Multi-Class Classification with Abstention Based on Crammer–Singer Surrogate with Linear Growth Rate

Hongyu Zhang
Kyoto University

Han Bao
The Institute of Statistical Mathematics
Tohoku University, RIKEN AIP

Junya Honda
Kyoto University
RIKEN AIP

Abstract

We study the problem of multi-class classification with abstention, where a learner can choose to abstain from making a prediction to avoid excessively uncertain predictions. In this problem, the predictor-rejector framework is known as one promising approach, in which a predictor and rejector are learned separately, and the abstention cost is explicitly taken into account. However, only nonconvex surrogate losses have been known previously due to their inherent difficulty in the multi-class setting. To tackle this difficulty, we propose a novel family of surrogate losses for multi-class classification with abstention based on the Crammer–Singer (CS) surrogate, which can constitute a convex loss that is easy to optimize. We show that the proposed surrogate losses lead to the optimal predictor and rejector, and prove excess error bounds for our surrogate losses, demonstrating a linear growth rate for a certain choice of losses.

1 Introduction

Learning with abstention (Chow, 1970; Herbei and Wegkamp, 2006; Cortes et al., 2016b; Mao et al., 2024a) has drawn lots of attention in recent years. It refers to a variant of a learning problem where a learner can abstain from making a prediction, which is illustrated in Figure 1. Learning with abstention can be used to control the uncertainty of prediction when abstention is better than random guessing (Silva Filho et al., 2023). It can be applied to many real-world applications that need reliable decision-making (Mozannar et al., 2023; Filippova, 2020). For example, in online question answering system, it is sometimes important to abstain from making an uncertain answer to avoid generating misleading or fake information. In multi-class classification with abstention, a learner can abstain from making a multi-class prediction to avoid serious misclassification. The abstention incurs a pre-defined cost $c \in (0, 1)$. Otherwise, the standard 0-1 loss is incurred.

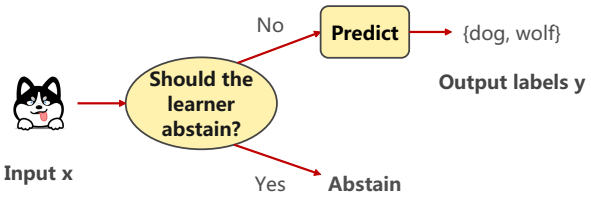


Figure 1: Process of learning with abstention.

Currently, many methods of multi-class classification with abstention have been proposed, and they can be divided into several categories. Confidence-based methods (Ramaswamy et al., 2018) are based on the idea of abstaining if the predictor's output score is lower than some threshold. For selective classification (Wiener and El-Yaniv, 2015), a predictor can choose to predict or reject according to the output probability of a selector. For the score-based formulation (Mozannar and Sontag, 2020), a learner will abstain if the score of an extra reject option is the highest. For the predictor-rejector formulation, a predictor and a rejector (Mao et al., 2024a) from different function families are learned.

In this paper, we focus on the predictor-rejector formulation. In this formulation, the surrogate loss minimization should lead to the optimal predictor and rejector, i.e., should be calibrated (Bartlett et al., 2006). Compared to existing methods, it can both model the abstention cost explicitly and avoid the possible failure from an uncalibrated predictor (Mao et al., 2024a). Nevertheless, Ni et al. (2019) show that for the predictor-rejector formulation, surrogate losses require an unnatural constraint to be calibrated in the multi-class case. To this end, Mao et al. (2024a) derive calibrated but nonconvex surrogate losses, and the

Workshop "Towards Trustworthy Predictions: Theory and Applications of Calibration for Modern AI" at AISTATS 2026, Tangier, Morocco.

analysis is specific to the form of their chosen surrogate.

To tackle this limitation, we propose a novel family of convex and calibrated surrogate losses based on the Crammer–Singer (CS) loss (Crammer and Singer, 2001). While the CS loss itself is well known in the work on multi-class classification, we reveal that the CS loss specially has a good property to satisfy the required constraint derived in Ni et al. (2019). With this observation, we make the following contributions:

- We propose a novel family of convex surrogate losses based on the CS surrogate for multi-class classification with abstention in the predictor-rejector framework, which enjoys better optimization properties than those in previous work.
- We show the proposed surrogate loss is calibrated and prove excess error bounds for our surrogate losses, demonstrating a linear growth rate for a certain choice of losses, which is the first result under the predictor-rejector framework to the best of our knowledge.
- Experiments empirically show the usefulness of our proposed surrogate losses.

Here, the notation of calibration is different from probability calibration, which is a way to assess and improve predicted probabilities (Silva Filho et al., 2023). Still, these notions are closely connected since a decent predictor from the viewpoint of probability calibration is also useful for appropriate abstention. Indeed, learning with abstention is an effective way to control the uncertainty and avoid overconfident prediction (Zhang et al., 2023), which is also a target of probability calibration (Silva Filho et al., 2023). Thus, they are both useful tools for handling the uncertainty and overconfident prediction in machine learning.

2 Problem Setup

In this section, we formulate the problem and introduce the predictor-rejector learning with abstention framework. Consider a multi-class classification scenario where $\mathcal{X}$ is the input space and $\mathcal{Y} = \{1, \cdots, K\}$ is the label space of K classes. Assume that samples $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_n, y_n)$ are drawn i.i.d. from a fixed and unknown distribution $p(\boldsymbol{x}, y)$ on $\mathcal{X} \times \mathcal{Y}$.

In the *predictor-rejector learning with abstention* framework, a predictor $\boldsymbol{g} : \mathcal{X} \to \mathbb{R}^K$ and rejector $r : \mathcal{X} \to \mathbb{R}$ are constructed and the abstention cost $c \in [0, 1]$ is explicitly taken into account. If $r(\boldsymbol{x}) \leqslant 0$, the predictor will choose to abstain. Otherwise predict label $f(\boldsymbol{x}) = \arg\max_{y \in \mathcal{Y}} g_y(\boldsymbol{x})$, where $g_y : \mathcal{X} \to \mathbb{R}$ is a score function for multiclass classification, and $\boldsymbol{g}(\boldsymbol{x}) = (g_1(\boldsymbol{x}), \cdots, g_K(\boldsymbol{x}))^\top$. Then the target predictor-rejector abstention loss is expressed as:

$$L_{0\text{-}1\text{-}c}(r, \boldsymbol{g}; \boldsymbol{x}, y) = \underbrace{\mathbb{I}_{f(\boldsymbol{x}) \neq y} \mathbb{I}_{r(\boldsymbol{x}) > 0}}_{\text{not to abstain}} + c \underbrace{\mathbb{I}_{r(\boldsymbol{x}) \leqslant 0}}_{\text{abstain}},$$

where $\mathbb{I}_{[\cdot]}$ denotes the indicator function. In the following part we assume abstention cost c is known to predictor $\boldsymbol{g}$ and $0 < c < \frac{1}{2}$, following Ni et al. (2019) and Cortes et al. (2016a). Given the target abstention loss function $L_{0\text{-}1\text{-}c}$, we can define its risk $R_{0\text{-}1\text{-}c}$ by

$$R_{0\text{-}1\text{-}c}(r, \boldsymbol{g}) = \mathbb{E}_{p(\boldsymbol{x}, y)} \left[L_{0\text{-}1\text{-}c}(r, \boldsymbol{g}; \boldsymbol{x}, y) \right], \tag{1}$$

and let $R^*_{0\text{-}1\text{-}c} = \inf_{\boldsymbol{g}, r} R_{0\text{-}1\text{-}c}(r, \boldsymbol{g})$ be its infimum taken over all measurable functions. However, it's intractable to optimize target risk $R_{0\text{-}1\text{-}c}$. Instead, we introduce a surrogate loss, which is easier to optimize. Denote the surrogate loss function for $L_{0\text{-}1\text{-}c}$ as L. Similarly to the case of $R_{0\text{-}1\text{-}c}$ in Eq. (1), we also define its risk R by

$$R(r, \boldsymbol{g}) = \mathbb{E}_{p(\boldsymbol{x}, y)} \left[L(r, \boldsymbol{g}; \boldsymbol{x}, y) \right],$$

and let $R^* = \inf_{\boldsymbol{g}, r} R(r, \boldsymbol{g})$ be the infimum of R taken over all measurable functions. Besides, we denote the empirical surrogate risk as

$$\widehat{R}(r, \boldsymbol{g}) = \frac{1}{n} \sum_{i=1}^{n} L(r, \boldsymbol{g}; \boldsymbol{x}_i, y_i).$$

The notions of calibration and an excess risk bound known in classification problems are also extended to the classification with abstention. Calibration is a way to justify the effectiveness of surrogate losses (Bartlett et al., 2006). A surrogate loss is called *calibrated* if any predictor sequence $(\boldsymbol{g}_m)_{m \geq 1}$ and any rejector sequence $(r_m)_{m \geq 1}$ minimizing the surrogate risk R also minimize the target risk $R_{0\text{-}1\text{-}c}$, that is, formally,

$$\begin{aligned} & R(r_m, \boldsymbol{g}_m) \xrightarrow{m \to \infty} R^* \\ \implies \quad & R_{0\text{-}1\text{-}c}(r_m, \boldsymbol{g}_m) \xrightarrow{m \to \infty} R^*_{0\text{-}1\text{-}c}. \end{aligned}$$

An *excess risk bound* (Zhang, 2004; Bartlett et al., 2006) is a relation to link the surrogate risk R and the target risk $R_{0\text{-}1\text{-}c}$, which has the form of

$$R_{0\text{-}1\text{-}c}(r, \boldsymbol{g}) - R^*_{0\text{-}1\text{-}c} \leqslant \Gamma(R(r, \boldsymbol{g}) - R^*),$$

for any $(r, \boldsymbol{g})$, where $\Gamma(\cdot)$ is a non-decreasing function called the *excess risk rate*. Here, $R_{0\text{-}1\text{-}c}(r, \boldsymbol{g}) - R^*_{0\text{-}1\text{-}c}$ is the suboptimality of the target loss, and $R(r, \boldsymbol{g}) - R^*$ is the suboptimality of the surrogate loss.

3 Surrogate Losses Based on the Crammer–Singer Loss

In this section, we introduce a family of surrogate losses based on the Crammer–Singer (CS) loss (Crammer and Singer, 2001). Informally, the CS loss is a multiclass loss, yet its form is close to binary classification losses, which makes the calibration analysis easier. Our losses can be divided into the multiplicative CS (MCS) and additive CS (ACS) losses. The MCS loss can be calibrated without dependence on the component of the surrogate function corresponding to the predictor, though the loss is not always convex with respect to $(r, \boldsymbol{g})$. The ACS loss can be calibrated and convex, while a case-by-case analysis is needed.

The MCS loss is of the following form:

$$L_{\mathrm{MCS}}(r, \boldsymbol{g}; \boldsymbol{x}, y) = \phi\left(g_y(\boldsymbol{x}) - \max_{y' \neq y} g_{y'}(\boldsymbol{x})\right) \psi_1\left(r(\boldsymbol{x})\right) + \psi_2\left(r(\boldsymbol{x})\right).$$

Here, $\phi \geqslant 0$ is a non-increasing convex function bounding $t \mapsto \mathbb{I}_{t \leqslant 0}$ from above. ψ_1 and ψ_2 are convex functions bounding $t \mapsto \mathbb{I}_{t>0}$ and $t \mapsto \mathbb{I}_{t \leqslant 0}$ from above, respectively.

The ACS loss is of the following form:

$$L_{\mathrm{ACS}}(r, \boldsymbol{g}; \boldsymbol{x}, y) = \phi\left(g_y(\boldsymbol{x}) - \max_{y' \neq y} g_{y'}(\boldsymbol{x}) - r(\boldsymbol{x})\right) + \psi(r(\boldsymbol{x})).$$

Here, $\phi \geqslant 0$ is a non-increasing convex function bounding $t \mapsto \mathbb{I}_{t \leqslant 0}$ from above. ψ is a convex function bounding $t \mapsto \mathbb{I}_{t \leqslant 0}$ from above.

Note that L_{MCS} is convex with respect to r and $\boldsymbol{g}$ separately, but not always convex jointly in $(r, \boldsymbol{g})$. On the other hand, L_{ACS} is always convex jointly in $(r, \boldsymbol{g})$.

4 Calibration Guarantee

In this section, we show the conditions for the MCS and ACS losses to be calibrated.

Theorem 4.1. *L_{MCS} is calibrated if and only if*

$$H_{1-c}\psi_1'(0) + \psi_2'(0) = 0,$$

where $H_\eta = \min_{\Delta \geqslant 0} \{\eta\phi(\Delta) + (1-\eta)\phi(-\Delta)\}$ for $0 \leqslant \eta \leqslant 1$.

Next, we consider the calibration condition for the ACS loss. If we consider the exponential loss for ϕ, then the ACS loss reduces to the MCS loss with $\psi_1(r) = \exp(r)$, and is immediately guaranteed to be calibrated. For other choices of ϕ, unlike the MCS loss, we do not have a general calibration result for the ACS loss, but we can still show that the ACS loss with hinge/logistic ϕ and generic ψ can be calibrated, as follows:

Theorem 4.2. *L_{ACS} with $\phi(\Delta) = \max\{1-\Delta, 0\}$ is calibrated if and only if*

$$\psi'(0) = -2c.$$

The analysis for the logistic loss is much more complicated than that for the hinge loss. Still, under the assumption of the differentiability of ψ, we have the following result.

Theorem 4.3. *Assume that ψ is a twice differentiable decreasing function. Then L_{ACS} with $\phi(\Delta) = \log\left(1 + \exp(-\Delta)\right)$ is calibrated if and only if*

$$\psi'(0) = -2c(1-c).$$

Compared with existing calibration results, we can take various ϕ and ψ while the existing ones Ni et al. (2019); Mao et al. (2024a) require specific forms of the losses.

5 Excess Risk Bounds

In this section, we derive excess risk bounds for the MCS and ACS losses. In the rest of this paper, we always assume that surrogate losses are taken so that the calibration condition in the last section is satisfied.

Denote the risk for the MCS loss and its infimum as R_{MCS} and R^*_{MCS}, respectively:

$$R_{\mathrm{MCS}}(r, \boldsymbol{g}) = \mathbb{E}_{p(\boldsymbol{x}, y)}\left[L_{\mathrm{MCS}}(r, \boldsymbol{g}; \boldsymbol{x}, y)\right], \quad R^*_{\mathrm{MCS}} = \inf_{\boldsymbol{g}, r} R_{\mathrm{MCS}}(r, \boldsymbol{g}).$$

Similarly, denote the risk for the ACS loss and its infimum as R_{ACS} and R^*_{ACS}, respectively. Then we have the following excess risk bounds, which are the main contribution of our paper.

Theorem 5.1. *The following excess risk bound holds for all $r, \boldsymbol{g}$ and any distribution $p(\boldsymbol{x}, y)$:*

$$R_{0\text{-}1\text{-}c}(r, \boldsymbol{g}) - R^*_{0\text{-}1\text{-}c} \leqslant \Gamma\left(R(r, \boldsymbol{g}) - R^*\right),$$

where

$$\Gamma(t) = O\left(\max\left\{t, \sqrt{t}\right\}\right)$$

for $R = R_{\mathrm{MCS}}$ with $\phi(\Delta) = \exp(-\Delta)$, $\psi_1(r) = \exp(r)$, and $\psi_2(r) = c\exp(-\alpha r)$, and

$$\Gamma(t) = O\left(t\right)$$

Table 1: Performance of our method and compared methods on the CIFAR-10 dataset.

Method	Misclassification error	Rejection ratio	Abstention loss
CE (no rej.)	10.72%	0.00%	10.72%
CS (no rej.)	14.82%	0.00%	14.82%
Ramaswamy et al. (2018) (CS)	15.09%	**17.66**%	14.19%
Mao et al. (2024a) (two-stage)	7.81%	75.13%	9.46%
Ours (MCS)	7.57%	70.05%	9.27%
Ours (ACS)	**6.40%**	71.62%	**8.98%**

for $R = R_{\mathrm{ACS}}$ *with* $\phi(\Delta) = \max\{1-\Delta, 0\}$, $\psi(r) = c\max\{1-\alpha r, 0\}$.

The excess risk rate function $\Gamma(t) = O(t)$ is linear and better than that in Mao et al. (2024a), in which they derive excess risk bound with excess risk rate function $O\left(\max\left\{t, \sqrt{t}\right\}\right)$. It is believed to be optimal as long as Γ is data-dependent (Cao et al., 2025).

6 Generalization Error Bounds

Next, we give generalization error bounds for the MCS loss and the ACS loss, where we use the concept of the Rademacher complexity (Bartlett and Mendelson, 2002).

Let $\mathfrak{G}$ be a family of functions $\boldsymbol{g} : \mathcal{X} \to \mathbb{R}^K$, $\mathcal{G}$ be a family of functions $g_y : \mathcal{X} \to \mathbb{R}$, and $\mathcal{R}$ be a family of functions $r : \mathcal{X} \to \mathbb{R}$. Assume ϕ, ψ_1, ψ_2, and ψ are all Lipschitz-continuous functions (with constants L_ϕ, L_{ψ_1}, L_{ψ_2}, and L_ψ, respectively), and all functions in the model class $\mathcal{G}$ and $\mathcal{R}$ are bounded. Note that the choices of exponential, hinge, and logistic functions over these $\mathcal{G}$ and $\mathcal{R}$ all satisfy this condition.

Theorem 6.1. *Define* $M_\phi = \sup_{x\in\mathcal{X}, g_y\in\mathcal{G}} \phi(g_y(\boldsymbol{x}))$, $M_{\psi_1}(r) = \sup_{x\in\mathcal{X}, r\in\mathcal{R}} \psi_1(r(\boldsymbol{x}))$, $M_{\psi_2}(r) = \sup_{x\in\mathcal{X}, r\in\mathcal{R}} \psi_2(r(\boldsymbol{x}))$, *and* $M_\psi(r) = \sup_{x\in\mathcal{X}, r\in\mathcal{R}} \psi(r(\boldsymbol{x}))$. *Let* $\mathfrak{R}_n(\mathcal{G}), \mathfrak{R}_n(\mathcal{R})$ *be the Rademacher complexity (Bartlett and Mendelson, 2002) of* $\mathcal{G}, \mathcal{R}$ *for data of size* n *drawn from* $p(\boldsymbol{x})$, *respectively. Then for any* $\delta \in (0,1)$, *with probability at least* $1-\delta$, *the following multi-class classification generalization bounds hold for all* $\boldsymbol{g} \in \mathfrak{G}$*:*

$$\begin{aligned} R_{\mathrm{MCS}}(\boldsymbol{g}) \leqslant & \widehat{R}_{\mathrm{MCS}}(\boldsymbol{g}) + \left(2M_\phi + M_{\psi_1}\right) L_\phi K^2 \mathfrak{R}_n(\mathcal{G}) \\ & + \left(\left(2M_\phi + M_{\psi_1}\right) L_{\psi_1} + 2L_{\psi_2}\right) \mathfrak{R}_n(\mathcal{R}) \\ & + \left(M_\phi M_{\psi_1} + M_{\psi_2}\right) \sqrt{\frac{1}{2n} \log\left(\frac{1}{\delta}\right)}, \\ R_{\mathrm{ACS}}(\boldsymbol{g}) \leqslant & \widehat{R}_{\mathrm{ACS}}(\boldsymbol{g}) + 2L_\phi K^2 \mathfrak{R}_n(\mathcal{G}) \\ & + 2\left(L_\phi + L_\psi\right) \mathfrak{R}_n(\mathcal{R}) \\ & + \left(M_{\psi_1} + M_{\psi_2}\right) \sqrt{\frac{1}{2n} \log\left(\frac{1}{\delta}\right)}. \end{aligned}$$

Theorem 6.1 ensures the generalization error bounds of our proposed MCS loss and ACS loss, which shows that when the empirical surrogate risk is minimized, the expected surrogate risk is also minimized under infinitely large data and $(\mathcal{G}, \mathcal{R})$ with asymptotically vanishing Rademacher complexities. This result can be proved by a standard technique for generalization error bounds, and it supports the validity of the proposed losses.

7 Experiments

In this section, we present experimental results for our losses in comparison with existing methods on CIFAR-10 (Krizhevsky and Hinton, 2009). We also give results for vehicle, satimage, and yeast datasets from UCI Machine Learning Repository (Bache and Lichman, 2013) in Appendix E. Note that the latter three datasets are also used in Ramaswamy et al. (2018). We compare the proposed losses with the vanilla cross-entropy loss (CE) and CS loss, the confidence-based method that also uses the CS loss proposed by Ramaswamy et al. (2018), and the two-stage predictor-rejector surrogate loss proposed by Mao et al. (2024a). In all of the experiments, for MCS loss we use hinge ψ_1 and ψ_2 of the forms $\psi_1(r) = \max\{0, 1-r\}$ and $\psi_2(r) = c\max\{0, 1-\alpha r\}$, where α is a hyperparameter and we can obtain that $\alpha = 2$ from Corollary D.1 for this setting of ψ_1 ad ψ_2. On the other hand, for the ACS loss we use hinge ψ of the form $\psi(r) = c\max\{0, 1-\alpha r\}$ and we can obtain that $\alpha = 2$ from Theorem 4.2 for this setting of ψ. More details on experimental details are shown in Appendix E.

Metrics. We use the abstention loss $L_{0\text{-}1\text{-}c}$ for the predictor-rejector method as an evaluation metric. The abstention loss can be seen as integrating the information of zero-one misclassification error on the accepted samples and rejection ratio, which provides a comprehensive comparison between different methods.

Performance. The performances of our method and compared methods on the above three metrics on CIFAR-10 are shown in Table 1. The best results among all the methods with abstention are boldfaced.

Firstly, we can observe from Table 1 that our losses perform better than the vanilla CS loss on misclassification error, and abstention loss on CIFAR-10, which validates the effectiveness of the mechanism of abstention. Besides, our ACS loss shows the best performance of the misclassification error and abstention loss among all the compared methods on CIFAR-10. The training dynamics of our method and the compared methods on CIFAR-10 and the impact of abstention cost c of our method on CIFAR-10 are examined in Appendix E.

Acknowledgements

HZ was supported by JST/SPRING (Grant Number JPMJSP2110). JH was supported by JST/CREST Innovative Measurement and Analysis (Grant Number JPMJCR2333). HB was supported by JST/PRESTO Mathematical Sciences for the Future (Grant Number JPMJPR24K6).

References

K. Bache and M. Lichman. UCI machine learning repository, 2013, 2013.

P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3 (Nov):463–482, 2002.

P. L. Bartlett and M. H. Wegkamp. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9(59):1823–1840, 2008.

P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Y. Cao, T. Cai, L. Feng, L. Gu, J. Gu, B. An, G. Niu, and M. Sugiyama. Generalizing consistent multi-class classification with rejection to be compatible with arbitrary losses. *Advances in Neural Information Processing Systems*, 35:521–534, 2022.

Y. Cao, H. Bao, L. Feng, and B. An. Establishing linear surrogate regret bounds for convex smooth losses via convolutional Fenchel-Young losses. *arXiv preprint arXiv:2505.09432*, 2025.

C. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.

C.-K. Chow. An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers*, (6):247–254, 1957.

C. Cortes, G. DeSalvo, and M. Mohri. Boosting with abstention. *Advances in Neural Information Processing Systems*, 29:1668–1676, 2016a.

C. Cortes, G. DeSalvo, and M. Mohri. Learning with rejection. In *International Conference on Algorithmic Learning Theory*, pages 67–82. Springer, 2016b.

K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2(Dec):265–292, 2001.

G. DeSalvo, M. Mohri, and U. Syed. Learning with deep cascades. In *International Conference on Algorithmic Learning Theory*, pages 254–269. Springer, 2015.

R. El-Yaniv and Y. Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(5):1605–1641, 2010.

K. Filippova. Controlled hallucinations: Learning to generate faithfully from noisy data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 864–870, 2020.

Y. Geifman and R. El-Yaniv. Selective classification for deep neural networks. *Advances in Neural Information Processing Systems*, 30:4885–4894, 2017.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

R. Herbei and M. H. Wegkamp. Classification with reject option. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, pages 709–721, 2006.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. 2009. URL `https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf`.

X. Li, S. Liu, C. Sun, and H. Wang. When no-rejection learning is consistent for regression with rejection. In *International Conference on Artificial Intelligence and Statistics*, pages 1126–1134. PMLR, 2024.

A. Mao, M. Mohri, and Y. Zhong. Predictor-rejector multi-class abstention: Theoretical analysis and algorithms. In *International Conference on Algorithmic Learning Theory*, pages 822–867. PMLR, 2024a.

A. Mao, M. Mohri, and Y. Zhong. Theoretically grounded loss functions and algorithms for score-based multi-class abstention. In *International Conference on Artificial Intelligence and Statistics*, pages 4753–4761. PMLR, 2024b.

C. Mohri, D. Andor, E. Choi, and M. Collins. Learning to reject with a fixed predictor: Application to decontextualization. In *International Conference on Learning Representations*, 2024.

M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning*. MIT press, 2018.

H. Mozannar and D. Sontag. Consistent estimators for learning to defer to an expert. In *International Conference on Machine Learning*, pages 7076–7087. PMLR, 2020.

H. Mozannar, H. Lang, D. Wei, P. Sattigeri, S. Das, and D. Sontag. Who should predict? exact algorithms for learning to defer to humans. In *International Conference on Artificial Intelligence and Statistics*, pages 10520–10545. PMLR, 2023.

C. Ni, N. Charoenphakdee, J. Honda, and M. Sugiyama. On the calibration of multiclass classification with rejection. *Advances in Neural Information Processing Systems*, 32, 2019.

H. G. Ramaswamy, A. Tewari, and S. Agarwal. Consistent algorithms for multiclass classification with an abstain option. *Electronic Journal of Statistics*, 12:530–554, 2018.

S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge university press, 2014.

T. Silva Filho, H. Song, M. Perello-Nieto, R. Santos-Rodriguez, M. Kull, and P. Flach. Classifier calibration: a survey on how to assess and improve predicted class probabilities. *Machine Learning*, 112 (9):3211–3260, 2023.

A. Soen, H. Husain, P. Schulz, and V. Nguyen. Rejection via learning density ratios. *Advances in Neural Information Processing Systems*, 37:51845–51884, 2024.

Y. Wiener and R. El-Yaniv. Agnostic pointwise-competitive selective classification. *Journal of Artificial Intelligence Research*, 52:171–201, 2015.

T. Zhang. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5(Oct):1225–1251, 2004.

X.-Y. Zhang, G.-S. Xie, X. Li, T. Mei, and C.-L. Liu. A survey on learning to reject. *Proceedings of the IEEE*, 111(2):185–215, 2023.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Yes
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. No

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes
   (b) Complete proofs of all theoretical results. Yes
   (c) Clear explanations of any assumptions. Yes

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). No
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Yes
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Yes

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. Yes
   (b) The license information of the assets, if applicable. Not Applicable
   (c) New assets either in the supplemental material or as a URL, if applicable. Not Applicable
   (d) Information about consent from data providers/curators. Not Applicable
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. Not Applicable
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable

A Detailed Related Work

Learning with abstention has long been studied, starting with the work of (Chow, 1970), in which the trade-off between accuracy and rejection rate is researched. In recent years, it has drawn lots of attention Cortes et al. (2016b); Ramaswamy et al. (2018); Ni et al. (2019); Mao et al. (2024a); Li et al. (2024); Soen et al. (2024). Currently, methods of learning with abstention can be categorized into different groups based on their learning paradigm.

