Deep Ensembles for Epistemic Uncertainty: A Frequentist Perspective

Anchit Jain
EECS, MIT
Cambridge MA, USA
ajain625@mit.edu

Stephen Bates
EECS, MIT
Cambridge MA, USA
s_bates@mit.edu

Abstract

Decomposing prediction uncertainty into aleatoric (irreducible) and epistemic (reducible) components is critical for the reliable deployment of machine learning systems. While the mutual information between the response variable and model parameters is a principled measure for epistemic uncertainty, it requires access to the parameter posterior, which is computationally challenging to approximate. Consequently, practitioners often rely on probabilistic predictions from deep ensembles to quantify uncertainty, which have demonstrated strong empirical performance. However, a theoretical understanding of their success from a frequentist perspective remains limited. We address this gap by first considering a bootstrap-based estimator for epistemic uncertainty, which we prove is asymptotically correct. Next, we connect deep ensembles to the bootstrap estimator by decomposing it into data variability and training stochasticity; specifically, we show that deep ensembles capture the training stochasticity component. Through empirical studies, we show that this stochasticity component constitutes the majority of epistemic uncertainty, thereby explaining the effectiveness of deep ensembles.

1 Introduction

While machine learning systems are increasingly accurate in tightly-controlled settings, a key challenge to their deployment in more complex, real-world environments is uncertainty awareness—algorithms should be able to accurately report on their level of confidence or reliability in different conditions. Moreover, there are different underlying reasons for uncertainty. Uncertainty may be due to inherent unpredictability in a population with the current feature set (aleatoric or irreducible uncertainty) or instead due to limited training data (epistemic or reducible uncertainty). Distinguishing these types of uncertainty is important in order to take the correct downstream action. High aleatoric uncertainty indicates that the current prediction task is inherently challenging, and we cannot have high confidence in any one prediction unless we are able to collect richer data (not just more data). On the other hand, high epistemic uncertainty signals that collecting more training data in this region is likely to improve the model, which makes it suitable for guiding model development and active learning [18, 10, 50]. We focus on this quantity in this work.

A popular measure of epistemic uncertainty is the mutual information (MI) between model parameters and the target label which quantifies the expected reduction in uncertainty from observing new data [18, 5]. MI is a fundamentally Bayesian quantity that requires access to the posterior distribution over model parameters, which generally renders its exact calculation intractable. As such, substantial effort has gone into developing approximate Bayesian computational techniques [38, 12, 55, 16, 9], but they are often difficult to scale, compromise predictive accuracy or have weak Bayesian justification.

In light of these challenges, practitioners often use deep ensembles [35] to obtain uncertainty estimates. Deep ensembles proceed by generating an ensemble of models by only varying the

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

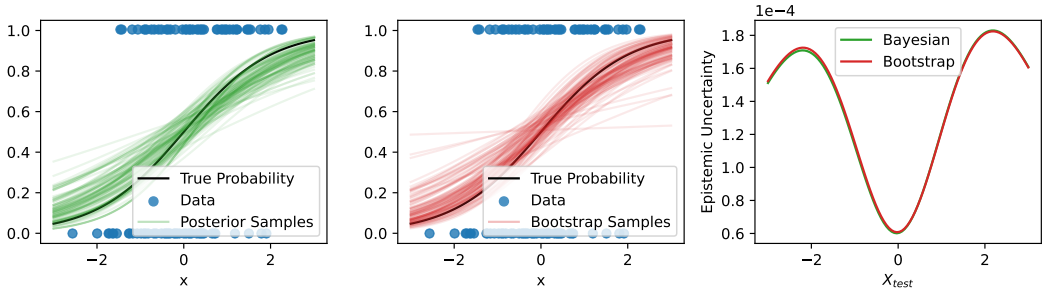

Figure 1: **Teaser Simulation:** *(left)* Training data, true label probability and posterior sample predictions. *(center)* Predictions from MLEs over bootstrapped datasets. *(right)* Epistemic uncertainty calculated through Bayesian inference (MCMC) versus the bootstrap estimate.

randomization in the training procedure (e.g. initial weights, data shuffling etc.). Previous works have noted that probabilistic predictions from the deep ensemble reliably quantify uncertainty and exhibit strong performance on downstream tasks [35, 41, 1]. While Bayesian interpretations have been proposed [56], a theoretical frequentist understanding behind the success of deep ensembles is lacking. Furthermore, various sources such as limited data, misspecified models or noisy optimization contribute to the uncertainty [20] and estimates may only account for some of these. Therefore, to correctly use an uncertainty estimate, it is crucial to understand the sources it captures. Our work aims to provide a frequentist perspective on the epistemic uncertainty captured by deep ensembles.

Our analysis proceeds in three parts. Firstly, we provide an asymptotic expansion for MI which links it to the Fisher information and shows how epistemic uncertainty differs from frequentist variance. We use this to motivate the construction of a bootstrap estimator which targets the MI. Secondly, since the epistemic uncertainty within a deep ensemble arises only due to the randomness in the training procedure, we decompose the bootstrap estimator into components arising due to data sampling and training randomness to facilitate comparison. Finally, through empirical studies, we note that deep ensembles capture the portion of epistemic uncertainty due to stochastic optimization, which in turn appears to be the majority component, lending theoretical justification to their success from a frequentist perspective. Our key contributions are—

- **Link between mutual information and frequentist Fisher information:** We provide an asymptotic expansion for MI and show how it depends not only on the variance in model predictions as measured by the Fisher information, but also on the degree of randomness in the true data generating process.
- **Contributions to MI from data variability and training randomness:** We use the asymptotic expansion to construct a bootstrap estimator of MI. To allow comparison with deep ensembles, we show how this bootstrap estimator separates into components arising from training stochasticity and variability in data sampling.
- **Empirical comparison with deep ensembles:** We conduct empirical studies across a variety of deep learning tasks that show that training stochasticity is dominant over variability in data sampling, with deep ensembles capturing the latter.

1.1 Illustration

To provide the key intuition motivating the bootstrap estimator, we provide the following simulation. Consider a binary logistic regression model, $\mathbb{P}(y = 1|x, \theta) = 1/(1 + e^{-\theta_1 x - \theta_0})$. We sample the features x from a standard normal and also impose standard normal priors on θ_0, θ_1. Fig. 1*(left)* visualizes the data and the true model along with 100 posterior samples from the true Bayesian posterior, obtained using a Markov chain Monte Carlo (MCMC) algorithm. Next, we calculate maximum likelihood estimates (MLEs) for the parameters across 100 bootstrapped datasets and show these in Fig. 1*(middle)*—note the remarkable similarity in the posterior and bootstrap samples. Finally, we use these MCMC and bootstrap samples to form our estimate of the epistemic uncertainty as per Algorithm 1 noting the excellent agreement in Fig. 1*(right)*. We leverage this connection to study Bayesian MI using bootstrapping, which then allows a comparison with deep ensembles.

2 Background

Mutual Information as Epistemic Uncertainty: A popular uncertainty decomposition [37, 18, 5] leverages information-theoretic quantities to propose measures of aleatoric and epistemic uncertainty. Let the random variables Y and X denote the labels and features respectively in a supervised learning setup. Let $\mathcal{D}_n$ be the data comprising of n realizations of (Y, X). Further, let θ be the random variable denoting the parameters of a prediction model. Information theory uses the Shannon entropy $\mathrm{H}(Y)$ to measure uncertainty in Y, and defines mutual information as

$$\underbrace{\mathrm{I}(Y;\theta|X,\mathcal{D}_n)}_{\text{Epistemic}} = \underbrace{\mathrm{H}(Y|X,\mathcal{D}_n)}_{\text{Total}} - \underbrace{\mathrm{H}(Y|\theta,X,\mathcal{D}_n)}_{\text{Aleatoric}}. \tag{1}$$

The quantification of epistemic uncertainty as the mutual information (MI) between the (test) label and parameters is conceptually pleasing since Eq. 1 reveals it to be the reduction in label uncertainty from knowing the model parameters. This MI is a fundamentally Bayesian object since it requires θ to be random.

Bootstrap: The original bootstrap procedure [6] was proposed to estimate the variance of an estimator under different draws of the data. Consider a setting where there are no repeated data points. Bootstrap first samples weights from a symmetric multinomial distribution with all event probabilities $1/n$. These weights are used to reweight the original data to create a "bootstrapped dataset" over which the estimator is recomputed. This procedure is replicated B times to obtain B "bootstrapped estimates". The key intuition behind the bootstrap—the "plug-in" principle—proposes that the distribution of these bootstrapped estimates approximates the distribution of the original estimator under redraws of the data. As such, we can obtain an estimate of the variance of our estimator as the empirical variance of the bootstrapped estimates.
We can also instead draw weights from a symmetric Dirichlet distribution [45]. We hypothesize that in a deep learning setting, these weights are more suitable since it allows the neural network to still see all the training instances and hence capture the training data better than multinomial weights which omit training points entirely. For our paper, we thus use Dirichlet weights for our experiments but emphasize that our theorems hold for multinomial weights too (Appendix A).

Deep Ensembles: Deep ensembles [35] are a simple but powerful uncertainty quantification technique for deep neural networks (DNNs). A deep ensemble is obtained by training multiple models with the exact same architecture and training data but with different random seeds controlling stochasticity in the training procedure (different parameter initializations, different data shuffling when using stochastic gradient descent etc.). The deep ensemble model prediction is simply the average of predictions of member models. They have shown success at a variety of uncertainty quantification tasks [35, 41, 1].

3 Algorithm and Theoretical Results

3.1 Problem Setup

We now specialize the previous notation to our exact setting. Consider a K-class supervised classification problem with labels $Y \in \{1, ..., K\}$. Denote the training data by $\mathcal{D}_n = \{(X_i, Y_i), i = 1, ..., n\}$. We wish to evaluate the epistemic uncertainty in the prediction for the label Y_{test} given the feature X_{test}. We denote our probability predictions for Y_{test} for a model parametrized by $\theta \in \Theta$ by $\hat{p}(X_{\text{test}};\theta)$. Further, denote by $\hat{p}_k(X_{\text{test}};\theta)$ the prediction probability assigned to the kth class. The model is specified such that $0 < \hat{p}_k(X_{\text{test}};\theta) < 1$ and $\sum_{k=1}^{K} \hat{p}_k(X_{\text{test}};\theta) = 1$. Finally, we also assume that the model is well-specified, i.e., there exists a $\theta_0 \in \Theta$ such that $(X_i, Y_i) \overset{iid}{\sim} P_X(x)\hat{p}_y(x;\theta_0) =: P_{\theta_0}$, where P_X denotes the marginal distribution of X.

