Mitigating Underfitting in Learning to Defer with Consistent Losses

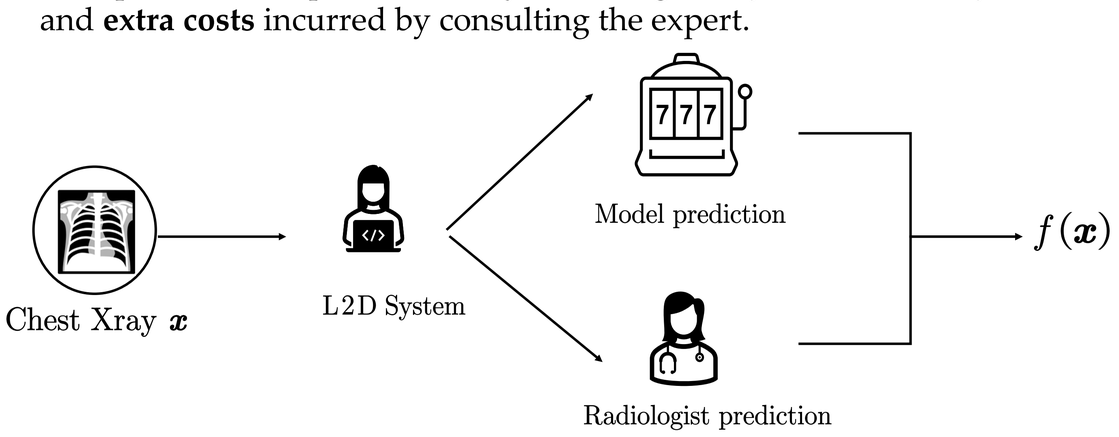
Shuqi Liu[1†] Yuzhou Cao[2†] Qiaozhen Zhang[1] Lei Feng[2] Bo An[3,2]

[1]Nankai University [2]Nanyang Technological University [3]Skywork AI

† Equal contribution

Learning to Defer (L2D)

Learning to defer (L2D) allows the classifier to defer its prediction to an expert for safer predictions, by balancing **the system's accuracy** and **extra costs** incurred by consulting the expert.



$$\text{Find } f^* = \operatorname*{argmin}_f \ \mathbb{E}_{p(\boldsymbol{x},y,m)}[\mathbb{I}(f(\boldsymbol{x})\neq\perp \text{ and } f(\boldsymbol{x})\neq y)+\mathbb{I}(f(\boldsymbol{x})=\perp \text{ and } m\neq y)]$$

$$+\mathbb{E}_{p(\boldsymbol{x})}[c\mathbb{I}(f(\boldsymbol{x})=\perp)] \Rightarrow \text{Extra deferral cost} \qquad \Downarrow \text{ Classification Error}$$

Goal:

- Optimize Classifier to adapt to Human's weaknesses and strengths.

Intuitive Explanation of *f:**

- ***f*(x)*** is ⊥, i.e., choose to defer to expert and output *m*, if expert's accuracy against classifier can offset the cost:

$$\boxed{\text{Acc}(m|\boldsymbol{x})=\Pr(m=y|\boldsymbol{x})\geq\max_y\ p(y|\boldsymbol{x})+c}$$

- Otherwise ***f*(x)*** is the Bayes optimal $y^*=\operatorname{argmax}_y\ p(y|\boldsymbol{x})$

Underfitting Issues in L2D

Existing L2D Surrogates

- Consistent loss formulation [5]

$$\min_{\boldsymbol{g}}\mathbb{E}_{p(\boldsymbol{x},y,m)}\left[\Psi(\boldsymbol{g}(\boldsymbol{x}),y)+c\mathbb{I}(m\neq y)\sum_{y'\in[K]}\Psi(\boldsymbol{g}(\boldsymbol{x}),y')+(1-c)\mathbb{I}(m=y)\Psi(\boldsymbol{g}(\boldsymbol{x}),K+1)\right]$$

