Classification Calibration Dimension for General Multiclass Losses

Harish G. Ramaswamy **Shivani Agarwal**
Department of Computer Science and Automation
Indian Institute of Science, Bangalore 560012, India
{harish_gurup,shivani}@csa.iisc.ernet.in

Abstract

We study consistency properties of surrogate loss functions for general multiclass classification problems, defined by a general loss matrix. We extend the notion of classification calibration, which has been studied for binary and multiclass 0-1 classification problems (and for certain other specific learning problems), to the general multiclass setting, and derive necessary and sufficient conditions for a surrogate loss to be classification calibrated with respect to a loss matrix in this setting. We then introduce the notion of *classification calibration dimension* of a multiclass loss matrix, which measures the smallest 'size' of a prediction space for which it is possible to design a convex surrogate that is classification calibrated with respect to the loss matrix. We derive both upper and lower bounds on this quantity, and use these results to analyze various loss matrices. In particular, as one application, we provide a different route from the recent result of Duchi et al. (2010) for analyzing the difficulty of designing 'low-dimensional' convex surrogates that are consistent with respect to pairwise subset ranking losses. We anticipate the classification calibration dimension may prove to be a useful tool in the study and design of surrogate losses for general multiclass learning problems.

1 Introduction

There has been significant interest and progress in recent years in understanding consistency of learning methods for various finite-output learning problems, such as binary classification, multiclass 0-1 classification, and various forms of ranking and multi-label prediction problems [1–15]. Such finite-output problems can all be viewed as instances of a general multiclass learning problem, whose structure is defined by a loss function, or equivalently, by a loss matrix. While the studies above have contributed to the understanding of learning problems corresponding to certain forms of loss matrices, a framework for analyzing consistency properties for a general multiclass learning problem, defined by a general loss matrix, has remained elusive.

In this paper, we analyze consistency of surrogate losses for general multiclass learning problems, building on the results of [3, 5–7] and others. We start in Section 2 with some background and examples that will be used as running examples to illustrate concepts throughout the paper, and formalize the notion of classification calibration with respect to a general loss matrix. In Section 3, we derive both necessary and sufficient conditions for classification calibration with respect to general multiclass losses; these are both of independent interest and useful in our later results. Section 4 introduces the notion of *classification calibration dimension* of a loss matrix, a fundamental quantity that measures the smallest 'size' of a prediction space for which it is possible to design a convex surrogate that is classification calibrated with respect to the loss matrix. We derive both upper and lower bounds on this quantity, and use these results to analyze various loss matrices. As one application, in Section 5, we provide a different route from the recent result of Duchi et al. [10] for analyzing the difficulty of designing 'low-dimensional' convex surrogates that are consistent with respect to certain pairwise subset ranking losses. We conclude in Section 6 with some future directions.

2 Preliminaries, Examples, and Background

Setup. We are given training examples $(X_1, Y_1), \ldots, (X_m, Y_m)$ drawn i.i.d. from a distribution D on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is an instance space and $\mathcal{Y} = [n] = \{1, \ldots, n\}$ is a finite set of *class labels*. We are also given a finite set $\mathcal{T} = [k] = \{1, \ldots, k\}$ of *target labels* in which predictions are to be made, and a *loss function* $\ell : \mathcal{Y} \times \mathcal{T} \to [0, \infty)$, where $\ell(y, t)$ denotes the loss incurred on predicting $t \in \mathcal{T}$ when the label is $y \in \mathcal{Y}$. In many common learning problems, $\mathcal{T} = \mathcal{Y}$, but in general, these could be different (e.g. when there is an'abstain' option available to a classifier, in which case $k = n + 1$).

We will find it convenient to represent the loss function ℓ as a *loss matrix* $\mathbf{L} \in \mathbb{R}_+^{n \times k}$ (here $\mathbb{R}_+ = [0, \infty)$), and for each $y \in [n], t \in [k]$, will denote by ℓ_{yt} the (y, t)-th element of $\mathbf{L}$, $\ell_{yt} = (\mathbf{L})_{yt} = \ell(y, t)$, and by $\boldsymbol{\ell}_t$ the t-th column of $\mathbf{L}$, $\boldsymbol{\ell}_t = (\ell_{1t}, \ldots, \ell_{nt})^\top \in \mathbb{R}^n$. Some examples follow:

Example 1 (0-1 loss)**.** *Here $\mathcal{Y} = \mathcal{T} = [n]$, and the loss incurred is 1 if the predicted label t is different from the actual class label y, and 0 otherwise: $\ell^{\text{0-1}}(y, t) = \mathbf{1}(t \neq y)$, where $\mathbf{1}(\cdot)$ is 1 if the argument is true and 0 otherwise. The loss matrix $\mathbf{L}^{\text{0-1}}$ for $n = 3$ is shown in Figure 1(a).*

Example 2 (Ordinal regression loss)**.** *Here $\mathcal{Y} = \mathcal{T} = [n]$, and predictions t farther away from the actual class label y are penalized more heavily, e.g. using absolute distance: $\ell^{\text{ord}}(y, t) = |t - y|$. The loss matrix $\mathbf{L}^{\text{ord}}$ for $n = 3$ is shown in Figure 1(b).*

Example 3 (Hamming loss)**.** *Here $\mathcal{Y} = \mathcal{T} = [2^r]$ for some $r \in \mathbb{N}$, and the loss incurred on predicting t when the actual class label is y is the number of bit-positions in which the r-bit binary representations of $t - 1$ and $y - 1$ differ: $\ell^{\text{Ham}}(y, t) = \sum_{i=1}^{r} \mathbf{1}((t-1)_i \neq (y-1)_i)$, where for any $z \in \{0, \ldots, 2^r - 1\}$, $z_i \in \{0, 1\}$ denotes the i-th bit in the r-bit binary representation of z. The loss matrix $\mathbf{L}^{\text{Ham}}$ for $r = 2$ is shown in Figure 1(c). This loss is used in sequence labeling tasks [16].*

Example 4 ('Abstain' loss)**.** *Here $\mathcal{Y} = [n]$ and $\mathcal{T} = [n+1]$, where $t = n+1$ denotes 'abstain'. One possible loss function in this setting assigns a loss of 1 to incorrect predictions in $[n]$, 0 to correct predictions, and $\frac{1}{2}$ for abstaining: $\ell^{(?)}(y, t) = \mathbf{1}(t \neq y)\,\mathbf{1}(t \in [n]) + \frac{1}{2}\mathbf{1}(t = n + 1)$. The loss matrix $\mathbf{L}^{(?)}$ for $n = 3$ is shown in Figure 1(d).*

The goal in the above setting is to learn from the training examples a function $h : \mathcal{X} \to [k]$ with low expected loss on a new example drawn from D, which we will refer to as the *ℓ-risk of h*:

$$\text{er}_D^\ell[h] \triangleq \mathbf{E}_{(X,Y)\sim D}\,\ell(Y, h(X)) = \mathbf{E}_X \sum_{y=1}^{n} p_y(X)\ell(y, h(X)) = \mathbf{E}_X \mathbf{p}(X)^\top \boldsymbol{\ell}_{h(X)}\,, \tag{1}$$

where $p_y(x) = \mathbf{P}(Y = y \mid X = x)$ under D, and $\mathbf{p}(x) = (p_1(x), \ldots, p_n(x))^\top \in \mathbb{R}^n$ denotes the conditional probability vector at x. In particular, the goal is to learn a function with ℓ-risk close to the *optimal ℓ-risk*, defined as

$$\text{er}_D^{\ell,*} \triangleq \inf_{h:\mathcal{X}\to[k]} \text{er}_D^\ell[h] = \inf_{h:\mathcal{X}\to[k]} \mathbf{E}_X \mathbf{p}(X)^\top \boldsymbol{\ell}_{h(X)} = \mathbf{E}_X \min_{t\in[k]} \mathbf{p}(X)^\top \ell_t\,. \tag{2}$$

Minimizing the discrete ℓ-risk directly is typically difficult computationally; consequently, one usually employs a *surrogate loss function* $\psi : \mathcal{Y} \times \widehat{\mathcal{T}} \to \mathbb{R}_+$ operating on a *surrogate target space* $\widehat{\mathcal{T}} \subseteq \mathbb{R}^d$ for some appropriate $d \in \mathbb{N}$,[1] and minimizes (approximately, based on the training sample) the *ψ-risk* instead, defined for a (vector) function $\mathbf{f} : \mathcal{X} \to \widehat{\mathcal{T}}$ as

$$\text{er}_D^\psi[\mathbf{f}] \triangleq \mathbf{E}_{(X,Y)\sim D}\,\psi(Y, \mathbf{f}(X)) = \mathbf{E}_X \sum_{y=1}^{n} p_y(X)\psi(y, \mathbf{f}(X))\,. \tag{3}$$

The learned function $\mathbf{f} : \mathcal{X} \to \widehat{\mathcal{T}}$ is then used to make predictions in $[k]$ via some transformation $\text{pred} : \widehat{\mathcal{T}} \to [k]$: the prediction on a new instance $x \in \mathcal{X}$ is given by $\text{pred}(\mathbf{f}(x))$, and the ℓ-risk incurred is $\text{er}_D^\ell[\text{pred} \circ \mathbf{f}]$. As an example, several algorithms for multiclass classification with respect to 0-1 loss learn a function of the form $\mathbf{f} : \mathcal{X} \to \mathbb{R}^n$ and predict according to $\text{pred}(\mathbf{f}(x)) = \text{argmax}_{t\in[n]} f_t(x)$.

Below we will find it useful to represent the surrogate loss function ψ via n real-valued functions $\psi_y : \widehat{\mathcal{T}} \to \mathbb{R}_+$ defined as $\psi_y(\hat{\mathbf{t}}) = \psi(y, \hat{\mathbf{t}})$ for $y \in [n]$, or equivalently, as a vector-valued function $\boldsymbol{\psi} : \widehat{\mathcal{T}} \to \mathbb{R}_+^n$ defined as $\boldsymbol{\psi}(\hat{\mathbf{t}}) = (\psi_1(\hat{\mathbf{t}}), \ldots, \psi_n(\hat{\mathbf{t}}))^\top$. We will also define the sets

$$\mathcal{R}_\psi \triangleq \left\{\boldsymbol{\psi}(\hat{\mathbf{t}}) \,:\, \hat{\mathbf{t}} \in \widehat{\mathcal{T}}\right\} \quad \text{and} \quad \mathcal{S}_\psi \triangleq \text{conv}(\mathcal{R}_\psi)\,, \tag{4}$$

where for any $\mathcal{A} \subseteq \mathbb{R}^n$, $\text{conv}(\mathcal{A})$ denotes the convex hull of $\mathcal{A}$.

[1] Equivalently, one can define $\psi : \mathcal{Y} \times \mathbb{R}^d \to \bar{\mathbb{R}}_+$, where $\bar{\mathbb{R}}_+ = \mathbb{R}_+ \cup \{\infty\}$ and $\psi(y, \hat{\mathbf{t}}) = \infty\ \forall \hat{\mathbf{t}} \notin \widehat{\mathcal{T}}$.

$$\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 1 & 2 \\ 1 & 0 & 1 \\ 2 & 1 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 1 & 1 & 2 \\ 1 & 0 & 2 & 1 \\ 1 & 2 & 0 & 1 \\ 2 & 1 & 1 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 1 & 1 & \frac{1}{2} \\ 1 & 0 & 1 & \frac{1}{2} \\ 1 & 1 & 0 & \frac{1}{2} \end{bmatrix}$$

(a) (b) (c) (d)

Figure 1: Loss matrices corresponding to Examples 1-4: (a) $\mathbf{L}^{\text{0-1}}$ for $n = 3$; (b) $\mathbf{L}^{\text{ord}}$ for $n = 3$; (c) $\mathbf{L}^{\text{Ham}}$ for $r = 2$ ($n = 4$); (d) $\mathbf{L}^{(?)}$ for $n = 3$.

