When and How Mixup Improves Calibration

Linjun Zhang [*1] **Zhun Deng** [*2] **Kenji Kawaguchi** [3] **James Zou** [4]

Abstract

In many machine learning applications, it is important for the model to provide confidence scores that accurately capture its prediction uncertainty. Although modern learning methods have achieved great success in predictive accuracy, generating calibrated confidence scores remains a major challenge. Mixup, a popular yet simple data augmentation technique based on taking convex combinations of pairs of training examples, has been empirically found to significantly improve confidence calibration across diverse applications. However, when and how Mixup helps calibration is still a mystery. In this paper, we theoretically prove that Mixup improves calibration in *high-dimensional* settings by investigating natural statistical models. Interestingly, the calibration benefit of Mixup increases as the model capacity increases. We support our theories with experiments on common architectures and datasets. In addition, we study how Mixup improves calibration in semi-supervised learning. While incorporating unlabeled data can sometimes make the model less calibrated, adding Mixup training mitigates this issue and provably improves calibration. Our analysis provides new insights and a framework to understand Mixup and calibration.

1. Introduction

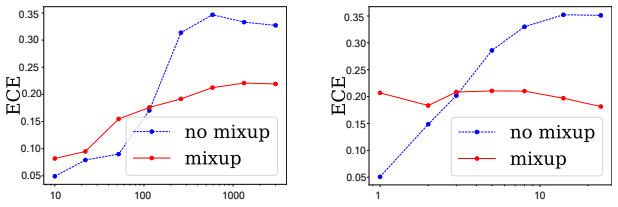
Modern machine learning methods have dramatically improved the predictive accuracy in many learning tasks (Simonyan & Zisserman, 2014; Srivastava et al., 2015; He et al., 2016a). The deployment of AI-based systems in high risk fields such as medical diagnosis (Jiang et al., 2012) requires a predictive model to be trustworthy, which makes the topic of accurately quantifying the predictive uncertainty an increasingly important problem (Thulasidasan et al., 2019). However, as pointed out by Guo et al. (2017), many popular modern architectures such as neural networks are very poorly calibrated. A variety of methods have been proposed for quantifying predictive uncertainty including training multiple probabilistic models with ensembling or bootstrap (Osband et al., 2016) and re-calibration of probabilities on a validation set through temperature scaling (Platt et al., 1999), which usually involves much more complicated procedures and extra computation. Meanwhile, recent work (Thulasidasan et al., 2019) has shown that models trained with Mixup (Zhang et al., 2017), a simple data augmentation technique based on taking convex combinations of pairs of examples and their labels, are significantly better calibrated. However, when and how Mixup helps calibration is still not well-understood, especially from a theoretical perspective.

As our **first contribution**, we demonstrate that the calibration improvement brought by Mixup is more significant in the *high-dimensional* settings, i.e. the number of parameters is comparable to the training sample size. Figure 1 shows a motivating experiment on CIFAR-10. The Expected Calibration Error (ECE), which is a standard measure of how un-calibrated a model is, is smaller with Mixup augmentation compared to those without Mixup augmentation, especially when the model is wider or deeper. We provide a theoretical explanation for this phenomenon under several natural statistical models. In particular, our theory holds when the data distribution can be described by a Gaussian generative model, which is very flexible and includes many generative adversarial networks (GANs). In a Gaussian generative model, a function is used to map a Gaussian random variable to an input vector of some models such as neural networks. Because the function used to map a Gaussian random variable is arbitrary and can be nonlinear, our theory is applicable to a very broad class of data distributions.

As our **second contribution**, we investigate how Mixup helps calibration in semi-supervised learning, which is relatively under-explored. Labeled data are usually expensive to obtain, and training models by combining a small amount of labeled data with abundant unlabeled data plays an important role in AI (Chapelle et al., 2009). In light of this, we investigate the effect of Mixup in semi-supervised learning, where we focus on the commonly used pseudo-labeling algorithm (Chapelle et al., 2009; Carmon et al., 2019). We

[*]Equal contribution [1]Rutgers University [2]Harvard University [3]National University of Singapore [4]Stanford University. Correspondence to: Linjun Zhang <linjun.zhang@rutgers.edu>, Zhun Deng <zhundeng@g.harvard.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.



(a) Network width (b) Network depth

Figure 1. Expected calibration error (ECE) calculated for a fully-connected neural network on CIFAR-10. In (a), we fix the depth and increase the width of the neural network; while in (b), we fix the width and increase the depth of the neural network. Mixup augmentation can reduce ECE especially for larger capacity models.

observe experimentally that the pseudo-labeling by itself can sometimes hurt calibration. However, combining Mixup with pseudo-labeling consistently improves calibration. We provide theories to explain these findings.

As our **third contribution**, we further extend our results to Maximum Calibration Error (MCE), which also demonstrates similar phenomena as those for ECE.

Outline of the paper. Section 2 discusses related works and introduces the notations. In Section 3, we present our main theoretical results for ECE by showing that Mixup improves calibration for classification problems in the high-dimensional regime. Section 4 investigates the semi-supervised learning setting and demonstrates the benefit of further applying Mixup to the pseudo-labeling algorithm. In Section 5, we extend our studies of calibration to MCE. Section 6 concludes with a discussion of future work. Proofs are deferred to the Appendix.

1.1. Related Work

Mixup is a popular data augmentation scheme that has been shown to improve a model's prediction accuracy (Zhang et al., 2017; Thulasidasan et al., 2019; Guo et al., 2019). Recent theoretical analysis shows that Mixup has an implicit regularization effect that enables models to better generalize (Zhang et al., 2020). The focus of our work is not on accuracy, but on calibration.

Modern learning models such as neural networks have achieved remarkable performance nowadays in optimization (Deng et al., 2020a; Ji et al., 2021b; Deng et al., 2021c; Ji et al., 2021a; Kawaguchi et al., 2022). Even though the generalization and prediction (Deng et al., 2020b; Zhang et al., 2020; Deng et al., 2021a) of neural networks are quite amazing, it has shown that neural networks tend to be over-confident. A well-calibrated predictive model is needed in many applications of machine learning, ranging from economics (Foster & Vohra, 1997), personalized medicine (Jiang et al., 2012), to weather forecasting (Gneiting & Raftery, 2005), to fraud detection (Bahnsen et al., 2014). The problem on producing a well-calibrated model has received increasing attention in recent years (Naeini et al., 2015; Lakshminarayanan et al., 2016; Guo et al., 2017; Zhao et al., 2020; Foster & Stine, 2004; Kuleshov et al., 2018; Wen et al., 2020; Huang et al., 2020). In real-world settings, the input distributions are sometimes shifted from the training distribution due to non-stationarity. The predictive uncertainty under such out-of-distribution condition was studied by Ovadia et al. (2019) and Chan et al. (2020). Mixup has been empirically shown to improve the calibration for deep neural networks in both the same and out-of-distribution domains (Thulasidasan et al., 2019; Tomani & Buettner, 2020). Ours is the first work to provide theoretical explanation for this phenomenon.

Semi-supervised learning is a broad field in machine learning concerned with learning from both labeled and unlabeled datasets (Chapelle et al., 2009). Prior work mostly focuses on improving the prediction accuracy with unlabeled data (Zhu et al., 2003; Zhu & Goldberg, 2009; Berthelot et al., 2019) and adversarial robustness (Carmon et al., 2019; Deng et al., 2021b). Recently, Chan et al. (2020) found that unlabeled data improves Bayesian uncertainty calibration in some experiments, but the relationship between using unlabeled data and calibration, especially from the theoretical perspective, is still largely unknown. All of the facts above motivate our theoretical exploration in this paper.

2. Preliminaries

In this section, We introduce the notations and briefly recap the mathematical formulation of Mixup and calibration measures considered in this paper.

2.1. Notations

We denote the training data set by $S = \{(x_1, y_1), \cdots, (x_n, y_n)\}$, where $x_i \in \mathcal{X} \subseteq \mathbb{R}^d$ and $y_i \in \mathcal{Y} \subseteq \mathbb{R}^m$ are drawn i.i.d. from a joint distribution $\mathcal{P}_{x,y}$. The general parameterized loss is denoted by $l(\theta, z)$, where $\theta \in \Theta \subseteq \mathbb{R}^p$ and $z_i = (x_i, y_i)$ denotes the input and output pair. Let $L(\theta) = \mathbb{E}_{z\sim\mathcal{P}_{x,y}} l(\theta, z)$ denote the standard population loss and $L_n^{std}(\theta, S) = \sum_{i=1}^n l(\theta, z_i)/n$ denote the standard empirical loss. In addition, we define $\tilde{x}_{i,j}(\lambda) = \lambda x_i + (1-\lambda)x_j$, $\tilde{y}_{i,j}(\lambda) = \lambda y_i + (1-\lambda)y_j$, and $\tilde{z}_{i,j}(\lambda) = (\tilde{x}_{i,j}(\lambda), \tilde{y}_{i,j}(\lambda))$ for $\lambda \in [0, 1]$. We use $t\mathcal{D}_1 + (1-t)\mathcal{D}_2$ for $t \in (0, 1)$ to denote the mixture distribution such that a sample coming from that distribution is drawn with probabilities t and $(1-t)$ from $\mathcal{D}_1$ and $\mathcal{D}_2$ respectively. In classification, the output y_i is the embedding of the class of x_i; i.e., $y_i \in \{0, 1\}^m$ is the one-hot encoding of the class (with all entries equal to zero except for the one corresponding to the class of x_i), where m is the total number of classes.

2.2. Mixup

Mixup is a data augmentation technique, which linearly interpolates the training sample pairs within the training data set to create a new data set $S^{mix}(\lambda) = \{(\tilde{z}_{i,j}(\lambda))\}_{i,j=1}^n$, with λ following a distribution $\mathcal{D}_\lambda$ supported on $[0,1]$. Throughout the paper, we consider the most commonly used $\mathcal{D}_\lambda$ — the Beta distribution $Beta(\alpha,\beta)$ for $\alpha,\beta>0$.

Typically, in a machine learning task, one wants to learn a function $f:\mathcal{X}\to\mathcal{Y}$ from a function class $\mathcal{F}$ using the training data set $S\in(\mathcal{X}\times\mathcal{Y})^n$. Such a function is usually parametrized as f_θ with some parameter θ. Let us denote the learned parameter by $\hat{\theta}=\mathcal{M}(S)$. In this paper, we consider learning the parameter by the Mixup training $\mathcal{M}(S^{mix}(\lambda))$. Due to the randomness in λ, we consider taking the expectation over λ. For example, a mapping could either be an estimator, such as the empirical mean of input: $\mathcal{M}(S)=\sum_{i=1}^n x_i/n$, or be the minimizer of a loss function: $\mathcal{M}(S,\theta)=\operatorname{argmin}_\theta\sum_{i=1}^n l(\theta,z_i)/n$. The corresponding transformed mappings obtained via Mixup are then $\mathbb{E}_{\lambda\sim\mathcal{D}_\lambda}\mathcal{M}(S^{mix}(\lambda))=\sum_{i,j=1}^n\mathbb{E}_{\lambda\sim\mathcal{D}_\lambda}x_{i,j}(\lambda)/n^2$ and $\mathbb{E}_{\lambda\sim\mathcal{D}_\lambda}\mathcal{M}(S^{mix}(\lambda),\theta)=\sum_{i,j=1}^n\mathbb{E}_{\lambda\sim\mathcal{D}_\lambda}l(\theta,\tilde{z}_{i,j}(\lambda))/n^2$ respectively.

2.3. Calibration for classification

For a classification problem, if there are K classes, typically, for an input x, a probability vector $\hat{h}(x)=(p_1(x),\cdots,p_K(x))^\top\in\mathbb{R}^K$ is obtained from the trained model, where p_i is the corresponding probability (or so-called confidence score) that x belongs to the class i, and $\sum_{i=1}^K p_i=1$. Then, the output is $\hat{y}=\operatorname{argmax}_i p_i(x)$. The hope is that, for instance, given 1000 samples, each with confidence 0.7, around 700 examples should be classified correctly. In other words, we expect for all $v\in[0,1]$, $\mathbb{P}(\hat{y}=y|\hat{p}=v)\approx v$, where $\hat{p}$ is the largest entry in $\hat{h}(x)$ and y is the true class x belongs to, which is termed as prediction confidence.

Expected Calibration Error (ECE). The most prevalent calibration metric is the Expected Calibration Error (Naeini et al., 2015), which is defined as,

$$ECE=\mathbb{E}_{v\sim\mathcal{D}_{\hat{p}}}\left[|\mathbb{P}(\hat{y}=y|\hat{p}=v)-v|\right], \tag{1}$$

where $\mathcal{D}_{\hat{p}}$ is the distribution of $\hat{p}$. While ECE is widely used, we note that recents works (Nixon et al., 2019; Kumar et al., 2019) found that some methods of estimating ECE in practice (such as the binning method) is sometimes undesirable and can produce biased estimator under some specially constructed data distributions. Throughout this paper, in our theories, *we mainly focus on the population version of calibration error as defined in (1), which does not suffer from any such bias.*

Maximum Calibration Error (MCE). Another widely used calibration metric is the Maximum Calibration Error (Naeini et al., 2015), which is defined as

$$MCE=\max_{v\in[0,1]}|\mathbb{P}(\hat{y}=y|\hat{p}=v)-v|.$$

Again, in our theory, we will only consider this population version of MCE. A predictor $\hat{p}$ with ECE/MCE equal to 0 is said to be perfectly calibrated.

3. Calibration in Supervised Learning

Although Mixup has been shown to improve the test accuracy (Zhang et al., 2017; Guo et al., 2019; Zhang et al., 2020), there has been much less understanding of how it affects model calibration [1]. In this section, we focus on investigating when and how Mixup improves calibration.

3.1. Problem set-up

As a confirmation of the phenomenon suggested in Figure 1 in the introduction, our theoretical results demonstrate that Mixup indeed improves calibration, and the improvement is especially significant in the ***high-dimensional regime***. Here, by high-dimensional regime, we mean when the number of parameters in the model, p, is comparable to the sample size n, i.e. $p/n>c$ for some constant $c>0$. In other words, the improvement in calibration by using Mixup is more significant in the over-parameterized case or when the ratio between p and n is a constant asymptotically larger than 0. Moreover, we also prove that Mixup helps calibration on out-of-domain data, which is critical for machine learning applications.

In order to derive tractable analysis, we first study the concrete and natural Gaussian model. The Gaussian model is a popular setting for understanding phenomena happening in more complex models due to its tractability in theory and its ability to partially capture some essence of the phenomena. Indeed, the Gaussian model has been widely used in theoretical investigations of more complex machine learning models such as neural networks in adversarial learning (Schmidt et al., 2018; Carmon et al., 2019; Dan et al., 2020; Deng et al., 2021b). ***We further extend our analysis to the very flexible Gaussian generative models*** in Section 3.4.

The Gaussian model. We consider a common model used for theoretical machine learning analysis: a mixture of two spherical Gaussians with one component per class (Carmon et al., 2019):

Definition 3.1 (Gaussian model)**.** *For $\theta^*\in\mathbb{R}^p$ and $\sigma>0$, the (θ^*,σ)-Gaussian model is defined as the following*

[1] Models with better test accuracy are not necessarily better calibrated.

distribution over $(x, y) \in \mathbb{R}^p \times \{1, -1\}$*:*

$$x \mid y \sim \mathcal{N}(y \cdot \theta^*, \sigma^2 I), \text{ for } i = 1, 2, ..., n,$$

and y *follows the Bernoulli distribution* $\mathbb{P}(y = 1) = \mathbb{P}(y = -1) = 1/2$.

