TOWARDS CALIBRATION FOR GENERAL CRAMMER–SINGER LOSSES IN MULTI-CLASS CLASSIFICATION WITH ABSTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

The Crammer–Singer (CS) loss has been a standard multi-class extension of base binary surrogate losses. Despite its solid empirical performance, classical theory of loss functions proves that the CS loss is not classification-calibrated to the 0-1 loss, that is, the minimization of the surrogate risk of the CS loss does not necessarily lead to the minimization of the 0-1 risk. The pitfall of the CS loss is caused by low-confidence samples with their maximum class probability less than $1/2$. This observation leads to the following question: is the CS loss calibrated when there are no low-confidence input samples? To this end, we consider the CS loss under multi-class classification *with abstention*, and in particular, the predictor-rejector framework, where a learner trains a class predictor as well as a rejector to abstain from making predictions for low-confidence samples. Under this setup, we show that general CS losses can be calibrated, verifying our hypothesis that the CS loss can be safely used even theoretically if all input samples are sufficiently confident. Theoretically, our calibration proof is interesting on its own right because it is applicable to broader losses composed of binary ones; the CS loss under classification with abstention has been considered previously, but the calibration property has been known only for polyhedral losses. Experiments on different datasets empirically validate the effectiveness of our calibrated general CS losses[1].

1 INTRODUCTION

Convex surrogate losses play a key role in classification problems due to the intractability of primary classification losses (Bartlett et al., 2006). For example, in multi-class classification, the logistic loss (known as cross-entropy loss in the multi-class case) and the Crammer–Singer (CS) loss (Crammer & Singer, 2001) have been classical instances of convex surrogate losses. Although the cross-entropy (CE) loss is more dominant in deep learning, the CS loss is also practical in scenarios where the number of classes is not too large because of its superior accuracy and noise robustness (Zhu et al., 2023). The CS loss is a family of multi-class surrogate losses that maximize the multi-class margin via binary losses. Despite its solid empirical performance, the CS loss is *not classification-calibrated* (Zhang, 2004a), that is, the minimization of the surrogate risk of CS loss does not necessarily lead to the minimization of the 0-1 risk[2]. This has been considered to be a minimally sufficient condition in the loss function theory, and many common examples, including the logistic loss, satisfy it (Zhang, 2004b). Thus, an alternative theory is necessary to characterize the superiority of the CS loss.

A crucial observation is that the CS loss would be calibrated if there were no low-confidence samples whose maximum class probability is no larger than $1/2$ (Zhang, 2004a). Ramaswamy et al. (2018) pointed it out and demonstrates that the CS loss is calibrated under multi-class classification *with abstention*, where a learner is allowed to abstain from making predictions on low-confidence samples (Ni et al., 2019; Mao et al., 2024a). Specifically, Ramaswamy et al. (2018) uses a predictor trained with the CS loss to reject its low-confidence samples. This formulation is often called a

[1]The Code is available at: https://anonymous.4open.science/r/Cal-Gen-CS-anonymized/.

[2]Note that the calibration we focus on in this paper is different from the notation of probability calibration, which is a way to assess and improve the output probabilities of the predictor (Silva Filho et al., 2023)

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

confidence-based method for classification with abstention (Chow, 1957; Bartlett & Wegkamp, 2008; Charoenphakdee et al., 2021). See Appendix A for a more thorough survey of related frameworks of classification with abstention.

Nevertheless, their positive result is applicable only to the CS loss generated by the binary hinge loss; this is because their calibration analysis relies on the so-called embedding framework (Finocchiaro et al., 2020), which assumes non-smooth polyhedral losses. In modern gradient-based learning, smooth surrogate losses typically yield well-behaved optimization trajectories. As displayed in Figure 1, the calibrated CS loss generated by the logistic loss could perform better in some settings with comparitively higher abstention cost, which can also be observed from Appendix E. Therefore, it remains a crucial challenge to demonstrate that the CS loss generated by smooth binary losses is calibrated.

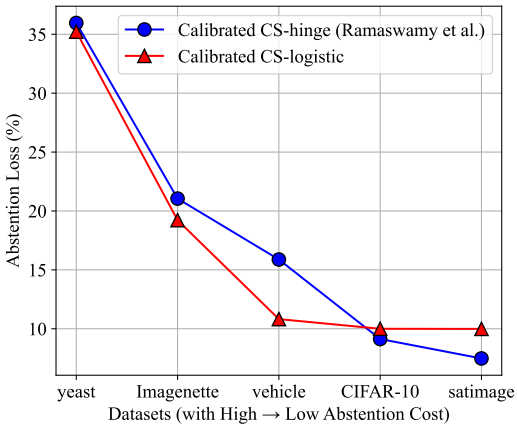


Figure 1: Empirical performance of the calibrated CS loss generated by the hinge loss (Ramaswamy et al., 2018), and the calibrated one generated by the logistic loss.

In this paper, we show that the CS loss under *predictor-rejector* framework is calibrated with broader binary losses. Under the predicotr-rejector framework, we train a predictor and rejector separately, and the rejector decides whether to predict or reject the input sample (Ni et al., 2019; Mao et al., 2024a), unlike confidence-based methods. In this way, we expand the applicability of the CS loss family to smooth loss functions, which are more friendly to gradient-based learning. Our contributions can be summarized as follows:

- We provide a new insight into calibrating general CS losses via the predictor-rejector framework. Under this framework, we show that general CS losses, including non-polyhedral losses, can be calibrated, verifying that the CS loss can be safely used even theoretically if all input samples are sufficiently confident.
- We prove excess error bounds for our calibrated general CS losses, showing that the surrogate excess risk controls the target excess risk at a linear rate for certain choices of losses, which is the first result under the predictor-rejector framework to the best of our knowledge.
- Experiments on different datasets empirically validate the effectiveness of our calibrated general CS losses.

2 Problem Setup

In this section, we formulate the predictor-rejector multi-class classification with abstention framework. Consider a supervised 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 probability 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 modeled separately. Let $g_y : \mathcal{X} \to \mathbb{R}$ be a score function for multi-class classification, and $\boldsymbol{g}(\boldsymbol{x}) = (g_1(x), \cdots, g_K(\boldsymbol{x}))^\top$. If $r(\boldsymbol{x}) \leqslant 0$, the predictor chooses to abstain. Otherwise, it returns $\arg\max_{y \in \mathcal{Y}} g_y(\boldsymbol{x})$ as the predicted label[3]. Let $c \in [0, 1]$ be the abstention cost. Then the target abstention loss is expressed in the following form (Bartlett & Wegkamp, 2008; Cortes et al., 2016a;

[3]Here we assume that the tie-breaking rule is fixed and $\arg\max_{y \in \mathcal{Y}} g_y(\boldsymbol{x})$ denotes the index of a maximizer rather than the set of maximizers for $g_1, g_2, \ldots, g_K$.

Ni et al., 2019):

$$L_{0\text{-}1\text{-}c}(r, \boldsymbol{g}; \boldsymbol{x}, y) = \underbrace{\mathbb{I}_{\arg\max_{y \in \mathcal{Y}} g_y(\boldsymbol{x}) \neq y} \mathbb{I}_{r(\boldsymbol{x})>0}}_{\text{not to abstain}} + c \underbrace{\mathbb{I}_{r(\boldsymbol{x}) \leqslant 0}}_{\text{abstain}}, \tag{1}$$

where $\mathbb{I}_{[\cdot]}$ denotes the indicator function. In the following part, we assume the 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 $L_{0\text{-}1\text{-}c}$, we define the target abstention risk $R_{0\text{-}1\text{-}c}$ by

$$R_{0\text{-}1\text{-}c}(r, \boldsymbol{g}) = \mathbb{E}_{p(\boldsymbol{x},y)} \left[L_{0\text{-}1\text{-}c}(r, \boldsymbol{g}; \boldsymbol{x}, y) \right], \tag{2}$$

and denote the infimum of $R_{0\text{-}1\text{-}c}$ taken over all measurable functions as $R^*_{0\text{-}1\text{-}c} = \inf_{\boldsymbol{g},r} R_{0\text{-}1\text{-}c}(r, \boldsymbol{g})$.

Since it is intractable to optimize the target abstention risk $R_{0\text{-}1\text{-}c}$, we instead introduce a surrogate loss that is easier to optimize. Denote the surrogate loss for $L_{0\text{-}1\text{-}c}$ as L, which is a continuous function bounding L from above. Similarly to the case of $R_{0\text{-}1\text{-}c}$ in Equation (2), we also define the surrogate risk R by

$$R(r, \boldsymbol{g}) = \mathbb{E}_{p(\boldsymbol{x},y)} \left[L(r, \boldsymbol{g}; \boldsymbol{x}, y) \right], \tag{3}$$

and denote the infimum of R taken over all measurable functions as $R^* = \inf_{\boldsymbol{g},r} R(r, \boldsymbol{g})$.

3 Calibration and Excess Risk Bound

The notions of calibration and an excess risk bound known in classification problems are also extended to 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 the minimization of the surrogate risk also leads to the minimization of the target risk. Following Ni et al. (2019), we can directly define calibration on a $(r, \boldsymbol{g})$ pair (defined in Section 2) by whether it can lead to Bayes-optimal target risk. A Bayes-optimal predictor and rejector (Cortes et al., 2016a; Ni et al., 2019) given probability distribution p satisfy:

$$\boldsymbol{g}^*(\boldsymbol{x}) \in \arg\max_y p(y \mid \boldsymbol{x}), \quad \operatorname{sign}\left(r^*(\boldsymbol{x})\right) = \operatorname{sign}\left(\max_y p(y \mid \boldsymbol{x}) - 1 + c \right)$$

for all $\boldsymbol{x} \in \mathcal{X}$. Then given probability distribution p, $(r, \boldsymbol{g})$ is calibrated if and only if there exists $(r^*, \boldsymbol{g}^*)$ such that

$$R_{0\text{-}1\text{-}c}(r, \boldsymbol{g}) = R_{0\text{-}1\text{-}c}(r^*, \boldsymbol{g}^*).$$

We can also define calibration on r and $\boldsymbol{g}$ separately by whether they can perform equivalently to some r^* and $\boldsymbol{g}^*$ following Ni et al. (2019). Given probability distribution p, r and $\boldsymbol{g}$ are calibrated separetely if and only if there exists r^* and $\boldsymbol{g}^*$ such that

$$\begin{aligned} &\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{g}^*(\boldsymbol{x}) \text{ almost everywhere on } \mathcal{X}, \\ &\operatorname{sign}\left(r(\boldsymbol{x})\right) = \operatorname{sign}\left(r^*(\boldsymbol{x})\right) \text{ for all } \boldsymbol{x} \in \mathcal{X} \text{ such that } r^*(\boldsymbol{x}) \neq 0. \end{aligned}$$

As stated in Section 2.2 in Ni et al. (2019), for the calibration of $(r, \boldsymbol{g})$, the calibration of r (which we call *rejection-calibration*) is necessary, while that of $\boldsymbol{g}$ is not. An intuitive explanation is that, when an instance is rejected, there is no longer a need to correctly predict its label.

Rejection-calibration of surrogate loss L can be defined on pointwise risk in a similar way to Ramaswamy & Agarwal (2016). Then we define the pointwise risk W of the surrogate loss L at $\boldsymbol{x}$ by

$$W\left(r, \boldsymbol{g}; \boldsymbol{\eta}\right) = \sum_y \eta_y(\boldsymbol{x}) L(r, \boldsymbol{g}; \boldsymbol{x}, y), \tag{4}$$

where $\boldsymbol{\eta}(\boldsymbol{x}) = \left(\eta_1(\boldsymbol{x}), \cdots, \eta_K(\boldsymbol{x})\right)^\top$ is class probability vector for $\eta_y(\boldsymbol{x}) = p(y|\boldsymbol{x})$, which lives in the probability simplex $\Delta_K = \left\{ \boldsymbol{\eta} \in \mathbb{R}^K_+ \mid \sum_{y=1}^K \eta_y = 1 \right\}$. For simplicity, we omit the notation of $\boldsymbol{x}$ and use $W\left(r, \boldsymbol{g}; \boldsymbol{\eta}\right)$ to denote $W\left(r(\boldsymbol{x}), \boldsymbol{g}(\boldsymbol{x}); \boldsymbol{\eta}(\boldsymbol{x})\right)$ for the pointwise risk, and use η_y to denote

$\eta_y(\boldsymbol{x})$ for the class probability. Then L is rejection-calibrated if and only if for any $\left(r^\dagger_{\boldsymbol{\eta}}, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}\right) \in \operatorname{argmin}_{r,\boldsymbol{g}} W(r, \boldsymbol{g}; \boldsymbol{\eta})$,

$$\operatorname{sign}(r^\dagger_{\boldsymbol{\eta}}) = \operatorname{sign}(\eta^* - 1 + c), \tag{5}$$

for all $\boldsymbol{\eta} \in \Delta_K$ such that $\eta^* \neq 1 - c$, where $\eta^* = \max_y \eta_y$.

An *excess risk bound* (Zhang, 2004a; Bartlett et al., 2006) is an inequality that upper-bounds the excess risk of target abstention loss $L_{0\text{-}1\text{-}c}$ with respect to the excess risk of its surrogate loss L, 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 $(r, \boldsymbol{g})$ and probability distribution p, where $\Gamma(\cdot)$ is a non-decreasing function satisfying $\Gamma(0) = 0$, which we call *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 abstention loss, and $R(r, \boldsymbol{g}) - R^*$ is the suboptimality of the surrogate loss.

4 General Calibrated Crammer–Singer Losses

In this section, we present our proposed general calibrated Crammer–Singer (CS) losses via the predictor-rejector framework, which can be seen as a family of surrogate losses for the target loss defined in Equation (1). Our proposed losses can be divided into multiplicative CS (MCS) and additive CS (ACS) losses. They enjoy better optimization properties than the previous ones for multi-class classification with abstention in the predictor-rejector framework (Mao et al., 2024a), which are non-convex.

4.1 Standard Crammer–Singer Loss

The CS loss is a multi-class loss, yet its form is close to binary classification losses, which leads to a special property available for calibration. It can be calibrated by rejecting low-confidence data with the maximum class probability smaller than $1/2$. The CS loss tries to maximize the prediction margin, which is defined by the difference between the output score of the label class and the highest prediction score except the label class. The general CS loss is defined as

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

where $g_y(\boldsymbol{x})$ is the prediction score of label class, and $\phi \geqslant 0$ is a non-increasing convex function bounding $t \mapsto \mathbb{I}_{t \leqslant 0}$ from above.

4.2 Multiplicative and Additive Crammer–Singer Losses

Our proposed calibrated losses can be divided into two forms: the multiplicative CS (MCS) and additive CS (ACS) losses. 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. Similarly, ψ_1 and ψ_2 are non-decreasing and non-increasing convex functions bounding $t \mapsto \mathbb{I}_{t > 0}$ and $t \mapsto \mathbb{I}_{t \leqslant 0}$ from above, respectively. We assume that ψ_1 and ψ_2 are differentiable at $r = 0$, and $\phi'(0) < 0, \psi_1'(0) > 0$. Typical choices of function ϕ are as follows:

- exponential function: $\phi(\Delta) = \exp(-\Delta)$;
- hinge function: $\phi(\Delta) = \max\{0, 1 - \Delta\}$;
- logistic function: $\phi(\Delta) = \log\left(1 + \exp(-\Delta)\right)$.

The choices of functions ψ_1 and ψ_2 are similar to thos. Although the MCS loss is not always jointly convex in $(\boldsymbol{g}, r)$, it is convex with respect to r and $\boldsymbol{g}$ separately, that is, biconvex. Besides, we will reveal that its calibration condition does not depend on a specific choice of ϕ.

The ACS loss is of the following form:

$$L_{\mathrm{ACS}}(r, \boldsymbol{g}; \boldsymbol{x}, y) = \phi\left(g_y(\boldsymbol{x}) - \max_{y' \neq y} g_{y'}(\boldsymbol{x}) - r(\boldsymbol{x})\right) + \psi(r(\boldsymbol{x})). \tag{6}$$

Here, $\phi \geqslant 0$ is a non-increasing convex function bounding $t \mapsto \mathbb{I}_{t \leqslant 0}$ from above. Similarly, ψ is a non-increasing convex function bounding $t \mapsto \mathbb{I}_{t \leqslant 0}$ from above. We assume that ψ is differentiable at $r = 0$. We will reveal that the calibration condition for the ACS loss depends on a specific choice of ϕ. However, the loss is always convex jointly in $(\boldsymbol{g}, r)$, which is a beneficial complement to the MCS loss (the proof is shown in Appendix B.1).

5 Calibration Conditions for Our Losses

In this section, we denote by

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

a minimizer of pointwise surrogate loss $W(r, \boldsymbol{g}; \boldsymbol{\eta})$ defined in Section 3. Before presenting calibration conditions for the MCS and ACS losses, we first give a modification of the calibration condition shown in Ni et al. (2019) so that the indifferentiable case is also covered and the treatment of the tie is formalized, which was slightly informal in the original proof.

Proposition 5.1. *(adapted from Ni et al., 2019, Theorem 4) Assume that L is a convex function with respect to r. Let $\partial W\left(r, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta}\right)$ be the subdifferential of $W\left(r, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta}\right)$. Then L is rejection-calibrated if and only if for any $\left(r^\dagger_{\boldsymbol{\eta}}, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}\right) \in \operatorname{argmin}_{r, \boldsymbol{g}} W(r, \boldsymbol{g}; \boldsymbol{\eta})$,*

$$\sup_{\boldsymbol{\eta}: \eta^* > 1-c} \sup_{s_{\boldsymbol{\eta}} \in \partial W\left(r, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta}\right)\big|_{r=0}} s_{\boldsymbol{\eta}} \leqslant 0 \leqslant \inf_{\boldsymbol{\eta}: \eta^* < 1-c} \inf_{s_{\boldsymbol{\eta}} \in \partial W\left(r, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta}\right)\big|_{r=0}} s_{\boldsymbol{\eta}},$$

and

$$0 \notin \partial W\left(r, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta}\right)\big|_{r=0}$$

for all $\boldsymbol{\eta} \in \Delta_K$ such that $\eta^* \neq 1 - c$, where $\eta^* = \max_y \eta_y$.

The proof of this proposition is shown in Appendix B.3. Ni et al. (2019) conjecture that the above unnatural constraint for a surrogate loss to be calibrated is hard to satisfy in the multi-class case and resort to confidence-based methods. Mao et al. (2024a) address this problem by using nonconvex surrogates. We will reveal that our losses can satisfy this constraint, that is, $W(r, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta})$ only depends on η^*, rather than the class probabilities of other classes. Next, we derive a useful result about the expression of $W(r, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta})$ for a general form of L which covers both of our proposed losses.

Proposition 5.2. *Consider surrogate losses of the following form:*

$$L(r, \boldsymbol{g}; \boldsymbol{x}, y) = \Phi\left(g_y(\boldsymbol{x}) - \max_{y' \neq y} g_{y'}(\boldsymbol{x}), r(\boldsymbol{x})\right),$$

where $\Phi(\Delta, r) : \mathbb{R}^2 \to \mathbb{R}$ is a function monotonically non-increasing in Δ. There exists $\boldsymbol{g}^\dagger_{\boldsymbol{\eta}} = \left(g^\dagger_1, g^\dagger_2, \ldots, g^\dagger_K\right)$ such that $g^\dagger_y = g^\dagger_{y^*} - \Delta$ for some $\Delta \geqslant 0$ and all $y \neq y^*$, where $y^* \in \operatorname{argmax}_y \eta_y$. Furthermore, for any $\boldsymbol{g}^\dagger_{\boldsymbol{\eta}}$ we can express $W(r, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta}) = h_{\eta^*}(r)$ 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 $0 \leqslant \eta \leqslant 1$.

This proposition is useful in the proof of calibration conditions and excess risk bounds for our proposed losses. The proof of this proposition is shown in Appendix B.2. Due to the special property of the CS loss such that it solely depends on the prediction margin Δ, we can partly reduce the analysis to the one close to the binary case. Next, we derive the conditions for the MCS and ACS losses to be calibrated. In the following part, unless stated explicitly, the assumptions of ψ_1, ψ_2, and ψ coincide with those in Section 4.2.

Theorem 5.3. *L_{MCS} is rejection-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 $0 \leqslant \eta \leqslant 1$.

The proof of this theorem is shown in Appendix D.1. During the proof, we use the non-increasing property of H_η and Proposition 5.1. By this theorem, we can see that the calibration condition of the MCS loss is related to H_{1-c}, and the condition is not limited to a specified ϕ, which reveals the advantages of the MCS loss. The specific forms of H_η for different ϕ are shown in Appendix C.1. Furthermore, we have

$$H_{1-c} = \begin{cases} 2\sqrt{c(1-c)} & \text{if } \phi(\Delta) = \exp(-\Delta), \\ 2c & \text{if } \phi(\Delta) = \max\{1-\Delta, 0\}, \\ -(1-c)\log(1-c) - c\log c & \text{if } \phi(\Delta) = \log(1+\exp(-\Delta)). \end{cases} \tag{7}$$

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 by Theorem 5.3. For other choices of ϕ, unlike the MCS loss, we do not have a general calibration result for the ACS loss because of the dependence of $\phi(\Delta - r)$ on r. However, we can still show that the ACS loss with hinge or logistic ϕ and generic ψ can be calibrated, as follows:

Theorem 5.4. *L_{ACS} with $\phi(x) = \max\{1-x, 0\}$ is rejection-calibrated if and only if*

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

The proof of this theorem is shown in Appendix D.2. During the proof, a case-by-case analysis and Proposition 5.1 are needed. 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 5.5. *Assume that ψ is a twice-differentiable decreasing function. Then L_{ACS} with $\phi(x) = \log(1+\exp(-x))$ is rejection-calibrated if and only if*

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

The proof of this theorem is shown in Appendix D.3. During the proof, a series of derivative analysis techniques are used. Compared with existing calibration results (Cortes et al., 2016b;a), we can take various ϕ and ψ, while they require specific forms of the losses. Although the calibration results for the ACS loss are a bit limited, the choices of ϕ still cover standard ones. Theorems 5.3 to 5.5 aim to indicate that for specific ϕ, ψ_1, ψ_2 or ϕ, ψ, we can determine the parameters of ψ_2 or ψ to calibrate L_{MCS} or L_{ACS} according to a given value of c. This point has been reflected in the experimental part.

