Diversity-enhanced Probabilistic Ensemble For Uncertainty Estimation

Hanjing Wang[1] **Qiang Ji**[1]

[1]ECSE, Rensselaer Polytechnic Institute, Troy, New York, USA

Abstract

Ensemble methods combine multiple individual models for prediction, which have demonstrated their effectiveness in accurate uncertainty quantification (UQ) and strong robustness. Obtaining a diverse ensemble set of model parameters results in better model averaging performance and better approximation of the true posterior distribution of these parameters. In this paper, we propose the diversity-enhanced probabilistic ensemble method with the adaptive uncertainty-guided ensemble learning strategy for better quantifying uncertainty and further improving the model robustness. Specifically, we construct the probabilistic ensemble model by building a Gaussian distribution of the model parameters for each ensemble component using Laplacian approximation in a post-processing manner. Then a mixture of Gaussian model is established with learnable and refinable parameters in an EM-like algorithm. During ensemble training, we leverage the uncertainty estimated from previous models as guidance when training the next one such that the new model will focus more on the less explored regions by previous models. Various experiments including out-of-distribution detection and image classification under distributional shifts have demonstrated better uncertainty estimation and improved model generalization ability of our proposed method.

1 INTRODUCTION

The real world is full of uncertainty. However, deterministic deep learning models might be overconfident in some predictions that they actually do not know due to the lack of knowledge of those data regions [Lakshminarayanan et al., 2017a]. Hence, establishing deep learning models in a probabilistic manner is very important for a trusted system, which will enable us to tell when the models will fail in their predictions and guide human behaviors with confidence.

There are mainly two types of uncertainty, namely, epistemic uncertainty and aleatoric uncertainty [Kendall and Gal, 2017]. Epistemic uncertainty represents the prediction uncertainty due to the lack of knowledge when building the models. Aleatoric uncertainty measures the inherent data noise in the distribution, which is irreducible. For quantifying those uncertainties, we can rely on Bayesian neural networks (BNNs) which aim at constructing the posterior distribution of the neural network parameters. However, the Bayesian inference performs marginalization over the posterior distribution, which is often intractable in practice.

Alternatively, the deep ensemble method [Lakshminarayanan et al., 2017a] trains an ensemble of deep neural networks from random initializations, which demonstrates great success in predictive uncertainty calibration and outperforms various approximate BNNs. Generating sufficient and diverse ensemble components can better approximate the complex posterior distribution. Ensemble diversity is also a good indicator of uncertainty quantification performance and model robustness Dusenberry et al. [2020]. Higher diversity enables different models to make independent errors such that their individual mistakes will be canceled out during majority voting and model averaging, leading to better prediction accuracy and improved generalization ability [Bian and Chen, 2021, Zhang et al., 2020]. However, traditional ensemble methods have limited diversity since each component is trained independently with only different initializations. Generating sufficient and diverse ensemble models requires many initializations and is hence computationally expensive. With limited computational resources, ensemble-based methods can only provide several modes to approximate the posterior distribution, which is hard to describe the complex posterior landscape. Moreover, previous methods often train each component independently and ignore the important knowledge from previous models when getting a new model. Finally, there are also multi-

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

ple resources that we can gain additional diversity during ensemble training besides random initializations.

To overcome the above limitations, we propose the diversity-enhanced probabilistic ensemble (PE) method, which has the following contributions.

- We leverage the PE, a Bayesian framework to model aleatoric and epistemic uncertainty by combining the ensemble method and Laplacian approximation (LA) [MacKay, 1992] for Bayesian inference. The diversity of ensemble components is enhanced through exploring the neighborhood of each ensemble member by LA, where performance guarantees are provided.
- Given the LA for ensemble members, a mixture of Gaussian (MoG) is constructed with learnable and refinable parameters in an EM-like algorithm, enabling a better posterior approximation of model parameters.
- We propose an adaptive uncertainty-guided ensemble training strategy (AUEL), where the new ensemble model is trained based on the knowledge of previous models with the guidance of uncertainty, leading to an improved ensemble diversity and a better joint model averaging performance.
- Various applications have been conducted including out-of-distribution detection and image classification under distributional shifts, which showcase the competitive performance of our method in uncertainty estimation and domain generalization.

2 RELATED WORK

Laplacian Approximation Laplace approximation assumes a Gaussian posterior distribution by performing Taylor expansion around the mode. However, constructing the Gaussian posterior for large models by LA is not applicable mainly because of the computational difficulty of the large covariance matrix for high-dimensional model parameters. Several methods are proposed to improve the efficiency of LA. For example, subnetwork LA [Daxberger et al., 2021b] and last-layer LA [Kristiadi et al., 2020] reduce the number of Bayesian parameters by constructing the posterior distribution only for partial neural network weights. Different Hessian matrix factorization methods are also proposed such as Kronecker-factored approximation curvature (KFAC) [Ritter et al., 2018] and low-rank KFAC [Lee et al., 2020]. Please refer to the survey paper [Daxberger et al., 2021a] for more information.

Ensemble Methods for Uncertainty Estimation Besides the deep ensemble method, different ensemble-based variants have been proposed to improve the UQ efficiency or accuracy. For improving efficiency, deep sub-ensemble [Valdenegro-Toro, 2019] ensembles only the layers close to the output. The snapshot ensemble [Huang et al., 2017] method collects different ensemble components in different epochs of one training attempt. Considering weight sharing, the batch-ensemble [Wen et al., 2020] method proposes a parameter-efficient representation of ensemble weights. For improving the accuracy, some ensemble methods further explore each ensemble subspace by an approximate posterior estimation such as Multi-SWAG [Wilson and Izmailov, 2020] and ensemble with subspace sampling [Fort et al., 2019]. Multi-SWAG combines the deep ensemble with SWAG to form a mixture of Gaussian distribution with uniform coefficients while Fort et al. [2019] built an ensemble model by training multiple variational BNNs with empirical analysis. Recently, Eschenhagen et al. [2021] connected ensemble methods with LA for better uncertainty quantification. Some ensemble techniques such as MIMO [Havasi et al., 2020] and Rank-1 BNN [Dusenberry et al., 2020] also use a mixture of approximate posteriors to capture ensemble components. However, they are mainly designed for training multiple subnetworks in one model's capacity, which is less accurate. Compared to the above methods, we focus on diversity-enhanced ensemble learning for improving UQ accuracy and proposed three sub-modules, including probabilistic ensemble, adaptive uncertainty-guided ensemble learning, and MoG refinement.

