Rethinking Aleatoric and Epistemic Uncertainty

Freddie Bickford Smith [1] **Jannik Kossen** [1] **Eleanor Trollope** [1]
Mark van der Wilk [1] **Adam Foster** [1] **Tom Rainforth** [1]

Abstract

The ideas of aleatoric and epistemic uncertainty are widely used to reason about the probabilistic predictions of machine-learning models. We identify incoherence in existing discussions of these ideas and suggest this stems from the aleatoric-epistemic view being insufficiently expressive to capture all the distinct quantities that researchers are interested in. To address this we present a decision-theoretic perspective that relates rigorous notions of uncertainty, predictive performance and statistical dispersion in data. This serves to support clearer thinking as the field moves forward. Additionally we provide insights into popular information-theoretic quantities, showing they can be poor estimators of what they are often purported to measure, while also explaining how they can still be useful in guiding data acquisition.

1. Introduction

When making decisions under uncertainty, it can be useful to reason about where that uncertainty comes from (Osband et al, 2023; Wen et al, 2022). Researchers commonly refer to the ideas of aleatoric (literal meaning: "relating to chance") and epistemic ("relating to knowledge") uncertainty, which have a long history in the study of probability (Hacking, 1975). Aleatoric uncertainty is typically associated with statistical dispersion in data (sometimes thought of as noise), while epistemic is associated with the internal information state of a model (Hüllermeier & Waegeman, 2021).

Concerningly given their scale of use, these ideas are not being discussed coherently in the literature. The line between model-based predictions and data-generating processes is repeatedly blurred (Amini et al, 2020; Ayhan & Berens, 2018; Immer et al, 2021; Kapoor et al, 2022; Smith & Gal, 2018; van Amersfoort et al, 2020). On top of this, tenuous assumptions are made about how uncertainty will decompose on unseen data (Seeböck et al, 2019; Wang & Aitchison, 2021), and misleading connections are drawn between uncertainty and predictive accuracy (Orlando et al, 2019; Wang et al, 2019). Meanwhile distinct mathematical quantities are used to refer to notionally the same concepts: epistemic uncertainty, for example, has been variously defined using density-based (Mukhoti et al, 2023; Postels et al, 2020), information-based (Gal et al, 2017) and variance-based (Gal, 2016; Kendall & Gal, 2017; McAllister, 2016) quantities.

We suggest this incoherence arises from the aleatoric-epistemic view being too simplistic in the context of machine learning. Researchers are looking for concrete notions of a model's predictive uncertainty and how that uncertainty might or might not change with more data (associated with a decomposition into irreducible and reducible components), but also related notions of predictive performance and data dispersion. The aleatoric-epistemic view cannot satisfy all these needs: many concepts stand to be defined, while the view fundamentally only has capacity for two concepts. Yet the current state of play is to nevertheless appeal to the aleatoric-epistemic view, with different researchers using it in different ways. A result of this conceptual overloading is to conflate quantities that ought to be recognised as distinct. Far from just a matter of semantics, this is having a meaningful effect on the field's progress: methods are being designed and evaluated based on shaky foundations.

To establish a clearer perspective, we draw on powerful yet underappreciated ideas from decision theory (Dawid, 1998; DeGroot, 1962; Neiswanger et al, 2022). Our starting point is a final decision of interest with an associated loss function. Given this, uncertainty in predictive beliefs can be formalised as the subjective expected loss of acting Bayes-optimally under those beliefs; this generalises quantities like variance and entropy. From there we show how reasoning about new data gives rise to a notion of expected uncertainty reduction, which we can use to identify a decomposition of uncertainty into irreducible and reducible components. Then we clarify the connection between uncertainty, predictive performance and data dispersion, linking to classic decompositions from statistics and information theory. Overall this provides a coherent synthesis of key quantities that researchers are interested in (Figure 1).

[1]University of Oxford. Correspondence to Freddie Bickford Smith <fbickfordsmith@cs.ox.ac.uk>.

Proceedings of the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

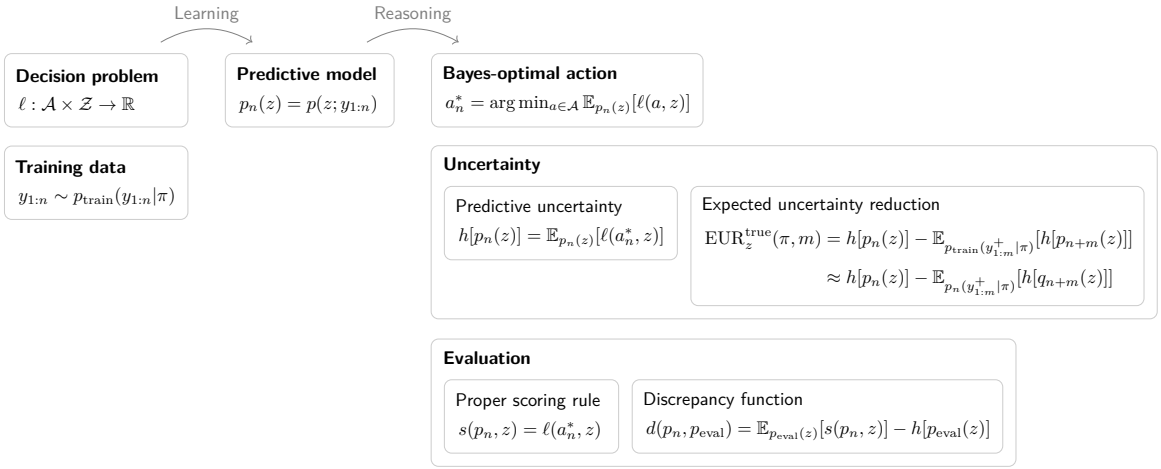

Figure 1 Our decision-theoretic view allows us to disentangle machine-learning concepts that have been conflated under the aleatoric-epistemic view. We consider taking an action, $a \in \mathcal{A}$, in light of imperfect knowledge of $z \in \mathcal{Z}$, with an action's consequences measured by a loss function, $\ell(a, z)$. Since z is unknown, we use any available training data, $y_{1:n}$, to build a predictive model, $p_n(z)$, with which we can reason over possible values of z and thus choose an action. We can also perform purely subjective reasoning to quantify various notions of model-based uncertainty, or evaluate the model using a ground-truth value of z or a reference distribution, $p_{\text{eval}}(z)$.

Bridging this generalised perspective back to how aleatoric and epistemic uncertainty have been discussed in past work, we provide new insights on BALD, a popular information-theoretic objective for data acquisition (Gal et al, 2017; Houlsby et al, 2011; Lindley, 1956). In particular we highlight that it should be seen not as a direct measure of long-run reducible predictive uncertainty, as has been suggested in the past, but instead as an estimator that can be highly inaccurate. Reconciling this with BALD's practical utility, we suggest it is often better understood as approximately measuring short-run reductions in parameter uncertainty. It can therefore be useful, albeit still suboptimal in prediction-oriented settings (Bickford Smith et al, 2023; 2024).

Our work thus serves to inform future work in two key ways. On the one hand it sheds light on the contradictions of the aleatoric-epistemic view and presents a coherent alternative perspective that allows clearer thinking about uncertainty in machine learning. On the other hand it provides more direct practical insights. It clarifies that what might have seemed like arbitrary choices for a decision-maker can instead be made by following well-defined logic: given some basic components and principles, it becomes clear how we should measure predictive uncertainty, how to assess models using reference systems and how to identify good future training data. It also highlights approximations that often have to be made in practice, revealing scope for suboptimal performance and therefore informing future methods research.

2. Background

The broad idea of the aleatoric-epistemic decomposition is to distinguish between different sources of uncertainty. If a model's prediction is uncertain, we might want to know whether that prediction is fundamentally uncertain for the given model class or instead due to a lack of data. This breakdown has clear utility in the context of seeking new data that will reduce predictive uncertainty (Bickford Smith et al, 2023; 2024; MacKay, 1992a;b). But it is also relevant elsewhere: in model selection, for example, we might want to quantify a model's scope for improvement by forecasting how its predictions will change given more data (Barbieri & Berger, 2004; Fong & Holmes, 2020; Geisser & Eddy, 1979; Kadane & Lazar, 2004; Laud & Ibrahim, 1995).

Uncertainty that resolves in light of new data can be thought of as "epistemic" in the sense that data conveys knowledge. Intuitively the corresponding irreducible uncertainty seems to be determined by not only the model class but also, among other things, an "inherent" level of uncertainty associated with the data source at hand, which is often thought of in terms of randomness or chance, hence the word "aleatoric".

While the concepts of aleatoric and epistemic uncertainty had previously been used in machine learning, for example by Lawrence (2012) and Senge et al (2014), their popularity grew following work by Gal (2016), Gal et al (2017) and Kendall & Gal (2017). The most widely used mathematical definitions of these ideas, which we will discuss in Section 4, are the information-theoretic quantities used by Gal et al (2017), building on earlier work on Bayesian experimental design (Lindley, 1956) and Bayesian active learning (Houlsby et al, 2011; MacKay, 1992a;b).