For simplicity, we first consider the case where σ is known, and the only unknown parameter is μ. The case where σ is unknown is a special example of the general Gaussian generative model that we will consider in Section 3.4.

Algorithms. In this section, we focus on studying the following linear classifier for the Gaussian classification. Specifically, the classifier follows the celebrated Fisher's rule (Johnson et al., 2002), or so-called linear discriminant analysis, which is also considered by Carmon et al. (2019) to study the adversarial robustness. The classifier is constructed as

$$\hat{\mathcal{C}}(x) = \text{sgn}(\hat{\theta}^\top x), \tag{2}$$

where $\hat{\theta} = \sum_{i=1}^n x_i y_i / n$. Given $\hat{\theta}$ and x, the output y obtained via classifier $\hat{C}$ can be equivalently defined by the following process: we first obtain the confidence vector $h(x) = (p_1(x), p_{-1}(x))^\top$, and then output $y = \hat{\mathcal{C}}(x) = \text{argmax}_{k \in \{-1,1\}} p_k(x)$. Here, for $k \in \{-1, 1\}$, the confidence score $p_k(x)$ represents an estimator of $\mathbb{P}(y = k|x)$ and therefore takes the following form:

$$p_k(x) = \frac{1}{e^{-2k \cdot \hat{\theta}^\top x_i / \sigma^2} + 1}. \tag{3}$$

In comparison, by applying Mixup to the above algorithm, we first obtain $\{\tilde{x}_{i,j}(\lambda), \tilde{y}_{i,j}(\lambda)\}_{i,j=1}^n$, which leads to another classifier

$$\hat{\mathcal{C}}^{mix}(x) = \text{sgn}(\hat{\theta}^{mix\top} x), \tag{4}$$

where $\hat{\theta}^{mix} = \mathbb{E}_{\lambda \sim \mathcal{D}_\lambda} \sum_{i,j=1}^n \tilde{x}_{i,j}(\lambda) \tilde{y}_{i,j}(\lambda) / n^2$. Here, given the randomness of λ, we take expectation with respect to λ in the same way as in the previous study (Zhang et al., 2017), though this is unnecessary in our theoretical analysis. The confidence score obtained by $\hat{\mathcal{C}}^{mix}$ can be obtained similarly to that in Eq. (3) with $\hat{\theta}$ being replaced by $\hat{\theta}^{mix}$.

3.2. Mixup helps calibration in classification

We follow the convention in high-dimensional statistics, where the parameter dimension p grows along with the sample size n, and state our theorem in the large n, p regime where both n and p goes to infinity.

Throughout the paper, we use the term "with high probability" to indicate that the event happens with probability at least $1 - o(1)$, where $o(1) \to 0$ as $n \to \infty$ and the randomness is taken over the training data set. In the following, we show that the condition $p/n = \Omega(1)$ is necessary and the fact that Mixup improves calibration is a high-dimensional phenomenon.

Let us denote the ECE calculated with respect to $\hat{\mathcal{C}}$ and $\hat{\mathcal{C}}^{mix}$ by ECE($\hat{\mathcal{C}}$) and ECE($\hat{\mathcal{C}}^{mix}$) respectively. Our first theorem states that Mixup indeed improves calibration for the above algorithm under the Gaussian model.

Theorem 3.1. *Under the settings described above, there exists* $c_2 > c_1 > 0$*, when* $p/n \in (c_1, c_2)$ *and* $\|\theta\|_2 < C$ *for some universal constants* $C > 0$ *(not depending on* n *and* p*), then for sufficiently large* p *and* n*, there exist* $\alpha, \beta > 0$*, such that when the distribution* $\mathcal{D}_\lambda$ *is chosen as* $Beta(\alpha, \beta)$*, with high probability,*

$$ECE(\hat{\mathcal{C}}^{mix}) < ECE(\hat{\mathcal{C}}).$$

The above theorem states that when p is comparable to n and p/n is not too small, applying Mixup leads to a better calibration than without applying Mixup. In the very next theorem, we further demonstrate that the condition " p and n are comparable" is necessary for Mixup to reach a smaller ECE.

Theorem 3.2. *There exists a threshold* $\tau = o(1)$ *such that if* $p/n \le \tau$ *and* $\|\theta\|_2 < C$ *for some universal constant* $C > 0$*, given any constants* $\alpha, \beta > 0$ *(not depending on* n *and* p*), when* n *is sufficiently large, we have, with high probability,*

$$ECE(\hat{\mathcal{C}}) < ECE(\hat{\mathcal{C}}^{mix}).$$

In Theorem 3.2, we can see if p is too small compared with n, then applying Mixup cannot have any gain and even hurts the calibration.

Usually, in the implementation of Mixup, we first fix α and β before training, and the above theorem reveals the fact that in the low-dimensional regime, where p/n is sufficiently close to 0, the Mixup could not help calibration with high probability. Moreover, combined with Theorem 3.3 stated below, which characterizes the monotonic relationship between p/n and the improvement brought by Mixup, we can see Mixup helps calibration more when the dimension is higher.

For the ease of presentation, for all $\beta > 0$, let us define $Beta(0, \beta)$ as the degenerated distribution which takes the only value at 0 with probability one. We also define $\hat{\mathcal{C}}^{mix}_{\alpha,\beta}$ as the classifier where we apply Mixup with distribution $\lambda \sim Beta(\alpha, \beta)$.

Theorem 3.3. *For any constant* $c_{\max} > 0$*,* $p/n \to c_{ratio} \in (0, c^{\max})$*, when* θ *is sufficiently large (still of a constant level), we have for any* $\beta > 0$*, with high probability, the change of ECE by using Mixup, characterized by*

$$\frac{d}{d\alpha} ECE(\hat{\mathcal{C}}^{mix}_{\alpha,\beta}) \mid_{\alpha \to 0+}$$

is negative, and monotonically decreasing with respect to c_{ratio}.

In Theorem 3.3, the derivative with respect to α is interpreted as follows. Since for any $\beta > 0$, $Beta(0, \beta)$ is the degenerated distribution at 0, $\hat{\theta}^{mix}(0, \beta)$ corresponds to the output without Mixup. Therefore, increasing α from 0 to some positive value implies applying Mixup. Thus, Theorem 3.3 suggests that in high-dimensions, increasing the interpolation range in Mixup decreases ECE.

Intuition behind our results. In the high-dimensional regime, especially in the over-parameterized case ($p > n$), the models have more flexibility to set the confidence vectors. For instance, for trained neural networks, the entries of the confidence vectors for many data points are all close to zero except for one entry, whose value is close to 1, because the model is trained to memorize the training labels. Mixup mitigates this problem by using linear interpolation that creates one-hot encoding terms with entry values lying between $(0, 1)$, which pushes the value of entries to diverge. This could be partially addressed in our analysis above, as the magnitude of the confidence is closely related to $\|\hat{\theta}\|$, i.e. when $\|\hat{\theta}\|$ is large, the confidence scores are more likely to be close to 0 or 1. Mixup, as a form of regularization (Zhang et al., 2020), could shrink $\|\hat{\theta}\|$ and avoid too extreme confidence scores.

Additional supporting experiments. To complement our theory, we further provide more experimental evidence on popular image classification data sets with neural networks. We used the fully-connected neural networks with various values of the width (i.e. the number of neurons per hidden layer) and the depth (i.e., the number of hidden layers). For the experiments on the effect of the width, we fixed the depth to be 8 and varied the width from 10 to 3000. For the experiments on the effect of the depth, the depth was varied from 1 to 24 (i.e., from 3 to 26 layers including input/output layers) by fixing the width to be 400 with data-augmentation and 80 without data-augmentation. We used the following standard data-augmentation operations using `torchvision.transforms` for both data sets: random crop (via `RandomCrop(32, padding=4)` and random horizontal flip (via `RandomHorizontalFlip`) for each image. In this experiment, we used the standard data sets — CIFAR-10 and CIFAR-100 (Krizhevsky & Hinton, 2009). We used stochastic gradient descent (SGD) with mini-batch size of 64. We set the learning rate to be 0.01 and momentum coefficient to be 0.9. We used the Beta distribution $Beta(\alpha, \alpha)$ with $\alpha = 1.0$ for Mixup. The results are reported in Figure 1 and 2 with a fully-connected neural network. Consistently across all the experiments, Mixup reduces ECE for larger capacity models and can hurt ECE for small models, which matches our theory. For reasons of space, since our focus is mainly on the calibration, the empirical results regarding test accuracy for each figure are deferred to the Appendix.

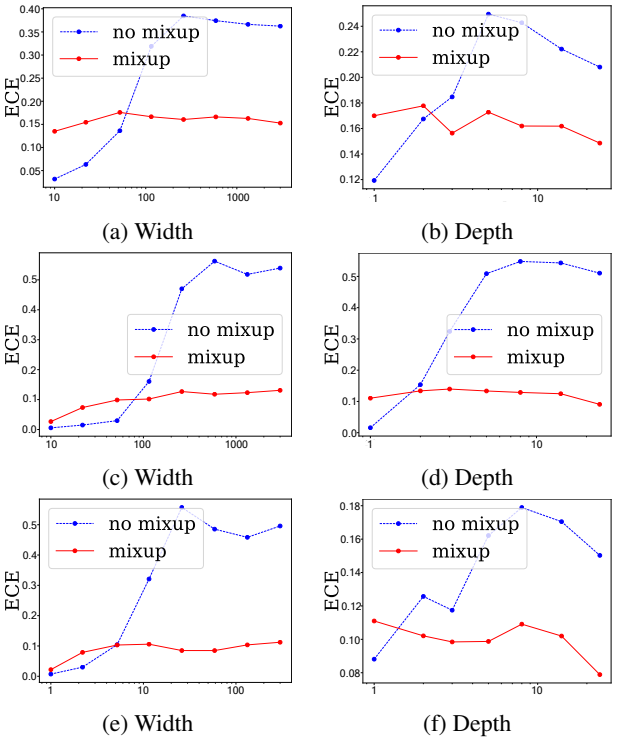


Figure 2. Expected calibration error (ECE). (a), (b): CIFAR-10 without data augmentation; (c), (d): CIFAR-100 with data augmentation; (e), (f): CIFAR-100 without data augmentation.

3.3. Improvement for out-of-domain data

In this section, we evaluate the quality of predictive uncertainty on out-of-domain inputs. It has been found empirically that in the out-of-domain setting, Mixup can also enhance the reliability of prediction and boost the performance in calibration comparing to the standard training (without Mixup) (Thulasidasan et al., 2019; Tomani & Buettner, 2020). To explain the above phenomenon, using the similar analysis as those for Theorem 3.1, we provide the following theorem.

Theorem 3.4. *Let us consider the ECE evaluated on the out-of-domain Gaussian model with mean parameter θ', that is, $\mathbb{P}(y = 1) = \mathbb{P}(y = -1) = 1/2$, and $x \mid y \sim \mathcal{N}(y \cdot \theta', \sigma^2 I)$, for $i = 1, 2, ..., n$. If we have $(\theta' - \theta^*)^\top \theta^* \leq p/(2n)$, then when p and n are sufficiently large, with high probability,*

$$ECE(\hat{\mathcal{C}}^{mix}; \theta', \sigma) < ECE(\hat{\mathcal{C}}; \theta', \sigma),$$

where $ECE(\cdot; \theta', \sigma)$ denotes the expected calibration error calculated with respect to the out-of-domain distribution – the Gaussian model with parameters θ' and σ, while $\hat{\mathcal{C}}^{mix}$

and $\hat{\mathcal{C}}$ are still obtained via (2) and (4) via the in-domain training data.

The above theorem states the continuity of the boosting effect of Mixup over the domain shift. As long as the domain shift is not too large, Mixup still helps calibration.

3.4. Gaussian generative model

Now let us consider a more general class of distributions, the Gaussian generative model, which is a flexible distribution and has been commonly considered in the machine learning literature. For example, many common deep generative models such as Generative Adversarial Nets (GANs) (Goodfellow et al., 2014) are Gaussian generative models, where the input is a Gaussian sample.

Definition 3.2 (Gaussian generative model)**.** *For $\theta^* \in \mathbb{R}^p$ and $g : \mathbb{R}^p \to \mathbb{R}^d$ ($d \geq p$), the (θ^*, g)-Gaussian model is defined as the following distribution over $(x, y) \in \mathbb{R}^d \times \{1, -1\}$, $x = g(z)$, where:*

$$z \mid y \sim \mathcal{N}(y \cdot \theta^*, I), \text{ for } i = 1, 2, ..., n,$$

and y follows the Bernoulli distribution $\mathbb{P}(y = 1) = \mathbb{P}(y = -1) = 1/2$.

Now suppose we can learn an $h \in \{h : h \circ g \text{ is an identity mapping in } \mathbb{R}^p\}$ approximately such that the estimator $\hat{h}$ satisfies the following condition.

Assumption 3.1. *For any given $v \in \mathbb{R}^p$, $k \in \{-1, 1\}$, there exists a $\theta^* \in \mathbb{R}^p$, such that given $y = k$, the probability density function of $R_1 = v^\top \hat{h}(x)$ and $R_2 = v^\top h(x) = v^\top z \sim N(k \cdot b^\top \theta^*, \|v\|^2)$ satisfies that $p_{R_1}(u) = p_{R_2}(u) \cdot (1 + \delta_u)$ for all $u \in \mathbb{R}$ where δ_u satisfies $\mathbb{E}_{R_1}[|\delta_u|] = o(1)$ when $n \to \infty$.*

As a special case of Definition 3.2, we consider the Gaussian model with unknown σ. Estimating σ by

$$\hat{\sigma} = \sqrt{\|\sum_{i=1}^{n}(x_i - y_i\hat{\theta})\|^2/pn}$$

will satisfy Assumption 3.1 when $\|\theta^*\| < C$ for some universal constant C. In practice, for more general cases, we can learn such h following the framework of GANs. For example,

$$\hat{h} = \text{argmin}_h \max_{k \in \{-1,1\}} \mathcal{W}(h(x), z \mid y),$$

where z is the Gaussian mixture defined in Definition 3.2 with $\theta^* = 1_p/\sqrt{p}$ and $\sigma = 1$, and $\mathcal{W}(\cdot, \cdot)$ denotes the Wasserstein distance. Due to the flexibility of h, the choice of θ^* and σ will not impact the training process.

Now we consider the following two classifiers:

$$\hat{\mathcal{C}}(x) = \text{sgn}(\hat{\theta}^\top \hat{h}(x)),$$

where $\hat{\theta} = \sum_{i=1}^{n} \hat{h}(x_i)y_i/n$, and

$$\hat{\mathcal{C}}^{mix}(x) = \text{sgn}(\hat{\theta}^{mix\top} \hat{h}(x))$$

where

$$\hat{\theta}^{mix} = \sum_{i,j=1}^{n} \mathbb{E}_{\lambda \sim \mathcal{D}_\lambda}(\lambda \hat{h}(x_i) + (1 - \lambda)\hat{h}(x_j)) \cdot \tilde{y}_{i,j}(\lambda)/n^2.$$

Similarly, for a generic $\hat{\theta}$, the confidence scores are given by

$$p_k(x) = 1/(e^{-2k \cdot \hat{\theta}^\top \hat{h}(x)} + 1).$$

We then have the following result showing that under the more general Gaussian generative model, the Mixup method could still provably lead to an improvement on the calibration.