Diversity-enhanced Ensemble Learning Diversity matters for improving ensemble performance. One line of work trains ensemble models with special diversity regularization [Zhang et al., 2020, Zaid et al., 2021, Jain et al., 2020, Liu and Yao, 1999, Pearce et al., 2018, Wabartha et al., 2021]. For example, Zhang et al. [2020] utilized the pairwise difference between classifiers as regularization. Zaid et al. [2021] created a diversity-promoted ensemble loss based on mutual information. Jain et al. [2020] leveraged out-of-distribution samples as regularization to increase ensemble diversity. Another line of work focuses on training each ensemble component with a subset of data so that each ensemble model has its own learning specialty to increase diversity [Lee et al., 2015, Zhou et al., 2018]. Moreover, EDST [Liu et al., 2021] and SeBayS [Jantre et al., 2022] make adjustments to the learning process to obtain ensembles sequentially from diverse models. Finally, Wenzel et al. [2020] tried to increase ensemble diversity by training with different hyperparameters. Zaidi et al. [2020] further constructed the ensemble models with different architectures and proposed a special selection procedure for choosing diverse ensemble members from a pre-trained ensemble model pool. Recently, several particle-based function-space variational inference methods [Tiulpin and Blaschko, 2021, D'Angelo and Fortuin, 2021, Yashima et al., 2022] try to utilize a finite number of models to approximate the Bayesian posterior distribution through optimization, where they often consider the interaction between models with an explicit diversity measurement as regularization. We exclude them from comparisons since our proposed method is in weight

space and is a randomization-based method, which is often more efficient than function-space methods. In the development of diversity-enhanced learning, different diversity metrics are studied [Wu et al., 2020, 2021]. There are also many related applications such as active learning [Tan et al., 2021] and computer vision tasks [Dvornik et al., 2019].

3 PROPOSED METHOD

3.1 BACKGROUND

General Notations and Assumptions. Denote the input as x, the target variable as y, the training data as $\mathcal{D} = \{x_m, y_m\}_{m=1}^M$. In this paper, we will focus on classification tasks. We denote $f(x, \theta) \in \mathcal{R}^C$ as the output of the neural network with input x parameterized by θ, which is the probability logit before the softmax layer. C represents the number of classes. When constructing the ensemble models, $f(x, \theta_i)$ represents the output of the ith ensemble component. β represents the hyperparameter of LA for the prior distribution of θ. $\mathbb{E}(\cdot)$ represents the expectation. $\mathcal{H}(\cdot)$ represents the entropy.

Laplacian Approximation. The LA constructs the posterior distribution $p(\theta|\mathcal{D}, \beta)$ by a Gaussian distribution around the MAP estimate θ_{map} where

$$\theta_{map} = \arg\max_{\theta} \log p(\theta|\mathcal{D}, \beta). \tag{1}$$

By taking the second-order Taylor expansion of $\log p(\theta|\mathcal{D}, \beta)$ around θ_{map}, we can observe that

$$p(\theta|\mathcal{D}, \beta) \approx \mathcal{N}(\theta_{map}, \Sigma) \tag{2}$$

where $\Sigma = -(H)^{-1}$ and $H = \nabla_\theta^2 \log p(\theta|\mathcal{D}, \beta)|_{\theta=\theta_{map}}$. Please refer to Appendix A.1 for more details. This paper utilizes the last-layer LA [Kristiadi et al., 2020] to achieve competitive accuracy with high efficiency.

Uncertainty Quantification. For classification problems, we estimate the epistemic uncertainty and the aleatoric uncertainty by the mutual information and the expected entropy [Depeweg et al., 2018]. Details can be found in Appendix A.2.

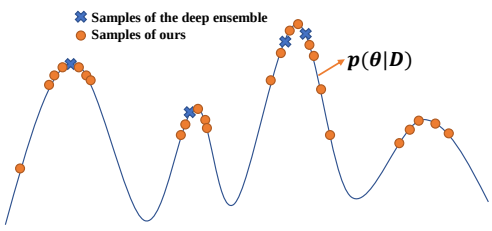


Figure 1: Posterior Approximation by Samples.

3.2 PROBABILISTIC ENSEMBLE

Given N pre-trained ensemble models, we denote θ_i as the MAP estimation of the ith ensemble component parameters. Inspired by [Eschenhagen et al., 2021], we perform the Laplacian approximation for each ensemble component as an approximation of the true posterior, denoted as $\mathcal{N}(\theta; \theta_i, \Sigma_i)$. A mixture of Gaussian model is constructed with coefficients $\{\lambda_i\}_{i=1}^N$ as shown in Eq. (3), which can better approximate the posterior distribution $p(\theta|\mathcal{D}, \beta)$.

$$p(\theta|\mathcal{D}, \beta) \approx \sum_{i=1}^{N} \lambda_i \mathcal{N}(\theta; \theta_i, \Sigma_i) \tag{3}$$

where $\lambda_i \in [0, 1], i = 1, 2, ..., N$ and $\sum_{i=1}^N \lambda_i = 1$.

As shown in Figure 1, PE can better approximate the posterior distribution than the deep ensemble method by exploring each ensemble subspace using LA. Given the probabilistic ensemble model, the Bayesian inference is performed shown in Eq. (4):

$$\begin{aligned} p(y|x, \mathcal{D}) &\approx \int p(y|x, \theta) \sum_{i=1}^{N} \lambda_i \mathcal{N}(\theta; \theta_i, \Sigma_i) d\theta \\ &\approx \frac{1}{S} \sum_{s=1}^{S} p(y|x, \theta^s). \end{aligned} \tag{4}$$

where $\theta^s \sim \sum_{i=1}^N \lambda_i \mathcal{N}(\theta; \theta_i, \Sigma_i)$ represents the sth sample from the Gaussian mixture model. While our suggested probabilistic ensemble approach may bear resemblances to the method outlined in [Eschenhagen et al., 2021], especially in the context of merging LA with ensemble models, our research is primarily driven by a focus on diversity-enhanced ensemble learning, backed with theoretical validations. The most notable distinctions between our strategies and [Eschenhagen et al., 2021] predominantly involve our process of securing the diverse modes $\{\theta_i\}_{i=1}^N$, and our methodology in formulating the Gaussian mixture.

Several propositions are shown to demonstrate the effectiveness of the PE model with theoretical guarantees. They are valid for the PE of any pre-trained deterministic ensemble models, regardless of their training methodology. All the proofs can be found in Appendix B. Specifically, approximation guarantees are shown in Proposition 3.1 and 3.2. When the sample size is large, Proposition 3.1 guarantees that the PE model converges to the true posterior distribution. Otherwise, Proposition 3.2 shows theoretical evidence that the PE model bridges the connection of the deep ensemble method to approximate Bayesian inference and has better posterior approximation than single LA.