A range of perspectives on aleatoric and epistemic uncertainty in machine learning have been put forward in recent years. These include a discussion of where uncertainty comes from in machine learning (Gruber et al, 2023); a case against Shannon entropy for notions of predictive un-

certainty (Wimmer et al, 2023); proposals for using alternative information-theoretic quantities (Schweighofer et al, 2023a;b; 2024); and various other suggestions for how to define uncertainty, such as in terms of frequentist risk (Lahlou et al, 2023), class-wise variance (Sale et al, 2023b; 2024b), credal sets (Hofman et al, 2024a; Sale et al, 2023a), distances between probability distributions (Sale et al, 2024a) and proper scoring rules (Hofman et al, 2024b). As we will show, our replacement for the aleatoric-epistemic view unifies and explains many of the ideas in this recent work.

3. Key concepts

Our aim in this work is to formalise and unify quantities that have been associated with the ideas of aleatoric and epistemic uncertainty in past work. In particular we look to identify a rigorous notion of predictive uncertainty and the extent to which it reduces as more data is observed, and also measures of predictive performance and statistical dispersion in data. We start by highlighting some foundational concepts that will be used throughout our discussion.

3.1. Reasoning should start with the decision of interest

We consider taking an action, $a \in \mathcal{A}$, under imperfect knowledge of a ground-truth variable, $z \in \mathcal{Z}$. Here z could for example be an output relating to a given input (if so, the input is left implicit in our notation) or a parameter in a model, and a could be a direct prediction of z, with $\mathcal{A} = \mathcal{Z}$ for point prediction, or $\mathcal{A} = \mathcal{P}(\mathcal{Z})$ for probabilistic prediction. We emphasise our choice to focus on this decision, in deliberate contrast with the more common starting point of learning a model from fixed data. We want a notion of predictive uncertainty that is grounded in actions and their consequences, and we need to reason about different possible datasets to rigorously think about reductions in uncertainty.

3.2. Actions induce losses that reflect preferences

We assume we can measure the consequences of taking action a in light of a realisation of z using a loss (or negative utility) function, $\ell : \mathcal{A} \times \mathcal{Z} \to \mathbb{R}$. In principle the specification of ℓ follows directly from having preferences that satisfy basic axioms of rationality (von Neumann & Morgenstern, 1947). In practice it can be hard to know what ℓ should be; options for dealing with this include using an intrinsic loss or a random loss (Robert, 1996; 2007).

3.3. Subjective expected loss enables decision-making

Since ℓ is a function of the unknown z, it cannot be used directly as an objective for selecting an action, a. A principled solution that we focus on here is to form subjective beliefs over possible values of z (conventionally this belief state would be a Bayesian prior or posterior), average over these to form an expected loss, then choose an action that minimises this subjective expected loss (Ramsey, 1926; Savage, 1951). Alternative approaches to decision-making include minimax (von Neumann, 1928; Wald, 1949), which involves acting so as to minimise the worst-case loss.

3.4. Machine learning allows data-driven prediction

Minimising subjective expected loss requires beliefs over z, and those beliefs can often be informed by some training data, $y_{1:n} \sim p_{\text{train}}(y_{1:n}|\pi)$, where $\pi \in \Pi$ is a policy that controls aspects of data generation. We want notions of uncertainty that reflect how we will actually learn from data, rather than assuming idealised updates that we cannot perform in practice. We therefore define our predictive beliefs, $p_n(z) = p(z; y_{1:n})$, to be the output of a generic machine-learning method applied to the training data (and the input of interest if there is one) for any given n, which lets us reason about actual changes in uncertainty as n varies. Conventional Bayesian inference—taking a generative model over possible data and conditioning on the observed data, giving $p_n(z) = p(z|y_{1:n})$—is one possible update method. Others include deep learning (LeCun et al, 2015), in-context learning (Brown et al, 2020) and non-Bayesian ensemble methods (Breiman, 2001). In some cases the predictive distribution is defined as $p_n(z) = \mathbb{E}_{p_n(\theta)}[p_n(z|\theta)]$ where $\theta \sim p_n(\theta) = p(\theta; y_{1:n})$ represents a set of stochastic parameters that we average over at prediction time.

If the model-updating scheme is itself stochastic, we take the convention that this stochasticity is implicitly absorbed into $y_{1:n}$. This is because we can mathematically consider our machine-learning method to take in both the training data and a random-number seed, from which the model update is a deterministic mapping. Thus, while randomness in training is an important source of variability in how uncertainty can reduce, this variability can be dealt with as part of the variability already present in what data we observe.

3.5. Bayes optimality is a subjective notion

An action taken by minimising subjective expected loss under $p_n(z)$ is referred to as Bayes optimal (Murphy, 2022); if the action is an estimator of some quantity of interest then it is known as a Bayes estimator. The notion of Bayes optimality assumes our beliefs, $p_n(z)$, represent our best knowledge of z, and ℓ reflects our preferences. It says nothing at all about how well our beliefs match reality, or about the actions we would take if we had different beliefs. Bayes-optimal actions can therefore be suboptimal as judged using realisations of z from somewhere other than $p_n(z)$, such as a system serving as a source of ground truth.

3.6. Predictions often do not match data generation

The correspondence between our predictions, $p_n(z)$, and the data-generating process, $p_{\text{train}}(y_i|\pi(y_{<i}), y_{<i})$, can be

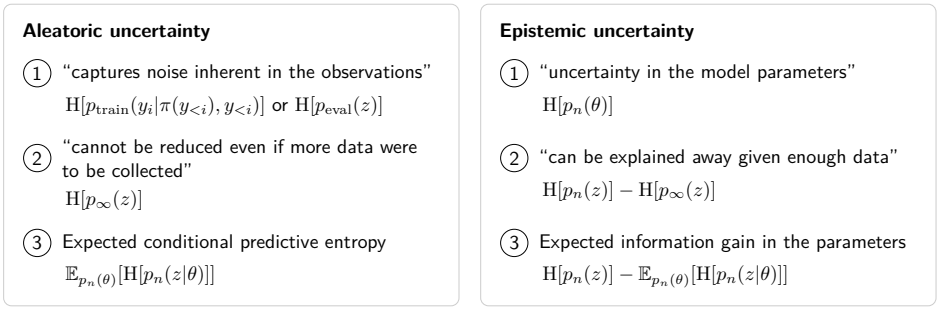

Figure 2 A popular view on aleatoric and epistemic uncertainty in machine learning attaches multiple mathematical quantities to each of the two concepts, rendering it incoherent and thus a likely source of conflations in the literature. Some of these quantities can coincide in particular cases but in the general case they are distinct. The quotations here are from Kendall & Gal (2017) and have been expressed mathematically as explicit information-theoretic quantities. The interpretation of Equation 1 is due to Gal (2016) and Gal et al (2017).

weak. One basic reason for this is that they might be defined over different event spaces. We could for example have $y_i \notin \mathcal{Z}$: perhaps we want to predict a coin's bias based on outcomes of coin tosses, or we want to predict a variable in one domain (eg, vision) based on data from another domain (eg, text). Even if that is not the case, $p_n(z)$ is a reflection of assumptions and design decisions based on incomplete knowledge (Box, 1976; Kleijn & van der Vaart, 2006), and there is no general guarantee that it will match reality.

3.7. Reference systems allow grounded evaluation

Because we expect $p_n(z)$ to be imperfect, we typically want to evaluate it against a reference system, $p_{\mathrm{eval}}(z)$, which could for example be a computer program, a human expert or a physical sensor. It is common in machine learning to evaluate models using an estimator of the frequentist risk (Berger, 1985), such as the mean squared error on finite data sampled from $p_{\mathrm{eval}}(z)$. Notably $p_{\mathrm{eval}}(z)$ could itself be imperfect (Fluri et al, 2023), so care is needed in designing and interpreting evaluations using reference systems.

4. Assessing a popular view

Having outlined intuitive descriptions of aleatoric and epistemic uncertainty and their motivation in Section 2, we turn to how they have been formalised in machine learning. Aleatoric and epistemic uncertainty are often thought of as additive components of predictive uncertainty. A popular way to formalise this for models with stochastic parameters, θ, is to relate three information-theoretic quantities:

$$\underbrace{\mathrm{EIG}_\theta}_{\text{“epistemic”}} = \underbrace{\mathrm{H}[p_n(z)]}_{\text{“total”}} - \underbrace{\mathbb{E}_{p_n(\theta)}[\mathrm{H}[p_n(z|\theta)]]}_{\text{“aleatoric”}} \quad (1)$$

where H denotes Shannon entropy and EIG_θ, also known as the BALD score (Houlsby et al, 2011), is the expected information gain in θ from observing z. Gal (2016) stated the "total = aleatoric + epistemic" relationship and the correspondence between $p_n(z)$ and total uncertainty, while Gal et al (2017) made the explicit link to Equation 1, informed by Houlsby et al (2011). Kendall & Gal (2017) discussed the aleatoric-epistemic view in the context of computer vision.

