A Proof of Theorem 1

We first prove that if there is an L2D probability estimator $\hat{\boldsymbol{\eta}}$ for L_ϕ, then we can reconstruct a probability estimator $\tilde{\boldsymbol{\eta}}$ for multiclass classification with ϕ if $\boldsymbol{\eta}$ is in $\tilde{\Delta}^{K+1}$, which is the collection of all the elements $\boldsymbol{\beta} \in \Delta^{K+1}$ with $\beta_{K+1} \leq \frac{1}{2}$. Then we show that according to the symmetry of ϕ, we can extend this estimator to Δ^{K+1}, and then construct an unbounded L2D probability estimator for L_ϕ.

Proof. Denote by $R_{L_\phi}(\boldsymbol{u}, \boldsymbol{\eta}(\boldsymbol{x}), \Pr(M = Y|X = \boldsymbol{x})) = \sum_{y=1}^{K} \eta(\boldsymbol{x})_y \phi(\boldsymbol{u}, y) + \Pr(M = Y|X = \boldsymbol{x})\phi(\boldsymbol{u}, K+1)$ the conditional risk for L2D. Denote by $\hat{\boldsymbol{\eta}}$ a probability estimator for L_ϕ that $\hat{\boldsymbol{\eta}}(\boldsymbol{u}^*) = [\boldsymbol{\eta}(\boldsymbol{x}); \Pr(M = Y|X = \boldsymbol{x})]$ for all $\boldsymbol{u}^* \in \operatorname{argmin}_{\boldsymbol{u}} R_{L_\phi}(\boldsymbol{u}, \boldsymbol{\eta}(\boldsymbol{x}), \Pr(M = Y|X = \boldsymbol{x}))$, we can learn that $\Delta^K \times [0, 1]$ is in the range of $\hat{\boldsymbol{\eta}}$. Then it is easy to verify that $\tilde{\boldsymbol{\eta}}(\boldsymbol{u}) = \frac{\hat{\boldsymbol{\eta}}(\boldsymbol{u})}{1+\hat{\eta}_{K+1}(\boldsymbol{u})} \in \tilde{\Delta}^{K+1}$ is a valid probability estimator for $K+1$-class multiclass classification with ϕ if the posterior probabilities $[p(y|\boldsymbol{x})]_{y=1}^{K+1}$ is in $\tilde{\Delta}^{K+1}$.

Then we construct a valid estimator for Δ^{K+1} based on $\tilde{\boldsymbol{\eta}}$. Denote by P a permutation matrix that exchanges the value of a vector's first and last dimension, then we have the following estimator $\tilde{\boldsymbol{\eta}}'$:

$$\tilde{\boldsymbol{\eta}}'(\boldsymbol{u}) = \begin{cases} \tilde{\boldsymbol{\eta}}(\boldsymbol{u}), \ u_{K+1} \neq \max_y u_y, \\ P\tilde{\boldsymbol{\eta}}(P\boldsymbol{u}), \ \text{else}. \end{cases}$$

We then prove that it is a valid estimator. Denote by Δ_+^{K+1} the set of all the $\boldsymbol{\beta} \in \Delta^{K+1}$ that $\beta_{K+1} \neq \max_y \beta_y$, we can learn that $\Delta_+^{K+1} \in \tilde{\Delta}^{K+1}$. Then for any $\boldsymbol{\beta} \in \Delta_+^{K+1}$, denote by $\boldsymbol{u}^*$ the minimizer of conditional risk *w.r.t.* $\boldsymbol{\beta}$ and ϕ, we can learn that $u^*_{K+1} \neq \max_y u^*_y$ due to the **consistency** of ϕ and thus $\tilde{\boldsymbol{\eta}}'(\boldsymbol{u}^*) = \tilde{\boldsymbol{\eta}}(\boldsymbol{u}^*) = \boldsymbol{\beta}$.

