Convex Calibration Dimension for Multiclass Loss Matrices

Harish G. Ramaswamy   HARISH_GURUP@CSA.IISC.ERNET.IN
Shivani Agarwal   SHIVANI@CSA.IISC.ERNET.IN
Department of Computer Science and Automation
Indian Institute of Science
Bangalore 560012, India

Abstract

We study consistency properties of surrogate loss functions for general multiclass learning problems, defined by a general multiclass 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 calibrated with respect to a loss matrix in this setting. We then introduce the notion of *convex calibration dimension* of a multiclass loss matrix, which measures the smallest 'size' of a prediction space in which it is possible to design a convex surrogate that is 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, we apply our framework to study various subset ranking losses, and use the convex calibration dimension as a tool to show both the existence and non-existence of various types of convex calibrated surrogates for these losses. Our results strengthen recent results of Duchi et al. (2010) and Calauzènes et al. (2012) on the non-existence of certain types of convex calibrated surrogates in subset ranking. We anticipate the convex calibration dimension may prove to be a useful tool in the study and design of surrogate losses for general multiclass learning problems.

Keywords: Statistical consistency, multiclass loss, loss matrix, surrogate loss, convex surrogates, calibrated surrogates, classification calibration, subset ranking.

1. Introduction

There has been significant interest and progress in recent years in understanding consistency properties of surrogate risk minimization algorithms for various learning problems, such as binary classification, multiclass 0-1 classification, and various forms of ranking and multi-label prediction problems (Lugosi and Vayatis, 2004; Jiang, 2004; Zhang, 2004a; Steinwart, 2005; Bartlett et al., 2006; Zhang, 2004b; Tewari and Bartlett, 2007; Steinwart, 2007; Cossock and Zhang, 2008; Xia et al., 2008; Duchi et al., 2010; Ravikumar et al., 2011; Buffoni et al., 2011; Gao and Zhou, 2011; Kotlowski et al., 2011). Any such problem that involves a finite number of class labels and predictions can be viewed as an instance of a general multiclass learning problem, whose structure is defined by a suitable loss matrix. While the above studies have enabled an understanding of learning problems corresponding to cer-

tain forms of loss matrices, a framework for analyzing consistency properties for a general multiclass problem, defined by a general loss matrix, has remained elusive.

In this paper, we develop a unified framework for studying consistency properties of surrogate losses for such general multiclass learning problems, defined by a general multiclass loss matrix. For algorithms minimizing a surrogate loss, the question of consistency with respect to the target loss matrix reduces to the question of *calibration* of the surrogate loss with respect to the target loss.[1] We start by giving both necessary and sufficient conditions for a surrogate loss function to be calibrated with respect to any given target loss matrix. These conditions generalize previous conditions for the multiclass 0-1 loss studied for example by Tewari and Bartlett (2007). We then introduce the notion of *convex calibration dimension* of a loss matrix, a fundamental quantity that measures the smallest 'size' of a prediction space in which it is possible to design a convex surrogate that is calibrated with respect to the given loss matrix. This quantity can be viewed as representing one measure of the intrinsic 'difficulty' of the loss, and has a non-trivial behavior in the sense that one can give examples of loss matrices defined on the same number of class labels that have very different values of the convex calibration dimension, ranging from one (in which case one can achieve consistency by learning a single real-valued function) to practically the number of classes (in which case one must learn as many real-valued functions as the number of classes). We give upper and lower bounds on this quantity in terms of various algebraic and geometric properties of the loss matrix, and apply these results to analyze various loss matrices.

As concrete applications of our framework, we use the convex calibration dimension as a tool to study various loss matrices that arise in subset ranking problems, including the normalized discounted cumulative gain (NDCG), pairwise disagreement (PD), and mean average precision (MAP) losses. A popular practice in subset ranking, where one needs to rank a set of r documents by relevance to a query, has been to learn r real-valued scoring functions by minimizing a convex surrogate loss in r dimensions, and to then sort the r documents based on these scores. As discussed recently by Duchi et al. (2010) and Calauzènes et al. (2012), such an approach cannot be consistent for the PD and MAP losses, since these losses do not admit convex calibrated surrogates in r dimensions that can be used together with the sorting operation. We obtain a stronger result; in particular, we show that the convex calibration dimension of these losses is lower bounded by a quadratic function of r, which means that if minimizing a convex surrogate loss, one necessarily needs to learn $\Omega(r^2)$ real-valued functions to achieve consistency for these losses.

1.1 Related Work

There has been much work in recent years on consistency and calibration of surrogate losses for various learning problems. We give a brief overview of this body of work here.

Initial work on consistency of surrogate risk minimization algorithms focused largely on binary classification. For example, Steinwart (2005) showed the consistency of support vector machines with universal kernels for the problem of binary classification; Jiang (2004)

---

1. Assuming the surrogate risk minimization procedure is itself consistent (with respect to the surrogate loss); in most cases, this can be achieved by minimizing the surrogate risk over a function class that approaches a universal function class as the training sample size increases, e.g. see Bartlett et al. (2006).

and Lugosi and Vayatis (2004) showed similar results for boosting methods. Bartlett et al. (2006) and Zhang (2004a) studied the calibration of margin-based surrogates for binary classification. In particular, in their seminal work, Bartlett et al. (2006) established that the property of 'classification calibration' of a surrogate loss is equivalent to its minimization yielding 0-1 consistency, and gave a simple necessary and sufficient condition for convex margin-based surrogates to be calibrated w.r.t. the binary 0-1 loss. More recently, Reid and Williamson (2010) analyzed the calibration of a general family of surrogates termed proper composite surrogates for binary classification. Variants of standard 0-1 binary classification have also been studied; for example, Yuan and Wegkamp (2010) studied consistency for the problem of binary classification with a reject option, and Scott (2012) studied calibrated surrogates for cost-sensitive binary classification.

Over the years, there has been significant interest in extending the understanding of consistency and calibrated surrogates to various multiclass learning problems. Early work in this direction, pioneered by Zhang (2004b) and Tewari and Bartlett (2007), considered mainly the multiclass 0-1 classification problem. This work generalized the framework of Bartlett et al. (2006) to the multiclass 0-1 setting and used these results to study calibration of various surrogates proposed for multiclass 0-1 classification, such as the surrogates of Weston and Watkins (1999), Crammer and Singer (2001), and Lee et al. (2004). In particular, while the multiclass surrogate of Lee et al. (2004) was shown to calibrated for multiclass 0-1 classification, it was shown that several other widely used multiclass surrogates are in fact not calibrated for multiclass 0-1 classification.

More recently, there has been much work on studying consistency and calibration for various other learning problems that also involve finite label and prediction spaces. For example, Gao and Zhou (2011) studied consistency and calibration for multi-label prediction with the Hamming loss. Another prominent class of learning problems for which consistency and calibration have been studied recently is that of subset ranking, where instances contain queries together with sets of documents, and the goal is to learn a prediction model that given such an instance ranks the documents by relevance to the query. Various subset ranking losses have been investigated in recent years. Cossock and Zhang (2008) studied subset ranking with the discounted cumulative gain (DCG) ranking loss, and gave a simple surrogate calibrated w.r.t. this loss; Ravikumar et al. (2011) further studied subset ranking with the normalized DCG (NDCG) loss. Xia et al. (2008) considered the 0-1 loss applied to permutations. Duchi et al. (2010) focused on subset ranking with the pairwise disagreement (PD) loss, and showed that several popular convex score-based surrogates used for this problem are in fact not calibrated w.r.t. this loss; they also conjectured that such surrogates may not exist. Calauzènes et al. (2012) showed conclusively that there do not exist any convex score-based surrogates that are calibrated w.r.t. the PD loss, or w.r.t. the mean average precision (MAP) or expected reciprocal rank (ERR) losses. Finally, in a more general study of subset ranking losses, Buffoni et al. (2011) introduced the notion of 'standardization' for subset ranking losses, and gave a way to construct convex calibrated score-based surrogates for subset ranking losses that can be 'standardized'; they showed that while the DCG and NDCG losses can be standardized, the MAP and ERR losses cannot be standardized.

We also point out that in a related but different context, consistency of ranking has also been studied in the instance ranking setting (Clémençon and Vayatis, 2007; Clémençon et al., 2008; Kotlowski et al., 2011; Agarwal, 2014).

Finally, Steinwart (2007) considered consistency and calibration in a very general setting. More recently, Pires et al. (2013) used Steinwart's techniques to obtain surrogate regret bounds for certain surrogates w.r.t. general multiclass losses, and Ramaswamy et al. (2013) showed how to design explicit convex calibrated surrogates for any low-rank loss matrix.

1.2 Contributions of this Paper

As noted above, we develop a unified framework for studying consistency and calibration for general multiclass (finite-output) learning problems, described by a general loss matrix. We give both necessary conditions and sufficient conditions for a surrogate loss to be calibrated w.r.t. a given multiclass loss matrix, and introduce the notion of *convex calibration dimension* of a loss matrix, which measures the smallest 'size' of a prediction space in which it is possible to design a convex surrogate that is calibrated with respect to the loss matrix. We derive both upper and lower bounds on this quantity in terms of certain algebraic and geometric properties of the loss matrix, and apply these results to study various subset ranking losses. In particular, we obtain stronger results on the non-existence of convex calibrated surrogates for certain types of subset ranking losses than previous results in the literature (and also positive results on the existence of convex calibrated surrogates for these losses in higher dimensions). The following is a summary of the main differences from the conference version of this paper (Ramaswamy and Agarwal, 2012):

- Enhanced definition of positive normal sets of a surrogate loss at a sequence of points (Definition 5; this is required for proofs of stronger versions of our earlier results).

- Stronger necessary condition for calibration (Theorem 7).

- Stronger versions of upper and lower bounds on the convex calibration dimension, with full proofs (Theorems 12, 16).

- Conditions under which the upper and lower bounds are tight (Section 4.3).

- Application to a more general setting of the PD loss (Section 5.2).

- Additional applications to the NDCG and MAP losses (Sections 5.1, 5.3).

- Additional examples and illustrations throughout (Examples 5, 6, 7, 9; Figures 3, 4).

- Minor improvements and changes in emphasis in notation and terminology.

1.3 Organization

We start in Section 2 with some preliminaries and examples that will be used as running examples to illustrate concepts throughout the paper, and formalize the notion of calibration with respect to a general multiclass loss matrix. In Section 3, we derive both necessary conditions and sufficient conditions for calibration with respect to general loss matrices; these are both of independent interest and useful in our later results. Section 4 introduces the notion of convex calibration dimension of a loss matrix and derives both upper and lower bounds on this quantity. In Section 5, we apply our results to study the convex calibration dimension of various subset ranking losses. We conclude with a brief discussion in Section 6. Shorter proofs are included in the main text; all longer proofs are collected in Section 7 so as to maintain easy readability of the main text. The only exception to

this is proofs of Lemma 2 and Theorem 3, which closely follow earlier proofs of Tewari and Bartlett (2007) and are included for completeness in Appendix A. Some calculations are given in Appendices B and C.

2. Preliminaries, Examples, and Background

In this section we set up basic notation (Section 2.1), give background on multiclass loss matrices and risks (Section 2.2) and on multiclass surrogates and calibration (Section 2.3), and then define certain properties associated with multiclass losses and surrogates that will be useful in our study (Section 2.4).

2.1 Notation

Throughout the paper, we denote $\mathbb{R} = (-\infty, \infty)$, $\mathbb{R}_+ = [0, \infty)$, $\overline{\mathbb{R}} = [-\infty, \infty]$, $\overline{\mathbb{R}}_+ = [0, \infty]$. Similarly, $\mathbb{Z}$ and $\mathbb{Z}_+$ denote the sets of all integers and non-negative integers, respectively. For $n \in \mathbb{Z}_+$, we denote $[n] = \{1, \ldots, n\}$. For a predicate ϕ, we denote by $\mathbf{1}(\phi)$ the indicator of ϕ, which takes the value 1 if ϕ is true and 0 otherwise. For $z \in \mathbb{R}$, we denote $z_+ = \max(0, z)$. For a vector $\mathbf{v} \in \mathbb{R}^n$, we denote $\|\mathbf{v}\|_0 = \sum_{i=1}^n \mathbf{1}(v_i \neq 0)$ and $\|\mathbf{v}\|_1 = \sum_{i=1}^n |v_i|$. For a set $\mathcal{A} \subseteq \mathbb{R}^n$, we denote by $\mathrm{relint}(\mathcal{A})$ the relative interior of $\mathcal{A}$, by $\mathrm{cl}(\mathcal{A})$ the closure of $\mathcal{A}$, by $\mathrm{span}(\mathcal{A})$ the linear span (or linear hull) of $\mathcal{A}$, by $\mathrm{aff}(\mathcal{A})$ the affine hull of $\mathcal{A}$, and by $\mathrm{conv}(\mathcal{A})$ the convex hull of $\mathcal{A}$. For a vector space $\mathcal{V}$, we denote by $\dim(\mathcal{V})$ the dimension of $\mathcal{V}$. For a matrix $\mathbf{M} \in \mathbb{R}^{m \times n}$, we denote by $\mathrm{rank}(\mathbf{M})$ the rank of $\mathbf{M}$, by $\mathrm{affdim}(\mathbf{M})$ the affine dimension of the set of columns of $\mathbf{M}$ (i.e. the dimension of the subspace parallel to the affine hull of the columns of $\mathbf{M}$), by $\mathrm{null}(\mathbf{M})$ the null space of $\mathbf{M}$, and by $\mathrm{nullity}(\mathbf{M})$ the nullity of $\mathbf{M}$ (i.e. the dimension of the null space of $\mathbf{M}$). We denote by Δ_n the probability simplex in $\mathbb{R}^n$: $\Delta_n = \{\mathbf{p} \in \mathbb{R}_+^n : \sum_{i=1}^n p_i = 1\}$. Finally, we denote by Π_n the set of all permutations of $[n]$, i.e. the set of all bijective mappings $\sigma : [n] \to [n]$; for a permutation $\sigma \in \Pi_n$ and element $i \in [n]$, $\sigma(i)$ therefore represents the position of element i under σ.

