On Classification-Calibration of Gamma-Phi Losses

Yutong Wang YUTONGW@UMICH.EDU
Clayton Scott CLAYSCOT@UMICH.EDU
University of Michigan

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

Gamma-Phi losses constitute a family of multiclass classification loss functions that generalize the logistic and other common losses, and have found application in the boosting literature. We establish the first general sufficient condition for the classification-calibration (CC) of such losses. To our knowledge, this sufficient condition gives the first family of nonconvex multiclass surrogate losses for which CC has been fully justified. In addition, we show that a previously proposed sufficient condition is in fact not sufficient. This contribution highlights a technical issue that is important in the study of multiclass CC but has been neglected in prior work.

Keywords: Loss functions, classification-calibration

1. Introduction

Multiclass classification into $k \geq 2$ categories is one of the most commonly encountered tasks in machine learning. To formulate the task mathematically, labelled training instances $\{(x_i, y_i)\}_{i=1}^n$ are drawn from a joint distribution P over $\mathcal{X} \times [k]$ where $[k] := \{1, \ldots, k\}$ and $\mathcal{X}$ is a space of unlabelled instances. The goal is to select a *classifier* $h : \mathcal{X} \to [k]$ that makes the fewest mistakes, i.e., the *01-risk* $R_{01,P}(h) := \mathbb{E}_{(X,Y)\sim P}\left[\mathbb{I}\{Y \neq h(X)\}\right]$ should be as low as possible. Here, $\mathbb{I}$ denotes the indicator function. A fundamental strategy for learning a classifier is *empirical risk minimization* (ERM), which selects an h minimizing the empirical 01-risk $\hat{R}_{01}^n(h) := \frac{1}{n}\sum_{i=1}^n \mathbb{I}\{y_i \neq h(x_i)\}$ over a class of functions, e.g., histogram classifiers.

Directly minimizing the empirical 01-risk is difficult due to the discrete nature of the objective. To address this, it is common to employ continuous-valued *class-score functions* $f = (f_1, \ldots, f_k) : \mathcal{X} \to \mathbb{R}^k$, whose components represent preference for the respective class, and are used in lieu of the discrete-valued h. The classifier associated to f is defined as $\arg\max \circ f(x) := \arg\max_{j=1,\ldots,k} f_j(x)$, where ties are broken arbitrarily. For selecting f, a *surrogate loss function* $\mathcal{L} : [k] \times \mathbb{R}^k \to \mathbb{R}_{\geq 0}$ is employed. The quantity $\mathcal{L}(y, f(x))$ is the loss incurred by f when instance x has label y. The *$\mathcal{L}$-risk* is defined to be $R_{\mathcal{L},P}(f) := \mathbb{E}_{(X,Y)\sim P}[\mathcal{L}(Y, f(X))]$, and f is selected by minimizing the *empirical* $\mathcal{L}$-risk $\hat{R}_{\mathcal{L}}^n(f)$ over a class of functions, e.g., neural networks.

It is desirable for the surrogate loss $\mathcal{L}$ to have the *consistency transfer property*: Let P be a distribution over $\mathcal{X} \times [k]$. Suppose that $\{\hat{f}^{(n)}\}_n$ is an arbitrary sequence of score functions, e.g., $\hat{f}^{(n)}$ is an empirical $\mathcal{L}$-risk minimizer, over a class of functions that may depend on n. Whenever $R_{\mathcal{L}}(\hat{f}^{(n)}) \to \inf_f R_{\mathcal{L}}(f)$ as $n \to \infty$, we have $R_{01}(\arg\max \circ \hat{f}^{(n)}) \to \inf_h R_{01}(h)$ as $n \to \infty$. The infimums are, respectively, over all measurable functions $f : \mathcal{X} \to \mathbb{R}^k$ and $h : \mathcal{X} \to [k]$.

This property justifies $\mathcal{L}$-risk minimization when minimizing the 01-risk is the actual target of interest. Establishing sufficient conditions for the property is a central goal of the theory of *classification-calibration* (CC) (Zhang, 2004; Tewari and Bartlett, 2007; Duchi et al., 2018).

The Gamma-Phi losses are a family of losses that have been successfully applied to the design and analysis of multiclass boosting algorithms (Beijbom et al., 2014; Saberian and Vasconcelos, 2019). Analysis of boosting algorithms have also been approached from the CC theory point of view (Bartlett and Traskin, 2006; Mukherjee and Schapire, 2013). However, no sufficient condition for the classification calibration of Gamma-Phi losses has been proposed that broadly encompasses a wide range of practical losses. This work aims to close this gap.

1.1. Our contributions

Informally stated, our contributions are:

1. Theorem 3.3: For γ with strictly positive derivative, and non-increasing ϕ with negative derivative at zero, the associated Gamma-Phi loss $\mathcal{L}$ is classification-calibrated.
2. Theorem 3.6: There exists strictly increasing γ, and non-increasing ϕ with negative derivative at zero whose associated Gamma-Phi loss is *not* classification-calibrated.

The positive result of this work, Theorem 3.3, establishes the first broadly applicable sufficient condition for classification calibration of Gamma-Phi loss. The negative result of this work, Theorem 3.6, shows that the condition that γ has strictly positive derivative cannot be significantly weakened, and that the sufficient condition conjectured by Pires and Szepesvári (2016) turns out to be *insufficient*.

The main impact of our theory is that for the first time it establishes CC of *nonconvex* multiclass loss functions, which have been shown to enjoy robustness to label noise (Masnadi-Shirazi and Vasconcelos, 2008; Amid et al., 2019) and adversarial contamination (Awasthi et al., 2021). For instance, our work establishes CC of the multiclass sigmoid loss ($\gamma =$ identity, $\phi =$ sigmoid). A second impact is that our work calls attention to the importance of considering the boundary of the probability simplex, not just the interior, when proving CC. Several prior works omit this case (Zhang, 2004; Zhang et al., 2009; Amid et al., 2019), and thus their proofs are incomplete.

Finally, our result provides insight on choosing Gamma-Phi losses that are suitable for multiclass classification. For instance, Gamma-Phi losses, which have been successful in (offline) multiclass boosting Saberian and Vasconcelos (2019), may be useful for the online regime as well (Raman and Tewari, 2022). See Jung et al. (2017) and the discussion in §4.1 therein.

1.2. Related works

Gamma-Phi losses were introduced and studied in a series of papers and have been shown to perform well in boosting (Saberian and Vasconcelos, 2019, 2011; Beijbom et al., 2014). Progress towards proving classification-calibration have been made for special instances of Gamma-Phi, namely for the *coherence loss* (Zhang et al., 2009) and the *pairwise-comparison* loss (Zhang, 2004)[1].

Non-convex multiclass loss functions. Many existing sufficient conditions for CC require the components of $\mathcal{L}$ to be convex, e.g., in Tewari and Bartlett (2007); Bartlett et al. (2006). Gamma-Phi losses in general are non-convex and thus our result Theorem 3.3, which does not require convexity, is complementary to these works. Non-convex loss have recently received attention in the context of robust learning (Huber, 2011), and learning with noisy labels (Amid et al., 2019).

Beyond classification. While our work focuses on classification, many works have developed theory for other learning tasks. Steinwart (2007) introduced the extension of loss calibration-theory to

1. See Remark 3.4. Our sufficient condition (Theorem 3.3) completes and subsumes these partial works.

cost-sensitive classification, regression and unsupervised learning tasks such as density estimation. Ramaswamy and Agarwal (2016) developed theory for learning with general discrete losses such as abstention (Ramaswamy et al., 2018). Finocchiaro et al. (2019) showed that there exists polyhedral losses that are calibrated with respect to arbitrary discrete losses.

Restricted class of score functions. The key result of CC theory relating $\mathcal{L}$-risk and 01-risk minimization assumes working with a sufficiently rich class of score function, i.e., when $f : \mathcal{X} \to \mathbb{R}^k$ range over all measure functions. Without this assumption, classification-calibration theory is of limited use (Mukherjee and Schapire, 2013, §9.1). To address this, Duchi et al. (2018) introduces the notion of *universal-equivalence of loss functions* and extends the key result of CC theory for certain restricted families of f's that are "quantized". Moreover, Long and Servedio (2013); Zhang and Agarwal (2020); Awasthi et al. (2022) study $\mathcal{H}$*-consistency* for the situation when f is restricted to some given family $\mathcal{H}$. Awasthi et al. (2022) derives generalized regret bounds based on the so-called *minimizability gap*.

1.3. Notations

Denote by $k \geq 2$ the number of classes and by $\Delta^k = \{\mathbf{p} \in \mathbb{R}^k_{\geq 0} : \sum_{j=1}^k p_j = 1\}$ the k-probability simplex. Let $\Delta^k_{\texttt{desc}} = \{\mathbf{p} \in \Delta^k : p_1 \geq \cdots \geq p_k\}$ denote the set of probability vectors whose entries are non-increasing (descending) with respect to the index.

Operations on vectors. Let the square bracket with subscript $[\cdot]_j$ be the projection of a vector onto its j-th component, i.e., $[\mathbf{v}]_j := v_j$ where $\mathbf{v} = (v_1, \ldots, v_k) \in \mathbb{R}^k$. Given two vectors $\mathbf{w}, \mathbf{v} \in \mathbb{R}^k$, we write $\mathbf{w} \geq \mathbf{v}$ if $w_j \geq v_j$ for all $j \in [k]$. Likewise, we write $\mathbf{w} > \mathbf{v}$ if $w_j > v_j$ for all $j \in [k]$.

Permutations. A bijection from $[k]$ to itself is called a *permutation on* $[k]$. Denote by $\texttt{Sym}(k)$ the set of all permutations on $[k]$. We often write $\sigma\sigma'$ instead of $\sigma \circ \sigma'$ for the compositions of two permutations $\sigma, \sigma' \in \texttt{Sym}(k)$. For $i, j \in [k]$, let $\tau_{(i,j)} \in \texttt{Sym}(k)$ denote the *transposition* which swaps i and j, leaving all other elements unchanged. More precisely, $\tau_{(i,j)}(i) = j$, $\tau_{(i,j)}(j) = i$ and $\tau_{(i,j)}(y) = y$ for $y \in [k] \setminus \{i, j\}$.

Permutation matrices. For each $\sigma \in \texttt{Sym}(k)$, let $\mathbf{S}_\sigma$ denote the permutation matrix corresponding to σ. In other words, if $\mathbf{v} \in \mathbb{R}^k$ is a vector, then $[\mathbf{S}_\sigma \mathbf{v}]_j = [\mathbf{v}]_{\sigma(j)} = v_{\sigma(j)}$. Below, we abuse notation and simply write $\sigma(\mathbf{v})$ instead of $\mathbf{S}_\sigma(\mathbf{v})$ when there is no confusion. Note that if $\sigma, \sigma' \in \texttt{Sym}(k)$, then $\mathbf{S}_{\sigma\sigma'} = \mathbf{S}_\sigma \mathbf{S}_{\sigma'}$.

2. Background

In this section, we review the definitions of Gamma-Phi losses and classification-calibration. In the introduction, a multiclass loss is denoted as $\mathcal{L} : [k] \times \mathbb{R}^k \to \mathbb{R}_{\geq 0}$. However, hereinafter, we will use a slightly modified, but mathematically equivalent notation that is more convenient.

Definition 2.1. A *k-ary multiclass loss function* $\mathcal{L} = (\mathcal{L}_1, \ldots, \mathcal{L}_k) : \mathbb{R}^k \to \mathbb{R}^k_+$ is a vector-valued function such that for all $\mathbf{v} \in \mathbb{R}^k$ and all $y, j \in [k]$, we have $v_y \geq v_j$ implies that $\mathcal{L}_y(\mathbf{v}) \leq \mathcal{L}_j(\mathbf{v})$. Given the score vector $\mathbf{v} \in \mathbb{R}^k$, the value $\mathcal{L}_y(\mathbf{v})$ is the loss incurred when the true label is $y \in [k]$. We say that $\mathcal{L}$ is *permutation equivariant*[2] if $\mathcal{L}(\mathbf{S}_\sigma(\mathbf{v})) = \mathbf{S}_\sigma(\mathcal{L}(\mathbf{v}))$ for all $\mathbf{v} \in \mathbb{R}^k$ and $\sigma \in \texttt{Sym}(k)$. In other words, the classes are viewed symmetrically from the loss function's perspective.

2. This property is sometimes referred to as *symmetric* in the literature. However, a symmetric function, say f, has the *in*variance property, i.e., $f(\mathbf{S}_\sigma(\mathbf{v})) = f(\mathbf{v})$, which is different than the notion of *equi*variance as used in Definition 2.1. See Bronstein et al. (2021, §3.1), for an in-depth discussion on invariance versus equivariance.

Definition 2.2 (Gamma-Phi losses)**.** Let $\gamma : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be non-decreasing and $\phi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ non-increasing functions. The *Gamma-Phi* loss associated to γ and ϕ is the loss $\mathcal{L} \equiv \mathcal{L}^{\gamma,\phi}$ whose y-th component is given by $\mathcal{L}_y(\mathbf{v}) := \gamma\Big(\sum_{y' \in [k] : y' \neq y} \phi(v_y - v_{y'})\Big)$.

