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

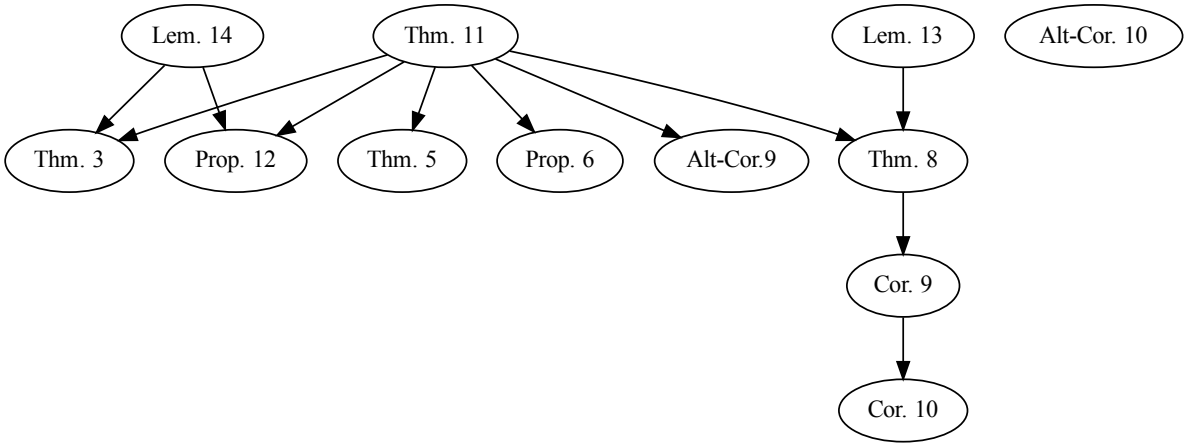

Figure 7. The dependency of the proofs. An arrow from node A to node B indicates that the result of node A is required to prove the result of node B. Alt-Cor. 9 (*resp.*, Alt-Cor. 10) denotes an alternative proof of Cor. 9 (*resp.*, Cor. 10).

A. Proofs

In this section, we provide the proofs of the results given in the main paper. To keep the notation uncluttered, sometimes we omit $\boldsymbol{x}$ and use η_i and q_i to denote the true class-posterior probability of class i and the score function of class i for the focal loss, where γ corresponds to q_i is also omitted but it can be straightforwardly inferred by the context. Figure 7 indicates the dependency of the proofs. For example, to prove Thm. 3, we can utilize the result of Lem. 14 and Thm. 11.

Proof index:

- Sec. A.1: Proof of Thm. 11: Recovering class-posterior probability from the focal loss minimizer
- Sec. A.2: Lem. 13: Properties of φ^γ
- Sec. A.3: Lem. 14: h^γ is a strictly increasing function
- Sec. A.4: Proof of Thm. 3: Focal loss is classification-calibrated
- Sec. A.5: Proof of Thm. 5: Focal loss is not strictly proper
- Sec. A.6: Proof of Prop. 6: Where risk minimizer correctly gives the true class-posterior probability
- Sec. A.7: Proof of Thm. 8: Focal loss gives under/overconfident classifier
- Sec. A.8: Proof of Cor. 9: Focal loss gives an underestimation of the true class-posterior probability
- Sec. A.9: Proof of Cor. 10: Focal loss gives underconfident classifier in binary classification
- Sec. A.10: Proof of Prop. 12: Transformation $\boldsymbol{\Psi}^\gamma$ preserves the decision rule
- Sec. A.11: Alternative proof of Cor. 9: $\boldsymbol{q}^{\gamma,*}$ is $\boldsymbol{\eta}$UC if $\frac{1}{2} \leq \max_y q_y^{\gamma,*}(\boldsymbol{x}) < 1$ and $\boldsymbol{q}^{\gamma,*}(\boldsymbol{x}) \notin \mathcal{S}^K$
- Sec. A.12: Alternative proof of Cor. 10: $\boldsymbol{\eta}$UC property for $\boldsymbol{q}^{\gamma,*}$ in binary classification where $K = 2$

A.1. Proof of Thm. 11: Recovering class-posterior probability from the focal loss minimizer

Proof. In order to derive a transformation $\boldsymbol{\Psi}^\gamma$ that recovers the true class-posterior probability $\eta_i = p(y = i|\boldsymbol{x})$ for all i from the focal loss minimizer $\boldsymbol{q}^*$[3], first consider the following optimization formulation which optimizes $W^{\ell_{\text{FL}}^\gamma}$:

$$\underset{\boldsymbol{q}}{\text{minimize}} -\sum_{i=1}^K \eta_i (1-q_i)^\gamma \log q_i \tag{14}$$

$$\text{subject to} \sum_{i=1}^K q_i = 1, \tag{15}$$

$$\boldsymbol{q} \geq \mathbf{0}_K. \tag{16}$$

Note that this optimization problem is convex with a bounded feasible set, thus an optimal solution exists. Recall that $\boldsymbol{q}^*$ denotes the minimizer of the above optimization problem. Without loss of generality, assume $\eta_i = p(y = i|\boldsymbol{x}) > 0$ for all i[4]. Observe that $\boldsymbol{q}^*$ must have $q_i^* > 0$ for all i since any $q_i^* = 0$ will make the objective goes to infinity. With this fact, we can say that the equality in (16) never holds at optimum. Therefore, by complementary slackness, the Lagrangian multipliers for constraints in (16) would be zero [4], and we can consider the following Lagrangian equation:

$$\mathcal{L}(\boldsymbol{q}, \lambda) = -\sum_{i=1}^K \eta_i (1-q_i)^\gamma \log q_i + \lambda \left(\sum_{i=1}^K q_i - 1 \right) \tag{17}$$

where λ is the Lagrangian multiplier for the equality constraint. Next, we take the derivative with respect to q_i and set to 0, then solve for λ at the optimum $\boldsymbol{q}^*$:

$$\frac{\partial}{\partial q_i} \mathcal{L}(\boldsymbol{q}, \lambda) \Big|_{\boldsymbol{q}=\boldsymbol{q}^*} = 0 = \eta_i \gamma (1-q_i^*)^{\gamma-1} \log q_i^* - \eta_i \frac{(1-q_i^*)^\gamma}{q_i^*} + \lambda \tag{18}$$

$$\lambda = \eta_i \left(\frac{(1-q_i^*)^\gamma - \gamma (1-q_i^*)^{\gamma-1} q_i^* \log q_i^*}{q_i^*} \right) \tag{19}$$

$$\eta_i = \frac{\lambda q_i^*}{(1-q_i^*)^\gamma - \gamma (1-q_i^*)^{\gamma-1} q_i^* \log q_i^*} \tag{20}$$

$$1 = \sum_{i=1}^K \eta_i = \lambda \sum_{i=1}^K \frac{q_i^*}{(1-q_i^*)^\gamma - \gamma (1-q_i^*)^{\gamma-1} q_i^* \log q_i^*} \tag{21}$$

$$\lambda = \frac{1}{\sum_{i=1}^K \frac{q_i^*}{(1-q_i^*)^\gamma - \gamma (1-q_i^*)^{\gamma-1} q_i^* \log q_i^*}}. \tag{22}$$

By replacing the above λ in (20), we can write η_i as a function of $\boldsymbol{q}^*$ as:

$$\eta_i = \frac{\frac{q_i^*}{(1-q_i^*)^\gamma - \gamma (1-q_i^*)^{\gamma-1} q_i^* \log q_i^*}}{\sum_{j=1}^K \frac{q_j^*}{(1-q_j^*)^\gamma - \gamma (1-q_j^*)^{\gamma-1} q_j^* \log q_j^*}} \tag{23}$$

$$= \frac{\frac{q_i^*}{\varphi^\gamma(q_i^*)}}{\sum_{j=1}^K \frac{q_j^*}{\varphi^\gamma(q_j^*)}} \tag{24}$$

$$= \boldsymbol{\Psi}_i^\gamma(q^*), \tag{25}$$

where $\varphi^\gamma(v) = (1-v)^\gamma - \gamma(1-v)^{\gamma-1} v \log v$ is the same function defined in (10). As a result, given $\boldsymbol{q}^{\gamma,*}$, one can recover the true class-posterior probability $\boldsymbol{\eta}$, by using the transformation $\boldsymbol{\Psi}^\gamma$. □

[3] We omit dependence on γ for brevity.

[4] If there exists a class j with $\eta_j = p(y = j|\boldsymbol{x}) = 0$, then we have $q_j^* = 0$. To see this, first let us define $\psi^\gamma(v) = -(1-v)^\gamma \log v$. We can see that $\frac{d}{dv}\psi^\gamma(v) < 0$ for $v > 0$. This means that if $q_j^* > 0$, then we can transfer q_j^* to other class k with $\eta_k > 0$, *e.g.*, $q_k^* := q_k^* + q_j^*$, then the objective in (14) would decrease, which means the original $\boldsymbol{q}^*$ is not the optimum.

A.2. Lem. 13: Properties of φ^γ

We present the following lemma, which describes the properties of the function $\varphi^\gamma : [0,1] \to \mathbb{R}$, defined as $\varphi(v) = (1-v)^\gamma - \gamma(1-v)^{\gamma-1} v \log v$, which plays a vital role in the analysis of the focal loss.

Lemma 13. *(Properties of φ^γ) The function $\varphi^\gamma : [0,1] \to \mathbb{R}$ for all $\gamma > 0$ has the following properties:*

1. *$\varphi^\gamma(0) = 1$ and $\varphi^\gamma(1) = 0$,*
2. *$\frac{d}{dv}\varphi^\gamma(v)$ changes sign from positive to negative only once at a point $\tilde{v} \in (0,1)$. In other words, there exists a unique $\tilde{v} \in (0,1)$ such that*
   (a) *$\frac{d}{dv}\varphi^\gamma(v) > 0$ for all $v < \tilde{v}$,*
   (b) *$\frac{d}{dv}\varphi^\gamma(v) = 0$ for $v = \tilde{v}$,*
   (c) *$\frac{d}{dv}\varphi^\gamma(v) < 0$ for all $v > \tilde{v}$.*
3. *There exists a unique $\acute{v} \in (0, 0.5)$ such that $\varphi^\gamma(\acute{v}) = 1$.*
4. *φ^γ has a unique maximum $\tilde{v}$, where $\tilde{v} \in (0, \acute{v})$.*

Proof. **Item 1**: We can see that

$$\varphi^\gamma(0) = (1-0)^\gamma - \gamma(1-0)^{\gamma-1} \cdot 0 \cdot \log 0 = 1 - 0 = 1, \tag{26}$$
$$\varphi^\gamma(1) = (1-1)^\gamma - \gamma(1-1)^{\gamma-1} \cdot 1 \cdot \log 1 = 0 - 0 = 0. \tag{27}$$

Item 2: To show that φ^γ changes sign from positive to negative only once in $(0,1)$, first we take its derivative and rearrange:

$$\frac{d}{dv}\varphi^\gamma(v) = -\gamma(1-v)^{\gamma-1} + \gamma(\gamma-1)(1-v)^{\gamma-2} v \log v - \gamma(1-v)^{\gamma-1} \log v - \gamma(1-v)^{\gamma-1} \tag{28}$$
$$= \gamma(1-v)^{\gamma-2}\left((v-1) + (\gamma-1)v \log v + (v-1)\log v + (v-1)\right) \tag{29}$$
$$= \underbrace{\gamma(1-v)^{\gamma-2}}_{=:t(v)\geq 0}\underbrace{(2v - 2 - \log v + \gamma v \log v)}_{=:s(v)}. \tag{30}$$

From the above, we can see that $t(v) > 0$ for $v \in (0,1)$, thus we only need to show that $s(v)$ changes sign only once in for $\frac{d}{dv}\varphi^\gamma(v)$ to also changes sign once in $(0,1)$. To see that, notice that $s(v)$ is convex since its second order derivative is always positive:

$$\frac{d}{dv}s(v) = 2 - \frac{1}{v} + \gamma \log v + \gamma, \tag{31}$$
$$\frac{d^2}{dv^2}s(v) = \frac{1}{v^2} + \frac{\gamma}{v} > 0 \text{ for all } v \in (0,1). \tag{32}$$

Also, we can compute the following:

$$s(0) = \infty, \tag{33}$$
$$s(1) = 0, \tag{34}$$
$$\frac{d}{dv}s(1) = 1 + \gamma > 0 \text{ for all } v \in (0,1). \tag{35}$$

From $s(1) = 0$ and $\frac{d}{dv}s(1) > 0$, we know that there exists $\hat{v} \in (0,1)$ such that $s(\hat{v}) < 0$. With such $\hat{v}$ and that $s(0) = \infty$, by the intermediate value theorem, there exists $\tilde{v} \in (0, \hat{v})$ such that $s(\tilde{v}) = 0$. Since $s(v)$ is convex in v, this $\tilde{v}$ is unique. Therefore, $s(v)$ changes sign only once (from positive to negative) in the range $(0,1)$ at $\tilde{v}$. As a result, $\frac{d}{dv}\varphi^\gamma(v)$ also changes sign only once (from positive to negative) at $\tilde{v} \in (0,1)$ (recall Eq. (30)). This also implies $\frac{d}{dv}\varphi^\gamma(v) = 0$ at $\tilde{v}$ and that φ^γ has a unique maximum at $\tilde{v}$.

Item 3: First, note that $\varphi^\gamma(0.5) = 1$ when $\gamma = 0$. Next, we can show that $\varphi^\gamma(0.5)$ is a decreasing function in γ:

$$\frac{d}{d\gamma}\varphi^\gamma(v)\Big|_{v=\frac{1}{2}} = \left[(1-v)^\gamma \log v - (1-v)^{\gamma-1} v \log v - \gamma(1-v)^{\gamma-1} v \log^2 v\right]_{v=\frac{1}{2}} \tag{36}$$

$$= -\gamma 0.5^\gamma \log^2 0.5 \tag{37}$$

$$\leq 0, \tag{38}$$

where the equality in Eq. (38) holds only when $\gamma = 0$. This implies that we have $\varphi^\gamma(0.5) < 1$ for all $\gamma > 0$. Since for all $\gamma > 0$, we have $\varphi^\gamma(0) = 1$ and $\frac{d}{dv}\varphi^\gamma(0) = \infty$, there exists $\check{v} \in (0, 1)$ where $\varphi^\gamma(\check{v}) > 1$. Thus, by the intermediate value theorem, there exists $\acute{v} \in (\check{v}, 0.5) \subset (0, 0.5)$ such that $\varphi^\gamma(\acute{v}) = 1$. Since $\frac{d}{dv}\varphi^\gamma(v)$ changes sign from positive to negative only once and together with the above result from the intermediate value theorem, this value $\acute{v}$ must lie on the descending side of φ^γ and thus has to be unique.

Item 4: Since from **Item 2**, we know that $\frac{d}{dv}\varphi^\gamma(v)$ changes sign from positive to negative only once at a point $\tilde{v} \in (0, 1)$, *i.e.*, $\frac{d}{dv}\varphi^\gamma(\tilde{v}) = 0$, thus φ^γ has a unique maximum at $\tilde{v}$. This fact together with that $\varphi^\gamma(0) = 1$ (**Item 1**) and that there exists $\acute{v} \in (0, 0.5)$ with $\varphi^\gamma(\acute{v}) = 1$ (**Item 3**), we conclude that the unique maximum $\tilde{v}$ of φ^γ must be in the range $(0, \acute{v})$. □

A.3. Lem. 14: h^γ is a strictly increasing function

We present the following lemma, which is highly useful for proving that the focal loss is classification-calibrated and the transformation $\mathbf{\Psi}^\gamma$ does not change the classifier's decision rule.

Lemma 14. *For any $\gamma > 0$ and $v \in (0, 1)$,*

$$h^\gamma(v) = \frac{v}{\varphi^\gamma(v)} = \frac{v}{(1-v)^\gamma - \gamma(1-v)^{\gamma-1} v \log v} \tag{39}$$

is a strictly increasing function, meaning that $h^\gamma(u) > h^\gamma(v)$ if and only if $u > v$.

Proof. Since h^γ is differentiable, it suffices to prove that h^γ is strictly increasing if $\frac{d}{dv}h^\gamma(v) > 0$. By taking the derivative of $h^\gamma(v)$, we have

$$\frac{d}{dv}h^\gamma(v) = \frac{1}{(1-v)^\gamma - \gamma(1-v)^{\gamma-1} v \log v} - \frac{-2\gamma v(1-v)^{\gamma-1} + (\gamma-1)\gamma v^2(1-v)^{\gamma-2}\log v - \gamma(1-v)^{\gamma-1} v \log v}{[(1-v)^\gamma - \gamma(1-v)^{\gamma-1} v \log v]^2} \tag{40}$$

$$= \frac{(1-v)^\gamma - \gamma(1-v)^{\gamma-1} v \log v + 2\gamma v(1-v)^{\gamma-1} - (\gamma-1)\gamma v^2(1-v)^{\gamma-2}\log v + \gamma(1-v)^{\gamma-1} v \log v}{[(1-v)^\gamma - \gamma(1-v)^{\gamma-1} v \log v]^2} \tag{41}$$

$$= \frac{(1-v)^\gamma + 2\gamma v(1-v)^{\gamma-1} - (\gamma-1)\gamma v^2(1-v)^{\gamma-2}\log v}{[(1-v)^\gamma - \gamma(1-v)^{\gamma-1} v \log v]^2}. \tag{42}$$

Since the goal is to show that $\frac{d}{dv}h^\gamma(v) > 0$, the denominator of $\frac{d}{dv}h^\gamma(v)$ can be ignored because $[(1-v)^\gamma - \gamma(1-v)^{\gamma-1} v \log v]^2 > 0$ for $v \in (0, 1)$. We denote ϕ_1^γ the numerator of $\frac{d}{dv}h^\gamma$ as follows:

$$\phi_1^\gamma(v) = (1-v)^\gamma + 2\gamma v(1-v)^{\gamma-1} - (\gamma-1)\gamma v^2(1-v)^{\gamma-2}\log v. \tag{43}$$

Now it suffices to show that $\phi_1^\gamma(v) > 0$ for all $\gamma > 0$ to prove that $\frac{d}{dv}h^\gamma(v) > 0$. We split the proof into two cases, which are the case where $\gamma \geq 1$ and $0 < \gamma < 1$.

Case 1: $\gamma \geq 1$.

It is straightforward to see that $\phi_1^\gamma(v) > 0$ because $(1-v)^\gamma > 0$, $2\gamma v(1-v)^{\gamma-1} > 0$ and $-(\gamma-1)\gamma v^2(1-v)^{\gamma-2}\log v \geq 0$ for $v \in (0, 1)$. This is because of the sum of two positive quantities and one nonnegative quantity must be positive.

Case 2: $0 \leq \gamma < 1$.

We begin by expressing $\phi_1^\gamma(v)$ as follows:

$$\phi_1^\gamma(v) = (1-v)^\gamma + 2\gamma v(1-v)^{\gamma-1} - (\gamma-1)\gamma v^2(1-v)^{\gamma-2}\log v \tag{44}$$

$$= (1-v)^{\gamma-2}[(1-v)^2 + 2\gamma v(1-v) - (\gamma-1)\gamma v^2 \log v] \tag{45}$$

$$= (1-v)^{\gamma-2}[(1-v)^2 + 2\gamma v(1-v) - \gamma^2 v^2 \log v + \gamma v^2 \log v]. \tag{46}$$

Since we want to show that $\phi_1^\gamma(v) > 0$, we can ignore a positive value $(1-v)^{\gamma-2}$ and prove that

$$(1-v)^2 + 2\gamma v(1-v) - \gamma^2 v^2 \log v + \gamma v^2 \log v > 0. \tag{47}$$

Since $-\gamma^2 v^2 \log v \geq 0$, we may ignore this term and it is sufficient to prove that

$$\phi_2^\gamma(v) = (1-v)^2 + 2\gamma v(1-v) + \gamma v^2 \log v > 0. \tag{48}$$

By substitution, we have

$$\phi_2^\gamma(0) = 1, \tag{49}$$
$$\phi_2^\gamma(1) = 0. \tag{50}$$

Then, we show that ϕ_2^γ is a decreasing function by showing that $\frac{d}{d_v}\phi_2^\gamma(v) < 0$.

The derivative of $\phi_2^\gamma(v)$ can be expressed as:

$$\frac{d}{d_v}\phi_2^\gamma(v) = \frac{d}{d_v}(1-v)^2 + \frac{d}{d_v}2\gamma v(1-v) + \frac{d}{d_v}\gamma v^2 \log v \tag{51}$$
$$= -2(1-v) + 2\gamma - 4\gamma v + \gamma v + 2\gamma v \log v \tag{52}$$
$$= 2v - 2 + 2\gamma - 3\gamma v + 2\gamma v \log v \tag{53}$$

By substitution, we have

$$\frac{d}{d_v}\phi_2^\gamma(0) = -2 + 2\gamma < 0, \tag{54}$$
$$\frac{d}{d_v}\phi_2^\gamma(1) = -\gamma < 0. \tag{55}$$

Moreover, $\frac{d}{d_v}\phi_2^\gamma$ is convex because

$$\frac{d^2}{dv^2}\phi_2^\gamma(v) = 2 - 3\gamma + 2\gamma[1 + \log v], \tag{56}$$
$$\frac{d^3}{dv^3}\phi_2^\gamma(v) = \frac{2\gamma}{v} > 0. \tag{57}$$

Based on the fact that $\frac{d}{d_v}\phi_2^\gamma(0) < 0$, $\frac{d}{d_v}\phi_2^\gamma(1) < 0$, and $\frac{d}{d_v}\phi_2^\gamma$ is convex, we can conclude that $\frac{d}{d_v}\phi_2^\gamma(v) < 0$ for $v \in (0,1)$ because it must be less than $\max(\frac{d}{d_v}\phi_2^\gamma(0), \frac{d}{d_v}\phi_2^\gamma(1))$ [4]. Therefore, $\frac{d}{d_v}\phi_2^\gamma(v) < 0$ and thus ϕ_2^γ is a decreasing function.

Next, because $\phi_2^\gamma(0) = 1$, $\phi_2^\gamma(1) = 0$, and ϕ_2^γ is a decreasing function, we know that $\phi_2^\gamma(v) > 0$ for $v \in (0,1)$, which proves that $\phi_1^\gamma(v) > 0$ for $0 < \gamma < 1$.

By combining the results of **Case 1** and **Case 2**, we have $\phi_1^\gamma(v) > 0$ for all $\gamma > 0$, which yields $\frac{d}{dv}h^\gamma(v) > 0$. Therefore, h^γ is a strictly increasing function. □

A.4. Proof of Thm. 3: Focal loss is classification-calibrated

Proof. To prove that the focal loss is classification-calibrated, we combine the result of Thm. 11 and the the existing result which suggests that a surrogate loss is classification-calibrated if it has satisfies the strictly order-preserving property [56].

The order-preserving property suggests that for any $\boldsymbol{x}$, the pointwise conditional risk W^ℓ has the risk minimizer $\boldsymbol{q}^{\ell,*}(\boldsymbol{x})$ such that $\eta_i(\boldsymbol{x}) < \eta_j(\boldsymbol{x}) \Rightarrow q_i^{\ell,*}(\boldsymbol{x}) < q_j^{\ell,*}(\boldsymbol{x})$, then a loss function ℓ is classification-calibrated [56].