In the Bayesian setup, we have a prior $p(\theta)$ over the parameters and we denote by $p(\theta|\mathcal{D}_n)$ our posterior. We use the notation $\mathbb{E}_Z[.]$ to denote expectations taken with respect to the random variable Z. As per Sec. 2, the epistemic uncertainty is

$$\begin{aligned}\mathrm{I}\left(Y_{\text{test}};\theta|X_{\text{test}},\mathcal{D}_n\right) &= \mathrm{H}\left(Y_{\text{test}}|X_{\text{test}},\mathcal{D}_n\right) - \mathrm{H}\left(Y_{\text{test}}|\theta,X_{\text{test}},\mathcal{D}_n\right) \\ &= \mathrm{H}\left(\mathbb{E}_{\theta|\mathcal{D}_n}\left[\hat{p}(X_{\text{test}};\theta)\right]\right) - \mathbb{E}_{\theta|\mathcal{D}_n}\left[\mathrm{H}(\hat{p}(X_{\text{test}};\theta))\right].\end{aligned}$$

3.2 Asymptotics of Mutual Information

We begin by deriving the asymptotic behavior of mutual information, connecting it with the Fisher information and frequentist sampling variation. This connection will allow us to repurpose frequentist methods for estimating variation to instead target the mutual information.

Theorem 3.1. *Under the assumptions needed for Bernstein von-Mises to hold and uniform integrability of the predicted probabilities (1-7 in Appendix A) we have the asymptotic expansion*

$$\mathrm{I}(Y_{\mathrm{test}};\theta|X_{\mathrm{test}},\mathcal{D}_n) = \frac{1}{2n}\sum_{k=1}^{K}\frac{\sigma_k^2}{\hat{p}_k(X_{\mathrm{test}};\theta_0)} + o_p(n^{-1}) \tag{2}$$

$\forall\ X_{\mathrm{test}}$*, where* σ_k^2 *is*

$$\sigma_k^2 = \left[\frac{\partial\hat{p}(X_{\mathrm{test}};\theta_0)}{\partial\theta}^T \mathcal{I}^{-1}(\theta_0)\frac{\partial\hat{p}(X_{\mathrm{test}};\theta_0)}{\partial\theta}\right]_{k,k}$$

for $\mathcal{I}^{-1}(\theta_0)$ *being the inverse Fisher information of the model at* θ_0*.*

Proof Sketch: Our analysis proceeds by Taylor expanding the entropy function about the mean posterior predicted probability. We then apply the functional delta method to the Bernstein von-Mises theorem to obtain the limiting law of our predictions. Recall that the Bernstein von-Mises theorem holds under correct model specification, sufficient smoothness, identifiability, and feasibility of θ_0 under the prior (e.g. Theorem 10.1 of van der Vaart (2000)). Finally, we use uniform integrability to obtain the limiting posterior moments. We defer the full proof to Appendix A.

We briefly digress to comment on the theorem and provide better intuition for epistemic uncertainty as measured by mutual information. Firstly, the theorem shows that mutual information decreases linearly in the number of samples. Next, it connects mutual information to the Fisher information and perhaps surprisingly shows how the first order asymptotics are governed by the inverse Fisher information divided by the true conditional probabilities.

Better intuition can be gained into the first order asymptotic term by considering the binary classification case in which the first order term simplifies to

$$\frac{\left(\partial\hat{p}_1(X_{\mathrm{test}};\theta_0)/\partial\theta\right)^T\left(n^{-1}\mathcal{I}^{-1}(\theta_0)\right)\left(\partial\hat{p}_1(X_{\mathrm{test}};\theta_0)/\partial\theta\right)}{2\hat{p}_1(X_{\mathrm{test}};\theta_0)(1-\hat{p}_1(X_{\mathrm{test}};\theta_0))}$$

as noted in MacKay (1992). Recall that $n^{-1}\mathcal{I}^{-1}(\theta_0)$ is the (asymptotic) variance of the MLE. The numerator in the expression above can then be interpreted as the variance of the model prediction at X_{test}. $1/2p(1-p)$ achieves a minimum at $1/2$ and tends to infinity at 0 and 1. The first order term is then directly proportional to the model variance but also inversely proportional to the ground truth certainty (i.e. how close $\hat{p}_1(X_{\mathrm{test}};\theta_0)$ is to 0 or 1). Thus, we see that this first order term matches the intuition behind epistemic uncertainty; if our predictions are uncertain but the true outcome is certain then we can expect to reduce our uncertainty and hence have high epistemic uncertainty.

3.3 A Bootstrap Estimator

Theorem 3.1 suggests that asymptotically correct estimation of the Fisher information and true model probabilities can thus be used to construct first order accurate estimators of mutual information. In light of this, we consider a bootstrap based estimator which we denote by $\mathrm{I}_b(X_{\mathrm{test}},\mathcal{D}_n)$.
Estimation proceeds by first sampling weights ξ from the symmetric Dirichlet with concentration parameter 1. Denote the MLE of θ over the bootstrapped dataset as $\hat{\theta}_b$ (for example, as obtained by minimizing cross-entropy loss). We denote the distribution of this estimate (induced by the randomness in the weights) given the data $\mathcal{D}_n$ as $p(\hat{\theta}_b|\mathcal{D}_n)$. Then,

$$\mathrm{I}_b(X_{\mathrm{test}},\mathcal{D}_n) := \mathrm{H}(\mathbb{E}_{\hat{\theta}_b|\mathcal{D}_n}[\hat{p}(X_{\mathrm{test}};\hat{\theta}_b)]) - \mathbb{E}_{\hat{\theta}_b|\mathcal{D}_n}[\mathrm{H}(\hat{p}(X_{\mathrm{test}};\hat{\theta}_b))].$$

Algorithm 1 makes this procedure explicit for clarity.

Theorem 3.2. *Under the assumptions needed for Bernstein von-Mises to hold, for asymptotic normality of the MLE to hold and uniform integrability of the predicted probabilities (1-12 in Appendix A), we have*

$$\frac{\mathrm{I}(Y_{\mathrm{test}};\theta|X_{\mathrm{test}},\mathcal{D}_n)}{\mathrm{I}_b(X_{\mathrm{test}},\mathcal{D}_n)} \xrightarrow{P_{\theta_0}^n} 1 \qquad \forall\ X_{\mathrm{test}}.$$

Algorithm 1 $\mathrm{I}_b(X_{\text{test}}, \mathcal{D}_n)$ estimation

Input: Number of bootstrap replications B, training dataset $\mathcal{D}_n$
for $b \leftarrow 1$ **to** B **do**
  sample $\xi \sim \mathrm{Dir}(1, ..., 1)$
  $\hat{\theta}_b \leftarrow \mathrm{argmax}_{\theta \in \Theta} \sum_{i=1}^{n} \xi_i \log(\hat{p}_{Y_i}(X_i; \theta))$
end for
$\mathrm{I}_b(X_{\text{test}}, \mathcal{D}_n) = \mathrm{H}\left(\frac{1}{B}\sum_{b=1}^{B} \hat{p}(X_{\text{test}}; \hat{\theta}_b)\right) - \frac{1}{B}\sum_{b=1}^{B} \mathrm{H}\left(\hat{p}(X_{\text{test}}; \hat{\theta}_b)\right)$
Output: $\mathrm{I}_b(X_{\text{test}}, \mathcal{D}_n)$

Proof Sketch: We proceed as in Theorem 3.1 but invoke bootstrap limit laws (e.g. Theorem 10.16 of Kosorok (2007)) instead of Bernstein von-Mises to obtain an identical asymptotic expansion for our estimator. We provide the complete proof in Appendix A.

We note that our theorem also holds if we use the classic multinomial weights for bootstrap instead of Dirichlet ones—see Appendix A. Theorem 3.2 proves the asymptotic validity of the bootstrap estimator and provides a frequentist path to studying the Bayesian MI.

3.4 Distinguishing Algorithmic Randomness from Statistical Uncertainty

Deep neural networks are generally trained with variants of stochastic gradient descent and a random parameter initialization. As such, the fitted model is random even with the training data fixed. Deep ensembles exploit this phenomenon, effectively capturing only the epistemic uncertainty arising from training stochasticity. The bootstrap estimator in Sec. 3.3 provides the crucial conceptual link between the ad-hoc operational technique of deep ensembles and principled Bayesian inference. While standard Bayesian formulations cannot account for stochasticity in model training, in this section we show how the bootstrap estimator naturally decomposes into distinct terms for data variability and training stochasticity, providing a mechanism to isolate the latter. This decomposition will then allow us to study deep ensembles by directly comparing them to the training stochasticity component of a theoretically grounded estimator.

In more detail, the parameter estimate can be viewed as a function of the data and the random seed. Thus, for the remainder of this section, we use the notation $\hat{p}(X_{\text{test}}; \hat{\theta}_b) = \hat{p}(\mathcal{D}_n^b, s)$ where $\mathcal{D}_n^b$ denotes the bootstrapped dataset and s denotes the seed used for training the DNN. Since $\hat{\theta}_b$ is deterministic given $(\mathcal{D}_n^b, s)$, $\hat{p}(\mathcal{D}_n^b, s)$ is a deterministic function. Normally, the selection of s is completely independent of everything else and hence $(s|\mathcal{D}_n^b, \mathcal{D}_n) \overset{d}{=} s$.

With this notation in place, we propose the following decomposition of the bootstrap epistemic uncertainty estimator

$$
\begin{aligned}
&\mathrm{I}_b(X_{\text{test}}, \mathcal{D}_n) \\
=&\mathrm{H}(\mathbb{E}_{(\mathcal{D}_n^b, s)|\mathcal{D}_n}[\hat{p}(\mathcal{D}_n^b, s)]) \pm \mathbb{E}_{\mathcal{D}_n^b|\mathcal{D}_n}[\mathrm{H}(\mathbb{E}_s[\hat{p}(\mathcal{D}_n^b, s)])] - \mathbb{E}_{(\mathcal{D}_n^b, s)|\mathcal{D}_n}[\mathrm{H}(\hat{p}(\mathcal{D}_n^b, s))] \\
=&\underbrace{\mathrm{H}(\mathbb{E}_{\mathcal{D}_n^b|\mathcal{D}_n}[\mathbb{E}_s[\hat{p}(\mathcal{D}_n^b, s)]]) - \mathbb{E}_{\mathcal{D}_n^b|\mathcal{D}_n}[\mathrm{H}(\mathbb{E}_s[\hat{p}(\mathcal{D}_n^b, s)])]}_{=:\mathrm{I}_b^{\text{resampling}}} + \underbrace{\mathbb{E}_{\mathcal{D}_n^b|\mathcal{D}_n}\left[\mathrm{H}(\mathbb{E}_s[\hat{p}(\mathcal{D}_n^b, s)]) - \mathbb{E}_s[\mathrm{H}(\hat{p}(\mathcal{D}_n^b, s))]\right]}_{=:\mathrm{I}_b^{\text{seeds}}}.
\end{aligned}
\tag{3}
$$

$\mathrm{I}_b^{\text{resampling}}$ can be viewed as an estimator of the epistemic uncertainty arising due to the randomness in the data sampling process. Intuitively, this represents how much the model prediction would change if different training data had been sampled from the true distribution. On the other hand, $\mathrm{I}_b^{\text{seeds}}$ captures the amount of the entropy difference explained by the variation in the random seeds. That is, this term represents the amount of variation due to the algorithmic randomness. Thus, our proposed estimator can be decomposed into a 'data sampling' component and a 'training randomness' component.