Confidence-based methods' (Chow, 1957, 1970; Ramaswamy et al., 2018; Bartlett and Wegkamp, 2008; Ni et al., 2019) main idea is to abstain if the score output by the predictor is lower than some threshold θ. Chow (1957, 1970) initiate the research on confidence-based methods. Bartlett and Wegkamp (2008) propose a convex surrogate loss depending on the abstention cost c. Ni et al. (2019) develop confidence-based methods for multi-classification setting.

Selective classification (El-Yaniv and Wiener, 2010; Wiener and El-Yaniv, 2015; Geifman and El-Yaniv, 2017) sets a predictor and a selector, and the predictor can choose to predict or reject according to the probability output by the selector. El-Yaniv and Wiener (2010) start the research of selective classification, in which they study the trade-off between the coverage and the accuracy of classifiers with a rejection option.Wiener and El-Yaniv (2015) study selective classification for specific hypothesis classes and families of distributions. Geifman and El-Yaniv (2017) consider selective classification in the context of deep neural networks.

Predictor-rejector formulation is based on learning a predictor and a rejector (Cortes et al., 2016b,a; Mohri et al., 2024; Mao et al., 2024a), which are from different function families. Cortes et al. (2016a) construct convex surrogate losses for the predictor-rejector formulation in the binary classification setting. Ni et al. (2019) show that calibration for the predictor-rejector formulation requires a quite unnatural constraint on the surrogate loss in the multi-class case. Mao et al. (2024a) derive calibrated but complicated nonconvex surrogate losses for the predictor-rejector formulation in the multi-class setting.

Score-based formulation (Mozannar and Sontag, 2020; Cao et al., 2022; Mao et al., 2024b) augments the class of multi-class classification (a class standing for rejection is added), and the learner will abstain if the score assigned to the augmentation class is the highest. Score-based formulation is introduced by Mozannar and Sontag (2020) in the setting of multi-class classification with abstention. Cao et al. (2022) propose a framework that can be equipped with an arbitrary surrogate loss function used in multi-class classification as long as they are calibrated. Mao et al. (2024b) introduce new families of surrogate loss with theoretical guarantees for score-based formulation.

B Useful Proposition

In this section, we show some useful propositions and their proofs.

B.1 Proof of Proposition B.1

Proposition B.1. *The additive CS (ACS) loss of the form in Section B.1 is convex jointly in* $(r, \boldsymbol{g})$.

Proof. Recall that the additive CS (ACS) loss with hinge ϕ can be expressed as the following form:

$$L_{\mathrm{ACS}}(r, \boldsymbol{g}) = \phi\left(g_y - \max_{y' \neq y} g_{y'} - r\right) + \psi(r).$$

Here, $\phi(x) \geqslant 0$ is a non-increasing convex function bounding $t \mapsto \mathbb{I}_{t \leqslant 0}$ from above. Similarly, $\psi(r)$ is a convex function bounding $t \mapsto \mathbb{I}_{t \leqslant 0}$ from above.

Let $\boldsymbol{u} = (\boldsymbol{g}, r)^\top$. Then for any $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^{K+1}$ and $0 \leqslant \theta \leqslant 1$, we have

$$
\begin{aligned}
L_{\mathrm{ACS}}\left(\theta\boldsymbol{u} + (1-\theta)\boldsymbol{v}\right) &= \phi\left(\left(\theta\boldsymbol{u} + (1-\theta)\boldsymbol{v}\right)_y - \max_{y' \in [K]/\{y\}} \left\{\left(\theta\boldsymbol{u} + (1-\theta)\boldsymbol{v}\right)_{y'}\right\} - \left(\theta\boldsymbol{u} + (1-\theta)\boldsymbol{v}\right)_{K+1}\right) \\
&\quad + \psi\left(\left(\theta\boldsymbol{u} + (1-\theta)\boldsymbol{v}\right)_{K+1}\right) \\
&\leqslant \phi\left(\left(\theta\boldsymbol{u} + (1-\theta)\boldsymbol{v}\right)_y - \theta \max_{y' \in [K]/\{y\}} \boldsymbol{u} - (1-\theta) \max_{y' \in [K]/\{y\}} \boldsymbol{v} - \left(\theta\boldsymbol{u} + (1-\theta)\boldsymbol{v}\right)_{K+1}\right) \\
&\quad + \psi\left(\left(\theta\boldsymbol{u} + (1-\theta)\boldsymbol{v}\right)_{K+1}\right) \qquad \text{(sub-additivity of } \max(\cdot)\text{, } \phi \text{ is nonincreasing)} \\
&= \phi\left(\theta\left(\boldsymbol{u}_y - \max_{y' \in [K]/\{y\}} \boldsymbol{u} - \boldsymbol{u}_{K+1}\right) + (1-\theta)\left(\boldsymbol{v}_y - \max_{y' \in [K]/\{y\}} \boldsymbol{v} - \boldsymbol{v}_{K+1}\right)\right) \\
&\quad + \psi\left(\theta\boldsymbol{u}_{K+1} + (1-\theta)\boldsymbol{v}_{K+1}\right) \\
&\leqslant \theta\phi\left(\boldsymbol{u}_y - \max_{y' \in [K]/\{y\}} \boldsymbol{u} - \boldsymbol{u}_{K+1}\right) + (1-\theta)\phi\left(\boldsymbol{v}_y - \max_{y' \in [K]/\{y\}} \boldsymbol{v} - \boldsymbol{v}_{K+1}\right) \\
&\quad + \theta\psi(\boldsymbol{u}_{K+1}) + (1-\theta)\psi(\boldsymbol{v}_{K+1}) \qquad (\phi, \psi \text{ are convex}) \\
&= \theta\left(\phi\left(\boldsymbol{u}_y - \max_{y' \in [K]/\{y\}} \boldsymbol{u} - \boldsymbol{u}_{K+1}\right) + \psi\left(\boldsymbol{u}_{K+1}\right)\right) \\
&\quad + (1-\theta)\left(\phi\left(\boldsymbol{v}_y - \max_{y' \in [K]/\{y\}} \boldsymbol{v} - \boldsymbol{v}_{K+1}\right) + \psi\left(\boldsymbol{u}_{K+1}\right)\right) \\
&= \theta L_{\mathrm{ACS}}(\boldsymbol{u}) + (1-\theta) L_{\mathrm{ACS}}(\boldsymbol{v}).
\end{aligned}
$$

Therefore, L_{ACS} is convex jointly in $(r, \boldsymbol{g})$.

□

B.2 Proof of Proposition B.2

We can define the pointwise risk $W_{\text{0-1-}c}$ of loss $L_{\text{0-1-}c}$ at $\boldsymbol{x}$ by

$$
W_{\text{0-1-}c}\left(r(\boldsymbol{x}), \boldsymbol{g}(\boldsymbol{x}); \boldsymbol{\eta}(\boldsymbol{x})\right) = \sum_y \eta_y(\boldsymbol{x}) L_{\text{0-1-}c}(r, \boldsymbol{g}; \boldsymbol{x}, y), \tag{2}
$$

where $\boldsymbol{\eta}(\boldsymbol{x}) = (\eta_1(\boldsymbol{x}), \cdots, \eta_K(\boldsymbol{x}))^\top$ is class probability vector for $\eta_y(\boldsymbol{x}) = p(y|\boldsymbol{x})$. Next, for simplicity, we omit the notation of $\boldsymbol{x}$ and use $W_{\text{0-1-}c}\left(r, \boldsymbol{g}; \boldsymbol{\eta}\right)$ to denote $W_{\text{0-1-}c}\left(r(\boldsymbol{x}), \boldsymbol{g}(\boldsymbol{x}); \boldsymbol{\eta}(\boldsymbol{x})\right)$ for the pointwise risk, and use η_y to denote $\eta_y(\boldsymbol{x})$ for the class probability.

Note that over the set of all measurable functions, the minimization of $R_{\text{0-1-}c}(r, \boldsymbol{g})$ in Eq. (1) with respect to $(r, \boldsymbol{g})$ is equivalent to the minimization of $W_{\text{0-1-}c}\left(r, \boldsymbol{g}; \boldsymbol{\eta}\right)$ with respect to $(r, \boldsymbol{g})$. Thus to minimize $R_{\text{0-1-}c}(r, \boldsymbol{g})$, it is just needed to consider minimizing pointwise risk $W_{\text{0-1-}c}\left(r, \boldsymbol{g}; \boldsymbol{\eta}\right)$.

We can also define the pointwise risk W of loss L at $\boldsymbol{x}$ in a similar way to Eq. (2) by

$$
W\left(r, \boldsymbol{g}; \boldsymbol{\eta}\right) = \sum_y \eta_y L(r, \boldsymbol{g}; \boldsymbol{x}, y). \tag{3}
$$

We denote by

$$
\left(r^*_{\boldsymbol{\eta}}, \boldsymbol{g}^*_{\boldsymbol{\eta}}\right) \in \operatorname*{argmin}_{r, \boldsymbol{g}} W(r, \boldsymbol{g}; \boldsymbol{\eta})
$$

a minimizer of $W(r, \boldsymbol{g}; \boldsymbol{\eta})$.

Proposition B.2. *There exists $\boldsymbol{g}^*_{\boldsymbol{\eta}} = (g^*_1, g^*_2, \ldots, g^*_K)$ such that $g_y = g^*_y - \Delta$ for some $\Delta \geqslant 0$ and all $y \neq y^*$, where $y^* = \operatorname{argmax}_y \eta_y$. Furthermore, for any $\boldsymbol{g}^*_{\boldsymbol{\eta}}$ we can express $W(r, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta}) = h_{\eta^*}(r)$ for $\eta^* = \max_y \eta_y$ and*

$$
h_\eta(r) = \min_{\Delta \geqslant 0} \left\{\eta\Phi(\Delta, r) + (1-\eta)\Phi(-\Delta, r)\right\},
$$

where $0 \leqslant \eta \leqslant 1$, and $\Phi(\Delta, r) : \mathbb{R}^2 \to \mathbb{R}$ is a function monotonically non-increasing in Δ, which can be in the form of $\Phi(\Delta, r) = \phi(\Delta)\psi_1(r) + \psi_2(r)$ or $\Phi(\Delta, r) = \phi(\Delta) + \psi(r)$, where $\phi(\Delta), \psi_1(r), \psi_2(r) \geqslant 0$ and ϕ is a non-increasing function.

Proof. Let $g^*_{(1)} \geqslant g^*_{(2)} \geqslant \ldots \geqslant g^*_{(K)}$ be the sorted sequence of $g^*_{\boldsymbol{\eta}} = (g^*_1, g^*_2, \cdots, g^*_K)$. Here letting $g_{(i+1)} := g_{(i)}$ $(i = 2, \cdots, K-1)$ never increases the objective function (3) and there exists $\boldsymbol{g}^*_{\boldsymbol{\eta}}$ such that g^*_y is the same for $y \neq \tilde{y}^*$, where $\tilde{y}^* \in \operatorname{argmax}_y g^*_y$. Then $W(r, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta})$ can be expressed as

$$\begin{aligned}
W(r, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta}) &= \sum_y \eta_y \Phi\left(g^*_y - \max_{y' \neq y} g^*_{y'}, r\right) \\
&= \eta_{\tilde{y}^*} \Phi\left(g^*_{(1)} - g^*_{(2)}, r\right) + \sum_{y \neq \tilde{y}^*} \eta_y \Phi\left(g^*_{(2)} - g^*_{(1)}, r\right) \\
&= \eta_{\tilde{y}^*} \Phi\left(\underbrace{g^*_{(1)} - g^*_{(2)}}_{\Delta^* \geqslant 0}, r\right) + (1 - \eta_{\tilde{y}^*}) \Phi\left(\underbrace{g^*_{(2)} - g^*_{(1)}}_{-\Delta^* \leqslant 0}, r\right) \\
&:= \min_{\Delta \geqslant 0} \left\{\eta_{\tilde{y}^*} \Phi(\Delta, r) + (1 - \eta_{\tilde{y}^*}) \Phi(-\Delta, r)\right\},
\end{aligned}$$

which is minimized by $\tilde{y}^* = y^*$ (i.e., $\operatorname{argmax}_y g^*_y = \operatorname{argmax}_y \eta_y$) and in this case $\eta_{\tilde{y}^*} = \eta^*$.

Specifically, for the form of $\Phi(\Delta, r) = \phi(\Delta)\psi_1(r) + \psi_2(r)$ or $\Phi(\Delta, r) = \phi(\Delta) + \psi(r)$, where $\phi(\Delta), \psi_1(r), \psi_2(r) \geqslant 0$ and ϕ is a non-increasing function, the above relationship holds. □

C Useful Lemmas

In this section, we show some useful lemmas and their proofs.

The following lemma gives specific forms of H_η in Theorem 4.1 for different ϕ.

Lemma C.1. *For $\phi(\Delta) = \exp(-\Delta)$, we have*

$$H_\eta = \begin{cases} 2\sqrt{\eta(1-\eta)} & \text{if } \eta \geqslant \frac{1}{2}, \\ 1 & \text{if } \eta < \frac{1}{2}. \end{cases}$$

For $\phi(\Delta) = \max\{0, 1 - \Delta\}$, we have

$$H_\eta = \begin{cases} 2(1-\eta) & \text{if } \eta \geqslant \frac{1}{2}, \\ 1 & \text{if } \eta < \frac{1}{2}. \end{cases}$$

For $\phi(\Delta) = \log(1 + \exp(-\Delta))$, we have

$$H_\eta = \begin{cases} -\eta \log \eta - (1-\eta)\log(1-\eta) & \text{if } \eta \geqslant \frac{1}{2}, \\ \log 2 & \text{if } \eta < \frac{1}{2}. \end{cases}$$

Proof. Recall that H_η is defined as

$$H_\eta = \min_{\Delta \geqslant 0} \left\{\eta\phi(\Delta) + (1-\eta)\phi(-\Delta)\right\},$$

where $0 \leqslant \eta \leqslant 1$. Let $F(\Delta) = \eta\phi(\Delta) + (1-\eta)\phi(-\Delta)$, where $0 \leqslant \eta \leqslant 1$ and $\Delta \geqslant 0$. Then, we analyze three cases to derive the expression of H_η.

(1) Case $\phi(\Delta) = \exp(-\Delta)$:

In this case, $F(\Delta) = \eta \exp(-\Delta) + (1-\eta)\exp(\Delta)$, which is convex in Δ. Its derivative is expressed as $F'(\Delta) = -\eta\exp(-\Delta) + (1-\eta)\exp(\Delta)$. Let $\Delta^\dagger \in \mathbb{R}$ be such that $F'(\Delta^\dagger) = 0$. Then we see that

$$\Delta^\dagger = \frac{1}{2}\log\left(\frac{\eta}{1-\eta}\right).$$

Due to the convexity, we have that $F'(\Delta)$ is increasing for all $\Delta \geqslant 0$. First consider the case $\eta \geqslant \frac{1}{2}$. In this case we have $\Delta^\dagger \geqslant 0$. Then $F(\Delta) \geqslant F(\Delta^\dagger) = 2\sqrt{\eta(1-\eta)},$ for $\Delta \geqslant 0$.

Next consider the case $\eta < \frac{1}{2}$. In this case we have $\Delta^\dagger < 0$. Then $F(\Delta) \geqslant F(0) = 1,$ for $\Delta \geqslant 0$.

Thus we have

$$H_\eta = \min_{\Delta \geqslant 0} F(\Delta) = \begin{cases} 2\sqrt{\eta(1-\eta)} & \text{if } \eta \geqslant \frac{1}{2}, \\ 1 & \text{if } \eta < \frac{1}{2}. \end{cases}$$

(2) Case $\phi(\Delta) = \max\{0, 1-\Delta\}$:

In this case, $F(\Delta) = \eta\max\{0,1-\Delta\} + (1-\eta)\max\{0,1+\Delta\}$. First consider the case $\eta \geqslant \frac{1}{2}$. In this case, we have $F(\Delta) \geqslant F(1) = 2(1-\eta)$, for $\Delta \geqslant 0$.

Next consider the case $\eta < \frac{1}{2}$. In this case, we have $F(\Delta) \geqslant F(0) = 1,$ for $\Delta \geqslant 0$.

Thus we have

$$H_\eta = \min_{\Delta \geqslant 0} F(\Delta) = \begin{cases} 2(1-\eta) & \text{if } \eta \geqslant \frac{1}{2}, \\ 1 & \text{if } \eta < \frac{1}{2}. \end{cases}$$

(3) Case $\phi(\Delta) = \log(1+\exp(-\Delta))$:

In this case, $F(\Delta) = \eta\log(1+\exp(-\Delta)) + (1-\eta)\log(1+\exp(\Delta))$, which is convex in Δ. Its derivative is expressed as $F'(\Delta) = -\eta\frac{1}{1+\exp(\Delta)} + (1-\eta)\frac{1}{1+\exp(-\Delta)}$. Let $\Delta^\dagger \in \mathbb{R}$ be such that $F'(\Delta^\dagger) = 0$. Then we see that

$$\Delta^\dagger = \log\left(\frac{\eta}{1-\eta}\right).$$

Due to the convexity, we have that $F'(\Delta)$ is increasing for all $\Delta \geqslant 0$. First consider the case $\eta^* \geqslant \frac{1}{2}$. In this case, we have $\Delta^\dagger \geqslant 0$. Then $F(\Delta) \geqslant F(\Delta^\dagger) = -\eta\log\eta - (1-\eta)\log(1-\eta),$ for $\Delta \geqslant 0$.

Next consider the case $\eta < \frac{1}{2}$. In this case we have $\Delta^\dagger < 0$. Then $F(\Delta) \geqslant F(0) = \log 2,$ for $\Delta \geqslant 0$.

Thus we have

$$H_\eta = \min_{\Delta \geqslant 0} F(\Delta) = \begin{cases} -\eta\log\eta - (1-\eta)\log(1-\eta) & \text{if } \eta \geqslant \frac{1}{2}, \\ \log 2 & \text{if } \eta < \frac{1}{2}. \end{cases}$$

□

The following lemma gives the minimum of $W_{\mathrm{MCS}}(r, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta})$ with respect to r, and the corresponding minimizer $r^*_{\boldsymbol{\eta}} \in \operatorname{argmin}_r W(r, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta})$ with specific choices of ϕ, ψ_1, and ψ_2.

Lemma C.2. *For MCS pointwise loss of the form*

$$W_{\mathrm{MCS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) = \sum_y \eta_y \phi\left(g_y - \max_{y' \neq y} g_{y'}\right)\psi_1(r) + \psi_2(r), \tag{4}$$

with $\phi(\Delta) = \exp(-\Delta)$, $\psi_1(r) = \exp(r)$, *and* $\psi_2(r) = c\exp(-\alpha r)$, *we have*

$$W_{\mathrm{MCS}}(r^*_{\boldsymbol{\eta}}, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta}) = \begin{cases} (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} \left(2\sqrt{\eta^*(1-\eta^*)}\right)^{\frac{\alpha}{\alpha+1}} & \text{if } \eta^* \geqslant \frac{1}{2}, \\ (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} & \text{if } \eta^* < \frac{1}{2}, \end{cases}$$

and

$$r^*_{\boldsymbol{\eta}} = \begin{cases} \frac{1}{\alpha+1}\log\left(\frac{\alpha c}{2\sqrt{\eta^*(1-\eta^*)}}\right) & \text{if } \eta^* \geqslant \frac{1}{2}, \\ \frac{1}{\alpha+1}\log(\alpha c) & \text{if } \eta^* < \frac{1}{2}. \end{cases}$$

Proof. From Eq. (4), the minimization of W with respect to $\boldsymbol{g}$ does not depend on r. Therefore, by Proposition B.2, we have

$$W_{\mathrm{MCS}}(r, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta}) = H_{\eta^*}\psi_1(r) + \psi_2(r),$$

where recall that

$$H_\eta = \min_{\Delta \geqslant 0} \{\eta\phi(\Delta) + (1-\eta)\phi(-\Delta)\},$$

for $0 \leqslant \eta \leqslant 1$. For $\phi(\Delta) = \exp(-\Delta)$, from Lemma C.1, we have

$$H_{\eta^*} = \begin{cases} 2\sqrt{\eta^*(1-\eta^*)} & \text{if } \eta^* \geqslant \frac{1}{2}, \\ 1 & \text{if } \eta^* < \frac{1}{2}, \end{cases}$$

where $\eta^* = \max_y \eta_y$. Thus for $\psi_1(r) = \exp(r)$, $\psi_2(r) = c\exp(-\alpha r)$, we have

$$W_{\mathrm{MCS}}(r, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta}) = \begin{cases} 2\sqrt{\eta^*(1-\eta^*)}\exp(r) + c\exp(-\alpha r) & \text{if } \eta^* \geqslant \frac{1}{2}, \\ \exp(r) + c\exp(-\alpha r) & \text{if } \eta^* < \frac{1}{2}, \end{cases} \tag{5}$$

which is a convex function respect to r. Its derivative is expressed as

$$\frac{\partial W_{\mathrm{MCS}}(r, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta})}{\partial r} = \begin{cases} 2\sqrt{\eta^*(1-\eta^*)}\exp(r) - \alpha c\exp(-\alpha r) & \text{if } \eta^* \geqslant \frac{1}{2}, \\ \exp(r) - \alpha c\exp(-\alpha r) & \text{if } \eta^* < \frac{1}{2}. \end{cases} \tag{6}$$

Since the minimizer $r^*_{\boldsymbol{\eta}}$ satisifies the first-order optimality condition $\frac{\partial W_{\mathrm{MCS}}(r, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta})}{\partial r} |_{r=r^*_{\boldsymbol{\eta}}} = 0$, we have

$$r^*_{\boldsymbol{\eta}} = \begin{cases} \frac{1}{\alpha+1}\log\left(\frac{\alpha c}{2\sqrt{\eta^*(1-\eta^*)}}\right) & \text{if } \eta^* \geqslant \frac{1}{2}, \\ \frac{1}{\alpha+1}\log(\alpha c) & \text{if } \eta^* < \frac{1}{2}. \end{cases} \tag{7}$$

By plugging it into $W_{\mathrm{MCS}}(r, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta})$, we have

$$W_{\mathrm{MCS}}(r^*_{\boldsymbol{\eta}}, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta}) = \begin{cases} (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} \left(2\sqrt{\eta^*(1-\eta^*)}\right)^{\frac{\alpha}{\alpha+1}} & \text{if } \eta^* \geqslant \frac{1}{2}, \\ (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} & \text{if } \eta^* < \frac{1}{2}. \end{cases} \tag{8}$$

□

The following lemma gives a beneficial result of exponential loss, which will be used in the proof of theorem about MCS loss.

Lemma C.3. *For function*

$$k(r) = Ae^{r} + Be^{-\alpha r},$$

where $A, B > 0$ *and* $\alpha > 0$*, we have*

$$k(r) > A + B$$

if $A \geqslant \alpha B, r > 0$*, or if* $A < \alpha B, r \leqslant 0$.

Proof. We can see that $k(r)$ is convex in r. Its derivative is expressed as

$$k'(r) = Ae^{r} - \alpha Be^{-\alpha r}.$$

Let $r^{\dagger} \in \mathbb{R}$ be such that $k'(r^{\dagger}) = 0$. Then we have

$$r^{\dagger} = -\frac{1}{\alpha + 1} \log \frac{A}{\alpha B}.$$

Due to the convexity, we have that $k'(r)$ is increasing for all r. Then we have the following results:

(**i**) If $A \geqslant \alpha B$: we have $r^{\dagger} \leqslant 0$. Then we have $k'(r) > k'(0) \geqslant 0$ for $r > 0$, which implies $k(r)$ is increasing when $r > 0$, then

$$k(r) > k(0) = A + B, \text{ for } r > 0.$$

(**ii**) If $A < \alpha B$: we have $r^{\dagger} > 0$, then we have $k'(r) \leqslant k'(0) < 0$ for $r \leqslant 0$, which implies $k(r)$ is decreasing when $r \leqslant 0$, then

$$k(r) > k(0) = A + B, \text{ for } r \leqslant 0.$$

□

The following lemma gives a useful lower bound for a function of α and c.

Lemma C.4. *For* $0 < c < \frac{1}{2}$ *and* $\alpha = 2\sqrt{\frac{1-c}{c}} > 2$*, we have*

$$(\alpha + 1)\alpha^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} > \frac{5}{12}.$$

Proof.

$$\begin{aligned}
(\alpha + 1)\alpha^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} &\geqslant (\alpha + 1)(\alpha + 1)^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} \\
&= ((\alpha + 1)c)^{\frac{1}{\alpha+1}} \\
&\geqslant 1 - \frac{1}{\alpha + 1}\left(\frac{1}{(\alpha+1)c} - 1\right) && (x^{-\frac{1}{\alpha+1}} \text{ is convex in } x \text{ for } x > 0) \\
&= 1 - \frac{\alpha^2 + 4}{4(\alpha+1)^2} - \frac{1}{\alpha + 1} && (c = \tfrac{4}{4+\alpha^2}) \\
&\geqslant 1 - \frac{(\alpha+1)^2}{4(\alpha+1)^2} - \frac{1}{\alpha + 1} && (\alpha > 2) \\
&= \frac{3}{4} - \frac{1}{\alpha + 1} > \frac{5}{12}. && (\alpha > 2)
\end{aligned}$$

□

The following lemma gives the infimum for function F of $0 \leqslant \eta \leqslant 1$ and $\Delta \geqslant 0$.

Lemma C.5. *Define $F(r, \Delta; \eta)$ as*

$$F(r, \Delta; \eta) = \eta \max\{1 - \Delta + r, 0\} + (1 - \eta) \max\{1 + \Delta + r, 0\}, \tag{9}$$

where $0 \leqslant \eta \leqslant 1$ and $\Delta \geqslant 0$. Then we have

$$\inf_{\Delta \geqslant 0} F(r, \Delta; \eta) = \begin{cases} 2(1-\eta)(r+1), & \text{if } \eta > \frac{1}{2} \text{ and } r \geqslant -1, \\ r + 1, & \text{if } \eta \leqslant \frac{1}{2} \text{ and } r \geqslant -1, \\ 0 & \text{if } r < -1. \end{cases}$$

Proof. First consider the case $\eta > \frac{1}{2}$. In this case, Δ to minimize $F(r, \Delta, \eta)$ is expressed as $\Delta = \max\{r + 1, 0\}$, and we have

$$\begin{aligned} F(r, \Delta; \eta) &\geqslant F(r, \max\{r+1, 0\}; \eta) \\ &= \begin{cases} 2(1-\eta)(r+1) & \text{if } r \geqslant -1, \\ 0 & \text{if } r < -1, \end{cases} \end{aligned}$$

where $\Delta \geqslant 0$.

Next consider the case $\eta \leqslant \frac{1}{2}$. In this case, the optimal Δ is $\Delta = \max\{-r - 1, 0\}$, and we have

$$\begin{aligned} F(r, \Delta; \eta) &\geqslant F(r, \max\{-r-1, 0\}; \eta) \\ &= \begin{cases} r + 1 & \text{if } r \geqslant -1, \\ 0 & \text{if } r < -1, \end{cases} \end{aligned}$$

where $\Delta \geqslant 0$.

Thus we have

$$\inf_{\Delta \geqslant 0} F(r, \Delta; \eta) = \begin{cases} 2(1-\eta)(r+1) & \text{if } \eta > \frac{1}{2} \text{ and } r \geqslant -1, \\ r + 1 & \text{if } \eta \leqslant \frac{1}{2} \text{ and } r \geqslant -1, \\ 0 & \text{if } r < -1. \end{cases}$$

□

The following lemma gives the minimum for $W_{\mathrm{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta})$ with hinge ϕ and different cases of ψ.

