Mixtures of Locally Bounded Langevin dynamics for Bayesian Model Averaging

Anonymous authors
Paper under double-blind review

Abstract

Properties of probability distributions change when going from low to high dimensions, to the extent that they admit counterintuitive behavior. Gaussian distributions intuitively illustrate a well-known effect of moving to higher dimensions, namely that the typical set almost surely does not contain the mean, which is the distribution's most probable point. This can be problematic in Bayesian Deep Learning, as the samples drawn from the high-dimensional posterior distribution are often used as Monte Carlo samples to estimate the integral of the predictive distribution. Here, the predictive distribution will reflect the behavior of the samples and, therefore, of the typical set. For instance, we cannot expect to sample networks close to the maximum a posteriori estimate after fitting a Gaussian approximation to the posterior using the Laplace method. In this paper, we introduce a method that aims to mitigate this typicality problem in high dimensions by sampling from the posterior with Langevin dynamics on a restricted support enforced by a reflective boundary condition. We demonstrate how this leads to improved posterior estimates by illustrating its capacity for fine-grained out-of-distribution (OOD) ranking on the Morpho-MNIST dataset.

1 Introduction

Estimation of epistemic uncertainty during inference time is crucial for many applications in machine learning. It is a key ingredient for identifying samples for which the algorithm's prediction cannot be trusted, which include both out-of-distribution (OOD) samples as well as challenging in-distribution (ID) samples. Even more so, an accurate estimation of epistemic uncertainty can play a crucial role in identifying domain shifts.

To this end, given a finite sample size and an overparameterized neural network model, which is often the case, identifying different sets of "optimal weights" is key to estimating epistemic uncertainty of the model. We find it important to explicitly define the "optimality" of these sets of weights. In this work, we adopt the following definition. If two sets of weights are optimal, they are both local minima of the loss function and yield similarly high accuracies on the training and/or validation sets. As such, it would not be possible to choose one over the other; thus, they are both optimal. Evaluating these different optimal models on the same sample would allow for approximating the epistemic uncertainty.

The current set of methods for estimating epistemic uncertainty can be divided into two groups. First, *the discrete support group*, estimates a finite number of parameter sets either through training multiple models and ensembling Lakshminarayanan et al. (2017), approximating ensembling through Monte-Carlo (MC) Dropout Gal & Ghahramani (2016), or directly sampling from a posterior distribution using an appropriate sampling, such as Hamiltonian Monte CarloBetancourt (2017). The second group, *the continuous support group*, estimates continuous posterior distributions for network parameters through using Laplace approximations MacKay (1992); Ritter et al. (2018); Daxberger et al. (2021) or variational inference Graves (2011); Bishop & Nasrabadi (2006).

Setting model ensembling aside for a moment, the current techniques rely on an underlying Bayesian model and posterior distribution of network weights given a dataset. This posterior is either approximated or sampled from. However, even in small networks, the number of parameters is very high, and therefore,

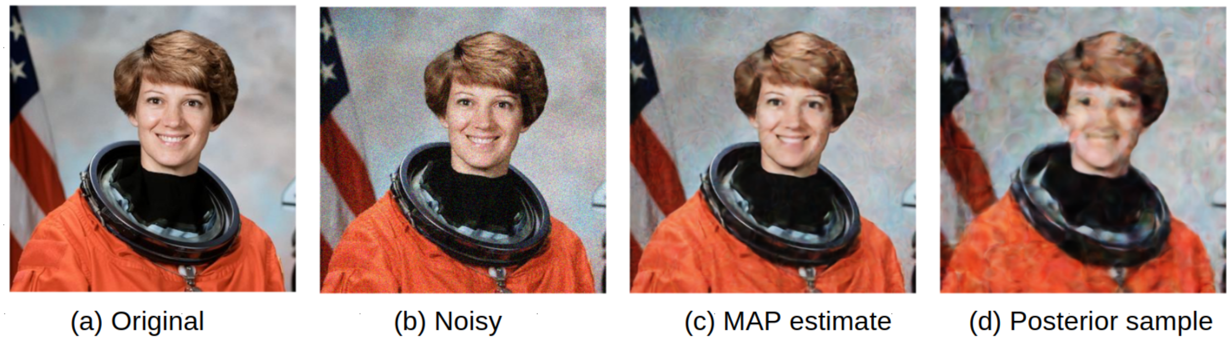

Figure 1: Illustration from Cagnotti (2023): A Deep decoder is trained to denoise the image shown in (b). (c) shows a MAP estimate obtained by training with Equation 4 using a Gaussian prior for weights, i.e., weight decay. (d) shows reconstruction with a sample drawn with MALA from the posterior distribution of the network weights after making sure the chain has converged. As explained below, this posterior sample likely comes from the typical set, and the reconstruction quality is clearly lower than the MAP estimate.

the underlying Bayesian model is very high-dimensional. As such, these Bayesian models are prone to counterintuitive effects of *typical sets* in high dimensions. An intuitive description of a typical set is given by Carpenter (2017) as *"...the central log density band into which almost all random draws from that distribution will fall"*. The main issue is that this band may be quite far from the modes of a distribution in high dimensions Kirby (1969). Thus, drawn samples will come from areas where the likelihood is potentially extremely low. When we consider the posterior distribution of network weights, this means sampled sets of weights may not be optimal as we defined above. They most likely have very low data likelihoods or in other words, they will not do an accurate prediction on the training set. Cagnotti (2023) et al. demonstrated this on the problem of image denoising with the Deep Decoder model Heckel & Hand (2019). Denoising an image with network weights that are sampled from the corresponding posterior distribution, using a Metropolis-adjusted Langevin algorithm (MALA), yielded much worse results compared to the Maximum-a-Posteriori (MAP) estimate, which is illustrated in Figure 1.