Next, notice that the estimate of the mutual information from deep ensembles is

$$
\mathrm{I}^{\text{deep ensemble}} := \mathrm{H}(\mathbb{E}_s[\hat{p}(\mathcal{D}_n, s)]) - \mathbb{E}_s[\mathrm{H}(\hat{p}(\mathcal{D}_n, s))]. \tag{4}
$$

This is parallel to the term $\mathrm{I}_b^{\text{seeds}}$ above, but with the full data instead of a bootstrap subsample. Thus, the bootstrap estimator above can be viewed as approximately the MI from random seeds plus an

additional term capturing variability in the dataset. In the experiments, we will see that the latter is often smaller than the former. In this way, $\mathrm{I}_b(X_{\mathrm{test}}, \mathcal{D}_n) \approx \mathrm{I}^{\text{deep ensemble}}$. This link gives a new frequentist motivation for deep ensembles. More precisely, since Sec. 3.3 proved that the bootstrap estimator is asymptotically correct for the true Bayesian target, the above relation indicates how deep ensembles theoretically succeed in capturing the Bayesian MI.

4 Experiments

Section 3.4 gives us a way to measure contributions to epistemic uncertainty from sampling variability in the training data and from stochasticity in the training procedure; as well as a measure for the epistemic uncertainty in a deep ensemble. Our empirical investigations now address two questions—

1. Does training randomness ($\mathrm{I}_b^{\text{seeds}}$ in eq. 3) contribute more to epistemic uncertainty than data sampling ($\mathrm{I}_b^{\text{resampling}}$ in eq. 3)?
2. Do deep ensembles ($\mathrm{I}^{\text{deep ensemble}}$ in eq. 4) capture the training randomness ($\mathrm{I}_b^{\text{seeds}}$)?

4.1 Experimental Setup

We will use the same experimental setup across various deep learning classification tasks. We first train a deep ensemble of B models on the entire training dataset. We use this to estimate $\mathrm{I}^{\text{deep ensemble}}$ for each test point by approximating expectations with empirical averages in eq. 4. Next, we construct B bootstrapped versions of the training dataset and train a deep ensemble of size B on each. We use these deep ensembles to similarly estimate $\mathrm{I}_b^{\text{seeds}}$ and $\mathrm{I}_b^{\text{resampling}}$ for each test point.

4.2 Evaluation Metrics

To answer the first question, we look at the slope of the line of best fit (m when $y = mx$) that regresses $\mathrm{I}_b^{\text{resampling}}$ on $\mathrm{I}_b^{\text{seeds}}$; we shall refer to this slope as $m_{\text{seeds}}^{\text{resampling}}$. Fig. 2 visualizes this line in red. This tells us how the magnitude of $\mathrm{I}_b^{\text{resampling}}$ compares to $\mathrm{I}_b^{\text{seeds}}$ on average. For the second question, we again look at the slope of the line of best fit that regresses $\mathrm{I}^{\text{deep ensemble}}$ on $\mathrm{I}_b^{\text{seeds}}$ to compare the relative magnitudes of the two; we shall refer to this slope as $m_{\text{seeds}}^{\text{deep ensemble}}$. Fig. 3 *(right)* visualizes this line in orange. We also look at the Spearman rank correlation between $\mathrm{I}^{\text{deep ensemble}}$ and $\mathrm{I}_b^{\text{seeds}}$ to evaluate if points with high relative epistemic uncertainty due to training stochasticity also exhibit high relative epistemic uncertainty when we use deep ensembles; we shall refer to this correlation as ρ. This is particularly insightful since performance on many downstream tasks (such as AUC-ROC in outlier detection tasks) often only depends on the ordering of test points rather the explicit magnitudes of the uncertainty estimates.

4.3 Deep Learning Tasks

To lend robustness to our empirical findings, we study performance across four tasks spanning various data domains (natural images, medical images, text), different dataset sizes (1000 classes with $>$ 1M training examples, 2 classes with 1000 training examples) and different deep learning regimes (training from scratch, fine-tuning, transfer learning). For each task, we defer hyperparameter details and accuracy metrics to Appendix B.1.

CIFAR-10: We train a ResNet-18 [15] from scratch on the CIFAR-10 dataset [33] which consists of 50,000 training examples spanning 10 classes of natural images. We use ensembles of size 25.

ImageNet: We train a ResNet-18 [15] from scratch on the ImageNet dataset [4] which consists of 1.28M training examples spanning 1000 classes of natural images. Due to computational limitations, we use ensembles of size 10.

ISIC Skin Lesion Dataset: The ISIC (International Skin Imaging Collaboration) 2016 skin lesion challenge [13] consists of a dataset of images of skin lesions. There are only 900 training images spanning two classes (benign and malignant) in this dataset. For such small datasets, transfer learning [42] is a common approach wherein a pretrained model from a similar task is used to extract features for the task of interest. More precisely, we use a pretrained ResNet-50 model from He et al. (2015) to

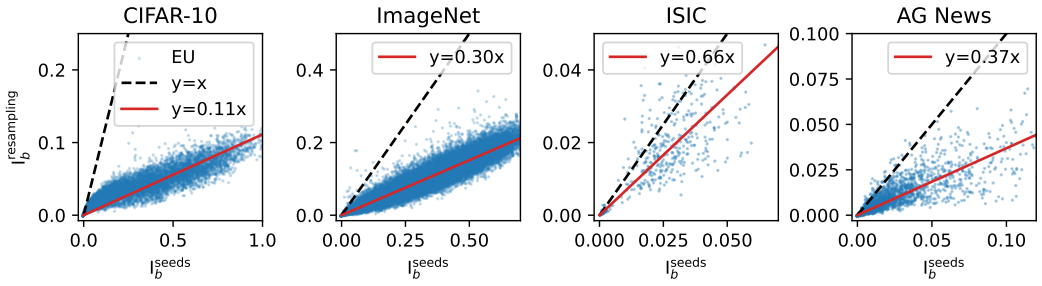

Figure 2: **Uncertainty Components of Bootstrap Estimator:** Scatter plot of epistemic uncertainty due to training stochasticity (I_b^{seeds}) versus epistemic uncertainty due to data resampling ($I_b^{resampling}$) along with line of best fit in red. See Sec. 4.2 for details.

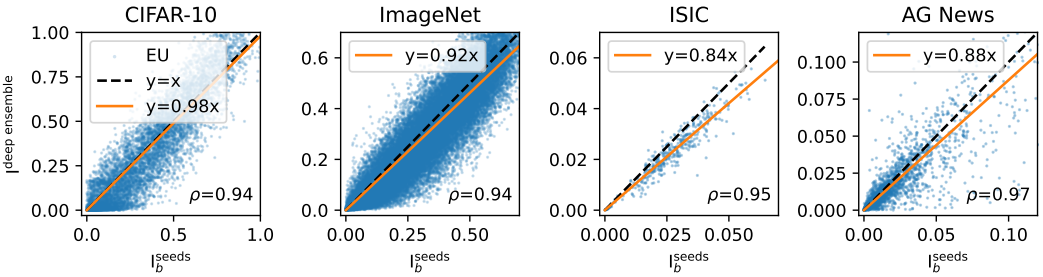

Figure 3: **Uncertainty due to Training Stochasticity:** Scatter plot of epistemic uncertainty due to training stochasticity (I_b^{seeds}) versus deep ensemble estimate of epistemic uncertainty ($I^{deep\ ensemble}$) along with line of best fit (orange) and spearman correlation (ρ). See Sec. 4.2 for details.

extract 2048-dimensional feature vectors. We then train a two layer MLP from random initialization on top of these feature vectors to perform binary classification. We use ensembles of size 25.

AG News Dataset: The AG news topic classification dataset [59] consists of 120,000 news articles (titles and descriptions) spanning 4 categories. As is common in natural language processing (NLP), we finetune a pretrained model to perform classification. Specifically, we finetune the TinyBERT [24] model which has a transformer architecture (14.5M parameters) for classification by adding a linear layer (initialized randomly) on top of the transformer layers to perform classification. We use ensembles of size 25.

4.4 Results

Fig. 2 visualizes scatter plots of I_b^{seeds} versus $I_b^{resampling}$ along with the line of best fit (slope = $m_{seeds}^{resampling}$) with numerical values of the slopes in Table 1. I_b^{seeds} always dominates over $I_b^{resampling}$, with the latter being just 11% of the former on average for CIFAR-10 for example. Therefore, training randomness contributes more to epistemic uncertainty than variability in data sampling. From a model development perspective, we believe that this suggests that if practitioners want to reduce their prediction uncertainty, it may be more fruitful to train a bigger deep ensemble to target the uncertainty due to stochastic training, than to collect more data to target the uncertainty due to variability in data sampling.

Fig. 3 visualizes scatter plots of I_b^{seeds} versus $I^{deep\ ensemble}$ along with the line of best fit (slope = $m_{seeds}^{deep\ ensemble}$) with numerical values of the slopes and correlation in Table 1. I_b^{seeds} and $I^{deep\ ensemble}$ have similar magnitudes (around 90% of each other) and correlate well ($\rho \approx 0.95$). Therefore, deep ensembles do indeed capture the uncertainty due to training randomness. Together, these observations allow us to lend theoretical backing to deep ensembles since they successfully approximate the majority portion (training randomness) of an asymptotically correct estimator ($I_b(X_{\text{test}}, \mathcal{D}_n)$) of the Bayesian epistemic uncertainty.

DATASET	$m_{\text{SEEDS}}^{\text{RESAMPLING}}$	$m_{\text{SEEDS}}^{\text{DEEP ENSEMBLE}}$	ρ
CIFAR-10	0.11	0.98	0.94
IMAGENET	0.30	0.92	0.94
ISIC	0.66	0.84	0.95
AG NEWS	0.37	0.88	0.97

Table 1: Slope of line of best fit from $\mathrm{I}_b^{\text{resampling}}$ on $\mathrm{I}_b^{\text{seeds}}$ ($m_{\text{seeds}}^{\text{resampling}}$), from $\mathrm{I}^{\text{deep ensemble}}$ on $\mathrm{I}_b^{\text{seeds}}$ ($m_{\text{seeds}}^{\text{deep ensemble}}$) and spearman correlation between $\mathrm{I}^{\text{deep ensemble}}$ and $\mathrm{I}_b^{\text{seeds}}$ (ρ). See Sec. 4.2 for details.

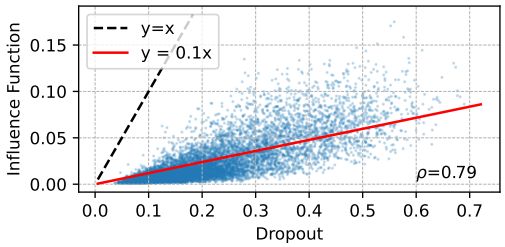

Figure 4: **Influence Function Approximation:** Estimates of epistemic uncertainty using Dropout, versus using influence functions to approximate predictions from bootstrapped datasets. See Sec. 6.2 for details.