2.2 Multiclass Losses and Risks

The general multiclass learning problem we consider can be described as follows: There is a finite set of *class labels* $\mathcal{Y}$ and a finite set of possible *predictions* $\widehat{\mathcal{Y}}$, which we take without loss of generality to be $\mathcal{Y} = [n]$ and $\widehat{\mathcal{Y}} = [k]$ for some $n, k \in \mathbb{Z}_+$. 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 the goal is to learn from these examples a prediction model $h : \mathcal{X} \to \widehat{\mathcal{Y}}$ which given a new instance $x \in \mathcal{X}$, makes a prediction $\widehat{y} = h(x) \in \widehat{\mathcal{Y}}$. In many common learning problems, the label and prediction spaces are the same, i.e. $\widehat{\mathcal{Y}} = \mathcal{Y}$, but in general, these could be different (e.g. when there is an 'abstain' option available to a classifier, in which case $k = n + 1$).

The performance of a prediction model is measured via a *loss function* $\ell : \mathcal{Y} \times \widehat{\mathcal{Y}} \to \mathbb{R}_+$, or equivalently, by a *loss matrix* $\mathbf{L} \in \mathbb{R}_+^{n \times k}$, with (y, t)-th element given by $\ell_{yt} = \ell(y, t)$; here $\ell_{yt} = \ell(y, t)$ defines the penalty incurred on predicting $t \in [k]$ when the true label is $y \in [n]$. We will use the notions of loss matrix and loss function interchangeably. Some examples of common multiclass loss functions and corresponding loss matrices are given below:

Example 1 (0-1 loss) *Here $\mathcal{Y} = \widehat{\mathcal{Y}} = [n]$, and the loss incurred is 1 if the predicted label t is different from the actual class label y, and 0 otherwise:*

$$\ell^{\text{0-1}}(y, t) \;=\; \mathbf{1}(t \neq y) \qquad \forall y, t \in [n].$$

The loss matrix $\mathbf{L}^{\text{0-1}}$ for $n = 3$ is shown in Figure 1(a). This is one of the most commonly used multiclass losses, and is suitable when all prediction errors are considered equal.

Example 2 (Ordinal regression loss) *Here $\mathcal{Y} = \widehat{\mathcal{Y}} = [n]$, and predictions t farther away from the actual class label y are penalized more heavily, e.g. using absolute distance:*

$$\ell^{\text{ord}}(y, t) \;=\; |t - y| \qquad \forall y, t \in [n].$$

The loss matrix $\mathbf{L}^{\text{ord}}$ for $n = 3$ is shown in Figure 1(b). This loss is often used when the class labels satisfy a natural ordinal property, for example in evaluating recommender systems that predict the number of stars (say out of 5) assigned to a product by user.

Example 3 (Hamming loss) *Here $\mathcal{Y} = \widehat{\mathcal{Y}} = [2^r]$ for some $r \in \mathbb{Z}_+$, 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:*

$$\ell^{\text{Ham}}(y, t) \;=\; \sum_{i=1}^{r} \mathbf{1}\big((t-1)_i \neq (y-1)_i\big) \qquad \forall y, t \in [2^r],$$

where for each $z \in \{0, \ldots, 2^r - 1\}$, $z_i \in \{0, 1\}$ 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 frequently used in sequence learning applications, where each element in $\mathcal{Y} = \widehat{\mathcal{Y}}$ is a binary sequence of length r, and the loss in predicting a sequence $\mathbf{t} \in \{0,1\}^r$ when the true label sequence is $\mathbf{y} \in \{0,1\}^r$ is simply the Hamming distance between the two sequences.

Example 4 ('Abstain' loss) *Here $\mathcal{Y} = [n]$ and $\widehat{\mathcal{Y}} = [n+1]$, where $t = n+1$ denotes a prediction of 'abstain' (or 'reject'). 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:*

$$\ell^{(?)}(y, t) \;=\; \mathbf{1}\big(t \in [n]\big) \cdot \mathbf{1}\big(t \neq y\big) + \frac{1}{2} \cdot \mathbf{1}\big(t = n+1\big) \qquad \forall y \in [n], t \in [n+1].$$

The loss matrix $\mathbf{L}^{(?)}$ for $n = 3$ is shown in Figure 1(d). This type of loss is suitable in applications where making an erroneous prediction is more costly than simply abstaining. For example, in medical diagnosis applications, when uncertain about the correct prediction, it may be better to abstain and request human intervention rather than make a misdiagnosis.

As noted above, given examples $(X_1, Y_1), \ldots, (X_m, Y_m)$ drawn i.i.d. from a distribution D on $\mathcal{X} \times [n]$, the goal is to learn a prediction model $h : \mathcal{X} \to [k]$. More specifically, given a target loss matrix $\mathbf{L} \in \mathbb{R}_+^{n \times k}$ with (y, t)-th element ℓ_{yt}, the goal is to learn a model $h : \mathcal{X} \to [k]$ with small expected loss on a new example drawn randomly from D, which we will refer to as the $\mathbf{L}$-*risk* or $\mathbf{L}$-*error* of h:

$$\text{er}_D^{\mathbf{L}}[h] \;\triangleq\; \mathbf{E}_{(X,Y) \sim D}\big[\ell_{Y, h(X)}\big]. \tag{1}$$

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

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

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

Clearly, denoting the t-th column of $\mathbf{L}$ as $\boldsymbol{\ell}_t = (\ell_{1t}, \ldots, \ell_{nt})^\top \in \mathbb{R}^n_+$, and the class probability vector at an instance $x \in \mathcal{X}$ under D as $\mathbf{p}(x) = (p_1(x), \ldots, p_n(x))^\top \in \Delta_n$, where $p_y(x) = \mathbf{P}(Y = y \mid X = x)$ under D, the $\mathbf{L}$-risk of h can be written as

$$
\mathrm{er}_D^{\mathbf{L}}[h] \;=\; \mathbf{E}_X\left[\sum_{y=1}^n p_y(X)\,\ell_{y,h(X)}\right] \;=\; \mathbf{E}_X\left[\mathbf{p}(X)^\top \boldsymbol{\ell}_{h(X)}\right]. \tag{2}
$$

The *optimal* $\mathbf{L}$-*risk* or *optimal* $\mathbf{L}$-*error* for a distribution D is then simply the smallest $\mathbf{L}$-risk or $\mathbf{L}$-error that can be achieved by any model h:

$$
\mathrm{er}_D^{\mathbf{L},*} \;\triangleq\; \inf_{h:\mathcal{X}\to[k]} \mathrm{er}_D^{\mathbf{L}}[h] \;=\; \inf_{h:\mathcal{X}\to[k]} \mathbf{E}_X\left[\mathbf{p}(X)^\top \boldsymbol{\ell}_{h(X)}\right] \;=\; \mathbf{E}_X\left[\min_{t\in[k]} \mathbf{p}(X)^\top \boldsymbol{\ell}_t\right]. \tag{3}
$$

Ideally, one would like to minimize (approximately) the $\mathbf{L}$-risk, e.g. by selecting a model that minimizes the average $\mathbf{L}$-loss on the training examples among some suitable class of models. However, minimizing the discrete $\mathbf{L}$-risk directly is typically computationally difficult. Consequently, one usually minimizes a (convex) *surrogate* risk instead.

2.3 Multiclass Surrogates and Calibration

Let $d \in \mathbb{Z}_+$, and let $\mathcal{C} \subseteq \mathbb{R}^d$ be a convex set. A *surrogate loss function* $\psi : \mathcal{Y} \times \mathcal{C} \to \mathbb{R}_+$ acting on the *surrogate prediction space* $\mathcal{C}$ assigns a penalty $\psi(y, \mathbf{u})$ on making a surrogate prediction $\mathbf{u} \in \mathcal{C}$ when the true label is $y \in [n]$. The ψ-*risk* or ψ-*error* of a surrogate prediction model $\mathbf{f} : \mathcal{X} \to \mathcal{C}$ w.r.t. a distribution D on $\mathcal{X} \times [n]$ is then defined as

$$
\mathrm{er}_D^{\psi}[\mathbf{f}] \;\triangleq\; \mathbf{E}_{(X,Y)\sim D}\left[\psi\big(Y, \mathbf{f}(X)\big)\right]. \tag{4}
$$

The surrogate ψ can be represented via n real-valued functions $\psi_y : \mathcal{C} \to \mathbb{R}_+$ for $y \in [n]$, defined as $\psi_y(\mathbf{u}) = \psi(y, \mathbf{u})$; equivalently, we can also represent the surrogate ψ as a vector-valued function $\boldsymbol{\psi} : \mathcal{C} \to \mathbb{R}^n_+$, defined as $\boldsymbol{\psi}(\mathbf{u}) = \big(\psi_1(\mathbf{u}), \ldots, \psi_n(\mathbf{u})\big)^\top$. Clearly, the ψ-risk of $\mathbf{f}$ can then be written as

$$
\mathrm{er}_D^{\psi}[\mathbf{f}] \;=\; \mathbf{E}_X\left[\sum_{y=1}^n p_y(X)\,\psi_y(\mathbf{f}(X))\right] \;=\; \mathbf{E}_X\left[\mathbf{p}(X)^\top \boldsymbol{\psi}(\mathbf{f}(X))\right]. \tag{5}
$$

The *optimal* ψ-*risk* or *optimal* ψ-*error* for a distribution D is then simply the smallest ψ-risk or ψ-error that can be achieved by any model $\mathbf{f}$:

$$
\mathrm{er}_D^{\psi,*} \;\triangleq\; \inf_{\mathbf{f}:\mathcal{X}\to\mathcal{C}} \mathrm{er}_D^{\psi}[\mathbf{f}] \;=\; \inf_{\mathbf{f}:\mathcal{X}\to\mathcal{C}} \mathbf{E}_X\left[\mathbf{p}(X)^\top \boldsymbol{\psi}(\mathbf{f}(X))\right] \;=\; \mathbf{E}_X\left[\inf_{\mathbf{u}\in\mathcal{C}} \mathbf{p}(X)^\top \boldsymbol{\psi}(\mathbf{u})\right]. \tag{6}
$$

We will find it convenient to define the sets

$$\mathcal{R}_\psi \;\overset{\triangle}{=}\; \boldsymbol{\psi}(\mathcal{C}) \;\subseteq\; \mathbb{R}^n_+ \tag{7}$$

$$\mathcal{S}_\psi \;\overset{\triangle}{=}\; \mathrm{conv}(\mathcal{R}_\psi) \;\subseteq\; \mathbb{R}^n_+ \,. \tag{8}$$

Clearly, the optimal ψ-risk can then also be written as

$$\mathrm{er}_D^{\psi,*} \;=\; \mathbf{E}_X\left[\inf_{\mathbf{z}\in\mathcal{R}_\psi} \mathbf{p}(X)^\top \mathbf{z} \right] \;=\; \mathbf{E}_X\left[\inf_{\mathbf{z}\in\mathcal{S}_\psi} \mathbf{p}(X)^\top \mathbf{z} \right] . \tag{9}$$

Example 5 (Crammer-Singer surrogate) *The Crammer-Singer surrogate was proposed as a hinge-like surrogate loss for 0-1 multiclass classification (Crammer and Singer, 2001). For $\mathcal{Y} = [n]$, the Crammer-Singer surrogate $\boldsymbol{\psi}^{\mathrm{CS}}$ acts on the surrogate prediction space $\mathcal{C} = \mathbb{R}^n$ and is defined as follows:*

$$\psi_y^{\mathrm{CS}}(\mathbf{u}) \;=\; \max_{y'\in[n],y'\neq y} \big(1 - (u_y - u_{y'})\big)_+ \qquad \forall y \in [n], \mathbf{u} \in \mathbb{R}^n \,.$$

A surrogate $\boldsymbol{\psi}$ is convex if ψ_y is convex $\forall y \in [n]$. As an example, the Crammer-Singer surrogate defined above is clearly convex. Given training examples $(X_1, Y_1), \ldots, (X_m, Y_m)$ drawn i.i.d. from a distribution D on $\mathcal{X} \times [n]$, a (convex) surrogate risk minimization algorithm using a (convex) surrogate loss $\boldsymbol{\psi} : \mathcal{C} \to \mathbb{R}^n_+$ learns a surrogate prediction model by minimizing (approximately, based on the training sample) the ψ-risk; the learned model $\mathbf{f} : \mathcal{X} \to \mathcal{C}$ is then used to make predictions in the original space $[k]$ via some transformation $\mathrm{pred} : \mathcal{C} \to [k]$. This yields a prediction model $h : \mathcal{X} \to [k]$ for the original multiclass problem given by $h = (\mathrm{pred} \circ \mathbf{f})$: the prediction on a new instance $x \in \mathcal{X}$ is given by $\mathrm{pred}(\mathbf{f}(x))$, and the $\mathbf{L}$-risk incurred is $\mathrm{er}_D^{\mathbf{L}}[\mathrm{pred} \circ \mathbf{f}]$. As an example, several surrogate risk minimizing algorithms for multiclass classification with respect to 0-1 loss (including that based on the Crammer-Singer surrogate) use a surrogate space $\mathcal{C} = \mathbb{R}^n$, learn a function of the form $\mathbf{f} : \mathcal{X} \to \mathbb{R}^n$, and predict according to $\mathrm{pred}(\mathbf{f}(x)) = \mathrm{argmax}_{t\in[n]} f_t(x)$.