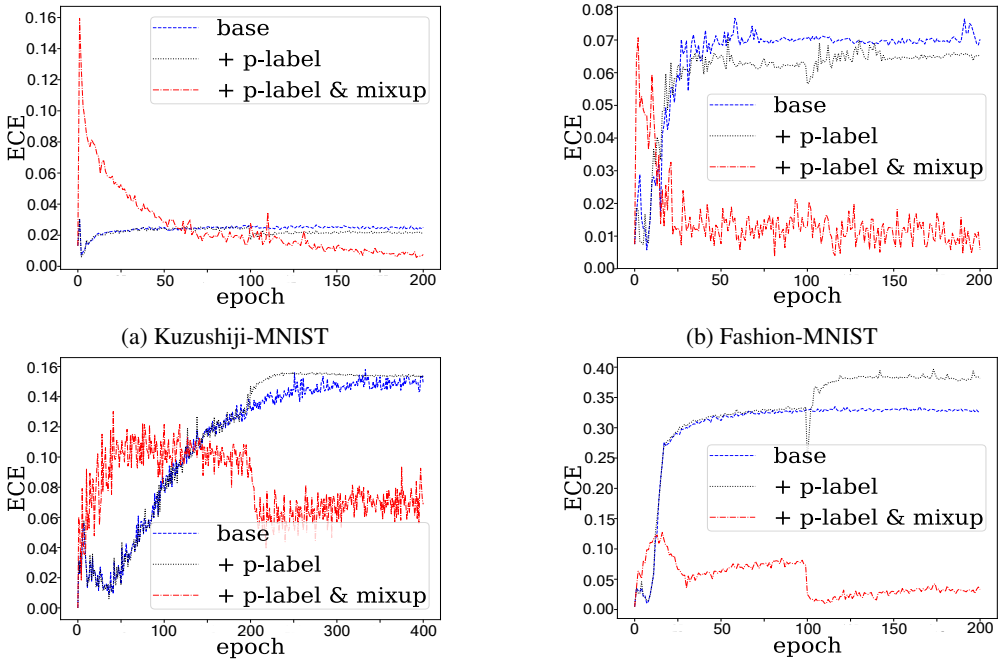
Theorem 3.5. *Under the settings described above with Assumption 3.1, there exists $c_2 > c_1 > 0$, when $p/n \in (c_1, c_2)$, $\hat{h}$ is L-Lipschitz, and $\|\theta\|_2 < C$ for some universal constants $L, C > 0$ (not depending on n and p), then for sufficiently large p and n, there exist $\alpha, \beta > 0$ for the Mixup distribution $\mathcal{D}_\lambda = Beta(\alpha, \beta)$, such that, with high probability,*

$$ECE(\hat{\mathcal{C}}^{mix}) < ECE(\hat{\mathcal{C}}).$$

4. Mixup Improves Calibration in Semi-supervised Learning

Data augmentation by incorporating cheap unlabeled data from multiple domains is a powerful way to improve prediction accuracy especially when there is limited labeled data. One of the commonly used semi-supervised learning algorithms is the pseudo-labeling algorithm (Chapelle et al., 2009), which first trains an initial classifier $\hat{\mathcal{C}}_{init}$ on the labeled data, then assigns pseudo-labels to the unlabeled data using the $\hat{\mathcal{C}}_{init}$. Lastly, using the combined labeled and pseudo-labeled data to perform supervised learning and obtain a final classifier $\hat{\mathcal{C}}_{final}$. Previous work has shown that the pseudo-labeling algorithm has many benefits such as improving prediction accuracy and robustness against adversarial attacks (Carmon et al., 2019). However, as we observe from Figure 3c and 3d, incorporating unlabeled data via the pseudo-labeling algorithm does not always improve calibration; sometimes pseudo-labeling even hurts calibration. We find that further applying Mixup at the last step of pseudo-labeling algorithm mitigates this issue and improves calibration as shown in Figure 3. The details of the experimental setup are included in the Appendix

We justify the empirical findings above by theoretically analyzing the calibration in the semi-supervised learning setting. Specifically, we assume we have n_l labeled data points $\{x_i, y_i\}_{i=1}^{n_l}$ and n_u unlabeled data points $\{x_i^u\}_{i=1}^{n_u}$



(a) Kuzushiji-MNIST (b) Fashion-MNIST

(c) CIFAR-10 (d) CIFAR-100

Figure 3. ECE calculated for ResNets on varieties of data sets. In (a) and (b), using only pseudo-label algorithm improves calibration, while in (c) and (d), using only pseudo-label algorithm hurts calibration. Further applying Mixup in the last step of pseudo-label algorithm promotes calibration in both cases. The pseudo-labels (or p-labels in short) are inserted into training at the midpoint of the entire training: i.e., at epoch = 100 for (a), (b) and (d) and epoch = 200 for (c).

i.i.d. sampled from the (θ^*, σ)-Gaussian model in Definition 3.1. The pseudo-labeling algorithm is the same as the one considered in Carmon et al. (2019), which is shown in Algorithm 1.

We then present two theorems. The first theorem demonstrates that when the labeled data is not sufficient, then under some mild conditions, the unlabeled data will help the calibration. The second theorem characterizes settings where the standard pseudo-labeling algorithm (Algorithm 1) makes calibration worse and increases ECE.

Algorithm 1 The pseudo-labeling algorithm

Step 1: Obtain an initial classifier

$$\hat{\mathcal{C}}_{init}(x) = sgn(\hat{\theta}_{init}^\top x),$$

where $\hat{\theta}_{init} = \sum_{i=1}^{n_l} x_i y_i / n_l$.

Step 2: Apply $\hat{\mathcal{C}}_{init}$ on the unlabeled data set $\{x_i^u\}_{i=1}^{n_u}$, and obtain pseudo-labels $y_i^u = \hat{\mathcal{C}}_{init}(x_i^u)$ for $i \in [n_u]$.

Step 3: Obtain the final classifier $\hat{\mathcal{C}}_{final}(x) = sgn(\hat{\theta}_{final}^\top x)$, where

$$\hat{\theta}_{final} = \frac{1}{n_l + n_u}\left(\sum_{i=1}^{n_l} x_i y_i + \sum_{i=1}^{n_u} x_i^u y_i^u\right)$$

Theorem 4.1. *Suppose $C_1\sqrt{p/n_l} \le \|\theta\| \le C_2\sqrt{p/n_l}$ for some universal constant $C_1 < 1/2$ and $C_2 > 2$, when p/n_l, $\|\theta\|$, n_u are sufficiently large, we have with high probability,*

$$ECE(\hat{\mathcal{C}}_{final}) < ECE(\hat{\mathcal{C}}_{init}).$$

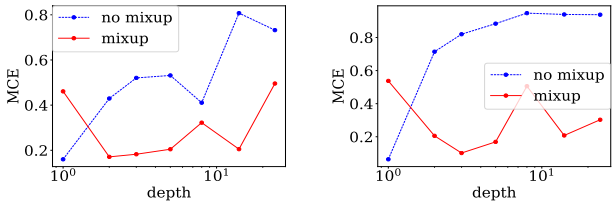
Meanwhile, in some cases, for instance, when the labeled data is sufficient, the pseudo-labeling algorithm may hurt the calibration, as shown in the following theorem.

Theorem 4.2. *If $C_1\|\theta\| < C_2$, $p < C_3$ for some constants $C_1, C_2, C_3 > 0$. Let n_l and $n_u \to \infty$, then with high probability,*

$$ECE(\hat{\mathcal{C}}_{init}) < ECE(\hat{\mathcal{C}}_{final}).$$

The above two theorems suggest that the pseudo-labeling algorithm is not able to robustly guarantee improvement in calibration. In the following, we show that we can mitigate this issue by applying Mixup to the last step in Algorithm 1. Specifically, we consider the following classifier with Mixup:

$$\hat{\mathcal{C}}_{mix,final}(x) = sgn(\hat{\theta}_{mix,final}^\top x),$$



(a) CIFAR-10 (b) CIFAR-100

Figure 4. Maximum Calibration Error (MCE) calculated with varying network depth. Mixup augmentation can reduce MCE especially for larger capacity models (deeper networks) compared to these models trained without Mixup.

where

$$\hat{\theta}_{final,mix}(\lambda) = \mathbb{E}_{\lambda\sim\mathcal{D}_\lambda}[\frac{1}{n_l+n_u}\sum_{i=1}^{n_l+n_u} x_{i,j}^{l,u}(\lambda)y_{i,j}^{l,u}(\lambda)].$$

Here $\{x_{i,j}^{l,u}(\lambda), y_{i,j}^{l,u}(\lambda)\}_{i,j=1}^{n_l+n_u}$ is the data set obtained by applying Mixup to the pooled data set by combining $\{x_i, y_i\}_{i=1}^{n_l}$ and $\{x_i^u, y_i^u\}_{i=1}^{n_u}$. We then have the following result showing Mixup helps the calibration in the semi-supervised setting.

Theorem 4.3. *Under the setup described above, and denote the ECE of $\hat{\mathcal{C}}_{final}$ and $\hat{\mathcal{C}}_{mix,final}$ by $ECE(\hat{\mathcal{C}}_{final})$ and $ECE(\hat{\mathcal{C}}_{mix,final})$ respectively. If $C_1 < \|\theta\| < C_2$ for some universal constants C_1, C_2 (not depending on n and p), then for sufficiently large p and n_l, n_u, there exists $\alpha, \beta > 0$, such that when the Mixup distribution $\lambda \sim Beta(\alpha, \beta)$, with high probability, we have*

$$ECE(\hat{\mathcal{C}}_{mix,final}) < ECE(\hat{\mathcal{C}}_{final}).$$

From Theorem 4.3, we can see that even though incorporating unlabeled data can sometimes make the model less calibrated, adding Mixup training consistently (i.e., under the same conditions of either Theorem 4.1 or Theorem 4.2) mitigates this issue and provably improves calibration.

5. Extension to Maximum Calibration Error

Here we further investigate how Mixup helps calibration under maximum calibration error. Similar conclusions can be reached for MCE as those for ECE in Section 3.2, demonstrating that the effects of Mixup can be found across common calibration metrics.

Theorem 5.1. *Under the settings described in Theorem 3.1, there exists $c_2 > c_1 > 0$, when $p/n \in (c_1, c_2)$ and $\|\theta\|_2 < C$ for some universal constants $C > 0$ (not depending on n and p), then for sufficiently large p and n, there exist $\alpha, \beta > 0$, such that when the distribution $\mathcal{D}_\lambda$ is chosen as $Beta(\alpha, \beta)$, with high probability,*

$$MCE(\hat{\mathcal{C}}^{mix}) < MCE(\hat{\mathcal{C}}).$$

From the above theorem, we can see Mixup can also help decrease the maximum calibration error. Comparing with ECE, from Figure 4, we can similarly observe that when the model capacity is small, Mixup does not really help. We here provide the following theorem to further illustrate that point.

Theorem 5.2. *There exists a threshold $\tau = o(1)$ such that if $p/n \le \tau$ and $\|\theta\|_2 < C$ for some universal constant $C > 0$, given any constants $\alpha, \beta > 0$ (not depending on n and p), when n is sufficiently large, we have, with high probability,*

$$MCE(\hat{\mathcal{C}}) < MCE(\hat{\mathcal{C}}^{mix}).$$

Lastly, we provide a similar theorem as Theorem 3.3 to further illustrate that Mixup helps in the high-dimensional (overparametrized) regime.

Theorem 5.3. *For any constant $c_{\max} > 0$, $p/n \to c_{ratio} \in (0, c^{\max})$, when θ is sufficiently large (still of a constant level), we have for any $\beta > 0$, with high probability, the change of ECE by using Mixup, characterized by*

$$\frac{d}{d\alpha} MCE(\hat{\mathcal{C}}_{\alpha,\beta}^{mix}) \mid_{\alpha\to 0+}$$

is negative, and monotonically decreasing with respect to c_{ratio}.

6. Conclusion and Discussion

Mixup is a popular data augmentation scheme and it has been empirically shown to improve calibration in machine learning. In this paper, we provide a theoretical point of view on how and when Mixup helps the calibration, by studying data generative models. We identify that the calibration improvement induced by Mixup is a high-dimensional phenomenon, and that such reduction in ECE becomes more substantial when the dimension is compared to the number of samples. This suggests that Mixup can be especially helpful for calibration in low sample regime where post-hoc calibration approaches like Platt-scaling are not commonly used. We further study the relationship between Mixup and calibration in a semi-supervised setting when there is an abundance of unlabeled data. Using unlabeled data alone can hurt calibration in some settings, while combining Mixup with pseudo-labeling can mitigate this issue.

Our work points to a few promising further directions. Since there are many variants of Mixup (Berthelot et al., 2019; Verma et al., 2019; Roady et al., 2020; Kim et al., 2020), it would be interesting to study how these extensions of Mixup affect calibration. Another interesting direction is to use the analysis and framework developed in this paper to study the semi-supervised setting where the unlabeled data come from a different domain than the target one. It would be interesting to study how the calibration will change by leveraging the out-of-domain unlabeled data.

Acknowledgements

The research of Linjun Zhang is partially supported by NSF DMS-2015378. The research of Zhun Deng is supported by the Sloan Foundation grants, the NSF grant 1763665, and the Simons Foundation Collaboration on the Theory of Algorithmic Fairness. James Zou is supported by funding from NSF CAREER and the Sloan Fellowship.

References

Bahnsen, A. C., Stojanovic, A., Aouada, D., and Ottersten, B. Improving credit card fraud detection with calibrated probabilities. In *Proceedings of the 2014 SIAM international conference on data mining*, pp. 677–685. SIAM, 2014.

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., and Raffel, C. Mixmatch: A holistic approach to semi-supervised learning. *arXiv preprint arXiv:1905.02249*, 2019.

Carmon, Y., Raghunathan, A., Schmidt, L., Liang, P., and Duchi, J. C. Unlabeled data improves adversarial robustness. *Advances in Neural Information Processing Systems*, 2019.

Chan, A., Alaa, A., Qian, Z., and Van Der Schaar, M. Unlabelled data improves bayesian uncertainty calibration under covariate shift. In *International Conference on Machine Learning*, pp. 1392–1402. PMLR, 2020.

Chapelle, O., Scholkopf, B., and Zien, A. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009.

Clanuwat, T., Bober-Irizar, M., Kitamoto, A., Lamb, A., Yamamoto, K., and Ha, D. Deep learning for classical japanese literature. In *NeurIPS Creativity Workshop 2019*, 2019.

Dan, C., Wei, Y., and Ravikumar, P. Sharp statistical guarantees for adversarially robust gaussian classification. In *International Conference on Machine Learning*, pp. 2345–2355. PMLR, 2020.

Deng, Z., Ding, F., Dwork, C., Hong, R., Parmigiani, G., Patil, P., and Sur, P. Representation via representations: Domain generalization via adversarially learned invariant representations. *arXiv preprint arXiv:2006.11478*, 2020a.

Deng, Z., He, H., Huang, J., and Su, W. Towards understanding the dynamics of the first-order adversaries. In *International Conference on Machine Learning*, pp. 2484–2493. PMLR, 2020b.

Deng, Z., Huang, J., and Kawaguchi, K. How shrinking gradient noise helps the performance of neural networks. In *2021 IEEE International Conference on Big Data (Big Data)*, pp. 1002–1007. IEEE, 2021a.

Deng, Z., Zhang, L., Ghorbani, A., and Zou, J. Improving adversarial robustness via unlabeled out-of-domain data. *International Conference on Artificial Intelligence and Statistics*, 2021b.

Deng, Z., Zhang, L., Vodrahalli, K., Kawaguchi, K., and Zou, J. Y. Adversarial training helps transfer learning via better representations. *Advances in Neural Information Processing Systems*, 34, 2021c.

Foster, D. P. and Stine, R. A. Variable selection in data mining: Building a predictive model for bankruptcy. *Journal of the American Statistical Association*, 99(466):303–313, 2004.

