Parametric ρ-Norm Scaling Calibration

Siyuan Zhang, Linbo Xie[*]

School of Internet of Things Engineering, Jiangnan University
siyuanz1996@gmail.com, xie_linbo@jiangnan.edu.cn

Abstract

Output uncertainty indicates whether the probabilistic properties reflect objective characteristics of the model output. Unlike most loss functions and metrics in machine learning, uncertainty pertains to individual samples, but validating it on individual samples is unfeasible. When validated collectively, it cannot fully represent individual sample properties, posing a challenge in calibrating model confidence in a limited data set. Hence, it is crucial to consider confidence calibration characteristics. To counter the adverse effects of the gradual amplification of the classifier output amplitude in supervised learning, we introduce a post-processing parametric calibration method, ρ-Norm Scaling, which expands the calibrator expression and mitigates overconfidence due to excessive amplitude while preserving accuracy. Moreover, bin-level objective-based calibrator optimization often results in the loss of significant instance-level information. Therefore, we include probability distribution regularization, which incorporates specific priori information that the instance-level uncertainty distribution after calibration should resemble the distribution before calibration. Experimental results demonstrate the substantial enhancement in the post-processing calibrator for uncertainty calibration with our proposed method.

Introduction

Model confidence calibration involves refining uncertainty estimates of the model outputs, enabling more accurate probability predictions that align closely with the objective characteristics of the output uncertainty. With the progressive expansion of model capacity, the modern models often demonstrate inadequately probability distributions. Specifically, these probability outputs display unwarranted over-confidence in comparison to the objective accuracy (Guo et al. 2017). Furthermore, researchers have identified that achieving high accuracy in classifiers and calibrating the model confidence in baseline are distinct objectives (Wenger, Kjellström, and Triebel 2020). This scenario emphasizes the pressing necessity to rectify the calibration of model output uncertainties in deep learning.

As one of the effective calibration methods, post-calibration methods have recently gained popularity, which operate independently of the model internal optimization, reconstructing and optimizing the output-probability mapping (Zadrozny and Elkan 2001). In contrast to other calibration techniques, post-processing calibration methods do not necessitate altering the original baseline, thereby preserving the model generalization ability for classification (Rahimi et al. 2020).

Among post-processing calibration methods, parametric techniques like Platt Scaling (Platt et al. 1999), Temperature Scaling (Tomani, Cremers, and Buettner 2022), and Beta Calibration (Kull, Silva Filho, and Flach 2017) necessitate parameter optimization using a validation set. The commonly used metric for assessing model confidence, known as Expected Calibration Error (ECE), computing the expected difference between confidence and accuracy within each bin (Naeini, Cooper, and Hauskrecht 2015). Confidence estimation captures individual sample characteristics, yet it cannot be validated on a per-sample basis. Meanwhile, validating bin-level uncertainty fails to capture the nuances of individual samples. This challenge distinguishes the assessment of model uncertainty and the design and optimization of parametric output-probability mapping. Specifically, the bin-level loss function, such as ECE, is more prone to converge to zero during optimization compared to an instance-level loss function like cross-entropy (Krishnan and Tickoo 2020). This characteristic makes calibrator optimization susceptible to overfitting, impeding generalization. To ensure calibrators learn output-probability mappings effectively, it is essential to incorporate specific priori knowledge to limit the optimization hypothesis space and to design instructive mapping and multi-level optimization objectives.

Inspired by the aforementioned research and questions, we meticulously considered the internal order-preserving property (Rahimi et al. 2020) to conserve the inherent uncertainty distribution. Additionally, we assessed the impact of amplitude on the classifier output, supported by prior evidence suggesting that excessive output amplitude might lead to unwarranted calibration error. Building upon this insight, we proposed parametric ρ-Norm Scaling calibration, addressing the expressivity limitation in TS (Tomani, Cremers, and Buettner 2022) and mitigating the negative effects by the output amplitudes. Besides, concerning the parameter optimization of the output-probability mapping, we proposed a multi-level loss by introducing a probability distri-

[*]Corresponding author.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

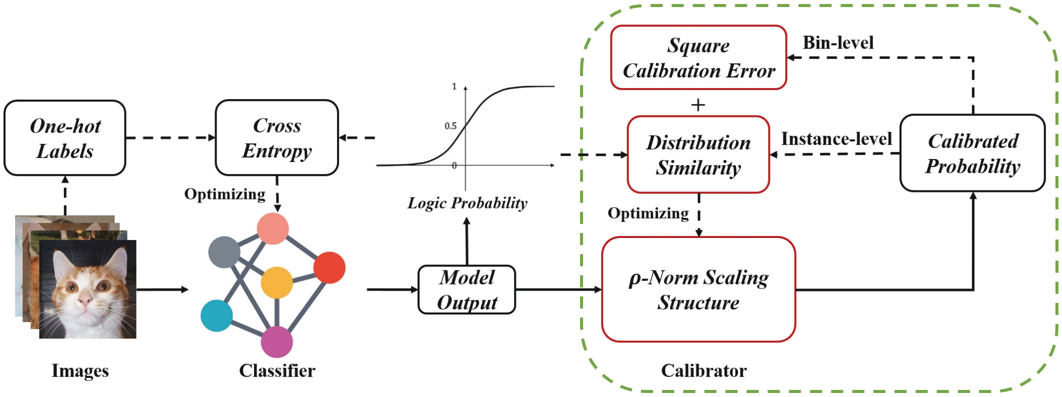


Figure 1: **Overview of our proposed post-hoc calibrator structure and optimization objective after pipeline of classifier optimization**: (1) Addressing the issue of output magnitude amplification during supervised learning, we introduce a ρ-Norm Scaling calibration within the post-calibration framework. (2) Uncertainty represents the entire dataset statistically, making its optimization prone to losing sample-level information. To address this, we incorporate probabilistic similarity between pre-calibration and post-calibration as a instance-level loss, combined with bin-level loss.

bution similarity regularization into Square Calibration Error. This regularization enhances the correlation between the original probability distribution and the calibrated distribution, aiming to prevent significant deviations from the original probability distribution and retain key properties of the original distribution.