From Thm. 11, we know that

$$\boldsymbol{\eta}(\boldsymbol{x}) = \boldsymbol{\Psi}^\gamma(\boldsymbol{q}^{\gamma,*}(\boldsymbol{x})), \tag{58}$$

where

$$\boldsymbol{\Psi}^\gamma(\boldsymbol{v}) = [\Psi_1^\gamma(\boldsymbol{v}), \ldots, \Psi_K^\gamma(\boldsymbol{v})]^\top, \tag{59}$$
$$\Psi_i^\gamma(\boldsymbol{v}) = \frac{h^\gamma(v_i)}{\sum_{l=1}^K h^\gamma(v_l)}, \tag{60}$$
$$h^\gamma(v) = \frac{v}{\varphi^\gamma(v)} = \frac{v}{(1-v)^\gamma - \gamma(1-v)^{\gamma-1} v \log v}. \tag{61}$$

From Lem. 14, we know that h^γ is a strictly increasing function. Thus, we have

$$h^\gamma(q_i^{\gamma,*}(\boldsymbol{x})) < h^\gamma(q_j^{\gamma,*}(\boldsymbol{x})) \Rightarrow q_i^{\gamma,*}(\boldsymbol{x}) < q_j^{\gamma,*}(\boldsymbol{x}). \tag{62}$$

Given $\boldsymbol{x}$, the denominator of $\Psi_i^\gamma(\boldsymbol{q}^{\gamma,*}(\boldsymbol{x}))$, i.e., $\sum_{l=1}^K q_i^{\gamma,*}(\boldsymbol{x})$ is identical for all classes. Also, the numerator of $\Psi_i^\gamma(\boldsymbol{q}^{\gamma,*}(\boldsymbol{x}))$ is a strictly increasing function $h^\gamma(q_i^{\gamma,*}(\boldsymbol{x}))$. Based on these facts, we have

$$\Psi_i^\gamma(\boldsymbol{q}^{\gamma,*}(\boldsymbol{x})) < \Psi_j^\gamma(\boldsymbol{q}^{\gamma,*}(\boldsymbol{x})) \Rightarrow q_i^{\gamma,*}(\boldsymbol{x}) < q_j^{\gamma,*}(\boldsymbol{x}). \tag{63}$$

Since $\boldsymbol{\Psi}^\gamma(\boldsymbol{q}^{\gamma,*})$ is equal to $\boldsymbol{\eta}(\boldsymbol{x})$ and note that $\Psi_i^\gamma(\boldsymbol{q}^{\gamma,*}(\boldsymbol{x})) = \eta_i(\boldsymbol{x})$, we have

$$\eta_i(\boldsymbol{x}) < \eta_j(\boldsymbol{x}) \Rightarrow q_i^{\gamma,*}(\boldsymbol{x}) < q_j^{\gamma,*}(\boldsymbol{x}). \tag{64}$$

Eq. (64) indicates that the focal loss satisfies the strictly order-preserving property for all $\gamma \geq 0$, which is sufficient to conclude that that the focal loss is classification-calibrated. □

Note that $\arg\max_y q_y^{\gamma,*}(\boldsymbol{x}) = \arg\max_y \eta_y(\boldsymbol{x})$ indicates that the decision rule of the focal risk minimizer is equivalent to that of the Bayes-optimal classifier. As a result, the Bayes-optimal classifier can be achieved by minimizing the focal risk, *i.e.*, $R^{\ell_{0\text{-}1}}(f^{\boldsymbol{q}^{\gamma,*}}) = R^{\ell_{0\text{-}1}}(f^{\ell_{0\text{-}1},*})$.

A.5. Proof of Thm. 5: Focal loss is not strictly proper

Proof. Recall 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 $\boldsymbol{u} = \boldsymbol{v}$ by the definition of strict properness. We will prove that the focal loss is not strictly proper for all $\gamma > 0$ by showing a counterexample that the focal loss can be minimized when $\boldsymbol{u} \neq \boldsymbol{v}$.

By the definition of the focal loss:

$$\ell_{\mathrm{FL}}^\gamma(\boldsymbol{u}, \boldsymbol{v}) = -\sum_{i=1}^K v_i(1-u_i)^\gamma \log(u_i). \tag{65}$$

For any $\boldsymbol{x}$, we have

$$\ell_{\mathrm{FL}}^\gamma(\boldsymbol{q}(\boldsymbol{x}), \boldsymbol{\eta}(\boldsymbol{x})) = -\sum_{y=1}^K \eta_y(\boldsymbol{x})(1-q_y(\boldsymbol{x}))^\gamma \log(q_y(\boldsymbol{x})) \tag{66}$$

$$= \sum_{y\in\mathcal{Y}} \eta_y(\boldsymbol{x}) \ell_{\mathrm{FL}}^\gamma\big(\boldsymbol{q}(\boldsymbol{x}), \boldsymbol{e}_y\big) \tag{67}$$

$$= W^{\ell_{\mathrm{FL}}^\gamma}\big(\boldsymbol{q}(\boldsymbol{x}); \boldsymbol{\eta}(\boldsymbol{x})\big). \tag{68}$$

It can be observed that $\ell_{\mathrm{FL}}^\gamma(\boldsymbol{q}(\boldsymbol{x}), \boldsymbol{\eta}(\boldsymbol{x}))$ coincides with the pointwise conditional risk w.r.t. the focal loss $W^{\ell_{\mathrm{FL}}^\gamma}\big(\boldsymbol{q}(\boldsymbol{x}); \boldsymbol{\eta}(\boldsymbol{x})\big)$. Note that the simplex $\boldsymbol{q}$ that minimizes $W^{\ell_{\mathrm{FL}}^\gamma}$ is the focal risk minimizer $\boldsymbol{q}^{\gamma,*}$. Based on Thm. 11, we know that although $\boldsymbol{q}^{\gamma,*}(\boldsymbol{x})$ minimizes $\ell_{\mathrm{FL}}^\gamma(\boldsymbol{q}(\boldsymbol{x}), \boldsymbol{\eta}(\boldsymbol{x}))$, we have $\boldsymbol{q}^{\gamma,*} \neq \boldsymbol{\eta}$ because $\boldsymbol{\Psi}^\gamma$ that transforms $\boldsymbol{q}^{\gamma,*}$ to the true class-posterior probability is not an identity function unless $\gamma \neq 0$. This counterexample is sufficient to conclude that the focal loss is not strictly proper for $\gamma > 0$ since the focal loss $\ell_{\mathrm{FL}}^\gamma(\boldsymbol{u}, \boldsymbol{v})$ can be minimized when $\boldsymbol{u} \neq \boldsymbol{v}$, which contradicts the definition of strict properness. □

A.6. Proof of Prop. 6: Where risk minimizer correctly gives the true class-posterior probability

Proof. Recall $\mathcal{S}^K = \{\boldsymbol{v} \in \Delta^K : v_i \in \{0, \max_j v_j\}\}$. From Thm. 11, we know that $\boldsymbol{\Psi}^\gamma(\boldsymbol{v}) = [\Psi_1^\gamma(\boldsymbol{v}), \ldots, \Psi_K^\gamma(\boldsymbol{v})]^\top$, $\Psi_i^\gamma(\boldsymbol{v}) = \frac{h^\gamma(v_i)}{\sum_{l=1}^K h^\gamma(v_l)}$, $h^\gamma(v) = \frac{v}{\varphi^\gamma(v)} = \frac{v}{(1-v)^\gamma - \gamma(1-v)^{\gamma-1} v \log v}$. We will show that If $\boldsymbol{q}^{\gamma,*}(\boldsymbol{x}) \in \mathcal{S}^K$, then $\boldsymbol{q}^{\gamma,*}(\boldsymbol{x}) = \boldsymbol{\eta}(\boldsymbol{x})$ by proving that $\Psi_i^\gamma(\boldsymbol{q}^{\gamma,*}(\boldsymbol{x})) = \eta_i(\boldsymbol{x}) = q_i^{\gamma,*}(\boldsymbol{x})$ for all $i \in \mathcal{Y}$.

Case 1: $\Psi_i^\gamma(\boldsymbol{q}^{\gamma,*}(\boldsymbol{x})) = q_i^{\gamma,*}(\boldsymbol{x}) = 0$.

Since

$$h^\gamma(q_i^{\gamma,*}(\boldsymbol{x})) = \frac{q_i^{\gamma,*}(\boldsymbol{x})}{\varphi^\gamma(q_i^{\gamma,*}(\boldsymbol{x}))} = \frac{0}{\varphi^\gamma(0)} = \frac{0}{1} = 0, \tag{69}$$

we have

$$\Psi_i^\gamma(\boldsymbol{q}^{\gamma,*}(\boldsymbol{x})) = \frac{h^\gamma(q_i^{\gamma,*}(\boldsymbol{x}))}{\sum_{l=1}^K h^\gamma(q_l^{\gamma,*}(\boldsymbol{x}))} \tag{70}$$

$$= \frac{h^\gamma(0)}{\sum_{l=1}^K h^\gamma(q_l^{\gamma,*}(\boldsymbol{x}))} \tag{71}$$

$$= 0 \tag{72}$$

$$= q_i^{\gamma,*}(\boldsymbol{x}). \tag{73}$$

Case 2: $\Psi_i^\gamma(\boldsymbol{q}^{\gamma,*}(\boldsymbol{x})) = q_i^{\gamma,*}(\boldsymbol{x}) = \max_j q_j^{\gamma,*}(\boldsymbol{x})$.

Since other q_y where $y \neq i$ can be only either $q_y = q_i = \max_j v_j$ or $q_y = 0$. Let $k \leq K$ be a number of classes that is non-zero. Thus, we have $\max_j q_j^{\gamma,*}(\boldsymbol{x}) = \frac{1}{k}$. It can be observed that

$$\sum_{l=1}^K h^\gamma(q_l^{\gamma,*}(\boldsymbol{x})) = kh^\gamma(\frac{1}{k}) + (K-k)h^\gamma(0) \tag{74}$$

$$= kh^\gamma(\frac{1}{k}) + (K-k)0 \tag{75}$$

$$= kh^\gamma(\frac{1}{k}). \tag{76}$$

Therefore, we have

$$\Psi_i^\gamma(\boldsymbol{q}^{\gamma,*}(\boldsymbol{x})) = \frac{h^\gamma(\frac{1}{k})}{\sum_{l=1}^K h^\gamma(q_l^{\gamma,*}(\boldsymbol{x}))} \tag{77}$$

$$= \frac{h^\gamma(\frac{1}{k})}{kh^\gamma(\frac{1}{k})} \tag{78}$$

$$= \frac{1}{k} \tag{79}$$

$$= \max_j q_j^{\gamma,*}(\boldsymbol{x}) \tag{80}$$

$$= q_i^{\gamma,*}(\boldsymbol{x}). \tag{81}$$

We can focus on a uniform vector over a subset of classes and zeros otherwise.

By combining **Case 1** and **Case 2**, we can conclude that if $\boldsymbol{q}^{\gamma,*}(\boldsymbol{x}) \in \mathcal{S}^K$, we have $\boldsymbol{q}^{\gamma,*}(\boldsymbol{x}) = \boldsymbol{\eta}(\boldsymbol{x})$.

□

A.7. Proof of Thm. 8: Focal loss gives under/overconfident classifier

Proof. 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 $\max_y q_y^{\gamma,*}(\boldsymbol{x}) \neq \frac{1}{K}$, we have

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

Item 1: $0 < \tau_{\mathrm{oc}}^\gamma < \tau_{\mathrm{uc}}^\gamma < 0.5$.

For any $\gamma > 0$, from Lem. 13 tells us that $\varphi^\gamma(0) = 1$, there exists a unique $\acute{v} \in (0, 0.5)$ such that $\varphi^\gamma(\acute{v}) = 1$, and $\frac{d}{dv}\varphi^\gamma(v) < 0$ for all $v > \tilde{v}$. φ^γ also has a unique maximum in the range $(0, \acute{v})$. We will use these facts prove that $0 < \tau_{\mathrm{oc}}^\gamma < \tau_{\mathrm{uc}}^\gamma < 0.5$ by showing that $\tau_{\mathrm{uc}}^\gamma < 0.5$ and $0 < \tau_{\mathrm{oc}}^\gamma < \tau_{\mathrm{uc}}^\gamma$.

Case 1: $\tau_{\mathrm{uc}}^\gamma < 0.5$.

From the definition of $\tau_{\mathrm{uc}}^\gamma$, we can conclude that $\tau_{\mathrm{uc}}^\gamma = \acute{v} < 0.5$ since $\acute{v} \in (0, 0.5)$. Note that there does not exist $v \in [0.5, 1)$ such that $\varphi^\gamma(v) = 1$ because $\varphi^\gamma(0.5) < 1$ for all $\gamma > 0$. This is due to the fact that $\frac{d}{dv}\varphi^\gamma(v) < 0$ for all $v > \tilde{v}$, i.e., φ^γ is a decreasing function when $v > \acute{v}$.

Case 2: $0 < \tau_{\text{oc}}^{\gamma} < \tau_{\text{uc}}^{\gamma}$.

From the definition of $\tau_{\text{oc}}^{\gamma}$, it is a unique maximum of φ^{γ} and therefore $\tau_{\text{oc}}^{\gamma} \in (0, \dot{v})$, which is equivalent to $\tau_{\text{oc}}^{\gamma} \in (0, \tau_{\text{uc}}^{\gamma})$. This implies that $0 < \tau_{\text{oc}}^{\gamma} < \tau_{\text{uc}}^{\gamma}$.

By combining **Case 1** and **Case 2**, we can conclude that $0 < \tau_{\text{oc}}^{\gamma} < \tau_{\text{uc}}^{\gamma} < 0.5$.

Item 2: $\boldsymbol{q}^{\gamma,*}$ is $\boldsymbol{\eta}$OC if $\max_y q_y^{\gamma,*}(\boldsymbol{x}) \in (\frac{1}{K}, \tau_{\text{oc}}^{\gamma}]$. Recall the definition of $\boldsymbol{\eta}$OC:

$$\max_y q_y^{\ell,*}(\boldsymbol{x}) - \max_y \eta_y(\boldsymbol{x}) > 0. \tag{82}$$

Without loss of generality let us define $i \in \arg\max_y q_y^{\ell,*}(\boldsymbol{x})$. Since the focal loss is classification-calibrated suggested in Thm. 3, the max-index of $q^{\ell,*}$ and $\boldsymbol{\eta}$ are identical.

From Thm. 11, we can rewrite $\eta_i(\boldsymbol{x})$ as

$$\begin{aligned}
\eta_i(\boldsymbol{x}) &= \Psi_i^{\gamma}(\boldsymbol{q}^{\gamma,*}(\boldsymbol{x})) && (83)\\
&= \Psi_i^{\gamma}(\boldsymbol{q}^{\gamma,*}(\boldsymbol{x})) && (84)\\
&= \frac{h^{\gamma}(q_i^{\gamma,*}(\boldsymbol{x}))}{\sum_{l=1}^{K} h^{\gamma}(q_l^{\gamma,*}(\boldsymbol{x}))}. && (85)
\end{aligned}$$

Note that

$$h^{\gamma}(v) = \frac{v}{\varphi^{\gamma}(v)} = \frac{v}{(1-v)^{\gamma} - \gamma(1-v)^{\gamma-1} v \log v}. \tag{86}$$

By the definition of $\boldsymbol{\eta}$OC, we have

$$\begin{aligned}
q_i^{\gamma,*}(\boldsymbol{x}) - \frac{h^{\gamma}(q_i^{\gamma,*}(\boldsymbol{x}))}{\sum_{l=1}^{K} h^{\gamma}(q_l^{\gamma,*}(\boldsymbol{x}))} > 0 && (87)\\
q_i^{\gamma,*}(\boldsymbol{x}) > \frac{h^{\gamma}(q_i^{\gamma,*}(\boldsymbol{x}))}{\sum_{l=1}^{K} h^{\gamma}(q_l^{\gamma,*}(\boldsymbol{x}))} && (88)\\
= \frac{\frac{q_i^{\gamma,*}(\boldsymbol{x})}{\varphi^{\gamma}(q_i^{\gamma,*}(\boldsymbol{x}))}}{\sum_{l=1}^{K} h^{\gamma}(q_l^{\gamma,*}(\boldsymbol{x}))} && (89)\\
&&& (90)
\end{aligned}$$

By dividing both sides by $q_i^{\gamma,*}(\boldsymbol{x})$, we have.

$$\begin{aligned}
1 > \frac{\frac{1}{\varphi^{\gamma}(q_i^{\gamma,*}(\boldsymbol{x}))}}{\sum_{l=1}^{K} h^{\gamma}(q_l^{\gamma,*}(\boldsymbol{x}))} && (91)\\
\sum_{l=1}^{K} h^{\gamma}(q_l^{\gamma,*}(\boldsymbol{x})) > \frac{1}{\varphi^{\gamma}(q_i^{\gamma,*}(\boldsymbol{x}))} && (92)
\end{aligned}$$

By the definition of h^{γ}, we have

$$\sum_{l=1}^{K} q_l^{\gamma,*}(\boldsymbol{x}) \frac{1}{\varphi^{\gamma}(q_l^{\gamma,*}(\boldsymbol{x}))} > \frac{1}{\varphi^{\gamma}(q_i^{\gamma,*}(\boldsymbol{x}))} \tag{93}$$

Next we will prove that Ineq. (93) is true to verify that $\boldsymbol{q}^{\gamma,*}(\boldsymbol{x})$ is $\boldsymbol{\eta}$OC.

Because the convex combination is not smaller the minimum value [4], to prove that Ineq. (93) holds, it suffices to show that for all $l \in \mathcal{Y}$:

$$\frac{1}{\varphi^{\gamma}(q_l^{\gamma,*}(\boldsymbol{x}))} \geq \frac{1}{\varphi^{\gamma}(q_i^{\gamma,*}(\boldsymbol{x}))}, \tag{94}$$

which is equivalent to

$$\varphi^{\gamma}(q_i^{\gamma,*}(\boldsymbol{x})) \geq \varphi^{\gamma}(q_l^{\gamma,*}(\boldsymbol{x})), \tag{95}$$

and there exists at least one l such that the strict inequality holds, i.e., $\varphi^{\gamma}(q_i^{\gamma,*}(\boldsymbol{x})) > \varphi^{\gamma}(q_l^{\gamma,*}(\boldsymbol{x}))$. Since $\boldsymbol{q}^{\gamma,*}(\boldsymbol{x}) \notin \mathcal{S}^K$, there exists l such that $q_l^{\gamma,*}(\boldsymbol{x})) \neq q_i^{\gamma,*}(\boldsymbol{x}))$ and $q_l^{\gamma,*}(\boldsymbol{x}))$ is non-zero. From Lem. 13, we know that $\frac{d}{dv}\varphi^{\gamma}(v) > 0$ for all $v \in (0, \tau_{\text{oc}}^{\gamma})$. Therefore φ^{γ} is an increasing function in $(0, \tau_{\text{oc}}^{\gamma}]$. Since $q_l^{\gamma,*}(\boldsymbol{x}) \leq q_i^{\gamma,*}(\boldsymbol{x})$, for $q_i^{\gamma,*}(\boldsymbol{x}) \in (\frac{1}{K}, \tau_{\text{oc}}^{\gamma}]$, we have

$$\varphi^{\gamma}(q_i^{\gamma,*}(\boldsymbol{x})) \geq \varphi^{\gamma}(q_l^{\gamma,*}(\boldsymbol{x})) \tag{96}$$

for all $l \in \mathcal{Y}$, where the equality holds only if $q_i^{\gamma,*}(\boldsymbol{x}) = q_j^{\gamma,*}(\boldsymbol{x})$. Thus, we have proven that Ineq.(93) holds, which indicates that $\boldsymbol{q}^{\gamma,*}$ is $\boldsymbol{\eta}$OC if $\max_y q_y^{\gamma,*}(\boldsymbol{x}) \in (\frac{1}{K}, \tau_{\text{oc}}^{\gamma}]$

Item 3: $\boldsymbol{q}^{\gamma,*}$ is $\boldsymbol{\eta}$UC if $\max_y q_y^{\gamma,*}(\boldsymbol{x}) \in [\tau_{\text{uc}}^{\gamma}, 1)$. Recall the definition of $\boldsymbol{\eta}$UC:

$$\max_y q_y^{\ell,*}(\boldsymbol{x}) - \max_y \eta_y(\boldsymbol{x}) < 0. \tag{97}$$

By using exactly the same technique for proving **Item 2** but we flip the sign to validate $\boldsymbol{\eta}$-underconfidence. We know that it suffices to prove that for all $l \in \mathcal{Y}$,

$$\varphi^{\gamma}(q_i^{\gamma,*}(\boldsymbol{x})) \leq \varphi^{\gamma}(q_l^{\gamma,*}(\boldsymbol{x})) \tag{98}$$

and there exists at least one l such that the strict inequality holds, i.e., $\varphi^{\gamma}(q_i^{\gamma,*}(\boldsymbol{x})) < \varphi^{\gamma}(q_l^{\gamma,*}(\boldsymbol{x}))$.

Recall Lem. 13 indicates that $\frac{d}{dv}\varphi^{\gamma}(v) > 0$ for all $v \in (0, \tau_{\text{oc}}^{\gamma})$. Thus, φ^{γ} is an increasing function in $(0, \tau_{\text{oc}}^{\gamma})$. Moreover, $\frac{d}{dv}\varphi^{\gamma}(v) < 0$ for all $v \in (\tau_{\text{oc}}^{\gamma}, 1)$. Thus, φ^{γ} is a decreasing function in $(\tau_{\text{oc}}^{\gamma}, 1)$. Also, we know that $\varphi^{\gamma}(0) = \varphi^{\gamma}(\tau_{\text{uc}}^{\gamma}) = 1$. Furthermore, Thm. 7 indicates that $\tau_{\text{oc}}^{\gamma} < \tau_{\text{uc}}^{\gamma}$. We will use these facts to prove our result.

Since $\varphi^{\gamma}(0) = \varphi^{\gamma}(\tau_{\text{uc}}^{\gamma}) = 1$, φ^{γ} is increasing in $(0, \tau_{\text{oc}}^{\gamma})$, and decreasing in $(\tau_{\text{oc}}^{\gamma}, 1)$, we have $\varphi^{\gamma}(v) > 1$ for $v \in (0, \tau_{\text{uc}}^{\gamma})$. Note that $q_l^{\gamma,*}(\boldsymbol{x}) \leq q_i^{\gamma,*}(\boldsymbol{x})$, for $q_i^{\gamma,*}(\boldsymbol{x})$. We know that $\boldsymbol{q}^{\gamma,*}(\boldsymbol{x}) \notin \mathcal{S}^K$, which implies that there exists l such that $q_l^{\gamma,*}(\boldsymbol{x})) \neq q_i^{\gamma,*}(\boldsymbol{x}))$ and $q_l^{\gamma,*}(\boldsymbol{x}))$ is non-zero. For such $q_l^{\gamma,*}(\boldsymbol{x})$, we have

$$\varphi^{\gamma}(q_i^{\gamma,*}(\boldsymbol{x})) < \varphi^{\gamma}(q_l^{\gamma,*}(\boldsymbol{x})). \tag{99}$$

As a result, for $q_i^{\gamma,*}(\boldsymbol{x}) \in [\tau_{\text{uc}}^{\gamma}, 1)$, we have

$$\varphi^{\gamma}(q_i^{\gamma,*}(\boldsymbol{x})) \leq \varphi^{\gamma}(q_l^{\gamma,*}(\boldsymbol{x})) \tag{100}$$

for all $l \in \mathcal{Y}$, and there are only two possibilities that the equality holds:

1. $q_i^{\gamma,*}(\boldsymbol{x}) = q_j^{\gamma,*}(\boldsymbol{x})$
2. $q_i^{\gamma,*}(\boldsymbol{x}) = \tau_{\text{uc}}^{\gamma}$ and $q_j^{\gamma,*}(\boldsymbol{x}) = 0$.

