-and-forth conversation. With the right intentions, they could serve as mediators in the dialogue between science and the public at a broader scale.

Despite the numerous cautionary tales in the preceding pages, we thus end this thesis on an optimistic note in reminiscence of a pioneering physicist and reasoner about uncertainty.

> "I can live with doubt and uncertainty and not knowing. (...) In order to make progress, one must leave the door to the unknown ajar." *Richard Feynman*

References

[1] T. Abe, E. K. Buchanan, G. Pleiss, R. Zemel, and J. P. Cunningham. Deep Ensembles Work, But Are They Necessary? In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[2] L. Adilova, M. Andriushchenko, M. Kamp, A. Fischer, and M. Jaggi. Layer-wise Linear Mode Connectivity. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024.

[3] B. Adlam, J. Snoek, and S. L. Smith. Cold Posteriors and Aleatoric Uncertainty. *Workshop on Uncertainty and Robustness in Deep Learning at the 37th International Conference on Machine Learning (ICML)*, 2020.

[4] C. C. Aggarwal, X. Kong, Q. Gu, J. Han, and P. S. Yu. Active learning: A survey. In C. C. Aggarwal, editor, *Data Classification*. CRC Press, 2014.

[5] L. Aichberger, K. Schweighofer, M. Ielanskyi, and S. Hochreiter. How many Opinions does your LLM have? Improving Uncertainty Estimation in NLG. In *Workshop on Secure and Trustworthy Large Language Models at the 12th International Conference on Learning Representations (ICLR)*, 2024.

[6] S. K. Ainsworth, J. Hayase, and S. Srinivasa. Git Re-Basin: Merging Models modulo Permutation Symmetries. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, 2023.

[7] L. Aitchison. A statistical theory of cold posteriors in deep neural networks. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021.

[8] F. Albertini and E. D. Sontag. Uniqueness of weights for neural networks. In R. J. Mammone, editor, *Artificial Neural Networks for Speech and Vision*, number 4 in Chapman & Hall Neural Computing Series. Chapman & Hall, London, first edition, 1994.

[9] A. Alexos, A. Boyd, and S. Mandt. Structured Stochastic Gradient MCMC. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022.

[10] C. Andrieu, N. de Freitas, A. Doucet, and M. I. Jordan. An Introduction to MCMC for Machine Learning. *Machine Learning*, 50:5–43, 2003.

[11] A. N. Angelopoulos and S. Bates. Conformal Prediction: A Gentle Introduction. *Foundations and Trends® in Machine Learning*, 16(4):494–591, 2023.

[12] J. Antorán, U. Bhatt, T. Adel, A. Weller, and J. M. Hernández-Lobato. Getting a CLUE: A Method for Explaining Uncertainty Estimates. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021.

[13] J. Antorán, D. Janz, J. U. Allingham, E. Daxberger, R. Barbano, E. Nalisnick, and J. M. Hernández-Lobato. Adapting the Linearised Laplace Model Evidence for Modern Deep Learning. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022.

[14] J. Arbel, K. Pitas, M. Vladimirova, and V. Fortuin. A Primer on Bayesian Neural Networks: Review and Debates, 2023.

[15] M. A. Armenta and P.-M. Jodoin. The Representation Theory of Neural Networks. *Mathematics*, 9(24), 2021.

[16] S. Arora, H. Khandeparkar, M. Khodak, O. Plevrakis, and N. Saunshi. A Theoretical Analysis of Contrastive Unsupervised Representation Learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

[17] S. Asmussen and P. W. Glynn. A new proof of convergence of MCMC via the ergodic theorem. *Statistics & Probability Letters*, 81(10):1482–1485, 2011.

[18] R. F. Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2): 455–482, 2020.

[19] R. F. Barber, E. J. Candes, A. Ramdas, and R. J. Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507, 2021.

[20] A. Bardes, J. Ponce, and Y. LeCun. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning. In *Proceedings of the 10th International Conference on Learning Representations (ICLR)*, 2022.

[21] M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine learning practice and the bias-variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32): 15849–15854, 2019.

[22] W. H. Beluch, T. Genewein, A. Nurnberger, and J. M. Kohler. The Power of Ensembles for Active Learning in Image Classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[23] V. Bengs, E. Hüllermeier, and W. Waegeman. Pitfalls of Epistemic Uncertainty Quantification through Loss Minimisation. In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[24] G. W. Benton, W. J. Maddox, S. Lotfi, and A. G. Wilson. Loss Surface Simplexes for Mode Connecting Volumes and Fast Ensembling. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

[25] J. M. Bernardo and A. F. M. Smith. *Bayesian Theory*. Wiley, 1994.

[26] M. Betancourt. A Conceptual Introduction to Hamiltonian Monte Carlo, 2018.

[27] F. Bickford Smith, J. Kossen, E. Trollope, M. van der Wilk, A. Foster, and T. Rainforth. Rethinking Aleatoric and Epistemic Uncertainty. In *Workshop on Bayesian Decision-making and Uncertainty at the 38th Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[28] R. Binkyte, I. Sheth, Z. Jin, M. Havaei, B. Schölkopf, and M. Fritz. Causality Is Key to Understand and Balance Multiple Goals in Trustworthy ML and Foundation Models, 2025.

[29] B. Bischl, M. Binder, M. Lang, T. Pielok, J. Richter, S. Coors, J. Thomas, T. Ullmann, M. Becker, A.-L. Boulesteix, D. Deng, and M. Lindauer. Hyperparameter Optimization: Foundations, Algorithms, Best Practices and Open Challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 13(2), 2023.

[30] C. M. Bishop and H. Bishop. *Deep Learning: Foundations and Concepts*. Springer International Publishing, Cham, 2024.

[31] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[32] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra. Weight Uncertainty in Neural Networks. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

[33] J. Y. Bo, S. Wan, and A. Anderson. To Rely or Not to Rely? Evaluating Interventions for Appropriate Reliance on Large Language Models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2025.

[34] J. Bona-Pellissier, F. Bachoc, and F. Malgouyres. Parameter identifiability of a deep feedforward ReLU neural network. *Machine Learning*, 112:4431–4493, 2023.

[35] J. Brea, B. Simsek, B. Illing, and W. Gerstner. Weight-space symmetry in deep networks gives rise to permutation saddles, connected by equal-loss valleys across the loss landscape, 2019.