Our main contributions in this work can be summarized as follows: (1) We propose a new family of parametric ρ-Norm Scaling calibration model for post-hoc calibration and the corresponding optimization strategy. (2) We provide a new multi-level objective for post-hoc parameter optimization by adding an instance-level regularization between original probability distribution and calibrated probability distribution into bin-level Square Calibration Error. (3) We perform extensive evaluations on multiple datasets and models, and our proposed method achieves state-of-the-art calibration performance.

Related Work

Strategies aimed at calibrating model uncertainty can be classified into the following approaches: Bayesian neural networks, training-based calibration, and post-processing calibration. The ability of neural network to quantify prediction uncertainty is limited, prompting the consideration of replacing a section of the model structure with a Bayesian inference process (Milios et al. 2018; Wen, Tran, and Ba 2019). Bayesian neural networks offer several advantages, including ease of implementation, parallelization feasibility, minimized hyperparameter adjustments, and the ability to offer precise estimates of predictive uncertainty (Gal and Ghahramani 2016; Calandra et al. 2016; Naeini, Cooper, and Hauskrecht 2015; Wenger, Kjellström, and Triebel 2020; Tran et al. 2019). Besides, training-based calibration methods have been explored to mitigate miscalibration risks in supervised learning. These approaches may involve techniques such as pre-training (Hendrycks, Lee, and Mazeika 2019), data augmentation (Thulasidasan et al. 2019), label smoothing (Menon et al. 2020), weight decay (Guo et al. 2017), and more. Furthermore, Tao investigated the limitations of early stopping and devised solutions to overfitting within specific network blocks concerning calibration metrics (Tao et al. 2023). The connection between model structure sparsity and model calibration was examined in (Lei et al. 2022). Additionally, researchers have proposed some innovative loss functions, such as MMCE (Kumar, Sarawagi, and Jain 2018), Correctness Ranking Loss (Moon et al. 2020), CALS (Liu et al. 2023), Focal loss (Lin et al. 2017; Tao, Dong, and Xu 2023; Mukhoti et al. 2020), and FLSD (Ghosh, Schaaf, and Gormley 2022), which simultaneously consider classification accuracy and confidence calibration. However, an excessive focus on model calibration during training might detrimentally affect overall model accuracy improvement. Furthermore, the calibration during training may compromise the efficacy of post-processing calibration methods (Wang, Feng, and Zhang 2021).

Post-processing calibration refers to reconstructing the output-probability mapping. One of its key advantages lies in decoupling classifier accuracy from calibration, thereby maintaining the original generalization without necessitating alterations to its training strategy. In the era preceding the ascendancy of deep learning, non-parametric post-processing calibration methods such as Histogram Binning (HB) (Zadrozny and Elkan 2001), Isotonic Regression (IR) (Zadrozny and Elkan 2002) and Bayesian processes (Gal and Ghahramani 2016) were prevalent. Unlike non-parametric methods, which calibrate a model confidence distribution using nonlinear logic, parametric calibration methods focus on establishing a parametric structure by learning from finite samples. Some commonly utilized parametric calibration structures include Platt Scaling (Platt et al. 1999), Temperature Scaling (Yu et al. 2022; Tomani, Cremers, and Buettner 2022), Beta Calibration (Kull, Silva Filho, and Flach 2017). Typically, parameters are learned through grid search or gradient-based optimization of Negative Log-

Likelihood (NLL) (Hastie et al. 2009). However, direct optimization of NLL often may compel the model output towards one-hot distribution, deviating from the intended calibration logic. To address this problem, Krishnan introduced the AvUC loss function (Krishnan and Tickoo 2020). Subsequently, Karandikar proposed soft calibration objective for optimizing the calibrator parameters (Karandikar et al. 2021). Additionally, in terms of mapping structure, Kull extended Beta Calibration and introduced Dirichlet calibration (Kull et al. 2019). Considering the flexibility of calibration mapping, Wang introduced Shape-Restricted Polynomial Regression as a parametric calibration method (Wang, Li, and Dang 2019). Furthermore, some studies propose a class of accuracy-preserving mappings (Tomani, Cremers, and Buettner 2022; Rahimi et al. 2020).

Methodology

Problem Formulation

Considering a dataset $\left\{(x^i, y^i)\right\}_{i=1}^{N} \subset \mathbf{R}^n \times \mathbf{R}^m$ and classifier f maps x to the outputs $z_j, j = 1, \ldots, m$ on m classes and $k = \arg\max_j z_j$. The ground-truth y and predicted labels $\hat{y}$ are formulated in one-hot format where $y_c = 1$ and $\hat{y}_k = 1$, where c represents the truth class. The confidence score of the predicted label in baseline is $\hat{p} = \max s_j(z), j = 1, \ldots, m$, where $s(\cdot)$ represents Softmax mapping $R^m \to R^m$. However, Softmax mapping probabilities are not accurately reflected in the properties of model output (Guo et al. 2017). To address this, the calibrator $g(\cdot)$ is introduced as a new output-probability mapping $g : z \to p$ for probability calibration.