Example 2.3. When $\gamma(x) := \log(1+x)$ and $\phi(x) := \exp(-x)$, we recover the logistic/cross entropy loss. When $\gamma(x) := T\log(1+x)$ and $\phi(x) := \exp((1-x)/T)$ where $T > 0$ is a hyperparameter, we recover the *coherence loss* used in the multiclass `GentleBoost.C` algorithm (Zhang et al., 2009). When γ is the identity, the Gamma-Phi loss reduces to the *pairwise comparison loss* (Zhang, 2004, Section 4.1). The multiclass exponential loss, used in `AdaBoost.MM` (Mukherjee and Schapire, 2013), is the pairwise comparison loss when $\phi(x) := \exp(-x)$. When $\gamma(x) := (x/(1+x))^2$ and $\phi(x) := \exp(-x)$, we recover the savage loss (Saberian and Vasconcelos, 2019).

We now review fundamental definitions and key result of the theory of classification-calibration.

Definition 2.4. Let $\mathbf{p} \in \Delta^k$. The *conditional risk* of $\mathcal{L}$ at $\mathbf{p}$ is the function $C^{\mathcal{L}}_{\mathbf{p}} : \mathbb{R}^k \to \mathbb{R}$ defined by $C^{\mathcal{L}}_{\mathbf{p}}(\mathbf{v}) = \sum_{y \in [k]} p_y \mathcal{L}_y(\mathbf{v})$. The *conditional Bayes risk* is defined as $C^{\mathcal{L},*}_{\mathbf{p}} := \inf_{\mathbf{v} \in \mathbb{R}^k} C^{\mathcal{L}}_{\mathbf{p}}(\mathbf{v})$. When there is no ambiguity about the loss function, we drop the superscript $\mathcal{L}$ and simply write $C_{\mathbf{p}}(\mathbf{v})$ and $C^*_{\mathbf{p}}$.

This "conditional" terminology was used in Bartlett et al. (2006). It was also called *inner $\mathcal{L}$-risk* by Steinwart (2007). The following is from Zhang (2004, Definition 1).

Definition 2.5. A loss $\mathcal{L}$ is *classification-calibrated* if for all $\mathbf{p} \in \Delta^k$ and y such that $p_y < \max_j p_j$, we have $C^{\mathcal{L},*}_{\mathbf{p}} < \inf\left\{C^{\mathcal{L}}_{\mathbf{p}}(\mathbf{v}) : \mathbf{v} \in \mathbb{R}^k,\ v_y = \max \mathbf{v}\right\}$.

Intuitively, the classification-calibration property states that when y is not the most probable class label, then outputing a score vector $\mathbf{v}$ maximized at y leads to sub-optimal conditional $\mathcal{L}$-risk. Next, we recall the definitions of the *01-risk* $R_{01,P}(h) := \mathbb{E}_{(X,Y)\sim P}\left[\mathbb{I}\{Y \neq h(X)\}\right]$ and the *$\mathcal{L}$-risk* $R_{\mathcal{L},P}(f) := \mathbb{E}_{(X,Y)\sim P}[\mathcal{L}(Y, f(X))]$. Finally, the key result in classification-calibration theory is

Theorem 2.6 (Zhang (2004))**.** *Let $\mathcal{L} : \mathbb{R}^k \to \mathbb{R}^k_+$ be a permutation equivariant loss function. Let $\mathcal{F}$ be the set of all Borel functions $\mathcal{X} \to \mathbb{R}^k$. If $\mathcal{L}$ is classification-calibrated, then $\mathcal{L}$ has the* consistency transfer property*: For all sequence of function classes $\{\mathcal{F}_n\}_n$ such that $\mathcal{F}_n \subseteq \mathcal{F}$, $\bigcup_n \mathcal{F}_n = \mathcal{F}$, all $\hat{f}_n \in \mathcal{F}_n$ and all probability distributions P on $\mathcal{X} \times [k]$*

$$R_{\mathcal{L},P}(\hat{f}_n) \xrightarrow{P} \inf_f R_{\mathcal{L},P}(f) \quad \textit{implies} \quad R_{01,P}(\arg\max \circ \hat{f}_n) \xrightarrow{P} \inf_h R_{01,P}(h)$$

where the infimums are taken over all Borel functions $f : \mathcal{X} \to \mathbb{R}^k$ and $h : \mathcal{X} \to [k]$, respectively.

In applications, $\hat{f}_n$ is often taken to be an $\mathcal{L}$-risk empirical minimizer over a training dataset of cardinality n. However, the above property holds for *any* sequence of functions $\hat{f}_n \in \mathcal{F}_n$.

Remark 2.7. Zhang (2004) refers to Definition 2.5 as *infinity-sample consistency* (ISC), while later work by Tewari and Bartlett (2007) considers a slightly different definition of *classification-calibration* (CC). Moreover, Tewari and Bartlett (2007, Theorem 2) shows that CC characterizes the consistency transfer property, while Zhang (2004, Theorem 3) only shows that ISC is sufficient. In fact, ISC is also necessary. While this fact is simple, it has not been explictly stated to the best of our knowledge. Therefore, we include its proof in Section A of the Appendix. Throughout this work, we will use the name "classification-calibration".

3. Main results

In this section, we consider the Gamma-Phi loss as in Definition 2.2.

Definition 3.1. Let $\gamma : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be a function satisfying $\sup_{x\in[0,\infty)} \gamma(x) = +\infty$. We say that γ satisfies condition[3] (`Gamma-SI`) if γ is strictly increasing, i.e., $\gamma(x) < \gamma(\tilde{x})$ if $x < \tilde{x}$, and condition (`Gamma-PD`) if γ is continuously differentiable and $\frac{d\gamma}{dx}(x) > 0$ for all $x \geq 0$.

Note that condition (`Gamma-PD`) implies condition (`Gamma-SI`), but the converse is not true.

Definition 3.2. Let $\phi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ be a function with the property that $\inf_{x\in\mathbb{R}} \phi(x) = 0$. We say that ϕ satisfies condition (`Phi-NDZ`) if ϕ is differentiable, $\frac{d\phi}{dx}(x) \leq 0$ for all $x \in \mathbb{R}$, and $\frac{d\phi}{dx}(0) < 0$.

Theorem 3.3. *Let $\mathcal{L} \equiv \mathcal{L}^{\gamma,\phi}$ be the Gamma-Phi loss where γ satisfies `Gamma-PD`, and ϕ satisfies `Phi-NDZ`. Then $\mathcal{L}$ is classification-calibrated.*

In light of Theorem 2.6, if $\mathcal{L}$ satisfies the conditions of Theorem 3.3, then $\mathcal{L}$ satisfies the transfer consistency property. As stated in the introduction, Theorem 3.3 establishes the first sufficient condition of CC for Gamma-Phi loss.

Remark 3.4. Both the coherence loss (Zhang et al., 2009) and pairwise comparison loss (Zhang, 2004) where ϕ satisfies (`Phi-NDZ`) satisfy the conditions of Theorem 3.3. On the other hand, the classification-calibration property for these losses has not been established previously due to omissions in the respective proofs. In both aforementioned works, the proofs only check the condition in Definition 2.5 for $\mathbf{p} \in \Delta^k$ such that $\mathbf{p} > 0$ elementwise. In this work, we address the case when $\mathbf{p}$ can have zero entries, which requires significant work.

Remark 3.5. The multiclass savage loss (Saberian and Vasconcelos, 2019) is a Gamma-Phi loss with $\gamma(x) = (x/(1+x))^2$ and $\phi(x) = \exp(-2x)$ which does not satisfy the condition of Theorem 3.3. More precisely, the condition $\sup_{x\in[0,\infty)} \gamma(x) = +\infty$ fails. While the binary savage loss is classification-calibrated (Masnadi-Shirazi and Vasconcelos, 2008), to the best of our knowledge it is unknown whether the multiclass savage loss is classification-calibrated.

We present an intuitive proof sketch of Theorem 3.3. The full proof can be found in Section 5.

Proof sketch of Theorem 3.3. Unwinding the definition, if the loss $\mathcal{L}$ is not classification-calibrated, then there exists a class-conditional distribution on the labels $\mathbf{p} \in \Delta^k$, a label $y \in [k]$, and a sequence $\{\mathbf{v}^t \in \mathbb{R}^k\}_{t=1,2,\ldots}$ such that *(i)* $p_y < \max_{j\in[k]} p_j$ is *not* maximal, *(ii)* $v^t_y = \max_{j\in[k]} v^t_j$ *is* maximal for all t and *(iii)* the conditional risk $C^{\mathcal{L}}_{\mathbf{p}}(\mathbf{v}^t) \to C^{\mathcal{L},*}_{\mathbf{p}}$ the Bayes conditional risk.

The "easy" case is when the sequence $\{\mathbf{v}^t\}_{t=1,2,\ldots}$ has a limit $\boldsymbol{\alpha} \in \mathbb{R}^k$ where $\alpha_y = \max_{j\in[k]} \alpha_j$ and $C^{\mathcal{L}}_{\mathbf{p}}(\boldsymbol{\alpha}) = C^{\mathcal{L},*}_{\mathbf{p}}$. Then it is straightforward[4] to derive a contradiction using the first derivative test for minimality. However, the challenge is that in general the sequence $\{\mathbf{v}^t\}_{t=1,2,\ldots}$ may diverge[5].

The major technical contribution of our work is the construction of a modified sequence $\{\tilde{\mathbf{v}}^t\}_t$, based on $\{\mathbf{v}^t\}_t$, with the following property: there exists some $\ell \in \{2,\ldots,k\}$ such that for all

3. "SI", "PD" and "NDZ" stand for strictly increasing, positive derivative and negative derivative at zero, respectively.
4. The "easy" case was previously addressed in Zhang (2004, Theorem 5). For the sake of completeness, we include the argument in Lemma 5.6.
5. Note that neither Zhang (2004, Theorem 5) nor Lemma 5.6 can be directly applied in the divergent case because doing so would amount to "taking a derivative at infinity". Our technique circumvents this by extracting a convergent subvector. It will be interesting to interpret this technique in the *astral space* formalism (Dudík et al., 2022).

$t = 1, 2, \ldots$ the "subvector" $(\tilde{v}_1^t, \ldots, \tilde{v}_\ell^t) = \boldsymbol{\alpha} \in \mathbb{R}^\ell$, i.e., the subvector consisting of the first ℓ entries of $\tilde{\mathbf{v}}^t$ is constant with respect to t and equals to $\boldsymbol{\alpha}$. Moreover, $C_{\mathbf{q}}^{\mathcal{L}}(\boldsymbol{\alpha}) = C_{\mathbf{q}}^{\mathcal{L},*}$, where $\mathbf{q} := (\sum_{j=1}^{\ell} p_j)^{-1} (p_1, \ldots, p_\ell) \in \Delta_{\texttt{desc}}^{\ell}$. Thus, we have reduced the problem to the "easy" case.

At a high level, our construction proceeds by first showing that the sequence $\{\mathbf{v}^t\}_t$ converges in the extended-real sense. Then, we proceed with a series of modifications to $\{\mathbf{v}^t\}_t$ to ensure that there exists an index $\ell \in [k]$ such that for all $1 \leq j \leq \ell$ we have v_j^t is equal to a finite quantity $\alpha_j \in \mathbb{R}$ for all t, and that for all $j > \ell$ we have $\lim_t v_j^t = \pm\infty$ diverges. We show that throughout these modification, the property $C_{\mathbf{p}}^{\mathcal{L}}(\mathbf{v}^t) \to C_{\mathbf{p}}^{\mathcal{L},*}$ continues to hold. □

Our analysis highlights the following intuition: if $\mathcal{L}$ is classification-calibrated for the k-category problem, then $\mathcal{L}$ must be classification-calibrated for every ℓ-category subproblem where $2 \leq \ell \leq k$.

Next, we show an example of a Gamma Phi loss that satisfies the conditions of Pires and Szepesvári (2016) and yet is not classification-calibrated. The paragraph before Pires and Szepesvári (2016, Section 3.4.2) conjectures that the Gamma-Phi loss is calibrated when γ is strictly increasing and ϕ satisfies the same condition as in Zhang (2004, Theorem 6), namely that ϕ is non-negative, non-increasing and $\phi'(0) < 0$. However, in the following example, we give a counterexample to the aforementioned statement.

Theorem 3.6. *Let $\mathcal{L}$ be the Gamma-Phi loss where $\phi(x) = \exp(-x)$ and*

$$\gamma(x) = \begin{cases} 1 - (x-1)^2 & : x < 1 \\ 2(x-1)^2 + 1 & : x \geq 1. \end{cases}$$

Then γ satisfies (`Gamma-SI`) and ϕ satisfies (`Phi-NDZ`), but $\mathcal{L}$ is not classification-calibrated.

Our proof relies on a careful analysis of the behavior of the loss function $\mathcal{L}(\mathbf{v})$ when the argument $\mathbf{v}$ approaches infinity. Moreover, our proof highlights the importance of verifying the main condition in the definition of classification-calibration (Definition 2.5) for $\mathbf{p}$ with zero entries. The following sections will prove results stated in this section. All omitted proofs of intermediate results are included in the appendix.

4. Conditional risks of permutation equivariant losses

In this section, we study some of the basic properties of the conditional risk (Definition 2.4) of permutation equivariant losses (Definition 2.1). All omitted proofs can be found in Section C of the Appendix.