Under suitable conditions, surrogate risk minimization 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 as the number of training examples m increases. This raises the natural question of whether, for a given loss matrix $\mathbf{L}$, there are surrogate losses ψ for which consistency with respect to the ψ-risk also guarantees consistency with respect to the $\mathbf{L}$-risk, i.e. guarantees convergence (in probability) to the optimal $\mathbf{L}$-risk (defined in Eq. (3)). As we shall see below, this amounts to the question of *calibration* of surrogate losses ψ w.r.t. a given target loss matrix $\mathbf{L}$, and 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 (Zhang, 2004b; Tewari and Bartlett, 2007). In this paper, we consider this question for general multiclass loss matrices $\mathbf{L} \in \mathbb{R}^{n\times k}_+$, including rectangular loss matrices with $k \neq n$. The only assumption we make on $\mathbf{L}$ is that for each $t \in [k]$, $\exists \mathbf{p} \in \Delta_n$ such that $\mathrm{argmin}_{t'\in[k]} \mathbf{p}^\top \boldsymbol{\ell}_{t'} = \{t\}$ (otherwise the element $t \in [k]$ never needs to be predicted and can simply be ignored).

We will need the following definitions and basic results, generalizing those of Zhang (2004b), Bartlett et al. (2006), and Tewari and Bartlett (2007). The notion of calibration will be central to our study; as Theorem 3 below shows, calibration of a surrogate loss ψ

w.r.t. $\mathbf{L}$ corresponds to the property that consistency w.r.t. ψ-risk implies consistency w.r.t. $\mathbf{L}$-risk. Proofs of Lemma 2 and Theorem 3 can be found in Appendix A.

Definition 1 ($(\mathbf{L}, \mathcal{P})$-calibration) *Let $\mathbf{L} \in \mathbb{R}_+^{n \times k}$ and $\mathcal{P} \subseteq \Delta_n$. A surrogate loss function $\psi : \mathcal{C} \to \mathbb{R}_+^n$ is said to be $(\mathbf{L}, \mathcal{P})$-calibrated if there exists a function $\mathrm{pred} : \mathcal{C} \to [k]$ such that*

$$\forall \mathbf{p} \in \mathcal{P} : \quad \inf_{\mathbf{u} \in \mathcal{C} : \mathrm{pred}(\mathbf{u}) \notin \mathrm{argmin}_t \, \mathbf{p}^\top \boldsymbol{\ell}_t} \mathbf{p}^\top \psi(\mathbf{u}) \; > \; \inf_{\mathbf{u} \in \mathcal{C}} \mathbf{p}^\top \psi(\mathbf{u}) \,.$$

If ψ is $(\mathbf{L}, \Delta_n)$-calibrated, we simply say ψ is $\mathbf{L}$-calibrated.

The above definition of calibration clearly generalizes that used in the binary case. For example, in the case of binary 0-1 classification, with $n = k = 2$ and $\mathbf{L}^{0\text{-}1} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, the probability simplex Δ_2 is equivalent to the interval $[0, 1]$, and it can be shown that one only need consider surrogates on the real line, $\mathcal{C} = \mathbb{R}$, and the predictor 'sign'; in this case one recovers the familiar definition of binary classification calibration of Bartlett et al. (2006), namely that a surrogate $\psi : \mathbb{R} \to \mathbb{R}_+^2$ is $(\mathbf{L}^{0\text{-}1}, \Delta_2)$-calibrated if

$$\forall p \in [0, 1], p \neq \frac{1}{2} : \quad \inf_{u \in \mathbb{R} : \mathrm{sign}(u) \neq \mathrm{sign}(p - \frac{1}{2})} p \, \psi_1(u) + (1 - p)\psi_2(u) \; > \; \inf_{u \in \mathbb{R}} p \, \psi_1(u) + (1 - p)\psi_2(u) \,.$$

Similarly, in the case of binary cost-sensitive classification, with $n = k = 2$ and $\mathbf{L}^c = \begin{bmatrix} 0 & 1-c \\ c & 0 \end{bmatrix}$ where $c \in (0, 1)$ is the cost of a false positive and $(1 - c)$ that of a false negative, one recovers the corresponding definition of Scott (2012), namely that a surrogate $\psi : \mathbb{R} \to \mathbb{R}_+^2$ is $(\mathbf{L}^c, \Delta_2)$-calibrated if

$$\forall p \in [0, 1], p \neq c : \quad \inf_{u \in \mathbb{R} : \mathrm{sign}(u) \neq \mathrm{sign}(p - c)} p \, \psi_1(u) + (1 - p)\psi_2(u) \; > \; \inf_{u \in \mathbb{R}} p \, \psi_1(u) + (1 - p)\psi_2(u) \,.$$

The following lemma gives a characterization of calibration similar to that used by Tewari and Bartlett (2007):

Lemma 2 *Let $\mathbf{L} \in \mathbb{R}_+^{n \times k}$ and $\mathcal{P} \subseteq \Delta_n$. Then a surrogate loss $\psi : \mathcal{C} \to \mathbb{R}_+^n$ is $(\mathbf{L}, \mathcal{P})$-calibrated iff there exists a function $\mathrm{pred}' : \mathcal{S}_\psi \to [k]$ such that*

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

In this paper, we will mostly be concerned with $(\mathbf{L}, \Delta_n)$-calibration, which as noted above, we refer to as simply $\mathbf{L}$-calibration. The following result, whose proof is a straightforward generalization of that of a similar result for the 0-1 loss given by Tewari and Bartlett (2007), explains why $\mathbf{L}$-calibration is useful:

Theorem 3 *Let $\mathbf{L} \in \mathbb{R}_+^{n \times k}$. A surrogate loss $\psi : \mathcal{C} \to \mathbb{R}_+^n$ is $\mathbf{L}$-calibrated iff there exists a function $\mathrm{pred} : \mathcal{C} \to [k]$ such that for all distributions D on $\mathcal{X} \times [n]$ and all sequences of (vector) functions $\mathbf{f}_m : \mathcal{X} \to \mathcal{C}$,*

$$\mathrm{er}_D^\psi[\mathbf{f}_m] \longrightarrow \mathrm{er}_D^{\psi,*} \qquad implies \qquad \mathrm{er}_D^{\mathbf{L}}[\mathrm{pred} \circ \mathbf{f}_m] \longrightarrow \mathrm{er}_D^{\mathbf{L},*} \,.$$

In particular, Theorem 3 implies that a surrogate ψ is $\mathbf{L}$-calibrated if and only if $\exists$ a mapping $\mathrm{pred} : \mathcal{C} \to [k]$ such that any ψ-consistent algorithm learning models of the form $\mathbf{f}_m : \mathcal{X} \to \mathcal{C}$ (from i.i.d. examples $(X_1, Y_1), \ldots, (X_m, Y_m)$) yields an $\mathbf{L}$-consistent algorithm learning models of the form $(\mathrm{pred} \circ \mathbf{f}_m) : \mathcal{X} \to [k]$.
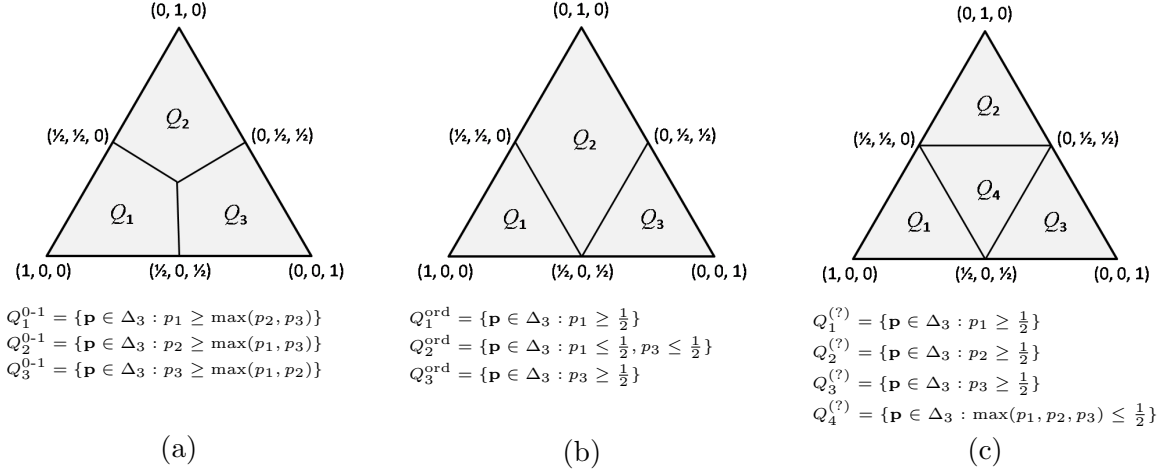
$Q_1^{0\text{-}1} = \{\mathbf{p} \in \Delta_3 : p_1 \geq \max(p_2, p_3)\}$
$Q_2^{0\text{-}1} = \{\mathbf{p} \in \Delta_3 : p_2 \geq \max(p_1, p_3)\}$
$Q_3^{0\text{-}1} = \{\mathbf{p} \in \Delta_3 : p_3 \geq \max(p_1, p_2)\}$

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

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

(a)   (b)   (c)

Figure 2: Trigger probability sets for (a) 0-1 loss $\mathbf{L}^{0\text{-}1}$; (b) ordinal regression loss $\mathbf{L}^{\mathrm{ord}}$; and (c) 'abstain' loss $\mathbf{L}^{(?)}$; all for $n = 3$, for which the probability simplex can be visualized easily. Calculations of these sets can be found in Appendix B.

2.4 Trigger Probabilities and Positive Normals

Our goal is to study conditions under which a surrogate loss $\boldsymbol{\psi} : \mathcal{C} \to \mathbb{R}_+^n$ is $\mathbf{L}$-calibrated for a target loss matrix $\mathbf{L} \in \mathbb{R}_+^{n \times k}$. To this end, we will now define certain properties of both multiclass loss matrices $\mathbf{L}$ and multiclass surrogates $\boldsymbol{\psi}$ that will be useful in relating the two. Specifically, we will define *trigger probability sets* associated with a multiclass loss matrix $\mathbf{L}$, and *positive normal sets* associated with a multiclass surrogate $\boldsymbol{\psi}$; in Section 3 we will use these to obtain both necessary and sufficient conditions for calibration.

Definition 4 (Trigger probability sets) *Let* $\mathbf{L} \in \mathbb{R}_+^{n \times k}$. *For each* $t \in [k]$, *the* trigger probability set *of* $\mathbf{L}$ *at* t *is defined as*

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

In words, the trigger probability set $\mathcal{Q}_t^{\mathbf{L}}$ is the set of class probability vectors for which predicting t is optimal in terms of minimizing $\mathbf{L}$-risk. Such sets have also been studied by Lambert and Shoham (2009) and O'Brien et al. (2008) in a different context. Lambert and Shoham (2009) show that these sets form what is called a power diagram, which is a generalization of the Voronoi diagram. Trigger probability sets for the 0-1, ordinal regression, and 'abstain' loss matrices (described in Examples 1, 2 and 4) are illustrated in Figure 2; the corresponding calculations can be found in Appendix B.

Definition 5 (Positive normal sets) *Let* $\boldsymbol{\psi} : \mathcal{C} \to \mathbb{R}_+^n$. *For each point* $\mathbf{z} \in \mathcal{S}_\psi$, *the* positive normal set *of* $\boldsymbol{\psi}$ *at* $\mathbf{z}$ *is defined as*[2]

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

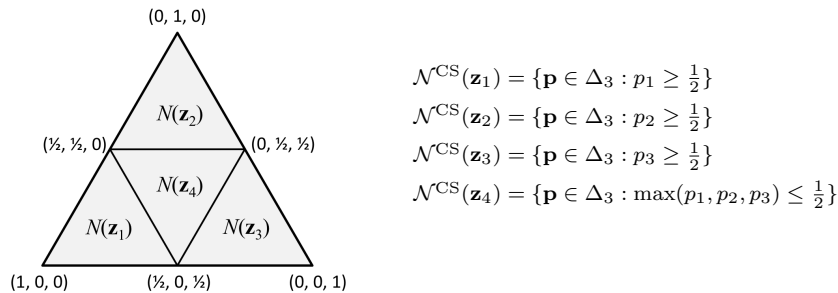2. For points $\mathbf{z}$ in the interior of $\mathcal{S}_\psi$, $\mathcal{N}^\psi(\mathbf{z})$ is empty.