Lemma C.6. *For ACS pointwise loss of the form*

$$W_{\mathrm{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) = \sum_y \eta_y \phi\left(g_y - \max_{y' \neq y} g_{y'} - r\right) + \psi(r), \tag{10}$$

with $\phi(\Delta) = \max\{1 - \Delta, 0\}$, we have

$$W_{\mathrm{ACS}}(r^*_{\boldsymbol{\eta}}, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta}) = \begin{cases} 2(1-\eta^*)\left(r^*_{\boldsymbol{\eta}} + 1\right) + \psi\left(r^*_{\boldsymbol{\eta}}\right) & \text{if } \eta > \frac{1}{2}, \\ \left(r^*_{\boldsymbol{\eta}} + 1\right) + \psi\left(r^*_{\boldsymbol{\eta}}\right) & \text{if } \eta^* \leqslant \frac{1}{2}, \end{cases}$$

where

$$r^*_{\boldsymbol{\eta}} = \begin{cases} (\psi')^{-1}(-2(1-\eta^*)) & \text{if } \eta^* > \frac{1}{2}, \\ (\psi')^{-1}(-1) & \text{if } \eta^* \leqslant \frac{1}{2}, \end{cases}$$

for $\psi(r) = c\exp(-\alpha r)$ *or* $\psi(r) = \log\left(1+\exp(-\alpha r)\right)$, *and*

$$r^*_{\boldsymbol{\eta}} = \begin{cases} -1 & \text{if } \frac{1}{2} < \eta^* \leqslant 1 - \frac{\alpha c}{2}, \\ \frac{1}{\alpha} & \text{if } \eta^* > \frac{1}{2} \text{ and } \eta^* > 1 - \frac{\alpha c}{2}, \\ -1 & \text{if } \eta^* \leqslant \frac{1}{2} \text{ and } \alpha c \leqslant 1, \\ \frac{1}{\alpha} & \text{if } \eta^* \leqslant \frac{1}{2} \text{ and } \alpha c > 1, \end{cases}$$

for $\psi(r) = c\max\{1-\alpha r, 0\}$.

Proof. By Proposition B.2, Eqs. (9) and (10), we have

$$\begin{aligned} W_{\mathrm{ACS}}\left(r, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta}\right) &= \inf_{\Delta \geqslant 0} \left\{\eta^* \max\{1-\Delta+r, 0\} + (1-\eta^*)\max(1+\Delta+r, 0)\right\} + \psi(r) \\ &= \inf_{\Delta \geqslant 0} F(r, \Delta; \eta^*) + \psi(r), \end{aligned}$$

Using Lemma C.5, we have

$$W_{\mathrm{ACS}}(r, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta}) = \begin{cases} 2(1-\eta^*)(r+1) + \psi(r) & \text{if } \eta^* > \frac{1}{2} \text{ and } r \geqslant -1, \\ r + 1 + \psi(r) & \text{if } \eta^* \leqslant \frac{1}{2} \text{ and } r \geqslant -1, \\ \psi(r) & \text{if } r < -1. \end{cases} \tag{11}$$

By taking the derivative, we obtain

$$\frac{\partial W_{\mathrm{ACS}}(r, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta})}{\partial r} = \begin{cases} 2(1-\eta^*) + \psi'(r) & \text{if } \eta^* > \frac{1}{2} \text{ and } r \geqslant -1, \\ 1 + \psi'(r) & \text{if } \eta^* \leqslant \frac{1}{2} \text{ and } r \geqslant -1, \\ \psi'(r) & \text{if } r < -1. \end{cases} \tag{12}$$

Then, we analyze two cases to obtain the expression of $W_{\mathrm{ACS}}(r^*_{\boldsymbol{\eta}}, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta})$.

(**i**) Case $\psi(r) = c\exp(-\alpha r)$ or $\phi(r) = \log\left(1+\exp(-\alpha r)\right)$:

In this case, for $\eta^* > \frac{1}{2}$ and $r \geqslant -1$, or $\eta^* \leqslant \frac{1}{2}$ and $r \geqslant -1$, the necessary and sufficient condition for r to be a minimizer of W_{ACS} is $\frac{\partial W_{\mathrm{ACS}}(r, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta})}{\partial r} = 0$, and then there is a unique minimizer

$$r^*_{\boldsymbol{\eta}} = \begin{cases} (\psi')^{-1}\left(-2(1-\eta^*)\right) & \text{if } \eta^* > \frac{1}{2} \text{ and } r \geqslant -1, \\ (\psi')^{-1}(-1) & \text{if } \eta^* \leqslant \frac{1}{2} \text{ and } r \geqslant -1. \end{cases} \tag{13}$$

By plugging it into $W_{\mathrm{ACS}}(r, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta})$, we have

$$\inf_{r \geqslant -1} W_{\mathrm{ACS}}(r, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta}) = \begin{cases} 2(1-\eta^*)\left(r^*_{\boldsymbol{\eta}} + 1\right) + \psi\left(r^*_{\boldsymbol{\eta}}\right) & \text{if } \eta^* > \frac{1}{2}, \\ \left(r^*_{\boldsymbol{\eta}} + 1\right) + \psi\left(r^*_{\boldsymbol{\eta}}\right) & \text{if } \eta^* \leqslant \frac{1}{2}. \end{cases} \tag{14}$$

For $r < -1$, since $\frac{\partial W_{\mathrm{ACS}}(r, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta})}{\partial r} < 0$, $W_{\mathrm{ACS}}(r, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta})$ is nonincreasing for $r < -1$, we have

$$\inf_{r < -1} W_{\mathrm{ACS}}(r, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta}) = W_{\mathrm{ACS}}(-1, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta}) = \psi(-1).$$

Since $\psi(r)$ is convex, it holds for r_1, r_2 that

$$\psi(r_1) \geqslant \psi(r_2) + \psi\prime(r_2)(r_1 - r_2).$$

Then for $r_1 = -1, r_2 = (\psi')^{-1}(-2(1-\eta^*))$, we have

$$\begin{aligned}\psi(-1) &\geqslant \psi\left((\psi')^{-1}(-2(1-\eta^*))\right) + \psi'\left((\psi')^{-1}(-2(1-\eta^*))\right)\left(-1-(\psi')^{-1}(-2(1-\eta^*))\right)\\ &= \psi\left((\psi')^{-1}(-2(1-\eta^*))\right) - 2(1-\eta^*)\left(-1-(\psi')^{-1}(-2(1-\eta^*))\right). \end{aligned} \tag{15}$$

Hence, we have

$$\begin{aligned}&2(1-\eta^*)\left(r_{\boldsymbol{\eta}}^*+1\right)+\psi\left(r_{\boldsymbol{\eta}}^*\right)\\ &= 2(1-\eta^*)\left((\psi')^{-1}(-2(1-\eta))+1\right)+\psi\left((\psi')^{-1}(-2(1-\eta^*))\right) && \text{(result by Eq. (13))}\\ &\leqslant \psi(-1),\end{aligned}$$

where $\frac{1}{2} < \eta^* \leqslant 1$.

For $(r_{\boldsymbol{\eta}}^*+1)+\psi(r_{\boldsymbol{\eta}}^*)$, note that $(r_{\boldsymbol{\eta}}^*+1)+\psi(r_{\boldsymbol{\eta}}^*) = \left((\psi')^{-1}(-1)+1\right)+\psi\left((\psi')^{-1}(-1)\right)$ is a special case of $2(1-\eta^*)\left((\psi')^{-1}(-2(1-\eta))+1\right)+\psi\left((\psi')^{-1}(-2(1-\eta^*))\right)$ when $\eta^* = \frac{1}{2}$. Then from Eq. (15), we have $(r_{\boldsymbol{\eta}}^*+1)+\psi(r_{\boldsymbol{\eta}}^*) \leqslant \psi(-1)$.

Therefore

$$\begin{aligned}W_{\mathrm{ACS}}(r_{\boldsymbol{\eta}}^*, \boldsymbol{g}_{\boldsymbol{\eta}}^*; \boldsymbol{\eta}) &= \begin{cases}\min\left\{2(1-\eta^*)\left(r_{\boldsymbol{\eta}}^*+1\right)+\psi\left(r_{\boldsymbol{\eta}}^*\right), \psi(-1)\right\} & \text{if } \eta^* > \frac{1}{2},\\ \min\left\{\left(r_{\boldsymbol{\eta}}^*+1\right)+\psi\left(r_{\boldsymbol{\eta}}^*\right), \psi(-1)\right\} & \text{if } \eta^* \leqslant \frac{1}{2},\end{cases}\\ &= \begin{cases}2(1-\eta^*)\left(r_{\boldsymbol{\eta}}^*+1\right)+\psi\left(r_{\boldsymbol{\eta}}^*\right) & \text{if } \eta^* > \frac{1}{2},\\ \left(r_{\boldsymbol{\eta}}^*+1\right)+\psi\left(r_{\boldsymbol{\eta}}^*\right) & \text{if } \eta^* \leqslant \frac{1}{2}.\end{cases}\end{aligned}$$

(**ii**) Case $\psi(r) = c\max\{1-\alpha r, 0\}$:

In this case, by Eq. (11) we have

$$W_{\mathrm{ACS}}(r, \boldsymbol{g}_{\boldsymbol{\eta}}^*; \boldsymbol{\eta}) = \begin{cases}2(1-\eta^*)(r+1)+c\max\{1-\alpha r, 0\} & \text{if } \eta^* > \frac{1}{2} \text{ and } r \geqslant -1,\\ r+1+c\max\{1-\alpha r, 0\} & \text{if } \eta^* \leqslant \frac{1}{2} \text{ and } r \geqslant -1,\\ c\max\{1-\alpha r, 0\} & \text{if } r < -1.\end{cases}$$

Then for $\frac{1}{2} < \eta^* \leqslant 1 - \frac{\alpha c}{2}$ and $r \geqslant -1$, or $\eta^* > \frac{1}{2}$ and $\eta^* > 1 - \frac{\alpha c}{2}$ and $r \geqslant -1$, we have

$$r_{\boldsymbol{\eta}}^* = \begin{cases}-1 & \text{if } \frac{1}{2} < \eta^* \leqslant 1 - \frac{\alpha c}{2} \text{ and } r \geqslant -1,\\ \frac{1}{\alpha} & \text{if } \eta^* > \frac{1}{2}, \eta^* > 1 - \frac{\alpha c}{2} \text{ and } r \geqslant -1.\end{cases} \tag{16}$$

For $\eta^* \leqslant \frac{1}{2}$ and $r \geqslant -1$, we have

$$r_{\boldsymbol{\eta}}^* = \begin{cases}-1 & \text{if } \eta^* \leqslant \frac{1}{2}, \alpha c \leqslant 1 \text{ and } r \geqslant -1,\\ \frac{1}{\alpha} & \text{if } \eta^* \leqslant \frac{1}{2}, \alpha c > 1 \text{ and } r \geqslant -1.\end{cases} \tag{17}$$

By plugging it into $W_{\mathrm{ACS}}(r, \boldsymbol{g}_{\boldsymbol{\eta}}^*; \boldsymbol{\eta})$, we have

$$\inf_{r > -1} W_{\mathrm{ACS}}(r, \boldsymbol{g}_{\boldsymbol{\eta}}^*; \boldsymbol{\eta}) = \begin{cases}2(1-\eta^*)\left(r_{\boldsymbol{\eta}}^*+1\right)+\psi\left(r_{\boldsymbol{\eta}}^*\right) & \text{if } \eta > \frac{1}{2},\\ \left(r_{\boldsymbol{\eta}}^*+1\right)+\psi\left(r_{\boldsymbol{\eta}}^*\right) & \text{if } \eta^* \leqslant \frac{1}{2}.\end{cases}$$

For $r < -1$, we have

$$\inf_{r < -1} W_{\mathrm{ACS}}(r, \boldsymbol{g}_{\boldsymbol{\eta}}^*; \boldsymbol{\eta}) = W_{\mathrm{ACS}}(-1, \boldsymbol{g}_{\boldsymbol{\eta}}^*; \boldsymbol{\eta}) = (\alpha+1)c.$$

We can see that $2(1-\eta^*)\left(r^*_{\boldsymbol{\eta}}+1\right)+\psi\left(r^*_{\boldsymbol{\eta}}\right)=2(1-\eta^*)(\frac{1}{\alpha}+1)\leqslant(\alpha+1)c$ holds for $\eta^*>\frac{1}{2}$ and $\eta^*>1-\frac{\alpha c}{2}$, and $2(1-\eta^*)\left(r^*_{\boldsymbol{\eta}}+1\right)+\psi\left(r^*_{\boldsymbol{\eta}}\right)=(\alpha+1)c$ for $\frac{1}{2}<\eta^*\leqslant 1-\frac{\alpha c}{2}$ from Eq. (16). Besides, we can also see that $\left(r^*_{\boldsymbol{\eta}}+1\right)+\psi\left(r^*_{\boldsymbol{\eta}}\right)=\min\left\{(\alpha+1)c,\frac{1}{\alpha}+1\right\}\leqslant(\alpha+1)c$ holds for $\eta^*\leqslant\frac{1}{2}$ from Eq. (17).

Therefore

$$\begin{aligned}W_{\mathrm{ACS}}(r^*_{\boldsymbol{\eta}},\boldsymbol{g}^*_{\boldsymbol{\eta}};\boldsymbol{\eta})&=\begin{cases}\min\left\{2(1-\eta^*)\left(r^*_{\boldsymbol{\eta}}+1\right)+\psi\left(r^*_{\boldsymbol{\eta}}\right),(\alpha+1)c\right\} & \text{if } \eta^*>\frac{1}{2},\\ \min\left\{\left(r^*_{\boldsymbol{\eta}}+1\right)+\psi\left(r^*_{\boldsymbol{\eta}}\right),(\alpha+1)c\right\} & \text{if } \eta^*\leqslant\frac{1}{2},\end{cases}\\&=\begin{cases}2(1-\eta^*)\left(r^*_{\boldsymbol{\eta}}+1\right)+\psi\left(r^*_{\boldsymbol{\eta}}\right) & \text{if } \eta^*>\frac{1}{2},\\ \left(r^*_{\boldsymbol{\eta}}+1\right)+\psi\left(r^*_{\boldsymbol{\eta}}\right) & \text{if } \eta^*\leqslant\frac{1}{2}.\end{cases}\end{aligned}$$

Overall,

$$W_{\mathrm{ACS}}(r^*_{\boldsymbol{\eta}},\boldsymbol{g}^*_{\boldsymbol{\eta}};\boldsymbol{\eta})=\begin{cases}2(1-\eta^*)\left(r^*_{\boldsymbol{\eta}}+1\right)+\psi\left(r^*_{\boldsymbol{\eta}}\right) & \text{if } \eta>\frac{1}{2},\\ \left(r^*_{\boldsymbol{\eta}}+1\right)+\psi\left(r^*_{\boldsymbol{\eta}}\right) & \text{if } \eta^*\leqslant\frac{1}{2},\end{cases}$$

where

$$r^*_{\boldsymbol{\eta}}=\begin{cases}(\psi')^{-1}\left(-2(1-\eta^*)\right) & \text{if } \eta^*>\frac{1}{2},\\ (\psi')^{-1}\left(-1\right) & \text{if } \eta^*\leqslant\frac{1}{2},\end{cases}$$

for $\psi(r)=c\exp(-\alpha r)$ or $\psi(r)=\log\left(1+\exp(-\alpha r)\right)$, and

$$r^*_{\boldsymbol{\eta}}=\begin{cases}-1 & \text{if } \frac{1}{2}<\eta^*\leqslant 1-\frac{\alpha c}{2},\\ \frac{1}{\alpha} & \text{if } \eta^*>\frac{1}{2} \text{ and } \eta^*>1-\frac{\alpha c}{2},\\ -1 & \text{if } \eta^*\leqslant\frac{1}{2} \text{ and } \alpha c\leqslant 1,\\ \frac{1}{\alpha} & \text{if } \eta^*\leqslant\frac{1}{2} \text{ and } \alpha c>1,\end{cases}$$

for $\psi(r)=c\max\{1-\alpha r,0\}$. □

We then introduce an elementary but beneficial result about $\psi(r)$, which is useful in the proof of the excess risk bound for ACS.

Lemma C.7. *For function*

$$l(r)=A(r+1)+\psi(r),$$

where $A\geqslant 1$ and $\psi(r)$ is a nonincreasing convex function, we have

$$l(r)\geqslant A+\psi(0)$$

if $A\geqslant-\psi'(0), r>0$, or if $A<-\psi'(0), r\leqslant 0$, where ψ' can be the subdifferential of ψ.

Proof. Since $l(r)$ is convex and $l'(0)=A+\psi'(0)$, the above result can be obtained. □

We then introduce a useful inequality about logarithmic function.

Lemma C.8. *For $a,b>0$, we have*

$$\frac{a-b}{\log a-\log b}<\frac{a+b}{2}. \tag{18}$$

Proof. Without loss of generality, let $a>b>0$. Then $\frac{a-b}{\log a-\log b}<\frac{a+b}{2}$ holds if and only if

$$\log a-\log b>\frac{2(a-b)}{a+b}\Leftrightarrow\log\frac{a}{b}>\frac{2(\frac{a}{b}-1)}{\frac{a}{b}+1}. \tag{19}$$

Let $\frac{a}{b} = q$ $(q > 1)$. Then to prove Inequity (19), we only need to prove

$$\log q > \frac{2(q-1)}{q+1}.$$

Let

$$h(q) = \log q - \frac{2(q-1)}{q+1} = \log q + \frac{4}{q+1} - 2.$$

Then by taking the derivatives of $h(q)$, we obtain

$$h'(q) = \frac{1}{q} - \frac{4}{(q+1)^2} = \frac{(q-1)^2}{q(q+1)^2} > 0.$$

Thus $h(q)$ is increasing for $q > 1$. Then we have $h(q) > h(1) = 0$, which implies $\log q > \frac{2(q-1)}{q+1}$. Therefore, we have $\frac{a-b}{\log a - \log b} < \frac{a+b}{2}$. □

Lemma C.9. *For function*

$$(1-\eta)\log\left(\frac{1-\eta}{c}\right) - (1-\eta) + c \geqslant (1-c-\eta)^2,$$

where $\eta > \frac{1}{2}, c < \frac{1}{2}$ *and* $1 - \eta > c$.

Proof.

$$\begin{aligned}
(1-\eta)\log\left(\frac{1-\eta}{c}\right) - (1-\eta) + c &\geqslant (1-\eta)\frac{2(1-\eta-c)}{1-\eta+c} - (1-\eta) + c \\
&\qquad \text{(result from Lemma C.8 when choosing } a = 1-\eta,\ b = c) \\
&= \frac{2(1-\eta)(1-\eta-c) - (1-\eta-c)(1-\eta+c)}{1-\eta+c} \\
&= \frac{(1-\eta)^2 - 2c(1-\eta) + c^2}{1-\eta+c} \\
&= \frac{(1-\eta-c)^2}{1-\eta+c} \\
&\geqslant (1-\eta-c)^2 \qquad (1-\eta+c < 1 \text{ for } \eta > \tfrac{1}{2} \text{ and } c < \tfrac{1}{2}) \\
&= (1-c-\eta)^2.
\end{aligned}$$

□

For ACS pointwise loss of the form

$$W_{\mathrm{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) = \sum_y \eta_y \phi\left(g_y - \max_{y' \neq y} g_{y'} - r\right) + \psi(r),$$

from Proposition B.2, $W(r, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta})$ can be writed as

$$W(r, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta}) = \min_{\Delta} \left\{\eta^* \phi\left(\Delta - r\right) + (1-\eta^*)\phi\left(-\Delta - r\right)\right\} + \psi(r).$$

Let us write $\tilde{W}(r, \Delta; \eta) = \eta\phi(\Delta - r) + (1-\eta)\phi(-\Delta - r) + \psi(r)$. Then we have $W(r, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta}) = \min_{\Delta \geqslant 0} \tilde{W}(r, \Delta; \eta^*)$. We write $(r_\eta, \Delta_\eta) \in \operatorname{argmin}_{r \in \mathbb{R}, \Delta \geqslant 0} \tilde{W}(r, \Delta; \eta)$.

Next, we introduce some beneficial results about the first or second order derivative of ϕ, which are useful in deriving the calibration result for L_{ACS} with logistic ϕ.

Lemma C.10. *Whenever $\Delta_\eta > 0$, for any twice differentiable decreasing function ψ , we have*

$$\eta\phi'(\Delta_\eta - r_\eta) = (1-\eta)\phi'(-\Delta_\eta - r_\eta) = \frac{\psi'(r_\eta)}{2}. \tag{20}$$

Proof. This lemma is straightforward since $\nabla \tilde{W}(r_\eta, \Delta_\eta) = \mathbf{0}$ is rewritten as

$$\begin{aligned} -\eta\phi'(\Delta_\eta - r_\eta) - (1-\eta)\phi'(-\Delta_\eta - r_\eta) &= -\psi'(r_\eta), \\ \eta\phi'(\Delta_\eta - r_\eta) - (1-\eta)\phi'(-\Delta_\eta - r_\eta) &= 0. \end{aligned}$$

□

We use a subscript to denote a derivative of W. For example,

$$\begin{aligned} \tilde{W}^{\mathrm{ACS}}_r(r, \Delta; \eta) &= \frac{\partial \tilde{W}_{\mathrm{ACS}}(r, \Delta; \eta)}{\partial r}, \\ \tilde{W}^{\mathrm{ACS}}_\Delta(r, \Delta; \eta) &= \frac{\partial \tilde{W}_{\mathrm{ACS}}(r, \Delta; \eta)}{\partial \Delta}, \\ \tilde{W}^{\mathrm{ACS}}_\eta(r, \Delta; \eta) &= \frac{\partial \tilde{W}_{\mathrm{ACS}}(r, \Delta; \eta)}{\partial \eta}. \end{aligned}$$

We use ∇ to denote the gradient with respect to (r, Δ). Then, for example,

$$\nabla \tilde{W}_{\mathrm{ACS}}(r, \Delta; \eta) = \begin{pmatrix} W^{\mathrm{ACS}}_r(r, \Delta; \eta) \\ W^{\mathrm{ACS}}_\Delta(r, \Delta; \eta) \end{pmatrix}.$$

Lemma C.11. *Assume $c < 1/2$ and that ϕ is decreasing. Then, r_η is monotonically nondecreasing in $\eta^* > 0$ if*

$$\eta^2\phi''(\Delta_\eta - r_\eta) - (1-\eta)^2\phi''(-\Delta_\eta - r_\eta) \geq 0.$$

for all $\eta > \frac{1}{2}$.

Proof. By taking the derivative respect to Δ and let $\Delta = 0$, we have

$$\tilde{W}^{\mathrm{ACS}}_\Delta(r, 0; \eta) = (2\eta - 1)\phi'(-r). \tag{21}$$

First consider the case $\eta \leqslant 1/2$. In this case, from Eq. (21) we have $\tilde{W}^{\mathrm{ACS}}_\Delta(r, 0; \eta) \geqslant 0$. Therefore $\tilde{W}_{\mathrm{ACS}}(r, \Delta; \eta)$ is increasing at $\Delta = 0$ for any r. Recall that Δ_η is the minimizer over $\Delta \geqslant 0$. Then we have $\Delta_\eta = 0$. Then r_η is the solution of

$$\tilde{W}^{\mathrm{ACS}}_r(r_\eta, 0; \eta) = -\phi'(-r_\eta) + \psi'(r_\eta) = 0.$$

Therefore r_η does not depend on η, which means that r_η is nondecreasing.

Next we consider the case $\eta > 1/2$. In this case from Eq. (21) we have $\tilde{W}^{\mathrm{ACS}}_\Delta(r, 0; \eta) < 0$, which implies that $\Delta_\eta > 0$ and

$$\nabla \tilde{W}_{\mathrm{ACS}}(r_\eta, \Delta_\eta; \eta) = \mathbf{0}.$$

By taking the derivative with respect to η, we have

$$\nabla^2 \tilde{W}_{\mathrm{ACS}}\left(r_\eta, \Delta_\eta; \eta\right) \begin{pmatrix} r'_\eta \\ \Delta'_\eta \end{pmatrix} + \nabla \tilde{W}_\eta^{\mathrm{ACS}}\left(r_\eta, \Delta_\eta; \eta\right) = \mathbf{0}, \tag{22}$$

where

$$\begin{pmatrix} r'_\eta \\ \Delta'_\eta \end{pmatrix} = \frac{\partial}{\partial \eta} \begin{pmatrix} r_\eta \\ \Delta_\eta \end{pmatrix}.$$

Here note that

$$\left(\nabla^2 \tilde{W}_{\mathrm{ACS}}\left(r_\eta, \Delta_\eta; \eta\right)\right)^{-1} = \frac{1}{D_\eta} \begin{pmatrix} \tilde{W}_{\Delta\Delta}^{\mathrm{ACS}}\left(r_\eta, \Delta_\eta; \eta\right) & -\tilde{W}_{r\Delta}^{\mathrm{ACS}}\left(r_\eta, \Delta_\eta; \eta\right) \\ -\tilde{W}_{r\Delta}^{\mathrm{ACS}}\left(r_\eta, \Delta_\eta; \eta\right) & \tilde{W}_{rr}^{\mathrm{ACS}}\left(r_\eta, \Delta_\eta; \eta\right) \end{pmatrix}, \tag{23}$$

where $D_\eta = |\nabla^2 \tilde{W}(r_\eta, \Delta_\eta; \eta)|$, which is positive from the convexity of W. From Eq. (22) and Eq. (23) we have

$$r'_\eta = -\frac{1}{D_\eta} \begin{pmatrix} \tilde{W}_{\Delta\Delta}^{\mathrm{ACS}}\left(r_\eta, \Delta_\eta; \eta\right) & -\tilde{W}_{r\Delta}^{\mathrm{ACS}}\left(r_\eta, \Delta_\eta; \eta\right) \end{pmatrix} \nabla \tilde{W}_\eta^{\mathrm{ACS}}\left(r_\eta, \Delta_\eta; \eta\right). \tag{24}$$

Recall that $\tilde{W}_{\mathrm{ACS}}(r, \Delta; \eta) = \eta\phi(\Delta - r) + (1-\eta)\phi(-\Delta - r) + \psi(r)$. Now, for notational simplicity, we write $\phi'_+ = \phi'(\Delta_\eta - r_\eta), \phi'_- = \phi'(-\Delta_\eta - r_\eta)$ and $\psi = \psi(r_\eta)$. We also define ϕ''_+ and ϕ''_- in the same way. Then Eq. (24) is expressed as

$$r'_\eta = -\frac{1}{D_\eta} \begin{pmatrix} \eta\phi''_+ + (1-\eta)\phi''_- & \eta\phi''_+ - (1-\eta)\phi''_- \end{pmatrix} \begin{pmatrix} -\phi'_+ + \phi'_- \\ \phi'_+ + \phi'_- \end{pmatrix}.$$

Since $\phi'_+ = \psi'/(2\eta)$, $\phi'_- = \psi'/(2(1-\eta))$ from Lemma C.10, we have

$$\begin{aligned} r'_\eta &= -\frac{\psi'}{2\eta(1-\eta)D_\eta} \begin{pmatrix} \eta\phi''_+ + (1-\eta)\phi''_- & \eta\phi''_+ - (1-\eta)\phi''_- \end{pmatrix} \begin{pmatrix} 2\eta - 1 \\ 1 \end{pmatrix} \\ &= -\frac{\psi'}{2\eta(1-\eta)D_\eta} \left(2\eta\left(\eta\phi''_+ + (1-\eta)\phi''_-\right) - 2(1-\eta)\phi''_-\right) \\ &= -\frac{\psi'}{\eta(1-\eta)D_\eta} \left(\eta^2\phi''_+ - (1-\eta)^2\phi''_-\right). \end{aligned}$$

Since $\psi' = \psi'(r_\eta) \leqslant 0$ from the assumption, we see that if $\eta^2\phi''_+ - (1-\eta)^2\phi''_- \geqslant 0$ then $r'_\eta \geqslant 0$, that is, r_η is nondecreasing. □

Definition C.12. (Rademacher Complexity) Let sample $S = \{z_1, \cdots, z_n\}$ is drawn i.i.d. from a probability distribution $\mathcal{D}$, and $\mathcal{H} = \{h : \mathcal{Z} \to \mathbb{R}\}$ be a family of measurable function. Then the Rademacher complexity of $\mathcal{H}$ is defined as

$$\mathfrak{R}_n(\mathcal{H}) = \mathop{\mathbb{E}}_{S,\boldsymbol{\sigma}} \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \sigma_i h(z_i)\right],$$

where $\boldsymbol{\sigma} = (\sigma_1, \cdots, \sigma_n)$ is a random vector, in which variables are independent uniform random variables taking values in $\{-1, +1\}$.

