# NECESSARY AND SUFFICIENT HYPOTHESIS OF CURVATURE: UNDERSTANDING CONNECTION BETWEEN OUT-OF-DISTRIBUTION GENERALIZATION AND CALIBRATION

**Hiroki Naganuma***
Mila, Université de Montréal
Montreal, Canada
naganuma.hiroki@mila.quebec

**Masanari Kimura***
ZOZO Research
Tokyo, Japan
masanari.kimura@zozo.com

## ABSTRACT

In this study, we address two significant issues that hinder the application of deep learning in real-world settings: Out-of-Distribution Generalization and Calibration. While both Out-of-Distribution Generalization and Calibration have been researched in different contexts, we propose a hypothesis that they can be considered through the lens of curvature. Our extensive experiments demonstrate that training with Sharpness-Aware Minimization, which achieves low curvature, results in well-calibrated models with high accuracy, even on Out-of-Distribution datasets. Finally, we provide theoretical analysis to show that low curvature models are well-calibrated.

## 1 INTRODUCTION

Recently, deep neural networks have been widely applied in various industrial applications such as image recognition and language processing due to their high performance (He et al., 2016; Devlin et al., 2018). There are, unfortunately, two problems in adapting DNNs to real-world applications.

**Out-of-Distribution (OOD) Generalization** (Arjovsky, 2021): The general supervised learning framework of machine learning assumes that training and test data are sampled from i.i.d. Therefore, by performing empirical risk minimization (ERM), which is a loss in the training data, one expects to minimize the loss in the test data (Vapnik, 1991). However, this i.i.d. assumption generally does not hold in real-world applications.

**Calibration**: Uncertainty, as well as ranking performance, is essential in how statistical models are evaluated. In other words, the level of confidence with which the estimation is made. Although recent DNNs have high-ranking performance, they tend to make overconfident estimates and are known to have low confidence, such that if there are 100 pathology images with a 99% confidence level, only about 50 of them can be correctly identified (Guo et al., 2017). This gap between confidence and accuracy (ranking performance) is called Expected Calibration Error (ECE), and the lower the ECE, the higher the confidence in the uncertainty. Correcting the confidence in uncertainty is called calibration (Naeini et al., 2015).

### 1.1 MOTIVATION

These two critical barriers to real-world applications are often discussed in different contexts. Our goal is to understand the connection between OOD and Calibration. In particular, we consider curvature to be a key factor in understanding the connection between OOD and Calibration, and we test the following hypothesis:

**Hypothesis 1** (Curvature as necessity and sufficiency for OOD generalization and calibration)**.** The following three operations are equivalent.

(a) Improvement of OOD generalization performance;
(b) Learning well-calibrated model;
(c) Reduction of loss curvature.

*These authors contributed equally to this work

### 1.2 CONTRIBUTIONS OF THIS PAPER

To investigate the Hypothesis 1, we addressed it both theoretically and empirically. First, when using sharpness-aware minimization (Foret et al., 2020), which is known to converge to the optimal solution with low curvature, we showed that it provides higher OOD generalization performance and better calibration than other training algorithms on several models in CIFAR10 and ImageNet-based OOD datasets. Next, we introduce theoretical results supporting the assumption that the well-calibrated model reduces curvature when learning with focal loss (Lin et al., 2017) that suppresses overconfidence.

We now summarize our contributions:

- We have demonstrated in sufficient experiments (4 datasets and 6 model architectures) that a model with smaller curvature obtained by using SAMs has better calibration performance and higher OOD accuracy (Figure 1, 3).
- We show in our experiments that increasing $\rho$, the strength of the regularization of the SAM to curvature, improves calibration and OOD accuracy (Figure 2).
- We provide several remarks through Theorem 1. These analytical results relate model calibration to curvature and suggest equivalence of their improvements.

## 2 PRELIMINARIES

Let $\mathcal{X}$ be the input space and $\mathcal{Y}$ be the output space. The goal of supervised learning is to obtain the parameter $\boldsymbol{\theta} \in \Theta$ which minimizes the following expected risk.

**Definition 1** (Expected risk)**.** For some loss function $\ell : \mathcal{X} \times \mathcal{Y} \times \Theta \to \mathbb{R}$, the expected risk $\mathcal{R}(\boldsymbol{\theta})$ is defined as $\mathcal{R}(\boldsymbol{\theta}) = \mathbb{E}_{(\boldsymbol{x},y)\sim q(\boldsymbol{x},y)}\left[\ell(\boldsymbol{x}, y; \boldsymbol{\theta})\right] = \int_{\mathcal{X}\times\mathcal{Y}} \ell(\boldsymbol{x}, y; \boldsymbol{\theta}) d\boldsymbol{x}dy$, where $q(\boldsymbol{x}, y)$ be a true data distribution.

Since $q(\boldsymbol{x}, y)$ is generally unknown, we cannot compute the expected risk directly. Then, we consider the ERM, which aims to minimize the following empirical risk (Vapnik, 1991).

**Definition 2.** Let $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ be the set of data point with sample size $N \in \mathbb{N}$. For some loss function $\ell : \mathcal{X} \times \mathcal{Y} \times \Theta \to \mathbb{R}$, the empirical risk $\hat{\mathcal{R}}(\boldsymbol{\theta})$ is defined as $\hat{\mathcal{R}}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^N \ell(\boldsymbol{x}_i, y_i; \boldsymbol{\theta})$.

Under the i.i.d. assumption, it is known that $\mathbb{E}[\hat{\mathcal{R}}(\boldsymbol{\theta})] = \mathcal{R}(\boldsymbol{\theta})$.

### 2.1 OUT-OF-DISTRIBUTION GENERALIZATION

For some $M \in \mathbb{N}$, let $\mathcal{E} = \{D_1, D_2, \ldots, D_M\}_{i=1}^M$ be a set of domains, and each data of domain $D_i$ is generated by $q_{D_i}(\boldsymbol{x}, y; \boldsymbol{\theta})$. Let $\mathcal{R}_{D_i}(\boldsymbol{\theta})$ be the expected risk under $q(\boldsymbol{x}, y; \boldsymbol{\theta})$. The goal of out of distribution generalization (Shen et al., 2021) is minimize $\mathcal{R}_{D_i}(\boldsymbol{\theta})$ for $i = 1, \ldots, N$ as well as $\mathcal{R}(\boldsymbol{\theta})$ under the source distribution $q(\boldsymbol{x}, y)$.

### 2.2 MODEL CALIBRATION

The goal of model calibration is to obtain the parameter such that

$$\mathbb{P}\left(\arg\max_y p(y|\boldsymbol{x}; \boldsymbol{\theta}) = y \middle| p(y|\boldsymbol{x}; \boldsymbol{\theta}) = s\right) = s, \quad \forall s \in [0, 1]. \tag{1}$$

Expected calibration error (ECE) is one of the most well-known metrics for the model calibration.

