On Focal Loss for Class-Posterior Probability Estimation: A Theoretical Perspective

Nontawat Charoenphakdee*
The University of Tokyo / RIKEN AIP
nontawat@ms.k.u-tokyo.ac.jp

Jayakorn Vongkulbhisal*
IBM Research
jayakornv@ibm.com

Nuttapong Chairatanakul
Tokyo Institute of Technology / RWBC-OIL, AIST
nuttapong.crtnk@gmail.com

Masashi Sugiyama
RIKEN AIP / The University of Tokyo
sugi@k.u-tokyo.ac.jp

Abstract

The focal loss has demonstrated its effectiveness in many real-world applications such as object detection and image classification, but its theoretical understanding has been limited so far. In this paper, we first prove that the focal loss is classification-calibrated, i.e., its minimizer surely yields the Bayes-optimal classifier and thus the use of the focal loss in classification can be theoretically justified. However, we also prove a negative fact that the focal loss is not strictly proper, i.e., the confidence score of the classifier obtained by focal loss minimization does not match the true class-posterior probability. This may cause the trained classifier to give an unreliable confidence score, which can be harmful in critical applications. To mitigate this problem, we prove that there exists a particular closed-form transformation that can recover the true class-posterior probability from the outputs of the focal risk minimizer. Our experiments show that our proposed transformation successfully improves the quality of class-posterior probability estimation and improves the calibration of the trained classifier, while preserving the same prediction accuracy.

1. Introduction

It is well-known that training classifiers with the same model architecture can have a huge performance difference if they are trained using different loss functions [3, 25, 16, 8]. To choose an appropriate loss function, it is highly useful to know theoretical properties of loss functions. For example, let us consider the hinge loss, which is related to the support vector machine [12, 3]. This loss function is known to be suitable for classification since minimizing this loss can achieve the Bayes-optimal classifier. However, it is also known that training with the hinge loss does not give the Bayes-optimal solutions for bipartite ranking [15, 47] and class-posterior probability estimation [38]. Such theoretical drawbacks of the hinge loss have been observed to be relevant in practice as well [37, 47]. Not only the hinge loss, but many other loss functions have also been analyzed and their theoretical results have been used as a guideline to choose an appropriate loss function for many problems, *e.g.*, classification from noisy labels [16, 8, 26], classification with rejection [55, 34], and direct optimization of linear-fractional metrics [2, 36].

Recently, the focal loss has been proposed as an alternative to the popular cross-entropy loss [25]. This loss function has been shown to be preferable over the cross-entropy loss when facing the class imbalance problem. Because of its effectiveness, it has been successfully applied in many applications, *e.g.*, medical diagnosis [48, 54, 41, 1], speech processing [46], and natural language processing [40]. Although the focal loss has been successfully applied in many real-world problems [48, 54, 41, 6, 27, 9, 39, 45, 43, 1], considerably less attention has been paid to the theoretical understanding of this loss function. For example, a fundamental question whether we can estimate a class-posterior probability from the classifier trained with the focal loss has remained unanswered. Knowing such a property is highly important when one wants to utilize the prediction confidence. For example, one may defer the decision to a human expert when a classifier has low prediction confidence [10, 55, 34, 29, 7], or one may use the prediction confidence to teach a new model, which has been studied in the literature of knowledge distillation [20, 51, 28].

Motivated by the usefulness of loss function analysis and the lack of theoretical understanding of the focal loss, the goal of this paper is to provide an extensive analysis of this loss function so that we can use it appropriately for the real-

*Nontawat and Jayakorn contributed equally.

world applications. Our contributions can be summarized as follows:

- In Sec. 3, we prove that the focal loss is classification-calibrated (Thm. 3), which theoretically confirms that the optimal classifier trained with the focal loss can achieve the Bayes-optimal classifier.
- In Sec. 4, we prove that learning with the focal loss can give both underconfident and overconfident classifiers (Thm. 8). Our result suggests that the simplex output of the classifier is not reliable as a class-posterior probability estimator (Thm. 5).
- In Sec. 5, we prove that the true class-posterior probability can be theoretically recovered from the focal risk minimizer by our proposed novel transformation $\boldsymbol{\Psi}^{\gamma}$ (Thm. 11). This allows us to calibrate the confidence score of the classifier, while maintaining the same decision rule (Prop. 12).

2. Preliminaries

In this section, we begin by describing the problem setting and notation we use in this paper. Then, we explain fundamental properties of loss functions used for classification, and end the section with a review of the focal loss.

2.1. Multiclass classification

Let $\mathcal{X}$ be an input space and $\mathcal{Y} = \{1, 2, \dots, K\}$ be a label space, where K denotes the number of classes[1]. In multiclass classification, we are given labeled examples $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ independently drawn from an unknown probability distribution (i.i.d.) over $\mathcal{X} \times \mathcal{Y}$ with density $p(\boldsymbol{x}, y)$. The goal of classification is to find a classifier $f\colon \mathcal{X} \to \mathcal{Y}$ that minimizes the following classification risk:

$$R^{\ell_{0\text{-}1}}(f) = \mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim p(\boldsymbol{x},y)}[\ell_{0\text{-}1}(f(\boldsymbol{x}), y)], \tag{1}$$

where $\ell_{0\text{-}1}$ is the zero-one loss $\ell_{0\text{-}1}(f(\boldsymbol{x}), y) = \mathbb{1}_{[f(\boldsymbol{x})\neq y]}$. Next, let us define the true class-posterior probability vector as $\boldsymbol{\eta}(\boldsymbol{x}) = [\eta_1(\boldsymbol{x}), \dots, \eta_K(\boldsymbol{x})]^\top$, where $\eta_y(\boldsymbol{x}) = p(y|\boldsymbol{x})$ denotes the true class-posterior probability for a class y. It is well-known that the Bayes-optimal classifier $f^{\ell_{0\text{-}1},*}$, which minimizes the expected classification risk in Eq. (1), can be defined as follows:.

Definition 1 (Bayes-optimal classifier [56])**.** The Bayes-optimal solution of multiclass classification, $f^{\ell_{0\text{-}1},*} = \arg\min_f R^{\ell_{0\text{-}1}}(f)$, can be expressed as

$$f^{\ell_{0\text{-}1},*}(\boldsymbol{x}) = \arg\max_y \eta_y(\boldsymbol{x}). \tag{2}$$

[1]Bold letters denote vectors, *e.g.*, $\boldsymbol{x}$. Non-bold letters denote scalars, *e.g.*, x. Subscripted letters denote vector elements, *e.g.*, x_i is element i of $\boldsymbol{x}$. $\mathbb{1}_{[\cdot]}$ denotes the indicator function. $\boldsymbol{x}^\top$ denotes the transpose of $\boldsymbol{x}$.

As suggested in Eq. (2), knowing the true class-posterior probability $\boldsymbol{\eta}$ can give the Bayes-optimal classifier but the converse is not necessarily true [3, 50]. The support vector machine [12] is a good example of a learning method that achieves the Bayes-optimal classifier but its confidence score is not guaranteed to obtain the true class-posterior probability [12, 37].

2.2. Surrogate loss

A common practice to learn a classifier using a neural network is to learn a mapping $\boldsymbol{q} : \mathcal{X} \to \Delta^K$, which maps an input to a K-dimensional vector. The simplex output $\boldsymbol{q}$ is often interpreted as a probability distribution over predicted output classes. We denote $\boldsymbol{q}(\boldsymbol{x}) = [q_1(\boldsymbol{x}), \dots, q_K(\boldsymbol{x})]^\top$, where $q_y\colon \mathcal{X} \to [0, 1]$ is a score for class y and $\sum_{y=1}^K q_y(\boldsymbol{x}) = 1$. One typical choice of a mapping $\boldsymbol{q}$ would be a deep convolutional neural network with a softmax function as the output layer. Given an example $\boldsymbol{x}$ and a trained mapping function $\boldsymbol{q}$, a decision rule $f^{\boldsymbol{q}}$ can be inferred by selecting a class with the largest score:

$$f^{\boldsymbol{q}}(\boldsymbol{x}) = \arg\max_y q_y(\boldsymbol{x}). \tag{3}$$

In classification, although the goal is to minimize the classification risk in Eq. (1), it is not straightforward to minimize the classification risk in practice. The first reason is we are given finite examples, not the full distribution. Another reason is minimizing the risk w.r.t *the zero-one loss* is known to be computationally infeasible [56, 3]. As a result, it is common to minimize an *empirical surrogate risk* [3, 49]. Let $\ell\colon \Delta^K \times \Delta^K \to \mathbb{R}$ be a *surrogate loss* and $\boldsymbol{e}_y \in \{0, 1\}^K$ be a one-hot vector with 1 at the y-th index and 0 otherwise. By following the empirical risk minimization approach [49], we minimize the following empirical surrogate risk:

$$\widehat{R}^{\ell}(\boldsymbol{q}) = \frac{1}{n}\sum_{i=1}^n \ell(\boldsymbol{q}(\boldsymbol{x}_i), \boldsymbol{e}_{y_i}), \tag{4}$$

where regularization can also be added to avoid overfitting.