Foster, D. P. and Vohra, R. V. Calibrated learning and correlated equilibrium. *Games and Economic Behavior*, 21(1-2):40, 1997.

Gneiting, T. and Raftery, A. E. Weather forecasting with ensemble methods. *Science*, 310(5746):248–249, 2005.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330. PMLR, 2017.

Guo, H., Mao, Y., and Zhang, R. Mixup as locally linear out-of-manifold regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3714–3722, 2019.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016a.

He, K., Zhang, X., Ren, S., and Sun, J. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, pp. 630–645. Springer, 2016b.

Huang, Y., Li, W., Macheret, F., Gabriel, R. A., and Ohno-Machado, L. A tutorial on calibration measurements and calibration models for clinical prediction models. *Journal of the American Medical Informatics Association*, 27(4): 621–633, 2020.

Ji, W., Deng, Z., Nakada, R., Zou, J., and Zhang, L. The power of contrast for feature learning: A theoretical analysis. *arXiv preprint arXiv:2110.02473*, 2021a.

Ji, W., Lu, Y., Zhang, Y., Deng, Z., and Su, W. J. An unconstrained layer-peeled perspective on neural collapse. *arXiv preprint arXiv:2110.02796*, 2021b.

Jiang, X., Osl, M., Kim, J., and Ohno-Machado, L. Calibrating predictive model estimates to support personalized medicine. *Journal of the American Medical Informatics Association*, 19(2):263–274, 2012.

Johnson, R. A., Wichern, D. W., et al. *Applied multivariate statistical analysis*, volume 5. Prentice hall Upper Saddle River, NJ, 2002.

Kawaguchi, K., Zhang, L., and Deng, Z. Understanding dynamics of nonlinear representation learning and its application. *Neural Computation*, 34(4):991–1018, 2022.

Kim, J.-H., Choo, W., and Song, H. O. Puzzle mix: Exploiting saliency and local statistics for optimal mixup. In *International Conference on Machine Learning*, pp. 5275–5285. PMLR, 2020.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Kuleshov, V., Fenner, N., and Ermon, S. Accurate uncertainties for deep learning using calibrated regression. In *International Conference on Machine Learning*, pp. 2796–2804. PMLR, 2018.

Kumar, A., Liang, P., and Ma, T. Verified uncertainty calibration. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 3792–3803, 2019.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *arXiv preprint arXiv:1612.01474*, 2016.

Naeini, M. P., Cooper, G., and Hauskrecht, M. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Nixon, J., Dusenberry, M. W., Zhang, L., Jerfel, G., and Tran, D. Measuring calibration in deep learning. In *CVPR Workshops*, volume 2, 2019.

Osband, I., Blundell, C., Pritzel, A., and Van Roy, B. Deep exploration via bootstrapped dqn. *arXiv preprint arXiv:1602.04621*, 2016.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., and Snoek, J. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *arXiv preprint arXiv:1906.02530*, 2019.

Platt, J. et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

Roady, R., Hayes, T. L., and Kanan, C. Improved robustness to open set inputs via tempered mixup. In *European Conference on Computer Vision*, pp. 186–201. Springer, 2020.

Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., and Madry, A. Adversarially robust generalization requires more data. *Advances in neural information processing systems*, 31, 2018.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Srivastava, R. K., Greff, K., and Schmidhuber, J. Highway networks. *arXiv preprint arXiv:1505.00387*, 2015.

Thulasidasan, S., Chennupati, G., Bilmes, J., Bhattacharya, T., and Michalak, S. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *arXiv preprint arXiv:1905.11001*, 2019.

Tomani, C. and Buettner, F. Towards trustworthy predictions from deep neural networks with fast adversarial calibration. *arXiv preprint arXiv:2012.10923*, 2020.

Verma, V., Lamb, A., Beckham, C., Najafi, A., Mitliagkas, I., Lopez-Paz, D., and Bengio, Y. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning*, pp. 6438–6447. PMLR, 2019.

Wen, Y., Jerfel, G., Muller, R., Dusenberry, M. W., Snoek, J., Lakshminarayanan, B., and Tran, D. Combining ensembles and data augmentation can harm your calibration. *arXiv preprint arXiv:2010.09875*, 2020.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Zhang, L., Deng, Z., Kawaguchi, K., Ghorbani, A., and Zou, J. How does mixup help with robustness and generalization? *arXiv preprint arXiv:2010.04819*, 2020.

Zhao, S., Ma, T., and Ermon, S. Individual calibration with randomized forecasting. In *International Conference on Machine Learning*, pp. 11387–11397. PMLR, 2020.

Zhu, X. and Goldberg, A. B. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1):1–130, 2009.

Zhu, X., Ghahramani, Z., and Lafferty, J. D. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International conference on Machine learning (ICML-03)*, pp. 912–919, 2003.

Appendix

A. Technical Details

A.1. Proof of Theorem 3.1

Theorem A.1 (Restatement of Theorem 3.1)**.** *Under the settings described in the main paper, if $p/n \to c$ and $\|\theta\|_2 < C$ for some universal constants $c, C > 0$ (not depending on n and p), then for sufficiently large p and n, there exist $\alpha, \beta > 0$, such that when the distribution $\mathcal{D}_\lambda$ is chosen as $Beta(\alpha, \beta)$, with high probability,*

$$ECE(\hat{\mathcal{C}}^{mix}) < ECE(\hat{\mathcal{C}}).$$

Proof. For the clarity of technical proofs, let us write the true parameter θ^* as θ, and denote $\hat{\theta}(0) = \frac{1}{n}\sum_{i=1}^n x_i y_i$. Additionally, since we assume σ is known in the main paper, without loss of generality (otherwise we can consider the data as x_i/σ), we let $\sigma = 1$ throughout the proof.

For the mixup estimator, we have

$$\begin{aligned}
\hat{\theta}(\lambda) =& \frac{1}{n^2}\sum_{i,j=1}^n (\lambda x_i + (1-\lambda)x_j)(y_i + (1-\lambda)y_j) \\
=& \frac{1}{n^2}\sum_{i,j=1}^n (\lambda^2 x_i y_i + (1-\lambda)^2 x_j y_j + \lambda(1-\lambda)x_i y_j + \lambda(1-\lambda)x_j y_i) \\
=& [1 - 2\lambda(1-\lambda)]\frac{1}{n}\sum_{i=1}^n x_i y_i + 2\lambda(1-\lambda)\frac{1}{n}\sum_{i=1}^n x_i \cdot \frac{1}{n}\sum_{i=1}^n y_i.
\end{aligned}$$

Then when $\lambda \sim Beta(\alpha, \beta)$, we have

$$\hat{\theta}^{mix} = \mathop{\mathbb{E}}_{\lambda}[\hat{\theta}(\lambda)] = \frac{(\alpha^2+\beta^2)(\alpha+\beta+1) + 2\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}\hat{\theta}(0) + \frac{2\alpha\beta(\alpha+\beta)}{(\alpha+\beta)^2(\alpha+\beta+1)}\frac{1}{2n}\sum_{i=1}^n x_i \cdot \frac{1}{n}\sum_{i=1}^n y_i$$

For the ease of presentation, we write

$$\hat{\theta}(t) = (1-t)\hat{\theta}(0) + t\epsilon,$$

where $\epsilon = \frac{1}{n}\sum_{i=1}^n x_i \cdot \frac{1}{n}\sum_{i=1}^n y_i$.

It is easy to see $t \in (0, 1]$ when $\alpha \in [0, \infty)$ and $\beta \in (0, \infty)$.

Under our model assumption, we have $\|\frac{1}{n}\sum_{i=1}^n x_i\| = O_p(\sqrt{\frac{p}{n}})$, $|\frac{1}{n}\sum_{i=1}^n y_i| = O_p(\sqrt{\frac{1}{n}})$, and therefore $\|\epsilon\| = O_p(\frac{\sqrt{p}}{n})$.

Now let us consider the expected calibration error

$$ECE = \mathop{\mathbb{E}}_{v=(\hat{\theta}(t))^\top X} |\,\mathbb{E}[Y = 1 \mid \hat{f}(X) = \frac{1}{e^{-2v}+1}] - \frac{1}{e^{-2v}+1}|.$$

We further expand this quantity as

$$\begin{aligned}
\mathbb{E}[Y = 1 \mid \hat{f}(X) = \frac{1}{e^{-2v}+1}] =& \mathbb{E}[Y = 1 \mid \hat{\theta}(t)^\top X = v] \\
=& \frac{\mathbb{P}(\hat{\theta}(t)^\top X = v \mid Y = 1)}{\mathbb{P}(\hat{\theta}^\top X = v \mid Y = 1) + \mathbb{P}(\hat{\theta}(t)^\top X = v \mid Y = -1)} \\
=& \frac{e^{-\frac{(v-\hat{\theta}(t)^\top\theta)^2}{2\|\hat{\theta}(t)\|^2}}}{e^{-\frac{(v-\hat{\theta}(t)^\top\theta)^2}{2\|\hat{\theta}(t)\|^2}} + e^{-\frac{(v+\hat{\theta}(t)^\top\theta)^2}{2\|\hat{\theta}(t)\|^2}}} \\
=& \frac{1}{e^{-\frac{2\hat{\theta}(t)^\top\theta}{\|\hat{\theta}(t)\|^2}\cdot v} + 1}.
\end{aligned}$$

Since $\frac{\hat{\theta}(t)^\top \theta}{\|\hat{\theta}(t)\|^2} = \frac{((1-t)\hat{\theta}(0)+t\epsilon)^\top \theta}{\|(1-t)\hat{\theta}(0)+t\epsilon)\|_2^2} = \frac{1}{1-t}\frac{\hat{\theta}(0)^\top \theta}{\|\hat{\theta}(0)\|^2} + O_P(\frac{\sqrt{p}}{n})$ and $\hat{\theta}(t)^\top X = (1-t)\hat{\theta}(0)^\top X + O_P(\frac{\sqrt{p}}{n})$, we then have

$$\begin{aligned}
ECE &= \mathop{\mathbb{E}}_{v=(\hat{\theta}(t))^\top X} |\, \mathbb{E}[Y=1 \mid \hat{f}(X) = \frac{1}{e^{-2v}+1}] - \frac{1}{e^{-2v}+1}| \\
&= \mathop{\mathbb{E}}_{v=(\hat{\theta}(t))^\top X} [|\frac{1}{e^{-\frac{2\hat{\theta}(t)^\top \theta}{\|\hat{\theta}(t)\|^2}\cdot v}+1} - \frac{1}{e^{-2v}+1}|] \\
&= \mathop{\mathbb{E}}_{v=(1-t)(\hat{\theta}(0))^\top X} [|\frac{1}{e^{-\frac{2\hat{\theta}(0)^\top \theta}{(1-t)\|\hat{\theta}(0)\|^2}\cdot v}+1} - \frac{1}{e^{-2v}+1}|] + O_P(\frac{\sqrt{p}}{n}) \\
&= \mathop{\mathbb{E}}_{v=(\hat{\theta}(0))^\top X} [|\frac{1}{e^{-\frac{2\hat{\theta}(0)^\top \theta}{\|\hat{\theta}(0)\|^2}\cdot v}+1} - \frac{1}{e^{-2(1-t)v}+1}|] + O_P(\frac{\sqrt{p}}{n}).
\end{aligned}$$

Now let us consider the quantity $\frac{\hat{\theta}(0)^\top \theta}{\|\hat{\theta}(0)\|^2}$.

Since we have $\hat{\theta}(0) = \theta + \epsilon_n$ with $\epsilon_n = \frac{1}{n}\sum_{i=1}^n x_i y_i - \theta$, this implies

$$\frac{\hat{\theta}(0)^\top \theta}{\|\hat{\theta}(0)\|^2} = \frac{\|\theta\|^2 + \epsilon_n^\top \theta}{\|\theta\|^2 + \|\epsilon_n\|^2 + 2\epsilon_n^\top \theta} \sim \frac{\|\theta\|^2 + \frac{1}{\sqrt{n}}\|\theta\|}{\|\theta\|^2 + \frac{p}{n} + O_P(\frac{\sqrt{p}}{n}) + \frac{1}{\sqrt{n}}\|\theta\|},$$

where the last equality uses the fact that $\|\epsilon_n\|^2 \overset{d}{=} \frac{\chi_p^2}{n} = \frac{p+O_P(\sqrt{p})}{n} = \frac{p}{n} + O_P(\frac{\sqrt{p}}{n})$.

By our assumption, we have $p/n \in (c_1, c_2)$ and $\|\theta\| < C$, implying that there exists a constant $c_0 \in (0,1)$, such that with high probability

$$\frac{\hat{\theta}(0)^\top \theta}{\|\hat{\theta}(0)\|^2} \le c_0.$$

Then on the event $\mathcal{E} = \{\frac{\hat{\theta}(0)^\top \theta}{\|\hat{\theta}(0)\|^2} \le c_0\}$, if we choose $t = 1 - c_0$, we will then have for any $v \in \mathbb{R}$,

$$|\frac{1}{e^{-\frac{2\hat{\theta}(0)^\top \theta}{\|\hat{\theta}(0)\|^2}\cdot v}+1} - \frac{1}{e^{-2(1-t)v}+1}| < |\frac{1}{e^{-\frac{2\hat{\theta}(0)^\top \theta}{\|\hat{\theta}(0)\|^2}\cdot v}+1} - \frac{1}{e^{-2v}+1}|,$$

and moreover, the difference is lower bounded by $|\frac{1}{e^{-2v}+1} - \frac{1}{e^{-2c_0 v}+1}|$. Use the fact that (since v and c_0 does not depend on n),

$$\mathop{\mathbb{E}}_{v=(\hat{\theta}(t))^\top X} |\frac{1}{e^{-2v}+1} - \frac{1}{e^{-2c_0 v}+1}| = \Omega(1),$$

we then have the desired result

$$ECE(\hat{\mathcal{C}}^{mix}) < ECE(\hat{\mathcal{C}}).$$

□

A.2. Proof of Theorem 3.2

Theorem A.2 (Restatement of Theorem 3.2)**.** *In the case where $p/n \to 0$ and $\|\theta\|_2 < C$ for some universal constant $C > 0$, given any constants $\alpha, \beta > 0$ (not depending on n and p), we have, with high probability,*

$$ECE(\hat{\mathcal{C}}) < ECE(\hat{\mathcal{C}}^{mix}).$$

Proof. According to the proof in Theorem 3.1, we have

$$ECE(\hat{\mathcal{C}}) = \mathop{\mathbb{E}}_{v=(\hat{\theta}(0))^\top X} [|\frac{1}{e^{-\frac{2\hat{\theta}(0)^\top \theta}{\|\hat{\theta}(0)\|^2}\cdot v}+1} - \frac{1}{e^{-2v}+1}|],$$

and

$$ECE(\hat{\mathcal{C}}^{mix}) = \mathop{\mathbb{E}}_{v=(\hat{\theta}(0))^\top X}[|\frac{1}{e^{-\frac{2\hat{\theta}(0)^\top\theta}{\|\hat{\theta}(0)\|^2}\cdot v}+1} - \frac{1}{e^{-2(1-t)v}+1}|] + O_P(\frac{\sqrt{p}}{n}),$$

where $t \in (0,1)$ is a fixed constant when $\alpha, \beta > 0$ are some fixed constants.