**Definition 3** (Expected calibration error (Naeini et al., 2015))**.** For $(\boldsymbol{x}, y)$, let $\hat{y} = \arg\max p(y|\boldsymbol{x}; \boldsymbol{\theta})$. The expected calibration error is defined as

$$\mathrm{ECE}(\boldsymbol{\theta}) := \mathbb{E}\left[\left|\mathbb{P}\left(\hat{y} = y \middle| p(y|\boldsymbol{x}; \boldsymbol{\theta}) = s\right) - s\right|\right]. \tag{2}$$

### 2.3 SHARPNESS-AWARE MINIMIZATION (SAM)

Sharpness-Aware Minimization (SAM) (Foret et al., 2020) prevents convergence to high curvature local minima. Its convergence towards smaller curvature solutions results in high validation and test performance on in-distribution (ID) environment. SAM searches for points where the loss is maximized within a neighborhood of $\rho$ and uses the gradient at that point for iterative optimization. The larger the $\rho$, the higher the effect of preventing convergence to high curvature local minima. We employ SAM to train models with low curvature.

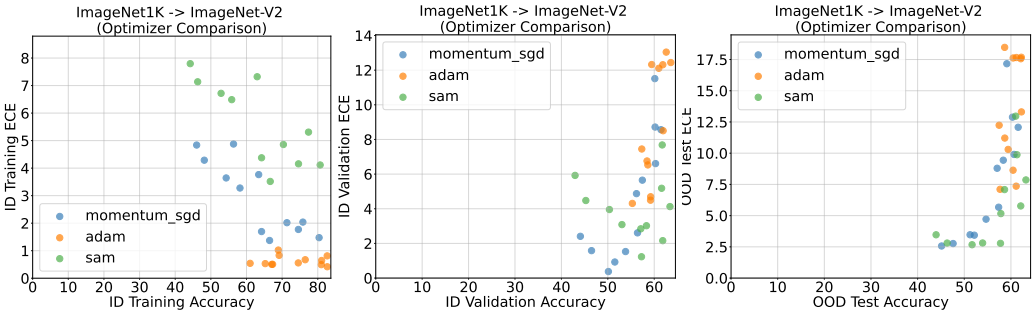


Figure 1: **Optimizer Comparison on ImageNet1K → ImageNet-V2**: The y-axis shows ECE and the x-axis shows Accuracy. (Left) evaluation on training data, (Center) evaluation on validation data in the same domain as training data, and (Right) evaluation on the OOD data set.

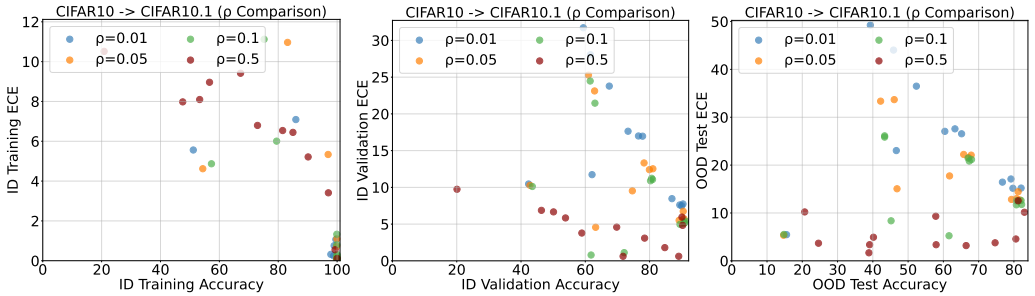


Figure 2: $\rho$ **Comparison on CIFAR10 → CIFAR10.1**: The y-axis shows ECE and the x-axis shows Accuracy. (Left) evaluation on training data, (Center) evaluation on validation data in the same domain as training data, and (Right) evaluation on the OOD data set.

## 3 SAM Leads Better OOD Generalization and Calibration

### 3.1 Experimental Protocol

We conducted a series of experiments using several model architectures, including Vision Transformer Small (ViT) (Lee et al., 2021), ResNet18, ResNet50 He et al. (2016), and Multi-Layer Perceptron (MLP). For the task, CIFAR10 (Krizhevsky, 2009) was used for training, and CIFAR10.1 (Recht et al., 2018) was used as OOD data for evaluation. As a more practical experimental setting, we also evaluated a task using ImageNet-1K (Russakovsky et al., 2015) as training data and ImageNet-V2 (Recht et al., 2019) as OOD data. Momentum SGD and Adam, the most practically used optimizer, were employed for comparison with SAM. Detailed information, such as learning rate and batch-size, is shown in Appendix C.

### 3.2 How SAM outperforms other optimizers

In the context of comparing two metrics, accuracy and ECE, SAM demonstrates competitive performance in both ID Validation and OOD Test environments (Figures 1, 3(in Appendix D)). Although SAM's high accuracy in ID Validation has been previously reported, this study is the first to evaluate its performance with respect to ECE and OOD. In the ID Validation setting, SAM exhibits similar behavior to Momentum SGD but ultimately outperforms it. Conversely, Adam demonstrates overfitting behavior in ID Training, which aligns with the findings of Naganuma et al. (2022). In OOD Test environments, SAM excels in both Accuracy and ECE metrics, further highlighting its superior performance.

### 3.3 Ablation Study: Effect of $\rho$ in SAM Optimizer

Figure 2 indicates that the behavior becomes more robust as the value of $\rho$ increases. This means that OOD accuracy and ECE performance improve as the flatness strength increases within the experimental scope range. Further ablation studies regarding other batch-size and models can be found in Appendix E.

# 4 CURVATURE CONNECTING OOD GENERALIZATION AND CALIBRATION

Our objective is to discern the shared principles that improve both calibration and out-of-distribution generalization performance. We specifically focus on the notion of curvature, otherwise known as the loss surface. Consequently, we have devised the **Hypothesis 1** as presented in Section 1.

Prior research (Wald et al., 2021) demonstrates the correlation between (a) and (b) by considering model calibration as a unique instance of invariant representation learning. In section 3, utilizing Sharpness-Aware Minimization (SAM), we have introduced empirical results for (c) $\rightarrow$ (a) and (c) $\rightarrow$ (b). In this section, we present the theoretical outcomes for (a) $\rightarrow$ (c) and (b) $\rightarrow$ (c).