Confidence Calibration: Perfect calibration of neural network can be realized when the confidence score reflects the real probability that the sample is classified correctly. Formally, the perfectly calibrated network satisfied $\mathrm{P}(\hat{y} = y | \hat{p} = p) = p$ for all $p \in [0, 1]$. However, in practical applications, the sample is divided into M bins $\{D_b\}_{b=1}^{M}$. The limited availability of data restricts our ability to accurately estimate the calibration error. According to their confidence scores and the calibration error, an approximation is calculated for each bins $\{D_b\}_{b=1}^{M}$. D_b contains all sample with $\hat{p} \in \left[\frac{b}{M}, \frac{b+1}{M}\right)$. Average confidence is computed as $conf(D_b) = \frac{1}{|D_b|}\sum_{i \in D_b} \hat{p}^i$ and the bin accuracy is computed as $acc(D_b) = \frac{1}{|D_b|}\sum_{i \in D_b} \mathrm{I}\left(y_c^i = \hat{y}_c^i\right)$. ECE (Naeini, Cooper, and Hauskrecht 2015) is calculated as follows.

$$ECE = \sum_{b=1}^{M} \frac{|D_b|}{N} \left|acc(D_b) - conf(D_b)\right| \tag{1}$$

To develop an effective output-probability mapping, we design the calibrator based on the relationship between output amplitude and confidence level. Additionally, we formulate the multi-level objective for calibrators parameter based on the characteristics of calibration error.

Output Magnitude: Our postulation suggests that overconfidence of modern deep model arises from the utilization of Softmax cross entropy optimization, particularly in high-capacity models, leading to an amplification of the output amplitude, as shown in Fig. 2, where the amplitude of a single sample's output is defined as $\left\|z^i\right\|_2$. The unsaturated regions of the original Softmax are present exclusively when the differences between the category outputs are relatively small. Consequently, a significant portion of samples tends to have probabilities that fall within the saturation region of Softmax, resulting in probability outputs close to 1. The similar conclusion can be found in (Zhang and Xiang 2023; Wei et al. 2022), where it is suggested that the magnitude of neural network output can be a culprit.

Calibration Characteristic: The optimization of the calibration error within the post-calibration structure appears to be challenging due to hard binning operation, as discussed in literature (Karandikar et al. 2021). We believe that another practical obstacle arises from the nature of the uncertainty estimation, which serves as collective binning properties for multiple samples $\sum_{i \in D_b} \hat{p}^i = \sum_{i \in D_b} \mathrm{I}\left(y_c^i = \hat{y}_c^i\right)$, rather than providing individual metric for each sample (Si et al. 2022; Widmann, Lindsten, and Zachariah 2019). Calibration error as loss metric disregards numerous sample-level output-probability mapping relationships. Consequently, the calibrated distribution significantly diverges from the original distribution, as shown in Fig. 4.

Parametric ρ-Norm Scaling

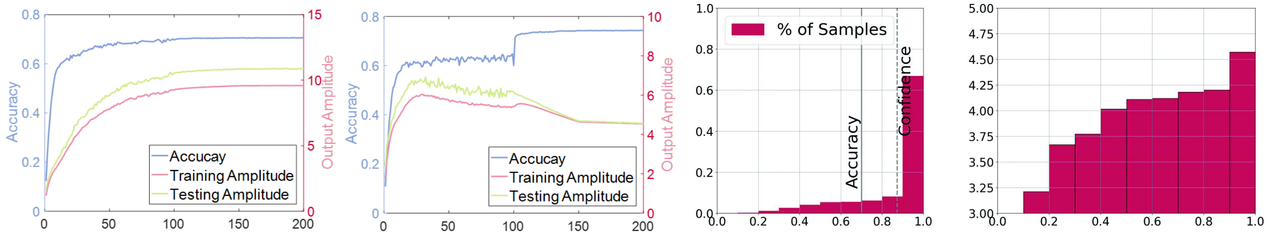
To regulate the influence of output amplitude on the scaling calibration and propose accurate calibrator, we enhance the expressive power by incorporating a parameterized ρ-norm normalization term into the output-probability mapping. The adopted calibration structure is represented below:

$$g_c(z) = \frac{e^{r_c}}{\sum_{j=1}^{m} e^{r_j}} \tag{2}$$

where $r_j(z) = \frac{z_j}{\gamma\|z\|_\rho + \beta}$, $\gamma > 0$ and $\beta > 0$. $\|\cdot\|_\rho$ represents ρ-norm, where $\|z\|_\rho = \left({z_1}^\rho + {z_2}^\rho + \cdots + {z_m}^\rho\right)^{1/\rho}$. ρ is defined as a learnable parameter in the algorithm that is used to control a learnable norm space to regulate the large output magnitude. In supervised learning, classifiers often produce outputs with substantial magnitudes, particularly for overconfident samples. When these outputs have excessively large magnitudes and are fed into a calibrator, they often fall within the saturation interval where the Softmax output converges to 1. Consequently, calibrating samples with such high-amplitude outputs becomes insensitive.

Proposition 1 *For any model output z and the probability by mapping of $g_c = \frac{e^{r_c}}{\sum_{j=1}^{m} e^{r_j}}$ where $r_j(z) = \frac{z_j}{\gamma\|z\|_\rho}$, the following inequalities holds.*

$$\begin{aligned} &\frac{1}{(m-1)\, e^{\frac{1}{\gamma}\left(\frac{1}{(m-1)^{1/\rho-1}}+1\right)^{\rho-1/\rho}} + 1} \leq g \\ &\leq \frac{1}{(m-1)\, e^{-\frac{1}{\gamma}\left(\frac{1}{(m-1)^{1/\rho-1}}+1\right)^{\rho-1/\rho}} + 1} \end{aligned} \tag{3}$$



(a) Amplitude without weight decay (b) Amplitude with weight decay (c) Confidence histogram (d) Amplitude histogram

Figure 2: **Amplitude changes in classifier optimization.** In these figures, the overall output magnitude of all samples is defined as $\frac{1}{Nm}\sum_{i=1}^{N}\left\|z^i\right\|_2$. During the supervised learning of the classifier, the output magnitude follows a specific pattern. (a) illustrates that in the absence of weight decay, the output amplitude steadily increases throughout the optimization process. Although this trend is alleviated in the presence of weight decay, as depicted in (b), the final magnitudes exhibit a positive correlation with the overall confidence distribution, shown in (c) and (d).