[36] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[37] S. Bubeck and M. Sellke. A Universal Law of Robustness via Isoperimetry. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[38] S. Budd, E. C. Robinson, and B. Kainz. A Survey on Active Learning and Human-in-the-Loop Deep Learning for Medical Image Analysis. *Medical Image Analysis*, 71, 2021.

[39] M. Cervera, R. Dätwyler, F. D'Angelo, H. Keurti, B. F. Grewe, and C. Henning. Uncertainty estimation under model misspecification in neural network regression. In *Workshop Your Model Is Wrong: Robustness and Misspecification in Probabilistic Modeling Ath the 35th Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[40] B. Charpentier, D. Zügner, and S. Günnemann. Posterior Network: Uncertainty Estimation without OOD Samples via Density-Based Pseudo-Counts. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[41] A. M. Chen, H.-m. Lu, and R. Hecht-Nielsen. On the Geometry of Feedforward Neural Network Error Surfaces. *Neural Computation*, 5(6):910–927, 1993.

[42] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

[43] X. Chen and K. He. Exploring Simple Siamese Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[44] T. Cinquin and R. Bamler. Regularized KL-Divergence for Well-Defined Function-Space Variational Inference in Bayesian neural networks. In *Proceedings of the 41st Conference on Uncertainty in Artificial Intelligence (UAI)*, 2025.

[45] A. D. Cobb and B. Jalaian. Scaling Hamiltonian Monte Carlo Inference for Bayesian Neural Networks with Symmetric Splitting. In *Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2021.

[46] T. S. Cohen and M. Welling. Group Equivariant Convolutional Networks. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, 2016.

[47] P. G. Constantine, C. Kent, and T. Bui-Thanh. Accelerating MCMC with active subspaces. *SIAM Journal on Scientific Computing*, 38(5):2779–2805, 2016.

[48] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 2 edition, 2006.

[49] A. Curth, A. Jeffares, and M. van der Schaar. A U-turn on Double Descent: Rethinking Parameter Counting in Statistical Learning. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[50] F. Cuzzolin. *The Geometry of Uncertainty. The Geometry of Imprecise Probabilities*. Artificial Intelligence: Foundations, Theory, and Algorithms. Springer, 2021.

[51] F. D'Angelo and V. Fortuin. Repulsive Deep Ensembles are Bayesian. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[52] S. Dasgupta and D. Hsu. Hierarchical Sampling for Active Learning. In *Proceedings of the 25th International Conference on Machine Learning (ICML)*, 2008.

[53] E. Daxberger, A. Kristiadi, A. Immer, R. Eschenhagen, M. Bauer, and P. Hennig. Laplace Redux – Effortless Bayesian Deep Learning. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[54] S. Depeweg, J. M. Hernández-Lobato, F. Doshi-Velez, and S. Udluft. Decomposition of Uncertainty in Bayesian Deep Learning for Efficient and Risk-sensitive Learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

[55] S. K. Deshpande, S. Ghosh, T. D. Nguyen, and T. Broderick. Are you using test log-likelihood correctly? *Transactions on Machine Learning Research*, 2024.

[56] B. Dherin, M. Munn, M. Rosca, and D. G. T. Barrett. Why neural networks find simple solutions: The many regularizers of geometric complexity. In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[57] G. Dionne and S. E. Harrington. *Foundations of Insurance Economics*. Number 14 in Huebner International Series on Risk, Insurance and Economic Security. Springer Dordrecht, 1992.

[58] D. Dold, J. Kobialka, N. Palm, E. Sommer, D. Rügamer, and O. Dürr. Paths and Ambient Spaces in Neural Loss Landscapes. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2025.

[59] D. Draper. Assessment and Propagation of Model Uncertainty. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 57(1):45–70, 1995.

[60] F. Draxler, K. Veschgini, M. Salmhofer, and F. A. Hamprecht. Essentially No Barriers in Neural Network Energy Landscape. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

[61] R. Duan, B. Caffo, H. X. Bai, H. I. Sair, and C. Jones. Evidential Uncertainty Quantification: A Variance-Based Perspective. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.

[62] S. Duane, A. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222, 1987.

[63] D. Dubois, H. Prade, and P. Smets. Representing partial ignorance. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 26(3):361–377, 1996.

[64] B. Efron. Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation. *Journal of the American Statistical Association*, 78(382), 1983.

[65] R. Entezari, H. Sedghi, O. Saukh, and B. Neyshabur. The Role of Permutation Invariance in Linear Mode Connectivity of Neural Networks. In *Proceedings of the 10th International Conference on Learning Representations (ICLR)*, 2022.

[66] R. Eschenhagen, E. Daxberger, P. Hennig, and A. Kristiadi. Mixtures of Laplace Approximations for Improved Post-Hoc Uncertainty in Deep Learning. In *Bayesian Deep Learning Workshop at the 35th Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[67] V. Eyring, W. D. Collins, P. Gentine, E. A. Barnes, M. Barreiro, T. Beucler, M. Bocquet, C. S. Bretherton, H. M. Christensen, K. Dagon, D. J. Gagne, D. Hall, D. Hammerling, S. Hoyer, F. Iglesias-Suarez, I. Lopez-Gomez, M. C. McGraw, G. A. Meehl, M. J. Molina, C. Monteleoni, J. Mueller, M. S. Pritchard, D. Rolnick, J. Runge, P. Stier, O. Watt-Meyer, K. Weigel, R. Yu, and L. Zanna. Pushing the frontiers in climate modelling and analysis with machine learning. *Nature Climate Change*, 14(9):916–928, 2024.

[68] S. Farquhar. *Understanding Approximation for Bayesian Inference in Neural Networks*. PhD thesis, University of Oxford, 2022.

[69] S. Farquhar, Y. Gal, and T. Rainforth. On Statistical Bias in Active Learning. How and When to Fix it. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021.

[70] M. Fellaji and F. Pennerath. The Epistemic Uncertainty Hole: An issue of Bayesian Neural Networks, 2024.

[71] Y. Feng and Y. Tu. The inverse variance–flatness relation in stochastic gradient descent is critical for finding flat minima. *Proceedings of the National Academy of Sciences*, 118(9), 2021.