- Treat deferral option ⊥ as a new class
- Ψ is any consistent multiclass loss
- $\boldsymbol{g}$ is a *K+1* dimensional scoring function. First K dimensions $\boldsymbol{g}_{[1:K]}$ is a classifier.
- Essence: Problem reduction to *K+1* class classification on distribution $\hat{p}$ [4]

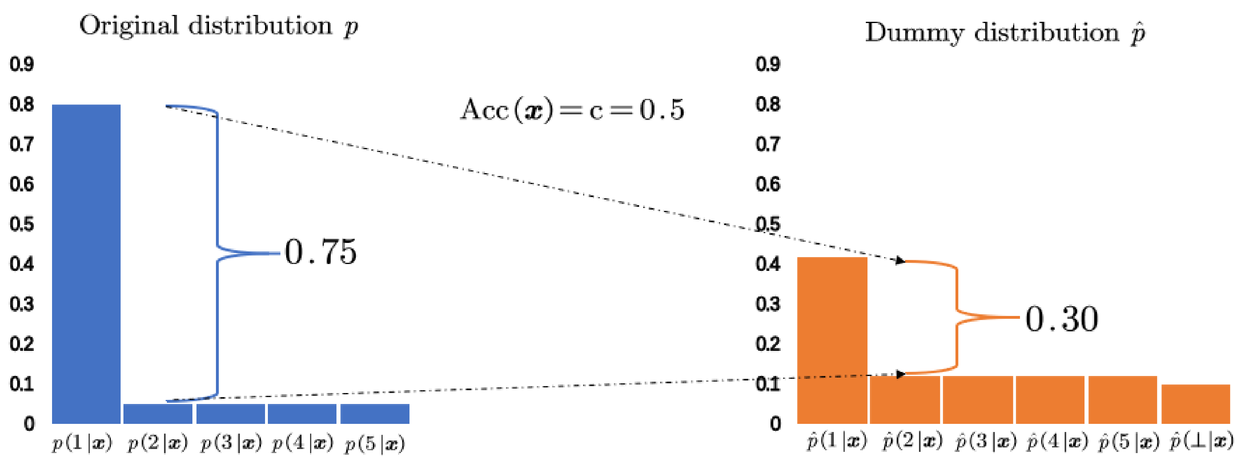
$$\{\hat{p}(y|\boldsymbol{x})\}_{y\in[K+1]}=\left\{\frac{p(y|\boldsymbol{x})+c(1-\text{Acc}(m|\boldsymbol{x}))}{\hat{\xi}(\boldsymbol{x})}\right\}_{y\in[K]}\cup\left\{\frac{(1-c)\,\text{Acc}(m|\boldsymbol{x})}{\hat{\xi}(\boldsymbol{x})}\right\}$$

Underfitting of Classifier [3]

Sample efficiency and classifier's performance $\boldsymbol{g}_{[1:K]}$ degrade rapidly with increasing extra expert cost *c*, i.e., underfitting occurs.

Cause of underfitting: Probability margin shrinks

$$p(y^*|\boldsymbol{x})-p(y|\boldsymbol{x})>\frac{p(y^*|\boldsymbol{x})-p(y|\boldsymbol{x})}{1+(1-c)\text{Acc}(\boldsymbol{x})+Kc(1-\text{Acc}(\boldsymbol{x}))}=\hat{p}(y^*|\boldsymbol{x})-\hat{p}(y|\boldsymbol{x})$$



$\hat{p}$ is more ambiguous!

