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

Nontawat Charoenphakdee*[1,2], Jayakorn Vongkulbhisal*[3], Nuttapong Chairatanakul[4,5], Masashi Sugiyama[2,1]

The University of Tokyo[1], RIKEN AIP[2], IBM Research[3], Tokyo Institute of Technology[4], RWBC-OIL (AIST)[5]

CVPR2021

Multiclass classification

$\mathcal{Y} = \{1, 2, \ldots, K\}$

$\boldsymbol{e}_y$: One-hot vector

Given: input-output pairs:

$$\{\boldsymbol{x}_i, y_i\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x}, y)$$

We train a classifier $g : \mathcal{X} \to \Delta^K$ by minimizing the empirical risk:

$$\hat{R}^{\ell}(g) = \frac{1}{n} \sum_{i=1}^{n} \ell\left(g(\boldsymbol{x}_i), e_{y_i}\right).$$

Loss function ℓ highly influences the behavior of the trained classifier.

- A good classifier should predict the most probable class.
- But is this enough?

Example

MNIST dataset

1

$p(y|\boldsymbol{x})$

0: 0.0
1: 1.0
2: 0.0
⋮
9: 0.0

1,2, or 7?

$p(y|\boldsymbol{x})$

0: 0.0
1: 0.3
2: 0.5
⋮
7: 0.2
⋮
9: 0.0

Medical decision

Need surgery?

$$p(y|\boldsymbol{x})$$

$$\begin{array}{l} y = +1 \\ y = -1 \end{array} \begin{pmatrix} .90 \\ .10 \end{pmatrix}$$

Class-posterior probability $p(y|\boldsymbol{x})$ provides confidence score.

Q: What loss can give $p(y|\boldsymbol{x})$?

Cross-entropy loss

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

$\boldsymbol{u} \in \Delta^K, \boldsymbol{v} \in \Delta^K$

* $\boldsymbol{u}$ is usually a one-hot label in practice

CE loss is **classification-calibrated**, i.e.,

CE risk minimizer gives the most probable class (Bayes-optimal):

$$\arg\max_y \boldsymbol{q}^*_{\mathrm{CE}}(\boldsymbol{x}) = \arg\max_y p(y|\boldsymbol{x}).$$

CE loss is **strictly proper**, i.e.,

CE risk minimizer is a class-posterior probability estimator:

$$\boldsymbol{q}^*_{\mathrm{CE}}(\boldsymbol{x}) = p(y|\boldsymbol{x}).$$

Q: What about focal loss?

Need surgery?

$$\begin{array}{ccc} & p(y|\boldsymbol{x}) & \boldsymbol{q}^*_{\mathrm{CE}}(\boldsymbol{x}) \\ \begin{array}{c} y=+1 \\ y=-1 \end{array} & \begin{pmatrix} .90 \\ .10 \end{pmatrix} & \begin{pmatrix} .90 \\ .10 \end{pmatrix} \end{array}$$

*Strictly properness is sufficient to guarantee classification-calibration.

Focal loss (Lin+, ICCV 2017)

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

Originally proposed for dense object detection.

CE loss is a special case when $\gamma = 0$.

Focal loss has been used in many applications, e.g.,

- Electrocardiogram classification (Al Rahhal+, 2019)
- Brain tumor segmentation (Chang+, 2019)
- Femur fractures classification. (Lotfy+, 2019)

Problem: theoretical understanding of focal loss is limited.

Q1: Is focal loss classification-calibrated?

Q2: Is focal loss strictly proper?

Main result

Focal loss is **classification-calibrated**:

$$\arg\max_y \boldsymbol{q}^*_{\mathrm{FL},\gamma}(\boldsymbol{x}) = \arg\max_y p(y|\boldsymbol{x}).$$

However, it is **not strictly proper for** $\gamma > 0$:

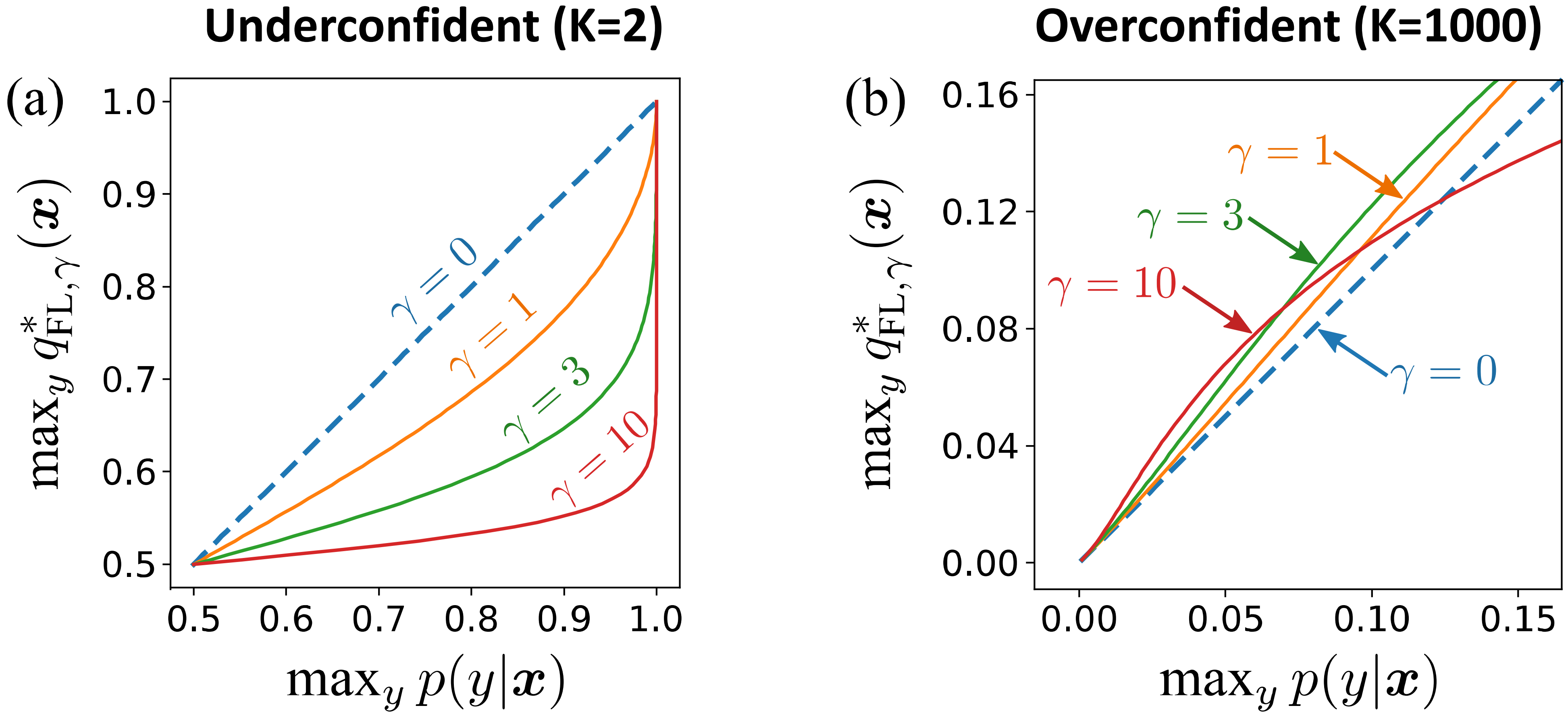
$$\boldsymbol{q}^*_{\mathrm{FL},\gamma}(\boldsymbol{x}) \neq p(y|\boldsymbol{x}).$$