For any $\boldsymbol{\beta} \notin \Delta_+^{K+1}$, we can learn that $P\boldsymbol{\beta} \in \Delta_+^{K+1}$. Denote by $\boldsymbol{u}^*$ the minimizer of conditional risk *w.r.t.* $\boldsymbol{\beta}$ and ϕ, then we can learn that $P\boldsymbol{u}^*$ must be the minimizer of $P\boldsymbol{\beta}$ according to the **symmetry** of ϕ. Then we have that $\tilde{\boldsymbol{\eta}}(P\boldsymbol{u}^*) = P\boldsymbol{\beta}$. Furthermore, notice that $PP\boldsymbol{\beta} = \boldsymbol{\beta}$, then we have $\tilde{\boldsymbol{\eta}}'(\boldsymbol{u}^*) = P\tilde{\boldsymbol{\eta}}(P\boldsymbol{u}^*) = PP\boldsymbol{\beta} = \boldsymbol{\beta}$.

Combining the two paragraphs above, we can learn that $\tilde{\boldsymbol{\eta}}'$ is a valid multiclass probability estimator for ϕ. Then we can construct an unbounded L2D estimator as in (3), which indicates the existence of an unbounded probability estimator. Furthermore, since the loss function is unchanged, the collection of all the minimizers of $R_{L_\phi}(\boldsymbol{u}, \boldsymbol{\eta}(\boldsymbol{x}), \Pr(M = Y|X = \boldsymbol{x}))$ are also unchanged, and thus the all the values should have the same value on $\mathcal{U}$. □

B Proof of Proposition 1

Proof. The boundedness of the first K dimensions is straightforward. The $K+1$th dimension's boundedness can also be directly proved by reformulating it as

$$\frac{\exp(u_{K+1})}{\exp(u_{K+1}) + \sum_{y'=1}^{K} \exp(u_{y'}) - \max_{y' \in \{1,\cdots,K\}} \exp(u_{y'})}.$$

Then we begin to prove its maxima-preserving. Denote by $y_1 = \operatorname{argmax}_{y \in \{1,\cdots,K\}} u_y$. It is easy to verify that $\tilde{\psi}_{y_1}(\boldsymbol{u}) > \tilde{\psi}_y(\boldsymbol{u})$ for any $y \in \{1, \cdots, K\}/\{y_1\}$ due to the property of softmax function. Then we focus on the relation between $\tilde{\psi}_{y_1}(\boldsymbol{u})$ and $\tilde{\psi}_{K+1}(\boldsymbol{u})$. We can learn that:

$$\frac{\exp(u_{K+1})}{\exp(u_{K+1}) + \sum_{y'=1}^{K} \exp(u_{y'}) - \exp(u_{y_1})} = \frac{1}{1 + \frac{\sum_{y'=1}^{K} \exp(u_{y'}) - \exp(u_{y_1})}{\exp(u_{K+1})}},$$

$$\frac{\exp(u_{y_1})}{\sum_{y'=1}^{K} \exp(u_{y'})} = \frac{1}{1 + \frac{\sum_{y'=1}^{K} \exp(u_{y'}) - \exp(u_{y_1})}{\exp(u_{y_1})}}.$$

When $u_{K+1} > u_{y_1}$, we can learn that $\frac{\sum_{y'=1}^{K} \exp(u_{y'}) - \exp(u_{y_1})}{\exp(u_{K+1})} < \frac{\sum_{y'=1}^{K} \exp(u_{y'}) - \exp(u_{y_1})}{\exp(u_{y_1})}$, then $\tilde{\psi}_{K+1}(\boldsymbol{u}) > \tilde{\psi}_{y_1}(\boldsymbol{u})$. Based on the formulations above, it is also easy to verify the cases when $u_{y_1} \geq u_{K+1}$. Combining the discussions above and we can conclude the proof. □

C Proof of Theorem 2

Proof. The consistency can be directly obtained by recovering the class probability and expert accuracy according to the maxima-preserving, then we first focus on how our proposed estimator recovers the posterior probabilities.

According to the property of log loss, it is easy to learn that $\tilde{\psi}_y(\boldsymbol{g}^*(\boldsymbol{x})) = p(y|\boldsymbol{x})$ for $y \in \{1, \cdots, K\}$. When $y = K+1$, we can learn from the range of our estimator and the property of binary log loss that $\tilde{\psi}_y(\boldsymbol{g}^*(\boldsymbol{x})) = \Pr(M = Y|X = \boldsymbol{x})$. Then we can conclude that $\tilde{\psi}(\boldsymbol{g}^*(\boldsymbol{x})) = \boldsymbol{\eta}(\boldsymbol{x}) \times \Pr(M = Y|X = \boldsymbol{x})$.