Model ensembling Lakshminarayanan et al. (2017) notably does not share this issue since the different sets of weights are all optimized and thus constitute modes in the posterior distribution when we view this approach from a Bayesian perspective using a flat prior Wilson & Izmailov (2020); Gustafsson et al. (2020); Pearce et al. (2018). While this is a good step, the small number of models that can be extracted with this approach may be underestimating the true epistemic uncertainty, as there can be many more optimal weight sets than the number of models in the ensemble. One direction towards this is to use Laplace approximations for each model in the ensemble Eschenhagen et al. (2021). Although the approach shows promising experimental results, the above-described problem of sampling in high dimensions persists here.

In this work, we propose a middle way between discrete and full continuous support by defining a mixture of *bounded support regions.* We propose to approximate the posterior locally around the weight configurations found by a deep ensemble using *locally bounded MALA* (lbMALA). We do so by defining a fixed-width hyperball around each ensemble weight parameter and setting a reflective boundary condition, effectively reducing the search space and support of the distribution to which the Markov Chain converges. While the normal distribution found by the standard Laplace approximation might struggle to pick up the complex statistical dependencies between weights, lbMALA inherently takes such dependencies into account.

To test whether lbMALA indeed leads to improved posterior estimates, we design a fine-grained OOD detection validation benchmark. This measures how sensitive different epistemic uncertainty methods are to samples that move gradually out of distribution. Finally, through the improved performance of lbMALA on this benchmark, we provide evidence that the typical set problem is indeed real and affecting the uncertainty estimates of state-of-the-art (SOTA) models.

2 Related Work

In probabilistic machine learning, the effects of concentration of measure and typicality appear both in the data and parameter space. The relevance is apparent: with larger models and higher-dimensional data, such as higher-resolution images, the studied spaces grow in dimension, reinforcing the issue. Nalisnick et al. (2018) and Choi et al. (2018) find that deep generative models might assign higher likelihoods to OOD data than to their training data. Based on this observation Nalisnick et al. (2019) argues that high-likelihood samples might not be part of the typical set of the high-dimensional distribution of images. They developed an OOD test based on typical set membership, where the entropy of new samples is compared to the entropy of the source distribution using the condition in equation 6. However, Zhang et al. (2021) finds the test unreliable, and Osada et al. (2024) attributes this to varying image complexity. Grathwohl et al. (2019) proposes a score that argues that data points with a high likelihood outside the typical set should have a higher gradient norm than ID samples. Recently, Abdi et al. (2024) has applied the above methods to a medical imaging setting with promising results in OOD detection.

The effects of typicality for Gaussian posterior approximations given by mean field variational inference in the parameter space of a neural network have been observed and discussed by Farquhar et al. (2020) and Farquhar (2022), which they termed "*soap bubble*" pathology. They propose an alternative probability distribution based on hyperspherical coordinates that forces probability mass to be close to the mean. However, such probability concentration may be 'artificial' as samples are biased towards the mean with less exploration.

The combination of multiple MAP estimates from a deep ensemble with local approximations of the posterior around each estimate is a well-established technique. Eschenhagen et al. (2021) use post-hoc Laplace approximations around independently trained neural networks, resulting in a Gaussian mixture model. The marginal likelihood is used to weight each distribution. Wilson & Izmailov (2020) extend Stochastic Weight Averaging (SWA) Izmailov et al. (2018) and Stochastic Weight Averaging Gaussian (SWAG) Maddox et al. (2019) to multiple neural networks of an ensemble. SWA averages the weights collected at different points, usually in later epochs, during training with a constant or cyclic learning rate. MultiSWA Wilson & Izmailov (2020) uses an ensemble of models found with SWA. SWAG extends SWA by not only averaging the weights but also fitting a Gaussian distribution to them, capturing the posterior distribution locally. MultiSWAG Wilson & Izmailov (2020) involves training multiple neural networks independently and then applying SWAG to each of these networks.

The Bayesian supervised learning framework distinguishes two types of uncertainty: Aleatoric, which captures the inherent variation in the data, and epistemic uncertainty, which informs about a trained model's state of knowledge in a certain area of the domain of the network. Different theoretically derived uncertainty measures capture aleatoric and epistemic uncertainty, and it is an open research question which of those measures captures the different types of uncertainty best Schweighofer et al. (2023a); Wimmer et al. (2023); Zepf et al. (2024); Schweighofer et al. (2023b). The success of the chosen uncertainty measure can depend on the application; therefore, mutual information, variance in predictions, the deviation of the posterior from the prior and predictive entropy Abdar et al. (2021), could vary in their suitability for downstream tasks like OOD detection. In addition, a strong correlation between measures of aleatoric and epistemic uncertainty has been found, posing the question whether decomposition is possible Kahl et al. (2024). Therefore, in practice, one often relies on the combined total uncertainty to circumvent the problem of decomposing uncertainties Yang et al. (2024), especially when either the likelihood or the Shannon entropy of the predictive distribution is the most frequently used measure for total uncertainty in the Bayesian framework, which in practice is approximated by sampling using the Monte-Carlo method.

3 Method: Mixtures of locally bounded Langevin dynamics

3.1 Bayesian Deep Learning

The Bayesian framework for supervised learning assumes a data-generating process

$$(x,y)_i \overset{\text{i.i.d.}}{\sim} p(x,y), \quad i = 1,\dots,N \tag{1}$$

from which a dataset $D = (x,y)_i^N$ is an independently and identically distributed sample. To infer y from x we assume a model of $p(y|x)$ with parameters θ and search for likely model parameters θ based on the data D. The predictive distribution marginalizes over the model parameters

$$p(y|x,D) = \int p(y,\theta|x,D)d\theta = \int p(y|x,\theta)p(\theta|D)d\theta, \tag{2}$$

where the data D and the model parameters θ are connected by Bayes' rule. The *posterior distribution* $p(\theta|D)$ of the parameters θ after observing the data D then decomposes into

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}, \tag{3}$$

with $p(D|\theta)$ being the *likelihood* of observing the data D given the parameters θ, $p(\theta)$ the *prior distribution* of θ, which represents the beliefs about θ before any data seen and, the *evidence* $p(D)$ also called marginal likelihood, that serves as a normalizing constant to ensure that the posterior distribution sums up to one. The integral $p(D) = \int p(D,\theta)d\theta = \int p(D|\theta)p(\theta)d\theta$ usually cannot be solved analytically and is intractable for numerical integration.