Proposition 3.1 (Convergence of PE)**.** *Denote the data samples as $\mathcal{D} = \{x_m, y_m\}_{m=1}^M$. Under mild regularity conditions [Gelman, 2011], as the sample size $M \to \infty$, the*

probabilistic ensemble representation of θ approaches its posterior distribution, i.e.,

$$\sup_{\theta} \left| p(\theta|\mathcal{D}, \beta) - \sum_{i=1}^{N} \lambda_i \mathcal{N}(\theta; \theta_i, \Sigma_i) \right| \to 0. \quad (5)$$

Proposition 3.2 (Better posterior approximation)**.** *PE models extend the deep ensemble method for approximate Bayesian inference. Denote $p_{PE}(\theta) = \sum_{i=1}^{N} \lambda_i \mathcal{N}(\theta; \theta_i, \Sigma_i), p_{LA}^{(i)}(\theta) = \mathcal{N}(\theta; \theta_i, \Sigma_i)$ as the PE approximation and the ith-network LA, respectively. The PE model has better posterior approximation compared to the single LA with a measure of KL divergence.*

$$KL(p(\theta|\mathcal{D}, \beta)||p_{PE}(\theta)) \leq \sum_{i=1}^{N} \lambda_i KL(p(\theta|\mathcal{D}, \beta)||p_{LA}^{(i)}(\theta)) \quad (6)$$

Proposition 3.3 (Error reduction of the PE and the role of diversity)**.** *Denote $\theta \sim \sum_{i=1}^{N} \lambda_i \mathcal{N}(\theta; \theta_i, \Sigma_i)$ as PE parameters, x as the input, and y^* as the corresponding label. The PE model fulfills*

$$\begin{aligned} -\log \mathbb{E}_\theta[p(y^*|x, \theta)] \leq & \mathbb{E}_\theta[-\log p(y^*|x, \theta)] \\ & - \inf_{\theta} \frac{1}{2p(y^*|x, \theta)^2} \mathbb{V}_\theta[p(y^*|x, \theta)] \end{aligned} \quad (7)$$

where $\inf_\theta \frac{1}{p(y^*|x,\theta)^2}$ is bounded given $p(y^*|x, \theta) \in [0, 1]$ and $\mathbb{V}_\theta[p(y^*|x, \theta)]$ is the variance of probabilistic ensemble model prediction.

$$\mathbb{V}_\theta[p(y^*|x, \theta)] = \mathbb{E}_\theta[(p(y^*|x, \theta) - \mathbb{E}_\theta[p(y^*|x, \theta)])^2] \quad (8)$$

Proposition 3.3 shows that the errors of the PE model are reduced compared to single models, which are also bounded by variance $\mathbb{V}_\theta[p(y^*|x, \theta)]$. The diversity measurement $\mathbb{V}_\theta[p(y^*|x, \theta)]$ can be applied for both regression and classification tasks since $p(y^*|x, \theta)$ is a scalar variable parameterized by θ given label y^*. With a larger variance, the upper bound of the negative log-likelihood (NLL) is reduced. As a result, we can theoretically show that enhancing diversity improves the prediction performance when $\mathbb{E}_\theta[-\log p(y^*|x, \theta)]$ remains similar. Moreover, Proposition 3.4 shows that PE has better diversity compared to deep ensemble method as the theoretical basis of the improved performance.

Proposition 3.4 (Enhanced diversity of PE)**.** *Let μ_D, Σ_D be the mean and covariance matrix of the deep ensemble representation $p_{DE}(\theta) = \sum_{i=1}^{N} \lambda_i \delta(\theta, \theta_i)$ where δ represents the delta function. Let μ_P, Σ_P be the mean and covariance matrix of $p_{PE}(\theta)$. We show that*

$$\mu_D = \mu_P \qquad \Sigma_P \geq \Sigma_D \quad (9)$$

where $\Sigma_P \geq \Sigma_D$ means $\Sigma_P - \Sigma_D$ is positive semi-definite. Compared to deep ensemble method, the PE model gains improved diversity.

Proposition 3.5 (Overconfidence reduction of PE)**.** *Given a probabilistic ensemble model with N components, let $f_{\theta_i} : R^{|x|} \to R^C$ be a ReLU network parameterized by θ_i. Let $|x|$ represent the dimension of x and $\theta \sim \sum_{i=1}^{N} \lambda_i \mathcal{N}(\theta; \theta_i, \Sigma_i)$. Then for any input x, the estimated probability based on multi-class probit approximation (see Appendix A.1) of the PE fulfills*

$$\lim_{\eta \to \infty} p_{PE}(y = c|\eta x) \leq \sum_{i=1}^{N} \frac{\lambda_i}{1 + \sum_{j \neq c} \exp\{-t_i^{(j)} - t_i^{(c)}\}} \quad (10)$$

where

$$t_i^{(k)} = \frac{||w_i^{(k)}||}{s_{min}(J_i^{(k)}) \sqrt{\frac{\pi}{8} \lambda_{min}(\Sigma_i)}} \quad k = 1, 2, \cdots, C$$

and $w_i = [w_i^{(1)}, w_i^{(2)}, \cdots, w_i^{(C)}] \in R^{|x| \times C}$ is a matrix that only depends on θ_i. $J_i^{(j)} = \frac{\partial w_i^{(j)}}{\partial \theta}|_{\theta=\theta_i}$ is the Jacobian matrix of $w_i^{(j)}$ at $\theta = \theta_i$. λ_{min} represents the minimum eigenvalue. s_{min} represents the minimum singular value.

Deterministic models suffer from the overconfidence issue such that the estimated probability is very high even if the input is far away from the data distribution. The Proposition 3.5 builds an upper bound for the predictive probability of samples $\{\eta x\}$ when $\eta \to \infty$, which prevents $p_{PE}(y = c|\eta x)$ to be extremely large. The Proposition 3.5 also shows that the confidence for far-away samples is upper bounded by the uncertainty estimated from LA. Especially, when the uncertainty is large, i.e., $\lambda_{min}(\Sigma_i) \to \infty, i = 1, 2, \cdots, N$, we have $\lim_{\eta \to \infty} p_{PE}(y = c|\eta x) \leq \frac{1}{C}$.

3.3 ADAPTIVE UNCERTAINTY-GUIDED ENSEMBLE LEARNING

The deep ensemble method trains ensemble models independently, which ignores the information obtained from previous models when getting a new one. This may cause knowledge redundancy that limits the diversity among ensemble models. The key idea of the proposed adaptive uncertainty-guided ensemble learning is to always make the new model focus on the regions which previous models have less explored, measured by uncertainty. Thus the new model will have the ability to provide complementary information to the previous models, which will improve the model averaging performance as well as implicitly enhance the diversity.

Given k trained deterministic models with parameters $\{\theta_i\}_{i=1}^{k}$, we perform the adaptive uncertainty-guided ensemble learning to get the $(k+1)$th model in the following steps. First, we construct the probabilistic ensemble illustrated in Sec. 3.2 such that $\theta \sim \sum_{i=1}^{k} \lambda_i \mathcal{N}(\theta; \theta_i, \Sigma_i)$. Then, the epistemic uncertainty $u(x)$ of each training data x is computed. Finally, we use the estimated epistemic uncertainty $u(x)$

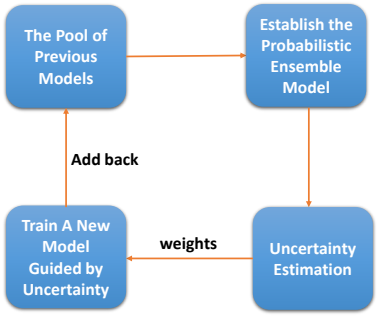


Figure 2: The Adaptive Uncertainty-guided Ensemble Learning Framework

from previous models as weights to guide the training of the $(k+1)$th model. Given a batch of data $\mathcal{D}_B = \{x_m, y_m\}_{m=1}^B$ of size B, the uncertainty-guided training loss can be expressed as

$$\mathcal{L}_{nll}(\theta) = -\frac{1}{B}\sum_{m=1}^{B} w(x_m)\log p(y_m|x_m,\theta) \tag{11}$$

where $w(x_m)$ is the weight for sample x_m as a function of $u(x_m)$, which is shown in Eq. (12).

$$w(x_m) = \frac{\exp(a*\log(u(x_m))+b)}{\sum_{j=1}^{B}\exp(a*\log(u(x_j))+b)} \tag{12}$$

Eq. (12) guarantees that samples with larger uncertainty will receive larger weights and the weights for a batch of data sum to 1. A log function is applied on $u(x_m)$ since the epistemic uncertainty is usually small. $a, b > 0$ are hyperparameters that can be tuned. The following propositions provide some theoretical evidence of the proposed method with proofs shown in Appendix C.