When $p/n \to 0$, then

$$\frac{\hat{\theta}(0)^\top\theta}{\|\hat{\theta}(0)\|^2} \sim \frac{\|\theta\|^2 + \frac{1}{\sqrt{n}}\|\theta\|}{\|\theta\|^2 + \frac{p}{n} + O_P(\frac{\sqrt{p}}{n}) + \frac{1}{\sqrt{n}}\|\theta\|} = 1 + O_P(\frac{p}{n}) = 1 + o_P(1).$$

Therefore, we have

$$\begin{aligned} ECE(\hat{\mathcal{C}}) =& \mathop{\mathbb{E}}_{v=(\hat{\theta}(0))^\top X}[|\frac{1}{e^{-\frac{2\hat{\theta}(0)^\top\theta}{\|\hat{\theta}(0)\|^2}\cdot v}+1} - \frac{1}{e^{-2v}+1}|] \\ =& \mathop{\mathbb{E}}_{v=(\hat{\theta}(0))^\top X}[|\frac{1}{e^{-2v+1}} - \frac{1}{e^{-2v}+1}|] + O_P(\frac{p}{n}) \\ =& O_P(\frac{p}{n}) = o_P(1) \end{aligned}$$

and

$$\begin{aligned} ECE(\hat{\mathcal{C}}^{mix}) =& \mathop{\mathbb{E}}_{v=(\hat{\theta}(0))^\top X}[|\frac{1}{e^{-\frac{2\hat{\theta}(0)^\top\theta}{\|\hat{\theta}(0)\|^2}\cdot v}+1} - \frac{1}{e^{-2(1-t)v}+1}|] + O_P(\frac{\sqrt{p}}{n}) \\ =& \mathop{\mathbb{E}}_{v=(\hat{\theta}(0))^\top X}[|\frac{1}{e^{-2v+1}} - \frac{1}{e^{-2(1-t)v}+1}|] + O_P(\frac{p}{n}) \end{aligned}$$

Since $\mathbb{E}_{v=(\hat{\theta}(0))^\top X}[|\frac{1}{e^{-2v+1}} - \frac{1}{e^{-2(1-t)v}+1}|] = \Omega(1)$ when $t \in (0,1)$ is a fixed constant, we then have the desired result that

$$ECE(\hat{\mathcal{C}}) < ECE(\hat{\mathcal{C}}^{mix}).$$

□

A.3. Proof of Theorem 3.3

Theorem A.3 (Restatement of Theorem 3.3)**.** *For any constant $c_{\max} > 0$, $p/n \to c_{ratio} \in (0, c^{\max})$, when θ is sufficiently large (still of a constant level), we have for any $\beta > 0$, with high probability, the change of ECE by using Mixup, characterized by*

$$\frac{d}{d\alpha} ECE(\hat{\mathcal{C}}^{mix}_{\alpha,\beta}) \mid_{\alpha \to 0+}$$

is negative, and monotonically decreasing with respect to c_{ratio}.

Proof. Recall that

$$ECE(\hat{\mathcal{C}}^{mix}) = \mathop{\mathbb{E}}_{v=(\hat{\theta}(0))^\top X}[|\frac{1}{e^{-\frac{2\hat{\theta}(0)^\top\theta}{\|\hat{\theta}(0)\|^2}\cdot v}+1} - \frac{1}{e^{-2(1-t)v}+1}|] + O_P(\frac{\sqrt{p}}{n}).$$

The case $\alpha = 0$ corresponds to the case where $t = 0$. Since $|\frac{1}{e^{-\frac{2\hat{\theta}(0)^\top\theta}{\|\hat{\theta}(0)\|^2}\cdot v}+1} - \frac{1}{e^{-2(1-t)v}+1}|$ as a function of v is symmetric around 0, we have that when t is sufficiently small (such that $1 - t > \frac{\hat{\theta}(0)^\top\theta}{\|\hat{\theta}(0)\|^2}$ with high probability)

$$\begin{aligned} \mathop{\mathbb{E}}_{v=(\hat{\theta}(0))^\top X}[|\frac{1}{e^{-\frac{2\hat{\theta}(0)^\top\theta}{\|\hat{\theta}(0)\|^2}\cdot v}+1} - \frac{1}{e^{-2(1-t)v}+1}|] =& \mathop{\mathbb{E}}_{v=|(\hat{\theta}(0))^\top X|}[|\frac{1}{e^{-\frac{2\hat{\theta}(0)^\top\theta}{\|\hat{\theta}(0)\|^2}\cdot v}+1} - \frac{1}{e^{-2(1-t)v}+1}|] \\ =& \mathop{\mathbb{E}}_{v=|(\hat{\theta}(0))^\top X|}[\frac{1}{e^{-\frac{2\hat{\theta}(0)^\top\theta}{\|\hat{\theta}(0)\|^2}\cdot v}+1} - \frac{1}{e^{-2(1-t)v}+1}]. \end{aligned}$$

Then let us take the derivative with respect to t, we get

$$\mathop{\mathbb{E}}_{v=|(\hat{\theta}(0))^\top X|}[-\frac{e^{-2(1-t)v}\cdot 2v}{(e^{-2(1-t)v}+1)^2}].$$

Therefore, the derivative evaluated at $t=0$ equals to

$$\mathop{\mathbb{E}}_{v=|(\hat{\theta}(0))^\top X|}[-\frac{e^{-2v}\cdot 2v}{(e^{-2v}+1)^2}],$$

which is negative.

We then only need to show it is monotonically decreasing in the rest of this proof.

Again, by symmetry, we only need to consider the distribution of X as $N(\theta, I)$. We have $\hat{\theta}(0)^\top X \sim N(\hat{\theta}(0)^\top\theta, \|\hat{\theta}(0)^\top\|^2)$. Since we have $\hat{\theta}(0)=\theta+\epsilon_n$ with $\epsilon_n=\frac{1}{n}\sum_{i=1}^n x_i y_i-\theta$, we then have $\hat{\theta}(0)^\top\theta=\|\theta\|^2+O_P(\frac{1}{\sqrt{n}})$, and $\|\hat{\theta}(0)\|^2=\|\theta\|^2+p/n+O(\frac{1}{\sqrt{n}})$. In order to show $\mathbb{E}_{v=|(\hat{\theta}(0))^\top X|}[-\frac{e^{-2v}\cdot 2v}{(e^{-2v}+1)^2}]$ is monotonically decreasing, it's sufficient to show that

$$\mathop{\mathbb{E}}_{Z\sim N(0,1)}\frac{\mu+\sigma Z}{e^{-2(\mu+\sigma Z)}+e^{2(\mu+\sigma Z)}+2}$$

is monotonically increasing in $\sigma\in(0, c_{\max})$ when μ is sufficiently large. Let us then take derivative with respect σ, we have

$$\mathop{\mathbb{E}}_{Z\sim N(0,1)}\frac{Z(e^{-2(\mu+\sigma Z)}+e^{2(\mu+\sigma Z)}+2)-(\mu+\sigma Z)\cdot 2Z\cdot(e^{2(\mu+\sigma Z)}-e^{-2(\mu+\sigma Z)})}{(e^{-2(\mu+\sigma Z)}+e^{2(\mu+\sigma Z)}+2)^2}.$$

It suffices to show when μ is sufficiently large, this term is positive.

In fact, when μ is sufficiently large, it suffices to look at dominating term $\mathbb{E}_{Z\sim N(0,1)}[\frac{-2\mu\cdot Z\cdot e^{2(\mu+\sigma Z)}}{e^{4(\mu+\sigma Z)}}]$, for which we have

$$\mathop{\mathbb{E}}_{Z\sim N(0,1)}[\frac{-2\mu\cdot Z\cdot e^{2(\mu+\sigma Z)}}{e^{4(\mu+\sigma Z)}}]=-2\mu\cdot\mathop{\mathbb{E}}_{Z\sim N(0,1)}[Z\cdot e^{-2(\mu+\sigma Z)}]>0.$$

We complete the proof.

□

A.4. Proof of Theorem 3.4

Theorem A.4 (Restatement of Theorem 3.4)**.** *Let us consider the ECE evaluated on the out-of-domain Gaussian model with mean parameter θ', that is, $\mathbb{P}(y=1)=\mathbb{P}(y=-1)=1/2$, and*

$$x\mid y\sim\mathcal{N}(y\cdot\theta', \sigma^2 I),\ \textit{for}\ i=1,2,...,n,$$

If we have $(\theta'-\theta^*)^\top\theta^*\le p/(2n)$, then when p and n are sufficiently large, there exist $\alpha,\beta>0$, such that when the distribution $\mathcal{D}_\lambda$ is chosen as $Beta(\alpha,\beta)$, with high probability,

$$ECE(\hat{\mathcal{C}}^{mix};\theta',\sigma)<ECE(\hat{\mathcal{C}};\theta',\sigma),$$

Proof. When the distribution has mean θ', using the same analysis before, we obtain

$$\begin{aligned}\mathbb{E}[Y=1\mid\hat{f}(X)=\frac{1}{e^{-2v}+1}]=&\mathbb{E}[Y=1\mid\hat{\theta}(t)^\top X=v]\\=&\frac{\mathbb{P}(\hat{\theta}(t)^\top X=v\mid Y=1)}{\mathbb{P}(\hat{\theta}^\top X=v\mid Y=1)+\mathbb{P}(\hat{\theta}(t)^\top X=v\mid Y=-1)}\\=&\frac{1}{e^{-\frac{2\hat{\theta}(t)^\top\theta'}{\|\hat{\theta}(t)\|^2}\cdot v}+1}.\end{aligned}$$

Again, following the same analysis above, it's suffices to show that with high probability,

$$\frac{\hat{\theta}(0)^\top \theta'}{\|\hat{\theta}(0)\|^2} < 1.$$

Again, we use $\hat{\theta}(0) = \theta + \epsilon_n$ with $\epsilon_n = \frac{1}{n}\sum_{i=1}^n x_i y_i - \theta$, we then have

$$\frac{\hat{\theta}(0)^\top \theta'}{\|\hat{\theta}(0)\|^2} = \frac{\theta^\top \theta'}{\|\theta\|^2 + p/n} + O_P(\frac{1}{\sqrt{n}}) = \frac{\theta^\top(\theta' - \theta) + \|\theta\|^2}{\|\theta\|^2 + p/n} + O_P(\frac{1}{\sqrt{n}}).$$

If we have $(\theta' - \theta^*)^\top \theta^* \le p/(2n)$, then when p and n are sufficiently large, there exist $\alpha, \beta > 0$, such that when the distribution $\mathcal{D}_\lambda$ is chosen as $Beta(\alpha, \beta)$, with high probability,

$$\frac{\hat{\theta}(0)^\top \theta'}{\|\hat{\theta}(0)\|^2} < 1.$$

and therefore

$$ECE(\hat{\mathcal{C}}^{mix}; \theta', \sigma) < ECE(\hat{\mathcal{C}}; \theta', \sigma),$$

□

A.5. Proof of Theorem 3.5

Theorem A.5 (Restatement of Theorem 3.5)**.** *Under the settings described above with Assumption 3.1, if $p/n \to c$, $g, \hat{h}$ is L-Lipschitz, and $\|\theta\|_2 < C$ for some universal constants $c, L, C > 0$ (not depending on n and p), then for sufficiently large p and n, there exist $\alpha, \beta > 0$ for the Mixup distribution $\mathcal{D}_\lambda = Beta(\alpha, \beta)$, such that, with high probability,*

$$ECE(\hat{\mathcal{C}}^{mix}) < ECE(\hat{\mathcal{C}}).$$

Proof. Let us first recall Assumption 3.1:

Assumption A.1 (Assumption 3.1 in the main text)**.** *For any given $v \in \mathbb{R}^p$, $k \in \{-1, 1\}$, there exists a $\theta^* \in \mathbb{R}^p$, such that given $y = k$, the probability density function of $R_1 = v^\top \hat{h}(x)$ and $R_2 = v^\top h(x) = v^\top z \sim N(k \cdot b^\top \theta^*, \|v\|^2)$ satisfies that $p_{R_1}(u) = p_{R_2}(u) \cdot (1 + \delta_u)$ for all $u \in \mathbb{R}$ where δ_u satisfies $\mathbb{E}_{R_1}[|\delta_u|^2] = o(1)$ when $n \to \infty$.*

Using the similar analysis from above, we have

$$\begin{aligned}
\mathbb{E}[Y = 1 \mid \hat{f}(X) = \frac{1}{e^{-2v} + 1}] =& \mathbb{E}[Y = 1 \mid \hat{\theta}(t)^\top \hat{h}(X) = v] \\
=& \frac{\mathbb{P}(\hat{\theta}(t)^\top \hat{h}(X) = v \mid Y = 1)}{\mathbb{P}(\hat{\theta}^\top \hat{h}(X) = v \mid Y = 1) + \mathbb{P}(\hat{\theta}(t)^\top \hat{h}(X) = v \mid Y = -1)} \\
=& \frac{e^{-\frac{(v - \hat{\theta}(t)^\top \theta)^2}{2\|\hat{\theta}(t)\|^2}}}{e^{-\frac{(v - \hat{\theta}(t)^\top \theta)^2}{2\|\hat{\theta}(t)\|^2}} + e^{-\frac{(v + \hat{\theta}(t)^\top \theta)^2}{2\|\hat{\theta}(t)\|^2}}}(1 + \delta_v) \\
=& \frac{1}{e^{-\frac{2\hat{\theta}(t)^\top \theta}{\|\hat{\theta}(t)\|^2} \cdot v} + 1}(1 + \delta_v).
\end{aligned}$$

Then by Assumption 3.1 and use the fact that $|\frac{1}{e^{-\frac{2\hat{\theta}(t)^\top \theta}{\|\hat{\theta}(t)\|^2} \cdot v} + 1}| \le 1$, we have the expected calibration error as

$$\begin{aligned}
ECE =& \mathop{\mathbb{E}}_{v = (\hat{\theta}(t))^\top X} |\mathbb{E}[Y = 1 \mid \hat{f}(X) = \frac{1}{e^{-2v} + 1}] - \frac{1}{e^{-2v} + 1}| \\
=& \mathop{\mathbb{E}}_{v = (\hat{\theta}(t))^\top X} [|\frac{1}{e^{-\frac{2\hat{\theta}(t)^\top \theta}{\|\hat{\theta}(t)\|^2} \cdot v} + 1} - \frac{1}{e^{-2v} + 1}|] + o(1)
\end{aligned}$$

Then, by Assumption 3.1, we have $\mathbb{E}[|v^\top(\hat{h}(x)-z)|] \to 0$, where $z \sim \frac{1}{2}N(-\theta, I) + \frac{1}{2}N(\theta, I)$, which implies $\|\frac{1}{n}\sum_{i=1}^n \hat{h}(x_i) - \frac{1}{n}\sum_{i=1}^n z_i\| = o(\sqrt{p})$. Since $\|\frac{1}{n}\sum_{i=1}^n z_i\| = O_P(\frac{\sqrt{p}}{n})$, and $|\frac{1}{n}\sum_{i=1}^n y_i| = O_p(\sqrt{\frac{1}{n}})$, we have

$$\|\frac{1}{n}\sum_{i=1}^n \hat{h}(x_i)\cdot\frac{1}{n}\sum_{i=1}^n y_i\| = o_p(\sqrt{\frac{p}{n}}) = o_p(1).$$

As a result, using the same analysis as those in Section A.1, we have

$$\hat{\theta}^{mix} = \sum_{i,j=1}^n \mathbb{E}_{\lambda\sim\mathcal{D}_\lambda}(\lambda\hat{h}(x_i) + (1-\lambda)\hat{h}(x_j))\cdot\tilde{y}_{i,j}(\lambda)/n^2 = (1-t)\hat{\theta}(0) + o_p(1),$$

and therefore $\frac{\hat{\theta}(t)^\top\theta}{\|\hat{\theta}(t)\|^2} = \frac{((1-t)\hat{\theta}(0)+t\epsilon)^\top\theta}{\|(1-t)\hat{\theta}(0)+t\epsilon)\|_2^2} = \frac{1}{1-t}\frac{\hat{\theta}(0)^\top\theta}{\|\hat{\theta}(0)\|^2} + o_P(1)$ and $\hat{\theta}(t)^\top X = (1-t)\hat{\theta}(0)^\top X + o_P(1)$, we then have

$$\begin{aligned}
ECE &= \mathop{\mathbb{E}}_{v=(\hat{\theta}(t))^\top X}|\,\mathbb{E}[Y=1\mid\hat{f}(X)=\frac{1}{e^{-2v}+1}] - \frac{1}{e^{-2v}+1}| \\
&= \mathop{\mathbb{E}}_{v=(\hat{\theta}(t))^\top X}[|\frac{1}{e^{-\frac{2\hat{\theta}(t)^\top\theta}{\|\hat{\theta}(t)\|^2}\cdot v}+1} - \frac{1}{e^{-2v}+1}|] \\
&= \mathop{\mathbb{E}}_{v=(1-t)(\hat{\theta}(0))^\top X}[|\frac{1}{e^{-\frac{2\hat{\theta}(0)^\top\theta}{(1-t)\|\hat{\theta}(0)\|^2}\cdot v}+1} - \frac{1}{e^{-2v}+1}|] + o_P(1) \\
&= \mathop{\mathbb{E}}_{v=(\hat{\theta}(0))^\top X}[|\frac{1}{e^{-\frac{2\hat{\theta}(0)^\top\theta}{\|\hat{\theta}(0)\|^2}\cdot v}+1} - \frac{1}{e^{-2(1-t)v}+1}|] + o_P(1).
\end{aligned}$$

Again, it boils down to studying the quantity $\frac{\hat{\theta}(0)^\top\theta}{\|\hat{\theta}(0)\|^2}$, and it suffices to show this quantity is smaller than 1 with high probability.