The standard way of training Neural Networks corresponds to Maximum a-posterior (MAP) inference in the Bayesian framework and yields a single parameter configuration for the posterior distribution:

$$\theta_{\text{MAP}} = \arg_\theta \max \sum_{i=1}^{N} \log p(y_i|x_i,\theta) + \log p(\theta). \tag{4}$$

$\log p(\theta)$ corresponds to a weight regularization, such as weight decay, and if no such regularization is used, it is a constant value corresponding to an improper flat prior. The goal of any approximate inference technique is to go beyond such point estimates to capture more characteristics of the posterior distribution.

3.2 Typicality

Consider a set of i.i.d. random variables $X_1, ..., X_n \overset{i.i.d.}{\sim} p(x)$. A sample of this set is a sequence of observed values $x^n = (x_1, ..., x_n)$. We call such a sequence typical Cover (1999) if it satisfies

$$2^{-n(H(x)+\epsilon)} \leq p(x_1, ..., x_n) \leq 2^{-n(H(x)-\epsilon)} \tag{5}$$

for $\epsilon > 0$ and the Shannon entropy $H(x)$ of $p(x)$. The *typical set* $A_\epsilon^{(n)}$ is defined as the set of all these typical sequences. If $x^n \in A_\epsilon^{(n)}$ we can rewrite the criterion in equation 5 into

$$H(x) - \epsilon \leq -\frac{1}{n}\sum_{i=1}^{n} \log_2 p(x_i) \leq H(x) + \epsilon, \tag{6}$$

where the Shannon entropy of a typical sequence is bounded from below and above by the Shannon entropy of the distribution $p(x)$ and defined by equation 7.

$$H[p(y|x,D)] = -\sum_y p(y|x,D) \log p(y|x,D). \tag{7}$$

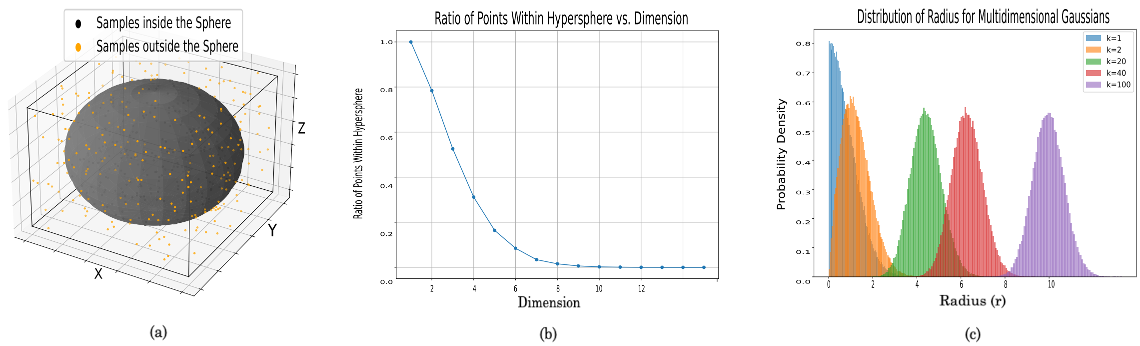

Figure 2: Illustrations of typical set "phenomena": (a) Samples of a uniform distribution on the unit cube in three dimensions. Samples that lie within the unit sphere are coloured grey, and samples outside are orange. (b) With growing dimensions of the unit cube, most samples lie outside of the respective unit sphere. (c) Probability density of a standard Gaussian distribution as a function of distance to the origin. In higher dimensions, most probability mass concentrates not at the mean (origin) but on a given radius.

Since $-\frac{1}{n}\sum_{i=1}^{n}\log_2 p(x_i)$ converges to $H(x)$ in probability for large n by the Asymptotic Equipartition property Cover (1999) it follows that

$$P(x^n \in A_\epsilon^{(n)}) > 1-\epsilon,$$

which means the probability that a sequence is part of the typical set is almost 1.

In high dimensions, most samples from a distribution will fall into this typical set due to the concentration of measure Carpenter (2017). In low dimensions d, Gaussian distributions have most of their probability mass close to the mean. However, for large d an origin-centered normal Gaussian $\mathcal{N}(0, \sigma\mathbb{I})$ has almost all of its mass located near a thin annulus with radius $\sigma\sqrt{d}$ (Gaussian Annulus Theorem Blum et al. (2020)). Considering the expected squared distance of a point $\mathbf{x} \sim \mathcal{N}(0, \mathbb{I})$ to the mean provides some intuition as to where to expect points when we sample from the distribution:

$$E(|\mathbf{x}|^2) = \sum_{i=1}^{d} E(x_i^2) = dE(x_1^2) = d.$$

While the Annulus theorem, as a result of the concentration of measure, describes the probability mass, the entropy-based condition of the typical set in Equation equation 6 leads to a similar result. For $n = 1$ a point x belongs to the typical set of the distribution $\mathcal{N}(0, \sigma\mathbb{I})$ if $||x-\mu||_2^2 = \sigma\sqrt{d}$.

As a consequence, posterior approximations of Bayesian neural networks based on high-dimensional Gaussians tend to under-represent the most-probable weight configurations Farquhar et al. (2020); Bishop & Nasrabadi (2006), since there is almost no probability mass in the hyper ball of radius smaller $\sqrt{d}$ and typical samples do almost certainly not lie in this region as illustrated in Figure 2. Assuming some degree of continuity in the weight space, i.e., weight configurations close to each other yield similarly performing functions, the likely weight configurations within the neighborhood of a MAP estimate would not be sampled by Gaussians in high dimensions.

3.3 Metropolis adjusted Langevin algorithm

The Metropolis-adjusted Langevin algorithm (MALA) is an MCMC method that combines Langevin dynamics Welling & Teh (2011) with the Metropolis-Hastings acceptance criterion to sample from a target distribution.