Proposition 2 *For any $\gamma > 0$ and $\beta > 0$ of ρ-Norm Scaling $g\left(z\right)$, the classification accuracy based one-versus-all classification keeps unchanged after the output-to-probability mapping.*

The results in Proposition 1 show that ρ-norm is able to restrict the confidence interval to prevent the confidence from a one-hot distribution, yielding a smoother confidence distribution. When $\gamma \rightarrow +\infty$, the calibrated probabilities g tend to $1/m$. When $\gamma \rightarrow +0$, all probabilities are adjusted to $[0, 1]$. Furthermore, Proposition 2 establishes that the ρ-Norm Scaling structure maintains decision invariance, ensuring that the pre-calibration classification results match the calibrated probability one-versus-all classification outcomes. This property serves as specific priori knowledge, preserving dataset distributional properties during optimization, aligning with calibration logic. Decision invariance is a crucial consideration in calibration structure design (Tomani, Cremers, and Buettner 2022; Rahimi et al. 2020).

Proposition 3 *For any output-to-probability mapping $g_j\left(z\right) = \frac{e^{z_j\sigma(z)}}{\sum_{j=1}^{m} e^{z_j\sigma(z)}}$, if the function $\sigma\left(z\right) > 0$ holds for any z, accuracy of model based one-versus-all classification decision-making keeps unchanged after the output-to-probability mapping.*

Proposition 3 extends the conclusion in Proposition 2 by presenting a general mapping-based criterion for satisfying decision invariance, for mapping design. When the function $\sigma\left(z\right)$ is defined as a single hyperparameter, this mapping degenerates into naive Softmax mapping with a temperature coefficient. Apart from calibrators, classifier optimization with different $\sigma\left(z\right)$ deserves further exploration.

Parameter Optimization

In this subsection, we address the challenge of optimizing the calibrator parameters from the desired calibration error. From a logical perspective, optimizing the calibrator with NLL does not differ from the optimization goal of the original classifier, Softmax cross entropy. However, it is essential to note that achieving high accuracy in classifiers using cross entropy as the objective and calibrating the model confidence represent distinct objectives (Wang, Feng, and Zhang 2021). There remains a bias in minimizing NLL compared to minimizing calibration error.

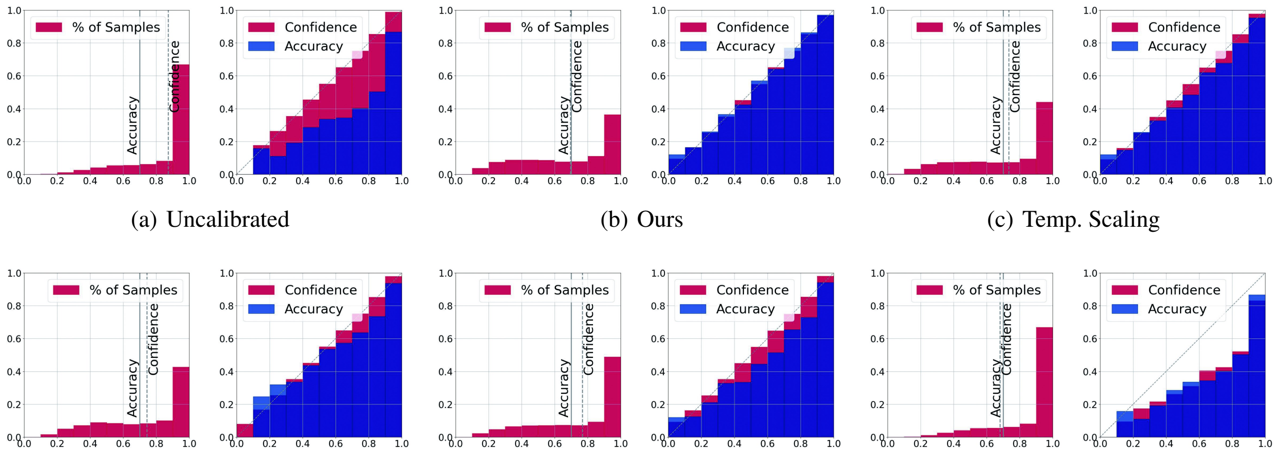
Algorithm 1: ρ-Norm Scaling Post-hoc Calibrator

```
Data: Validation set $\{\,(x_i, y_i)|\, i = 1, \ldots, N\}$;
      Classifier $f$; Learning rate $\lambda$; Batch size $N_B$.
Result: $\rho$-Norm Scaling calibrator $g\left(z, \rho^*, \theta^*\right)$
Initialize $\theta$, $\theta^*$, $\rho^*$, $ECE^*$;
$z \leftarrow f\left(x\right)$;
while $\rho \in \{1, \ldots, 3\}$ do
    while $t < T_{max}$ do
        $D_{N_B} \leftarrow \{z^i\}_i^{N_B}$ ;
        $l_{SCE} \leftarrow$ Computing by Eq.(4);
        $l_{KL} \leftarrow$ Computing by Eq.(5);
        $\theta \leftarrow \theta - \lambda \frac{\partial l}{\partial \theta}$;
    end
    $ECE_{val} \leftarrow ECE$ on validation set by Eq.(1);
    if $ECE_{val} < ECE^*$ then
        $\rho^* \leftarrow \rho$;
        $\theta^* \leftarrow \theta$;
        $ECE^* \leftarrow ECE_{val}$;
    end
end
```