Using the same analysis above, recall that we have $\mathbb{E}[|v^\top(\hat{h}(x)-z)|] \to 0$ for any $\boldsymbol{v}$ with $\|\boldsymbol{v}\| < C$. Let $\tilde{\theta} = \mathbb{E}_x[\hat{h}(x)]$ and plugging in $v = \tilde{\theta}$, we then obtain $|\tilde{\theta}^\top(\tilde{\theta}-\theta)| = o(1)$. Also, plugging in $v=\theta$, we obtain $|\theta^\top(\tilde{\theta}-\theta)| = o(1)$. Combining these two pieces, we obtain

$$\|\tilde{\theta}-\theta\| = o(1).$$

As a result, we have

$$\|\hat{\theta}(0)\|^2 = \|\frac{1}{2n}\sum_{i=1}^n \hat{y}_i h(x_i)\| \geq \|\frac{1}{2n}\sum_{i=1}^n y_i\hat{h}(x_i) - \mathbb{E}[\hat{h}(x)]\| + \|\,\mathbb{E}[\hat{h}(x)]\| = \Omega_P(\sqrt{\frac{p}{n}}) + \|\theta\| + o(1),$$

where the term $\Omega_P(\sqrt{\frac{p}{n}})$ is derived as follows.

First of all, we write

$$\|\frac{1}{2n}\sum_{i=1}^n \hat{y}_i h(x_i) - \mathbb{E}[\hat{h}(x)]\|^2 = \sum_{j=1}^p(\frac{1}{2n}\sum_{i=1}^n \hat{y}_i h_j(x_i) - \mathbb{E}[\hat{h}_j(x)])^2.$$

For each coordinate, we have $Var(z_j) = 1$ and

$$|Var(\hat{h}_j(x_i)) - 1| = o(1). \tag{5}$$

Additionally, since $\hat{h}(x)$ is sub-gaussian, combining with the inequality (5), we have that $\hat{h}_j(x_i)$ has subgaussian norm lower bounded by some constant, which implies

$$\|\frac{1}{n}\sum_{i=1}^n \hat{y}_i h(x_i) - \mathbb{E}[\hat{h}(x)]\|^2 = \sum_{j=1}^p(\frac{1}{n}\sum_{i=1}^n \hat{y}_i h_j(x_i) - \mathbb{E}[\hat{h}_j(x)])^2 = \Omega_P(\frac{p}{n}).$$

Additionally, we have

$$\|\hat{\theta}(0)^\top\theta\| \leq \|\theta\| + o(1).$$

Therefore, we have with high probability,

$$\frac{\hat{\theta}(0)^\top\theta}{\|\hat{\theta}(0)\|^2} < 1.$$

Verification of the unknown σ case When σ is unknown, we estimate σ by $\hat{\sigma} = \sqrt{\|\sum_{i=1}^n (x_i - y_i\hat{\theta})\|^2/pn}$. It's easy to see $|\hat{\sigma} - \sigma| = O_P(1/\sqrt{n})$. We then let $\hat{h}(x) = x/\hat{\sigma}$, and verify for any $v \in \mathbb{R}^p$ with $\|v\| \le C$, $R_1 = v^\top \hat{h}(x) = v^\top x/\hat{\sigma}$ and $R_2 = v^\top x/\sigma$ satisfies that $p_{R_1}(u) = p_{R_2}(u) \cdot (1 + \delta_u)$ for all $u \in \mathbb{R}$ where δ_u satisfies $\mathbb{E}_{R_1}[|\delta_u|^2] = o(1)$ when $n \to \infty$. We have $\hat{h}$ and g are all Lipschitz with constant 2σ.

When y=1, we have

$$p_{R_1}(u) = \frac{1}{\sqrt{2\pi}\sigma/\hat{\sigma}} \exp\{-\frac{(u - v^\top\theta)^2}{2\sigma^2/\hat{\sigma}^2}\}, p_{R_2}(u) = \frac{1}{\sqrt{2\pi}} \exp\{-\frac{(u - v^\top\theta)^2}{2}\}$$

Denote $g(a) = \frac{1}{\sqrt{2\pi}a} \exp\{-\frac{(u-v^\top\theta)^2}{2a^2}\}$, we then have $g'(a) = -\frac{1}{\sqrt{2\pi}a^2} \exp\{-\frac{(u-v^\top\theta)^2}{2a^2}\} + \frac{1}{\sqrt{2\pi}a} \exp\{-\frac{(u-v^\top\theta)^2}{2a^2}\} \cdot \frac{(u-v^\top\theta)^2}{a^3}$ and therefore $g'(1) = -\frac{1}{\sqrt{2\pi}} \exp\{-\frac{(u-v^\top\theta)^2}{2}\} + \frac{1}{\sqrt{2\pi}} \exp\{-\frac{(u-v^\top\theta)^2}{2}\} \cdot (u - v^\top\theta)^2$.

We then have

$$\delta_u = \frac{p_{R_1}(u) - p_{R_2}(u)}{p_{R_2}(u)} = [(u - v^\top\theta)^2 - 1] \cdot (\frac{\sigma^2}{\hat{\sigma}^2} - 1) = O_P(\frac{(u - v^\top\theta)^2 - 1}{\sqrt{n}}).$$

Since $\mathbb{E}_{u\sim R_1}[((u - v^\top\theta)^2 - 1)^2] = O(1)$, we have $\mathbb{E}_{R_1}[|\delta_u|^2] = o(1)$ when $n \to \infty$. □

A.6. Proof of Theorem 4.1

Theorem A.6 (Restatement of Theorem 4.1)**.** *Suppose $C_1\sqrt{p/n_l} \le \|\theta\| \le C_2\sqrt{p/n_l}$ for some universal constant $C_1 < 1/2$ and $C_2 > 2$, when p/n_l, $\|\theta\|$, n_u are sufficiently large, we have with high probability,*

$$ECE(\hat{\mathcal{C}}_{final}) < ECE(\hat{\mathcal{C}}_{init}).$$

Proof. According to the proof in the above section, we have

$$ECE(\hat{C}_{final}) = \mathop{\mathbb{E}}_{v=(\hat{\theta}_{final})^\top X}[|\frac{1}{e^{-\frac{2\hat{\theta}_{final}^\top\theta}{\|\hat{\theta}_{final}\|^2}\cdot v} + 1} - \frac{1}{e^{-2v} + 1}|],$$

and

$$ECE(\hat{C}_{init}) = \mathop{\mathbb{E}}_{v=(\hat{\theta}_{init})^\top X}[|\frac{1}{e^{-\frac{2\hat{\theta}_{init}^\top\theta}{\|\hat{\theta}_{init}\|^2}\cdot v} + 1} - \frac{1}{e^{-2v} + 1}|].$$

For the initial estimator, we use $\hat{\theta}_{init} = \hat{\theta}(0) = \theta + \epsilon_n$ with $\epsilon_n = \frac{1}{n}\sum_{i=1}^n x_i y_i - \theta$, we then have

$$\frac{\hat{\theta}(0)^\top\theta}{\|\hat{\theta}(0)\|^2} = \frac{\|\theta\|^2}{\|\theta\|^2 + p/n_l} + O_P(\frac{1}{\sqrt{n_l}}).$$

When $C_1\sqrt{p/n_l} \le \|\theta\| \le C_2\sqrt{p/n_l}$, we have

$$\frac{\hat{\theta}_{init}^\top\theta}{\|\hat{\theta}_{init}\|^2} = \frac{\hat{\theta}(0)^\top\theta}{\|\hat{\theta}(0)\|^2} \le \frac{C_2^2}{C_2^2 + 1}.$$

In the case where we combine the unlabeled data, we follow the similar analysis of Carmon et al. (2019) to study the property of y_i^u. Let b_i be the indicator that the i-th pseudo-label is incorrect, so that $x_i^u \sim N((1 - 2b_i)y_i^u\theta, I) := (1 - 2b_i)y_i^u\theta + \epsilon_i^u$. Then we can write

$$\hat{\theta}_{final} = \gamma\theta + \tilde{\delta},$$

where $\gamma = \frac{1}{n_u}\sum_{i=1}^{n_u}(1 - 2b_i)$ and $\tilde{\delta} = \frac{1}{n_u}\sum_{i=1}^{n_u}\epsilon_i^u y_i^u$.

We then derive concentration bounds for $\|\tilde{\delta}\|^2$ and $\theta^\top\tilde{\delta}$. Recall that $y_i^u = sgn(\hat{\theta}_{init}^\top x_i^u)$, we choose a coordinate system such that the first coordinate is in the direction of $\hat{\theta}_{init}$, and let $v^{(i)}$ denote the i-th entry of vector v in this coordinate system. Then $y_i^u = sgn(x_i^{u(1)}) = sgn(\theta^\top\hat{\theta}_{init} + \epsilon_i^{u(1)})$.

Under this coordinate system, for $j \geq 2$, we have $\epsilon_i^{u(j)}$ are independent with y_i^u and therefore $\epsilon_i^{u(j)} y_i^u \sim N(0,1)$ for all $j \geq 2$. For the first coordinate, since $\theta^\top \hat{\theta}_{init} = \Omega_P(1)$, we have $|\mathbb{E}[\epsilon_i^{u(1)} y_i^u]| = |\mathbb{E}[\epsilon_i^{u(1)} sgn(\theta^\top \hat{\theta}_{init} + \epsilon_i^{u(1)})]| = \Omega_P(1)$.

Then we have

$$\sum_{j=1}^{p} (\frac{1}{n_u} \sum_{i=1}^{n_u} \epsilon_i^{u(j)} y_i^u)^2 = (\frac{1}{n_u} \sum_{i=1}^{n_u} \epsilon_i^{u(1)} y_i^u)^2 + \sum_{j=2}^{p} (\frac{1}{n_u} \sum_{i=1}^{n_u} \epsilon_i^{u(j)} y_i^u)^2 = \Omega_P(1) + \frac{p + O_p(\sqrt{p})}{n_u}.$$

The same analysis also yields

$$|\tilde{\delta}^\top \theta| = |\frac{1}{n_u} \sum_{i=1}^{n_u} \epsilon_i^{u(1)} y_i^u \theta_1 + \sum_{j=2}^{p} \frac{1}{n_u} \sum_{i=1}^{n_u} \epsilon_i^{u(j)} y_i^u \theta_j| = \Omega_P(|\theta_1|) + O_P(\|\theta\|_2 / \sqrt{n}).$$

Moreover, the proportion of misclassified samples converge the misclassification error produced by $\hat{C}_{init}$:

$$\gamma = \frac{1}{n_u} \sum_{i=1}^{n_u} (1 - 2b_i) \to 1 - \exp(-c\|\theta\|^2) + O_P(\frac{1}{n}).$$

This implies

$$\frac{\hat{\theta}_{final}^\top \theta}{\|\hat{\theta}_{final}\|^2} = \frac{\gamma \|\theta\|^2 + \tilde{\delta}^\top \theta}{\gamma^2 \|\theta\|^2 + \|\tilde{\delta}\|^2 + 2\tilde{\delta}^\top \theta} \sim \frac{\gamma \|\theta\|^2 + \Omega_P(|\theta_1|)}{\gamma^2 \|\theta\|^2 + \Omega_P(\frac{n_u + p}{n_u}) + \Omega_P(|\theta_1|)}.$$

When $\|\theta\|$ is sufficiently large (which implies p/n_l is sufficiently large), we have

$$\frac{C_2^2}{C_2^2 + 1} < \frac{\hat{\theta}_{final}^\top \theta}{\|\hat{\theta}_{final}\|^2} < 1,$$

implying

$$\frac{\hat{\theta}_{init}^\top \theta}{\|\hat{\theta}_{init}\|^2} < \frac{\hat{\theta}_{final}^\top \theta}{\|\hat{\theta}_{final}\|^2} < 1,$$

and therefore

$$ECE(\hat{\mathcal{C}}_{final}) < ECE(\hat{\mathcal{C}}_{init}).$$

□

A.7. Proof of Theorem 4.2

Theorem A.7 (Restatement of Theorem 4.2)**.** *If $\|\theta\| < C$ for some constant $C > 2$, given fixed p and let n_l and $n_u \to \infty$ with p fixed, then with high probability probability,*

$$ECE(\hat{\mathcal{C}}_{init}) < ECE(\hat{\mathcal{C}}_{final}).$$

Proof. Using the same analysis as in the above section, we have

$$ECE(\hat{\mathcal{C}}_{final}) = \mathop{\mathbb{E}}_{v=(\hat{\theta}_{final})^\top X} [|\frac{1}{e^{-\frac{2\hat{\theta}_{final}^\top \theta}{\|\hat{\theta}_{final}\|^2} \cdot v} + 1} - \frac{1}{e^{-2v} + 1}|],$$

and

$$ECE(\hat{\mathcal{C}}_{init}) = \mathop{\mathbb{E}}_{v=(\hat{\theta}_{init})^\top X} [|\frac{1}{e^{-\frac{2\hat{\theta}_{init}^\top \theta}{\|\hat{\theta}_{init}\|^2} \cdot v} + 1} - \frac{1}{e^{-2v} + 1}|].$$

For the initial estimator, we use $\hat{\theta}_{init} = \hat{\theta}(0) = \theta + \epsilon_n$ with $\epsilon_n = \frac{1}{n}\sum_{i=1}^n x_i y_i - \theta$, we then have

$$\frac{\hat{\theta}(0)^\top \theta}{\|\hat{\theta}(0)\|^2} = \frac{\|\theta\|^2}{\|\theta\|^2 + p/n_l} + O_P(\frac{1}{\sqrt{n_l}}).$$

When $\|\theta\| < C$ for some constant $C > 2$, given p fixed and $n_l \to \infty$, we have

$$\frac{\hat{\theta}_{init}^\top \theta}{\|\hat{\theta}_{init}\|^2} = \frac{\hat{\theta}(0)^\top \theta}{\|\hat{\theta}(0)\|^2} \to 1.$$