Under suitable conditions, algorithms that approximately minimize the ψ-risk based on a training sample are known to be consistent with respect to the ψ-risk, i.e. to converge (in probability) to the *optimal ψ-risk*, defined as

$$\text{er}_D^{\psi,*} \triangleq \inf_{\mathbf{f}:\mathcal{X}\to\widehat{\mathcal{T}}} \text{er}_D^{\psi}[\mathbf{f}] = \inf_{\mathbf{f}:\mathcal{X}\to\widehat{\mathcal{T}}} \mathbf{E}_X \mathbf{p}(X)^\top \boldsymbol{\psi}(\mathbf{f}(X)) = \mathbf{E}_X \inf_{\mathbf{z}\in\mathcal{R}_\psi} \mathbf{p}(X)^\top \mathbf{z} = \mathbf{E}_X \inf_{\mathbf{z}\in\mathcal{S}_\psi} \mathbf{p}(X)^\top \mathbf{z} \,. \tag{5}$$

This raises the natural question of whether, for a given loss ℓ, there are surrogate losses ψ for which consistency with respect to the ψ-risk also guarantees consistency with respect to the ℓ-risk, i.e. guarantees convergence (in probability) to the optimal ℓ-risk (defined in Eq. (2)). This question has been studied in detail for the 0-1 loss, and for square losses of the form $\ell(y,t) = a_y \mathbf{1}(t \neq y)$, which can be analyzed similarly to the 0-1 loss [6, 7]. In this paper, we consider this question for general multiclass losses $\ell : [n] \times [k] \to \mathbb{R}_+$, including rectangular losses with $k \neq n$. The only assumption we make on ℓ is that for each $t \in [k]$, $\exists \mathbf{p} \in \Delta_n$ such that $\text{argmin}_{t'\in[k]} \mathbf{p}^\top \ell_{t'} = \{t\}$ (otherwise the label t never needs to be predicted and can simply be ignored).[2]

Definitions and Results. We will need the following definitions and basic results, generalizing those of [5–7]. The notion of classification calibration will be central to our study; as Theorem 3 below shows, classification calibration of a surrogate loss ψ w.r.t. ℓ corresponds to the property that consistency w.r.t. ψ-risk implies consistency w.r.t. ℓ-risk. Proofs of these results are straightforward generalizations of those in [6, 7] and are omitted.

Definition 1 (Classification calibration)**.** *A surrogate loss function $\psi : [n] \times \widehat{\mathcal{T}} \to \mathbb{R}_+$ is said to be* classification calibrated *with respect to a loss function $\ell : [n] \times [k] \to \mathbb{R}_+$ over $\mathcal{P} \subseteq \Delta_n$ if there exists a function* pred : $\widehat{\mathcal{T}} \to [k]$ *such that*

$$\forall \mathbf{p} \in \mathcal{P} : \quad \inf_{\hat{\mathbf{t}}\in\widehat{\mathcal{T}}:\text{pred}(\hat{\mathbf{t}})\notin\text{argmin}_t \mathbf{p}^\top \boldsymbol{\ell}_t} \mathbf{p}^\top \boldsymbol{\psi}(\hat{\mathbf{t}}) \;>\; \inf_{\hat{\mathbf{t}}\in\widehat{\mathcal{T}}} \mathbf{p}^\top \boldsymbol{\psi}(\hat{\mathbf{t}}) \,.$$

Lemma 2. *Let $\ell : [n] \times [k] \to \mathbb{R}_+$ and $\psi : [n] \times \widehat{\mathcal{T}} \to \mathbb{R}_+$. Then ψ is classification calibrated with respect to ℓ over $\mathcal{P} \subseteq \Delta_n$ iff there exists a function* pred$'$: $\mathcal{S}_\psi \to [k]$ *such that*

$$\forall \mathbf{p} \in \mathcal{P} : \quad \inf_{\mathbf{z}\in\mathcal{S}_\psi:\text{pred}'(\mathbf{z})\notin\text{argmin}_t \mathbf{p}^\top \boldsymbol{\ell}_t} \mathbf{p}^\top \mathbf{z} \;>\; \inf_{\mathbf{z}\in\mathcal{S}_\psi} \mathbf{p}^\top \mathbf{z} \,.$$

Theorem 3. *Let $\ell : [n] \times [k] \to \mathbb{R}_+$ and $\psi : [n] \times \widehat{\mathcal{T}} \to \mathbb{R}_+$. Then ψ is classification calibrated with respect to ℓ over Δ_n iff $\exists$ a function* pred : $\widehat{\mathcal{T}} \to [k]$ *such that for all distributions D on $\mathcal{X} \times [n]$ and all sequences of random (vector) functions $\mathbf{f}_m : \mathcal{X} \to \widehat{\mathcal{T}}$ (depending on $(X_1, Y_1), \ldots, (X_m, Y_m)$),*[3]

$$\text{er}_D^{\psi}[\mathbf{f}_m] \xrightarrow{P} \text{er}_D^{\psi,*} \quad \textit{implies} \quad \text{er}_D^{\ell}[\text{pred} \circ \mathbf{f}_m] \xrightarrow{P} \text{er}_D^{\ell,*} \,.$$

Definition 4 (Positive normals)**.** *Let $\psi : [n] \times \widehat{\mathcal{T}} \to \mathbb{R}_+$. For each point $\mathbf{z} \in \mathcal{S}_\psi$, the set of* positive normals at $\mathbf{z}$ *is defined as*[4]

$$\mathcal{N}_{\mathcal{S}_\psi}(\mathbf{z}) \triangleq \left\{ \mathbf{p} \in \Delta_n : \mathbf{p}^\top (\mathbf{z} - \mathbf{z}') \leq 0 \; \forall \mathbf{z}' \in \mathcal{S}_\psi \right\} .$$

Definition 5 (Trigger probabilities)**.** *Let $\ell : [n] \times [k] \to \mathbb{R}_+$. For each $t \in [k]$, the set of* trigger probabilities of t *with respect to ℓ is defined as*

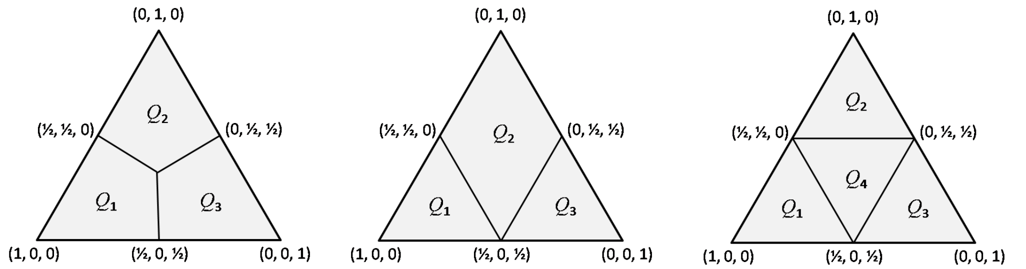
$$\mathcal{Q}_t^\ell \triangleq \left\{ \mathbf{p} \in \Delta_n : \mathbf{p}^\top (\boldsymbol{\ell}_t - \boldsymbol{\ell}_{t'}) \leq 0 \; \forall t' \in [k] \right\} \;=\; \left\{ \mathbf{p} \in \Delta_n : t \in \text{argmin}_{t'\in[k]} \mathbf{p}^\top \boldsymbol{\ell}_{t'} \right\} .$$

Examples of trigger probability sets for various losses are shown in Figure 2.

[2] Here Δ_n denotes the probability simplex in $\mathbb{R}^n$, $\Delta_n = \{\mathbf{p} \in \mathbb{R}^n : p_i \geq 0 \; \forall \, i \in [n], \sum_{i=1}^n p_i = 1\}$.

[3] Here $\xrightarrow{P}$ denotes convergence in probability.

[4] The set of positive normals is non-empty only at points $\mathbf{z}$ in the boundary of $\mathcal{S}_\psi$.



$Q_1^{\text{0-1}} = \{\mathbf{p} \in \Delta_3 : p_1 \geq \max(p_2, p_3)\}$
$Q_2^{\text{0-1}} = \{\mathbf{p} \in \Delta_3 : p_2 \geq \max(p_1, p_3)\}$
$Q_3^{\text{0-1}} = \{\mathbf{p} \in \Delta_3 : p_3 \geq \max(p_1, p_2)\}$

(a)

$Q_1^{\text{ord}} = \{\mathbf{p} \in \Delta_3 : p_1 \geq \frac{1}{2}\}$
$Q_2^{\text{ord}} = \{\mathbf{p} \in \Delta_3 : p_1 \leq \frac{1}{2}, p_3 \leq \frac{1}{2}\}$
$Q_3^{\text{ord}} = \{\mathbf{p} \in \Delta_3 : p_3 \geq \frac{1}{2}\}$

(b)

$Q_1^{(?)} = \{\mathbf{p} \in \Delta_3 : p_1 \geq \frac{1}{2}\}$
$Q_2^{(?)} = \{\mathbf{p} \in \Delta_3 : p_2 \geq \frac{1}{2}\}$
$Q_3^{(?)} = \{\mathbf{p} \in \Delta_3 : p_3 \geq \frac{1}{2}\}$
$Q_4^{(?)} = \{\mathbf{p} \in \Delta_3 : \max(p_1, p_2, p_3) \leq \frac{1}{2}\}$

(c)

Figure 2: Trigger probability sets for (a) 0-1 loss $\ell^{\text{0-1}}$; (b) ordinal regression loss ℓ^{ord}; and (c) 'abstain' loss $\ell^{(?)}$; all for $n = 3$, for which the probability simplex can be visualized easily. Calculations of these sets can be found in the appendix. We note that such sets have also been studied in [17, 18].

3 Necessary and Sufficient Conditions for Classification Calibration

We start by giving a necessary condition for classification calibration of a surrogate loss ψ with respect to any multiclass loss ℓ over Δ_n, which requires the positive normals of all points $\mathbf{z} \in \mathcal{S}_\psi$ to be 'well-behaved' w.r.t. ℓ and generalizes the 'admissibility' condition used for 0-1 loss in [7]. All proofs not included in the main text can be found in the appendix.

Theorem 6. *Let $\psi : [n] \times \widehat{\mathcal{T}} \to \mathbb{R}_+$ be classification calibrated with respect to $\ell : [n] \times [k] \to \mathbb{R}_+$ over Δ_n. Then for all $\mathbf{z} \in \mathcal{S}_\psi$, there exists some $t \in [k]$ such that $\mathcal{N}_{\mathcal{S}_\psi}(\mathbf{z}) \subseteq \mathcal{Q}_t^\ell$.*

We note that, as in [7], it is possible to give a necessary and sufficient condition for classification calibration in terms of a similar property holding for positive normals associated with projections of $\mathcal{S}_\psi$ in lower dimensions. Instead, below we give a different sufficient condition that will be helpful in showing classification calibration of certain surrogates. In particular, we show that for a surrogate loss ψ to be classification calibrated with respect to ℓ over Δ_n, it is sufficient for the above property of positive normals to hold only at a finite number of points in $\mathcal{R}_\psi$, as long as their positive normal sets jointly cover Δ_n:

Theorem 7. *Let $\ell : [n] \times [k] \to \mathbb{R}_+$ and $\psi : [n] \times \widehat{\mathcal{T}} \to \mathbb{R}_+$. Suppose there exist $r \in \mathbb{N}$ and $\mathbf{z}_1, \ldots, \mathbf{z}_r \in \mathcal{R}_\psi$ such that $\bigcup_{j=1}^r \mathcal{N}_{\mathcal{S}_\psi}(\mathbf{z}_j) = \Delta_n$ and for each $j \in [r]$, $\exists t \in [k]$ such that $\mathcal{N}_{\mathcal{S}_\psi}(\mathbf{z}_j) \subseteq \mathcal{Q}_t^\ell$. Then ψ is classification calibrated with respect to ℓ over Δ_n.*

Computation of $\mathcal{N}_{\mathcal{S}_\psi}(\mathbf{z})$. The conditions in the above results both involve the sets of positive normals $\mathcal{N}_{\mathcal{S}_\psi}(\mathbf{z})$ at various points $\mathbf{z} \in S_\psi$. Thus in order to use the above results to show that a surrogate ψ is (or is not) classification calibrated with respect to a loss ℓ, one needs to be able to compute or characterize the sets $\mathcal{N}_{S_\psi}(\mathbf{z})$. Here we give a method for computing these sets for certain surrogate losses ψ and points $\mathbf{z} \in \mathcal{S}_\psi$.