Then we begin to prove that there is no unbounded probability estimator by contradiction. Suppose there exists an unbounded estimator ψ. For any $\boldsymbol{g}$, it must be the solution of a distribution and expert whose posterior probability is $\tilde{\psi}(\boldsymbol{g}(\boldsymbol{x}))$ for each point $\boldsymbol{x}$. However, for a $\boldsymbol{g}$ that there exists $\boldsymbol{x}$ that $\psi(\boldsymbol{g}(\boldsymbol{x})) \notin \Delta^K \times [0,1]$, we can learn that it cannot be the solution of any distribution and expert according to the definition of probability estimator. We can learn from this contradiction that ψ is not a probability estimator as long as its range is not $\Delta^K \times [0,1]$. □

D Proof of Corollary 1

Proof. We can set the multi-class loss to $\phi_{\tilde{\psi}}$ to get our proposed loss:

$$\phi_{\tilde{\psi}}(\boldsymbol{u}, y) = \begin{cases} -\log\left(\frac{\exp(u_y)}{\sum_{y'=1}^{K'-1}\exp(u_{y'})}\right), \ y \neq K', \\ -\log\left(\frac{\exp(u_{K'})}{\sum_{y'=1}^{K'}\exp(u_{y'}) - \max_{y' \in \{1,\cdots,K'-1\}}\exp_{u_{y'}}}\right), \ \text{else}. \end{cases}$$

A similar result can be deduced for the OvA-based surrogate by considering the following consistent multi-class loss with a strictly proper binary composite loss ξ:

$$\phi_{\mathrm{OvA}}(\boldsymbol{u}, y) = \begin{cases} \xi(u_y) + \sum_{y' \neq y, K'} \xi(-u_{y'}), \ y \neq K', \\ \xi(u_{K'}) - \xi(-u_{k'}), \ \text{else}. \end{cases}$$

We then begin to prove their consistency. It is easy to verify that $\phi_{\tilde{\psi}}$ is minimized when $\tilde{\psi}(\boldsymbol{g}(\boldsymbol{x})) = \frac{p(y|\boldsymbol{x})}{1-p(K'|\boldsymbol{x})}$, and we can conclude its consistency using the maxima-preserving property of $\tilde{\psi}$.

For ϕ_{OvA}, denote by ψ_{OvA} the inverse link of ξ. We can learn that $\boldsymbol{g}^*(\boldsymbol{x}) = \psi_{\mathrm{OvA}}(\frac{p(y|\boldsymbol{x})}{1-p(K'|\boldsymbol{x})})$, and we can learn the consistency since ξ is strictly proper and thus ψ_{OvA} is increasing. □

E Proof of Theorem 3

Proof. We first apply Pinsker's inequality. We can learn that:

$$R_{L_{\tilde{\psi}}}(\boldsymbol{g}) - R^*_{L_{\tilde{\psi}}} \geq \mathbb{E}_{p(\boldsymbol{x})}\left[\frac{1}{2}\|\tilde{\psi}_{1:K}(\boldsymbol{g}(\boldsymbol{x})) - \boldsymbol{\eta}(x)\|_1^2 + 2(\tilde{\psi}_{K+1}(\boldsymbol{g}(\boldsymbol{x}) - \Pr(M = Y|X = \boldsymbol{x}))^2\right].$$

We can learn that $R_{L_{\tilde{\psi}}}(\boldsymbol{g}) - R^*_{L_{\tilde{\psi}}} \geq \mathbb{E}_{p(\boldsymbol{x})}\left[\frac{1}{2}\|\tilde{\psi}_{1:K}(\boldsymbol{g}(\boldsymbol{x})) - \boldsymbol{\eta}(x)\|_1^2\right]$ immediately and learn that the bound for R_{01} following the analysis of cross-entropy loss in ordinary classification. Then we move to analyze the 0-1-deferral risk. We can further learn that :

$$\begin{aligned} R_{L_{\tilde{\psi}}}(\boldsymbol{g}) - R^*_{L_{\tilde{\psi}}} &\geq \mathbb{E}_{p(\boldsymbol{x})}\left[\frac{1}{2}\|\tilde{\psi}_{1:K}(\boldsymbol{g}(\boldsymbol{x})) - \boldsymbol{\eta}(x)\|_1^2 + 2(\tilde{\psi}_{K+1}(\boldsymbol{g}(\boldsymbol{x})) - \Pr(M = Y|X = \boldsymbol{x}))^2\right] \\ &\geq \frac{1}{2}\mathbb{E}_{p(\boldsymbol{x})}\left[\|\tilde{\psi}_{1:K}(\boldsymbol{g}(\boldsymbol{x})) - \boldsymbol{\eta}(x)\|_1^2 + (\tilde{\psi}_{K+1}(\boldsymbol{g}(\boldsymbol{x})) - \Pr(M = Y|X = \boldsymbol{x}))^2\right] \end{aligned}$$