Lemma 4.1. *Let $\mathcal{L}$ be a permutation equivariant loss. Let $\sigma \in \texttt{Sym}(k)$, $\mathbf{v} \in \mathbb{R}^k$ and $\mathbf{p} \in \Delta^k$ be arbitrary. Then $C_{\mathbf{p}}(\mathbf{v}) = C_{\sigma(\mathbf{p})}(\sigma(\mathbf{v}))$. Furthermore, we have $C_{\mathbf{p}}^* = C_{\sigma(\mathbf{p})}^*$.*

Lemma 4.2. *Suppose that $\mathcal{L}$ is permutation equivariant. Let $\mathbf{p} \in \Delta^k$, $y, y' \in [k]$ and $\mathbf{v} \in \mathbb{R}^k$. Let $\tau \in \texttt{Sym}(k)$ be the* transposition *of y and y', i.e., $\tau(y) = y'$, $\tau(y') = y$ and $\tau(j) = j$ for all $j \in [k] \setminus \{y, y'\}$. Then $C_{\mathbf{p}}(\mathbf{v}) - C_{\mathbf{p}}(\tau(\mathbf{v})) = (p_y - p_{y'})(\mathcal{L}_y(\mathbf{v}) - \mathcal{L}_{y'}(\mathbf{v}))$.*

Proposition 4.3. *Let $\mathbf{p} \in \Delta_{\texttt{desc}}^k$. Let $\mathbf{v} \in \mathbb{R}^k$ be arbitrary. Let $\sigma \in \texttt{Sym}(k)$ be such that $v_{\sigma(1)} \geq v_{\sigma(2)} \geq \cdots \geq v_{\sigma(k)}$. Then $C_{\mathbf{p}}(\mathbf{v}) \geq C_{\mathbf{p}}(\sigma(\mathbf{v}))$.*

Proof. This proof is essentially Lemma S3.8 from Wang and Scott (2020) Supplemental Materials. First, we note that if $\tilde{\sigma} \in \mathtt{Sym}(k)$ is another permutation such that $v_{\tilde{\sigma}(1)} \geq v_{\tilde{\sigma}(2)} \geq \cdots \geq v_{\tilde{\sigma}(k)}$, then $\tilde{\sigma}(\mathbf{v}) = \sigma(\mathbf{v})$. Thus, it suffices to prove the result while assuming that the permutation σ that sorts $\mathbf{v}$ is given by the *bubble sort* algorithm:

L1. Initialize the iteration index $t \leftarrow 0$ and $\mathbf{v}^0 := \mathbf{v}$,
L2. While there exists $i \in [k]$ such that $v_i^t < v_{i+1}^t$, do
   (a) Let $\tau^t = \tau_{(i,i+1)} \in \mathtt{Sym}(k)$ be the permutation that swaps i and $i+1$, leaving other indices unchanged.
   (b) $\mathbf{v}^{t+1} \leftarrow \tau^t(\mathbf{v}^t)$
   (c) $t \leftarrow t+1$
L3. Output $\mathbf{v}^T$, where $T \leftarrow t$ is the final iteration index.

Let $\langle \cdot, \cdot \rangle$ be the ordinary dot product on $\mathbb{R}^k$. Note that $C_{\mathbf{p}}(\mathbf{v}) = \langle \mathbf{p}, \mathcal{L}(\mathbf{v}) \rangle$. Furthermore, at termination, there exists $\sigma \in \mathtt{Sym}(k)$ such that $\mathbf{v}^T = \sigma(\mathbf{v})$ is sorted as in the statement of Proposition 4.3. We claim that at every intermediate step $t \in \{0, \ldots, T\}$, we have $\langle \mathbf{p}, \mathcal{L}(\mathbf{v}^t) \rangle \geq \langle \mathbf{p}, \mathcal{L}(\mathbf{v}^{t+1}) \rangle$. This would prove Proposition 4.3, since $\langle \mathbf{p}, \mathcal{L}(\mathbf{v}^0) \rangle = C_{\mathbf{p}}(\mathbf{v})$ and $\langle \mathbf{p}, \mathcal{L}(\mathbf{v}^T) \rangle = C_{\mathbf{p}}(\sigma(\mathbf{v}))$.

Towards proving our claim, let t be an intermediate iteration of the above "bubble sort" algorithm, and let $i \in [k]$ be as in L2. Then we have

$$\begin{aligned}
&\langle \mathbf{p}, \mathcal{L}(\mathbf{v}^t) \rangle - \langle \mathbf{p}, \mathcal{L}(\mathbf{v}^{t+1}) \rangle \\
&= \langle \mathbf{p}, \mathcal{L}(\mathbf{v}^t) \rangle - \langle \mathbf{p}, \mathcal{L}(\tau^t(\mathbf{v}^t)) \rangle \qquad \because \text{Definition on L2.(b)} \\
&= (p_i - p_{i+1})(\mathcal{L}_i(\mathbf{v}^t) - \mathcal{L}_{i+1}(\mathbf{v}^t)) \geq 0, \qquad \text{Lemma 4.2}
\end{aligned}$$

as desired. The last inequality follows immediately from Definition 2.1 for multiclass losses. □

5. Proof of Theorem 3.3

In this section, we develop the machinery which will be put together at the end of the section to prove Theorem 3.3. The first goal is to prove Proposition 5.4, whose proof sketch was introduced above in Section 3. Roughly speaking, Proposition 5.4 derives properties about sequences $\{\mathbf{v}^t\}_t$ such that $\lim_t C_{\mathbf{p}}(\mathbf{v}^t) = C_{\mathbf{p}}^*$. Assuming that $\mathcal{L}$ is not classification-calibrated, these properties together with Lemmas 5.6 to 5.7 are then used to derive a contradiction, thereby proving Theorem 3.3.

Now, to prepare for the proof of Proposition 5.4, we state several helper results:

Lemma 5.1. *In the situation of Theorem 3.3, let $\{\mathbf{v}^t\}_t \subseteq \mathbb{R}^k$ be a totally convergent sequence and $\mathbf{p} \in \Delta^k$. Then $\lim_t C_{\mathbf{p}}(\mathbf{v}^t)$ exists and is $\in [0, +\infty]$. If $\{\tilde{\mathbf{v}}^t\}_t \subseteq \mathbb{R}^k$ is another totally convergent sequence such that $\lim_t v_y^t - v_j^t = \lim_t \tilde{v}_y^t - \tilde{v}_j^t$ for all $y, j \in [k]$, then $\lim_t C_{\mathbf{p}}(\mathbf{v}^t) = \lim_t C_{\mathbf{p}}(\tilde{\mathbf{v}}^t)$.*

The proof of Lemma 5.1 can be found in Section D of the Appendix.

Corollary 5.2. *In the situation of Theorem 3.3, let $\{\mathbf{v}^t\}_t \subseteq \mathbb{R}^k$ be a totally convergent sequence and $S \subseteq [k]$ be a set such that $\lim_t v_y^t \in \mathbb{R}$ for all $y \in S$. Define $\{\tilde{\mathbf{v}}^t\}_t \subseteq \mathbb{R}^k$ by $\tilde{v}_j^t := v_j^t$ if $j \notin S$ and $\tilde{v}_j^t := \lim_t v_j^t$ if $j \in S$. Then $\lim_t C_p(\mathbf{v}^t) = \lim_t C_p(\tilde{\mathbf{v}}^t)$ as elements of $[0, +\infty]$.*

Proof. Note that $\{\mathbf{v}^t\}_t$ and $\{\tilde{\mathbf{v}}^t\}_t$ satisfy the conditions of Lemma 5.1. □

Lemma 5.3. *In the situation of Theorem 3.3, let $\mathcal{L}$ be the Gamma-Phi loss as in Definition 2.2 where γ satisfies (Gamma-PD) and ϕ satisfies (Phi-NDZ). Let $\mathbf{p} \in \Delta^k$ and $y, y' \in [k]$ be such that $p_{y'} > p_y$. Suppose $\{\mathbf{v}^t\}_t \subseteq \mathbb{R}^k$ is a sequence where $\liminf_t v_y^t - v_{y'}^t > 0$ and $\lim_t C_{\mathbf{p}}(\mathbf{v}^t) < +\infty$ exists. Then $\lim_t C_{\mathbf{p}}(\mathbf{v}^t) > C_{\mathbf{p}}^*$.*

Proof. Suppose that $\lim_t C_{\mathbf{p}}(\mathbf{v}^t) = C_{\mathbf{p}}^*$. We show that this leads to a contradiction. Since $p_{y'} > p_y \geq 0$ and $\lim_t C_{\mathbf{p}}(\mathbf{v}^t) < +\infty$, we have $\limsup_t p_{y'}\gamma\left(\sum_{j\in[k]\setminus\{y'\}} \phi(v_{y'}^t - v_j^t)\right) < \infty$. By our assumption on γ, we have $M := \limsup_t \sum_{j\in[k]\setminus\{y'\}} \phi(v_{y'}^t - v_j^t) < \infty$.

Next, by assumption, there exists $\epsilon > 0$ such that $v_y^t \geq v_{y'}^t + \epsilon$ for all $t \gg 0$. Below, we assume t is in this sufficiently large regime. Hence, for all $j \in [k]$ we have $v_y^t - v_j^t > v_{y'}^t - v_j^t$ and consequently $\phi(v_y^t - v_j^t) \leq \phi(v_{y'}^t - v_j^t)$. Furthermore, $v_y^t - v_{y'}^t \geq \epsilon > 0 > -\epsilon \geq v_{y'}^t - v_y^t$ and so $\phi(v_y^t - v_{y'}^t) \leq \phi(\epsilon) < \phi(-\epsilon) \leq \phi(v_{y'}^t - v_y^t)$. Let $a^t := \sum_{j\in[k]\setminus\{y'\}} \phi(v_{y'}^t - v_j^t)$ and $b^t := \sum_{j\in[k]\setminus\{y\}} \phi(v_y^t - v_j^t)$. Furthermore, define $\tilde{a}^t := \phi(-\epsilon) + \sum_{j\in[k]\setminus\{y,y'\}} \phi(v_{y'}^t - v_j^t)$ and $\tilde{b}^t := \phi(\epsilon) + \sum_{j\in[k]\setminus\{y,y'\}} \phi(v_y^t - v_j^t)$. Observe that

$$\tilde{a}^t - \tilde{b}^t = \phi(-\epsilon) - \phi(\epsilon) + \sum_{j\in[k]\setminus\{y,y'\}} \phi(v_{y'}^t - v_j^t) - \phi(v_y^t - v_j^t) \geq \phi(-\epsilon) - \phi(\epsilon).$$

In summary, we have $0 \leq b^t \leq \tilde{b}^t \leq \tilde{a}^t \leq a^t \leq M < \infty$. Let $\tau \in \mathtt{Sym}(k)$ be the permutation that swaps y and y'. By Lemma 4.2, we have

$$C_p(\mathbf{v}^t) - C_p(\tau(\mathbf{v}^t)) = (p_y - p_{y'})(\mathcal{L}_y(\mathbf{v}^t) - \mathcal{L}_{y'}(\mathbf{v}^t)) = (p_y - p_{y'})(\gamma(a^t) - \gamma(b^t)).$$

By the Fundamental Theorem of Calculus, we have

$$\begin{aligned}\gamma(a^t) - \gamma(b^t) = \int_{b^t}^{a^t} \gamma'(x)dx \geq \int_{\tilde{b}^t}^{\tilde{a}^t} \gamma'(x)dx &\geq (\tilde{a}^t - \tilde{b}^t) \inf_{x\in[\tilde{b}^t,\tilde{a}^t]} \gamma'(x) \\ &\geq (\phi(-\epsilon) - \phi(\epsilon)) \inf_{x\in[0,M]} \gamma'(x).\end{aligned}$$

Since γ satisfies (Gamma-PD), we have $\gamma' > 0$ and so $\delta := \inf_{x\in[0,M]} \gamma'(x) > 0$. Thus,

$$\lim_{t\to\infty} C_{\mathbf{p}}(\mathbf{v}^t) - C_{\mathbf{p}}(\tau(\mathbf{v}^t)) \geq (p_y - p_{y'})(\phi(-\epsilon) - \phi(\epsilon))\delta > 0$$

where the right hand side is a positive quantity independent of t. Therefore, we conclude that $\lim_{t\to\infty} C_{\mathbf{p}}(\mathbf{v}^t) > \lim_{t\to\infty} C_{\mathbf{p}}(\tau(\mathbf{v}^t))$, which is a contradiction of $\lim_{t\to\infty} C_{\mathbf{p}}(\mathbf{v}^t) = C_{\mathbf{p}}^*$. □

Before proceeding, we adopt the notation $\{v^t\}_t \equiv \alpha$ to denote that $v^t = \alpha$ for all t, where $\{v^t\}_t \subseteq \mathbb{R}$ is a sequence of real numbers and $\alpha \in \mathbb{R}$ is a constant. The first part of the following proposition applies under slightly weaker assumptions than that of Theorem 3.3, namely the (Gamma-PD) assumption is replaced by the weaker (Gamma-SI). The proposition will be used again in Section 6 in the proof of Theorem 3.6.