## 4.1 WELL-CALIBRATED MODEL REDUCES CURVATURE

Here, we discuss the relationship between a well-calibrated model and loss curvature and aim to show (b) $\rightarrow$ (c) of Hypothesis 1. To achieve this goal, we consider the Focal loss, a well-known loss function that is useful for model calibration.

**Definition 4** (Focal loss (Lin et al., 2017))**.** Let $p(y|\boldsymbol{x};\boldsymbol{\theta})$ be a model parameterized by $\boldsymbol{\theta} \in \Theta$ and $q(y|\boldsymbol{x})$ be a true distribution. For $\gamma \geq 0$, the Focal loss $\mathcal{L}_f^\gamma(\boldsymbol{\theta})$is defined as follows:

$$\mathcal{L}_f^\gamma(\boldsymbol{\theta}) \coloneqq -\int_{\mathcal{X}\times\mathcal{Y}} (1-p(y|\boldsymbol{x};\boldsymbol{\theta}))^\gamma q(y|\boldsymbol{x}) \ln p(y|\boldsymbol{x};\boldsymbol{\theta}) d\boldsymbol{x}dy. \tag{3}$$

Focal loss is well-known to prevent overconfidence of the model. Therefore, theoretical investigation of the behavior of focal loss should provide insight into what is needed to learn a well-calibrated model.

Here, we can rewrite Focal loss as the regularized cross-entropy loss.

**Proposition 1.** Let $p(y|\boldsymbol{x};\boldsymbol{\theta})$ be a model parameterized by $\boldsymbol{\theta} \in \Theta$ and $q(y|\boldsymbol{x})$ be a true distribution. For $\gamma \geq 0$, we have

$$\mathcal{L}_f^\gamma(\boldsymbol{\theta}) = \mathcal{L}_c(\boldsymbol{\theta}) - \gamma H(y|\boldsymbol{x},\boldsymbol{\theta}), \tag{4}$$

where $\mathcal{L}_c(\boldsymbol{\theta})$ is the cross-entropy loss and $H(y|\boldsymbol{x},\boldsymbol{\theta})$ is the conditional entropy.

From equation 4, learning procedure with Focal loss can be regarded as maximizing conditional entropy term under the constraint

$$\int_\Theta p(\boldsymbol{\theta})\mathcal{L}_c(\boldsymbol{\theta})d\theta = \int_\Theta p(\boldsymbol{\theta}) \left( -\int_{\mathcal{X}\times\mathcal{Y}} q(y|\boldsymbol{x}) \ln p(y|\boldsymbol{x};\boldsymbol{\theta}) d\boldsymbol{x}dy \right) d\boldsymbol{\theta} = \delta_\gamma \leq \delta, \tag{5}$$

for some $\delta \geq 0$ and $\delta_\gamma \geq 0$.

**Theorem 1.** Among all distributions defined on $\Theta$ with a given $\delta_\gamma$, the distribution with the largest entropy is the Maxwell-Boltzmann distribution

$$\tilde{p}(\boldsymbol{\theta}) = \alpha \cdot e^{-\beta \cdot \mathcal{L}_c(\boldsymbol{\theta})} = \alpha \cdot \exp\left\{ \beta \int_{\mathcal{X}\times\mathcal{Y}} q(y|\boldsymbol{x}) \ln p(y|\boldsymbol{x};\boldsymbol{\theta}) d\boldsymbol{x}dy \right\}, \tag{6}$$

where the constants $\alpha$ and $\beta$ are determined from the following constraints

$$\int_\Theta p(\boldsymbol{\theta})d\boldsymbol{\theta} = 1 \Longleftrightarrow \alpha = \left( \int_\Theta e^{-\beta \cdot \mathcal{L}_c(\boldsymbol{\theta})} d\boldsymbol{\theta} \right)^{-1}, \tag{7}$$

$$\int_\Theta \mathcal{L}_c(\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} = \delta_\gamma \Longleftrightarrow \int_\Theta \mathcal{L}_c(\boldsymbol{\theta}) e^{-\beta \cdot \mathcal{L}_c(\boldsymbol{\theta})} d\boldsymbol{\theta} = \frac{\delta_\gamma}{\alpha}. \tag{8}$$

Using equation 6 as the prior of parameter distribution as $\pi = \tilde{p}$, we give the following PAC-Bayes bound by using Thiemann's bound(Thiemann et al., 2017).

**Remark 1.** Maxwell-Boltzmann distribution induces the following bound

$$\mathbb{P}\left[ \forall \varsigma \in P(\Theta), \mathbb{E}_{\boldsymbol{\theta}\sim\varsigma}\left[\mathcal{R}(\boldsymbol{\theta})\right] \leq \frac{\mathbb{E}_{\boldsymbol{\theta}\sim\varsigma}[\hat{R}(\boldsymbol{\theta})]}{1-\frac{\lambda}{2}} + \frac{D_{KL}[\varsigma\|\pi] + \log\frac{2\sqrt{n}}{\epsilon}}{n\lambda(1-\frac{\lambda}{2})} \right] \leq \epsilon \tag{9}$$

for some $\epsilon > 0$ and $\lambda \geq 0$. The minimum of the right-hand side is achieved by

$$\varsigma_\lambda(\boldsymbol{\theta}) = \frac{\pi(\boldsymbol{\theta})e^{-\lambda n\hat{R}(\boldsymbol{\theta})}}{\mathbb{E}_{\boldsymbol{\theta}\sim\pi}\left[e^{-\lambda n\hat{R}(\boldsymbol{\theta})}\right]}, \quad \text{(fixed } \lambda\text{)}, \quad \lambda = \frac{2}{\sqrt{\frac{2n\mathbb{E}_\varsigma\left[\hat{\mathcal{R}}(\boldsymbol{\theta})\right]}{D_{KL}[\varsigma\|\pi]+\ln\frac{2\sqrt{n}}{\delta}}} + 1 + 1} \quad \text{(fixed } \varsigma\text{)}. \tag{10}$$

**Remark 2.** Focal loss equips the following regularizer:

$$D_{KL}[\varsigma\|\pi] = \int_{\Theta} \varsigma(\boldsymbol{\theta}) \log \frac{\varsigma(\boldsymbol{\theta})}{c \cdot \exp\left\{\beta \int_{\mathcal{X}\times\mathcal{Y}} q(\boldsymbol{x}, y) \log p(y|\boldsymbol{x}; \boldsymbol{\theta})\right\}} d\boldsymbol{\theta}. \tag{11}$$

**Proposition 2.** Learning with Focal loss reduces the local sharpness of the likelihood, if prior $\pi$ and posterior $\varsigma$ are close enough.

Since cross-entropy is the expectation of the negative log-likelihood function, we have the following remarks.