5 Related Work

5.1 Uncertainty Decomposition

Mutual information as a measure of epistemic uncertainty has demonstrated success in active learning [18, 10], out-of-distribution and misclassification detection tasks [32]. Beyond the specific information theoretic uncertainty decomposition considered in our work, a variety of other ways of quantifying epistemic uncertainty have also been proposed [47, 27, 17, 34, 46, 51] and MI does have drawbacks [57]. See Hüllermeier and Waegeman (2021) for a comprehensive review of the topic. Indeed, epistemic and aleatoric uncertainty are often not cleanly separable and take on different meanings in different contexts [29]. We use MI for our investigations because of its widespread use, strong conceptual foundation and practical success; our work is not intended to address the optimality of MI as an epistemic uncertainty assessment.

Similar to our work, Huang et al. (2023) and Neal et al. (2019) also decompose the variance in model predictions for regression tasks into data variability and training randomness components. However, quite different to our contribution, Huang et al. (2023) used this decomposition to study ways of theoretically eliminating training randomness using the NTK approximation [23], while Neal et al. (2019) used this decomposition to investigate the bias-variance tradeoff.

5.2 Bayesian and Frequentist Equivalence

Rubin (1981) considers bootstrap from a Bayesian lens, although this interpretation is hard to reconcile with traditional Bayesian inference since it considers priors over the data rather than the parameters. Closest to our theoretical results is the work by Newton and Raftery (1994) which studied the use of bootstrap as a way of generating samples from a Bayesian posterior. Our work expands upon this idea by specifically leveraging this connection to connect frequentist and Bayesian notions of uncertainty. Moreover, we operationalize this idea in a modern context with DNNs and use it to study deep ensembles. Finally, Fong et al. (2023) also connects Bayesian parameter uncertainty to frequentist uncertainty arising due to lack of data. While their work conceptually shares connections to ours, it considers different practical applications.

5.3 Deep Ensembles

The uncertainty quantification capabilities of deep ensembles [35] have been studied in depth [41, 1]. Broadly speaking, these studies observed that deep ensembles have well calibrated probabilistic forecasts both in and out-of-domain, and exhibit strong performance on downstream tasks such as out-of-distribution detection and misclassification detection.

Several reasons behind their success have been put forth [8], including Bayesian ones [56]. These works primarily note that members of a deep ensemble capture different modes of the loss landscape and can be seen as high quality samples from the posterior. Our work is complementary to this perspective and provides a frequentist understanding of what uncertainties deep ensembles capture.

5.4 Bootstrapping for Uncertainties in Deep Learning

The idea of using bootstrap to estimate uncertainties in machine learning has been explored by several works in different contexts. Jürgens et al. (2024) investigates the suitability of evidential deep learning [48] for uncertainty quantification by comparing them to the bootstrapped estimates of the model parameters, which they believe accurately capture uncertainty in model predictions due to randomness in the training data. Jiménez et al. (2025) investigates variance based measures of uncertainty in a regression setting to also note that deep ensembles only capture uncertainty due to training variability which often dominates over uncertainty due to randomness in training data, similar to our findings in Sec. 4. Lastly, Shaker and Hüllermeier (2020) have also considered approximating the posterior distribution by the trees in a random forest, which can be seen as bootstrap estimates, for the purposes of mutual information estimation as in our work.

6 Discussion

6.1 Limitations

Theorem 3.2 only proves the asymptotic validity of the bootstrap estimator which can weaken the theoretical link between deep ensembles and Bayesian inference in low data regimes. Furthermore, our theorems also rely on assumptions that can be unrealistic in practical settings. In particular, correct specification of the model and identifiability of parameters can certainly appear tenuous for DNNs. Finally, while we aimed to cover a wide range of deep learning settings in our empirical studies, we recognize that more evaluation can always be carried out to lend robustness to our conclusions. We do provide a few additional ablations against dataset size and choice of bootstrap weights in Appendix B.3 to lend further evidence towards the robustness of our qualitative conclusions.

6.2 Future Directions

6.2.1 Data Attribution Methods

We identify an interesting link between the bootstrap estimator in Sec. 3.3 and data attribution methods such as influence functions [14, 30, 11] and datamodels [21, 43] which aim to identify training data points most responsible for a model's predictions. These methods allow for estimation of predictions under perturbation of training data and can therefore be used to approximate the predictions of a model trained on bootstrapped data needed for our proposed measure. Any data attribution algorithm that yields a mapping from training data to model predictions can be leveraged to estimate the variance terms in (2), and hence give epistemic uncertainty estimates. We investigate the performance of such an MI estimate based on influence functions; see Appendix B.2 for a similar experiment with datamodels and hyperparameter details.

We train a single convolutional network called the LeNet-5 on MNIST data [36]. Next, we calculate the influence functions for each data point as described in Koh and Liang (2017) for the model parameters in the last two fully connected layers. For each data point, we multiply the influence function by the change in the weighting under bootstrap and add all these products to the parameters to obtain an approximation for model parameters under a bootstrap reweighting. We obtain predictions from 100 such approximated models and use Algorithm 1 to obtain MI estimates. To obtain a Bayesian estimate of MI, we use the Bayesian interpretation of Monte Carlo Dropout [9] wherein Dropout is enabled at inference time and stochastic forward passes are performed to obtain samples from the Bayesian predictive distribution.

Fig. 4 shows the influence function approximation and Dropout estimates. While there is correlation, influence function estimates are an order of magnitude smaller. We hypothesize that this is because the influence function approximation only uses a single trained model and hence fails to capture the randomness due to training stochasticity. Nonetheless, our experiment serves as proof-of-concept for studying "uncertainty attribution"—identifying data points or modalities responsible for model uncertainty—by estimating their influence function. Since the variance of an estimator is equal to variance of its influence function [53], we can note that training data points with high influence on a test prediction drive up the variance of the influence function and hence lead to higher epistemic uncertainty. We hope to explore this more in a future work.

6.2.2 Laplace Approximation

We now note the connection between the Laplace approximation for Bayesian inference [38] and the asymptotic expansion for mutual information in Theorem 3.1. The Laplace approximation models the posterior $p(\theta|\mathcal{D}_n)$ as a Gaussian $\mathcal{N}(\hat{\theta}_{\text{MAP}}, \Sigma_{\text{Laplace}})$ centered at the maximum a posteriori (MAP) estimate $\hat{\theta}_{\text{MAP}}$, with precision matrix $\Sigma_{\text{Laplace}}^{-1}$ given by the negative Hessian of the log-posterior (the observed Fisher information plus the prior Hessian) evaluated at $\hat{\theta}_{\text{MAP}}$. Under suitable regularity conditions, the prior contribution is $O(1)$ while the normalized observed Fisher information converges to $\mathcal{I}(\theta_0)$, so that the Laplace posterior covariance Σ_{Laplace} directly estimates the asymptotic covariance $n^{-1}\mathcal{I}^{-1}(\theta_0)$ in eq. 2. Moreover, in finite samples where the empirical Fisher information may be ill-conditioned, the prior precision in $\Sigma_{\text{Laplace}}^{-1}$ provides natural regularization without altering the $O(n^{-1})$ asymptotics.

Consequently, when using the Laplace approximation, eq. 2 can be used to directly provide an estimate of the epistemic uncertainty without explicit sampling from the posterior by evaluating $\hat{p}_k$ and its gradients at $\hat{\theta}_{\text{MAP}}$ and substituting Σ_{Laplace} for $n^{-1}\mathcal{I}^{-1}(\theta_0)$. Empirically investigating the practicality of this idea and leveraging developments in computationally feasible Laplace approximations for deep learning [44, 22, 3] to provide plug-in estimators of epistemic uncertainty is a promising future avenue.

7 Conclusion

Our paper aims to investigate the uncertainty quantification capabilities of deep ensembles from a frequentist perspective. Specifically, we study the epistemic uncertainty captured by deep ensembles, as measured by the mutual information between the model parameters and the data label. We derive an asymptotic expansion for this mutual information and use it to motivate the construction of an asymptotically correct bootstrap based estimator. Bayesian formulations cannot account for randomness in model parameter selection, while deep ensembles exclusively make use of this randomness. To bridge this gap, we make use of the bootstrap estimator—we show how the bootstrap estimate can be seen as an additive composition of the uncertainty from training randomness and uncertainty due to data sampling. We then conduct empirical studies which establish that training randomness dominates over data sampling, and that deep ensemble estimates of epistemic uncertainty closely capture the training randomness. Therefore, our investigations provide a theoretical frequentist justification to deep ensembles by noting that they approximate the dominant contribution to an asymptotically correct estimate of the true Bayesian epistemic uncertainty.

Acknowledgments

The authors thank Willem Waegeman, Mira Jürgens and Sebastián Jiménez for their insightful comments and pointers to relevant literature. The authors would also like to thank Dan Kluger for his kind help and insights into the proofs. Finally, the authors would also like to thank the anonymous reviewers for their feedback which helped improve this paper.

Funding Transparency and Competing Interests Statement

This research was, in part, funded by the Advanced Research Projects Agency for Health (ARPA-H). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the United States Government. A.J. was also partly supported by the Michael Athans Fellowship. The authors declare no competing interests.

References

[1] Arsenii Ashukha, Alexander Lyzhov, Dmitry Molchanov, and Dmitry Vetrov. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. In *International Conference on Learning Representations*, 2020.

[2] Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter. A downsampled variant of imagenet as an alternative to the cifar datasets. *arXiv preprint arXiv:1707.08819*, 2017.

[3] Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux - effortless bayesian deep learning. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2009.

[5] Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *International conference on machine learning*. PMLR, 2018.

[6] Brad Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1), 1979.

[7] Edwin Fong, Chris Holmes, and Stephen G Walker. Martingale posterior distributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(5), 2023.

[8] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019.

[9] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*. PMLR, 2016.

[10] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International conference on machine learning*. PMLR, 2017.

[11] Ryan Giordano, William Stephenson, Runjing Liu, Michael Jordan, and Tamara Broderick. A swiss army infinitesimal jackknife. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*. PMLR, 16–18 Apr 2019.

[12] Alex Graves. Practical variational inference for neural networks. *Advances in neural information processing systems*, 24, 2011.

[13] David Gutman, Noel C. F. Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic), 2016.

[14] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, and W.A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. Wiley Series in Probability and Statistics. Wiley, 2011. ISBN 9781118150689.

[15] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[16] Jose Miguel Hernandez-Lobato and Ryan Adams. Probabilistic backpropagation for scalable learning of bayesian neural networks. In *International conference on machine learning*. PMLR, 2015.