Note that the choice of a surrogate loss is not straightforward and can highly influence the performance of a trained classifier. Necessarily, we should use a surrogate loss that is easier to minimize than the zero-one loss. Moreover, the surrogate risk minimizer should also minimize the expected classification risk in Eq. (1) as well.

2.3. Focal loss

In this paper, we focus on a surrogate loss $\ell\colon \Delta^K \times \Delta^K \to \mathbb{R}$ that receives two simplex vectors as arguments. Let $\boldsymbol{u}, \boldsymbol{v} \in \Delta^K$, and $\gamma \geq 0$ be a nonnegative scalar. The focal loss $\ell_{\mathrm{FL}}^{\gamma} : \Delta^K \times \Delta^K \to \mathbb{R}$ is defined as follows [25]:

$$\ell_{\mathrm{FL}}^{\gamma}(\boldsymbol{u}, \boldsymbol{v}) = -\sum_{i=1}^K v_i(1 - u_i)^{\gamma}\log(u_i). \tag{5}$$

It can be observed that the focal loss with $\gamma = 0$ is equivalent to the well-known cross-entropy loss, *i.e.*, [25]:

$$\ell_{\mathrm{CE}}(\boldsymbol{u}, \boldsymbol{v}) = -\sum_{i=1}^{K} v_i \log(u_i). \tag{6}$$

Unlike the cross-entropy loss that has been studied extensively [56, 5, 50], we are not aware of any theoretical analysis on the fundamental properties of the focal loss. Most analyses of the focal loss are based on an analysis of its gradient and empirical observation [25, 30]. In this paper, we will study the properties of *classification-calibration* [3, 44] (Sec. 3) and *strict properness* [42, 5, 17] (Sec. 4) to provide a theoretical foundation to the focal loss.

3. Focal loss is classification-calibrated

In this section, we theoretically prove that minimizing the focal risk $R^{\ell^{\gamma}_{\mathrm{FL}}}$ can give the Bayes-optimal classifier, which guarantees to maximize the expected accuracy in classification [56]. We show this fact by proving that the focal loss is *classification-calibrated* [3, 44].

First, let us define the pointwise conditional risk W^{ℓ} of an input $\boldsymbol{x}$ with its class-posterior probability $\boldsymbol{\eta}(\boldsymbol{x})$:

$$W^{\ell}\big(\boldsymbol{q}(\boldsymbol{x}); \boldsymbol{\eta}(\boldsymbol{x})\big) = \sum_{y \in \mathcal{Y}} \eta_y(\boldsymbol{x}) \ell\big(\boldsymbol{q}(\boldsymbol{x}), \boldsymbol{e}_y\big). \tag{7}$$

Intuitively, the pointwise conditional risk W^{ℓ} corresponds to the expected penalty for a data point $\boldsymbol{x}$ when using $\boldsymbol{q}(\boldsymbol{x})$ as a score function. Next, we give the definition of a classification-calibrated loss.

Definition 2 (Classification-calibrated loss [3, 44])**.** *Let $\boldsymbol{q}^{\ell,*} = \arg\min_{\boldsymbol{q}} W^{\ell}\big(\boldsymbol{q}(\boldsymbol{x}); \boldsymbol{\eta}(\boldsymbol{x})\big)$ be the minimizer of the pointwise conditional risk. If $R^{\ell_{0\text{-}1}}(f^{\boldsymbol{q}^{\ell,*}}) = R^{\ell_{0\text{-}1}}(f^{\ell_{0\text{-}1},*})$, then a loss ℓ is classification-calibrated.*

Classification-calibration guarantees that the minimizer of the pointwise conditional risk of a surrogate loss will give the Bayes-optimal classifier. Definition 2 suggests that by minimizing a classification-calibrated loss, even if $\boldsymbol{q}^{\ell,*}(\boldsymbol{x})$ is not equal to the true class-posterior probability $\boldsymbol{\eta}(\boldsymbol{x})$, we can still achieve the Bayes-optimal classifier from $\boldsymbol{q}^{\ell,*}(\boldsymbol{x})$ as long as their decision rule matches.

For notational simplicity, we use $\boldsymbol{q}^{\gamma,*}$ to denote $\boldsymbol{q}^{\ell^{\gamma}_{\mathrm{FL}},*}$, *i.e.*, the focal risk minimizer with the parameter γ. The following theorem guarantees that the focal loss is classification-calibrated (its proof is given in Appx. A.4).

Theorem 3. *For any $\gamma \geq 0$, the focal loss $\ell^{\gamma}_{\mathrm{FL}}$ is classification-calibrated.*

Our proof is based on showing that the focal loss has the *strictly order-preserving property*, which is sufficient for classification-calibration [56]. The order-preserving property suggests that for $\boldsymbol{\eta}(\boldsymbol{x})$, the pointwise conditional risk $W^{\ell^{\gamma}_{\mathrm{FL}}}$ has the minimizer $\boldsymbol{q}^{\gamma,*}(\boldsymbol{x})$ such that $\eta_i(\boldsymbol{x}) < \eta_j(\boldsymbol{x}) \Rightarrow q^{\gamma,*}_i(\boldsymbol{x}) < q^{\gamma,*}_j(\boldsymbol{x})$. Since $\boldsymbol{q}^{\gamma,*}$ preserves the order of $\boldsymbol{\eta}$, it is straightforward to see that the focal risk minimizer achieves the Bayes-optimal risk, *i.e.*, $R^{\ell_{0\text{-}1}}(f^{\boldsymbol{q}^{\gamma,*}}) = R^{\ell_{0\text{-}1}}(f^{\ell_{0\text{-}1},*})$. Our result agrees with the empirical effectiveness observed in the previous work [25], where evaluation metrics are based on accuracy or ranking such as mean average precision.

4. On confidence score of classifier trained with focal loss

In this section, we analyze the focal loss for the class-posterior probability estimation problem. We theoretically prove that the simplex output of the focal risk minimizer $\boldsymbol{q}^{\gamma,*}$ does not give the true class-posterior probability. Further, we reveal that the focal loss can yield both underestimation and overestimation of the true class-posterior probability.

4.1. Focal loss is *not* strictly proper

To ensure that a surrogate loss is appropriate for class-posterior probability estimation, it is required that a surrogate loss is *strictly proper*, which is defined as follows.

Definition 4 (Strictly proper loss [42, 5, 17])**.** *We say that a loss $\ell : \Delta^K \times \Delta^K \to \mathbb{R}$ is strictly proper if $\ell(\boldsymbol{u}, \boldsymbol{v})$ is minimized if and only if $\boldsymbol{u} = \boldsymbol{v}$.*

The notion of strict properness can be seen as a natural requirement of a loss when one wants to estimate the true class-posterior probability [52]. When comparing between the ground truth probability $\boldsymbol{v}$ and its estimate $\boldsymbol{u}$, we want a loss function to be minimized if and only if $\boldsymbol{u} = \boldsymbol{v}$, meaning that the probability estimation is correct. Note that strict properness is a stronger requirement of a loss than classification-calibration because all strictly proper losses are classification-calibrated but the converse is false [38, 52].

Here, we prove that the focal loss is not strictly proper in general (its proof is given in Appx. A.5). In fact, it is strictly proper if and only if $\gamma = 0$, *i.e.*, when it coincides with the cross-entropy loss.