6 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 the 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 6.1. *For L_{MCS}, the following excess risk bound holds for all $r, \boldsymbol{g}$ and any distribution $p(\boldsymbol{x}, y)$:*

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

where

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

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

The proof of this theorem is shown in Appendix D.4. During the proof, the infimum of $W_{\text{0-1-}c}$ and pointwise surrogate risk W_{MCS} are considered, and discussions of different cases of η^* and r are required. In our analysis, we use several techniques to bound the difference between the pointwise surrogate risk and its infimum in various cases of η^* and r, such as bounding based on the property of the complicated power function resulting from the calibrated MCS loss.

Remark 6.2. The techniques used in the proof are different from those in the proof for Mao et al. (2024a, Theorem 2), mainly because the infimum of the pointwise risk in Theorem 6.1 cannot be simply expressed as a linear function of $1 - \eta^*$ (as shown in Appendix C.2). Besides, our bound does not depend on the number of classes K, which is also different from Mao et al. (2024a, Theorem 2). This is because we can partly reduce the problem to one close to the binary case during the proof, as stated in Proposition 5.2.

Theorem 6.3. *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 $r, \boldsymbol{g}$ *and any distribution* $p(\boldsymbol{x}, y)$*, where*

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

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

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

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

The proof of this theorem is shown in Appendix D.5. During the proof, the infinums of pointwise target risk $W_{\text{0-1-}c}$ and pointwise surrogate risk W_{ACS} are used. Besides, similar to the case of L_{MCS}, the proof also requires discussion of different η^* and r. A difference is that in the proof of Theorem 6.3, due to the inherent property of hinge loss, we need more analysis for the pointwise minimum of piecewise linear functions. For L_{ACS} with hinge ϕ and ψ, a linear excess risk rate of $O(t)$ can be reached, which is better than that of $O\left(\max\left\{t, \sqrt{t}\right\}\right)$ shown in Mao et al. (2024a, Theorem 2) (where $\sqrt{t}$ becomes dominant when $t \to 0$). This means that the target excess risk can be controlled more tightly by the surrogate excess risk and is believed to be optimal as long as Γ is data-dependent (see Cao et al., 2025, Section 1).

7 Generalization Error Bounds

In this section, we give generalization error bounds for the MCS and ACS losses. Although these bounds can be proved by standard techniques and are not the main contribution of this paper, they still support the validity of the proposed losses. During the proof, we use the concept of the Rademacher complexity (Mohri et al., 2018; Bartlett et al., 2005). 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 $r : \mathcal{X} \to \mathbb{R}$. Denote the empirical surrogate risk of L as

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

Assume that 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 the above condition.

Definition 7.1. (Rademacher Complexity) Let $S = \{z_1, \cdots, z_n\}$ be a sample set 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

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

Table 1: Settings for the four variants of our methods.

METHOD	ϕ	ψ	ψ_1	ψ_2
MCS-HINGE	$\max\{0, 1-\Delta\}$	–	$\max\{0, 1-r\}$	$c\max\{0, 1-\alpha r\}$
ACS-HINGE	$\max\{0, 1-\Delta\}$	$c\max\{0, 1-\alpha r\}$	–	–
MCS-LOGISTIC	$\log(1+\exp(-\Delta))$	–	$\log(1+\exp(-r))$	$c\log(1+\exp(-\alpha r))$
ACS-LOGISTIC	$\log(1+\exp(-\Delta))$	$c\log(1+\exp(-\alpha r))$	–	–

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

Theorem 7.2 (Generalization Error Bounds for L_{MCS} and L_{ACS})**.** *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* $p(\boldsymbol{x}, y)$*, respectively. Then for any* $\delta \in (0, 1)$*, with probability at least* $1-\delta$*, the following multi-class classification generalization bounds hold for all* $\boldsymbol{g} \in \mathfrak{G}$*:*

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

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

The proof of Theorem 7.2 is shown in Appendix D.6. Theorem 7.2 ensures the generalization error bounds of the MCS and ACS losses, which indicate that when the empirical surrogate risks are minimized, the expected surrogate risks are also minimized under infinitely large data and ($\mathcal{G}$, $\mathcal{R}$) with asymptotically vanishing Rademacher complexities.

8 EXPERIMENTS

In this section, we present experimental results for our losses in comparison with existing methods on the vehicle dataset, which is from the UCI Machine Learning Repository (Bache & Lichman, 2013), and Imagenette (Howard, 2019), which is a smaller subset of ImageNet (Russakovsky et al., 2015). Note that due to the inherent attributes of CS loss, that is, it only considers the output score of the label class and the highest one except the label class, it exhibits superior classification performance when the number of classes is not too large (as discussed in Zhu et al., 2023). Thus, to provide a reasonable practical scenario for CS losses, we adopt datasets with a moderate number of classes ($\leqslant 10$) to validate our losses. We also give results for CIFAR-10 (Krizhevsky & Hinton, 2009), satimage and yeast datasets in Appendix E. Note that the latter two are both from the UCI Machine Learning Repository; they and the vehicle dataset are also used in Ramaswamy et al. (2018).

To validate the effectiveness of our calibrated general CS losses, we compare four variants of the proposed losses with the vanilla CE loss and the vanilla CS loss. As mentioned in Appendix A, the score-based formulation (Mozannar & Sontag, 2020) and selective classification (El-Yaniv & Wiener, 2010) are also classical paradigms of classification with abstention, while they might fail to achieve calibration. Confidence-based methods could calibrate the CS loss in the scope of polyhedral losses. Thus, we also with calibrated polyhedral losses via confidence-based methods, e.g., the calibrated CS loss generated by the hinge loss and the binary encoded predictions (BEP) loss presented in Ramaswamy et al. (2018). Besides, Mao et al. (2024a) is also compared to exhibit the superiority of our method within the predictor-rejector formulation. The settings for four variants of our methods are displayed in Table 1, where α is a hyperparameter depending on c, which is an environment setting given externally. In experiments, we choose c around the best misclassification error with no abstention to moderately predict and reject (Mao et al., 2024a, Appendix E). For both the vehicle dataset and Imagenette we set c to 0.2. More details on experiments are in Appendix E.

Metrics. We use the abstention loss $L_{0\text{-}1\text{-}c}$ for the predictor-rejector method as an evaluation metric. $L_{0\text{-}1\text{-}c}$ can be interpreted as the combination of two parts: the zero-one misclassification error when a

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

.

METHOD	REJECTION RATIO	MISCLASSIFICATION ERROR	ABSTENTION LOSS
CE (NO REJ.)	0.00%	19.41%	19.41%
CS (NO REJ.)	0.00%	20.00%	20.00%
RAMASWAMY ET AL. (2018) (CS-HINGE)	8.82%	15.48%	15.88%
RAMASWAMY ET AL. (2018) (BEP)	1.76%	23.95%	23.88%
MAO ET AL. (2024A) (TWO-STAGE)	44.12%	7.37%	12.94%
OURS (MCS-HINGE)	38.24%	5.71%	11.18 %
OURS (ACS-HINGE)	89.41%	0.00%	17.88%
OURS (MCS-LOGISTIC)	25.88%	10.32%	12.82 %
OURS (ACS-LOGISTIC)	39.41%	4.85%	**10.82**%

Table 3: Performance of our method and the compared methods on Imagenette.

METHOD	REJECTION RATIO	MISCLASSIFICATION ERROR	ABSTENTION LOSS
CE (NO REJ.)	0.00%	21.99%	21.99%
CS (NO REJ.)	0.00%	25.68%	25.68%
RAMASWAMY ET AL. (2018) (CS-HINGE)	39.19%	21.73%	21.05%
RAMASWAMY ET AL. (2018) (BEP)	8.62%	21.00%	20.91%
MAO ET AL. (2024A) (TWO STAGE)	89.88%	15.13%	19.51%
OURS (MCS-HINGE)	12.69%	22.74%	22.40%
OURS (ACS-HINGE)	88.88%	7.05%	**18.56**%
OURS (MCS-LOGISTIC)	4.78%	22.46%	22.34%
OURS (ACS-LOGISTIC)	45.69%	18.56%	19.22%

sample is accepted and the abstention cost when a sample is rejected. Therefore, the abstention loss can be seen as integrating the information of zero-one misclassification error on the accepted samples and the rejection ratio, which provides a comprehensive comparison between different methods.

Performance on three metrics. The performance of our method and the compared methods on the above three metrics on the vehicle dataset and Imagenette are displayed in Tables 2 and 3, respectively. Note that misclassification error and rejection ratio are supplementary information, and the main target is the abstention loss. It is possible that lower abstention losses are accompanied by higher rejection ratios, which do not necessarily mean that related methods are overly rejecting, but rather might imply that others were overconfident. The best abstention losses among all methods with abstention are boldfaced. Firstly, we can observe from Tables 2 and 3 that our MCS and ACS losses perform better than the vanilla CS loss on misclassification error and abstention loss on the vehicle dataset and Imagenette, which validates the effectiveness of our calibration strategy via abstention. Besides, as displayed in Tables 2 and 3, our ACS-logistic loss and ACS-hinge loss exhibit the best performance in misclassification error and abstention loss among all compared methods with abstention on the vehicle dataset and Imagenette, respectively, which validates the superiority of calibrating general CS losses via the predictor-rejector framework. On Imagenette, logistic variants can also perform competitively with hinge variants, which reveals the potential of non-polyhedral CS losses to be calibrated. We recommend choosing logistic variants in the setting of comparatively higher ($\geqslant 0.2$) c (e.g., the vehicle dataset) while choosing the hinge variants otherwise. That is because we speculate that in the case of comparatively lower c (e.g., CIFAR-10 with $c = 0.1$), they might be underconfident. Detailed discussions can be found in Appendix E), where the performances of our method and the compared methods on CIFAR-10, satimage, and yeast datasets are examined.

9 CONCLUSION

In this paper, we provided a new insight into calibrating general CS losses via the predictor-rejector framework. Under this framework, we showed that general CS losses, including non-polyhedral losses, are calibrated, verifying that the CS loss can be safely used even theoretically if all input samples are sufficiently confident. Experiments on different datasets empirically validated the effectiveness of our calibrated general CS losses.

AI USE STATEMENT

In this work, we used generative AI tools for text correcting/improving and source/searching for information. We have not used generative AI tools for thereotical models developing, synthetic dataset generating, code creating and so on. We have reviewed all AI-assisted work. For example, we checked the text correcting/improving results returned by AI and referenced credible sources with the assistance of the retrieved results of AI. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have attached a link to an anonymous GitHub repository in the footnote of the abstract; we give clear statement for the assumptions needed in the paper, which is included in the main body and the appendix; we give detailed proof for our claims, which is included in the appendix; the setting for hyperparameters can be found in the appendix and the code repository; the description for the data processing steps can be found in the code repository.

REFERENCES

Kevin Bache and Moshe Lichman. UCI machine learning repository, 2013.

Peter L Bartlett and Marten H Wegkamp. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9(8), 2008.

Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005.

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

Jonathan Borwein and Adrian Lewis. *Convex Analysis and Nonlinear Optimization: Theory and Examples*. Springer, 2006.

Yuzhou Cao, Tianchi Cai, Lei Feng, Lihong Gu, Jinjie Gu, Bo An, Gang Niu, and Masashi Sugiyama. Generalizing consistent multi-class classification with rejection to be compatible with arbitrary losses. *Advances in Neural Information Processing Systems*, 35:521–534, 2022.

Yuzhou Cao, Han Bao, Lei Feng, and Bo An. Establishing linear surrogate regret bounds for convex smooth losses via convolutional fenchel-young losses. *arXiv preprint arXiv:2505.09432*, 2025.

Nontawat Charoenphakdee, Zhenghang Cui, Yivan Zhang, and Masashi Sugiyama. Classification with rejection based on cost-sensitive classification. In *International Conference on Machine Learning*, pp. 1507–1517. PMLR, 2021.

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

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

Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Boosting with abstention. *Advances in Neural Information Processing Systems*, 29:1668–1676, 2016a.

Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Learning with rejection. In *International Conference on Algorithmic Learning Theory*, pp. 67–82. Springer, 2016b.

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

Giulia DeSalvo, Mehryar Mohri, and Umar Syed. Learning with deep cascades. In *International Conference on Algorithmic Learning Theory*, pp. 254–269. Springer, 2015.

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

Jessie Finocchiaro, Rafael Frongillo, and Bo Waggoner. Embedding dimension of polyhedral losses. In *Conference on Learning Theory*, pp. 1558–1585. PMLR, 2020.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

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

Jeremy Howard. Imagenette: A smaller subset of 10 easily classified classes from Imagenet, March 2019. URL `https://github.com/fastai/imagenette`.

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

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

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Predictor-rejector multi-class abstention: Theoretical analysis and algorithms. In *International Conference on Algorithmic Learning Theory*, pp. 822–867. PMLR, 2024a.

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

Christopher Mohri, Daniel Andor, Eunsol Choi, and Michael Collins. Learning to reject with a fixed predictor: Application to decontextualization. In *International Conference on Learning Representations*, 2024.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT press, 2018.

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

Chenri Ni, Nontawat Charoenphakdee, Junya Honda, and Masashi Sugiyama. On the calibration of multiclass classification with rejection. *Advances in Neural Information Processing Systems*, 32, 2019.

Harish G Ramaswamy and Shivani Agarwal. Convex calibration dimension for multiclass loss matrices. *Journal of Machine Learning Research*, 17(14):1–45, 2016.

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

R Tyrrell Rockafellar. *Convex Analysis*, volume 28. Princeton university press, 1997.

Ralph Rockafellar. On the maximal monotonicity of subdifferential mappings. *Pacific Journal of Mathematics*, 33(1):209–216, 1970.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

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

Telmo Silva Filho, Hao Song, Miquel Perello-Nieto, Raul Santos-Rodriguez, Meelis Kull, and Peter Flach. Classifier calibration: a survey on how to assess and improve predicted class probabilities. *Machine Learning*, 112(9):3211–3260, 2023.

Alexander Soen, Hisham Husain, Philip Schulz, and Vu 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.

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

Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85, 2004b.

Dixian Zhu, Yiming Ying, and Tianbao Yang. Label distributionally robust losses for multi-class classification: Consistency, robustness and adaptivity. In *International Conference on Machine Learning*, pp. 43289–43325. PMLR, 2023.

A Detailed Related Work

Classification with abstention (Chow, 1970; Herbei & Wegkamp, 2006; Cortes et al., 2016b; Ramaswamy et al., 2018; Ni et al., 2019; Mao et al., 2024a; Soen et al., 2024) refers to a variant of a classification problem where a learner can abstain from making a prediction. In classification with abstention, the abstention from prediction incurs a pre-defined cost $c \in (0, 1)$. Otherwise, the standard 0-1 loss is incurred. Classification 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 a lot of attention (Cortes et al., 2016b; Ramaswamy et al., 2018; Ni et al., 2019; Mao et al., 2024a; Soen et al., 2024). Currently, methods of classification with abstention can be categorized into different groups based on their learning paradigm.

The main idea of confidence-based methods (Chow, 1957; 1970; Ramaswamy et al., 2018; Bartlett & Wegkamp, 2008; Ni et al., 2019) 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 & 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. Charoenphakdee et al. (2021) construct a cost-sensitive surrogate loss for the multi-class case.

Selective classification (El-Yaniv & Wiener, 2010; Wiener & El-Yaniv, 2015; Geifman & 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 & Wiener (2010) start the research on selective classification, in which they study the trade-off between the coverage and the accuracy of classifiers with a rejection option.Wiener & El-Yaniv (2015) study selective classification for specific hypothesis classes and families of distributions. Geifman & 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 & 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 & 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 it is calibrated. Mao et al. (2024b) introduce new families of surrogate losses with theoretical guarantees for score-based formulation.

Note that in the above paradigms, both the score-based formulation and the predictor-rejector formulation explicitly consider the abstention cost c in the learning process, and assign a loss of c when the learner chooses to abstain. However, the score-based formulation might fail to reach the Bayes-optimal solution unless resorting to more complex functions (Mao et al., 2024a, Section 3). Besides, Selective classification considers another objective other than Bayes-optimality. Thus, we choose the predictor-rejector framework to calibrate the general CS loss in this paper.

B Useful Propositions

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

B.1 Proof of Proposition B.1

Proposition B.1. *The ACS loss of the form in Equation* (6) *is convex jointly in* $(\boldsymbol{g}, r)$.

Proof. Recall that the ACS loss can be expressed in 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 non-increasing convex function bounding $t \mapsto \mathbb{I}_{t \leqslant 0}$ from above.

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

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

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

B.2 Proof of Proposition 5.2

Proposition 5.2 (restated) *Given surrogate losses of the following form:*

$$L(r, \boldsymbol{g}; \boldsymbol{x}, y) = \Phi\left(g_y(\boldsymbol{x}) - \max_{y' \neq y} g_{y'}(\boldsymbol{x}), r(\boldsymbol{x})\right),$$

where $\Phi(\Delta, r) : \mathbb{R}^2 \to \mathbb{R}$ is a function monotonically non-increasing in Δ. There exists $\boldsymbol{g}^\dagger_{\boldsymbol{\eta}} = \left(g^\dagger_1, g^\dagger_2, \ldots, g^\dagger_K\right)$ such that $g^\dagger_y = g^\dagger_{y^*} - \Delta$ for some $\Delta \geqslant 0$ and all $y \neq y^*$, where $y^* \in \operatorname{argmax}_y \eta_y$. Furthermore, for any $\boldsymbol{g}^\dagger_{\boldsymbol{\eta}}$ we can express $W(r, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta}) = h_{\eta^*}(r)$ 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 $0 \leqslant \eta \leqslant 1$.

Proof. Let $g^\dagger_{(1)} \geqslant g^\dagger_{(2)} \geqslant \cdots \geqslant g^\dagger_{(K)}$ be the sorted sequence of $g^\dagger_{\boldsymbol{\eta}} = (g^\dagger_1, g^\dagger_2, \ldots, g^\dagger_K)$. Here letting $g_{(i+1)} := g_{(i)}$ $(i = 2, \ldots, K-1)$ never increases the objective function Equation (4) and there exists $\boldsymbol{g}^\dagger_{\boldsymbol{\eta}}$ such that $g^\dagger_y$ is the same for $y \neq \tilde{y}^\dagger$, where $\tilde{y}^\dagger = \operatorname{argmax}_y g^\dagger_y$. Then $W(r, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta})$ can be expressed as

$$W(r, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta}) = \sum_y \eta_y \Phi\left(g^\dagger_y - \max_{y' \neq y} g^\dagger_{y'}, r\right)$$

$$
\begin{aligned}
&= \eta_{\tilde{y}^\dagger} \Phi\left(g^\dagger_{(1)} - g^\dagger_{(2)}, r\right) + \sum_{y \neq \tilde{y}^\dagger} \eta_y \Phi\left(g^\dagger_{(2)} - g^\dagger_{(1)}, r\right) \\
&= \eta_{\tilde{y}^\dagger} \Phi\left(\underbrace{g^\dagger_{(1)} - g^\dagger_{(2)}}_{\Delta^\dagger \geqslant 0}, r\right) + (1 - \eta_{\tilde{y}^\dagger}) \Phi\left(\underbrace{g^\dagger_{(2)} - g^\dagger_{(1)}}_{-\Delta^\dagger \leqslant 0}, r\right) \\
&:= \min_{\Delta \geqslant 0} \left\{ \eta_{\tilde{y}^\dagger} \Phi(\Delta, r) + (1 - \eta_{\tilde{y}^\dagger}) \Phi(-\Delta, r) \right\},
\end{aligned}
$$

which is minimized by $\tilde{y}^\dagger = y^*$ and in this case $\eta_{\tilde{y}^\dagger} = \eta^*$. □

B.3 Proof of Proposition 5.1

Proposition 5.1 (restated) *Assume that L is a convex function with respect to r. Let $\partial W\left(r, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta}\right)$ be the subdifferential of $W\left(r, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta}\right)$. Then L is rejection-calibrated if and only if for any $\left(r^\dagger_{\boldsymbol{\eta}}, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}\right) \in \operatorname{argmin}_{r, \boldsymbol{g}} W(r, \boldsymbol{g}; \boldsymbol{\eta})$,*

$$
\sup_{\boldsymbol{\eta}: \eta^* > 1-c} \sup_{s_{\boldsymbol{\eta}} \in \partial W\left(r, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta}\right)\big|_{r=0}} s_{\boldsymbol{\eta}} \leqslant 0 \leqslant \inf_{\boldsymbol{\eta}: \eta^* < 1-c} \inf_{s_{\boldsymbol{\eta}} \in \partial W\left(r, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta}\right)\big|_{r=0}} s_{\boldsymbol{\eta}},
$$

and

$$
0 \notin \partial W\left(r, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta}\right)\big|_{r=0}
$$

for all $\boldsymbol{\eta} \in \Delta_K$ such that $\eta^* \neq 1 - c$, where $\eta^* = \max_y \eta_y$.