1. Langevin Proposal:

$$X' = X_t + \tau\nabla\log\pi(X_t) + \tau Z$$

Where X' is the proposed move, X_t is the current position, $\nabla \log \pi(X_t)$ is the gradient of the log of the target distribution at the current position, τ is a step size, and Z is a standard normal random variable.

2. Metropolis-Hastings Acceptance Criterion:

$$\alpha(X_t, X') = \min\left(1, \frac{\pi(X')q(X' \to X_t)}{\pi(X_t)q(X_t \to X')}\right)$$

Where π is the target distribution and q is the transition kernel of the Langevin dynamics. The term $q(X' \to X_t)$ represents the probability of transitioning from the proposed move X' back to the current position X_t and vice-versa for $q(X_t \to X')$.

The algorithm will accept the proposed move X' with probability $\alpha(X_t, X')$. If accepted, the chain moves to X', otherwise it remains at X_t.

3.4 Domain Restriction through Reflective Boundary Condition

A hyperball is defined around the θ_{MAP} with radius r, where $r < \sigma\sqrt{K}$ (Gaussian Annulus). Here, σ is the standard deviation of a Laplace approximation, i.e., a function of the curvature in the θ_{MAP}. MALA is run while enforcing the boundary reflection in the constrained region with radius r. This ensures the Markov Chain is constrained to a feasible space, as any samples exceeding the boundary will reflect back. Samples are then generated from the chain. In this way, we still converge to a stationary distribution but within a restricted domain. This ensures more stable outcomes and confines the samples to regions with the highest probability, thereby ensuring an ergodic chain Oliviero-Durmus & Moulines (2024).

3.5 Mixtures of Locally Bounded Langevin dynamics

The approach starts with an ensemble of initial models, represented from a constrained region of high probability. Each model is run under a domain-restricted MALA, yielding a locally bounded stationary distribution in the region. This results in a mixture of the distributions (lbMALA) in multiple high-probability regions.

3.6 Implementation

lbMALA was implemented in three stages, as illustrated in Algorithm 1:

Step 1 A pre-trained baseline model is required and used to initialise the chain with a starting point estimate, Z_0, i.e., the MAP estimate. The associated model parameters are loaded and hyperparameters defined; step size τ, chain length L, burn-in B, and boundary conditions (domain$_{\text{min}}$ and domain$_{\text{max}}$). The local curvature was estimated using the Hessian of the negative log-likelihood to dynamically calculate the reflective boundaries and constrain the sampling to a high probability region of the posterior. A buffer size $S = 5000$ was also initialised to store a sufficient number of samples while searching within the boundary conditions.

Step 2 aimed to accelerate convergence but through a constrained MALA adapted with the reflective boundary condition to provide an efficient sampling approach. With each iteration, a mini-batch of data, $\mathcal{D}$ is drawn, the gradient $\nabla \log \pi(Z)$ computed, and a new Langevin sample is proposed (Z'). Each proposal is checked against pre-defined boundary constraints, and if a proposed movement is outside this region, it is reflected within a valid vicinity as detailed in Section 3.3. This ensures better definition in the constrained region. Thereafter, the continuous acceptance and rejection rates are monitored and assessed as convergence occurs to ensure stability and convergence in the bounded region.

Step 3 allows $n = 5$ posterior samples from the chain, with the sample intialising a separate model which was fine-tuned and trained for a further 10 epochs. To ensure consistency with SWAG and Multi-SWAG, the baseline MAP at 10 epochs was utilised. A final set of parameters produced the optimal lbMALA approach with step size τ of $1e-6$, chain length L of 400, burn-in B of 1000 samples, with the learning rate for the final optimsation set to $lr = 1e-4$.

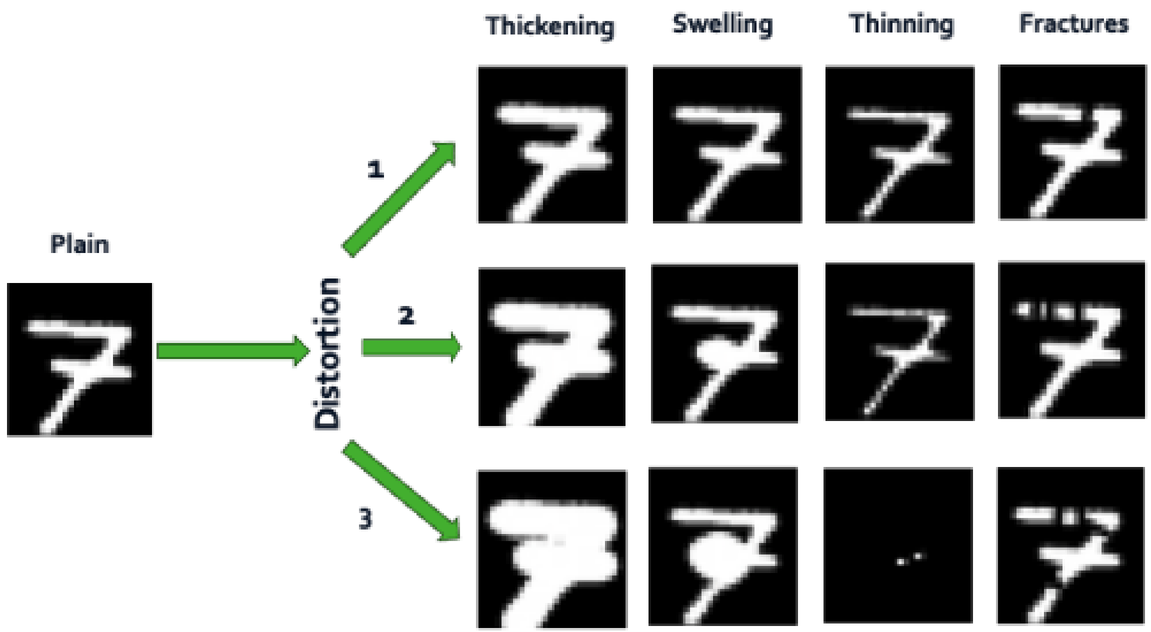

Figure 3: An illustration for the digit 7 across the types of perturbations, across each distortion placed on the original "*plain*" MNIST dataset.