[17] Paul Hofman, Yusuf Sale, and Eyke Hüllermeier. Quantifying aleatoric and epistemic uncertainty: A credal approach. In *ICML 2024 Workshop on Structured Probabilistic Inference {\&} Generative Modeling*, 2024.

[18] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

[19] Ziyi Huang, Henry Lam, and Haofeng Zhang. Efficient uncertainty quantification and reduction for over-parameterized neural networks. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36. Curran Associates, Inc., 2023.

[20] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110(3), 2021. ISSN 1573-0565.

[21] Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. Datamodels: Understanding predictions with data and data with predictions. In *International Conference on Machine Learning*. PMLR, 2022.

[22] Alexander Immer, Maciej Korzepa, and Matthias Bauer. Improving predictions of bayesian neural nets via local linearization. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 703–711. PMLR, 13–15 Apr 2021.

[23] Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, 2018.

[24] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. TinyBERT: Distilling BERT for natural language understanding. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, November 2020.

[25] Sebastián Jiménez, Mira Jürgens, and Willem Waegeman. Position: Epistemic uncertainty estimation methods are fundamentally incomplete. *arXiv preprint arXiv:2505.23506*, 2025.

[26] Mira Jürgens, Nis Meinert, Viktor Bengs, Eyke Hüllermeier, and Willem Waegeman. Is epistemic uncertainty faithfully represented by evidential deep learning methods? In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[27] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

[28] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[29] Michael Kirchhof, Gjergji Kasneci, and Enkelejda Kasneci. Reexamining the aleatoric and epistemic uncertainty dichotomy. In *ICLR Blogposts 2025*, 2025.

[30] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17. JMLR.org, 2017.

[31] M.R. Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Springer Series in Statistics. Springer New York, 2007.

[32] Nikita Kotelevskii, Vladimir Kondratyev, Martin Takáč, Eric Moulines, and Maxim Panov. From risk to uncertainty: Generating predictive uncertainty measures via bayesian estimation. In *The Thirteenth International Conference on Learning Representations*, 2025.

[33] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. Unpublished technical report.

[34] Salem Lahlou, Moksh Jain, Hadi Nekoei, Victor I Butoi, Paul Bertin, Jarrid Rector-Brooks, Maksym Korablyov, and Yoshua Bengio. DEUP: Direct epistemic uncertainty prediction. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. Expert Certification.

[35] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017.

[36] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 1998.

[37] David JC MacKay. Information-based objective functions for active data selection. *Neural computation*, 4(4), 1992.

[38] David JC MacKay. Probable networks and plausible predictions-a review of practical bayesian methods for supervised neural networks. *Network: computation in neural systems*, 6(3), 1995.

[39] Brady Neal, Sarthak Mittal, Aristide Baratin, Vinayak Tantia, Matthew Scicluna, Simon Lacoste-Julien, and Ioannis Mitliagkas. A modern take on the bias-variance tradeoff in neural networks, 2019.

[40] Michael A. Newton and Adrian E. Raftery. Approximate bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(1), 1994. ISSN 00359246.

[41] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019.

[42] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10), 2010.

[43] Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. Trak: attributing model behavior at scale. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[44] Hippolyt Ritter, Aleksandar Botev, and David Barber. A scalable laplace approximation for neural networks. In *International Conference on Learning Representations*, 2018.

[45] Donald B. Rubin. The Bayesian Bootstrap. *The Annals of Statistics*, 9(1), 1981.

[46] Kajetan Schweighofer, Lukas Aichberger, Mykyta Ielanskyi, and Sepp Hochreiter. Introducing an improved information-theoretic measure of predictive uncertainty. *arXiv preprint arXiv:2311.08309*, 2023.

[47] Robin Senge, Stefan Bösner, Krzysztof Dembczyński, Jörg Haasenritter, Oliver Hirsch, Norbert Donner-Banzhoff, and Eyke Hüllermeier. Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty. *Information Sciences*, 255, 2014.

[48] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[49] Mohammad Hossein Shaker and Eyke Hüllermeier. Aleatoric and epistemic uncertainty with random forests. In Michael R. Berthold, Ad Feelders, and Georg Krempl, editors, *Advances in Intelligent Data Analysis XVIII*. Springer International Publishing, 2020. ISBN 978-3-030-44584-3.

[50] Freddie Bickford Smith, Andreas Kirsch, Sebastian Farquhar, Yarin Gal, Adam Foster, and Tom Rainforth. Prediction-oriented bayesian active learning. In Francisco J. R. Ruiz, Jennifer G. Dy, and Jan-Willem van de Meent, editors, *International Conference on Artificial Intelligence and Statistics, 25-27 April 2023, Palau de Congressos, Valencia, Spain*, volume 206 of *Proceedings of Machine Learning Research*. PMLR, 2023.

[51] Freddie Bickford Smith, Jannik Kossen, Eleanor Trollope, Mark van der Wilk, Adam Foster, and Tom Rainforth. Rethinking aleatoric and epistemic uncertainty. In *Forty-second International Conference on Machine Learning*, 2025.

[52] Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, volume 11006. SPIE, 2019.

[53] A. Tsiatis. *Semiparametric Theory and Missing Data*. Springer Series in Statistics. Springer New York, 2007. ISBN 9780387373454.

[54] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 2000.

[55] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011.

[56] Andrew Gordon Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, 2020.

[57] Lisa Wimmer, Yusuf Sale, Paul Hofman, Bernd Bischl, and Eyke Hüllermeier. Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures? In *Uncertainty in artificial intelligence*. PMLR, 2023.

[58] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2020.

[59] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

A Proofs

A.1 Problem Setup and Notation:

We refresh the problem setup and notation below. Consider a supervised learning K-class classification problem. Specifically,

1. We assume that the training data $\mathcal{D}_n = \{(X_i, Y_i), i = 1, ..., n\}$ consists of feature-label pairs, where the features $X \in \mathcal{X}$ and the labels $Y \in \{1, ..., K\}$.
2. We wish to evaluate the epistemic uncertainty in the prediction for the label Y_{test} for the features X_{test}.
3. We denote our predictions by $\hat{p}(X_{\text{test}}; \theta)$ where $\theta \in \Theta$ our model parameters. Further, denote by $\hat{p}_k(X_{\text{test}}; \theta)$ the prediction probability assigned to the kth class. The model is specified such that $0 < \hat{p}_k(X_{\text{test}}; \theta) < 1$ and $\sum_{k=1}^{K} \hat{p}_k(X_{\text{test}}; \theta) = 1$.
4. In the Bayesian case, we have a prior $p(\theta)$ over the parameters and $p(\theta|\mathcal{D}_n)$ is our posterior. Recall that the epistemic uncertainty measure in the Bayesian case is then
$$\begin{aligned}\mathrm{I}(Y_{\text{test}}; \theta | X_{\text{test}}, \mathcal{D}_n) &= \mathrm{H}(Y_{\text{test}} | X_{\text{test}}, \mathcal{D}_n) - \mathrm{H}(Y_{\text{test}} | \theta, X_{\text{test}}, \mathcal{D}_n) \\ &= \mathrm{H}(\mathbb{E}_{\theta \sim p(\theta|\mathcal{D}_n)}[\hat{p}(X_{\text{test}}; \theta)]) - \mathbb{E}_{\theta \sim p(\theta|\mathcal{D}_n)}[\mathrm{H}(\hat{p}(X_{\text{test}}; \theta))].\end{aligned}$$
5. In the bootstrap case, we denote our estimate of θ over the bootstrapped dataset as $\hat{\theta}_b$. We further denote the distribution of this estimate (induced by the randomness in the choice of bootstrap weights) given the data $\mathcal{D}_n$ as $p(\hat{\theta}_b|\mathcal{D}_n)$. Recall that our proposed epistemic uncertainty measure is then
$$\mathrm{I}_b(X_{\text{test}}, \mathcal{D}_n) = \mathrm{H}(\mathbb{E}_{\hat{\theta}_b \sim p_b(\hat{\theta}_b|\mathcal{D}_n)}[\hat{p}(X_{\text{test}}; \hat{\theta}_b)]) - \mathbb{E}_{\hat{\theta}_b \sim p_b(\hat{\theta}_b|\mathcal{D}_n)}[\mathrm{H}(\hat{p}(X_{\text{test}}; \hat{\theta}_b))].$$
The bootstrap weights may be drawn from either a symmetric Dirichlet with concentration parameter 1 or a symmetric multinomial distribution.

A.2 Assumptions

1. $\Theta \subset \mathbb{R}^p$ is open.
2. There exists a $\theta_0 \in \Theta$ such that $(X_i, Y_i) \overset{iid}{\sim} P_X(x)\hat{p}_y(x; \theta_0) =: P_{\theta_0}$.
3. Assume
$$\psi_\theta(x, y) := \left.\frac{\partial \ln(\hat{p}_y(x; \theta))}{\partial \theta}\right|_{\theta=\theta}$$
exists everywhere. Further assume $\mathbb{E}_{(x,y)\sim P_{\theta_0}}[\psi_{\theta_0}(x, y)] = 0$.
4. $\mathbb{E}_{(x,y)\sim P_{\theta_0}}[||\psi_{\theta_0}(x, y)||^2] < \infty$. Further, $\mathbb{E}_{(x,y)\sim P_{\theta_0}}[\psi_\theta(x, y)]$ is differentiable at θ_0 with non singular derivative matrix $-\mathcal{I}(\theta_0)$. The model is sufficiently smooth such that $\mathbb{E}_{(x,y)\sim P_{\theta_0}}[\psi_{\theta_0}(x, y)\psi_{\theta_0}(x, y)^T] = \mathcal{I}(\theta_0)$.
5. The prior $p(\theta)$ is absolutely continuous with respect to the Lebesgue measure in a neighborhood of θ_0, with a continuous positive density at θ_0.
6. For every $\varepsilon > 0$ there exists a sequence of tests ϕ_n such that
$$P^n_{\theta_0}(\phi_n(\mathcal{D}_n)) \to 0, \qquad \sup_{||\theta - \theta_0|| > \varepsilon} P^n_\theta(1 - \phi_n(\mathcal{D}_n)) \to 0.$$
7. We further make the following assumption which implies uniform integrability to ensure that the first three moments of the posterior distribution converge
$$\sup_n \mathbb{E}_{\theta \sim p(\theta|\mathcal{D}_n)}[|\sqrt{n}(\hat{p}_k(X_{\text{test}}; \theta) - \hat{p}_k(X_{\text{test}}; \theta_0)|^{3+\delta}] < \infty \quad \forall\, k \quad \text{for some } \delta > 0.$$
8. For any sequence $\{\theta_n\} \in \Theta$, $\mathbb{E}_{(x,y)\sim P_{\theta_0}}[\psi_{\theta_n}(x, y)] \to 0$ implies $||\theta_n - \theta_0|| \to 0$.
9. The class $\{\psi_\theta : \theta \in \Theta\}$ is strong Glivenko-Cantelli.