Proof. Since L is convex with respect to r and $W(r, \boldsymbol{g}; \boldsymbol{\eta})$ is a convex combination of L, we have $W(r, \boldsymbol{g}; \boldsymbol{\eta})$ is also convex with respect to r. Then using the property of subdifferential at optimality (Borwein & Lewis, 2006, Proposition 3.1.5), $r^\dagger_{\boldsymbol{\eta}}$ is a minimizer of $W\left(r, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta}\right)$ if and only if

$$
0 \in \partial W\left(r^\dagger_{\boldsymbol{\eta}}, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta}\right). \tag{8}
$$

First, consider the case $\eta^* > 1 - c$. In this case, from Equation (5), L is rejection-calibrated if and only if $r^\dagger_{\boldsymbol{\eta}} > 0$ for any $\left(r^\dagger_{\boldsymbol{\eta}}, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}\right) \in \operatorname{argmin}_{r, \boldsymbol{g}} W(r, \boldsymbol{g}; \boldsymbol{\eta})$. Next, we show that it holds if and only if

$$
\sup_{\boldsymbol{\eta}: \eta^* > 1-c} \sup_{s_{\boldsymbol{\eta}} \in \partial W\left(r, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta}\right)\big|_{r=0}} s_{\boldsymbol{\eta}} \leqslant 0, \tag{9}
$$

and

$$
0 \notin \partial\, W\left(r, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta}\right)\big|_{r=0}.
$$

For the sufficiency, assume $r^\dagger_{\boldsymbol{\eta}} > 0$ for any $\left(r^\dagger_{\boldsymbol{\eta}}, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}\right) \in \operatorname{argmin}_{r, \boldsymbol{g}} W(r, \boldsymbol{g}; \boldsymbol{\eta})$. Since the convexity of $W(r, \boldsymbol{g}; \boldsymbol{\eta})$ with respect to r, for any r_1 and r_2, we have

$$
(s_1 - s_2)(r_1 - r_2) \geqslant 0, \quad \text{for any } s_1 \in \partial W\left(r, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta}\right)\big|_{r=r_1}, s_2 \in \partial\, W\left(r, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta}\right)\big|_{r=r_2},
$$

where we use the monotonicity of subdifferential mappings (Rockafellar, 1970, Theorem A).

Then for any $\left(r^\dagger_{\boldsymbol{\eta}}, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}\right) \in \operatorname{argmin}_{r, \boldsymbol{g}} W(r, \boldsymbol{g}; \boldsymbol{\eta})$ and $r = 0$, we have

$$
(0 - s_{\boldsymbol{\eta}})(r^\dagger_{\boldsymbol{\eta}} - 0) \geqslant 0, \quad \text{for any } s_{\boldsymbol{\eta}} \in \partial W\left(r, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta}\right)\big|_{r=0}.
$$

Then we have that for any $s_{\boldsymbol{\eta}} \in \partial W\left(r, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta}\right)\big|_{r=0}, s_{\boldsymbol{\eta}} \leqslant 0$ for $r^\dagger_{\boldsymbol{\eta}} > 0$, which implies $\sup_{\boldsymbol{\eta}: \eta^* > 1-c} \sup_{s_{\boldsymbol{\eta}} \in \partial W\left(r, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta}\right)\big|_{r=0}} s_{\boldsymbol{\eta}} \leqslant 0$. Besides, since that $r^\dagger_{\boldsymbol{\eta}} \neq 0$ for any $r^\dagger_{\boldsymbol{\eta}} \in \operatorname{argmin}_r W\left(r, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta}\right)$, and that $0 \in \partial W\left(r^\dagger_{\boldsymbol{\eta}}, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta}\right)$ if and only if $r^\dagger_{\boldsymbol{\eta}} \in \operatorname{argmin}_r W\left(r, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta}\right)$, we have $0 \notin \partial W\left(r, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta}\right)\big|_{r=0}$ in this case (which can be validated by contradiction).

For the necessity, assume that for any $\left(r_{\boldsymbol{\eta}}^{\dagger}, \boldsymbol{g}_{\boldsymbol{\eta}}^{\dagger}\right) \in \operatorname{argmin}_{r,\boldsymbol{g}} W(r, \boldsymbol{g}; \boldsymbol{\eta})$, $\sup_{\boldsymbol{\eta}:\eta^*>1-c} \sup_{s_{\boldsymbol{\eta}} \in \partial W\left(r, \boldsymbol{g}_{\boldsymbol{\eta}}^{\dagger}; \boldsymbol{\eta}\right)\big|_{r=0}} s_{\boldsymbol{\eta}} \leqslant 0$, and $0 \notin \partial W\left(r, \boldsymbol{g}_{\boldsymbol{\eta}}^{\dagger}; \boldsymbol{\eta}\right)\big|_{r=0}$ for $\eta^* > 1-c$. The first inequality implies $\forall s_{\boldsymbol{\eta}} \in \partial W\left(r, \boldsymbol{g}_{\boldsymbol{\eta}}^{\dagger}; \boldsymbol{\eta}\right)\big|_{r=0}, s_{\boldsymbol{\eta}} \leqslant 0$. Combined with the second equality, we have $\forall s_{\boldsymbol{\eta}} \in \partial W\left(r, \boldsymbol{g}_{\boldsymbol{\eta}}^{\dagger}; \boldsymbol{\eta}\right)\big|_{r=0}, s_{\boldsymbol{\eta}} < 0$. From the convexity of $W(r, \boldsymbol{g}; \boldsymbol{\eta})$ with respect to r, for any $\left(r_{\boldsymbol{\eta}}^{\dagger}, \boldsymbol{g}_{\boldsymbol{\eta}}^{\dagger}\right) \in \operatorname{argmin}_{r,\boldsymbol{g}} W(r, \boldsymbol{g}; \boldsymbol{\eta})$, we have

$$(0 - s_{\boldsymbol{\eta}})(r_{\boldsymbol{\eta}}^{\dagger} - 0) \geqslant 0, \quad \text{for any } s_{\boldsymbol{\eta}} \in \partial W\left(r, \boldsymbol{g}_{\boldsymbol{\eta}}^{\dagger}; \boldsymbol{\eta}\right)\big|_{r=0}. \tag{10}$$

Since that $\forall s_{\boldsymbol{\eta}} \in \partial W\left(r, \boldsymbol{g}_{\boldsymbol{\eta}}^{\dagger}; \boldsymbol{\eta}\right)\big|_{r=0}, s_{\boldsymbol{\eta}} \neq 0$ and $0 \in \partial W\left(r_{\boldsymbol{\eta}}^{\dagger}, \boldsymbol{g}_{\boldsymbol{\eta}}^{\dagger}; \boldsymbol{\eta}\right)$ if and only if $r_{\boldsymbol{\eta}}^{\dagger} \in \operatorname{argmin}_r W\left(r, \boldsymbol{g}_{\boldsymbol{\eta}}^{\dagger}; \boldsymbol{\eta}\right)$, we have $r_{\boldsymbol{\eta}}^{\dagger} \neq 0$. Then combined with Equation (10), we have that for any $\left(r_{\boldsymbol{\eta}}^{\dagger}, \boldsymbol{g}_{\boldsymbol{\eta}}^{\dagger}\right) \in \operatorname{argmin}_{r,\boldsymbol{g}} W(r, \boldsymbol{g}; \boldsymbol{\eta})$, $r_{\boldsymbol{\eta}}^{\dagger} > 0$.

Therefore, the sufficient and necessary condition is proved.

Similary, for the case $\eta^* < 1-c$, for any $\left(r_{\boldsymbol{\eta}}^{\dagger}, \boldsymbol{g}_{\boldsymbol{\eta}}^{\dagger}\right) \in \operatorname{argmin}_{r,\boldsymbol{g}} W(r, \boldsymbol{g}; \boldsymbol{\eta})$, we have

$$\inf_{\boldsymbol{\eta}:\eta^*>1-c} \inf_{s_{\boldsymbol{\eta}} \in \partial W\left(r, \boldsymbol{g}_{\boldsymbol{\eta}}^{\dagger}; \boldsymbol{\eta}\right)\big|_{r=0}} s_{\boldsymbol{\eta}} \geqslant 0,$$

and

$$0 \notin \partial W\left(r, \boldsymbol{g}_{\boldsymbol{\eta}}^{\dagger}; \boldsymbol{\eta}\right).$$

Therefore, we have L is rejection-calibrated if and only if for any $\left(r_{\boldsymbol{\eta}}^{\dagger}, \boldsymbol{g}_{\boldsymbol{\eta}}^{\dagger}\right) \in \operatorname{argmin}_{r,\boldsymbol{g}} W(r, \boldsymbol{g}; \boldsymbol{\eta})$,

$$\sup_{\boldsymbol{\eta}:\eta^*>1-c} \sup_{s_{\boldsymbol{\eta}} \in \partial W\left(r, \boldsymbol{g}_{\boldsymbol{\eta}}^{\dagger}; \boldsymbol{\eta}\right)\big|_{r=0}} s_{\boldsymbol{\eta}} \leqslant 0 \leqslant \inf_{\boldsymbol{\eta}:\eta<1-c} \inf_{s_{\boldsymbol{\eta}} \in \partial W\left(r, \boldsymbol{g}_{\boldsymbol{\eta}}^{\dagger}; \boldsymbol{\eta}\right)\big|_{r=0}} s_{\boldsymbol{\eta}},$$

and

$$0 \notin \partial W\left(r, \boldsymbol{g}_{\boldsymbol{\eta}}^{\dagger}; \boldsymbol{\eta}\right)$$

for all $\boldsymbol{\eta} \in \Delta$ such that $\eta^* \neq 1-c$, where $\eta^* = \max_y \eta_y$.

□

C Useful Lemmas

C.1 Lemmas for the MCS Loss

In this section, we show some useful lemmas for the MCS loss and their proofs. The following lemma gives specific forms of H_η in Theorem 5.3 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

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

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 $\Delta^\circ \in \mathbb{R}$ be such that $F'(\Delta^\circ) = 0$. Then we see that

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

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 $\Delta^\circ \geqslant 0$. Then $F(\Delta) \geqslant F(\Delta^\circ) = 2\sqrt{\eta(1-\eta)},$ for $\Delta \geqslant 0$.

Next consider the case $\eta < \frac{1}{2}$. In this case, we have $\Delta^\circ < 0$. 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 $\Delta^\circ \in \mathbb{R}$ be such that $F'(\Delta^\circ) = 0$. Then we see that

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

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 $\Delta^\circ \geqslant 0$. Then $F(\Delta) \geqslant F(\Delta^\circ) = -\eta\log\eta - (1-\eta)\log(1-\eta),$ for $\Delta \geqslant 0$.

Next consider the case $\eta < \frac{1}{2}$. In this case, we have $\Delta^\circ < 0$. 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 $W_{\mathrm{MCS}}(r, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta})$ with respect to r, and the corresponding minimizer $r^\dagger_{\boldsymbol{\eta}} \in \operatorname{argmin}_r W(r, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta})$ with specific choices of ϕ, ψ_1, and ψ_2.

Lemma C.2. *Define the MCS pointwise loss as*

$$W_{\mathrm{MCS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) = \sum_y \eta_y \phi\left(g_y - \max_{y' \neq y} g_{y'}\right)\psi_1(r) + \psi_2(r). \tag{11}$$

For W_{MCS} with $\phi(\Delta) = \exp(-\Delta)$, $\psi_1(r) = \exp(r)$, and $\psi_2(r) = c\exp(-\alpha r)$, we have

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

and

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

Proof. From Equation (11), the minimization of W_{MCS} with respect to $\boldsymbol{g}$ does not depend on r. Therefore, by Proposition 5.2, we have

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

where recall that

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

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

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

Thus for $\psi_1(r) = \exp(r)$, $\psi_2(r) = c\exp(-\alpha r)$, we have

$$W_{\mathrm{MCS}}(r, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta}) = \begin{cases} 2\sqrt{\eta^*(1-\eta^*)}\exp(r) + c\exp(-\alpha r) & \text{if } \eta^* \geqslant \frac{1}{2}, \\ \exp(r) + c\exp(-\alpha r) & \text{if } \eta^* < \frac{1}{2}, \end{cases} \tag{12}$$

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

$$\frac{\partial W_{\mathrm{MCS}}(r, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta})}{\partial r} = \begin{cases} 2\sqrt{\eta^*(1-\eta^*)}\exp(r) - \alpha c\exp(-\alpha r) & \text{if } \eta^* \geqslant \frac{1}{2}, \\ \exp(r) - \alpha c\exp(-\alpha r) & \text{if } \eta^* < \frac{1}{2}. \end{cases} \tag{13}$$

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

$$r^\dagger_{\boldsymbol{\eta}} = \begin{cases} \frac{1}{\alpha+1}\log\left(\frac{\alpha c}{2\sqrt{\eta^*(1-\eta^*)}}\right) & \text{if } \eta^* \geqslant \frac{1}{2}, \\ \frac{1}{\alpha+1}\log\left(\alpha c\right) & \text{if } \eta^* < \frac{1}{2}. \end{cases} \tag{14}$$

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

$$W_{\mathrm{MCS}}(r^\dagger_{\boldsymbol{\eta}}, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta}) = \begin{cases} (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} \left(2\sqrt{\eta^*(1-\eta^*)}\right)^{\frac{\alpha}{\alpha+1}} & \text{if } \eta^* \geqslant \frac{1}{2}, \\ (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} & \text{if } \eta^* < \frac{1}{2}. \end{cases} \tag{15}$$

□

The following lemma gives the minimum of some function with respect to r.

Lemma C.3. *Given*

$$I(r; \eta) = 2\sqrt{\eta(1-\eta)}e^r + ce^{-\alpha r},$$

where $\frac{1}{2} \leqslant \eta \leqslant 1$. *Let* $r_\eta \in \operatorname{argmin}_r I(r; \eta)$ *be a minimizer of* $I(r; \eta)$. *Then we have*

$$I(r_\eta, \eta) = (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} \left(2\sqrt{\eta(1-\eta)}\right)^{\frac{\alpha}{\alpha+1}}.$$

Proof. We can see that $I(r; \eta)$ is a convex function respect to r. Its derivative is expressed as

$$\frac{\partial I(r; \eta)}{\partial r} = 2\sqrt{\eta(1-\eta)}\exp(r) - \alpha c\exp(-\alpha r).$$

Since the minimizer r_η satisifies the first-order optimality condition $\frac{\partial I(r;\eta)}{\partial r} \mid_{r = r_\eta} = 0$, we have

$$r_\eta = \frac{1}{\alpha+1}\log\left(\frac{\alpha c}{2\sqrt{\eta(1-\eta)}}\right).$$

By plugging it into $I(r; \eta)$, we have

$$I(r_\eta, \eta) = (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} \left(2\sqrt{\eta(1-\eta)}\right)^{\frac{\alpha}{\alpha+1}}.$$

□

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

Lemma C.4. *For function*

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

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

$$k(r) \geqslant A + B$$

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

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

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

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

Due to the convexity, we have that $k'(r)$ is non-decreasing for all r. First, consider the case of $A \geqslant \alpha B$. In this case, we have $r^{\circ} \leqslant 0$. Then we have $k'(r) \geqslant k'(0) \geqslant 0$ for $r > 0$, which implies that $k(r)$ is non-decreasing when $r > 0$. Thus,

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

Next, consider the case $A < \alpha B$. In this case, we have $r^{\circ} > 0$. Then we have $k'(r) \leqslant k'(0) \leqslant 0$ for $r \leqslant 0$, which implies that $k(r)$ is non-increasing when $r \leqslant 0$. Thus,

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

□

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

Lemma C.5. *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.

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

□

C.2 LEMMAS FOR THE ACS LOSS

In this section, we show some useful lemmas for the ACS loss and their proofs. The following lemma gives the infimum of function $F(r, \Delta; \eta)$ for $0 \leqslant \eta \leqslant 1$ and $\Delta \geqslant 0$.

Lemma C.6. *Define $F(r, \Delta; \eta)$ as*

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

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 strictly convex ψ.

Lemma C.7. *Define the ACS pointwise loss as*

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

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

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

where

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

for strictly convex $\psi(r)$, and

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

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

Proof. Recall $F(r,\Delta;\eta) = \eta\max\{1-\Delta+r,0\} + (1-\eta)\max\{1+\Delta+r,0\}$. By Proposition 5.2, we have

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

Using Lemma C.6, we have

$$W_{\mathrm{ACS}}(r, \boldsymbol{g}_{\boldsymbol{\eta}}^{\dagger}; \boldsymbol{\eta}) = \begin{cases} 2(1-\eta^*)(r+1) + \psi(r) & \text{if } \eta^* > \frac{1}{2} \text{ and } r \geqslant -1, \\ r+1+\psi(r) & \text{if } \eta^* \leqslant \frac{1}{2} \text{ and } r \geqslant -1, \\ \psi(r) & \text{if } r < -1. \end{cases} \tag{16}$$

For $\eta^* > \frac{1}{2}$ and $r \geqslant -1$, or $\eta^* \leqslant \frac{1}{2}$ and $r \geqslant -1$, define $\tilde{r}_{\boldsymbol{\eta}} \in \inf_{r \geqslant -1} W_{\mathrm{ACS}}(r, \boldsymbol{g}_{\boldsymbol{\eta}}^{\dagger}; \boldsymbol{\eta})$ as a minimizer. Then, we analyze two cases to obtain the expression of $W_{\mathrm{ACS}}(r_{\boldsymbol{\eta}}^{\dagger}, \boldsymbol{g}_{\boldsymbol{\eta}}^{\dagger}; \boldsymbol{\eta})$.

(**i**) Case $\psi(r)$ is strictly convex:

In this case, by taking the derivative, we obtain

$$\frac{\partial W_{\mathrm{ACS}}(r, \boldsymbol{g}_{\boldsymbol{\eta}}^{\dagger}; \boldsymbol{\eta})}{\partial r} = \begin{cases} 2(1-\eta^*) + \psi'(r) & \text{if } \eta^* > \frac{1}{2} \text{ and } r \geqslant -1, \\ 1+\psi'(r) & \text{if } \eta^* \leqslant \frac{1}{2} \text{ and } r \geqslant -1, \\ \psi'(r) & \text{if } r < -1. \end{cases} \tag{17}$$

Then the necessary and sufficient condition for $\tilde{r}_{\boldsymbol{\eta}}$ to be a minimizer of W_{ACS} for $r \geqslant -1$ is $\left.\frac{\partial W_{\mathrm{ACS}}(r, \boldsymbol{g}_{\boldsymbol{\eta}}^{\dagger}; \boldsymbol{\eta})}{\partial r}\right|_{r=\tilde{r}_{\boldsymbol{\eta}}} = 0$ for $r \geqslant -1$. Then we have

$$\tilde{r}_{\boldsymbol{\eta}} = \begin{cases} \left(\psi'\right)^{-1}\left(-2(1-\eta^*)\right) & \text{if } \eta^* > \frac{1}{2} \text{ and } r \geqslant -1, \\ \left(\psi'\right)^{-1}(-1) & \text{if } \eta^* \leqslant \frac{1}{2} \text{ and } r \geqslant -1. \end{cases} \tag{18}$$

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

$$\inf_{r \geqslant -1} W_{\mathrm{ACS}}(r, \boldsymbol{g}_{\boldsymbol{\eta}}^{\dagger}; \boldsymbol{\eta}) = \begin{cases} 2(1-\eta^*)\left(\tilde{r}_{\boldsymbol{\eta}}+1\right) + \psi\left(\tilde{r}_{\boldsymbol{\eta}}\right) & \text{if } \eta^* > \frac{1}{2}, \\ \left(\tilde{r}_{\boldsymbol{\eta}}+1\right) + \psi\left(\tilde{r}_{\boldsymbol{\eta}}\right) & \text{if } \eta^* \leqslant \frac{1}{2}. \end{cases} \tag{19}$$

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

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

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

$$\begin{aligned} \psi(-1) &\geqslant \psi\left(\left(\psi'\right)^{-1}\left(-2(1-\eta^*)\right)\right) + \psi\prime\left(\left(\psi'\right)^{-1}\left(-2(1-\eta^*)\right)\right)\left(-1-\left(\psi'\right)^{-1}\left(-2(1-\eta^*)\right)\right) \\ &= \psi\left(\left(\psi'\right)^{-1}\left(-2(1-\eta^*)\right)\right) - 2(1-\eta^*)\left(-1-\left(\psi'\right)^{-1}\left(-2(1-\eta^*)\right)\right). \end{aligned} \tag{20}$$

Hence, we have

$$\begin{aligned} &2(1-\eta^*)\left(\tilde{r}_{\boldsymbol{\eta}}+1\right) + \psi\left(\tilde{r}_{\boldsymbol{\eta}}\right) \\ &= 2(1-\eta^*)\left(\left(\psi'\right)^{-1}\left(-2(1-\eta^*)\right)+1\right) + \psi\left(\left(\psi'\right)^{-1}\left(-2(1-\eta^*)\right)\right) \qquad \text{(by Equation (21))} \end{aligned}$$

$$\leqslant \psi(-1),$$

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

For $(\tilde{r}_{\boldsymbol{\eta}} + 1) + \psi(\tilde{r}_{\boldsymbol{\eta}})$, note that $(\tilde{r}_{\boldsymbol{\eta}} + 1) + \psi(\tilde{r}_{\boldsymbol{\eta}}) = \left((\psi')^{-1}(-1) + 1\right) + \psi\left((\psi')^{-1}(-1)\right)$ 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 Equation (20), we have $(\tilde{r}_{\boldsymbol{\eta}} + 1) + \psi(\tilde{r}_{\boldsymbol{\eta}}) \leqslant \psi(-1)$. Thus

$$\begin{aligned}
\inf_r W_{\mathrm{ACS}}(r, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta}) &= \min\left\{\inf_{r \geqslant -1} W_{\mathrm{ACS}}(r, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta}), \inf_{r < -1} W_{\mathrm{ACS}}(r, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta})\right\} \\
&= \min\left\{\inf_{r \geqslant -1} W_{\mathrm{ACS}}(r, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta}), \phi(-1))\right\} \\
&= \inf_{r \geqslant -1} W_{\mathrm{ACS}}(r, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta}),
\end{aligned}$$

and we have $r^\dagger_{\boldsymbol{\eta}} = \tilde{r}_{\boldsymbol{\eta}}$.