While that work successfully captured some of the intuitions from Section 2, we highlight that it also overloaded the ideas of aleatoric and epistemic uncertainty with multiple meanings, introducing a number of spurious associations (Figure 2). The competing definitions of aleatoric uncertainty conflate $\mathrm{H}[p_\infty(z)]$, measuring the uncertainty in $p_n(z)$ as $n \to \infty$ (this depends on the data-generating process, as we will discuss in Section 5.2), with three separate quantities:

(a) $\mathrm{H}[p_{\mathrm{train}}(y_i|\pi(y_{<i}), y_{<i})]$, the entropy in training-data generation. Issue: $p_\infty(z)$ is a subjective belief state that need not match $p_{\mathrm{train}}(y_i|\pi(y_{<i}), y_{<i})$ (Section 3.6).

(b) $\mathrm{H}[p_{\mathrm{eval}}(z)]$, the entropy of the reference system used in evaluation. Issue: $p_\infty(z)$ is a subjective belief state that need not match $p_{\mathrm{eval}}(z)$ (Sections 3.6 and 3.7).

(c) $\mathbb{E}_{p_n(\theta)}[\mathrm{H}[p_n(z|\theta)]]$, the expected conditional predictive entropy. Issue: for finite n the expected conditional predictive entropy is only an estimator of $\mathrm{H}[p_\infty(z)]$, and it can be highly inaccurate (Section 5.5).

Meanwhile the multiple definitions of epistemic uncertainty mix up $\mathrm{H}[p_n(z)] - \mathrm{H}[p_\infty(z)]$, the predictive-entropy reduction from updating on infinite new data, with two quantities:

(a) $\mathrm{H}[p_n(\theta)]$, the entropy of the model's stochastic parameters. Issue: the mapping from parameters to predictions is typically not invertible, so $\mathrm{H}[p_n(\theta)]$ will not necessarily relate to the reduction in predictive entropy.

(b) $\mathrm{H}[p_n(z)] - \mathbb{E}_{p_n(\theta)}[\mathrm{H}[p_n(z|\theta)]]$, the expected information gain in the model parameters. Issue: for finite n this EIG is only an estimator of $\mathrm{H}[p_n(z)] - \mathrm{H}[p_\infty(z)]$, and the estimation error can be large (Section 5.5).

Other sources of confusion in this aleatoric-epistemic view include an incorrect association between a model's subjective uncertainty and frequentist measures of performance,

such as classification accuracy (Figure 2 in Kendall & Gal (2017)), along with misleading implications about how a model's uncertainty will behave with varying n (Figure 6.11-6.12 in Gal (2016) and Table 3 in Kendall & Gal (2017)) and varying distance from the training data ("Aleatoric uncertainty does not increase for out-of-data examples... whereas epistemic uncertainty does" in Kendall & Gal (2017)).

5. An alternative perspective

We now present a coherent, general synthesis of key ideas used in existing discussions of aleatoric and epistemic uncertainty (Figure 1). We begin by reasoning about the subjective expected loss of acting Bayes-optimally under a given belief state, which leads to a decision-grounded measure of predictive uncertainty. By thinking about how that uncertainty would change in light of new data, we then identify a notion of expected uncertainty reduction, which we use to explain uncertainty decomposition. Then, shifting our focus from purely subjective reasoning to externally grounded evaluation, we highlight the distinction between uncertainty, predictive performance and data dispersion. Finally we return to the BALD score discussed in Section 4, providing insights on its utility as a data-acquisition objective.

5.1. Predictive uncertainty can be derived from the final decision of interest and the associated loss function

We first deal with the question of how to measure predictive uncertainty, to which many different answers have been put forward (Sections 1 and 2). Revisiting past work, we show that minimising subjective expected loss (Ramsey, 1926; Savage, 1951) directly leads to a loss-grounded measure of uncertainty that reflects our preferences about model behaviour in the final decision of interest. We thus clarify that a decision-maker does not face an arbitrary choice over uncertainty measures: if they specify a loss function based on their preferences, a rigorous uncertainty measure follows.

If $p_n(z)$ represents our beliefs over z then we can identify the Bayes-optimal action, a_n^*, in our final decision of interest by minimising the expected loss under those beliefs:

$$a_n^* = \arg\min_{a \in \mathcal{A}} \mathbb{E}_{p_n(z)}[\ell(a, z)] \,. \tag{2}$$

Now we can reason about the loss we expect (under our belief state) to incur by taking this Bayes-optimal action. An important, underappreciated result is that this minimal expected loss provides a way to measure uncertainty in $p_n(z)$ (Dawid, 1998; DeGroot, 1962; Neiswanger et al, 2022):

$$h[p_n(z)] = \mathbb{E}_{p_n(z)}[\ell(a_n^*, z)] \,.$$

A crucial implication of this is that any two decision-makers should not necessarily use the same uncertainty measure, depending on their decisions of interest and loss functions. One might use variance (Hastie et al, 2009) while the other uses entropy (Shannon, 1948), as Examples 1 and 2 show.

Example 1 (Dawid, 1998) *Point prediction with $\mathcal{A} = \mathcal{Z}$ and $\ell(a, z) = (a - z)^2$ corresponds to measuring uncertainty in our beliefs, $p_n(z)$, using variance.*

Proof The optimal action is the mean of $p_n(z)$:

$$a_n^* = \arg\min_{a \in \mathcal{A}} \mathbb{E}_{p_n(z)}\left[(a - z)^2\right] = \mathbb{E}_{p_n(z)}[z] \,.$$

The subjective expected loss of taking this action is the variance of $p_n(z)$:

$$h[p_n(z)] = \mathbb{E}_{p_n(z)}\left[(\mathbb{E}_{p_n(z)}[z] - z)^2\right] = \mathbb{V}_{p_n(z)}[z] \,. \; \square$$

Example 2 (Dawid, 1998) *Probabilistic prediction with $\mathcal{A} = \mathcal{P}(\mathcal{Z})$ and $\ell(a, z) = -\log a(z)$ corresponds to measuring uncertainty in our beliefs, $p_n(z)$, using entropy.*

Proof The optimal action is $p_n(z)$:

$$a_n^* = \arg\min_{a \in \mathcal{A}} -\mathbb{E}_{p_n(z)}[\log a(z)] = p_n(z).$$

The subjective expected loss of taking this action is the Shannon entropy of $p_n(z)$:

$$h[p_n(z)] = -\mathbb{E}_{p_n(z)}[\log p_n(z)] = \mathrm{H}[p_n(z)] \,. \qquad \square$$

5.2. Decomposing predictive uncertainty requires accounting for the data-generating process

Next, to formalise the popular idea of decomposing predictive uncertainty into irreducible and reducible components (Sections 2 and 4), we reason about how predictive uncertainty changes in light of new data. We show that, as long as we explicitly account for the data-gathering process, we can write down a notion of expected uncertainty reduction that is well defined for any method that maps from data to a predictive distribution (Section 3.4). From this we identify a rigorous irreducible-reducible decomposition.

Characterising the reducibility of uncertainty initially seems as simple as considering new data, $y^+_{1:m} = y_{(n+1):(n+m)}$, and measuring the corresponding uncertainty reduction,

$$\mathrm{UR}_z(y^+_{1:m}) = h[p_n(z)] - h[p_{n+m}(z)] \,,$$

which we note could be negative. But this uncertainty reduction depends on exactly what the new data is, and that in turn depends on the process by which the data is generated. We therefore need to explicitly account for the data-generating process to produce a well-defined notion of reducibility. If our model-updating scheme is stochastic then this also needs to be taken into account, but this stochasticity can be absorbed into the definition of $y^+_{1:m}$ (Section 3.4).

Revisiting $p_{\text{train}}(y^+_{1:m}|\pi)$ from Section 3.4, we define the distribution over y^+_i to depend both on decisions made by the data-acquisition policy, π, and on the previous data:

$$p_{\text{train}}(y^+_{1:m}|\pi) = \prod_{i=1}^{m} p_{\text{train}}(y^+_i|\pi(y_{<i}), y_{<i}).$$

With this we can define the true expected uncertainty reduction (EUR) in z under a given policy, π, as

$$\text{EUR}^{\text{true}}_z(\pi, m) = \mathbb{E}_{p_{\text{train}}(y^+_{1:m}|\pi)}[\text{UR}_z(y^+_{1:m})]\,. \quad (3)$$

This allows us to shift from thinking about a specific realisation of data to the range of possible data that might be generated. Working from this EUR to an uncertainty decomposition, we consider the limit of $m \to \infty$:

$$\underbrace{\text{EUR}^{\text{true}}_z(\pi, \infty)}_{\text{reducible}} = \underbrace{h[p_n(z)]}_{\text{total}} - \underbrace{\mathbb{E}_{p_{\text{train}}(y^+_{1:\infty}|\pi)}[h[p_\infty(z)]]}_{\text{irreducible}}\,.$$

Thus we see that three components—a loss function, a machine-learning method mapping from data to a predictive distribution, and a data-acquisition policy—fully specify a rigorous measure of expected uncertainty reduction and an associated irreducible-reducible decomposition. This contrasts with the decomposition in Equation 1, which requires stochastic parameters and exact Bayesian updating.

It is worth noting that in some restricted cases the dependency on the data-acquisition policy, π, of the infinite-data terms in this uncertainty decomposition can disappear. For example, if we are in a supervised-learning setting where the policy's decisions are based on inputs to acquire labels for, and if we are using a well-specified Bayesian model and exact Bayesian updates, $p_\infty(z)$ should be independent of π as long as π produces dense samples across the input space (Kleijn & van der Vaart, 2012). However, the requirements for this are very strict, with any model misspecification or error in belief updating reintroducing the dependency.