A straightforward approach is to utilize the calibration error as a loss function for optimizing the parametric calibration mapping. With calibration error as loss function, we treat the entire batch as a bin D in each iteration of the optimization, randomly sampling a bin from the validation set. This approach helps alleviate the issue where finite data fail to fully reflect overall uncertainty. Modified SCE (Square Calibration Error) is shown as follows.

$$l_{SCE} = \left(acc\left(D\right) - conf\left(D\right)\right)^2 \tag{4}$$

where $conf\left(D\right) = \frac{1}{|D|}\sum_{z^i \in D} \kappa \log \sum_{j}^{m} e^{g_j(z_i)/\kappa}$, $g_j =$



(a) Uncalibrated (b) Ours (c) Temp. Scaling

(d) Vector Scaling (e) TS-AvUC (f) Hist. Bin.

Figure 3: **Confidence histograms and reliability diagrams for different post-hoc calibration methods with ResNet35 on CIFAR-100.** Confidence histograms display the sample count within each bin, whereas reliability diagrams illustrate the difference between the average confidence (marked in red) and the accuracy (indicated in blue) in each bin.

$\frac{e^{r_j}}{\sum_{j=1}^{m} e^{r_j}}$ and $r_j\left(z\right) = \frac{z_j}{\gamma^2\|z\|_\rho+\beta^2}$. The function for max confidence in each class is smoothed by sum-exp-up, and the coefficient κ is set relatively small, such as 10^{-4}, since the probability outputs are less than 1. The small κ ensures the confidence closely approximates the max predictive probability. Furthermore, to satisfy Proposition 2 in optimization, we replace β and γ in the original ρ-Norm Scaling with β^2 and γ^2 to the constraint that the hyperparameter is greater than zero, respectively.

However, the model uncertainty typically represents a statistic of the dataset, lacking the ability to adequately characterize individual samples. When directly using the calibration error as the optimization objective $\sum_{i\in D}\hat{p}^i = \sum_{i\in D}\mathrm{I}\left(y_c^i = \hat{y}_c^i\right)$, a substantial amount of sample-level information is lost in accurately characterizing the output-probability mapping $p^i \leftarrow z^i$, which impedes finding a solution with strong generalization ability. So, we introduce instance-level KL divergence for the probability distribution between pre-calibration $s\left(z\right)$ and post-calibration $g\left(z\right)$, serving as a regularization technique to maintain the probability distribution similarity between probability distributions before and after calibration. This addition term ensures that the calibrated distribution retains specific distributional characteristics of the original one. It can provide important instance-level information for optimization.

$$l_{KL} = \sum_{z^i\in D}\sum_{j=1}^{m} g_j(z^i)\left(\log g_j\left(z^i\right) - \log s_j\left(z^i\right)\right) \qquad (5)$$

The final multi-level objective is represented below.

$$l = l_{SCE} + \alpha l_{KL} \qquad (6)$$

Relying solely on the bin-level SCE often leads to significant deviations between the pre- and post-calibration probability distributions due to its statistical nature, particularly when calibration structure has strong expressive capacity, as shown in Fig. 4. To mitigate this, we introduce the instance-level KL divergence of the original output distribution $s\left(z\right)$ concerning the calibrated probability distribution $g\left(z\right)$ as a regularization. When KL divergence converges to 0, $g\left(z\right)$ equals $s\left(z\right)$. Therefore, $s\left(z\right)$ replaces the one-hot label in NLL, similar to label smoothing.

In parameter optimization, the proposed algorithm employs a two-stage optimization strategy. In the outer loop, the algorithm conducts a grid search within the intervals $\{1, 1.25, \ldots, 3\}$ to determine ρ, using ECE as the metric. Meanwhile, in the inner loop, parameters γ and β are optimized by the small batch gradient-based methods with the proposed loss (6).

Experiments

We evaluate our methods on multiple DNNs, including ResNet and VGG series. Our experiments are conducted on SVHN, CIFAR-10/100, 102 Flower, and Tiny-ImageNet for post-hoc calibration performance. Different ablation experiments are designed to evaluate efficiency of the ρ-Norm Scaling calibration structure and the multi-level objective. In tables, the best results and relative improvements over 2^{nd} best result in each section are in bold. Results are averaged over five runs with different seeds.

Baselines: In experiments, we compare our methods with different calibration methods, such as non-parametric Hist. Binning, TS, Vector Scaling (Niculescu-Mizil and Caruana 2005). Above all parametric structure are optimized by gradient descent based on NLL. TS-AvUC represents Tem. Scaling with NLL-AvUC as objective (Krishnan and Tickoo 2020). The ρ in ρ-Norm Scaling is selected by grid search on $\{1, 1.25, \ldots, 3\}$ and other parameter are optimized based on Eq.(6). In ablation experiments, different structures are optimized by different optimization objectives,