4 Experimental Analysis

Our experiments are designed to highlight the ability of different methods to perform fine-grained ID versus OOD detection. To this end, we utilize the Morpho-MNIST toolkit Castro et al. (2019) to create versions of the MNIST digit classification dataset that are increasingly distorted. Using these increasingly OOD datasets, we quantify the ability of different posteriors to rank these datasets according to their level of OOD distortion.

4.1 Design and Setup

The Morpho-MNIST toolkit Castro et al. (2019) provides four different distortions that can be applied to the original MNIST digits to produce either a thickened, swelled, thinned, or fractured (broken continuity) version of any MNIST digit, see Fig. 3 for an illustration. As shown in the figures, the level of distortion is controlled by a perturbation factor that influences the severity of a distortion. We used the original digits as an ID (Plain) dataset for training and ID testing, and generated, for each of the four distortions, three increasingly OOD test sets by using each distortion with three increasing severity levels.

To evaluate the performance of the lbMALA, it was benchmarked against several baseline methods for epistemic uncertainty quantification. All methods were tested with two different backbones: a lightweight convolutional neural network (CNN) and a more complex ResNet18 model. The CNN model consists of 2 convolutional layers with ReLU + max pooling, followed by 2 fully connected layers producing logits for the 10 digits (classes). The ResNet18 model comprises 18 layers, 4 residual blocks, global pooling, and a final fully connected layer. Both models were trained on the original MNIST (Plain) dataset, consisting of 28×28 grayscale images, of which 60.000 examples were used for training (90%) and validation (10%,) and 10.000 were used for testing. The same hyperparameters lr and batch size, were used for both the CNN and ResNet architectures.

Algorithm 1 lbMALA Algorithm

Step 1: Initialization
Load pretrained model with MAP estimate Z_0
Define hyperparameters: τ, L, B, $\text{domain}_{\text{min}}$, $\text{domain}_{\text{max}}$, buffer size S
Estimate boundary using local Hessian curvature
Initialize $Z = Z_0$, rejection counter $= 0$, buffer $\mathcal{S}$
for $i = 1$ to $L + B$ **do**
  Sample mini-batch $(x, y) \sim \mathcal{D}$
  Compute $\nabla \log \pi(Z)$
  Propose $Z' = Z + \tau \nabla \log \pi(Z) + \sqrt{2\tau} \cdot \eta$, $\eta \sim \mathcal{N}(0, I)$
  if $Z' < \text{domain}_{\text{min}}$ **then**
    $Z' = 2 \cdot \text{domain}_{\text{min}} - Z'$
  end if
  if $Z' > \text{domain}_{\text{max}}$ **then**
    $Z' = 2 \cdot \text{domain}_{\text{max}} - Z'$
  end if
  Compute acceptance probability: $\alpha = \min\left(1, \frac{\pi(Z')q(Z' \to Z)}{\pi(Z)q(Z \to Z')}\right)$
  Sample $u \sim \mathcal{U}(0, 1)$
  if $u \leq \alpha$ **then**
    Accept $Z = Z'$
  else
    Reject, increment rejection counter
  end if
  Append Z to $\mathcal{S}$
  if $i > B$ and $i \bmod k = 0$ **then**
    Save Z
  end if
end for
Step 3: Extract and Retrain
for $s = 1$ to S **do**
  Extract S samples $\mathcal{S}_s$
  Retrain model with $\mathcal{S}_s$ for more epochs
end for
Output: Ensemble of retrained models

Baselines To evaluate the performance of the **lbMALA** method for Bayesian posterior sampling and uncertainty quantification, we performed comparisons against a range of widely utilised baselines. All baselines were selected to cover a range of methods, including deterministic point-estimate models and Bayesian inference approaches. Each method was implemented under consistent training conditions to ensure a robust and fair comparison.

Maximum A Posteriori (MAP) was used as the main baseline. Here, the cross-entropy loss is optimised and a single point-based estimate of model weights is generated as a result of the corresponding zero-mean Gaussian prior over model parameters. This provides a computationally efficient approach, but as mentioned earlier provides overconfident outcomes as there is no measure of uncertainty. Thus, it is used as an initial starting point for subsequent approaches.

The **MC dropout**, Gal & Ghahramani (2016) was implemented to compute an approximation of variational inference by randomly dropping units in the model to prevent overfitting (with a dropout probability of p=0.3). However, even though it was easy to implement, it has been known to not provide the best representation of model uncertainty.

Deep Ensembles, Lakshminarayanan et al. (2017) utilises multiple independently trained baseline MAP models, while computationally more expensive, typically outperform MC dropout, whose weight configurations tend to be less diverse Durasov et al. (2021). We trained 5 models, each initialized and trained from scratch on the same training data. All three of these approaches were trained over 20 epochs and with a learning rate of $lr = 1e-3$ using stochastic gradient descent (SGD) and a batch size of 64. All subsequent models trained and discussed utilised the same batch size of 64.

SWAG provides a structured and more rigorous approach to Bayesian approximation. Here, SWAG computes multiple weights when training using SGD and uses the mean and covariance of the collected weights to estimate a Gaussian distribution; samples from this distribution are used to provide scalable yet high-quality uncertainty estimates Maddox et al. (2019). We used the baseline MAP model at 10 epochs as the initial starting MAP point. However, even with a better approximation, SWAG is centered around the MAP estimate and thus may still not capture model diversity. We therefore extended the approach through **Multi-SWAG**, which utilises an ensemble of SWAG models Wilson & Izmailov (2020) to improve posterior coverage and improve prediction uncertainty with a more diverse posterior. Both SWAG and Multi-SWAG utilised a learning rate of $lr = 1e-2$.