10. For some $\eta > 0$, the class $\mathcal{F} := \{\psi_\theta : \theta \in \Theta, ||\theta - \theta_0|| \leq \eta\}$ is Donsker and $\mathbb{E}_{(x,y)\sim P_{\theta_0}}[||\psi_\theta(x,y) - \psi_{\theta_0}(x,y)||^2] \to 0$ as $||\theta - \theta_0|| \to 0$.
11. Denote our parameter estimate over the complete dataset as $\hat{\theta}$. Then, $\frac{1}{n}\sum_{i=1}^n \psi_{\hat{\theta}}(X_i, Y_i) = o_p(n^{-1/2})$. Likewise, for our bootstrap estimate, $\sum_{i=1}^n \xi_i \psi_{\hat{\theta}_b}(X_i, Y_i) = o_p(n^{-1/2})$ where $\xi_i, i = 1, ..., n$ are the bootstrap weights.
12. We again make the following assumption which implies uniform integrability to ensure that the first three moments of the bootstrap distributions converge:

$$\sup_n \mathbb{E}_{\hat{\theta}_b \sim p(\hat{\theta}_b|\mathcal{D}_n)}[(\sqrt{n}|\hat{p}_k(X_{\text{test}}; \hat{\theta}_b) - \hat{p}_k(X_{\text{test}}; \theta_0)|^{3+\delta}] < \infty \quad \forall\, k \quad \text{for some } \delta > 0.$$

A.3 Theorem 1

Under the assumptions 1-7, we have the asymptotic expansion

$$\mathrm{I}(Y_{\text{test}}; \theta | X_{\text{test}}, \mathcal{D}_n) = \frac{1}{2n}\sum_{k=1}^K \frac{\sigma_k^2}{\hat{p}_k(X_{\text{test}}; \theta_0)} + o_p(n^{-1}) \qquad \forall\, X_{\text{test}},$$

where σ_k^2 is

$$\sigma_k^2 = \left[\frac{\partial \hat{p}(X_{\text{test}}; \theta_0)}{\partial \theta}^T \mathcal{I}^{-1}(\theta_0) \frac{\partial \hat{p}(X_{\text{test}}; \theta_0)}{\partial \theta}\right]_{k,k},$$

for $\mathcal{I}^{-1}(\theta_0)$ being the inverse Fisher information of the model at θ_0.

A.3.1 Proof

Assume that we have some prior $p(\theta)$ over our parameters and denote our posterior by $p(\theta|\mathcal{D}_n)$. Note that our epistemic uncertainty measure is then

$$\begin{aligned}\mathrm{I}(Y_{\text{test}}; \theta | X_{\text{test}}, \mathcal{D}_n) &= \mathrm{H}(Y_{\text{test}}|X_{\text{test}}, \mathcal{D}_n) - \mathrm{H}(Y_{\text{test}}|\theta, X_{\text{test}}, \mathcal{D}_n) \\ &= \mathrm{H}(\mathbb{E}_{\theta\sim p(\theta|\mathcal{D}_n)}[\hat{p}(X_{\text{test}}; \theta)]) - \mathbb{E}_{\theta\sim p(\theta|\mathcal{D}_n)}[\mathrm{H}(\hat{p}(X_{\text{test}}; \theta))].\end{aligned}$$

Our proof proceeds in two parts, we first use Taylor's theorem to find an expansion of the epistemic uncertainty and then we find the asymptotic limit.

Note that H is infinitely differentiable with first three derivatives given by

$$\begin{aligned}\frac{\partial \mathrm{H}(\hat{p})}{\partial \hat{p}_i} &= -1 - \log(\hat{p}_i), \\ \frac{\partial^2 \mathrm{H}(\hat{p})}{\partial \hat{p}_i \partial \hat{p}_j} &= \mathbb{1}\{i = j\}\left(\frac{-1}{\hat{p}_i}\right), \\ \frac{\partial^3 \mathrm{H}(\hat{p})}{\partial \hat{p}_i \partial \hat{p}_j \partial \hat{p}_k} &= \mathbb{1}\{i = j = k\}\left(\frac{1}{\hat{p}_i^2}\right).\end{aligned}$$

Taylor expanding $\mathrm{H}(\hat{p}(X_{\text{test}}; \theta))$ around $\mathbb{E}_{\theta\sim p(\theta|\mathcal{D}_n)}[\hat{p}(X_{\text{test}}; \theta)]$ and using the Lagrange form of the remainder yields

$$\begin{aligned}\mathrm{H}(\hat{p}(X_{\text{test}}; \theta)) =& \mathrm{H}(\mathbb{E}_{\theta\sim p(\theta|\mathcal{D}_n)}[\hat{p}(X_{\text{test}}; \theta)]) \\ &+ \sum_{k=1}^K (\hat{p}_k(X_{\text{test}}; \theta) - \mathbb{E}_{\theta\sim p(\theta|\mathcal{D}_n)}[\hat{p}_k(X_{\text{test}}; \theta)])(-1 - \log(\mathbb{E}_{\theta\sim p(\theta|\mathcal{D}_n)}[\hat{p}_k(X_{\text{test}}; \theta)]) \\ &+ \frac{1}{2}\sum_{k=1}^K (\hat{p}_k(X_{\text{test}}; \theta) - \mathbb{E}_{\theta\sim p(\theta|\mathcal{D}_n)}[\hat{p}_k(X_{\text{test}}; \theta)])^2 \left(\frac{-1}{\mathbb{E}_{\theta\sim p(\theta|\mathcal{D}_n)}[\hat{p}_k(X_{\text{test}}; \theta)]}\right) \\ &+ \frac{1}{6}\sum_{k=1}^K (\hat{p}_k(X_{\text{test}}; \theta) - \mathbb{E}_{\theta\sim p(\theta|\mathcal{D}_n)}[\hat{p}_k(X_{\text{test}}; \theta)])^3 \left(\frac{1}{\hat{p}_k'^2}\right),\end{aligned}$$

where $\hat{p}_k'$ is some value between $\hat{p}_k(X_{\text{test}}; \theta)$ and $\mathbb{E}_{\theta\sim p(\theta|\mathcal{D}_n)}[\hat{p}_k(X_{\text{test}}; \theta)]$ for $k = 1, ..., K$.

Substituting this back into the expression for epistemic uncertainty yields

$$
\begin{aligned}
&\mathrm{I}(Y_{\text{test}};\theta|X_{\text{test}},\mathcal{D}_n)\\
&=\mathrm{H}(\mathbb{E}_{\theta\sim p(\theta|\mathcal{D}_n)}[\hat{p}(X_{\text{test}};\theta)])\\
&-\mathbb{E}_{\theta\sim p(\theta|\mathcal{D}_n)}[\mathrm{H}(\mathbb{E}_{\theta\sim p(\theta|\mathcal{D}_n)}[\hat{p}(X_{\text{test}};\theta)])]\\
&-\mathbb{E}_{\theta\sim p(\theta|\mathcal{D}_n)}\left[\sum_{k=1}^{K}(\hat{p}_k(X_{\text{test}};\theta)-\mathbb{E}_{\theta\sim p(\theta|\mathcal{D}_n)}[\hat{p}_k(X_{\text{test}};\theta)])(-1-\log(\mathbb{E}_{\theta\sim p(\theta|\mathcal{D}_n)}[\hat{p}_k(X_{\text{test}};\theta)])\right]\\
&-\frac{1}{2}\mathbb{E}_{\theta\sim p(\theta|\mathcal{D}_n)}\left[\sum_{k=1}^{K}(\hat{p}_k(X_{\text{test}};\theta)-\mathbb{E}_{\theta\sim p(\theta|\mathcal{D}_n)}[\hat{p}_k(X_{\text{test}};\theta)])^2\left(\frac{-1}{\mathbb{E}_{\theta\sim p(\theta|\mathcal{D}_n)}[\hat{p}_k(X_{\text{test}};\theta)]}\right)\right]\\
&-\frac{1}{6}\mathbb{E}_{\theta\sim p(\theta|\mathcal{D}_n)}\left[\sum_{k=1}^{K}(\hat{p}_k(X_{\text{test}};\theta)-\mathbb{E}_{\theta\sim p(\theta|\mathcal{D}_n)}[\hat{p}_k(X_{\text{test}};\theta)])^3\left(\frac{1}{\hat{p}_k'^2}\right)\right].
\end{aligned}
$$

Note that the first and second term cancel, and the third term is equal to zero. Thus,

$$
\begin{aligned}
\mathrm{I}(Y_{\text{test}};\theta|X_{\text{test}},\mathcal{D}_n)=&\frac{1}{2}\sum_{k=1}^{K}\left(\frac{\mathbb{E}_{\theta\sim p(\theta|\mathcal{D}_n)}\left[(\hat{p}_k(X_{\text{test}};\theta)-\mathbb{E}_{\theta\sim p(\theta|\mathcal{D}_n)}[\hat{p}_k(X_{\text{test}};\theta)])^2\right]}{\mathbb{E}_{\theta\sim p(\theta|\mathcal{D}_n)}[\hat{p}_k(X_{\text{test}};\theta)]}\right)\\
&-\frac{1}{6}\sum_{k=1}^{K}\left(\mathbb{E}_{\theta\sim p(\theta|\mathcal{D}_n)}\left[\frac{(\hat{p}_k(X_{\text{test}};\theta)-\mathbb{E}_{\theta\sim p(\theta|\mathcal{D}_n)}[\hat{p}_k(X_{\text{test}};\theta)])^3}{\hat{p_k'^2}}\right]\right).
\end{aligned}
$$

Note that the first term above is the variance of the probability predictions divided by the mean prediction under the posterior distribution, i.e.

$$
\begin{aligned}
\mathrm{I}(Y_{\text{test}};\theta|X_{\text{test}},\mathcal{D}_n)=&\frac{1}{2}\sum_{k=1}^{K}\left(\frac{\mathbb{V}\mathrm{ar}_{\theta\sim p(\theta|\mathcal{D}_n)}\left[\hat{p}_k(X_{\text{test}};\theta)\right]}{\mathbb{E}_{\theta\sim p(\theta|\mathcal{D}_n)}[\hat{p}_k(X_{\text{test}};\theta)]}\right)\\
&-\frac{1}{6}\sum_{k=1}^{K}\left(\mathbb{E}_{\theta\sim p(\theta|\mathcal{D}_n)}\left[\frac{(\hat{p}_k(X_{\text{test}};\theta)-\mathbb{E}_{\theta\sim p(\theta|\mathcal{D}_n)}[\hat{p}_k(X_{\text{test}};\theta)])^3}{\hat{p_k'^2}}\right]\right).
\end{aligned}
$$

Next, we shall evaluate the limit.
Under assumptions 1- 7, we can apply the Bernstein-von Mises theorem (Theorem 10.1 of van der Vaart (2000)). Note that Bernstein von-Mises implies convergence in total variation distance which directly implies convergence in distribution. Hence, we have

$$
\sqrt{n}\,(\theta-\theta_0)\,\big|\mathcal{D}_n\xrightarrow{d}\mathcal{N}(0,\mathcal{I}^{-1}(\theta_0))\qquad\text{in } P_{\theta_0}^n\text{-probability.}
$$

In particular, the above is short-hand for

$$
\sup_{h\in BL_1}\left|\mathbb{E}_{\theta\sim p(\theta|\mathcal{D}_n)}\left[h\left(\sqrt{n}(\theta-\theta_0)\right)\right]-\mathbb{E}_{Z\sim\mathcal{N}(0,\mathcal{I}^{-1}(\theta_0))}[h(Z)]\right|\xrightarrow{P_{\theta_0}^n}0,
$$

where BL_1 is the space of functions $h:\mathbb{R}^p\mapsto\mathbb{R}$ with Lipschitz norm bounded by 1. Recall that if both expectations on the left hand side were unconditional, limit of the left hand side going to 0 is the standard definition of convergence in distribution. See Sec. 23.2.1 of van der Vaart (2000) for further discussion.