- **Direct fix:** Post-hoc methods train base model with c=0, then retrain/ reformulate deferral rule to account for non-zero c [3]

Novel Surrogate Loss

Motivation: Eliminate redundancy in previous problem reduction

- Previous deferral rule: compare each label's acc to expert acc

$$\text{Acc}(m|\boldsymbol{x})\overset{?}{\geq}p(1|\boldsymbol{x})+c,\ \text{Acc}(m|\boldsymbol{x})\overset{?}{\geq}p(2|\boldsymbol{x})+c,\ \cdots,\ \text{Acc}(m|\boldsymbol{x})\overset{?}{\geq}p(K|\boldsymbol{x})+c,$$

- Ideal deferral rule: only compare the Bayes optimal one

$$\Downarrow \qquad \text{Acc}(m|\boldsymbol{x})\overset{?}{\geq}p(y^*|\boldsymbol{x})+c$$

- Induced loss formulation: (Bayes optimal label)

$$\min_{\boldsymbol{g}}\mathbb{E}_{p(\boldsymbol{x},y,m)}[\Psi(\boldsymbol{g}(\boldsymbol{x}),y)+c\mathbb{I}(m\neq y)\Psi(\boldsymbol{g}(\boldsymbol{x}),y^*)+(1-c)\mathbb{I}(m=y)\Psi(\boldsymbol{g}(\boldsymbol{x}),K+1)]$$

- Essence: Problem reduction to *K+1* class classification on distribution $\tilde{p}$:

$$\{\tilde{p}(y|\boldsymbol{x})\}=\left\{\frac{p(y|\boldsymbol{x})+c(1-\text{Acc}(m|\boldsymbol{x}))\mathbb{I}(y=y^*)}{\tilde{\xi}(\boldsymbol{x})}\right\}_{y\in[K]}\cup\left\{\frac{(1-c)\,\text{Acc}(m|\boldsymbol{x})}{\tilde{\xi}(\boldsymbol{x})}\right\}$$

Result 1: Better probability margin!

$$\tilde{p}(y^*|\boldsymbol{x})-\tilde{p}(y|\boldsymbol{x})>\hat{p}(y^*|\boldsymbol{x})-\hat{p}(y|\boldsymbol{x})$$

