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]

CVPR VIRTUAL JUNE 19-25

Summary

Theoretical analysis of focal loss with practical use.

Q1: Does focal risk minimizer give Bayes-optimal classifier
Yes!

Q2: Does focal risk minimizer match class-posterior probability $p(y|\boldsymbol{x})$?
No! Directly using model's output gives **unreliable confidence**.

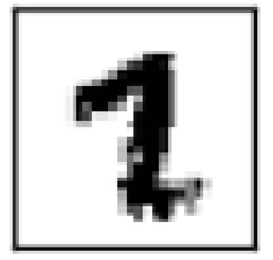
Q3: Following Q2, can we do anything about it?
Yes! We discovered a closed-form transformation $\boldsymbol{\Psi}^\gamma$ that can recover $p(y|\boldsymbol{x})$ **with theoretical guarantee!**

Introduction

MNIST classification

Medical decision

$\boldsymbol{x}$



y 1,2, or 7? Need surgery?

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

$\boldsymbol{q}^*_\ell = \mathrm{argmin}_{\boldsymbol{q}}\, \mathbb{E}_{y\sim p(y|\boldsymbol{x})}[\ell(\boldsymbol{q}(\boldsymbol{x}), \boldsymbol{e}_y)]$
$\boldsymbol{q}^*_{\mathrm{CE}}$: Cross-entropy risk minimizer
$\boldsymbol{q}^*_{\mathrm{FL},\gamma}$: Focal risk minimizer

$\boldsymbol{q}(\boldsymbol{x}) \in \Delta^K$
$\boldsymbol{e}_y$: One-hot vector

- Bayes-optimal classifier predicts the most probable class $\arg\max_y p(y|\boldsymbol{x})$.
- Class-posterior probability $p(y|\boldsymbol{x})$ provides useful confidence score.
- ***Loss function highly influences the behavior of the trained model.***

Example: The well-studied cross-entropy (CE) loss for K-class classification:

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

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

$\boldsymbol{u}$ is often a one-hot vector

CE loss is
- **classification-calibrated**: CE risk minimizer $\boldsymbol{q}^*_{\mathrm{CE}}(\boldsymbol{x})$ gives Bayes optimal classifier.
- **strictly proper**: CE risk minimizer $\boldsymbol{q}^*_{\mathrm{CE}}(\boldsymbol{x})$ gives class-posterior probability.

Q: What about theoretical properties of focal loss?

Focal loss

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

$\gamma \geq 0$

(Lin+, 2017)

- Originally proposed for dense object detection.
- Many practical applications in the medical field.

(Al Rahhal+, 2019, Chang+, 2018, Lotfy+, 2019, Sun+, 2019)

Main results

Focal loss is **classification-calibrated** for $\gamma \geq 0$:

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

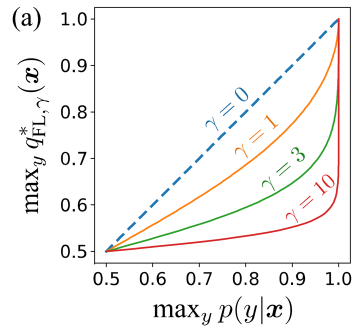
However, it is **not strictly proper for $\gamma > 0$**:

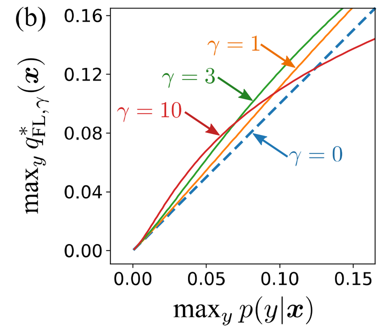
$$\boldsymbol{q}^*_{\mathrm{FL},\gamma}(\boldsymbol{x}) \neq p(y|\boldsymbol{x}).$$

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

Example:

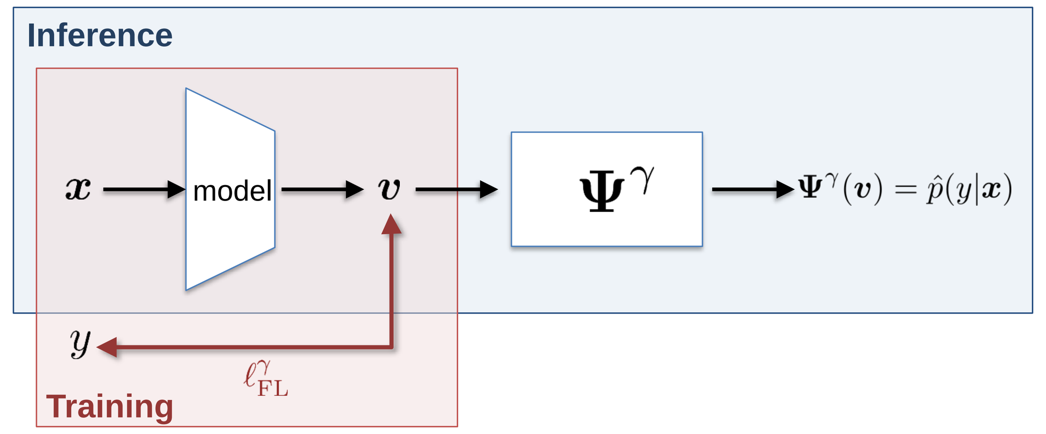
Underconfident (K=2) **Overconfident (K=1000)**





(Please see our paper for more detail.)