Therefore

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

where

$$r^\dagger_{\boldsymbol{\eta}} = \begin{cases} (\psi')^{-1}(-2(1-\eta^*)) & \text{if } \eta^* > \frac{1}{2} \text{ and } r \geqslant -1, \\ (\psi')^{-1}(-1) & \text{if } \eta^* \leqslant \frac{1}{2} \text{ and } r \geqslant -1. \end{cases} \tag{21}$$

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

In this case, by Equation (16) we have

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

Recall that for $\eta^* > \frac{1}{2}$ and $r \geqslant -1$, or $\eta^* \leqslant \frac{1}{2}$ and $r \geqslant -1$, $\tilde{r}_{\boldsymbol{\eta}} \in \inf_{r \geqslant -1} W_{\mathrm{ACS}}(r, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta})$. 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

$$\tilde{r}_{\boldsymbol{\eta}} = \begin{cases} -1 & \text{if } \frac{1}{2} < \eta^* \leqslant 1 - \frac{\alpha c}{2} \text{ and } r \geqslant -1, \\ \frac{1}{\alpha} & \text{if } \eta^* > \frac{1}{2}, \eta^* > 1 - \frac{\alpha c}{2} \text{ and } r \geqslant -1. \end{cases} \tag{22}$$

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

$$\tilde{r}_{\boldsymbol{\eta}} = \begin{cases} -1 & \text{if } \eta^* \leqslant \frac{1}{2}, \alpha c \leqslant 1 \text{ and } r \geqslant -1, \\ \frac{1}{\alpha} & \text{if } \eta^* \leqslant \frac{1}{2}, \alpha c > 1 \text{ and } r \geqslant -1. \end{cases} \tag{23}$$

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

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

For $r < -1$, we have

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

We can see that $2(1-\eta^*)\left(\tilde{r}_{\boldsymbol{\eta}} + 1\right) + \psi\left(\tilde{r}_{\boldsymbol{\eta}}\right) = 2(1-\eta^*)(\frac{1}{\alpha} + 1) \leqslant (\alpha+1)c$ holds for $\eta^* > \frac{1}{2}$ and $\eta^* > 1 - \frac{\alpha c}{2}$, and $2(1-\eta^*)\left(\tilde{r}_{\boldsymbol{\eta}} + 1\right) + \psi\left(\tilde{r}_{\boldsymbol{\eta}}\right) = (\alpha+1)c$ for $\frac{1}{2} < \eta^* \leqslant 1 - \frac{\alpha c}{2}$ from Equation (22).

Besides, we can also see that $(\tilde{r}_{\boldsymbol{\eta}}+1)+\psi\left(\tilde{r}_{\boldsymbol{\eta}}\right)=\min\left\{(\alpha+1)c, \frac{1}{\alpha}+1\right\} \leqslant (\alpha+1)c$ holds for $\eta^* \leqslant \frac{1}{2}$ from Equation (23). Thus

$$\begin{aligned}\inf_r W_{\mathrm{ACS}}(r, \boldsymbol{g}_{\boldsymbol{\eta}}^\dagger; \boldsymbol{\eta}) &= \min\left\{\inf_{r\geqslant -1} W_{\mathrm{ACS}}(r, \boldsymbol{g}_{\boldsymbol{\eta}}^\dagger; \boldsymbol{\eta}), \inf_{r<-1} W_{\mathrm{ACS}}(r, \boldsymbol{g}_{\boldsymbol{\eta}}^\dagger; \boldsymbol{\eta})\right\} \\ &= \min\left\{\inf_{r\geqslant -1} W_{\mathrm{ACS}}(r, \boldsymbol{g}_{\boldsymbol{\eta}}^\dagger; \boldsymbol{\eta}), (\alpha+1)c\right\} \\ &= \inf_{r\geqslant -1} W_{\mathrm{ACS}}(r, \boldsymbol{g}_{\boldsymbol{\eta}}^\dagger; \boldsymbol{\eta}),\end{aligned}$$

and we have $r_{\boldsymbol{\eta}}^\dagger = \tilde{r}_{\boldsymbol{\eta}}$.

Therefore

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

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

Overall,

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

where

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

for strictly convex $\psi(r)$, and

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

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

The following lemma gives the minimum of some function with respect to r.

Lemma C.8. *Given*

$$I(r;\eta) = 2(1-\eta)(r+1)+\psi(r),$$

where $\frac{1}{2} < \eta \leqslant 1$. *Let* $r_\eta \in \operatorname{argmin} I(r;\eta)$ *be a minimizer of* $I(r;\eta)$. *We have*

$$I(r_\eta;\eta) = 2(1-\eta)\left(r_\eta+1\right)+\psi\left(r_\eta\right),$$

where

$$r_\eta = \left(\psi'\right)^{-1}\left(-2(1-\eta)\right),$$

for strictly convex $\psi(r)$, *and*

$$r_\eta = \begin{cases} -1 & \text{if } \frac{1}{2} < \eta \leqslant 1-\frac{\alpha c}{2}, \\ \frac{1}{\alpha} & \text{if } \eta > \frac{1}{2} \text{ and } \eta > 1-\frac{\alpha c}{2}, \end{cases}$$

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

Proof. Define $r_\eta \in \inf_r I(r;\eta)$ as a minimizer. Then, we analyze two cases to obtain the expression of $I(r_\eta;\eta)$.

(**i**) Case $\psi(r)$ is strictly convex:

In this case, by taking the derivative, we obtain

$$\frac{\partial I(r;\eta)}{\partial r} = 2(1-\eta) + \psi'(r).$$

Then the necessary and sufficient condition for r_η to be a minimizer of W_{ACS} is $\frac{\partial I(r;\eta)}{\partial r}\Big|_{r=r_\eta} = 0$. Then we have

$$r_\eta = (\psi')^{-1}(-2(1-\eta)).$$

By plugging it into $I(r_\eta;\eta)$, we have

$$I(r_\eta;\eta) = 2(1-\eta)(r_\eta+1) + \psi(r_\eta).$$

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

In this case, we have

$$I(r;\eta) = 2(1-\eta)(r+1) + c\max\{1-\alpha r, 0\},$$

where $\eta > \frac{1}{2}$.

Recall that $r_\eta \in \inf_r I(r;\eta)$. Then for $\frac{1}{2} < \eta \leqslant 1 - \frac{\alpha c}{2}$, or $\eta > \frac{1}{2}$ and $\eta > 1 - \frac{\alpha c}{2}$, we have

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

By plugging it into $I(r;\eta)$, we have

$$I(r_\eta;\eta) = 2(1-\eta)(r_\eta+1) + \psi(r_\eta).$$

Overall,

$$I(r_\eta;\eta) = 2(1-\eta)(r_\eta+1) + \psi(r_\eta),$$

where

$$r_\eta = (\psi')^{-1}(-2(1-\eta)),$$

for strictly convex $\psi(r)$, and

$$r_\eta = \begin{cases} -1 & \text{if } \frac{1}{2} < \eta \leqslant 1 - \frac{\alpha c}{2}, \\ \frac{1}{\alpha} & \text{if } \eta > \frac{1}{2} \text{ and } \eta > 1 - \frac{\alpha c}{2}, \end{cases}$$

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.9. *For function*

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

where $A \geqslant 1$ and $\psi(r)$ is a non-increasing convex function differentiable at $r = 0$, we have

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

if $A \geqslant -\psi'(0), r > 0$, or if $A < -\psi'(0), r \leqslant 0$.

Proof. From the property of subdifferential (Rockafellar, 1997, Theorem 23.8), we have $\partial l(r) = \{A\} + \partial\psi(r)$. Then Since $\psi(r)$ is differentiable at $r = 0$, for any $s_0 \in \partial l(0) = \{A\} + \partial\psi(0)$, we have $s_0 = A + \psi'(0)$. Then, combined with the convexity of $l(r)$, the above result can be obtained. □

We then introduce a useful result about a function of $\eta > \frac{1}{2}$ and $c < \frac{1}{2}$, which is useful in the proof of the excess risk bound for ACS.

Lemma C.10. *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.

$$\begin{aligned}
(1-\eta)\log\left(\frac{1-\eta}{c}\right) - (1-\eta) + c &\geqslant (1-\eta)\frac{2(1-\eta-c)}{1-\eta+c} - (1-\eta) + c \\
&\qquad\qquad \text{(by } \log x > \tfrac{2(x-1)}{x+1} \text{ for } x > 1\text{)} \\
&= \frac{2(1-\eta)(1-\eta-c) - (1-\eta-c)(1-\eta+c)}{1-\eta+c} \\
&= \frac{(1-\eta)^2 - 2c(1-\eta) + c^2}{1-\eta+c} \\
&= \frac{(1-\eta-c)^2}{1-\eta+c} \\
&\geqslant (1-\eta-c)^2 \qquad \left(1-\eta+c < 1 \text{ for } \eta > \tfrac{1}{2} \text{ and } c < \tfrac{1}{2}\right) \\
&= (1-c-\eta)^2.
\end{aligned}$$

□

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. Before that, we introduce a general form of $W_{\mathrm{ACS}}(r, \boldsymbol{g}_{\boldsymbol{\eta}}^{\dagger}; \boldsymbol{\eta})$. Recall that the ACS pointwise loss is 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 5.2, $W_{\mathrm{ACS}}(r, \boldsymbol{g}_{\boldsymbol{\eta}}^{\dagger}; \boldsymbol{\eta})$ can be writed as

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

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

First, we introduce the result about the first-order derivative of ϕ.

Lemma C.11. *Whenever* $\Delta_\eta > 0$*, for any twice differentiable decreasing function* ψ*, we have*

$$\eta\phi'\left(\Delta_\eta - r_\eta\right) = (1-\eta)\phi'\left(-\Delta_\eta - r_\eta\right) = \frac{\psi'\left(r_\eta\right)}{2}. \tag{24}$$

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

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

□

Next, we introduce the result about the second order derivative of ϕ. In the proof of the following lemma, we use a subscript to denote a derivative of $\tilde{W}_{\mathrm{ACS}}$. For example,

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

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

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

Lemma C.12. *Assume $c < 1/2$ and that ϕ is decreasing and twice differentiable. Then, r_η is monotonically increasing in $\eta > 0$ if*

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

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

Proof. For $\tilde{W}_{\mathrm{ACS}}$ we have

$$\tilde{W}_\Delta^{\mathrm{ACS}}(r,0;\eta) = (2\eta - 1)\phi'(-r). \tag{25}$$

First consider the case $\eta \leqslant 1/2$. In this case, from Equation (25) we have $\tilde{W}_\Delta^{\mathrm{ACS}}(r,0;\eta) \geqslant 0$. Therefore $\tilde{W}_{\mathrm{ACS}}(r,\Delta;\eta)$ is increasing at $\Delta = 0$ for any r. Recall that Δ_η is the minimizer of $\tilde{W}_{\mathrm{ACS}}$ over $\Delta \geqslant 0$. Then we have $\Delta_\eta = 0$. Then r_η is the solution of

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

Therefore r_η does not depend on η, which means that r_η is non-decreasing.

Next we consider the case $\eta > 1/2$. In this case from Equation (25) we have $\tilde{W}_\Delta^{\mathrm{ACS}}(r,0;\eta) < 0$, 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

$$\nabla^2 \tilde{W}_{\mathrm{ACS}}(r_\eta,\Delta_\eta;\eta) \begin{pmatrix} r'_\eta \\ \Delta'_\eta \end{pmatrix} + \nabla \tilde{W}_\eta^{\mathrm{ACS}}(r_\eta,\Delta_\eta;\eta) = \mathbf{0}, \tag{26}$$

where

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

Here note that

$$\left(\nabla^2 \tilde{W}_{\mathrm{ACS}}(r_\eta,\Delta_\eta;\eta)\right)^{-1} = \frac{1}{D_\eta} \begin{pmatrix} \tilde{W}_{\Delta\Delta}^{\mathrm{ACS}}(r_\eta,\Delta_\eta;\eta) & -\tilde{W}_{r\Delta}^{\mathrm{ACS}}(r_\eta,\Delta_\eta;\eta) \\ -\tilde{W}_{r\Delta}^{\mathrm{ACS}}(r_\eta,\Delta_\eta;\eta) & \tilde{W}_{rr}^{\mathrm{ACS}}(r_\eta,\Delta_\eta;\eta) \end{pmatrix}, \tag{27}$$

where $D_\eta = |\nabla^2 \tilde{W}(r_\eta,\Delta_\eta;\eta)|$, which is positive from the convexity of W. From Equation (26) and Equation (27) we have

$$r'_\eta = -\frac{1}{D_\eta} \left(\tilde{W}_{\Delta\Delta}^{\mathrm{ACS}}(r_\eta,\Delta_\eta;\eta) \quad - \tilde{W}_{r\Delta}^{\mathrm{ACS}}(r_\eta,\Delta_\eta;\eta) \right) \nabla \tilde{W}_\eta^{\mathrm{ACS}}(r_\eta,\Delta_\eta;\eta). \tag{28}$$

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 Equation (28) 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.11, 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 $\psi' = \psi'(r_\eta) < 0$ from the assumption, we see that if $\eta^2\phi''_+ - (1-\eta)^2\phi''_- > 0$ then $r'_\eta > 0$, that is, r_η is increasing. □

C.3 Lemmas for Generalization Bounds

In this section, we show some useful lemmas for the proof of generalization bounds. Recall definition of $\mathcal{G}$ in Section 7, and we define $\widetilde{\mathcal{G}}$, $\phi \circ \widetilde{\mathcal{G}}$, $\mathcal{L}_{\mathrm{MCS}}$ as:

$$\begin{aligned} \widetilde{\mathcal{G}} &= \left\{\boldsymbol{x} \mapsto \left(g_y(\boldsymbol{x}) - \max_{y'\neq y} g_{y'}(\boldsymbol{x})\right) \mid g_1, \ldots, g_K \in \mathcal{G}\right\}, \\ \phi \circ \widetilde{\mathcal{G}} &= \left\{\boldsymbol{x} \mapsto \phi(\tilde{g}(\boldsymbol{x})) \mid \tilde{g} \in \widetilde{\mathcal{G}}\right\}, \\ \mathcal{L}_{\mathrm{MCS}} &= \{(\boldsymbol{x}, y) \mapsto L_{\mathrm{MCS}}(r, \boldsymbol{g}; \boldsymbol{x}, y) \mid g_1, \ldots, g_K \in \mathcal{G}\}, \qquad (29) \\ \mathcal{L}_{\mathrm{ACS}} &= \{(\boldsymbol{x}, y) \mapsto L_{\mathrm{ACS}}(r, \boldsymbol{g}; \boldsymbol{x}, y) \mid g_1, \ldots, g_K \in \mathcal{G}\}. \qquad (30) \end{aligned}$$

Recall the Rademacher complexity of $\mathcal{H}$ is defined as $\mathfrak{R}_n(\mathcal{H})$ in Definition 7.1.

Next, we introduce a result about the 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 $\mathcal{H}_1, \ldots, \mathcal{H}_K$ be K hypothesis classes ($K \geqslant 1$) of hypotheses mapping $\mathcal{X}$ to $\mathbb{R}$, and let $\mathcal{H} = \{\max\{h_1, \ldots, h_K\} \mid h_i \in \mathcal{H}_i, i \in [K]\}$. 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 the Rademacher complexity of the hypothesis class of functions that are inner products of other functions. A similar lemma can be found in DeSalvo et al. (2015); the main difference is the codomain of hypotheses.

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:

$$\begin{aligned} \mathfrak{R}_n(\mathcal{H}) &\leqslant \frac{1}{2}(M_1 + M_2)\mathfrak{R}_n(\mathcal{H}_1 + \mathcal{H}_2) + \frac{1}{2}M_1\mathfrak{R}_n(\mathcal{H}_1 - \mathcal{H}_2) \\ &\qquad \text{(Talagrand's contraction lemma , see Mohri et al., 2018, Lemma 5.7)} \\ &\leqslant \frac{1}{2}(M_1 + M_2)\mathfrak{R}_n(\mathcal{H}_1) + \frac{1}{2}(M_1 + M_2)\mathfrak{R}_n(\mathcal{H}_2) + \frac{1}{2}M_1\mathfrak{R}_n(\mathcal{H}_1) + \frac{1}{2}M_1\mathfrak{R}_n(-\mathcal{H}_2) \\ &\qquad \text{(sub-additivity of } \sup(\cdot)) \end{aligned}$$

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

□

Next, we introduce a useful inequality that bounds the Rademacher complexity of $\mathcal{L}_{\mathrm{MCS}}$ with that of $\mathcal{G}$ and $\mathcal{R}$.

Lemma C.15. *Let $\mathfrak{R}_n(\mathcal{L}_{\mathrm{MCS}})$ be the Rademacher complexity of $\mathcal{L}_{\mathrm{MCS}}$ for $S = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$ 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 $\mathfrak{R}_n(\mathcal{L}_{\mathrm{MCS}}) \leqslant 2M_\phi L_\phi K^2 \mathfrak{R}_n(\mathcal{G}) + (2M_{\psi_1}L_{\psi_1} + L_{\psi_2})\mathfrak{R}_n(\mathcal{R})$.*

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\ldots,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\ldots,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\ldots,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\qquad \text{(sub-additivity of } \sup(\cdot)\text{)} \\
&\leqslant \mathop{\mathbb{E}}_{S,\boldsymbol{\sigma}}\left[\sup_{g_1\ldots,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\ldots,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\ldots,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\ldots,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\ldots,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{(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\ldots,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{(by Lemma C.13)} \\
&\leqslant K^2\mathfrak{R}_n(\mathcal{G}). \qquad (31)
\end{aligned}$$

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

$$\begin{aligned}
&\mathfrak{R}_n(\mathcal{L}_{\mathrm{MCS}}) \\
&= \mathop{\mathbb{E}}_{S,\boldsymbol{\sigma}}\left[\sup_{L\in\mathcal{L}_{\mathrm{MCS}}} \frac{1}{n}\sum_{i=1}^n \sigma_i L_{\mathrm{MCS}}(r, \boldsymbol{g}; \boldsymbol{x}_i, y_i)\right] \\
&= \mathop{\mathbb{E}}_{S,\boldsymbol{\sigma}}\left[\sup_{r\in\mathcal{R},g_1\ldots,g_K\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n \sigma_i \left(\phi\left(g_{y_i}(\boldsymbol{x}_i) - \max_{y'\neq y_i} g_{y'}(\boldsymbol{x}_i)\right)\psi_1(r(\boldsymbol{x}_i)) + \psi_2(r(\boldsymbol{x}_i))\right)\right] \\
&\leqslant \mathop{\mathbb{E}}_{S,\boldsymbol{\sigma}}\left[\sup_{r\in\mathcal{R},g_1\ldots,g_K\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n \sigma_i \phi\left(g_{y_i}(\boldsymbol{x}_i) - \max_{y'\neq y_i} g_{y'}(\boldsymbol{x}_i)\right)\psi_1(r(\boldsymbol{x}_i))\right] \\
&\quad + \mathop{\mathbb{E}}_{S,\boldsymbol{\sigma}}\left[\sup_{r\in\mathcal{R}} \frac{1}{n}\sum_{i=1}^n \sigma_i \psi_2(r(\boldsymbol{x}_i))\right] \qquad \text{(sub-additivity of } \sup(\cdot)\text{)} \\
&= \mathfrak{R}_n((\phi\circ\widetilde{\mathcal{G}})(\psi_1\circ\mathcal{R})) + \mathfrak{R}_n(\psi_2\circ\mathcal{R}) \\
&\qquad \text{(where we denote } \{h_2h_2 \mid h_1\in\mathcal{H}_1, h_2\in\mathcal{H}_2\} \text{ as } \mathcal{H}_1\mathcal{H}_2\text{)} \\
&\leqslant \left(M_\phi + \frac{1}{2}M_{\psi_1}\right)\left(\mathfrak{R}_n(\phi\circ\widetilde{\mathcal{G}}) + \mathfrak{R}_n(\psi_1\circ\mathcal{R})\right) + \mathfrak{R}_n(\psi_2\circ\mathcal{R}) \qquad \text{(by Lemma C.14)}
\end{aligned}$$

$$\leqslant \left(M_\phi + \frac{1}{2}M_{\psi_1}\right) L_\phi \mathfrak{R}_n(\widetilde{\mathcal{G}}) + \left(M_\phi + \frac{1}{2}M_{\psi_1}\right) L_{\psi_1}\mathfrak{R}_n(\mathcal{R}) + L_{\psi_2}\mathfrak{R}_n(\mathcal{R})$$