5.3. Practical estimation of expected uncertainty reduction relies on approximations

Now we turn to estimating expected uncertainty reduction (EUR) in practice. The decomposition in Section 5.2 is well defined but the infinite-data quantities within it are typically not practically obtainable. While this might seem problematic, we suggest the takeaway should in fact be to deemphasise the decomposition in the context of real-world machine learning, where we do not have infinite data. There is more concrete value (eg, for data acquisition) in estimating the EUR in Equation 3 for finite m. We now point out key approximations required for practical estimation.

Since we typically do not know the true data-generating process, $p_{\text{train}}(y^+_{1:m}|\pi)$, a core approximation is to use a model over new data, $p_n(y^+_{1:m}|\pi)$, as a proxy. On top of this, we might also need to approximate how our beliefs over z update when we obtain new data. In principle the EUR is defined with respect to whichever update scheme we are using, but in practice the true update can be too expensive to perform within an expectation over new data. This can be addressed by using some $q_{n+m}(z)$ in place of $p_{n+m}(z)$. A common approach is to assume $q_{n+m}(z) \propto p_n(y^+_{1:m}|z)p_n(z)$, even if the true update is not Bayesian (Bickford Smith et al, 2023; 2024; Gal et al, 2017; Kirsch et al, 2019; 2023). Notably this applies to methods based on Bayesian models that do not permit exact inference. If the true update is Bayesian and can be performed exactly then this is of course not an approximation.

Combining model-based data simulation with an approximate updating scheme, we can estimate the true EUR using

$$\text{EUR}'_z(\pi, m) = h[p_n(z)] - \mathbb{E}_{p_n(y^+_{1:m}|\pi)}[h[q_{n+m}(z)]]\,.$$

The accuracy of this estimator depends on the mismatch between $p_n(y^+_{1:m}|\pi)$ and $p_{\text{train}}(y^+_{1:m}|\pi)$ as well as the mismatch between $q_{n+m}(z)$ and $p_{n+m}(z)$, both of which are likely to be greater for larger m. Estimation therefore requires careful tradeoffs to mitigate these mismatches.

The practical relevance of this becomes clearer upon appreciating that the EUR estimator generalises a number of existing data-acquisition objectives. Under the exact-Bayesian-update assumption it is equivalent to what has variously been called the "expected value of [additional/sample] information" (Bernardo & Smith, 1994; Raiffa & Schlaifer, 1961), the "expected $H_{\ell,\mathcal{A}}$-information gain" (Neiswanger et al, 2022) and the "expected decision utility gain" (Huang et al, 2024). From that objective we can then recover the expected information gain in z, which corresponds to the BALD score (Gal et al, 2017; Houlsby et al, 2011) if $z = \theta$ represents a set of stochastic model parameters, or the expected predictive information gain (Bickford Smith et al, 2023) if $z = (x_*, y_*)$ represents a target input and its output. We note that $\text{EUR}'_z(\pi, m)$ is also closely related to the idea of the martingale posterior (Fong et al, 2023) as $p_n(y^+_{1:m}|\pi)$ and our updating scheme together imply a joint distribution from which a martingale posterior can be derived.

5.4. Model-based uncertainty should be used with care, and externally grounded evaluation is crucial

Next we clarify the relationship between the predictive uncertainty we have discussed so far and quantities commonly associated with it (Sections 2 and 4): measures of predictive performance and data dispersion. We identify how a model could be used to estimate those quantities but emphasise the limitations of that approach, underlining the key role to be played by externally grounded evaluation (Section 3.7).

We consider assessing a predictive distribution, $p_n(z)$, either

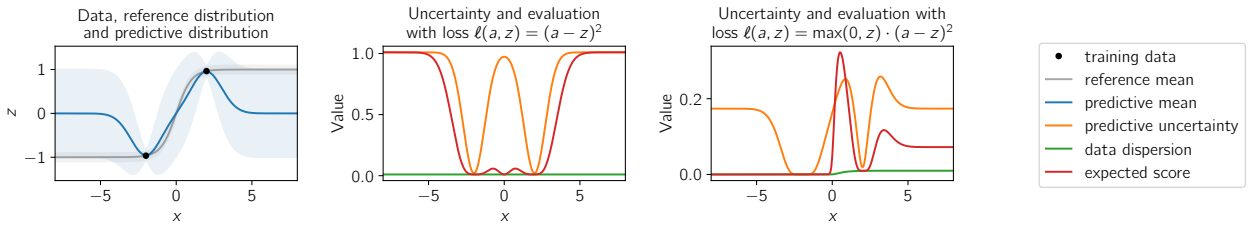

Figure 3 Taking a decision-theoretic perspective allows us to disentangle model-based uncertainty from predictive performance and data dispersion. Here we have a model's predictive distribution over an unknown variable, z, along with a reference distribution serving as a source of evaluation data. Using the model alone, we compute a measure of predictive uncertainty using a loss function, $\ell(a, z)$, where a is an action. This differs from the data dispersion, which describes the reference distribution. Connecting the two distributions, the expected score (lower is better) measures the predictive performance of the model as judged using data from the reference distribution.

using a single ground-truth value of z or using a reference distribution over z. If we have a single z, we can evaluate $p_n(z)$ using a proper scoring rule (Savage, 1971) of the form

$$s(p_n, z) = \ell(a_n^*, z)$$

where a_n^* is defined as in Equation 2 (Dawid, 1998). This scoring rule measures the loss incurred by the Bayes-optimal action under $p_n(z)$ when the ground truth is z.

If we instead have a reference distribution, $p_{\text{eval}}(z)$, we can evaluate $p_n(z)$ using a discrepancy function (Dawid, 1998):

$$d(p_n, p_{\text{eval}}) = \mathbb{E}_{p_{\text{eval}}(z)}[\ell(a_n^*, z) - \ell(a_{\text{eval}}^*, z)] \quad (4)$$
$$= \mathbb{E}_{p_{\text{eval}}(z)}[s(p_n, z)] - h[p_{\text{eval}}(z)] \quad (5)$$

where a_{eval}^* is defined analogously to a_n^* but with $p_{\text{eval}}(z)$ as the predictive distribution. Equation 4 highlights that the discrepancy function measures the expected excess loss from acting based on $p_n(z)$ rather than $p_{\text{eval}}(z)$ when $p_{\text{eval}}(z)$ represents ground truth, which is connected to the idea of regret in learning theory (Szepesvári, 2010). Equation 5 shows this is equivalent to the expected loss under $p_{\text{eval}}(z)$ of the Bayes-optimal action derived from $p_n(z)$, minus the uncertainty measure from Section 5.1 applied to $p_{\text{eval}}(z)$.

We can rearrange Equation 5 to highlight a decomposition:

$$\underbrace{\mathbb{E}_{p_{\text{eval}}(z)}[s(p_n, z)]}_{\text{expected score}} = \underbrace{d(p_n, p_{\text{eval}})}_{\text{discrepancy}} + \underbrace{h[p_{\text{eval}}(z)]}_{\text{data dispersion}}$$

where the expected score (lower is better) measures predictive performance and the data dispersion measures the spread in evaluation data drawn from the reference distribution. This generalises classic decompositions from statistics (Rice, 2007) and information theory (Cover & Thomas, 2005), as demonstrated in Examples 3 and 4.

Because past work has sometimes drawn connections between model-based uncertainty, predictive performance and data dispersion, we now explain how such connections can come about and emphasise that in the general case they should not be taken to hold. In particular, if we assume a specific model setup and estimation loss then we can derive Bayes estimators that generalise the predictive entropy and the expected conditional predictive entropy in Equation 1, but we emphasise that these assumptions will often not apply and that the estimators can be inaccurate.

Proposition 1 (Berger, 1985) *Let F be a quantity of interest, and let $f(\theta)$ represent subjective beliefs over F, derived from a pushforward of a distribution on model parameters $\theta \sim p_n(\theta)$. Under a quadratic estimation loss, $\ell_\eta(\eta, \theta) = (\eta - f(\theta))^2$, the Bayes estimator of F is $\eta^* = \mathbb{E}_{p_n(\theta)}[f(\theta)]$.*

Proposition 2 *Assume $p_n(z) = \mathbb{E}_{p_n(\theta)}[p_n(z|\theta)]$ is a model intended to directly approximate $p_{\text{eval}}(z)$. Then the model's predictive uncertainty, $h[p_n(z)]$, is a Bayes estimator of $\mathbb{E}_{p_{\text{eval}}(z)}[s(p_n, z)]$, the expected loss from acting Bayes-optimally under $p_n(z)$ when z is in fact drawn from $p_{\text{eval}}(z)$.*

Proof Applying Proposition 1 with $F = \mathbb{E}_{p_{\text{eval}}(z)}[s(p_n, z)]$ and $f(\theta) = \mathbb{E}_{p_n(z|\theta)}[s(p_n, z)]$ gives $\eta^* = h[p_n(z)]$. □

Proposition 3 *Assume the same model as in Proposition 2. Then the expected conditional predictive uncertainty, $\mathbb{E}_{p_n(\theta)}[h[p_n(z|\theta)]]$, is a Bayes estimator of $h[p_{\text{eval}}(z)]$, the dispersion in data drawn from a reference distribution.*

Proof Applying Proposition 1 with $F = h[p_{\text{eval}}(z)]$ and $f(\theta) = h[p_n(z|\theta)]$ gives $\eta^* = \mathbb{E}_{p_n(\theta)}[h[p_n(z|\theta)]]$. □

Three things are important to note in regard to Propositions 2 and 3. First, while they both suppose that $p_n(z)$ is designed to directly approximate $p_{\text{eval}}(z)$, this need not always be the case. Because $p_{\text{eval}}(z)$ might itself be imperfect, our evaluation might just be serving to provide a rough signal of the model's predictive performance (Section 3.7). Second, they both assume a quadratic estimation loss, which might not reflect our preferences. An alternative loss would result in different Bayes estimators. For example, an absolute loss would lead us to use medians rather than expectations (Berger, 1985). Third, they present estimators that are derived from subjective models and so have no accuracy guarantee in the general case (Section 3.5).