Note that there exists $q_l^{\gamma,*}(\boldsymbol{x}))$ such that $\varphi^{\gamma}(q_i^{\gamma,*}(\boldsymbol{x})) < \varphi^{\gamma}(q_l^{\gamma,*}(\boldsymbol{x}))$ since $\boldsymbol{q}^{\gamma,*}(\boldsymbol{x}) \notin \mathcal{S}^K$. Thus, we can conclude that $\boldsymbol{q}^{\gamma,*}$ is $\boldsymbol{\eta}$UC if $\max_y q_y^{\gamma,*}(\boldsymbol{x}) \in [\tau_{\text{uc}}^{\gamma}, 1)$.

□

A.8. Proof of Cor. 9: Focal loss gives an underestimation of the true class-posterior probability

Proof. From Thm. 8, we know that for all $\gamma > 0$, we have $0 \leq \tau_{\text{oc}}^{\gamma} < \tau_{\text{uc}}^{\gamma} < 0.5$. Moreover, Thm. 8 also tells us that $\boldsymbol{q}^{\gamma,*}$ is $\boldsymbol{\eta}$UC if $\max_y q_y^{\gamma,*}(\boldsymbol{x}) \in [\tau_{\text{uc}}^{\gamma}, 1)$. Since $\tau_{\text{uc}}^{\gamma} < 0.5$, it is straightforward to see that $(0.5, 1) \subset [\tau_{\text{uc}}^{\gamma}, 1)$ and therefore $\boldsymbol{q}^{\gamma,*}$ is $\boldsymbol{\eta}$UC if $\max_y q_y^{\gamma,*}(\boldsymbol{x}) \in (0.5, 1)$. □

A.9. Proof of Cor. 10: Focal loss gives underconfident classifier in binary classification

Proof. We know that $\max_y \boldsymbol{q}^{\gamma,*}(\boldsymbol{x}) \geq 0.5$ in binary classification since $\max_y \boldsymbol{q}^{\gamma,*}(\boldsymbol{x}) \geq \frac{1}{K}$, and $K = 2$. As a result, unless the label distribution is uniform, i.e., $\max_y \eta_y(\boldsymbol{x}) = 0.5$ or the label distribution is deterministic, i.e., $\max_y \eta_y(\boldsymbol{x}) = 1$, $\boldsymbol{q}^{\gamma,*}$ must always be $\boldsymbol{\eta}$UC since $\max_y \eta_y(\boldsymbol{x}) \in (0.5, 1)$ as proven in Cor. 9. □

A.10. Proof of Prop. 12: Transformation Ψ^γ preserves the decision rule

Proof. From Thm. 11, we know that $\boldsymbol{\Psi}^\gamma(\boldsymbol{v}) = [\Psi_1^\gamma(\boldsymbol{v}), \dots, \Psi_K^\gamma(\boldsymbol{v})]^\top$, $\Psi_i^\gamma(\boldsymbol{v}) = \frac{h^\gamma(v_i)}{\sum_{l=1}^K h^\gamma(v_l)}$, and $h^\gamma(v) = \frac{v}{\varphi^\gamma(v)}$. Note that the denominator of $\Psi_i^\gamma(\boldsymbol{v}$, i.e., $\sum_{l=1}^K v_i(\boldsymbol{v})$ is identical for all classes. Thus, it suffices to look at the numerator to determine which index has the largest value. It can be observed that the numerator $\Psi_i^\gamma(\boldsymbol{q}^{\gamma,*}(\boldsymbol{x}))$ is a strictly increasing function $h^\gamma(q_i^{\gamma,*}(\boldsymbol{x}))$, as proven in Lem. 14. Based on these facts, we have

$$\arg\max_i \ \Psi_i^\gamma(\boldsymbol{v}) = \arg\max_i v_i. \tag{101}$$

Note that this holds for any simplex input, not only the output of the focal risk minimizer $\boldsymbol{q}^{\gamma,*}$. □

A.11. Alternative proof of Cor. 9: $\boldsymbol{q}^{\gamma,*}$ is $\boldsymbol{\eta}$UC if $\frac{1}{2} \leq \max_y q_y^{\gamma,*}(\boldsymbol{x}) < 1$ and $\boldsymbol{q}^{\gamma,*}(\boldsymbol{x}) \notin \mathcal{S}^K$

Here, we present an alternative proof of Cor. 9, which is independent from Thm. 8, Lem. 13 and Lem. 14.

Proof. Let $q = \max_y q_y^{\gamma,*}(\boldsymbol{x})$ be the highest score in the simplex output from the focal risk minimizer $\boldsymbol{q}^{\gamma,*}$ and $\eta = \max_y \eta_y(\boldsymbol{x})$ be the true class-posterior probability of the most probable class.

In the multiclass cases, we show that $\boldsymbol{q}^{\gamma,*}$ is $\boldsymbol{\eta}$UC when $\frac{1}{2} \leq q < 1$. That is, for $q \in [\frac{1}{2}, 1)$, we have

$$\eta - q > 0. \tag{102}$$

From Thm. 11, we can replace η and rewrite the above inequality as

$$\frac{\frac{q}{(1-q)^\gamma - \gamma(1-q)^{\gamma-1} q \log q}}{\sum_{l=1}^K \frac{q_l}{(1-q_l)^\gamma - \gamma(1-q_l)^{\gamma-1} q_l \log q_l}} - q > 0 \tag{103}$$

$$\frac{q}{(1-q)^\gamma - \gamma(1-q)^{\gamma-1} q \log q} > q \sum_{l=1}^K \frac{q_l}{(1-q_l)^\gamma - \gamma(1-q_l)^{\gamma-1} q_l \log q_l} \tag{104}$$

$$\frac{1}{\varphi(q)} > \sum_{l=1}^K \frac{q_l}{\varphi(q_l)}, \tag{105}$$

where we $\varphi(q) = (1-q)^\gamma - \gamma(1-q)^{\gamma-1} q \log q$. Note that since $q_l < 1$, we have $\varphi(q_l) > 0$. In words, Ineq. (105) says it suffices to show that $\frac{1}{\varphi(q)}$ is larger than the *convex* combination of $\frac{1}{\varphi(q_l)}$. This is true when $\frac{1}{\varphi(q)} \geq \frac{1}{\varphi(q_l)}$ for all l, and at least one strict inequality holds for some $l \in \mathcal{Y}$. Note that this is a sufficient condition *but not a necessary condition* to prove that $\boldsymbol{q}^{\gamma,*}$ is $\boldsymbol{\eta}$UC. Nevertheless, we will show that this condition holds and thus Cor. 9 can be proven based on this condition. Thus, in the following proof, we focus on showing that when $q \in [\frac{1}{2}, 1)$ then for all l, we have

$$\varphi(q_l) \geq \varphi(q). \tag{106}$$

To show the above inequality, we split the proof into two cases:

- **Case 1**: $\varphi(a) > \varphi(b)$ when $1 > b > a \geq \frac{1}{2}$, *i.e.*, φ is a decreasing function when its argument is greater or equal to $\frac{1}{2}$.
- **Case 2**: $\varphi(a) > \varphi(\frac{1}{2})$ when $a \in [0, \frac{1}{2})$, *i.e.*, when $a < \frac{1}{2}$, $\varphi(a)$ is lowerbounded by $\varphi(\frac{1}{2})$.

If the above two cases hold, then Ineq. (106) holds. Next, we present the proof for each case.

Case 1: $\varphi(a) > \varphi(b)$ when $1 > b > a \geq \frac{1}{2}$.

To show $\varphi(a)$ is decreasing for $a \in [\frac{1}{2}, 1)$, we will show that its derivative is smaller than zero. This can be seen by manipulating the derivative of φ as follows:

$$\frac{d}{da}\varphi(a) = -\gamma(1-a)^{\gamma-1} + \gamma(\gamma-1)(1-a)^{\gamma-2} a \log a - \gamma(1-a)^{\gamma-1} \log a - \gamma(1-a)^{\gamma-1} \tag{107}$$

$$= \gamma(1-a)^{\gamma-2}(-(1-a) + (\gamma-1)a \log a - (1-a)\log a - (1-a)) \tag{108}$$

$$= \gamma(1-a)^{\gamma-2}((\gamma a - 1)\log a - 2 + 2a) \tag{109}$$

$$= \underbrace{\gamma(1-a)^{\gamma-2}}_{>0}(\underbrace{\gamma a \log a}_{<0} + \underbrace{(-\log a - 2 + 2a)}_{<0}) \tag{110}$$

$$< 0. \tag{111}$$

In Eq. (110), we know $-\log a - 2 + 2a \leq 0$ by noting its derivative is $2 - \frac{1}{a} \geq 0$ for $a \in [\frac{1}{2}, 1)$, meaning it is an increasing function in $a \in [\frac{1}{2}, 1)$. Therefore, its supremum must be at 1 with the value $-\log 1 - 2 + 2(1) = 0$. Since its supremum is zero, all other values in the range must be smaller than zero, making the term negative. This suffices to show that $\varphi(a)$ is decreasing for $a \in [\frac{1}{2}, 1)$, and concludes that **Case 1** holds.

Case 2: $\varphi(a) > \varphi(\frac{1}{2})$ when $a \in [0, \frac{1}{2})$.

Let us reexpress the above expression,

$$\varphi(a) > \varphi(\frac{1}{2}) \tag{112}$$

$$(1-a)^\gamma - \gamma(1-a)^{\gamma-1} a \log a > \frac{1}{2^\gamma} - \gamma \frac{1}{2^\gamma} \log \frac{1}{2} \tag{113}$$

$$(1-a)^\gamma (1 - \gamma(1-a)^{-1} a \log a) > \frac{1}{2^\gamma} \left(1 - \gamma \log \frac{1}{2}\right) \tag{114}$$

$$\frac{(1-a)^\gamma (1 - \gamma(1-a)^{-1} a \log a)}{\frac{1}{2^\gamma}\left(1 - \gamma \log \frac{1}{2}\right)} > 1 \tag{115}$$

$$\underbrace{\frac{(1-a)^\gamma \left(\frac{1}{2}+a\right)^\gamma}{\frac{1}{2^\gamma}}}_{=:s_1} \underbrace{\frac{(1 - \gamma(1-a)^{-1} a \log a)}{\left(\frac{1}{2}+a\right)^\gamma \left(1 - \gamma \log \frac{1}{2}\right)}}_{=:s_2} > 1, \tag{116}$$

where in the last inequality we multiply $\left(\frac{1}{2}+a\right)^\gamma$ to both numerator and denominator[5]. Next, we show that both $s_1 > 1$ and $s_2 > 1$.

For s_1, we can see that for $a \in (0, \frac{1}{2})$,

$$s1 = \frac{(1-a)^\gamma \left(\frac{1}{2}+a\right)^\gamma}{\frac{1}{2^\gamma}} \tag{117}$$

$$= \frac{(1-a)^\gamma (1+2a)^\gamma}{\underbrace{\frac{1}{2^\gamma} 2^\gamma}_{=1}} \tag{118}$$

$$= (1-a)^\gamma (1+2a)^\gamma \tag{119}$$

$$= ((1-a)(1+2a))^\gamma \tag{120}$$

$$= (1 - a + 2a - 2a^2)^\gamma \tag{121}$$

$$= (1 + a - 2a^2)^\gamma \tag{122}$$

$$> 1. \tag{123}$$

The last inequality comes from the fact that $1 + a - 2a^2 > 0$, since it is quadratic with a negative coefficient on the a^2 term (*i.e.*, an upside-down U-curve) and it takes value of 1 at $a = 0$ and $a = \frac{1}{2}$. Thus, this quadratic term is larger than 1 in the range $a \in (0, \frac{1}{2})$. This shows that in Ineq. (116) we have $s_1 > 1$ for $a \in (0, \frac{1}{2})$.

For s_2, our goal is to show that for $a \in (0, \frac{1}{2})$, we have

$$\frac{1 - \gamma(1-a)^{-1} a \log a}{\left(\frac{1}{2}+a\right)^\gamma \left(1 - \gamma \log \frac{1}{2}\right)} > 1. \tag{124}$$

This can be rearranged as

$$1 - \frac{\gamma a \log a}{(1-a)} > \left(\frac{1}{2}+a\right)^\gamma \left(1 - \gamma \log \frac{1}{2}\right) \tag{125}$$

$$1 - \frac{\gamma a \log a}{(1-a)} - \left(\frac{1}{2}+a\right)^\gamma \left(1 - \gamma \log \frac{1}{2}\right) > 0 \tag{126}$$

$$\frac{1}{\left(\frac{1}{2}+a\right)^\gamma} - \frac{\gamma a \log a}{(1-a)\left(\frac{1}{2}+a\right)^\gamma} - \left(1 - \gamma \log \frac{1}{2}\right) > 0. \tag{127}$$

[5]We came up with $\left(\frac{1}{2}+a\right)^\gamma$ by trial and error.

Let us denote the left-hand side of the above in equality as the following:

$$g(\gamma) = \frac{1}{\left(\frac{1}{2}+a\right)^{\gamma}} - \frac{\gamma a \log a}{(1-a)\left(\frac{1}{2}+a\right)^{\gamma}} - \left(1 - \gamma \log \frac{1}{2}\right). \tag{128}$$

Based on the above derivation, showing Ineq. (124) is equivalent to showing that $g(\gamma) > 0$ for $\gamma > 0$ and $0 < a < \frac{1}{2}$. To show $g(\gamma) > 0$ for $\gamma > 0$, we observe the following properties of g:

- g is convex in γ. This is because it is a sum of convex functions, *i.e.*, $\left(\frac{1}{2}+a\right)^{-\gamma}$ and $-\frac{\gamma a \log a}{(1-a)\left(\frac{1}{2}+a\right)^{\gamma}}$ can be shown to be convex by noting their second order derivatives are positive for all $\gamma > 0$, while $-(1 - \gamma \log \frac{1}{2})$ is linear in γ.
- $g(0) = 0$.

With these properties, we only need to show that the derivative of g at 0 is nonnegative, *i.e.*, $\frac{dg}{d\gamma}(0) \geq 0$, since convexity guarantees that the derivative of g would only increase in γ for $\gamma \geq 0$ and since $g(0) = 0$ we will have $g(\gamma) \geq 0$ for all $\gamma \geq 0$. Thus, next, we show that

$$0 \leq \frac{d}{d\gamma} g(0) \tag{129}$$

$$= \left[-\frac{\log(\frac{1}{2}+a)}{(\frac{1}{2}+a)^{\gamma}} - \frac{a \log a}{(1-a)(\frac{1}{2}+a)^{\gamma}} + \frac{\gamma a \log a \log(\frac{1}{2}+a)}{(1-a)(\frac{1}{2}+a)^{\gamma}} + \log \frac{1}{2} \right]_{\gamma=0} \tag{130}$$

$$= -\log\left(\frac{1}{2}+a\right) - \frac{a \log a}{(1-a)} + \log \frac{1}{2} \tag{131}$$

$$= -\log(1+2a) - \frac{a \log a}{(1-a)}. \tag{132}$$

Showing the above inequality is equivalent to showing

$$-(1-a)\log(1+2a) - a \log a \geq 0. \tag{133}$$

Let us denote the above expression as $\phi(a) = -(1-a)\log(1+2a) - a \log a$, where its derivative and second order derivative are

$$\frac{d}{da}\phi(a) = \log(1+2a) - \frac{2(1-a)}{1+2a} - \log a - 1 \tag{134}$$

$$= \log(1+2a) - \frac{3}{1+2a} - \log a, \text{ and} \tag{135}$$

$$\frac{d^2}{da^2}\phi(a) = \frac{2}{1+2a} + \frac{6}{(1+2a)^2} - \frac{1}{a} \tag{136}$$

$$= \frac{4a-1}{a(1+2a)^2}. \tag{137}$$

From Eq. (137), we can see that $\frac{d^2}{da^2}\phi(a) \geq 0$ for $a \in [\frac{1}{4}, \frac{1}{2}]$ and $\frac{d^2}{da^2}\phi(a) \leq 0$ for $a \in (0, \frac{1}{4}]$. This means $\frac{d\phi}{da}$ is increasing for $a \in [\frac{1}{4}, \frac{1}{2}]$ and decreasing for $a \in (0, \frac{1}{4}]$. Also, we see that $\frac{d}{da}\phi(0) = \infty$, $\frac{d}{da}\phi(\frac{1}{4}) = -0.2082$, and $\frac{d}{da}\phi(\frac{1}{2}) = -0.1137$. By the intermediate value theorem, there exists an $a_0 \in (0, \frac{1}{4})$ such that $\frac{d}{da}\phi(a_0) = 0$, which means ϕ is increasing for $a \in (0, a_0)$ and decreasing for $a \in (a_0, \frac{1}{2})$. Therefore, the minimum of ϕ in $[0, \frac{1}{2}]$ can only be at either 0 or $\frac{1}{2}$. By computing the values at both ends, we have

$$\min\left\{\phi(0), \phi\left(\frac{1}{2}\right)\right\} = \min\{0, 0\} = 0 \geq 0. \tag{138}$$

Thus, $\phi(a) \geq 0$ for $a \in [0, \frac{1}{2}]$. To wrap up, this result means Ineq. (129), *i.e.*, $\frac{dg}{d\gamma}(0) \geq 0$, holds. Together with g being convex and $g(0) = 0$, this means $g(a) \geq 0$ for $a \in [0, \frac{1}{2}]$, making Ineq. (127) and therefore Ineq. (124) hold. This completes the proof for s_2.

With the above lowerbounds of s_1 and s_2, we conclude that **Case 2** holds.

By combining **Case 1** and **Case 2**, we have proven Cor. 9, which states that $\boldsymbol{q}^{\gamma,*}(\boldsymbol{x})$ is ηUC when $\frac{1}{2} \leq q < 1$ if $\boldsymbol{q}^{\gamma,*}(\boldsymbol{x}) \notin \mathcal{S}^K$. □

A.12. Alternative proof of Cor. 10: ηUC property for $\boldsymbol{q}^{\gamma,*}$ in binary classification where $K = 2$

Here, we present an alternative proof of Cor. 10, which is independent from Thm. 11, Thm. 8, Lem. 13 and Lem. 14. Therefore, this proof does not explicitly utilize the function φ.

Proof. Let $q = \max_y q_y^{\gamma,*}(\boldsymbol{x})$ be the highest score in the simplex output from the focal risk minimizer $\boldsymbol{q}^{\gamma,*}$ and $\eta = \max_y \eta_y(\boldsymbol{x})$ be the true class-posterior probability of the most probable class. Note that $q > \frac{1}{2}$ since $\boldsymbol{q}^{\gamma,*}(\boldsymbol{x}) \notin \mathcal{S}^K$ and there are only two classes. Also, in binary classification, it is restricted that the output of the simplex of the other class must be $1 - q$ since $\boldsymbol{q}^{\gamma,*}(\boldsymbol{x}) \in \Delta^K$.

First, we relate η to the focal risk minimizer $\boldsymbol{q}^{\gamma,*}(\boldsymbol{x})$. Since we know that the focal loss is classification calibrated from our Thm. 3, we have $\arg\max_y q_y^{\gamma,*}(\boldsymbol{x}) = \arg\max_y \eta_y(\boldsymbol{x})$. As a result, the pointwise conditional risk w.r.t. the focal risk minimizer $\boldsymbol{q}^{\gamma,*}$ can be written as follows:

$$W^{\ell_{\mathrm{FL}}^\gamma}\big(\boldsymbol{q}^{\gamma,*}(\boldsymbol{x}); \boldsymbol{\eta}(\boldsymbol{x})\big) = \sum_{y\in\mathcal{Y}} \eta_y(\boldsymbol{x}) \ell_{\mathrm{FL}}^\gamma\big(\boldsymbol{q}(\boldsymbol{x}), \boldsymbol{e}_y\big) \tag{139}$$

$$= \eta(-(1-q)^\gamma \log q) + (1-\eta)(-q^\gamma \log(1-q)). \tag{140}$$

For simplicity in the binary case, let us define

$$W^{\ell_{\mathrm{FL}}^\gamma,*}\big(q, \eta\big) = \eta(-(1-q)^\gamma \log q) + (1-\eta)(-q^\gamma \log(1-q)) \tag{141}$$

Since the focal loss is differentiable, $W^{\ell_{\mathrm{FL}}^\gamma,*}\big(q, \eta\big)$ is also differentiable, and we can express $\frac{d}{dq} W^{\ell_{\mathrm{FL}}^\gamma,*}$ as

$$\frac{d}{dq} W^{\ell_{\mathrm{FL}}^\gamma,*}\big(q, \eta\big) = \eta\left[(1-q)^{\gamma-1}\gamma\log q - \frac{(1-q)^\gamma}{q}\right] + (1-\eta)\left[\frac{q^\gamma}{1-q} - q^{\gamma-1}\gamma\log(1-q)\right] \tag{142}$$

$$= \eta\left[(1-q)^{\gamma-1}\gamma\log q - \frac{(1-q)^\gamma}{q} - \frac{q^\gamma}{1-q} + q^{\gamma-1}\gamma\log(1-q)\right] + \frac{q^\gamma}{1-q} - q^{\gamma-1}\gamma\log(1-q) \tag{143}$$

Since $\boldsymbol{q}^{\gamma,*}$ minimizes $W^{\ell_{\mathrm{FL}}^\gamma}$ by the definition of the focal risk minimizer. Also, the binary pointwise conditional risk $W^{\ell_{\mathrm{FL}}^\gamma,*}\big(q, \eta\big)$ is convex in q given η. We know that $\frac{d}{dq} W^{\ell_{\mathrm{FL}}^\gamma,*}\big(q, \eta\big) = 0$ and the following equation holds:

$$\eta\left[(1-q)^{\gamma-1}\gamma\log q - \frac{(1-q)^\gamma}{q} - \frac{q^\gamma}{1-q} + q^{\gamma-1}\gamma\log(1-q)\right] + \frac{q^\gamma}{1-q} - q^{\gamma-1}\gamma\log(1-q) = 0 \tag{144}$$

$$\eta\left[(1-q)^{\gamma-1}\gamma\log q - \frac{(1-q)^\gamma}{q} - \frac{q^\gamma}{1-q} + q^{\gamma-1}\gamma\log(1-q)\right] = q^{\gamma-1}\gamma\log(1-q) - \frac{q^\gamma}{1-q} \tag{145}$$

$$\eta = \frac{q^{\gamma-1}\gamma\log(1-q) - \frac{q^\gamma}{1-q}}{(1-q)^{\gamma-1}\gamma\log q - \frac{(1-q)^\gamma}{q} - \frac{q^\gamma}{1-q} + q^{\gamma-1}\gamma\log(1-q)} \tag{146}$$

$$\eta = \frac{\frac{q^\gamma}{1-q} - q^{\gamma-1}\gamma\log(1-q)}{-(1-q)^{\gamma-1}\gamma\log q + \frac{(1-q)^\gamma}{q} + \frac{q^\gamma}{1-q} - q^{\gamma-1}\gamma\log(1-q)}. \tag{147}$$

Therefore, we can relate η and q as follows:

$$\eta = \frac{\frac{q^\gamma}{1-q} - \gamma q^{\gamma-1}\log(1-q)}{\frac{q^\gamma}{1-q} - \gamma q^{\gamma-1}\log(1-q) + \frac{(1-q)^\gamma}{q} - \gamma(1-q)^{\gamma-1}\log q}. \tag{148}$$

As a result, the true class-posterior probability of the most probable class η can be recovered from the highest score of the simplex output from the focal risk minimizer q by Eq. (148).