(Talagrand's contraction lemma, L_ϕ, L_{ψ_1}, L_{ψ_2} are Lipschitz constants of ϕ, ψ_1, ψ_2, respectively))

$$\leqslant \left(M_\phi + \frac{1}{2}M_{\psi_1}\right) L_\phi K^2 \mathfrak{R}_n(\mathcal{G}) + \left(\left(M_\phi + \frac{1}{2}M_{\psi_1}\right) L_{\psi_1} + L_{\psi_2}\right) \mathfrak{R}_n(\mathcal{R}).$$

(by Equation (31))

□

Next, we introduce a useful inequality that bounds the 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 $S = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$ 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\ldots,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\ldots,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\ldots,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\ldots,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\ldots,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\ldots,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\ldots,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\ldots,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{(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{(by Lemma C.13)} \\
&\leqslant K^2\mathfrak{R}_n(\mathcal{G}) + \mathfrak{R}_n(\mathcal{R}). && (32)
\end{aligned}$$

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

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

$$
\begin{aligned}
&= \mathfrak{R}_n(\phi \circ \widetilde{\mathcal{G}}) + \mathfrak{R}_n(\psi \circ \mathcal{R}) \\
&\leqslant L_\phi \mathfrak{R}_n(\widetilde{\mathcal{G}}) + L_\psi \mathfrak{R}_n(\mathcal{R}) && \text{(by Talagrand's contraction lemma)} \\
&\leqslant L_\phi K^2 \mathfrak{R}_n(\mathcal{G}) + (L_\phi + L_\psi)\, \mathfrak{R}_n(\mathcal{R}). && \text{(by Equation (32))}
\end{aligned}
$$

□

C.4 Lemmas for Excess Risk Bounds

In this section, we show some useful lemmas for the proof of excess risk bounds. Define the pointwise risk $W_{0\text{-}1\text{-}c}$ of the target abstention loss $L_{0\text{-}1\text{-}c}$ at $\boldsymbol{x}$ in a similar way to Equation (4) by

$$
W_{0\text{-}1\text{-}c}\left(r, \boldsymbol{g}; \boldsymbol{\eta}\right) = \sum_y \eta_y L(r, \boldsymbol{g}; \boldsymbol{x}, y). \tag{33}
$$

Denote by

$$
\left(r^*_{\boldsymbol{\eta}}, \boldsymbol{g}^*_{\boldsymbol{\eta}}\right) \in \operatorname*{argmin}_{r, \boldsymbol{g}} W_{0\text{-}1\text{-}c}(r, \boldsymbol{g}; \boldsymbol{\eta})
$$

a minimizer of $W_{0\text{-}1\text{-}c}$. Lemma C.17 in the following 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 Equation* (33)*, we have*

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

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 $\tilde{y} = \operatorname{argmax}_y g_y$. Then $W_{0\text{-}1\text{-}c}(r, \boldsymbol{g}; \boldsymbol{\eta}) = \sum_y \eta_y \mathbb{I}_{\tilde{y} \neq y} = 1 - \eta_{\tilde{y}}$ if $r > 0$, and $W_{0\text{-}1\text{-}c}(r, \boldsymbol{g}; \boldsymbol{\eta}) = c$ if $r \leqslant 0$. Let $\tilde{Y} = \left\{\operatorname{argmax}_y g_y \mid g_1, \ldots, g_K \in \mathcal{G}\right\}$, where recall that $\mathcal{G}$ is a family of functions $g_y : \mathcal{X} \to \mathbb{R}$. Then 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

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

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

□

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

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

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

Proof.

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

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

D Proofs of Theorems and Corollaries

D.1 Proof of Theorem 5.3

Theorem 5.3 (restated) *$L_{\mathrm{MCS}}(r, \boldsymbol{g}; \boldsymbol{x}, y)$ is rejection-calibrated if and only if*

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

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

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

$$L_{\mathrm{MCS}}(r, \boldsymbol{g}; \boldsymbol{x}, y) = \phi\left(g_y - \max_{y' \neq y} g_{y'}\right) \psi_1\left(r\right) + \psi_2\left(r\right), \tag{34}$$

where $\phi(\Delta) \geqslant 0$ is a non-increasing function, $\psi_1(r)$, $\psi_2(r)$ are convex function differentiable at $r = 0$, and $\phi'(0) < 0, \psi_1'(0) > 0$. From the expression of Equation (34), the minimization of $W_{\mathrm{MCS}}\left(r, \boldsymbol{g}; \boldsymbol{\eta}\right) = \sum_y \eta_y L_{\mathrm{MCS}}(r, \boldsymbol{g}; \boldsymbol{x}, y)$ with respect to $\boldsymbol{g}$ does not depend on r. Therefore, by Proposition 5.2 with $\Phi(\Delta, r) = \phi(\Delta)\psi_1(r) + \psi_2(r)$, for any $(r^\dagger_{\boldsymbol{\eta}}, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}) \in \operatorname{argmin}_{r, \boldsymbol{g}} W(r, \boldsymbol{g}; \boldsymbol{\eta})$, we have

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

where

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

for $0 \leqslant \eta \leqslant 1$.

Using the property of subdifferential (Rockafellar, 1997, Theorem 23.8), we obtain

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

where for set A and B, we define $A + B = \{a + b \mid a \in A, b \in B\}$. Because hinge is differentiable at $r = 0$, we have

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

Thus we obtain

$$\inf_{s_{\boldsymbol{\eta}} \in \partial W_{\mathrm{MCS}}(r, \boldsymbol{g}_{\boldsymbol{\eta}}^{\dagger}; \boldsymbol{\eta})\big|_{r=0}} s_{\boldsymbol{\eta}} = \sup_{s_{\boldsymbol{\eta}} \in \partial W_{\mathrm{MCS}}(r, \boldsymbol{g}_{\boldsymbol{\eta}}^{\dagger}; \boldsymbol{\eta})\big|_{r=0}} s_{\boldsymbol{\eta}} = H_{\eta^*}\psi_1'(0) + \psi_2'(0).$$

Since ϕ is a non-increasing function, we can see that H_η is non-increasing in η. Thus, by Proposition 5.1, L_{MCS} is rejection-calibrated if and only if for any $(r_{\boldsymbol{\eta}}^{\dagger}, \boldsymbol{g}_{\boldsymbol{\eta}}^{\dagger}) \in \operatorname{argmin}_{r,\boldsymbol{g}} W(r, \boldsymbol{g}; \boldsymbol{\eta})$,

$$\inf_{\boldsymbol{\eta}:\eta^* < 1-c} \inf_{s_{\boldsymbol{\eta}} \in \partial W_{\mathrm{MCS}}(r, \boldsymbol{g}_{\boldsymbol{\eta}}^{\dagger}; \boldsymbol{\eta})\big|_{r=0}} s_{\boldsymbol{\eta}} = \sup_{\boldsymbol{\eta}:\eta^* > 1-c} \sup_{s_{\boldsymbol{\eta}} \in \partial W_{\mathrm{MCS}}(r, \boldsymbol{g}_{\boldsymbol{\eta}}^{\dagger}; \boldsymbol{\eta})\big|_{r=0}} s_{\boldsymbol{\eta}} = H_{1-c}\psi_1'(0) + \psi_2'(0) = 0, \tag{35}$$

and

$$0 \notin \partial W\left(r, \boldsymbol{g}_{\boldsymbol{\eta}}^{\dagger}; \boldsymbol{\eta}\right)\big|_{r=0} \tag{36}$$

for all $\boldsymbol{\eta} \in \Delta_K$ such that $\eta^* \neq 1 - c$. Next, we show that Equation (36) can be implied by Equation (35) and the condition that $\phi_1'(0) < 0, \psi_1'(0) > 0$.

For $W_{\mathrm{MCS}}(r, \boldsymbol{g}_{\boldsymbol{\eta}}^{\dagger}; \boldsymbol{\eta})$, we have

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

From Equation (35), we have $\psi_2'(0) = -H_{1-c}\psi_1'(0)$. Thus

$$\begin{aligned}
\partial W_{\mathrm{MCS}}(r, \boldsymbol{g}_{\boldsymbol{\eta}}^{\dagger}; \boldsymbol{\eta})\big|_{r=0} &= \left\{H_{\eta^*}\psi_1'(0) - H_{1-c}\psi_1'(0)\right\} \\
&= \left\{\left(H_{\eta^*} - H_{1-c}\right)\psi_1'(0)\right\}.
\end{aligned}$$

.

Since $\phi'(0) < 0$, we have $\phi(\Delta) < \phi(-\Delta)$ for $\Delta > 0$, thus H_η is decreasing in η. Then we have $H_{\eta^*} \neq H_{1-c}$ for $\eta^* \neq 1 - c$. Combined with the condition that $\psi_1'(0) > 0$, we have $0 \notin \partial W_{\mathrm{MCS}}(r, \boldsymbol{g}_{\boldsymbol{\eta}}^{\dagger}; \boldsymbol{\eta})\big|_{r=0}$ for all $\boldsymbol{\eta} \in \Delta_K$ such that $\eta^* \neq 1 - c$. □

D.2 Proof of Theorem 5.4

Theorem 5.4 (restated) *$L_{\mathrm{ACS}}(r, \boldsymbol{g}; \boldsymbol{x}, y)$ with $\phi(x) = \max\{1 - x, 0\}$ is rejection-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 $r = 0$. Using Lemma C.6 and Proposition 5.2, for any $(r_{\boldsymbol{\eta}}^{\dagger}, \boldsymbol{g}_{\boldsymbol{\eta}}^{\dagger}) \in \operatorname{argmin}_{r,\boldsymbol{g}} W(r, \boldsymbol{g}; \boldsymbol{\eta})$, we have

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

Using the property of subdifferential (Rockafellar, 1997, Theorem 23.8), we obtain

$$\partial W_{\mathrm{ACS}}(r, \boldsymbol{g}_{\boldsymbol{\eta}}^{\dagger}; \boldsymbol{\eta}) = \begin{cases} \{2(1-\eta^*)\} + \partial\psi(r) & \text{if } \eta^* > \frac{1}{2} \text{ and } r \geqslant -1, \\ \{1\} + \partial\psi(r) & \text{if } \eta^* \leqslant \frac{1}{2} \text{ and } r \geqslant -1, \\ \partial\psi(r) & \text{if } r < -1. \end{cases} \tag{37}$$

Then because hinge is differentiable at $r = 0$, we have

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

Thus we obtain

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

We can see that $\left.\frac{\partial W_{\mathrm{ACS}}(r, \boldsymbol{g}_{\boldsymbol{\eta}}^{\dagger}; \boldsymbol{\eta})}{\partial r}\right|_{r=0}$ is non-increasing in η^*. Thus, L_{ACS} is rejection-calibrated if and only if for any $(r_{\boldsymbol{\eta}}^{\dagger}, \boldsymbol{g}_{\boldsymbol{\eta}}^{\dagger}) \in \operatorname{argmin}_{r, \boldsymbol{g}} W(r, \boldsymbol{g}; \boldsymbol{\eta})$,

$$\inf_{\boldsymbol{\eta}: \eta^* < 1-c} \inf_{s_{\boldsymbol{\eta}} \in \partial W_{\mathrm{ACS}}(r, \boldsymbol{g}_{\boldsymbol{\eta}}^{\dagger}; \boldsymbol{\eta})\big|_{r=0}} s_{\boldsymbol{\eta}} = \sup_{\boldsymbol{\eta}: \eta^* > 1-c} \sup_{s_{\boldsymbol{\eta}} \in \partial W_{\mathrm{ACS}}(r, \boldsymbol{g}_{\boldsymbol{\eta}}^{\dagger}; \boldsymbol{\eta})\big|_{r=0}} s_{\boldsymbol{\eta}} = 2c + \psi'(0) = 0, \tag{38}$$

and

$$0 \notin \partial W_{\mathrm{ACS}}\left(r, \boldsymbol{g}_{\boldsymbol{\eta}}^{\dagger}; \boldsymbol{\eta}\right) \tag{39}$$

for all $\boldsymbol{\eta} \in \Delta_K$ such that $\eta^* \neq 1 - c$. Next, we show that Equation (39) can be implied by Equation (38).

For $W_{\mathrm{ACS}}(r, \boldsymbol{g}_{\boldsymbol{\eta}}^{\dagger}; \boldsymbol{\eta})$, we have

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

From Equation (38), we have $\psi'(0) = -2c$. Thus

$$\partial W_{\mathrm{MCS}}(r, \boldsymbol{g}_{\boldsymbol{\eta}}^{\dagger}; \boldsymbol{\eta})\big|_{r=0} = \begin{cases} \{2(1-\eta^*) - 2c\} & \text{if } \eta^* > \frac{1}{2}, \\ \{1 - 2c\} & \text{if } \eta^* \leqslant \frac{1}{2}, \end{cases}$$

.

Then we have $0 \notin \partial W_{\mathrm{MCS}}(r, \boldsymbol{g}_{\boldsymbol{\eta}}^{\dagger}; \boldsymbol{\eta})\big|_{r=0}$ for all $\boldsymbol{\eta} \in \Delta_K$ such that $\eta^* \neq 1 - c$. □

D.3 Proof of Theorem 5.5

Theorem 5.5 (restated) *Assume that ϕ is a twice differentiable decreasing function. Then L_{ACS} with $\phi(x) = \log(1 + \exp(-x))$ is rejection-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 $\phi(x) = \log(1 + \exp(-x))$, we have

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

Recall that Lemma C.12 shows that if $\eta^2\phi''(\Delta_\eta - r_\eta) - (1-\eta)^2\phi''(-\Delta_\eta - r_\eta) > 0$ for all $\eta > \frac{1}{2}$, then r_η is monotonically increasing in $\eta > 0$. Under the same notation as the proof of Lemma C.12, 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

$$\begin{aligned}\eta^2\phi''_+ - (1-\eta)^2\phi''_- &= \left((1-\eta)^2\phi'_- - (\eta)^2\phi'_+\right) + \left((1-\eta)^2\left(\phi'_-\right)^2 - \eta^2\left(\phi'_+\right)^2\right) \\ &= \frac{\psi'}{2}(1-2\eta) \qquad (40) \\ &> 0,\end{aligned}$$

where Equation (40) follows from Lemma C.11 and the inequality holds since ψ is decreasing.

Now we can apply Lemma C.12 and we see that r_η is increasing. From Equation (5), to prove that the loss is rejection-calibrated, we need to prove $r_\eta > 0$ for $\eta > 1 - c$ and $r_\eta < 0$ for $\eta < 1 - c$. Therefore, to prove that the loss is rejection calibrated, it suffices to show $r_{1-c} = 0$. Recall that Lemma C.11 shows that $\eta\phi'(\Delta_\eta - r_\eta) = (1-\eta)\phi'(-\Delta_\eta - r_\eta) = \frac{\psi'(r_\eta)}{2}$. 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 6.1

Theorem 6.1 (restated) *The following excess risk bound holds for L_{MCS}:*

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

for all $r, \boldsymbol{g}$ and any distribution $p(\boldsymbol{x}, y)$, where

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

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

Proof. First, from Lemma C.17, we have

$$\begin{aligned}\Delta W_{0\text{-}1\text{-}c}(r, \boldsymbol{g}; \boldsymbol{\eta}) &= W_{0\text{-}1\text{-}c}(r, \boldsymbol{g}; \boldsymbol{\eta}) - W_{0\text{-}1\text{-}c}(r^*_{\boldsymbol{\eta}}, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta}), \\ &= \begin{cases} \eta^* - \eta_{\tilde{y}} & \text{if } \eta^* > (1-c), \quad r > 0, \\ 1 - c - \eta_{\tilde{y}} & \text{if } \eta^* \leqslant (1-c), \quad r > 0, \\ 0 & \text{if } \eta^* \leqslant (1-c), \quad r \leqslant 0, \\ \eta^* - 1 + c & \text{if } \eta^* > (1-c), \quad r \leqslant 0. \end{cases} \qquad (41)\end{aligned}$$

In the following, we introduce some useful results for bounding $\Delta W_{\mathrm{MCS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) = W_{\mathrm{MCS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) - W_{\mathrm{MCS}}(r^\dagger_{\boldsymbol{\eta}}, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta})$. 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

$$\sum_y \eta_y \exp\left(-\left(g_y - \max_{y' \neq y} g_{y'}\right)\right)$$

$$\geqslant \eta_{\tilde{y}} \exp(-\Delta) + (1-\eta_{\tilde{y}}) \exp(\Delta) := G(\Delta; \eta_{\tilde{y}}), \tag{42}$$

where $\tilde{y} = f(\boldsymbol{g}) \in \operatorname{argmax}_y g_y$, $0 \leqslant \eta_{\tilde{y}} \leqslant 1$, 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. Its derivative is expressed as $G'(\Delta; \eta_{\tilde{y}}) = -\eta_{\tilde{y}} \exp(-\Delta) + (1-\eta_{\tilde{y}}) \exp(\Delta)$. Let $\Delta^\circ \in \mathbb{R}$ be such that $G'(\Delta^\circ) = 0$. Then we see that

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

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

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

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

$$G(\Delta; \eta_{\tilde{y}}) \geqslant G(0; \eta_{\tilde{y}}) = 1, \text{ for } \Delta \geqslant 0. \tag{43}$$

Thus we have

$$\inf_{\Delta \geqslant 0} G(\Delta; \eta_{\tilde{y}}) \geqslant \min \left\{ 2\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})}, 1 \right\} = 2\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})}, \tag{44}$$

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

Since ψ_1, ψ_2 satisfy the calibration condition in Theorem 5.3, combined with Equation (7), 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 Equation (41) we have $\Delta W_{0\text{-}1\text{-}c}(r, \boldsymbol{g}; \boldsymbol{\eta}) = \eta^* - \eta_{\tilde{y}}$. For multiplicative CS loss, we have $\Delta W_{\mathrm{MCS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) = W_{\mathrm{MCS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) - W_{\mathrm{MCS}}(r^\dagger_{\boldsymbol{\eta}}, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta})$, where $W_{\mathrm{MCS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) \geqslant G(\Delta; \eta_{\tilde{y}})$ (by Equation (42)). For $G(\Delta; \eta_{\tilde{y}})$ we have $\inf_{\Delta \geqslant 0} G(\Delta; \eta_{\tilde{y}}) \geqslant 2\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})}$, where $\eta_{\tilde{y}} \geqslant \frac{1}{2}$ (by Equation (44)). By Lemma C.2, we have $W_{\mathrm{MCS}}(r^\dagger_{\boldsymbol{\eta}}, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta}) = (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} \left(2\sqrt{\eta^*(1-\eta^*)} \right)^{\frac{\alpha}{\alpha+1}}$ for $\eta^* > 1-c > \frac{1}{2}$. Thus we have

$$\begin{aligned}
&\Delta W_{\mathrm{MCS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) \\
&= \sum_y \eta_y \exp\left(-\left(g_y - \max_{y' \neq y} g_{y'} \right) \right) e^r + c e^{-\alpha r} - (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} \left(2\sqrt{\eta^*(1-\eta^*)} \right)^{\frac{\alpha}{\alpha+1}} \\
&\geqslant G(\Delta; \eta_{\tilde{y}}) e^r + c e^{-\alpha r} - (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} \left(2\sqrt{\eta^*(1-\eta^*)} \right)^{\frac{\alpha}{\alpha+1}} \\
&\qquad \left(\textstyle\sum_y \eta_y \exp\left(-\left(g_y - \max_{y' \neq y} g_{y'} \right) \right) \geqslant G(\Delta; \eta_{\tilde{y}}) \text{ in Equation (42)} \right) \\
&\geqslant 2\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})} e^r + c e^{-\alpha r} - (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} \left(2\sqrt{\eta^*(1-\eta^*)} \right)^{\frac{\alpha}{\alpha+1}} \\
&\qquad \left(\inf_{\Delta \geqslant 0} G(\Delta) \geqslant 2\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})}, \text{ where } \eta_{\tilde{y}} \geqslant \tfrac{1}{2} \text{ in Equation (44)} \right) \\
&\geqslant (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} \left(\left(2\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})} \right)^{\frac{\alpha}{\alpha+1}} - \left(2\sqrt{\eta^*(1-\eta^*)} \right)^{\frac{\alpha}{\alpha+1}} \right) \qquad (45) \\
&\geqslant (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} \left(2\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})} - \left(2\sqrt{\eta^*(1-\eta^*)} \right)^{\frac{\alpha}{\alpha+1}} \right) \\
&\qquad \left(2\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})} \leqslant 1, \tfrac{\alpha}{\alpha+1} \leqslant 1 \text{ for } \alpha > 2 \right)
\end{aligned}$$

$$\geqslant (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} \left(2\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})} - \left(1 + \frac{\alpha}{\alpha+1}\left(2\sqrt{\eta^*(1-\eta^*)} - 1\right)\right)\right)$$

$$\left(x^{\frac{\alpha}{\alpha+1}} \text{ is concave in } x \geqslant 0\right)$$

$$= (\alpha+1)\alpha^{\frac{1}{\alpha+1}} c^{\frac{1}{\alpha+1}} \left(\frac{1}{\alpha}\left(2\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})} - 1\right) - \frac{1}{\alpha+1}\left(2\sqrt{\eta^*(1-\eta^*)} - 1\right)\right)$$

$$\geqslant (\alpha+1)\alpha^{\frac{1}{\alpha+1}} c^{\frac{1}{\alpha+1}} \left(\frac{1}{\alpha}\left(2\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})} - 2\sqrt{\eta^*(1-\eta^*)}\right)\right) \qquad \left(\tfrac{1}{\alpha} > \tfrac{1}{\alpha+1}\right)$$

$$= (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} \left(2\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})} - 2\sqrt{\eta^*(1-\eta^*)}\right)$$

$$\geqslant \frac{3}{4}\left(2\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})} - 2\sqrt{\eta^*(1-\eta^*)}\right) \qquad \text{(by Lemma C.5)}$$

$$= \frac{3}{2}\frac{\eta_{\tilde{y}}(1-\eta_{\tilde{y}}) - \eta^*(1-\eta^*)}{\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})} + \sqrt{\eta^*(1-\eta^*)}} \qquad (46)$$

$$= \frac{3}{2}\frac{(\eta^* - \eta_{\tilde{y}})(\eta^* + \eta_{\tilde{y}} - 1)}{\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})} + \sqrt{\eta^*(1-\eta^*)}}$$

$$\geqslant \frac{3}{4}(1-2c)\frac{\eta^* - \eta_{\tilde{y}}}{\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})} + \sqrt{\eta^*(1-\eta^*)}} \qquad \left(\eta^* > 1-c, \eta_{\tilde{y}} \geqslant \tfrac{1}{2}\right)$$

$$\geqslant \frac{3}{4}(1-2c)\left(\eta^* - \eta_{\tilde{y}}\right) \qquad \left(\eta^* \geqslant \eta_{\tilde{y}},\ \sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})} \leqslant \tfrac{1}{2}, \text{ and } \sqrt{\eta^*(1-\eta^*)} \leqslant \tfrac{1}{2}\right)$$

$$= W_c^{(1)}\left(\eta^* - \eta_{\tilde{y}}\right) \qquad (47)$$

$$= W_c^{(1)} W_{\text{0-1-}c}(r_{\boldsymbol{\eta}}, \boldsymbol{g}_{\boldsymbol{\eta}}; \boldsymbol{\eta}).$$

In Equation (45), consider function $I(r;\eta) = 2\sqrt{\eta(1-\eta)}e^r + ce^{-\alpha r}$, where $\frac{1}{2} \leqslant \eta \leqslant 1$. Let $r_\eta \in \operatorname{argmin}_r I(r;\eta)$ be a minimizer of $I(r;\eta)$. Then by Lemma C.3, we have $I(r_{\eta_{\tilde{y}}};\eta_{\tilde{y}}) = (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} \left(2\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})}\right)^{\frac{\alpha}{\alpha+1}}$. $W_c^{(1)} = \frac{3}{4}(1-2c)$ is a positive constant that depends on $c < \frac{1}{2}$.