We therefore stress that model-based uncertainty should by default be considered separate from measures of predictive

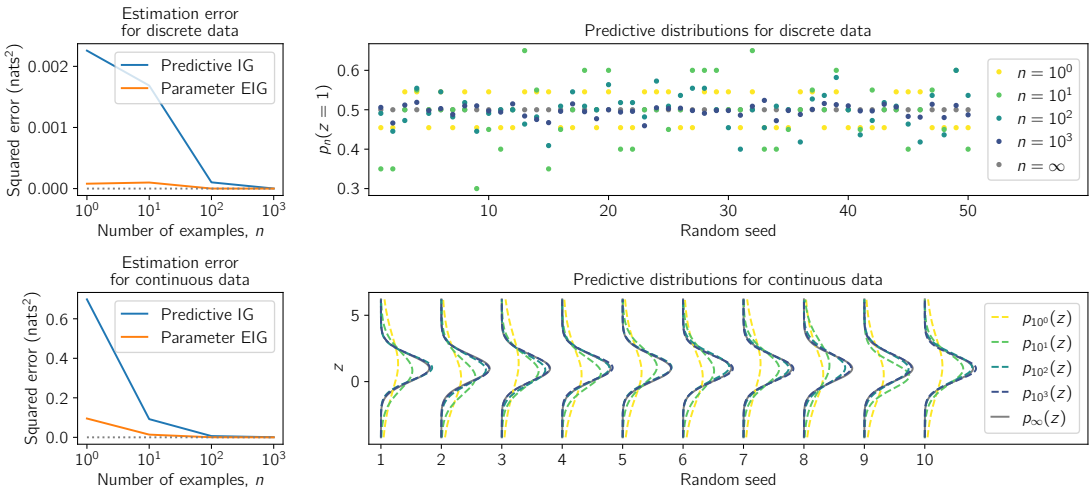

Figure 4 BALD's correspondence with infinite-step predictive information gain (a long-run reduction in a model's predictive entropy) can be weak, such that it is often better thought of as an estimator of a "true" one-step expected information gain in the model parameters. Here we show the behaviour of conjugate models trained on discrete data (top) and continuous data (bottom). For $n \in (1, 10, 100, 1000)$ we computed two estimation errors: $\varepsilon_z = (\mathrm{EIG}_\theta - \mathrm{IG}_z(y^+_{1:\infty}))^2$ and $\varepsilon_\theta = (\mathrm{EIG}_\theta - \mathrm{EIG}^{\mathrm{true}}_\theta)^2$. We show these errors (left; mean over 50 random seeds) along with the evolution of the predictive distributions (right). Both estimation errors are due to inaccurately forecasting future data, an issue that is resolved as n increases and the model's predictive distribution converges to the true data-generating process.

performance and data dispersion, with Figure 3 providing a concrete demonstration of their differences (see Appendix C for details). Uncertainty alone is not a reliable indicator of whether we can trust a model. Some kind of external grounding is crucial for well-informed practical deployment.

5.5. Popular information-theoretic quantities are best understood as imperfect estimators

Finally we return to the information-theoretic quantities in Equation 1, which have been central to many existing discussions of aleatoric and epistemic uncertainty. Principally we highlight that BALD (that is, EIG_θ, the expected information gain in a set of stochastic model parameters, θ) can be understood as an estimator of two separate unknown quantities: the infinite-step information gain in the model predictions and a "true" one-step expected information gain in the model parameters. We also suggest the relative magnitudes of the two corresponding estimation errors might help explain BALD's utility as a data-acquisition objective.

First we show that, under assumptions on the data and model that result in convergence to a single setting of θ (Doob, 1949; Freedman, 1963; 1965), BALD can be understood as an estimator of the infinite-step predictive information gain,

$$\mathrm{IG}_z(y^+_{1:\infty}) = \mathrm{H}[p_n(z)] - \mathrm{H}[p_n(z|y^+_{1:\infty})]\,,$$

measuring the reduction in the model's predictive entropy from a Bayesian update on infinite new data, $y^+_{1:\infty}$.

Proposition 4 *Let $y^+_{1:m}$ and $p_n(y|\theta)p_n(z|\theta)p_n(\theta)$ be a combination of data sequence and generative model that yield $p_n(\theta|y^+_{1:m}) \to \delta_{\theta_\infty}(\theta)$ as $m \to \infty$. Then the expected conditional predictive entropy, $\mathbb{E}_{p_n(\theta)}[\mathrm{H}[p_n(z|\theta)]]$, is a Bayes estimator of $\mathrm{H}[p_n(z|y^+_{1:\infty})]$, the marginal predictive entropy after a Bayesian update on infinite new data, $y^+_{1:\infty}$.*

Proposition 5 *Assume the data and model from Proposition 4. Then the expected information gain in the model parameters, EIG_θ, from observing z is a Bayes estimator of the infinite-step predictive information gain, $\mathrm{IG}_z(y^+_{1:\infty})$.*

In Figure 4 we demonstrate that the approximation $\mathrm{EIG}_\theta \approx \mathrm{IG}_z(y^+_{1:\infty})$ from Proposition 5 can be coarse. These results were produced with extremely simple setups within which we can perform exact inference and we are sure to recover the true data-generating process in the limit of infinite data (see Appendix C for details). We therefore know that the estimation error is due to a failure of the model to accurately forecast future data, which in turn is due to n being finite.

This behaviour appears to align with existing results (with the caveat that past studies did not match the assumptions of Propositions 4 and 5). Figure 2 in Bickford Smith et al (2024) and Figure 5 in Wimmer et al (2023) show small-n estimates of EIG_θ that differ substantially from the changes in predictive entropy that actually occurred in practice. Those results even suggest it would have been more accurate to assume $\mathrm{IG}_z(y^+_{1:\infty}) = \mathrm{H}[p_n(z)]$, with $\mathrm{H}[p_n(z|y^+_{1:\infty})] = 0$, than to estimate it using EIG_θ. Meanwhile Mucsányi et al (2024) and Valdenegro-Toro & Saromo-Mori (2022) emphasised the "entanglement" of aleatoric- and epistemic-uncertainty estimators, which can be understood as the estimators themselves having an "epistemic" component—or, in our terminology, being inaccurate finite-data estimators.

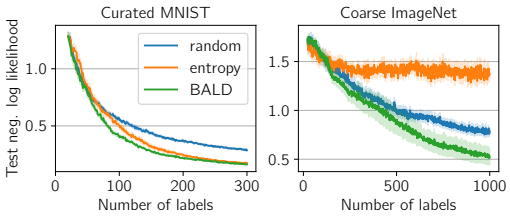

Figure 5 BALD outperforms predictive entropy as a data-acquisition objective in active learning, even though BALD tends to be a worse estimator of long-run predictive information gain in the setups studied. These results were produced using experimental setups described in Bickford Smith et al (2023; 2024).

This raises the question of why BALD has proven practically useful as a data-acquisition objective in active learning (Gal et al, 2017; Houlsby et al, 2011; Osband et al, 2023). If our intuition is that BALD's utility stems from its correspondence with infinite-step predictive information gain (that is, a long-run reduction in a model's predictive entropy) and we consider setups in which the current predictive entropy is a better estimator than BALD, then we would expect that using the predictive entropy as a data-acquisition objective would lead to better predictive performance in active learning. Yet this is not what we see in practice (Figure 5).

We suggest another perspective on BALD might address this question. In particular it can be understood (given assumptions on the data and model) as an estimator of the true one-step expected information gain in the model parameters,

$$\mathrm{EIG}_\theta^{\mathrm{true}} = \mathrm{H}[p_n(\theta)] - \mathbb{E}_{p_{\mathrm{train}}(z)}[\mathrm{H}[p_n(\theta|z)]]\,,$$

where the expectation is over observations from the true data-generating process, not model-simulated observations.