[72] D. Ferbach, B. Goujaud, G. Gidel, and A. Dieuleveut. Proving Linear Mode Connectivity of Neural Networks via Optimal Transport. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023.

[73] T. S. Filho, H. Song, M. Perello-Nieto, R. Santos-Rodriguez, M. Kull, and P. Flach. Classifier Calibration: A survey on how to assess and improve predicted class probabilities. *Machine Learning*, 112(9):3211–3260, 2023.

[74] K. Fisher and Y. Marzouk. Can Bayesian Neural Networks Make Confident Predictions? In *Mathematics of Modern Machine Learning Workshop at the 38th Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[75] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur. Sharpness-Aware Minimization for Efficiently Improving Generalization. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021.

[76] V. Fortuin. Priors in Bayesian Deep Learning: A Review. *International Statistical Review*, 90(3):563–591, 2022.

[77] V. Fortuin, A. Garriga-Alonso, S. W. Ober, F. Wenzel, G. Rätsch, R. E. Turner, M. van der Wilk, and L. Aitchison. Bayesian Neural Network Priors Revisited. In *Proceedings of the 10th International Conference on Learning Representations (ICLR)*, 2022.

[78] J. Frankle, G. K. Dziugaite, D. M. Roy, and M. Carbin. Linear Mode Connectivity and the Lottery Ticket Hypothesis. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

[79] C. D. Freeman and J. Bruna. Topology and Geometry of Half-Rectified Network Optimization. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, 2017.

[80] Y. Gal and Z. Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, 2016.

[81] Y. Gal, R. Islam, and Z. Ghahramani. Deep Bayesian Active Learning with Image Data. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

[82] T. Garipov, P. Izmailov, D. Podoprikhin, D. P. Vetrov, and A. G. Wilson. Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs. In *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

[83] J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, M. Shahzad, W. Yang, R. Bamler, and X. X. Zhu. A Survey of Uncertainty in Deep Neural Networks. *Artificial Intelligence Review*, 56:1513–1589, 2023.

[84] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. Shortcut Learning in Deep Neural Networks. *Nature Machine Intelligence*, 2 (11):665–673, 2020.

[85] Y. Gelberg, T. F. van der Ouderaa, M. van der Wilk, and Y. Gal. Variational Inference Failures Under Model Symmetries: Permutation Invariant Posteriors for Bayesian Neural Networks. In *Workshop on Geometry-grounded Representation Learning and Generative Modeling at the 41st International Conference on Machine Learning (ICML)*, 2024.

[86] A. Gelman and K. Shirley. *Inference from Simulations and Monitoring Convergence*, pages 162–174. CRC Press, New York, 1st edition, 2011.

[87] A. Gelman and Y. Yao. Holes in Bayesian Statistics. *Journal of Physics G: Nuclear and Particle Physics*, 48(1), 2021.

[88] A. Gelman, J. Hwang, and A. Vehtari. Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6):997–1016, 2014.

[89] A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin. *Bayesian Data Analysis*. CRC Press, 2021.

[90] Z. Ghahramani. Should all Machine Learning be Bayesian? Should all Bayesian models be non-parametric?, 2008.

[91] T. Gneiting and A. E. Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[92] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and Harnessing Adversarial Examples. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.

[93] L. Grenioux, A. O. Durmus, É. Moulines, and M. Gabrié. On Sampling with Approximate Transport Maps. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

[94] C. Gruber, P. O. Schenk, M. Schierholz, F. Kreuter, and G. Kauermann. Sources of Uncertainty in Supervised Machine Learning – A Statisticians' View (forthcoming publication), 2025.

[95] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

[96] D. Hansen, S. Devic, P. Nakkiran, and V. Sharan. When is Multicalibration Post-Processing Necessary? In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[97] M. Havasi, R. Jenatton, S. Fort, J. Z. Liu, J. Snoek, B. Lakshminarayanan, A. M. Dai, and D. Tran. Training independent subnetworks for robust prediction. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021.

[98] H. He, G. Huang, and Y. Yuan. Asymmetric Valleys: Beyond Sharp and Flat Local Minima. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[99] R. Hecht-Nielsen. On the Algebraic Structure of Feedforward Network Weight Spaces. In *Advanced Neural Computers*. Elsevier, 1990.

[100] D. Hendrycks, M. Mazeika, S. Kadavath, and D. Song. Using Self-Supervised Learning Can Improve Model Robustness and Uncertainty. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[101] G. E. Hinton and R. Salakhutdinov. Reducing the Dimensionality of Data with Neural Networks. *Science*, 313(5786):504–507, 2006.

[102] A. Hoarau, S. Destercke, Y. Sale, P. Hofman, and E. Hullermeier. Measures of Uncertainty: A Quantitative Analysis, 2025.

[103] L. Hodgkinson, C. van der Heide, R. Salomone, F. Roosta, and M. W. Mahoney. The Interpolating Information Criterion for Overparameterized Models, 2023.

[104] M. D. Hoffman and A. Gelman. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15:1351–1381, 2014.

[105] S. Hoffmann, F. Schönbrodt, R. Elsas, R. Wilson, U. Strasser, and A.-L. Boulesteix. The multiplicity of analysis strategies jeopardizes replicability: Lessons learned across disciplines. *Royal Society Open Science*, 8(4), 2021.

[106] P. Hofman, Y. Sale, and E. Hüllermeier. Quantifying Aleatoric and Epistemic Uncertainty: A Credal Approach. In *Workshop on Structured Probabilistic Inference & Generative Modeling Workshop at the 41st International Conference on Machine Learning (ICML)*, 2024.

[107] N. Hollmann, S. Müller, K. Eggensperger, and F. Hutter. TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, 2023.

[108] N. Houlsby, F. Huszár, Z. Ghahramani, and M. Lengyel. Bayesian Active Learning for Classification and Preference Learning, 2011.

[109] Z. Huang, H. Lam, and H. Zhang. Efficient Uncertainty Quantification and Reduction for Over-Parameterized Neural Networks. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[110] M. Huh, H. Mobahi, R. Zhang, B. Cheung, P. Agrawal, and P. Isola. The Low-Rank Simplicity Bias in Deep Networks. *Transactions on Machine Learning Research*, 3, 2023.