We shall henceforth continue with the "$(.)\xrightarrow{d}(.)$ in $P_{\theta_0}^n$ probability" notation for ease of exposition.

Note that 3 allows us to apply the functional delta method (see chapter 20 of van der Vaart (2000)) to evaluate the limit of the function $\theta\mapsto\hat{p}_k(X_{\text{test}};\theta)$ under the posterior. This yields,

$$
\sqrt{n}\,(\hat{p}_k(X_{\text{test}};\theta)-\hat{p}_k(X_{\text{test}};\theta_0))\,\big|\mathcal{D}_n\xrightarrow{d}\mathcal{N}(0,\sigma_k^2)\qquad\text{in } P_{\theta_0}^n\text{-probability,}\tag{5}
$$

where $\sigma_k^2 = \left[\frac{\partial \hat{p}(X_{\text{test}};\theta_0)}{\partial \theta}^T \mathcal{I}^{-1}(\theta_0) \frac{\partial \hat{p}(X_{\text{test}};\theta_0)}{\partial \theta}\right]_{k,k}$. Finally, 7 implies uniform integrability allowing us to conclude that the limit of the moment is the moment of the limiting distribution:

$$\mathbb{E}_{\theta \sim p(\theta|\mathcal{D}_n)}[\hat{p}_k(X_{\text{test}};\theta)] \overset{P^n_{\theta_0}}{\to} \hat{p}_k(X_{\text{test}};\theta_0),$$

$$\mathbb{V}\text{ar}_{\theta \sim p(\theta|\mathcal{D}_n)}\left[\sqrt{n}\hat{p}_k(X_{\text{test}};\theta)\right] \overset{P^n_{\theta_0}}{\to} \sigma_k^2.$$

Next, we shall consider the remainder term from the Taylor expansion. Note that $\hat{p}'_k \in [\hat{p}_k(X_{\text{test}};\theta), \mathbb{E}_{\theta \sim p(\theta|\mathcal{D}_n)}[\hat{p}_k(X_{\text{test}};\theta)]]$. By assumption, $0 < \hat{p}_k(X_{\text{test}};\theta) < 1$ and hence $1/\hat{p'}_k^2 = O_p(1)$. Furthermore, we also have from equation 5 that $(\hat{p}_k(X_{\text{test}};\theta) - \hat{p}_k(X_{\text{test}};\theta_0))\,\big|\mathcal{D}_n = O_p\left(\frac{1}{\sqrt{n}}\right)$. Therefore,

$$n\frac{(\hat{p}_k(X_{\text{test}};\theta) - \mathbb{E}_{\theta \sim p(\theta|\mathcal{D}_n)}[\hat{p}_k(X_{\text{test}};\theta)])^3}{\hat{p'}_k^2} = nO_p\left(\frac{1}{n^{3/2}}\right)O_p(1) = O_p\left(\frac{1}{\sqrt{n}}\right).$$

Again, by the uniform integrability assumption

$$\mathbb{E}_{\theta \sim p(\theta|\mathcal{D}_n)}\left[n\frac{(\hat{p}_k(X_{\text{test}};\theta) - \mathbb{E}_{\theta \sim p(\theta|\mathcal{D}_n)}[\hat{p}_k(X_{\text{test}};\theta)])^3}{\hat{p'}_k^2}\right] \overset{P^n_{\theta_0}}{\to} 0.$$

Therefore,

$$n\text{I}(Y_{\text{test}};\theta|X_{\text{test}},\mathcal{D}_n) \overset{P^n_{\theta_0}}{\to} \frac{1}{2}\sum_{k=1}^{K}\frac{\sigma_k^2}{\hat{p}_k(X_{\text{test}};\theta_0)}.$$

A.4 Theorem 2

Under the assumptions 1-12,

$$\frac{\text{I}(Y_{\text{test}};\theta|X_{\text{test}},\mathcal{D}_n)}{\text{I}_b(X_{\text{test}},\mathcal{D}_n)} \overset{P^n_{\theta_0}}{\to} 1 \qquad \forall\, X_{\text{test}},$$

In particular, note that I_b may be calculated using either symmetric Dirichlet weights with concentration parameter 1, or multinomial weights.

A.4.1 Proof

Recall that our proposed epistemic uncertainty measure for the bootstrap case is

$$\text{I}_b(X_{\text{test}},\mathcal{D}_n) = \text{H}(\mathbb{E}_{\hat{\theta}_b \sim p_b(\hat{\theta}_b|\mathcal{D}_n)}[\hat{p}(X_{\text{test}};\hat{\theta}_b)]) - \mathbb{E}_{\hat{\theta}_b \sim p_b(\hat{\theta}_b|\mathcal{D}_n)}[\text{H}(\hat{p}(X_{\text{test}};\hat{\theta}_b))].$$

We can proceed exactly as in the Bayesian case to obtain

$$\begin{aligned}\text{I}_b(X_{\text{test}},\mathcal{D}_n) =& \frac{1}{2}\sum_{k=1}^{K}\left(\frac{\mathbb{V}\text{ar}_{\hat{\theta}_b \sim p_b(\hat{\theta}_b|\mathcal{D}_n)}\left[\hat{p}_k(X_{\text{test}};\hat{\theta}_b)\right]}{\mathbb{E}_{\hat{\theta}_b \sim p_b(\hat{\theta}_b|\mathcal{D}_n)}[\hat{p}_k(X_{\text{test}};\hat{\theta}_b)]}\right) \\ &- \frac{1}{6}\sum_{k=1}^{K}\left(\mathbb{E}_{\hat{\theta}_b \sim p_b(\hat{\theta}_b|\mathcal{D}_n)}\left[\frac{(\hat{p}_k(X_{\text{test}};\hat{\theta}_b) - \mathbb{E}_{\hat{\theta}_b \sim p_b(\hat{\theta}_b|\mathcal{D}_n)}[\hat{p}_k(X_{\text{test}};\hat{\theta}_b)])^3}{\hat{p'}_k^2}\right]\right),\end{aligned}$$

where $\hat{p}'_k$ is now some value between $\hat{p}_k(X_{\text{test}};\hat{\theta}_b)$ and $\mathbb{E}_{\hat{\theta}_b \sim p_b(\hat{\theta}_b|\mathcal{D}_n)}[\hat{p}_k(X_{\text{test}};\hat{\theta}_b)]$ for $k = 1, ..., K$. Next, from Theorem 10.16 of Kosorok (2007), we have

$$\sqrt{n}\left(\hat{\theta}_b - \hat{\theta}\big|\mathcal{D}_n\right) \overset{d}{\to} \mathcal{N}(0, \mathcal{I}^{-1}(\theta_0)) \qquad \text{in } P^n_{\theta_0}\text{-probability}$$

and

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \overset{d}{\to} \mathcal{N}(0, \mathcal{I}^{-1}(\theta_0)),$$

where $\hat{\theta}$ is the MLE over the complete data. Notably, both the statements above hold for either multinomial or Dirichlet weights.

We can now apply the delta method for bootstrap (van der Vaart [54] Theorem 23.9) to this and proceed exactly as in the Bayesian case to conclude

$$n\mathrm{I}_b(X_{\mathrm{test}}, \mathcal{D}_n) \overset{P^n_{\theta_0}}{\to} \frac{1}{2}\sum_{k=1}^{K} \frac{\sigma_k^2}{\hat{p}_k(X_{\mathrm{test}}; \theta_0)}.$$

Combining the Bayesian and bootstrap results by the continuous mapping theorem,

$$\frac{n\mathrm{I}(Y_{\mathrm{test}}; \theta | X_{\mathrm{test}}, \mathcal{D}_n)}{n\mathrm{I}_b(X_{\mathrm{test}}, \mathcal{D}_n)} \overset{P^n_{\theta_0}}{\to} 1.$$

B Additional Experiments and Experimental Details

All experiments were conducted on an Ubuntu 22.04.5 Linux cluster using 2 NVIDIA V100 Tensor Core GPUs. All experiments needed no more than 32GB of GPU memory and 40GB of CPU RAM. Training the ImageNet models for a single seed took roughly 8 hours per seed, CIFAR-10 took about 30 minutes per seed, AGNews took less than 20 minutes per seed and ISIC took less than a minute per seed. Total time to reproduce results in the paper is no more than 30 days of computation with resources listed above and total computation time used by the research project is no more than 100 days.

B.1 Epistemic Uncertainty Decomposition

B.1.1 CIFAR-10

Experimental Details: We use the CIFAR-10 dataset [33] (publicly available granting all rights to users) with the original training and test set folds. We use a standard ResNet-18 [15] architecture but modify the first convolutional layer to have a kernel size of 3, stride of 1, padding of 1 to allow the network to work with 32 $\times$32 size images. Further, we modify the last fully connected layer to only have 10 output neurons for a 10 class classification task. We use cross entropy loss and Adam [28] as an optimizer with learning rate of 0.001 and betas of 0.9 and 0.999. We use a batch size of 128 and train for 15 epochs. The rest of the experiment proceeds as described in the main text.

Accuracy Metrics: Models within the deep ensemble achieve an accuracy of 0.83 on average, whereas the bootstrap models achieve an accuracy of 0.80 on average. The deep ensemble itself (i.e. averaged model predictions) achieves an accuracy of 0.89. Deep ensembles over the bootstrapped datasets (i.e. model predictions averaged over the random seeds only) achieves an accuracy of 0.88 on average across the bootstrapped datasets.

B.1.2 ImageNet

Experimental Details: We use the ImageNet dataset [4] (available for non-commercial use subject to terms of usage listed at `https://www.image-net.org/`) with the original training and test set folds. For computational ease, we use a downsampled version of the images downsampled to 32×32 pixels as in Chrabaszcz et al. (2017). We use a standard ResNet-18 [15] architecture but modify the first convolutional layer to have a kernel size of 3, stride of 1, padding of 1 to allow the network to work with 32×32 size images. We use cross entropy loss and SGD as an optimizer with learning rate of 0.1, momentum of 0.9 and weight decay of 0.0001. We use a batch size of 128 and train for 40 epochs. We use the one cycle learning rate policy from Smith and Topin (2019). The rest of the experiment proceeds as described in the main text.