Theorem 5. *For any $\gamma > 0$, the focal loss $\ell^{\gamma}_{\mathrm{FL}}$ is not strictly proper.*

Our Thm. 5 suggests that to minimize the focal loss, the simplex output of a classifier does not necessarily need to coincide with the true class-posterior probability. Surprisingly, a recent work [30] suggested that training with the focal loss can give a classifier with reliable confidence. Although their finding seems to contradict with the fact

that the focal loss is not strictly proper, we will discuss in Sec. 6.3 that this phenomenon could occur in practice due to the fact that deep neural networks (DNNs) can suffer from overconfident estimation of the true class-posterior probability [18].

4.2. Focal loss gives under/overconfident classifier

Here, we take a closer look at the behavior of the simplex output of the focal risk minimizer $\boldsymbol{q}^{\gamma,*}$. We begin by pointing out that there exists the case where $\boldsymbol{q}^{\gamma,*}(\boldsymbol{x})$ coincides with $\boldsymbol{\eta}(\boldsymbol{x})$ (its proof can be found in Appx. A.6).

Proposition 6. *Define* $\mathcal{S}^K = \{\boldsymbol{v} \in \Delta^K : v_i \in \{0, \max_j v_j\}\}$. *If* $\boldsymbol{q}^{\gamma,*}(\boldsymbol{x}) \in \mathcal{S}^K$, *then* $\boldsymbol{q}^{\gamma,*}(\boldsymbol{x}) = \boldsymbol{\eta}(\boldsymbol{x})$.

The set $\mathcal{S}^K$ is the set of probability vectors where a subset of classes has uniform probability and the rest has zero probability, *e.g.*, the uniform vector and one-hot vectors. Prop. 6 indicates that, although the focal loss is not strictly proper, the focal risk minimizer can give the true class-posterior probability if $\boldsymbol{q}^{\gamma,*}(\boldsymbol{x}) \in \mathcal{S}^K$.

For the rest of this section, we assume that $\boldsymbol{q}^{\gamma,*}(\boldsymbol{x}) \notin \mathcal{S}^K$ for readability. Next, to analyze the focal loss behavior in general, we propose the notion of $\boldsymbol{\eta}$-underconfidence and $\boldsymbol{\eta}$-overconfidence of the risk minimizer $\boldsymbol{q}^{\ell,*}$ as follows.

Definition 7 ($\boldsymbol{\eta}$-under/overconfidence of risk minimizer)**.** *We say that the risk minimizer* $\boldsymbol{q}^{\ell,*}$ *is* $\boldsymbol{\eta}$*-underconfident (*$\boldsymbol{\eta}$*UC) at* $\boldsymbol{x}$ *if*

$$\max_y q_y^{\ell,*}(\boldsymbol{x}) - \max_y \eta_y(\boldsymbol{x}) < 0. \tag{8}$$

Similarly, $\boldsymbol{q}^{\ell,*}$ *is said to be* $\boldsymbol{\eta}$*-overconfident (*$\boldsymbol{\eta}$*OC) at* $\boldsymbol{x}$ *if*

$$\max_y q_y^{\ell,*}(\boldsymbol{x}) - \max_y \eta_y(\boldsymbol{x}) > 0. \tag{9}$$

Def. 7 can be interpreted as follows. If $\boldsymbol{q}^{\ell,*}$ is $\boldsymbol{\eta}$UC (*resp.*, $\boldsymbol{\eta}$OC) at $\boldsymbol{x}$, then the confidence score $\max_y q_y^{\ell,*}(\boldsymbol{x})$ for the predicted class must be lower (*resp.*, higher) than that of the true class-posterior probability $\max_y \eta_y(\boldsymbol{x})$. It is straightforward to see that the risk minimizer of any strictly proper loss does not give an $\boldsymbol{\eta}$UC/$\boldsymbol{\eta}$OC classifier because $\boldsymbol{q}^{\ell,*}$ must be equal to the true class-posterior probability $\boldsymbol{\eta}$. Thus, Def. 7 is not useful for characterizing strictly proper losses but it is highly useful for analyzing the behavior of the focal loss.

We emphasize that the notion of $\boldsymbol{\eta}$-under/overconfidence of the risk minimizer is significantly different from the notion of *overconfidence* that has been used in the literature of confidence-calibration [18, 22, 30]. In that literature, *overconfidence* was used to describe the *empirical performance* of modern neural networks [13, 35], where a classifier outputs an average confidence score higher than its average accuracy for a set of data points. In our case, $\boldsymbol{\eta}$OC and $\boldsymbol{\eta}$UC

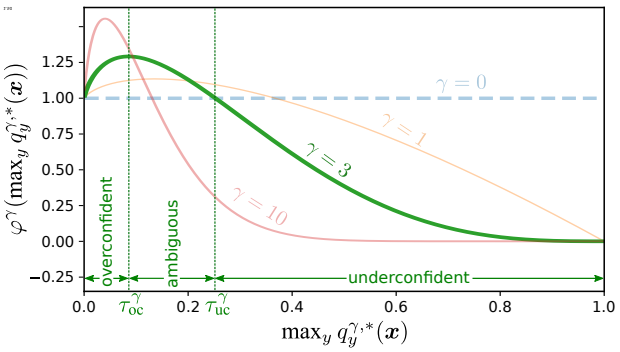


Figure 1. The function $\varphi^\gamma(v)$ for various γ. Visualization of the region where $\boldsymbol{q}^{\gamma,*}$ can be $\boldsymbol{\eta}$UC and $\boldsymbol{\eta}$OC are emphasized for $\gamma = 3$. Whether $\boldsymbol{q}^{\gamma,*}$ is $\boldsymbol{\eta}$OC or $\boldsymbol{\eta}$UC can be largely determined by the relation between φ^γ and the maximum predicted score $\max_y q_y^{\gamma,*}(\boldsymbol{x})$. See details in Thm. 8.

are based on the behavior of the risk minimizer of the loss function, which does not concern with the empirical validation.

To study the behavior of $\boldsymbol{q}^{\gamma,*}$, let us define a function $\varphi^\gamma : [0,1] \to \mathbb{R}$ as

$$\varphi^\gamma(v) = (1-v)^\gamma - \gamma(1-v)^{\gamma-1} v \log v. \tag{10}$$

This function plays a key role in characterizing if $\boldsymbol{q}^{\gamma,*}$ is $\boldsymbol{\eta}$UC/$\boldsymbol{\eta}$OC. See Appx. A for more details on how φ^γ was derived. Next, we state our main theorem that characterizes the $\boldsymbol{\eta}$UC/$\boldsymbol{\eta}$OC behaviors of the risk minimizer of the focal loss $\boldsymbol{q}^{\gamma,*}$ (its proof is given in Appx. A.7).

Theorem 8. *Consider the focal loss* $\ell_{\mathrm{FL}}^\gamma$ *where* $\gamma > 0$. *Define* $\tau_{\mathrm{oc}}^\gamma = \arg\max_v \varphi^\gamma(v)$ *and* $\tau_{\mathrm{uc}}^\gamma \in (0,1)$ *such that* $\varphi^\gamma(\tau_{\mathrm{uc}}^\gamma) = 1$. *If* $\boldsymbol{q}^{\gamma,*}(\boldsymbol{x}) \notin \mathcal{S}^K$, *we have*

1. $0 < \tau_{\mathrm{oc}}^\gamma < \tau_{\mathrm{uc}}^\gamma < 0.5$.
2. $\boldsymbol{q}^{\gamma,*}$ *is* $\boldsymbol{\eta}$*OC if* $\max_y q_y^{\gamma,*}(\boldsymbol{x}) \in (0, \tau_{\mathrm{oc}}^\gamma]$.
3. $\boldsymbol{q}^{\gamma,*}$ *is* $\boldsymbol{\eta}$*UC if* $\max_y q_y^{\gamma,*}(\boldsymbol{x}) \in [\tau_{\mathrm{uc}}^\gamma, 1)$.