The last comparative approach implemented was the **mixture of Laplace approximations** Eschenhagen et al. (2021), applied to the baseline of ensemble MAP models to represent the complexity of the posterior distribution using multiple local optima in the loss landscape. This approach allows a more accurate quantification of uncertainty using the multimodal nature of the posterior. A smaller noise perturbation was used (0.01) on model parameters, as higher values created very noisy and destabilised predictions.

4.2 Results

To investigate the relative performance of lbMALA against the Bayesian SOTA approaches for fine-grained ID versus OOD detection, the entropy values, reflecting epistemic uncertainty, were used to classify samples as ID or OOD. For a well-estimated posterior distribution, we would expect our ability to detect distorted samples of either kind as being OOD, to increase with increasing level of distortion. We illustrate our performance on ID vs OOD performance, quantified using the Area Under the Curve (AUC) metric, for various distortions in Figure 4, with CNN results in the top row and ResNet18 on the bottom row. Finally, all log entropy plots over the ID and OOD test sets across all distortions are represented in Appendix A.

We note that while all the tested methods show the desired trend of increased AUC for ID vs OOD classification as the distortion level increases, our lbMALA performs consistently well across both distortion types and backbones.

5 Discussion

Quantification of epistemic uncertainty is critical in many applications, such as the OOD detection problem tackled here. Many Bayesian approaches rely on posterior approximations based on posterior samples, and are therefore sensitive to the fact that typical sets of very high-dimensional distributions can fall far from the posterior modes. As a result, these samples do not represent sets of optimal parameters for the task at hand. Our proposed lbMALA approach, which utilizes sampling within a restricted space around MAP estimates, will generate samples closer to the MAP, which are more likely to be optimal for the ID datasets. These samples should yield stronger differences between ID and OOD samples. The results presented in Figure 4 confirm this assumption, as lbMALA has a more consistent and reliable performance across both backbones. In particular, the Figures suggest that entropy-based measures assist in identifying OOD cases of increasing complexity, where the distortions are increasingly pronounced but the digit itself is still recognisable.

Differences to existing methods. We compared lbMALA to a range of popular and recent methods for epistemic uncertainty quantification on our entropy-based fine-grained ID vs OOD detection task. MC dropout and Ensembles are well-known and widely used methods for epistemic uncertainty quantification. Despite providing only discrete support, they are both able to separate ID and OOD data in a fine-grained manner. However, their performance is not consistent between the 3 different distortions, as illustrated in

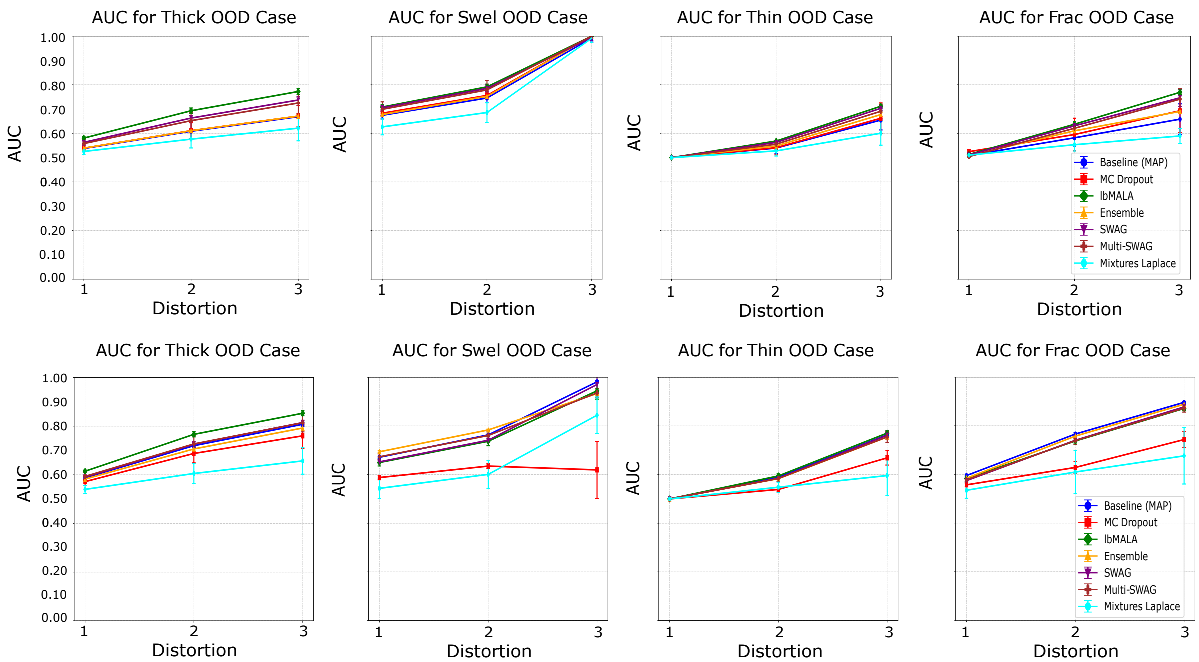

Figure 4: The AUC performance for the CNN (top) and ResNet18 (bottom) architectures across all methods for fine-grained ID and OOD separation for (from left to right) Thickening, Swelling, Fractures, and Thinning datasets for 5 random seed initialisations for reliability.

Figure 4. For instance, while MC Dropout for the CNN architecture performed best for the Thickening dataset, it comes out as consistently worst on all distortions using the ResNet18 backbone, see Figure 4.

For the Thinning dataset in Figure 4, the Baseline MAP model exhibits a sharp deterioration in AUC performance. This may indicate that the ResNet18 architecture assigns high softmax scores to an incorrect class, leading to a reduced entropy value.

For the CNN backbone, lbMALA came out consistently top or second for the Thickening distortion. It gets more competition for the ResNet18 backbone, where SWAG and Ensemble also perform well. Interestingly, while lbMALA demonstrates consistently strong performance, the other methods vary more in their performance across backbones – in particular, as we expect the (fine-tuned) ResNet18 to represent a more over-fitted model than the simpler CNN.