Dataset	Model	Metric	Uncalibrated	Hist. Binning	TS	Vector Scaling	TS-AvUC	Ours
CIFAR-100	ResNet18	ECE	0.160±0.025	0.025±0.006	0.033±0.006	0.061±0.012	0.028±0.004	**0.009(↓ *0.016*)**
		MCE	0.344±0.055	0.078±0.012	0.059 ±0.011	0.138 ±0.022	**0.052(↓ *0.007*)**	0.098±0.021
		AdaECE	0.160±0.023	-	0.030±0.007	0.061±0.011	0.027±0.006	**0.007(↓ *0.020*)**
CIFAR-100	ResNet50	ECE	0.186±0.031	0.025±0.004	0.030±0.013	0.073±0.021	0.052±0.012	**0.007(↓ *0.018*)**
		MCE	0.407±0.101	0.110±0.015	**0.091(↓ *0.009*)**	0.153±0.036	0.116±0.021	0.100±0.023
		AdaECE	0.186±0.029	-	0.029±0.012	0.071±0.028	0.052±0.010	**0.006(↓ *0.023*)**
CIFAR-100	VGG16	ECE	0.240±0.106	0.035±0.002	0.029±0.003	0.035±0.006	0.044±0.008	**0.019(↓ *0.010*)**
		MCE	0.508±0.151	**0.042(↓ *0.001*)**	0.093±0.029	0.084±0.009	0.101±0.026	0.043±0.003
		AdaECE	0.240±0.106	-	0.029±0.004	0.035±0.006	0.044±0.008	**0.019(↓ *0.010*)**
CIFAR-10	ResNet35	ECE	0.054±0.010	0.011±0.001	0.015±0.002	0.014±0.003	0.015±0.006	**0.007(↓ *0.004*)**
		MCE	0.300±0.085	0.255±0.102	0.121±0.026	**0.077(↓ *0.030*)**	0.121±0.021	0.107±0.019
		AdaECE	0.054±0.011	-	0.014±0.004	0.013±0.002	0.013±0.005	**0.010(↓ *0.003*)**
SVHN	ResNet18	ECE	0.021±0.006	0.016±0.002	0.009±0.003	**0.007(↓ *0.001*)**	0.010±0.003	0.008↑±0.002
		MCE	0.286±0.053	**0.251(↓ *0.035*)**	0.313±0.052	0.313±0.069	0.315±0.080	0.438±0.103
		AdaECE	0.021±0.006	-	0.010±0.003	0.009±0.002	0.013±0.005	**0.008(↓ *0.001*)**
102 Flower	ResNet50	ECE	0.101±0.018	0.084±0.012	0.086±0.011	0.093±0.015	0.075±0.009	**0.046(↓ *0.029*)**
		MCE	0.231±0.048	0.365±0.066	0.180±0.041	0.163±0.043	0.165±0.044	**0.152(↓ *0.011*)**
		AdaECE	0.100±0.017	-	0.089±0.012	0.098±0.019	0.079±0.011	**0.048(↓ *0.031*)**
Tiny-ImageNet	ResNet35	ECE	0.144±0.022	0.033±0.005	0.017±0.003	0.053±0.007	0.017±0.003	**0.007(↓ *0.010*)**
		MCE	0.236±0.052	0.055±0.016	0.035±0.010	0.093±0.021	**0.030(↓ *0.001*)**	0.031±0.004
		AdaECE	0.143±0.021	-	0.017±0.004	0.054±0.008	0.016±0.002	**0.006(↓ *0.010*)**

Table 1: The calibration performance of different post-hoc calibration methods.

Model	Metrics	NLL		Ours	
		Temp. Scaling	ρ-Norm Scaling	Temp. Scaling	ρ-Norm Scaling
ResNet35	ECE	0.026	0.011(↓ ***0.015***)	0.024	0.009(↓ ***0.015***)
	AdaECE	0.027	0.011(↓ ***0.016***)	0.024	0.007(↓ ***0.017***)
ResNet50	ECE	0.048	0.006(↓ ***0.042***)	0.042	0.007(↓ ***0.035***)
	AdaECE	0.048	0.008(↓ ***0.040***)	0.042	0.006(↓ ***0.036***)

Table 2: The ablation study of calibrator structure and optimization objective on CIFAR-100.

such as NLL, SCE, AvUC and SB-ECE (Karandikar et al. 2021). In Hist. Binning, ECE, MCE and AdaECE (Nguyen and O'Connor 2015), the number of bins is 10. In all experiments for CIFAR-10/100, the learning rate was set to 0.1, the momentum to 0.9, the weight clipping to Norm=3, and the batch size to 128. The learning rate decreased to 10% at 40% and 80% of the iterations. The weight decay was set to 10^{-4} and the iteration number was 200. For the Tiny-ImageNet, the learning rate was set to 0.01 and batch size was 64. The hyperparameter α is set to 1.

The Efficiency of Our Method: Table 1 presents the outcomes of various calibration techniques. Our method substantially enhances the performance of post-calibration in both ECE and AdaECE, outperforming classical methods. However, ρ-Norm Scaling does not yield superior results in MCE. The confidence histograms and reliability diagrams depicted in Fig. 3 reveal a bias in bin $[0, 0.1]$ with a minimal number of samples, which does not significantly impact the overall calibration outcomes in ECE and AdaECE. TS, utilizing a single hyperparameter to control the smoothing of uncertainty distribution, achieves superior results compared to post-processing calibrations employing multiple hyperparameters like Vector Scaling. Furthermore, both TS and our proposed method demonstrate accuracy-preserving property, preserving the order of probability outputs across different categories for individual samples (Rahimi et al. 2020). The notably smaller calibration errors of these mapping, in contrast to Vector Scaling, underscore the pivotal role of maintaining decision invariance as a fundamental prerequisite in calibrator design.