For the semi-supervised classifier, when $\|\theta\| < C$ and $n_u \to \infty$, we have

$$\frac{\hat{\theta}_{final}^\top \theta}{\|\hat{\theta}_{final}\|^2} = \frac{\gamma\|\theta\|^2 + \tilde{\delta}^\top\theta}{\gamma^2\|\theta\|^2 + \|\tilde{\delta}\|^2 + 2\tilde{\delta}^\top\theta} \sim \frac{\gamma\|\theta\|^2 + \Omega_P(|\theta_1|)}{\gamma^2\|\theta\|^2 + \Omega_P(\frac{n_u+p}{n_u}) + \Omega_P(|\theta_1|)} < 1.$$

As a result,

$$\frac{\hat{\theta}_{final}^\top \theta}{\|\hat{\theta}_{final}\|^2} < \frac{\hat{\theta}_{init}^\top \theta}{\|\hat{\theta}_{init}\|^2} \le 1,$$

and therefore

$$ECE(\hat{\mathcal{C}}_{init}) < ECE(\hat{\mathcal{C}}_{final}).$$

□

A.8. Proof of Theorem 4.3

Theorem A.8 (Restatement of Theorem 4.3). *Under the setup described above, and denote the ECE of $\hat{\mathcal{C}}_{final}$ and $\hat{\mathcal{C}}_{mix,final}$ by $ECE(\hat{\mathcal{C}}_{final})$ and $ECE(\hat{\mathcal{C}}_{mix,final})$ respectively. If $C_1 < \|\theta\|_2 < C_2$ for some universal constants C_1, C_2 (not depending on n and p), then for sufficiently large p and n_l, n_u, there exists $\alpha, \beta > 0$, such that when the Mixup distribution $\lambda \sim Beta(\alpha, \beta)$, with high probability, we have*

$$ECE(\hat{\mathcal{C}}_{mix,final}) < ECE(\hat{\mathcal{C}}_{final}).$$

Proof. Using the same analysis as those in Section A.1, we have

Using the similar analysis from above, we have

$$\begin{aligned}\mathbb{E}[Y = 1 \mid \hat{f}(X) = \frac{1}{e^{-2v}+1}] =& \mathbb{E}[Y = 1 \mid \hat{\theta}^\top X = v] \\ =& \frac{\mathbb{P}(\hat{\theta}^\top X = v \mid Y = 1)}{\mathbb{P}(\hat{\theta}^\top X = v \mid Y = 1) + \mathbb{P}(\hat{\theta}^\top X = v \mid Y = -1)} \\ =& \frac{1}{e^{-\frac{2\hat{\theta}^\top\theta}{\|\hat{\theta}\|^2}\cdot v} + 1}.\end{aligned}$$

Then we have the expected calibration error as

$$\begin{aligned}ECE =& \mathop{\mathbb{E}}_{v=(\hat{\theta}(t))^\top X} |\,\mathbb{E}[Y = 1 \mid \hat{f}(X) = \frac{1}{e^{-2v}+1}] - \frac{1}{e^{-2v}+1}| \\ =& \mathop{\mathbb{E}}_{v=(\hat{\theta}(t))^\top X} [|\frac{1}{e^{-\frac{2\hat{\theta}(t)^\top\theta}{\|\hat{\theta}(t)\|^2}\cdot v} + 1} - \frac{1}{e^{-2v}+1}|] + o(1)\end{aligned}$$

Then, since $\|\frac{1}{n_u}\sum_{i=1}^{n_u} x_i^u\| = O_P(\frac{\sqrt{p}}{n_u})$, and $|\frac{1}{n}\sum_{i=1}^n y_i^u| = O_p(\sqrt{\frac{1}{n_u}})$, we have

$$\|\frac{1}{n_u}\sum_{i=1}^{n_u} x_i^u \cdot \frac{1}{n_u}\sum_{i=1}^{n_u} y_i^u\| = O_p(\frac{\sqrt{p}}{n_u}) = o_p(1).$$

As a result, using the same analysis as those in Section A.1, and denote $\hat{\theta}(0) = \hat{\theta}_{final}$ we have

$$\hat{\theta}_{final,mix} = \sum_{i,j=1}^{n} \mathbb{E}_{\lambda\sim\mathcal{D}_\lambda}(\lambda\hat{h}(x_i) + (1-\lambda)\hat{h}(x_j)) \cdot \tilde{y}_{i,j}(\lambda)/n^2 = (1-t)\hat{\theta}(0) + o_p(1),$$

and therefore $\frac{\hat{\theta}(t)^\top\theta}{\|\hat{\theta}(t)\|^2} = \frac{1}{1-t}\frac{\hat{\theta}(0)^\top\theta}{\|\hat{\theta}(0)\|^2} + o_P(1)$ and $\hat{\theta}(t)^\top X = (1-t)\hat{\theta}(0)^\top X + o_P(1)$, we then have

$$\begin{aligned}
ECE &= \mathop{\mathbb{E}}_{v=(\hat{\theta}(t))^\top X} |\,\mathbb{E}[Y=1 \mid \hat{f}(X) = \frac{1}{e^{-2v}+1}] - \frac{1}{e^{-2v}+1}| \\
&= \mathop{\mathbb{E}}_{v=(\hat{\theta}(t))^\top X} [|\frac{1}{e^{-\frac{2\hat{\theta}(t)^\top\theta}{\|\hat{\theta}(t)\|^2}\cdot v}+1} - \frac{1}{e^{-2v}+1}|] \\
&= \mathop{\mathbb{E}}_{v=(1-t)(\hat{\theta}(0))^\top X} [|\frac{1}{e^{-\frac{2\hat{\theta}(0)^\top\theta}{(1-t)\|\hat{\theta}(0)\|^2}\cdot v}+1} - \frac{1}{e^{-2v}+1}|] + o_P(1) \\
&= \mathop{\mathbb{E}}_{v=(\hat{\theta}(0))^\top X} [|\frac{1}{e^{-\frac{2\hat{\theta}(0)^\top\theta}{\|\hat{\theta}(0)\|^2}\cdot v}+1} - \frac{1}{e^{-2(1-t)v}+1}|] + o_P(1).
\end{aligned}$$

Again, it boils down to studying the quantity $\frac{\hat{\theta}(0)^\top\theta}{\|\hat{\theta}(0)\|^2}$, and it suffices to show this quantity is smaller than 1 with high probability.

To see this, when $C_1 < \|\theta\|_2 < C_2$ and n_u, p sufficiently large, we have with high probability,

$$\frac{\hat{\theta}_{final}^\top\theta}{\|\hat{\theta}_{final}\|^2} = \frac{\gamma\|\theta\|^2 + \tilde{\delta}^\top\theta}{\gamma^2\|\theta\|^2 + \|\tilde{\delta}\|^2 + 2\tilde{\delta}^\top\theta} \sim \frac{\gamma\|\theta\|^2 + \Omega_P(|\theta_1|)}{\gamma^2\|\theta\|^2 + \Omega_P(\frac{n_u+p}{n_u}) + \Omega_P(|\theta_1|)} < 1.$$

Therefore, there exists $\alpha, \beta > 0$, such that when the Mixup distribution $\lambda \sim Beta(\alpha, \beta)$, with high probability, we have

$$ECE(\hat{\mathcal{C}}_{mix,final}) < ECE(\hat{\mathcal{C}}_{final}).$$

□

A.9. Proof of Theorem 5.1

Theorem A.9 (Restatement of Theorem 5.1)**.** *Under the settings described in Theorem 3.1, there exists $c_2 > c_1 > 0$, when $p/n \in (c_1, c_2)$ and $\|\theta\|_2 < C$ for some universal constants $C > 0$ (not depending on n and p), then for sufficiently large p and n, there exist $\alpha, \beta > 0$, such that when the distribution $\mathcal{D}_\lambda$ is chosen as $Beta(\alpha, \beta)$, with high probability,*

$$MCE(\hat{\mathcal{C}}^{mix}) < MCE(\hat{\mathcal{C}}).$$

Proof. Following the calculation of Theorem 3.1, we consider the maximum calibration error,

$$MCE = \max_v |\,\mathbb{E}[Y=1 \mid \hat{f}(X) = \frac{1}{e^{-2v}+1}] - \frac{1}{e^{-2v}+1}|.$$

In addition, we have

$$\mathbb{E}[Y=1 \mid \hat{f}(X) = \frac{1}{e^{-2v}+1}] = \frac{1}{e^{-\frac{2\hat{\theta}(t)^\top\theta}{\|\hat{\theta}(t)\|^2}\cdot v}+1},$$

where

$$t = \frac{2\alpha\beta(\alpha+\beta)}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Let us denote $\rho = \frac{\hat{\theta}(t)^\top\theta}{\|\hat{\theta}(t)\|^2}$. Since $\frac{\hat{\theta}(t)^\top\theta}{\|\hat{\theta}(t)\|^2} = \frac{((1-t)\hat{\theta}(0)+t\epsilon)^\top\theta}{\|(1-t)\hat{\theta}(0)+t\epsilon)\|_2^2} = \frac{1}{1-t}\frac{\hat{\theta}(0)^\top\theta}{\|\hat{\theta}(0)\|^2} + O_P(\frac{\sqrt{p}}{n})$. As a result,

$$
\begin{aligned}
MCE &= \max_v |\,\mathbb{E}[Y=1 \mid \hat{f}(X) = \frac{1}{e^{-2v}+1}] - \frac{1}{e^{-2v}+1}| \\
&= \max_v |\frac{1}{e^{-\frac{2\hat{\theta}(t)^\top\theta}{\|\hat{\theta}(t)\|^2}\cdot v}+1} - \frac{1}{e^{-2v}+1}| \\
&= \max_v |\frac{1}{e^{-\frac{2\hat{\theta}(0)^\top\theta}{(1-t)\|\hat{\theta}(0)\|^2}\cdot v}+1} - \frac{1}{e^{-2v}+1}| + O_P(\frac{\sqrt{p}}{n}) \\
&= \max_v |\frac{1}{e^{-\frac{2\hat{\theta}(0)^\top\theta}{\|\hat{\theta}(0)\|^2}\cdot v}+1} - \frac{1}{e^{-2(1-t)v}+1}| + O_P(\frac{\sqrt{p}}{n}).
\end{aligned}
$$

Now let us consider the quantity $\frac{\hat{\theta}(0)^\top\theta}{\|\hat{\theta}(0)\|^2}$.

Since we have $\hat{\theta}(0) = \theta + \epsilon_n$ with $\epsilon_n = \frac{1}{n}\sum_{i=1}^n x_i y_i - \theta$, this implies

$$
\frac{\hat{\theta}(0)^\top\theta}{\|\hat{\theta}(0)\|^2} = \frac{\|\theta\|^2 + \epsilon_n^\top\theta}{\|\theta\|^2 + \|\epsilon_n\|^2 + 2\epsilon_n^\top\theta} \sim \frac{\|\theta\|^2 + \frac{1}{\sqrt{n}}\|\theta\|}{\|\theta\|^2 + \frac{p}{n} + O_P(\frac{\sqrt{p}}{n}) + \frac{1}{\sqrt{n}}\|\theta\|},
$$

where the last equality uses the fact that $\|\epsilon_n\|^2 \overset{d}{=} \frac{\chi_p^2}{n} = \frac{p+O_P(\sqrt{p})}{n} = \frac{p}{n} + O_P(\frac{\sqrt{p}}{n})$.

By our assumption, we have $p/n \in (c_1, c_2)$ and $\|\theta\| < C$, implying that there exists a constant $c_0 \in (0,1)$, such that with high probability

$$
\frac{\hat{\theta}(0)^\top\theta}{\|\hat{\theta}(0)\|^2} \le c_0.
$$

Then on the event $\mathcal{E} = \{\frac{\hat{\theta}(0)^\top\theta}{\|\hat{\theta}(0)\|^2} \le c_0\}$, if we choose $t = 1 - c_0$, we will then have for any $v \in \mathbb{R}$,

$$
|\frac{1}{e^{-\frac{2\hat{\theta}(0)^\top\theta}{\|\hat{\theta}(0)\|^2}\cdot v}+1} - \frac{1}{e^{-2(1-t)v}+1}| < |\frac{1}{e^{-\frac{2\hat{\theta}(0)^\top\theta}{\|\hat{\theta}(0)\|^2}\cdot v}+1} - \frac{1}{e^{-2v}+1}|,
$$

and moreover, the difference between LHS and RHS is lower bounded by $|\frac{1}{e^{-2v}+1} - \frac{1}{e^{-2c_0 v}+1}|$. Thus, we have

$$
MCE(\hat{\mathcal{C}}^{mix}) < MCE(\hat{\mathcal{C}}).
$$

□

A.10. Proof of Theorem 5.2

Theorem A.10 (Restatement of Theorem 5.2)**.** *There exists a threshold $\tau = o(1)$ such that if $p/n \le \tau$ and $\|\theta\|_2 < C$ for some universal constant $C > 0$, given any constants $\alpha, \beta > 0$ (not depending on n and p), when n is sufficiently large, we have, with high probability,*

$$
MCE(\hat{\mathcal{C}}) < MCE(\hat{\mathcal{C}}^{mix}).
$$

Proof. According to the proof in Theorem 5.1, we have

$$
MCE(\hat{\mathcal{C}}) = \max_v |\frac{1}{e^{-\frac{2\hat{\theta}(0)^\top\theta}{\|\hat{\theta}(0)\|^2}\cdot v}+1} - \frac{1}{e^{-2v}+1}|,
$$

and

$$
MCE(\hat{\mathcal{C}}^{mix}) = \max_v |\frac{1}{e^{-\frac{2\hat{\theta}(0)^\top\theta}{\|\hat{\theta}(0)\|^2}\cdot v}+1} - \frac{1}{e^{-2(1-t)v}+1}| + O_P(\frac{\sqrt{p}}{n}),
$$

where $t \in (0,1)$ is a fixed constant when $\alpha, \beta > 0$ are some fixed constants.

When $p/n \to 0$, then

$$\frac{\hat{\theta}(0)^\top \theta}{\|\hat{\theta}(0)\|^2} \sim \frac{\|\theta\|^2 + \frac{1}{\sqrt{n}}\|\theta\|}{\|\theta\|^2 + \frac{p}{n} + O_P(\frac{\sqrt{p}}{n}) + \frac{1}{\sqrt{n}}\|\theta\|} = 1 + O_P(\frac{p}{n}) = 1 + o_P(1).$$

Therefore, we have

$$\begin{aligned} MCE(\hat{\mathcal{C}}) =& \max_v [|\frac{1}{e^{-\frac{2\hat{\theta}(0)^\top \theta}{\|\hat{\theta}(0)\|^2}\cdot v} + 1} - \frac{1}{e^{-2v}+1}|] \\ =& \max_v [|\frac{1}{e^{-2v+1}} - \frac{1}{e^{-2v}+1}|] + O_P(\frac{p}{n}) \\ =& O_P(\frac{p}{n}) = o_P(1) \end{aligned}$$

and

$$\begin{aligned} MCE(\hat{\mathcal{C}}^{mix}) =& \max_v |\frac{1}{e^{-\frac{2\hat{\theta}(0)^\top \theta}{\|\hat{\theta}(0)\|^2}\cdot v} + 1} - \frac{1}{e^{-2(1-t)v}+1}| + O_P(\frac{\sqrt{p}}{n}) \\ =& |\frac{1}{e^{-2v+1}} - \frac{1}{e^{-2(1-t)v}+1}| + O_P(\frac{p}{n}) \end{aligned}$$

Since $\max_v |\frac{1}{e^{-2v+1}} - \frac{1}{e^{-2(1-t)v}+1}| = \Omega(1)$ when $t \in (0,1)$ is a fixed constant, we then have the desired result that

$$MCE(\hat{\mathcal{C}}) < MCE(\hat{\mathcal{C}}^{mix}).$$

□

A.11. Proof of Theorem 5.3

Theorem A.11 (Restatement of Theorem 5.3)**.** *For any constant $c_{\max} > 0$, $p/n \to c_{ratio} \in (0, c^{\max})$, when θ is sufficiently large (still of a constant level), we have for any $\beta > 0$, with high probability, the change of ECE by using Mixup, characterized by*

$$\frac{d}{d\alpha} MCE(\hat{\mathcal{C}}^{mix}_{\alpha,\beta}) \mid_{\alpha \to 0+}$$

is negative, and monotonically decreasing with respect to c_{ratio}.