10

Figure 3: Positive normal sets for the Crammer-Singer surrogate $\boldsymbol{\psi}^{\mathrm{CS}}$ for $n = 3$, at 4 points $\mathbf{z}_i = \boldsymbol{\psi}^{\mathrm{CS}}(\mathbf{u}_i) \in \mathbb{R}^3_+$ $(i \in [4])$ for $\mathbf{u}_1 = (1, 0, 0)^\top$, $\mathbf{u}_2 = (0, 1, 0)^\top$, $\mathbf{u}_3 = (0, 0, 1)^\top$, and $\mathbf{u}_4 = (0, 0, 0)^\top$. Calculations of these sets are based on Lemma 9 and can be found in Appendix C.

For any sequence of points $\{\mathbf{z}_m\}$ in $\mathcal{S}_\psi$, the positive normal set of ψ at $\{\mathbf{z}_m\}$ is defined as[3]

$$\mathcal{N}^\psi(\{\mathbf{z}_m\}) \;\triangleq\; \left\{ \mathbf{p} \in \Delta_n : \lim_{m \to \infty} \mathbf{p}^\top \mathbf{z}_m = \inf_{\mathbf{z}' \in \mathcal{S}_\psi} \mathbf{p}^\top \mathbf{z}' \right\}.$$

In words, the positive normal set $\mathcal{N}^\psi(\mathbf{z})$ at a point $\mathbf{z} = \boldsymbol{\psi}(\mathbf{u}) \in \mathcal{R}_\psi$ is the set of class probability vectors for which predicting $\mathbf{u}$ is optimal in terms of minimizing ψ-risk. Such sets were also studied by Tewari and Bartlett (2007). The extension to sequences of points in $\mathcal{S}_\psi$ is needed for technical reasons in some of our proofs. Note that for $\mathcal{N}^\psi(\{\mathbf{z}_m\})$ to be well-defined, the sequence $\{\mathbf{z}_m\}$ need not converge itself; however if the sequence $\{\mathbf{z}_m\}$ does converge to some point $\mathbf{z} \in \mathcal{S}_\psi$, then $\mathcal{N}^\psi(\{\mathbf{z}_m\}) = \mathcal{N}^\psi(\mathbf{z})$. Positive normal sets for the Crammer-Singer surrogate (described in Example 5) at 4 points are illustrated in Figure 3; the corresponding calculations can be found in Appendix C.

3. Conditions for Calibration

In this section we give both necessary conditions (Section 3.1) and sufficient conditions (Section 3.2) for a surrogate ψ to be calibrated w.r.t. an arbitrary target loss matrix $\mathbf{L}$. Both sets of conditions involve the trigger probability sets of $\mathbf{L}$ and the positive normal sets of ψ; in Section 3.3 we give a result that facilitates computation of positive normal sets for certain classes of surrogates ψ.

3.1 Necessary Conditions for Calibration

We start by deriving necessary conditions for $\mathbf{L}$-calibration of a surrogate loss ψ. Consider what happens if for some point $\mathbf{z} \in \mathcal{S}_\psi$, the positive normal set of ψ at $\mathbf{z}$, $\mathcal{N}^\psi(\mathbf{z})$, has a non-empty intersection with the interiors of two trigger probability sets of $\mathbf{L}$, say $\mathcal{Q}^\mathbf{L}_1$ and $\mathcal{Q}^\mathbf{L}_2$ (see Figure 4 for an illustration), which means $\exists \mathbf{q}_1, \mathbf{q}_2 \in \mathcal{N}^\psi(\mathbf{z})$ with $\operatorname{argmin}_{t \in [k]} \mathbf{q}_1^\top \boldsymbol{\ell}_t = \{1\}$ and $\operatorname{argmin}_{t \in [k]} \mathbf{q}_2^\top \boldsymbol{\ell}_t = \{2\}$. If ψ is $\mathbf{L}$-calibrated, then by Lemma 2, we have $\exists \mathrm{pred}' : \mathcal{S}_\psi \to [k]$

---

3. For sequences $\{\mathbf{z}_m\}$ for which $\lim_{m \to \infty} \mathbf{p}^\top \mathbf{z}_m$ does not exist for any $\mathbf{p}$, $\mathcal{N}^\psi(\mathbf{z})$ is empty.
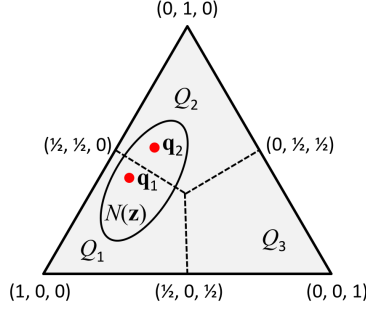
Figure 4: Visual proof of Theorem 6. If a surrogate ψ is such that its positive normal set $\mathcal{N}^\psi(\mathbf{z})$ at some point $\mathbf{z}$ has non-empty intersection with the interiors of two trigger probability sets (say $\mathcal{Q}_1^{\mathbf{L}}$ and $\mathcal{Q}_2^{\mathbf{L}}$) of $\mathbf{L}$, then ψ cannot be $\mathbf{L}$-calibrated.

such that

$$
\inf_{\mathbf{z}'\in\mathcal{S}_\psi:\text{pred}'(\mathbf{z}')\neq 1} \mathbf{q}_1^\top \mathbf{z}' \;=\; \inf_{\mathbf{z}'\in\mathcal{S}_\psi:\text{pred}'(\mathbf{z}')\notin\text{argmin}_t \mathbf{q}_1^\top \boldsymbol{\ell}_t} \mathbf{q}_1^\top \mathbf{z}' \;>\; \inf_{\mathbf{z}'\in\mathcal{S}_\psi} \mathbf{q}_1^\top \mathbf{z}' \;=\; \mathbf{q}_1^\top \mathbf{z}
$$

$$
\inf_{\mathbf{z}'\in\mathcal{S}_\psi:\text{pred}'(\mathbf{z}')\neq 2} \mathbf{q}_2^\top \mathbf{z}' \;=\; \inf_{\mathbf{z}'\in\mathcal{S}_\psi:\text{pred}'(\mathbf{z}')\notin\text{argmin}_t \mathbf{q}_2^\top \boldsymbol{\ell}_t} \mathbf{q}_2^\top \mathbf{z}' \;>\; \inf_{\mathbf{z}'\in\mathcal{S}_\psi} \mathbf{q}_2^\top \mathbf{z}' \;=\; \mathbf{q}_2^\top \mathbf{z} \,.
$$

The first inequality above implies $\text{pred}'(\mathbf{z}) = 1$; the second inequality implies $\text{pred}'(\mathbf{z}) = 2$, leading to a contradiction. This gives us the following necessary condition for $\mathbf{L}$-calibration of ψ, which requires the positive normal sets of ψ at all points $\mathbf{z} \in \mathcal{S}_\psi$ to be 'well-behaved' w.r.t. $\mathbf{L}$ in the sense of being contained within individual trigger probability sets of $\mathbf{L}$ and generalizes the 'admissibility' condition used for 0-1 loss by Tewari and Bartlett (2007):

Theorem 6 *Let $\mathbf{L} \in \mathbb{R}_+^{n\times k}$, and let $\psi : \mathcal{C}\to\mathbb{R}_+^n$ be $\mathbf{L}$-calibrated. Then for all points $\mathbf{z} \in \mathcal{S}_\psi$, there exists some $t \in [k]$ such that $\mathcal{N}^\psi(\mathbf{z}) \subseteq \mathcal{Q}_t^{\mathbf{L}}$.*

Proof See above discussion. ∎

In fact, we have the following stronger necessary condition, which requires the positive normal sets of ψ not only at all points $\mathbf{z} \in \mathcal{S}_\psi$ but also at all sequences $\{\mathbf{z}_m\}$ in $\mathcal{S}_\psi$ to be contained within individual trigger probability sets of $\mathbf{L}$:

Theorem 7 *Let $\mathbf{L} \in \mathbb{R}_+^{n\times k}$, and let $\psi : \mathcal{C}\to\mathbb{R}_+^n$ be $\mathbf{L}$-calibrated. Then for all sequences $\{\mathbf{z}_m\}$ in $\mathcal{S}_\psi$, there exists some $t \in [k]$ such that $\mathcal{N}^\psi(\{\mathbf{z}_m\}) \subseteq \mathcal{Q}_t^{\mathbf{L}}$.*

Proof Assume for the sake of contradiction that there is some sequence $\{\mathbf{z}_m\}$ in $\mathcal{S}_\psi$ for which $\mathcal{N}^\psi(\{\mathbf{z}_m\})$ is not contained in $\mathcal{Q}_t^{\mathbf{L}}$ for any $t \in [k]$. Then $\forall t \in [k]$, $\exists \mathbf{q}_t \in \mathcal{N}^\psi(\{\mathbf{z}_m\})$ such that $\mathbf{q}_t \notin \mathcal{Q}_t^{\mathbf{L}}$, i.e. such that $t \notin \text{argmin}_{t'} \mathbf{q}_t^\top \boldsymbol{\ell}_{t'}$. Now, since ψ is $\mathbf{L}$-calibrated, by Lemma 25, there exists a function $\text{pred}' : \mathcal{S}_\psi\to[k]$ such that for all $\mathbf{p} \in \mathcal{N}^\psi(\{\mathbf{z}_m\})$, we have $\text{pred}'(\mathbf{z}_m) \in \arg\min_{t'} \mathbf{p}^\top \boldsymbol{\ell}_{t'}$ for all large enough m. In particular, for $\mathbf{p} = \mathbf{q}_t$, we get $\text{pred}'(\mathbf{z}_m) \in \text{argmin}_{t'} \mathbf{q}_t^\top \boldsymbol{\ell}_{t'}$ ultimately. Since this is true for each $t \in [k]$, we get $\text{pred}'(\mathbf{z}_m) \in \cap_{t\in[k]} \text{argmin}_{t'} \mathbf{q}_t^\top \boldsymbol{\ell}_{t'}$ ultimately. However by choice of $\mathbf{q}_t$, this intersection

is empty, thus yielding a contradiction. This completes the proof. ∎

Note that Theorem 7 includes Theorem 6 as a special case, since $\mathcal{N}^\psi(\mathbf{z}) = \mathcal{N}^\psi(\{\mathbf{z}_m\})$ for the constant sequence $\mathbf{z}_m = \mathbf{z} \; \forall m$. We stated Theorem 6 separately above since it had a simple, direct proof that helps build intuition.

Example 6 (Crammer-Singer surrogate is not calibrated for 0-1 loss) *Looking at the positive normal sets of the Crammer-Singer surrogate $\boldsymbol{\psi}^{\mathrm{CS}}$ (for $n = 3$) shown in Figure 3 and the trigger probability sets of the 0-1 loss $\mathbf{L}^{\text{0-1}}$ shown in Figure 2(a), we see that $\mathcal{N}^{\mathrm{CS}}(\mathbf{z}_4)$ is not contained in any single trigger probability set of $\mathbf{L}^{\text{0-1}}$, and therefore applying Theorem 6, it is immediately clear that $\boldsymbol{\psi}^{\mathrm{CS}}$ is not $\mathbf{L}^{\text{0-1}}$-calibrated (this was also established by Tewari and Bartlett (2007) and Zhang (2004b)).*

3.2 Sufficient Condition for Calibration

We now give a sufficient condition for $\mathbf{L}$-calibration of a surrogate loss $\boldsymbol{\psi}$ that will be helpful in showing calibration of various surrogates. In particular, we show that for a surrogate loss $\boldsymbol{\psi}$ to be $\mathbf{L}$-calibrated, it is sufficient for the above property of positive normal sets of $\boldsymbol{\psi}$ being contained in trigger probability sets of $\mathbf{L}$ to hold for only a finite number of points in $\mathcal{S}_\psi$, as long as the corresponding positive normal sets jointly cover Δ_n:

Theorem 8 *Let $\mathbf{L} \in \mathbb{R}_+^{n \times k}$ and $\boldsymbol{\psi} : \mathcal{C} \to \mathbb{R}_+^n$. Suppose there exist $r \in \mathbb{Z}_+$ and $\mathbf{z}_1, \ldots, \mathbf{z}_r \in \mathcal{S}_\psi$ such that $\bigcup_{j=1}^r \mathcal{N}^\psi(\mathbf{z}_j) = \Delta_n$ and for each $j \in [r]$, $\exists t \in [k]$ such that $\mathcal{N}^\psi(\mathbf{z}_j) \subseteq \mathcal{Q}_t^{\mathbf{L}}$. Then $\boldsymbol{\psi}$ is $\mathbf{L}$-calibrated.*

Example 7 (Crammer-Singer surrogate is calibrated for $\mathbf{L}^{(?)}$ and $\mathbf{L}^{\mathrm{ord}}$ for $n = 3$) *Inspecting the positive normal sets of the Crammer-Singer surrogate $\boldsymbol{\psi}^{\mathrm{CS}}$ (for $n = 3$) in Figure 3 and the trigger probability sets of the 'abstain' loss matrix $\mathbf{L}^{(?)}$ in Figure 2(c), we see that $\mathcal{N}^{\mathrm{CS}}(\mathbf{z}_i) = \mathcal{Q}_i^{(?)} \; \forall i \in [4]$, and therefore by Theorem 8, the Crammer-Singer surrogate $\boldsymbol{\psi}^{\mathrm{CS}}$ is $\mathbf{L}^{(?)}$-calibrated. Similarly, looking at the trigger probability sets of the ordinal regression loss matrix $\mathbf{L}^{\mathrm{ord}}$ in Figure 2(b) and again applying Theorem 8, we see that the Crammer-Singer surrogate $\boldsymbol{\psi}^{\mathrm{CS}}$ is also $\mathbf{L}^{\mathrm{ord}}$-calibrated!*

Some additional examples of applications of Theorems 6 and 8 are provided in Section 3.3 below. Both the necessary and sufficient conditions above will also be used when we study the convex calibration dimension of a loss matrix $\mathbf{L}$ in Section 4.