Limitations. lbMALA does exhibit some limitations, such as the high computational costs for the two-step gradient evaluation (first proposing and then accepting a sample). While this is manageable for simpler problems like digit classification or lightweight architectures, it could impact more complex problems. We learn, however, from lbMALA's positive performance, **that the typicality problem is real and affects the performance of epistemic uncertainty quantification**. This is critical knowledge and a necessary starting point for developing more efficient methods whose solutions are not limited to the typical set.

Need for fine-grained OOD detection. In this paper, we have introduced a notion of fine-grained ID vs OOD detection, including a validation benchmark based on the Morpho-MNIST toolkit. This is a contrast to the standard validation schemes found in uncertainty quantification papers, where the performance of OOD detection is often demonstrated by showing that methods can recognize that data comes from an entirely different dataset – e.g., MNIST version Fashion-MNIST. We stress that fine-grained OOD detection is of crucial importance, e.g. in healthcare AI Cui & Wang (2022) where "typical" model failure does not come as a clear breakdown on obviously incorrect data – but rather as somewhat reduced performance on underrepresented population groups or disease subtypes. Often, these somewhat reduced performances are only visible when aggregating performances across an entire group, which is rarely done in everyday practice. As a result, fine-grained OOD detection is important for warning users of a potential decrease in the reliability of the model.

6 Conclusion

Ensuring more robust discrimination between ID and OOD data is fundamental for developing reliable deep learning models, more so for applications in high-risk settings where misclassification of samples may have a pronounced impact.

In this paper, we propose lbMALA, a Bayesian approach utilising a reflective boundary condition to enhance the ability to localise the approximation of the posterior weight configurations. To evaluate the performance of the method, various comparative SOTA approaches were implemented on the problem of fine-grained ID and OOD detection. Our analysis demonstrated promising outcomes with the novelty of lbMALA outperforming SOTA in terms of reliability and consistency across OOD datasets to separate data on a finer scale, as reflected by the AUC values through logistic regression analysis.

However, even with the noticeable improvements, some limitations require further exploration to determine the impact on computational costs and performance on more complex datasets. Nevertheless, it would be valuable to apply the lbMALA approach to a larger dataset or a more complex domain (medical applications) to explore further and evaluate its benefit for reliable uncertainty quantification.

7 Impact statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, one of which we would like to emphasise. If the uncertainty estimates of our predictions are inaccurate and users trust them, it could mislead users and impact their work.

References

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, 76:243–297, 2021.

Lemar Abdi, MM Amaan Valiuddin, Christiaan GA Viviers, Peter HN de With, and Fons van der Sommen. Typicality excels likelihood for unsupervised out-of-distribution detection in medical imaging. In *International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*, pp. 149–159. Springer, 2024.

Michael Betancourt. A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*, 2017.

Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

Avrim Blum, John Hopcroft, and Ravindran Kannan. *Foundations of data science*. Cambridge University Press, 2020.

Roberto Cagnotti. Exploring epistemic uncertainty in untrained networks for the inverse problem. Student paper, ETH Zurich, 2023. URL `https://doi.org/10.3929/ethz-b-000736348`. Examiner: Ender Konukoglu.

Bob Carpenter. Typical sets and the curse of dimensionality, 2017. URL `https://mc-stan.org/learn-stan/case-studies/curse-dims.html`.

Daniel C Castro, Jeremy Tan, Bernhard Kainz, Ender Konukoglu, and Ben Glocker. Morpho-mnist: Quantitative assessment and diagnostics for representation learning. *Journal of Machine Learning Research*, 20 (178):1–29, 2019.

Hyunsun Choi, Eric Jang, and Alexander A Alemi. Waic, but why? generative ensembles for robust anomaly detection. *arXiv preprint arXiv:1810.01392*, 2018.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

Peng Cui and Jinjia Wang. Out-of-distribution (ood) detection based on deep learning: A review. *Electronics*, 11(21):3500, 2022.

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*, 34:20089–20103, 2021.

Nikita Durasov, Timur Bagautdinov, Pierre Baque, and Pascal Fua. Masksembles for uncertainty estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13539–13548, 2021.

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.

Sebastian Farquhar. Understanding approximation for bayesian inference in neural networks. *arXiv preprint arXiv:2211.06139*, 2022.

Sebastian Farquhar, Michael A Osborne, and Yarin Gal. Radial bayesian neural networks: Beyond discrete support in large-scale bayesian deep learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 1352–1362. PMLR, 2020.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1050–1059, New York, New York, USA, 20–22 Jun 2016. PMLR. URL `https://proceedings.mlr.press/v48/gal16.html`.

Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. *arXiv preprint arXiv:1912.03263*, 2019.

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

Fredrik K Gustafsson, Martin Danelljan, and Thomas B Schon. Evaluating scalable bayesian deep learning methods for robust computer vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 318–319, 2020.

Reinhard Heckel and Paul Hand. Deep decoder: Concise image representations from untrained non-convolutional networks, 2019. URL `https://arxiv.org/abs/1810.03982`.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

Kim-Celine Kahl, Carsten T Lüth, Maximilian Zenk, Klaus Maier-Hein, and Paul F Jaeger. Values: A framework for systematic validation of uncertainty estimation in semantic segmentation. *arXiv preprint arXiv:2401.08501*, 2024.

Robion C Kirby. Stable homeomorphisms and the annulus conjecture. *Annals of Mathmatics*, 89(2):575–582, 1969.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

David JC MacKay. A practical bayesian framework for backpropagation networks. *Neural computation*, 4 (3):448–472, 1992.

Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning. *Advances in neural information processing systems*, 32, 2019.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? *arXiv preprint arXiv:1810.09136*, 2018.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, and Balaji Lakshminarayanan. Detecting out-of-distribution inputs to deep generative models using typicality. *arXiv preprint arXiv:1906.02994*, 2019.

Alain Oliviero-Durmus and Éric Moulines. On geometric convergence for the metropolis-adjusted langevin algorithm under simple conditions. *Biometrika*, 111(1):273–289, 2024.