Recall definition of $\mathcal{G}$ in Section 6, and we define $\widetilde{\mathcal{G}}$, $\phi \circ \widetilde{\mathcal{G}}$, $\mathcal{L}_{\mathrm{MCS}}$ as:

$$\begin{aligned} \widetilde{\mathcal{G}} &= \left\{\boldsymbol{x} \mapsto g_y(\boldsymbol{x}) - \max_{y' \neq y} g_{y'}(\boldsymbol{x}) \mid g_1, \cdots, g_k \in \mathcal{G}\right\}, \\ \phi \circ \widetilde{\mathcal{G}} &= \left\{\boldsymbol{x} \mapsto \phi(\tilde{g}(\boldsymbol{x})) \mid \tilde{g} \in \widetilde{\mathcal{G}}\right\}, \\ \mathcal{L}_{\mathrm{MCS}} &= \left\{(\boldsymbol{x}, y) \mapsto L_{\mathrm{MCS}}(r, \boldsymbol{g}; \boldsymbol{x}, y) \mid g_1, \cdots, g_k \in \mathcal{G}\right\}, \qquad (25) \\ \mathcal{L}_{\mathrm{ACS}} &= \left\{(\boldsymbol{x}, y) \mapsto L_{\mathrm{ACS}}(r, \boldsymbol{g}; \boldsymbol{x}, y) \mid g_1, \cdots, g_k \in \mathcal{G}\right\}. \qquad (26) \end{aligned}$$

Next, we introduce a result about Rademacher complexity of the hypothesis class of functions that are the pointwise maximum of other functions.

Lemma C.13. *(Mohri et al. (2018), Lemma 9.1) Let $\mathcal{H}_1, \cdots, \mathcal{H}_K$ be K hypothesis classes ($K \geqslant 1$) mapping $\mathcal{X}$ to $\mathbb{R}$, and let $\mathcal{H} = \{\max\{h_1, \cdots, h_K\} \mid h_i \in \mathcal{H}_i, i \in [K]\}$. Then, for any sample S of size n, the Rademacher complexity of $\mathcal{H}$ can be upper bounded as follows:*

$$\mathfrak{R}_n(\mathcal{H}) \leqslant \sum_{j=1}^{K} \mathfrak{R}_n(\mathcal{H}_j).$$

Next, we introduce a result about Rademacher complexity of the hypothesis class of functions that are inner products of other functions.

Lemma C.14. *Let $\mathcal{H}_1 = \{h_1 : \mathcal{X} \to \mathbb{R}\}$, $\mathcal{H}_2 = \{h_2 : \mathcal{X} \to \mathbb{R}\}$ be two hypothesis classes of functions mapping $\mathcal{X}$ to $[0, M_1]$ and $[0, M_2]$ respectively. Let $\mathcal{H} = \{h_2 h_2 \mid h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2\}$. Then, for any sample S of size n, the Rademacher complexity of $\mathcal{H}$ can be upper bounded as follows:*

$$\mathfrak{R}_n(\mathcal{H}) \leqslant (M_1 + \frac{1}{2}M_2)\left(\mathfrak{R}_n(\mathcal{H}_1) + \mathfrak{R}_n(\mathcal{H}_2)\right).$$

Proof. For any $h_1 \in \mathcal{H}_1$ and $h_2 \in \mathcal{H}_2$, we can write $h_1 h_2 = \frac{1}{4}\left[(h_1 + h_2)^2 - (h_1 - h_2^2)\right]$. For the term $(h_1 + h_2)^2$, note that the function $x \mapsto \frac{1}{4}x^2$ is Lipschitz-continous with constant $\frac{1}{2}(M_1 + M_2)$ over $[0, M_1 + M_2]$. For the term $(h_1 - h_2)^2$, observe that the function $x \mapsto \frac{1}{4}x^2$ is Lipschitz-continous with consant $\frac{1}{2}M_1$ over $[0, M_1]$. Thus, the following holds:

$$\begin{aligned}
\mathfrak{R}_n(\mathcal{H}) &\leqslant \frac{1}{2}(M_1 + M_2)\mathfrak{R}_n(\mathcal{H}_1 + \mathcal{H}_2) + \frac{1}{2}M_1\mathfrak{R}_n(\mathcal{H}_1 - \mathcal{H}_2) \\
&\qquad\qquad \text{(Talagrand's contraction lemma (Mohri et al., 2018))} \\
&\leqslant \frac{1}{2}(M_1 + M_2)\mathfrak{R}_n(\mathcal{H}_1) + \frac{1}{2}(M_1 + M_2)\mathfrak{R}_n(\mathcal{H}_2) + \frac{1}{2}M_1\mathfrak{R}_n(\mathcal{H}_1) + \frac{1}{2}M_1\mathfrak{R}_n(-\mathcal{H}_2) \\
&\qquad\qquad \text{(sub-additivity of } \sup(\cdot)) \\
&= (M_1 + \frac{1}{2}M_2)\left(\mathfrak{R}_n(\mathcal{H}_1) + \mathfrak{R}_n(\mathcal{H}_2)\right).
\end{aligned}$$

□

A similar lemma can be found in DeSalvo et al. (2015); the main difference is the co-domain of hypotheses.

Lemma C.15. *Let $\mathfrak{R}_n(\mathcal{L}_{\text{MCS}})$ be the Rademacher complexity of $\mathcal{L}_{\text{MCS}}$ for $S = \{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n)\}$ from $p(\boldsymbol{x}, y)$, and $\mathfrak{R}_n(\mathcal{G}), \mathfrak{R}_n(\mathcal{R})$ be the Rademacher complexity for data of size n drawn from $p(\boldsymbol{x})$. Then we have $\mathfrak{R}_n(\mathcal{L}_{\text{MCS}}) \leqslant 2M_\phi L_\phi K^2 \mathfrak{R}_n(\mathcal{G}) + (2M_{\psi_1}L_{\psi_1} + L_{\psi_2})\,\mathfrak{R}_n(\mathcal{R})$.*

Proof. By definition, we can bound $\mathfrak{R}_n(\widetilde{\mathcal{G}})$ as follows:

$$
\begin{aligned}
\mathfrak{R}_n(\widetilde{\mathcal{G}}) &= \mathop{\mathbb{E}}_{S,\boldsymbol{\sigma}}\left[\sup_{g_1\cdots,g_K\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\left(g_{y_i}(\boldsymbol{x_i})-\max_{y'\neq y_i}g_{y'_i}(\boldsymbol{x})\right)\right]\\
&\leqslant \mathop{\mathbb{E}}_{S,\boldsymbol{\sigma}}\left[\sup_{g_1\cdots,g_K\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i g_{y_i}(\boldsymbol{x_i})\right]+\mathop{\mathbb{E}}_{S,\boldsymbol{\sigma}}\left[\sup_{g_1\cdots,g_K\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}-\sigma_i\max_{y'\neq y_i}g_{y'_i}(\boldsymbol{x})\right] \qquad \text{(sub-additivity of } \sup(\cdot))\\
&\leqslant \mathop{\mathbb{E}}_{S,\boldsymbol{\sigma}}\left[\sup_{g_1\cdots,g_K\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i g_{y_i}(\boldsymbol{x_i})\right]+\mathop{\mathbb{E}}_{S,\boldsymbol{\sigma}}\left[\sup_{g_1\cdots,g_K\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\max_{y'\neq y_i}g_{y'_i}(\boldsymbol{x})\right]\\
&\leqslant \mathop{\mathbb{E}}_{S,\boldsymbol{\sigma}}\left[\sup_{g_1\cdots,g_K\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\max_{y\in\mathcal{Y}}g_{y}(\boldsymbol{x_i})\right]+\mathop{\mathbb{E}}_{S,\boldsymbol{\sigma}}\left[\sup_{g_1\cdots,g_K\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\max_{y'\neq y_i}g_{y'_i}(\boldsymbol{x})\right]\\
&\leqslant K\mathfrak{R}_n(\mathcal{G})+\mathop{\mathbb{E}}_{S,\boldsymbol{\sigma}}\left[\sup_{g_1\cdots,g_K\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\max_{y'\neq y_i}g_{y'_i}(\boldsymbol{x})\right] \qquad \text{(result by Lemma C.13)}\\
&\leqslant K\mathfrak{R}_n(\mathcal{G})+\sum_{y\in\mathcal{Y}}\mathop{\mathbb{E}}_{S,\boldsymbol{\sigma}}\left[\sup_{g_1\cdots,g_K\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\max_{y'\neq y}g_{y'}(\boldsymbol{x})\right]\\
&\leqslant K\mathfrak{R}_n(\mathcal{G})+K(K-1)\mathfrak{R}_n(\mathcal{G}) \qquad \text{(result by Lemma C.13)}\\
&\leqslant K^2\mathfrak{R}_n(\mathcal{G}).
\end{aligned}
$$

Then, we can bound $\mathfrak{R}_n(\mathcal{L}_{\mathrm{MCS}})$ as follows:

$$
\begin{aligned}
\mathfrak{R}_n(\mathcal{L}_{\mathrm{MCS}}) &= \mathop{\mathbb{E}}_{S,\boldsymbol{\sigma}}\left[\sup_{L\in\mathcal{L}_{\mathrm{MCS}}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i L_{\mathrm{MCS}}(r,\boldsymbol{g};\boldsymbol{x}_i,y_i)\right]\\
&= \mathop{\mathbb{E}}_{S,\boldsymbol{\sigma}}\left[\sup_{r\in\mathcal{R},g_1\cdots,g_K\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\left(\phi\left(g_{y_i}(\boldsymbol{x}_i)-\max_{y'\neq y_i}g_{y'}(\boldsymbol{x}_i)\right)\psi_1(r(\boldsymbol{x}_i))+\psi_2(r(\boldsymbol{x}_i))\right)\right]\\
&\leqslant \mathop{\mathbb{E}}_{S,\boldsymbol{\sigma}}\left[\sup_{r\in\mathcal{R},g_1\cdots,g_K\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\phi\left(g_{y_i}(\boldsymbol{x}_i)-\max_{y'\neq y_i}g_{y'}(\boldsymbol{x}_i)\right)\psi_1(r(\boldsymbol{x}_i))\right]\\
&\quad+\mathop{\mathbb{E}}_{S,\boldsymbol{\sigma}}\left[\sup_{r\in\mathcal{R}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\psi_2(r(\boldsymbol{x}_i))\right] \qquad \text{(sub-additivity of } \sup(\cdot))\\
&= \mathfrak{R}_n((\phi\circ\widetilde{\mathcal{G}})(\psi_1\circ\mathcal{R}))+\mathfrak{R}_n(\psi_2\circ\mathcal{R})\\
&\leqslant \left(M_\phi+\frac{1}{2}M_{\psi_1}\right)\left(\mathfrak{R}_n(\phi\circ\widetilde{\mathcal{G}})+\mathfrak{R}_n(\psi_1\circ\mathcal{R})\right)+\mathfrak{R}_n(\psi_2\circ\mathcal{R}) \qquad \text{(result from Lemma C.14)}\\
&\leqslant \left(M_\phi+\frac{1}{2}M_{\psi_1}\right)L_\phi\mathfrak{R}_n(\widetilde{\mathcal{G}})+\left(M_\phi+\frac{1}{2}M_{\psi_1}\right)L_{\psi_1}\mathfrak{R}_n(\mathcal{R})+L_{\psi_2}\mathfrak{R}_n(\mathcal{R})\\
&\qquad\qquad \text{(Talagrand's contraction lemma)}\\
&\leqslant \left(M_\phi+\frac{1}{2}M_{\psi_1}\right)L_\phi K^2\mathfrak{R}_n(\mathcal{G})+\left(\left(M_\phi+\frac{1}{2}M_{\psi_1}\right)L_{\psi_1}+L_{\psi_2}\right)\mathfrak{R}_n(\mathcal{R}). \qquad \text{(result from (27))}
\end{aligned}
$$

□

Next, we introduce a useful inequity that bound Rademacher complexity of $\mathcal{L}_{\mathrm{ACS}}$ with that of $\mathcal{G}$ and $\mathcal{R}$.

Lemma C.16. *Let $\mathfrak{R}_n(\mathcal{L}_{\mathrm{ACS}})$ be the Rademacher complexity of $\mathcal{L}_{\mathrm{ACS}}$ for $S=\{(\boldsymbol{x}_1,y_1),\cdots,(\boldsymbol{x}_n,y_n)\}$ from $p(\boldsymbol{x},y)$, and $\mathfrak{R}_n(\mathcal{G}),\mathfrak{R}_n(\mathcal{R})$ be the Rademacher complexity for data of size n drawn from $p(\boldsymbol{x})$. Then we have $\mathfrak{R}_n(\mathcal{L}_{\mathrm{ACS}})\leqslant L_\phi K^2\mathfrak{R}_n(\mathcal{G})+(L_\phi+L_\psi)\mathfrak{R}_n(\mathcal{R})$.*

Proof. By definition, we can bound $\mathfrak{R}_n(\widetilde{\mathcal{G}})$ as follows:

$$\begin{aligned}
\mathfrak{R}_n(\widetilde{\mathcal{G}}) &= \mathop{\mathbb{E}}_{S,\boldsymbol{\sigma}}\left[\sup_{g_1\cdots,g_K\in\mathcal{G},r\in\mathcal{R}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\left(g_{y_i}(\boldsymbol{x_i})-\max_{y'\neq y_i}g_{y'_i}(\boldsymbol{x})-r(\boldsymbol{x}_i)\right)\right]\\
&\leqslant \mathop{\mathbb{E}}_{S,\boldsymbol{\sigma}}\left[\sup_{g_1\cdots,g_K\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i g_{y_i}(\boldsymbol{x_i})\right]+\mathop{\mathbb{E}}_{S,\boldsymbol{\sigma}}\left[\sup_{g_1\cdots,g_K\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}-\sigma_i\max_{y'\neq y_i}g_{y'_i}(\boldsymbol{x}_i)\right]\\
&\quad+\mathop{\mathbb{E}}_{S,\boldsymbol{\sigma}}\left[\sup_{r\in\mathcal{R}}\frac{1}{n}\sum_{i=1}^{n}-\sigma_i r(\boldsymbol{x}_i)\right] && \text{(sub-additivity of } \sup(\cdot))\\
&= \mathop{\mathbb{E}}_{S,\boldsymbol{\sigma}}\left[\sup_{g_1\cdots,g_K\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i g_{y_i}(\boldsymbol{x_i})\right]+\mathop{\mathbb{E}}_{S,\boldsymbol{\sigma}}\left[\sup_{g_1\cdots,g_K\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\max_{y'\neq y_i}g_{y'_i}(\boldsymbol{x}_i)\right]\\
&\quad+\mathop{\mathbb{E}}_{S,\boldsymbol{\sigma}}\left[\sup_{r\in\mathcal{R}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i r(\boldsymbol{x}_i)\right]\\
&\leqslant \mathop{\mathbb{E}}_{S,\boldsymbol{\sigma}}\left[\sup_{g_1\cdots,g_K\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\max_{y\in\mathcal{Y}}g_y(\boldsymbol{x_i})\right]+\mathop{\mathbb{E}}_{S,\boldsymbol{\sigma}}\left[\sup_{g_1\cdots,g_K\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\max_{y'\neq y_i}g_{y'_i}(\boldsymbol{x}_i)\right]\\
&\quad+\mathop{\mathbb{E}}_{S,\boldsymbol{\sigma}}\left[\sup_{r\in\mathcal{R}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i r(\boldsymbol{x}_i)\right]\\
&\leqslant K\mathfrak{R}_n(\mathcal{G})+\sum_{y\in\mathcal{Y}}\mathop{\mathbb{E}}_{S,\boldsymbol{\sigma}}\left[\sup_{g_1\cdots,g_K\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\max_{y'\neq y}g_{y'}(\boldsymbol{x}_i)\right]+\mathfrak{R}_n(\mathcal{R}) && \text{(result by Lemma C.13)}\\
&\leqslant K\mathfrak{R}_n(\mathcal{G})+K(K-1)\mathfrak{R}_n(\mathcal{G})+\mathfrak{R}_n(\mathcal{R}) && \text{(result by Lemma C.13)}\\
&\leqslant K^2\mathfrak{R}_n(\mathcal{G})+\mathfrak{R}_n(\mathcal{R}). && (27)
\end{aligned}$$

Then, we can bound $\mathfrak{R}_n(\mathcal{L}_{\text{ACS}})$ as follows:

$$\begin{aligned}
\mathfrak{R}_n(\mathcal{L}_{\text{ACS}}) &= \mathop{\mathbb{E}}_{S,\boldsymbol{\sigma}}\left[\sup_{L\in\mathcal{L}_{\text{MCS}}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i L_{\text{ACS}}(r,\boldsymbol{g};\boldsymbol{x}_i,y_i)\right]\\
&= \mathop{\mathbb{E}}_{S,\boldsymbol{\sigma}}\left[\sup_{r\in\mathcal{R},g_1\cdots,g_K\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\left(\phi\left(g_{y_i}(\boldsymbol{x}_i)-\max_{y'\neq y_i}g_{y'}(\boldsymbol{x}_i)-r(\boldsymbol{x}_i)\right)+\psi(r(\boldsymbol{x}_i))\right)\right]\\
&\leqslant \mathop{\mathbb{E}}_{S,\boldsymbol{\sigma}}\left[\sup_{r\in\mathcal{R},g_1\cdots,g_K\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\phi\left(g_{y_i}(\boldsymbol{x}_i)-\max_{y'\neq y_i}g_{y'}(\boldsymbol{x}_i)-r(\boldsymbol{x}_i)\right)\right]\\
&\quad+\mathop{\mathbb{E}}_{S,\boldsymbol{\sigma}}\left[\sup_{r\in\mathcal{R}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\psi(r(\boldsymbol{x}_i))\right] && \text{(sub-additivity of } \sup(\cdot))\\
&= \mathfrak{R}_n(\phi\circ\widetilde{\mathcal{G}})+\mathfrak{R}_n(\psi\circ\mathcal{R})\\
&\leqslant L_\phi\mathfrak{R}_n(\widetilde{\mathcal{G}})+L_\psi\mathfrak{R}_n(\mathcal{R}) && \text{(Talagrand's contraction lemma)}\\
&\leqslant L_\phi K^2\mathfrak{R}_n(\mathcal{G})+(L_\phi+L_\psi)\,\mathfrak{R}_n(\mathcal{R}). && \text{(result from (27))}
\end{aligned}$$

□

The following Lemma C.17 gives the infimum of $W_{0\text{-}1\text{-}c}$ and the gap between $W_{0\text{-}1\text{-}c}$ and its infimum.

Lemma C.17. *For $W_{0\text{-}1\text{-}c}$ defined in Eq. (2), we have:*

$$W_{0\text{-}1\text{-}c}(r^*_{\boldsymbol{\eta}},\boldsymbol{g}^*_{\boldsymbol{\eta}};\boldsymbol{\eta})=1-\max\{\eta^*,1-c\},$$

and

$$\begin{aligned}\Delta W_{0\text{-}1\text{-}c}(r,\boldsymbol{g};\boldsymbol{\eta}) &= W_{0\text{-}1\text{-}c}(r,\boldsymbol{g};\boldsymbol{\eta}) - W_{0\text{-}1\text{-}c}(r^*_{\boldsymbol{\eta}},\boldsymbol{g}^*_{\boldsymbol{\eta}};\boldsymbol{\eta}),\\ &= \begin{cases}\eta^* - \eta_{\tilde{y}} & \text{if } \eta^* > (1-c), \quad r>0,\\ 1-c-\eta_{\tilde{y}} & \text{if } \eta^* \leqslant (1-c), \quad r>0,\\ 0 & \text{if } \eta^* \leqslant (1-c), \quad r\leqslant 0,\\ \eta^* - 1 + c & \text{if } \eta^* > (1-c), \quad r\leqslant 0.\end{cases}\end{aligned}$$

Proof. First, let us consider expressions of $W_{0\text{-}1\text{-}c}(r,\boldsymbol{g};\boldsymbol{\eta})$ and $W_{0\text{-}1\text{-}c}(r^*_{\boldsymbol{\eta}},\boldsymbol{g}^*_{\boldsymbol{\eta}};\boldsymbol{\eta})$.
According to the definition, we have

$$\begin{aligned}W_{0\text{-}1\text{-}c}(r,\boldsymbol{g};\boldsymbol{\eta}) &= \sum_y \eta_y L_{0\text{-}1\text{-}c}(r,\boldsymbol{g};y)\\ &= \sum_y \eta_y \mathbb{I}_{\tilde{y}\neq y}\mathbb{I}_{r>0} + c\mathbb{I}_{r\leqslant 0},\end{aligned}$$

for $\tilde{y}\in \operatorname{argmax}_y g_y$. For $W_{0\text{-}1\text{-}c}(r,\boldsymbol{g};\boldsymbol{\eta})$, $W_{0\text{-}1\text{-}c}(r,\boldsymbol{g};\boldsymbol{\eta}) = \sum_y \eta_y \mathbb{I}_{\tilde{y}\neq y} = 1-\eta_{\tilde{y}}$ if $r>0$, $W_{0\text{-}1\text{-}c}(r,\boldsymbol{g};\boldsymbol{\eta}) = c$ if $r\leqslant 0$.
For $W_{0\text{-}1\text{-}c}(r^*_{\boldsymbol{\eta}},\boldsymbol{g}^*_{\boldsymbol{\eta}};\boldsymbol{\eta})$, let $\tilde{Y} = \left\{\operatorname{argmax}_y g_y \mid \boldsymbol{g}\in\mathfrak{G}\right\}$, we have

$$\begin{aligned}W_{0\text{-}1\text{-}c}(r^*_{\boldsymbol{\eta}},\boldsymbol{g}^*_{\boldsymbol{\eta}};\boldsymbol{\eta}) &= \min\left\{\min_{y\in\tilde{Y}}(1-\eta_y), c\right\}\\ &= 1-\max\left\{\max_{y\in\tilde{Y}}\eta_y, 1-c\right\}\\ &= 1-\max\left\{\eta^*, 1-c\right\}.\end{aligned}$$

Then, $W_{0\text{-}1\text{-}c}(r^*_{\boldsymbol{\eta}},\boldsymbol{g}^*_{\boldsymbol{\eta}};\boldsymbol{\eta}) = 1-\eta^*$ if $\eta^* > 1-c$, and $W_{0\text{-}1\text{-}c}(r^*_{\boldsymbol{\eta}},\boldsymbol{g}^*_{\boldsymbol{\eta}};\boldsymbol{\eta}) = c$ if $\eta^*\leqslant 1-c$.
Therefore, we have

$$\begin{aligned}\Delta W_{0\text{-}1\text{-}c}(r,\boldsymbol{g};\boldsymbol{\eta}) &= W_{0\text{-}1\text{-}c}(r,\boldsymbol{g};\boldsymbol{\eta}) - W_{0\text{-}1\text{-}c}(r^*_{\boldsymbol{\eta}},\boldsymbol{g}^*_{\boldsymbol{\eta}};\boldsymbol{\eta}),\\ &= \begin{cases}\eta^* - \eta_{\tilde{y}} & \text{if } \eta^* > (1-c), \quad r>0,\\ 1-c-\eta_{\tilde{y}} & \text{if } \eta^* \leqslant (1-c), \quad r>0,\\ 0 & \text{if } \eta^* \leqslant (1-c), \quad r\leqslant 0,\\ \eta^* - 1 + c & \text{if } \eta^* > (1-c), \quad r\leqslant 0.\end{cases}\end{aligned}$$

□

Lemma C.18. *For non-increasing function $\phi(x)\geqslant 0$ and $\Delta\geqslant 0$, we have*

$$\sum_y \eta_y \phi\left(g_y - \max_{y'\neq y} g_{y'}\right) \geqslant \eta_{\tilde{y}}\phi(\Delta) + (1-\eta_{\tilde{y}})\phi(-\Delta),$$

where $\tilde{y}\in\operatorname{argmax}_y g_y$ and $0\leqslant\eta_{\tilde{y}}\leqslant 1$.

Proof.

$$\begin{aligned}
&\sum_y \eta_y \phi\left(g_y - \max_{y' \neq y} g_{y'}\right) \\
&\geqslant \eta_{\tilde{y}} \phi\left(g_{(1)} - g_{(2)}\right) + \sum_{y \neq \tilde{y}} \eta_y \phi\left(g_{(2)} - g_{(1)}\right) \\
&= \eta_{\tilde{y}} \phi\left(\underbrace{g_{(1)} - g_{(2)}}_{\Delta \geqslant 0}\right) + (1 - \eta_{\tilde{y}}) \phi\left(\underbrace{g_{(2)} - g_{(1)}}_{-\Delta \leqslant 0}\right) \\
&= \eta_{\tilde{y}} \phi(\Delta) + (1 - \eta_{\tilde{y}}) \phi(-\Delta),
\end{aligned}$$

where $g^{(1)} = \max_y g_y$, $\tilde{y} \in \operatorname{argmax}_y g_y$, $g^{(2)} = \max_{y \neq \tilde{y}} g_y$, and $0 \leqslant \eta_{\tilde{y}} \leqslant 1$. □

D Proofs of Theorems and Corollaries

D.1 Proof of Theorem 4.1

Theorem 4.1 (restated) *$L_{\mathrm{MCS}}(r, \boldsymbol{g}; \boldsymbol{x}, y)$ is calibrated if and only if*

$$H_{1-c} \psi_1'(0) + \psi_2'(0) = 0,$$

where $H_\eta = \min_{\Delta \geqslant 0} \{\eta \phi(\Delta) + (1 - \eta) \phi(-\Delta)\} \geq 0$ for $0 \leqslant \eta \leqslant 1$.

Proof. Recall that the multiplicative CS (MCS) loss is of the following form:

$$L_{\mathrm{MCS}}(r, \boldsymbol{g}; \boldsymbol{x}, y) = \phi\left(g_y - \max_{y' \neq y} g_{y'}\right) \psi_1(r) + \psi_2(r), \tag{28}$$

where $\phi(\Delta) \geqslant 0$ is a nonincreasing function, and $\psi_1(r)$, $\psi_2(r)$ are convex function differentiable at $r = 0$.