For any $\boldsymbol{x}$, when $\boldsymbol{g}(\boldsymbol{x})$ can induce the Bayes optimal solution for it, the excess risk is zero and the bound holds naturally. When it is not optimal, denote by $\eta_{K+1}(\boldsymbol{x}) = \Pr(M = Y|X = \boldsymbol{x})$. y' is the

dimension with the largest value of $\boldsymbol{g}(\boldsymbol{x})$ and that of $\boldsymbol{\eta}(\boldsymbol{x})$ is y''. Then we can learn that:

$$\begin{aligned}
&\frac{1}{2}\left(\|\tilde{\psi}_{1:K}(\boldsymbol{g}(\boldsymbol{x})) - \boldsymbol{\eta}(x)\|_1^2 + (\tilde{\psi}_{K+1}(\boldsymbol{g}(\boldsymbol{x})) - \Pr(M = Y|X = \boldsymbol{x}))^2\right) \\
&\geq \frac{1}{2}\left(\tilde{\psi}_{y'}(\boldsymbol{g}(\boldsymbol{x})) - \eta_{y'}(\boldsymbol{x}) - \tilde{\psi}_{y''}(\boldsymbol{g}(\boldsymbol{x})) + \eta_{y''}(\boldsymbol{x})\right)^2 \\
&\geq \frac{1}{2}\left(\eta_{y'}(\boldsymbol{x}) - \eta_{y''}(\boldsymbol{x})\right)^2
\end{aligned}$$

The last step is obtained according to the maxima-preserving property. Further generalizing y' and y'' to be instance-dependent ($y'(\boldsymbol{x})$ and $y''(\boldsymbol{x})$), we can learn the following inequality using Jensen's inequality:

$$R_{L_{\tilde{\psi}}}(\boldsymbol{g}) - R^*_{L_{\tilde{\psi}}} \geq \frac{1}{2}(\mathbb{E}_{p(\boldsymbol{x})}[|\eta_{y'}(\boldsymbol{x}) - \eta_{y''}(\boldsymbol{x})|])^2,$$

which concludes the proof since the second expectation term is $R^{\perp}_{01}(\boldsymbol{g}) - R^{\perp *}_{01}$ □

F Details of Experiments

Details of Model and Optimizer: For all the methods on different datasets, we use the 28-layer WideResNet that is the same as those used in Mozannar and Sontag [28], Charusaie et al. [9]. The optimizer is SGD with cosine annealing, where the learning rate is 1e-1 and weight decay is 5e-4. We conduct the experiments on 8 NVIDIA GeForce 3090 GPUs and the batch size is 1024 (128 on each GPU). The training epoch on CIFAR100 is set to 200 and 400 on CIFAR10H, respectively.

Details of Evaluation Metrics: The reported Error is the sample mean of $\ell^{\perp}_{01}$, and Coverage is the ratio of undeferred samples. The ECE of expert accuracy is defined below:

$$\mathrm{ECE} = \sum_{i=1}^{N} b_i |p_i - c_i|,$$

where b_i is the ratio of predictions whose confidences fall into the ith bin. p_i is the average confidence and c_i is the average accuracy in this bin. We set the bin number to 15. The budgeted error is calculated as below: if the coverage is lower than $1 - x\%$, we will use the classifier's prediction instead of the expert's for those samples whose estimated expert accuracy is lower to make the coverage equal to $1 - x\%$.

G Limitations and Broader Impact

Limitations: This work is designed for L2D without extra constraints on the number of expert queries. We believe that combining it with selective learning, i.e., adding explicit constraints on the ratio of deferred samples, can be a promising future direction.

Broader Impact: When applied in real-world applications, the frequency of expert queries may be imbalanced due to the performance differences of the expert among samples. This is a common impact shared by all the L2D methods. We believe that introducing fairness targets into L2D can be another promising direction.