Need surgery?

$$\begin{array}{cccccc} & p(y|\boldsymbol{x}) & \boldsymbol{q}^*_{\mathrm{CE}}(\boldsymbol{x}) & \boldsymbol{q}^*_{\mathrm{FL},1}(\boldsymbol{x}) & \boldsymbol{q}^*_{\mathrm{FL},3}(\boldsymbol{x}) & \boldsymbol{q}^*_{\mathrm{FL},5}(\boldsymbol{x}) \\ \begin{array}{c} y=+1 \\ y=-1 \end{array} & \begin{pmatrix} .90 \\ .10 \end{pmatrix} & \begin{pmatrix} .90 \\ .10 \end{pmatrix} & \begin{pmatrix} .78 \\ .22 \end{pmatrix} & \begin{pmatrix} .65 \\ .35 \end{pmatrix} & \begin{pmatrix} .60 \\ .40 \end{pmatrix} \end{array}$$

We can predict the most probable class, but **confidence score is unreliable**!

Focal risk minimizer can be both under/overconfident



Q: How to solve this problem?

(Please see our paper for more detail.)

Solution: Recover $p(y|\boldsymbol{x})$ from $\boldsymbol{q}^*_{\mathrm{FL},\gamma}(\boldsymbol{x})$ via $\boldsymbol{\Psi}^\gamma$

Need surgery?

$$
\begin{array}{cccccc}
 & p(y|\boldsymbol{x}) & \boldsymbol{q}^*_{\mathrm{CE}}(\boldsymbol{x}) & \boldsymbol{q}^*_{\mathrm{FL},1}(\boldsymbol{x}) & \boldsymbol{q}^*_{\mathrm{FL},3}(\boldsymbol{x}) & \boldsymbol{q}^*_{\mathrm{FL},5}(\boldsymbol{x}) \\
\begin{array}{c} y=+1 \\ y=-1 \end{array} & \begin{pmatrix} .90 \\ .10 \end{pmatrix} & \begin{pmatrix} .90 \\ .10 \end{pmatrix} & \begin{pmatrix} .78 \\ .22 \end{pmatrix} & \begin{pmatrix} .65 \\ .35 \end{pmatrix} & \begin{pmatrix} .60 \\ .40 \end{pmatrix} \\
 & & & \boldsymbol{\Psi}^1 \downarrow & \boldsymbol{\Psi}^3 \downarrow & \boldsymbol{\Psi}^5 \downarrow \\
 & & & \begin{pmatrix} .90 \\ .10 \end{pmatrix} & \begin{pmatrix} .90 \\ .10 \end{pmatrix} & \begin{pmatrix} .90 \\ .10 \end{pmatrix}
\end{array}
$$

Define $\boldsymbol{\Psi}^\gamma(\boldsymbol{v}) = [\Psi^\gamma_1(\boldsymbol{v}), \ldots, \Psi^\gamma_K(\boldsymbol{v})]^\top$,

where $\Psi^\gamma_i(\boldsymbol{v}) = \frac{h^\gamma(v_i)}{\sum_{l=1}^K h^\gamma(v_l)}$,

and $h^\gamma(v_i) = \frac{v_i}{(1-v_i)^\gamma - \gamma(1-v_i)^{\gamma-1} v_i \log v_i}$.

- **Closed-form**
- **No hyperparameter**
- **Theoretically justified**
- **Preserves accuracy**
- **No additional training required**

(Please see our paper for experiments on benchmark datasets.)

Conclusions

Theoretical analysis of focal loss with practical use.

Q1: Is focal loss classification-calibrated?

Yes!

Q2: Is focal loss strictly proper?

No! Directly using model's output gives **unreliable confidence**.

Q3: Following Q2, can we do anything about it?

Yes! We discovered a closed-form transformation $\mathbf{\Psi}^{\gamma}$ that can recover $p(y|\boldsymbol{x})$ **with theoretical guarantee!**