Accuracy Metrics: Models within the deep ensemble achieve top-1 and top-5 accuracy of 0.55 and 0.79 respectively on average, whereas the bootstrap models achieve top-1 and top-5 accuracies of 0.51 and 0.75 respectively on average. The deep ensemble itself (i.e. averaged model predictions) achieves top-1 and top-5 accuracy of 0.59 and 0.81 respectively. Deep ensembles over the bootstrapped datasets (i.e. model predictions averaged over the random seeds only) achieves top-1 and top-5 accuracy of 0.55 and 0.78 on average across the bootstrapped datasets.

B.1.3 ISIC Dataset

Experimental Details: We use the ISIC Skin Lesion challenge (2016) dataset [13] (publicly available for non-commerical research and educational use) with the original training and test set folds. We pass all images through the trained ResNet-50 from He et al. (2015) and extract the activations from the last layer to obtain 2048-dimensional feature vectors. We then train a two layer neural network from scratch on top of these feature vectors with 512 hidden units, ReLU activation and Dropout with a rate of 0.5. We use Adam [28] as an optimizer with a learning rate of 0.001 and betas of 0.9 and 0.999. We use a batch size of 32 and train for 15 epochs minimizing cross entropy loss weighted by prevalence of the ground truth class in the training data to account for class imbalance. The rest of the experiment proceeds as described in the main text.

Accuracy Metrics: Models within the deep ensemble achieve an AUC-ROC of 0.81 on average, whereas the bootstrap models achieve an accuracy of 0.79 on average. The deep ensemble itself (i.e.

averaged model predictions) achieves an accuracy of 0.81. Deep ensembles over the bootstrapped datasets (i.e. model predictions averaged over the random seeds only) achieves an accuracy of 0.80 on average across the bootstrapped datasets.

B.1.4 AG News

Experimental Details: We use the AG News dataset [59] (publicly available for non-commerical research and educational use) with the original training and test set folds. We finetune the TinyBERT model from Jiao et al. (2020). We add a linear layer on top of the final 312 dimensional output from TinyBERT to perform 4 class classification (parameters initialized randomly). We use Adam [28] as an optimizer with a learning rate of $2e-5$ and betas of 0.9 and 0.999. We also use a linear learning rate schedule with warmup as provided by the transformers [58] library. We train for 3 epochs with a batch size of 32. We tokenize each example up to 128 tokens, padding or truncating as necessary. The rest of the experiment proceeds as described in the main text.

Accuracy Metrics: Models within the deep ensemble achieve an accuracy of 0.94 on average, whereas the bootstrap models achieve an accuracy of 0.93 on average. The deep ensemble itself (i.e. averaged model predictions) achieves an accuracy of 0.94. Deep ensembles over the bootstrapped datasets (i.e. model predictions averaged over the random seeds only) achieves an accuracy of 0.93 on average across the bootstrapped datasets.

B.2 Data Attribution

Experimental Details: We use the MNIST dataset [36] (available under a CC BY-SA 3.0 license). For training, we randomly select 600 images from each class to form the training dataset and keep the test dataset the same as the original.

We use a two layer convolutional neural network with 3 fully connected layers. We use Adam [28] as an optimizer with a learning rate of 0.001, betas of 0.9 and 0.999, weight decay of 0.0001, and a batch size of 32. We train for 10 epochs.

To calculate the influence function, we freeze all but the final two fully connected layers and calculate the hessian and gradients only with respect to the parameters in these layers. Specifically, as in Koh and Liang (2017), the change in model parameters (for the final two layers) for an ϵ increase in the weighting of the ith training data point is $-H_{\hat{\theta}}^{-1}\nabla_\theta \log(\hat{p}_{Y_i}(X_i;\hat{\theta}))$ where $\hat{\theta}$ is the MLE on the original (unweighted) dataset, $H_{\hat{\theta}} = \frac{1}{n}\sum_{i=1}^{n}\nabla_\theta^2 \log(\hat{p}_{Y_i}(X_i;\hat{\theta}))$ is the hessian and ∇_θ denotes derivative with respect to θ. We use multinomial bootstrap weights. For practical calculation, we add the identity matrix scaled by 10^{-5} to the Hessian to ensure invertibility in cases with non-negative definite hessian—see Koh and Liang (2017) for discussion.

For the Dropout estimate, we add Dropout layers with probability 0.5 between the fully connected layers and train with the same hyperparameters above.

The rest of the experiment proceeds as in the main text.

B.2.1 Datamodels Experiment

We now also investigate the use of datamodels [21] as a means of approximating our measure.

Experimental Details: We use the CIFAR-10 dataset [33] and ResNet-9 architecture as in Ilyas et al. (2022). Each test point is associated with a datamodel which consists of a weight vector of length 50001—the number of training data points and a bias term. The datamodel is such that it predicts the model output for that test data point given the weights assigned to the training data as the dot product between the training data weights and the datamodel weights. The model output considered in Ilyas et al. (2022) is unfortunately not the entire probability vector, but simply the difference between the correct logit and the highest incorrect logit. For our experiment, we use the datamodels for CIFAR-10 provided as part of the paper—see Ilyas et al. (2022) for details. We then draw 100 sets of Dirichlet weights and obtain estimates of model outputs using the datamodel. We then pass the model output through the sigmoid function to obtain a probability and use this for calculation of the EU. Note that this is equivalent to treating the classification task as binary between the correct class vs the highest incorrect class. While this is not equivalent to the correct calculation of EU over the 10 class probability vector, it is the best we can do with the existing datamodel.

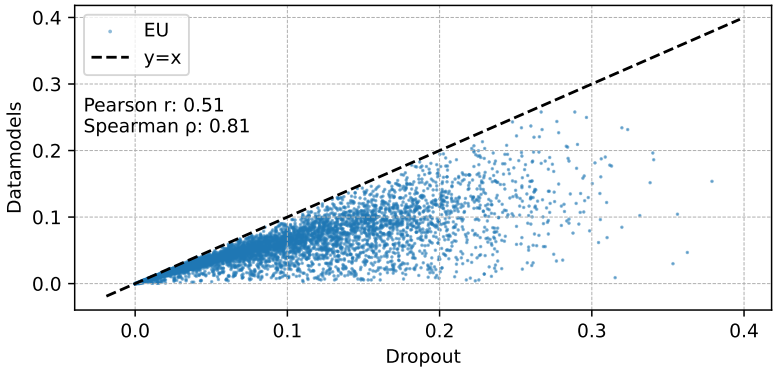

Figure 5: **Datamodels Approximation:** Datamodel approximation to our measure versus mutual information estimates obtained using Dropout on CIFAR-10 test points. See text for details.

For the dropout estimate, we train a ResNet-9 with Dropout layers added with probability 0.5 with the exact same hyperparameters used to create the datamodel.

Results: Fig. 5 shows that the datamodels approximation to our measure and the Dropout epistemic uncertainty scores have a high rank correlation and moderate Pearson correlation. The datamodel estimates are also consistently lower than the Dropout estimates. This may be because of the reduction to a binary classification task in the Datamodels approximation versus the full 10 class classification in the Dropout calculation. Nonetheless, this is a useful proof of concept experiment which opens the door to using data attribution methods for a more nuanced understanding of sources of uncertainty.

B.3 Ablations

B.3.1 Choice of Bootstrap Weights

The experimental results in Sec. 4 were conducted with Dirichlet weights for the bootstrap instead of the more common multinomial weights. As discussed in Sec. 2, we hypothesize that these weights are more suitable since it allows the neural network to still see all the training instances and hence capture the training data better than multinomial weights which omit training points entirely. We now ablate against this choice of weights. Specifically, we now perform bootstrapping by drawing bootstrap datasets by performing i.i.d. draws from the empirical distribution. This is the same as using bootstrap weights drawn from the multinomial distribution with event probabilities all being symmetrically $1/n$ where n is the size of the dataset (assuming no repeated datapoints).

	Dataset	$m_{\text{seeds}}^{\text{resampling}}$	$m_{\text{seeds}}^{\text{deep ensemble}}$	ρ
Multinomial	CIFAR-10	0.24	0.82	0.89
	AG News	0.47	0.84	0.96
	ISIC	0.61	0.77	0.88
Dirichlet	CIFAR-10	0.11	0.98	0.94
	AG News	0.37	0.88	0.97
	ISIC	0.66	0.84	0.95

Table 2: Comparison of slopes of line of best fit and correlations when using multinomial versus dirichlet weights across datasets.

Table 2 reports the metrics defined in Sec. 4.2 for the two different choices of weights. Since multinomial weights omit datapoints entirely from training, they tend to lead to more significant variation when bootstrapping and hence the resampling component becomes more significant while

the correlation between the deep ensemble estimate and seeds component worsens. Yet, qualitatively, the seeds component still dominates and correlates well with the deep ensemble estimate.

B.3.2 Dataset Size

We additionally ablate against the size of the dataset to investigate whether our conclusions still hold in low data regimes where the resampling component may be more dominant. On the CIFAR-10 dataset, we subsample to 5% and 25% of the full training data and then repeat our experiments. Table 3 reports values of the metrics defined in Sec. 4.2.

	Dataset Size		
	5%	25%	100%
$m_{\text{seeds}}^{\text{resampling}}$	0.21	0.22	0.11
$m_{\text{seeds}}^{\text{deep ensemble}}$	1.05	1.26	0.98
ρ	0.79	0.87	0.94

Table 3: Comparison of slopes of line of best fit and correlation when using subsampled versions of the CIFAR-10 dataset.

The contribution to uncertainty from resampling the data does grow slightly as expected, but continues to be clearly dominated by the seeds contribution (even at just 5% of the original data). Further, while the correlation between the deep ensemble estimate and the seeds component worsens, it still stays quite high. Our conclusions thus seem robust to the dataset size.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Sec. 3 presents all theoretical results and Sec. 4 combines with experimental findings to prove main claims.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Sec. 6

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The main assumptions and proof sketch are given when presenting the theorems in Sec. 3 itself, with formal mathematical statements for all assumptions and rigorous proofs in the Supplement Sec. A.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: See Supplement Sec. B and code is also provided separately for reproducibility.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Code is provided separately for reproducibility and all datasets used are publicly available.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: See supplement Sec. B

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The relevant numerical values are obtained by averaging over at least 10 seeds, usually 25 for stability. The exact numerical values are unimportant and only qualitative behavior is relevant to our conclusions. Further, repeating these experiments for 5 seeds for example would require training over $5 \times 650 \times 4$ models which is computationally prohibitive.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See supplement Sec. B

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: There are no major ethical concerns

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The work is intended to better understand uncertainty quantification capabilities of Deep Ensembles and does not have any major societal impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The work is intended to better understand uncertainty quantification capabilities of Deep Ensembles and does not have any major societal impacts

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: See Sec. B.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: No new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No crowdsourcing used.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were not used.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.