3.3 Computation of Positive Normal Sets

Both the necessary and sufficient conditions for calibration above involve the positive normal sets $\mathcal{N}^\psi(\mathbf{z})$ at various points $\mathbf{z} \in \mathcal{S}_\psi$. Thus in order to use the above results to show that a surrogate $\boldsymbol{\psi}$ is (or is not) $\mathbf{L}$-calibrated, one needs to be able to compute or characterize the sets $\mathcal{N}^\psi(\mathbf{z})$. Here we give a method for computing these sets for certain surrogates $\boldsymbol{\psi}$ at certain points $\mathbf{z} \in \mathcal{S}_\psi$. Specifically, the following result gives an explicit method for computing $\mathcal{N}^\psi(\mathbf{z})$ for convex surrogate losses $\boldsymbol{\psi}$ operating on a convex surrogate space $\mathcal{C} \subseteq \mathbb{R}^d$, at points $\mathbf{z} = \boldsymbol{\psi}(\mathbf{u}) \in \mathcal{R}_\psi$ for which the subdifferential $\partial \psi_y(\mathbf{u})$ for each $y \in [n]$

13

can be described as the convex hull of a finite number of points in $\mathbb{R}^d$; this is particularly applicable for piecewise linear surrogates.[4,5]

Lemma 9 *Let $\mathcal{C} \subseteq \mathbb{R}^d$ be a convex set and let $\boldsymbol{\psi} : \mathcal{C} \to \mathbb{R}_+^n$ be convex. Let $\mathbf{z} = \boldsymbol{\psi}(\mathbf{u})$ for some $\mathbf{u} \in \mathcal{C}$ such that $\forall y \in [n]$, the subdifferential of ψ_y at $\mathbf{u}$ can be written as*

$$\partial\psi_y(\mathbf{u}) = \mathrm{conv}\left(\{\mathbf{w}_1^y, \ldots, \mathbf{w}_{s_y}^y\}\right)$$

for some $s_y \in \mathbb{Z}_+$ and $\mathbf{w}_1^y, \ldots, \mathbf{w}_{s_y}^y \in \mathbb{R}^d$. 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

$$\mathcal{N}^\psi(\mathbf{z}) = \left\{\mathbf{p} \in \Delta_n : \mathbf{p} = \mathbf{B}\mathbf{q} \text{ for some } \mathbf{q} \in \mathrm{null}(\mathbf{A}) \cap \Delta_s\right\},$$

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

The proof makes use of the fact that a convex function $\phi : \mathbb{R}^d \to \mathbb{R}$ 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 Bertsekas et al. (2003)). We will also make use of the fact that if $\phi_1, \phi_2 : \mathbb{R}^d \to \mathbb{R}$ 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 (Proof of Lemma 9)

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

$$
\begin{aligned}
\mathbf{p} \in \mathcal{N}^\psi(\boldsymbol{\psi}(\mathbf{u})) \iff & \mathbf{p} \in \Delta_n, \ \mathbf{p}^\top\boldsymbol{\psi}(\mathbf{u}) \leq \mathbf{p}^\top\mathbf{z}' \ \forall \mathbf{z}' \in \mathcal{S}_\psi \\
\iff & \mathbf{p} \in \Delta_n, \ \mathbf{p}^\top\boldsymbol{\psi}(\mathbf{u}) \leq \mathbf{p}^\top\mathbf{z}' \ \forall \mathbf{z}' \in \mathcal{R}_\psi \\
\iff & \mathbf{p} \in \Delta_n, \text{ and the convex function } \phi(\mathbf{u}') = \mathbf{p}^\top\boldsymbol{\psi}(\mathbf{u}') = \textstyle\sum_{y=1}^n p_y\psi_y(\mathbf{u}') \\
& \text{achieves its minimum at } \mathbf{u}' = \mathbf{u} \\
\iff & \mathbf{p} \in \Delta_n, \ \mathbf{0} \in \sum_{y=1}^n p_y\partial\psi_y(\mathbf{u}) \\
\iff & \mathbf{p} \in \Delta_n, \ \mathbf{0} = \sum_{y=1}^n p_y \sum_{j=1}^{s_y} v_j^y\mathbf{w}_j^y \text{ for some } \mathbf{v}^y \in \Delta_{s_y} \\
\iff & \mathbf{p} \in \Delta_n, \ \mathbf{0} = \sum_{y=1}^n \sum_{j=1}^{s_y} q_j^y\mathbf{w}_j^y \text{ for some } \mathbf{q}^y = p_y\mathbf{v}^y, \quad \mathbf{v}^y \in \Delta_{s_y} \\
\iff & \mathbf{p} \in \Delta_n, \mathbf{A}\mathbf{q} = \mathbf{0} \text{ for some } \mathbf{q} = (p_1\mathbf{v}^1, \ldots, p_n\mathbf{v}^n)^\top \in \Delta_s, \quad \mathbf{v}^y \in \Delta_{s_y} \\
\iff & \mathbf{p} = \mathbf{B}\mathbf{q} \text{ for some } \mathbf{q} \in \mathrm{null}(\mathbf{A}) \cap \Delta_s.
\end{aligned}
$$

---

4. Recall that a vector function is convex if all its component functions are convex.
5. Recall that the subdifferential of a convex function $\phi : \mathbb{R}^d \to \overline{\mathbb{R}}$ 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 Bertsekas et al. (2003)).

■

We now give examples of computation of positive normal sets using Lemma 9 for two convex surrogates, both of which operate on the one-dimensional surrogate space $\mathcal{C} = \mathbb{R}$, and as we shall see, turn out to be calibrated w.r.t. the ordinal regression loss $\mathbf{L}^{\mathrm{ord}}$ but not w.r.t. the 0-1 loss $\mathbf{L}^{0\text{-}1}$ or the 'abstain' loss $\mathbf{L}^{(?)}$. As another example of an application of Lemma 9, calculations showing computation of positive normal sets of the Crammer-Singer surrogate (as shown in Figure 3) are given in the appendix.

Example 8 (Positive normal sets of 'absolute' surrogate) *Let $n = 3$, and let $\mathcal{C} = \mathbb{R}$. Consider the 'absolute' surrogate $\boldsymbol{\psi}^{\mathrm{abs}} : \mathbb{R} \to \mathbb{R}^3_+$ defined as follows:*

$$\psi_y^{\mathrm{abs}}(u) = |u - y| \qquad \forall y \in [3],\, u \in \mathbb{R}. \tag{10}$$

Clearly, $\boldsymbol{\psi}^{\mathrm{abs}}$ is a convex function (see Figure 5). Moreover, we have

$$\mathcal{R}_{\mathrm{abs}} = \boldsymbol{\psi}^{\mathrm{abs}}(\mathbb{R}) = \left\{ (|u - 1|, |u - 2|, |u - 3|)^\top : u \in \mathbb{R} \right\} \subset \mathbb{R}^3_+.$$

Now let $u_1 = 1$, $u_2 = 2$, and $u_3 = 3$, and let

$$\begin{aligned}
\mathbf{z}_1 &= \boldsymbol{\psi}^{\mathrm{abs}}(u_1) = \boldsymbol{\psi}^{\mathrm{abs}}(1) = (0, 1, 2)^\top \in \mathcal{R}_{\mathrm{abs}} \\
\mathbf{z}_2 &= \boldsymbol{\psi}^{\mathrm{abs}}(u_2) = \boldsymbol{\psi}^{\mathrm{abs}}(2) = (1, 0, 1)^\top \in \mathcal{R}_{\mathrm{abs}} \\
\mathbf{z}_3 &= \boldsymbol{\psi}^{\mathrm{abs}}(u_3) = \boldsymbol{\psi}^{\mathrm{abs}}(3) = (2, 1, 0)^\top \in \mathcal{R}_{\mathrm{abs}}.
\end{aligned}$$

Let us consider computing the positive normal sets of $\boldsymbol{\psi}^{\mathrm{abs}}$ at the 3 points $\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3$ above. To see that $\mathbf{z}_1$ satisfies the conditions of Lemma 9, note that

$$\begin{aligned}
\partial \psi_1^{\mathrm{abs}}(u_1) &= \partial \psi_1^{\mathrm{abs}}(1) = [-1, 1] = \mathrm{conv}(\{+1, -1\}); \\
\partial \psi_2^{\mathrm{abs}}(u_1) &= \partial \psi_2^{\mathrm{abs}}(1) = \{-1\} = \mathrm{conv}(\{-1\}); \\
\partial \psi_3^{\mathrm{abs}}(u_1) &= \partial \psi_3^{\mathrm{abs}}(1) = \{-1\} = \mathrm{conv}(\{-1\}).
\end{aligned}$$

Therefore, we can use Lemma 9 to compute $\mathcal{N}^{\mathrm{abs}}(\mathbf{z}_1)$. Here $s = 4$, and

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

This gives

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

It is easy to see that $\mathbf{z}_2$ and $\mathbf{z}_3$ also satisfy the conditions of Lemma 9; similar computations then yield

$$\begin{aligned}
\mathcal{N}^{\mathrm{abs}}(\mathbf{z}_2) &= \left\{ \mathbf{p} \in \Delta_3 : p_1 \leq \tfrac{1}{2}, p_3 \leq \tfrac{1}{2} \right\} \\
\mathcal{N}^{\mathrm{abs}}(\mathbf{z}_3) &= \left\{ \mathbf{p} \in \Delta_3 : p_3 \geq \tfrac{1}{2} \right\}.
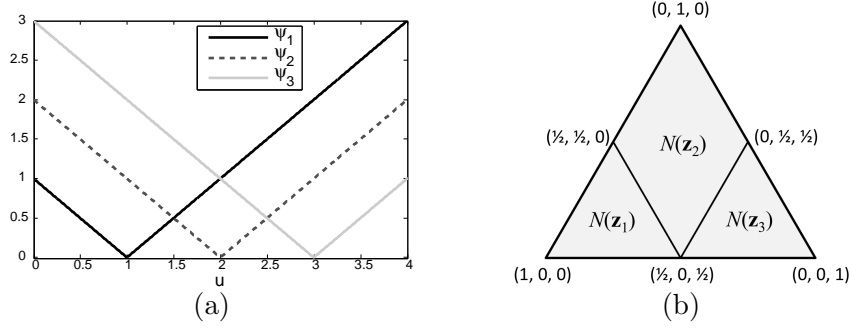\end{aligned}$$

Figure 5: (a) The 'absolute' surrogate $\boldsymbol{\psi}^{\mathrm{abs}} : \mathbb{R} \to \mathbb{R}_+^3$ (for $n = 3$), and (b) its positive normal sets at 3 points $\mathbf{z}_i = \boldsymbol{\psi}^{\mathrm{abs}}(u_i) \in \mathbb{R}_+^3$ ($i \in [3]$) for $u_1 = 1, u_2 = 2, u_3 = 3$. See Example 8 for details.

The positive normal sets above are shown in Figure 5. Comparing these with the trigger probability sets in Figure 2, we have by Theorem 8 that $\boldsymbol{\psi}^{\mathrm{abs}}$ is $\mathbf{L}^{\mathrm{ord}}$-calibrated, and by Theorem 6 that $\boldsymbol{\psi}^{\mathrm{abs}}$ is not calibrated w.r.t. $\mathbf{L}^{0\text{-}1}$ or $\mathbf{L}^{(?)}$.