Proposition 3.6 (Prediction error bound)**.** *The prediction error is bounded by the total uncertainty. The epistemic uncertainty is positively correlated with the prediction error.*

Proposition 3.7 (Striking the right balance with uncertainty [Khan et al., 2019])**.** *For the imbalanced classification problems, minimizing the empirical loss results in a hypothesis that the classification boundary is towards the minority classes, leading to a larger classification region for the majority ones.*

Proposition 3.6 provides theoretical support for uncertainty-guided learning. By putting higher weights on problematic samples, the proposed method can reduce their uncertainty through adaptive ensemble learning to improve overall accuracy. Proposition 3.7 shows that a single model tends to sacrifice minority samples to obtain a good overall performance. It motivates us to adaptively learn complementary models focusing on minority samples, in order to achieve better ensemble performance.

Although our method is similar to boosting methods [Freund and Schapire, 1997, Hastie et al., 2009] in terms of reweighing the samples, they, however, are fundamentally different. As a discriminative model, boosting methods build an ensemble classifier by combining a set of weak classifiers to better classify the data. In contrast, we construct a generative ensemble that better models the posterior distribution of model parameters, through which we perform uncertainty quantification. Moreover, instead of using classification errors to weigh the samples, we use epistemic uncertainty to weigh the training samples. As epistemic uncertainty inversely measures training sample density, training of the next model will focus more on the samples that are not well represented by previous models.

3.4 MIXTURE OF GAUSSIAN REFINEMENT

In this section, we will establish an EM-like algorithm for refining the mixture of Gaussian parameters. To our knowledge, most of the ensemble methods assume that each ensemble component has the same importance, which may not be the case for real-world applications. Denote $\phi = \{\{\lambda_i\}_{i=1}^N, \{\theta_i\}_{i=1}^N, \{\Sigma_i\}_{i=1}^N\}$ as the mixture of Gaussian parameters, $\phi^0 = \{\{\lambda_i^0\}_{i=1}^N, \{\theta_i^0\}_{i=1}^N, \{\Sigma_i^0\}_{i=1}^N\}$ as the previous learned parameters before the refinement, and the training data as $\mathcal{D} = \{\mathcal{D}_m\}_{m=1}^M = \{x_m, y_m\}_{m=1}^M$. Let $Z \sim Cat(\lambda_1, \lambda_2, \ldots, \lambda_N)$ be the latent variable indicating membership of (x, y) belonging to which ensemble component. We learn non-uniform $\{\lambda_i\}_{i=1}^N$ and refine $\{\{\theta_i\}_{i=1}^N, \{\Sigma_i\}_{i=1}^N\}$ in the following EM steps.

E-step: construct the loss function $Q(\phi|\phi^0, \mathcal{D})$ as the expected value of the log-likelihood function of ϕ with respect to the current conditional distribution of Z given ϕ^0 and $\mathcal{D}$.

$$\begin{aligned}
\log p(\mathcal{D}|\phi) &= \sum_{m=1}^{M}\log p(\mathcal{D}_m|\phi) \\
&= \sum_{m=1}^{M}\log\sum_{i=1}^{N}\frac{p(Z=i|\mathcal{D}_m,\phi^0)}{p(Z=i|\mathcal{D}_m,\phi^0)}p(\mathcal{D}_m, Z=i|\phi) \\
&\geq \sum_{m=1}^{M}\sum_{i=1}^{N}p(Z=i|\mathcal{D}_m,\phi^0)\log\frac{p(\mathcal{D}_m, Z=i|\phi)}{p(Z=i|\mathcal{D}_m,\phi^0)} \\
&:= Q(\phi|\phi^0,\mathcal{D})
\end{aligned} \tag{13}$$

M-step: maximize $Q(\phi|\phi^0, \mathcal{D})$ with respect to ϕ.

$$\phi^* = \arg\max_{\phi} Q(\phi|\phi^0, \mathcal{D}) \tag{14}$$

Optimizing Eq. (14) returns a close-form expression of

$\{\lambda_i^*\}_{i=1}^N$.

$$\lambda_i^* = \frac{\sum_{m=1}^M p(Z=i|\mathcal{D}_m, \phi^0)}{\sum_{m=1}^M \sum_{j=1}^N p(Z=j|\mathcal{D}_m, \phi^0)} \tag{15}$$

Letting $p_m(\theta) = p(y_m|x_m, \theta)$,

$$p(Z=i|\mathcal{D}_m, \phi^0) = \frac{\lambda_i^0 \int p_m(\theta)\mathcal{N}(\theta; \theta_i^0, \Sigma_i^0)d\theta}{\sum_{j=1}^N \lambda_j^0 \int p_m(\theta)\mathcal{N}(\theta; \theta_j^0, \Sigma_j^0)d\theta} \tag{16}$$

Then given the distribution $Z \sim Cat(\{\lambda_i^*\}_{i=1}^N)$, we assign each data samples to its top l nearest components based on their weighted log-likelihood (i.e., $l = N/2$). The refinement is conducted by fine-tuning the existing ensemble components on the data samples they receive to further strengthen the specialty and diversity of each ensemble model. Details can be found in Appendix D.

3.5 PROBABILISTIC ENSEMBLE TRAINING STRATEGY

In this paper, three sub-modules are proposed: the probabilistic ensemble built by LA, the uncertainty-guided ensemble learning, and the mixture of Gaussian refinement. The pseudocode of the overall proposed method is shown in Algorithm 1, consisting of four steps. Although the final refinement step can further improve performance, it is not required.

During training, we admit that AUEL requires sequential training, which takes more time than parallel training. However, our methods can achieve similar UQ results with fewer ensemble components, compared to other ensemble baselines in Sec. 4.4. It could be more useful when there are limited capacities for parallel training or when there exist parallelly trained ensemble models with low diversity and we want to add a new one for providing complementary information. PE can be applied to any trained ensemble models with high efficiency. The last-layer LA is efficient whose complexity is $O(m + c^3 + p^3)$, where m, c, p represents the total number of parameters, the number of classes, and the number of last-layer parameters. For the inference complexity of PE, we can generate an arbitrary number of samples from the mixture of Gaussian. Compared to the deep ensemble method, the additional cost to obtain one more sample is $O(p)$, which is minimal since we only sample the last-layer parameters and reuse the intermediate outputs. More importantly, each sub-module can be applied to other ensemble methods separately to make further improvements. Although incorporating all sub-modules leads to the best performance, only applying PE could be an alternative way for efficient training.