From the expression of Eq. (28), the minimization of $W_{\mathrm{MCS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) = \sum_y \eta_y L_{\mathrm{MCS}}(r, \boldsymbol{g}; \boldsymbol{x}, y)$ with respect to $\boldsymbol{g}$ does not depend on r. Therefore, by Proposition B.2 with $\Phi(\Delta, r) = \phi(\Delta)\psi_1(r) + \psi_2(r)$, we have

$$W_{\mathrm{MCS}}\left(r, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta}\right) = \min_{\Delta \geqslant 0} \{\eta \Phi(\Delta, r) + (1 - \eta) \Phi(-\Delta, r)\} = H_{\eta^*} \psi_1(r) + \psi_2(r),$$

where

$$H_\eta = \min_{\Delta \geqslant 0} \{\eta \phi(\Delta) + (1 - \eta) \phi(-\Delta)\} \geqslant 0.$$

By taking the derivative and let $r = 0$, we obtain

$$\left.\frac{\partial W_{\mathrm{MCS}}(r, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta})}{\partial r}\right|_{r=0} = H_{\eta^*} \psi_1'(0) + \psi_2'(0).$$

Since ϕ is a nonincreasing function, we can see that H_{η^*} is nonincreasing in η^*, and thus

$$\inf_{\boldsymbol{\eta}: \eta^* \leqslant 1-c} \left.\frac{\partial W_{\mathrm{MCS}}\left(r, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta}\right)}{\partial r}\right|_{r=0} = \sup_{\boldsymbol{\eta}: \eta^* \geqslant 1-c} \left.\frac{\partial W_{\mathrm{MCS}}\left(r, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta}\right)}{\partial r}\right|_{r=0} = H_{1-c} \psi_1'(0) + \psi_2'(0) = 0.$$

From Theorem 4 in Ni et al. (2019), we obtain that it holds if and only if $L_{\mathrm{MCS}}(r, \boldsymbol{g}; \boldsymbol{x}, y)$ is calibrated. □

Corollary D.1. *For $\phi(\Delta) = \exp(-\Delta)$, $L_{\mathrm{MCS}}(r, \boldsymbol{g}; \boldsymbol{x}, y)$ is calibrated if and only if*

$$2\sqrt{c(1-c)}\psi_1'(0) + \psi_2'(0) = 0.$$

For $\phi(\Delta) = \max\{1-\Delta, 0\}$, $L_{\mathrm{MCS}}(r, \boldsymbol{g}; \boldsymbol{x}, y)$ is calibrated if and only if

$$2c\psi_1'(0) + \psi_2'(0) = 0.$$

For $\phi(\Delta) = \log(1+\exp(-\Delta))$, $L_{\mathrm{MCS}}(r, \boldsymbol{g}; \boldsymbol{x}, y)$ is calibrated if and only if

$$\left(-(1-c)\log(1-c) - c\log c\right)\psi_1'(0) + \psi_2'(0) = 0.$$

Proof. Then, we analyze three cases to derive calibration results of L_{MCS}.

(**1**) Case $\phi(\Delta) = \exp(-\Delta)$:

In this case, for $\eta^* = 1-c$ and $0 < c < \frac{1}{2}$, using Lemma C.1, we have

$$H_{1-c} = 2\sqrt{c(1-c)}.$$

Then from Theorem 4.1, we have the loss is calibrated if and only if

$$2\sqrt{c(1-c)}\psi_1'(0) + \psi_2'(0) = 0.$$

(**2**) Case $\phi(\Delta) = \max\{0, 1-\Delta\}$:

In this case, for $\eta^* = 1-c$ and $0 < c < \frac{1}{2}$, using Lemma C.1, we have

$$H_{1-c} = 2c.$$

Then from Theorem 4.1, we have the loss is calibrated if and only if

$$2c\psi_1'(0) + \psi_2'(0) = 0.$$

(**3**) Case $\phi(\Delta) = \log(1+\exp(-\Delta))$:

In this case, for $\eta^* = 1-c$ and $0 < c < \frac{1}{2}$, using Lemma C.1, we have

$$H_{1-c} = -(1-c)\log(1-c) - c\log c.$$

Then from Theorem 4.1, we have the loss is calibrated if and only if

$$\left(-(1-c)\log(1-c) - c\log c\right)\psi_1'(0) + \psi_2'(0) = 0.$$

□

D.2 Proof of Theorem 4.2

Theorem 4.2 (restated) *$L_{\mathrm{ACS}}(r, \boldsymbol{g}; \boldsymbol{x}, y)$ with $\phi(\Delta) = \max\{1-\Delta, 0\}$ is calibrated if and only if*

$$\psi'(0) = -2c.$$

Proof. Recall that the additive CS (ACS) loss with hinge ϕ is of the following form:

$$L_{\mathrm{ACS}}(r, \boldsymbol{g}; \boldsymbol{x}, y) = \max\left\{0, 1 - \left(g_y - \max_{y' \neq y} g_{y'} - r\right)\right\} + \psi\left(r\right),$$

where $\psi(r)$ is convex function differentiable at $r = 0$.

From (12) we have

$$\left.\frac{\partial W_{\mathrm{ACS}}(r, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta})}{\partial r}\right|_{r=0} = \begin{cases} 2(1-\eta^*) + \psi'(0) & \text{if } \eta^* > \frac{1}{2}, \\ 1 + \psi'(0) & \text{if } \eta^* \leqslant \frac{1}{2}. \end{cases}$$

We can see that $\left.\frac{\partial W_{\mathrm{ACS}}(r, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta})}{\partial r}\right|_{r=0}$ is nonincreasing in η^*, and thus

$$\inf_{\boldsymbol{\eta}: \eta^* \leqslant 1-c} \left.\frac{\partial W\left(r, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta}\right)}{\partial r}\right|_{r=0} = \sup_{\boldsymbol{\eta}: \eta^* \geqslant 1-c} \left.\frac{\partial W\left(r, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta}\right)}{\partial r}\right|_{r=0} = 2c + \psi'(0) = 0.$$

Then we have $\psi'(0) = -2c$. □

D.3 Proof of Theorem 4.3

Theorem 4.3 (restated) *Assume that ϕ is a twice differentiable decreasing function. Then L_{ACS} with $\phi(\Delta) = \log\left(1 + \exp(-\Delta)\right)$ is calibrated if and only if*

$$\psi'(0) = -2c(1-c).$$

Proof. Recall that $\tilde{W}_{\mathrm{ACS}}(r, \Delta; \eta) = \eta\phi(\Delta - r) + (1-\eta)\phi(-\Delta - r) + \psi(r)$ and $(r_\eta, \Delta_\eta) = \operatorname{argmin}_{r \in \mathbb{R}, \Delta \geqslant 0} \tilde{W}_{\mathrm{ACS}}(r, \Delta; \eta)$.

For $\phi(x) = \log\left(1 + \exp(-x)\right)$, we have

$$\begin{aligned} \phi'(x) &= \frac{-\exp(-x)}{1 + \exp(-x)} \\ &= -\frac{1}{1 + \exp(x)}, \\ \phi''(x) &= \frac{\exp(x)}{\left(1 + \exp(x)\right)^2} \\ &= -\left(\phi'(x) + \left(\phi'(x)\right)^2\right). \end{aligned}$$

From Lemma C.11, if we can show that $\eta^2\phi''\left(\Delta_\eta - r_\eta\right) - (1-\eta)^2\phi''\left(-\Delta_\eta - r_\eta\right) \geqslant 0$ for all $\eta > \frac{1}{2}$, then r_η is monotonically nondecreasing in $\eta > 0$. Under the same notation as the proof of Lemma Lemma C.11, we write $\phi'_+ = \phi'(\Delta_\eta - r_\eta), \phi'_- = \phi'(-\Delta_\eta - r_\eta)$ and $\psi = \psi(r_\eta)$. Therefore, for $\eta > 1/2$ we have

$$\begin{aligned} \eta^2\phi''_+ - (1-\eta)^2\phi''_- &= \left((1-\eta)^2\phi'_- - (\eta)^2\phi'_+\right) + \left((1-\eta)^2\left(\phi'_-\right)^2 - \eta^2\left(\phi'_+\right)^2\right) \\ &= \frac{\psi'}{2}(1-2\eta) \qquad (29) \\ &\geqslant 0, \end{aligned}$$

where Eq. (29) follows from Lemma C.10 and the inequality holds since ψ is decreasing.

Now we can apply Lemma C.11 and we see that r_η is nondecreasing. Therefore, to prove that the loss is rejection calibrated, it suffices to show $r_{1-c} = 0$, sothat we have $r_\eta \geqslant 0$ for $\eta \geqslant 1-c$ and $r_\eta < 0$ for $\eta < 1-c$, which means

$\operatorname{sign}(r^*_{\boldsymbol{\eta}}) = \operatorname{sign}(\eta^* - (1-c))$. Recall that from Lemma C.10, we have $\eta\phi'(\Delta_\eta - r_\eta) = (1-\eta)\phi'(-\Delta_\eta - r_\eta) = \frac{\psi'(r_\eta)}{2}$. For the logistic loss with $\eta = 1-c$, this is equivalent to

$$-\frac{1-c}{1+\exp(\Delta_\eta - r_\eta)} = -\frac{c}{1+\exp(-\Delta_\eta - r_\eta)} = \frac{\psi'(r_\eta)}{2},$$

which is satisfied for $(r_\eta, \Delta_\eta) = (0, \log\frac{1-c}{c})$ if $\psi'(0) = -2c(1-c)$.

□

D.4 Proof of Theorem D.4

Theorem D.4 (restated) *For L_{MCS}, the following excess risk bound holds for all $r, \boldsymbol{g}$ and any distribution $p(\boldsymbol{x}, y)$:*

$$R_{\text{0-1-}c}(r, \boldsymbol{g}) - R^*_{\text{0-1-}c} \leqslant \Gamma\left(R(r, \boldsymbol{g}) - R^*\right),$$

where

$$\Gamma(t) = O\left(\max\left\{t, \sqrt{t}\right\}\right)$$

for $R = R_{\mathrm{MCS}}$ with $\phi(\Delta) = \exp(-\Delta)$, $\psi_1(r) = \exp(r)$, and $\psi_2(r) = c\exp(-\alpha r)$, and

$$\Gamma(t) = O(t)$$

for $R = R_{\mathrm{ACS}}$ with $\phi(\Delta) = \max\{1-\Delta, 0\}$, $\psi(r) = c\max\{1-\alpha r, 0\}$.

Proof. Firstly, we proof the following excess risk bound holds for L_{MCS}:

$$R_{\text{0-1-}c}(r, \boldsymbol{g}) - R^*_{\text{0-1-}c} \leqslant \Gamma\left(R_{\mathrm{MCS}}(r, \boldsymbol{g}) - R^*_{\mathrm{MCS}}\right),$$

for all $r, \boldsymbol{g}$, where

$$\Gamma(t) = O\left(\max\left\{t, \sqrt{t}\right\}\right)$$

for $\phi(r) = \exp(-r)$, $\psi_1(r) = \exp(r)$, and $\psi_2(r) = c\exp(-\alpha r)$.

First, from Lemma C.17, we have:

$$\begin{aligned}
\Delta W_{\text{0-1-}c}(r, \boldsymbol{g}; \boldsymbol{\eta}) &= W_{\text{0-1-}c}(r, \boldsymbol{g}; \boldsymbol{\eta}) - W_{\text{0-1-}c}(r^*_{\boldsymbol{\eta}}, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta}), \\
&= \begin{cases} \eta^* - \eta_{\tilde{y}} & \text{if } \eta^* > (1-c), \quad r(x) > 0, \\ 1 - c - \eta_{\tilde{y}} & \text{if } \eta^* \leqslant (1-c), \quad r(x) > 0, \\ 0 & \text{if } \eta^* \leqslant (1-c), \quad r(x) \leqslant 0, \\ \eta^* - 1 + c & \text{if } \eta^* > (1-c), \quad r(x) \leqslant 0. \end{cases}
\end{aligned} \tag{30}$$

In the following, we introduce some useful results for bounding $\Delta W_{\mathrm{MCS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) = W_{\mathrm{MCS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) - W_{\mathrm{MCS}}(r^*_{\boldsymbol{\eta}}, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta})$. For $W_{\mathrm{MCS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) = \sum_y \eta_y \exp\left(-\left(g_y - \max_{y' \neq y} g_{y'}\right)\right) e^r + ce^{-\alpha r}$, using Lemma C.18, we can preliminarily bound $\sum_y \eta_y \exp\left(-\left(g_y - \max_{y' \neq y} g_{y'}\right)\right)$ as

$$\begin{aligned}
&\sum_y \eta_y \exp\left(-\left(g_y - \max_{y' \neq y} g_{y'}\right)\right) \\
&\geqslant \eta_{\tilde{y}} \exp(-\Delta) + (1 - \eta_{\tilde{y}}) \exp(\Delta) := G(\Delta; \eta_{\tilde{y}}),
\end{aligned} \tag{31}$$

where $\tilde{y} \in \operatorname{argmax}_y g_y$, $0 \leqslant \eta_{\tilde{y}} \leqslant 1$, and $\Delta \geqslant 0$.

Next, we can derive the infimum for $G(\Delta; \eta_{\tilde{y}})$ to further bound $W_{\mathrm{MCS}}(r, \boldsymbol{g}; \boldsymbol{\eta})$. We can see that $G(\Delta; \eta_{\tilde{y}})$ is convex in Δ. Let Δ^* be the minimizer of $G(\Delta; \eta_{\tilde{y}})$. By taking the derivative, we obtain $G'(\Delta; \eta_{\tilde{y}}) = -\eta_{\tilde{y}} \exp(-\Delta) + (1 - \eta_{\tilde{y}}) \exp(\Delta)$. Then we have that $G'(\Delta^*; \eta_{\tilde{y}}) = 0$ holds if

$$\Delta^* = \frac{1}{2} \log\left(\frac{\eta_{\tilde{y}}}{1 - \eta_{\tilde{y}}}\right).$$

Due to the convexity, we have that $G'(\Delta;\eta_{\tilde{y}})$ is nondecreasing for all Δ. First consider the case $\eta_{\tilde{y}} \geqslant \frac{1}{2}$. In this case we have $\Delta^* \geqslant 0$. Then we have

$$G(\Delta;\eta_{\tilde{y}}) \geqslant G(\Delta^*;\eta_{\tilde{y}}) = 2\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})}, \text{ for } \Delta \geqslant 0.$$

Next consider the case $\eta_{\tilde{y}} < \frac{1}{2}$. In this case we have $\Delta^* < 0$. Then we have

$$G(\Delta;\eta_{\tilde{y}}) \geqslant G(0;\eta_{\tilde{y}}) = 1, \text{ for } \Delta \geqslant 0. \tag{32}$$

Thus we have

$$\inf_{\Delta \geqslant 0} G(\Delta;\eta_{\tilde{y}}) = \min\left\{2\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})}, 1\right\} = 2\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})}, \tag{33}$$

where $\eta_{\tilde{y}} \geqslant \frac{1}{2}$.

Since ψ_1, ψ_2 satisfy the calibration condition in Theorem D.1, we have $\alpha = 2\sqrt{\frac{1-c}{c}} > 2$ for $0 < c < \frac{1}{2}$.

Next, we will analyze five cases to derive the excess risk bound for L_{MCS}.

(**i**) If $\eta^* > (1-c)$ and $r > 0$: In this case, by Eq. (30) we have $\Delta W_{0\text{-}1\text{-}c}(r, \boldsymbol{g}; \boldsymbol{\eta}) = \eta^* - \eta_{\tilde{y}}$. For multiplicative CS loss, we have $\Delta W_{\mathrm{MCS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) = W_{\mathrm{MCS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) - W_{\mathrm{MCS}}(r^*_{\boldsymbol{\eta}}, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta})$, where $W_{\mathrm{MCS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) \geqslant G(\Delta;\eta_{\tilde{y}})$ (result by Eq. (31)). For $G(\Delta;\eta_{\tilde{y}})$ we have $\inf_{\Delta \geqslant 0} G(\Delta;\eta_{\tilde{y}}) = 2\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})}$, where $\eta_{\tilde{y}} \geqslant \frac{1}{2}$ (result by Eq. (33)). By Lemma C.2, we have $W_{\mathrm{MCS}}(r^*_{\boldsymbol{\eta}}, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta}) = (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} \left(2\sqrt{\eta^*(1-\eta^*)}\right)^{\frac{\alpha}{\alpha+1}}$ for $\eta^* > 1-c > \frac{1}{2}$. Thus we have

$$
\begin{aligned}
\Delta W_{\mathrm{MCS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) &= \sum_y \eta_y \exp\left(-\left(g_y - \max_{y' \neq y} g_{y'}\right)\right) e^r + ce^{-\alpha r} - (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} \left(2\sqrt{\eta^*(1-\eta^*)}\right)^{\frac{\alpha}{\alpha+1}} \\
&\geqslant G(\Delta; \eta_{\tilde{y}}) e^r + ce^{-\alpha r} - (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} \left(2\sqrt{\eta^*(1-\eta^*)}\right)^{\frac{\alpha}{\alpha+1}} \\
&\qquad \left(\textstyle\sum_y \eta_y \exp\left(-\left(g_y - \max_{y' \neq y} g_{y'}\right)\right) \geqslant G(\Delta; \eta_{\tilde{y}}) \text{ in Eq. (31)}\right) \\
&\geqslant 2\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})} e^r + ce^{-\alpha r} - (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} \left(2\sqrt{\eta^*(1-\eta^*)}\right)^{\frac{\alpha}{\alpha+1}} \\
&\qquad \left(\inf_{\Delta \geqslant 0} G(\Delta) = 2\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})}, \text{ where } \eta_{\tilde{y}} \geqslant \tfrac{1}{2} \text{ in Eq. (33)}\right) \\
&\geqslant (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} \left(\left(2\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})}\right)^{\frac{\alpha}{\alpha+1}} - \left(2\sqrt{\eta^*(1-\eta^*)}\right)^{\frac{\alpha}{\alpha+1}}\right) \qquad (34) \\
&\geqslant (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} \left(2\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})} - \left(2\sqrt{\eta^*(1-\eta^*)}\right)^{\frac{\alpha}{\alpha+1}}\right) \\
&\qquad \left(2\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})} \leqslant 1, \tfrac{\alpha}{\alpha+1} \leqslant 1 \text{ for } \alpha > 2\right) \\
&\geqslant (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} \left(2\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})} - \left(1 + \frac{\alpha}{\alpha+1}\left(2\sqrt{\eta^*(1-\eta^*)} - 1\right)\right)\right) \\
&\qquad \left(x^{\frac{\alpha}{\alpha+1}} \text{ is concave in } x \geqslant 0\right) \\
&= (\alpha+1)\alpha^{\frac{1}{\alpha+1}} c^{\frac{1}{\alpha+1}} \left(\frac{1}{\alpha}\left(2\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})} - 1\right) - \frac{1}{\alpha+1}\left(2\sqrt{\eta^*(1-\eta^*)} - 1\right)\right) \\
&\geqslant (\alpha+1)\alpha^{\frac{1}{\alpha+1}} c^{\frac{1}{\alpha+1}} \left(\frac{1}{\alpha}\left(2\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})} - 2\sqrt{\eta^*(1-\eta^*)}\right)\right) \qquad \left(\tfrac{1}{\alpha} > \tfrac{1}{\alpha+1}\right) \\
&= (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} \left(2\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})} - 2\sqrt{\eta^*(1-\eta^*)}\right) \\
&\geqslant \frac{5}{12}\left(2\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})} - 2\sqrt{\eta^*(1-\eta^*)}\right) \qquad \text{(result by Lemma C.4)} \\
&= \frac{5}{6}\frac{\eta_{\tilde{y}}(1-\eta_{\tilde{y}}) - \eta^*(1-\eta^*)}{\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})} + \sqrt{\eta^*(1-\eta^*)}} = \frac{5}{6}\frac{(\eta^* - \eta_{\tilde{y}})(\eta^* + \eta_{\tilde{y}} - 1)}{\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})} + \sqrt{\eta^*(1-\eta^*)}} \\
&\geqslant \frac{5}{6}\frac{\eta^* - \eta_{\tilde{y}}}{\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})} + \sqrt{\eta^*(1-\eta^*)}} \qquad \left(\eta^* > \tfrac{1}{2}, \eta_{\tilde{y}} \geqslant \tfrac{1}{2}\right) \\
&\geqslant \frac{5}{6}(\eta^* - \eta_{\tilde{y}}) \qquad \left(\eta^* \geqslant \eta_{\tilde{y}},\ \sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})} \leqslant \tfrac{1}{2}, \text{ and } \sqrt{\eta^*(1-\eta^*)} \leqslant \tfrac{1}{2}\right) \\
&= \frac{5}{6} W_{0\text{-}1\text{-}c}(r_{\boldsymbol{\eta}}, \boldsymbol{g}_{\boldsymbol{\eta}}; \boldsymbol{\eta}),
\end{aligned}
$$

In Eq. (34), consider function $I(r;\eta) = 2\sqrt{\eta(1-\eta)}e^r + ce^{-\alpha r}$, where $0 \leqslant \eta \leqslant 1$. Let $r_\eta \in \operatorname{argmin}_r I(r;\eta)$ be a minimizer of $I(r;\eta)$. Then by the same argument as Lemma C.2 with η^* replaced with $\eta_{\tilde{y}}$ and condition that $\eta_{\tilde{y}} \geqslant \frac{1}{2}$, we have $I(r_{\eta_{\tilde{y}}}; \eta_{\tilde{y}}) = (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} \left(2\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})}\right)^{\frac{\alpha}{\alpha+1}}$.

Therefore,

$$
\begin{aligned}
R_{0\text{-}1\text{-}c}(r, \boldsymbol{g}; \boldsymbol{\eta}) - R_{0\text{-}1\text{-}c}(r^*_{\boldsymbol{\eta}}, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta}) &= \mathbb{E}\left[\Delta W_{0\text{-}1\text{-}c}(r, \boldsymbol{g}; \boldsymbol{\eta})\right] \\
&\leqslant \mathbb{E}\left[\Gamma_1\left(\Delta W_{\mathrm{MCS}}(r, \boldsymbol{g}; \boldsymbol{\eta})\right)\right] \\
&\leqslant \Gamma_1\left(\mathbb{E}\left[\Delta W_{\mathrm{MCS}}(r, \boldsymbol{g}; \boldsymbol{\eta})\right]\right) \quad (\Gamma_1 \text{ is concave}) \\
&\leqslant \Gamma_1\left(R_{\mathrm{MCS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) - R_{\mathrm{MCS}}(r^*_{\boldsymbol{\eta}}, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta})\right)
\end{aligned}
$$

where $\Gamma_1(t) = \frac{6}{5}t$.

(**ii**) If $\frac{1}{2} \leqslant \eta^* \leqslant (1-c)$ and $r > 0$: In this case, by Eq. (30) we have $\Delta W_{0\text{-}1\text{-}c}(r, \boldsymbol{g}; \boldsymbol{\eta}) = 1 - c - \eta_{\tilde{y}}$. Consider $\Delta W_{\mathrm{MCS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) = W_{\mathrm{MCS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) - W_{\mathrm{MCS}}(r^*_{\boldsymbol{\eta}}, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta})$, where $W_{\mathrm{MCS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) \geqslant G(\Delta; \eta_{\tilde{y}})$ (result by Eq. (31)). For

$G(\Delta;\eta_{\tilde{y}})$ we have $\inf_{\Delta\geqslant 0} G(\Delta;\eta_{\tilde{y}}) = 2\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})}$, where $\eta_{\tilde{y}} \geqslant \frac{1}{2}$ (result by Eq. (33)). By Lemma C.2, we have $W_{\mathrm{MCS}}(r^*_{\boldsymbol{\eta}}, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta}) = (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} \left(2\sqrt{\eta^*(1-\eta^*)}\right)^{\frac{\alpha}{\alpha+1}}$ for $\eta^* \geqslant \frac{1}{2}$. Thus we have

$$
\begin{aligned}
\Delta W_{\mathrm{MCS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) &= \sum_y \eta_y \exp\left(-\left(g_y - \max_{y'\neq y} g_{y'}\right)\right) e^r + ce^{-\alpha r} - (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} \left(2\sqrt{\eta^*(1-\eta^*)}\right)^{\frac{\alpha}{\alpha+1}} \\
&\geqslant \left(\eta_{\tilde{y}} \exp(-\Delta) + (1-\eta_{\tilde{y}}) \exp(\Delta)\right) e^r + ce^{-\alpha r} - (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} \left(2\sqrt{\eta^*(1-\eta^*)}\right)^{\frac{\alpha}{\alpha+1}} \\
&\qquad \left(\textstyle\sum_y \eta_y \exp\left(-\left(g_y - \max_{y'\neq y} g_{y'}\right)\right) \geqslant G(\Delta;\eta_{\tilde{y}}) \text{ in Eq. (31)}\right) \\
&\geqslant \eta_{\tilde{y}} \exp(-\Delta) + (1-\eta_{\tilde{y}}) \exp(\Delta) + c - (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} \left(2\sqrt{\eta^*(1-\eta^*)}\right)^{\frac{\alpha}{\alpha+1}} \qquad (35) \\
&\geqslant 2\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})} + c - (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} \left(2\sqrt{\eta^*(1-\eta^*)}\right)^{\frac{\alpha}{\alpha+1}} \\
&\qquad \left(\inf_{\Delta\geqslant 0} G(\Delta) = 2\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})}, \text{ where } \eta_{\tilde{y}} \geqslant \tfrac{1}{2} \text{ in Eq. (33)}\right) \\
&= 2\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})} + c - (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} \left(2\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})}\right)^{\frac{\alpha}{\alpha+1}} \\
&\qquad \left(2\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})} \geqslant 2\sqrt{\eta^*(1-\eta^*)} \text{ for } \eta^* \geqslant \eta_{\tilde{y}} \geqslant \tfrac{1}{2}\right) \\
&\geqslant \frac{c\alpha(\alpha+1)}{2} \left(1 - \left(\frac{\alpha c}{2\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})}}\right)^{\frac{1}{\alpha+1}}\right)^2 \qquad (36) \\
&\geqslant \frac{c\alpha(\alpha+1)}{2} \left(1 - \left(1 + \frac{1}{\alpha+1}\left(\frac{\alpha c}{2\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})}} - 1\right)\right)\right)^2 \\
&\qquad \left(x^{\frac{1}{\alpha+1}} \text{ is concave in } x>0,\ 2\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})} \geqslant \alpha c \text{ for } \eta^* \geqslant \eta_{\tilde{y}} \geqslant \tfrac{1}{2}\right) \\
&= \frac{\alpha c}{2(\alpha+1)} \left(1 - \frac{\alpha c}{2\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})}}\right)^2 \\
&\geqslant \frac{\alpha c}{2(\alpha+1)} \left(1 - \frac{\alpha c}{\frac{1-c-\eta_{\tilde{y}}}{\frac{1}{2}-c} + \left(1 - \frac{1-c-\eta_{\tilde{y}}}{\frac{1}{2}-c}\right)\alpha c}\right)^2 \qquad (37) \\
&= \frac{\alpha c}{2(\alpha+1)} \left(1 - \frac{\alpha c(\frac{1}{2}-c)}{(\alpha c - 1)\eta_{\tilde{y}} + 1 - \frac{1}{2}\alpha c - c}\right)^2 \\
&\geqslant \frac{\alpha c}{2(\alpha+1)} \left(1 - \alpha c(\frac{1}{2}-c)\left(\frac{1-c-\eta_{\tilde{y}}}{\frac{1}{2}-c}\frac{1}{\frac{1}{2}-c} + \left(1 - \frac{1-c-\eta_{\tilde{y}}}{\frac{1}{2}-c}\right)\frac{1}{\alpha c(\frac{1}{2}-c)}\right)\right)^2 \qquad (38) \\
&= \frac{2\alpha c(1-\alpha c)^2}{(\alpha+1)(1-2c)^2} (1-c-\eta_{\tilde{y}})^2 \\
&= W_c^{(1)} (1-c-\eta_{\tilde{y}})^2 \qquad (39) \\
&= W_c^{(1)} \Delta W_{0\text{-}1\text{-}c}(r, \boldsymbol{g}; \boldsymbol{\eta})^2,
\end{aligned}
$$

where in Eq. (35) we used Lemma C.3 and the fact that

$$
\begin{aligned}
\eta_{\tilde{y}} \exp(-\Delta) + (1-\eta_{\tilde{y}}) \exp(\Delta) &\geqslant \eta^* \exp(-\Delta) + (1-\eta^*) \exp(\Delta) \\
&\geqslant (1-c) \exp(-\Delta) + c \exp(\Delta) \\
&\geqslant 2\sqrt{c(1-c)} = \alpha c
\end{aligned}
$$

for $c = \frac{4}{4+\alpha^2}$. In Eq. (36), let $S_{\eta_{\tilde{y}}} = 2\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})}$, $g(x) = S_{\eta_{\tilde{y}}} x + cx^{-\alpha} (x>0)$, $x^*_{\tilde{y}} \in \operatorname{argmin}_{x>0}\ g(x)$. Note that $g(1) = 2\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})} + c$, and we will show that $g(x^*_{\eta_{\tilde{y}}}) = (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} \left(2\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})}\right)^{\frac{\alpha}{\alpha+1}}$. By taking the

derivative, we obtain $g'(x) = S_{\eta_{\tilde{y}}} - c\alpha x^{-\alpha-1}$. Then we have that $g'(x^*_{\eta_{\tilde{y}}}) = 0$ holds if

$$x^*_{\eta_{\tilde{y}}} = \left(\frac{\alpha c}{S_{\eta_{\tilde{y}}}}\right)^{\frac{1}{\alpha+1}} \leqslant 1,$$

where $S_{\eta_{\tilde{y}}} = 2\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})} \geqslant 2\sqrt{c(1-c)} = \alpha c$ (note that $\frac{1}{2} \leqslant \eta_{\tilde{y}} \leqslant \eta^* \leqslant (1-c)$). Since $g(x)$ is convex, we have $g(x) \geqslant g(x^*_{\eta_{\tilde{y}}}) = (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} \left(S_{\eta_{\tilde{y}}}\right)^{\frac{\alpha}{\alpha+1}}$. Since $g'(x^*_{\eta_{\tilde{y}}}) = 0$ and $g''(x) = c\alpha(\alpha+1)x^{-\alpha-2}$ is decreasing, it holds for $x^*_{\eta_{\tilde{y}}} \leqslant 1$ that

$$\begin{aligned} g(1) - g(x^*_{\eta_{\tilde{y}}}) &\geqslant \frac{g''(1)}{2}(1 - x^*_{\eta_{\tilde{y}}})^2 \\ &\geqslant \frac{c\alpha(\alpha+1)}{2}\left(1 - \left(\frac{\alpha c}{S_{\eta_{\tilde{y}}}}\right)^{\frac{1}{\alpha+1}}\right)^2. \end{aligned}$$

The proof of Eq. (36) is completed.