Example 9 (Positive normal sets of 'ϵ-insensitive' surrogate) *Let $n = 3$, and let $\mathcal{C} = \mathbb{R}$. Let $\epsilon \in [0, 0.5)$, and consider the 'ϵ-insensitive' surrogate $\boldsymbol{\psi}^\epsilon : \mathbb{R} \to \mathbb{R}_+^3$ defined as follows:*

$$\psi_y^\epsilon(u) = \big(|u - y| - \epsilon\big)_+ \qquad \forall y \in [3], u \in \mathbb{R}. \tag{11}$$

For $\epsilon = 0$, we have $\boldsymbol{\psi}^\epsilon = \boldsymbol{\psi}^{\mathrm{abs}}$. Clearly, $\boldsymbol{\psi}^\epsilon$ is a convex function (see Figure 6). Moreover, we have

$$\mathcal{R}_\epsilon = \boldsymbol{\psi}^\epsilon(\mathbb{R}) = \Big\{ \big((|u - 1| - \epsilon)_+, (|u - 2| - \epsilon)_+, (|u - 3| - \epsilon)_+\big)^\top : u \in \mathbb{R} \Big\} \subset \mathbb{R}_+^3.$$

For concreteness, we will take $\epsilon = 0.25$ below, but similar computations hold $\forall \epsilon \in (0, 0.5)$. Let $u_1 = 1 + \epsilon = 1.25$, $u_2 = 2 - \epsilon = 1.75$, $u_3 = 2 + \epsilon = 2.25$, and $u_4 = 3 - \epsilon = 2.75$, and let

$$\begin{aligned}
\mathbf{z}_1 &= \boldsymbol{\psi}^{0.25}(u_1) = \boldsymbol{\psi}^{0.25}(1.25) = (0, 0.5, 1.5)^\top \in \mathcal{R}_{0.25} \\
\mathbf{z}_2 &= \boldsymbol{\psi}^{0.25}(u_2) = \boldsymbol{\psi}^{0.25}(1.75) = (0.5, 0, 1)^\top \in \mathcal{R}_{0.25} \\
\mathbf{z}_3 &= \boldsymbol{\psi}^{0.25}(u_3) = \boldsymbol{\psi}^{0.25}(2.25) = (1, 0, 0.5)^\top \in \mathcal{R}_{0.25} \\
\mathbf{z}_4 &= \boldsymbol{\psi}^{0.25}(u_4) = \boldsymbol{\psi}^{0.25}(2.75) = (1.5, 0.5, 0)^\top \in \mathcal{R}_{0.25}.
\end{aligned}$$

Let us consider computing the positive normal sets of $\boldsymbol{\psi}^{0.25}$ at the 4 points $\mathbf{z}_i$ ($i \in [4]$) above. To see that $\mathbf{z}_1$ satisfies the conditions of Lemma 9, note that

$$\begin{aligned}
\partial \psi_1^{0.25}(u_1) &= \partial \psi_1^{0.25}(1.25) = [0, 1] = \mathrm{conv}(\{0, 1\}); \\
\partial \psi_2^{0.25}(u_1) &= \partial \psi_2^{0.25}(1.25) = \{-1\} = \mathrm{conv}(\{-1\}); \\
\partial \psi_3^{0.25}(u_1) &= \partial \psi_3^{0.25}(1.25) = \{-1\} = \mathrm{conv}(\{-1\}).
\end{aligned}$$

16

Therefore, we can use Lemma 9 to compute $\mathcal{N}^{0.25}(\mathbf{z}_1)$. Here $s = 4$, and

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

This gives

$$
\begin{aligned}
\mathcal{N}^{0.25}(\mathbf{z}_1) &= \left\{ \mathbf{p} \in \Delta_3 : \mathbf{p} = (q_1 + q_2, q_3, q_4) \text{ for some } \mathbf{q} \in \Delta_4, \ q_2 - q_3 - q_4 = 0 \right\} \\
&= \left\{ \mathbf{p} \in \Delta_3 : \mathbf{p} = (q_1 + q_2, q_3, q_4) \text{ for some } \mathbf{q} \in \Delta_4, \ q_1 + q_2 \geq q_3 + q_4 \right\} \\
&= \left\{ \mathbf{p} \in \Delta_3 : p_1 \geq \tfrac{1}{2} \right\}.
\end{aligned}
$$

Similarly, to see that $\mathbf{z}_2$ satisfies the conditions of Lemma 9, note that

$$
\begin{aligned}
\partial \psi_1^{0.25}(u_2) &= \partial \psi_1^{0.25}(1.75) &=& \ \{1\} = \mathrm{conv}(\{1\}); \\
\partial \psi_2^{0.25}(u_2) &= \partial \psi_2^{0.25}(1.75) &=& \ [-1, 0] = \mathrm{conv}(\{-1, 0\}); \\
\partial \psi_3^{0.25}(u_2) &= \partial \psi_3^{0.25}(1.75) &=& \ \{-1\} = \mathrm{conv}(\{-1\}).
\end{aligned}
$$

Again, we can use Lemma 9 to compute $\mathcal{N}^{0.25}(\mathbf{z}_2)$; here $s = 4$, and

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

This gives

$$
\begin{aligned}
\mathcal{N}^{0.25}(\mathbf{z}_2) &= \left\{ \mathbf{p} \in \Delta_3 : \mathbf{p} = (q_1, q_2 + q_3, q_4) \text{ for some } \mathbf{q} \in \Delta_4, \ q_1 - q_2 - q_4 = 0 \right\} \\
&= \left\{ \mathbf{p} \in \Delta_3 : p_1 \geq p_3, \ p_1 \leq \tfrac{1}{2} \right\}.
\end{aligned}
$$

It is easy to see that $\mathbf{z}_3$ and $\mathbf{z}_4$ also satisfy the conditions of Lemma 9; similar computations then yield

$$
\begin{aligned}
\mathcal{N}^{0.25}(\mathbf{z}_3) &= \left\{ \mathbf{p} \in \Delta_3 : p_1 \leq p_3, \ p_3 \leq \tfrac{1}{2} \right\} \\
\mathcal{N}^{0.25}(\mathbf{z}_4) &= \left\{ \mathbf{p} \in \Delta_3 : p_3 \geq \tfrac{1}{2} \right\}.
\end{aligned}
$$

The positive normal sets above are shown in Figure 6. Comparing these with the trigger probability sets in Figure 2, we have by Theorem 8 that $\psi^{0.25}$ is $\mathbf{L}^{\mathrm{ord}}$-calibrated, and by Theorem 6 that $\psi^{0.25}$ is not calibrated w.r.t. $\mathbf{L}^{0\text{-}1}$ or $\mathbf{L}^{(?)}$.

4. Convex Calibration Dimension

We now turn to the study of a fundamental quantity associated with the property of $\mathbf{L}$-calibration. Specifically, in Examples 6 and 7 above, we saw that to develop a surrogate calibrated w.r.t. to the ordinal regression loss $\mathbf{L}^{\mathrm{ord}}$ for $n = 3$, it was sufficient to consider a surrogate prediction space $\mathcal{C} = \mathbb{R}$, with dimension $d = 1$; in addition, the surrogates we considered were convex, and can therefore be used in developing computationally efficient
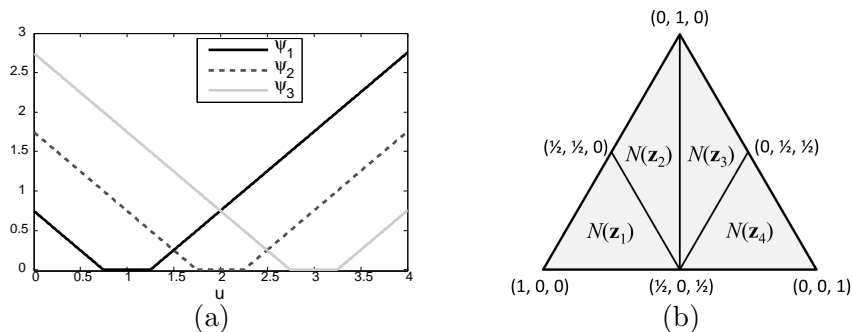
Figure 6: (a) The 'ϵ-insensitive' surrogate $\boldsymbol{\psi}^\epsilon : \mathbb{R} \to \mathbb{R}_+^3$ for $\epsilon = 0.25$ (and $n = 3$), and (b) its positive normal sets at 4 points $\mathbf{z}_i = \boldsymbol{\psi}^\epsilon(u_i) \in \mathbb{R}_+^3$ ($i \in [4]$) for $u_1 = 1.25, u_2 = 1.75, u_3 = 2.25, u_4 = 2.75$. See Example 9 for details.

algorithms. In fact the same surrogate prediction space with $d = 1$ can be used to develop similar convex surrogate losses calibrated w.r.t. the $\mathbf{L}^{\mathrm{ord}}$ for *any* $n \in \mathbb{Z}_+$. However not all multiclass loss matrices $\mathbf{L}$ have such 'low-dimensional' convex surrogates. This raises the natural question of what is the smallest dimension d that supports a convex $\mathbf{L}$-calibrated surrogate for a given multiclass loss $\mathbf{L}$, and leads us to the following definition:

Definition 10 (Convex calibration dimension) *Let* $\mathbf{L} \in \mathbb{R}_+^{n \times k}$. *Define the* convex calibration dimension (CC dimension) *of* $\mathbf{L}$ *as*

$$\mathrm{CCdim}(\mathbf{L}) \overset{\triangle}{=} \min \big\{ d \in \mathbb{Z}_+ : \exists \text{ a convex set } \mathcal{C} \subseteq \mathbb{R}^d \text{ and a convex surrogate } \boldsymbol{\psi} : \mathcal{C} \to \mathbb{R}_+^n$$
$$\text{that is } \mathbf{L}\text{-calibrated} \big\},$$

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

The CC-dimension of a loss matrix $\mathbf{L}$ provides an important measure of the 'complexity' of designing convex calibrated surrogates for $\mathbf{L}$. Indeed, while the computational complexity of minimizing a surrogate loss, as well as that of converting surrogate predictions into target predictions, can depend on factors other than the dimension d of the surrogate space $\mathcal{C} \subseteq \mathbb{R}^d$, in the absence of other guiding factors, one would in general prefer to use a surrogate in a lower dimension d since this involves learning a smaller number of real-valued functions.

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

4.1 Upper Bounds on the Convex Calibration Dimension

We start with a simple result that establishes that the CC dimension of any multiclass loss matrix $\mathbf{L}$ is finite, and in fact is strictly smaller than the number of class labels n.

Lemma 11 *Let* $\mathbf{L} \in \mathbb{R}_+^{n \times k}$. *Let* $\mathcal{C} = \left\{ \mathbf{u} \in \mathbb{R}_+^{n-1} : \sum_{j=1}^{n-1} u_j \leq 1 \right\}$, *and for each* $y \in [n]$, *let* $\psi_y : \mathcal{C} \rightarrow \mathbb{R}_+$ *be given by*

$$\psi_y(\mathbf{u}) \;=\; \mathbf{1}(y \neq n)\,(u_y - 1)^2 + \sum_{j \in [n-1], j \neq y} {u_j}^2 \,.$$

Then $\boldsymbol{\psi}$ *is* $\mathbf{L}$-*calibrated. In particular, since* $\boldsymbol{\psi}$ *is convex,* $\mathrm{CCdim}(\mathbf{L}) \leq n - 1$.

It may appear surprising that the convex surrogate $\boldsymbol{\psi}$ in the above lemma, operating on a surrogate space $\mathcal{C} \subset \mathbb{R}^{n-1}$, is $\mathbf{L}$-calibrated for *all* multiclass losses $\mathbf{L}$ 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 (Vernet et al., 2011).

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 prediction space in a lower-dimensional space. Next we give a different upper bound on the CC dimension that depends on the loss $\mathbf{L}$, and for certain losses, can be significantly tighter than the general bound above.

Theorem 12 *Let* $\mathbf{L} \in \mathbb{R}_+^{n \times k}$. *Then* $\mathrm{CCdim}(\mathbf{L}) \leq \mathrm{affdim}(\mathbf{L})$.

Proof Let $\mathrm{affdim}(\mathbf{L}) = d$. We will construct a convex $\mathbf{L}$-calibrated surrogate loss $\boldsymbol{\psi}$ with surrogate prediction space $\mathcal{C} \subseteq \mathbb{R}^d$.

Let $\mathcal{V} \subseteq \mathbb{R}^n$ denote the (d-dimensional) subspace parallel to the affine hull of the column vectors of $\mathbf{L}$, and let $\mathbf{r} \in \mathbb{R}^n$ be the corresponding translation vector, so that $\mathcal{V} = \mathrm{aff}(\{\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_k\}) + \mathbf{r}$. Let $\mathbf{v}_1, \ldots, \mathbf{v}_d \in \mathcal{V}$ be d linearly independent vectors in $\mathcal{V}$. Let $\{\mathbf{e}_1, \ldots, \mathbf{e}_d\}$ denote the standard basis in $\mathbb{R}^d$, and define a linear function $\widetilde{\boldsymbol{\psi}} : \mathbb{R}^d \rightarrow \mathbb{R}^n$ by

$$\widetilde{\boldsymbol{\psi}}(\mathbf{e}_j) = \mathbf{v}_j \quad \forall j \in [d] \,.$$

Then for each $\mathbf{v} \in \mathcal{V}$, there exists a unique vector $\mathbf{u} \in \mathbb{R}^d$ such that $\widetilde{\boldsymbol{\psi}}(\mathbf{u}) = \mathbf{v}$. In particular, since $\boldsymbol{\ell}_t + \mathbf{r} \in \mathcal{V} \ \forall t \in [k]$, there exist unique vectors $\mathbf{u}_1, \ldots, \mathbf{u}_k \in \mathbb{R}^d$ such that for each $t \in [k]$, $\widetilde{\boldsymbol{\psi}}(\mathbf{u}_t) = \boldsymbol{\ell}_t + \mathbf{r}$. Let $\mathcal{C} = \mathrm{conv}(\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}) \subseteq \mathbb{R}^d$, and define $\boldsymbol{\psi} : \mathcal{C} \rightarrow \mathbb{R}_+^n$ as

$$\boldsymbol{\psi}(\mathbf{u}) = \widetilde{\boldsymbol{\psi}}(\mathbf{u}) - \mathbf{r} \quad \forall \mathbf{u} \in \mathcal{C} \,.$$

To see that $\boldsymbol{\psi}(\mathbf{u}) \in \mathbb{R}_+^n \ \forall \mathbf{u} \in \mathcal{C}$, note that for any $\mathbf{u} \in \mathcal{C}$, $\exists \boldsymbol{\alpha} \in \Delta_k$ such that $\mathbf{u} = \sum_{t=1}^k \alpha_t \mathbf{u}_t$, which gives $\boldsymbol{\psi}(\mathbf{u}) = \widetilde{\boldsymbol{\psi}}(\mathbf{u}) - \mathbf{r} = \left(\sum_{t=1}^k \alpha_t \widetilde{\boldsymbol{\psi}}(\mathbf{u}_t) \right) - \mathbf{r} = \left(\sum_{t=1}^k \alpha_t (\boldsymbol{\ell}_t + \mathbf{r}) \right) - \mathbf{r} = \sum_{t=1}^k \alpha_t \boldsymbol{\ell}_t$ (and $\boldsymbol{\ell}_t \in \mathbb{R}_+^n \ \forall t \in [k]$). The function $\boldsymbol{\psi}$ is clearly convex. To show $\boldsymbol{\psi}$ is $\mathbf{L}$-calibrated, we will use Theorem 8. 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 $\mathcal{S}_\psi = \mathrm{conv}(\boldsymbol{\psi}(\mathcal{C})) = \mathrm{conv}(\{\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_k\})$; from the definitions of positive normals and trigger probabilities, it then follows that $\mathcal{N}^\psi(\mathbf{z}_t) = \mathcal{N}^\psi(\boldsymbol{\ell}_t) = \mathcal{Q}_t^{\mathbf{L}}$ for all $t \in [k]$. Thus by Theorem 8, $\boldsymbol{\psi}$ is $\mathbf{L}$-calibrated. ∎

Since affdim($\mathbf{L}$) is equal to either rank($\mathbf{L}$) or rank($\mathbf{L}$) $- 1$, this immediately gives us the following corollary:

Corollary 13 *Let* $\mathbf{L} \in \mathbb{R}_+^{n \times k}$. *Then* $\mathrm{CCdim}(\mathbf{L}) \leq \mathrm{rank}(\mathbf{L})$.