Lemma 8. *Let $\widehat{\mathcal{T}} \subseteq \mathbb{R}^d$ be a convex set and let $\boldsymbol{\psi} : \widehat{\mathcal{T}} \to \mathbb{R}_+^n$ be convex.[5] Let $\mathbf{z} = \boldsymbol{\psi}(\hat{\mathbf{t}})$ for some $\hat{\mathbf{t}} \in \widehat{\mathcal{T}}$ such that for each $y \in [n]$, the subdifferential of ψ_y at $\hat{\mathbf{t}}$ can be written as $\partial\psi_y(\hat{\mathbf{t}}) = \text{conv}(\{\mathbf{w}_1^y, \ldots, \mathbf{w}_{s_y}^y\})$ for some $s_y \in \mathbb{N}$ and $\mathbf{w}_1^y, \ldots, \mathbf{w}_{s_y}^y \in \mathbb{R}^d$.[6] Let $s = \sum_{y=1}^n s_y$, and let*

$$\mathbf{A} = \left[\mathbf{w}_1^1 \ldots \mathbf{w}_{s_1}^1 \mathbf{w}_1^2 \ldots \mathbf{w}_{s_2}^2 \ldots \ldots \mathbf{w}_1^n \ldots \mathbf{w}_{s_n}^n\right] \in \mathbb{R}^{d \times s}; \qquad \mathbf{B} = [b_{yj}] \in \mathbb{R}^{n \times s},$$

where b_{yj} is 1 if the j-th column of $\mathbf{A}$ came from $\{\mathbf{w}_1^y, \ldots, \mathbf{w}_{s_y}^y\}$ and 0 otherwise. Then

$$\mathcal{N}_{S_\psi}(\mathbf{z}) = \left\{\mathbf{p} \in \Delta_n : \mathbf{p} = \mathbf{B}\mathbf{q} \textit{ for some } \mathbf{q} \in \text{Null}(\mathbf{A}) \cap \Delta_s\right\},$$

where $\text{Null}(\mathbf{A}) \subseteq \mathbb{R}^s$ *denotes the null space of the matrix* $\mathbf{A}$.

[5] A vector function is convex if all its component functions are convex.

[6] Recall that the subdifferential of a convex function $\phi : \mathbb{R}^d \to \bar{\mathbb{R}}_+$ at a point $\mathbf{u}_0 \in \mathbb{R}^d$ is defined as $\partial\phi(\mathbf{u}_0) = \left\{\mathbf{w} \in \mathbb{R}^d : \phi(\mathbf{u}) - \phi(\mathbf{u}_0) \geq \mathbf{w}^\top(\mathbf{u} - \mathbf{u}_0)\ \forall \mathbf{u} \in \mathbb{R}^d\right\}$ and is a convex set in $\mathbb{R}^d$ (e.g. see [19]).

We give an example illustrating the use of Theorem 7 and Lemma 8 to show classification calibration of a certain surrogate loss with respect to the ordinal regression loss ℓ^{ord} defined in Example 2:

Example 5 (Classification calibrated surrogate for ordinal regression loss)**.** *Consider the ordinal regression loss ℓ^{ord} defined in Example 2 for $n = 3$. Let $\widehat{\mathcal{T}} = \mathbb{R}$, and let $\psi : \{1, 2, 3\} \times \mathbb{R} \to \mathbb{R}_+$ be defined as (see Figure 3)*

$$\psi(y, \hat{t}) = |\hat{t} - y| \qquad \forall y \in \{1, 2, 3\},\ \hat{t} \in \mathbb{R}\,. \tag{6}$$

Thus $\mathcal{R}_\psi = \big\{\boldsymbol{\psi}(\hat{t}) = \big(|\hat{t}-1|, |\hat{t}-2|, |\hat{t}-3|\big)^\top : \hat{t} \in \mathbb{R}\big\}$. We will show there are 3 points in $\mathcal{R}_\psi$ satisfying the conditions of Theorem 7. Specifically, consider $\hat{t}_1 = 1$, $\hat{t}_2 = 2$, and $\hat{t}_3 = 3$, giving $\mathbf{z}_1 = \boldsymbol{\psi}(\hat{t}_1) = (0, 1, 2)^\top$, $\mathbf{z}_2 = \boldsymbol{\psi}(\hat{t}_2) = (1, 0, 1)^\top$, and $\mathbf{z}_3 = \boldsymbol{\psi}(\hat{t}_3) = (2, 1, 0)^\top$ in $\mathcal{R}_\psi$. Observe that $\widehat{\mathcal{T}}$ here is a convex set and $\boldsymbol{\psi} : \widehat{\mathcal{T}} \to \mathbb{R}^3$ is a convex function. Moreover, for $\hat{t}_1 = 1$, we have

$$\begin{aligned}
\partial\psi_1(1) &= [-1, 1] = \text{conv}(\{+1, -1\})\,; \\
\partial\psi_2(1) &= \{-1\} = \text{conv}(\{-1\})\,; \\
\partial\psi_3(1) &= \{-1\} = \text{conv}(\{-1\})\,.
\end{aligned}$$

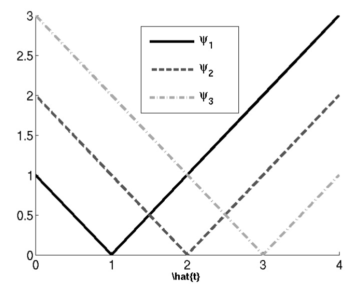


Figure 3: The surrogate ψ

Therefore, we can use Lemma 8 to compute $\mathcal{N}_{\mathcal{S}_\psi}(\mathbf{z}_1)$. Here $s = 4$, and

$$\mathbf{A} = \begin{bmatrix} +1 & -1 & -1 & -1 \end{bmatrix}\,; \qquad \mathbf{B} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

This gives

$$\begin{aligned}
\mathcal{N}_{\mathcal{S}_\psi}(\mathbf{z}_1) &= \big\{\mathbf{p} \in \Delta_3 : \mathbf{p} = (q_1 + q_2, q_3, q_4) \text{ for some } \mathbf{q} \in \Delta_4,\ q_1 - q_2 - q_3 - q_4 = 0\big\} \\
&= \big\{\mathbf{p} \in \Delta_3 : \mathbf{p} = (q_1 + q_2, q_3, q_4) \text{ for some } \mathbf{q} \in \Delta_4,\ q_1 = \tfrac{1}{2}\big\} \\
&= \big\{\mathbf{p} \in \Delta_3 : p_1 \geq \tfrac{1}{2}\big\} \\
&= \mathcal{Q}_1^{\text{ord}}\,.
\end{aligned}$$

A similar procedure yields $\mathcal{N}_{\mathcal{S}_\psi}(\mathbf{z}_2) = \mathcal{Q}_2^{\text{ord}}$ and $\mathcal{N}_{\mathcal{S}_\psi}(\mathbf{z}_3) = \mathcal{Q}_3^{\text{ord}}$. Thus, by Theorem 7, we get that ψ is classification calibrated with respect to ℓ^{ord} over Δ_3.

We note that in general, computational procedures such as Fourier-Motzkin elimination [20] can be helpful in computing $\mathcal{N}_{\mathcal{S}_\psi}(\mathbf{z})$ via Lemma 8.

4 Classification Calibration Dimension

We now turn to the study of a fundamental quantity associated with the property of classification calibration with respect to a general multiclass loss ℓ. Specifically, in the above example, we saw that to develop a classification calibrated surrogate loss w.r.t. the ordinal regression loss for $n = 3$, it was sufficient to consider a surrogate target space $\widehat{\mathcal{T}} = \mathbb{R}$, with dimension $d = 1$; in addition, this yielded a convex surrogate $\boldsymbol{\psi} : \mathbb{R} \to \mathbb{R}^3_+$ which can be used in developing computationally efficient algorithms. In fact the same surrogate target space with $d = 1$ can be used to develop a similar convex, classification calibrated surrogate loss w.r.t. the ordinal regression loss for *any* $n \in \mathbb{N}$. However not all losses ℓ have such 'low-dimensional' surrogates. This raises the natural question of what is the smallest dimension d that supports a convex classification calibrated surrogate for a given multiclass loss ℓ, and leads us to the following definition:

Definition 9 (Classification calibration dimension)**.** *Let $\ell : [n] \times [k] \to \mathbb{R}_+$. Define the* classification calibration dimension (CC dimension) *of ℓ as*

$$\text{CCdim}(\ell) \ \stackrel{\triangle}{=}\ \min\big\{d \in \mathbb{N} : \exists \text{ a convex set } \widehat{\mathcal{T}} \subseteq \mathbb{R}^d \text{ and a convex surrogate } \boldsymbol{\psi} : \widehat{\mathcal{T}} \to \mathbb{R}^n_+ \text{ that is classification calibrated w.r.t. } \ell \text{ over } \Delta_n\big\}\,,$$

if the above set is non-empty, and $\text{CCdim}(\ell) = \infty$ *otherwise.*

From the above discussion, $\text{CCdim}(\ell^{\text{ord}}) = 1$ for all n. In the following, we will be interested in developing an understanding of the CC dimension for general losses ℓ, and in particular in deriving upper and lower bounds on this.

4.1 Upper Bounds on the Classification Calibration Dimension

We start with a simple result that establishes that the CC dimension of any multiclass loss ℓ is finite, and in fact is strictly smaller than the number of class labels n.

Lemma 10. *Let* $\ell : [n] \times [k] \to \mathbb{R}_+$. *Let* $\widehat{\mathcal{T}} = \left\{ \hat{\mathbf{t}} \in \mathbb{R}_+^{n-1} : \sum_{j=1}^{n-1} \hat{t}_j \leq 1 \right\}$, *and for each* $y \in [n]$, *let* $\psi_y : \widehat{\mathcal{T}} \to \mathbb{R}_+$ *be given by*

$$\psi_y(\hat{\mathbf{t}}) = \mathbf{1}(y \neq n)\, (\hat{t}_y - 1)^2 + \sum_{j \in [n-1], j \neq y} \hat{t}_j^{\,2} \,.$$

Then $\boldsymbol{\psi}$ *is classification calibrated with respect to* ℓ *over* Δ_n. *In particular, since* $\boldsymbol{\psi}$ *is convex,* $\text{CCdim}(\ell) \leq n - 1$.

It may appear surprising that the convex surrogate $\boldsymbol{\psi}$ in the above lemma is classification calibrated with respect to *all* multiclass losses ℓ on n classes. However this makes intuitive sense, since in principle, for any multiclass problem, if one can estimate the conditional probabilities of the n classes accurately (which requires estimating $n-1$ real-valued functions on $\mathcal{X}$), then one can predict a target label that minimizes the expected loss according to these probabilities. Minimizing the above surrogate effectively corresponds to such class probability estimation. Indeed, the above lemma can be shown to hold for any surrogate that is a strictly proper composite multiclass loss [21].

In practice, when the number of class labels n is large (such as in a sequence labeling task, where n is exponential in the length of the input sequence), the above result is not very helpful; in such cases, it is of interest to develop algorithms operating on a surrogate target space in a lower-dimensional space. Next we give a different upper bound on the CC dimension that depends on the loss ℓ, and for certain losses, can be significantly tighter than the general bound above.

Theorem 11. *Let* $\ell : [n] \times [k] \to \mathbb{R}_+$. *Then* $\text{CCdim}(\ell) \leq \text{rank}(\mathbf{L})$, *the rank of the loss matrix* $\mathbf{L}$.

Proof. Let $\text{rank}(\mathbf{L}) = d$. We will construct a convex classification calibrated surrogate loss ψ for ℓ with surrogate target space $\widehat{\mathcal{T}} \subseteq \mathbb{R}^d$.