**Remark 3.** Focal loss reduces the local curvature of the loss landscape.

**Remark 4.** Reducing the local curvature of the loss landscape improves the calibration error.

### 4.2 OOD GENERALIZATION MODEL REDUCES CURVATURE

Next, we introduce the relationship between OOD generalization and curvature, which indicates (a) $\rightarrow$ (c) of Hypothesis 1. Previous studies have revealed the following.

Rame et al. (2022) propose an "inconsistency score," a metric that quantifies the degree of inconsistency between different domains in a model. Their approach aims to capture how well the model generalizes across various domains by measuring the difference in risks between them. According to their findings, good weights should be optimal for all domains and difficult to change. They suggest a model is expected to generalize better across different domains by minimizing this inconsistency score.

**Proposition 3** (Rame et al. (2022))**.** Under the quadratic bowl Assumption with positive definite Hessians, for small $\epsilon > 0$,

$$\begin{aligned} \mathcal{I}^{\epsilon}(\hat{\boldsymbol{\theta}}) &= \max_{(A,B)\in\mathcal{E}^2} \max_{\boldsymbol{\theta}\in\mathcal{N}^{\epsilon}(A,\hat{\boldsymbol{\theta}})} \left|\mathcal{R}_B(\boldsymbol{\theta}) - \mathcal{R}_A(\hat{\boldsymbol{\theta}})\right| \\ &= \max_{(A,B)\in\mathcal{E}^2} \left(\mathcal{R}_A(\hat{\boldsymbol{\theta}}) - \mathcal{R}_A(\hat{\boldsymbol{\theta}}) + \max_{\frac{1}{2}\boldsymbol{\theta}^{\top}\mathcal{H}_A\boldsymbol{\theta}\leq\epsilon} \frac{1}{2}\boldsymbol{\theta}^{\top}\mathcal{H}_B\boldsymbol{\theta}\right), \end{aligned} \tag{12}$$

where $\mathcal{E} = \{A, B\}$ be a set of domains, $\mathcal{N}^{\epsilon}(A, \hat{\boldsymbol{\theta}})$ be the neighborhood of $\hat{\boldsymbol{\theta}}$ and $\mathcal{H}_e = \nabla^2_{\boldsymbol{\theta}}\mathcal{R}_e(\boldsymbol{\theta})$ be the Hessian for $e \in \mathcal{E}$.

Proposition 3 indicates the relationship between OOD generalization and curvature.

Another related work relates to the curvature of the model manifold and covariate shift assumption, which is one of the OOD problems (Kimura & Hino, 2022).

## 5 DISCUSSION AND CONCLUSION

We theoretically and empirically showed the connection between OOD generalization and Calibration through the lens of curvature.

Finally, we would like to mention the limitations of our work. One limitation is that we could have studied a more comprehensive range of distribution shift datasets. Evaluation on datasets such as DomainBed (Gulrajani & Lopez-Paz, 2021), WILDS (Koh et al., 2021), and SHIFT15M (Kimura et al., 2021) is for future work. Another limitation is that we demonstrated through only experiments that low curvature leads to high OOD accuracy. For a more detailed discussion, it would be helpful to investigate how curvature depends on the convergence rate of the OOD generalization. That is, for some function $\varphi(r)$ where $r$ is a quantity expressing curvature, we expect the OOD generalization error $\mathcal{R}_{OOD}$ to be written as $\mathcal{R}_{OOD}(\boldsymbol{\theta}) \leq \hat{\mathcal{R}}(\boldsymbol{\theta}) + \varphi(r)$. Such an analysis guarantees the effectiveness of the learning algorithm for OOD generalization, which depends on the curvature.

### ACKNOWLEDGMENTS

We are sincerely grateful to the meta-reviewer and the three anonymous reviewers for helping us improve the original manuscript.

## REFERENCES


Martin Arjovsky. Out of distribution generalization in machine learning, 2021. URL https://arxiv.org/abs/2103.02667.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330. PMLR, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Alexander Immer, Matthias Bauer, Vincent Fortuin, Gunnar Rätsch, and Khan Mohammad Emtiyaz. Scalable marginal likelihood estimation for model selection in deep learning. In *International Conference on Machine Learning*, pp. 4563–4573. PMLR, 2021.

Masanari Kimura and Hideitsu Hino. Information geometrically generalized covariate shift adaptation. *Neural Computation*, 34(9):1944–1977, 2022.

Masanari Kimura, Takuma Nakamura, and Yuki Saito. Shift15m: multiobjective large-scale fashion dataset with distributional shifts. *arXiv preprint arXiv:2108.12992*, 2021.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664. PMLR, 2021.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

Seung Hoon Lee, Seunghyun Lee, and Byung Cheol Song. Vision transformer for small-size datasets. *arXiv preprint arXiv:2112.13492*, 2021.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip Torr, and Puneet Dokania. Calibrating deep neural networks using focal loss. *Advances in Neural Information Processing Systems*, 33:15288–15299, 2020.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

Hiroki Naganuma, Kartik Ahuja, Ioannis Mitliagkas, Shiro Takagi, Tetsuya Motokawa, Rio Yokota, Kohta Ishikawa, and Ikuro Sato. Empirical study on optimizer selection for out-of-distribution generalization. *arXiv preprint arXiv:2211.08583*, 2022.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

Alexandre Rame, Corentin Dancette, and Matthieu Cord. Fishr: Invariant gradient variances for out-of-distribution generalization. In *International Conference on Machine Learning*, pp. 18347–18377. PMLR, 2022.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do cifar-10 classifiers generalize to cifar-10? 2018. https://arxiv.org/abs/1806.00451.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pp. 5389–5400. PMLR, 2019.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021.

Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021.

Niklas Thiemann, Christian Igel, Olivier Wintenberger, and Yevgeny Seldin. A strongly quasiconvex pac-bayesian bound. In *International Conference on Algorithmic Learning Theory*, pp. 466–492. PMLR, 2017.

Vladimir Vapnik. Principles of risk minimization for learning theory. *Advances in neural information processing systems*, 4, 1991.

Yoav Wald, Amir Feder, Daniel Greenfeld, and Uri Shalit. On calibration and out-of-domain generalization. *Advances in Neural Information Processing Systems*, 34, 2021.

Jize Zhang, Bhavya Kailkhura, and T Yong-Jin Han. Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning. In *International conference on machine learning*, pp. 11117–11128. PMLR, 2020.