Next, to prove that $\boldsymbol{q}^{\gamma^*}$ is $\boldsymbol{\eta}$UC, it suffices to prove that $\eta - q > 0$ for $q > \frac{1}{2}$, which means the minimizer of the focal loss underestimates the true class-posterior probability of the most probable class, *i.e.*, $q < \eta$. Thus for $q > \frac{1}{2}$, we have

$$\frac{\frac{q^\gamma}{1-q} - \gamma q^{\gamma-1}\log(1-q)}{\frac{q^\gamma}{1-q} - \gamma q^{\gamma-1}\log(1-q) + \frac{(1-q)^\gamma}{q} - \gamma(1-q)^{\gamma-1}\log q} - q > 0 \tag{149}$$

$$\frac{\frac{q^\gamma}{1-q} - \gamma q^{\gamma-1}\log(1-q) - q\left(\frac{q^\gamma}{1-q} - \gamma q^{\gamma-1}\log(1-q) + \frac{(1-q)^\gamma}{q} - \gamma(1-q)^{\gamma-1}\log q\right)}{\frac{q^\gamma}{1-q} - \gamma q^{\gamma-1}\log(1-q) + \frac{(1-q)^\gamma}{q} - \gamma(1-q)^{\gamma-1}\log q} > 0. \tag{150}$$

Since every term in the denominator of Ineq. (150) is positive, it can be ignored because it is sufficient to only check if the numerator is positive to prove that the fraction is positive. Therefore, we have

$$\frac{q^\gamma}{1-q} - \gamma q^{\gamma-1}\log(1-q) - q\left(\frac{q^\gamma}{1-q} - \gamma q^{\gamma-1}\log(1-q) + \frac{(1-q)^\gamma}{q} - \gamma(1-q)^{\gamma-1}\log q\right) > 0 \tag{151}$$

$$(1-q)\left[\frac{q^\gamma}{1-q} - \gamma q^{\gamma-1}\log(1-q)\right] - q\left[\frac{(1-q)^\gamma}{q} - \gamma(1-q)^{\gamma-1}\log q\right] > 0 \tag{152}$$

$$(1-q)\left[\frac{q^\gamma}{1-q} - \gamma q^{\gamma-1}\log(1-q)\right] - \left[(1-q)^\gamma - \gamma q(1-q)^{\gamma-1}\log q\right] > 0 \tag{153}$$

$$q^\gamma - \gamma(1-q)q^{\gamma-1}\log(1-q) - (1-q)^\gamma + \gamma q(1-q)^{\gamma-1}\log q > 0. \tag{154}$$

Next, we will show that Ineq. (154) always holds when $q > \frac{1}{2}$ for all $\gamma > 0$. We split the proof into two cases, the first case when $\gamma \geq 1$ and the second case when $0 < \gamma < 1$.

Case 1 ($\gamma \geq 1$): We can also express Ineq. (154) as

$$q^\gamma - \gamma(1-q)q^{\gamma-1}\log(1-q) - (1-q)^\gamma + \gamma q(1-q)^{\gamma-1}\log q > 0 \tag{155}$$

$$q^\gamma - \gamma(1-q)q^{\gamma-1}\log(1-q) > (1-q)^\gamma - \gamma q(1-q)^{\gamma-1}\log q \tag{156}$$

$$\frac{q^{\gamma-1}}{(1-q)^{\gamma-1}}\frac{[q-\gamma(1-q)\log(1-q)]}{[(1-q)-\gamma q\log q]} > 1. \tag{157}$$

Because $q > \frac{1}{2} > 1-q$ and $\gamma >= 1$, we have

$$\frac{q^{\gamma-1}}{(1-q)^{\gamma-1}} \geq 1. \tag{158}$$

Next, we will show that the following inequality

$$\frac{q-\gamma(1-q)\log(1-q)}{(1-q)-\gamma q\log q} > 1 \tag{159}$$

can be rewritten as

$$\frac{q-\gamma(1-q)\log(1-q)}{(1-q)-\gamma q\log q} > 1 \tag{160}$$

$$q-\gamma(1-q)\log(1-q) > (1-q)-\gamma q\log q \tag{161}$$

$$(q-\gamma(1-q)\log(1-q)) - ((1-q)-\gamma q\log q) > 0 \tag{162}$$

$$\underbrace{2q-1}_{>0} + \gamma \underbrace{(-(1-q)\log(1-q) + q\log q)}_{>0} > 0. \tag{163}$$

Therefore, to prove that Ineq. (159) holds, it suffices to prove that $2q-1 > 0$ and $-(1-q)\log(1-q) + q\log q > 0$. Since

$q > \frac{1}{2}$, it is straightforward to see that $2q - 1 > 0$ for all $q > \frac{1}{2}$. Next, we prove that $-(1-q)\log(1-q) + q\log q > 0$:

$$-(1-q)\log(1-q) + q\log q > 0 \tag{164}$$
$$-\log(1-q) + \frac{1}{1-q} q\log q > 0 \tag{165}$$
$$-\frac{1}{q}\log(1-q) + \frac{1}{1-q}\log q > 0 \tag{166}$$
$$-\frac{1}{q}\left[-\sum_{i=1}^{\infty}\frac{q^i}{i}\right] + \frac{1}{1-q}\left[-\sum_{i=1}^{\infty}\frac{(1-q)^i}{i}\right] > 0 \tag{167}$$
$$\left[\sum_{i=1}^{\infty}\frac{q^{i-1}}{i}\right] + \left[-\sum_{i=1}^{\infty}\frac{(1-q)^{i-1}}{i}\right] > 0 \tag{168}$$
$$\sum_{i=1}^{\infty}\frac{q^{i-1} - (1-q)^{i-1}}{i} > 0. \tag{169}$$
$$\sum_{i=2}^{\infty}\frac{q^{i-1} - (1-q)^{i-1}}{i} > 0. \tag{170}$$

Note that we can write $\log(1-q) = \sum_{i=1}^{\infty} -\frac{q^i}{i}$ (Maclaurin series) and $q^{i-1} - (1-q)^{i-1} > 0$ since $q > \frac{1}{2}$ for $i \geq 2$. Thus, by combining Ineqs. (158) and (159), we prove that Ineq. (154) holds for the case where $\gamma \geq 1$. Therefore, we can conclude that if $\boldsymbol{q}^{\gamma,*}(\boldsymbol{x}) \notin \mathcal{S}^K$, then $\boldsymbol{q}^{\gamma,*}$ is $\boldsymbol{\eta}$UC for all $\gamma \geq 1$.

Case 2 ($0 < \gamma < 1$): First, we add $\gamma q - \gamma q + \gamma(1-q) - \gamma(1-q)$, which equals to zero, to Ineq. (154):

$$q^\gamma - \gamma(1-q)q^{\gamma-1}\log(1-q) - (1-q)^\gamma + \gamma q(1-q)^{\gamma-1}\log q + \gamma q - \gamma q + \gamma(1-q) - \gamma(1-q) > 0. \tag{171}$$

We then split Ineq. (171) into two parts, which are

$$q^\gamma - (1-q)^\gamma - \gamma q + \gamma(1-q) > 0, \tag{172}$$

and

$$-\gamma(1-q)q^{\gamma-1}\log(1-q) + \gamma q(1-q)^{\gamma-1}\log q + \gamma q - \gamma(1-q) > 0 \tag{173}$$
$$-(1-q)q^{\gamma-1}\log(1-q) + q(1-q)^{\gamma-1}\log q + q - (1-q) > 0. \tag{174}$$

We will prove **Case 2** by showing that both Ineqs. (172) and (174) hold simultaneously, which suggests that Ineq. (154) must also hold.

Next, we will show that Ineq. (172) holds. First, it can be observed that the left-hand side of Ineq. (172) is zero when $q = \frac{1}{2}$. Next, the derivative with respect to q can be expressed as follows:

$$\frac{d}{dq}\left[q^\gamma - (1-q)^\gamma - \gamma q + \gamma(1-q)\right] = \gamma q^{\gamma-1} + \gamma(1-q)^{\gamma-1} - 2\gamma q, \tag{175}$$

which can be shown to always be positive for $0 < \gamma < 1$:

$$\gamma q^{\gamma-1} + \gamma(1-q)^{\gamma-1} - 2\gamma q > 0 \tag{176}$$
$$q^{\gamma-1} + (1-q)^{\gamma-1} - 2q > 0 \tag{177}$$
$$\frac{1}{q^{1-\gamma}} - q + \frac{1}{(1-q)^{1-\gamma}} - q > 0 \tag{178}$$

because $\frac{1}{q^{1-\gamma}} > 1$ and $\frac{1}{(1-q)^{1-\gamma}} > 1$ for $0 < \gamma < 1$. Therefore, $\frac{1}{q^{1-\gamma}} + \frac{1}{(1-q)^{1-\gamma}} > 2 > 2q$, which indicates that Ineq. (178) holds. By combining the fact that the left-hand side of Ineq. (172) is zero when $q = \frac{1}{2}$, and the left-hand side of Ineq. (172) is increasing as q increases. Ineq. (172) must be more than zero for all $q > \frac{1}{2}$. Thus, we can conclude that Ineq. (172) holds for all $q > \frac{1}{2}$.

Next, we will show that Ineq. (174) holds by showing that its left-hand side of Ineq. (174) is an increasing function w.r.t. γ. The following derivative

$$\frac{d}{d\gamma}\left[-(1-q)q^{\gamma-1}\log(1-q)+q(1-q)^{\gamma-1}\log q+q-(1-q)\right] \tag{179}$$

is equal to

$$-(1-q)q^{\gamma-1}\log q\log(1-q)+q(1-q)^{\gamma-1}\log(1-q)\log q. \tag{180}$$

We can see that

$$-(1-q)q^{\gamma-1}\log q\log(1-q)+q(1-q)^{\gamma-1}\log(1-q)\log q>0 \tag{181}$$

$$-(1-q)q^{\gamma-1}+q(1-q)^{\gamma-1}>0. \tag{182}$$

Ineq. (182) holds because $q>1-q$ and $(1-q)^{\gamma-1}>q^{\gamma-1}$ for $0<\gamma<1$. This suggests that the left-hand side of Ineq. (159) is an increasing function w.r.t. γ. Next, we show that Ineq. (171) holds for $\gamma=0$. By substitute $\gamma=0$ to Ineq. (174), we have

$$-\frac{(1-q)\log(1-q)}{q}+\frac{q\log q}{(1-q)}+q-(1-q)>0 \tag{183}$$

$$-(1-q)\left(\frac{\log(1-q)}{q}+1\right)+q\left(\frac{\log q}{(1-q)}+1\right)>0. \tag{184}$$

By using the fact that $\log(q)=-\sum_{i=1}^{\infty}\frac{(1-q)^i}{i}$, we have

$$-(1-q)\left(\frac{-\sum_{i=1}^{\infty}\frac{q^i}{i}}{q}+1\right)+q\left(\frac{-\sum_{i=1}^{\infty}\frac{(1-q)^i}{i}}{(1-q)}+1\right)>0. \tag{185}$$

Dividing by $q(1-q)$ gives

$$-\frac{1}{q}\left(\frac{-\sum_{i=1}^{\infty}\frac{q^i}{i}}{q}+1\right)+\frac{1}{1-q}\left(\frac{-\sum_{i=1}^{\infty}\frac{(1-q)^i}{i}}{(1-q)}+1\right)>0 \tag{186}$$

$$-\frac{1}{q^2}\left(-\left(\sum_{i=1}^{\infty}\frac{q^i}{i}\right)+q\right)+\frac{1}{(1-q)^2}\left(\left(-\sum_{i=1}^{\infty}\frac{(1-q)^i}{i}\right)+(1-q)\right)>0 \tag{187}$$

$$-\frac{1}{q^2}\left(-\sum_{i=2}^{\infty}\frac{q^i}{i}\right)+\frac{1}{(1-q)^2}\left(-\sum_{i=2}^{\infty}\frac{(1-q)^i}{i}\right)>0 \tag{188}$$

$$\left(\sum_{i=2}^{\infty}\frac{q^{i-2}}{i}\right)+\left(-\sum_{i=2}^{\infty}\frac{(1-q)^{i-2}}{i}\right)>0 \tag{189}$$

$$\left(\sum_{i=2}^{\infty}\frac{q^{i-2}-(1-q)^{i-2}}{i}\right)>0. \tag{190}$$

Since $q>1-q$ and $i>=2$, Ineq. (190) holds, which indicates that Ineq. (174) holds for $\gamma=0$. And we know that the left-hand size of Ineq. (174) is an increasing function as γ increases for $0<\gamma<1$ from Ineq. (182). As a result, Ineq. (190) holds for $0<\gamma<1$. Therefore, we can conclude that if $\boldsymbol{q}^{\gamma,*}(\boldsymbol{x})\notin\mathcal{S}^K$, then $\boldsymbol{q}^{\gamma,*}(\boldsymbol{x})$ is $\boldsymbol{\eta}$UC for all $0<\gamma<1$.

By combining **Case 1** and **Case 2**, we have proven Cor. 10, which states that $\boldsymbol{q}^{\gamma,*}(\boldsymbol{x})$ is ηUC unless it is uniform or a one-hot vector. □

B. Evaluation metrics

In practice, hard labels are given. As a result, it is not straightforward to measure the quality of class-posterior probability estimation. In this section, we review evaluation metrics that are used in this paper to evaluate the quality of prediction confidence given hard labels.

B.1. Expected calibration error (ECE)

The expected calibration error can be defined as follows [32, 18]:

$$\text{ECE} = \frac{1}{n}\sum_{j=1}^{N_B} |B_j| \left|\text{acc}(B_j) - \text{conf}(B_j)\right|, \tag{191}$$

where $|B_j|$ is the number of samples in bin j, $\text{acc}(B_j)$ and $\text{conf}(B_j)$ are the average accuracy and the average confidence of data in bin j, and N_B is the number of bins. In this paper, we used $N_B = 10$. To allocate the data in the different bins, we use the following procedure. First, we rank the test data by the prediction confidence of our classifier. Then we put the test data in the bin based on their confidence scores. More precisely, the first bin is for data with prediction confidence in the range of $0 - 0.1$, the second one is in the range of $0.1 - 0.2$, and the tenth one is in the range of $0.9 - 1.0$. Then, we can calculate ECE by calculating the average accuracy and the average confidence of data in each bin and then combine them based on Eq. (191).

B.2. Negative log-likelihood (NLL)

Given data with hard labels $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ and a classifier $\boldsymbol{q}$, one can calculate NLL as follows:

$$\text{NLL} = -\sum_{i=1}^{n} \log(q_{y_i}(\boldsymbol{x}_i)). \tag{192}$$

It is simple since it can be calculated in a pointwise manner.

B.3. Classwise expected calibration error (CW-ECE)

CW-ECE is defined as follows [22]:

$$\text{CW-ECE} = \frac{1}{K}\sum_{l=1}^{K}\sum_{j=1}^{N_B} \frac{|B_{j,l}|}{n} \left|\text{prop}_l(B_{j,l}) - \text{conf}_l(B_{j,l})\right|, \tag{193}$$

where $|B_{j,l}|$ denotes the number of samples in bin (j, l), $\text{prop}_l(B_{j,l})$ denotes the proportion of class l in bin $B_{j,l}$, and $\text{conf}_l(B_{j,l})$ denotes the average confidence of predicting class l in bin $B_{j,l}$. Unlike ECE, we rank the confidence score based on each class, not the maximum confidence score. Moreover, CW-ECE does not use the average accuracy but simply the proportion of the class in each bin to compare with the average confidence of the class in that bin. The original motivation of CW-ECE is to mitigate the problem of ECE that it only focuses on the class with the highest confidence.

Table 1. Mean and standard error of the classification error under different datasets, training paradigms, models, and losses. The purpose of showing this table is for reference.

Dataset	Training paradigm	Model	CE	FL-1	FL-2	FL-3
SVHN	**Standard/TS**	ResNet8	3.83(0.07)	3.98(0.08)	4.06(0.15)	4.19(0.11)
		ResNet20	2.01(0.07)	2.05(0.05)	2.09(0.05)	2.16(0.05)
		ResNet44	1.92(0.07)	1.92(0.08)	2.02(0.05)	2.05(0.06)
		ResNet110	1.90(0.09)	1.99(0.08)	2.00(0.08)	2.07(0.07)
	LS	ResNet8	3.83(0.10)	3.90(0.14)	3.90(0.06)	3.94(0.07)
		ResNet20	1.98(0.05)	2.04(0.06)	2.02(0.07)	2.05(0.06)
		ResNet44	1.87(0.04)	1.88(0.06)	1.90(0.07)	1.94(0.08)
		ResNet110	1.89(0.11)	1.89(0.05)	1.90(0.03)	1.90(0.09)
CIFAR10	**Standard/TS**	ResNet8	12.58(0.39)	12.83(0.47)	13.28(0.45)	13.95(0.59)
		ResNet20	7.42(0.40)	7.53(0.31)	7.74(0.42)	8.01(0.29)
		ResNet44	6.21(0.44)	6.17(0.36)	6.47(0.31)	6.78(0.37)
		ResNet110	5.68(0.32)	5.87(0.36)	5.82(0.33)	6.42(0.61)
	LS	ResNet8	12.58(0.30)	12.51(0.33)	13.06(0.37)	13.12(0.25)
		ResNet20	7.53(0.32)	7.53(0.42)	7.46(0.38)	7.69(0.32)
		ResNet44	6.41(0.29)	6.35(0.28)	6.49(0.36)	6.67(0.31)
		ResNet110	5.82(0.33)	5.86(0.39)	5.75(0.33)	5.96(0.38)
CIFAR10-s	**Standard/TS**	ResNet8	25.56(0.63)	26.10(0.89)	26.45(0.80)	27.36(0.93)
		ResNet20	21.44(0.51)	21.98(0.66)	22.47(0.52)	23.01(0.48)
		ResNet44	20.97(0.52)	21.60(0.39)	22.11(0.61)	23.44(1.13)
		ResNet110	24.91(1.54)	27.08(2.28)	27.78(2.36)	28.12(1.76)
	LS	ResNet8	24.98(0.82)	25.48(0.78)	26.16(0.54)	26.46(0.90)
		ResNet20	21.20(0.43)	21.37(0.58)	21.80(0.47)	22.15(0.54)
		ResNet44	21.09(0.37)	21.28(0.39)	21.95(0.71)	22.48(0.68)
		ResNet110	27.88(2.65)	29.25(3.23)	27.92(1.51)	28.45(2.07)
CIFAR100	**Standard/TS**	ResNet8	41.05(0.53)	41.40(0.50)	41.55(0.52)	42.07(0.33)
		ResNet20	31.71(0.46)	32.38(0.40)	32.96(0.34)	33.40(0.45)
		ResNet44	28.88(0.42)	29.47(0.41)	29.30(0.37)	29.93(0.45)
		ResNet110	25.10(0.68)	25.30(0.91)	25.58(0.81)	26.22(0.53)
	LS	ResNet8	41.42(0.42)	41.44(0.31)	41.72(0.44)	42.15(0.38)
		ResNet20	31.99(0.20)	32.25(0.37)	32.70(0.40)	32.94(0.26)
		ResNet44	28.89(0.38)	28.84(0.50)	29.17(0.43)	29.25(0.47)
		ResNet110	25.12(0.70)	24.84(0.42)	25.15(0.37)	24.90(0.50)

C. Additional experimental results

In this section, we provide additional experimental results. We report the classification error of all ResNets we used on SVHN, CIFAR10, CIFAR10-s, and CIFAR100. Then, we report ECE and accuracy on additional datasets to support our claim in Sec. 6.3 using a linear-in-input model and neural networks with one hidden layer. Next, we report reliability diagrams for all ResNet models we used for SVHN, CIFAR10, CIFAR10-s, and CIFAR100. Finally, we report the performance of all ResNet models we used for SVHN, CIFAR10, CIFAR10-s, and CIFAR100 using three evaluation metrics, which are the expected calibration error (ECE), negative log-likelihood (NLL), and classwise expected calibration error (CW-ECE).

C.1. Classification error of different datasets and models

Table 1 shows the classification error of all ResNet models we used with different training paradigms on SVHN, CIFAR10, CIFAR10-s, and CIFAR100. Since the purpose of showing this table is for reference, we do not bold any numbers.

C.2. Experiments on additional datasets

Here, we conducted experiments on additional 30 datasets on models that arguably to be simpler than ResNet [19], which are a linear-in-input model and a neural network with one hidden layer.

Datasets: We used datasets from the UCI Data Repository [24]. We also used MNIST [23], Kuzushiji-MNIST (KMNIST) [11], and Fashion-MNIST [53].

Models: We used two models in our experiments on additional datasets, which are a linear-in-input model and a neural network with one hidden layer. Note that we do not compare the performance across the model but to validate the effectiveness of using the transformation $\mathbf{\Psi}^\gamma$ for all settings.

Methods: We compared the models that use $\mathbf{\Psi}^\gamma$ after the softmax layer to those that do not. Note that both methods have the same accuracy since $\mathbf{\Psi}^\gamma$ does not affect the decision rule (Prop. 12). We used the focal loss with $\gamma \in \{0, 1, 2, 3\}$ in this experiment, 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]. We used 10 as the number of bins. CE denotes the method that uses the cross entropy loss and FL-γ denotes the method that uses the focal loss with γ. We denote FL-γ-$\mathbf{\Psi}^\gamma$ for a method that applies $\mathbf{\Psi}^\gamma$ to the output of the softmax layer before evaluating the confidence score when training with the focal loss. Note that CE is equivalent to FL-0 and thus $\mathbf{\Psi}^\gamma$ does not change the output of the trained classifier.

Hyperparameters: For both linear-in-input model and a neural network, the number of epochs was 50 for all datasets and the batch size was 64. We used SGD with momentum of 0.9, where the learning rate was 0.01. The weight decay parameter was 10^{-3}. For a neural network model, the number of nodes in a hidden layer was 64.

Discussion: Tables 2 and 4 show ECE on all datasets using the same model for different losses. It can be observed that $\mathbf{\Psi}^\gamma$ can effectively improve the performance of the classifier trained with the focal loss in most cases although not all the cases. We can also see that as the γ increases, ECE also increases for most datasets as well. Tables 3 and 5 show the classification error for all methods for reference.

Table 2. Mean and standard error of ECE over ten trials (rescaled to $0-100$). We used **a linear-in-input model**. Outperforming methods are highlighted in boldface using one-sided t-test with the significance level 5%.