Proposition 6 *Let $z = y_{n+1}$. Assume $y_{1:n}$ are independent and identically distributed, with $y_i \sim p_{\mathrm{train}}(y)$, and assume $p_n(z) = \mathbb{E}_{p_n(\theta)}[p_n(z|\theta)]$ is a model intended to directly approximate $p_{\mathrm{train}}(z)$. Then the expected information gain in the model parameters, EIG_θ, from observing z is a Bayes estimator of the true one-step expected information gain, $\mathrm{EIG}_\theta^{\mathrm{true}}$, where the expectation is with respect to $p_{\mathrm{train}}(z)$.*

Returning to the same experimental setup as before, we find (Figure 4) that the approximation $\mathrm{EIG}_\theta \approx \mathrm{EIG}_\theta^{\mathrm{true}}$ is more accurate than $\mathrm{EIG}_\theta \approx \mathrm{IG}_z(y^+_{1:\infty})$. In other words, BALD more closely tracks short-run changes in parameter uncertainty than it does long-run changes in predictive uncertainty. We do not claim this is a general result that will hold in all settings, but it is consistent with BALD being useful as a data-acquisition objective. The data-acquisition horizons in active learning are typically very short, so it is the short-run notion of information gain that matters, not the asymptotic notion. And while targeting predictions rather than parameters can be even more effective (Bickford Smith et al, 2023; 2024), maximising short-run parameter information gain is still often preferable over random acquisition.

One takeaway here is that information-theoretic quantities should not be confused with the quantities they estimate. Another is that expected information gain in a variable of interest is a well motivated objective for data acquisition, assuming the action of interest is a probabilistic prediction of z and we use a negative-log-likelihood loss function (Example 2), but it also crucially depends on the model's ability to simulate future data, and it assumes a Bayesian update that might not match the true update (Section 5.3).

6. Conclusion

We have argued that the aleatoric-epistemic view on uncertainty does not serve machine-learning researchers' needs: its lack of expressive capacity has led to conceptual overloading and confusion. To address this we have presented a decision-theoretic view that unifies many concepts of interest to researchers. This provides clarity on five key points:

(a) Measures of predictive uncertainty need not be an arbitrary choice but can instead be derived from a decision of interest with an associated loss function.

(b) If we explicitly account for how training data is generated, we can identify a decomposition of uncertainty into reducible and irreducible components for any method that maps from data to a predictive distribution.

(c) In practice we can typically only produce an approximate notion of expected uncertainty reduction that relies on a proxy for the true data-generating process and possibly also an approximation of model updating.

(d) Predictive uncertainty should be assumed to be separate from measures of predictive performance and data dispersion, and externally grounded evaluation is therefore key for building trust in a model's predictions.

(e) BALD does not directly measure predictive-uncertainty reduction but is an (often inaccurate) estimator of it.

We suggest this new decision-theoretic view should be used in place of the predominant aleatoric-epistemic view as the field moves forward. Our hope is that this will support more productive discourse and methodological development.

A recurring point in our work is that in practice we almost always have to approximate quantities we are interested in, building on assumptions and design decisions. However, these approximations can be inaccurate, thereby requiring significant caution around the interpretation and use of practical quantities, noting that the accuracy of the approximation might vary between different interpretations of the original quantity. We thus believe it is important future work on quantifying sources of uncertainty is grounded in practical decision-making scenarios, such as active learning and model selection, so that such approximations can be judged in terms of the utility they provide in concrete problems.

Impact statement

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgements

We thank Philip Dawid for sharing a technical report with us, and Andreas Kirsch and Mike Osborne for useful discussions and feedback. Freddie Bickford Smith is supported by the EPSRC Centre for Doctoral Training in Autonomous Intelligent Machines and Systems (EP/L015897/1). Tom Rainforth is supported by EPSRC grant EP/Y037200/1.

References

Amini, Schwarting, Soleimany, & Rus (2020). Deep evidential regression. *Conference on Neural Information Processing Systems*.

Ayhan & Berens (2018). Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. *Conference on Medical Imaging with Deep Learning*.

Barbieri & Berger (2004). Optimal predictive model selection. *Annals of Statistics*.

Berger (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer.

Bernardo & Smith (1994). *Bayesian Theory*. John Wiley and Sons.

Bickford Smith, Foster, & Rainforth (2024). Making better use of unlabelled data in Bayesian active learning. *International Conference on Artificial Intelligence and Statistics*.

Bickford Smith, Kirsch, Farquhar, Gal, Foster, & Rainforth (2023). Prediction-oriented Bayesian active learning. *International Conference on Artificial Intelligence and Statistics*.

Box (1976). Science and statistics. *Journal of the American Statistical Association*.

Breiman (2001). Random forests. *Machine Learning*.

Brown, Mann, Ryder, Subbiah, Kaplan, Dhariwal, Neelakantan, Shyam, Sastry, Askell, Agarwal, Herbert-Voss, Krueger, Henighan, Child, Ramesh, Ziegler, Wu, Winter, Hesse, Chen, Sigler, Litwin, Gray, Chess, Clark, Berner, McCandlish, Radford, Sutskever, & Amodei (2020). Language models are few-shot learners. *Conference on Neural Information Processing Systems*.

Cover & Thomas (2005). *Elements of Information Theory*. John Wiley and Sons.

Dawid (1998). Coherent measures of discrepancy, uncertainty and dependence, with applications to Bayesian predictive experimental design. Technical report, University College London.

DeGroot (1962). Uncertainty, information, and sequential experiments. *Annals of Mathematical Statistics*.

Doob (1949). Application of the theory of martingales. *Le Calcul des Probabilités et ses Applications*.

Fluri, Paleka, & Tramèr (2023). Evaluating superhuman models with consistency checks. *arXiv*.

Fong & Holmes (2020). On the marginal likelihood and cross-validation. *Biometrika*.

Fong, Holmes, & Walker (2023). Martingale posterior distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Freedman (1963). On the asymptotic behavior of bayes estimates in the discrete case. *Annals of Mathematical Statistics*.

Freedman (1965). On the asymptotic behavior of bayes estimates in the discrete case II. *Annals of Mathematical Statistics*.

Gal (2016). *Uncertainty in deep learning*. PhD thesis, University of Cambridge.

Gal, Islam, & Ghahramani (2017). Deep Bayesian active learning with image data. *International Conference on Machine Learning*.

Geisser & Eddy (1979). A predictive approach to model selection. *Journal of the American Statistical Association*.

Gruber, Schenk, Schierholz, Kreuter, & Kauermann (2023). Sources of uncertainty in machine learning—a statisticians' view. *arXiv*.

Hacking (1975). *The Emergence of Probability*. Cambridge University Press.

Hastie, Tibshirani, Friedman, & Friedman (2009). *The Elements of Statistical Learning*. Springer.

Hofman, Sale, & Hüllermeier (2024a). Quantifying aleatoric and epistemic uncertainty: a credal approach. *Workshop on "Structured Probabilistic Inference and Generative Modeling", International Conference on Machine Learning*.

Hofman, Sale, & Hüllermeier (2024b). Quantifying aleatoric and epistemic uncertainty with proper scoring rules. *arXiv*.

Houlsby, Huszár, Ghahramani, & Lengyel (2011). Bayesian active learning for classification and preference learning. *arXiv*.

Huang, Guo, Acerbi, & Kaski (2024). Amortized Bayesian experimental design for decision-making. *Conference on Neural Information Processing Systems*.

Hüllermeier & Waegeman (2021). Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*.

Immer, Bauer, Fortuin, Rätsch, & Khan (2021). Scalable marginal likelihood estimation for model selection in deep learning. *International Conference on Machine Learning*.

Kadane & Lazar (2004). Methods and criteria for model selection. *Journal of the American Statistical Association*.

Kapoor, Maddox, Izmailov, & Wilson (2022). On uncertainty, tempering, and data augmentation in Bayesian classification. *Conference on Neural Information Processing Systems*.

Kendall & Gal (2017). What uncertainties do we need in Bayesian deep learning for computer vision? *Conference on Neural Information Processing Systems*.

Kirsch, Farquhar, Atighehchian, Jesson, Branchaud-Charron, & Gal (2023). Stochastic batch acquisition: a simple baseline for deep active learning. *Transactions on Machine Learning Research*.

Kirsch, van Amersfoort, & Gal (2019). BatchBALD: efficient and diverse batch acquisition for deep Bayesian active learning. *Conference on Neural Information Processing Systems*.

Kleijn & van der Vaart (2006). Misspecification in infinite-dimensional Bayesian statistics. *Annals of Statistics*.

Kleijn & van der Vaart (2012). The Bernstein-von-Mises theorem under misspecification. *Electronic Journal of Statistics*.

Lahlou, Jain, Nekoei, Butoi, Bertin, Rector-Brooks, Korablyov, & Bengio (2023). DEUP: direct epistemic uncertainty prediction. *Transactions on Machine Learning Research*.

Laud & Ibrahim (1995). Predictive model selection. *Journal of the Royal Statistical Society: Series B (Methodological)*.

Lawrence (2012). What is machine learning? videolectures.net/mlss2012_lawrence_machine_learning.

LeCun, Bengio, & Hinton (2015). Deep learning. *Nature*.

Lindley (1956). On a measure of the information provided by an experiment. *Annals of Mathematical Statistics*.

MacKay (1992a). The evidence framework applied to classification networks. *Neural Computation*.

MacKay (1992b). Information-based objective functions for active data selection. *Neural Computation*.

McAllister (2016). *Bayesian learning for data-efficient control*. PhD thesis, University of Cambridge.