The possible parallel training extensions may include: (1) Train one deterministic model using LA for UQ, then parallelly train other models with varying uncertainty-driven weights from Eq. 11 with different hyperparameters a, b; (2) Train all models in parallel, compute LA for each near completion to build PE, and use uncertainty-guided weights to refine the models in their final training phase. We will investigate those possibilities in our future research. It is also worth noting that the proposed method can be applied to autoregressive ensemble training Havasi et al. [2020], Dusenberry et al. [2020]. Uncertainty-guided weights can promote diversity in MIMO sub-networks, and LA can be used to construct a probabilistic ensemble model after training.

Algorithm 1 Probabilistic Ensemble with Adaptive Uncertainty-guided Ensemble Learning

Input: Training data $\mathcal{D} = \{x_m, y_m\}_{m=1}^M$. Initialize the model pool $P = \{\}$
Ouput: The probabilistic ensemble model parameters $\theta \sim \sum_{i=1}^N \lambda_i \mathcal{N}(\theta; \theta_i, \Sigma_i)$
Step 1 (single model): Train the first model using NLL loss to obtain θ_1; $P = P + \{\theta_1\}$
Step 2 (AUEL): Perform the adaptive uncertainty-guided ensemble learning to obtain $\{\theta_i\}_{i=2}^N$
for $k = 2 : N$ **do**
(1) Given P, construct the probabilistic ensemble model with uniform weights.
$\theta \sim p_{k-1}(\theta) = \frac{1}{k-1} \sum_{i=1}^{k-1} \mathcal{N}(\theta; \theta_i, \Sigma_i)$
(2) Estimate the epistemic uncertainty $\{u(x_m)\}_{m=1}^M$ using $\theta \sim p_{k-1}(\theta)$
(3) Use the weighted loss in Eq. (11) to train θ_k
(4) Update the model pool: $P = P + \{\theta_k\}$
end for
Step 3 (AUEL+PE): Based on current $P = \{\theta_i\}_{i=1}^N$, construct the probabilistic ensemble model
Step 4 (AUEL+RPE): We refine the Gaussian mixture model parameters based on Sec. 3.4

4 EXPERIMENT

4.1 OUT-OF-DISTRIBUTION DETECTION

Out-of-distribution (OOD) detection tries to detect anomalous data that is inconsistent with the training data distribution. Utilizing epistemic uncertainty as a measure for out-of-distribution detection is one of the major applications for demonstrating the quality of UQ performance. We evaluate our methods on benchmark image classification datasets MNIST [Deng, 2012] and CIFAR-10 (C10) [Krizhevsky et al., 2014], respectively. We choose Omniglot [Lake et al., 2015], EMNIST [Cohen et al., 2017], and KMNIST [Clanuwat et al., 2018] as OOD datasets for MNIST. For C10 dataset, the SVHN [Netzer et al., 2011], LSUN [Yu et al., 2015], and CIFAR-100 (C100) [Krizhevsky et al., 2009] are the OOD datasets. We compare our proposed

Table 1: OOD Detection Results for AUROC (%) and AUPR (%) on MNIST-related and C10-related Datasets with Epistemic Uncertainty. Each experiment result is aggregated over 3 independent runs.

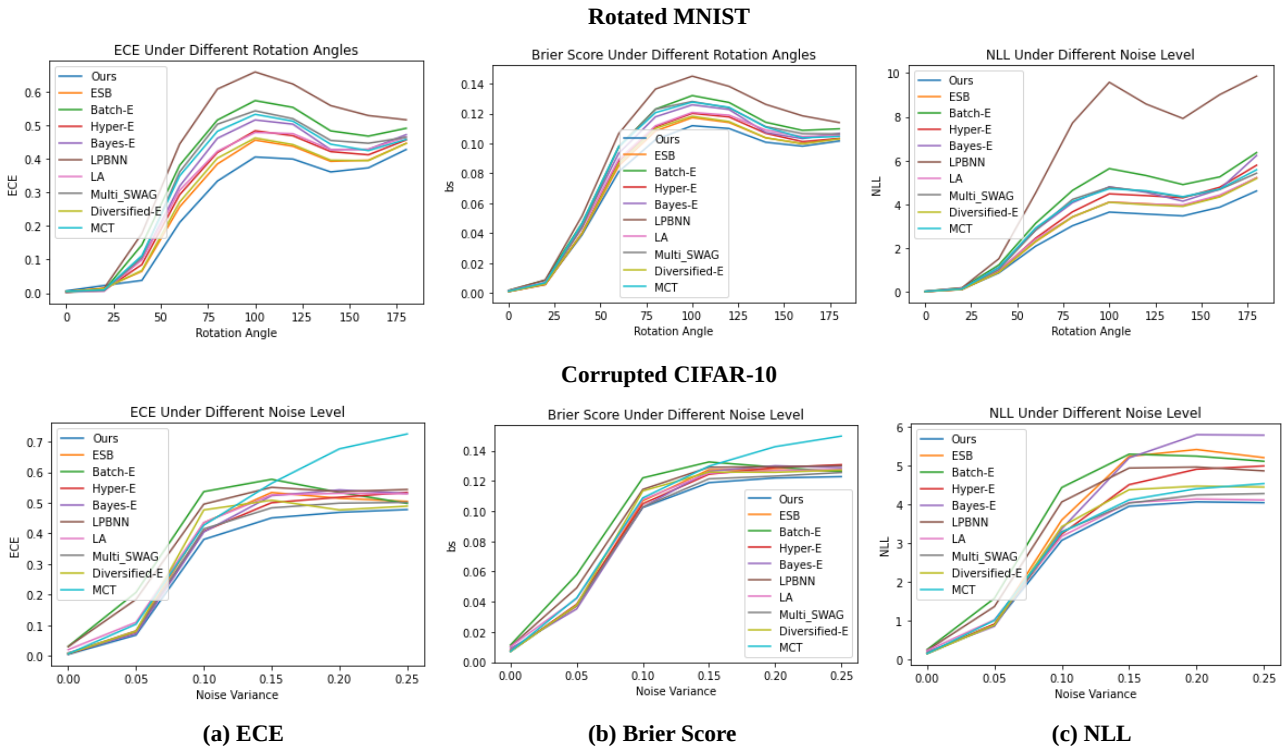
Method	MNIST → Omniglot		MNIST → EMNIST		MNIST→ KMNIST	
	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR
Ours	**98.49** ± 0.01	**98.23** ± 0.03	**98.01** ± 0.07	**97.26** ± 0.08	**98.39** ± 0.11	**97.98** ± 0.08
ESB	97.92 ± 0.25	97.33 ± 0.34	97.32 ± 0.14	96.10 ± 0.46	97.92 ± 0.10	97.13 ± 0.27
Batch-E	95.95 ± 0.17	94.74 ± 0.22	95.79 ± 0.70	93.76 ± 0.75	96.59 ± 0.45	94.72 ± 0.43
Hyper-E	97.97 ± 0.26	97.55 ± 0.24	97.56 ± 0.31	96.68 ± 0.51	97.92 ± 0.43	97.32 ± 0.53
Bayes-E	97.42 ± 0.28	96.94 ± 0.46	97.07 ± 0.29	95.86 ± 0.33	97.73 ± 0.06	96.72 ± 0.14
LPBNN	95.94 ± 0.52	94.41 ± 0.57	92.84 ± 0.69	92.54 ± 0.39	97.40 ± 0.71	95.96 ± 0.95
LA	97.87 ± 0.39	97.49 ± 0.37	97.72 ± 0.48	97.02 ± 0.44	98.11 ± 0.19	97.54 ± 0.17
Multi-SWAG	96.52 ± 0.37	94.56 ± 0.84	95.81 ± 0.60	90.64 ± 1.70	96.70 ± 0.42	94.34 ± 0.98
Diversified-E	97.92 ± 0.19	97.21 ± 0.23	94.40 ± 0.16	96.21 ± 0.37	97.93 ± 0.12	97.01 ± 0.32
MCT	97.04 ± 0.34	95.62 ± 0.93	96.65 ± 0.46	95.61 ± 0.82	97.31 ± 0.10	95.81 ± 0.56
Method	C10 → SVHN		C10 → LSUN		C10→ C100	
	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR
Ours	93.88 ± 0.57	90.58 ± 1.58	**89.57** ± 0.08	**86.81** ± 0.14	**93.80** ± 0.11	91.67 ± 0.36
ESB	91.23 ± 1.35	86.16 ± 1.73	88.42 ± 0.85	84.99 ± 0.65	91.87 ± 0.58	88.69 ± 0.55
Batch-E	90.40 ± 1.62	85.12 ± 2.64	86.10 ± 0.24	81.42 ± 0.40	90.15 ± 0.18	85.48 ± 0.49
Hyper-E	91.11 ± 0.32	85.86 ± 0.46	88.82 ± 0.15	85.29 ± 0.25	92.59 ± 0.24	89.65 ± 0.71
Bayes-E	90.96 ± 3.35	86.57 ± 5.27	87.85 ± 1.22	84.56 ± 1.01	91.80 ± 0.45	88.83 ± 0.02
LPBNN	89.99 ± 2.44	85.18 ± 4.00	86.87 ± 0.01	82.14 ± 0.49	90.80 ± 0.22	85.62 ± 1.59
LA	93.39 ± 0.46	91.17 ± 0.98	87.27 ± 0.19	85.77 ± 0.21	93.45 ± 1.17	**92.59** ± 1.47
Multi-SWAG	**94.06** ± 0.54	**93.92** ± 0.59	87.23 ± 0.29	85.44 ± 0.61	90.24 ± 0.86	88.05 ± 1.02
Diversified-E	92.56 ± 1.36	88.04 ± 3.29	89.06 ± 0.09	85.53 ± 0.20	92.90 ± 0.07	90.01 ± 0.16
MCT	91.04 ± 0.44	84.73 ± 0.35	88.71 ± 0.16	84.86 ± 0.18	92.18 ± 0.03	88.67 ± 0.24