Dataset	CE	FL-1	FL-2	FL-3	FL-1-$\mathbf{\Psi}^\gamma$	FL-2-$\mathbf{\Psi}^\gamma$	FL-3-$\mathbf{\Psi}^\gamma$
Australian	$\mathbf{4.10(1.40)}$	$8.83(2.37)$	$15.47(1.46)$	$19.46(1.85)$	$\mathbf{4.45(1.84)}$	$\mathbf{4.55(0.84)}$	$\mathbf{4.67(1.42)}$
Phishing	$\mathbf{0.89(0.16)}$	$6.48(0.53)$	$12.05(0.42)$	$16.74(0.67)$	$\mathbf{0.83(0.23)}$	$\mathbf{0.96(0.16)}$	$\mathbf{0.94(0.38)}$
Spambase	$3.11(0.40)$	$7.28(0.58)$	$13.07(0.97)$	$16.67(0.64)$	$\mathbf{2.45(0.29)}$	$\mathbf{2.76(0.52)}$	$\mathbf{2.65(0.69)}$
Waveform	$\mathbf{1.42(0.19)}$	$5.83(0.86)$	$10.33(0.90)$	$14.36(0.94)$	$\mathbf{1.61(0.53)}$	$1.72(0.36)$	$1.75(0.20)$
Twonorm	$\mathbf{0.48(0.11)}$	$2.51(0.35)$	$6.13(0.61)$	$10.05(0.55)$	$0.56(0.12)$	$\mathbf{0.52(0.09)}$	$\mathbf{0.43(0.16)}$
Adult	$\mathbf{1.21(0.22)}$	$6.12(0.38)$	$11.76(0.42)$	$15.82(0.34)$	$1.64(0.41)$	$1.94(0.42)$	$2.16(0.41)$
Banknote	$4.37(0.31)$	$9.93(0.56)$	$15.69(0.54)$	$21.07(0.69)$	$3.43(0.24)$	$\mathbf{2.99(0.31)}$	$\mathbf{2.92(0.30)}$
Phoneme	$\mathbf{4.37(0.62)}$	$7.25(0.51)$	$11.92(0.67)$	$14.60(0.58)$	$\mathbf{4.29(0.71)}$	$\mathbf{4.52(0.33)}$	$\mathbf{4.66(0.44)}$
Magic	$\mathbf{1.56(0.35)}$	$9.30(0.58)$	$14.71(0.59)$	$18.29(0.16)$	$\mathbf{1.53(0.22)}$	$\mathbf{1.64(0.49)}$	$\mathbf{1.37(0.31)}$
Gisette	$1.52(0.18)$	$\mathbf{1.41(0.23)}$	$\mathbf{1.47(0.20)}$	$\mathbf{1.50(0.24)}$	$2.18(0.23)$	$2.75(0.21)$	$3.07(0.24)$
USPS	$\mathbf{0.37(0.11)}$	$1.82(0.23)$	$4.80(0.38)$	$8.19(0.56)$	$\mathbf{0.36(0.10)}$	$\mathbf{0.32(0.09)}$	$\mathbf{0.35(0.12)}$
Splice	$\mathbf{2.98(0.97)}$	$6.37(0.87)$	$13.41(0.97)$	$16.98(1.69)$	$\mathbf{2.84(0.76)}$	$\mathbf{2.31(0.66)}$	$3.43(1.10)$
Banana	$\mathbf{6.90(2.41)}$	$8.94(3.09)$	$\mathbf{6.85(2.69)}$	$\mathbf{6.26(1.97)}$	$8.07(3.86)$	$\mathbf{7.11(3.73)}$	$\mathbf{5.09(2.49)}$
Ringnorm	$3.43(0.49)$	$9.86(0.97)$	$14.15(0.87)$	$17.04(0.50)$	$\mathbf{3.15(0.72)}$	$\mathbf{3.08(0.46)}$	$\mathbf{3.00(0.49)}$
Image	$5.45(0.77)$	$10.64(1.47)$	$15.05(0.72)$	$19.27(1.24)$	$5.35(0.77)$	$\mathbf{4.50(0.63)}$	$\mathbf{4.63(0.94)}$
Coil20	$1.39(0.26)$	$5.24(0.59)$	$9.47(0.72)$	$14.53(1.34)$	$1.20(0.24)$	$\mathbf{0.94(0.49)}$	$\mathbf{0.87(0.33)}$
Basehock	$\mathbf{1.16(0.30)}$	$2.32(0.40)$	$3.76(0.67)$	$5.82(0.69)$	$\mathbf{1.39(0.47)}$	$2.39(0.54)$	$2.48(0.53)$
Isolet	$2.80(0.22)$	$7.11(0.35)$	$11.93(0.46)$	$16.92(0.54)$	$2.03(0.29)$	$\mathbf{1.61(0.16)}$	$\mathbf{1.51(0.29)}$
W8a	$\mathbf{0.36(0.11)}$	$2.45(0.32)$	$5.73(0.44)$	$8.69(0.49)$	$0.49(0.16)$	$0.55(0.15)$	$0.70(0.15)$
Mushroom	$0.09(0.04)$	$1.06(0.05)$	$3.18(0.14)$	$5.96(0.26)$	$0.06(0.01)$	$\mathbf{0.04(0.02)}$	$\mathbf{0.04(0.01)}$
Artificial-character	$\mathbf{3.52(0.39)}$	$3.99(0.51)$	$4.70(0.47)$	$5.95(0.64)$	$\mathbf{3.56(0.74)}$	$\mathbf{3.27(0.53)}$	$4.81(0.72)$
Gas-drift	$5.65(0.55)$	$9.05(0.57)$	$12.65(0.53)$	$15.98(0.38)$	$4.54(0.50)$	$\mathbf{4.23(0.51)}$	$\mathbf{4.12(0.49)}$
Japanesevowels	$4.28(0.14)$	$8.88(0.64)$	$12.88(0.30)$	$16.57(0.26)$	$2.89(0.51)$	$1.90(0.18)$	$\mathbf{1.35(0.20)}$
Letter	$10.19(0.33)$	$14.82(0.42)$	$18.26(0.24)$	$20.88(0.60)$	$7.95(0.43)$	$6.13(0.27)$	$\mathbf{5.19(0.43)}$
Pendigits	$4.72(0.33)$	$11.14(0.44)$	$16.29(0.33)$	$20.61(0.55)$	$3.71(0.38)$	$2.68(0.24)$	$\mathbf{1.84(0.36)}$
Satimage	$\mathbf{2.16(0.36)}$	$6.39(0.37)$	$10.19(0.92)$	$13.65(0.50)$	$\mathbf{1.92(0.29)}$	$\mathbf{2.19(0.69)}$	$2.64(0.36)$
Vehicle	$11.27(1.72)$	$16.79(3.20)$	$19.05(2.31)$	$23.62(1.11)$	$10.39(2.16)$	$\mathbf{6.95(1.80)}$	$\mathbf{8.00(1.53)}$
MNIST	$1.27(0.09)$	$5.82(0.14)$	$10.27(0.16)$	$14.24(0.23)$	$\mathbf{0.93(0.15)}$	$1.09(0.15)$	$1.43(0.14)$
KMNIST	$5.51(0.22)$	$\mathbf{1.91(0.25)}$	$5.41(0.35)$	$9.17(0.40)$	$6.45(0.48)$	$7.88(0.39)$	$8.90(0.44)$
Fashion-MNIST	$\mathbf{1.60(0.31)}$	$3.38(0.53)$	$7.46(0.42)$	$10.78(0.95)$	$2.89(0.42)$	$4.07(0.36)$	$4.85(0.92)$

Table 3. Mean and standard error of the classification error over ten trials (rescaled to $0 - 100$). We used **a linear-in-input model**. The purpose of showing this table is for reference.

Dataset	CE	FL-1	FL-2	FL-3
Australian	$14.03(1.47)$	$13.54(1.27)$	$13.36(1.33)$	$13.65(1.34)$
Phishing	$7.24(0.20)$	$7.22(0.30)$	$7.47(0.20)$	$7.41(0.24)$
Spambase	$7.86(0.45)$	$8.21(0.34)$	$7.96(0.62)$	$8.52(0.48)$
Waveform	$12.00(0.37)$	$12.24(0.44)$	$12.39(0.61)$	$12.55(0.84)$
Twonorm	$2.19(0.15)$	$2.28(0.16)$	$2.22(0.20)$	$2.31(0.20)$
Adult	$15.46(0.17)$	$15.75(0.20)$	$15.71(0.26)$	$16.00(0.26)$
Banknote-authentication	$2.33(0.39)$	$1.90(0.64)$	$2.24(0.21)$	$2.13(0.43)$
Phoneme	$24.89(0.48)$	$24.98(0.68)$	$25.14(0.38)$	$25.22(0.47)$
Magic	$20.91(0.17)$	$20.98(0.37)$	$21.16(0.35)$	$20.90(0.19)$
Gisette	$2.62(0.25)$	$2.98(0.26)$	$3.37(0.24)$	$3.53(0.24)$
USPS	$1.26(0.15)$	$1.32(0.09)$	$1.22(0.08)$	$1.22(0.11)$
Splice	$16.04(0.92)$	$16.28(0.60)$	$15.65(0.59)$	$16.20(1.06)$
Banana	$43.91(3.69)$	$46.16(3.22)$	$44.34(1.44)$	$45.23(2.21)$
Ringnorm	$23.85(0.44)$	$23.95(0.65)$	$24.56(0.71)$	$24.26(0.39)$
Image	$17.81(1.13)$	$18.03(0.98)$	$19.11(0.52)$	$18.85(1.14)$
Coil20	$1.53(0.68)$	$1.17(0.78)$	$1.61(0.71)$	$1.56(0.65)$
Basehock	$3.38(0.54)$	$3.85(0.44)$	$4.77(0.66)$	$4.65(0.82)$
Isolet	$1.28(0.34)$	$1.09(0.42)$	$1.24(0.48)$	$1.18(0.36)$
W8a	$1.38(0.06)$	$1.47(0.09)$	$1.49(0.08)$	$1.64(0.06)$
Mushroom	$0.03(0.05)$	$0.01(0.01)$	$0.01(0.04)$	$0.01(0.03)$
Artificial-character	$64.55(0.44)$	$64.97(0.51)$	$64.71(0.47)$	$65.51(0.61)$
Gas-drift	$2.09(0.36)$	$2.05(0.42)$	$1.98(0.31)$	$2.06(0.42)$
Japanesevowels	$7.01(0.13)$	$6.71(0.46)$	$6.64(0.24)$	$6.80(0.27)$
Letter	$23.56(0.32)$	$23.49(0.30)$	$24.03(0.23)$	$24.69(0.46)$
Pendigits	$6.74(0.21)$	$6.25(0.23)$	$6.37(0.19)$	$6.66(0.31)$
Satimage	$15.38(0.40)$	$14.63(0.31)$	$14.93(0.50)$	$14.91(0.35)$
Vehicle	$26.12(1.33)$	$25.34(2.33)$	$26.78(1.91)$	$25.79(0.99)$
MNIST	$7.42(0.08)$	$7.53(0.07)$	$7.73(0.09)$	$7.99(0.14)$
KMNIST	$30.23(0.21)$	$30.20(0.40)$	$30.64(0.29)$	$31.06(0.41)$
Fashion-MNIST	$15.83(0.28)$	$16.21(0.40)$	$16.82(0.44)$	$16.90(0.83)$

Table 4. Mean and standard error of ECE over ten trials (rescaled to $0-100$). We used **neural networks with one hidden layer** as a model. Outperforming methods are highlighted in boldface using one-sided t-test with the significance level 5%.

Dataset	CE	FL-1	FL-2	FL-3	FL-1-$\mathbf{\Psi}^\gamma$	FL-2-$\mathbf{\Psi}^\gamma$	FL-3-$\mathbf{\Psi}^\gamma$
Australian	$\mathbf{4.31(1.21)}$	$8.15(2.10)$	$12.74(1.53)$	$18.08(1.96)$	$\mathbf{4.62(1.03)}$	$\mathbf{4.48(1.15)}$	$\mathbf{4.83(1.22)}$
Phishing	$\mathbf{0.80(0.17)}$	$3.14(0.33)$	$7.28(0.41)$	$12.16(0.40)$	$\mathbf{0.84(0.29)}$	$\mathbf{0.82(0.23)}$	$1.03(0.26)$
Spambase	$\mathbf{1.81(0.43)}$	$3.92(0.62)$	$9.61(0.51)$	$14.78(0.66)$	$\mathbf{1.79(0.51)}$	$\mathbf{1.52(0.25)}$	$\mathbf{1.52(0.29)}$
Waveform	$\mathbf{2.03(0.47)}$	$3.76(0.36)$	$7.95(0.84)$	$11.84(0.58)$	$\mathbf{1.79(0.38)}$	$\mathbf{1.74(0.57)}$	$\mathbf{1.95(0.46)}$
Twonorm	$\mathbf{0.87(0.24)}$	$1.26(0.23)$	$4.27(0.53)$	$8.02(0.70)$	$\mathbf{0.98(0.19)}$	$\mathbf{0.95(0.22)}$	$\mathbf{0.84(0.20)}$
Adult	$4.36(0.27)$	$\mathbf{2.32(0.47)}$	$8.00(0.28)$	$12.83(0.36)$	$3.99(0.41)$	$3.56(0.20)$	$3.06(0.30)$
Banknote	$1.81(0.13)$	$5.20(0.49)$	$10.99(0.69)$	$17.50(1.18)$	$\mathbf{1.64(0.19)}$	$\mathbf{1.78(0.20)}$	$2.18(0.36)$
Phoneme	$\mathbf{2.39(0.59)}$	$8.04(0.92)$	$12.88(1.03)$	$16.47(0.76)$	$\mathbf{1.97(0.55)}$	$\mathbf{2.09(0.29)}$	$\mathbf{2.01(0.54)}$
Magic	$\mathbf{1.00(0.25)}$	$8.54(0.54)$	$14.76(0.73)$	$19.69(0.52)$	$1.19(0.18)$	$\mathbf{1.04(0.21)}$	$1.24(0.30)$
Gisette	$1.20(0.21)$	$\mathbf{0.53(0.14)}$	$2.32(0.19)$	$5.03(0.23)$	$1.29(0.23)$	$1.32(0.23)$	$1.55(0.21)$
USPS	$\mathbf{0.35(0.06)}$	$0.44(0.08)$	$1.93(0.19)$	$4.60(0.25)$	$\mathbf{0.31(0.06)}$	$\mathbf{0.30(0.08)}$	$\mathbf{0.31(0.07)}$
Splice	$4.76(0.62)$	$\mathbf{1.60(0.29)}$	$5.40(0.83)$	$12.11(1.34)$	$5.07(0.74)$	$5.14(0.65)$	$4.80(0.82)$
Banana	$3.31(0.77)$	$11.17(0.62)$	$18.78(0.37)$	$24.61(0.50)$	$\mathbf{2.79(0.45)}$	$3.59(0.64)$	$5.07(0.72)$
Ringnorm	$\mathbf{0.59(0.17)}$	$4.17(0.32)$	$9.55(0.30)$	$15.71(0.67)$	$\mathbf{0.71(0.18)}$	$\mathbf{0.72(0.14)}$	$1.02(0.26)$
Image	$\mathbf{2.78(0.63)}$	$8.64(0.84)$	$15.56(0.96)$	$21.21(0.63)$	$\mathbf{2.91(0.47)}$	$\mathbf{3.04(0.55)}$	$\mathbf{3.07(0.64)}$
Coil20	$\mathbf{0.15(0.04)}$	$1.20(0.16)$	$3.49(0.27)$	$7.75(0.42)$	$\mathbf{0.18(0.07)}$	$\mathbf{0.15(0.10)}$	$\mathbf{0.14(0.03)}$
Basehock	$\mathbf{1.26(0.19)}$	$4.74(0.55)$	$9.94(0.65)$	$14.59(0.66)$	$\mathbf{1.11(0.31)}$	$1.42(0.32)$	$\mathbf{1.22(0.23)}$
Isolet	$\mathbf{0.60(0.18)}$	$1.56(0.21)$	$3.81(0.43)$	$7.43(0.35)$	$\mathbf{0.51(0.15)}$	$\mathbf{0.56(0.09)}$	$\mathbf{0.60(0.22)}$
W8a	$\mathbf{0.38(0.13)}$	$1.31(0.22)$	$3.65(0.32)$	$6.64(0.37)$	$\mathbf{0.38(0.11)}$	$\mathbf{0.39(0.08)}$	$\mathbf{0.30(0.14)}$
Mushroom	$\mathbf{0.05(0.02)}$	$0.80(0.08)$	$3.36(0.07)$	$7.25(0.10)$	$\mathbf{0.06(0.05)}$	$\mathbf{0.07(0.05)}$	$\mathbf{0.05(0.02)}$
Artificial-character	$\mathbf{3.34(0.93)}$	$4.58(0.78)$	$7.13(0.93)$	$9.08(0.80)$	$\mathbf{3.27(0.39)}$	$3.97(0.97)$	$4.79(0.72)$
Gas-drift	$0.77(0.08)$	$3.23(0.22)$	$6.48(0.41)$	$10.18(0.36)$	$\mathbf{0.64(0.13)}$	$\mathbf{0.73(0.20)}$	$0.82(0.08)$
Japanesevowels	$1.24(0.20)$	$4.22(0.17)$	$7.41(0.21)$	$10.53(0.45)$	$0.84(0.12)$	$\mathbf{0.60(0.11)}$	$0.70(0.14)$
Letter	$2.80(0.34)$	$7.66(0.33)$	$11.61(0.25)$	$15.23(0.45)$	$1.99(0.27)$	$\mathbf{1.42(0.21)}$	$\mathbf{1.59(0.25)}$
Pendigits	$1.01(0.10)$	$3.98(0.35)$	$8.38(0.39)$	$12.79(0.43)$	$0.80(0.19)$	$\mathbf{0.74(0.12)}$	$\mathbf{0.65(0.10)}$
Satimage	$\mathbf{1.30(0.36)}$	$5.62(0.84)$	$11.22(0.77)$	$15.24(1.14)$	$1.72(0.43)$	$1.68(0.48)$	$1.99(0.52)$
Vehicle	$\mathbf{5.13(1.44)}$	$8.58(1.46)$	$13.03(1.14)$	$16.70(1.03)$	$5.22(1.29)$	$\mathbf{4.43(1.12)}$	$\mathbf{4.37(0.68)}$
MNIST	$1.04(0.07)$	$\mathbf{0.42(0.05)}$	$1.67(0.07)$	$3.39(0.14)$	$1.16(0.06)$	$1.22(0.09)$	$1.34(0.09)$
KMNIST	$8.17(0.28)$	$5.42(0.15)$	$3.27(0.24)$	$\mathbf{1.22(0.24)}$	$8.12(0.16)$	$8.62(0.25)$	$9.02(0.29)$
Fashion-MNIST	$3.92(0.24)$	$\mathbf{0.92(0.20)}$	$4.14(0.59)$	$7.36(0.79)$	$4.04(0.35)$	$4.47(0.46)$	$5.05(0.66)$

Table 5. Mean and standard error of the classification error over ten trials (rescaled to $0-100$). We used **neural networks with one hidden layer** as a model. The purpose of showing this table is for reference.

Dataset	CE	FL-1	FL-2	FL-3
Australian	$13.42(1.24)$	$13.68(0.95)$	$13.80(1.46)$	$13.80(1.28)$
Phishing	$4.46(0.21)$	$4.66(0.36)$	$4.77(0.25)$	$5.49(0.16)$
Spambase	$6.20(0.40)$	$6.15(0.43)$	$6.33(0.36)$	$6.36(0.40)$
Waveform	$9.05(0.26)$	$9.21(0.26)$	$9.08(0.52)$	$9.24(0.48)$
Twonorm	$2.52(0.18)$	$2.58(0.15)$	$2.61(0.14)$	$2.54(0.19)$
Adult	$16.27(0.20)$	$16.17(0.25)$	$16.02(0.24)$	$15.72(0.22)$
Banknote-authentication	$0.60(0.50)$	$0.71(0.55)$	$0.71(0.54)$	$1.18(0.90)$
Phoneme	$16.33(0.81)$	$16.64(0.76)$	$17.11(0.54)$	$17.95(0.83)$
Magic	$13.23(0.32)$	$13.29(0.31)$	$13.69(0.49)$	$13.66(0.14)$
Gisette	$2.16(0.25)$	$2.34(0.23)$	$2.38(0.29)$	$2.61(0.21)$
USPS	$0.58(0.12)$	$0.59(0.07)$	$0.58(0.07)$	$0.58(0.09)$
Splice	$9.68(0.49)$	$10.15(0.74)$	$10.70(1.06)$	$11.58(1.17)$
Banana	$10.63(0.71)$	$10.48(0.38)$	$10.45(0.33)$	$10.77(0.35)$
Ringnorm	$2.30(0.17)$	$2.25(0.16)$	$2.25(0.22)$	$2.50(0.24)$
Image	$5.85(0.84)$	$6.16(1.14)$	$6.61(1.01)$	$7.89(0.95)$
Coil20	$0.06(0.13)$	$0.03(0.06)$	$0.03(0.08)$	$0.01(0.04)$
Basehock	$3.17(0.51)$	$3.49(0.41)$	$3.83(0.73)$	$4.07(0.67)$
Isolet	$0.73(0.24)$	$0.51(0.18)$	$0.72(0.28)$	$0.92(0.27)$
W8a	$1.05(0.04)$	$1.05(0.05)$	$1.06(0.05)$	$1.04(0.05)$
Mushroom	$0.02(0.02)$	$0.04(0.07)$	$0.06(0.07)$	$0.03(0.02)$
Artificial-character	$38.93(0.91)$	$38.90(0.74)$	$38.90(0.75)$	$39.49(0.84)$
Gas-drift	$0.87(0.14)$	$0.95(0.23)$	$0.93(0.12)$	$0.91(0.10)$
Japanesevowels	$3.07(0.33)$	$3.17(0.20)$	$3.27(0.14)$	$3.69(0.26)$
Letter	$9.54(0.41)$	$9.50(0.31)$	$9.71(0.25)$	$10.34(0.41)$
Pendigits	$1.07(0.14)$	$1.10(0.15)$	$1.16(0.17)$	$1.24(0.14)$
Satimage	$10.70(0.26)$	$11.00(0.42)$	$11.28(0.41)$	$11.11(0.52)$
Vehicle	$23.45(1.93)$	$24.44(2.03)$	$24.26(1.29)$	$25.51(1.11)$
MNIST	$2.35(0.07)$	$2.43(0.07)$	$2.40(0.08)$	$2.54(0.11)$
KMNIST	$13.03(0.25)$	$12.91(0.17)$	$13.28(0.27)$	$13.61(0.30)$
Fashion-MNIST	$11.89(0.27)$	$11.92(0.31)$	$12.07(0.14)$	$12.70(0.71)$

C.3. Additional reliability diagrams for ResNet models

Here, we present reliability diagrams for all ResNet models we used in this paper on the SVHN, CIFAR10, CIFAR10-s, and CIFAR100 datasets.