Therefore,

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

where $\Gamma_1(t) = \sqrt{\frac{1}{W_c^{(1)}}}\sqrt{t}$ ($W_c^{(1)}$ is a positive constant that depends on c).

(**ii**) If $\frac{1}{2} \leqslant \eta^* \leqslant (1-c)$ and $r > 0$: In this case, by Equation (41) we have $\Delta W_{\text{0-1-}c}(r, \boldsymbol{g}; \boldsymbol{\eta}) = 1 - c - \eta_{\tilde{y}}$. Consider $\Delta W_{\text{MCS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) = W_{\text{MCS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) - W_{\text{MCS}}(r^\dagger_{\boldsymbol{\eta}}, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta})$, where $W_{\text{MCS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) \geqslant G(\Delta; \eta_{\tilde{y}})$ (by Equation (42)). For $G(\Delta; \eta_{\tilde{y}})$ we have $\inf_{\Delta \geqslant 0} G(\Delta; \eta_{\tilde{y}}) \geqslant 2\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})}$, where $\eta_{\tilde{y}} \geqslant \frac{1}{2}$ (by Equation (44)). By Lemma C.2, we have $W_{\text{MCS}}(r^\dagger_{\boldsymbol{\eta}}, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta}) = (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} \left(2\sqrt{\eta^*(1-\eta^*)}\right)^{\frac{\alpha}{\alpha+1}}$ for $\eta^* \geqslant \frac{1}{2}$. Thus we have

$$\begin{aligned}
&\Delta W_{\text{MCS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) \\
&= \sum_y \eta_y \exp\left(-\left(g_y - \max_{y' \neq y} g_{y'}\right)\right) e^r + ce^{-\alpha r} - (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} \left(2\sqrt{\eta^*(1-\eta^*)}\right)^{\frac{\alpha}{\alpha+1}} \\
&\geqslant \left(\eta_{\tilde{y}} \exp(-\Delta) + (1-\eta_{\tilde{y}})\exp(\Delta)\right) e^r + ce^{-\alpha r} - (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} \left(2\sqrt{\eta^*(1-\eta^*)}\right)^{\frac{\alpha}{\alpha+1}} \\
&\qquad \left(\textstyle\sum_y \eta_y \exp\left(-\left(g_y - \max_{y' \neq y} g_{y'}\right)\right) \geqslant G(\Delta; \eta_{\tilde{y}}) \text{ in Equation (42)}\right) \\
&\geqslant \eta_{\tilde{y}} \exp(-\Delta) + (1-\eta_{\tilde{y}})\exp(\Delta) + c - (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} \left(2\sqrt{\eta^*(1-\eta^*)}\right)^{\frac{\alpha}{\alpha+1}} \qquad (48)
\end{aligned}$$

$$
\begin{aligned}
&\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) \geqslant 2\sqrt{\eta_{\tilde{y}}(1-\eta_{\tilde{y}})}, \text{ where } \eta_{\tilde{y}} \geqslant \tfrac{1}{2} \text{ in Equation (44)}\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 (49) \\
&\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}, \text{ where } \alpha = 2\sqrt{\tfrac{1-c}{c}}\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 (50) \\
&= \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 (51) \\
&= \frac{2\alpha c(1-\alpha c)^2}{(\alpha+1)(1-2c)^2}\left(1-c-\eta_{\tilde{y}}\right)^2 \\
&= W_c^{(2)}\left(1-c-\eta_{\tilde{y}}\right)^2 \qquad (52) \\
&= W_c^{(2)} \Delta W_{\text{0-1-}c}(r, \boldsymbol{g}; \boldsymbol{\eta})^2,
\end{aligned}
$$

where in Equation (48) we used Lemma C.4 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 Equation (49), 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

$$
g(1) - g(x^*_{\eta_{\tilde{y}}}) \geqslant \frac{g''(1)}{2}(1 - x^*_{\eta_{\tilde{y}}})^2
$$

$$\geqslant \frac{c\alpha(\alpha+1)}{2}\left(1-\left(\frac{\alpha c}{S_{\eta_{\tilde{y}}}}\right)^{\frac{1}{\alpha+1}}\right)^2.$$

The proof of Equation (49) is completed.

Equation (50) 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 Equation (51). Equation (52) 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 $k(\eta_{\tilde{y}})=\frac{1}{(\alpha c-1)\eta_{\tilde{y}}+1-\frac{1}{2}\alpha c-c}$. $W_c^{(2)}=\frac{2c\alpha(1-\alpha c)^2}{(\alpha+1)(1-2c)^2}$ 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,

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

where $\Gamma_2(t)=\sqrt{\frac{1}{W_c^{(2)}}}\sqrt{t}$ ($W_c^{(2)}$ is a positive constant that depends on c).

(**iii**) If $\eta^*<\frac{1}{2}<(1-c)$ and $r>0$: In this case, by Equation (41) we have $\Delta W_{\text{0-1-}c}(r,\boldsymbol{g};\boldsymbol{\eta})=1-c-\eta_{\tilde{y}}$. Consider $\Delta W_{\text{MCS}}(r,\boldsymbol{g};\boldsymbol{\eta})=W_{\text{MCS}}(r,\boldsymbol{g};\boldsymbol{\eta})-W_{\text{MCS}}(r^\dagger_{\boldsymbol{\eta}},\boldsymbol{g}^\dagger_{\boldsymbol{\eta}};\boldsymbol{\eta})$, 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

$$\begin{aligned}
\Delta W_{\text{MCS}}(r_{\boldsymbol{\eta}},\boldsymbol{g}_{\boldsymbol{\eta}};\boldsymbol{\eta}) &= \sum_y \eta_y \exp\left(-\left(g_y-\max_{y'\neq y}g_{y'}\right)\right)e^r+ce^{-\alpha r}-(\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}}c^{\frac{1}{\alpha+1}}\\
&\geqslant \left(\eta_{\tilde{y}}\exp(-\Delta)+(1-\eta_{\tilde{y}})\exp(\Delta)\right)e^r+ce^{-\alpha r}-(\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}}c^{\frac{1}{\alpha+1}}\\
&\qquad\left(\textstyle\sum_y \eta_y\exp\left(-\left(g_y-\max_{y'\neq y}g_{y'}\right)\right)\geqslant G(\Delta;\eta_{\tilde{y}})\text{ in Equation (43)}\right)\\
&\geqslant \eta_{\tilde{y}}\exp(-\Delta)+(1-\eta_{\tilde{y}})\exp(\Delta)+c-(\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}}c^{\frac{1}{\alpha+1}} \qquad (53)\\
&\geqslant 1+c-(\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}}c^{\frac{1}{\alpha+1}}\\
&\qquad\left(\inf_{\Delta\geqslant 0}G(\Delta)=1\text{ for }\eta_{\tilde{y}}\leqslant\eta^*<\tfrac{1}{2}\text{ in Equation (43)}\right)\\
&= \frac{1+c-(\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}}c^{\frac{1}{\alpha+1}}}{1-c}(1-c)\\
&= W_c^{(3)}(1-c) \qquad (54)\\
&\geqslant W_c^{(3)}(1-c-\eta_{\tilde{y}}) \qquad (\eta_{\tilde{y}}\geqslant 0)\\
&= W_c^{(3)}\Delta W_{\text{0-1-}c}(r,\boldsymbol{g};\boldsymbol{\eta}),
\end{aligned}$$

where in Equation (53) we used Lemma C.4 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 Equation (54), $W_c^{(3)}=\frac{1+c-(\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}}c^{\frac{1}{\alpha+1}}}{1-c}$ is a positive constant that depends on $c<\frac{1}{2}$, where $\alpha=2\sqrt{\frac{1-c}{c}}$. Note that

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

$$
\begin{aligned}
&= c - \frac{\alpha c - 1}{\alpha + 1} \\
&= \frac{c+1}{\alpha+1} \\
&> 0. && (c, \alpha > 0)
\end{aligned}
$$

Therefore,

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

where $\Gamma_3(t) = \frac{1}{W_c^{(3)}} t$ ($W_c^{(3)}$ is a positive constant that depends on c).

(**iv**) If $\eta^* \leqslant (1-c)$ and $r \leqslant 0$: In this case, by Equation (41) we have $\Delta W_{\text{0-1-}c}(r_{\boldsymbol{\eta}}, \boldsymbol{g}_{\boldsymbol{\eta}}; \boldsymbol{\eta}) = 0$, which implies that $W_{\text{0-1-}c}(r_{\boldsymbol{\eta}}, \boldsymbol{g}_{\boldsymbol{\eta}}; \boldsymbol{\eta}) - W_{\text{0-1-}c}(r^*_{\boldsymbol{\eta}}, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta}) = \Delta W_{\text{0-1-}c}(r_{\boldsymbol{\eta}}, \boldsymbol{g}_{\boldsymbol{\eta}}; \boldsymbol{\eta}) = 0 \leqslant \Gamma\left(W_{\text{MCS}}(r_{\boldsymbol{\eta}}, \boldsymbol{g}_{\boldsymbol{\eta}}; \boldsymbol{\eta}) - W_{\text{MCS}}(r^\dagger_{\boldsymbol{\eta}}, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta})\right)$ for any $\Gamma \geqslant 0$.

(**v**) If $\eta^* > (1-c)$ and $r \leqslant 0$: In this case, by Equation (41) we have $\Delta W_{\text{0-1-}c}(r_{\boldsymbol{\eta}}, \boldsymbol{g}_{\boldsymbol{\eta}}; \boldsymbol{\eta}) = \eta^* - 1 + c$. Consider $\Delta W_{\text{MCS}}(r_{\boldsymbol{\eta}}, \boldsymbol{g}_{\boldsymbol{\eta}}; \boldsymbol{\eta}) = W_{\text{MCS}}(r_{\boldsymbol{\eta}}, \boldsymbol{g}_{\boldsymbol{\eta}}; \boldsymbol{\eta}) - W_{\text{MCS}}(r^\dagger_{\boldsymbol{\eta}}, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta})$, by Lemma C.2, we have $W_{\text{MCS}}(r^\dagger_{\boldsymbol{\eta}}, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta}) = (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}}\left(2\sqrt{\eta^*(1-\eta^*)}\right)^{\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}}$ for $\eta^* > (1-c) > \frac{1}{2}$. Thus we have

$$
\begin{aligned}
&\Delta W_{\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 + c e^{-\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 + c e^{-\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}} \\
&\qquad\qquad \text{(by Equation (12))} \\
&\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}} && (55) \\
&= S_{\eta^*} + c - (\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}}\left(S_{\eta^*}\right)^{\frac{\alpha}{\alpha+1}} c^{\frac{1}{\alpha+1}} && \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 && (56) \\
&= \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\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 && (57) \\
&= \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
\end{aligned}
$$

$$
\begin{aligned}
&\geqslant \frac{\alpha c}{8(\alpha+1)}\left(\frac{S_{\eta^*}}{\alpha c}-1\right)^2 && \left(S_{\eta^*}<2\sqrt{c(1-c)}=\alpha c\right)\\
&= \frac{1}{8c\alpha(\alpha+1)}\left(2\sqrt{\eta^*(1-\eta^*)}-\alpha c\right)^2 && \left(S_{\eta^*}=2\sqrt{\eta^*(1-\eta^*)}\right)\\
&\geqslant \frac{1}{8c\alpha(\alpha+1)}\left(\frac{4\eta^*(1-\eta^*)-\alpha^2c^2}{2\sqrt{\eta^*(1-\eta^*)}+\alpha c}\right)^2 &&\\
&\geqslant \frac{1}{32c^2\alpha^2(\alpha+1)}\left(4\eta^*(1-\eta^*)-\alpha^2c^2\right)^2 && \left(2\sqrt{\eta^*(1-\eta^*)}<2\sqrt{c(1-c)}=\alpha c\right)\\
&\geqslant \frac{1}{32c^2\alpha^2(\alpha+1)}\left(4(2c-1)\eta^*+4(2c-1)(c-1)\right)^2 && (58)\\
&= \frac{(1-2c)^2}{2c^2\alpha^2(\alpha+1)}\left(\eta^*-1+c\right)^2 &&\\
&= W_c^{(5)}\left(\eta^*-1+c\right)^2 && (59)\\
&= W_c^{(5)}\Delta W_{\text{0-1-}c}(r,\boldsymbol{g};\boldsymbol{\eta})^2, &&
\end{aligned}
$$

where in Equation (55) we used Lemma C.4 (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 Equation (56), 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 $h(x)\geqslant h(x^*_{\eta^*})=(\alpha+1)\alpha^{-\frac{\alpha}{\alpha+1}}c^{\frac{1}{\alpha+1}}\left(S_{\eta^*}\right)^{\frac{\alpha}{\alpha+1}}$. 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 Equation (56) is completed.

Equation (57) 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 Equation (58), we use the tangent line of $k(\eta^*;\alpha,c)=4\eta^*(1-\eta^*)-\alpha^2c^2$ 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 Equation (59), $W_c^{(5)}=\frac{(1-2c)^2}{2c^2\alpha^2(\alpha+1)}$ is a positive constant that depends on $c<\frac{1}{2}$, where $\alpha=2\sqrt{\frac{1-c}{c}}$.

Therefore,

$$
R_{\text{0-1-}c}(r,\boldsymbol{g};\boldsymbol{\eta})-R_{\text{0-1-}c}(r^*_{\boldsymbol{\eta}},\boldsymbol{g}^*_{\boldsymbol{\eta}};\boldsymbol{\eta})=\mathbb{E}\left[\Delta W_{\text{0-1-}c}(r,\boldsymbol{g};\boldsymbol{\eta})\right]
$$

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

where $\Gamma_5(t) = \sqrt{\frac{1}{W_c^{(5)}}}\sqrt{t}$ ($W_c^{(5)}$ is a positive constant that depends on c).

Overall, we obtain

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

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

D.5 Proof of Theorem 6.3

Theorem 6.3 (restated) The following excess risk bound holds for L_{ACS}:

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

for all $r, \boldsymbol{g}$ and any distribution $p(\boldsymbol{x}, y)$, where

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

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

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

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

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

$$
\begin{aligned}
&\sum_y \eta_y \max\left\{1 - \left(g_y - \max_{y' \neq y} g_{y'} - r\right), 0\right\} \\
&\geqslant \eta_{\tilde{y}} \max\{1 - \Delta + r, 0\} + (1 - \eta_{\tilde{y}}) \max\{1 + \Delta + r, 0\} := G(r, \Delta; \eta_{\tilde{y}}),
\end{aligned} \tag{60}
$$

where $\tilde{y} = f(\boldsymbol{g}) \in \operatorname{argmax}_y g_y$, $0 \leqslant \eta_{\tilde{y}} \leqslant 1$, and $\Delta \geqslant 0$. Then using Lemma C.6, we have

$$
\inf_{\Delta \geqslant 0} G(r, \Delta; \eta_{\tilde{y}}) = \begin{cases} 2(1 - \eta_{\tilde{y}})(r + 1) & \text{if } \eta_{\tilde{y}} > \frac{1}{2} \text{ and } r \geqslant -1, \\ r + 1 & \text{if } \eta_{\tilde{y}} \leqslant \frac{1}{2} \text{ and } r \geqslant -1, \\ 0 & \text{if } r < -1. \end{cases} \tag{61}
$$

Thus we have

$$
\begin{aligned}
\inf_{\Delta \geqslant 0} G(r, \Delta; \eta_{\tilde{y}}) &\geqslant \begin{cases} \min\{2(1 - \eta_{\tilde{y}})(r + 1), r + 1\} & \text{if } r \geqslant -1, \\ 0 & \text{if } r < -1. \end{cases} \\
&= \begin{cases} 2(1 - \eta_{\tilde{y}})(r + 1) & \text{if } r \geqslant -1, \\ 0 & \text{if } r < -1, \end{cases}
\end{aligned} \tag{62}
$$

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

Since ψ satisfy the calibration condition in Theorem 5.4, 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 Equation (41) we have $\Delta W_{0\text{-}1\text{-}c}(r, \boldsymbol{g}; \boldsymbol{\eta}) = \eta^* - \eta_{\tilde{y}}$. Consider $\Delta W_{\mathrm{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) = W_{\mathrm{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) - W_{\mathrm{ACS}}(r^\dagger_{\boldsymbol{\eta}}, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta})$, where $W_{\mathrm{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) \geqslant G(r, \Delta; \eta_{\tilde{y}})$ (by Equation (60)). For $G(r, \Delta; \eta_{\tilde{y}})$ we have $\inf_{\Delta \geqslant 0} G(r, \Delta; \eta_{\tilde{y}}) \geqslant 2(1 - \eta_{\tilde{y}})(r + 1)$,

where $\eta_{\tilde{y}} \geqslant \frac{1}{2}$, for $r > 0$ (by Equation (62)). By Lemma C.7, we have $W_{\mathrm{ACS}}(r^\dagger_{\boldsymbol{\eta}}, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta}) = 2(1-\eta^*)(r^\dagger_{\boldsymbol{\eta}}+1) + \psi(r^\dagger_{\boldsymbol{\eta}})$ for $\eta^* > (1-c) > \frac{1}{2}$. Thus we have

$$
\begin{aligned}
\Delta W_{\mathrm{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) &= W_{\mathrm{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) - W_{\mathrm{ACS}}(r^\dagger_{\boldsymbol{\eta}}, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta}) \\
&\geqslant G(r, \Delta; \eta_{\tilde{y}}) + ce^{-\alpha r} - \left(2(1-\eta^*)(r^\dagger_{\boldsymbol{\eta}}+1) + \psi(r^\dagger_{\boldsymbol{\eta}})\right) \\
&\qquad \left(\textstyle\sum_y \eta_y \max\left\{1 - (g_y - \max_{y' \neq y} g_{y'} - r), 0\right\} \geqslant G(\Delta, r; \eta_{\tilde{y}}) \text{ in Equation (60)}\right) \\
&\geqslant 2(1-\eta_{\tilde{y}})(r+1) + \psi(r) - \left(2(1-\eta^*)(r^\dagger_{\boldsymbol{\eta}}+1) + \psi(r^\dagger_{\boldsymbol{\eta}})\right) \\
&\qquad \left(\inf_{\Delta \geqslant 0} G(r, \Delta; \eta_{\tilde{y}}) = 2(1-\eta_{\tilde{y}})(r+1) \text{ where } \eta_{\tilde{y}} > \tfrac{1}{2} \text{ for } r > 0 \text{ in Equation (62)}\right) \\
&\geqslant \inf_{r>0} \left\{2(1-\eta_{\tilde{y}})(r+1) + \psi(r)\right\} - \left(2(1-\eta^*)(r^\dagger_{\boldsymbol{\eta}}+1) + \psi(r^\dagger_{\boldsymbol{\eta}})\right) \\
&:= f_1(\eta^*, \eta_{\tilde{y}}).
\end{aligned}
$$

For $\psi(r) = c\exp(-\alpha r)$, Let $I(r; \eta) = 2(1-\eta_{\tilde{y}})(r+1) + c\exp(-\alpha r)$, which is convex in r. Its derivative is expressed as $I'(r) = 2(1-\eta_{\tilde{y}}) - \alpha c \exp(-\alpha r)$. Let $r^\circ \in \mathbb{R}$ be such that $I'(r^\circ) = 0$. Then we see that

$$
r^\circ = \frac{1}{\alpha} \log\left(\frac{\alpha c}{2(1-\eta_{\tilde{y}})}\right).
$$

Due to the convexity, we have that $I'(r)$ is increasing for all $r > 0$. First consider the case $(1-\eta_{\tilde{y}}) < \frac{\alpha c}{2}$. In this case, we have $r^\circ > 0$. Then $I(r) \geqslant I(r^\circ) = 2(1-\eta_{\tilde{y}})\frac{1}{\alpha}\log\left(\frac{c}{1-\eta_{\tilde{y}}}\right) + 3(1-\eta_{\tilde{y}}),$ for $r > 0$.