Thm. 8 suggests that **training with focal loss can lead to both $\boldsymbol{\eta}$UC and $\boldsymbol{\eta}$OC classifiers**. It also indicates that we can determine if $\boldsymbol{q}^{\gamma,*}$ is $\boldsymbol{\eta}$OC or $\boldsymbol{\eta}$UC at $\boldsymbol{x}$ if $\max_y q_y^{\gamma,*}(\boldsymbol{x})$ is in $(0, \tau_{\mathrm{oc}}^\gamma]$ or $[\tau_{\mathrm{uc}}^\gamma, 1)$. For $\max_y q_y^{\gamma,*}(\boldsymbol{x}) \in (\tau_{\mathrm{oc}}^\gamma, \tau_{\mathrm{uc}}^\gamma)$, we may require the knowledge of $q_{y'}$ for all $y' \in \mathcal{Y}$ to determine if $\boldsymbol{q}^{\gamma,*}$ is $\boldsymbol{\eta}$UC or $\boldsymbol{\eta}$OC. Nevertheless, in Sec. 5, we will show that given any $\boldsymbol{q}^{\gamma,*}(\boldsymbol{x})$, $\boldsymbol{\eta}$UC and $\boldsymbol{\eta}$OC can be determined everywhere *including the ambiguous region* $(\tau_{\mathrm{oc}}^\gamma, \tau_{\mathrm{uc}}^\gamma)$ by using our novel transformation $\boldsymbol{\Psi}^\gamma$. Fig. 1 illustrates the overconfident, ambiguous, and underconfident regions of $\boldsymbol{q}^{\gamma,*}$. Interestingly, the fact that $\boldsymbol{q}^{\gamma,*}$ can be $\boldsymbol{\eta}$OC cannot be explained by the previous analysis [30], which only implicitly suggested that $\boldsymbol{q}^{\gamma,*}$ is $\boldsymbol{\eta}$UC by interpreting focal loss minimization as the minimization of an upper bound of the regularized Kullback-Leibler divergence.

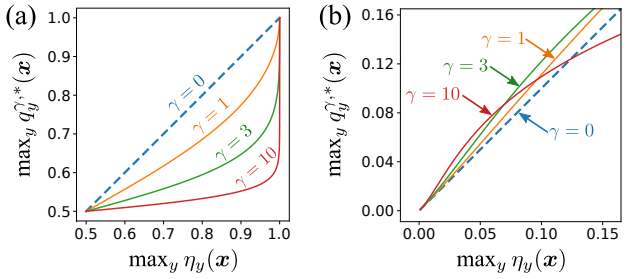


Figure 2. Relation between $\max_y \eta_y(\boldsymbol{x})$ and $\max_y q_y^{\gamma,*}(\boldsymbol{x})$ under different γ. (a) shows the relation in the binary case, where $\boldsymbol{q}^{\gamma,*}$ can be $\boldsymbol{\eta}$UC and the effect of $\boldsymbol{\eta}$UC is stronger as γ increases. On the other hand, (b) shows the relation of a 1000-way classification where the focal loss can also induce $\boldsymbol{\eta}$OC.

Since calculating $\tau_{\text{oc}}^{\gamma}$ and $\tau_{\text{uc}}^{\gamma}$ is not straightforward because their simple close-form solutions may not exist for all γ, we provide the following corollary to show that there exists a region where $\boldsymbol{q}^{\gamma,*}$ is always $\boldsymbol{\eta}$UC regardless of the choice of γ (its proof is given in Appx. A.8).

Corollary 9. *For all $\gamma > 0$, $\boldsymbol{q}^{\gamma,*}$ is $\boldsymbol{\eta}$UC if $\max_y q_y^{\gamma,*}(\boldsymbol{x}) \in (0.5, 1)$.*

Cor. 9 suggests that $\boldsymbol{q}^{\gamma,*}$ is $\boldsymbol{\eta}$UC when the label is not too ambiguous. In practice, a classifier is more likely to be $\boldsymbol{\eta}$UC but it still could be $\boldsymbol{\eta}$OC when the number of classes K is large and γ is small. Fig. 2b demonstrates that $\boldsymbol{q}^{\gamma,*}$ can be $\boldsymbol{\eta}$OC when having 1000 classes for different γ.[2]

We also provide the following corollary, which is an immediate implication from Cor. 9 for the binary classification scenario (its proof is given in Appx. A.9).

Corollary 10. *For all $\gamma > 0$, $\boldsymbol{q}^{\gamma,*}$ is always $\boldsymbol{\eta}$UC in binary classification unless $\boldsymbol{q}^{\gamma,*}(\boldsymbol{x})$ is uniform or a one-hot vector.*

Fig. 2a demonstrates that $\boldsymbol{q}^{\gamma,*}$ is $\boldsymbol{\eta}$UC in binary classification, where a larger γ causes a larger gap between $\max_y q_y^{\gamma,*}$ and the true class-posterior probability.

5. Recovering class-posterior probability from classifiers trained with focal loss

In this section, we propose a novel transformation $\boldsymbol{\Psi}^{\gamma}$ to recover the true class-posterior probability from the focal risk minimizer with theoretical justification. Then, we provide a numerical example to demonstrate its effectiveness.

5.1. Proposed transformation $\boldsymbol{\Psi}^{\gamma}$

Our following theorem reveals that there exists a transformation that can be computed *in a closed form* to recover the true class-posterior probability from the focal risk minimizer (its proof is given in Appx. A.1).

[2]We numerically found that $\boldsymbol{q}^{\gamma,*}$ can be $\boldsymbol{\eta}$OC with as minimum as $K = 5$ classes with $\gamma \leq 0.03$ when $\max_y q_y^{\gamma,*}(\boldsymbol{x}) \to 1/K$.

Theorem 11. *Let $\boldsymbol{\eta}(\boldsymbol{x})$ be the true class-posterior probability of an input $\boldsymbol{x}$ and $\boldsymbol{q}^{\gamma,*} = \arg\min_{\boldsymbol{q}} W^{\ell_{\text{FL}}^{\gamma}}(\boldsymbol{q}(\boldsymbol{x}), \boldsymbol{\eta}(\boldsymbol{x}))$ be the focal risk minimizer, where $\gamma \geq 0$. Then, the true class-posterior probability $\boldsymbol{\eta}(\boldsymbol{x})$ can be recovered from $\boldsymbol{q}^{\gamma,*}$ with the transformation $\boldsymbol{\Psi}^{\gamma} : \Delta^K \to \Delta^K$,* i.e.,

$$\boldsymbol{\eta}(\boldsymbol{x}) = \boldsymbol{\Psi}^{\gamma}(\boldsymbol{q}^{\gamma,*}(\boldsymbol{x})), \tag{11}$$

where $\boldsymbol{\Psi}^{\gamma}(\boldsymbol{v}) = [\Psi_1^{\gamma}(\boldsymbol{v}), \ldots, \Psi_K^{\gamma}(\boldsymbol{v})]^{\top}$, and

$$\Psi_i^{\gamma}(\boldsymbol{v}) = \frac{h^{\gamma}(v_i)}{\sum_{l=1}^{K} h^{\gamma}(v_l)}, \tag{12}$$

$$h^{\gamma}(v) = \frac{v}{\varphi^{\gamma}(v)} = \frac{v}{(1-v)^{\gamma} - \gamma(1-v)^{\gamma-1} v \log v}. \tag{13}$$

For completeness, we also define $\Psi_i^{\gamma}(\boldsymbol{v}) = v_i$ if $\boldsymbol{v}$ is a one-hot vector. Note that if $\gamma = 0$, then $\eta_i(\boldsymbol{x}) = \Psi_i^{\gamma}(\boldsymbol{q}^{\gamma,*}(\boldsymbol{x})) = q_i^{\gamma,*}(\boldsymbol{x})$, which coincides with the known analysis that the cross-entropy loss is strictly proper [14, 17]. On the other hand, if $\gamma > 0$, an additional step of applying $\boldsymbol{\Psi}^{\gamma}$ is required to recover the true class-posterior probability. We also want to emphasize that for any given $\max_y q_y^{\gamma,*}(\boldsymbol{x})$ in the ambiguous region (see Fig. 1), one can easily determine if it is $\boldsymbol{\eta}$UC or $\boldsymbol{\eta}$OC by comparing $\max_y q_y^{\gamma,*}(\boldsymbol{x})$ and $\max_y \Psi_y^{\gamma}(\boldsymbol{q}^{\gamma,*}(\boldsymbol{x}))$.