Let $\boldsymbol{\ell}_{t_1}, \ldots, \boldsymbol{\ell}_{t_d}$ be linearly independent columns of $\mathbf{L}$. Let $\{\mathbf{e}_1, \ldots, \mathbf{e}_d\}$ denote the standard basis in $\mathbb{R}^d$. We can define a linear function $\tilde{\boldsymbol{\psi}} : \mathbb{R}^d \to \mathbb{R}^n$ by

$$\tilde{\boldsymbol{\psi}}(\mathbf{e}_j) = \boldsymbol{\ell}_{t_j} \ \ \forall j \in [d] \,.$$

Then for each $\mathbf{z}$ in the column space of $\mathbf{L}$, there exists a unique vector $\mathbf{u} \in \mathbb{R}^d$ such that $\tilde{\boldsymbol{\psi}}(\mathbf{u}) = \mathbf{z}$. In particular, there exist unique vectors $\mathbf{u}_1, \ldots, \mathbf{u}_k \in \mathbb{R}^d$ such that for each $t \in [k]$, $\tilde{\boldsymbol{\psi}}(\mathbf{u}_t) = \boldsymbol{\ell}_t$. Let $\widehat{\mathcal{T}} = \text{conv}(\{\mathbf{u}_1, \ldots, \mathbf{u}_k\})$, and define $\boldsymbol{\psi} : \widehat{\mathcal{T}} \to \mathbb{R}_+^n$ as

$$\boldsymbol{\psi}(\hat{\mathbf{t}}) = \tilde{\boldsymbol{\psi}}(\hat{\mathbf{t}}) \,;$$

we note that the resulting vectors are always in $\mathbb{R}_+^n$, since by definition, for any $\hat{\mathbf{t}} = \sum_{t=1}^k \alpha_t \mathbf{u}_t$ for $\boldsymbol{\alpha} \in \Delta_k$, $\boldsymbol{\psi}(\hat{\mathbf{t}}) = \sum_{t=1}^k \alpha_t \boldsymbol{\ell}_t$, and $\boldsymbol{\ell}_t \in \mathbb{R}_+^n \ \forall t \in [k]$. The function $\boldsymbol{\psi}$ is clearly convex. To show $\boldsymbol{\psi}$ is classification calibrated w.r.t. ℓ over Δ_n, we will use Theorem 7. Specifically, consider the k points $\mathbf{z}_t = \boldsymbol{\psi}(\mathbf{u}_t) = \boldsymbol{\ell}_t \in \mathcal{R}_\psi$ for $t \in [k]$. By definition of $\boldsymbol{\psi}$, we have $\mathcal{S}_\psi = \text{conv}(\{\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_k\})$; from the definitions of positive normals and trigger probabilities, it then follows that $\mathcal{N}_{\mathcal{S}_\psi}(\mathbf{z}_t) = \mathcal{N}_{\mathcal{S}_\psi}(\boldsymbol{\ell}_t) = \mathcal{Q}_t^\ell$ for all $t \in [k]$. Thus by Theorem 7, $\boldsymbol{\psi}$ is classification calibrated w.r.t. ℓ over Δ_n. □

Example 6 (CC dimension of Hamming loss)**.** *Consider the Hamming loss* ℓ^{Ham} *defined in Example 3, for* $n = 2^r$. *For each* $i \in [r]$, *define* $\boldsymbol{\sigma}_i \in \mathbb{R}^n$ *as*

$$\sigma_{iy} = \begin{cases} +1 & \text{if } (y-1)_i \text{, the } i\text{-th bit in the } r\text{-bit binary representation of } (y-1) \text{, is } 1 \\ -1 & \text{otherwise.} \end{cases}$$

Then the loss matrix $\mathbf{L}^{\text{Ham}}$ *satisfies*

$$\mathbf{L}^{\text{Ham}} = \frac{r}{2} \mathbf{e}\mathbf{e}^\top - \frac{1}{2} \sum_{i=1}^r \boldsymbol{\sigma}_i {\boldsymbol{\sigma}_i}^\top \,,$$

where $\mathbf{e}$ *is the* $n \times 1$ *all ones vector. Thus* $\text{rank}(\mathbf{L}^{\text{Ham}}) \leq r + 1$, *giving us* $\text{CCdim}(\ell^{\text{Ham}}) \leq r + 1$. *For* $r \geq 3$, *this is a significantly tighter upper bound than the bound of* $2^r - 1$ *given by Lemma 10.*

We note that the upper bound of Theorem 11 need not always be tight: for example, for the ordinal regression loss, for which we already know $\text{CCdim}(\ell^{\text{ord}}) = 1$, the theorem actually gives an upper bound of n, which is even weaker than that implied by Lemma 10.

4.2 Lower Bound on the Classification Calibration Dimension

In this section we give a lower bound on the CC dimension of a loss function ℓ and illustrate it by using it to calculate the CC dimension of the 0-1 loss. Section 5 we will explore consequences of the lower bound for classification calibrated surrogates for certain types of ranking losses. We will need the following definition:

Definition 12. *The* feasible subspace dimension *of a convex set $\mathcal{C}$ at $\mathbf{p} \in \mathcal{C}$, denoted by $\mu_{\mathcal{C}}(\mathbf{p})$, is defined as the dimension of the subspace $\mathcal{F}_{\mathcal{C}}(\mathbf{p}) \cap (-\mathcal{F}_{\mathcal{C}}(\mathbf{p}))$, where $\mathcal{F}_{\mathcal{C}}(\mathbf{p})$ is the cone of feasible directions of $\mathcal{C}$ at $\mathbf{p}$.*[7]

The following gives a lower bound on the CC dimension of a loss ℓ in terms of the feasible subspace dimension of the trigger probability sets $\mathcal{Q}_t^\ell$ at certain points $\mathbf{p} \in \mathcal{Q}_t^\ell$:

Theorem 13. *Let $\ell : [n] \times [k] \to \mathbb{R}_+$. Then for all $\mathbf{p} \in \text{relint}(\Delta_n)$ and $t \in \arg\min_{t'} \mathbf{p}^\top \ell_{t'}$ (i.e. such that $\mathbf{p} \in \mathcal{Q}_t^\ell$):*[8]

$$\text{CCdim}(\ell) \geq n - \mu_{\mathcal{Q}_t^\ell}(\mathbf{p}) - 1\,.$$

The proof requires extensions of the definition of positive normals and the necessary condition of Theorem 6 to sequences of points in $\mathcal{S}_\psi$ and is quite technical. In the appendix, we provide a proof in the special case when $\mathbf{p} \in \text{relint}(\Delta_n)$ is such that $\inf_{\mathbf{z}\in\mathcal{S}_\psi} \mathbf{p}^\top \mathbf{z}$ is achieved in $\mathcal{S}_\psi$, which does not require these extensions. Full proof details will be provided in a longer version of the paper. Both the proof of the lower bound and its applications make use of the following lemma, which gives a method to calculate the feasible subspace dimension for certain convex sets $\mathcal{C}$ and points $\mathbf{p} \in \mathcal{C}$:

Lemma 14. *Let $\mathcal{C} = \big\{\mathbf{u} \in \mathbb{R}^n : \mathbf{A}^1\mathbf{u} \leq \mathbf{b}^1, \mathbf{A}^2\mathbf{u} \leq \mathbf{b}^2, \mathbf{A}^3\mathbf{u} = \mathbf{b}^3\big\}$. Let $\mathbf{p} \in \mathcal{C}$ be such that $\mathbf{A}^1\mathbf{p} = \mathbf{b}^1$, $\mathbf{A}^2\mathbf{p} < \mathbf{b}^2$. Then $\mu_{\mathcal{C}}(\mathbf{p}) = \text{nullity}\left(\begin{bmatrix}\mathbf{A}^1\\ \mathbf{A}^3\end{bmatrix}\right)$, the dimension of the null space of $\begin{bmatrix}\mathbf{A}^1\\ \mathbf{A}^3\end{bmatrix}$.*

The above lower bound allows us to calculate precisely the CC dimension of the 0-1 loss:

Example 7 (CC dimension of 0-1 loss)**.** *Consider the 0-1 loss $\ell^{\text{0-1}}$ defined in Example 1. Take $\mathbf{p} = (\frac{1}{n}, \ldots, \frac{1}{n})^\top \in \text{relint}(\Delta_n)$. Then $\mathbf{p} \in \mathcal{Q}_t^{\text{0-1}}$ for all $t \in [k] = [n]$ (see Figure 2); in particular, we have $\mathbf{p} \in \mathcal{Q}_1^{\text{0-1}}$. Now $\mathcal{Q}_1^{\text{0-1}}$ can be written as*

$$\begin{aligned}\mathcal{Q}_1^{\text{0-1}} &= \big\{\mathbf{q} \in \Delta_n : q_1 \geq q_y \ \forall y \in \{2, \ldots, n\}\big\}\\ &= \big\{\mathbf{q} \in \mathbb{R}^n : \big[-\mathbf{e}_{n-1}\ \ \mathbf{I}_{n-1}\big]\mathbf{q} \leq \mathbf{0}, -\mathbf{q} \leq \mathbf{0}, \mathbf{e}_n^\top \mathbf{q} = 1\big\}\,,\end{aligned}$$

where $\mathbf{e}_{n-1}$, $\mathbf{e}_n$ denote the $(n-1)\times 1$ and $n \times 1$ all ones vectors, respectively, and $\mathbf{I}_{n-1}$ denotes the $(n-1)\times(n-1)$ identity matrix. Moreover, we have $\big[-\mathbf{e}_{n-1}\ \ \mathbf{I}_{n-1}\big]\mathbf{p} = 0$, $-\mathbf{p} < \mathbf{0}$. Therefore, by Lemma 14, we have

$$\mu_{\mathcal{Q}_1^{\text{0-1}}}(\mathbf{p}) = \text{nullity}\left(\begin{bmatrix}-\mathbf{e}_{n-1} & \mathbf{I}_{n-1}\\ & \mathbf{e}_n^\top\end{bmatrix}\right) = \text{nullity}\left(\begin{bmatrix}-1 & 1 & 0 & \ldots & 0\\ -1 & 0 & 1 & \ldots & 0\\ & & \vdots & & \\ -1 & 0 & 0 & \ldots & 1\\ 1 & 1 & 1 & \ldots & 1\end{bmatrix}\right) = 0\,.$$

Thus by Theorem 13, we get $\text{CCdim}(\ell^{\text{0-1}}) \geq n-1$. Combined with the upper bound of Lemma 10, this gives $\text{CCdim}(\ell^{\text{0-1}}) = n-1$.

[7] For a set $\mathcal{C} \subseteq \mathbb{R}^n$ and point $\mathbf{p} \in \mathcal{C}$, the cone of feasible directions of $\mathcal{C}$ at $\mathbf{p}$ is defined as $\mathcal{F}_{\mathbf{A}}(\mathbf{p}) = \{\mathbf{v} \in \mathbb{R}^n :\ \exists \epsilon_0 > 0 \text{ such that } \mathbf{p} + \epsilon\mathbf{v} \in \mathcal{C}\ \forall \epsilon \in (0, \epsilon_0)\}$.

[8] Here $\text{relint}(\Delta_n)$ denotes the relative interior of Δ_n: $\text{relint}(\Delta_n) = \{\mathbf{p} \in \Delta_n : p_y > 0\ \forall y \in [n]\}$.

5 Application to Pairwise Subset Ranking

We consider an application of the above framework to analyzing certain types of subset ranking problems, where each instance $x \in \mathcal{X}$ consists of a query together with a set of r documents (for simplicity, $r \in \mathbb{N}$ here is fixed), and the goal is to learn a predictor which given such an instance will return a ranking (permutation) of the r documents [8]. Duchi et al. [10] showed recently that for certain pairwise subset ranking losses ℓ, finding a predictor that minimizes the ℓ-risk is an NP-hard problem. They also showed that several common pairwise convex surrogate losses that operate on $\widehat{\mathcal{T}} = \mathbb{R}^r$ (and are used to learn scores for the r documents) fail to be classification calibrated with respect to such losses ℓ, even under some low-noise conditions on the distribution, and proposed an alternative convex surrogate, also operating on $\widehat{\mathcal{T}} = \mathbb{R}^r$, that is classification calibrated under certain conditions on the distribution (i.e. over a strict subset of the associated probability simplex).

Here we provide an alternative route to analyzing the difficulty of obtaining consistent surrogates for such pairwise subset ranking problems using the classification calibration dimension. Specifically, we will show that even for a simple setting of such problems, the classification calibration dimension of the underlying loss ℓ is greater than r, and therefore no convex surrogate operating on $\widehat{\mathcal{T}} \subseteq \mathbb{R}^r$ can be classification calibrated w.r.t. such a loss over the full probability simplex.

Formally, we will identify the set of class labels $\mathcal{Y}$ with a set $\mathcal{G}$ of 'preference graphs', which are simply directed acyclic graphs (DAGs) over r vertices; for each directed edge (i,j) in a preference graph $\mathbf{g} \in \mathcal{G}$ associated with an instance $x \in \mathcal{X}$, the i-th document in the document set in x is preferred over the j-th document. Here we will consider a simple setting where each preference graph has exactly one edge, so that $|\mathcal{Y}| = |\mathcal{G}| = r(r-1)$; in this setting, we can associate each $\mathbf{g} \in \mathcal{G}$ with the edge (i,j) it contains, which we will write as $\mathbf{g}_{(i,j)}$. The target labels consist of permutations over r objects, so that $\mathcal{T} = S_r$ with $|\mathcal{T}| = r!$. Consider now the following simple pairwise loss $\ell^{\text{pair}} : \mathcal{Y} \times \mathcal{T} \rightarrow \mathbb{R}_+$:

$$\ell^{\text{pair}}(\mathbf{g}_{(i,j)}, \sigma) = \mathbf{1}\big(\sigma(i) > \sigma(j)\big) . \tag{7}$$

Let $\mathbf{p} = (\frac{1}{r(r-1)}, \ldots, \frac{1}{r(r-1)})^\top \in \text{relint}(\Delta_{r(r-1)})$, and observe that $\mathbf{p}^\top \boldsymbol{\ell}^{\text{pair}}_\sigma = \frac{1}{2}$ for all $\sigma \in \mathcal{T}$. Thus $\mathbf{p}^\top(\boldsymbol{\ell}^{\text{pair}}_\sigma - \boldsymbol{\ell}^{\text{pair}}_{\sigma'}) = 0 \; \forall \sigma, \sigma' \in \mathcal{T}$, and so $\mathbf{p} \in \mathcal{Q}^{\text{pair}}_\sigma \; \forall \sigma \in \mathcal{T}$.