Next consider the case $(1-\eta_{\tilde{y}}) \geqslant \frac{\alpha c}{2}$. In this case, we have $r^\circ \leqslant 0$. Then $I(r) > I(0) = 2(1-\eta_{\tilde{y}}) + c,$ for $r > 0$. Thus, combined with Theorem 5.4, we have

$$
\inf_{r>0} \left\{2(1-\eta_{\tilde{y}})(r+1) + \psi(r)\right\} = \begin{cases} (1-\eta_{\tilde{y}}) \log\left(\frac{c}{1-\eta_{\tilde{y}}}\right) + 3(1-\eta_{\tilde{y}}) & \text{if } \eta_{\tilde{y}} > 1-c, \\ 2(1-\eta_{\tilde{y}}) + c & \text{if } \eta_{\tilde{y}} \leqslant 1-c. \end{cases}
$$

By Theorem 5.4, we also have $2(1-\eta^*)(r^\dagger_{\boldsymbol{\eta}}+1) + \psi(r^\dagger_{\boldsymbol{\eta}}) = (1-\eta^*)\log\left(\frac{c}{1-\eta^*}\right) + 3(1-\eta^*)$.

Then we discuss cases of $\eta_{\tilde{y}}$ to bound $f_1(\eta^*, \eta_{\tilde{y}})$ further. First consider the case $\eta_{\tilde{y}} > 1-c$. In this case, we have

$$
\begin{aligned}
f_1(\eta^*, \eta_{\tilde{y}}) &= (1-\eta_{\tilde{y}}) \log\left(\frac{c}{1-\eta_{\tilde{y}}}\right) + 3(1-\eta_{\tilde{y}}) - \left(2(1-\eta^*)(r^\dagger_{\boldsymbol{\eta}}+1) + \psi(r^\dagger_{\boldsymbol{\eta}})\right) \\
&= (1-\eta_{\tilde{y}}) \log\left(\frac{c}{1-\eta_{\tilde{y}}}\right) - (1-\eta^*)\log\left(\frac{c}{1-\eta^*}\right) + 3(\eta^* - \eta_{\tilde{y}}) \\
&\qquad \left(r^\dagger_{\boldsymbol{\eta}} = \tfrac{1}{\alpha}\log\left(\tfrac{\alpha c}{2(1-\eta^*)}\right), \alpha = 2 \text{ by Theorem 5.4}\right) \\
&\geqslant (1-\eta^*)\log\left(\frac{1-\eta^*}{1-\eta_{\tilde{y}}}\right) + 3(\eta^* - \eta_{\tilde{y}}) \qquad (1-\eta_{\tilde{y}} < c, \eta^* \geqslant \eta_{\tilde{y}}) \\
&\geqslant (1-\eta^*)\left(1 - \frac{1-\eta_{\tilde{y}}}{1-\eta^*}\right) + 3(\eta^* - \eta_{\tilde{y}}) \qquad (63) \\
&= 2(\eta^* - \eta_{\tilde{y}}) \\
&= 2\Delta W_{\text{0-1-}c}(r, \boldsymbol{g}; \boldsymbol{\eta}),
\end{aligned}
$$

where Equation (63) follows from the fact that $\log x \geqslant 1 - \frac{1}{x}$ $(x > 0)$.

Next consider the case $\eta_{\tilde{y}} \leqslant 1-c$. In this case, we have

$$
\begin{aligned}
f_1(\eta^*, \eta_{\tilde{y}}) &= 2(1-\eta_{\tilde{y}}) + c - \left(2(1-\eta^*)(r^\dagger_{\boldsymbol{\eta}}+1) + \psi(r^\dagger_{\boldsymbol{\eta}})\right) \\
&\geqslant 2(1-\eta_{\tilde{y}}) + c - \inf_{r>0}\left\{2(1-\eta^*)(r+1) + \psi(r)\right\} \\
&\geqslant 2(1-\eta_{\tilde{y}}) + c - (2(1-\eta^*) + c) \qquad (\eta^* > 1-c) \\
&= 2(\eta^* - \eta_{\tilde{y}})
\end{aligned}
$$

$$= 2\Delta W_{\text{0-1-}c}(r, \boldsymbol{g}; \boldsymbol{\eta}).$$

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

$$\begin{aligned}
f_1(\eta^*, \eta_{\tilde{y}}) &= \min\left\{2(1-\eta_{\tilde{y}})+c, 2(1-\eta_{\tilde{y}})(\frac{1}{\alpha}+1)\right\} - 2(1-\eta^*)\cdot\left(\frac{1}{\alpha}+1\right) \\
&= \min\{2(1-\eta_{\tilde{y}})+c, 3(1-\eta_{\tilde{y}})\} - 3(1-\eta^*) \qquad (\alpha = 2 \text{ from Theorem 5.4}) \\
&\geqslant \min\{2(\eta^*-\eta_{\tilde{y}})+(\eta^*-1+c), 3(\eta^*-\eta_{\tilde{y}})\} \\
&\geqslant \min\{2(\eta^*-\eta_{\tilde{y}}), 3(\eta^*-\eta_{\tilde{y}})\} \\
&= 2\Delta W_{\text{0-1-}c}(r, \boldsymbol{g}; \boldsymbol{\eta}).
\end{aligned}$$

Therefore,

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

where $\Gamma_1(t) = \frac{1}{2}t$ for $\psi(r) = c\exp(-\alpha r)$, or $\Gamma_1(t) = \frac{1}{2}t$ 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 Equation (41) we have $\Delta W_{\text{0-1-}c}(r, \boldsymbol{g}; \boldsymbol{\eta}) = 1-c-\eta_{\tilde{y}}$. Consider $\Delta W_{\text{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) = W_{\text{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) - W_{\text{ACS}}(r^\dagger_{\boldsymbol{\eta}}, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta})$, where $W_{\text{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) \geqslant G(\Delta, r; \eta_{\tilde{y}})$ (by Equation (60)). 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 $r > 0$ (by Equation (62)). By Lemma C.7, we have $W_{\text{ACS}}(r^\dagger_{\boldsymbol{\eta}}, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta}) = 2(1-\eta^*)(r^\dagger_{\boldsymbol{\eta}}+1) + \psi(r^\dagger_{\boldsymbol{\eta}})$. Thus we have

$$\begin{aligned}
\Delta W_{\text{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) &= W_{\text{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) - W_{\text{ACS}}(r^\dagger_{\boldsymbol{\eta}}, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta}) \\
&\geqslant G(r, \Delta; \eta_{\tilde{y}}) + ce^{-\alpha r} - \left(2(1-\eta^*)(r^\dagger_{\boldsymbol{\eta}}+1) + \psi(r^\dagger_{\boldsymbol{\eta}})\right) \\
&\quad \left(\textstyle\sum_y \eta_y \max\{1-(g_y - \max_{y'\neq y} g_{y'} - r), 0\} \geqslant G(\Delta, r; \eta_{\tilde{y}}) \text{ in Equation (60)}\right) \\
&\geqslant 2(1-\eta_{\tilde{y}})(r+1) + \psi(r) - \left(2(1-\eta^*)(r^\dagger_{\boldsymbol{\eta}}+1) + \psi(r^\dagger_{\boldsymbol{\eta}})\right) \\
&\quad \left(\inf_{\Delta\geqslant 0} G(r, \Delta; \eta_{\tilde{y}}) \geqslant 2(1-\eta_{\tilde{y}})(r+1) \text{ where } \eta_{\tilde{y}} > \tfrac{1}{2} \text{ for } r > 0 \text{ in Equation (62)}\right) \\
&\geqslant 2(1-\eta_{\tilde{y}}) + \psi(0) - \left(2(1-\eta^*)(r^\dagger_{\boldsymbol{\eta}}+1) + \psi(r^\dagger_{\boldsymbol{\eta}})\right) \qquad (64) \\
&\geqslant 2(1-\eta_{\tilde{y}}) + \psi(0) - \left(2(1-\eta_{\tilde{y}})(r_{\eta_{\tilde{y}}}+1) + \psi(r_{\eta_{\tilde{y}}})\right) \qquad (65) \\
&\geqslant -2(1-\eta_{\tilde{y}})r_{\eta_{\tilde{y}}} + \psi(0) - \psi(r_{\eta_{\tilde{y}}}) \\
&:= f_2(\eta_{\tilde{y}}),
\end{aligned}$$

where in Equation (64) we used Lemma C.9 (note that $2(1-\eta_{\tilde{y}}) \geqslant 2(1-\eta^*) \geqslant 2c = -\psi'(0)$). In Equation (65), consider function $I(r; \eta) = 2(1-\eta)(r+1) + \psi(r)$, where $\eta > \frac{1}{2}$ and $r > 0$. Let $r_\eta \in \operatorname{argmin}\ I(r; \eta)$ be a minimizer of $I(r; \eta)$. Then by Lemma C.8, we have

$$I(r_{\eta_{\tilde{y}}}; \eta_{\tilde{y}}) = 2(1-\eta_{\tilde{y}})\left(r_{\eta_{\tilde{y}}}+1\right) + \psi\left(r_{\eta_{\tilde{y}}}\right),$$

where

$$r_{\eta_{\tilde{y}}} = (\psi')^{-1}\left(-2(1-\eta_{\tilde{y}})\right),$$

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

$$r_{\eta_{\tilde{y}}} = \begin{cases} -1 & \text{if } \frac{1}{2} < \eta_{\tilde{y}} \leqslant 1 - \frac{\alpha c}{2}, \\ \frac{1}{\alpha} & \text{if } \eta_{\tilde{y}} > \frac{1}{2} \text{ and } \eta_{\tilde{y}} > 1 - \frac{\alpha c}{2}, \end{cases}$$

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

For $\psi(r) = c\exp(-\alpha r)$, $r_{\eta_{\tilde{y}}} = \frac{1}{\alpha}\log\left(\frac{\alpha c}{2(1-\eta_{\tilde{y}})}\right)$. Then we have

$$f_2(\eta_{\tilde{y}}) = -2(1-\eta_{\tilde{y}})\cdot\frac{1}{\alpha}\log\left(\frac{\alpha c}{2(1-\eta_{\tilde{y}})}\right) + c - (1-\eta_{\tilde{y}})$$

$$
\begin{aligned}
&= (1-\eta_{\tilde{y}}) \log\left(\frac{1-\eta_{\tilde{y}}}{c}\right) - (1-\eta_{\tilde{y}}) + c && (\alpha = 2 \text{ from Theorem 5.4}) \\
&\geqslant (1-c-\eta_{\tilde{y}})^2 && (\text{by Lemma C.10}) \\
&= \Delta W_{\text{0-1-}c}(r, \boldsymbol{g}; \boldsymbol{\eta})^2.
\end{aligned}
$$

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

$$
\begin{aligned}
f_2(\eta_{\tilde{y}}) &= -2(1-\eta_{\tilde{y}}) \cdot (-1) + c - (\alpha+1)c \\
&= 2(1-\eta_{\tilde{y}}) + c - 3c && (\alpha = 2 \text{ from Theorem 5.4}) \\
&= 2(1-c-\eta_{\tilde{y}}) \\
&= 2\Delta W_{\text{0-1-}c}(r, \boldsymbol{g}; \boldsymbol{\eta}).
\end{aligned}
$$

Therefore,

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

where $\Gamma_2(t) = \sqrt{t}$ for $\psi(r) = c\exp(-\alpha r)$, or $\Gamma_2(t) = \frac{1}{2}t$ 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 Equation (41) we have $\Delta W_{\text{0-1-}c}(r, \boldsymbol{g}; \boldsymbol{\eta}) = 1 - c - \eta_{\tilde{y}}$. Consider $\Delta W_{\text{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) = W_{\text{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) - W_{\text{ACS}}(r^\dagger_{\boldsymbol{\eta}}, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta})$, by Lemma C.7, we have $W_{\text{ACS}}(r^\dagger_{\boldsymbol{\eta}}, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta}) = (r^\dagger_{\boldsymbol{\eta}} + 1) + \psi(r^\dagger_{\boldsymbol{\eta}})$ for $\eta^* \leqslant \frac{1}{2}$. Thus we have

$$
\begin{aligned}
\Delta W_{\text{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) &= W_{\text{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) - W_{\text{ACS}}(r^\dagger_{\boldsymbol{\eta}}, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta}) \\
&\geqslant G(r, \Delta; \eta_{\tilde{y}}) + ce^{-\alpha r} - \left((r^\dagger_{\boldsymbol{\eta}} + 1) + \psi(r^\dagger_{\boldsymbol{\eta}})\right) \\
&\quad \left(\textstyle\sum_y \eta_y \max\{1 - (g_y - \max_{y' \neq y} g_{y'} - r), 0\} \geqslant G(\Delta, r; \eta_{\tilde{y}}) \text{ in Equation (60)}\right) \\
&\geqslant (r+1) + \psi(r) - \left((r^\dagger_{\boldsymbol{\eta}} + 1) + \psi(r^\dagger_{\boldsymbol{\eta}})\right) \\
&\quad \left(\inf_{\Delta \geqslant 0} G(r, \Delta; \eta_{\tilde{y}}) = r + 1 \text{ for } \eta_{\tilde{y}} \leqslant \eta^* \leqslant \tfrac{1}{2}, r > 0 \text{ in (61)}\right) \\
&\geqslant 1 + \psi(0) - \left((r^\dagger_{\boldsymbol{\eta}} + 1) + \psi(r^\dagger_{\boldsymbol{\eta}})\right) && (66) \\
&= \psi(0) - \left(r^\dagger_{\boldsymbol{\eta}} + \psi(r^\dagger_{\boldsymbol{\eta}})\right) \\
&:= f_3(c),
\end{aligned}
$$

where in (66), we used Lemma C.9 (note that $1 > 2c = -\psi'(0)$).

For $\psi(r) = c\exp(-\alpha r)$, $r^\dagger_{\boldsymbol{\eta}} = -\frac{1}{\alpha}\log\frac{1}{\alpha c}$. Then we have

$$
\begin{aligned}
f_3(c) &= c - \left(-\frac{1}{\alpha}\log\frac{1}{\alpha c} + \frac{1}{\alpha}\right) \\
&= c - \left(\frac{1}{2}\log(2c) + \frac{1}{2}\right) && (\alpha = 2) \\
&= \frac{c - \frac{1}{2}\log(2c) - \frac{1}{2}}{1-c} 1 - c \\
&= W_c^{(1)}(1-c) && (67) \\
&\geqslant W_c^{(1)}(1-c-\eta_{\tilde{y}}) && (\eta_{\tilde{y}} \geqslant 0) \\
&= W_c^{(1)} \Delta W_{\text{0-1-}c}(r, \boldsymbol{g}; \boldsymbol{\eta}),
\end{aligned}
$$

where in (67), $W_c^{(1)} = \frac{c - \frac{1}{2}\log(2c) - \frac{1}{2}}{1-c}$ is a positive constant that depends on $c < 1/2$ (note that $c - \frac{1}{2}\log(2c) - \frac{1}{2} > c - \frac{1}{2}(2c-1) - \frac{1}{2} = 0$, where we use the relationship $\log x \leqslant x - 1$ $(x > 0)$).

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

$$
f_3(c) = c - (-1 + (\alpha+1)c)
$$

$$
\begin{aligned}
&= 1 - 2c && (\alpha = 2)\\
&= \frac{1-2c}{1-c} 1 - c \\
&= W_c^{(2)}(1-c) && (68)\\
&\geqslant W_c^{(2)}(1 - c - \eta_{\tilde{y}}) && (\eta_{\tilde{y}} \geqslant 0)\\
&= W_c^{(2)} \Delta W_{\text{0-1-}c}(r, \boldsymbol{g}; \boldsymbol{\eta}),
\end{aligned}
$$

where in (68), $W_c^{(2)} = 1 - 2c$ is a positive constant that depends on $c < 1/2$.

Therefore,

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

where $\Gamma_3(t) = \frac{1}{W_c^{(1)}} t$ for $\psi(r) = c\exp(-\alpha r)$, or $\Gamma_3(t) = \frac{1}{W_c^{(2)}} t$ 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 Equation (41) we have $\Delta W_{\text{0-1-}c}(r, \boldsymbol{g}; \boldsymbol{\eta}) = 0$, which implies that $W_{\text{0-1-}c}(r, \boldsymbol{g}; \boldsymbol{\eta}) - W_{\text{0-1-}c}(r^*_{\boldsymbol{\eta}}, \boldsymbol{g}^*_{\boldsymbol{\eta}}; \boldsymbol{\eta}) = \Delta W_{\text{0-1-}c}(r, \boldsymbol{g}; \boldsymbol{\eta}) = 0 \leqslant \Gamma\left(W_{\text{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) - W_{\text{ACS}}(r^\dagger_{\boldsymbol{\eta}}, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta})\right)$ for any $\Gamma \geqslant 0$.

(**v**) If $\eta^* > (1-c)$ and $-1 \leqslant r \leqslant 0$: In this case, by Equation (41) we have $\Delta W_{\text{0-1-}c}(r, \boldsymbol{g}; \boldsymbol{\eta}) = \eta^* - 1 + c$. Consider $\Delta W_{\text{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) = W_{\text{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) - W_{\text{ACS}}(r^\dagger_{\boldsymbol{\eta}}, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta})$, by Lemma C.7, we have $W_{\text{ACS}}(r^\dagger_{\boldsymbol{\eta}}, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta}) = 2(1-\eta^*)(r^\dagger_{\boldsymbol{\eta}} + 1) + \psi(r^\dagger_{\boldsymbol{\eta}})$ for $\eta^* > (1-c) > \frac{1}{2}$. Thus we have

$$
\begin{aligned}
\Delta W_{\text{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) &= W_{\text{ACS}}(r, \boldsymbol{g}; \boldsymbol{\eta}) - W_{\text{ACS}}(r^\dagger_{\boldsymbol{\eta}}, \boldsymbol{g}^\dagger_{\boldsymbol{\eta}}; \boldsymbol{\eta}) \\
&\geqslant 2(1-\eta^*)(r+1) + \psi(r) - \left(2(1-\eta^*)(r^\dagger_{\boldsymbol{\eta}} + 1) + \psi(r^\dagger_{\boldsymbol{\eta}})\right) \\
&\qquad \text{(by Equation (16))} \\
&\geqslant 2(1-\eta^*) + \psi(0) - \left(2(1-\eta^*)(r^\dagger_{\boldsymbol{\eta}} + 1) + \psi(r^\dagger_{\boldsymbol{\eta}})\right) && (69)\\
&= -2(1-\eta^*) r^\dagger_{\boldsymbol{\eta}} + \psi(0) - \psi(r^\dagger_{\boldsymbol{\eta}}) \\
&:= f_5(\eta^*),
\end{aligned}
$$

where in (69), we used Lemma C.9 (note that $2(1-\eta^*) < 2c = -\psi'(0)$).

For $\psi(r) = c\exp(-\alpha r)$, $r^\dagger_{\boldsymbol{\eta}} = \frac{1}{\alpha}\log\left(\frac{\alpha c}{2(1-\eta^*)}\right)$. Then we have

$$
\begin{aligned}
f_5(\eta^*) &= -2(1-\eta^*) \cdot \frac{1}{\alpha} \log\left(\frac{\alpha c}{2(1-\eta^*)}\right) + c - (1-\eta^*) \\
&= (1-\eta^*)\log\left(\frac{1-\eta^*}{c}\right) - (1-\eta^*) + c && (\alpha = 2)\\
&\geqslant (1 - \eta^* - c)^2 && \text{(by Lemma C.10)} \\
&= (\eta^* - 1 + c)^2 \\
&= \Delta W_{\text{0-1-}c}(r, \boldsymbol{g}; \boldsymbol{\eta})^2.
\end{aligned}
$$

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

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

Therefore,

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

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

(**vi**) If $\eta^*>(1-c)$ and $r<-1$: In this case, by Equation (41) we have $\Delta W_{\text{0-1-}c}(r,\boldsymbol{g};\boldsymbol{\eta})=\eta^*-1+c$. For $\Delta W_{\text{ACS}}(r,\boldsymbol{g};\boldsymbol{\eta})=W_{\text{ACS}}(r,\boldsymbol{g};\boldsymbol{\eta})-W_{\text{ACS}}(r^\dagger_{\boldsymbol{\eta}},\boldsymbol{g}^\dagger_{\boldsymbol{\eta}};\boldsymbol{\eta})$, we have

$$\begin{aligned}\Delta W_{\text{ACS}}(r,\boldsymbol{g};\boldsymbol{\eta}) &= W_{\text{ACS}}(r,\boldsymbol{g};\boldsymbol{\eta})-W_{\text{ACS}}(r^\dagger_{\boldsymbol{\eta}},\boldsymbol{g}^\dagger_{\boldsymbol{\eta}};\boldsymbol{\eta})\\ &\geqslant \psi(r)-\left(2(1-\eta^*)(r^\dagger_{\boldsymbol{\eta}}+1)+\psi(r^\dagger_{\boldsymbol{\eta}})\right) \qquad \text{(by Equation (16))}\\ &\geqslant \psi(-1)-\left(2(1-\eta^*)(r^\dagger_{\boldsymbol{\eta}}+1)+\psi(r^\dagger_{\boldsymbol{\eta}})\right) \quad (\psi(r) \text{ is nonincreasing for } r<-1)\\ &= \psi(-1)-2(1-\eta^*)(r^\dagger_{\boldsymbol{\eta}}+1)-\psi(r^\dagger_{\boldsymbol{\eta}})\\ &:= f_6(\eta^*).\end{aligned}$$

For $\psi(r)=c\exp(-\alpha r)$, $r^\dagger_{\boldsymbol{\eta}}=\frac{1}{\alpha}\log\left(\frac{\alpha c}{2(1-\eta^*)}\right)$. Then we have

$$\begin{aligned}f_6(\eta^*) &= ce^{\alpha}-2(1-\eta^*)\left(\frac{1}{\alpha}\log\left(\frac{\alpha c}{2(1-\eta^*)}\right)+1\right)-(1-\eta^*)\\ &\geqslant c(\alpha+1)-2(1-\eta^*)\left(\frac{1}{\alpha}\log\left(\frac{\alpha c}{2(1-\eta^*)}\right)+1\right)-(1-\eta^*) \qquad (e^x\geqslant x+1)\\ &= 3c+(1-\eta^*)\log\left(\frac{1-\eta^*}{c}\right)-3(1-\eta^*) \qquad (\alpha=2)\\ &= c\frac{(1-\eta^*)}{c}\log\left(\frac{1-\eta^*}{c}\right)-3(1-\eta^*)+3c\\ &\geqslant c\left(\frac{1-\eta^*}{c}-1\right)-3(1-\eta^*)+3c \qquad (70)\\ &= 2(\eta^*-1+c)\\ &= 2\Delta W_{\text{0-1-}c}(r,\boldsymbol{g};\boldsymbol{\eta}),\end{aligned}$$

where (70) follows from the fact that $x\log x\geqslant x-1$ $(x>0)$.