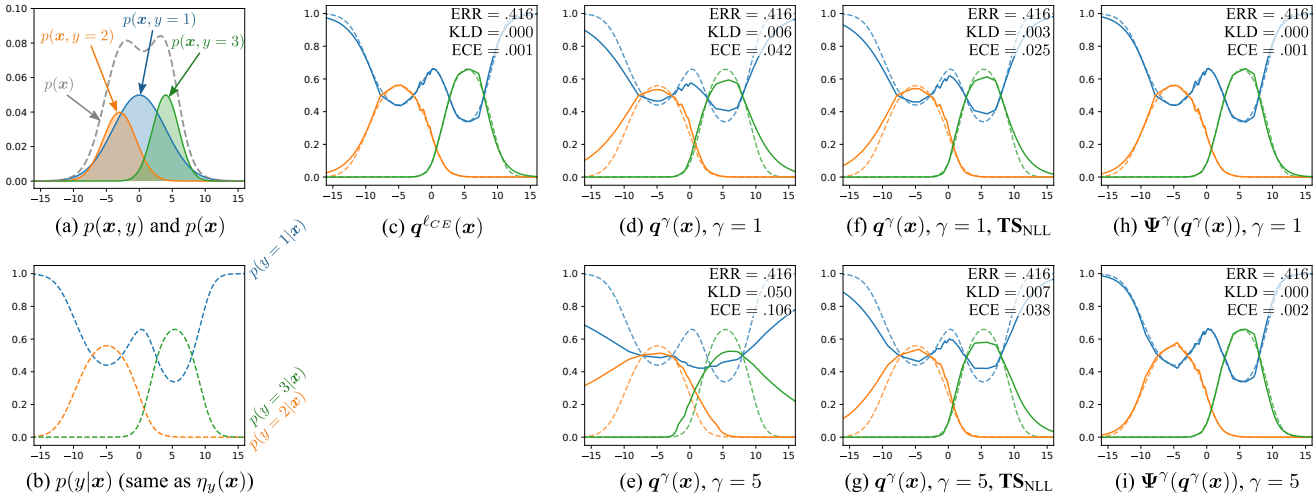
Next, we confirm that our proposed transformation $\boldsymbol{\Psi}^{\gamma}$ does not degrade the classification performance of the classifier by proving that $\boldsymbol{\Psi}^{\gamma}$ preserves the decision rule (its proof is given in Appx. A.10).

Proposition 12. *Given $\boldsymbol{v} \in \Delta^K$ and $\gamma \geq 0$, we have $\arg\max_i \Psi_i^{\gamma}(\boldsymbol{v}) = \arg\max_i v_i$.*

In summary, if one wants to recover the true class-posterior probability from the focal risk minimizer with $\gamma \neq 0$, an additional step of applying $\boldsymbol{\Psi}^{\gamma}$ is suggested by Thm. 11. However, if one only wants to know which class has the highest prediction probability, then applying $\boldsymbol{\Psi}^{\gamma}$ is unneeded since it does not change the prediction result. We want to emphasize that that using the transformation $\boldsymbol{\Psi}^{\gamma}$ to recover the true class-posterior probability is significantly different and orthogonal from using a heuristic technique such as Platt scaling [37]. The differences are: (1) Using $\boldsymbol{\Psi}^{\gamma}$ is theoretically guaranteed given the risk minimizer and (2) No additional training is involved since the transformation $\boldsymbol{\Psi}^{\gamma}$ does not contain any tuning parameter, whereas Platt scaling requires additional training, which can be computationally expensive when using a large training dataset.

5.2. Numerical illustration

Here, we use synthetic data to demonstrate the $\boldsymbol{\eta}$UC property of the focal loss and show that applying $\boldsymbol{\Psi}^{\gamma}$ can successfully recover the true class-posterior probability. The purpose of using the synthetic data is because



(a) $p(\boldsymbol{x}, y)$ and $p(\boldsymbol{x})$ (c) $\boldsymbol{q}^{\ell_{CE}}(\boldsymbol{x})$ (d) $\boldsymbol{q}^{\gamma}(\boldsymbol{x})$, $\gamma = 1$ (f) $\boldsymbol{q}^{\gamma}(\boldsymbol{x})$, $\gamma = 1$, $\mathbf{TS}_{\mathrm{NLL}}$ (h) $\boldsymbol{\Psi}^{\gamma}(\boldsymbol{q}^{\gamma}(\boldsymbol{x}))$, $\gamma = 1$

(b) $p(y|\boldsymbol{x})$ (same as $\eta_y(\boldsymbol{x})$) (e) $\boldsymbol{q}^{\gamma}(\boldsymbol{x})$, $\gamma = 5$ (g) $\boldsymbol{q}^{\gamma}(\boldsymbol{x})$, $\gamma = 5$, $\mathbf{TS}_{\mathrm{NLL}}$ (i) $\boldsymbol{\Psi}^{\gamma}(\boldsymbol{q}^{\gamma}(\boldsymbol{x}))$, $\gamma = 5$

Figure 3. Demonstration of the underconfident ($\boldsymbol{\eta}$UC) property of the focal loss and the result of the transformation $\boldsymbol{\Psi}^{\gamma}$. (a) and (b) show $p(\boldsymbol{x})$, $p(\boldsymbol{x}, y)$, and $p(y|\boldsymbol{x})$ used for training the MLPs. For (c-i), solid graphs are the raw or transformed predicted scores from the MLPs while dashed graphs are $p(y|\boldsymbol{x})$ (same as (b)). The ERR, KLD, and ECE values are reported on the top-right of each subfigure. (c) shows $\boldsymbol{q}^{\ell}_{\mathrm{CE}}(\boldsymbol{x})$ of an MLP trained with ℓ_{CE} while (d) and (e) show $\boldsymbol{q}^{\gamma}(\boldsymbol{x})$ of MLPs trained with $\ell^{\gamma}_{\mathrm{FL}}$ with $\gamma = 1$ and 5. (f) and (g) show the scores after processing with $\mathbf{TS}_{\mathrm{NLL}}$. (h) and (i) show the scores after using the our proposed $\boldsymbol{\Psi}^{\gamma}$ in Eq. (11). See Sec. 5.2 for details.

we know the true class-posterior probability $\boldsymbol{\eta}$ in this problem. Unlike many real-world datasets where only hard labels are given, we can directly evaluate the quality of class-posterior probability estimation using the Kullback-Leibler divergence (KLD), which is defined as $\mathrm{KL}(\boldsymbol{\eta}(\boldsymbol{x})||\boldsymbol{q}(\boldsymbol{x})) = \sum_{i=1}^{K} \eta_i(\boldsymbol{x}) \log \frac{\eta_i(\boldsymbol{x})}{q_i(\boldsymbol{x})}$.

We simulate a 1-dimensional 3-class classification problem with the distribution given in Fig. 3a. We then trained three-layer multilayered perceptrons (MLPs) with ℓ_{CE} and $\ell^{\gamma}_{\mathrm{FL}}$ ($\gamma = 1$ and 5) using data sampled from the distribution. The estimated confidence scores $q^{\ell}_{y}(\boldsymbol{x})$ of all losses are shown in Fig. 3c,d,e. We can see that all MLPs can correctly identify the class having the highest class-posterior probability for the whole $\mathcal{X}$ and achieve roughly the same classification error (ERR), which corresponds to the fact that both ℓ_{CE} and $\ell^{\gamma}_{\mathrm{FL}}$ are classification-calibrated. However, while $q^{\ell_{\mathrm{CE}}}_{y}(\boldsymbol{x})$ in Fig. 3c could correctly estimate $\eta_y(\boldsymbol{x})$, $q^{\gamma}_{y}(\boldsymbol{x})$ in Fig. 3d,e do not match $\eta_y(\boldsymbol{x})$, which agrees with our result that the focal loss is not strictly proper. More precisely, the value of the $\max_y q^{\gamma}_{y}(\boldsymbol{x})$ is lower than $\max_y \eta_y(\boldsymbol{x})$, which indicates that $\boldsymbol{q}^{\gamma}(\boldsymbol{x})$ is $\boldsymbol{\eta}$UC. With a larger γ, we can observe this trend more significantly by looking at KLD and the expected calibration error (ECE) [32, 18], where low ECE indicates good *empirical confidence*.