Reliability diagram index:

- Figures 8-11: Reliability diagrams for SVHN using ResNet8, ResNet20, ResNet44, and ResNet110.
- Figures 12-15: Reliability diagrams for CIFAR10 using ResNet8, ResNet20, ResNet44, and ResNet110.
- Figures 16-19: Reliability diagrams for CIFAR10-s using ResNet8, ResNet20, ResNet44, and ResNet110.
- Figures 20-23: Reliability diagrams for CIFAR100 using ResNet8, ResNet20, ResNet44, and ResNet110.

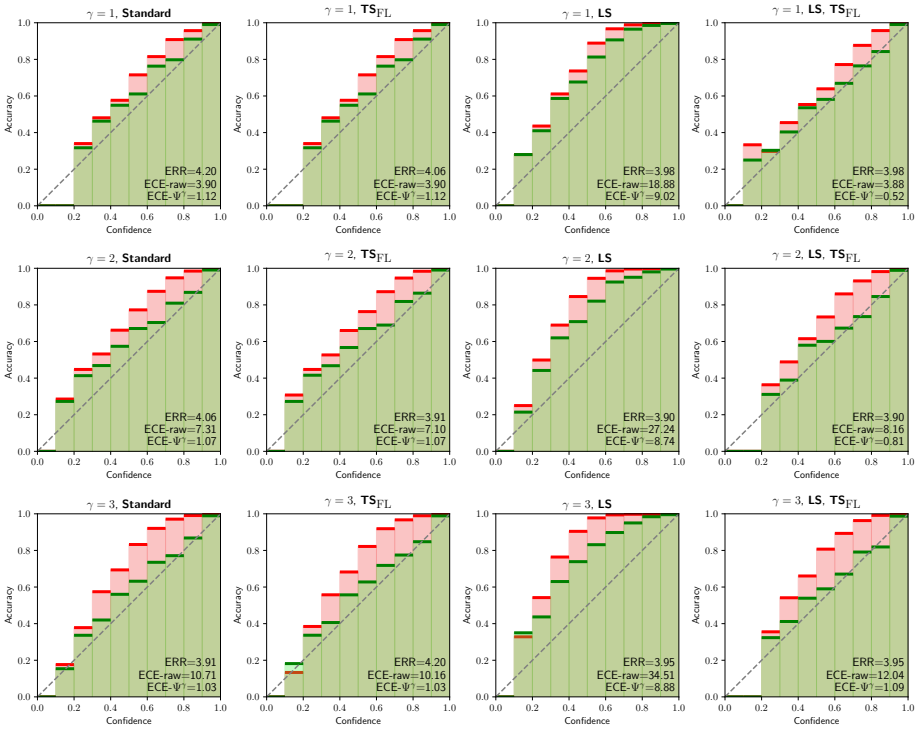

Figure 8. Reliability diagrams for SVHN and ResNet8. See each graph's title for the particular details.

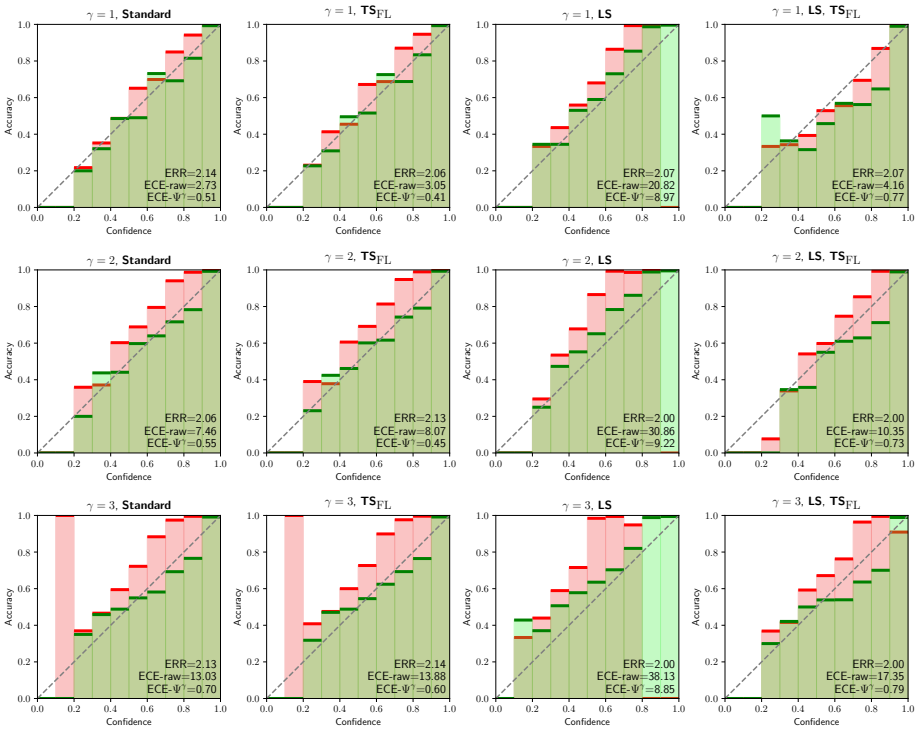

Figure 9. Reliability diagrams for SVHN and ResNet20. See each graph's title for the particular details.

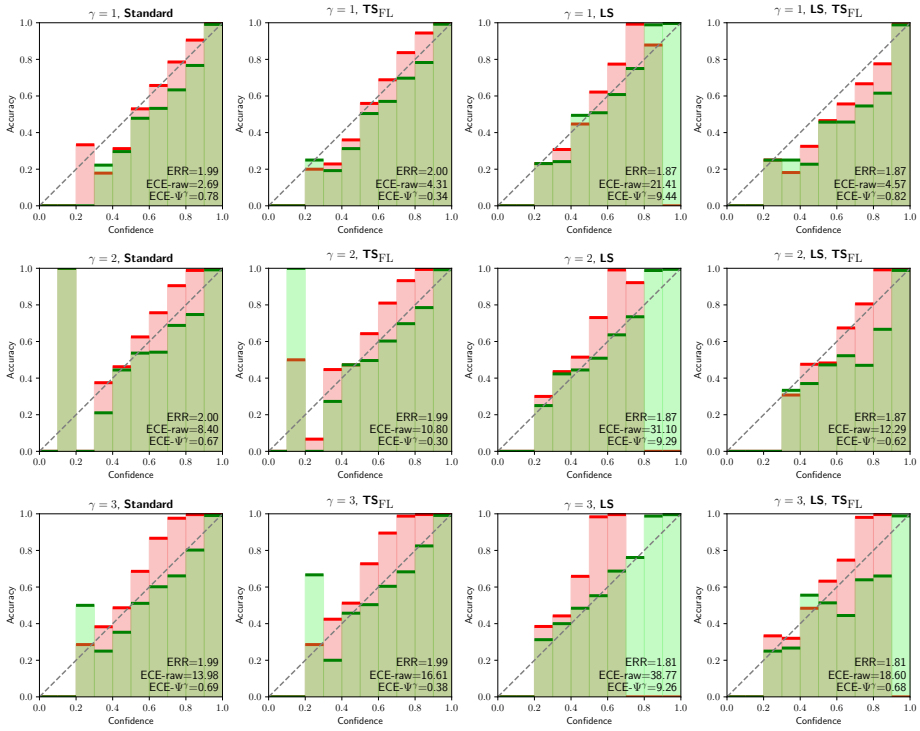

Figure 10. Reliability diagrams for SVHN and ResNet44. See each graph's title for the particular details.

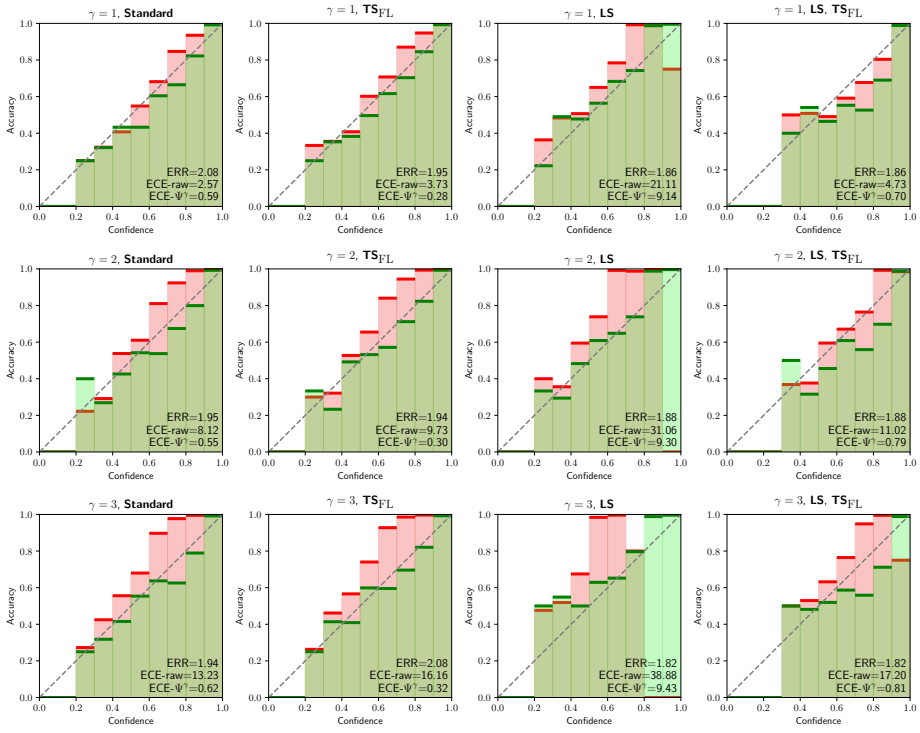

Figure 11. Reliability diagrams for SVHN and ResNet110. See each graph's title for the particular details.

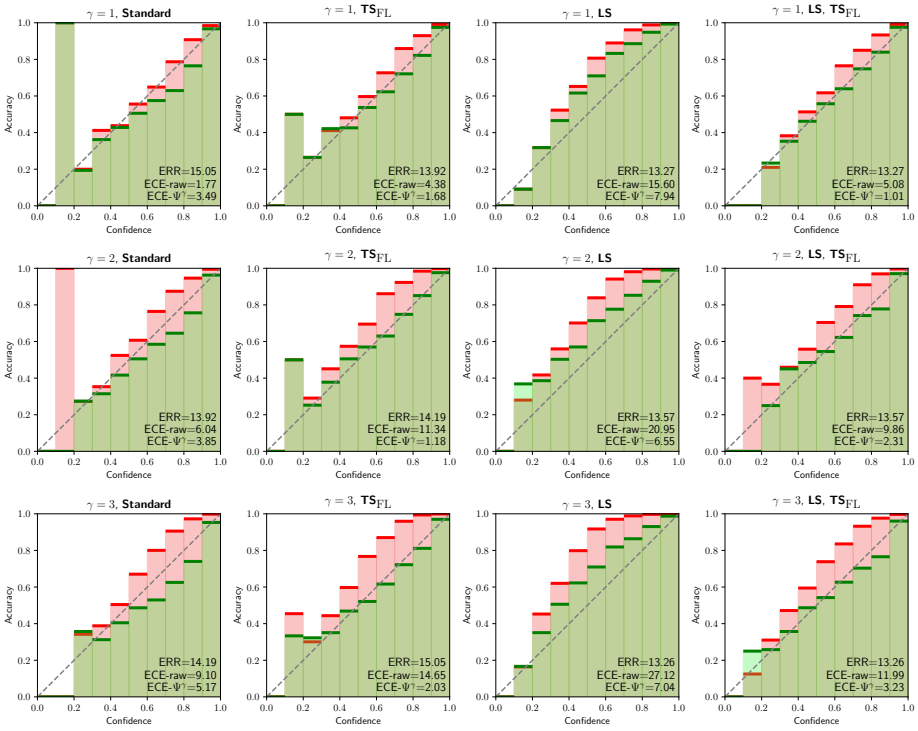

Figure 12. Reliability diagrams for CIFAR10 and ResNet8. See each graph's title for the particular details.

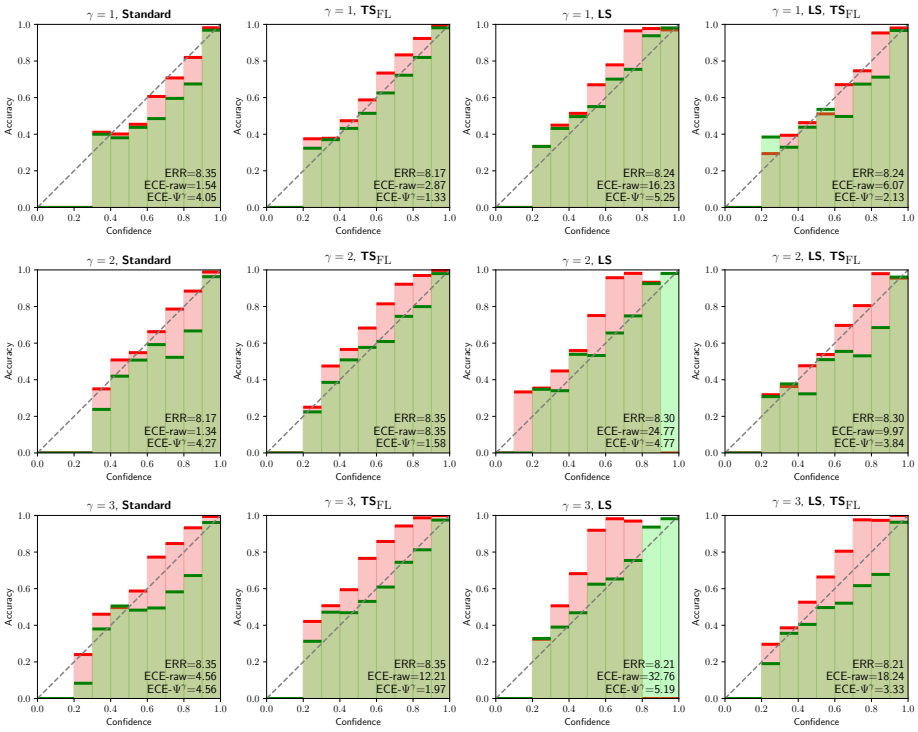

Figure 13. Reliability diagrams for CIFAR10 and ResNet20. See each graph's title for the particular details.

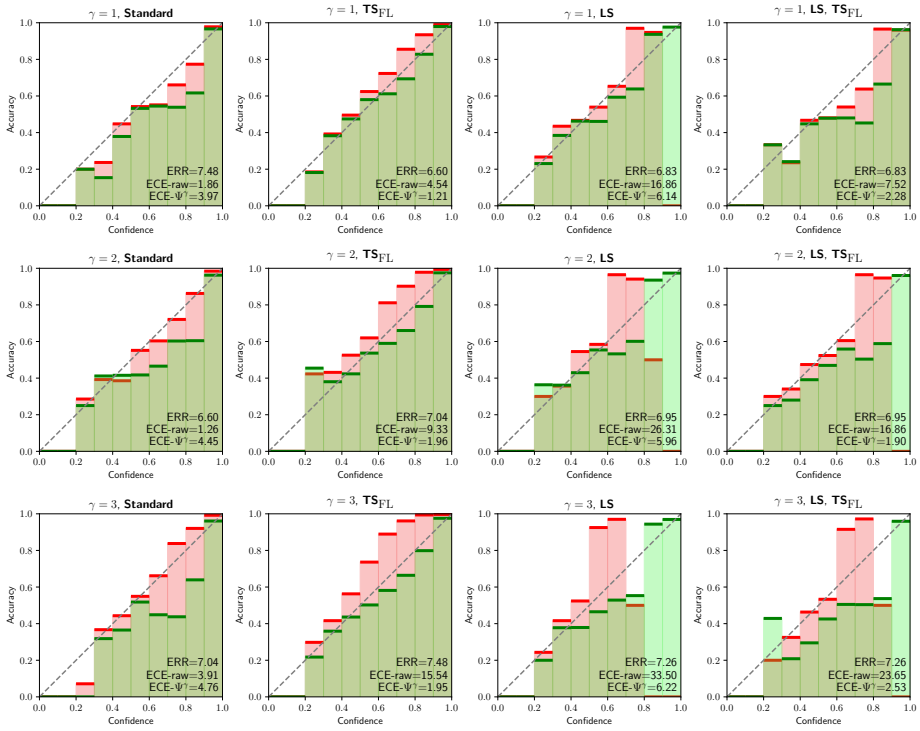

Figure 14. Reliability diagrams for CIFAR10 and ResNet44. See each graph's title for the particular details.

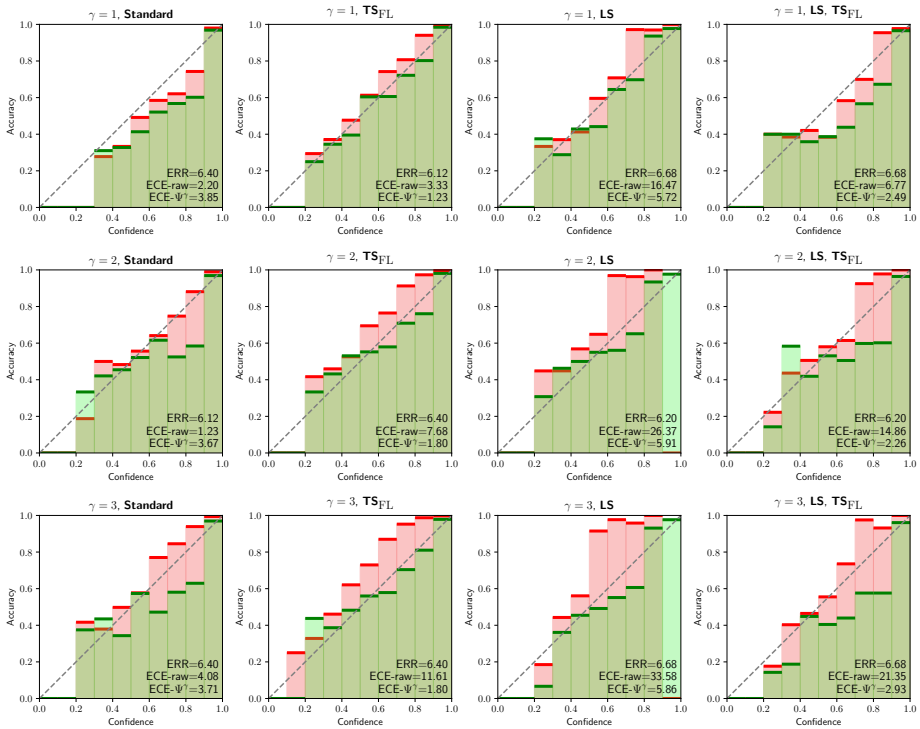

Figure 15. Reliability diagrams for CIFAR10 and ResNet110. See each graph's title for the particular details.

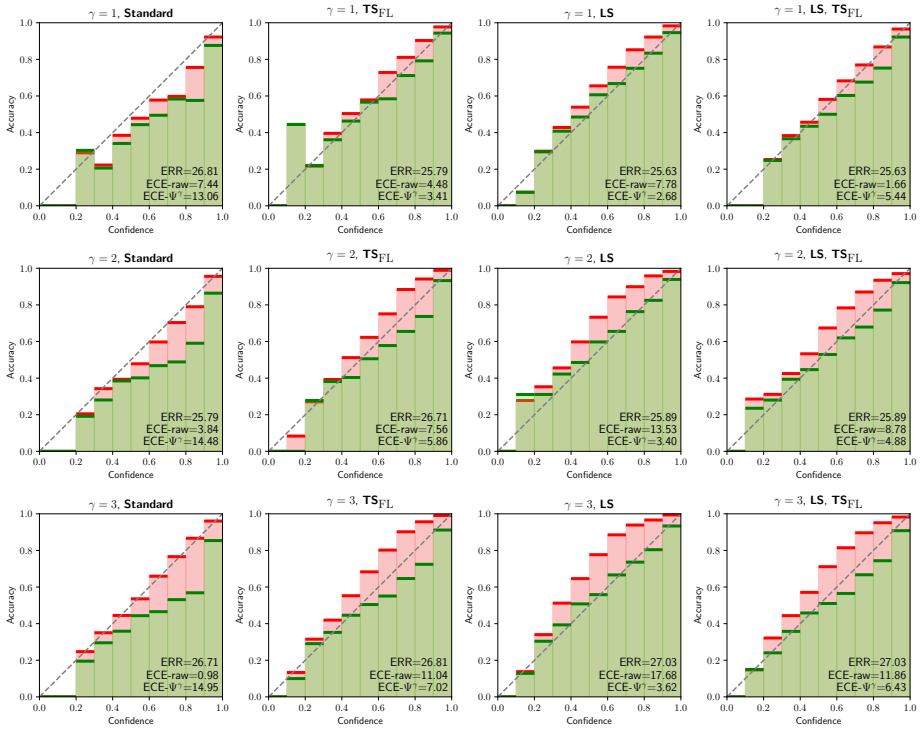

Figure 16. Reliability diagrams for CIFAR10-s and ResNet8. See each graph's title for the particular details.

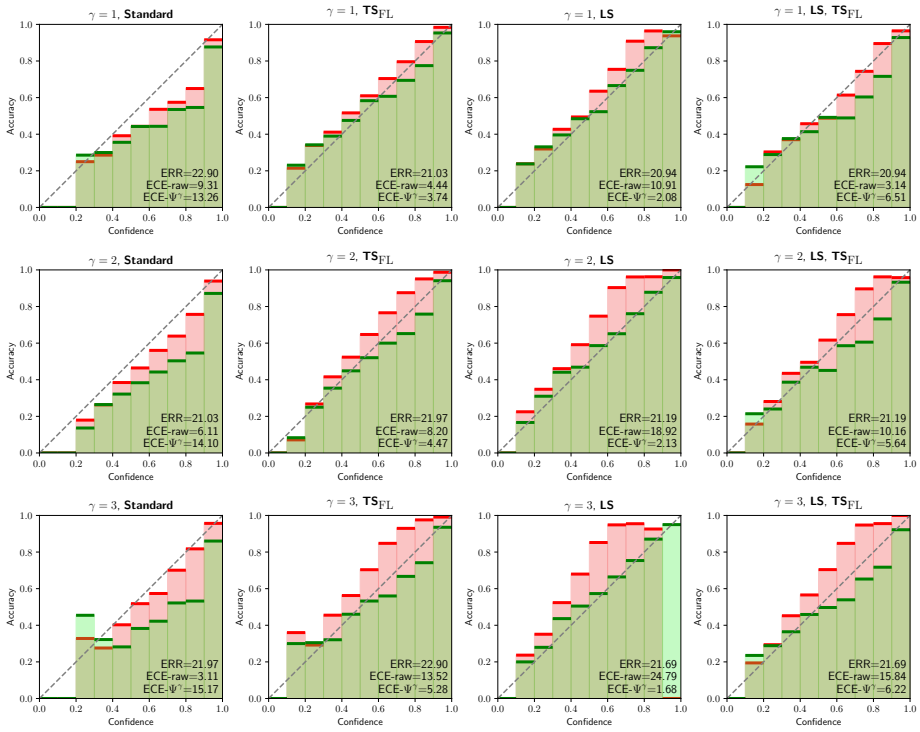

Figure 17. Reliability diagrams for CIFAR10-s and ResNet20. See each graph's title for the particular details.