Eq. (37) follows from the fact that $h\left(t\cdot\frac{1}{2} + (1-t)(1-c)\right) \geqslant t\cdot h\left(\frac{1}{2}\right) + (1-t)h(1-c)$, where $t = \frac{1-c-\eta_{\tilde{y}}}{\frac{1}{2}-c} \in [0,1]$, for concave function $h(\eta_{\tilde{y}}) = 2\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})}$. In Eq. (38). Eq. (39) follows from the fact that $k\left(t\cdot\frac{1}{2} + (1-t)(1-c)\right) \leqslant t\cdot k\left(\frac{1}{2}\right) + (1-t)k(1-c)$, where $t = \frac{1-c-\eta_{\tilde{y}}}{\frac{1}{2}-c} \in [0,1]$, for convex function $k(\eta_{\tilde{y}}) = \frac{1}{(\alpha c-1)\eta_{\tilde{y}} + 1 - \frac{1}{2}\alpha c - c}$. $W_c^{(1)} = \frac{2c\alpha(1-\alpha c)^2}{(\alpha+1)(1-2c)^2}$ is a positive constant that depends on $c < \frac{1}{2}$, where $\alpha = 2\sqrt{\frac{1-c}{c}}$. Note that $\alpha c = 2\sqrt{c(1-c)} < 1$. Therefore,

$$\begin{aligned} R_{\text{0-1-}c}(r, \boldsymbol{g}; \boldsymbol{\eta}) - R_{\text{0-1-}c}(r^*_{\boldsymbol{\eta}}, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta}) &= \mathbb{E}\left[\Delta W_{\text{0-1-}c}(r, \boldsymbol{g}; \boldsymbol{\eta})\right] \\ &\leqslant \mathbb{E}\left[\Gamma_2\left(\Delta W_{\text{MCS}}(r, \boldsymbol{g}; \boldsymbol{\eta})\right)\right] \\ &\leqslant \Gamma_2\left(\mathbb{E}\left[\Delta W_{\text{MCS}}(r, \boldsymbol{g}; \boldsymbol{\eta}]\right)\right) \quad (\Gamma_2 \text{ is concave}) \\ &\leqslant \Gamma_2\left(R_{\text{MCS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) - R_{\text{MCS}}(r^*_{\boldsymbol{\eta}}, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta})\right), \end{aligned}$$

where $\Gamma_2(t) = \sqrt{\frac{1}{W_c^{(1)}}}\sqrt{t}$ ($W_c^{(1)}$ is a positive constant that depends on c).

(**iii**) If $\eta^* < \frac{1}{2} < (1-c)$ and $r > 0$: In this case, by Eq. (30) we have $\Delta W_{\text{0-1-}c}(r, \boldsymbol{g}; \boldsymbol{\eta}) = 1 - c - \eta_{\tilde{y}}$. Consider $\Delta W_{\text{MCS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) = W_{\text{MCS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) - W_{\text{MCS}}(r^*_{\boldsymbol{\eta}}, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta})$, by Lemma C.2, we have $W_{\text{MCS}}(r^*_{\boldsymbol{\eta}}, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta}) = (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}}$ for $\eta^* < \frac{1}{2}$. Thus we have

$$\begin{aligned} \Delta W_{\text{MCS}}(r_{\boldsymbol{\eta}}, \boldsymbol{g}_{\boldsymbol{\eta}}; \boldsymbol{\eta}) &= \sum_y \eta_y \exp\left(-\left(g_y - \max_{y' \neq y} g_{y'}\right)\right) e^r + ce^{-\alpha r} - (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} \\ &\geqslant \left(\eta_{\tilde{y}} \exp(-\Delta) + (1-\eta_{\tilde{y}})\exp(\Delta)\right) e^r + ce^{-\alpha r} - (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} \\ &\qquad \left(\textstyle\sum_y \eta_y \exp\left(-\left(g_y - \max_{y'\neq y} g_{y'}\right)\right) \geqslant G(\Delta; \eta_{\tilde{y}}) \text{ in Eq. (31)}\right) \\ &\geqslant \eta_{\tilde{y}} \exp(-\Delta) + (1-\eta_{\tilde{y}})\exp(\Delta) + c - (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} \qquad (40) \\ &\geqslant 1 + c - (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} \qquad \left(\inf_{\Delta \geqslant 0} G(\Delta) = 1 \text{ for } \eta_{\tilde{y}} \leqslant \eta^* < \tfrac{1}{2} \text{ in Eq. (32)}\right) \\ &= \frac{1 + c - (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}}}{1-c}(1-c) \\ &= W_c^{(2)}(1-c) \qquad (41) \\ &\geqslant W_c^{(2)}(1-c-\eta_{\tilde{y}}) \qquad (\eta_{\tilde{y}} \geqslant 0) \\ &= W_c^{(2)} \Delta W_{\text{0-1-}c}(r, \boldsymbol{g}; \boldsymbol{\eta}), \end{aligned}$$

where in Eq. (40) we used Lemma C.3 and the fact that

$$\begin{aligned}\eta_{\tilde{y}}\exp(-\Delta)+(1-\eta_{\tilde{y}})\exp(\Delta) &\geqslant \eta^*\exp(-\Delta)+(1-\eta^*)\exp(\Delta)\\ &\geqslant (1-c)\exp(-\Delta)+c\exp(\Delta)\\ &\geqslant 2\sqrt{c(1-c)}=\alpha c\end{aligned}$$

for $c=\frac{4}{4+\alpha^2}$. In Eq. (41), $W_c^{(2)}=\frac{1+c-(\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}}c^{\frac{1}{\alpha+1}}}{1-c}$ is a positive constant that depends on $c<\frac{1}{2}$, where $\alpha=2\sqrt{\frac{1-c}{c}}$. Note that

$$\begin{aligned}1+c-(\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}}c^{\frac{1}{\alpha+1}} &\geqslant 1+c-(\alpha c)^{\frac{1}{\alpha+1}}\\ &\geqslant 1+c-\left(1+\frac{1}{\alpha+1}(\alpha c-1)\right) && \left(x^{\frac{1}{\alpha+1}}\text{ is concave in } x>0\right)\\ &= c-\frac{\alpha c-1}{\alpha+1}\\ &= \frac{c+1}{\alpha+1}\\ &> 0. && (c,\alpha>0)\end{aligned}$$

Therefore,

$$\begin{aligned}R_{0\text{-}1\text{-}c}(r,\boldsymbol{g};\boldsymbol{\eta})-R_{0\text{-}1\text{-}c}(r^*_{\boldsymbol{\eta}},\boldsymbol{g}^*_{\boldsymbol{\eta}};\boldsymbol{\eta}) &= \mathbb{E}\left[\Delta W_{0\text{-}1\text{-}c}(r,\boldsymbol{g};\boldsymbol{\eta})\right]\\ &\leqslant \mathbb{E}\left[\Gamma_3\left(\Delta W_{\mathrm{MCS}}(r,\boldsymbol{g};\boldsymbol{\eta})\right)\right]\\ &\leqslant \Gamma_3\left(\mathbb{E}\left[\Delta W_{\mathrm{MCS}}(r,\boldsymbol{g};\boldsymbol{\eta})\right]\right) \quad (\Gamma_2\text{ is concave})\\ &\leqslant \Gamma_3\left(R_{\mathrm{MCS}}(r,\boldsymbol{g};\boldsymbol{\eta})-R_{\mathrm{MCS}}(r^*_{\boldsymbol{\eta}},\boldsymbol{g}^*_{\boldsymbol{\eta}};\boldsymbol{\eta})\right),\end{aligned}$$

where $\Gamma_3(t)=\frac{1}{W_c^{(2)}}t$ ($W_c^{(2)}$ is a positive constant that depends on c).

(**iv**) If $\eta^*\leqslant(1-c)$ and $r\leqslant 0$: In this case, by Eq. (30) we have $\Delta W_{0\text{-}1\text{-}c}(r_{\boldsymbol{\eta}},\boldsymbol{g}_{\boldsymbol{\eta}};\boldsymbol{\eta})=0$, which implies that $W_{0\text{-}1\text{-}c}(r_{\boldsymbol{\eta}},\boldsymbol{g}_{\boldsymbol{\eta}};\boldsymbol{\eta})-W_{0\text{-}1\text{-}c}(r^*_{\boldsymbol{\eta}},\boldsymbol{g}^*_{\boldsymbol{\eta}};\boldsymbol{\eta})=\Delta W_{0\text{-}1\text{-}c}(r_{\boldsymbol{\eta}},\boldsymbol{g}_{\boldsymbol{\eta}};\boldsymbol{\eta})=0\leqslant\Gamma\left(W_{\mathrm{MCS}}(r_{\boldsymbol{\eta}},\boldsymbol{g}_{\boldsymbol{\eta}};\boldsymbol{\eta})-W_{\mathrm{MCS}}(r^*_{\boldsymbol{\eta}},\boldsymbol{g}^*_{\boldsymbol{\eta}};\boldsymbol{\eta})\right)$ for any $\Gamma\geqslant 0$.

(**v**) If $\eta^*>(1-c)$ and $r\leqslant 0$: In this case, by Eq. (30) we have $\Delta W_{0\text{-}1\text{-}c}(r_{\boldsymbol{\eta}},\boldsymbol{g}_{\boldsymbol{\eta}};\boldsymbol{\eta})=\eta^*-1+c$. Consider $\Delta W_{\mathrm{MCS}}(r_{\boldsymbol{\eta}},\boldsymbol{g}_{\boldsymbol{\eta}};\boldsymbol{\eta})=W_{\mathrm{MCS}}(r_{\boldsymbol{\eta}},\boldsymbol{g}_{\boldsymbol{\eta}};\boldsymbol{\eta})-W_{\mathrm{MCS}}(r^*_{\boldsymbol{\eta}},\boldsymbol{g}^*_{\boldsymbol{\eta}};\boldsymbol{\eta})$, by Lemma C.2, we have $W_{\mathrm{MCS}}(r^*_{\boldsymbol{\eta}},\boldsymbol{g}^*_{\boldsymbol{\eta}};\boldsymbol{\eta})=(\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}}\left(2\sqrt{\eta^*(1-\eta^*)}\right)^{\frac{\alpha}{\alpha+1}}c^{\frac{1}{\alpha+1}}$ for $\eta^*>(1-c)>\frac{1}{2}$. Thus we have

$$
\begin{aligned}
\Delta W_{\mathrm{MCS}}(r_{\boldsymbol{\eta}}, \boldsymbol{g}_{\boldsymbol{\eta}}; \boldsymbol{\eta}) &= \sum_y \eta_y \exp\left(-\left(g_y - \max_{y' \neq y} g_{y'}\right)\right) e^r + ce^{-\alpha r} - (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} \left(2\sqrt{\eta^*(1-\eta^*)}\right)^{\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} \\
&\geqslant \left(2\sqrt{\eta^*(1-\eta^*)}\right) e^r + ce^{-\alpha r} - (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} \left(2\sqrt{\eta^*(1-\eta^*)}\right)^{\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} \quad \text{(result by Eq. (5))} \\
&\geqslant 2\sqrt{\eta^*(1-\eta^*)} + c - (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} \left(2\sqrt{\eta^*(1-\eta^*)}\right)^{\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} \qquad (42) \\
&= S_{\eta^*} + c - (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} \left(S_{\eta^*}\right)^{\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} \qquad \left(S_{\eta^*} = 2\sqrt{\eta^*(1-\eta^*)}\right) \\
&\geqslant \frac{c\alpha(\alpha+1)}{2} \left(\frac{S_{\eta^*}}{\alpha c}\right)^{\frac{\alpha+2}{\alpha+1}} \left(1 - \left(\frac{\alpha c}{S_{\eta^*}}\right)^{\frac{1}{\alpha+1}}\right)^2 \qquad (43) \\
&= \frac{c\alpha(\alpha+1)}{2} \left(\left(\frac{S_{\eta^*}}{\alpha c}\right)^{\frac{\alpha+2}{2(\alpha+1)}} - \left(\frac{S_{\eta^*}}{\alpha c}\right)^{\frac{\alpha}{2(\alpha+1)}}\right)^2 \\
&\geqslant \frac{c\alpha(\alpha+1)}{2} \left(\left(\frac{S_{\eta^*}}{\alpha c}\right)^{\frac{\alpha+2}{2(\alpha+1)}} - \left(\frac{S_{\eta^*}}{\alpha c}\right)^{\frac{1}{2}}\right)^2 \\
&\qquad \left(S_{\eta^*} < 2\sqrt{c(1-c)} = \alpha c, \left(\frac{S_{\eta^*}}{\alpha c}\right)^{\frac{\alpha+2}{2(\alpha+1)}} < \left(\frac{S_{\eta^*}}{\alpha c}\right)^{\frac{1}{2}} < \left(\frac{S_{\eta^*}}{\alpha c}\right)^{\frac{\alpha}{2(\alpha+1)}}\right) \\
&\geqslant \frac{c\alpha(\alpha+1)}{2} \left(\left(1 + \frac{\alpha+2}{\alpha+1}\left(\left(\frac{S_{\eta^*}}{\alpha c}\right)^{\frac{1}{2}} - 1\right)\right) - \left(\frac{S_{\eta^*}}{\alpha c}\right)^{\frac{1}{2}}\right)^2 \qquad (44) \\
&= \frac{\alpha c}{2(\alpha+1)} \left(\left(\frac{S_{\eta^*}}{\alpha c}\right)^{\frac{1}{2}} - 1\right)^2 \\
&= \frac{\alpha c}{2(\alpha+1)} \left(\frac{\frac{S_{\eta^*}}{\alpha c} - 1}{\left(\frac{S_{\eta^*}}{\alpha c}\right)^{\frac{1}{2}} + 1}\right)^2 \\
&\geqslant \frac{\alpha c}{8(\alpha+1)} \left(\frac{S_{\eta^*}}{\alpha c} - 1\right)^2 \qquad \left(S_{\eta^*} < 2\sqrt{c(1-c)} = \alpha c\right) \\
&= \frac{1}{8c\alpha(\alpha+1)} \left(2\sqrt{\eta^*(1-\eta^*)} - \alpha c\right)^2 \qquad \left(S_{\eta^*} = 2\sqrt{\eta^*(1-\eta^*)}\right) \\
&\geqslant \frac{1}{8c\alpha(\alpha+1)} \left(\frac{4\eta^*(1-\eta^*) - \alpha^2 c^2}{2\sqrt{\eta^*(1-\eta^*)} + \alpha c}\right)^2 \\
&\geqslant \frac{1}{32c^2\alpha^2(\alpha+1)} \left(4\eta^*(1-\eta^*) - \alpha^2 c^2\right)^2 \qquad \left(2\sqrt{\eta^*(1-\eta^*)} < 2\sqrt{c(1-c)} = \alpha c\right) \\
&\geqslant \frac{1}{32c^2\alpha^2(\alpha+1)} \left(4(2c-1)\eta^* + 4(2c-1)(c-1)\right)^2 \qquad (45) \\
&= \frac{(1-2c)^2}{2c^2\alpha^2(\alpha+1)} \left(\eta^* - 1 + c\right)^2 \\
&= W_c^{(3)} \left(\eta^* - 1 + c\right)^2 \qquad (46) \\
&= W_c^{(3)} \Delta W_{\text{0-1-}c}(r, \boldsymbol{g}; \boldsymbol{\eta})^2,
\end{aligned}
$$

where in Eq. (42) we used Lemma C.3 (note that $2\sqrt{\eta^*(1-\eta^*)} < 2\sqrt{c(1-c)} = \alpha c$ for $\eta^* > 1-c$, $c = \frac{4}{4+\alpha^2}$). In Eq. (43), let $h(x) = S_{\eta^*}x + cx^{-\alpha}(x > 0)$, $x^*_{\eta^*} \in \operatorname*{argmin}_{x>0} h(x)$. Note that $h(1) = S_{\eta^*} + c$, and we will show that $h(x^*_{\eta^*}) = (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} \left(S_{\eta^*}\right)^{\frac{\alpha}{\alpha+1}}$. By taking the derivative, we obtain $h'(x) = S_{\eta^*} - c\alpha x^{-\alpha-1}$. Then we

have that $h'(x^*_{\eta^*}) = 0$ holds if

$$x^*_{\eta^*} = \left(\frac{\alpha c}{S_{\eta^*}}\right)^{\frac{1}{\alpha+1}} > 1,$$

where $S_{\eta^*} = 2\sqrt{\eta^*(1-\eta^*)} < 2\sqrt{c(1-c)} = \alpha c$ (note that $\eta^* > (1-c)$). Since $h(x)$ is convex, we have $h(x) \geqslant h(x^*_{\eta^*}) = (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} (S_{\eta^*})^{\frac{\alpha}{\alpha+1}}$. Since $h'(x^*_{\eta^*}) = 0$ and $h''(x) = c\alpha(\alpha+1)x^{-\alpha-2}$ is decreasing, it holds for $x^*_{\eta^*} \leqslant 1$ that

$$\begin{aligned} h(1) - h(x^*_{\eta^*}) &\geqslant \frac{h''(x^*_{\eta^*})}{2}(1 - x^*_{\eta^*})^2 \\ &\geqslant \frac{c\alpha(\alpha+1)}{2}\left(\frac{S_{\eta^*}}{\alpha c}\right)^{\frac{\alpha+2}{\alpha+1}}\left(1 - \left(\frac{\alpha c}{S_{\eta^*}}\right)^{\frac{1}{\alpha+1}}\right)^2. \end{aligned}$$

The proof of Eq. (43) is completed.

Eq. (44) follows from the fact that $S_{\eta^*} < 2\sqrt{c(1-c)} = \alpha c$ and

$$\left(\frac{S_{\eta^*}}{\alpha c}\right)^{\frac{\alpha+2}{2(\alpha+1)}} < \left(1 + \frac{\alpha+2}{\alpha+1}\left(\left(\frac{S_{\eta^*}}{\alpha c}\right)^{\frac{1}{2}} - 1\right)\right) < \left(\frac{S_{\eta^*}}{\alpha c}\right)^{\frac{1}{2}}.$$

Here note that $\left(\frac{S_{\eta^*}}{\alpha c}\right)^{\frac{\alpha+2}{2(\alpha+1)}}$ is concave and $\left(1 + \frac{\alpha+2}{\alpha+1}\left(\left(\frac{S_{\eta^*}}{\alpha c}\right)^{\frac{1}{2}} - 1\right)\right) - \left(\frac{S_{\eta^*}}{\alpha c}\right)^{\frac{1}{2}} = \frac{1}{\alpha+1}\left(\frac{S_{\eta^*}}{\alpha c}\right)^{\frac{1}{2}} - \frac{1}{\alpha+1} < 0$.

In Eq. (45), we use the tangent line of $k(\eta^*; \alpha, c) = 4\eta^*(1-\eta^*) - \alpha^2 c^2$ through $(1-c, 0)$ (note that $k'(\eta^*; \alpha, c) = -8\eta^* + 4$ and the tangent line $4(2c-1)\eta^* + 4(2c-1)(c-1) < 0$ for $\eta^* > (1-c)$). In Eq. (46), $W_c^{(1)} = \frac{(1-2c)^2}{2c^2\alpha^2(\alpha+1)}$ is a positive constant that depends on $c < \frac{1}{2}$, where $\alpha = 2\sqrt{\frac{1-c}{c}}$.

Therefore,

$$\begin{aligned} R_{\text{0-1-}c}(r, \boldsymbol{g}; \boldsymbol{\eta}) - R_{\text{0-1-}c}(r^*_{\boldsymbol{\eta}}, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta}) &= \mathbb{E}\left[\Delta W_{\text{0-1-}c}(r, \boldsymbol{g}; \boldsymbol{\eta})\right] \\ &\leqslant \mathbb{E}\left[\Gamma_5\left(\Delta W_{\text{MCS}}(r, \boldsymbol{g}; \boldsymbol{\eta})\right)\right] \\ &\leqslant \Gamma_5\left(\mathbb{E}\left[\Delta W_{\text{MCS}}(r, \boldsymbol{g}; \boldsymbol{\eta}]\right)\right) \quad (\Gamma_5 \text{ is concave}) \\ &\leqslant \Gamma_5\left(R_{\text{MCS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) - R_{\text{MCS}}(r^*_{\boldsymbol{\eta}}, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta})\right), \end{aligned}$$

where $\Gamma_5(t) = \sqrt{\frac{1}{W_c^{(3)}}}\sqrt{t}$ ($W_c^{(3)}$ is a positive constant that depends on c).

Overall, we obtain

$$R_{\text{0-1-}c}(r_{\boldsymbol{\eta}}, \boldsymbol{g}_{\boldsymbol{\eta}}; \boldsymbol{\eta}) - R_{\text{0-1-}c}(r^*_{\boldsymbol{\eta}}, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta}) \leqslant \Gamma\left(R_{\text{MCS}}(r_{\boldsymbol{\eta}}, \boldsymbol{g}_{\boldsymbol{\eta}}; \boldsymbol{\eta}) - R_{\text{MCS}}(r^*_{\boldsymbol{\eta}}, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta})\right),$$

where $\Gamma(t) = \max\left\{\frac{6}{5}t, \sqrt{\frac{1}{W_c^{(1)}}}\sqrt{t}, \frac{1}{W_c^{(2)}}t, \sqrt{\frac{1}{W_c^{(3)}}}\sqrt{t}\right\} = O\left(\max\left\{t, \sqrt{t}\right\}\right)$ ($W_c^{(1)}$, $W_c^{(2)}$, and $W_c^{(3)}$ are positive constants that depends on c).

Next, we prove the following excess risk bound holds for L_{ACS}:

$$R_{\text{0-1-}c}(r, \boldsymbol{g}) - R^*_{\text{0-1-}c} \leqslant \Gamma\left(R_{\text{ACS}}(r, \boldsymbol{g}) - R^*_{\text{ACS}}\right)$$

for all $r, \boldsymbol{g}$, where

$$\Gamma(t) = O(t)$$

for $\phi(r) = \max\{1 - x, 0\}$, $\psi(r) = c\max\{1 - \alpha r, 0\}$.

Proof. In the following, we introduce some useful results for bounding $\Delta W_{\mathrm{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) = W_{\mathrm{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) - W_{\mathrm{ACS}}(r^*_{\boldsymbol{\eta}}, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta})$. For $W_{\mathrm{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) = \sum_y \eta_y \max\{1 - (g_y - \max_{y' \neq y} g_{y'} - r), 0\} + \psi(r)$, using Theorem C.18, we can preliminarily bound $\sum_y \eta_y \max\{1 - (g_y - \max_{y' \neq y} g_{y'} - r), 0\}$ as

$$\begin{aligned}
&\sum_y \eta_y \max\left\{1 - \left(g_y - \max_{y' \neq y} g_{y'} - r\right), 0\right\} \\
&\geqslant \eta_{\tilde{y}} \max\{1 - \Delta + r, 0\} + (1 - \eta_{\tilde{y}}) \max\{1 + \Delta + r, 0\} := G(r, \Delta; \eta_{\tilde{y}}),
\end{aligned} \tag{47}$$

where $\tilde{y} \in \operatorname{argmax}_y g_y$, $0 \leqslant \eta_{\tilde{y}} \leqslant 1$, and $\Delta \geqslant 0$. Then using Lemma C.5, we have

$$\inf_{\Delta \geqslant 0} G(r, \Delta; \eta_{\tilde{y}}) = \begin{cases} 2(1 - \eta_{\tilde{y}})(r + 1) & \text{if } \eta_{\tilde{y}} > \frac{1}{2} \text{ and } r \geqslant -1, \\ r + 1 & \text{if } \eta_{\tilde{y}} \leqslant \frac{1}{2} \text{ and } r \geqslant -1, \\ 0 & \text{if } r < -1. \end{cases} \tag{48}$$

Thus we have

$$\begin{aligned}
\inf_{\Delta \geqslant 0} G(r, \Delta; \eta_{\tilde{y}}) &= \begin{cases} \min\{2(1 - \eta_{\tilde{y}})(r + 1), r + 1\} & \text{if } r \geqslant -1, \\ 0 & \text{if } r < -1. \end{cases} \\
&= \begin{cases} 2(1 - \eta_{\tilde{y}})(r + 1) & \text{if } r \geqslant -1, \\ 0 & \text{if } r < -1, \end{cases}
\end{aligned} \tag{49}$$

where $\eta_{\tilde{y}} > \frac{1}{2}$.

Since ψ satisfy the calibration condition in Theorem 4.2, we have $\alpha = 2$ for $\psi(r) = c \exp(-\alpha r)$ or $\psi(r) = c \max\{1 - \alpha r, 0\}$.