[111] E. Hüllermeier and W. Waegeman. Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods. *Machine Learning*, 110:457–506, 2021.

[112] A. Ito, M. Yamada, and A. Kumagai. Linear Mode Connectivity between Multiple Models modulo Permutation Symmetries. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

[113] N. Iyer, V. Thejas, N. Kwatra, R. Ramjee, and M. Sivathanu. Wide-minima Density Hypothesis and the Explore-Exploit Learning Rate Schedule. *Journal of Machine Learning Research*, 24, 2023.

[114] P. Izmailov, W. J. Maddox, P. Kirichenko, T. Garipov, D. Vetrov, and A. G. Wilson. Subspace Inference for Bayesian Deep Learning. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2020.

[115] P. Izmailov, P. Nicholson, S. Lotfi, and A. G. Wilson. Dangers of Bayesian Model Averaging under Covariate Shift. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[116] P. Izmailov, S. Vikram, M. D. Hoffman, and A. G. Wilson. What Are Bayesian Neural Network Posteriors Really Like? In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

[117] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon. A Survey on Contrastive Self-Supervised Learning. *Technologies*, 9(2), 2020.

[118] S. Jiménez, M. Jürgens, and W. Waegeman. Why Machine Learning Models Fail to Fully Capture Epistemic Uncertainty, 2025.

[119] D. R. Jones and M. Schonlau. Efficient Global Optimization of Expensive Black-Box Functions. *Journal of Global Optimization*, 13:455–492, 1998.

[120] K. Jordan, H. Sedghi, O. Saukh, R. Entezari, and B. Neyshabur. REPAIR: Renormalizing Permuted Activations for Interpolation Repair. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, 2023.

[121] M. Kängsepp, K. Valk, and M. Kull. On the Usefulness of the Fit-on-the-Test View on Evaluating Calibration of Classifiers. *Machine Learning*, 114, 2025.

[122] S. Kapoor, W. J. Maddox, P. Izmailov, and A. G. Wilson. On Uncertainty, Tempering, and Data Augmentation in Bayesian Classification. In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[123] Y. Kato, D. M. Tax, and M. Loog. A View on Model Misspecification in Uncertainty Quantification. In *Proceedings of BNAIC/BeNeLearn*, 2022.

[124] A. Kendall and Y. Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS)*, 2017.

[125] L. Klarner, T. G. J. Rudner, M. Reutlinger, T. Schindler, G. M. Morris, C. M. Deane, and Y. W. Teh. Drug Discovery under Covariate Shift withDomain-Informed Prior Distributions over Functions. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

[126] J. Knoblauch, J. Jewson, and T. Damoulas. An Optimization-centric View on Bayes' Rule: Reviewing and Generalizing Variational Inference. *Journal of Machine Learning Research*, 23, 2022.

[127] J. Kobialka, E. Sommer, J. Kwon, D. Dold, and D. Rügamer. Approximate Posteriors in Neural Networks: A Sampling Perspective. In *7th Symposium on Advances in Approximate Bayesian Inference (AABI)*, 2025.

[128] C. Kolb, T. Weber, B. Bischl, and D. Rügamer. Deep Weight Factorization: Sparse Learning Through the Lens of Artificial Symmetries. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*, 2025.

[129] B. Kompa, J. Snoek, and A. L. Beam. Empirical Frequentist Coverage of Deep Learning Uncertainty Quantification Procedures. *Entropy*, 23(12), 2021.

[130] A.-K. Kopetzki, B. Charpentier, D. Zügner, S. Giri, and S. Günnemann. Evaluating Robustness of Predictive Uncertainty Estimation: Are Dirichlet-based Models Reliable? In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

[131] N. Kotelevskii and M. Panov. Predictive Uncertainty Quantification via Risk Decompositions for Strictly Proper Scoring Rules, 2024.

[132] A. Kristiadi, M. Hein, and P. Hennig. Being Bayesian, Even Just a Bit, Fixes Overconfidence in ReLU Networks. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

[133] A. Kucukelbir, D. Tran, R. Ranganath, A. Gelman, and D. M. Blei. Automatic Differentiation Variational Inference. *Journal of Machine Learning Research*, 18, 2017.

[134] R. Kuditipudi, X. Wang, H. Lee, Y. Zhang, Z. Li, W. Hu, S. Arora, and R. Ge. Explaining Landscape Connectivity of Low-cost Solutions for Multilayer Nets. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[135] S. Kulinski and D. I. Inouye. Towards Explaining Distribution Shifts. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

[136] M. Kull and P. Flach. Patterns of dataset shift. In *Workshop on Learning over Multiple Contexts at the Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD)*, 2014.

[137] A. Kumar, P. Liang, and T. Ma. Verified Uncertainty Calibration. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[138] D. Kunin, J. Sagastuy-Brena, S. Ganguli, D. L. K. Yamins, and H. Tanaka. Neural Mechanics: Symmetry and Broken Conservation Laws in Deep Learning Dynamics. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021.

[139] V. Kůrková and P. C. Kainen. Functionally Equivalent Feedforward Neural Networks. *Neural Computation*, 6(3):543–558, 1994.

[140] S. M. Kwon, Z. Zhang, D. Song, L. Balzano, and Q. Qu. Efficient Low-Dimensional Compression of Overparameterized Models. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2024.

[141] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS)*, 2017.

[142] O. Laurent, E. Aldea, and G. Franchi. A Symmetry-Aware Exploration of Bayesian Neural Network Posteriors. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024.

[143] D.-H. Lee. Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks. In *Workshop on Challenges in Representation Learning at the 30th International Conference on Machine Learning (ICML)*, 2013.

[144] N. Lehmann, N. M. Gottschling, J. Gawlikowski, A. J. Stewart, S. Depeweg, and E. Nalisnick. Lightning UQ Box: Uncertainty Quantification for Neural Networks. *Journal of Machine Learning Research*, 26, 2025.

[145] C. T. Lewis and C. Short. *A Latin Dictionary.* Oxford: Clarendon Press, 1879.

[146] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein. Visualizing the Loss Landscape of Neural Nets. In *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

[147] J. Li, M. Zichen, Q. Qiu, and R. Zhang. Training Bayesian Neural Networks with Sparse Subspace Variational Inference. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024.