Let $(\sigma_1, \ldots, \sigma_{r!})$ be any fixed ordering of the permutations in $\mathcal{T}$, and consider $\mathcal{Q}^{\text{pair}}_{\sigma_1}$, defined by the intersection of $r! - 1$ half-spaces of the form $\mathbf{q}^\top(\boldsymbol{\ell}^{\text{pair}}_{\sigma_1} - \boldsymbol{\ell}^{\text{pair}}_{\sigma_t}) \leq 0$ for $t = 2, \ldots, r!$ and the simplex constraints $\mathbf{q} \in \Delta_{r(r-1)}$. Moreover, from the above observation, $\mathbf{p} \in \mathcal{Q}^{\text{pair}}_{\sigma_1}$ satisfies $\mathbf{p}^\top(\boldsymbol{\ell}^{\text{pair}}_{\sigma_1} - \boldsymbol{\ell}^{\text{pair}}_{\sigma_t}) = 0 \; \forall t = 2, \ldots, r!$. Therefore, by Lemma 14, we get

$$\mu_{\mathcal{Q}^{\text{pair}}_{\sigma_1}}(\mathbf{p}) = \text{nullity}\Big(\big[(\boldsymbol{\ell}^{\text{pair}}_{\sigma_1} - \boldsymbol{\ell}^{\text{pair}}_{\sigma_2}), \ldots, (\boldsymbol{\ell}^{\text{pair}}_{\sigma_1} - \boldsymbol{\ell}^{\text{pair}}_{\sigma_{r!}}), \mathbf{e}\big]^\top\Big) , \tag{8}$$

where $\mathbf{e}$ is the $r(r-1) \times 1$ all ones vector. It is not hard to see that the set $\{\boldsymbol{\ell}^{\text{pair}}_\sigma : \sigma \in \mathcal{T}\}$ spans a $\frac{r(r-1)}{2}$ dimensional space, and hence the nullity of the above matrix is at most $r(r-1) - \big(\frac{r(r-1)}{2} - 1\big)$. Thus by Theorem 13, we get that $\text{CCdim}(\ell^{\text{pair}}) \geq r(r-1) - \big(\frac{r(r-1)}{2} + 1\big) - 1 = \frac{r(r-1)}{2} - 2$. In particular, for $r \geq 5$, this gives $\text{CCdim}(\ell^{\text{pair}}) > r$, and therefore establishes that no convex surrogate ψ operating on a surrogate target space $\widehat{\mathcal{T}} \subseteq \mathbb{R}^r$ can be classification calibrated with respect to ℓ^{pair} on the full probability simplex $\Delta_{r(r-1)}$.

6 Conclusion

We developed a framework for analyzing consistency for general multiclass learning problems defined by a general loss matrix, introduced the notion of classification calibration dimension of a multiclass loss, and used this to analyze consistency properties of surrogate losses for various general multiclass problems. An interesting direction would be to develop a generic procedure for designing consistent convex surrogates operating on a 'minimal' surrogate target space according to the classification calibration dimension of the loss matrix. It would also be of interest to extend the results here to account for noise conditions as in [9, 10].

Acknowledgments

We would like to thank the anonymous reviewers for helpful comments. HG thanks Microsoft Research India for a travel grant. This research is supported in part by a Ramanujan Fellowship to SA from DST and an Indo-US Joint Center Award from the Indo-US Science & Technology Forum.

References

[1] Gábor Lugosi and Nicolas Vayatis. On the bayes-risk consistency of regularized boosting methods. *Annals of Statistics*, 32(1):30–55, 2004.

[2] Wenxin Jiang. Process consistency for AdaBoost. *Annals of Statistics*, 32(1):13–29, 2004.

[3] Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, 32(1):56–134, 2004.

[4] Ingo Steinwart. Consistency of support vector machines and other regularized kernel classifiers. *IEEE Transactions on Information Theory*, 51(1):128–142, 2005.

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

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

[7] Ambuj Tewari and Peter Bartlett. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8:1007–1025, 2007.

[8] David Cossock and Tong Zhang. Statistical analysis of bayes optimal subset ranking. *IEEE Transactions on Information Theory*, 54(11):5140–5154, 2008.

[9] Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. Listwise approach to learning to rank: Theory and algorithm. In *International Conference on Machine Learning*, 2008.

[10] John Duchi, Lester Mackey, and Michael Jordan. On the consistency of ranking algorithms. In *International Conference on Machine Learning*, 2010.

[11] Pradeep Ravikumar, Ambuj Tewari, and Eunho Yang. On NDCG consistency of listwise ranking methods. In *International Conference on Artificial Intelligence and Statistics(AISTATS)*, volume 15. JMLR: W&CP, 2011.

[12] David Buffoni, Clément Calauzènes, Patrick Gallinari, and Nicolas Usunier. Learning scoring functions with order-preserving losses and standardized supervision. In *International Conference on Machine Learning*, 2011.

[13] Wei Gao and Zhi-Hua Zhou. On the consistency of multi-label learning. In *Conference on Learning Theory*, 2011.

[14] Wojciech Kotlowski, Krzysztof Dembczynski, and Eyke Huellermeier. Bipartite ranking through minimization of univariate loss. In *International Conference on Machine Learning*, 2011.

[15] Ingo Steinwart. How to compare different loss functions and their risks. *Constructive Approximation*, 26:225–287, 2007.

[16] Ben Taskar, Carlos Guestrin, and Daphne Koller. Max-margin Markov networks. In *Neural Information Processing Systems*, 2003.

[17] Deirdre O'Brien, Maya Gupta, and Robert Gray. Cost-sensitive multi-class classification from probability estimates. In *International Conference on Machine Learning*, 2008.

[18] Nicolas Lambert and Yoav Shoham. Eliciting truthful answers to multiple-choice questions. In *ACM Conference on Electronic Commerce*, 2009.

[19] Dimitri Bertsekas, Angelia Nedic, and Asuman Ozdaglar. *Convex Analysis and Optimization*. Athena Scientific, 2003.

[20] Jean Gallier. Notes on convex sets, polytopes, polyhedra, combinatorial topology, Voronoi diagrams and Delaunay triangulations. Technical report, Department of Computer and Information Science, University of Pennsylvania, 2009.

[21] Elodie Vernet, Robert C. Williamson, and Mark D. Reid. Composite multiclass losses. In *Neural Information Processing Systems*, 2011.

Appendix: *Classification Calibration Dimension for General Multiclass Losses*

Calculation of Trigger Probability Sets for Figure 2

(a) 0-1 loss $\ell^{\text{0-1}}$ ($n = 3$).

$$\boldsymbol{\ell}_1 = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} ; \ \boldsymbol{\ell}_2 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} ; \ \boldsymbol{\ell}_3 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} .$$

$$\begin{aligned} \mathcal{Q}_1^{\text{0-1}} &= \{\mathbf{p} \in \Delta_3 : \mathbf{p}^\top \boldsymbol{\ell}_1 \le \mathbf{p}^\top \boldsymbol{\ell}_2, \ \mathbf{p}^\top \boldsymbol{\ell}_1 \le \mathbf{p}^\top \boldsymbol{\ell}_3\} \\ &= \{\mathbf{p} \in \Delta_3 : p_2 + p_3 \le p_1 + p_3, \ p_2 + p_3 \le p_1 + p_2\} \\ &= \{\mathbf{p} \in \Delta_3 : p_2 \le p_1, \ p_3 \le p_1\} \\ &= \{\mathbf{p} \in \Delta_3 : p_1 \ge \max(p_2, p_3)\} \end{aligned}$$

By symmetry,

$$\begin{aligned} \mathcal{Q}_2^{\text{0-1}} &= \{\mathbf{p} \in \Delta_3 : p_2 \ge \max(p_1, p_3)\} \\ \mathcal{Q}_3^{\text{0-1}} &= \{\mathbf{p} \in \Delta_3 : p_3 \ge \max(p_1, p_2)\} \end{aligned}$$

(b) Ordinal regression loss ℓ^{ord} ($n = 3$).

$$\boldsymbol{\ell}_1 = \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix} ; \ \boldsymbol{\ell}_2 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} ; \ \boldsymbol{\ell}_3 = \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix} .$$

$$\begin{aligned} \mathcal{Q}_1^{\text{ord}} &= \{\mathbf{p} \in \Delta_3 : \mathbf{p}^\top \boldsymbol{\ell}_1 \le \mathbf{p}^\top \boldsymbol{\ell}_2, \ \mathbf{p}^\top \boldsymbol{\ell}_1 \le \mathbf{p}^\top \boldsymbol{\ell}_3\} \\ &= \{\mathbf{p} \in \Delta_3 : p_2 + 2p_3 \le p_1 + p_3, \ p_2 + 2p_3 \le 2p_1 + p_2\} \\ &= \{\mathbf{p} \in \Delta_3 : p_2 + p_3 \le p_1, \ p_3 \le p_1\} \\ &= \{\mathbf{p} \in \Delta_3 : 1 - p_1 \le p_1\} \\ &= \{\mathbf{p} \in \Delta_3 : p_1 \ge \tfrac{1}{2}\} \end{aligned}$$

By symmetry,

$$\mathcal{Q}_3^{\text{ord}} = \{\mathbf{p} \in \Delta_3 : p_3 \ge \tfrac{1}{2}\}$$

Finally,

$$\begin{aligned} \mathcal{Q}_2^{\text{ord}} &= \{\mathbf{p} \in \Delta_3 : \mathbf{p}^\top \boldsymbol{\ell}_2 \le \mathbf{p}^\top \boldsymbol{\ell}_1, \ \mathbf{p}^\top \boldsymbol{\ell}_2 \le \mathbf{p}^\top \boldsymbol{\ell}_3\} \\ &= \{\mathbf{p} \in \Delta_3 : p_1 + p_3 \le p_2 + 2p_3, \ p_1 + p_3 \le 2p_1 + p_2\} \\ &= \{\mathbf{p} \in \Delta_3 : p_1 \le p_2 + p_3, \ p_3 \le p_1 + p_2\} \\ &= \{\mathbf{p} \in \Delta_3 : p_1 \le 1 - p_1, \ p_3 \le 1 - p_3\} \\ &= \{\mathbf{p} \in \Delta_3 : p_1 \le \tfrac{1}{2}, \ p_3 \le \tfrac{1}{2}\} \end{aligned}$$

(c) 'Abstain' loss $\ell^{(?)}$ ($n = 3$).

$$\boldsymbol{\ell}_1 = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} ; \ \boldsymbol{\ell}_2 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} ; \ \boldsymbol{\ell}_3 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} ; \ \boldsymbol{\ell}_4 = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \end{pmatrix} .$$

$$\begin{aligned} \mathcal{Q}_1^{(?)} &= \{\mathbf{p} \in \Delta_3 : \mathbf{p}^\top \boldsymbol{\ell}_1 \le \mathbf{p}^\top \boldsymbol{\ell}_2, \ \mathbf{p}^\top \boldsymbol{\ell}_1 \le \mathbf{p}^\top \boldsymbol{\ell}_3, \ \mathbf{p}^\top \boldsymbol{\ell}_1 \le \mathbf{p}^\top \boldsymbol{\ell}_4\} \\ &= \{\mathbf{p} \in \Delta_3 : p_2 + p_3 \le p_1 + p_3, \ p_2 + p_3 \le p_1 + p_2, \ p_2 + p_3 \le \tfrac{1}{2}(p_1 + p_2 + p_3)\} \\ &= \{\mathbf{p} \in \Delta_3 : p_2 \le p_1, \ p_3 \le p_1, \ p_2 + p_3 \le \tfrac{1}{2}\} \\ &= \{\mathbf{p} \in \Delta_3 : p_1 \ge \tfrac{1}{2}\} \end{aligned}$$

By symmetry,

$$\begin{aligned}\mathcal{Q}_2^{(?)} &= \{\mathbf{p} \in \Delta_3 : p_2 \geq \tfrac{1}{2}\} \\ \mathcal{Q}_3^{(?)} &= \{\mathbf{p} \in \Delta_3 : p_3 \geq \tfrac{1}{2}\}\end{aligned}$$