Proposition 5.4. *Let $\mathcal{L}$ be a Gamma-Phi loss where γ satisfies (Gamma-SI) and ϕ satisfies (Phi-NDZ). Let $\mathbf{p} \in \Delta_{\mathtt{desc}}^k$ and $z \in [k]$ be such that $C_{\mathbf{p}}^* = \inf\{C_{\mathbf{p}}(\mathbf{v}) : \mathbf{v} \in \mathbb{R}^k, v_z = \max \mathbf{v}\}$. Then there exists a sequence $\{\mathbf{v}^t\}_t \subseteq \mathbb{R}^{k-1}$ satisfying the following properties:*

1. $\lim_{t\to\infty} C_{\mathbf{p}}^{\mathcal{L}}(\mathbf{v}^t) = C_{\mathbf{p}}^*$
2. *there exists an index* $\ell \in [k]$ *and a vector* $\boldsymbol{\alpha} := (\alpha_1, \ldots, \alpha_\ell) \in \mathbb{R}^\ell$ *such that for each* $j \in \{1, \ldots, \ell\}$ *we have* $\{v_j^t\} \equiv \alpha_j$ *and* $\lim_t v_j^t = -\infty$ *for* $j > \ell$. *In addition,* $\alpha_1 = 0$.
3. *Let* $\mathbf{q} := (\sum_{j=1}^{\ell} p_j)^{-1} (p_1, \ldots, p_\ell) \in \Delta_{\texttt{desc}}^{\ell}$. *Then* $C_{\mathbf{q}}(\boldsymbol{\alpha}) = C_{\mathbf{q}}^*$.

Furthermore, suppose $z > 1$, $p_{z-1} > p_z$, *and* γ *satisfies (*`Gamma-PD`*). Then* $\{\mathbf{v}^t\}$ *can be chosen such that* $\alpha_j = 0$ *for all* $j \in [z]$.

Proof. Let $\{\mathbf{v}^t\}_t \subseteq \mathbb{R}^{k-1}$ be a sequence such that $\lim_{t\to\infty} C_{\mathbf{p}}^{\mathcal{L}}(\mathbf{v}^t) = C_{\mathbf{p}}^*$ and $v_z^t = \max \mathbf{v}^t$ for all $t \in \mathbb{N}$. A sequence $\{\mathbf{v}^t\}_t$ satisfying this preceding condition is said to have property $\texttt{P}_0$. Throughout this proof, t denotes the index of the sequence where "for all t" means "for all $t \in \mathbb{N}$". We will repeatedly modify the sequence $\mathbf{v}^t$ until all properties $\texttt{P}_1$ to $\texttt{P}_3$ below are met in addition to $\texttt{P}_0$. Under the assumptions in the "Furthermore" part of the Proposition, we will continue to modify the sequence until properties $\texttt{P}_4$ and $\texttt{P}_5$ marked by "$*$" are further satisfied.

Properties:

$\texttt{P}_1$. $\max \mathbf{v}^t = 0$ for all t

$\texttt{P}_2$. $\{\mathbf{v}^t\}_t$ is totally convergent and $\{v_j^t\}_t$ has a limit in $[-\infty, 0]$ for each $j \in [k]$

$\texttt{P}_3$. there exists an $\ell \in [k]$ such that for each $j \in [\ell]$, we have $\{v_j^t\} \equiv \alpha_j$ where $\alpha_j \in (-\infty, 0]$ and for each $j \in [k] \setminus [\ell] := \{\ell+1, \ldots, k\}$, we have $\lim_t v_j^t = -\infty$. In fact, $\ell \in [k]$ is the largest index such that $\lim_t v_\ell^t > -\infty$.

$\texttt{P}_4*$. the sequence $\{v_j^t\}_t \equiv 0$ for each $j \in [z]$

$\texttt{P}_5*$. $\ell \geq z$.

Properties 1 and 2. First, note that $C_p(\mathbf{v}) = C_p(\mathbf{v} - c\mathbf{1})$ for any $c \in \mathbb{R}$ and any $\mathbf{v} \in \mathbb{R}^k$. Replacing each $\mathbf{v}^t$ by $\mathbf{v}^t - (\max \mathbf{v}^t)\mathbf{1}$ for all t, we may assume $v_z^t = \max \mathbf{v}^t = 0$ for all t. In particular, $v_j^t \in (-\infty, 0]$ for all $j \in [k]$ and t. Passing to a subsequence if necessary, we may assume that $\{\mathbf{v}^t\}_t$ is totally convergent (Lemma B.6) whose components have limits in $[-\infty, 0]$.

Property 3. Let $\sigma^t \in \texttt{Sym}(k)$ be the permutation that sorts $\mathbf{v}^t$ in non-increasing order as in Proposition 4.3, i.e., $v_{\sigma^t(1)}^t \geq \cdots \geq v_{\sigma^t(k)}^t$. By Proposition 4.3, $C_{\mathbf{p}}(\sigma^t(\mathbf{v}^t)) \leq C_{\mathbf{p}}(\mathbf{v}^t)$ and hence $\lim_t C_{\mathbf{p}}(\sigma^t(\mathbf{v}^t)) = C_{\mathbf{p}}^*$ as well. We now replace $\mathbf{v}^t$ by $\sigma^t(\mathbf{v}^t)$. Note that $\texttt{P}_1$ is preserved by the sorting, but $\texttt{P}_2$ may no longer hold. Replacing by a subsequence if necessary, we can that $\mathbf{v}^t$ is totally convergent (Lemma B.6). Since $\texttt{P}_1$ holds, we have $\lim_t v_j^t \in [-\infty, 0]$ and hence $\texttt{P}_2$ holds.

By Property 2. By the sorting in the preceding paragraph, we have that $\lim_t v_1^t \geq \cdots \geq \lim_t v_k^t$. Now, let $\ell \in [k]$ be the largest index such that $\lim_t v_\ell^t > -\infty$. Such an index exists because $\lim_t v_1^t = \lim_t \max \mathbf{v}^t = 0$ (due to $\texttt{P}_1$). Let $\alpha_j := \lim_t v_j^t \in (-\infty, 0]$ for each $j \in \{1, \ldots, \ell\}$. Define $\tilde{\mathbf{v}}^t$ such that $\{\tilde{v}_j^t\}_t \equiv \alpha_j$ for $j \in \{1, \ldots, \ell\}$ and $\{\tilde{v}_j^t\}_t = \{v_j^t\}_t$ for $j > \ell$. Then by Corollary 5.2, we have $\tilde{\mathbf{v}}^t$ is totally convergent, and $\lim_t C_{\mathbf{p}}(\tilde{\mathbf{v}}^t) = \lim_t C_{\mathbf{p}}(\mathbf{v}^t)$. Replace v^t by $\tilde{v}^t$. Thus, we have that $\{\mathbf{v}^t\}_t$ satisfies $\texttt{P}_1$ to $\texttt{P}_3$.

Property 4. By $\texttt{P}_1$, we already have $v_z^t = \max \mathbf{v}^t = 0$. By the assumption that $p_{z-1} > p_z$ and that $\mathbf{p} \in \Delta_{\texttt{desc}}^k$, we have $p_j > p_z$ for each $j \in [z-1]$. Furthermore, by $\texttt{P}_2$, $\{\mathbf{v}^t\}_t$ is totally convergent. Hence, for a fixed $j \in [z-1]$ by definition $v_z^t - v_j^t = -v_j^t$ has a limit in $[0, \infty]$. By Lemma 5.3,

$\lim_t -v_j^t \notin (0,\infty]$ and thus $\lim_t -v_j^t = 0$. Now, define the sequence $\{\tilde{\mathbf{v}}^t\}_t$ by

$$\tilde{v}_j^t := \begin{cases} 0 & : j \in \{1,\dots,z-1\} \\ v_j^t & : j \in \{z,\dots,k\} \end{cases}$$

for all t. By Corollary 5.2, $\{\tilde{\mathbf{v}}^t\}_t$ is also totally convergent and $\lim_t C_{\mathbf{p}}(\tilde{\mathbf{v}}^t) = \lim_t C_{\mathbf{p}}(\mathbf{v}^t)$. Thus, $\lim_t C_{\mathbf{p}}(\tilde{\mathbf{v}}^t) = C_{\mathbf{p}}^*$. Replacing $\mathbf{v}^t$ by $\tilde{\mathbf{v}}^t$, we have that $\mathtt{P}_4$ holds. Clearly, $\mathtt{P}_0$, $\mathtt{P}_1$ and $\mathtt{P}_2$ all hold. Note that $\lim_t \tilde{\mathbf{v}}^t = \lim_t \mathbf{v}^t$ by construction. Thus, $\mathtt{P}_3$ still holds. Moreover, the ℓ defined in $\mathtt{P}_3$ is not changed when $\mathbf{v}^t$ is replaced by $\tilde{\mathbf{v}}^t$.

Property 5. By $\mathtt{P}_4$, $\{v_j^t\} \equiv 0$ for each $j \in [z]$. Hence, by the definition of ℓ in $\mathtt{P}_3$, we have $\ell \geq z$.

We now proceed with the rest of the proof for Proposition 5.4. Consider the sequence $\{\mathbf{v}^t\}_t$ constructed as above satisfying $\mathtt{P}_0$ through $\mathtt{P}_3$. Then items 1 and 2 of Proposition 5.4 are implied by $\mathtt{P}_0$ and $\mathtt{P}_3$ respectively. Now, if the assumptions in the "Furthermore" part hold, then the conclusion of the "Furthermore" part holds by $\mathtt{P}_4$. It only remains to check item 3 of Proposition 5.4. Below, we write $[k] \setminus [\ell] := \{\ell+1,\dots,k\}$. Now, note that

$$\begin{aligned}
&\lim_t C_{\mathbf{p}}(\mathbf{v}^t) \\
&= \sum_{y\in[k]} p_y \gamma\Big(\sum_{j\in[k]:j\neq y} \phi(\lim_t v_y^t - v_j^t)\Big) \\
&= \sum_{y\in[\ell]} p_y \gamma\Big(\sum_{j\in[k]:j\neq y} \phi(\lim_t v_y^t - v_j^t)\Big) + \underbrace{\sum_{y\in[k]\setminus[\ell]} p_y \gamma\Big(\sum_{j\in[k]:j\neq y} \phi(\lim_t v_y^t - v_j^t)\Big)}_{=:A} \qquad (1) \\
&= \underbrace{(p_1+\cdots+p_\ell)}_{=:S} \sum_{y\in[\ell]} q_y \gamma\Big(\sum_{j\in[k]:j\neq y} \phi(\lim_t v_y^t - v_j^t)\Big) + A. \qquad (2)
\end{aligned}$$

Note that A is defined as a limit. At this point, it is unclear if this limit exists and is $\in [0,\infty)$. This will become clear after Equation (3) below. But first, we focus on the $\lim_t v_y^t - v_j^t$ case by case[6]:

$$\lim_t v_y^t - v_j^t = \begin{cases} \alpha_y - \alpha_j & : y \in [\ell],\ j \in [\ell] \\ \alpha_y - \lim_t v_j^t = +\infty & : y \in [\ell],\ j \in [k]\setminus[\ell] \\ \lim_t v_y^t - \alpha_j = -\infty & : y \in [k]\setminus[\ell],\ j \in [\ell]. \end{cases}$$

Now,

$$\phi(\lim_t v_y^t - v_j^t) = \begin{cases} \phi(\alpha_y - \alpha_j) & : j \in [\ell] \\ \phi(+\infty) = 0 & : j \in \{\ell+1,\dots,k\}. \end{cases}$$

6. The case $y \in [k]\setminus[\ell]$, $j \in [k]\setminus[\ell]$ is omitted from the expression because $\lim_t v_y^t - v_j^t$ cannot be simplified further.

Putting it all together, we have

$$\begin{aligned}\lim_t C_{\mathbf{p}}(\mathbf{v}^t) &= S\sum_{y\in[\ell]} q_y\gamma\Big(\sum_{j\in[k]:j\neq y}\phi(\lim_t v_y^t - v_j^t)\Big) + A \\ &= S\sum_{y\in[\ell]} q_y\gamma\Big(\sum_{j\in[\ell]:j\neq y}\phi(\alpha_y-\alpha_j) + \sum_{j\in[k]\setminus[\ell]:j\neq y}\phi(+\infty)\Big) + A \\ &= S\sum_{y\in[\ell]} q_y\gamma\Big(\sum_{j\in[\ell]:j\neq y}\phi(\alpha_y-\alpha_j)\Big) + A \\ &= S\cdot C_{\mathbf{q}}(\boldsymbol{\alpha}) + A. \end{aligned} \tag{3}$$

Note that $C_{\mathbf{p}}^* = \lim_t C_{\mathbf{p}}(\mathbf{v}^t) \in [0,\infty)$ and $S\cdot C_{\mathbf{q}}(\boldsymbol{\alpha}) \in [0,\infty)$. Thus, the limit and defines A exists and is $\in [0,\infty)$. Now, let $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_\ell) \in \mathbb{R}^\ell$ be arbitrary and define a sequence $\{\mathbf{w}^t\} \subseteq \mathbb{R}^k$ by

$$w_j^t := \begin{cases}\beta_j & : j \in [\ell] \\ v_j^t & : j \in [k]\setminus[\ell].\end{cases}$$

Then analogous to Equation (1) above, we have the decomposition

$$\lim_t C_{\mathbf{p}}(\mathbf{w}^t) = \sum_{y\in[\ell]} p_y\gamma\Big(\sum_{j\in[k]:j\neq y}\phi(\lim_t w_y^t - w_j^t)\Big) + \underbrace{\sum_{y\in[k]\setminus[\ell]} p_y\gamma\Big(\sum_{j\in[k]:j\neq y}\phi(\lim_t w_y^t - w_j^t)\Big)}_{=:B}.$$