Proof. Recall that

$$MCE(\hat{\mathcal{C}}^{mix}) = \max_v |\frac{1}{e^{-\frac{2\hat{\theta}(0)^\top \theta}{\|\hat{\theta}(0)\|^2}\cdot v} + 1} - \frac{1}{e^{-2(1-t)v}+1}| + O_P(\frac{\sqrt{p}}{n}).$$

The case $\alpha = 0$ corresponds to the case where $t = 0$. Since $|\frac{1}{e^{-\frac{2\hat{\theta}(0)^\top \theta}{\|\hat{\theta}(0)\|^2}\cdot v}+1} - \frac{1}{e^{-2(1-t)v}+1}|$ as a function of v is symmetric around 0, we have that when t is sufficiently small (such that $1 - t > \frac{\hat{\theta}(0)^\top \theta}{\|\hat{\theta}(0)\|^2}$ with high probability)

$$\begin{aligned} \max_v |\frac{1}{e^{-\frac{2\hat{\theta}(0)^\top \theta}{\|\hat{\theta}(0)\|^2}\cdot v} + 1} - \frac{1}{e^{-2(1-t)v}+1}| =& \max_{v>0} |\frac{1}{e^{-\frac{2\hat{\theta}(0)^\top \theta}{\|\hat{\theta}(0)\|^2}\cdot v} + 1} - \frac{1}{e^{-2(1-t)v}+1}| \\ =& \max_{v>0} [\frac{1}{e^{-\frac{2\hat{\theta}(0)^\top \theta}{\|\hat{\theta}(0)\|^2}\cdot v} + 1} - \frac{1}{e^{-2(1-t)v}+1}]. \end{aligned}$$

Let us denote

$$v^* = \text{argmax}_{v>0} [\frac{1}{e^{-\frac{2\hat{\theta}(0)^\top \theta}{\|\hat{\theta}(0)\|^2}\cdot v} + 1} - \frac{1}{e^{-2(1-t)v}+1}].$$

For the term

$$\xi(t) = \frac{1}{e^{-\frac{2\hat{\theta}(0)^\top \theta}{\|\hat{\theta}(0)\|^2}\cdot v^*}+1} - \frac{1}{e^{-2(1-t)v^*}+1},$$

let us take the derivative with respect to t, we get

$$\frac{d}{dt}\xi(t) = -\frac{e^{-2(1-t)v^*}\cdot 2v^*}{(e^{-2(1-t)v^*}+1)^2}.$$

Therefore, the derivative evaluated at $t=0$ equals to

$$-\frac{e^{-2v^*}\cdot 2v^*}{(e^{-2v*}+1)^2},$$

for $v^* > 0$, which is negative.

Recall that we have $\hat{\theta}(0) = \theta + \epsilon_n$ with $\epsilon_n = \frac{1}{n}\sum_{i=1}^n x_i y_i - \theta$, this implies

$$\frac{\hat{\theta}(0)^\top \theta}{\|\hat{\theta}(0)\|^2} = \frac{\|\theta\|^2 + \epsilon_n^\top \theta}{\|\theta\|^2 + \|\epsilon_n\|^2 + 2\epsilon_n^\top \theta} \sim \frac{\|\theta\|^2 + \frac{1}{\sqrt{n}}\|\theta\|}{\|\theta\|^2 + \frac{p}{n} + O_P(\frac{\sqrt{p}}{n}) + \frac{1}{\sqrt{n}}\|\theta\|},$$

where the last equality uses the fact that $\|\epsilon_n\|^2 \overset{d}{=} \frac{\chi_p^2}{n} = \frac{p+O_P(\sqrt{p})}{n} = \frac{p}{n} + O_P(\frac{\sqrt{p}}{n})$.

Consider the case when $c_{ratio} = c_1$ and $c_{ratio} = c_2$, where $0 < c_1 < c_2$, we want to prove that,

$$-\frac{e^{-2v^*(c_1)}\cdot 2v^*(c_1)}{(e^{-2v^*(c_1)}+1)^2} > -\frac{e^{-2v^*(c_2)}\cdot 2v^*(c_2)}{(e^{-2v^*(c_2)}+1)^2},$$

where $v^*(c_1) > 0$ and $v^*(c_2) > 0$ are the maximizers of $\frac{1}{e^{-\frac{2\hat{\theta}(0)^\top \theta}{\|\hat{\theta}(0)\|^2}\cdot v}+1} - \frac{1}{e^{-2(1-t)v}+1}$ when $c_{ratio} = c_1$ and $c_{ratio} = c_2$ respectively.

Since

$$-\frac{e^{-2v^*}\cdot 2v^*}{(e^{-2v*}+1)^2}$$

is a decreasing function of v^*, and with high probability

$$\frac{\hat{\theta}(0)^\top \theta}{\|\hat{\theta}(0)\|^2}\Big|_{c_{ratio}=c_1} > \frac{\hat{\theta}(0)^\top \theta}{\|\hat{\theta}(0)\|^2}\Big|_{c_{ratio}=c_2}.$$

Thus, we only need to show that the maximizer $v^*(\rho)$ defined by

$$v^*(\rho) = \text{argmax}_{v>0}\big[\frac{1}{e^{-2\rho\cdot v}+1} - \frac{1}{e^{-2(1-t)v}+1}\big]\Big|_{t\to 0^+}$$

is an decreasing function of ρ for $\rho \in [0,1)$ (since $\frac{\hat{\theta}(0)^\top \theta}{\|\hat{\theta}(0)\|^2} \in [0,1)$).

As we know that $v^*(\rho)$ is the solution of the following equation:

$$\frac{2\rho v^*(\rho)}{e^{2\rho v^*(\rho)} + e^{-2\rho v^*(\rho)} + 2} = \frac{2v^*(\rho)}{e^{2v^*(\rho)} + e^{-2v^*(\rho)} + 2}.$$

From Figure 5, we can directly see that $v^*(\rho)$ increases as ρ decreases. We complete the proof.

□

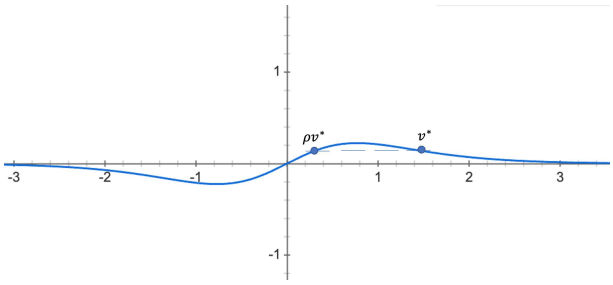


Figure 5. Illustration plot for the function $2x/(e^2 + e^{-2x} + 2)$.

B. Experimental setup and additional numerical results

To complement our theory, we further provide more experimental evidence on popular image classification data sets with neural networks. In Figures 1 and 2, we used fully-connected neural networks and ResNets with various values of the width (*i.e.,* the number of neurons per hidden layer) and the depth (*i.e.*, the number of hidden layers). For the experiments on the effect of the width, we fixed the depth to be 8 and varied the width from 10 to 3000. For the experiments on the effect of the depth, the depth was varied from 1 to 24 (*i.e.*, from 3 to 26 layers including input/output layers) by fixing the width to be 400 with data-augmentation and 80 without data-augmentation. We used the following standard data-augmentation operations using `torchvision.transforms` for both data sets: random crop (via `RandomCrop(32, padding=4)`) and random horizontal flip (via `RandomHorizontalFlip`) for each image. We used the standard data sets — CIFAR-10 and CIFAR-100 (Krizhevsky & Hinton, 2009). We employed SGD with mini-batch size of 64. We set the learning rate to be 0.01 and momentum coefficient to be 0.9. We used the Beta distribution $Beta(\alpha, \alpha)$ with $\alpha = 1.0$ for Mixup.

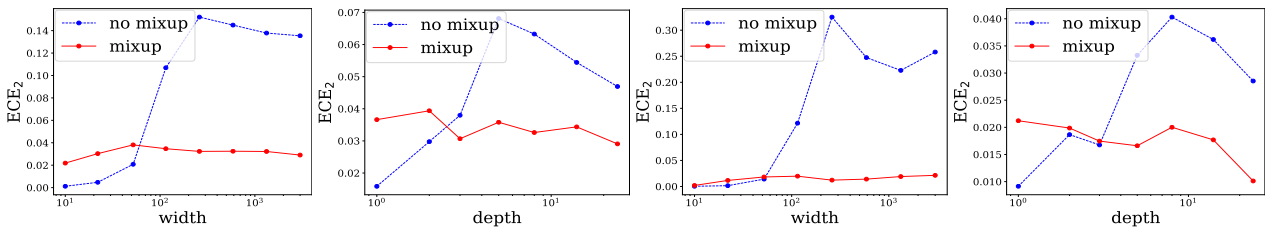
In Figure 3, we adopted the standard data sets, Kuzushiji-MNIST (Clanuwat et al., 2019), Fashion-MNIST (Xiao et al., 2017), and CIFAR-10 and CIFAR-100 (Krizhevsky & Hinton, 2009). We used SGD with mini-batch size of 64 and the learning rate of 0.01. The Beta distribution $Beta(\alpha, \alpha)$ with $\alpha = 1.0$ was used for Mixup. We used the standard pre-activation ResNet with 18 layers and ReLU activations (He et al., 2016b). For each data set, we randomly divided each training data (100%) into a labeled training data (50%) and a unlabeled training data (50%) with the 50-50 split. Following the theoretical analysis, we first trained the ResNet with labeled data until the half of the last epoch in each figure. Then, the pseudo-labels were generated by the ResNet and used for the final half of the training.

We run experiments with a machine with 10-Core 3.30 GHz Intel Core i9-9820X and four NVIDIA RTX 2080 Ti GPUs with 11 GB GPU memory.

Figure 6 shows that Mixup also tend to reduce test loss for larger capacity models. The experimental setting of Figure 6 is the exactly same as that of Figures 1 and 2. Here, relative test loss of a particular case is defined by $\frac{\text{test loss of a particular case}}{\text{test loss of no mixup base case}}$.

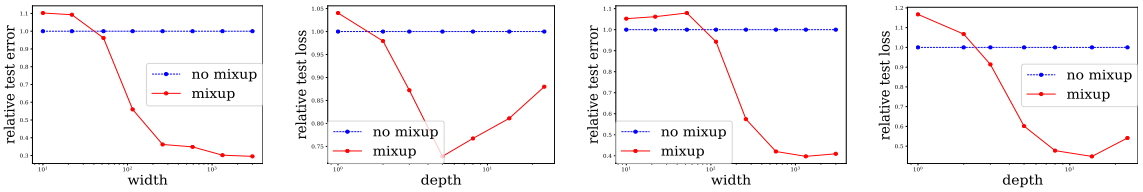
Figures 7–8 show that Mixup can reduce ECE_2 particularly for larger capacity models.

Figure 4 uses the same setting as that of Figures 1 and 2. Similarly to Figure 4, Figure 9 below shows that Mixup can reduce MCE particularly for larger capacity models with varying degrees of depth and width. The setting of Figure 9 is the same as that of Figures 1 and 2 with the data-augmentation.

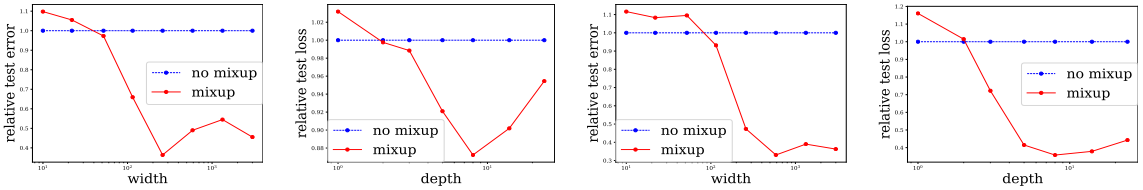


(a) CIFAR-10: width (b) CIFAR-10: depth (c) CIFAR-100: width (d) CIFAR-100: depth

Figure 8. ECE_2 without data-augmentation

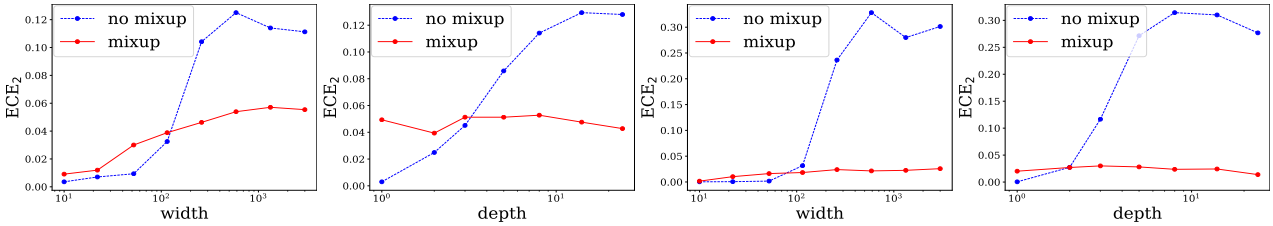


(a) Width (b) Depth (c) Width (d) Depth



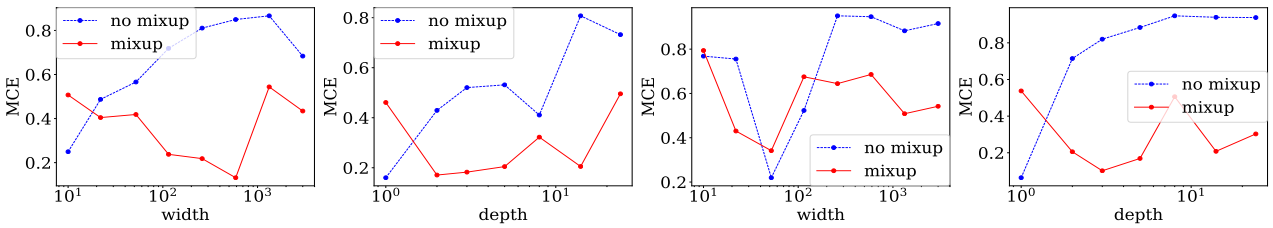
(e) Width (f) Depth (g) Width (h) Depth

Figure 6. Relative test loss: (a), (b): CIFAR-10 without data augmentation; (c), (d): CIFAR-10 with data augmentation; (e), (f): CIFAR-100 without data augmentation; (g), (h): CIFAR-100 with data augmentation.



(a) CIFAR-10: width (b) CIFAR-10: depth (c) CIFAR-100: width (d) CIFAR-100: depth

Figure 7. ECE_2 with data-augmentation



(a) CIFAR-10: width (b) CIFAR-10: depth (c) CIFAR-100: width (d) CIFAR-100: depth

Figure 9. Maximum Calibration Error (MCE)