Finally,

$$\begin{aligned}\mathcal{Q}_4^{(?)} &= \{\mathbf{p} \in \Delta_3 : \mathbf{p}^\top \boldsymbol{\ell}_4 \leq \mathbf{p}^\top \boldsymbol{\ell}_1,\ \mathbf{p}^\top \boldsymbol{\ell}_4 \leq \mathbf{p}^\top \boldsymbol{\ell}_2,\ \mathbf{p}^\top \boldsymbol{\ell}_4 \leq \mathbf{p}^\top \boldsymbol{\ell}_2\} \\ &= \{\mathbf{p} \in \Delta_3 : \tfrac{1}{2}(p_1 + p_2 + p_3) \leq \min(p_2 + p_3, p_1 + p_3, p_1 + p_2)\} \\ &= \{\mathbf{p} \in \Delta_3 : \tfrac{1}{2} \leq 1 - \max(p_1, p_2, p_3)\} \\ &= \{\mathbf{p} \in \Delta_3 : \max(p_1, p_2, p_3) \leq \tfrac{1}{2}\}\end{aligned}$$

Proof of Theorem 6

Proof. Since ψ is classification calibrated w.r.t. ℓ over Δ_n, by Lemma 2, $\exists \text{pred}' : S_\psi \rightarrow [k]$ such that

$$\forall \mathbf{p} \in \Delta_n : \quad \inf_{\mathbf{z}' \in \mathcal{S}_\psi : \text{pred}'(\mathbf{z}') \notin \text{argmin}_t \mathbf{p}^\top \boldsymbol{\ell}_t} \mathbf{p}^\top \mathbf{z}' \ > \ \inf_{\mathbf{z}' \in \mathcal{S}_\psi} \mathbf{p}^\top \mathbf{z}' \,. \tag{9}$$

Now suppose there is some $\mathbf{z} \in \mathcal{S}_\psi$ such that $\mathcal{N}_{\mathcal{S}_\psi}(\mathbf{z})$ is not contained in $\mathcal{Q}_t^\ell$ for any $t \in [k]$. Then $\forall t \in [k]$, $\exists \mathbf{q} \in \mathcal{N}_{S_\psi}(\mathbf{z})$ such that $\mathbf{q} \notin \mathcal{Q}_t^\ell$, i.e. such that $t \notin \text{argmin}_{t'} \mathbf{q}^\top \boldsymbol{\ell}_{t'}$. In particular, for $t = \text{pred}'(\mathbf{z})$, $\exists \mathbf{q} \in \mathcal{N}_{S_\psi}(\mathbf{z})$ such that $\text{pred}'(\mathbf{z}) \notin \text{argmin}_{t'} \mathbf{q}^\top \boldsymbol{\ell}_{t'}$.

Since $\mathbf{q} \in \mathcal{N}_{\mathcal{S}_\psi}(\mathbf{z})$, we have

$$\mathbf{q}^\top \mathbf{z} \ = \ \inf_{\mathbf{z}' \in \mathcal{S}_\psi} \mathbf{q}^\top \mathbf{z}' \,. \tag{10}$$

Moreover, since $\text{pred}'(\mathbf{z}) \notin \text{argmin}_{t'} \mathbf{q}^\top \boldsymbol{\ell}_{t'}$, we have

$$\inf_{\mathbf{z}' \in \mathcal{S}_\psi : \text{pred}'(\mathbf{z}') \notin \text{argmin}_{t'} \mathbf{q}^\top \boldsymbol{\ell}_{t'}} \mathbf{q}^\top \mathbf{z}' \ \leq \ \mathbf{q}^\top \mathbf{z} \ = \ \inf_{\mathbf{z}' \in \mathcal{S}_\psi} \mathbf{q}^\top \mathbf{z}' \,. \tag{11}$$

This contradicts Eq. (9). Thus it must be the case that $\forall \mathbf{z} \in \mathcal{S}_\psi$, $\exists t \in [k]$ with $\mathcal{N}_{\mathcal{S}_\psi}(\mathbf{z}) \subseteq \mathcal{Q}_t^\ell$. ☐

Proof of Theorem 7

The proof uses the following technical lemma:

Lemma 15. *Let $\psi : [n] \times \widehat{\mathcal{T}} \rightarrow \mathbb{R}_+$. Suppose there exist $r \in \mathbb{N}$ and $\mathbf{z}_1, \ldots, \mathbf{z}_r \in \mathcal{R}_\psi$ such that $\bigcup_{j=1}^r \mathcal{N}_{\mathcal{S}_\psi}(\mathbf{z}_j) = \Delta_n$. Then any element $\mathbf{z} \in \mathcal{S}_\psi$ can be written as $\mathbf{z} = \mathbf{z}' + \mathbf{z}''$ for some $\mathbf{z}' \in \text{conv}(\{\mathbf{z}_1, \ldots, \mathbf{z}_r\})$ and $\mathbf{z}'' \in \mathbb{R}_+^n$.*

Proof. Let $\mathcal{S}' = \{\mathbf{z}' + \mathbf{z}'' : \mathbf{z}' \in \text{conv}(\{\mathbf{z}_1, \ldots, \mathbf{z}_r\}), \mathbf{z}'' \in \mathbb{R}_+^n\}$, and suppose there exists a point $\mathbf{z} \in \mathcal{S}_\psi$ which cannot be decomposed as claimed, i.e. such that $\mathbf{z} \notin \mathcal{S}'$. Then by the Hahn-Banach theorem (e.g. see [20], corollary 3.10), there exists a hyperplane that strictly separates $\mathbf{z}$ from $\mathcal{S}'$, i.e. $\exists \mathbf{w} \in \mathbb{R}^n$ such that $\mathbf{w}^\top \mathbf{z} < \mathbf{w}^\top \mathbf{a}\ \forall \mathbf{a} \in \mathcal{S}'$. It is easy to see that $\mathbf{w} \in \mathbb{R}_+^n$ (since a negative component in $\mathbf{w}$ would allow us to choose an element $\mathbf{a}$ from $\mathcal{S}'$ with arbitrarily small $\mathbf{w}^\top \mathbf{a}$).

Now consider the vector $\mathbf{q} = \mathbf{w} / \sum_{i=1}^n w_i \in \Delta_n$. Since $\bigcup_{j=1}^r \mathcal{N}_{\mathcal{S}_\psi}(\mathbf{z}_j) = \Delta_n$, $\exists j \in [r]$ such that $\mathbf{q} \in \mathcal{N}_{\mathcal{S}_\psi}(\mathbf{z}_j)$. By definition of positive normals, this gives $\mathbf{q}^\top \mathbf{z}_j \leq \mathbf{q}^\top \mathbf{z}$, and therefore $\mathbf{w}^\top \mathbf{z}_j \leq \mathbf{w}^\top \mathbf{z}$. But this contradicts our construction of $\mathbf{w}$ (since $\mathbf{z}_j \in \mathcal{S}'$). Thus it must be the case that every $\mathbf{z} \in \mathcal{S}_\psi$ is also an element of $\mathcal{S}'$. ☐

Proof. (Proof of Theorem 7)

We will show classification calibration of ψ w.r.t. ℓ (over Δ_n) via Lemma 2. For each $j \in [r]$, let

$$T_j = \left\{ t \in [k] : \mathcal{N}_{\mathcal{S}_\psi}(\mathbf{z}_j) \subseteq \mathcal{Q}_t^\ell \right\};$$

by assumption, $T_j \neq \emptyset \; \forall j \in [r]$. By Lemma 15, for every $\mathbf{z} \in \mathcal{S}_\psi$, $\exists \boldsymbol{\alpha} \in \Delta_r, \mathbf{u} \in \mathbb{R}^n_+$ such that $\mathbf{z} = \sum_{j=1}^r \alpha_j \mathbf{z}_j + \mathbf{u}$. For each $\mathbf{z} \in \mathcal{S}_\psi$, arbitrarily fix a unique $\boldsymbol{\alpha}^{\mathbf{z}} \in \Delta_r$ and $\mathbf{u}^{\mathbf{z}} \in \mathbb{R}^n_+$ satisfying the above, i.e. such that

$$\mathbf{z} = \sum_{j=1}^r \alpha_j^{\mathbf{z}} \mathbf{z}_j + \mathbf{u}^{\mathbf{z}} .$$

Now define $\text{pred}' : \mathcal{S}_\psi \to [k]$ as

$$\text{pred}'(\mathbf{z}) = \min \left\{ t \in [k] : \exists j \in [r] \text{ such that } \alpha_j^{\mathbf{z}} \geq \tfrac{1}{r} \text{ and } t \in T_j \right\} .$$

We will show pred′ satisfies the condition for classification calibration.

Fix any $\mathbf{p} \in \Delta_n$. Let

$$J_{\mathbf{p}} = \Big\{ j \in [r] : \mathbf{p} \in \mathcal{N}_{\mathcal{S}_\psi}(\mathbf{z}_j) \Big\} ;$$

since $\Delta_n = \bigcup_{j=1}^r \mathcal{N}_{\mathcal{S}_\psi}(\mathbf{z}_j)$, we have $J_{\mathbf{p}} \neq \emptyset$. Clearly,

$$\forall j \in J_{\mathbf{p}} : \mathbf{p}^\top \mathbf{z}_j = \inf_{\mathbf{z} \in \mathcal{S}_\psi} \mathbf{p}^\top \mathbf{z} \tag{12}$$

$$\forall j \notin J_{\mathbf{p}} : \mathbf{p}^\top \mathbf{z}_j > \inf_{\mathbf{z} \in \mathcal{S}_\psi} \mathbf{p}^\top \mathbf{z} \tag{13}$$

Moreover, from definition of T_j, we have

$$\forall j \in J_{\mathbf{p}} : \quad t \in T_j \implies \mathbf{p} \in \mathcal{Q}_t^\ell \implies t \in \text{argmin}_{t'} \mathbf{p}^\top \boldsymbol{\ell}_{t'} .$$

Thus we get

$$\forall j \in J_{\mathbf{p}} : \quad T_j \subseteq \text{argmin}_{t'} \mathbf{p}^\top \boldsymbol{\ell}_{t'} . \tag{14}$$

Now, for any $\mathbf{z} \in \mathcal{S}_\psi$ for which $\text{pred}'(\mathbf{z}) \notin \arg\min_{t'} \mathbf{p}^\top \boldsymbol{\ell}_{t'}$, we must have $\alpha_j^{\mathbf{z}} \geq \frac{1}{r}$ for at least one $j \notin J_{\mathbf{p}}$ (otherwise, we would have $\text{pred}'(\mathbf{z}) \in T_j$ for some $j \in J_{\mathbf{p}}$, giving $\text{pred}'(\mathbf{z}) \in \arg\min_{t'} \mathbf{p}^\top \boldsymbol{\ell}_{t'}$, a contradiction). Thus we have

$$\begin{aligned}
\inf_{\mathbf{z} \in \mathcal{S}_\psi : \text{pred}'(\mathbf{z}) \notin \text{argmin}_{t'} \mathbf{p}^\top \boldsymbol{\ell}_{t'}} \mathbf{p}^\top \mathbf{z} &= \inf_{\mathbf{z} \in \mathcal{S}_\psi : \text{pred}'(\mathbf{z}) \notin \text{argmin}_{t'} \mathbf{p}^\top \boldsymbol{\ell}_{t'}} \sum_{j=1}^r \alpha_j^{\mathbf{z}} \mathbf{p}^\top \mathbf{z}_j + \mathbf{p}^\top \mathbf{u}^{\mathbf{z}} && (15) \\
&\geq \inf_{\boldsymbol{\alpha} \in \Delta_r : \alpha_j \geq \frac{1}{r} \text{ for some } j \notin J_{\mathbf{p}}} \sum_{j=1}^r \alpha_j \mathbf{p}^\top \mathbf{z}_j && (16) \\
&\geq \min_{j \notin J_{\mathbf{p}}} \inf_{\alpha_j \in [\frac{1}{r}, 1]} \alpha_j \mathbf{p}^\top \mathbf{z}_j + (1 - \alpha_j) \inf_{\mathbf{z} \in \mathcal{S}_\psi} \mathbf{p}^\top \mathbf{z} && (17) \\
&> \inf_{\mathbf{z} \in \mathcal{S}_\psi} \mathbf{p}^\top \mathbf{z} , && (18)
\end{aligned}$$

where the last inequality follows from Eq. (13). Since the above holds for all $\mathbf{p} \in \Delta_n$, by Lemma 2, we have that ψ is classification calibrated w.r.t. ℓ over Δ_n. □