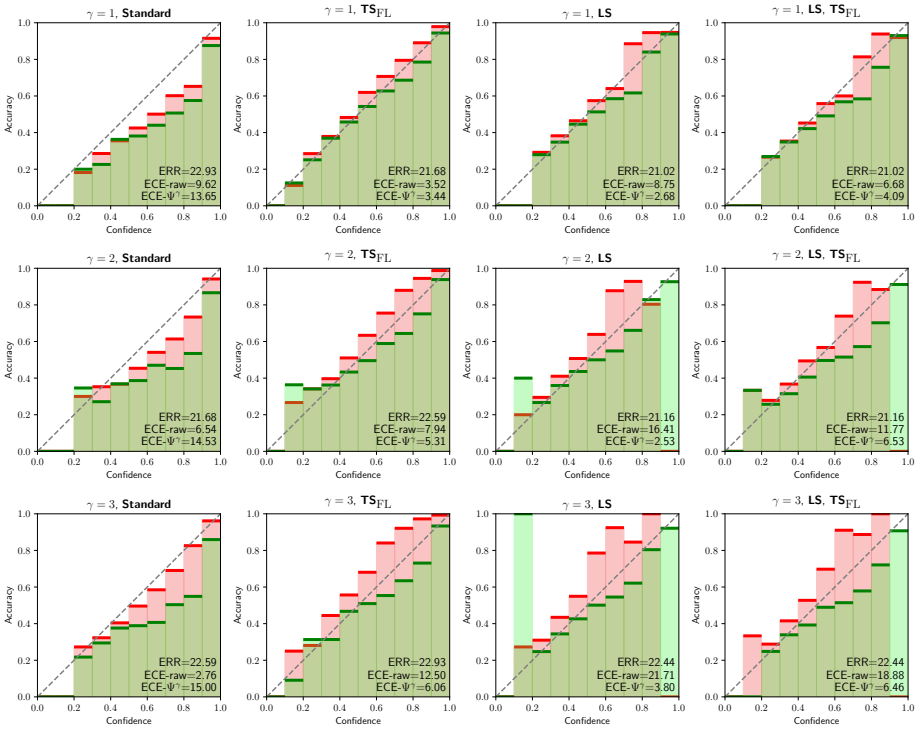

Figure 18. Reliability diagrams for CIFAR10-s and ResNet44. See each graph's title for the particular details.

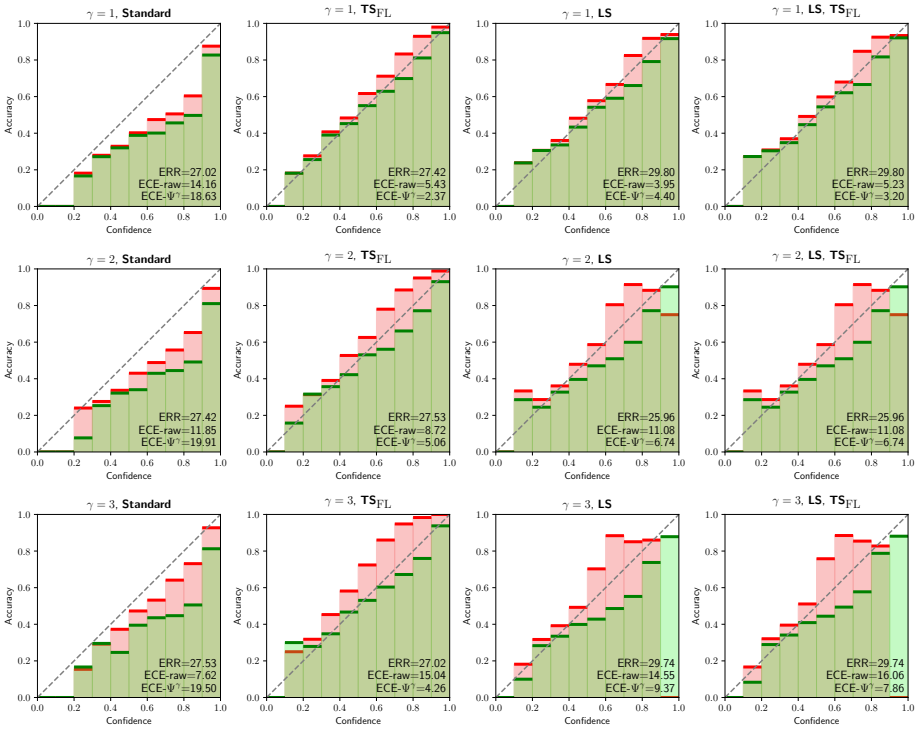

Figure 19. Reliability diagrams for CIFAR10-s and ResNet110. See each graph's title for the particular details.

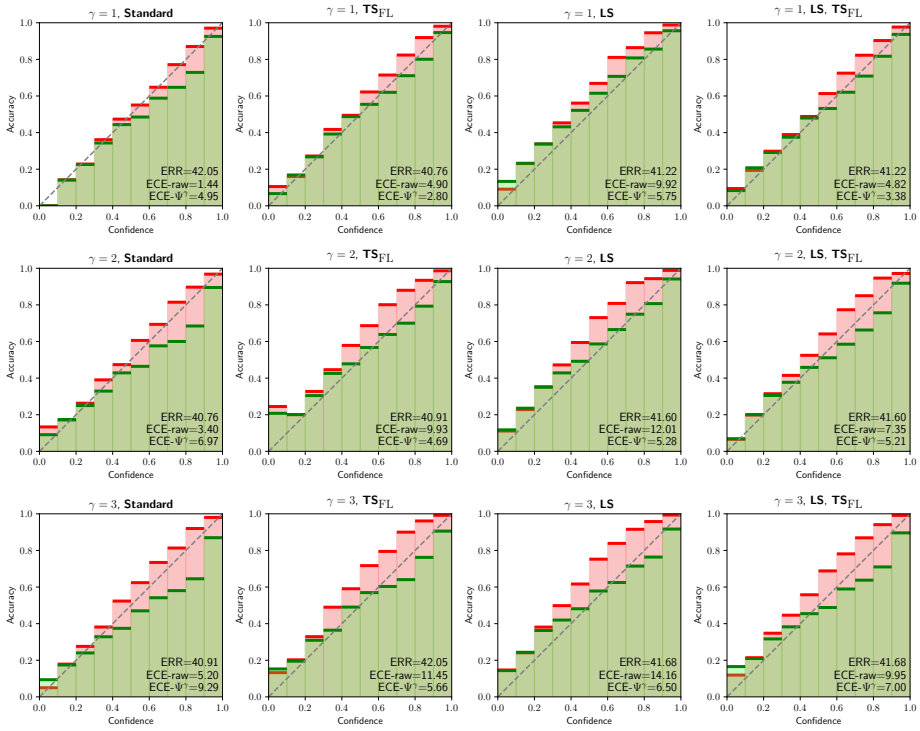

Figure 20. Reliability diagrams for CIFAR100 and ResNet8. See each graph's title for the particular details.

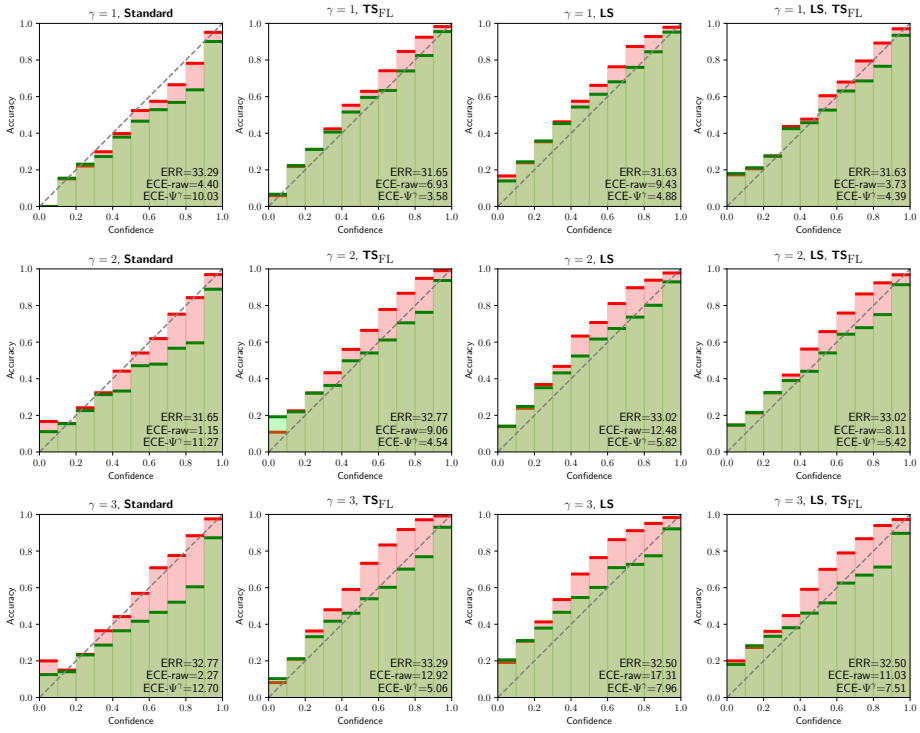

Figure 21. Reliability diagrams for CIFAR100 and ResNet20. See each graph's title for the particular details.

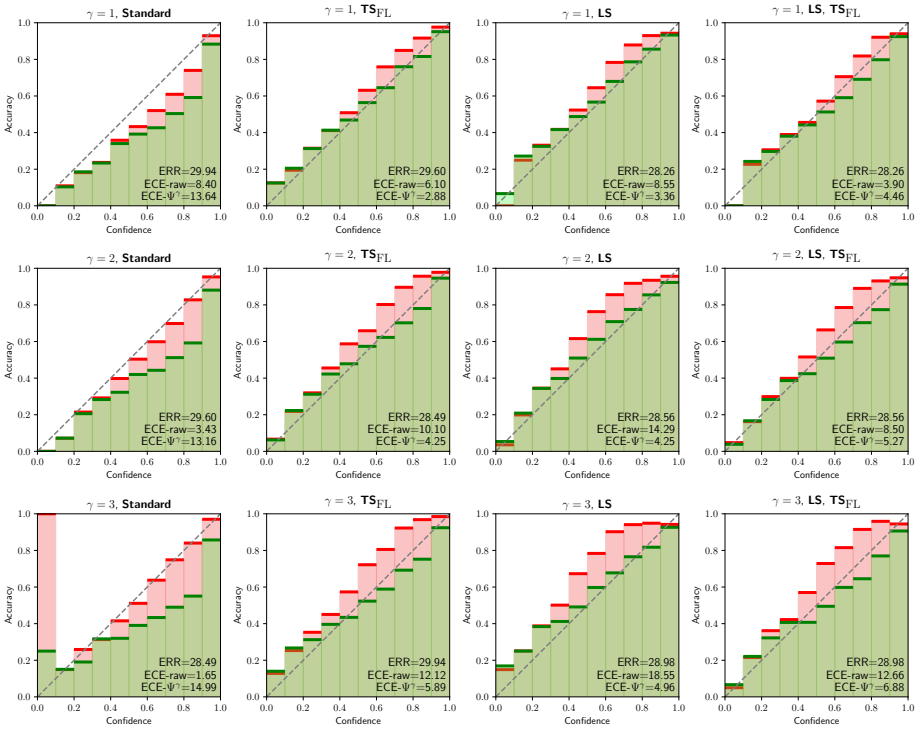

Figure 22. Reliability diagrams for CIFAR100 and ResNet44. See each graph's title for the particular details.

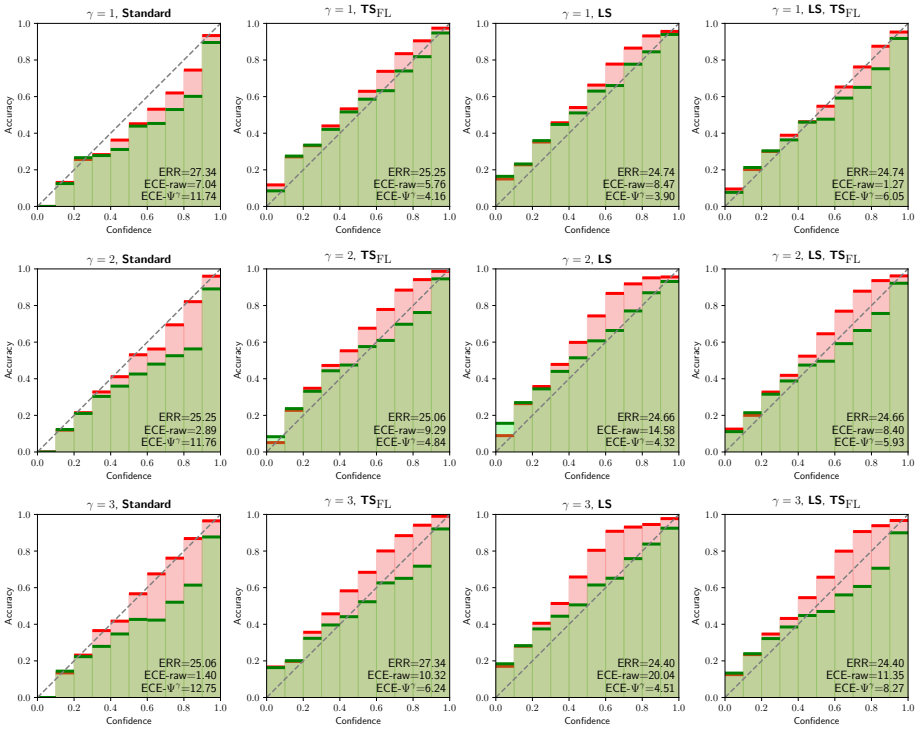

Figure 23. Reliability diagrams for CIFAR100 and ResNet110. See each graph's title for the particular details.

C.4. Additional box plots of ECE, NLL, and CW-ECE for ResNet models

Here, we present box plots of ECE, NLL, and CW-ECE for all ResNet models we used in this paper on the SVHN, CIFAR10, CIFAR10-s, and CIFAR100 datasets. More details on evaluation metrics can be found in Appx. B.

Box plot index:

- Figures 24-27: Box plots for SVHN using ResNet8, ResNet20, ResNet44, and ResNet110.
- Figures 28-31: Box plots for CIFAR10 using ResNet8, ResNet20, ResNet44, and ResNet110.
- Figures 32-35: Box plots for CIFAR10-s using ResNet8, ResNet20, ResNet44, and ResNet110.
- Figures 36-39: Box plots for CIFAR100 using ResNet8, ResNet20, ResNet44, and ResNet110.

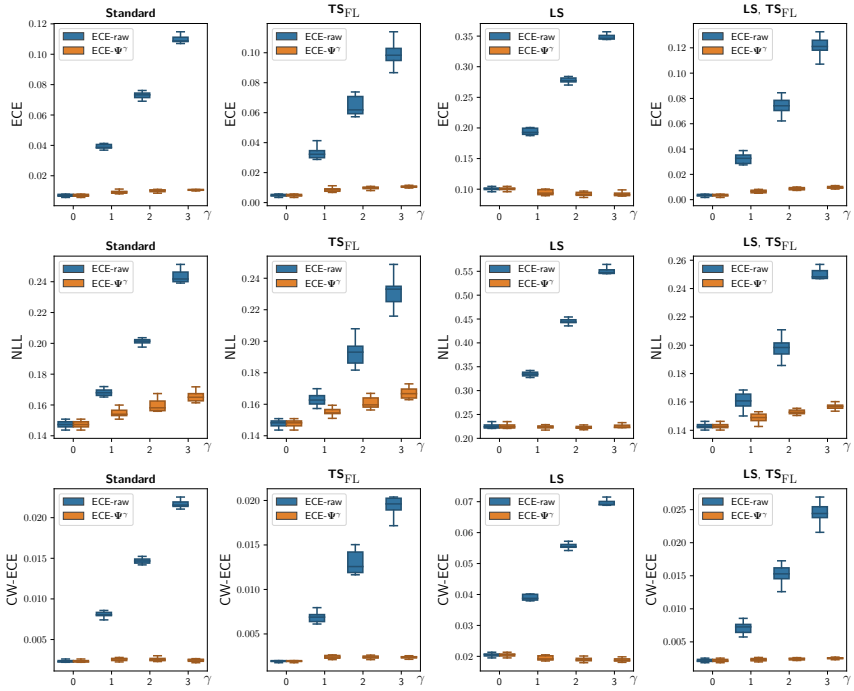

Figure 24. Box plots of ECE, NLL, and CW-ECE for SVHN and ResNet8. See each graph's title for the particular details.

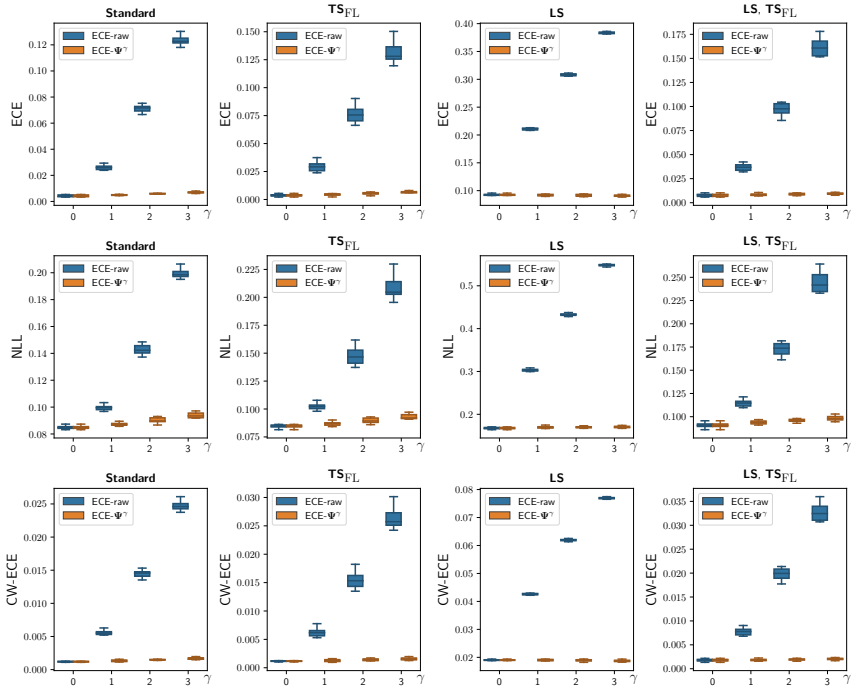

Figure 25. Box plots of ECE, NLL, and CW-ECE for SVHN and ResNet20. See each graph's title for the particular details.

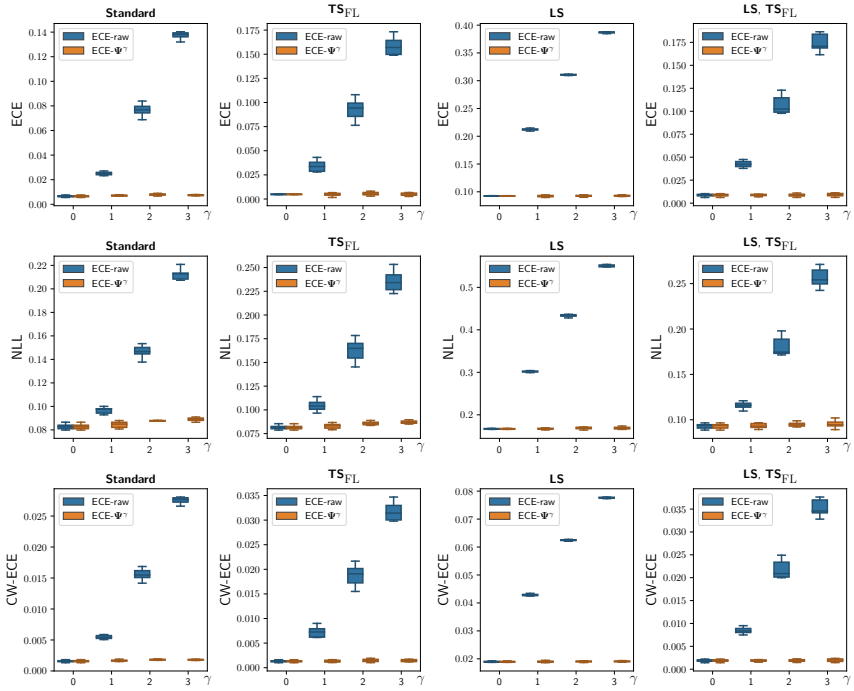

Figure 26. Box plots of ECE, NLL, and CW-ECE for SVHN and ResNet44. See each graph's title for the particular details.

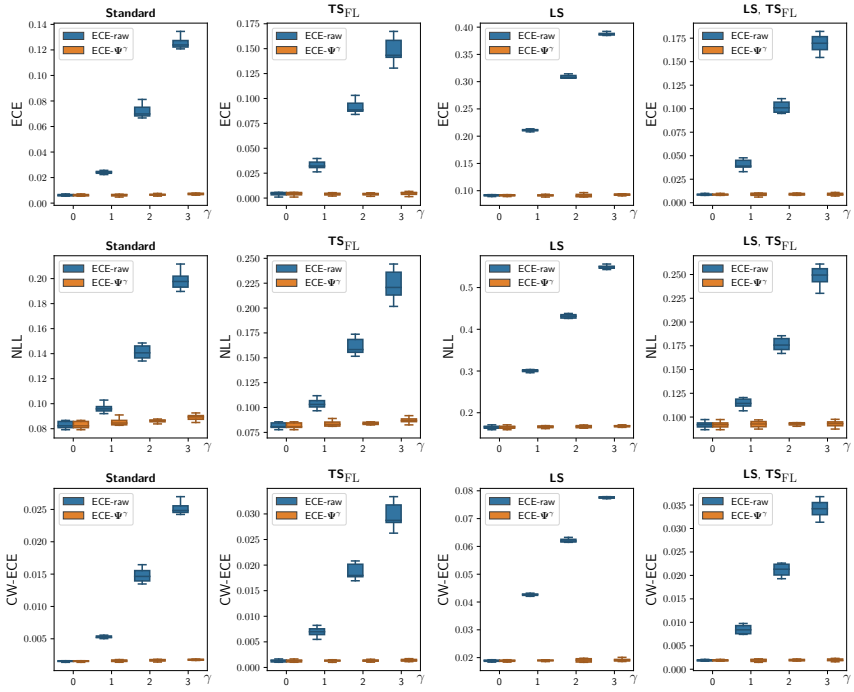

Figure 27. Box plots of ECE, NLL, and CW-ECE for SVHN and ResNet110. See each graph's title for the particular details.