One well-known approach to improve confidence estimation in neural networks is temperature scaling (TS) [18]. We applied TS with negative log-likelihood (NLL) as the validation objective ($\mathbf{TS}_{\mathrm{NLL}}$) to the MLPs trained with the focal loss. We can see from Fig. 3f,g that while $\mathbf{TS}_{\mathrm{NLL}}$ made the $q^{\gamma}_{y}(\boldsymbol{x})$ move closer to $\eta_y(\boldsymbol{x})$, a large gap between them still exists, suggesting that $\mathbf{TS}_{\mathrm{NLL}}$ fails to obtain the true class-posterior probability.

By using the transformation $\boldsymbol{\Psi}^{\gamma}$, we can plot Fig. 3h,i and see that $\boldsymbol{\Psi}^{\gamma}(\boldsymbol{q}^{\gamma}(\boldsymbol{x}))$ can improve the quality of the estimation, where both KLD and ECE are almost zero. Recall that $\boldsymbol{\Psi}^{\gamma}$ can be applied without any additional data or changing decision rule, thus the ERR remains exactly the same. This synthetic experiment demonstrates that the simplex outputs of neural networks trained with the focal loss is likely to be $\boldsymbol{\eta}$UC, and this can be effectively fixed using the transformation $\boldsymbol{\Psi}^{\gamma}$.

6. Experimental results

In this section, we perform experiments to study the behavior of the focal loss and validate the effectiveness of $\boldsymbol{\Psi}^{\gamma}$ under different training paradigms. To do so, we use the CIFAR10 [21] and SVHN [33] datasets as the benchmark datasets. The details of the experiments are as follows.

Models: To see the influence of the model complexity, we used ResNet$L$ [19] with $L \in \{8, 20, 44, 110\}$, where complexity increases as L increases.

Methods: We compared the networks that use $\boldsymbol{\Psi}^{\gamma}$ after the softmax layer to those that do not. Note that both methods have the same accuracy since $\boldsymbol{\Psi}^{\gamma}$ does not affect the decision rule (Prop. 12). We used $\gamma \in \{0, 1, 2, 3\}$ and conducted 10 trials for each experiment setting.

Evaluation metrics: Since true class-posterior probability labels are not available, a common practice is to use ECE to evaluate the quality of prediction confidence [32, 18]. In this paper, we used 10 as the number of bins. ECE-$\boldsymbol{\Psi}^{\gamma}$ (*resp.*, ECE-raw) denotes the ECE of the networks that use (*resp.*, do not use) $\boldsymbol{\Psi}^{\gamma}$. ERR denotes the classification er-

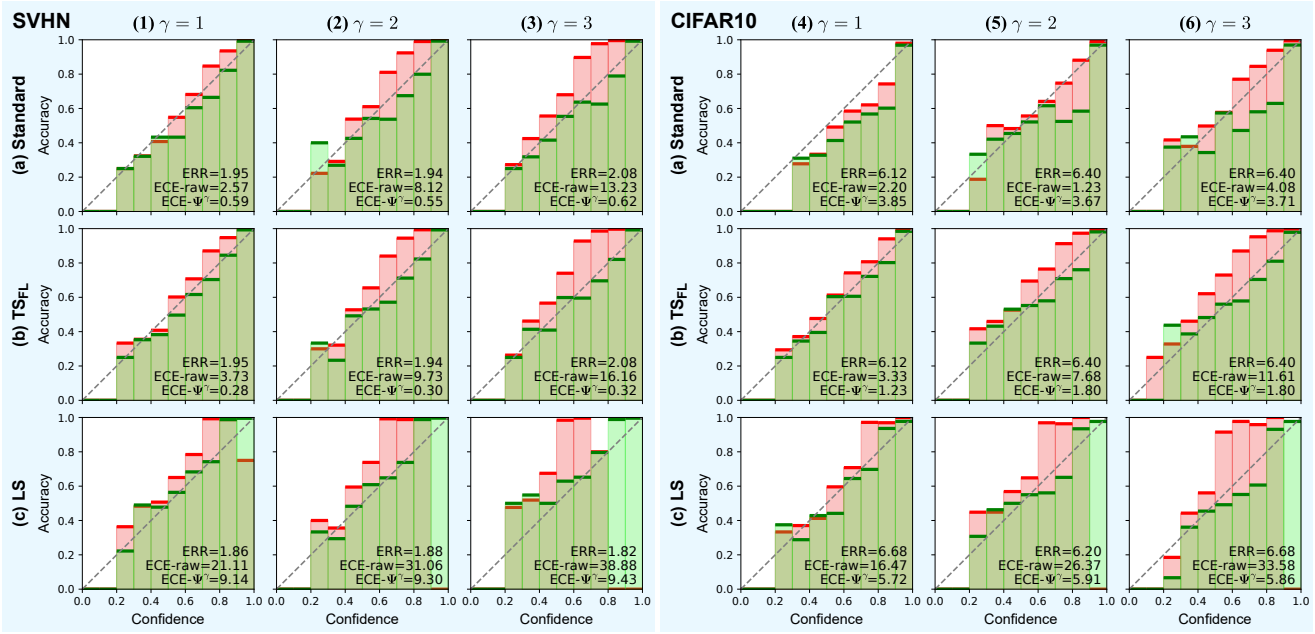


Figure 4. Reliability diagrams of ResNet110 trained with $\ell^\gamma_{\text{FL}}, \gamma = 1, 2, 3$ on SVHN and CIFAR10 datasets. ECE-$\boldsymbol{\Psi}^\gamma$ (*resp.*, ECE-raw) denotes the ECE of the networks that use (*resp.*, do not use) $\boldsymbol{\Psi}^\gamma$ and their diagrams are plotted in green (*resp.*, red). ERR denotes the classification error. Each row shows the results of different training paradigms: (a) **Standard**, (b) **TS**$_{\text{FL}}$, and (c) **LS**. See Sec. 6.1 for details.

ror. We scale the values of ECE-$\boldsymbol{\Psi}^\gamma$, ECE-raw, and ERR to $0 - 100$ for readability in Fig. 4. We report full results on more evaluation metrics and models in Appx. C.

Hyperparameters: For all models, the number of epochs was 200 for CIFAR10 and 50 for SVHN. The batch size was 128. We used SGD with momentum of 0.9, where the initial learning rate was 0.1, which was then divided by 10 at epoch 80 and 150 for CIFAR10 and at epoch 25 and 40 for SVHN. The weight decay parameter was 5×10^{-4}.

6.1. ECE of different training paradigms

We trained models using three different paradigms: (1) **Standard** uses one-hot ground truth vectors, which is known to be susceptible to overconfidence [18]; (2) **TS**$_{\text{FL}}$ post-processes the output of **Standard** with TS that uses the focal loss in the validation objective; and (3) **LS** uses label smoothing to smoothen one-hot labels to soft labels, which has been reported to alleviate the overconfidence issue in DNNs [31]. The label smoothing parameter was 0.1.

Fig. 4 shows the reliability diagrams for ResNet110 trained with the focal loss using different γ. We can see that $\boldsymbol{\Psi}^\gamma$ substantially improves ECE for most settings. This demonstrates that our theoretically-motivated transformation $\boldsymbol{\Psi}^\gamma$ can be highly relevant in practice. For **LS**, ECE-raw drastically increases as γ increases, whereas the value of γ does not significantly affect ECE-$\boldsymbol{\Psi}^\gamma$. Next, in **TS**$_{\text{FL}}$, if $\boldsymbol{\Psi}^\gamma$ is not applied, we can see that ECE-raw degrades compared with that of **Standard**. On the other hand, our transformation $\boldsymbol{\Psi}^\gamma$ can further improve the performance of **Standard**. This could be due to **TS**$_{\text{FL}}$ giving a more accurate estimate of the focal risk minimizer $\boldsymbol{q}^{\gamma,*}$, but $\boldsymbol{q}^{\gamma,*}$ does not coincide with the true class-posterior probability $\boldsymbol{\eta}$ if $\boldsymbol{\Psi}^\gamma$ is not applied, as proven in Thm. 11. Apart from **Standard** in CIFAR10, underconfident bins (*i.e.*, the bins that align above the diagonal of the reliability diagram) can be observed especially when γ is large. The results indicate that the focal loss is susceptible to be underconfident as γ increases, which agrees with our analysis that the focal loss is not strictly proper (Thm. 5) and prone to $\boldsymbol{\eta}$UC (Cor. 9).