We claim that $A = B$ and $\lim_t C_{\mathbf{p}}(w^t) = S\cdot C_{\mathbf{q}}(\boldsymbol{\beta}) + A$. We first prove that $A = B$. To this end, observe that

$$\lim_t w_y^t - w_j^t = \begin{cases}\beta_y - \beta_j & : y\in[\ell],\, j\in[\ell] \\ \beta_y - \lim_t v_j^t = +\infty & : y\in[\ell],\, j\in[k]\setminus[\ell] \\ \lim_t v_y^t - \beta_j = -\infty & : y\in[k]\setminus[\ell],\, j\in[\ell] \\ \lim_t v_y^t - v_j^t & : y\in[k]\setminus[\ell],\, j\in[k]\setminus[\ell].\end{cases}$$

In particular, for $y\in[k]\setminus[\ell], j\in[\ell]$, we have $\lim_t w_y^t - w_j^t = -\infty = \lim_t v_y^t - v_j^t$. Thus,

$$\begin{aligned} B &= \sum_{y\in[k]\setminus[\ell]} p_y\gamma\Big(\sum_{j\in[\ell]:j\neq y}\phi(\lim_t w_y^t - w_j^t) + \sum_{j\in[k]\setminus[\ell]:j\neq y}\phi(\lim_t w_y^t - w_j^t)\Big) \\ &= \sum_{y\in[k]\setminus[\ell]} p_y\gamma\Big(\sum_{j\in[\ell]:j\neq y}\phi(-\infty) + \sum_{j\in[k]\setminus[\ell]:j\neq y}\phi(\lim_t v_y^t - v_j^t)\Big) \\ &= A. \end{aligned}$$

Next, we have

$$\begin{aligned}\lim_t C_{\mathbf{p}}(\mathbf{w}^t) &= \sum_{y\in[\ell]} p_y\gamma\Big(\sum_{j\in[k]:j\neq y}\phi(\lim_t w_y^t - w_j^t)\Big) + A \\ &= S\sum_{y\in[\ell]} q_y\gamma\Big(\sum_{j\in[\ell]:j\neq y}\phi(\beta_y-\beta_j) + \sum_{j\in[k]\setminus[\ell]:j\neq y}\phi(+\infty)\Big) + A \\ &= S\cdot C_{\mathbf{q}}(\boldsymbol{\beta}) + A.\end{aligned}$$

Since $\lim_t C_{\mathbf{p}}(\mathbf{w}^t) \geq \lim_t C_{\mathbf{p}}(\mathbf{v}^t) = C^*_{\mathbf{p}}$, we have $C_{\mathbf{q}}(\boldsymbol{\beta}) \geq C_{\mathbf{q}}(\boldsymbol{\alpha})$. Since $\boldsymbol{\beta}$ is arbitrary, this proves that $C_{\mathbf{q}}(\boldsymbol{\alpha}) = C^*_{\mathbf{q}}$. □

Corollary 5.5. *Let $\mathcal{L}$ be a Gamma-Phi loss where γ satisfies (`Gamma-SI`) and ϕ satisfies (`Phi-NDZ`). Let $\{\mathbf{v}^t\}_t$ be any sequence satisfying items 1, 2 and 3 of Proposition 5.4. If $p_y = 0$ for each $y > \ell$, then $C_{\mathbf{q}}(\boldsymbol{\alpha}) = \lim_t C_{\mathbf{p}}(\mathbf{v}^t)$.*

Proof. In Equation (3), we showed that $\lim_t C_{\mathbf{p}}(\mathbf{v}^t) = S \cdot C_{\mathbf{q}}(\boldsymbol{\alpha}) + A$ where S and A are defined on Equations (2) and (1) respectively. If $p_y = 0$ for all $y > \ell$, then clearly $S = 1$ and $A = 0$. □

Lemma 5.6. *In the situation of Theorem 3.3, suppose that $\mathbf{q} \in \Delta^\ell$ and $\boldsymbol{\alpha} \in \mathbb{R}^\ell$ are such that $\boldsymbol{\alpha}$ is a minimizer of $C_{\mathbf{q}}(\cdot)$ and $\alpha_1 = \alpha_2$. Then $q_1 = q_2$.*

Lemma 5.7. *Suppose $\mathcal{L}$ is not classification-calibrated. Then there exists a probability vector $\mathbf{p} \in \Delta^k_{\texttt{desc}}$ and an index $z \in \{2, \dots, k\}$ satisfying 1) $p_{z-1} > p_z$ and 2) $C^*_{\mathbf{p}} = \inf\{C_{\mathbf{p}}(\mathbf{v}) : \mathbf{v} \in \mathbb{R}^k,\ v_z = \max \mathbf{v}\}$.*

Proofs of both lemmas can be found in Section D of the Appendix. Now, we conclude with the

Proof of Theorem 3.3. Let $\mathbf{p} \in \Delta^k_{\texttt{desc}}$ and $z \in \{2, \dots, k\}$ be as in Lemma 5.7, which states that $\mathbf{p}$ and z satisfy the conditions of Proposition 5.4. Next, let $\ell \in [k]$, $\boldsymbol{\alpha} \in \mathbb{R}^\ell$, and $\mathbf{q} \in \Delta^\ell_{\texttt{desc}}$ be as in Proposition 5.4, which satisfy $C_{\mathbf{q}}(\boldsymbol{\alpha}) = C^*_{\mathbf{q}}$ and $q_z < q_{z-1} \leq q_1 = \max \mathbf{q}$. Let $\tau \in \texttt{Sym}(\ell)$ be the permutation which swaps z and 2 leaving all elements in $[\ell] \setminus \{2, z\}$ unchanged. Then

$$C^*_{\tau(\mathbf{q})} = C^*_{\mathbf{q}} = C_{\mathbf{q}}(\boldsymbol{\alpha}) = C_{\tau(\mathbf{q})}(\tau(\boldsymbol{\alpha})).$$

Let $\tilde{\mathbf{q}} := \tau(\mathbf{q})$ and $\tilde{\boldsymbol{\alpha}} := \tau(\boldsymbol{\alpha})$. Then $[\tilde{\boldsymbol{\alpha}}]_1 = [\boldsymbol{\alpha}]_{\tau(1)} = \alpha_1 = 0$ and $[\tilde{\boldsymbol{\alpha}}]_2 = [\boldsymbol{\alpha}]_{\tau(2)} = \alpha_z = 0$. In particular, $\tilde{\alpha}_1 = \tilde{\alpha}_2$. Thus, by Lemma 5.6, we have $\tilde{q}_1 = \tilde{q}_2$. However, $\tilde{q}_1 = [\mathbf{q}]_{\tau(1)} = q_1$ and $\tilde{q}_2 = [\mathbf{q}]_{\tau(2)} = q_z$. Since $q_z < q_1$, we have a contradiction. □

6. Proof of Theorem 3.6: A Gamma-Phi loss that is not classification-calibrated

For $r \in [\frac{1}{2}, 1]$, define $\mathbf{p} := [r, 1-r, 0, \dots, 0] \in \Delta^k_{\texttt{desc}}$. Thus, for $\mathbf{v} \in \mathbb{R}^k$, we have

$$C_{\mathbf{p}}(\mathbf{v}) = r\gamma\Big(\sum_{j \in [k] \setminus \{1\}} \phi(v_1 - v_j)\Big) + (1-r)\gamma\Big(\sum_{j \in [k] \setminus \{2\}} \phi(v_2 - v_j)\Big). \tag{4}$$

Below, fix $r \in (\frac{1}{2}, \frac{2}{3}]$ once and for all. Denote by `SEQ` the set of all sequences $\{\mathbf{v}^t\}_t$ satisfying Proposition 5.4 items 1, 2 and 3. For each sequence $\{\mathbf{v}^t\}_t \in \texttt{SEQ}$, there exists an $\ell \in [k]$ (defined as in Proposition 5.4 item 2) such that $\lim_t v^t_j = -\infty$ if and only if $j \in [k]$ satisfies $j > \ell$. Below, we choose a particular sequence $\{\mathbf{v}^t\}_t$:

$$\text{Fix a sequence } \{\mathbf{v}^t\}_t \in \texttt{SEQ} \text{ such that } \ell \text{ is as small as possible.} \tag{5}$$

Furthermore, let $\boldsymbol{q} \in \Delta^\ell_{\texttt{desc}}$, $\boldsymbol{\alpha} \in \mathbb{R}^\ell$ be from Proposition 5.4 item 3. Recall that we have $\lim_t C_{\mathbf{p}}(\mathbf{v}^t) = C^*_{\mathbf{p}}$ and that $C_{\mathbf{q}}(\boldsymbol{\alpha}) = C^*_{\mathbf{q}}$. Furthermore, Proposition 5.4 asserts that $\alpha_1 = v^t_1 = 0$.

We prove in Lemma E.1 in the Appendix that in fact $\ell = 2$ and $\boldsymbol{\alpha} = (0, 0)$. To sketch the main idea here briefly, the key step is showing that $F(x) := r\gamma\left(\phi(x)\right) + (1-r)\gamma\left(\phi(-x)\right)$ has a unique minimum at $x = 0$. The γ and ϕ in Theorem 3.6 is chosen specifically to achieve this.

Now, define another sequence $\{\mathbf{w}^t\}_{t=1}^{\infty} \subseteq \mathbb{R}^k$ where $\mathbf{w}^t = (0, \frac{1}{t}, -t, \ldots, -t)$. Then by construction we have $\lim_t C_{\mathbf{p}}(\mathbf{w}^t) = C_{\mathbf{q}}(\boldsymbol{\alpha}) = C^*_{\mathbf{q}} = C^*_{\mathbf{p}}$ and $\arg\max_{j\in[k]} w^t_j = 2$ for all t. Thus, we have constructed an example of $\mathbf{p}$ and $y \in [k]$ where $C^*_{\mathbf{p}} = \inf\{C_{\mathbf{p}}(\mathbf{v}) : \mathbf{v} \in \mathbb{R}^k, v_y = \max \mathbf{v}\}$. This shows that $\mathcal{L}$ is not classification-calibrated (Definition 2.5). Thus, we have proven Theorem 3.6. □

7. Discussion

In this work, we establish the first sufficient condition for the classification calibration of a Gamma-Phi loss in terms of the functional properties of γ and ϕ in Theorem 3.3. We also showed that our sufficient condition cannot be significantly weakened (in terms of the condition on γ) via our counter-example in Proposition 3.6.

For future work, studying the connection between non-convex Gamma-Phi losses and $\mathcal{H}$-consistency (Awasthi et al., 2022) may be fruitful. Another interesting direction is the relationship between non-convex Gamma-Phi losses and learning with label noise (Amid et al., 2019). Finally, an important future direction is to establish a concrete regret/excess risk bound. While Zhang (2004) guarantees the existence of such a risk bound, the proof is not constructive. Deriving concrete bounds is likely to require additional assumptions on γ and ϕ.

Acknowledgments

The authors were supported in part by the National Science Foundation under awards 1838179 and 2008074, and by the Department of Defense, Defense Threat Reduction Agency under award HDTRA1-20-2-0002. YW is supported by the Eric and Wendy Schmidt AI in Science Postdoctoral Fellowship, a Schmidt Futures program.

REFERENCES

Ehsan Amid, Manfred K Warmuth, Rohan Anil, and Tomer Koren. Robust bi-tempered logistic loss based on Bregman divergences. *Advances in Neural Information Processing Systems*, 32, 2019.

Pranjal Awasthi, Natalie Frank, Anqi Mao, Mehryar Mohri, and Yutao Zhong. Calibration and consistency of adversarial surrogate losses. *Advances in Neural Information Processing Systems*, 34:9804–9815, 2021.

Pranjal Awasthi, Anqi Mao, Mehryar Mohri, and Yutao Zhong. Multi-Class H-Consistency Bounds. In *Advances in Neural Information Processing Systems*, 2022.

Peter Bartlett and Mikhail Traskin. AdaBoost is consistent. *Advances in Neural Information Processing Systems*, 19, 2006.

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.

Oscar Beijbom, Mohammad Saberian, David Kriegman, and Nuno Vasconcelos. Guess-averse loss functions for cost-sensitive multiclass boosting. In *International Conference on Machine Learning*, pages 586–594. PMLR, 2014.

Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

John Duchi, Khashayar Khosravi, and Feng Ruan. Multiclass classification, information, divergence and surrogate risk. *The Annals of Statistics*, 46(6B):3246–3275, 2018.

Miroslav Dudík, Ziwei Ji, Robert E Schapire, and Matus Telgarsky. Convex analysis at infinity: An introduction to astral space. *arXiv preprint arXiv:2205.03260*, 2022.

Jessica Finocchiaro, Rafael Frongillo, and Bo Waggoner. An embedding framework for consistent polyhedral surrogates. *Advances in Neural Information Processing Systems*, 32, 2019.

Peter J Huber. Robust statistics. In *International Encyclopedia of Statistical Science*, pages 1248–1251. Springer, 2011.

Young Hun Jung, Jack Goetz, and Ambuj Tewari. Online multiclass boosting. *Advances in Neural Information Processing Systems*, 30, 2017.

Phil Long and Rocco Servedio. Consistency versus realizable h-consistency for multiclass classification. In *International Conference on Machine Learning*, pages 801–809. PMLR, 2013.

Hamed Masnadi-Shirazi and Nuno Vasconcelos. On the design of loss functions for classification: theory, robustness to outliers, and savageboost. *Advances in Neural Information Processing Systems*, 21, 2008.

Indraneel Mukherjee and Robert E Schapire. A theory of multiclass boosting. *Journal of Machine Learning Research*, 2013.

J Tinsley Oden and Leszek F Demkowicz. *Applied functional analysis*. Chapman and Hall/CRC, 2017.