Proof Follows immediately from Theorem 12 and the fact that affdim($\mathbf{L}$) $\leq$ rank($\mathbf{L}$). ∎

Example 10 (CC dimension of Hamming loss) *Let* $n = 2^r$ *for some* $r \in \mathbb{Z}_+$, *and consider the Hamming loss* $\mathbf{L}^{\mathrm{Ham}} \in \mathbb{R}_+^{n \times n}$ *defined in Example 3. As in Example 3, for each* $z \in \{0, \ldots, 2^r - 1\}$, *let* $z_i \in \{0, 1\}$ *denote the i-th bit in the r-bit binary representation of z. For each* $y \in [n]$, *define* $\boldsymbol{\sigma}_y \in \{\pm 1\}^r$ *as*

$$\sigma_{yi} = 2(y-1)_i - 1 = \begin{cases} +1 & \text{if } (y-1)_i = 1 \\ -1 & \text{otherwise.} \end{cases}$$

Then we have

$$\begin{aligned} \ell_{yt}^{\mathrm{Ham}} &= \sum_{i=1}^{r} \mathbf{1}\big((y-1)_i \neq (t-1)_i\big) \\ &= \sum_{i=1}^{r} \left(\frac{1 - \sigma_{yi}\sigma_{ti}}{2} \right) \\ &= \frac{r}{2} - \sum_{i=1}^{r} \frac{\sigma_{yi}\sigma_{ti}}{2} \quad \forall y, t \in [n]. \end{aligned}$$

Thus affdim($\mathbf{L}^{\mathrm{Ham}}$) $\leq r$, *and therefore by Theorem 12, we have* $\mathrm{CCdim}(\mathbf{L}^{\mathrm{Ham}}) \leq r$. *This is a significantly tighter upper bound than the bound of* $2^r - 1$ *given by Lemma 11.*

4.2 Lower Bound on the Convex Calibration Dimension

In this section we give a lower bound on the CC dimension of a loss matrix $\mathbf{L}$ and illustrate it by using it to calculate the CC dimension of the 0-1 loss. In Section 5 we will explore applications of the lower bound to obtaining impossibility results on the existence of convex calibrated surrogates in low-dimensional surrogate spaces for certain types of subset ranking losses. We will need the following definition:

Definition 14 (Feasible subspace dimension) *The* feasible subspace dimension *of a convex set* $\mathcal{Q} \subseteq \mathbb{R}^n$ *at a point* $\mathbf{p} \in \mathcal{Q}$, *denoted by* $\mu_{\mathcal{Q}}(\mathbf{p})$, *is defined as the dimension of the subspace* $\mathcal{F}_{\mathcal{Q}}(\mathbf{p}) \cap (-\mathcal{F}_{\mathcal{Q}}(\mathbf{p}))$, *where* $\mathcal{F}_{\mathcal{Q}}(\mathbf{p})$ *is the cone of feasible directions of* $\mathcal{Q}$ *at* $\mathbf{p}$.[6]

In essence, the feasible subspace dimension of a convex set Q at a point $\mathbf{p} \in \mathcal{Q}$ is simply the dimension of the smallest face of $\mathcal{Q}$ containing $\mathbf{p}$; see Figure 7 for an illustration.

---

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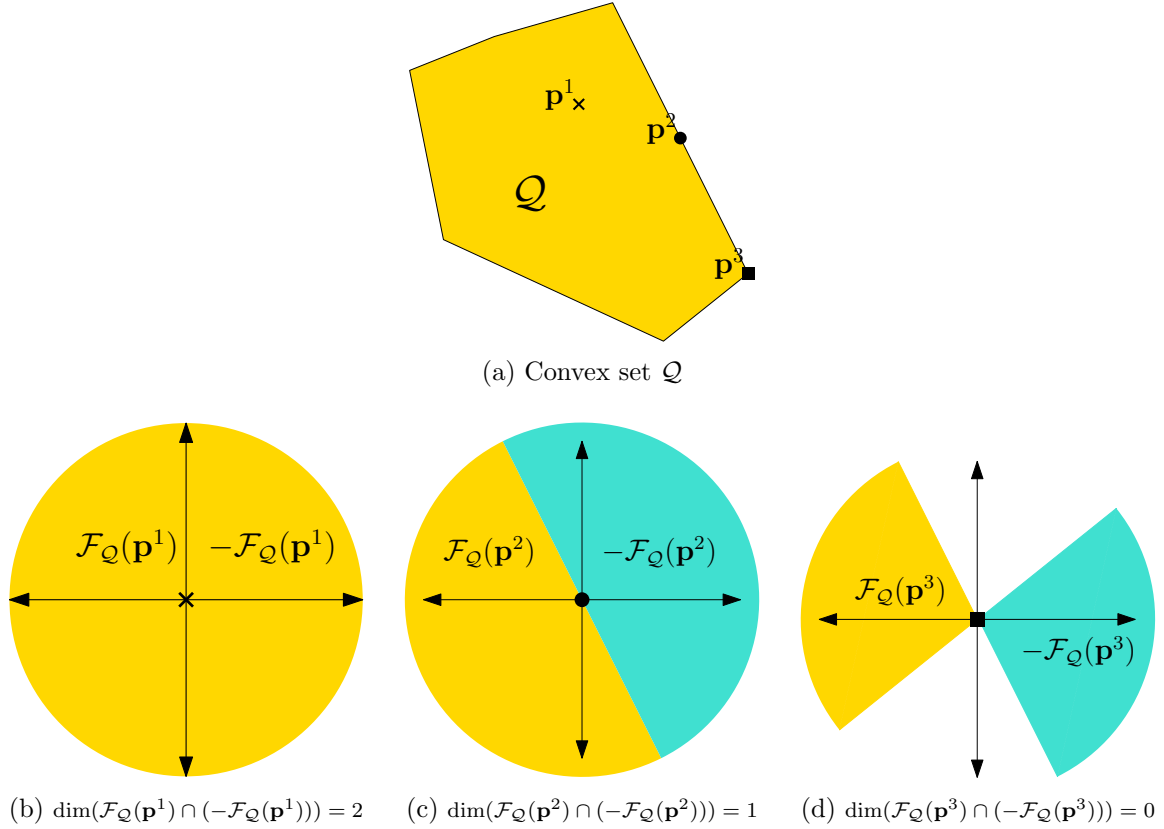
(a) Convex set $\mathcal{Q}$

(b) $\dim(\mathcal{F}_{\mathcal{Q}}(\mathbf{p}^1) \cap (-\mathcal{F}_{\mathcal{Q}}(\mathbf{p}^1))) = 2$ (c) $\dim(\mathcal{F}_{\mathcal{Q}}(\mathbf{p}^2) \cap (-\mathcal{F}_{\mathcal{Q}}(\mathbf{p}^2))) = 1$ (d) $\dim(\mathcal{F}_{\mathcal{Q}}(\mathbf{p}^3) \cap (-\mathcal{F}_{\mathcal{Q}}(\mathbf{p}^3))) = 0$

Figure 7: Illustration of feasible subspace dimension $\mu_{\mathcal{Q}}(\mathbf{p})$ of a 2-dimensional convex set $\mathcal{Q}$ at three points $\mathbf{p} = \mathbf{p}^1, \mathbf{p}^2, \mathbf{p}^3$. Here $\mu_{\mathcal{Q}}(\mathbf{p}^1) = 2$, $\mu_{\mathcal{Q}}(\mathbf{p}^2) = 1$, and $\mu_{\mathcal{Q}}(\mathbf{p}^3) = 0$.

Both the proof of the lower bound we will provide below and its applications make use of the following lemma, which gives a method to calculate the feasible subspace dimension for certain convex sets $\mathcal{Q}$ and points $\mathbf{p} \in \mathcal{Q}$:

Lemma 15 *Let* $\mathcal{Q} = \left\{\mathbf{q} \in \mathbb{R}^n : \mathbf{A}^1\mathbf{q} \leq \mathbf{b}^1, \mathbf{A}^2\mathbf{q} \leq \mathbf{b}^2, \mathbf{A}^3\mathbf{q} = \mathbf{b}^3\right\}$. *Let* $\mathbf{p} \in \mathcal{Q}$ *be such that* $\mathbf{A}^1\mathbf{p} = \mathbf{b}^1$, $\mathbf{A}^2\mathbf{p} < \mathbf{b}^2$. *Then* $\mu_{\mathcal{Q}}(\mathbf{p}) = \text{nullity}\left(\begin{bmatrix}\mathbf{A}^1 \\ \mathbf{A}^3\end{bmatrix}\right)$.

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

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

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

Thus $\mathbf{v} \in \mathcal{F}_{\mathcal{Q}}(\mathbf{p})$. Similarly, we can show $-\mathbf{v} \in \mathcal{F}_{\mathcal{Q}}(\mathbf{p})$. Thus $\mathbf{v} \in \mathcal{F}_{\mathcal{Q}}(\mathbf{p}) \cap (-\mathcal{F}_{\mathcal{Q}}(\mathbf{p}))$, giving null$\left(\begin{bmatrix} \mathbf{A}^1 \\ \mathbf{A}^3 \end{bmatrix}\right) \subseteq \mathcal{F}_{\mathcal{Q}}(\mathbf{p}) \cap (-\mathcal{F}_{\mathcal{Q}}(\mathbf{p}))$. Now let $\mathbf{v} \in \mathcal{F}_{\mathcal{Q}}(\mathbf{p}) \cap (-\mathcal{F}_{\mathcal{Q}}(\mathbf{p}))$. 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 $\mathcal{F}_{\mathcal{Q}}(\mathbf{p}) \cap (-\mathcal{F}_{\mathcal{Q}}(\mathbf{p})) \subseteq$ null$\left(\begin{bmatrix} \mathbf{A}^1 \\ \mathbf{A}^3 \end{bmatrix}\right)$. ∎

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

Theorem 16 *Let $\mathbf{L} \in \mathbb{R}_+^{n \times k}$. Let $\mathbf{p} \in \Delta_n$ and $t \in \arg\min_{t'} \mathbf{p}^\top \boldsymbol{\ell}_{t'}$ (equivalently, let $\mathbf{p} \in \mathcal{Q}_t^{\mathbf{L}}$). Then*

$$\mathrm{CCdim}(\mathbf{L}) \geq \|\mathbf{p}\|_0 - \mu_{\mathcal{Q}_t^{\mathbf{L}}}(\mathbf{p}) - 1.$$

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

Example 11 (CC dimension of 0-1 loss) *Let $n \in \mathbb{Z}_+$, and consider the 0-1 loss $\mathbf{L}^{0\text{-}1} \in \mathbb{R}_+^{n \times n}$ defined in Example 1. Take $\mathbf{p} = (\frac{1}{n}, \ldots, \frac{1}{n})^\top \in \Delta_n$. Then $\mathbf{p} \in \mathcal{Q}_t^{0\text{-}1}$ for all $t \in [k] = [n]$ (see Figure 2); in particular, we have $\mathbf{p} \in \mathcal{Q}_1^{0\text{-}1}$. Now $\mathcal{Q}_1^{0\text{-}1}$ can be written as*

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

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 $\begin{bmatrix} -\mathbf{e}_{n-1} & \mathbf{I}_{n-1} \end{bmatrix}\mathbf{p} = 0$, $-\mathbf{p} < \mathbf{0}$. Therefore, by Lemma 15, we have

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

Moreover, $\|\mathbf{p}\|_0 = n$. Thus by Theorem 16, we have $\mathrm{CCdim}(\mathbf{L}^{0\text{-}1}) \geq n - 1$. Combined with the upper bound of Lemma 11, this gives $\mathrm{CCdim}(\mathbf{L}^{0\text{-}1}) = n - 1$.