6.2. Why does Ψ^γ not always improve ECE?

In Fig. 4, although our transformation $\boldsymbol{\Psi}^\gamma$ can greatly improve the performance for **Standard** in SVHN, it worsens the performance for **Standard** in CIFAR10. This demonstrates that our proposed transformation does not always improve the performance in practice, which could occur when the focal risk minimizer $\boldsymbol{q}^{\gamma,*}$ is not successfully learned. Note that if $\boldsymbol{q}^{\gamma,*}$ is obtained, the transformation $\boldsymbol{\Psi}^\gamma$ is the only mapping to obtain the true class-posterior probability $\boldsymbol{\eta}$ from $\boldsymbol{q}^{\gamma,*}$ (Thm. 11).

Here, we take a closer look at the scenario where $\boldsymbol{\Psi}^\gamma$ could be less effective. We hypothesize that there are two potential reasons: (1) DNNs can overfit the one-hot vector, which leads to overconfident prediction [18]. By using one-hot vectors as labels, perfectly minimizing the empirical risk implies making the confidence score close to a one-hot vector. (2) The amount of data could be insufficient for correctly estimating the true class-posterior probability.

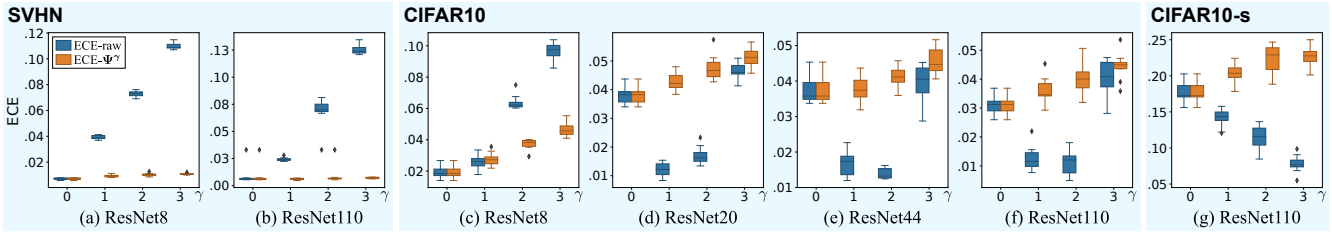


Figure 5. Box plots of ECEs for **Standard** with varying γ using different models on (a-b) SVHN, (c-f) CIFAR10, and (g) CIFAR10-s.

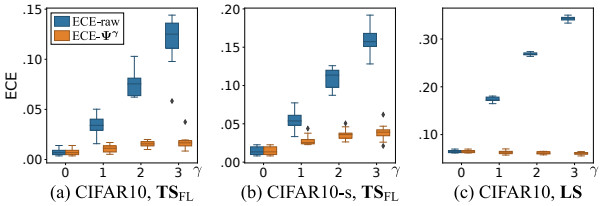


Figure 6. Box plots of ECEs with varying γ for ResNet110 using $\mathbf{TS}_{\mathrm{FL}}$ for CIFAR10 and CIFAR10-s, and **LS** for CIFAR10. It can be observed that using the transformation $\boldsymbol{\Psi}^{\gamma}$ is preferable.

To justify our claim, we conducted experiments with different models on SVHN, CIFAR10, and CIFAR10-s, where CIFAR10-s is CIFAR10 that uses only 10% of training data for each class. Note that SVHN has a larger number of data than CIFAR10, and ResNet$L$ is more complex as L increases. Fig. 5 illustrates ECEs with different dataset size and models. In Fig. 5b,f,g, where the same model was used, we can observe that $\boldsymbol{\Psi}^{\gamma}$ becomes less effective as the dataset size gets smaller. Also, in Fig. 5c-f, where the different models were used in CIFAR10, $\boldsymbol{\Psi}^{\gamma}$ becomes less effective as the model becomes more complex. Therefore, the results agree with our hypotheses that the model complexity and dataset size play a role in the effectiveness of $\boldsymbol{\Psi}^{\gamma}$. Note that $\boldsymbol{\Psi}^{\gamma}$ is still effective in SVHN regardless of the model size in our experiments, as can be seen in Fig. 5a,b, since the size of SVHN may be sufficiently large to accurately estimate $\boldsymbol{q}^{\gamma,*}$ for complex models.

It is also insightful to observe the best value of γ in different settings. More precisely, we see from Fig. 5c-g that the best γ for CIFAR10 is $\gamma = 0$ for ResNet8, $\gamma = 1$ for ResNet20, and $\gamma = 2$ for ResNet110, while the best γ for CIFAR10-s and ResNet110 is $\gamma = 3$. Therefore, we can conclude that the best γ increases as the data size decreases or the model becomes more complex. Nevertheless, For **LS** and $\mathbf{TS}_{\mathrm{FL}}$, we observe that a larger γ always leads to worse performance and $\boldsymbol{\Psi}^{\gamma}$ can effectively mitigate this problem for every dataset, as illustrated in Fig. 6.

6.3. Discussion

Recently, Mukhoti *et al.* [30] studied the relation between the focal loss and the confidence issue of DNNs, and showed that without post-processing, training with the focal loss can achieve lower ECE than that of the cross-entropy loss. Our results indicate that this is not always the case (see SVHN for example). In Appx. C, we provide additional experimental results on 30 datasets to show that the focal loss is less desirable compared with the cross-entropy loss in most datasets, and that $\boldsymbol{\Psi}^{\gamma}$ can successfully improve ECE to be comparable with that of the cross-entropy loss. Nevertheless, the focal loss can also outperform the cross-entropy loss as shown in Fig. 5, which agrees with the previous work [30]. This could occur when classifiers (especially DNNs) suffer from overconfidence due to empirical estimation [18]. Since the focal loss tends to give an $\boldsymbol{\eta}$UC classifier, there may exist a sweet spot for $\gamma > 0$ that gives the best ECE because the overconfident and underconfident effects cancel each other out. In addition, it has been observed that applying TS w.r.t. NLL or ECE on a classifier trained with the focal loss can be empirically effective to reduce ECE [18, 30]. Nevertheless, for a classifier trained with the focal loss, Fig. 3 illustrates that using such heuristics may fail to recover the true class-posterior probability. Theoretically, since TS only tunes one scalar to optimize the validation objective, it may suffer from model misspecification and could fail to achieve the optimal NLL/ECE w.r.t. all measurable functions [38, 52].

7. Conclusions

We proved that the focal loss is classification-calibrated but not strictly proper. We further investigated and pointed out that focal loss can give both underconfident and overconfident classifiers. Then, we proposed a transformation that can theoretically recover the true class-posterior probability from the focal risk minimizer. Experimental results showed that the proposed transformation can improve the performance of class-posterior probability estimation.

Acknowledgements

We would like to thank Zhenguo Wu, Yivan Zhang, Han Bao, and Zhenghang Cui for helpful discussion. Nontawat Charoenphakdee was supported by MEXT scholarship and Google PhD Fellowship program. Nuttapong Chairatanakul was supported by MEXT scholarship. Part of this work is conducted as research activities of AIST-Tokyo Tech Real World Big-Data Computation Open Innovation Laboratory (RWBC-OIL). Masashi Sugiyama was supported by JST CREST Grant Number JPMJCR18A2.