Bernardo Ávila Pires and Csaba Szepesvári. Multiclass classification calibration functions. *arXiv preprint arXiv:1609.06385*, 2016.

Vinod Raman and Ambuj Tewari. Online agnostic multiclass boosting. In *Advances in Neural Information Processing Systems*, 2022.

Harish G Ramaswamy and Shivani Agarwal. Convex calibration dimension for multiclass loss matrices. *The Journal of Machine Learning Research*, 17(1):397–441, 2016.

Harish G Ramaswamy, Ambuj Tewari, and Shivani Agarwal. Consistent algorithms for multiclass classification with an abstain option. *Electronic Journal of Statistics*, 12(1):530 – 554, 2018.

Mohammad Saberian and Nuno Vasconcelos. Multiclass boosting: Theory and algorithms. *Advances in Neural Information Processing Systems*, 24, 2011.

Mohammad J Saberian and Nuno Vasconcelos. Multiclass boosting: Margins, codewords, losses, and algorithms. *Journal of Machine Learning Research*, 20(137):1–68, 2019.

Ingo Steinwart. How to compare different loss functions and their risks. *Constructive Approximation*, 26(2):225–287, 2007.

Ambuj Tewari and Peter L Bartlett. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8(5), 2007.

Yutong Wang and Clayton Scott. Weston-Watkins hinge loss and ordered partitions. *Advances in Neural Information Processing Systems*, 33:19873–19883, 2020.

Mingyuan Zhang and Shivani Agarwal. Bayes consistency vs. H-consistency: The interplay between surrogate loss functions and the scoring function class. *Advances in Neural Information Processing Systems*, 33:16927–16936, 2020.

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

Zhihua Zhang, Michael Jordan, Wu-Jun Li, and Dit-Yan Yeung. Coherence functions for multicategory margin-based classification methods. In *Artificial Intelligence and Statistics*, pages 647–654. PMLR, 2009.

Appendix A. The ISC property characterizes consistency transfer property

Proposition A.1. *Let $\mathcal{L} : \mathbb{R}^k \to \mathbb{R}^k_+$ be a multiclass loss function that does not have the ISC property, and $\mathcal{F}$ be the set of Borel functions $\mathcal{X} \to \mathbb{R}^k$. Then $\mathcal{L}$ does not have the consistency transfer property, namely: There exists a sequence of functions $\hat{f}_n \in \mathcal{F}$ and a probability distribution P on $\mathcal{X} \times [k]$ such that*

$$R_{\mathcal{L},P}(\hat{f}_n) \xrightarrow{P} \inf_f R_{\mathcal{L},P}(f) \text{ holds but } R_{01,P}(\arg\max \circ \hat{f}_n) \xrightarrow{P} \inf_h R_{01,P}(h) \text{ fails.}$$

Here, the infimums are taken over all Borel functions $f : \mathcal{X} \to \mathbb{R}^k$ and $h : \mathcal{X} \to [k]$, respectively.

The above proposition is essentially the multiclass analog of the result (Bartlett et al., 2006, Theorem 1, part 3c $\implies$ part 3a). Our proof below is a simple extension to the multiclass case of the argument in the paragraph before §2.4 in Bartlett et al. (2006).

Proof. Since $\mathcal{L}$ does not have the ISC property, there exists $\mathbf{p} \in \Delta^k$ and $y \in [k]$ such that y such that $p_y < \max_j p_j$, and $C^{\mathcal{L},*}_{\mathbf{p}} = \inf\left\{C^{\mathcal{L}}_{\mathbf{p}}(\mathbf{v}) : \mathbf{v} \in \mathbb{R}^k,\, v_y = \max \mathbf{v}\right\}$. Let $\{\mathbf{v}^n\}_n$ be a sequence such that $v^n_y = \max \mathbf{v}^n$ and $\lim_{n\to\infty} C^{\mathcal{L}}_{\mathbf{p}}(\mathbf{v}^n) = C^*_{\mathbf{p}}$. Below, fix some arbitrary $x \in \mathcal{X}$. Define P on $\mathcal{X} \times [k]$ such that $P(X = x) = 1$ and $P(X = x, Y = y') = p_{y'}$ for each $y' \in [k]$.

Next, we define a pair of maps $\boldsymbol{V} : \mathbb{R}^k \to \mathcal{F}$ and $\boldsymbol{F} : \mathcal{F} \to \mathbb{R}^k$ as follows. Given $f \in \mathcal{F}$, let $\boldsymbol{V}(f) := (f_1(x), \ldots, f_k(x)) \in \mathbb{R}^k$. Given $\mathbf{v} \in \mathbb{R}^k$ and $x' \in X$, define $\boldsymbol{F}(\mathbf{v})(x') := \mathbf{v}$ to be the constant-valued map with value $\mathbf{v}$. Since $P(X = x) = 1$, we have for all $f \in \mathcal{F}$

$$C^{\mathcal{L}}_{\mathbf{p}}(\boldsymbol{V}(f)) = R_{\mathcal{L},P}(f) \quad \text{and} \quad 1 - \max p_{\arg\max \boldsymbol{V}(f)} = R_{01,P}(\arg\max \circ f).$$

Now, for each n, let $\hat{f}^n := \boldsymbol{F}(\mathbf{v}^n) \in \mathcal{F}$. Then by construction we have $\boldsymbol{V}(\hat{f}^n) = \mathbf{v}^n$ and

$$\begin{aligned}
\lim_{n\to\infty} R_{\mathcal{L},P}(\hat{f}^n) &= \lim_{n\to\infty} C^{\mathcal{L}}_{\mathbf{p}}(\boldsymbol{V}(\hat{f}^n)) && \because C^{\mathcal{L}}_{\mathbf{p}}(\boldsymbol{V}(f)) = R_{\mathcal{L},P}(f),\ \forall f \in \mathcal{F} \\
&= \lim_{n\to\infty} C^{\mathcal{L}}_{\mathbf{p}}(\mathbf{v}^n) \\
&= C^{\mathcal{L},*}_{\mathbf{p}} && \because \text{assumption on } \mathbf{v}^n \\
&= \inf_{\mathbf{v}\in\mathbb{R}^k} C^{\mathcal{L}}_{\mathbf{p}}(\mathbf{v}) && \because \text{definition} \\
&= \inf_{f\in\mathcal{F}} C^{\mathcal{L}}_{\mathbf{p}}(\boldsymbol{V}(f)) && \because \boldsymbol{V} \text{ is surjective onto } \mathbb{R}^k \\
&= \inf_{f\in\mathcal{F}} R_{\mathcal{L},P}(f) && \because C^{\mathcal{L}}_{\mathbf{p}}(\boldsymbol{V}(f)) = R_{\mathcal{L},P}(f),\ \forall f \in \mathcal{F}
\end{aligned}$$

On the other hand, since the $\arg\max$ breaks tie arbitrarily (Tewari and Bartlett, 2007, Lemma 4), we can choose $\arg\max \boldsymbol{V}(\hat{f}^n) = \arg\max \mathbf{v}^n = y$. Therefore, we have

$$R_{01,P}(\arg\max \circ \hat{f}^n) = 1 - p_y > 1 - \max_{j\in[k]} p_j = \inf_h R_{01,P}(h).$$

Thus, we have constructed a sequence $\hat{f}_n$ such that

$$R_{\mathcal{L},P}(\hat{f}_n) \xrightarrow{P} \inf_f R_{\mathcal{L},P}(f) \text{ holds but } R_{01,P}(\arg\max \circ \hat{f}_n) \xrightarrow{P} \inf_h R_{01,P}(h) \text{ fails.}$$

as desired. □

Appendix B. Extended reals

In this section, we review results on the extended real numbers. For reference, see Oden and Demkowicz (2017, §1.16).

Definition B.1 (Convergence in extended reals)**.** Let $\overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$ and $\overline{\mathbb{R}}_{\geq 0} = \mathbb{R}_{\geq 0} \cup \{+\infty\}$. A sequence $\{z^t\}_t \subseteq \mathbb{R}$ has a limit in $\overline{\mathbb{R}}$ if one of the following holds: 1) $\{z^t\}$ has a limit in the usual sense, 2) for all $c \in \mathbb{R}$, we have $z^t \geq c$ (resp. $z^t \leq c$) for all t sufficiently large in which case we say $\lim_t z^t = +\infty$ (resp. $\lim_t z^t = -\infty$).

The following are elementary properties of convergence in the extended reals:

Lemma B.2. *Let $\{z^t\}$ and $\{\tilde{z}^t\}$ be sequences in $\mathbb{R}$ with limits in $\overline{\mathbb{R}}$. Then $z^t + \tilde{z}^t$ has a limit in $\overline{\mathbb{R}}$ equal to $\lim_t z^t + \lim_t \tilde{z}^t$ if any of the following holds:*

1. *at least one of $\lim_t z^t$ or $\lim_t \tilde{z}^t$ is finite, i.e., $\in \mathbb{R}$,*
2. *$\{z^t\}_t$ and $\{\tilde{z}^t\}_t$ are both $\subseteq [0, \infty)$,*
3. *$\{z^t\}_t$ and $\{\tilde{z}^t\}_t$ are both $\subseteq (-\infty, 0]$.*

Proof. The results follow respectively from the following standard properties of the extended real numbers: 1. $\pm\infty + c = \pm\infty$ for any $c \in \mathbb{R}$, 2. $+\infty + \infty = +\infty$ and 3. $-\infty - \infty = -\infty$. □

Definition B.3. A function $f : \mathbb{R} \to \mathbb{R}$ is *monotone non-increasing* (resp. non-decreasing) if $f(x) \geq f(y)$ for all $x, y \in \mathbb{R}$ such that $x \leq y$ (resp. $x \geq y$).

Lemma B.4. *Let $f : \mathbb{R} \to \mathbb{R}$ be continuous and monotone non-increasing. Suppose that $\{z^t\}_t \subseteq \mathbb{R}$ has a limit $z^* \in \overline{\mathbb{R}}$. Then $f(z^t)$ has a limit $\in \overline{\mathbb{R}}$ and*

$$\lim_t f(z^t) = \begin{cases} f(z^*) & : z^* \in \mathbb{R} \\ \inf_{x \in \mathbb{R}} f(x) & : z^* = +\infty \\ \sup_{x \in \mathbb{R}} f(x) & : z^* = -\infty. \end{cases} \tag{6}$$

*Thus, the statement $\lim_t f(z^t) = f(\lim_t z^t)$ is correct. When f is monotone non-*decreasing*, Equation (6) holds with the* inf *and* sup *swapped.*

Proof. If $z^* \in \mathbb{R}$, then the result is simply the definition of continuity. Next, suppose that $z^* = +\infty$. Our goal is to show that $\lim_t f(z^t)$ exists and converges to $I := \inf_{x \in \mathbb{R}} f(x)$.

Consider the case that $I = -\infty$. Then for any $U \in \mathbb{R}$, there exists $u \in \mathbb{R}$ such that $f(u) \leq U$. Since $z^* = +\infty$, $z_t \geq u$ for all $t \gg 0$ sufficiently large, and in which case $f(z_t) \leq f(u) \leq U$. Since $U \in \mathbb{R}$ is arbitrary, we have that $\lim_t f(z^t) = -\infty$ (Definition B.1).

Now, consider the case that $I \in \mathbb{R}$. Then by definition $f(z^t) \geq I$ for all t. Furthermore, for any $\epsilon > 0$, there exists u such that $f(u) \leq I + \epsilon$. Again, since $z^* = +\infty$, $z_t \geq u$ for all $t \gg 0$ sufficiently large, in which case $f(z_t) \leq f(u) \leq I + \epsilon$. Since $\epsilon > 0$ is arbitrary, this proves that $\lim_t f(z^t) = I$. The proof for the case when $z^* = -\infty$ is completely analogous. Furthermore, when f is monotone non-decreasing, the roles of $\inf$ and $\sup$ are clearly swapped. □

Definition B.5. A sequence of vectors $\{\mathbf{v}^t\}_t \in \mathbb{R}^k$ is *totally convergent* if for all $y, j \in [k]$, both sequences of real numbers $\{v_y^t\}$ and $\{v_y^t - v_j^t\}$ have limits in $\overline{\mathbb{R}}$.