# APPENDIX

## A PROOFS

### A.1 PROOF FOR PROPOSITION 1

*Proof.*

$$\begin{aligned}
\mathcal{L}_f^\gamma(\boldsymbol{\theta}) &= -\int_{\mathcal{X}\times\mathcal{Y}} (1-p(y|\boldsymbol{x};\boldsymbol{\theta}))^\gamma q(y|\boldsymbol{x}) \ln p(y|\boldsymbol{x};\boldsymbol{\theta}) d\boldsymbol{x}dy \\
&\geq -\int_{\mathcal{X}\times\mathcal{Y}} (1-\gamma p(y|\boldsymbol{x};\boldsymbol{\theta})) q(y|\boldsymbol{x}) \ln p(y|\boldsymbol{x};\boldsymbol{\theta}) d\boldsymbol{x}dy \\
&\geq -\int_{\mathcal{X}} \left\{ \int_{\mathcal{Y}} q(y|\boldsymbol{x}) \ln p(y|\boldsymbol{x};\boldsymbol{\theta}) dy - \gamma \max_j q(j|\boldsymbol{x};\boldsymbol{\theta}) \int_{\mathcal{Y}} |p(y|\boldsymbol{x};\boldsymbol{\theta}) \ln p(y|\boldsymbol{x};\boldsymbol{\theta})|\, dy \right\} d\boldsymbol{x} \\
&\geq -\int_{\mathcal{X}\times\mathcal{Y}} q(y|\boldsymbol{x}) \ln p(y|\boldsymbol{x};\boldsymbol{\theta}) d\boldsymbol{x}dy + \gamma \int_{\mathcal{X}\times\mathcal{Y}} p(y|\boldsymbol{x};\boldsymbol{\theta}) \ln p(y|\boldsymbol{x};\boldsymbol{\theta}) d\boldsymbol{x}dy \\
&= \mathcal{L}_c(\boldsymbol{\theta}) - \gamma H(y|\boldsymbol{x};\boldsymbol{\theta}).
\end{aligned}$$

□

### A.2 PROOF FOR THEOREM 1

*Proof.* From chain rule, the conditional entropy is written as

$$H(y|\boldsymbol{x};\boldsymbol{\theta}) = H(y,\boldsymbol{\theta}|\boldsymbol{x}) - H(\boldsymbol{\theta}).$$

Then, maximizing conditional entropy $H(y|\boldsymbol{x};\boldsymbol{\theta})$ is equal to minimizing entropy $H(\boldsymbol{\theta})$. In order to minimize the entropy $(\boldsymbol{\theta}) = -\int_\Theta p(\boldsymbol{\theta}) d(\boldsymbol{\theta})$ subject to constraints

$$\int_\Theta p(\boldsymbol{\theta}) d\boldsymbol{\theta} = 1 \tag{13}$$

$$\int_\Theta p(\boldsymbol{\theta}) \mathcal{L}_c(\boldsymbol{\theta}) d\boldsymbol{\theta} = \delta_\gamma, \tag{14}$$

consider the Lagrangian

$$L = -p(\boldsymbol{\theta}) \ln p(\boldsymbol{\theta}) - \beta \mathcal{L}_c(\boldsymbol{\theta}) p(\boldsymbol{\theta}) - \eta p(\boldsymbol{\theta}). \tag{15}$$

The Euler-Lagrange equation for equation 15 is

$$\ln p(\boldsymbol{\theta}) = \beta \mathcal{L}_c(\boldsymbol{\theta}) - \eta + 1, \tag{16}$$

with solution

$$p(\boldsymbol{\theta}) = \alpha \cdot e^{-\beta \cdot \mathcal{L}_c(\boldsymbol{\theta})}, \tag{17}$$

$$\alpha = e^{1-\eta}. \tag{18}$$

From equation 17 and equation 18, we have

$$\int_\Theta e^{-\beta \cdot \mathcal{L}_c(\boldsymbol{\theta})} d\boldsymbol{\theta} \int_\Theta \mathcal{L}_c(\boldsymbol{\theta}) e^{-\beta \mathcal{L}_c(\boldsymbol{\theta})} d\boldsymbol{\theta} = \delta_\gamma. \tag{19}$$

Consider

$$\Phi(\beta) = \int_\Theta e^{\beta \mathcal{L}_c(-\boldsymbol{\theta})} d\boldsymbol{\theta} \int_\Theta \mathcal{L}_c(\boldsymbol{\theta}) e^{-\beta \mathcal{L}_c(\boldsymbol{\theta})} d\boldsymbol{\theta}, \tag{20}$$

which is an decreasing function of $\beta$. Since

$$\lim_{\beta \to +\infty} \Phi(\beta) = 0, \tag{21}$$

$$\lim_{\beta \to -\infty} \Phi(\beta) = +\infty, \tag{22}$$

the continuity of $\Phi(\beta)$ implies that equation 19 has a solution and this is unique. Hence, a unique pair $(\alpha, \beta)$ satisfies the problem constraints. Therefore, Maxwell-Boltzmann distribution is unique. □

### A.3 PROOF FOR PROPOSITION 2

*Proof.* From assumption, let

$$\Delta\xi_i = \xi_i^{\pi} - \xi_i^{\varsigma}, \tag{23}$$

where $\boldsymbol{\xi}^{\varsigma}$ and $\boldsymbol{\xi}^{\pi}$ are parameters of $\varsigma$ and $\pi$. Consider the quadratic approximation of $d_{kl}(\boldsymbol{\xi}) = D_{KL}[\boldsymbol{\xi}^{\varsigma} \| \boldsymbol{\xi}]$ as

$$d_{kl}(\boldsymbol{x}) = d_{kl}(\boldsymbol{\xi}^{\varsigma}) + \sum_i \frac{\partial d_{kl}}{\partial \xi_i}(\xi^{\varsigma})\Delta\xi_i + \frac{1}{2}\sum_{i,j}\frac{\partial^2 d_{kl}}{\partial\xi_i\xi_j}(\boldsymbol{x}^{\pi})\Delta\xi_i\Delta\xi_j - o(\|\Delta\boldsymbol{\xi}\|^2). \tag{24}$$

First, since $D_{KL}[p\|q] = 0$ if and $p = q$, we have

$$d_{kl}(\boldsymbol{\xi}^{\varsigma}) = D_{KL}[\boldsymbol{\xi}^{\varsigma}\|\boldsymbol{\xi}^{\varsigma}] = 0. \tag{25}$$