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



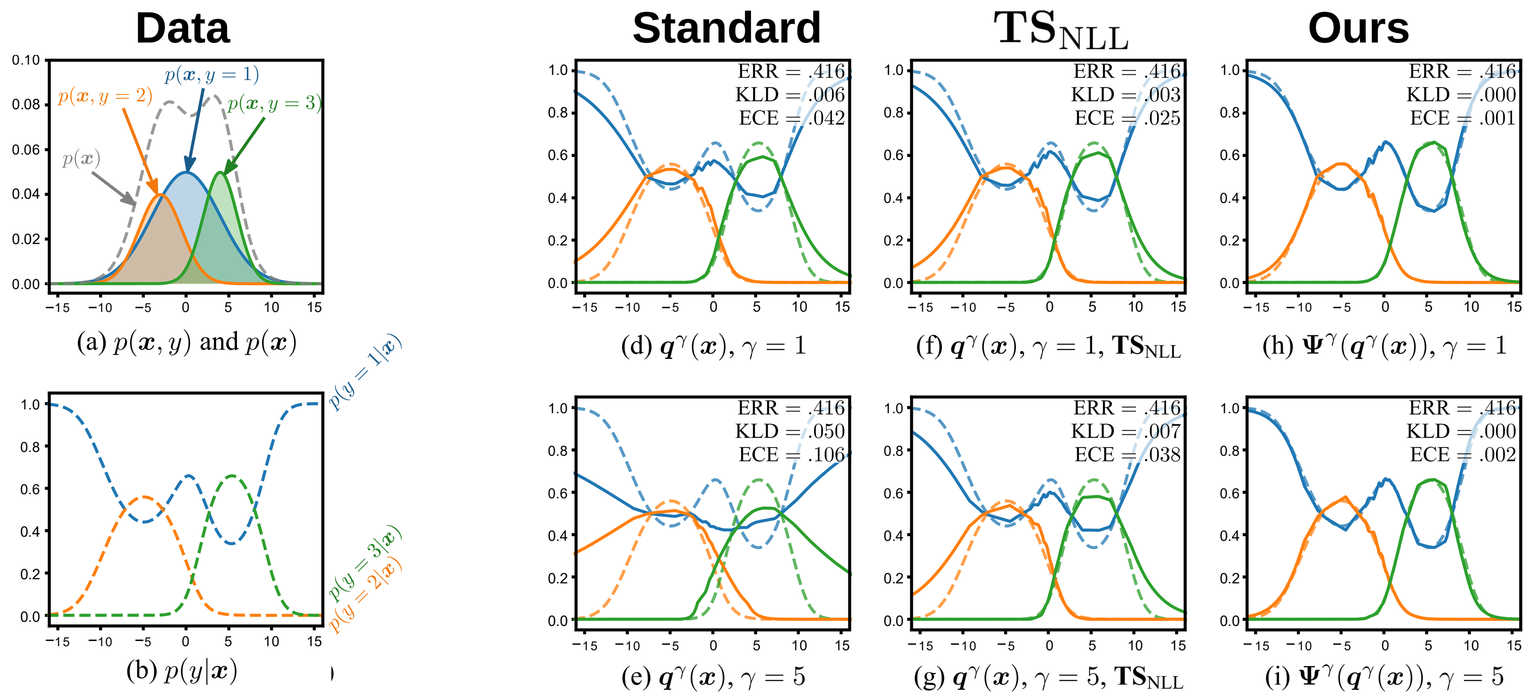
Define $\boldsymbol{\Psi}^\gamma(\boldsymbol{v}) = [\Psi^\gamma_1(\boldsymbol{v}), \ldots, \Psi^\gamma_K(\boldsymbol{v})]^\top$,