method (AUEL+PE) with general ensemble-based methods (i.e., ESB [Lakshminarayanan et al., 2017b], Batch-E [Wen et al., 2020], Bayes-E [Pearce et al., 2018]), Diversity-promoted ensemble methods (Hyper-E [Wenzel et al., 2020], Multi-SWAG [Wilson and Izmailov, 2020], Diversified-E [Zhang et al., 2020], MCT [Lee et al., 2015]), and approximate BNNs (i.e., LPBNN [Franchi et al., 2020], LA). We exclude sequential ensemble methods (EDST [Liu et al., 2021], SeBayS [Jantre et al., 2022]) and other mixture posterior approximation methods (MIMO [Havasi et al., 2020], Rank-1 BNN [Dusenberry et al., 2020]) for comparison since they are shown to perform worse than ESB method. The evaluation metrics include the area under the receiver operating characteristic curve (AUROC ↑) and the area under the precision-recall curve (AUPR ↑). All ensemble-based methods have size 5. The experiment settings and implementation details can be found in Appendix E.

The out-of-distribution detection performance is shown in Table 1. It is obvious that the proposed method (AUEL+PE) can achieve significant improvement over recent ensemble-based methods on various OOD detection tasks. Additional OOD detection experiments are shown in Sec. 4.2 for MNIST and C10 under different levels of distributional shifts. Since the post-processing refinement of the MoG is not required, we will show the effectiveness of the refinement (AUEL+RPE) in Sec. 4.4. Compared to diversity-enhanced ensemble learning such as Hyper-E and Multi-SWAG, our better OOD detection performance also indicates enhanced diversity.

4.2 IMAGE CLASSIFICATION UNDER DISTRIBUTIONAL SHIFT

Bayesian models marginalize all possible solutions for the final prediction, leading to improved robustness. In this section, we will demonstrate the effectiveness of the proposed method for image classification tasks on MNIST and C10 with synthetic distributional shifts. For MNIST, we create the synthetic rotated MNIST dataset, where we increasingly rotate the MNIST testing data from 0° to 180° with a step of 20°. For the C10 dataset, we add the Gaussian noise with 0 mean and variance ranging from 0 to 0.25 with a step of 0.05 to the testing data as the corrupted C10 dataset. Additional adversarial shifts can be found in Sec. 4.4. Note that we keep the original training strategy on MNIST/C10 training data but test on the shifted testing data. During the evaluation, the uncertainty calibration metrics include negative log-likelihood (NLL ↓), accuracy (ACC ↑), expected calibration error (ECE ↓), maximum calibration error (MCE ↓), and brier score (BS ↓). We also provide the OOD detection results of MNIST → Rotated MNIST and C10 → Corrupted C10 in terms of AUROC and AUPR. The comparisons are



(a) ECE **(b) Brier Score** **(c) NLL**

Figure 3: Predictive Calibration Analysis of Rotated MINST and Corrupted C10 Datasets. The first row shows the results for MNIST while the second row represents C10. There are three different metrics (ECE, Brier Score, NLL) that are analyzed in each column, respectively. Each experiment result is aggregated over 3 independent runs.

conducted under the same experiment settings as Sec. 4.1. In Figure 3, partial results for ECE, BS, and NLL are shown. Additional analysis can be found in Appendix F.

Based on Figure 3, we can observe that the probabilistic ensemble method can achieve better calibration performance for both rotated MNIST and corrupted C10 datasets. As the shift level increases, our proposed method consistently outperforms other ensemble-based methods, which demonstrates the great potential of our method in better generalization ability. Besides improved robustness of uncertainty quantification, comparable within-dataset performance can be found in Appendix F.1.

Table 2: Diversity Analysis of Ensemble-based Methods Trained on C10 dataset

Method	QS	BD	FK	KW
Ours	**0.174**	**0.538**	**0.383**	**0.857**
ESB	0.185	0.552	0.404	0.860
Batch-E	0.284	0.576	0.422	0.868
Hyper-E	0.199	0.554	0.406	0.861
Bayes-E	**0.174**	0.548	0.406	0.859
LPBNN	0.209	0.557	0.405	0.862
Multi-SWAG	0.246	0.566	0.418	0.865
Diversified-E	0.184	0.552	0.402	0.860
MCT	0.190	0.553	0.405	0.861

4.3 DIVERSITY ANALYSIS

In addition to the theoretical confirmation of augmented diversity exhibited in Sections 3.2 and 3.3, we also supply empirical analysis underscoring the diversity benefits derived from our proposed methodology. The diversity metrics we employ originate from Wu et al. [2020], featuring both pairwise diversity measures like Q Statistics (QS) and Binary Disagreement (BD), and non-pairwise metrics such as Fleiss' Kappa (FK) and Kohavi-Wolpert Variance (KW).