Next, diagonal part of the first variation of the KL-divergence is

$$\begin{aligned}\frac{\partial}{\partial\xi_i}D_{KL}[\boldsymbol{\xi}^{\varsigma}\|\boldsymbol{\xi}] &= \frac{\partial}{\partial\xi_i}\int_{Theta} p(\boldsymbol{\theta};\boldsymbol{\xi}^{\varsigma})\ln p(\boldsymbol{\theta};\boldsymbol{\xi}^{\varsigma})d\boldsymbol{\theta} - \frac{\partial}{\partial\xi_i}\int_{\Theta} p(\boldsymbol{\theta};\boldsymbol{\xi}^{\varsigma})\ln p(\boldsymbol{\theta};\boldsymbol{\xi})d\boldsymbol{\theta} \\ &= -\int_{\Theta} p(\boldsymbol{\theta};\boldsymbol{\xi}^{\varsigma})\frac{\partial}{\partial\xi_i}\ln p(\boldsymbol{\theta};\boldsymbol{\xi})d\boldsymbol{\theta}. \end{aligned} \tag{26}$$

$$\frac{\partial}{\partial\xi_i}D_{KL}[\boldsymbol{\xi}^{\varsigma}\|\boldsymbol{\xi}]_{|\boldsymbol{\xi}=\boldsymbol{\xi}^{\varsigma}} = -\int_{\Theta} p(\boldsymbol{\theta};\boldsymbol{\xi}^{\varsigma})\frac{\partial}{\partial\xi_i}\ln p(\boldsymbol{\theta};\boldsymbol{\xi}^{\varsigma})d\boldsymbol{\theta} = -\mathbb{E}_{\boldsymbol{\xi}^{\varsigma}}\left[\frac{\partial}{\partial\xi_i}\ln p(\boldsymbol{\theta};\boldsymbol{\xi}^{\varsigma})\right] = 0. \tag{27}$$

Finally, the diagonal part of the Hessian of the KL-divergence is

$$\begin{aligned}\frac{\partial^2}{\partial\xi_i\partial\xi_j}D_{KL}[\boldsymbol{\xi}^{\varsigma}\|\boldsymbol{\xi}] &= \frac{\partial^2}{\partial\xi_i\xi_j}\int_{\Theta} p(\boldsymbol{\theta};\boldsymbol{\xi}^{\varsigma})\ln p(\boldsymbol{\theta};\boldsymbol{\xi}^{\varsigma})d\boldsymbol{\theta} - \int_{\Theta} p(\boldsymbol{\theta};\boldsymbol{\xi}^{\varsigma})\ln p(\boldsymbol{\theta};\boldsymbol{\xi})d\boldsymbol{\theta} \\ &= -\int_{\Theta} p(\boldsymbol{\theta};\boldsymbol{\xi}^{\varsigma})\frac{\partial^2}{\partial\xi_i\partial\xi_j}\ln p(\boldsymbol{\theta};\boldsymbol{\xi})d\boldsymbol{\theta}. \end{aligned} \tag{28}$$

$$\begin{aligned}\frac{\partial^2}{\partial\xi_i\partial\xi_j}D_{KL}[\boldsymbol{\xi}^{\varsigma}\|\boldsymbol{\xi}]_{|\boldsymbol{\xi}=\boldsymbol{\xi}^{\varsigma}} &= -\int_{\Theta} p(\boldsymbol{\theta};\boldsymbol{\xi}^{\varsigma})\frac{\partial^2}{\partial\xi_i\partial\xi_j}\ln p(\boldsymbol{\theta};\boldsymbol{\xi})d\boldsymbol{\theta} \\ &= -\mathbb{E}_{\boldsymbol{\xi}^{\varsigma}}\left[\frac{\partial^2}{\partial\xi_i\partial\xi_j}\ln p(\boldsymbol{\theta};\boldsymbol{\xi}^{\varsigma})\right] = g_{ij}(\boldsymbol{\xi}^{\varsigma}), \end{aligned} \tag{29}$$

where $I(\boldsymbol{\xi}) = (g_{ij}(\boldsymbol{\xi}))$ is the Fisher information matrix. Then, we have

$$D_{KL}[\varsigma\|\pi] = \frac{1}{2}\sum_{i,j} g_{ij}(\boldsymbol{\xi}^{\varsigma})\Delta\xi_i\Delta\xi_j + o(\|\Delta\boldsymbol{\xi}\|^2). \tag{30}$$

From equation 30 and equation 9,

$$\mathbb{P}\left[\forall\varsigma \in P(\Theta), \mathbb{E}_{\boldsymbol{\theta}\sim\varsigma}\left[\mathcal{R}(\boldsymbol{\theta})\right] \leq \frac{\mathbb{E}_{\boldsymbol{\theta}\sim\varsigma}[\hat{R}(\boldsymbol{\theta})]}{1-\frac{\lambda}{2}} + \frac{\frac{1}{2}\sum_{i,j} g_{ij}(\boldsymbol{\xi}^{\varsigma})\Delta\xi_i\Delta\xi_j + \log\frac{2\sqrt{n}}{\epsilon}}{n\lambda(1-\frac{\lambda}{2})}\right] \leq \epsilon$$

with $o(\|\Delta\boldsymbol{\xi}\|^2)$, and the second term of the right-hand side measures the curvature of the log-likelihood. □

## B RELATED WORKS

### B.1 CALIBRATION

In the current operation of deep learning techniques, post-hoc methods are usually used for calibration. Methods such as isotonic regression and temperature scaling are ways to improve calibration without degrading ranking performance. Methods such as Ensemble learningOvadia et al. (2019) and Bayesian inferenceImmer et al. (2021) are also known to improve calibration. They are effective for uncertainty estimation, but their computational cost, which can be several times higher, is a problem in practical applications.

Several studies have been conducted on calibration, but most focus on uncertainty issues in contexts such as computer visionMukhoti et al. (2020). Calibration methods in recommendation systems have also been proposedZhang et al. (2020).

### B.2 Calibration and Out-of-Distribution Generalization

Different approaches have been taken for calibration and OOD, and no research has practically addressed the calibration problem in an OOD environment. Recent theoretical studies have shown that domain-specific calibration is equivalent to Invariant Risk Minimization (IRM) Arjovsky et al. (2019), a typical learning method for OOD, in learning multiple domains Wald et al. (2021).

## C Experimental Settings

### C.1 Datasets

**CIFAR10.1** Recht et al. (2018): This dataset contains approximately 2,000 new test images. This dataset was collected after multiple years of research on the original CIFAR-10 dataset and was designed to minimize distribution shifts. The images in CIFAR-10.1 are a subset of the TinyImages dataset.

**ImageNet-V2** Recht et al. (2019): This dataset is a new benchmark for testing image recognition models, providing 10,000 new images for each of the three test sets. This dataset is unique because it was collected after ten years of evolution of the original ImageNet dataset, ensuring that accuracy scores are not biased by adaptive overfitting. The data collection process is designed to maintain similarity to the original ImageNet dataset. The repository provides code for working with ImageNetV2, a pool of candidate images, and rich metadata. ImageNet-V2 consists of three categories: TopImages, Threshold0.7, and MatchedFrequency. We evaluated all of them and averaged them to report as Test OOD accuracy and Test OOD ECE.