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

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

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

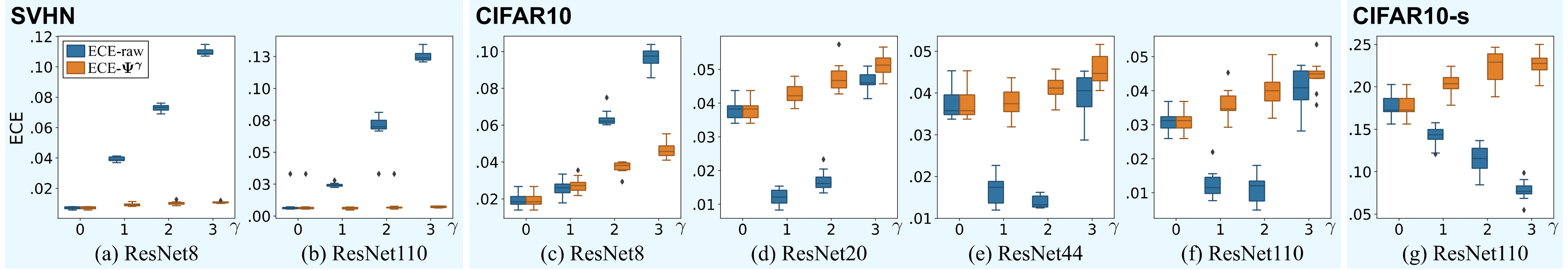
Numerical simulation



- Bigger γ makes the network $\boldsymbol{q}^\gamma$ more prone to be underconfident in **Standard**.
- Using temperature scaling $\mathbf{TS}_{\mathrm{NLL}}$ (Guo+, 2017) is insufficient to recover $p(y|\boldsymbol{x})$.
- With our proposed $\boldsymbol{\Psi}^\gamma$, we can recover $p(y|\boldsymbol{x})$ (almost) perfectly.

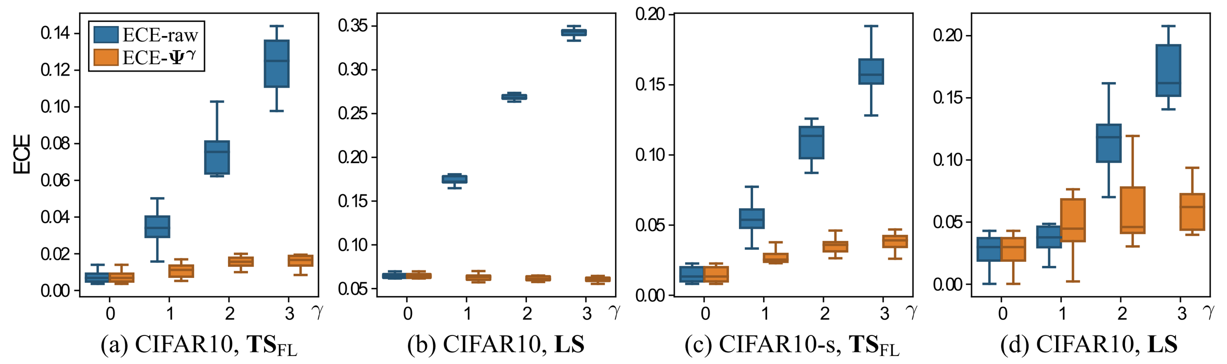
Benchmark experiments

Evaluation metric: Expected calibration error (ECE) (Naeini+, 2015)



- Using $\boldsymbol{\Psi}^\gamma$ is effective when we have good approximation of $\boldsymbol{q}^*_{\mathrm{FL},\gamma}(\boldsymbol{x})$ (Fig. a-c)
- But it is less effective when having small data or model architecture is too large (Fig. d-g)

With focal-loss-based temperature scaling $\mathbf{TS}_{\mathrm{FL}}$ or label smoothing $\mathbf{LS}$:



Using $\boldsymbol{\Psi}^\gamma$ is preferable for both cases.

**We used ResNet110 for Fig. a-d. Same trend can be observed for all models in our paper (ResNet8-110).*

References

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In ICML, pages 1321–1330. JMLR, 2017.
Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. Focal loss for dense object detection. ICCV, 2017.
Naeini, M.P., Cooper, G. and Hauskrecht, M. Obtaining well calibrated probabilities using bayesian binning. AAAI, 2015.
Al Rahhal, M.M., Bazi, Y., Almubarak, H., Alajlan, N. and Al Zuair, M. Dense convolutional networks with focal loss and image generation for electrocardiogram classification. IEEE Access, 2019.
Chang, J., Zhang, X., Ye, M., Huang, D., Wang, P. and Yao, C. Brain tumor segmentation based on 3D Unet with multi-class focal loss. CISP-BMEI, 2018.
Lotfy, M., Shubair, R.M., Navab, N. and Albarqouni, S. Investigation of Focal Loss in Deep Learning Models For Femur Fractures Classification. ICECTA, 2019.
Sun, X., Dong, K., Ma, L., Sutcliffe, R., He, F., Chen, S. and Feng, J. Drug-drug interaction extraction via recurrent hybrid convolutional neural networks with an improved focal loss. Entropy, 2019.