Mucsányi, Kirchhof, & Oh (2024). Benchmarking uncertainty disentanglement: specialized uncertainties for specialized tasks. *Conference on Neural Information Processing Systems*.

Mukhoti, Kirsch, van Amersfoort, Torr, & Gal (2023). Deep deterministic uncertainty: a new simple baseline. *Conference on Computer Vision and Pattern Recognition*.

Murphy (2022). *Probabilistic Machine Learning: An Introduction*. MIT Press.

Neiswanger, Yu, Zhao, Meng, & Ermon (2022). Generalizing Bayesian optimization with decision-theoretic entropies. *Conference on Neural Information Processing Systems*.

Orlando, Seeböck, Bogunović, Klimscha, Grechenig, Waldstein, Gerendas, & Schmidt-Erfurth (2019). U2-net: a Bayesian U-net model with epistemic uncertainty feedback for photoreceptor layer segmentation in pathological OCT scans. *International Symposium on Biomedical Imaging*.

Osband, Wen, Asghari, Dwaracherla, Ibrahimi, Lu, & Van Roy (2023). Epistemic neural networks. *Conference on Neural Information Processing Systems*.

Postels, Blum, Cadena, Siegwart, Van Gool, & Tombari (2020). Quantifying aleatoric and episteic uncertainty using density estimation in latent space. *arXiv*.

Raiffa & Schlaifer (1961). *Applied Statistical Decision Theory*. Division of Research, Harvard Business School.

Ramsey (1926). Truth and probability. *Studies in Subjective Probability*.

Rice (2007). *Mathematical Statistics and Data Analysis*. Thomson Brooks/Cole.

Robert (1996). Intrinsic losses. *Theory and Decision*.

Robert (2007). *The Bayesian Choice*. Springer.

Sale, Bengs, Caprio, & Hüllermeier (2024a). Second-order uncertainty quantification: a distance-based approach. *International Conference on Machine Learning*.

Sale, Caprio, & Hüllermeier (2023a). Is the volume of a credal set a good measure for epistemic uncertainty? *Conference on Uncertainty in Artificial Intelligence*.

Sale, Hofman, Löhr, Wimmer, Nagler, & Hüllermeier (2024b). Label-wise aleatoric and epistemic uncertainty quantification. *Conference on Uncertainty in Artificial Intelligence*.

Sale, Hofman, Wimmer, Hüllermeier, & Nagler (2023b). Second-order uncertainty quantification: variance-based measures. *arXiv*.

Savage (1951). The theory of statistical decision. *Journal of the American Statistical Association*.

Savage (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*.

Schweighofer, Aichberger, Ielanskyi, & Hochreiter (2023a). Introducing an improved information-theoretic measure of predictive uncertainty. *Workshop on "Mathematics of Modern Machine Learning", Conference on Neural Information Processing Systems*.

Schweighofer, Aichberger, Ielanskyi, & Hochreiter (2024). On information-theoretic measures of predictive uncertainty. *arXiv*.

Schweighofer, Aichberger, Ielanskyi, Klambauer, & Hochreiter (2023b). Quantification of uncertainty with adversarial models. *Conference on Neural Information Processing Systems*.

Seeböck, Orlando, Schlegl, Waldstein, Bogunović, Klimscha, Langs, & Schmidt-Erfurth (2019). Exploiting epistemic uncertainty of anatomy segmentation for anomaly detection in retinal OCT. *Transactions on Medical Imaging*.

Senge, Bösner, Dembczyński, Haasenritter, Hirsch, Donner-Banzhoff, & Hüllermeier (2014). Reliable classification: learning classifiers that distinguish aleatoric and epistemic uncertainty. *Information Sciences*.

Shannon (1948). A mathematical theory of communication. *The Bell System Technical Journal*.

Smith & Gal (2018). Understanding measures of uncertainty for adversarial example detection. *Conference on Uncertainty in Artificial Intelligence*.

Szepesvári (2010). *Algorithms for Reinforcement Learning*. Morgan and Claypool.

Valdenegro-Toro & Saromo-Mori (2022). A deeper look into aleatoric and epistemic uncertainty disentanglement. *Workshop on "LatinX in CV Research", Conference on Computer Vision and Pattern Recognition*.

van Amersfoort, Smith, Teh, & Gal (2020). Uncertainty estimation using a single deep deterministic neural network. *International Conference on Machine Learning*.

von Neumann (1928). Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen*.

von Neumann & Morgenstern (1947). *Theory of Games and Economic Behavior*. Princeton University Press.

Wald (1949). Statistical decision functions. *Annals of Mathematical Statistics*.

Wang & Aitchison (2021). Bayesian OOD detection with aleatoric uncertainty and outlier exposure. *arXiv*.

Wang, Li, Aertsen, Deprest, Ourselin, & Vercauteren (2019). Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*.

Wen, Osband, Qin, Lu, Ibrahimi, Dwaracherla, Asghari, & Van Roy (2022). From predictions to decisions: the importance of joint predictive distributions. *arXiv*.

Wimmer, Sale, Hofman, Bischl, & Hüllermeier (2023). Quantifying aleatoric and epistemic uncertainty in machine learning: are conditional entropy and mutual information appropriate measures? *Conference on Uncertainty in Artificial Intelligence*.

A. Examples

Example 3 *Evaluating $p_n(z)$ against a reference distribution, $p_{\text{eval}}(z)$, with $\mathcal{A} = \mathcal{Z}$ and $\ell(a, z) = (a - z)^2$ corresponds to measuring the predictive performance of $p_n(z)$ using mean squared error, measuring the discrepancy between $p_n(z)$ and $p_{\text{eval}}(z)$ using squared bias and measuring dispersion in $p_{\text{eval}}(z)$ using variance.*

Proof Starting from Equation 4 with optimal actions $a_n^* = \mathbb{E}_{p_n(z)}[z] = \mu_n$ and $a_{\text{eval}}^* = \mathbb{E}_{p_{\text{eval}}(z)}[z] = \mu_{\text{eval}}$ from Example 1, the discrepancy between $p_n(z)$ and $p_{\text{eval}}(z)$ is

$$d(p_n, p_{\text{eval}}) = \mathbb{E}_{p_{\text{eval}}(z)}\left[(\mu_n - z)^2 - (\mu_{\text{eval}} - z)^2\right] = (\mu_n - \mu_{\text{eval}})^2.$$

Now starting from Equation 5, it can also be written as

$$d(p_n, p_{\text{eval}}) = \mathbb{E}_{p_{\text{eval}}(z)}\left[(\mu_n - z)^2\right] - \mathbb{V}_{p_{\text{eval}}(z)}[z]\,.$$

Equating these two expressions for the discrepancy leads to a standard bias-variance decomposition:

$$\underbrace{\mathbb{E}_{p_{\text{eval}}(z)}\left[(\mu_n - z)^2\right]}_{\text{mean squared error}} = \underbrace{(\mu_n - \mu_{\text{eval}})^2}_{\text{squared bias}} + \underbrace{\mathbb{V}_{p_{\text{eval}}(z)}[z]}_{\text{variance}}$$

where the mean squared error measures predictive performance and the variance measures data dispersion. □

Example 4 *Evaluating $p_n(z)$ against a reference distribution, $p_{\text{eval}}(z)$, with $\mathcal{A} = \mathcal{P}(\mathcal{Z})$ and $\ell(a, z) = -\log a(z)$ corresponds to measuring the predictive performance of $p_n(z)$ using cross entropy, measuring discrepancy between $p_n(z)$ and $p_{\text{eval}}(z)$ using Kullback-Leibler divergence and measuring dispersion in $p_{\text{eval}}(z)$ using Shannon entropy.*

Proof Starting from Equation 4 with optimal actions $a_n^* = p_n(z)$ and $a_{\text{eval}}^* = p_{\text{eval}}(z)$ (Example 2), the discrepancy between $p_n(z)$ and $p_{\text{eval}}(z)$ is

$$d(p_n, p_{\text{eval}}) = \mathbb{E}_{p_{\text{eval}}(z)}[-\log p_n(z) + \log p_{\text{eval}}(z)] = \mathrm{KL}[p_{\text{eval}}(z) \,\|\, p_n(z)]\,.$$

Now starting from Equation 5, it can also be written as

$$d(p_n, p_{\text{eval}}) = -\mathbb{E}_{p_{\text{eval}}(z)}[\log p_n(z)] - \mathrm{H}[p_{\text{eval}}(z)] = \mathrm{H}[p_{\text{eval}}(z) \,\|\, p_n(z)] - \mathrm{H}[p_{\text{eval}}(z)]\,.$$

Equating these two expressions for the discrepancy leads to a standard information-theoretic decomposition:

$$\underbrace{\mathrm{H}[p_{\text{eval}}(z) \,\|\, p_n(z)]}_{\text{cross entropy}} = \underbrace{\mathrm{KL}[p_{\text{eval}}(z) \,\|\, p_n(z)]}_{\text{KL divergence}} + \underbrace{\mathrm{H}[p_{\text{eval}}(z)]}_{\text{entropy}}$$

where the cross entropy measures predictive performance and the entropy measures data dispersion. □