### C.2 Hyperparameters

In our experiments on both CIFAR10 and ImageNet, the batch size was standardized at 256. The learning rate was explored within the range of ”0.0005”, ”0.001”, ”0.005”, ”0.01” using grid search. The parameter rho was also grid searched within the range of ”0.005”, ”0.01”, ”0.05”, ”0.1”, ”0.5”. The number of training epochs was 400 for CIFAR10 and 90 for ImageNet experiments.

## D Omitted Experimental Results in Main Paper

Here we present experimental results that could not be included in the main paper due to paper space limitations. Figure 3 shows similar results with Figure 1 (The only experimental difference is the data set and model).

Figures 4, 5 show the transition of each metric during training in the experiments in Figure 3, 1. The solid line shows the average of the results for the three hyperparameters with the highest final validation accuracy for each optimization method. The colored regions indicate the variance.

During the initial stages of learning, for SAM, the Train ECE is seen to increase while the other optimizers suddenly drop (Figure 4). For other optimizers, this sudden drop may be attributed to the

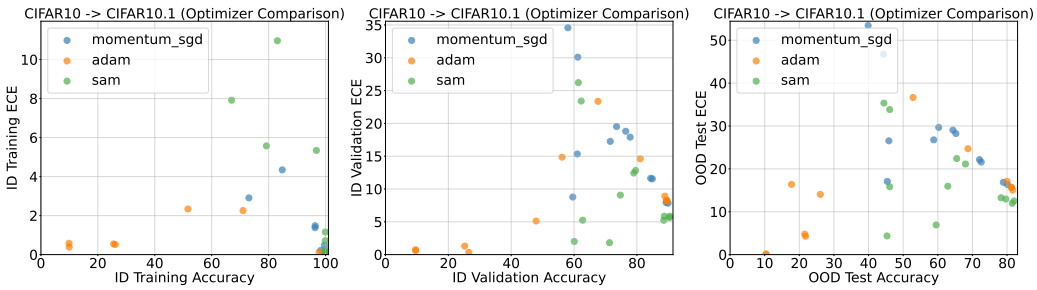


Figure 3: **Optimizer Comparison on CIFAR10 → CIFAR10.1**: The y-axis shows ECE and the x-axis shows Accuracy. (Left) evaluation on training data, (Center) evaluation on validation data in the same domain as training data, and (Right) evaluation on the OOD data set.

poor performance of Val ECE and Test ECE. To mitigate this issue, it may be effective to implement regularization techniques that prevent over-reduction of Train ECE during the early stages of learning.

In the ImageNet experiment (Figure 5), Train ECE remained high in the case of SAM, which could lead the better performance in the OOD environment.

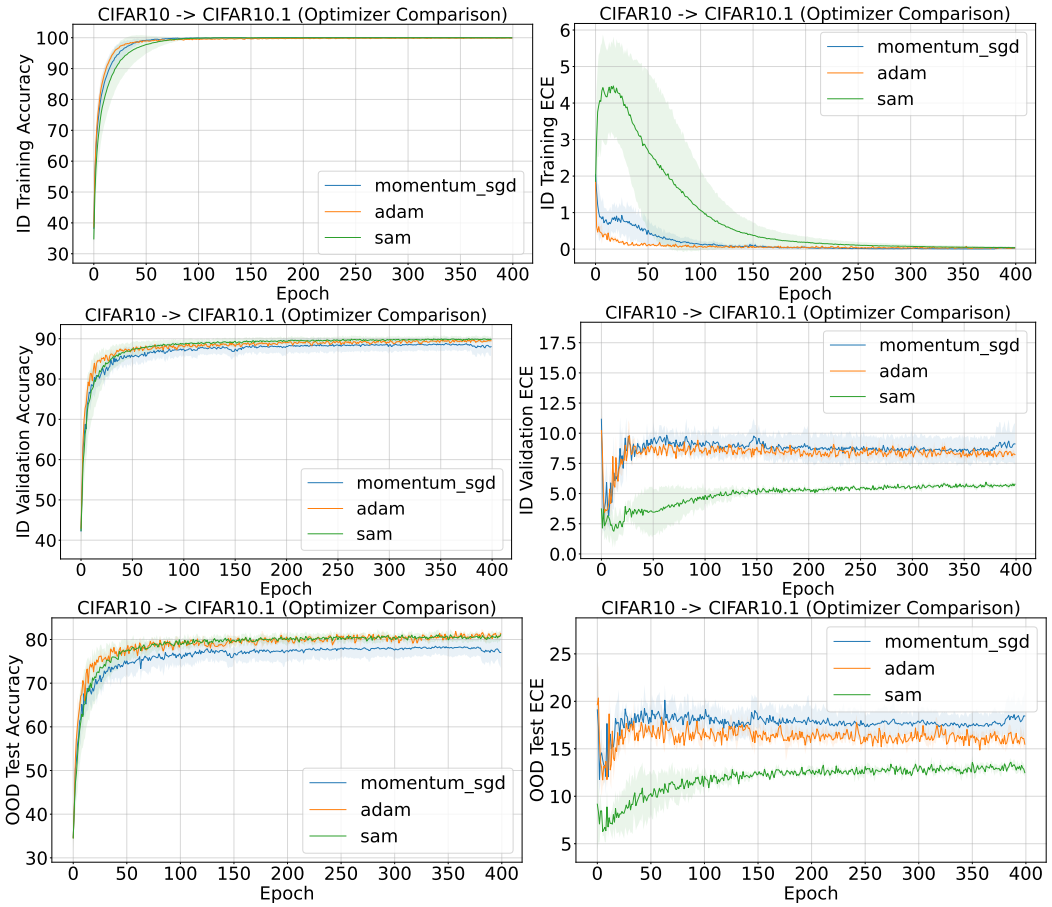


Figure 4: **Optimizer Comparison on CIFAR10 → CIFAR10.1**: The y-axis shows each metric (Accuracy or ECE) and the x-axis shows Epochs. (Top) evaluation on training data, (Center) evaluation on validation data in the same domain as training data, and (Bottom) evaluation on the OOD data set.