References

[1] Mohamad Mahmoud Al Rahhal, Yakoub Bazi, Haidar Almubarak, Naif Alajlan, and Mansour Al Zuair. Dense convolutional networks with focal loss and image generation for electrocardiogram classification. *IEEE Access*, 7:182225–182237, 2019. 1

[2] Han Bao and Masashi Sugiyama. Calibrated surrogate maximization of linear-fractional utility in binary classification. In *International Conference on Artificial Intelligence and Statistics*, pages 2337–2347. PMLR, 2020. 1

[3] Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006. 1, 2, 3

[4] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004. 12, 15, 18

[5] Andreas Buja, Werner Stuetzle, and Yi Shen. Loss functions for binary class probability estimation and classification: Structure and applications. *Working draft*, 2005. 3

[6] Jie Chang, Xiaoci Zhang, Minquan Ye, Daobin Huang, Peipei Wang, and Chuanwen Yao. Brain tumor segmentation based on 3d unet with multi-class focal loss. In *2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–5. IEEE, 2018. 1

[7] Nontawat Charoenphakdee, Zhenghang Cui, Yivan Zhang, and Masashi Sugiyama. Classification with rejection based on cost-sensitive classification. *arXiv preprint arXiv:2010.11748*, 2020. 1

[8] Nontawat Charoenphakdee, Jongyeong Lee, and Masashi Sugiyama. On symmetric losses for learning from corrupted labels. *ICML*, 2019. 1

[9] Mingqiang Chen, Lin Fang, and Huafeng Liu. Fr-net: Focal loss constrained deep residual networks for segmentation of cardiac mri. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 764–767. IEEE, 2019. 1

[10] C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on information theory*, 16(1):41–46, 1970. 1

[11] Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical japanese literature. *arXiv preprint arXiv:1812.01718*, 2018. 29

[12] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995. 1, 2

[13] Morris H DeGroot and Stephen E Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983. 4

[14] Andrey Feuerverger and Sheikh Rahman. Some aspects of probability forecasting. *Communications in statistics-theory and methods*, 21(6):1615–1632, 1992. 5

[15] Wei Gao and Zhi-Hua Zhou. On the consistency of AUC pairwise optimization. In *IJCAI*, pages 939–945, 2015. 1

[16] Aritra Ghosh, Himanshu Kumar, and PS Sastry. Robust loss functions under label noise for deep neural networks. In *AAAI*, pages 1919–1925, 2017. 1

[17] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007. 3, 5

[18] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. *ICML*, 2017. 4, 6, 7, 8, 27, 29

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 6, 29

[20] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 1

[21] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009. 6

[22] Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. In *NeurIPS*, pages 12316–12326, 2019. 4, 27

[23] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998. 29

[24] Moshe Lichman et al. UCI machine learning repository, 2013. 29

[25] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *CVPR*, 2017. 1, 2, 3

[26] Yang Liu and Hongyi Guo. Peer loss functions: Learning from noisy labels without knowing noise rates. *ICML*, 2020. 1

[27] Mayar Lotfy, Raed M Shubair, Nassir Navab, and Shadi Albarqouni. Investigation of focal loss in deep learning models for femur fractures classification. In *2019 International Conference on Electrical and Computing Technologies and Applications (ICECTA)*, pages 1–4. IEEE, 2019. 1

[28] Aditya Krishna Menon, Ankit Singh Rawat, Sashank J Reddi, Seungyeon Kim, and Sanjiv Kumar. Why distillation helps: a statistical perspective. *arXiv preprint arXiv:2005.10419*, 2020. 1

[29] Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. *ICML*, 2020. 1

[30] Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip HS Torr, and Puneet K Dokania. Calibrating deep neural networks using focal loss. *NeurIPS*, 2020. 3, 4, 8

[31] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? In *NeurIPS*, pages 4694–4703, 2019. 7

[32] Mahdi Pakdaman Naeini, Gregory F Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *AAAI*, volume 2015, page 2901. NIH Public Access, 2015. 6, 27, 29

[33] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011. 6

[34] Chenri Ni, Nontawat Charoenphakdee, Junya Honda, and Masashi Sugiyama. On the calibration of multiclass classifi-

cation with rejection. In *NeurIPS*, pages 2586–2596, 2019. 1

[35] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *ICML*, pages 625–632, 2005. 4

[36] Marcus Nordström, Han Bao, Fredrik Löfman, Henrik Hult, Atsuto Maki, and Masashi Sugiyama. Calibrated surrogate maximization of dice. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 269–278. Springer, 2020. 1

[37] John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999. 1, 2, 5

[38] Mark D Reid and Robert C Williamson. Composite binary losses. *JMLR*, 11:2387–2422, 2010. 1, 3, 8

[39] Taissir Fekih Romdhane and Mohamed Atri Pr. Electrocardiogram heartbeat classification based on a deep convolutional neural network and focal loss. *Computers in Biology and Medicine*, 123:103866, 2020. 1

[40] Yunsheng Shi, Jun Meng, Jian Wang, Hongfei Lin, and Yumeng Li. A normalized encoder-decoder model for abstractive summarization using focal loss. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 383–392. Springer, 2018. 1

[41] Wenting Shu, Shaoyu Wang, Qiang Chen, Yun Hu, Zhengwei Cai, and Runlong Lin. Pathological image classification of breast cancer based on residual network and focal loss. In *Proceedings of the 2019 3rd International Conference on Computer Science and Artificial Intelligence*, pages 211–214, 2019. 1

[42] Emir H Shuford, Arthur Albert, and H Edward Massengill. Admissible probability measurement procedures. *Psychometrika*, 31(2):125–145, 1966. 3

[43] Xia Sun, Ke Dong, Long Ma, Richard Sutcliffe, Feijuan He, Sushing Chen, and Jun Feng. Drug-drug interaction extraction via recurrent hybrid convolutional neural networks with an improved focal loss. *Entropy*, 21(1):37, 2019. 1

[44] Ambuj Tewari and Peter L Bartlett. On the consistency of multiclass classification methods. *JMLR*, 8(May):1007–1025, 2007. 3

[45] Chao Tong, Baoyu Liang, Mengze Zhang, Rongshan Chen, Arun Kumar Sangaiah, Zhigao Zheng, Tao Wan, Chenyang Yue, and Xinyi Yang. Pulmonary nodule detection based on isodata-improved faster rcnn and 3d-cnn with focal loss. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(1s):1–9, 2020. 1

[46] Suraj Tripathi, Abhay Kumar, Abhiram Ramesh, Chirag Singh, and Promod Yenigalla. Focal loss based residual convolutional neural network for speech emotion recognition. *arXiv preprint arXiv:1906.05682*, 2019. 1

[47] Kazuki Uematsu and Yoonkyung Lee. On theoretically optimal ranking functions in bipartite ranking. *Journal of the American Statistical Association*, 112(519):1311–1322, 2017. 1

[48] Gustavo Ulloa, Alejandro Veloz, Héctor Allende-Cid, and Héctor Allende. Improving multiple sclerosis lesion boundaries segmentation by convolutional neural networks with focal learning. In *International Conference on Image Analysis and Recognition*, pages 182–192. Springer, 2020. 1

[49] Vladimir Vapnik. *Statistical learning theory. 1998*, volume 3. Wiley, New York, 1998. 2

[50] Elodie Vernet, Mark D Reid, and Robert C Williamson. Composite multiclass losses. In *NeurIPS*, pages 1224–1232, 2011. 2, 3

[51] Jayakorn Vongkulbhisal, Phongtharin Vinayavekhin, and Marco Visentini-Scarzanella. Unifying heterogeneous classifiers with distillation. In *CVPR*, pages 3175–3184, 2019. 1

[52] Robert C Williamson, Elodie Vernet, and Mark D Reid. Composite multiclass losses. *JMLR*, 17(1):7860–7911, 2016. 3, 8

[53] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. 29

[54] Guoping Xu, Hanqiang Cao, Youli Dong, Chunyi Yue, Kexin Li, and Yubing Tong. Focal loss function based deeplabv3+ for pathological lymph node segmentation on pet/ct. In *Proceedings of the 2020 2nd International Conference on Intelligent Medicine and Image Processing*, pages 24–28, 2020. 1

[55] Ming Yuan and Marten Wegkamp. Classification methods with reject option based on convex risk minimization. *JMLR*, 11:111–130, 2010. 1

[56] Tong Zhang. Statistical analysis of some multi-category large margin classification methods. *JMLR*, 5(Oct):1225–1251, 2004. 2, 3, 15