We undertake a normalization process for all these scores to ensure that lower values (↓) signify a higher degree of diversity. As illustrated in Table 2, our approach surpasses other ensemble-based techniques in relation to diversity.

It is crucial to highlight that our proposed method leverages uncertainty-guided learning via AUEL to generate diverse modes, utilizes LA for neighborhood exploration to yield diverse samples, and employs an EM-like refinement strategy to further boost diversity. On the other hand, the baseline methods generally concentrate on fostering diversity in a single area.

Table 3: Ablation Studies: OOD Detection Results and Robustness Analysis on MNIST/C10 Datasets. The first table shows the effectiveness of the sub-modules. The second table shows the improvement when PE serves as a plug-and-play module. Each experiment result is aggregated over 3 independent runs.

Method	MNIST → Omniglot		C10 → SVHN		Rotated MNIST 60°		Noisy C10 Level 0.1	
	AUROC	AUPR	AUROC	AUPR	NLL	ECE	NLL	ECE
Ensemble	97.92	97.33	91.23	86.16	2.30	0.256	3.58	0.435
AUEL	98.02	97.50	92.98	88.97	2.24	0.243	3.28	0.394
AUEL+PE	98.49	98.23	93.88	90.58	2.09	0.210	3.06	0.380
AUEL+RPE	**98.95**	**98.90**	**93.93**	**91.93**	**1.92**	**0.163**	**3.02**	**0.375**

Method	MNIST → Omniglot		C10 → SVHN		Rotated MNIST 120°		Noisy C10 Level 0.1	
	AUROC	AUPR	AUROC	AUPR	NLL	ECE	NLL	ECE
Hyper-E	97.97	97.55	91.11	85.86	4.40	0.468	3.23	0.407
Hyper-E + PE	**98.56**	**98.37**	**92.09**	**87.52**	**3.72**	**0.416**	**2.66**	**0.342**
Bayes-E	97.42	96.94	90.96	86.57	4.55	0.502	3.36	0.404
Bayes-E + PE	**98.21**	**98.02**	**93.27**	**90.41**	**3.68**	**0.450**	**2.40**	**0.298**

Aside from numerical findings, we provide visualizations of both parameter space and prediction space diversity in Appendix G. Essentially, we represent the neural network parameters and the predictive logits for MNIST testing data within a two-dimensional space, utilizing principal component analysis (PCA).

4.4 ABLATION STUDIES AND FURTHER ANALYSIS

Effectiveness of Sub-modules. In this section, we evaluate the effectiveness of each step illustrated in Algorithm 1. Each proposed sub-module helps further improve the OOD detection and uncertainty calibration performance. The MNIST and CIFAR-10 related experiments are shown in Table 3. More analysis for various experiment settings with different metrics can be found in Appendix H.1.

Probabilistic Ensemble as a Plug-and-Play Module. Our method can be a plug-and-play module for easily applying to other ensemble methods with further improvements. Given trained ensemble models from other ensemble methods, we can apply the PE module to construct the mixture of Gaussian model in the post-processing way. For example, we combine the Hyper Ensemble (Hyper-E) with PE and the Bayesian Ensemble (Bayes-E) with PE to show further improvements in Table 3. Additional analysis can be found in Appendix H.2.

Efficiency Analysis. In Appendix H.3, we present a thorough theoretical and practical evaluation of our methodology's efficiency against various ensemble baselines. In addition, we extend our analysis to compare ensemble baseline models with varying numbers of components. The findings demonstrate that our approach necessitates a smaller number of ensemble components to reach comparable outcomes.

Application to Larger Datasets. In Appendix I, we demonstrate the suitability of our techniques for handling larger datasets, such as CIFAR-100 and TinyImagenet [CS231N, 2017]. Our approach can effectively scale with a large number of parameters in the last layer. We can utilize diagonal or block-diagonal covariance matrices for LA, which scale impressively while maintaining competitive accuracy, as per [Daxberger et al., 2021a].

Other Distributional Shifts. In Appendix J, we perform adversarial perturbations on C10 testing dataset using the fast gradient sign method [Goodfellow et al., 2014]. Then, we compute the ACC and NLL of our proposed methods on the perturbed images compared to various ensemble baselines, indicating the effectiveness of our method against adversarial attacks.

Synthetic Experiments. In Appendix K, we provide some toy examples of the one-dimensional regression problem and the two-moon classification problem. These examples show that the estimated epistemic uncertainty of the PE model inversely matches well with the training data density.

5 CONCLUSION

In this paper, we propose the probabilistic ensemble method with adaptive uncertainty-guided ensemble training to construct the Gaussian mixture model with learnable and refinable parameters. Both theoretical and empirical evidence is provided to show that our proposed method can achieve a better approximation of the posterior distribution with enhanced diversity. Moreover, the proposed method has demonstrated better uncertainty quantification performance as well as improved uncertainty calibration ability for various applications including out-of-distribution detection and image classification under different distributional shifts.

References

Yijun Bian and Huanhuan Chen. When does diversity help generalization in classification ensembles? *IEEE Transactions on Cybernetics*, 2021.

Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical japanese literature. *arXiv preprint arXiv:1812.01718*, 2018.

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. Emnist: Extending mnist to handwritten letters. In *2017 international joint conference on neural networks (IJCNN)*, pages 2921–2926. IEEE, 2017.

Stanford CS231N. Tiny imagenet. 2017. URL `https://tiny-imagenet.herokuapp.com/`.

Francesco D'Angelo and Vincent Fortuin. Repulsive deep ensembles are bayesian. *Advances in Neural Information Processing Systems*, 34:3451–3465, 2021.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux-effortless bayesian deep learning. *Advances in Neural Information Processing Systems*, 34, 2021a.

Erik Daxberger, Eric Nalisnick, James U Allingham, Javier Antorán, and José Miguel Hernández-Lobato. Bayesian deep learning via subnetwork inference. In *International Conference on Machine Learning*, pages 2510–2521. PMLR, 2021b.

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

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*, pages 1184–1193. PMLR, 2018.

Michael Dusenberry, Ghassen Jerfel, Yeming Wen, Yian Ma, Jasper Snoek, Katherine Heller, Balaji Lakshminarayanan, and Dustin Tran. Efficient and scalable bayesian neural nets with rank-1 factors. In *International conference on machine learning*, pages 2782–2792. PMLR, 2020.

Nikita Dvornik, Cordelia Schmid, and Julien Mairal. Diversity with cooperation: Ensemble methods for few-shot classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3723–3731, 2019.

Runa Eschenhagen, Erik Daxberger, Philipp Hennig, and Agustinus Kristiadi. Mixtures of laplace approximations for improved post-hoc uncertainty in deep learning. *arXiv preprint arXiv:2111.03577*, 2021.