[148] Y. L. Li, D. Lu, P. Kirichenko, S. Qiu, T. G. J. Rudner, C. B. Bruss, and A. G. Wilson. Position: Supervised Classifiers Answer the Wrong Questions for OOD Detection. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

[149] D. Lim. The Empirical Impact of Neural Parameter Symmetries, or Lack Thereof. In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[150] G. Loaiza-Ganem, V. Villecroze, and Y. Wang. Deep Ensembles Secretly Perform Empirical Bayes, 2025.

[151] E. Lobacheva, N. Chirkova, M. Kodryan, and D. Vetrov. On Power Laws in Deep Ensembles. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[152] S. Lotfi, M. Finzi, S. Kapoor, A. Potapczynski, M. Goldblum, and A. G. Wilson. PAC-Bayes Compression Bounds So Tight That They Can Explain Generalization. In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[153] D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms.* Cambridge University Press, 2003.

[154] W. J. Maddox, P. Izmailov, T. Garipov, D. P. Vetrov, and A. G. Wilson. A Simple Baseline for Bayesian Uncertainty in Deep Learning. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[155] A. Malinin and M. Gales. Predictive Uncertainty Estimation via Prior Networks. In *Proceedings of the 32nd Conference of Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[156] S. K. Manchingal, M. Mubashar, K. Wang, and F. Cuzzolin. A Unified Evaluation Framework for Epistemic Predictions. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2025.

[157] C. C. Margossian, M. D. Hoffman, P. Sountsov, L. Riou-Durand, A. Vehtari, and A. Gelman. Nested $\hat{R}$: Assessing the convergence of Markov chain Monte Carlo when running many short chains, 2022.

[158] C. H. Martin and M. W. Mahoney. Implicit Self-Regularization in Deep Neural Networks: Evidence from Random Matrix Theory and Implications for Learning. *Journal of Machine Learning Research*, 22:1–73, 2021.

[159] G. M. Martin, D. T. Frazier, and C. P. Robert. Computing Bayes: From Then 'Til Now. *Statistical Science*, 39(1), 2024.

[160] F. Martinelli, A. V. Meegen, B. Şimşek, W. Gerstner, and J. Brea. Flat Channels to Infinity in Neural Loss Landscapes, 2025.

[161] A. R. Masegosa. Learning under Model Misspecification: Applications to Variational and Ensemble methods. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[162] Q. Meng, S. Zheng, H. Zhang, W. Chen, Z.-M. Ma, and T.-Y. Liu. $\mathcal{G}$-SGD: Optimizing ReLU Neural Networks in its Positively Scale-Invariant Space. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, 2019.

[163] M. Miani, H. Roy, and S. Hauberg. Bayes without Underfitting: Fully Correlated Deep Learning Posteriors via Alternating Projections. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2025.

[164] M. Minderer, J. Djolonga, R. Romijnders, F. Hubis, X. Zhai, N. Houlsby, D. Tran, and M. Lucic. Revisiting the Calibration of Modern Neural Networks. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[165] T. P. Minka. Bayesian model averaging is not model combination, 2002.

[166] B. Mlodozeniec, R. E. Turner, and D. Krueger. Implicitly Bayesian Prediction Rules in Deep Learning. In *Proceedings of the 6th Symposium on Advances in Approximate Bayesian Inference*, volume 253, pages 79–110. PMLR, 2024.

[167] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera. A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1):521–530, 2012.

[168] B. Mucsányi, M. Kirchhof, E. Nguyen, A. Rubinstein, and S. J. Oh. *Trustworthy Machine Learning. Theory, Applications, Intuitions.* arXiv, 2023.

[169] B. Mucsányi, M. Kirchhof, and S. J. Oh. Benchmarking Uncertainty Disentanglement: Specialized Uncertainties for Specialized Tasks. In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[170] J. Mukhoti, V. Kulharia, A. Sanyal, S. Golodetz, P. H. S. Torr, and P. K. Dokania. Calibrating Deep Neural Networks using Focal Loss. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[171] K. P. Murphy. *Machine Learning: A Probabilistic Perspective.* Adaptive Computation and Machine Learning Series. MIT Press, Cambridge, 2012.

[172] K. P. Murphy. *Probabilistic Machine Learning: An Introduction.* Adaptive Computation and Machine Learning Series. MIT Press, Cambridge, 2022.

[173] C. Murray, J. U. Allingham, J. Antorán, and J. M. Hernández-Lobato. Addressing Bias in Active Learning with Depth Uncertainty Networks... or Not. In *I (Still) Can't Believe It's Not Better Workshop at the 35th Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[174] M. P. Naeini, G. Cooper, and M. Hauskrecht. Obtaining Well Calibrated Probabilities Using Bayesian Binning. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, 2015.

[175] T. Nagler and D. Rügamer. Uncertainty Quantification for Prior-Data Fitted Networks using Martingale Posteriors. In *Workshop on Frontiers in Probabilistic Inference: Learning Meets Sampling at the 13th International Conference on Learning Representations (ICLR)*, 2025.

[176] V. Nair and G. E. Hinton. Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, 2010.

[177] E. T. Nalisnick. *On Priors for Bayesian Neural Networks*. PhD thesis, University of California, Irvine, 2018.

[178] G. Nam, J. Yoon, Y. Lee, and J. Lee. Diversity Matters When Learning From Ensembles. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[179] A. Navon, A. Shamsian, I. Achituve, E. Fetaya, G. Chechik, and H. Maron. Equivariant Architectures for Learning in Deep Weight Spaces. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

[180] R. M. Neal. MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. Jones, and X.-L. Meng, editors, *Handbook of Markov Chain Monte Carlo*. CRC Press, 2011.

[181] B. Neyshabur, R. Salakhutdinov, and N. Srebro. Path-SGD: Path-Normalized Optimization in Deep Neural Networks. In *Proceedings of the 29th Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2015.

[182] V.-L. Nguyen, M. H. Shaker, and E. Hüllermeier. How to measure uncertainty in uncertainty sampling for active learning. *Machine Learning*, 111:89–122, 2022.

[183] M. Nickel. Epistemic limits of passive data collection in complex social systems. In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[184] D. Nix and A. Weigend. Estimating the Mean and Variance of the Target Probability Distribution. In *Proceedings of the 1994 IEEE International Conference on Neural Networks (ICNN)*, 1994.