Next, we will analyze six cases to derive the excess risk bound for L_{ACS}.

(**i**) If $\eta^* > (1 - c)$ and $r > 0$: In this case, by Eq. (30) we have $\Delta W_{0\text{-}1\text{-}c}(r, \boldsymbol{g}; \boldsymbol{\eta}) = \eta^* - \eta_{\tilde{y}}$. Consider $\Delta W_{\mathrm{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) = W_{\mathrm{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) - W_{\mathrm{ACS}}(r^*_{\boldsymbol{\eta}}, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta})$, where $W_{\mathrm{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) \geqslant G(r, \Delta; \eta_{\tilde{y}})$ (result by Eq. (47)). For $G(r, \Delta; \eta_{\tilde{y}})$ we have $\inf_{\Delta \geqslant 0} G(r, \Delta; \eta_{\tilde{y}}) = 2(1 - \eta_{\tilde{y}})(r + 1)$, where $\eta_{\tilde{y}} \geqslant \frac{1}{2}$, for $r > 0$ (result by Eq. (49)). By Lemma C.6, we have $W_{\mathrm{ACS}}(r^*_{\boldsymbol{\eta}}, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta}) = 2(1 - \eta^*)(r^*_{\boldsymbol{\eta}} + 1) + \psi(r^*_{\boldsymbol{\eta}})$ for $\eta^* > (1 - c) > \frac{1}{2}$. Thus we have

$$\begin{aligned}
\Delta W_{\mathrm{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) &= W_{\mathrm{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) - W_{\mathrm{ACS}}(r^*_{\boldsymbol{\eta}}, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta}) \\
&\geqslant G(r, \Delta; \eta_{\tilde{y}}) + ce^{-\alpha r} - \left(2(1 - \eta^*)(r^*_{\boldsymbol{\eta}} + 1) + \psi(r^*_{\boldsymbol{\eta}})\right) \\
&\qquad \left(\textstyle\sum_y \eta_y \max\{1 - (g_y - \max_{y' \neq y} g_{y'} - r), 0\} \geqslant G(\Delta, r; \eta_{\tilde{y}}) \text{ in Eq. (47)}\right) \\
&\geqslant 2(1 - \eta_{\tilde{y}})(r + 1) + \psi(r) - \left(2(1 - \eta^*)(r^*_{\boldsymbol{\eta}} + 1) + \psi(r^*_{\boldsymbol{\eta}})\right) \\
&\qquad \left(\inf_{\Delta \geqslant 0} G(r, \Delta; \eta_{\tilde{y}}) = 2(1 - \eta_{\tilde{y}})(r + 1) \text{ where } \eta_{\tilde{y}} > \tfrac{1}{2} \text{ for } r > 0 \text{ in Eq. (49)}\right) \\
&\geqslant \inf_{r > 0} \{2(1 - \eta_{\tilde{y}})(r + 1) + \psi(r)\} - \left(2(1 - \eta^*)(r^*_{\boldsymbol{\eta}} + 1) + \psi(r^*_{\boldsymbol{\eta}})\right) \\
&:= f_1(\eta^*, \eta_{\tilde{y}}).
\end{aligned} \tag{50}$$

For $\psi(r) = c \max\{1 - \alpha r, 0\}$, $r^*_{\boldsymbol{\eta}} = \frac{1}{\alpha}$, $\inf_{r > 0} \{2(1 - \eta_{\tilde{y}})(r + 1) + \psi(r)\} = \min\{2(1 - \eta_{\tilde{y}}) + c, 3(1 - \eta_{\tilde{y}})\}$. Then we have

$$\begin{aligned}
f_1(\eta^*, \eta_{\tilde{y}}) &= \min\left\{2(1 - \eta_{\tilde{y}}) + c, 2(1 - \eta_{\tilde{y}})(\frac{1}{\alpha} + 1)\right\} - 2(1 - \eta^*) \cdot \left(\frac{1}{\alpha} + 1\right) \\
&= \min\{2(1 - \eta_{\tilde{y}}) + c, 3(1 - \eta_{\tilde{y}})\} - 3(1 - \eta^*) \qquad (\alpha = 2 \text{ from Theorem 4.2}) \\
&\geqslant \min\{2(\eta^* - \eta_{\tilde{y}}) + (\eta^* - 1 + c), 3(\eta^* - \eta_{\tilde{y}})\} \\
&\geqslant \min\{2(\eta^* - \eta_{\tilde{y}}), 3(\eta^* - \eta_{\tilde{y}})\} \\
&= 2\Delta W_{0\text{-}1\text{-}c}(r, \boldsymbol{g}; \boldsymbol{\eta}).
\end{aligned}$$

Therefore,

$$\begin{aligned} R_{\text{0-1-}c}(r,\boldsymbol{g};\boldsymbol{\eta}) - R_{\text{0-1-}c}(r^*_{\boldsymbol{\eta}},\boldsymbol{g}^*_{\boldsymbol{\eta}};\boldsymbol{\eta}) &= \mathbb{E}\left[\Delta W_{\text{0-1-}c}(r,\boldsymbol{g};\boldsymbol{\eta})\right] \\ &\leqslant \mathbb{E}\left[\Gamma_1\left(\Delta W_{\text{ACS}}(r,\boldsymbol{g};\boldsymbol{\eta})\right)\right] \\ &\leqslant \Gamma_1\left(\mathbb{E}\left[\Delta W_{\text{ACS}}(r,\boldsymbol{g};\boldsymbol{\eta}]\right)\right) \quad (\Gamma_1 \text{ is concave}) \\ &\leqslant \Gamma_1\left(R_{\text{ACS}}(r,\boldsymbol{g};\boldsymbol{\eta}) - R_{\text{ACS}}(r^*_{\boldsymbol{\eta}},\boldsymbol{g}^*_{\boldsymbol{\eta}};\boldsymbol{\eta})\right), \end{aligned}$$

where $\Gamma_1(t) = \frac{1}{2}t$ for $\psi(r) = c\max\{1-\alpha r, 0\}$.

(**ii**) If $\frac{1}{2} < \eta^* \leqslant (1-c)$ and $r > 0$: In this case, by Eq. (30) we have $\Delta W_{\text{0-1-}c}(r,\boldsymbol{g};\boldsymbol{\eta}) = 1 - c - \eta_{\tilde{y}}$. Consider $\Delta W_{\text{ACS}}(r,\boldsymbol{g};\boldsymbol{\eta}) = W_{\text{ACS}}(r,\boldsymbol{g};\boldsymbol{\eta}) - W_{\text{ACS}}(r^*_{\boldsymbol{\eta}},\boldsymbol{g}^*_{\boldsymbol{\eta}};\boldsymbol{\eta})$, where $W_{\text{ACS}}(r,\boldsymbol{g};\boldsymbol{\eta}) \geqslant G(\Delta, r; \eta_{\tilde{y}})$ (result by Eq. (47)). For $G(\Delta;\eta_{\tilde{y}})$ we have $\inf_{\Delta\geqslant 0} G(\Delta;\eta_{\tilde{y}}) = 2(1-\eta_{\tilde{y}})(r+1)$, where $\eta_{\tilde{y}} \geqslant \frac{1}{2}$, for $r > 0$ (result by Eq. (49)). By Lemma C.6, we have $W_{\text{ACS}}(r^*_{\boldsymbol{\eta}},\boldsymbol{g}^*_{\boldsymbol{\eta}};\boldsymbol{\eta}) = 2(1-\eta^*)(r^*_{\boldsymbol{\eta}}+1) + \psi(r^*_{\boldsymbol{\eta}})$. Thus we have

$$\begin{aligned} \Delta W_{\text{ACS}}(r,\boldsymbol{g};\boldsymbol{\eta}) &= W_{\text{ACS}}(r,\boldsymbol{g};\boldsymbol{\eta}) - W_{\text{ACS}}(r^*_{\boldsymbol{\eta}},\boldsymbol{g}^*_{\boldsymbol{\eta}};\boldsymbol{\eta}) \\ &\geqslant G(r,\Delta;\eta_{\tilde{y}}) + ce^{-\alpha r} - \left(2(1-\eta^*)(r^*_{\boldsymbol{\eta}}+1) + \psi(r^*_{\boldsymbol{\eta}})\right) \\ &\qquad \left(\textstyle\sum_y \eta_y \max\{1-(g_y - \max_{y'\neq y} g_{y'} - r), 0\} \geqslant G(\Delta, r;\eta_{\tilde{y}}) \text{ in Eq. (47)}\right) \\ &\geqslant 2(1-\eta_{\tilde{y}})(r+1) + \psi(r) - \left(2(1-\eta^*)(r^*_{\boldsymbol{\eta}}+1) + \psi(r^*_{\boldsymbol{\eta}})\right) \\ &\quad \left(\inf_{\Delta\geqslant 0} G(r,\Delta;\eta_{\tilde{y}}) = 2(1-\eta_{\tilde{y}})(r+1) \text{ where } \eta_{\tilde{y}} > \tfrac{1}{2} \text{ for } r > 0 \text{ in Eq. (49)}\right) \\ &\geqslant 2(1-\eta_{\tilde{y}}) + \psi(0) - \left(2(1-\eta^*)(r^*_{\boldsymbol{\eta}}+1) + \psi(r^*_{\boldsymbol{\eta}})\right) \qquad (51) \\ &\geqslant 2(1-\eta_{\tilde{y}}) + \psi(0) - \left(2(1-\eta_{\tilde{y}})(r^*_{\eta_{\tilde{y}}}+1) + \psi(r^*_{\eta_{\tilde{y}}})\right) \\ &\geqslant -2(1-\eta_{\tilde{y}})r^*_{\eta_{\tilde{y}}} + \psi(0) - \psi(r^*_{\eta_{\tilde{y}}}) \\ &:= f_2(\eta_{\tilde{y}}), \end{aligned}$$

where in (51) we used Lemma C.7 (note that $2(1-\eta_{\tilde{y}}) \geqslant 2(1-\eta^*) \geqslant 2c = -\psi'(0)$).

For $\psi(r) = c\max\{1-\alpha r, 0\}$, $r^*_{\eta_{\tilde{y}}} = -1$. Then we have

$$\begin{aligned} f_2(\eta_{\tilde{y}}) &= -2(1-\eta_{\tilde{y}})\cdot(-1) + c - (\alpha+1)c \\ &= 2(1-\eta_{\tilde{y}}) + c - 3c \qquad (\alpha = 2 \text{ from Theorem 4.2}) \\ &= 2(1-c-\eta_{\tilde{y}}) \\ &= 2\Delta W_{\text{0-1-}c}(r,\boldsymbol{g};\boldsymbol{\eta}). \end{aligned}$$

Therefore,

$$\begin{aligned} R_{\text{0-1-}c}(r,\boldsymbol{g};\boldsymbol{\eta}) - R_{\text{0-1-}c}(r^*_{\boldsymbol{\eta}},\boldsymbol{g}^*_{\boldsymbol{\eta}};\boldsymbol{\eta}) &= \mathbb{E}\left[\Delta W_{\text{0-1-}c}(r,\boldsymbol{g};\boldsymbol{\eta})\right] \\ &\leqslant \mathbb{E}\left[\Gamma_2\left(\Delta W_{\text{ACS}}(r,\boldsymbol{g};\boldsymbol{\eta})\right)\right] \\ &\leqslant \Gamma_2\left(\mathbb{E}\left[\Delta W_{\text{ACS}}(r,\boldsymbol{g};\boldsymbol{\eta}]\right)\right) \quad (\Gamma_2 \text{ is concave}) \\ &\leqslant \Gamma_2\left(R_{\text{ACS}}(r,\boldsymbol{g};\boldsymbol{\eta}) - R_{\text{ACS}}(r^*_{\boldsymbol{\eta}},\boldsymbol{g}^*_{\boldsymbol{\eta}};\boldsymbol{\eta})\right), \end{aligned}$$

where $\Gamma_2(t) = \frac{1}{2}t$ for $\psi(r) = c\max\{1-\alpha r, 0\}$.

(**iii**) If $\eta^* \leqslant \frac{1}{2} < (1-c)$ and $r > 0$: In this case, by Eq. (30) we have $\Delta W_{\text{0-1-}c}(r,\boldsymbol{g};\boldsymbol{\eta}) = 1 - c - \eta_{\tilde{y}}$. Consider $\Delta W_{\text{ACS}}(r,\boldsymbol{g};\boldsymbol{\eta}) = W_{\text{ACS}}(r,\boldsymbol{g};\boldsymbol{\eta}) - W_{\text{ACS}}(r^*_{\boldsymbol{\eta}},\boldsymbol{g}^*_{\boldsymbol{\eta}};\boldsymbol{\eta})$, by Lemma C.6, we have $W_{\text{ACS}}(r^*_{\boldsymbol{\eta}},\boldsymbol{g}^*_{\boldsymbol{\eta}};\boldsymbol{\eta}) = (r^*_{\boldsymbol{\eta}}+1) + \psi(r^*_{\boldsymbol{\eta}})$ for $\eta^* \leqslant \frac{1}{2}$. Thus we have

$$
\begin{aligned}
\Delta W_{\mathrm{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) &= W_{\mathrm{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) - W_{\mathrm{ACS}}(r^*_{\boldsymbol{\eta}}, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta}) \\
&\geqslant G(r, \Delta; \eta_{\tilde{y}}) + ce^{-\alpha r} - \left((r^*_{\boldsymbol{\eta}} + 1) + \psi(r^*_{\boldsymbol{\eta}})\right) \\
&\quad \left(\textstyle\sum_y \eta_y \max\left\{1 - (g_y - \max_{y' \neq y} g_{y'} - r), 0\right\} \geqslant G(\Delta, r; \eta_{\tilde{y}}) \text{ in Eq. (47)}\right) \\
&\geqslant (r+1) + \psi(r) - \left((r^*_{\boldsymbol{\eta}} + 1) + \psi(r^*_{\boldsymbol{\eta}})\right) \\
&\qquad \left(\inf_{\Delta \geqslant 0} G(r, \Delta; \eta_{\tilde{y}}) = r + 1 \text{ for } \eta_{\tilde{y}} \leqslant \eta^* \leqslant \tfrac{1}{2}, r > 0 \text{ in (48)}\right) \\
&\geqslant 1 + \psi(0) - \left((r^*_{\boldsymbol{\eta}} + 1) + \psi(r^*_{\boldsymbol{\eta}})\right) \qquad (52) \\
&= \psi(0) - \left(r^*_{\boldsymbol{\eta}} + \psi(r^*_{\boldsymbol{\eta}})\right) \\
&:= f_3(c),
\end{aligned}
$$

where in (52), we used Lemma C.7 (note that $1 > 2c = -\psi'(0)$).

For $\psi(r) = c\max\{1 - \alpha r, 0\}$, $r^*_{\eta_{\tilde{y}}} = -1$. Then we have

$$
\begin{aligned}
f_3(c) &= c - (-1 + (\alpha + 1)c) \\
&= 1 - 2c && (\alpha = 2) \\
&= \frac{1 - 2c}{1 - c} 1 - c \\
&= W_c^{(2)}(1 - c) && (53) \\
&\geqslant W_c^{(2)}(1 - c - \eta_{\tilde{y}}) && (\eta_{\tilde{y}} \geqslant 0) \\
&= W_c^{(2)} \Delta W_{\text{0-1-}c}(r, \boldsymbol{g}; \boldsymbol{\eta}),
\end{aligned}
$$

where in (53), $W_c^{(2)} = 1 - 2c$ is a positive constant that depends on $c < 1/2$.

Therefore,

$$
\begin{aligned}
R_{\text{0-1-}c}(r, \boldsymbol{g}; \boldsymbol{\eta}) - R_{\text{0-1-}c}(r^*_{\boldsymbol{\eta}}, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta}) &= \mathbb{E}\left[\Delta W_{\text{0-1-}c}(r, \boldsymbol{g}; \boldsymbol{\eta})\right] \\
&\leqslant \mathbb{E}\left[\Gamma_3\left(\Delta W_{\mathrm{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta})\right)\right] \\
&\leqslant \Gamma_3\left(\mathbb{E}\left[\Delta W_{\mathrm{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta})\right]\right) \quad (\Gamma_3 \text{ is concave}) \\
&\leqslant \Gamma_3\left(R_{\mathrm{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) - R_{\mathrm{ACS}}(r^*_{\boldsymbol{\eta}}, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta})\right),
\end{aligned}
$$

where $\Gamma_3(t) = \frac{1}{W_c^{(1)}} t$ for $\psi(r) = c\max\{1 - \alpha r, 0\}$ ($W_c^{(1)}$, $W_c^{(2)}$ are positive constants that depend on c).

(**iv**) If $\eta^* \leqslant (1 - c)$ and $r \leqslant 0$: In this case, by Eq. (30) we have $\Delta W_{\text{0-1-}c}(r, \boldsymbol{g}; \boldsymbol{\eta}) = 0$, which implies that $W_{\text{0-1-}c}(r, \boldsymbol{g}; \boldsymbol{\eta}) - W_{\text{0-1-}c}(r^*_{\boldsymbol{\eta}}, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta}) = \Delta W_{\text{0-1-}c}(r, \boldsymbol{g}; \boldsymbol{\eta}) = 0 \leqslant \Gamma\left(W_{\mathrm{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) - W_{\mathrm{ACS}}(r^*_{\boldsymbol{\eta}}, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta})\right)$ for any $\Gamma \geqslant 0$.

(**v**) If $\eta^* > (1 - c)$ and $-1 \leqslant r \leqslant 0$: In this case, by Eq. (30) we have $\Delta W_{\text{0-1-}c}(r, \boldsymbol{g}; \boldsymbol{\eta}) = \eta^* - 1 + c$. Consider $\Delta W_{\mathrm{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) = W_{\mathrm{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) - W_{\mathrm{ACS}}(r^*_{\boldsymbol{\eta}}, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta})$, by Lemma C.6, we have $W_{\mathrm{ACS}}(r^*_{\boldsymbol{\eta}}, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta}) = 2(1 - \eta^*)(r^*_{\boldsymbol{\eta}} + 1) + \psi(r^*_{\boldsymbol{\eta}})$ for $\eta^* > (1 - c) > \frac{1}{2}$. Thus we have

$$
\begin{aligned}
\Delta W_{\mathrm{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) &= W_{\mathrm{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) - W_{\mathrm{ACS}}(r^*_{\boldsymbol{\eta}}, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta}) \\
&\geqslant 2(1 - \eta^*)(r + 1) + \psi(r) - \left(2(1 - \eta^*)(r^*_{\boldsymbol{\eta}} + 1) + \psi(r^*_{\boldsymbol{\eta}})\right) && \text{(result by Eq. (11))} \\
&\geqslant 2(1 - \eta^*) + \psi(0) - \left(2(1 - \eta^*)(r^*_{\boldsymbol{\eta}} + 1) + \psi(r^*_{\boldsymbol{\eta}})\right) && (54) \\
&= -2(1 - \eta^*) r^*_{\boldsymbol{\eta}} + \psi(0) - \psi(r^*_{\boldsymbol{\eta}}) \\
&:= f_5(\eta_{\tilde{y}}),
\end{aligned}
$$

where in (54), we used Lemma C.7 (note that $2(1 - \eta^*) < 2c = -\psi'(0)$).

For $\psi(r) = c\max\{1-\alpha r, 0\}$, $r^*_{\boldsymbol{\eta}} = \frac{1}{\alpha}$. Then we have

$$\begin{aligned} f_5(\eta^*) &= -2(1-\eta_{\tilde{y}})\cdot\frac{1}{\alpha} + c - 0 \\ &= \eta^* - 1 + c \qquad (\alpha = 2) \\ &= \Delta W_{\text{0-1-}c}(r, \boldsymbol{g}; \boldsymbol{\eta}). \end{aligned}$$

Therefore,

$$\begin{aligned} R_{\text{0-1-}c}(r, \boldsymbol{g}; \boldsymbol{\eta}) - R_{\text{0-1-}c}(r^*_{\boldsymbol{\eta}}, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta}) &= \mathbb{E}\left[\Delta W_{\text{0-1-}c}(r, \boldsymbol{g}; \boldsymbol{\eta})\right] \\ &\leqslant \mathbb{E}\left[\Gamma_5\left(\Delta W_{\text{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta})\right)\right] \\ &\leqslant \Gamma_5\left(\mathbb{E}\left[\Delta W_{\text{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta})\right]\right) \quad (\Gamma_5 \text{ is concave}) \\ &\leqslant \Gamma_5\left(R_{\text{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) - R_{\text{ACS}}(r^*_{\boldsymbol{\eta}}, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta})\right), \end{aligned}$$

where $\Gamma_5(t) = \sqrt{t}$ for $\psi(r) = c\max\{1-\alpha r, 0\}$.

(**vi**) If $\eta^* > (1-c)$ and $r < -1$: In this case, by Eq. (30) we have $\Delta W_{\text{0-1-}c}(r, \boldsymbol{g}; \boldsymbol{\eta}) = \eta^* - 1 + c$. For $\Delta W_{\text{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) = W_{\text{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) - W_{\text{ACS}}(r^*_{\boldsymbol{\eta}}, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta})$, we have

$$\begin{aligned} \Delta W_{\text{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) &= W_{\text{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) - W_{\text{ACS}}(r^*_{\boldsymbol{\eta}}, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta}) \\ &\geqslant \psi(r) - \left(2(1-\eta^*)(r^*_{\boldsymbol{\eta}} + 1) + \psi(r^*_{\boldsymbol{\eta}})\right) \qquad \text{(result by Eq. (11))} \\ &\geqslant \psi(-1) - \left(2(1-\eta^*)(r^*_{\boldsymbol{\eta}} + 1) + \psi(r^*_{\boldsymbol{\eta}})\right) \qquad (\psi(r) \text{ is nonincreasing for } r < -1) \\ &= \psi(-1) - 2(1-\eta^*)(r^*_{\boldsymbol{\eta}} + 1) - \psi(r^*_{\boldsymbol{\eta}}) \\ &:= f_6(\eta_{\tilde{y}}). \end{aligned}$$

For $\psi(r) = c\max\{1-\alpha r, 0\}$, $r^*_{\boldsymbol{\eta}} = \frac{1}{\alpha}$. Then we have

$$\begin{aligned} f_6(\eta^*) &= (\alpha+1)c - 2(1-\eta^*)(\frac{1}{\alpha} + 1) \\ &= 3c - 3(1-\eta^*) \qquad (\alpha = 2) \\ &= 3(\eta^* - 1 + c) \\ &= \Delta W_{\text{0-1-}c}(r, \boldsymbol{g}; \boldsymbol{\eta}). \end{aligned}$$

Therefore,

$$\begin{aligned} R_{\text{0-1-}c}(r, \boldsymbol{g}; \boldsymbol{\eta}) - R_{\text{0-1-}c}(r^*_{\boldsymbol{\eta}}, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta}) &= \mathbb{E}\left[\Delta W_{\text{0-1-}c}(r, \boldsymbol{g}; \boldsymbol{\eta})\right] \\ &\leqslant \mathbb{E}\left[\Gamma_6\left(\Delta W_{\text{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta})\right)\right] \\ &\leqslant \Gamma_6\left(\mathbb{E}\left[\Delta W_{\text{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta})\right]\right) \quad (\Gamma_6 \text{ is concave}) \\ &\leqslant \Gamma_6\left(R_{\text{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) - R_{\text{ACS}}(r^*_{\boldsymbol{\eta}}, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta})\right), \end{aligned}$$

where $\Gamma_6(t) = \frac{1}{2}t$ for $\psi(r) = c\max\{1-\alpha r, 0\}$.

Overall, we obtain

$$R_{\text{0-1-}c}(r, \boldsymbol{g}; \boldsymbol{\eta}) - R_{\text{0-1-}c}(r^*_{\boldsymbol{\eta}}, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta}) \leqslant \Gamma\left(R_{\text{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) - R_{\text{ACS}}(r^*_{\boldsymbol{\eta}}, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta})\right),$$

where $\Gamma(t) = \max\{t, \frac{1}{W_c^{(2)}}t\} = O(t)$ for $\psi(r) = c\max\{1-\alpha r, 0\}$ ($W_c^{(1)}$, $W_c^{(2)}$ are positive constants that depends on c). □

D.5 Proof of Theorem 6.1

Theorem 6.1 (restated) *Define* $M_\phi = \sup_{x\in\mathcal{X},g_y\in\mathcal{G}} \phi(g_y(\boldsymbol{x}))$, $M_{\psi_1}(r) = \sup_{x\in\mathcal{X},r\in\mathcal{R}} \psi_1(r(\boldsymbol{x}))$, $M_{\psi_2}(r) = \sup_{x\in\mathcal{X},r\in\mathcal{R}} \psi_2(r(\boldsymbol{x}))$, *and* $M_\psi(r) = \sup_{x\in\mathcal{X},r\in\mathcal{R}} \psi(r(\boldsymbol{x}))$. *Let* $\mathfrak{R}_n(\mathcal{G}), \mathfrak{R}_n(\mathcal{R})$ *be the Rademacher complexity of* $\mathcal{G}, \mathcal{R}$ *for data of size* n *drawn from* $p(\boldsymbol{x})$. *Then for any* $\delta \in (0,1)$, *with probability at least* $1-\delta$, *the following multi-class classification generalization bounds hold for all* $\boldsymbol{g} \in \mathfrak{G}$:

$$
\begin{aligned}
R_{\mathrm{MCS}}(\boldsymbol{g}) \leqslant & \widehat{R}_{\mathrm{MCS}}(\boldsymbol{g}) + \left(2M_\phi + M_{\psi_1}\right) L_\phi K^2 \mathfrak{R}_n(\mathcal{G}) \\
& + \left(\left(2M_\phi + M_{\psi_1}\right) L_{\psi_1} + 2L_{\psi_2}\right) \mathfrak{R}_n(\mathcal{R}) \\
& + \left(M_\phi M_{\psi_1} + M_{\psi_2}\right) \sqrt{\frac{1}{2n} \log\left(\frac{1}{\delta}\right)}, \\
R_{\mathrm{ACS}}(\boldsymbol{g}) \leqslant & \widehat{R}_{\mathrm{ACS}}(\boldsymbol{g}) + 2L_\phi K^2 \mathfrak{R}_n(\mathcal{G}) \\
& + 2\left(L_\phi + L_\psi\right) \mathfrak{R}_n(\mathcal{R}) \\
& + \left(M_{\psi_1} + M_{\psi_2}\right) \sqrt{\frac{1}{2n} \log\left(\frac{1}{\delta}\right)}.
\end{aligned}
$$

Proof. Firstly, we proof the bound for $R_{\mathrm{MCS}}(\boldsymbol{g})$.