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

Gianni Franchi, Andrei Bursuc, Emanuel Aldea, Séverine Dubuisson, and Isabelle Bloch. Encoding the latent posterior of bayesian neural networks for uncertainty quantification. *arXiv preprint arXiv:2012.02818*, 2020.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55 (1):119–139, 1997.

Andrew Gelman. Induction and deduction in bayesian data analysis. 2011.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Trevor Hastie, Saharon Rosset, Ji Zhu, and Hui Zou. Multi-class adaboost. *Statistics and its Interface*, 2(3):349–360, 2009.

Marton Havasi, Rodolphe Jenatton, Stanislav Fort, Jeremiah Zhe Liu, Jasper Snoek, Balaji Lakshminarayanan, Andrew M Dai, and Dustin Tran. Training independent subnetworks for robust prediction. *arXiv preprint arXiv:2010.06610*, 2020.

Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E Hopcroft, and Kilian Q Weinberger. Snapshot ensembles: Train 1, get m for free. *arXiv preprint arXiv:1704.00109*, 2017.

Siddhartha Jain, Ge Liu, Jonas Mueller, and David Gifford. Maximizing overall diversity for improved uncertainty estimates in deep ensembles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4264–4271, 2020.

Sanket Jantre, Sandeep Madireddy, Shrijita Bhattacharya, Tapabrata Maiti, and Prasanna Balaprakash. Sequential bayesian neural subnetwork ensembles. *arXiv preprint arXiv:2206.00794*, 2022.

Alex Kendall and Yarin Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? (Nips), 2017. URL `http://arxiv.org/abs/1703.04977`.

Salman Khan, Munawar Hayat, Syed Waqas Zamir, Jianbing Shen, and Ling Shao. Striking the right balance with uncertainty. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition, pages 103–112, 2019.

Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Being bayesian, even just a bit, fixes overconfidence in relu networks. In *International Conference on Machine Learning*, pages 5436–5446. PMLR, 2020.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. The cifar-10 dataset. *online: http://www. cs. toronto. edu/kriz/cifar. html*, 55(5), 2014.

Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6402–6413. Curran Associates, Inc., 2017a. URL http://papers.nips.cc/paper/7219-simple-and-scalable-predictive-uncertainty-estimation-using-deep-ensembles.pdf.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6402–6413. Curran Associates, Inc., 2017b. URL http://papers.nips.cc/paper/7219-simple-and-scalable-predictive-uncertainty-estimation-using-deep-ensembles.pdf.

Jongseok Lee, Matthias Humt, Jianxiang Feng, and Rudolph Triebel. Estimating model uncertainty of neural networks in sparse information form. In *International Conference on Machine Learning*, pages 5702–5713. PMLR, 2020.

Stefan Lee, Senthil Purushwalkam, Michael Cogswell, David Crandall, and Dhruv Batra. Why m heads are better than one: Training a diverse ensemble of deep networks. *arXiv preprint arXiv:1511.06314*, 2015.

Shiwei Liu, Tianlong Chen, Zahra Atashgahi, Xiaohan Chen, Ghada Sokar, Elena Mocanu, Mykola Pechenizkiy, Zhangyang Wang, and Decebal Constantin Mocanu. Deep ensembling with no overhead for either training or testing: The all-round blessings of dynamic sparsity. *arXiv preprint arXiv:2106.14568*, 2021.

Yong Liu and Xin Yao. Ensemble learning via negative correlation. *Neural networks*, 12(10):1399–1404, 1999.

David J. C. MacKay. A Practical Bayesian Framework for Backpropagation Networks. *Neural Computation*, 4(3):448–472, 1992. ISSN 0899-7667. doi: 10.1162/neco.1992.4.3.448. URL http://www.mitpressjournals.org/doi/10.1162/neco.1992.4.3.448.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

Tim Pearce, Felix Leibfried, Alexandra Brintrup, Mohamed Zaki, and Andy Neely. Uncertainty in neural networks: Approximately bayesian ensembling. 2018.

Hippolyt Ritter, Aleksandar Botev, and David Barber. A scalable laplace approximation for neural networks. In *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, 2018.

Wei Tan, Lan Du, and Wray Buntine. Diversity enhanced active learning with strictly proper scoring rules. *Advances in Neural Information Processing Systems*, 34, 2021.

Aleksei Tiulpin and Matthew B Blaschko. Greedy bayesian posterior approximation with deep ensembles. *arXiv preprint arXiv:2105.14275*, 2021.

Matias Valdenegro-Toro. Deep sub-ensembles for fast uncertainty estimation in image classification. *arXiv preprint arXiv:1910.08168*, 2019.

Maxime Wabartha, Audrey Durand, Vincent Francois-Lavet, and Joelle Pineau. Handling black swan events in deep learning with diversely extrapolated neural networks. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 2140–2147, 2021.

Yeming Wen, Dustin Tran, and Jimmy Ba. Batchensemble: an alternative approach to efficient ensemble and lifelong learning. *arXiv preprint arXiv:2002.06715*, 2020.

Florian Wenzel, Jasper Snoek, Dustin Tran, and Rodolphe Jenatton. Hyperparameter ensembles for robustness and uncertainty quantification. *arXiv preprint arXiv:2006.13570*, 2020.

Andrew G Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. *Advances in neural information processing systems*, 33:4697–4708, 2020.

Yanzhao Wu, Ling Liu, Zhongwei Xie, Juhyun Bae, Ka-Ho Chow, and Wenqi Wei. Promoting high diversity ensemble learning with ensemblebench. In *2020 IEEE Second International Conference on Cognitive Machine Intelligence (CogMI)*, pages 208–217. IEEE, 2020.

Yanzhao Wu, Ling Liu, Zhongwei Xie, Ka-Ho Chow, and Wenqi Wei. Boosting ensemble accuracy by revisiting ensemble diversity metrics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16469–16477, 2021.

Shingo Yashima, Teppei Suzuki, Kohta Ishikawa, Ikuro Sato, and Rei Kawakami. Feature space particle inference for neural network ensembles. *arXiv preprint arXiv:2206.00944*, 2022.

Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *CoRR*, abs/1506.03365, 2015. URL `http://dblp.uni-trier.de/db/journals/corr/corr1506.html#YuZSSX15`.

Gabriel Zaid, Lilian Bossuet, Amaury Habrard, and Alexandre Venelli. Efficiency through diversity in ensemble models applied to side-channel attacks:–a case study on public-key algorithms–efficiency through diversity in ensemble models applied to side-channel attacks:–a case study on public-key algorithms–. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, pages 60–96, 2021.

Sheheryar Zaidi, Arber Zela, Thomas Elsken, Chris Holmes, Frank Hutter, and Yee Whye Teh. Neural ensemble search for uncertainty estimation and dataset shift. 2020.

Shaofeng Zhang, Meng Liu, and Junchi Yan. The diversified ensemble neural network. *Advances in Neural Information Processing Systems*, 33:16001–16011, 2020.

Tianyi Zhou, Shengjie Wang, and Jeff A Bilmes. Diverse ensemble evolution: Curriculum data-model marriage. *Advances in Neural Information Processing Systems*, 31, 2018.