[185] J. Nixon, M. Dusenberry, G. Jerfel, T. Nguyen, J. Liu, L. Zhang, and D. Tran. Measuring Calibration in Deep Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019.

[186] L. Noci, K. Roth, G. Bachmann, S. Nowozin, and T. Hofmann. Disentangling the Roles of Curation, Data-Augmentation and the Prior in the Cold Posterior Effect. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[187] S. J. Oh, K. Murphy, J. Pan, J. Roth, F. Schroff, and A. Gallagher. Modeling Uncertainty with Hedged Instance Embedding. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, 2019.

[188] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. V. Dillon, B. Lakshminarayanan, and J. Snoek. Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[189] T. Papamarkou, J. Hinkle, M. T. Young, and D. Womble. Challenges in Markov Chain Monte Carlo for Bayesian Neural Networks. *Statistical Science*, 37(3), 2022.

[190] T. Papamarkou, M. Skoularidou, K. Palla, L. Aitchison, J. Arbel, D. Dunson, M. Filippone, V. Fortuin, P. Hennig, J. M. Hernandez-Lobato, A. Hubin, A. Immer, T. Karaletsos, M. E. Khan, A. Kristiadi, Y. Li, S. Mandt, C. Nemeth, M. A. Osborne, T. G. J. Rudner, and R. Zhang. Position: Bayesian Deep Learning is Needed in the Age of Large-Scale AI. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

[191] P. Patil, J.-H. Du, and R. J. Tibshirani. Revisiting Optimism and Model Complexity in the Wake of Overparameterized Machine Learning, 2024.

[192] A. Peleg and M. Hein. Bias of Stochastic Gradient Descent or the Architecture: Disentangling the Effects of Overparameterization of Neural Networks. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

[193] M. Phuong and C. H. Lampert. Functional vs. Parametric Equivalence of ReLU Networks. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, 2020.

[194] J. Piironen and A. Vehtari. Comparison of Bayesian predictive methods for model selection. *Statistics and Computing*, 27(3):711–735, 2017.

[195] R. Pinsler, J. Gordon, E. Nalisnick, and J. M. Hernández-Lobato. Bayesian Batch Active Learning as Sparse Subset Approximation. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[196] F. Pittorino, A. Ferraro, G. Perugini, C. Feinauer, C. Baldassi, and R. Zecchina. Deep networks on toroids: Removing symmetries reveals the structure of flat regions in the landscape geometry. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022.

[197] G. Pituk, V. Shirvaikar, and T. Rainforth. Do Bayesian Neural Networks Actually Behave Like Bayesian Models? In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

[198] A. A. Pourzanjani, R. M. Jiang, and L. R. Petzold. Improving the Identifiability of Neural Networks for Bayesian Inference. In *Workshop on Bayesian Deep Learning at the 31st Conference on Neural Information Processing Systems (NIPS)*, 2017.

[199] A. F. Psaros, X. Meng, Z. Zou, L. Guo, and G. E. Karniadakis. Uncertainty Quantification in Scientific Machine Learning: Methods, Metrics, and Comparisons. *Journal of Computational Physics*, 477(C), 2023.

[200] A. Rame and M. Cord. DICE: Diversity in Deep Ensembles via Conditional Redundancy Adversarial Estimation. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021.

[201] L. Riou-Durand, P. Sountsov, J. Vogrinc, and C. C. Margossian. Adaptive Tuning for Metropolis Adjusted Langevin Trajectories. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023.

[202] C. P. Robert. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer, 2nd edition, 2007.

[203] J. Rodemann, J. Goschenhofer, E. Dorigatti, T. Nagler, and T. Augustin. Approximately Bayes-Optimal Pseudo-Label Selection. In *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2023.

[204] R. Roelofs, N. Cain, J. Shlens, and M. C. Mozer. Mitigating Bias in Calibration Error Estimation. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

[205] D. Rolnick and K. P. Körding. Reverse-Engineering Deep ReLU Networks. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

[206] H. Roy, M. Miani, C. H. Ek, P. Hennig, M. Pförtner, L. Tatzel, and S. Hauberg. Reparameterization invariance in approximate Bayesian inference. In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[207] T. G. J. Rudner, S. Kapoor, S. Qiu, and A. G. Wilson. Function-Space Regularization in Neural Networks: A Probabilistic Perspective. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

[208] T. G. J. Rudner, Y. S. Zhang, A. G. Wilson, and J. Kempe. Mind the GAP: Improving Robustness to Subpopulation Shifts with Group-Aware Priors. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2024.

[209] D. Rundel, E. Sommer, and B. Bischl. Efficiently Warmstarting MCMC for BNNs. In *Workshop on Frontiers in Probabilistic Inference: Learning Meets Sampling at the 13th International Conference on Learning Representations (ICLR)*, 2025.

[210] Y. Sale, P. Hofman, T. Löhr, L. Wimmer, T. Nagler, and E. Hüllermeier. Label-wise Aleatoric and Epistemic Uncertainty Quantification. In *Proceedings of the 40th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2024.

[211] R. E. Schapire. The strength of weak learnability. *Machine Learning*, 5(2):197–227, 1990.

[212] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. Towards Causal Representation Learning. *Proceedings of the IEEE*, 109(5), 2021.

[213] K. Schürholt. *Hyper-Representations: Learning from Populations of Neural Networks*. PhD thesis, University of St. Gallen, 2024.

[214] K. Schweighofer, L. Aichberger, M. Ielanskyi, and S. Hochreiter. On Information-Theoretic Measures of Predictive Uncertainty. In *Proceedings of the 41st Conference on Uncertainty in Artificial Intelligence (UAI)*, 2025.

[215] T. Seidenfeld. Entropy and Uncertainty. *Philosophy of Science*, 53(4):467–491, 1986.

[216] R. Senge, S. Bösner, K. Dembczyński, J. Haasenritter, O. Hirsch, N. Donner-Banzhoff, and E. Hüllermeier. Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty. *Information Sciences*, 255:16–29, 2014.

[217] B. Settles. Active Learning Literature Survey. Technical Report 1648, University of Wisconsin–Madison, 2010.

[218] C. E. Shannon. A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27:379–423, 1948.