To begin with, we first bound the change of $\sup_{g_1\cdots,g_K\in\mathcal{G}}(R_{\mathrm{MCS}}(\boldsymbol{g}) - \widehat{R}_{\mathrm{MCS}}(\boldsymbol{g}))$ when a single entry $z_i = (\boldsymbol{x}_i, y_i)$ of $(z_1, \cdots, z_i, \cdots, z_n)$ is replaced with $z_i' = (\boldsymbol{x}_i', y_i')$. Define $A(z_1, \cdots, z_n) = \sup_{g_1\cdots,g_K\in\mathcal{G}}(R_{\mathrm{MCS}}(\boldsymbol{g}) - \widehat{R}_{\mathrm{MCS}}(\boldsymbol{g}))$. Then it holds that

$$
\begin{aligned}
& A(z_1, \cdots, z_i, \cdots, z_n) - A(z_1, \cdots, z_i', \cdots, z_n) \\
& = \sup_{r\in\mathcal{R},\boldsymbol{g}\in\mathfrak{G}} \left[\mathbb{E}_{p(\boldsymbol{x},y)}\left[L_{\mathrm{MCS}}(r,\boldsymbol{g},\boldsymbol{x},y)\right] - \frac{1}{n}\sum_{j=1}^n L_{\mathrm{MCS}}(r,\boldsymbol{g};\boldsymbol{x}_j,y_j)\right] \\
& \quad - \sup_{r'\in\mathcal{R},\boldsymbol{g}'\in\mathfrak{G}} \left[\mathbb{E}_{p(\boldsymbol{x},y)}\left[L_{\mathrm{MCS}}(r',\boldsymbol{g}',\boldsymbol{x},y)\right] - \frac{1}{n} L_{\mathrm{MCS}}(r',\boldsymbol{g}';\boldsymbol{x}_i',y_i') - \frac{1}{n}\sum_{j\in[n]\setminus\{i\}} L_{\mathrm{MCS}}(r',\boldsymbol{g}';\boldsymbol{x}_j,y_j)\right] \\
& = \sup_{r\in\mathcal{R},\boldsymbol{g}\in\mathfrak{G}} \inf_{r'\in\mathcal{R},\boldsymbol{g}'\in\mathfrak{G}} \left[\mathbb{E}_{p(\boldsymbol{x},y)}\left[L_{\mathrm{MCS}}(r,\boldsymbol{g},\boldsymbol{x},y)\right] - \frac{1}{n}\sum_{j=1}^n L_{\mathrm{MCS}}(r,\boldsymbol{g};\boldsymbol{x}_j,y_j)\right. \\
& \quad \left. -\mathbb{E}_{p(\boldsymbol{x},y)}\left[L_{\mathrm{MCS}}(r',\boldsymbol{g}',\boldsymbol{x},y)\right] + \frac{1}{n} L_{\mathrm{MCS}}(r',\boldsymbol{g}';\boldsymbol{x}_i',y_i') + \frac{1}{n}\sum_{j\in[n]\setminus\{i\}} L_{\mathrm{MCS}}(r',\boldsymbol{g}';\boldsymbol{x}_j,y_j)\right] \\
& \leqslant \sup_{r\in\mathcal{R},\boldsymbol{g}\in\mathfrak{G}} \left[\mathbb{E}_{p(\boldsymbol{x},y)}\left[L_{\mathrm{MCS}}(r,\boldsymbol{g},\boldsymbol{x},y)\right] - \frac{1}{n}\sum_{j=1}^n L_{\mathrm{MCS}}(r,\boldsymbol{g};\boldsymbol{x}_j,y_j)\right. \\
& \quad \left. -\mathbb{E}_{p(\boldsymbol{x},y)}\left[L_{\mathrm{MCS}}(r,\boldsymbol{g},\boldsymbol{x},y)\right] + \frac{1}{n} L_{\mathrm{MCS}}(r,\boldsymbol{g};\boldsymbol{x}_i',y_i') + \frac{1}{n}\sum_{j\in[n]\setminus\{i\}} L_{\mathrm{MCS}}(r,\boldsymbol{g};\boldsymbol{x}_j,y_j)\right] \\
& = \sup_{r\in\mathcal{R},\boldsymbol{g}\in\mathfrak{G}} \left[\frac{1}{n} L_{\mathrm{MCS}}(r,\boldsymbol{g};\boldsymbol{x}_i',y_i') - \frac{1}{n} L_{\mathrm{MCS}}(r,\boldsymbol{g};\boldsymbol{x}_i,y_i)\right] \\
& = \frac{1}{n} \sup_{r\in\mathcal{R},g_1,\cdots,g_K\in\mathcal{G}} \left[\phi\left(g_{y_i'}(\boldsymbol{x}_i') - \max_{y''\neq y_i'} g_{y''}(\boldsymbol{x}_i')\right)\psi_1(r(\boldsymbol{x}_i')) + \psi_2(r(\boldsymbol{x}_i'))\right. \\
& \quad \left. -\phi\left(g_{y_i}(\boldsymbol{x}_i) - \max_{y'\neq y_i} g_{y'}(\boldsymbol{x}_i)\right)\psi_1(r(\boldsymbol{x}_i)) - \psi_2(r(\boldsymbol{x}_i))\right] \\
& \leqslant \frac{M_\phi M_{\psi_1} + M_{\psi_2}}{n},
\end{aligned}
$$

where $M_\phi = \sup_{x\in\mathcal{X},g_y\in\mathcal{G}} \phi(g_y(\boldsymbol{x}))$, $M_{\psi_1}(r) = \sup_{x\in\mathcal{X},r\in\mathcal{R}} \psi_1(r(\boldsymbol{x}))$, $M_{\psi_2}(r) = \sup_{x\in\mathcal{X},r\in\mathcal{R}} \psi_2(r(\boldsymbol{x}))$.

In the same way, we can get $A(z_1, \cdots, z_i', \cdots, z_n) - A(z_1, \cdots, z_i, \cdots, z_n) \geqslant -\frac{M_\phi M_{\psi_1} + M_{\psi_2}}{n}$. Then $|A(z_1, \cdots, z_i, \cdots, z_n) - A(z_1, \cdots, z_i', \cdots, z_n)| \leqslant c = \frac{M_\phi M_{\psi_1} + M_{\psi_2}}{n}$. Thus, for any $\delta \in (0,1)$, with probability at least $1-\delta$ we have

$$
\begin{aligned}
&\sup_{g_1 \cdots, g_K \in \mathcal{G}} (R_{\mathrm{ACS}}(\boldsymbol{g}) - \widehat{R}_{\mathrm{ACS}}(\boldsymbol{g})) \\
&\leqslant \mathbb{E}_S \left[\sup_{g_1 \cdots, g_K \in \mathcal{G}} (R_{\mathrm{ACS}}(\boldsymbol{g}) - \widehat{R}_{\mathrm{ACS}}(\boldsymbol{g})) \right] + c \sqrt{\frac{n}{2} \log\left(\frac{1}{\delta}\right)} \\
&\qquad\qquad \text{(McDiarmid inequality Shalev-Shwartz and Ben-David (2014))} \\
&\leqslant \mathbb{E}_S \left[\sup_{g_1 \cdots, g_K \in \mathcal{G}} (R_{\mathrm{ACS}}(\boldsymbol{g}) - \widehat{R}_{\mathrm{ACS}}(\boldsymbol{g})) \right] + (M_\phi + M_\psi) \sqrt{\frac{1}{2n} \log\left(\frac{1}{\delta}\right)} \qquad \left(c = \tfrac{M_\phi + M_\psi}{n}\right) \\
&\leqslant 2\mathfrak{R}_n(\mathcal{L}_{\mathrm{ACS}}) + (M_\phi + M_\psi) \sqrt{\frac{1}{2n} \log\left(\frac{1}{\delta}\right)} \\
&\qquad\qquad \text{(symmetrization Shalev-Shwartz and Ben-David (2014), } \mathcal{L}_{\mathrm{ACS}} \text{ is defined in Eq. (26))} \\
&\leqslant 2L_\phi K^2 \mathfrak{R}_n(\mathcal{G}) + 2\left(L_\phi + L_\psi\right) \mathfrak{R}_n(\mathcal{R}) \\
&+ (M_\phi + M_\psi) \sqrt{\frac{1}{2n} \log\left(\frac{1}{\delta}\right)}. \qquad \text{(result by Lemma C.16)}
\end{aligned}
$$

Overall, we obtain for any $\delta \in (0,1)$, with probability at least $1-\delta$, the following multi-class classification generalization bound holds for all $\boldsymbol{g} \in \mathfrak{G}$:

$$
\begin{aligned}
R_{\mathrm{MCS}}(\boldsymbol{g}) \leqslant \widehat{R}_{\mathrm{MCS}}(\boldsymbol{g}) &+ \left(2M_\phi + M_{\psi_1}\right) L_\phi K^2 \mathfrak{R}_n(\mathcal{G}) + \left(\left(2M_\phi + M_{\psi_1}\right) L_{\psi_1} + 2L_{\psi_2}\right) \mathfrak{R}_n(\mathcal{R}), \\
&+ \left(M_\phi M_{\psi_1} + M_{\psi_2}\right) \sqrt{\frac{1}{2n} \log\left(\frac{1}{\delta}\right)},
\end{aligned}
$$

where L_ϕ, L_{ψ_1}, and L_{ψ_2} are Lipschitz constants, and $M_\phi = \sup_{x \in \mathcal{X}, g_y \in \mathcal{G}} \phi(g_y(\boldsymbol{x}))$, $M_{\psi_1}(r) = \sup_{x \in \mathcal{X}, r \in \mathcal{R}} \psi_1(r(\boldsymbol{x}))$, $M_{\psi_2}(r) = \sup_{x \in \mathcal{X}, r \in \mathcal{R}} \psi_2(r(\boldsymbol{x}))$.

Next, we proof the bound for $R_{\mathrm{ACS}}(\boldsymbol{g})$.

To begin with, we first bound the change of $\sup_{g_1 \cdots, g_K \in \mathcal{G}} (R_{\mathrm{ACS}}(\boldsymbol{g}) - \widehat{R}_{\mathrm{ACS}}(\boldsymbol{g}))$ when a single entry $z_i = (\boldsymbol{x}_i, y_i)$ of $(z_1, \cdots, z_i, \cdots, z_n)$ is replaced with $z_i' = (\boldsymbol{x}_i', y_i')$. Define $A(z_1, \cdots, z_i, \cdots, z_n) = \sup_{g_1 \cdots, g_K \in \mathcal{G}} (R_{\mathrm{ACS}}(\boldsymbol{g}) - \widehat{R}_{\mathrm{ACS}}(\boldsymbol{g}))$. By the same argument as previous with L_{MCS} replaced with L_{ACS}, we have

$$
\begin{aligned}
&A(z_1, \cdots, z_i, \cdots, z_n) - A(z_1, \cdots, z_i', \cdots, z_n) \\
&= \sup_{r \in \mathcal{R}, \boldsymbol{g} \in \mathfrak{G}} \left[\frac{1}{n} L_{\mathrm{ACS}}(r, \boldsymbol{g}; \boldsymbol{x}_i', y_i') - \frac{1}{n} L_{\mathrm{ACS}}(r, \boldsymbol{g}; \boldsymbol{x}_i, y_i) \right].
\end{aligned} \tag{55}
$$

Then we have

$$
\begin{aligned}
&A(z_1, \cdots, z_i, \cdots, z_n) - A(z_1, \cdots, z_i', \cdots, z_n) \\
&= \frac{1}{n} \sup_{r \in \mathcal{R}, g_1, \cdots, g_K \in \mathcal{G}} \left[\phi\left(g_{y_i'}(\boldsymbol{x}_i') - \max_{y'' \neq y_i'} g_{y''}(\boldsymbol{x}_i') + r(\boldsymbol{x}_i') \right) \right) + \psi(r(\boldsymbol{x}_i')) \\
&\quad - \phi\left(g_{y_i}(\boldsymbol{x}_i) - \max_{y' \neq y_i} g_{y'}(\boldsymbol{x}_i) - r(\boldsymbol{x}_i) \right) - \psi(r(\boldsymbol{x}_i)) \Big] \\
&\leqslant \frac{M_\phi + M_\psi}{n},
\end{aligned}
$$

where $M_\phi = \sup_{x \in \mathcal{X}, g_y \in \mathcal{G}} \phi(g_y(\boldsymbol{x}))$, $M_\psi(r) = \sup_{x \in \mathcal{X}, r \in \mathcal{R}} \psi(r(\boldsymbol{x}))$.

In the same way, we can get $A(z_1, \cdots, z_i', \cdots, z_n) - A(z_1, \cdots, z_i, \cdots, z_n) \geqslant -\frac{M_\phi + M_{\psi_2}}{n}$. Then $|A(z_1, \cdots, z_i, \cdots, z_n) - A(z_1, \cdots, z_i', \cdots, z_n)| \leqslant c = \frac{M_\phi + M_{\psi_2}}{n}$. Thus, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ we have

$$
\begin{aligned}
&\sup_{g_1 \cdots, g_K \in \mathcal{G}} (R_{\mathrm{ACS}}(\boldsymbol{g}) - \widehat{R}_{\mathrm{ACS}}(\boldsymbol{g})) \\
&\leqslant \mathbb{E}_S \left[\sup_{g_1 \cdots, g_K \in \mathcal{G}} (R_{\mathrm{ACS}}(\boldsymbol{g}) - \widehat{R}_{\mathrm{ACS}}(\boldsymbol{g})) \right] + c \sqrt{\frac{n}{2} \log\left(\frac{1}{\delta}\right)} \\
&\qquad \text{(McDiarmid inequality Shalev-Shwartz and Ben-David (2014))} \\
&\leqslant \mathbb{E}_S \left[\sup_{g_1 \cdots, g_K \in \mathcal{G}} (R_{\mathrm{ACS}}(\boldsymbol{g}) - \widehat{R}_{\mathrm{ACS}}(\boldsymbol{g})) \right] + (M_\phi + M_\psi) \sqrt{\frac{1}{2n} \log\left(\frac{1}{\delta}\right)} \qquad \left(c = \tfrac{M_\phi + M_\psi}{n}\right) \\
&\leqslant 2\mathfrak{R}_n(\mathcal{L}_{\mathrm{ACS}}) + (M_\phi + M_\psi) \sqrt{\frac{1}{2n} \log\left(\frac{1}{\delta}\right)} \\
&\qquad \text{(symmetrization Shalev-Shwartz and Ben-David (2014), } \mathcal{L}_{\mathrm{ACS}} \text{ is defined in Eq. (26))} \\
&\leqslant 2L_\phi K^2 \mathfrak{R}_n(\mathcal{G}) + 2\left(L_\phi + L_\psi\right) \mathfrak{R}_n(\mathcal{R}) \\
&+ (M_\phi + M_\psi) \sqrt{\frac{1}{2n} \log\left(\frac{1}{\delta}\right)}. \qquad \text{(result by Lemma C.16)}
\end{aligned}
$$

Overall, we obtain for any $\delta \in (0, 1)$, with probability at least $1 - \delta$, the following multi-class classification generalization bound holds for all $\boldsymbol{g} \in \mathfrak{G}$:

$$
\begin{aligned}
R_{\mathrm{ACS}}(\boldsymbol{g}) \leqslant \widehat{R}_{\mathrm{ACS}}(\boldsymbol{g}) &+ 2L_\phi K^2 \mathfrak{R}_n(\mathcal{G}) + 2\left(L_\phi + L_\psi\right) \mathfrak{R}_n(\mathcal{R}), \\
&+ (M_\phi + M_\psi) \sqrt{\frac{1}{2n} \log\left(\frac{1}{\delta}\right)},
\end{aligned}
$$

where L_ϕ, L_ψ are Lipschitz constants, and $M_\phi = \sup_{x \in \mathcal{X}, g_y \in \mathcal{G}} \phi(g_y(\boldsymbol{x}))$, $M_\psi(r) = \sup_{x \in \mathcal{X}, r \in \mathcal{R}} \psi_2(r(\boldsymbol{x}))$. □

E Experiment Details

Setup. For all the datasets, we first split 80% of the samples to form the training set, and then split 20% of the samples from the training set to form the validation set. Specially, for Mao et al. 2024a (two stage), we split non-overlapped datasets for both stages and train for 200 epochs for each non-overlapped dataset. We adopt ResNet-32 (He et al., 2016), a 32-layer deep residual network with ReLU activations for CIFAR-10, and adopt a one-hidden-layer neural network with a ReLU activation for vehicle, satimage, and *yeast* datasets. We train for 200 epochs using Adam method (Kingma and Ba, 2015). During the training, we adopt the learning rate decay strategy, using an initial learning rate of 0.01 and a learning rate decay rate of 0.9. For vehicle, satimage, and *yeast* datasets, the batch size is set to 128, and for CIFAR-10, the batch size is set to 1024. We use a weight decay strategy, and the weight decay rate is 1×10^{-4}. We decay the learning rate after 10 epochs. Specifically, for Mao et al. 2024a (two-stage), we split non-overlapped datasets for each stage, and train for 200 epochs for each part. We adopt the following data augmentation strategy for training: 4 pixels padding on each side and a 32×32 random crop sampled from the padded image or its horizontal flip. We set the abstention cost c to 0.2 for the vehicle dataset, 0.4 for the *yeast* dataset, 0.1 for CIFAR-10, and the satimage dataset. We choose these values of c because they are not too far from the best misclassification error obtained by vanilla CE/CS with no abstention, which can encourage a reasonable number of samples to be abstained. The GPU for our experiments is NVIDIA RTX A6000.

Performance on three metrics. The performances of our method and comparison methods on three metrics on vehicle, satimage, and yeast datasets from UCI Machine Learning Repository are shown in Table 2–4. The best results among all the methods with abstention are boldfaced.

Table 2: Performance of our method and comparison methods on the vehicle dataset.

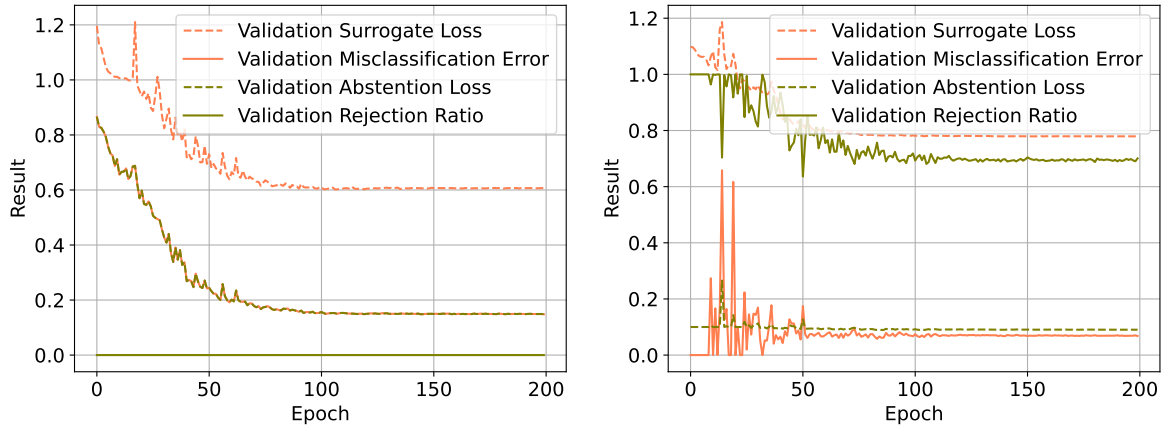
Method	Misclassification error	Rejection ratio	Abstention loss
CE (no rej.)	19.41%	0.00%	19.41%
CS (no rej.)	20.00%	0.00%	20.00%
Ramaswamy et al. (2018)	17.37%	**1.76%**	17.41%
Mao et al. (2024a) (two-stage)	7.37%	44.12%	12.94%
Ours (MCS)	5.71%	38.24%	**11.18** %
Ours (ACS)	**0.00%**	89.41%	17.88%

Table 3: Performance of our method and comparison methods on the satimage dataset.

Method	Misclassification error	Rejection ratio	Abstention loss
CE (no rej.)	10.33%	0.00%	10.33%
CS (no rej.)	10.49%	0.00%	10.49%
Ramaswamy et al. (2018)	9.51%	**1.94%**	9.52%
Mao et al. (2024a) (two stage)	3.19%	43.90%	6.18%
Ours (MCS)	4.18%	31.16%	**5.99**%
Ours (ACS)	**0.23**%	66.74%	6.75%

We can observe from Table 2–4 that our MCS and ACS perform better than vanilla CS loss on misclassification error and abstention loss on UCL datasets, which validates the effectiveness of machanism of abstention. Besides, as shown in Table 2–4, our MCS loss shows the best performance of abstention loss, and our ACS loss shows the best performance of misclassification error among all the comparison methods on UCL datasets, respectively.

Training dynamics. The training dynamics of CS and our MCS loss on the CIFAR-10 dataset are shown in Figure 2. For the case of CS, as shown in Figure 2(a), with the decreasing of validation surrogate loss, validation abstention loss also tends to be decreasing. Since the rejection ratio is 0 in this setting, the validation abstention loss equals to the validation misclassification error. For the case of MCS, as shown in Figure 2(b), with the decreasing of validation surrogate loss, validation abstention loss decreases sharply at early epochs and tends to keep a comparatively low level in the later epochs, which validates the calibration of our MCS loss. The



(a) Performance of CS (no rejection). (b) Performance of our method (MCS) ($c = 0.1$, $\alpha = 6.0$).

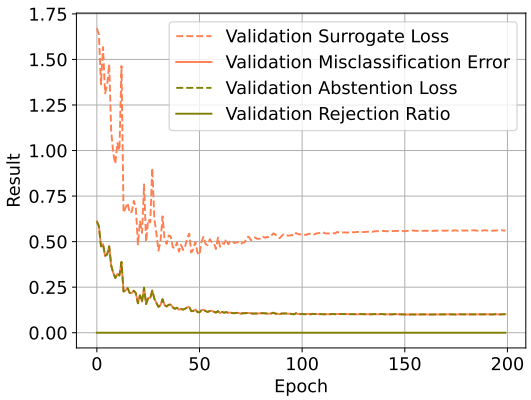
Figure 2: The training dynamics of CS and our method.

training dynamics of CE, Ramaswamy et al. 2018, Mao et al. 2024a, and ours (ACS) on CIFAR-10 dataset are shown in Figure 3. For the case of CE, as shown in Figure 3(a), with the decreasing of validation surrogate loss, validation abstention loss also tends to be decreasing. And because for this case the rejection ratio is 0, the validation abstention loss equals to the validation misclassification error. For the case of Ramaswamy et al. 2018,

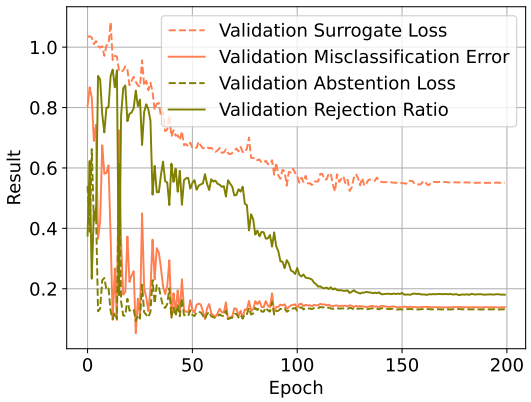
Table 4: Performance of our method and comparison methods on the yeast dataset.

Method	Misclassification error	Rejection ratio	Abstention loss
CE (no rej.)	40.40%	0.00%	40.40%
CS (no rej.)	40.40%	0.00%	40.40%
Ramaswamy et al. (2018)	39.37%	**3.37%**	39.39%
Mao et al. (2024a) (two stage)	27.59%	70.71%	36.36%
Ours (MCS)	29.55%	55.56%	**35.69%**
Ours (ACS)	**22.22%**	96.97%	39.46%

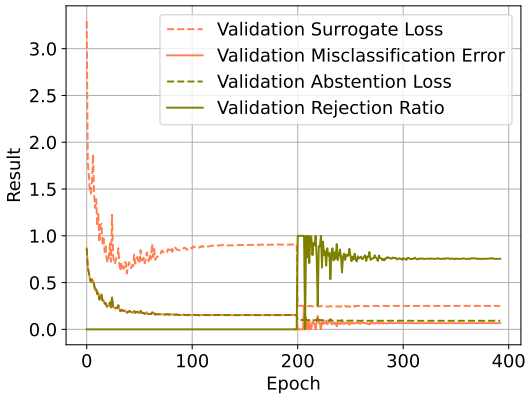
as shown in Figure 3(b), with the decreasing of validation surrogate loss, validation abstention loss also tends to be decreasing. Besides, as shown in Figure 3(b), validation abstention loss can be regarded as an integrating metric of validation misclassification error and validation rejection ratio. For the case of Mao et al. 2024a, as shown in Figure 3(c), at the first stage the validation rejection ratio is 0, whereas it increases sharply at the second stage and validation surrogate loss decreases sharply at the second stage. For our ACS loss, as shown in Figure 3(d), with the decreasing of validation surrogate loss, validation abstention loss decreases sharply at early epochs and tends to keep a comparatively low level in the later epochs, which validates the calibration of our ACS loss.



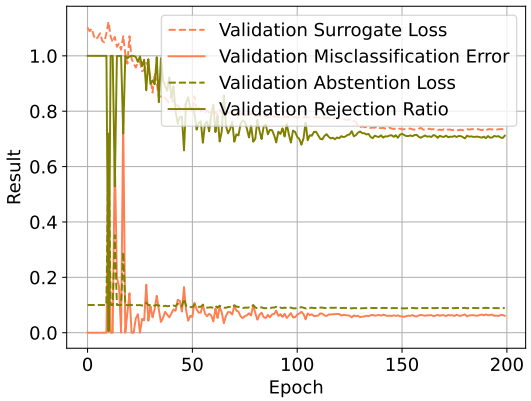
(a) Performance of CE (no rejection).



(b) Performance of Ramaswamy et al. 2018 ($c = 0.1$).



(c) Performance of Mao et al. 2024a ($c = 0.1$, $\alpha = 1.0$).

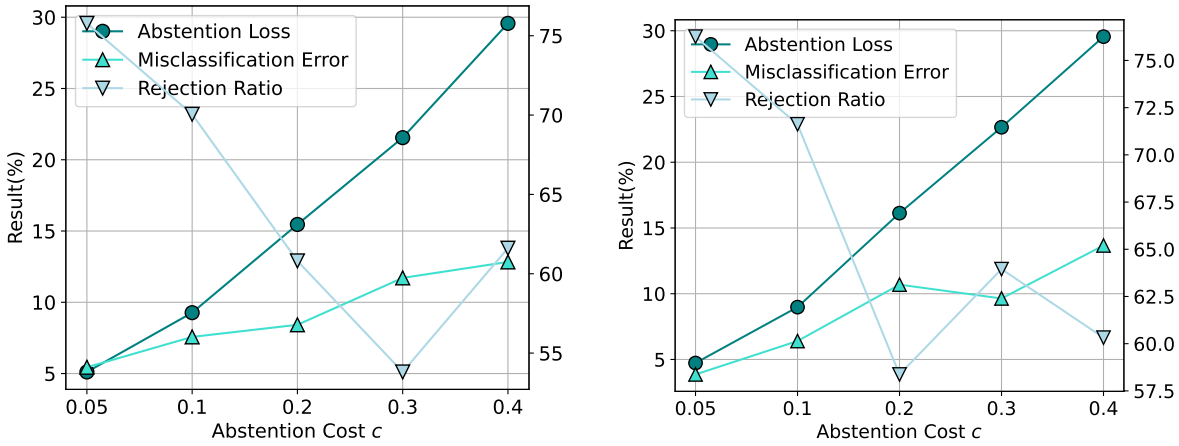


(d) Performance of ours (ACS) ($c = 0.1$, $\alpha = 2.0$).

Figure 3: The training dynamics of CE, Ramaswamy et al. 2018, Mao et al. 2024a, and ours (ACS) on CIFAR-10 dataset.

Impact of abstention cost c. The impacts of abstention cost c for our MCS loss and ACS loss on CIFAR-10 dataset are shown in Figure 4. As shown in Figure 4(a), for our MCS loss, with the increase in abstention cost c,

the abstention loss and misclassification error tend to increase, which validates the sensitivity of the abstention cost c. Our ACS loss also shows a similar behaviour, as shown in Figure 4(b).



(a) Impact of abstention cost c for MCS loss.

(b) Impact of abstention cost c for ACS loss.

Figure 4: The impact of abstention cost c for our method.