Lemma B.6. *Every sequence $\{\mathbf{v}^t\}_t \in \mathbb{R}^k$ has a subsequence that is totally convergent.*

Proof. Every sequence of real numbers has a convergent subsequence with limit in $\mathbb{R} \cup \{\pm\infty\}$.[7] By repeatedly passing to convergent subsequences, first for all $j \in [k]$, then for all pairs $j, j' \in [k]$ with $j < j'$, we get the desired result. □

Appendix C. Omitted proofs of results from Section 4

Proof of Lemma 4.1. To prove that $C_{\mathbf{p}}(\mathbf{v}) = C_{\sigma(\mathbf{p})}(\sigma(\mathbf{v}))$, we note that

$$C_{\mathbf{p}}(\mathbf{v}) = \sum_{y \in [k]} p_y \mathcal{L}_y(\mathbf{v}) = \sum_{y \in [k]} p_{\sigma(y)} \mathcal{L}_{\sigma(y)}(\mathbf{v}) = \sum_{y \in [k]} [\sigma(\mathbf{p})]_y \mathcal{L}_y(\sigma(\mathbf{v})) = C_{\sigma(\mathbf{p})}(\sigma(\mathbf{v})).$$

For the "Furthermore" part, note that $\mathbf{v} \mapsto \sigma^{-1}(\mathbf{v})$ is a bijection from $\mathbb{R}^k$ to itself. Hence,

$$\begin{aligned}
C^*_{\mathbf{p}} &= \inf\{C_{\mathbf{p}}(\mathbf{v}) : \mathbf{v} \in \mathbb{R}^k\} && \because \text{Definition 2.4} \\
&= \inf\{C_{\mathbf{p}}(\sigma^{-1}(\mathbf{v})) : \mathbf{v} \in \mathbb{R}^k\} && \because \mathbf{v} \mapsto \sigma^{-1}(\mathbf{v}) \text{ is a bijection} \\
&= \inf\{C_{\sigma(\mathbf{p})}(\sigma(\sigma^{-1}(\mathbf{v}))) : \mathbf{v} \in \mathbb{R}^k\} && \because C_{\mathbf{p}}(\mathbf{v}) = C_{\sigma(\mathbf{p})}(\sigma(\mathbf{v})) \\
&= \inf\{C_{\sigma(\mathbf{p})}(\mathbf{v}) : \mathbf{v} \in \mathbb{R}^k\} && \because \sigma(\sigma^{-1}(\mathbf{v})) = \mathbf{v}
\end{aligned}$$

The right hand side is equal to $\inf\{C_{\sigma(\mathbf{p})}(\mathbf{v}) : \mathbf{v} \in \mathbb{R}^k\} = C^*_{\sigma(\mathbf{p})}$ since $\sigma(\sigma^{-1}(\mathbf{v})) = \mathbf{v}$ □

Proof of Lemma 4.2. Since $\mathcal{L}$ is permutation equivariant, we have for all $j \in [k]$ that

$$\mathcal{L}_j(\mathbf{S}_\tau(\mathbf{v})) = [\mathcal{L}(\mathbf{S}_\tau(\mathbf{v}))]_j = [\mathbf{S}_\tau(\mathcal{L}(\mathbf{v}))]_j = [\mathcal{L}(\mathbf{v})]_{\tau(j)} = \mathcal{L}_{\tau(j)}(\mathbf{v}). \tag{7}$$

To finish the proof, we have

$$\begin{aligned}
&C_{\mathbf{p}}(\mathbf{v}) - C_{\mathbf{p}}(\tau(\mathbf{v})) \\
&= \sum_{j \in [k]} p_j \left(\mathcal{L}_j(\mathbf{v}) - \mathcal{L}_j(\tau(\mathbf{v}))\right) \\
&= \sum_{j \in [k]} p_j \left(\mathcal{L}_j(\mathbf{v}) - \mathcal{L}_{\tau(j)}(\mathbf{v})\right) && \because \text{Equation (7)} \\
&= (p_y \mathcal{L}_y(\mathbf{v}) + p_{y'} \mathcal{L}_{y'}(\mathbf{v})) - (p_y \mathcal{L}_{y'}(\mathbf{v}) + p_{y'} \mathcal{L}_y(\mathbf{v})) && \because \tau \text{ is a transposition} \\
&= p_y(\mathcal{L}_y(\mathbf{v}) - \mathcal{L}_{y'}(\mathbf{v})) + p_{y'}(\mathcal{L}_{y'}(\mathbf{v}) - \mathcal{L}_y(\mathbf{v})) \\
&= (p_y - p_{y'})(\mathcal{L}_y(\mathbf{v}) - \mathcal{L}_{y'}(\mathbf{v})),
\end{aligned}$$

as desired. □

Appendix D. Omitted proof of results from Section 5

Proof of Lemma 5.1. Let $y, j \in [k]$ be arbitrary. Since $\mathbf{v}^t$ is totally convergent, $v^t_y - v^t_j$ has a limit in $\mathbb{R} \cup \{\pm\infty\}$. Next, since ϕ is monotone and non-negative by condition (`Phi-NDZ`), we have by Lemma B.4 that $\phi(v^t_y - v^t_j)$ has a limit in $[0, +\infty]$. Now, that $\lim_t C_{\mathbf{p}}(\mathbf{v}^t)$ has a limit in $[0, +\infty]$ follows immediately from Lemma B.2.

Next, define $a^t_y := \sum_{j \in [k]: j \neq y} \phi(v^t_y - v^t_j)$ and $\tilde{a}^t_y := \sum_{j \in [k]: j \neq y} \phi(\tilde{v}^t_y - \tilde{v}^t_j)$. We proceed stepwise as follows:

7. This is follows from $\mathbb{R} \cup \{\pm\infty\}$ being the compactification of $\mathbb{R}$. See Oden and Demkowicz (2017, §1.16).

Step 1: $\lim_t \phi(v_y^t - v_j^t) = \lim_t \phi(\tilde{v}_y^t - \tilde{v}_j^t)$ as elements of $[0, +\infty]$,
Step 2: $\lim_t a_y^t = \lim_t \tilde{a}_y^t$ as elements of $[0, +\infty]$,
Step 3: $\lim_t \gamma(a_y^t) = \lim_t \gamma(\tilde{a}_y^t)$ as elements of $[0, +\infty]$
Step 4: $\lim_t \sum_{y\in[k]} p_y \gamma(a_y^t) = \lim_t \sum_{y\in[k]} p_y \gamma(\tilde{a}_y^t)$

Proof of Step 1. From Lemma B.4 and the fact that ϕ is monotone and continuous, we get that $\lim_t \phi(v_y^t - v_j^t) = \phi(\lim_t v_y^t - v_j^t)$ and $\lim_t \phi(\tilde{v}_y^t - \tilde{v}_j^t) = \phi(\lim_t \tilde{v}_y^t - \tilde{v}_j^t)$. Note that Lemma B.4 also guarantees that these limits exist. Non-negativity of the limit values follows from the non-negativity of ϕ.

Step 2. From Lemma B.2 and the non-negativity of ϕ, we have

$$\lim a_y^t = \sum_{j\in[k]\setminus\{y\}} \lim_t \phi(v_y^t - v_j^t) = \sum_{j\in[k]\setminus\{y\}} \lim_t \phi(\tilde{v}_y^t - \tilde{v}_j^t) = \lim \tilde{a}_y^t$$

where the equality in the middle follows from Step 1. Note that Lemma B.2 also guarantees that these limits exist.

Step 3. This follows from Step 2, Lemma B.4 and the non-negativity of γ on $[0, \infty)$.

Step 4. This follows from Step 3 and Lemma B.2. □

Proof of Lemma 5.6. Throughout this proof, let $\gamma'(x) := \frac{d\gamma}{dx}(x)$ and $\phi'(x) := \frac{d\phi}{dx}(x)$. Recall that

$$C_{\mathbf{q}}(\mathbf{v}) = \sum_{y\in[\ell]} q_y \gamma\Big(\sum_{j\in[k]\setminus\{y\}} \phi(v_y - v_j)\Big).$$

For each y, define $\Gamma_y(\mathbf{v}) := \gamma'\Big(\sum_{j\in[k]\setminus\{y\}} \phi(v_y - v_j)\Big)$. Thus

$$\frac{\partial C_{\mathbf{q}}}{\partial v_y}(\mathbf{v}) = \Big(q_y \Gamma_y(\mathbf{v}) \sum_{j\in[k]\setminus\{y\}} \phi'(v_y - v_j)\Big) - \Big(\sum_{j\in[k]\setminus\{y\}} q_j \Gamma_j(\mathbf{v}) \phi'(v_j - v_y)\Big). \tag{8}$$

The vanishing of the first two partial derivatives $\left[\frac{\partial C_{\mathbf{q}}}{\partial v_1}(\mathbf{v}) \quad \frac{\partial C_{\mathbf{q}}}{\partial v_2}(\mathbf{v})\right] = 0$ (i.e., Equation (8) where $y = 1, 2$) can be cast in matrix form equivalently as follows:

$$\begin{bmatrix} q_1\Gamma_1(\mathbf{v}) \\ q_2\Gamma_2(\mathbf{v}) \\ q_3\Gamma_3(\mathbf{v}) \\ \vdots \\ q_k\Gamma_k(\mathbf{v}) \end{bmatrix}^\top \begin{bmatrix} \sum_{j\in[k]\setminus\{1\}} \phi'(v_1 - v_j) & -\phi'(v_1 - v_2) \\ -\phi'(v_2 - v_1) & \sum_{j\in[k]\setminus\{2\}} \phi'(v_2 - v_j) \\ -\phi'(v_3 - v_1) & -\phi'(v_3 - v_2) \\ \vdots & \vdots \\ -\phi'(v_k - v_1) & -\phi'(v_k - v_2) \end{bmatrix} = \mathbf{0}.$$

The above equation is satisfied at $\mathbf{v} = \boldsymbol{\alpha}$, which satisfies $\alpha_1 = \alpha_2$ by assumption.

$$\begin{bmatrix} q_1\Gamma_1(\boldsymbol{\alpha}) \\ q_2\Gamma_2(\boldsymbol{\alpha}) \\ q_3\Gamma_3(\boldsymbol{\alpha}) \\ \vdots \\ q_k\Gamma_k(\boldsymbol{\alpha}) \end{bmatrix}^\top \begin{bmatrix} \sum_{j\in[k]\setminus\{1\}} \phi'(\alpha_1 - \alpha_j) & -\phi'(0) \\ -\phi'(0) & \sum_{j\in[k]\setminus\{2\}} \phi'(\alpha_2 - \alpha_j) \\ -\phi'(\alpha_3 - \alpha_1) & -\phi'(\alpha_3 - \alpha_1) \\ \vdots & \vdots \\ -\phi'(\alpha_k - \alpha_1) & -\phi'(\alpha_k - \alpha_1) \end{bmatrix} = \mathbf{0}.$$

Equivalently, we can rearrange the above equation as

$$\begin{bmatrix} q_1\Gamma_1(\boldsymbol{\alpha}) \\ q_2\Gamma_2(\boldsymbol{\alpha}) \end{bmatrix}^\top \begin{bmatrix} \sum_{j\in[k]\setminus\{1\}} \phi'(\alpha_1-\alpha_j) & -\phi'(0) \\ -\phi'(0) & \sum_{j\in[k]\setminus\{2\}} \phi'(\alpha_2-\alpha_j) \end{bmatrix}$$
$$= \underbrace{\begin{bmatrix} q_3\Gamma_3(\boldsymbol{\alpha}) \\ \vdots \\ q_k\Gamma_k(\boldsymbol{\alpha}) \end{bmatrix}^\top \begin{bmatrix} \phi'(\alpha_3-\alpha_1) \\ \vdots \\ \phi'(\alpha_k-\alpha_1) \end{bmatrix}}_{=:d} \begin{bmatrix} 1 & 1 \end{bmatrix} = d\mathbf{1}^\top$$

Furthermore, note that

$$\begin{aligned} \sum_{j\in[k]\setminus\{1\}} \phi'(\alpha_1-\alpha_j) &= \phi'(\alpha_1-\alpha_2) + \sum_{j\in[k]\setminus\{1,2\}} \phi'(\alpha_1-\alpha_j) \\ &= \phi'(0) + \sum_{j\in[k]\setminus\{1,2\}} \phi'(\alpha_1-\alpha_j) \\ &= \phi'(0) + \sum_{j\in[k]\setminus\{1,2\}} \phi'(\alpha_2-\alpha_j) \\ &= \sum_{j\in[k]\setminus\{2\}} \phi'(\alpha_2-\alpha_j). \end{aligned}$$

Likewise, $\Gamma_1(\boldsymbol{\alpha}) = \gamma'(\phi(0)+\sum_{j\in[k]\setminus\{1,2\}} \phi(v_1-v_j)) = \Gamma_2(\boldsymbol{\alpha})$. Let $a := \phi'(0)$, $b := \sum_{j\in[k]\setminus\{1,2\}} \phi'(\alpha_1-\alpha_j)$, and $c := \Gamma_1(\boldsymbol{\alpha})$. Since $\gamma'(\cdot) > 0$, we have $c > 0$ and so

$$c\begin{bmatrix} q_1 \\ q_2 \end{bmatrix}^\top \begin{bmatrix} a+b & -a \\ -a & a+b \end{bmatrix} = d\mathbf{1}^\top \implies \begin{bmatrix} a+b & -a \\ -a & a+b \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \end{bmatrix} = \frac{d}{c}\mathbf{1}.$$

Note that since $\phi' \le 0$ and $\phi'(0) \neq 0$, we have $a \in (-\infty, 0)$ and $b \in (-\infty, 0]$. Next, subtract the second equation from the first one in the above linear system, we get $(2a+b)(q_1-q_2) = 0$. Since $2a+b < 0$, we have that $q_1 = q_2$. □

Proof of Lemma 5.7. By Definition 2.5, there exists some $\mathbf{q} \in \Delta^k$ and $y \in [k]$ such that $q_y < \max_{j\in[k]} q_j$ and

$$C^*_{\mathbf{q}} = \inf\{C_{\mathbf{q}}(\mathbf{v}) : \mathbf{v} \in \mathbb{R}^k,\ v_y = \max_{j\in[k]} v_j\}.$$

The above implies that there exists a sequence $\{\mathbf{v}^t\}_t \subseteq \mathbb{R}^k$ such that $\lim_t C_{\mathbf{q}}(\mathbf{v}^t) = C^*_{\mathbf{q}}$ and $v^t_y = \max_{j\in[k]} v^t_j$ for all t. Let $\sigma \in \mathtt{Sym}(k)$ be such that $\sigma(\mathbf{q}) \in \Delta^k_{\mathtt{desc}}$. Let $\tilde{y} := \sigma^{-1}(y)$ and $z \in [k]$ be the smallest index such that $q_{\sigma(z)} = q_{\sigma(\tilde{y})}$ (note that $\sigma(\tilde{y}) = y$ by definition). Furthermore, we have that $z > 1$ since $q_{\sigma(1)} = \max \mathbf{q} > q_y = q_{\sigma(z)}$.