For $\psi(r)=c\max\{1-\alpha r,0\}$, $r^\dagger_{\boldsymbol{\eta}}=\frac{1}{\alpha}$. 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,

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

where $\Gamma_6(t)=\frac{1}{2}t$ for $\psi(r)=c\exp(-\alpha r)$, or $\Gamma_6(t)=\frac{1}{3}t$ for $\psi(r)=c\max\{1-\alpha r,0\}$.

Overall, we obtain

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

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

D.6 Proof of Theorem 7.2

Theorem 7.2 (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 $p(\boldsymbol{x}, y)$, respectively. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$, the following multi-class classification generalization bounds hold for all $g \in \mathfrak{G}$:*

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

Proof. Firstly, we prove the generalization bound for L_{MCS}. To begin with, we 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]
\end{aligned}
$$

$$\leqslant \frac{M_\phi M_{\psi_1} + M_{\psi_2}}{n},$$

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

$$\begin{aligned}
&\sup_{g_1\cdots, g_K\in\mathcal{G}} (R_{\mathrm{MCS}}(\boldsymbol{g}) - \widehat{R}_{\mathrm{MCS}}(\boldsymbol{g})) \\
&\leqslant \mathbb{E}_S\left[\sup_{g_1\cdots, g_K\in\mathcal{G}} (R_{\mathrm{MCS}}(\boldsymbol{g}) - \widehat{R}_{\mathrm{MCS}}(\boldsymbol{g}))\right] + c\sqrt{\frac{n}{2}\log\left(\frac{1}{\delta}\right)} \\
&\qquad \text{(McDiarmid inequality, see Shalev-Shwartz \& Ben-David, 2014, Lemma 26.4)} \\
&\leqslant \mathbb{E}_S\left[\sup_{g_1\cdots, g_K\in\mathcal{G}} (R_{\mathrm{MCS}}(\boldsymbol{g}) - \widehat{R}_{\mathrm{MCS}}(\boldsymbol{g}))\right] + (M_\phi M_{\psi_1} + M_{\psi_2})\sqrt{\frac{1}{2n}\log\left(\frac{1}{\delta}\right)} \\
&\qquad \left(c = \tfrac{M_\phi M_{\psi_i}+M_{\psi_2}}{n}\right) \\
&\leqslant 2\mathfrak{R}_n(\mathcal{L}_{\mathrm{MCS}}) + (M_\phi M_{\psi_1} + M_{\psi_2})\sqrt{\frac{1}{2n}\log\left(\frac{1}{\delta}\right)} \\
&\qquad \text{(Shalev-Shwartz \& Ben-David, 2014, Lemma 26.2; } \mathcal{L}_{\mathrm{MCS}} \text{ is defined in Equation (29))} \\
&\leqslant (2M_\phi + M_{\psi_1})\, L_\phi K^2 \mathfrak{R}_n(\mathcal{G}) + ((2M_\phi + M_{\psi_1})\, L_{\psi_1} + 2L_{\psi_2})\, \mathfrak{R}_n(\mathcal{R}) \\
&+ (M_\phi M_{\psi_1} + M_{\psi_2})\sqrt{\frac{1}{2n}\log\left(\frac{1}{\delta}\right)}. \qquad \text{(by Lemma C.15)}
\end{aligned}$$

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

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

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 prove the generalization bound for L_{ACS}. To begin with, we 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 the case of L_{MCS} with L_{MCS} replaced with L_{ACS}, we have

$$\begin{aligned}
&A(z_1, \cdots, z_i, \cdots, z_n) - A(z_1, \cdots, z_i', \cdots, z_n) \\
&= \sup_{r\in\mathcal{R}, \boldsymbol{g}\in\mathfrak{G}} \left[\frac{1}{n} L_{\mathrm{ACS}}(r, \boldsymbol{g}; \boldsymbol{x}_i', y_i') - \frac{1}{n} L_{\mathrm{ACS}}(r, \boldsymbol{g}; \boldsymbol{x}_i, y_i)\right].
\end{aligned} \tag{71}$$

Then we have

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

Table 4: Performance of our method and the compared methods on CIFAR-10.

METHOD	REJECTION RATIO	MISCLASSIFICATION ERROR	ABSTENTION LOSS
CE (NO REJ.)	0.00%	10.45%	10.45%
CS (NO REJ.)	0.00%	16.22%	16.23%
RAMASWAMY ET AL. (2018) (CS-HINGE)	64.25%	7.58%	9.13%
RAMASWAMY ET AL. (2018) (BEP)	58.77%	13.91%	11.61%
MAO ET AL. (2024A) (TWO STAGE)	75.13%	7.81%	9.46%
OURS (MCS-HINGE)	64.54%	7.10%	8.97%
OURS (ACS-HINGE)	61.68%	5.96%	**8.45%**
OURS (MCS-LOGISTIC)	82.29%	6.02%	9.30%
OURS (ACS-LOGISTIC)	100.00%	–	10.00%

$$\leqslant \frac{M_\phi + M_\psi}{n},$$

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

$$\begin{aligned}
&\sup_{g_1\cdots,g_K\in\mathcal{G}} (R_{\mathrm{ACS}}(\boldsymbol{g}) - \widehat{R}_{\mathrm{ACS}}(\boldsymbol{g})) \\
&\leqslant \mathbb{E}_S\left[\sup_{g_1\cdots,g_K\in\mathcal{G}} (R_{\mathrm{ACS}}(\boldsymbol{g}) - \widehat{R}_{\mathrm{ACS}}(\boldsymbol{g}))\right] + c\sqrt{\frac{n}{2}\log\left(\frac{1}{\delta}\right)} \\
&\qquad \text{(McDiarmid inequality, see Shalev-Shwartz \& Ben-David, 2014, Lemma 26.4)} \\
&\leqslant \mathbb{E}_S\left[\sup_{g_1\cdots,g_K\in\mathcal{G}} (R_{\mathrm{ACS}}(\boldsymbol{g}) - \widehat{R}_{\mathrm{ACS}}(\boldsymbol{g}))\right] + (M_\phi+M_\psi)\sqrt{\frac{1}{2n}\log\left(\frac{1}{\delta}\right)} \qquad \left(c = \tfrac{M_\phi+M_\psi}{n}\right) \\
&\leqslant 2\mathfrak{R}_n(\mathcal{L}_{\mathrm{ACS}}) + (M_\phi+M_\psi)\sqrt{\frac{1}{2n}\log\left(\frac{1}{\delta}\right)} \\
&\qquad \text{(Shalev-Shwartz \& Ben-David, 2014, Lemma 26.2; } \mathcal{L}_{\mathrm{ACS}} \text{ is defined in Equation (30))} \\
&\leqslant 2L_\phi K^2\mathfrak{R}_n(\mathcal{G}) + 2\left(L_\phi+L_\psi\right)\mathfrak{R}_n(\mathcal{R}) \\
&+ (M_\phi+M_\psi)\sqrt{\frac{1}{2n}\log\left(\frac{1}{\delta}\right)}. \qquad \text{(by Lemma C.16)}
\end{aligned}$$

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. Specifically, for Mao et al. 2024a (two-stage), we split non-overlapping datasets for both stages and train for 200 epochs for each non-overlapping dataset. We adopt ResNet-32 (He et al., 2016), a 32-layer deep residual network with ReLU activations for CIFAR-10 and Imagenette. We adopt a one-hidden-layer neural network with a

Table 5: Performance of our method and comparison methods on the satimage dataset.

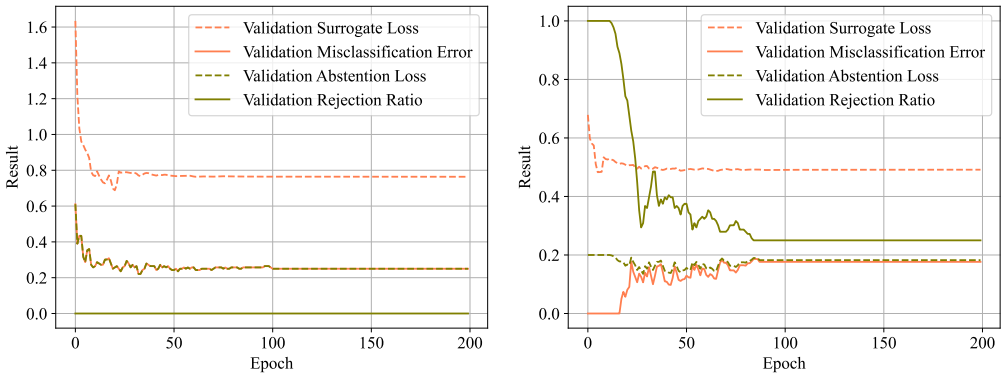
METHOD	REJECTION RATIO	MISCLASSIFICATION ERROR	ABSTENTION LOSS
CE (NO REJ.)	0.00%	10.33%	10.33%
CS (NO REJ.)	0.00%	10.72%	10.72%
RAMASWAMY ET AL. (2018) (CS-HINGE)	8.00%	7.26%	7.48%
RAMASWAMY ET AL. (2018) (BEP)	3.50%	10.23%	10.22%
MAO ET AL. (2024A) (TWO-STAGE)	43.90%	2.91%	6.02%
OURS (MCS-HINGE)	30.46%	4.25%	**6.00%**
OURS (ACS-HINGE)	67.13%	0.24%	6.79%
OURS (MCS-LOGISTIC)	61.69%	1.42%	6.71%
OURS (ACS-LOGISTIC)	99.92%	–	9.99%

Table 6: Performance of our method and comparison methods on the yeast dataset.

METHOD	REJECTION RATIO	MISCLASSIFICATION ERROR	ABSTENTION LOSS
CE (NO REJ.)	0.00%	40.40%	40.40%
CS (NO REJ.)	0.00%	41.41%	41.41%
RAMASWAMY ET AL. (2018) (CS-HINGE)	20.88%	34.89%	35.96%
RAMASWAMY ET AL. (2018) (BEP)	10.44%	37.97%	38.18%
MAO ET AL. (2024A) (TWO-STAGE)	70.71%	27.59%	36.36%
OURS (MCS-HINGE)	55.56%	29.55%	35.35%
OURS (ACS-HINGE)	96.97%	22.22%	39.46%
OURS (MCS-LOGISTIC)	33.00%	34.67%	36.43%
OURS (ACS-LOGISTIC)	29.97%	33.17%	**35.22%**

ReLU activation for the vehicle, satimage, and yeast datasets. We train for 200 epochs using the Adam method (Kingma & Ba, 2015). During 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 the vehicle, satimage, and *yeast* datasets, the batch size is set to 128. For CIFAR-10 and Imagenette, 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. For CIFAR-10, we adopt the following data augmentation strategy for training: 4-pixel padding on each side and a 32×32 random crop sampled from the padded image or its horizontal flip. For Imagenette, except that the padding is 0 and the images are resized to a size of 64×64 beforehand, other augmentations are the same as CIFAR-10. We set the abstention cost c to 0.2 for Imagenette and the vehicle dataset, 0.4 for the yeast dataset, and 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 with no abstention, which can encourage a reasonable number of samples to be abstained (Mao et al., 2024a). For a given dataset and abstention cost c, we can calculate its corresponding calibration parameter α for MCS-hinge, ACS-hinge, MCS-logistic, and ACS-logistic via Theorems 5.3 to 5.5 (combined with Equation (7)). For CIFAR-10, we have $\alpha = 2$, $\alpha = 2$, $\alpha = 3.25$, and $\alpha = 3.6$ for MCS-hinge, ACS-hinge, MCS-logistic, and ACS-logistic, respectively; For Imagenette, we have $\alpha = 2$, $\alpha = 2$, $\alpha = 2.5$, and $\alpha = 3.2$ for MCS-hinge, ACS-hinge, MCS-logistic, and ACS-logistic, respectively; or vehicle, we have $\alpha = 2$, $\alpha = 2$, $\alpha = 2.5$, and $\alpha = 3.2$ for MCS-hinge, ACS-hinge, MCS-logistic, and ACS-logistic, respectively; For satimage dataset, we have $\alpha = 2$, $\alpha = 2$, $\alpha = 3.25$, and $\alpha = 3.6$ for MCS-hinge, ACS-hinge, MCS-logistic, and ACS-logistic, respectively; For yeast dataset, we have $\alpha = 2$, $\alpha = 2$, $\alpha = 1.68$, and $\alpha = 2.4$ for MCS-hinge, ACS-hinge, MCS-logistic, and ACS-logistic, respectively. The GPU for our experiments is an NVIDIA RTX A6000 Ada Generation with 48 GB of memory.

Performance on three metrics. The performance of our method and comparison methods on three metrics on CIFAR-10, satimage, and yeast datasets from the UCI Machine Learning Repository are shown in Tables 4 to 6. The abstention losses among all the methods with abstention are boldfaced. We can observe from Tables 4 to 6 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 our calibration strategy via abstention. Besides, as displayed in Tables 4 to 6, our MCS-hinge loss or ACS-hinge shows the best performance in misclassification error and abstention loss among all the comparison methods on CIFAR-10 and the satimage dataset (with comparatively lower c), and



(a) Performance of CS (no rejection).

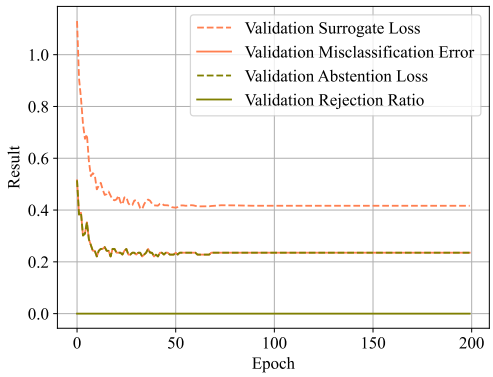
(b) Performance of our method (MCS-logistic) ($c = 0.2$, $\alpha = 2.5$).

Figure 2: The training dynamics of CS and ours (MCS-logistic) on the vehicle dataset.

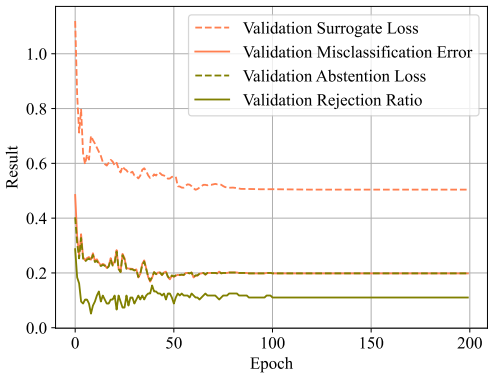
our ACS-logistic shows the best performance in misclassification error and abstention loss among all the comparison methods on the yeast dataset (with comparatively higher c). This validates the superiority of calibrating general CS losses via the predictor-learning framework and the potential for the non-polyhedral CS losses to be calibrated in the setting of higher abstention cost. Note that we speculate that the logistic variants more conservatively accept low-confidence samples, so that in the case of comparatively lower c (e.g., CIFAR-10 and satimage), they might be underconfident and choose to abstain to obtain lower losses. Thus, we recommend choosing logistic variants in the setting of comparatively higher ($\geqslant 0.2$) c while choosing the hinge variants otherwise.

Training dynamics. The training dynamics of CS and our MCS-logistic loss on the vehicle dataset are shown in Figure 2. For the case of CS, as shown in Figure 2(a), as the validation surrogate loss decreases, the validation abstention loss also tends to decrease. And because in this case the rejection ratio is 0, the validation abstention loss equals the validation misclassification error. For the case of MCS-logistic, as shown in Figure 2(b), with the decrease in validation surrogate loss, validation abstention loss decreases at early epochs and tends to keep a comparatively stable level in the late epochs, which validates the calibration of our MCS-logistic loss. The training dynamics of CE, Ramaswamy et al. 2018 (CS), Mao et al. 2024a, and ours (ACS-logistic) on the vehicle dataset are shown in Figure 3. For the case of CE, as shown in Figure 3(a), as the validation surrogate loss decreases, validation abstention loss also tends to decrease. And because in this case the rejection ratio is 0, the validation abstention loss equals the validation misclassification error. For the case of Ramaswamy et al. 2018 (CS), as shown in Figure 3(b), with the decrease in validation surrogate loss, validation abstention loss also tends to decrease. Besides, as shown in Figure 3(b), validation abstention loss can be regarded as an integrated 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-logisitc loss, as shown in Figure 3(d), with the decrease in validation surrogate loss, validation abstention loss decreases at early epochs and tends to keep a comparatively low level in the late epochs, which validates the calibration of our ACS-logistic loss.

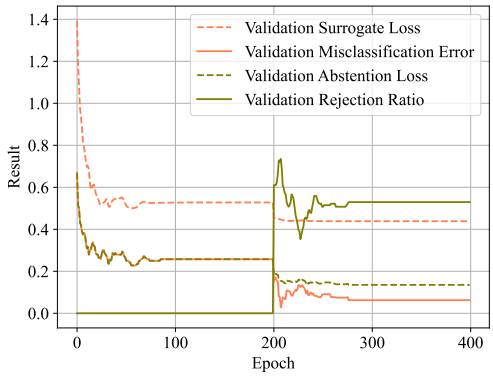
Impact of abstention cost c The impacts of abstention cost c for our MCS-hinge and ACS-hinge losses on CIFAR-10 are shown in Figure 4. As shown in Figure 4(a), for our MCS-hinge loss, with an increase in abstention cost c, the abstention loss and misclassification error tend to increase, which validates the sensitivity to the abstention cost c. Our ACS-hinge loss also shows a similar behaviour, as shown in Figure 4(b).



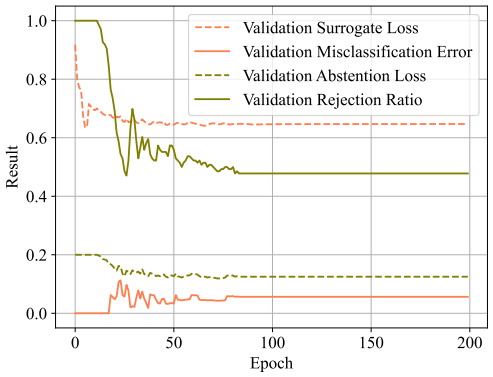
(a) Performance of CE (no rejection).



(b) Performance of Ramaswamy et al. 2018 (CS) ($c = 0.2$).

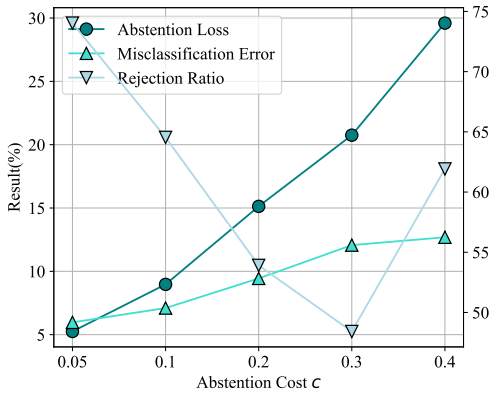


(c) Performance of Mao et al. 2024a ($c = 0.2$, $\alpha = 1.0$).

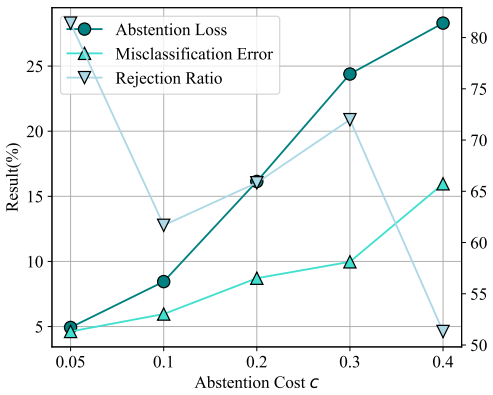


(d) Performance of ours (ACS-logistic) ($c = 0.2$, $\alpha = 3.2$).

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



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



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

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