Genki Osada, Tsubasa Takahashi, and Takashi Nishide. Understanding likelihood of normalizing flow and image complexity through the lens of out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 21492–21500, 2024.

Tim Pearce, Mohamed Zaki, Alexandra Brintrup, N Anastassacos, and A Neely. Uncertainty in neural networks: Bayesian ensembling. *stat*, 1050:12, 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*, volume 6. International Conference on Representation Learning, 2018.

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

Kajetan Schweighofer, Lukas Aichberger, Mykyta Ielanskyi, Günter Klambauer, and Sepp Hochreiter. Quantification of uncertainty with adversarial models. *Advances in Neural Information Processing Systems*, 36: 19446–19484, 2023b.

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)*, pp. 681–688. Citeseer, 2011.

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.

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*, pp. 2282–2292. PMLR, 2023.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132(12):5635–5662, 2024.

Kilian Zepf, Selma Wanna, Marco Miani, Juston Moore, Jes Frellsen, Søren Hauberg, Frederik Warburg, and Aasa Feragen. Laplacian segmentation networks improve epistemic uncertainty quantification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 349–359. Springer, 2024.

Lily Zhang, Mark Goldstein, and Rajesh Ranganath. Understanding failures in out-of-distribution detection with deep generative models. In *International Conference on Machine Learning*, pp. 12427–12436. PMLR, 2021.

A Appendix

In this appendix, we provide all the detailed log softmax entropy graphs for each of the methods discussed in this paper.

A.1 Illustrations ID vs OOD

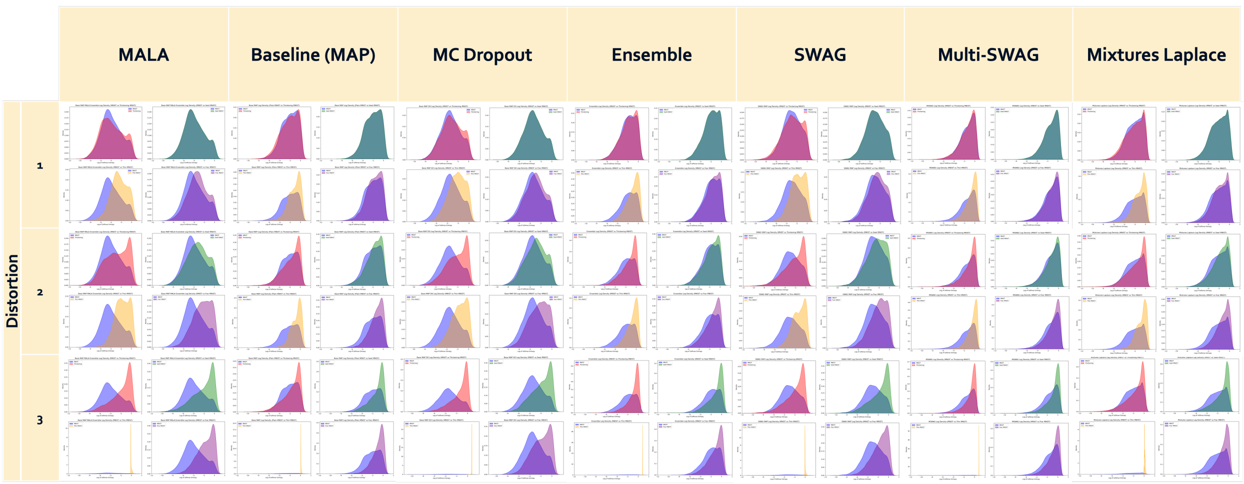

Figure 5: All log softmax entropy distributions indicating the ID and OOD datasets (Thickening, Swelling, Thinning, and Fracture datasets respectively) across all implemented methods for the CNN architecture.

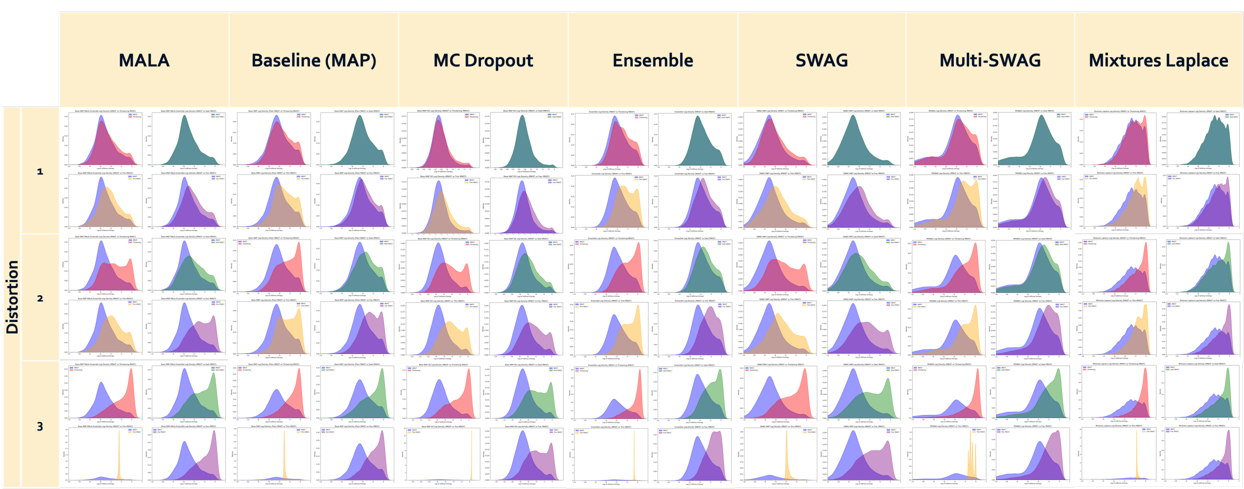

Figure 6: All log softmax entropy distributions indicating the ID and OOD datasets (Thickening, Swelling, Thinning, and Fracture datasets respectively) across all implemented methods for the ResNet18 architecture.