Let $\tau \in \mathtt{Sym}(k)$ be the permutation that swaps z and $\tilde{y}$ while leaving all other elements of $[k]$ unchanged. Note that if $z = \tilde{y}$, then τ is the trivial permutation, i.e., the identity map on $[k]$. Define $\mathbf{p} := \tau(\sigma(\mathbf{q}))$, and $\mathbf{w}^t := \tau(\sigma(\mathbf{v}^t))$. Observe that $\mathbf{p} = \tau(\sigma(\mathbf{q})) = \sigma(\mathbf{q})$ and thus $\mathbf{p} \in \Delta^k_{\mathtt{desc}}$ as well. We claim that $p_{z-1} > p_z$. To see this, note that

$$p_{z-1} = [\tau(\sigma(\mathbf{q}))]_{z-1} = [\sigma(\mathbf{q})]_{\tau(z-1)} = [\sigma(\mathbf{q})]_{z-1} = q_{\sigma(z-1)} > q_{\sigma(z)} = q_y$$

and

$$p_z = [\tau(\sigma(\mathbf{q}))]_z = [\sigma(\mathbf{q})]_{\tau(z)} = [\sigma(\mathbf{q})]_{\tilde{y}} = q_{\sigma(\tilde{y})} = q_y.$$

By Lemma 4.1, we have

$$\lim_t C_{\mathbf{p}}(\mathbf{w}^t) = \lim_t C_{\tau(\sigma(\mathbf{q}))}(\tau(\sigma(\mathbf{v}^t))) = \lim_t C_{\mathbf{q}}(\mathbf{v}^t) = C^*_{\mathbf{q}} = C^*_{\tau(\sigma(\mathbf{q}))} = C^*_{\mathbf{p}}.$$

Furthermore, we have $\max \mathbf{v}^t = \max \sigma(\mathbf{v}^t) = \max \mathbf{w}^t$ and so

$$w^t_z = [\mathbf{w}^t]_z = [\tau(\sigma(\mathbf{v}^t))]_z = [\sigma(\mathbf{v}^t)]_{\tau(z)} = v^t_{\sigma(\tau(z))} = v^t_{\sigma(\tilde{y})} = v^t_y = \max \mathbf{v}^t = \max \mathbf{w}^t.$$

In summary, we have an index $z \in [k]$ where $z > 1$ and a probability vector $\mathbf{p} \in \Delta^k_{\texttt{desc}}$ such that $p_{z-1} > p_z$. Furthermore, we have a sequence $\{\mathbf{w}^t\}_t$ such that $\lim_t C_{\mathbf{p}}(\mathbf{w}^t) = C^*_{\mathbf{p}}$ and $w^t_z = \max \mathbf{w}^t$. This implies the desired condition in the statement of Lemma 5.7. □

Appendix E. Omitted proof of results from Section 6

Lemma E.1. *In the setting of Section 6, we have $\ell = 2$ and $\boldsymbol{\alpha} = (0, 0)$.*

Proof of Lemma E.1. Note that since $\alpha_1 = 0$ already, for the "$\boldsymbol{\alpha} = (0,0)$" part we only need to show that $\alpha_2 = 0$. Now, to proceed, we first show that $\ell = 2$. To this end, we show that assuming $\ell \in \{1, 3, \ldots, k\}$ leads to a contradiction. First, assume that $\ell = 1$. Then we have $\lim_t v^t_2 = \cdots = \lim_t v^t_k = -\infty$. Since γ is increasing and $\phi \geq 0$, from Equation (4) we have for any $\mathbf{v} \in \mathbb{R}^k$ that

$$C_{\mathbf{p}}(\mathbf{v}) \geq r\gamma\Big(\sum_{j \in [k]\setminus\{1\}} \phi(v_1 - v_j)\Big) + (1-r)\gamma\left(\phi(v_2 - v_1)\right).$$

Since $v^t_1 = 0$ for all t, we have

$$\begin{aligned}
\lim_t C_{\mathbf{p}}(\mathbf{v}^t) &\geq \lim_t r\gamma\Big(\sum_{j \in [k]\setminus\{1\}} \phi(-v_j)\Big) + (1-r)\gamma\left(\phi(v_2)\right) \\
&= r\gamma\left((k-1)\phi(+\infty)\right) + (1-r)\gamma\left(\phi(-\infty)\right) \\
&= r\gamma(0) + (1-r)\gamma(+\infty) \\
&\geq +\infty. \qquad \because \gamma(+\infty) = +\infty \text{ (Definition 3.1)}
\end{aligned}$$

This is a contradiction since $C_{\mathbf{p}}(\mathbf{0}) = \gamma\left((k-1)\phi(0)\right) < +\infty$.

Next, we assume that $\ell \in \{3, \ldots, k\}$ and derive a contradiction. Recall our definition that $\mathbf{q} = (r, 1-r, 0, \ldots, 0)$. Now, for a generic $\mathbf{w} \in \mathbb{R}^\ell$, recall that $C_{\mathbf{q}}(\mathbf{w}) = r\mathcal{L}_1(\mathbf{w}) + (1-r)\mathcal{L}_2(\mathbf{w})$ where for $y \in \{1, 2\}$, we have

$$\mathcal{L}_y(\mathbf{w}) = \gamma\Big(\sum_{j \in [\ell]\setminus\{y\}} \phi(w_y - w_j)\Big).$$

Let $\epsilon > 0$ and define $\boldsymbol{\beta} \in \mathbb{R}^\ell$ by

$$\beta_j = \begin{cases} \alpha_j & : j \neq \ell \\ \alpha_\ell - \epsilon & : j = \ell. \end{cases}$$

For $y \in \{1, 2\}$, since $\beta_\ell < \alpha_\ell$ and $\beta_j = \alpha_j$ for $j \in [k] \setminus \{\ell\}$, we have

$$\begin{aligned}
&\begin{cases} \beta_y - \beta_j = \alpha_y - \alpha_j & : j \neq \ell \\ \beta_y - \beta_\ell > \alpha_y - \alpha_\ell & : j = \ell \end{cases} \\
&\implies \begin{cases} \phi(\beta_y - \beta_j) = \alpha_y - \alpha_j & : j \neq \ell \\ \phi(\beta_y - \beta_\ell) \leq \phi(\alpha_y - \alpha_\ell) & : j = \ell \end{cases} \\
&\implies \mathcal{L}_y(\boldsymbol{\beta}) = \gamma\Big(\sum_{j \in [\ell] \setminus \{y\}} \phi(\beta_y - \beta_j) \Big) \leq \gamma\Big(\sum_{j \in [\ell] \setminus \{y\}} \phi(\alpha_y - \alpha) \Big) = \mathcal{L}_y(\boldsymbol{\alpha}).
\end{aligned}$$

Thus, $C_{\mathbf{q}}(\boldsymbol{\alpha}) \geq C_{\mathbf{q}}(\boldsymbol{\beta})$ and so $C_{\mathbf{q}}(\boldsymbol{\alpha}) \geq \lim_{\epsilon \to \infty} C_{\mathbf{q}}(\boldsymbol{\beta})$ as well. By Lemma 5.5 and that $p_y = 0$ for $y \geq 2$, we have $\lim_t C_{\mathbf{p}}(\mathbf{v}^t) = C_{\mathbf{q}}(\boldsymbol{\alpha})$. Now, define $\{\tilde{\mathbf{v}}^t\}_t \subseteq \mathbb{R}^k$ by

$$\tilde{v}^t_j := \begin{cases} v^t_j & : j \neq \ell \\ -t & : j = \ell. \end{cases}$$

By construction we have $\lim_t C_{\mathbf{p}}(\tilde{\mathbf{v}}^t) = \lim_{\epsilon \to \infty} C_{\mathbf{q}}(\boldsymbol{\beta})$ and $\{\tilde{\mathbf{v}}^t\}_t \in \texttt{SEQ}$. Furthermore, since $\lim_t \tilde{v}^t_\ell = -\infty$, we have a contradiction of the minimality of ℓ (Equation 5).

Below, we can assume that $\ell = 2$, where we have $\mathbf{q} = [r, 1 - r] \in \Delta^2_{\texttt{desc}}$ and so

$$C_{\mathbf{q}}(\boldsymbol{\alpha}) = r\gamma\left(\phi(-\alpha_2)\right) + (1 - r)\gamma\left(\phi(\alpha_2)\right) = \inf_{\mathbf{w} \in \mathbb{R}^2} C_{\mathbf{q}}(\mathbf{w}). \tag{9}$$

Recall that our goal is to show that $\alpha_2 = 0$. To this end, consider the function

$$F(x) := r\gamma\left(\phi(x)\right) + (1 - r)\gamma\left(\phi(-x)\right).$$

Thus we have $C_{\mathbf{q}}(\boldsymbol{\alpha}) = \inf_{x \in \mathbb{R}} F(x)$. To finish the proof, it suffices to prove that F has a unique minimum at 0. We begin by computing the derivative $F'(x)$ of $F(x)$. Using the chain rule, we have

$$F'(x) = r\gamma'(\phi(x))\phi'(x) - (1 - r)\gamma'(\phi(-x))\phi'(-x).$$

Now, $\phi'(x) = -\exp(-x)$ and

$$\gamma'(x) = \begin{cases} -2(x - 1) & : x < 1 \\ 4(x - 1) & : x \geq 1. \end{cases}$$

If $x > 0$, then $\phi(x) < 1$ and $\phi(-x) > 1$. Thus, when $x > 0$, we have

$$\begin{aligned}
F'(x) &= r(-2(\exp(-x) - 1))(-\exp(-x)) - (1 - r)(4(\exp(x) - 1))(-\exp(x)) \\
&= 2r(\exp(-x) - 1)\exp(-x) + 4(1 - r)(\exp(x) - 1)\exp(x) \\
&=: G_+(x).
\end{aligned}$$

If $x \leq 0$, then $\phi(x) \geq 1$ and $\phi(-x) \leq 1$. Thus, when $x \leq 0$, we have

$$\begin{aligned}
F'(x) &= r(4(\exp(-x) - 1))(-\exp(-x)) - (1 - r)(-2(\exp(x) - 1))(-\exp(x)) \\
&= -4r(\exp(-x) - 1)\exp(-x) - 2(1 - r)(\exp(x) - 1)\exp(x) \\
&=: G_-(x).
\end{aligned}$$

Thus, by definition, we have

$$F'(x) = \begin{cases} G_+(x) & : x > 0 \\ G_-(x) & : x < 0 \\ 0 & : x = 0. \end{cases}$$

Finally, we prove that $F'(x)$ vanishes only at $x = 0$ under the assumption that $r \in [\frac{1}{3}, \frac{2}{3}]$. We now consider the zeros of both $G_+(x)$ and $G_-(x)$, i.e., $x \in \mathbb{R}$ where the functions vanish. Clearly, both functions vanish at $x = 0$. For $x \neq 0$, we compute

$$\begin{aligned} 0 = G_+(x) &= 2r(\exp(-x) - 1)\exp(-x) + 4(1-r)(\exp(x) - 1)\exp(x) \\ \iff \tfrac{r}{2(1-r)} &= -\tfrac{\exp(x)(\exp(x)-1)}{\exp(-x)(\exp(-x)-1)}. \end{aligned}$$

Simplifying the right hand side, we have

$$\begin{aligned} -\tfrac{\exp(x)(\exp(x)-1)}{\exp(-x)(\exp(-x)-1)} &= -\exp(2x)\tfrac{\exp(x)-1}{\exp(-x)-1} \\ &= -\exp(2x)\exp(x)\tfrac{1-\exp(-x)}{\exp(-x)-1} \\ &= \exp(3x). \end{aligned}$$

Thus, $0 = G_+(x)$ iff $\frac{1}{3}\ln\left(\frac{r}{2(1-r)}\right) = x$. Similarly, $0 = G_-(x)$ iff $\frac{1}{3}\ln\left(\frac{2r}{1-r}\right) = x$. Thus, $G_+(x)$ has a zero on $x > 0$ if and only if

$$\tfrac{1}{3}\ln\left(\tfrac{r}{2(1-r)}\right) > 0 \iff \tfrac{r}{2(1-r)} > 1 \iff r > 2/3.$$

Similarly, $G_-(x)$ has a zero on $x < 0$ if and only if

$$\tfrac{1}{3}\ln\left(\tfrac{2r}{1-r}\right) < 0 \iff \tfrac{2r}{1-r} < 1 \iff r < 1/3.$$

Taken together, we see that if $r \in [\frac{1}{3}, \frac{2}{3}]$, then $F'(x)$ only vanishes at $x = 0$. Moreover,

$$F'(\ln(2)) = G_+(\ln(2)) = 8 - \tfrac{17}{2}r \geq 8 - \tfrac{17}{2} \cdot \tfrac{2}{3} > 0$$

and likewise $F'(-\ln(2)) = G_-(-\ln(2)) = \frac{1}{2}(1 - 17r) \leq \frac{1}{2}(1 - 17\frac{1}{3}) < 0$. Thus, $F(x)$ is decreasing on $x < 0$ and increasing on $x > 0$. This proves that F has a unique minimizer at $x = 0$ and concludes the proof of Lemma E.1. □