B. Proofs of Propositions 4 to 6

Proposition 4 *Let $y^+_{1:m}$ and $p_n(y|\theta)p_n(z|\theta)p_n(\theta)$ be a combination of data sequence and generative model that yield $p_n(\theta|y^+_{1:m}) \to \delta_{\theta_\infty}(\theta)$ as $m \to \infty$. Then the expected conditional predictive entropy, $\mathbb{E}_{p_n(\theta)}[\mathrm{H}[p_n(z|\theta)]]$, is a Bayes estimator of $\mathrm{H}[p_n(z|y^+_{1:\infty})]$, the marginal predictive entropy after a Bayesian update on infinite new data, $y^+_{1:\infty}$.*

Proof Since $y^+_{1:\infty}$ recovers a single setting of θ, reasoning about θ is equivalent to reasoning about $y^+_{1:\infty}$, following the argument presented in Fong et al (2023). This allows us to apply Proposition 1 with $F = \mathrm{H}[p_n(z|y^+_{1:\infty})]$ and $f(\theta) = \mathrm{H}[p_n(z|\theta)]$, which gives $\eta^* = \mathbb{E}_{p_n(\theta)}[\mathrm{H}[p_n(z|\theta)]]$. □

Proposition 5 *Assume the data and model from Proposition 4. Then the expected information gain in the model parameters, EIG_θ, from observing z is a Bayes estimator of the infinite-step predictive information gain, $\mathrm{IG}_z(y^+_{1:\infty})$.*

Proof The information gain to be estimated, $\mathrm{IG}_z(y^+_{1:\infty})$, is defined as the reduction in the model's predictive entropy from a Bayesian update on infinite new data, $y^+_{1:\infty}$:

$$\mathrm{IG}_z(y^+_{1:\infty}) = \mathrm{H}[p_n(z)] - \mathrm{H}[p_n(z|y^+_{1:\infty})]\,.$$

Combining the known $\mathrm{H}[p_n(z)]$ with the Bayes estimator of $\mathrm{H}[p_n(z|y^+_{1:\infty})]$ from Proposition 4 gives

$$\mathrm{EIG}_\theta = \mathrm{H}[p_n(z)] - \mathbb{E}_{p_n(\theta)}[\mathrm{H}[p_n(z|\theta)]]$$

as a Bayes estimator of $\mathrm{IG}_z(y^+_{1:\infty})$. □

Proposition 6 *Let $z = y_{n+1}$. Assume $y_{1:n}$ are independent and identically distributed, with $y_i \sim p_{\mathrm{train}}(y)$, and assume $p_n(z) = \mathbb{E}_{p_n(\theta)}[p_n(z|\theta)]$ is a model intended to directly approximate $p_{\mathrm{train}}(z)$. Then the expected information gain in the model parameters, EIG_θ, from observing z is a Bayes estimator of the true one-step expected information gain, $\mathrm{EIG}^{\mathrm{true}}_\theta$, where the expectation is with respect to $p_{\mathrm{train}}(z)$.*

Proof The expected information gain to be estimated, $\mathrm{EIG}^{\mathrm{true}}_\theta$, is defined as the reduction in the model's parameter entropy from a Bayesian update on new data, z, where z is drawn from $p_{\mathrm{train}}(z)$:

$$\mathrm{EIG}^{\mathrm{true}}_\theta = \mathrm{H}[p_n(\theta)] - \mathbb{E}_{p_{\mathrm{train}}(z)}[\mathrm{H}[p_n(\theta|z)]]\,.$$

The second term here can be estimated by applying Proposition 1 with $F = \mathbb{E}_{p_{\mathrm{train}}(z)}[\mathrm{H}[p_n(\theta|z_{1:m})]]$ and $f(\theta) = \mathbb{E}_{p_n(z|\theta)}[\mathrm{H}[p_n(\theta|z)]]$. The Bayes estimator that results from this, $\eta^* = \mathbb{E}_{p_n(z)}[\mathrm{H}[p_n(\theta|z)]]$, can be combined with the known current entropy, $\mathrm{H}[p_n(\theta)]$, to produce

$$\mathrm{EIG}_\theta = \mathrm{H}[p_n(\theta)] - \mathbb{E}_{p_n(z)}[\mathrm{H}[p_n(\theta|z)]]$$

as a Bayes estimator of $\mathrm{EIG}^{\mathrm{true}}_\theta$. □

C. Implementation details

C.1. Figure 3

We consider predicting an output, $z \in \mathbb{R}$, corresponding to an input, $x \in \mathbb{R}$. The training data, $y_{1:n}$, comprises $n = 2$ input-label pairs: $y_1 = (-2, \tanh(-2))$ and $y_2 = (2, \tanh(2))$. We use this to compute a Gaussian-process predictive posterior, $p_n(z|x) = p(z|x, y_{1:n})$, based on a generative model comprising a Gaussian likelihood function, $p(z|x, \theta) = \text{Normal}(z|\theta(x), \sigma^2)$, where $\sigma = 0.1$, and a Gaussian-process prior, $\theta \sim \text{GP}(0, k)$, where $k(x, x') = \exp(-(x - x')^2/2)$. We compare this with a reference distribution, $p_{\text{eval}}(z|x) = \text{Normal}(y|\tanh(x), \sigma^2)$. Using $p_n(z|x)$ and $p_{\text{eval}}(z|x)$, we compute three quantities for $x \in [-8, 8]$: the predictive uncertainty, $h[p_n(z|x)]$; the data dispersion, $h[p_{\text{eval}}(z|x)]$; and the expected score, $\mathbb{E}_{p_{\text{eval}}(z|x)}[s(p_n, z)]$. We do this for two loss functions: $\ell(a, z) = (a-z)^2$ and $\ell(a, z) = \max(0, z)\cdot(a-z)^2$.

C.2. Figure 4

We consider two cases of predicting $z = y_{n+1}$: a discrete case and a continuous case. In the discrete case we have $y \in \{0, 1\}$, data generated from $p_{\text{train}}(y) = \text{Bernoulli}(y|\eta = 0.5)$, and the Bayesian generative model is

$$\begin{aligned} p(y, \eta|\alpha, \beta) &= p(y|\eta)p(\eta|\alpha, \beta) \\ p(y|\eta) &= \text{Bernoulli}(y|\eta) \\ p(\eta|\alpha, \beta) &= \text{Beta}(\eta|\alpha, \beta). \end{aligned}$$

In the continuous case we have $y \in \mathbb{R}$, data generated from $p_{\text{train}}(y) = \text{Normal}(y|\mu = 1, \sigma^2 = 1)$, and the Bayesian generative model is

$$\begin{aligned} p(y, \mu, \lambda|\alpha, \beta, \kappa, m) &= p(y|\mu, \lambda)p(\mu|m, \kappa, \lambda)p(\lambda|\alpha, \beta) \\ p(y|\mu, \lambda) &= \text{Normal}(y|\mu, \lambda^{-1}) \\ p(\mu|m, \kappa, \lambda) &= \text{Normal}(\mu|m, (\kappa\lambda)^{-1}) \\ p(\lambda|\alpha, \beta) &= \text{Gamma}(\lambda|\alpha, \beta). \end{aligned}$$

In both cases we can compute exact Bayesian posteriors (Murphy, 2022), and there is some n for which $p_n(z) = p_{\text{train}}(z)$ (Doob, 1949; Freedman, 1963; 1965). For each case we sample four datasets, $y_{1:n}$, with $y_i \sim p_{\text{train}}(y)$ and $n \in (1, 10, 100, 1000)$. On each dataset we compute the Bayesian parameter posterior, $p_n(\theta) = p(\theta|y_{1:n})$, where $\theta = (\alpha, \beta)$ or $\theta = (\alpha, \beta, \kappa, m)$, and then compute two quadratic estimation errors. The first is the error from approximating the "true" expected information gain in θ with the standard, model-based expected information gain in θ:

$$\begin{aligned} \sqrt{\varepsilon_\theta} &= \text{EIG}_\theta - \text{EIG}_\theta^{\text{true}} \\ &= (\text{H}[p_n(\theta)] - \mathbb{E}_{p_n(z)}[\text{H}[p_n(\theta|z)]]) - (\text{H}[p_n(\theta)] - \mathbb{E}_{p_{\text{train}}(z)}[\text{H}[p_n(\theta|z)]]) \\ &= \mathbb{E}_{p_{\text{train}}(z)}[\text{H}[p_n(\theta|z)]] - \mathbb{E}_{p_n(z)}[\text{H}[p_n(\theta|z)]]\,. \end{aligned}$$

The second is the error from approximating the infinite-step information gain in z with the expected information gain in θ:

$$\begin{aligned} \sqrt{\varepsilon_z} &= \text{EIG}_\theta - \text{IG}_z(y_{1:\infty}^+) \\ &= (\text{H}[p_n(z)] - \mathbb{E}_{p_n(\theta)}[\text{H}[p_n(z|\theta)]]) - (\text{H}[p_n(z)] - \text{H}[p_{\text{train}}(z)]) \\ &= \text{H}[p_{\text{train}}(z)] - \mathbb{E}_{p_n(\theta)}[\text{H}[p_n(z|\theta)]]\,. \end{aligned}$$

We average over 50 repeats with different random-number seeds.