Proof of Lemma 8

Recall that a convex function $\phi : \mathbb{R}^d \to \bar{\mathbb{R}}$ (where $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$) attains its minimum at $\mathbf{u}_0 \in \mathbb{R}^d$ iff the subdifferential $\partial\phi(\mathbf{u}_0)$ contains $\mathbf{0} \in \mathbb{R}^d$ (e.g. see [19]). Also, if $\phi_1, \phi_2 : \mathbb{R}^d \to \bar{\mathbb{R}}$ are convex functions, then the subdifferential of their sum $\phi_1 + \phi_2$ at $\mathbf{u}_0$ is is equal to the Minkowski sum of the subdifferentials of ϕ_1 and ϕ_2 at $\mathbf{u}_0$:

$$\partial(\phi_1 + \phi_2)(\mathbf{u}_0) \;=\; \left\{ \mathbf{w}_1 + \mathbf{w}_2 : \mathbf{w}_1 \in \partial\phi_1(\mathbf{u}_0), \mathbf{w}_2 \in \partial\phi_2(\mathbf{u}_0) \right\} .$$

Proof. We have for all $\mathbf{p} \in \mathbb{R}^n$,

$$\begin{aligned}
\mathbf{p} \in \mathcal{N}_{\mathcal{S}_\psi}(\boldsymbol{\psi}(\hat{\mathbf{t}})) \quad &\iff \quad \mathbf{p} \in \Delta_n,\ \mathbf{p}^\top \boldsymbol{\psi}(\hat{\mathbf{t}}) \le \mathbf{p}^\top \mathbf{z}'\ \forall \mathbf{z}' \in \mathcal{S}_\psi \\
&\iff \quad \mathbf{p} \in \Delta_n,\ \mathbf{p}^\top \boldsymbol{\psi}(\hat{\mathbf{t}}) \le \mathbf{p}^\top \mathbf{z}'\ \forall \mathbf{z}' \in \mathcal{R}_\psi \\
&\iff \quad \mathbf{p} \in \Delta_n,\ \text{and the convex function } \phi(\hat{\mathbf{t}}') = \mathbf{p}^\top \boldsymbol{\psi}(\hat{\mathbf{t}}') = \textstyle\sum_{y=1}^n p_y \psi_y(\hat{\mathbf{t}}') \\
&\qquad\ \ \text{achieves its minimum at } \hat{\mathbf{t}}' = \hat{\mathbf{t}} \\
&\iff \quad \mathbf{p} \in \Delta_n,\ \mathbf{0} \in \sum_{y=1}^n p_y \partial \psi_y(\hat{\mathbf{t}}) \\
&\iff \quad \mathbf{p} \in \Delta_n,\ \mathbf{0} = \sum_{y=1}^n p_y \sum_{j=1}^{s_y} v_j^y \mathbf{w}_j^y \text{ for some } \mathbf{v}^y \in \Delta_{s_y} \\
&\iff \quad \mathbf{p} \in \Delta_n,\ \mathbf{0} = \sum_{y=1}^n \sum_{j=1}^{s_y} q_j^y \mathbf{w}_j^y \text{ for some } \mathbf{q}^y = p_y \mathbf{v}^y,\ \ \mathbf{v}^y \in \Delta_{s_y} \\
&\iff \quad \mathbf{p} \in \Delta_n, \mathbf{A}\mathbf{q} = \mathbf{0} \text{ for some } \mathbf{q} = (p_1 \mathbf{v}^1, \ldots, p_n \mathbf{v}^n)^\top \in \Delta_s,\ \ \mathbf{v}^y \in \Delta_{s_y} \\
&\iff \quad \mathbf{p} = \mathbf{B}\mathbf{q} \text{ for some } \mathbf{q} \in \text{Null}(\mathbf{A}) \cap \Delta_s\,.
\end{aligned}$$

□

Proof of Lemma 10

Proof. For each $\hat{\mathbf{t}} \in \widehat{\mathcal{T}}$, define $\mathbf{p}^{\hat{\mathbf{t}}} = \begin{pmatrix} \hat{\mathbf{t}} \\ 1 - \sum_{j=1}^{n-1} \hat{t}_j \end{pmatrix} \in \Delta_n$. Define pred : $\widehat{\mathcal{T}} \to [k]$ as

$$\text{pred}(\hat{\mathbf{t}}) \ = \ \min \left\{ t \in [k] : \mathbf{p}^{\hat{\mathbf{t}}} \in Q_t^\ell \right\}.$$

We will show that pred satisfies the condition of Definition 1.

Fix $\mathbf{p} \in \Delta_n$. It can be seen that

$$\mathbf{p}^\top \boldsymbol{\psi}(\hat{\mathbf{t}}) \ = \ \sum_{j=1}^{n-1} \Big(p_j (\hat{t}_j - 1)^2 + (1 - p_j)\, \hat{t}_j^{\,2} \Big).$$

Minimizing the above over $\hat{\mathbf{t}}$ yields the unique minimizer $\hat{\mathbf{t}}^* = (p_1, \ldots, p_{n-1})^\top \in \widehat{\mathcal{T}}$, which after some calculation gives

$$\inf_{\hat{\mathbf{t}} \in \widehat{\mathcal{T}}} \mathbf{p}^\top \boldsymbol{\psi}(\hat{\mathbf{t}}) \ = \ \mathbf{p}^\top \boldsymbol{\psi}(\hat{\mathbf{t}}^*) \ = \ \sum_{j=1}^{n-1} p_j (1 - p_j)\,.$$

Now, for each $t \in [k]$, define

$$\text{regret}_{\mathbf{p}}^\ell(t) \ \triangleq \ \mathbf{p}^\top \boldsymbol{\ell}_t - \min_{t' \in [k]} \mathbf{p}^\top \boldsymbol{\ell}_{t'}\,.$$

Clearly, $\text{regret}_{\mathbf{p}}^\ell(t) = 0 \iff \mathbf{p} \in \mathcal{Q}_t^\ell$. Note also that $\mathbf{p}^{\hat{\mathbf{t}}^*} = \mathbf{p}$, and therefore $\text{regret}_{\mathbf{p}}^\ell(\text{pred}(\hat{\mathbf{t}}^*)) = 0$. Let

$$\epsilon \ = \ \min_{t \in [k] : \mathbf{p} \notin \mathcal{Q}_t^\ell} \text{regret}_{\mathbf{p}}^\ell(t) \ > \ 0\,.$$

Then we have

$$\begin{aligned}
\inf_{\hat{\mathbf{t}} \in \widehat{\mathcal{T}}, \text{pred}(\hat{\mathbf{t}}) \notin \text{argmin}_t \mathbf{p}^\top \boldsymbol{\ell}_t} \mathbf{p}^\top \boldsymbol{\psi}(\hat{\mathbf{t}}) \ &= \ \inf_{\hat{\mathbf{t}} \in \widehat{\mathcal{T}}, \text{regret}_{\mathbf{p}}^\ell(\text{pred}(\hat{\mathbf{t}})) \ge \epsilon} \mathbf{p}^\top \boldsymbol{\psi}(\hat{\mathbf{t}}) \qquad (19) \\
&= \ \inf_{\hat{\mathbf{t}} \in \widehat{\mathcal{T}}, \text{regret}_{\mathbf{p}}^\ell(\text{pred}(\hat{\mathbf{t}})) \ge \text{regret}_{\mathbf{p}}^\ell(\text{pred}(\hat{\mathbf{t}}^*)) + \epsilon} \mathbf{p}^\top \boldsymbol{\psi}(\hat{\mathbf{t}})\,. \qquad (20)
\end{aligned}$$

Now, we claim that the mapping $\hat{\mathbf{t}} \mapsto \text{regret}_{\mathbf{p}}^\ell(\text{pred}(\hat{\mathbf{t}}))$ is continuous at $\hat{\mathbf{t}} = \hat{\mathbf{t}}^*$. To see this, suppose the sequence $\hat{\mathbf{t}}_m$ converges to $\hat{\mathbf{t}}^*$. Then it is easy to see that $\mathbf{p}^{\hat{\mathbf{t}}_m}$ converges to $\mathbf{p}^{\hat{\mathbf{t}}^*} = \mathbf{p}$, and therefore for each $t \in [k]$, $(\mathbf{p}^{\hat{\mathbf{t}}_m})^\top \boldsymbol{\ell}_t$ converges to $\mathbf{p}^\top \boldsymbol{\ell}_t$. Since by definition of pred we have that for all m, $\text{pred}(\hat{\mathbf{t}}_m) \in \text{argmin}_t (\mathbf{p}^{\hat{\mathbf{t}}_m})^\top \boldsymbol{\ell}_t$, this implies that for all large enough m, $\text{pred}(\hat{\mathbf{t}}_m) \in \text{argmin}_t \mathbf{p}^\top \boldsymbol{\ell}_t$. Thus for all large enough m, $\text{regret}^\ell_{\mathbf{p}}(\text{pred}(\hat{\mathbf{t}}_m)) = 0$; i.e. the sequence $\text{regret}^\ell_{\mathbf{p}}(\text{pred}(\hat{\mathbf{t}}_m))$ converges to $\text{regret}^\ell_{\mathbf{p}}(\text{pred}(\hat{\mathbf{t}}^*))$, yielding continuity at $\hat{\mathbf{t}}^*$. In particular, this implies $\exists \delta > 0$ such that

$$\|\hat{\mathbf{t}} - \hat{\mathbf{t}}^*\| < \delta \implies \text{regret}^\ell_{\mathbf{p}}(\text{pred}(\hat{\mathbf{t}})) - \text{regret}^\ell_{\mathbf{p}}(\text{pred}(\hat{\mathbf{t}}^*)) < \epsilon .$$

This gives

$$\inf_{\hat{\mathbf{t}} \in \widehat{\mathcal{T}}, \text{regret}^\ell_{\mathbf{p}}(\text{pred}(\hat{\mathbf{t}})) \geq \text{regret}^\ell_{\mathbf{p}}(\text{pred}(\hat{\mathbf{t}}^*)) + \epsilon} \mathbf{p}^\top \boldsymbol{\psi}(\hat{\mathbf{t}}) \;\geq\; \inf_{\hat{\mathbf{t}} \in \widehat{\mathcal{T}}, \|\hat{\mathbf{t}} - \hat{\mathbf{t}}^*\| \geq \delta} \mathbf{p}^\top \boldsymbol{\psi}(\hat{\mathbf{t}}) \tag{21}$$

$$> \inf_{\hat{\mathbf{t}} \in \widehat{\mathcal{T}}} \mathbf{p}^\top \boldsymbol{\psi}(\hat{\mathbf{t}}) , \tag{22}$$

where the last inequality holds since $\mathbf{p}^\top \boldsymbol{\psi}(\hat{\mathbf{t}})$ is a strictly convex function of $\hat{\mathbf{t}}$ and $\hat{\mathbf{t}}^*$ is its unique minimizer. The above sequence of inequalities give us that

$$\inf_{\hat{\mathbf{t}} \in \widehat{\mathcal{T}}, \text{pred}(\hat{\mathbf{t}}) \notin \text{argmin}_t \mathbf{p}^\top \boldsymbol{\ell}_t} \mathbf{p}^\top \boldsymbol{\psi}(\hat{\mathbf{t}}) \;>\; \inf_{\hat{\mathbf{t}} \in \widehat{\mathcal{T}}} \mathbf{p}^\top \boldsymbol{\psi}(\hat{\mathbf{t}}) . \tag{23}$$

Since this holds for all $\mathbf{p} \in \Delta_n$, we have that ψ is classification calibrated w.r.t. ℓ over Δ_n. □

Proof of Theorem 13

The proof uses the following lemma:

Lemma 16. *Let $\ell : [n] \times [k] \to \mathbb{R}^n_+$. Let $\mathbf{p} \in \text{relint}(\Delta_n)$. Then for any $t_1, t_2 \in \arg\min_{t'} \mathbf{p}^\top \boldsymbol{\ell}_{t'}$ (i.e. such that $\mathbf{p} \in \mathcal{Q}^\ell_{t_1} \cap \mathcal{Q}^\ell_{t_2}$),*

$$\mu_{\mathcal{Q}^\ell_{t_1}}(\mathbf{p}) = \mu_{\mathcal{Q}^\ell_{t_2}}(\mathbf{p}) .$$

Proof. Let $t_1, t_2 \in \arg\min_{t'} \mathbf{p}^\top \boldsymbol{\ell}_{t'}$ (i.e. $\mathbf{p} \in \mathcal{Q}^\ell_{t_1} \cap \mathcal{Q}^\ell_{t_2}$). Now

$$\mathcal{Q}^\ell_{t_1} = \left\{ \mathbf{q} \in \mathbb{R}^n : -\mathbf{q} \leq \mathbf{0}, \mathbf{e}^\top \mathbf{q} = 1, (\boldsymbol{\ell}_{t_1} - \boldsymbol{\ell}_t)^\top \mathbf{q} \leq 0 \; \forall t \in [k] \right\} .$$