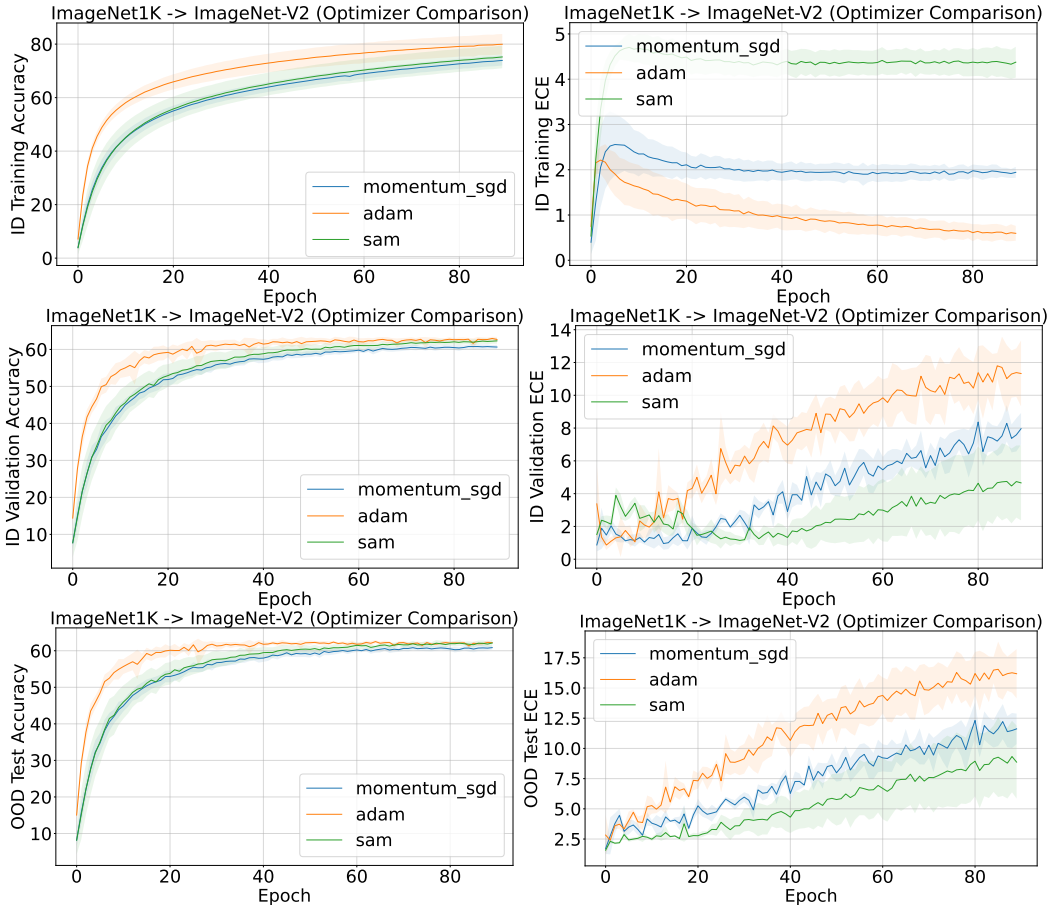


Figure 5: **Optimizer Comparison on ImageNet1K → ImageNet-V2**: The y-axis shows each metric (Accuracy or ECE) and the x-axis shows Epochs. (Top) evaluation on training data, (Center) evaluation on validation data in the same domain as training data, and (Bottom) evaluation on the OOD data set.

# E Ablation Study

## E.1 Model Comparison

We conducted ablation studies with several model architectures to investigate differences in behavior by model architecture. In our comparative analysis, we observed that a general trend is that if a model performs well on OOD data, it is likely to perform well on ECE (Figure 6).

It should be noted that data augmentation was not added in the CIFAR10 experiments. The experimental results showed that ResNet performed well in calibration and accuracy in both training and OOD environments. MLP underperformed the other model architectures in the training environment and the OOD. ViT outperformed ResNet in the training environment but underperformed ResNet in the OOD environment. This finding aligns with previous observations that ViT requires data augmentation to achieve optimal performance Steiner et al. (2021).

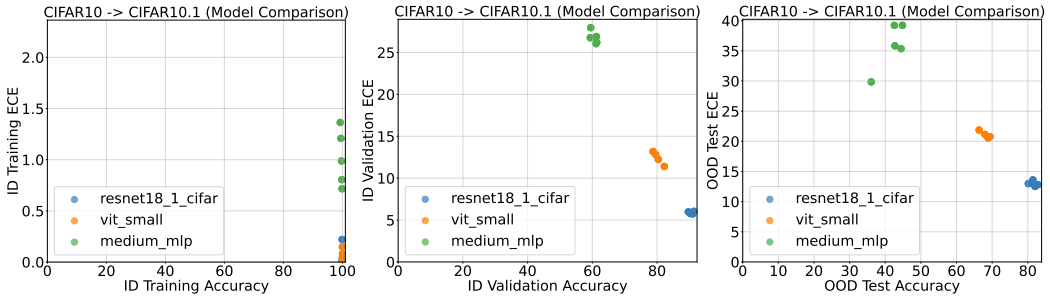


Figure 6: **Model Comparison on CIFAR10 → CIFAR10.1**: The y-axis shows ECE and the x-axis shows Accuracy. (Left) evaluation on training data, (Center) evaluation on validation data in the same domain as training data, and (Right) evaluation on the OOD data set.

### E.2 Batch Size Comparison

We conducted experiments with different batch sizes using multiple model architectures as an ablation study. For learning rates, the square root scaling rule was used. Our experimental results showed that there was no significant difference between the batch sizes within the range of our experiments. Figure 7 shows scatter plots of the relationship between accuracy and ECE when training with CIFAR10 and testing with CIFAR10.1. Each data point represents a different hyperparameter configuration (difference in model and initial values).

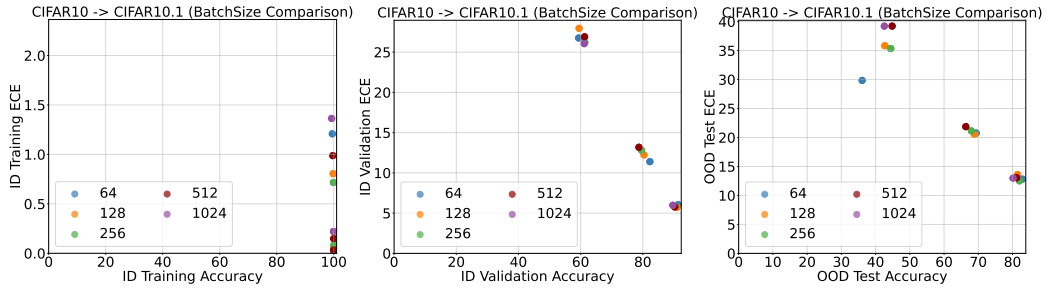


Figure 7: **batch-Size Comparison on CIFAR10 → CIFAR10.1**: The y-axis shows ECE and the x-axis shows Accuracy. (Left) evaluation on training data, (Center) evaluation on validation data in the same domain as training data, and (Right) evaluation on the OOD data set.