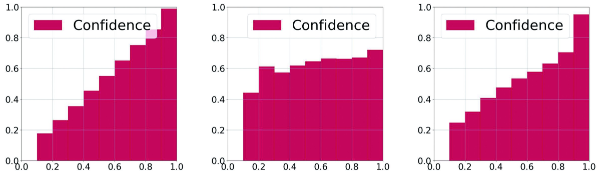
The Ablation Study of Calibrator Structure: To mitigate the influence of the loss function on the experimental outcomes, we conducted ablation experiments and presented the results in Table 2. Notably, in Table 2, ρ-Norm Scaling continues to enhance calibration performance when compared to TS under the same optimization objective.

Metrics	SCE	KL	SCE+KL	NLL	NLL+KL	NLL-AvUC	SB-ECE	SB-ECE+KL	Uncalibrated
ECE	0.173	0.161	0.041(↓ ***0.120***)	0.056	0.039(↓ ***0.008***)	0.047	0.156	0.039 (↓ ***0.117***)	0.172
AdaECE	0.172	0.161	0.043(↓ ***0.118***)	0.053	0.040(↓ ***0.005***)	0.045	0.157	0.036 (↓ ***0.121***)	0.172
KL	0.109	0.001	0.003	0.006	0.004	0.002	0.097	0.002	-

Table 3: The ablation study with Vector Scaling on CIFAR-100.

Dataset	Metrics	Different ρ in ρ-Norm Scaling					Uncalibrated	Temp. Scaling
		1.5	1.75	2	2.25	2.5		
CIFAR-100	ECE	0.028	**0.006**(↓ ***0.003***)	0.010	0.009	0.010	0.172	0.026
	AdaECE	0.028	0.008	0.009	**0.007**(↓ ***0.001***)	0.009	0.172	0.027
Tiny-ImageNet	ECE	0.052	**0.007**(↓ ***0.011***)	0.018	0.040	0.044	0.144	0.017
	AdaECE	0.051	**0.006**(↓ ***0.012***)	0.018	0.041	0.044	0.143	0.017

Table 4: The calibration performance on ResNet35 of different norm in ρ-Norm Scaling.



(a) Uncalibrated (b) SCE (c) SCE+KL

Figure 4: **Coincidence distribution of different optimization objective in Vector Scaling.** In (a), samples are categorized into bins based on confidence levels through Softmax. Each sample in (b) and (c) belongs to the same bin as in (a). Using sample-level SCE alone in post-calibration results in a significant deviation from the original distribution. This challenge is mitigated by the bin-level KL.

This observation suggests that the supervised optimization using Softmax cross entropy as the objective leads to a larger amplitude in model output, negatively impacting calibration. Furthermore, it fails to ensure both classifier accuracy and uncertainty estimates derived from Softmax mapping. Consequently, ρ-Norm Scaling can realize better calibration performance than TS.

The Ablation Study of Optimization Objective: The KL divergence regularity introduces instance-level information into calibrator optimization, ensuring that the calibrated probability retains certain properties of the uncalibrated distribution, as emphasized by the results in Table 3. The bin-level calibration error as a statistical representation of collective binning properties, compressing sample-level information significantly. Furthermore, Vector Scaling offers a broader assumption space and greater expressive power in comparison to Temp. Scaling. When SCE or SB-ECE is used solely as the loss function, the calibrated results deviate markedly from the original results as shown in Fig. 4 and the KL divergence remains relatively large. In addition, using KL divergence alone does not yield improved results. However, better outcomes are achieved when bin-level loss and instance-level loss jointly optimize parameters. KL divergence, acting as a regularity term, guides the calibrated model to learn distributions mirroring the properties of the original distributions, while SCE fine-tunes the mapping parameters and refines the calibrator model.

The Ablation Study of Different Norms: Table 4 displays calibration results across various norms. The data illustrates that smoother outcomes are achieved when ρ is close to 2, though ρ does not guarantee optimality. Different learning paradigms enable the exploration of diverse spaces, facilitating the acquisition of more precise output-probability mappings. This reflects the significance of searching for the appropriate norm, rather than directly replacing it with the conventional Euclidean norm.

Conclusion

The desired calibration metric, based on sample set statistics, captures the dataset general characteristics but overlooks sample-level nuances. Relying solely on the calibration error as the optimization objective fails to yield a well-generalized output-probability mapping within a broad assumption space. Consequently, integrating specific priori knowledge becomes imperative when designing and optimizing the post-hoc calibrator. In this paper, we introduce a ρ-Norm Scaling to mitigate the adverse impact of amplified output amplitude in supervised learning while preserving accuracy. Simultaneously, an instance-level probability distribution regularization is proposed in the optimization, which incorporates specific priori knowledge and emphasizes the need for the uncertainty distribution after calibration to keep some characteristics of the pre-calibration distribution. The experimental results show the significant enhancement in uncertainty calibration performance through ρ-Norm Scaling and multi-level objective. They also underscore the necessity for precise calibrator design to guide the model effectively in learning an ideal calibration mapping.

References

Calandra, R.; Peters, J.; Rasmussen, C. E.; and Deisenroth, M. P. 2016. Manifold Gaussian processes for regression. In *International joint conference on neural networks (IJCNN)*, 3338–3345. IEEE.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, 1050–1059. PMLR.

Ghosh, A.; Schaaf, T.; and Gormley, M. 2022. AdaFocal: Calibration-aware adaptive focal loss. In *Advances in Neural Information Processing Systems*, volume 35, 1583–1595.

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

Hastie, T.; Tibshirani, R.; Friedman, J. H.; and Friedman, J. H. 2009. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Hendrycks, D.; Lee, K.; and Mazeika, M. 2019. Using pre-training can improve model robustness and uncertainty. In *International Conference on Machine Learning*, 2712–2721. PMLR.