Moreover, we have $-\mathbf{p} < \mathbf{0}$, and $(\boldsymbol{\ell}_{t_1} - \boldsymbol{\ell}_t)^\top \mathbf{p} = 0$ iff $\mathbf{p} \in \mathcal{Q}^\ell_t$. Let $\left\{ t \in [k] : \mathbf{p} \in \mathcal{Q}^\ell_t \right\} = \left\{ \tilde{t}_1, \ldots, \tilde{t}_r \right\}$ for some $r \in [k]$. Then by Lemma 14, we have

$$\mu_{Q^\ell_{t_1}} = \text{nullity}(\mathbf{A}_1) ,$$

where $\mathbf{A}_1 \in \mathbb{R}^{(r+1) \times n}$ is a matrix containing r rows of the form $(\boldsymbol{\ell}_{t_1} - \boldsymbol{\ell}_{\tilde{t}_j})^\top, j \in [r]$ and the all ones row. Similarly, we get

$$\mu_{Q^\ell_{t_2}} = \text{nullity}(\mathbf{A}_2) ,$$

where $\mathbf{A}_2 \in \mathbb{R}^{(r+1) \times n}$ is a matrix containing r rows of the form $(\boldsymbol{\ell}_{t_2} - \boldsymbol{\ell}_{\tilde{t}_j})^\top, j \in [r]$ and the all ones row. It can be seen that the subspaces spanned by the first r rows of $\mathbf{A}_1$ and $\mathbf{A}_2$ are both equal to the subspace parallel to the affine space containing $\boldsymbol{\ell}_{\tilde{t}_1}, \ldots, \boldsymbol{\ell}_{\tilde{t}_r}$. Thus both $\mathbf{A}_1$ and $\mathbf{A}_2$ have the same row space and hence the same null space and nullity, and therefore $\mu_{\mathcal{Q}^\ell_{t_1}}(\mathbf{p}) = \mu_{\mathcal{Q}^\ell_{t_2}}(\mathbf{p})$. □

Proof. (Proof of Theorem 13 for $\mathbf{p} \in \text{relint}(\Delta_n)$ such that $\inf_{\mathbf{z} \in \mathcal{S}_\psi} \mathbf{p}^\top \mathbf{z}$ is achieved in $\mathcal{S}_\psi$)

Let $d \in \mathbb{N}$ be such that there exists a convex surrogate target space $\widehat{\mathcal{T}} \subseteq \mathbb{R}^d$ and a convex surrogate loss $\boldsymbol{\psi} : \widehat{\mathcal{T}} \to \mathbb{R}^n_+$ that is classification calibrated with respect to ℓ over Δ_n. As noted previously, we can equivalently view $\boldsymbol{\psi}$ as being defined as $\boldsymbol{\psi} : \mathbb{R}^d \to \bar{\mathbb{R}}^n_+$, with $\psi_y(\hat{\mathbf{t}}) = \infty$ for $\hat{\mathbf{t}} \notin \widehat{\mathcal{T}}$ (and all $y \in [n]$). If $d \geq n-1$, we are done. Therefore in the following, we assume $d < n-1$.

Let $\mathbf{p} \in \text{relint}(\Delta_n)$. Note that $\inf_{\mathbf{z} \in \mathcal{S}_\psi} \mathbf{p}^\top \mathbf{z}$ always exists (since both $\mathbf{p}$ and ψ are non-negative). It can be shown that this infimum is attained in $\text{cl}(\mathcal{S}_\psi)$, i.e. $\exists \mathbf{z}^* \in \text{cl}(\mathcal{S}_\psi)$ such that $\inf_{\mathbf{z} \in \mathcal{S}_\psi} \mathbf{p}^\top \mathbf{z} = \mathbf{p}^\top \mathbf{z}^*$. In the following, we give a proof for the case when this infimum is attained within $\mathcal{S}_\psi$; the proof for the general case where the infimum is attained in $\text{cl}(\mathcal{S}_\psi)$ is similar but more technical, requiring extensions of the positive normals and the necessary condition ofTheorem 6 to sequences of points in $\mathcal{S}_\psi$ (complete details will be provided in a longer version of the paper).

For the rest of the proof, we assume $\mathbf{p}$ is such that the infimum $\inf_{\mathbf{z}\in\mathcal{S}_\psi} \mathbf{p}^\top \mathbf{z}$ is achieved in $\mathcal{S}_\psi$. In this case, it is easy to see that the infimum must then be achieved in $\mathcal{R}_\psi$ (e.g. see [19]). Thus $\exists \mathbf{z}^* = \boldsymbol{\psi}(\hat{\mathbf{t}}^*)$ for some $\hat{\mathbf{t}}^* \in \widehat{\mathcal{T}}$ such that $\inf_{\mathbf{z}\in\mathcal{S}_\psi} \mathbf{p}^\top \mathbf{z} = \mathbf{p}^\top \mathbf{z}^*$, and therefore $\mathbf{p} \in \mathcal{N}_{\mathcal{S}_\psi}(\mathbf{z}^*)$. This gives (e.g. see discussion before proof of Lemma 8)

$$\mathbf{0} \;\in\; \partial(\mathbf{p}^\top \boldsymbol{\psi}(\hat{\mathbf{t}}^*)) = \sum_{y=1}^n p_y \partial \psi_y(\hat{\mathbf{t}}^*)\,.$$

Thus for each $y \in [n]$, $\exists \mathbf{w}_y \in \partial\psi_y(\hat{\mathbf{t}}^*)$ such that $\sum_{y=1}^n p_y \mathbf{w}_y = \mathbf{0}$. Now let $\mathbf{A} = \big[\mathbf{w}_1 \dots \mathbf{w}_n\big] \in \mathbb{R}^{d\times n}$, and let

$$\mathcal{H} \;\;=\;\; \big\{\mathbf{q} \in \Delta_n : \mathbf{A}\mathbf{q} = \mathbf{0}\big\} \;=\; \big\{\mathbf{q} \in \mathbb{R}^n : \mathbf{A}\mathbf{q} = \mathbf{0}, \mathbf{e}^\top\mathbf{q} = 1, -\mathbf{q} \leq \mathbf{0}\big\}\,,$$

where $\mathbf{e}$ is the $n \times 1$ all ones vector. We have $\mathbf{p} \in \mathcal{H}$, and moreover, $-\mathbf{p} < \mathbf{0}$. Therefore, by Lemma 14, we have

$$\mu_{\mathcal{H}}(\mathbf{p}) \;=\; \text{nullity}\Big(\begin{bmatrix}\mathbf{A}\\ \mathbf{e}^\top\end{bmatrix}\Big) \;\geq\; n - (d+1)\,.$$

Now,

$$\mathbf{q} \in \mathcal{H} \implies \mathbf{A}\mathbf{q} = \mathbf{0} \implies \mathbf{0} \in \sum_{y=1}^n q_y \partial\psi_y(\hat{\mathbf{t}}^*) \implies \mathbf{q}^\top\mathbf{z}^* = \inf_{\mathbf{z}\in\mathcal{S}_\psi} \mathbf{q}^\top \mathbf{z} \implies \mathbf{q} \in \mathcal{N}_{\mathcal{S}_\psi}(\mathbf{z}^*)\,,$$

which gives $\mathcal{H} \subseteq \mathcal{N}_{\mathcal{S}_\psi}(\mathbf{z}^*)$. Moreover, by Theorem 6, we have that $\exists t_0 \in [k]$ such that $\mathcal{N}_{\mathcal{S}_\psi}(\mathbf{z}^*) \subseteq \mathcal{Q}^\ell_{t_0}$. This gives $\mathcal{H} \subseteq \mathcal{Q}^\ell_{t_0}$, and therefore

$$\mu_{\mathcal{Q}^\ell_{t_0}}(\mathbf{p}) \;\geq\; \mu_{\mathcal{H}}(\mathbf{p}) \;\geq\; n - d - 1\,.$$

By Lemma 16, we then have that for all t such that $\mathbf{p} \in \mathcal{Q}^\ell_t$,

$$\mu_{\mathcal{Q}^\ell_t}(\mathbf{p}) \;=\; \mu_{\mathcal{Q}^\ell_{t_0}}(\mathbf{p}) \;\geq\; n - d - 1\,,$$

which gives

$$d \;\geq\; n - \mu_{\mathcal{Q}^\ell_t}(\mathbf{p}) - 1\,.$$

This completes the proof for the case when $\inf_{\mathbf{z}\in\mathcal{S}_\psi} \mathbf{p}^\top\mathbf{z}$ is achieved in $\mathcal{S}_\psi$. As noted above, the proof for the case when this infimum is attained in $\text{cl}(\mathcal{S}_\psi)$ but not in $\mathcal{S}_\psi$ requires more technical details which will be provided in a longer version of the paper. □

Proof of Lemma 14

Proof. We will show that $\mathcal{F}_{\mathcal{C}}(\mathbf{p}) \cap (-\mathcal{F}_{\mathcal{C}}(\mathbf{p})) = \text{Null}\Big(\begin{bmatrix}\mathbf{A}^1\\ \mathbf{A}^3\end{bmatrix}\Big)$, from which the lemma follows.

First, let $\mathbf{v} \in \text{Null}\Big(\begin{bmatrix}\mathbf{A}^1\\ \mathbf{A}^3\end{bmatrix}\Big)$. Then for $\epsilon > 0$, we have

$$\begin{aligned}
\mathbf{A}^1(\mathbf{p}+\epsilon\mathbf{v}) \;&=\; \mathbf{A}^1\mathbf{p} + \epsilon\mathbf{A}^1\mathbf{v} \;=\; \mathbf{A}^1\mathbf{p} + \mathbf{0} \;=\; \mathbf{b}^1\\
\mathbf{A}^2(\mathbf{p}+\epsilon\mathbf{v}) \;&<\; \mathbf{b}^2 \;\;\text{for small enough } \epsilon\text{, since } \mathbf{A}^2\mathbf{p} < \mathbf{b}^2\\
\mathbf{A}^3(\mathbf{p}+\epsilon\mathbf{v}) \;&=\; \mathbf{A}^3\mathbf{p} + \epsilon\mathbf{A}^3\mathbf{v} \;=\; \mathbf{A}^3\mathbf{p} + \mathbf{0} \;=\; \mathbf{b}^3\,.
\end{aligned}$$

Thus $\mathbf{v} \in \mathcal{F}_{\mathcal{C}}(\mathbf{p})$. Similarly, we can show $-\mathbf{v} \in \mathcal{F}_{\mathcal{C}}(\mathbf{p})$. Thus $\mathbf{v} \in \mathcal{F}_{\mathcal{C}}(\mathbf{p}) \cap (-\mathcal{F}_{\mathcal{C}}(\mathbf{p}))$, giving $\text{Null}\Big(\begin{bmatrix}\mathbf{A}^1\\ \mathbf{A}^3\end{bmatrix}\Big) \subseteq \mathcal{F}_{\mathcal{C}}(\mathbf{p}) \cap (-\mathcal{F}_{\mathcal{C}}(\mathbf{p}))$.

Now let $\mathbf{v} \in \mathcal{F}_{\mathcal{C}}(\mathbf{p}) \cap (-\mathcal{F}_{\mathcal{C}}(\mathbf{p}))$. Then for small enough $\epsilon > 0$, we have both $\mathbf{A}^1(\mathbf{p}+\epsilon\mathbf{v}) \leq \mathbf{b}^1$ and $\mathbf{A}^1(\mathbf{p}-\epsilon\mathbf{v}) \leq \mathbf{b}^1$. Since $\mathbf{A}^1\mathbf{p} = \mathbf{b}^1$, this gives $\mathbf{A}^1\mathbf{v} = \mathbf{0}$. Similarly, for small enough $\epsilon > 0$, we have $\mathbf{A}^3(\mathbf{p}+\epsilon\mathbf{v}) = \mathbf{b}^3$; since $\mathbf{A}^3\mathbf{p} = \mathbf{b}^3$, this gives $\mathbf{A}^3\mathbf{v} = \mathbf{0}$. Thus $\begin{bmatrix}\mathbf{A}^1\\ \mathbf{A}^3\end{bmatrix}\mathbf{v} = \mathbf{0}$, giving $\mathcal{F}_{\mathcal{C}}(\mathbf{p}) \cap (-\mathcal{F}_{\mathcal{C}}(\mathbf{p})) \subseteq \text{Null}\Big(\begin{bmatrix}\mathbf{A}^1\\ \mathbf{A}^3\end{bmatrix}\Big)$. □