However, there is no access to the Bayes optimal label

Approximate Bayes optimal with intermediate learning results *y′*

Def. 1: Redundancys — free Loss Formulation $R_\Psi^\perp(\boldsymbol{g})$

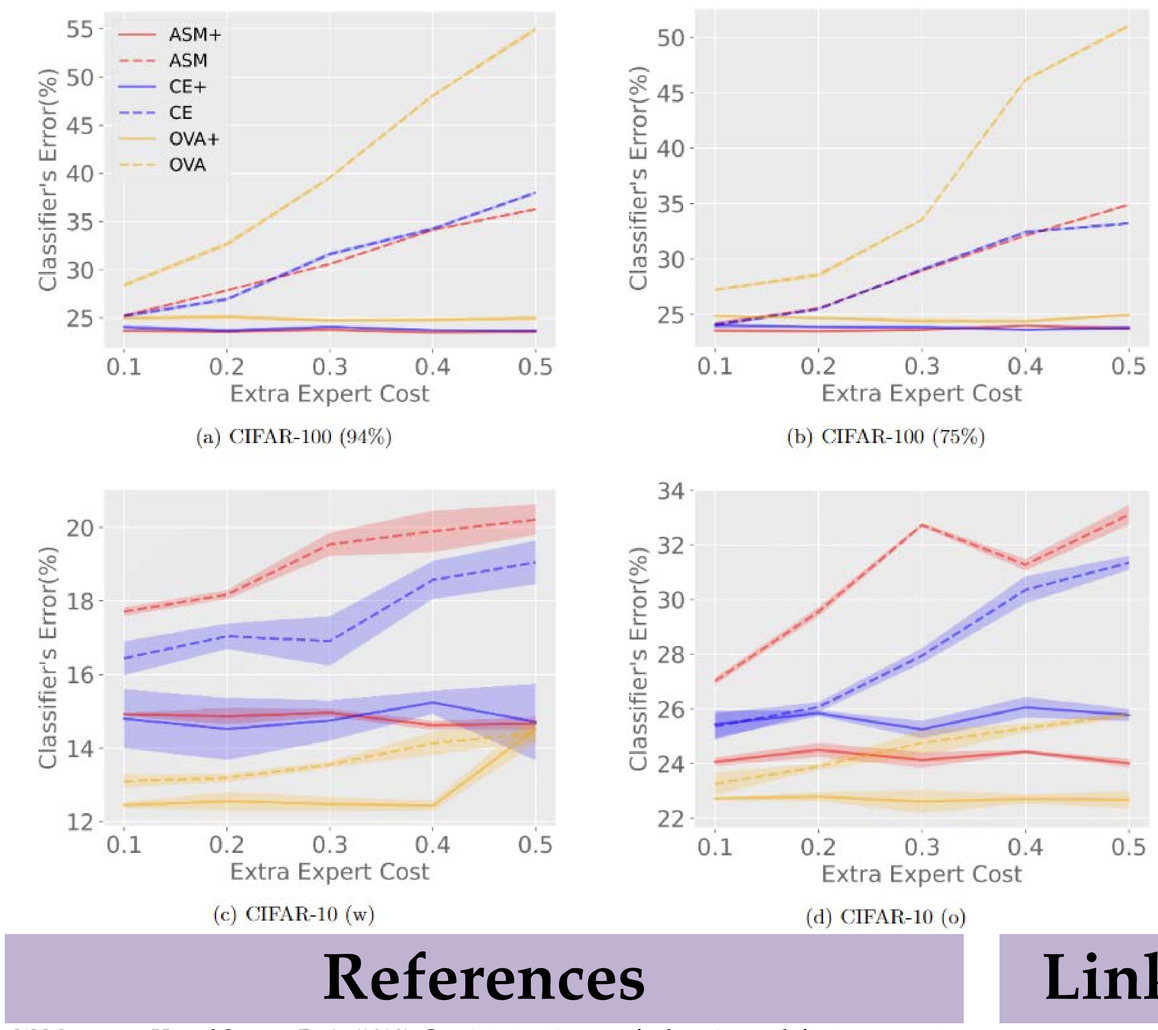
$$\mathbb{E}_{p(\boldsymbol{x},y,m)}\left[\Psi(\boldsymbol{g}(\boldsymbol{x}),y)+c\mathbb{I}(m\neq y)\min_{y'\in[K]}\ \Psi(\boldsymbol{g}(\boldsymbol{x}),y')+(1-c)\mathbb{I}(m=y)\Psi(\boldsymbol{g}(\boldsymbol{x}),K+1)\right]$$

Result 2: Consistency analysis for Ψ in [1,2,6]

$$\text{Thm. 1: } \forall\,\boldsymbol{g}^*\in\operatorname{argmin}_{\boldsymbol{g}}R_\Psi^\perp(\boldsymbol{g}),\ f_{\boldsymbol{g}^*}=f^*$$

Experiments

Proposed loss mitigates underfitting and generates more accurate predictions!



(a) CIFAR-100 (94%) (b) CIFAR-100 (75%)

(c) CIFAR-10 (w) (d) CIFAR-10 (o)

References

[1] Mozannar, H. and Sontag, D. A. (2020). Consistent estimators for learning to defer to an expert. In ICML.

[2] Verma, R. and Nalisnick, E. T. (2022). Calibrated learning to defer with one-vs-all classifiers. In ICML.

[3] Narasimhan, H., Jitkrittum, W., Menon, A. K., Rawat, A. S., and Kumar, S. (2022). Post-hoc estimators for learning to defer to an expert. In NeurIPS.

[4] Cao, Y., Cai, T., Feng, L., Gu, L., GU, J., An, B., Niu, G., and Sugiyama, M. (2022). Generalizing consistent multi-class classification with rejection to be compatible with arbitrary losses. In NeurIPS

[5] Charusaie, M., Mozannar, H., Sontag, D. A., and Samadi, S. (2022). Sample efficient learning of predictors that complement humans. In ICML.

[6] Cao, Y., Mozannar, H., Feng, L., Wei, H., and An, B. (2023). In defense of softmax parametrization for calibrated and consistent learning to defer. In NeurIPS.

Link

QR code for the paper

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia,Spain.