[219] E. Sharma, D. Kwok, T. Denton, D. M. Roy, D. Rolnick, and G. K. Dziugaite. Simultaneous linear connectivity of neural networks modulo permutation. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD)*. Springer International Publishing, 2024.

[220] M. Sharma, S. Farquhar, E. Nalisnick, and T. Rainforth. Do Bayesian Neural Networks Need To Be Fully Stochastic? In *Proceedings of the 26th International Conference on Artificial Intel Ligence and Statistics (AISTATS)*, 2023.

[221] Y. Shen, N. Daheim, B. Cong, P. Nickl, G. M. Marconi, C. Bazan, R. Yokota, I. Gurevych, D. Cremers, M. E. Khan, and T. Möllenhoff. Variational Learning is Effective for Large Deep Networks. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

[222] M. Shoja and E. S. Soofi. Uncertainty, information, and disagreement of economic forecasters. *Econometric Reviews*, 36(6-9):796–817, 2017.

[223] R. Shwartz-Ziv, M. Goldblum, H. Souri, S. Kapoor, C. Zhu, Y. LeCun, and A. G. Wilson. Pre-Train Your Loss: Easy Bayesian Transfer Learning with Informative Priors. In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[224] J. B. Simon, D. Karkada, N. Ghosh, and M. Belkin. More is Better in Modern Machine Learning: When Infinite Overparameterization is Optimal and Overfitting is Obligatory. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024.

[225] B. Simsek, F. Ged, A. Jacot, F. Spadaro, C. Hongler, W. Gerstner, and J. Brea. Geometry of the Loss Landscape in Overparameterized Neural Networks: Symmetries and Invariances. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

[226] L. Sluijterman, E. Cator, and T. Heskes. How to evaluate uncertainty estimates in machine learning for regression? *Neural Networks*, 173(11), 2024.

[227] E. Sommer, J. Robnik, G. Nozadze, U. Seljak, and D. Rügamer. Microcanonical Langevin Ensembles: Advancing the Sampling of Bayesian Neural Networks. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2025.

[228] H. Song, T. Diethe, M. Kull, and P. Flach. Distribution Calibration for Regression. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

[229] J. S. Speagle. A Conceptual Introduction to Markov Chain Monte Carlo Methods, 2020.

[230] D. J. Spiegelhalter. *The Art of Uncertainty: How to Navigate Chance, Ignorance, Risk and Luck*. Random House, 2024.

[231] S. Stanton, W. Maddox, and A. G. Wilson. Bayesian Optimization with Conformal Coverage Guarantees. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023.

[232] D. Steinmann, F. Divo, M. Kraus, A. Wüst, L. Struppek, F. Friedrich, and K. Kersting. Navigating Shortcuts, Spurious Correlations, and Confounders: From Origins via Detection to Mitigation, 2024.

[233] M. Steup and N. Ram. Epistemology. In E. N. Zalta and U. Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Stanford University, 2024.

[234] A. C. Stickland and I. Murray. Diverse Ensembles Improve Calibration. In *Workshop on Uncertainty and Robustness in Deep Learning at the 37th International Conference on Machine Learning (ICML)*, 2020.

[235] P. Stock, B. Graham, R. Gribonval, and H. Jégou. Equi-normalization of Neural Networks. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, 2019.

[236] E. Štrumbelj, A. Bouchard-Côté, J. Corander, A. Gelman, H. Rue, L. Murray, H. Pesonen, M. Plummer, and A. Vehtari. Past, Present and Future of Software for Bayesian Inference. *Statistical Science*, 39(1), 2024.

[237] S. Sun, G. Zhang, J. Shi, and R. Grosse. Functional Variational Bayesian Neural Networks. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, 2019.

[238] H. J. Sussmann. Uniqueness of the weights for minimal feedforward nets with a given input-output map. *Neural Networks*, 5(4):589–593, 1992.

[239] V. Tamames-Rodero, A. Moya, L. M. Sarro, and R. J. López-Sastre. Unveiling the power of uncertainty: A journey into Bayesian Neural Networks for stellar dating. *Astronomy and Computing*, 52, 2025.

[240] G. Tata, G. K. Gudur, G. Chennupati, and M. E. Khan. Can Calibration Improve Sample Prioritization? In *Has It Trained Yet? Workshop at the 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[241] A. Theus, A. Cabodi, S. Anagnostidis, A. Orvieto, S. P. Singh, and V. Boeva. Generalized Linear Mode Connectivity for Transformers, 2025.

[242] Y. Tian, Z. Al-Ars, M. Kitsak, and P. Hofstee. Low-Loss Space in Neural Networks is Continuous and Fully Connected, 2025.

[243] B.-H. Tran, S. Rossi, D. Milios, and M. Filippone. All You Need is a Good Functional Prior for Bayesian Deep Learning. *Journal of Machine Learning Research*, 23, 2022.

[244] D. Tran, J. Liu, M. W. Dusenberry, D. Phan, M. Collier, J. Ren, K. Han, Z. Wang, Z. Mariet, H. Hu, N. Band, T. G. J. Rudner, K. Singhal, Z. Nado, J. van Amersfoort, A. Kirsch, R. Jenatton, N. Thain, H. Yuan, K. Buchanan, K. Murphy, D. Sculley, Y. Gal, Z. Ghahramani, J. Snoek, and B. Lakshminarayanan. Plex: Towards Reliability using Pretrained Large Model Extensions. In *Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward at the 38th International Conference on Machine Learning (ICML)*, 2022.

[245] M. O. Turkoglu, A. Becker, H. A. Gündüz, M. Rezaei, B. Bischl, R. C. Daudt, S. D'Aronco, J. D. Wegner, and K. Schindler. FiLM-Ensemble: Probabilistic Deep Learning via Feature-wise Linear Modulation. In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[246] J. Vaicenavicius, D. Widmann, C. Andersson, and F. Lindsten. Evaluating model calibration in classification. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

[247] J. van Amersfoort, L. Smith, Y. W. Teh, and Y. Gal. Uncertainty Estimation Using a Single Deep Deterministic Neural Network. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

[248] A. M. Van Der Bles, S. Van Der Linden, A. L. J. Freeman, J. Mitchell, A. B. Galvao, L. Zaval, and D. J. Spiegelhalter. Communicating uncertainty about facts, numbers and science. *Royal Society Open Science*, 6(5):181870, 2019.