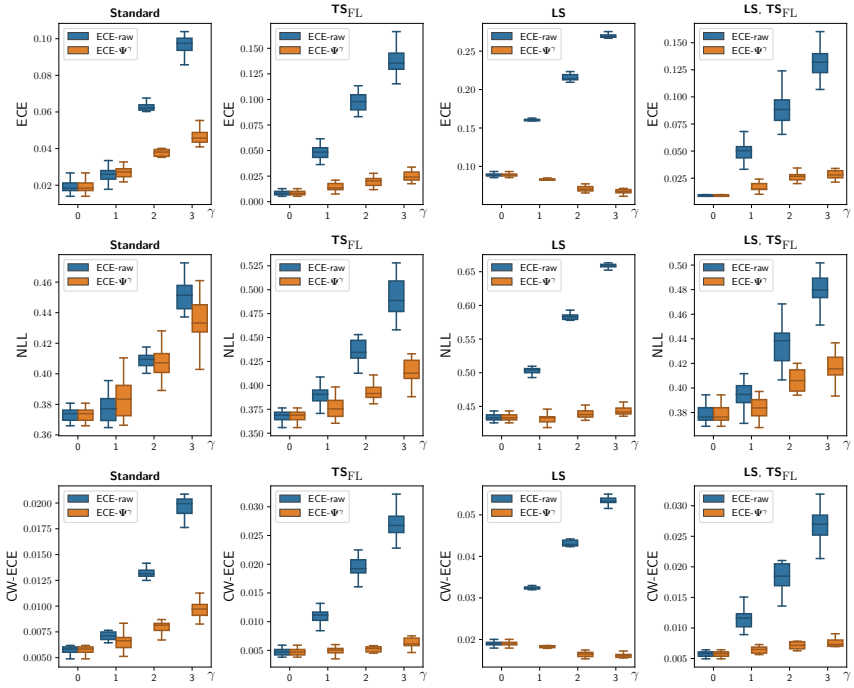

Figure 28. Box plots of ECE, NLL, and CW-ECE for CIFAR10 and ResNet8. See each graph's title for the particular details.

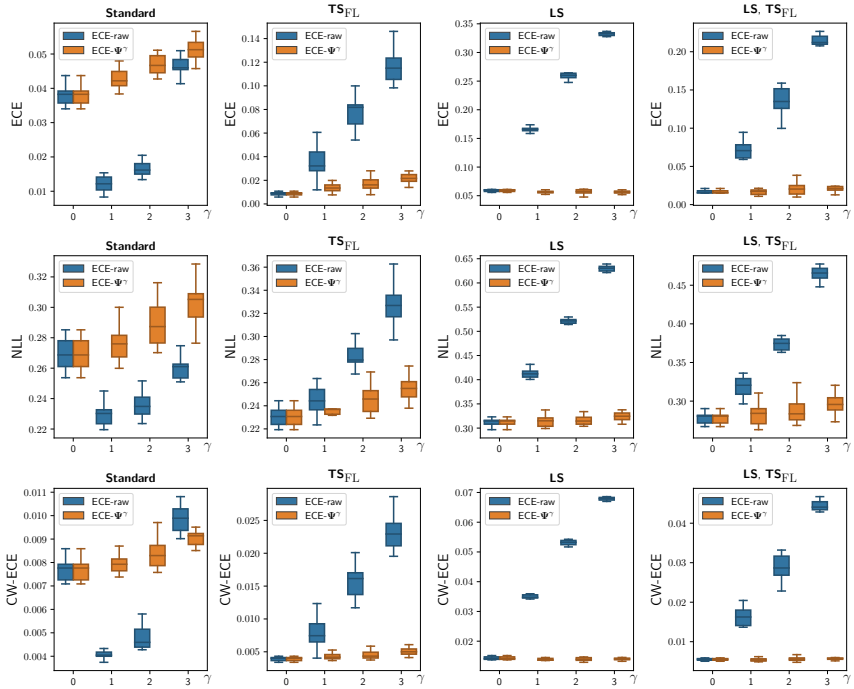

Figure 29. Box plots of ECE, NLL, and CW-ECE for CIFAR10 and ResNet20. See each graph's title for the particular details.

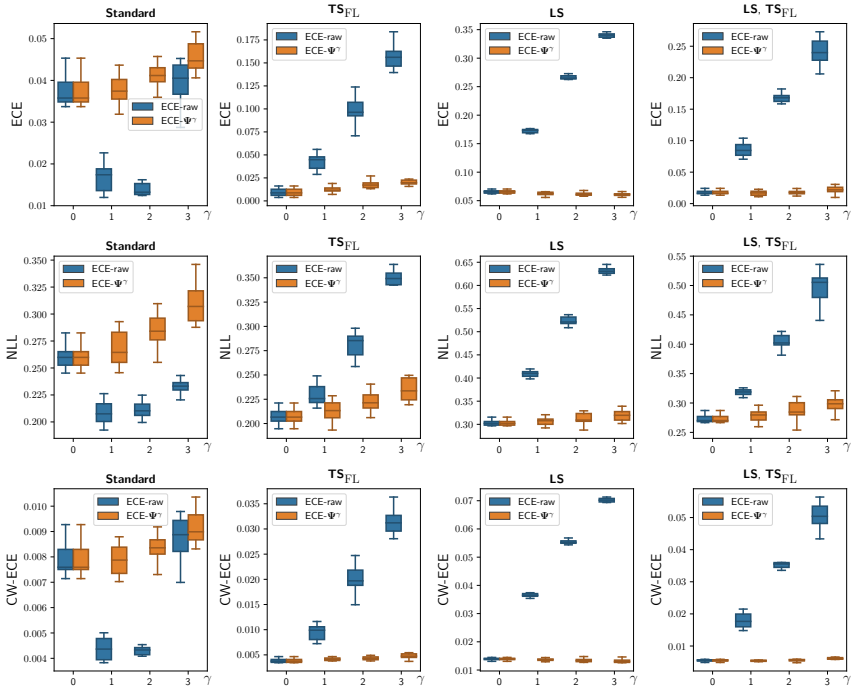

Figure 30. Box plots of ECE, NLL, and CW-ECE for CIFAR10 and ResNet44. See each graph's title for the particular details.

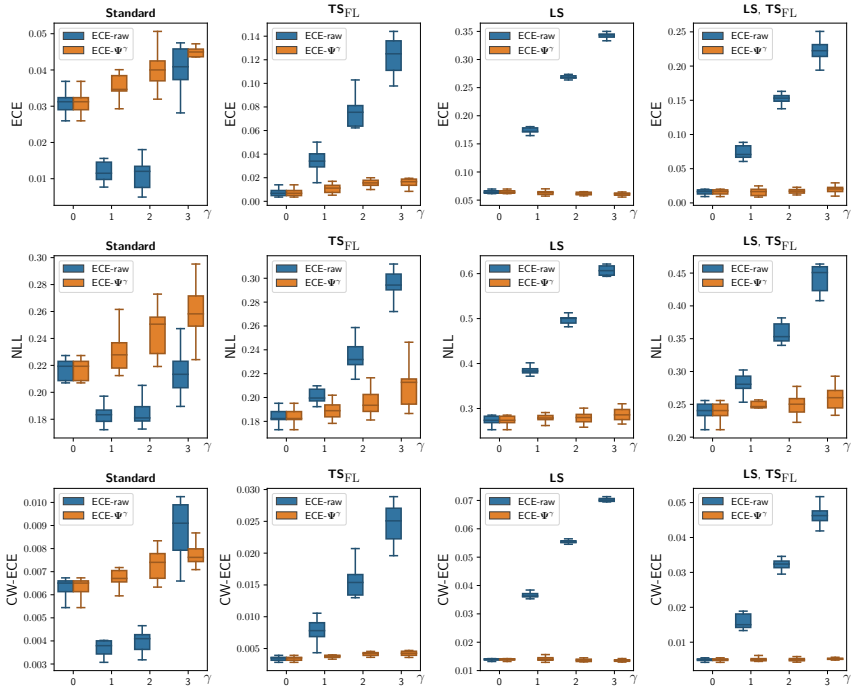

Figure 31. Box plots of ECE, NLL, and CW-ECE for CIFAR10 and ResNet110. See each graph's title for the particular details.

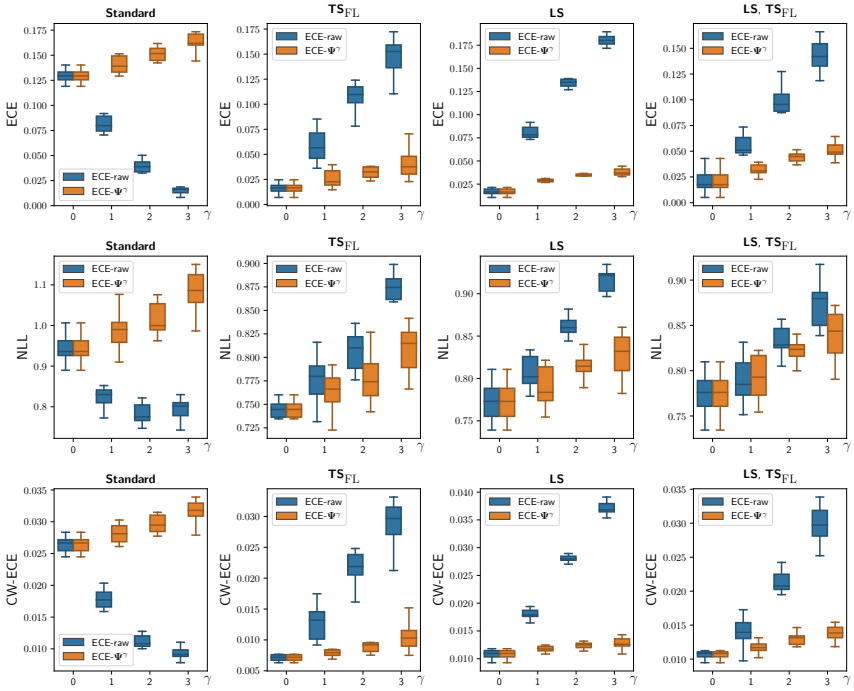

Figure 32. Box plots of ECE, NLL, and CW-ECE for CIFAR10-s and ResNet8. See each graph's title for the particular details.

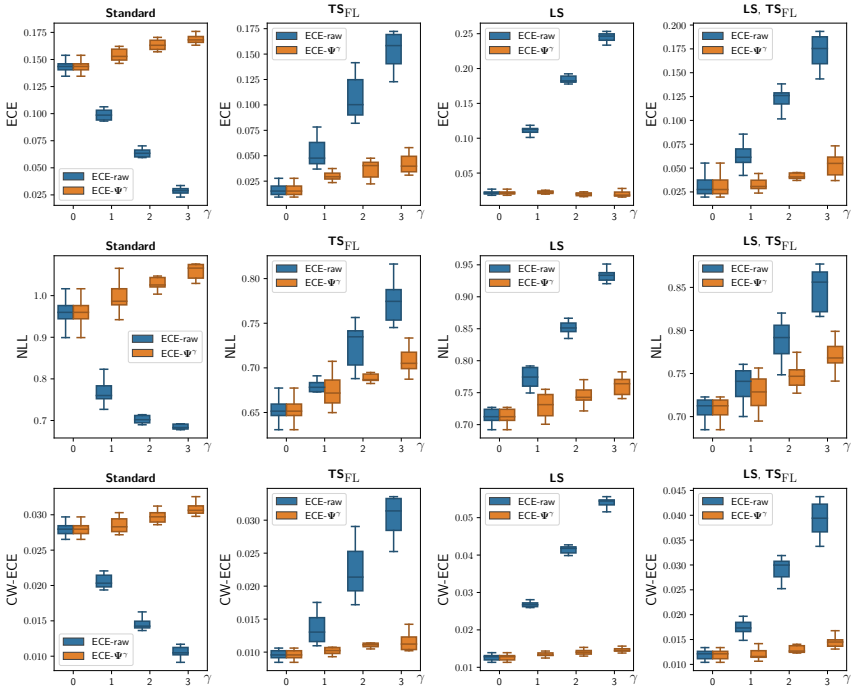

Figure 33. Box plots of ECE, NLL, and CW-ECE for CIFAR10-s and ResNet20. See each graph's title for the particular details.

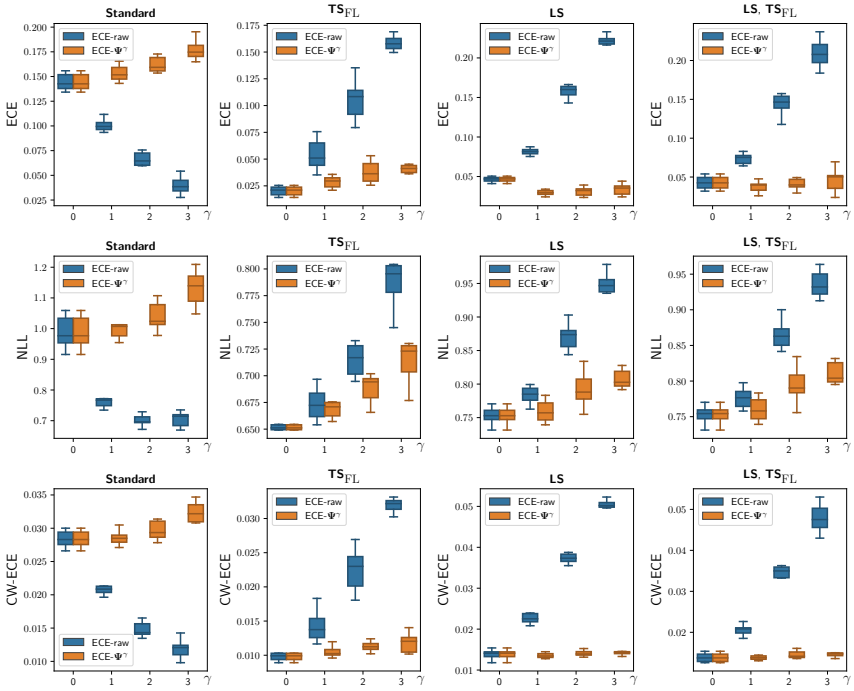

Figure 34. Box plots of ECE, NLL, and CW-ECE for CIFAR10-s and ResNet44. See each graph's title for the particular details.

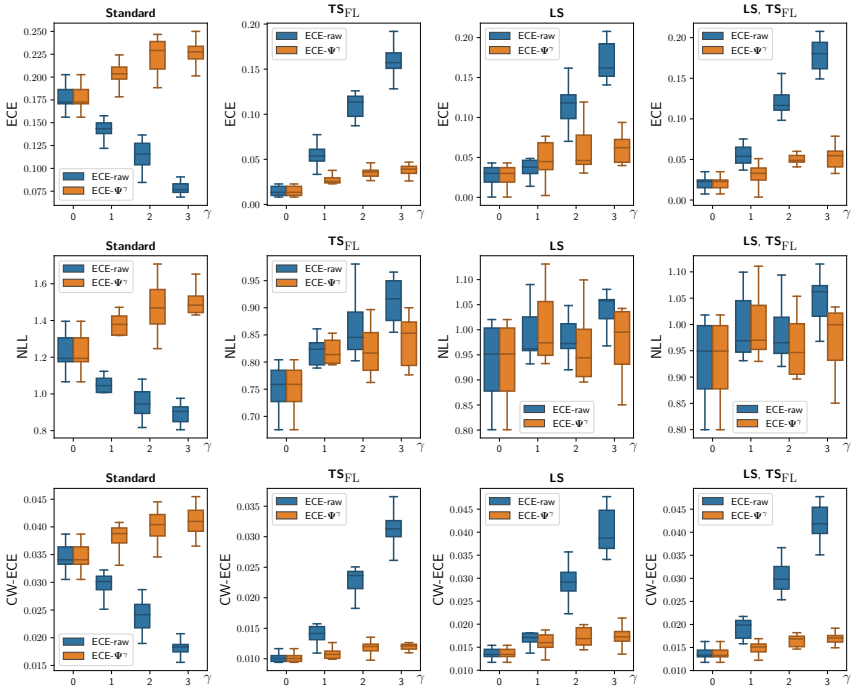

Figure 35. Box plots of ECE, NLL, and CW-ECE for CIFAR10-s and ResNet110. See each graph's title for the particular details.

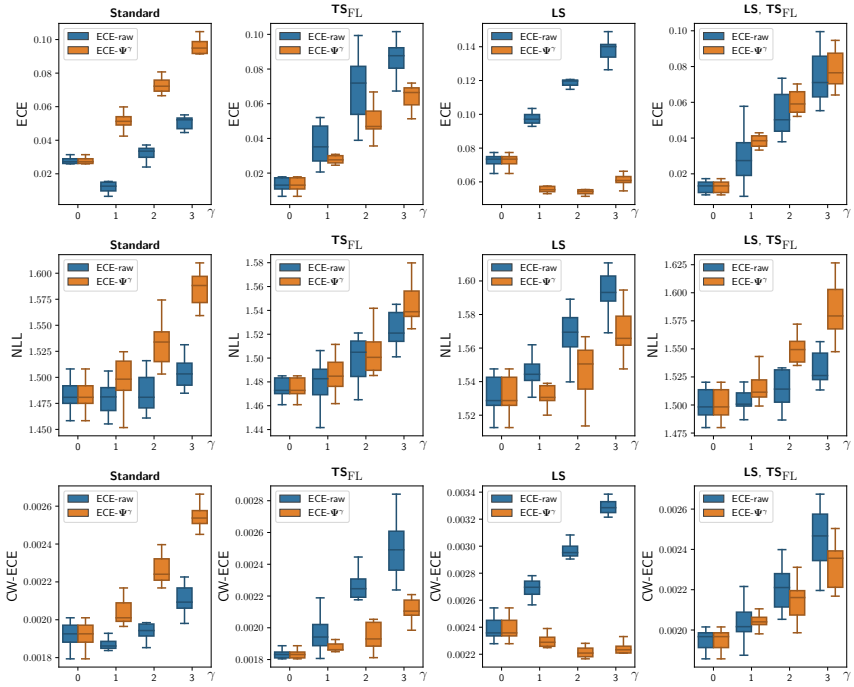

Figure 36. Box plots of ECE, NLL, and CW-ECE for CIFAR100 and ResNet8. See each graph's title for the particular details.

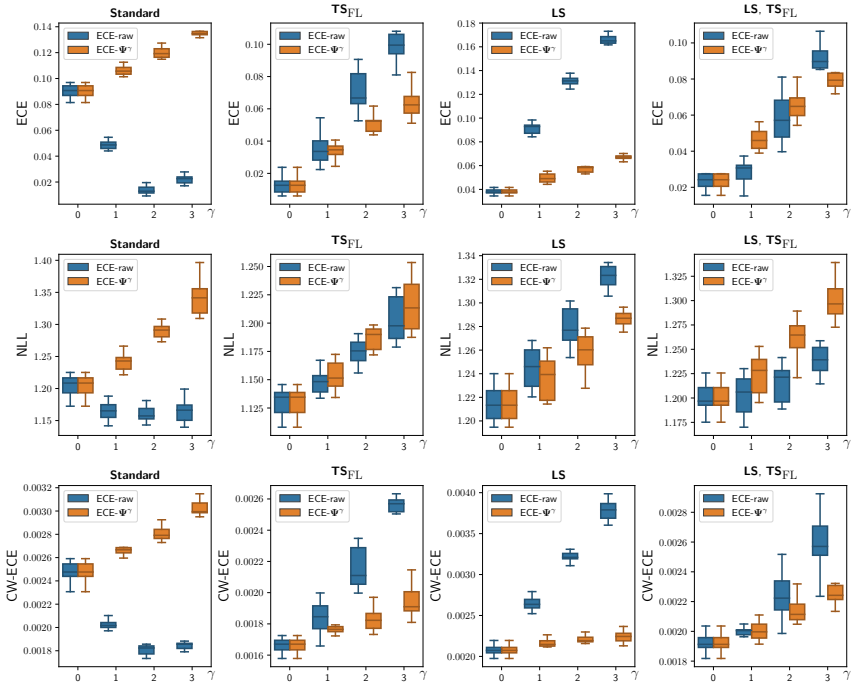

Figure 37. Box plots of ECE, NLL, and CW-ECE for CIFAR100 and ResNet20. See each graph's title for the particular details.

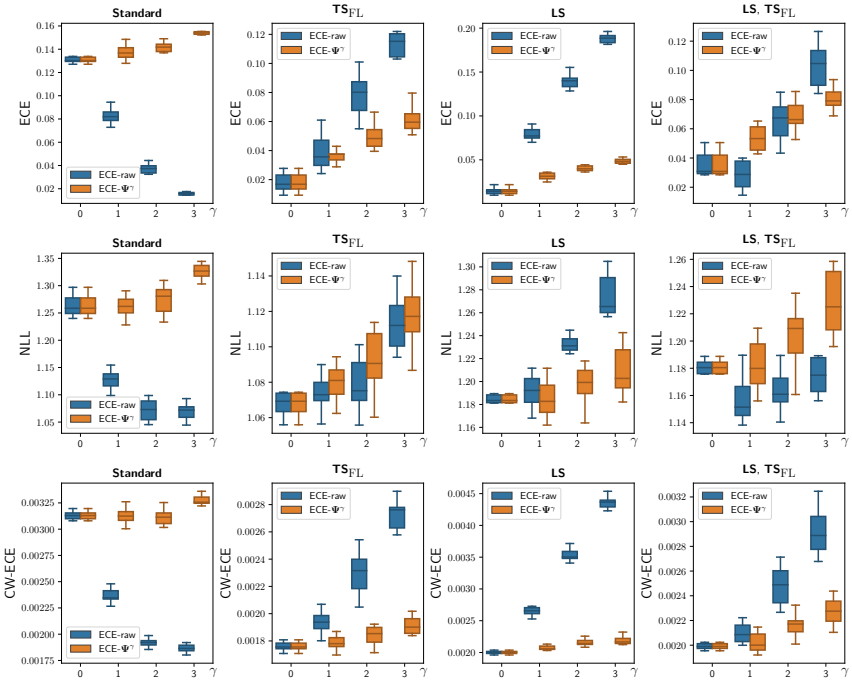

Figure 38. Box plots of ECE, NLL, and CW-ECE for CIFAR100 and ResNet44. See each graph's title for the particular details.

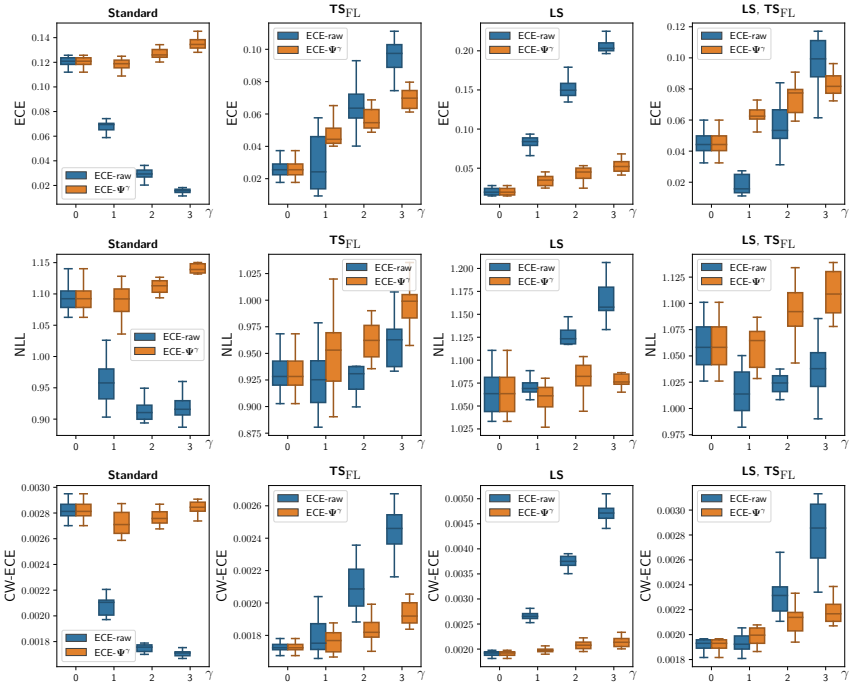

Figure 39. Box plots of ECE, NLL, and CW-ECE for CIFAR100 and ResNet110. See each graph's title for the particular details.