Karandikar, A.; Cain, N.; Tran, D.; Lakshminarayanan, B.; Shlens, J.; Mozer, M. C.; and Roelofs, B. 2021. Soft calibration objectives for neural networks. In *Advances in Neural Information Processing Systems*, volume 34, 29768–29779.

Krishnan, R.; and Tickoo, O. 2020. Improving model calibration with accuracy versus uncertainty optimization. In *Advances in Neural Information Processing Systems*, volume 33, 18237–18248.

Kull, M.; Perello Nieto, M.; Kängsepp, M.; Silva Filho, T.; Song, H.; and Flach, P. 2019. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. In *Advances in Neural Information Processing Systems*, volume 32.

Kull, M.; Silva Filho, T.; and Flach, P. 2017. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Artificial Intelligence and Statistics*, 623–631. PMLR.

Kumar, A.; Sarawagi, S.; and Jain, U. 2018. Trainable calibration measures for neural networks from kernel mean embeddings. In *International Conference on Machine Learning*, 2805–2814. PMLR.

Lei, B.; Zhang, R.; Xu, D.; and Mallick, B. 2022. Calibrating the Rigged Lottery: Making All Tickets Reliable. In *International Conference on Learning Representations*.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2980–2988.

Liu, B.; Rony, J.; Galdran, A.; Dolz, J.; and Ben Ayed, I. 2023. Class adaptive network calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16070–16079.

Menon, A. K.; Rawat, A. S.; Reddi, S. J.; Kim, S.; and Kumar, S. 2020. Why distillation helps: a statistical perspective. *arXiv preprint arXiv:2005.10419*.

Milios, D.; Camoriano, R.; Michiardi, P.; Rosasco, L.; and Filippone, M. 2018. Dirichlet-based gaussian processes for large-scale calibrated classification. In *Advances in Neural Information Processing Systems*, volume 31.

Moon, J.; Kim, J.; Shin, Y.; and Hwang, S. 2020. Confidence-aware learning for deep neural networks. In *International Conference on Machine Learning*, 7034–7044. PMLR.

Mukhoti, J.; Kulharia, V.; Sanyal, A.; Golodetz, S.; Torr, P.; and Dokania, P. 2020. Calibrating deep neural networks using focal loss. In *Advances in Neural Information Processing Systems*, volume 33, 15288–15299.

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

Nguyen, K.; and O'Connor, B. 2015. Posterior calibration and exploratory analysis for natural language processing models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1587–1598.

Niculescu-Mizil, A.; and Caruana, R. 2005. Predicting good probabilities with supervised learning. In *International Conference on Machine Learning*, 625–632.

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

Rahimi, A.; Shaban, A.; Cheng, C.-A.; Hartley, R.; and Boots, B. 2020. Intra order-preserving functions for calibration of multi-class neural networks. In *Advances in Neural Information Processing Systems*, volume 33, 13456–13467.

Si, C.; Zhao, C.; Min, S.; and Boyd-Graber, J. 2022. Re-Examining calibration: The case of question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2814–2829.

Tao, L.; Dong, M.; Liu, D.; Sun, C.; and Xu, C. 2023. Calibrating a deep neural network with its predecessors. *arXiv preprint arXiv:2302.06245*.

Tao, L.; Dong, M.; and Xu, C. 2023. Dual focal loss for calibration. In *International Conference on Machine Learning*.

Thulasidasan, S.; Chennupati, G.; Bilmes, J. A.; Bhattacharya, T.; and Michalak, S. 2019. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32.

Tomani, C.; Cremers, D.; and Buettner, F. 2022. Parameterized temperature scaling for boosting the expressive power in post-hoc uncertainty calibration. In *European Conference on Computer Vision*, 555–569. Springer.

Tran, G.-L.; Bonilla, E. V.; Cunningham, J.; Michiardi, P.; and Filippone, M. 2019. Calibrating deep convolutional gaussian processes. In *International Conference on Artificial Intelligence and Statistics*, 1554–1563. PMLR.

Wang, D.-B.; Feng, L.; and Zhang, M.-L. 2021. Rethinking calibration of deep neural networks: Do not be afraid of overconfidence. In *Advances in Neural Information Processing Systems*, volume 34, 11809–11820.

Wang, Y.; Li, L.; and Dang, C. 2019. Calibrating classification probabilities with shape-restricted polynomial regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8): 1813–1827.

Wei, H.; Xie, R.; Cheng, H.; Feng, L.; An, B.; and Li, Y. 2022. Mitigating neural network overconfidence with logit normalization. In *International Conference on Machine Learning*, 23631–23644. PMLR.

Wen, Y.; Tran, D.; and Ba, J. 2019. BatchEnsemble: an alternative approach to efficient ensemble and lifelong learning. In *International Conference on Learning Representations*.

Wenger, J.; Kjellström, H.; and Triebel, R. 2020. Nonparametric calibration for classification. In *International Conference on Artificial Intelligence and Statistics*, 178–190. PMLR.

Widmann, D.; Lindsten, F.; and Zachariah, D. 2019. Calibration tests in multi-class classification: A unifying framework. In *Advances in Neural Information Processing Systems*, volume 32.

Yu, Y.; Bates, S.; Ma, Y.; and Jordan, M. 2022. Robust Calibration with Multi-domain Temperature Scaling. In *Advances in Neural Information Processing Systems*, volume 35, 27510–27523. Curran Associates, Inc.

Zadrozny, B.; and Elkan, C. 2001. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *International Conference on Machine Learning*, volume 1, 609–616.

Zadrozny, B.; and Elkan, C. 2002. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 694–699.

Zhang, Z.; and Xiang, X. 2023. Decoupling maxlogit for out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3388–3397.