[249] A. W. Van Der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1st edition, 1998.

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

[251] S. Villar, D. W. Hogg, W. Yao, G. A. Kevrekidis, and B. Schölkopf. Towards fully covariant machine learning, 2023.

[252] L. Vilnis and A. McCallum. Word Representations via Gaussian Embedding. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.

[253] V. Vlačić and H. Bölcskei. Affine symmetries and neural network identifiability. *Advances in Mathematics*, 376:107485, 2021.

[254] M. J. Wainwright and M. I. Jordan. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends in Machine Learning*, 1:1–305, 2008.

[255] Y. Wald, A. Feder, D. Greenfeld, and U. Shalit. On Calibration and Out-of-domain Generalization. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[256] Y. Wang and X. Wang. Agree to Disagree: Demystifying Homogeneous Deep Ensembles through Distributional Equivalence. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*, 2025.

[257] D. S. Watson and J. O'Hara. Explaining Predictive Uncertainty with Information Theoretic Shapley Values. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[258] F. Wenzel, K. Roth, B. S. Veeling, J. Światkowski, L. Tran, S. Mandt, J. Snoek, T. Salimans, R. Jenatton, and S. Nowozin. How Good is the Bayes Posterior in Deep Neural Networks Really? In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

[259] D. Widmann, F. Lindsten, and D. Zachariah. Calibration tests in multi-class classification: A unifying framework. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[260] V. D. Wild, S. Ghalebikesabi, D. Sejdinovic, and J. Knoblauch. A Rigorous Link between Deep Ensembles and (Variational) Bayesian Methods. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[261] A. G. Wilson. The Case for Bayesian Deep Learning, 2020.

[262] A. G. Wilson. Deep Learning is Not So Mysterious or Different, 2025.

[263] A. G. Wilson and P. Izmailov. Bayesian Deep Learning and a Probabilistic Perspective of Generalization. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[264] A. G. Wilson, S. Lotfi, S. Vikram, M. D. Hoffman, Y. Gal, Y. Li, M. F. Pradier, A. Foong, S. Farquhar, and P. Izmailov. Evaluating Approximate Inference in Bayesian Deep Learning. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[265] K. Xia, K.-Z. Lee, Y. Bengio, and E. Bareinboim. The Causal-Neural Connection: Expressiveness, Learnability, and Inference. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[266] Z. Xiao, J. Shen, X. Zhen, L. Shao, and C. G. M. Snoek. A Bit More Bayesian: Domain-Invariant Learning with Uncertainty. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

[267] Y. A. Yadkori, I. Kuzborskij, A. György, and C. Szepesvári. To Believe or Not to Believe Your LLM: Iterative Prompting for Estimating Epistemic Uncertainty. In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[268] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny. Barlow Twins: Self-Supervised Learning via Redundancy Reduction. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

[269] C. Zhang, S. Bengio, and Y. Singer. Are All Layers Created Equal? *Journal of Machine Learning Research*, 23:1–28, 2022.

[270] D. W. Zhang, M. Kofinas, Y. Zhang, Y. Chen, G. J. Burghouts, and C. G. M. Snoek. Neural Networks Are Graphs! Graph Neural Networks for Equivariant Processing of Neural Networks. In *Workshop on Topology, Algebra, and Geometry in Machine Learning (TAG-ML) at the 40th International Conference on Machine Learning (ICML)*, 2023.

[271] X. Zhang, Y. Li, W. Li, K. Guo, and Y. Shao. Personalized Federated Learning via Variational Bayesian Inference. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022.

[272] Y. Zhang, Y.-S. Wu, L. A. Ortega, and A. R. Masegosa. The Cold Posterior Effect Indicates Underfitting, and Cold Posteriors Represent a Fully Bayesian Method to Mitigate It. *Transactions on Machine Learning Research*, 8, 2024.

[273] B. Zhao, I. Ganev, R. Walters, R. Yu, and N. Dehmamy. Symmetries, flat minima, and the conserved quantities of gradient flow. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, 2023.

[274] B. Zhao, R. M. Gower, R. Walters, and R. Yu. Improving Convergence and Generalization Using Parameter Symmetries. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024.

[275] B. Zhao, R. Walters, and R. Yu. Symmetry in Neural Network Parameter Spaces, 2025.

[276] A. Zhou and S. Levine. Training on Test Data with Bayesian Adaptation for Covariate Shift. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[277] A. Zhou, K. Yang, K. Burns, Y. Jiang, S. Sokota, J. Z. Kolter, and C. Finn. Permutation Equivariant Neural Functionals. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[278] L. Ziyin. Symmetry Induces Structure and Constraint of Learning. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

[279] L. Ziyin, M. Wang, and L. Wu. The Implicit Bias of Gradient Noise: A Symmetry Perspective, 2024.

[280] L. Ziyin, Y. Xu, and I. Chuang. Remove Symmetries to Control Model Expressivity and Improve Optimization. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*, 2025.

Eidesstattliche Versicherung

(Siehe Promotionsordnung vom 12. Juli 2011, § 8 Abs. 2 Pkt. 5)

Hiermit erkläre ich an Eides statt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

Während des Erstellungsprozesses habe ich Grammarly verwendet, um Rechtschreibung und grammatikalische Fehler zu korrigieren. Für das Verfassen von Code habe ich GitHub Copilot zur Vervollständigung häufig auftretender Komponenten genutzt, wobei Vorschläge stets sorgfältig überprüft und nur mit großer Vorsicht in überarbeiteter Fassung übernommen wurden. Die Konzeption und Struktur von Programmcode habe ich vollkommen eigenständig verantwortet. Außerdem wurde im einleitenden Teil dieser Arbeit ChatGPT (GPT-4.5) als Hilfsmittel verwendet, um an einzelnen Stellen die von mir erdachten Beispiele und Erklärungen zu überprüfen. Sämtlicher Text, soweit nicht als wörtliches Zitat gekennzeichnet, wurde von mir verfasst. Insbesondere habe ich die Gliederung, Literaturrecherche und wissenschaftlichen Beiträge dieser Arbeit, im Rahmen der in Teil II aufgeführten Verantwortlichkeiten, eigenständig und ohne fremde Hilfe ausgefertigt.

München, den 02.08.2025 Lisa Wimmer