4.3 Tightness of Bounds

The upper and lower bounds above are not necessarily tight in general. For example, for the n-class ordinal regression loss of Example 2, we know that $\mathrm{CCdim}(\mathbf{L}^{\mathrm{ord}}) = 1$; however the upper bound of Theorem 12 only gives $\mathrm{CCdim}(\mathbf{L}^{\mathrm{ord}}) \leq n - 1$. Similarly, for the n-class abstain loss of Example 4, it can be shown that $\mathrm{CCdim}(\mathbf{L}^{(?)}) = O(\ln n)$ (in fact we conjecture it to be $\Theta(\ln n)$) (Ramaswamy et al., 2015), whereas the upper bound of Theorem 12 gives $\mathrm{CCdim}(\mathbf{L}^{(?)}) \leq n$, and the lower bound of Theorem 16 yields only $\mathrm{CCdim}(\mathbf{L}^{(?)}) \geq 1$. However, as we show below, for certain losses $\mathbf{L}$, the bounds of Theorems 12 and 16 are in fact tight (upto an additive constant of 1).

Lemma 17 *Let $\mathbf{L} \in \mathbb{R}_+^{n \times k}$. Let $\mathbf{p} \in \mathrm{relint}(\Delta_n)$ and $c \in \mathbb{R}_+$ be such that $\mathbf{p}^\top \boldsymbol{\ell}_t = c \ \forall t \in [k]$. Then $\forall t \in [k]$,*

$$\mu_{\mathcal{Q}_t^{\mathbf{L}}}(\mathbf{p}) \leq n - \mathrm{affdim}(\mathbf{L}) \,.$$

Proof Since $\mathbf{p}^\top \boldsymbol{\ell}_t = c \ \forall t \in [k]$, we have $\mathbf{p} \in \mathcal{Q}_t^{\mathbf{L}} \ \forall t \in [k]$. In particular, we have $\mathbf{p} \in \mathcal{Q}_1^{\mathbf{L}}$. Now

$$\mathcal{Q}_1^{\mathbf{L}} = \left\{ \mathbf{q} \in \mathbb{R}^n : \begin{bmatrix} (\boldsymbol{\ell}_2 - \boldsymbol{\ell}_1)^\top \\ \vdots \\ (\boldsymbol{\ell}_k - \boldsymbol{\ell}_1)^\top \end{bmatrix} \mathbf{q} \geq \mathbf{0}, -\mathbf{q} \leq \mathbf{0}, \mathbf{e}_n^\top \mathbf{q} = 1 \right\} \,.$$

Moreover, $\begin{bmatrix} (\boldsymbol{\ell}_2 - \boldsymbol{\ell}_1)^\top \\ \vdots \\ (\boldsymbol{\ell}_k - \boldsymbol{\ell}_1)^\top \end{bmatrix} \mathbf{p} = \mathbf{0}$ and $-\mathbf{p} < \mathbf{0}$. Therefore, by Lemma 15, we have

$$\mu_{\mathcal{Q}_1^{\mathbf{L}}}(\mathbf{p}) = \mathrm{nullity}\left(\begin{bmatrix} (\boldsymbol{\ell}_2 - \boldsymbol{\ell}_1)^\top \\ \vdots \\ (\boldsymbol{\ell}_k - \boldsymbol{\ell}_1)^\top \\ \mathbf{e}_n^\top \end{bmatrix} \right) = n - \mathrm{rank}\left(\begin{bmatrix} (\boldsymbol{\ell}_2 - \boldsymbol{\ell}_1)^\top \\ \vdots \\ (\boldsymbol{\ell}_k - \boldsymbol{\ell}_1)^\top \\ \mathbf{e}_n^\top \end{bmatrix} \right) \leq n - \mathrm{affdim}(\mathbf{L}) \,.$$

A similar proof holds for $\mu_{\mathcal{Q}_t^{\mathbf{L}}}(\mathbf{p})$ for all other $t \in [k]$. ∎

Combining the above result with Theorem 16 immediately gives the following:

Theorem 18 *Let $\mathbf{L} \in \mathbb{R}_+^{n \times k}$. If $\exists \mathbf{p} \in \mathrm{relint}(\Delta_n), c \in \mathbb{R}_+$ such that $\mathbf{p}^\top \boldsymbol{\ell}_t = c \ \forall t \in [k]$, then*

$$\mathrm{CCdim}(\mathbf{L}) \geq \mathrm{affdim}(\mathbf{L}) - 1 \,.$$

Proof Follows immediately from Theorem 16 and Lemma 17. ∎

Intuitively, the condition that $\exists \mathbf{p} \in \mathrm{relint}(\Delta_n), c \in \mathbb{R}_+$ such that $\mathbf{p}^\top \boldsymbol{\ell}_t = c \ \forall t \in [k]$ in Lemma 17 and Theorem 18 above captures the essence of a hard problem: in this case, if the underlying label probability distribution $\mathbf{p}'$ is very close to $\mathbf{p}$, then it becomes hard to decide which element $t \in [k]$ is an optimal prediction. This is essentially what leads the lower bound on the CC-dimension to become tight in this case.

A particularly useful application of Theorem 18 is to losses $\mathbf{L}$ whose columns $\boldsymbol{\ell}_t$ can be obtained from one another by permuting entries:

Corollary 19 *Let $\mathbf{L} \in \mathbb{R}_+^{n \times k}$ be such that all columns of $\mathbf{L}$ can be obtained from one another by permuting entries, i.e. $\forall t_1, t_2 \in [k], \exists \sigma \in \Pi_n$ such that $\ell_{y,t_2} = \ell_{\sigma(y),t_1} \ \forall y \in [n]$. Then*

$$\mathrm{CCdim}(\mathbf{L}) \geq \mathrm{affdim}(\mathbf{L}) - 1 \,.$$

Proof Let $\mathbf{p} = \left(\frac{1}{n}, \ldots, \frac{1}{n}\right)^\top \in \mathrm{relint}(\Delta_n)$. Let $c = \frac{\|\boldsymbol{\ell}_1\|_1}{n}$. Then under the given condition, $\mathbf{p}^\top \boldsymbol{\ell}_t = c \ \forall t \in [k]$. The result then follows from Theorem 18. ∎

We will use the above corollary in establishing lower bounds on the CC dimension of certain subset ranking losses below.

5. Applications to Subset Ranking

We now consider applications of the above framework to analyzing various subset ranking problems, where each instance $x \in \mathcal{X}$ consists of a query together with a set of r documents (for simplicity, $r \in \mathbb{Z}_+$ here is fixed), and the goal is to learn a prediction model which given such an instance predicts a ranking (permutation) of the r documents (Cossock and Zhang, 2008).[7] We consider three popular losses used for subset ranking: the normalized discounted cumulative gain (NDCG) loss, the pairwise disagreement (PD) loss, and the mean average precision (MAP) loss.[8] Each of these subset ranking losses can be viewed as a specific type of multiclass loss acting on a certain label space $\mathcal{Y}$ and prediction space $\widehat{\mathcal{Y}}$. In particular, for the NDCG loss, the label space $\mathcal{Y}$ contains r-dimensional multi-valued relevance vectors; for PD loss, $\mathcal{Y}$ contains directed acyclic graphs on r nodes; and for MAP loss, $\mathcal{Y}$ contains r-dimensional binary relevance vectors. In each case, the prediction space $\widehat{\mathcal{Y}}$ is the set of permutations of r objects: $\widehat{\mathcal{Y}} = \Pi_r$. We study the convex calibration dimension of these losses below. Specifically, we show that the CC dimension of the NDCG loss is upper bounded by r (Section 5.1), and that of both the PD and MAP losses is lower bounded by a quadratic function of r (Sections 5.2 and 5.3). Our result on the CC dimension of the NDCG loss is consistent with previous results in the literature showing the existence of r-dimensional convex calibrated surrogates for NDCG (Ravikumar et al., 2011; Buffoni et al., 2011); our results on the CC dimension of the PD and MAP losses strengthen previous results of Calauzènes et al. (2012), who showed non-existence of r-dimensional convex calibrated surrogates (with a fixed argsort predictor) for PD and MAP.

5.1 Normalized Discounted Cumulative Gain (NDCG)

The NDCG loss is widely used in information retrieval applications (Järvelin and Kekäläinen, 2000). Here $\mathcal{Y}$ is the set of r-dimensional relevance vectors with say s relevance levels, $\mathcal{Y} = \{0, 1, \ldots, s-1\}^r$, and $\widehat{\mathcal{Y}}$ is the set of permutations of r objects, $\widehat{\mathcal{Y}} = \Pi_r$ (thus here $n = |\mathcal{Y}| = s^r$ and $k = |\widehat{\mathcal{Y}}| = r!$). The loss on predicting a permutation $\sigma \in \Pi_r$ when the true label is $\mathbf{y} \in \{0, 1, \ldots, s-1\}^r$ is given by

$$\ell^{\text{NDCG}}(\mathbf{y}, \sigma) = 1 - \frac{1}{z(\mathbf{y})} \sum_{i=1}^{r} \frac{2^{y_i} - 1}{\log_2(\sigma(i) + 1)} \,,$$

where $z(\mathbf{y})$ is a normalizer that ensures the loss is non-negative and depends only on $\mathbf{y}$. The NDCG loss can therefore be viewed as a multiclass loss matrix $\mathbf{L}^{\text{NDCG}} \in \mathbb{R}_+^{s^r \times r!}$. Clearly, $\text{affdim}(\mathbf{L}^{\text{NDCG}}) \leq r$, and therefore by Theorem 12, we have

$$\text{CCdim}(\mathbf{L}^{\text{NDCG}}) \leq r \,.$$

Indeed, previous results in the literature have shown the existence of r-dimensional convex calibrated surrogates for NDCG (Ravikumar et al., 2011; Buffoni et al., 2011).

---

7. The term 'subset ranking' here refers to the fact that in a query-based setting, each instance involves a different 'subset' of documents to be ranked; see (Cossock and Zhang, 2008).

8. Note that NDCG and MAP are generally expressed as gains, where a higher value corresponds to better performance; we can express them as non-negative losses by subtracting them from a suitable constant.

5.2 Pairwise Disagreement (PD)

Here the label space $\mathcal{Y}$ is the set of all directed acyclic graphs (DAGs) on r vertices, which we shall denote as $\mathcal{G}_r$; for each directed edge (i, j) in a graph $G \in \mathcal{G}_r$ 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. The prediction space $\widehat{\mathcal{Y}}$ is again the set of permutations of r objects, $\widehat{\mathcal{Y}} = \Pi_r$. The loss on predicting a permutation $\sigma \in \Pi_r$ when the true label is $G \in \mathcal{G}_r$ is given by

$$
\begin{aligned}
\ell^{\mathrm{PD}}(G, \sigma) &= \sum_{(i,j) \in G} \mathbf{1}\big(\sigma(i) > \sigma(j)\big) \\
&= \sum_{i=1}^{r} \sum_{j=1}^{r} \mathbf{1}\big((i,j) \in G\big) \cdot \mathbf{1}\big(\sigma(i) > \sigma(j)\big) \\
&= \sum_{i=1}^{r} \sum_{j=1}^{i-1} \mathbf{1}\big((i,j) \in G\big) \cdot \mathbf{1}\big(\sigma(i) > \sigma(j)\big) + \mathbf{1}\big((j,i) \in G\big) \cdot \Big(1 - \mathbf{1}\big(\sigma(i) > \sigma(j)\big)\Big) \\
&= \sum_{i=1}^{r} \sum_{j=1}^{i-1} \Big(\mathbf{1}\big((i,j) \in G\big) - \mathbf{1}\big((j,i) \in G\big)\Big) \cdot \mathbf{1}\big(\sigma(i) > \sigma(j)\big) + \sum_{i=1}^{r} \sum_{j=1}^{i-1} \mathbf{1}\big((j,i) \in G\big) .
\end{aligned}
$$

The PD loss can be viewed as a multiclass loss matrix $\mathbf{L}^{\mathrm{PD}} \in \mathbb{R}_+^{|\mathcal{G}_r| \times r!}$. Note that the second term in the sum above depends only the label G; removing this term amounts to simply subtracting a fixed vector from each column of the loss matrix, which does not change the properties of the minimizer of the loss or its CC dimension. We can therefore consider the following loss instead:

$$
\widetilde{\ell}^{\mathrm{PD}}(G, \sigma) = \sum_{i=1}^{r} \sum_{j=1}^{i-1} \Big(\mathbf{1}\big((i,j) \in G\big) - \mathbf{1}\big((j,i) \in G\big)\Big) \cdot \mathbf{1}\big(\sigma(i) > \sigma(j)\big) .
$$

The resulting loss matrix $\widetilde{\mathbf{L}}^{\mathrm{PD}}$ clearly has rank at most $\frac{r(r-1)}{2}$. Therefore, by Corollary 13, we have

$$
\mathrm{CCdim}(\mathbf{L}^{\mathrm{PD}}) = \mathrm{CCdim}(\widetilde{\mathbf{L}}^{\mathrm{PD}}) \leq \frac{r(r-1)}{2} .
$$

In fact one can show that the rank of $\widetilde{\mathbf{L}}^{\mathrm{PD}}$ is exactly $\frac{r(r-1)}{2}$:

Proposition 20 $\mathrm{rank}(\widetilde{\mathbf{L}}^{\mathrm{PD}}) = \frac{r(r-1)}{2}$.

Moreover, it is easy to see that the columns of $\widetilde{\mathbf{L}}^{\mathrm{PD}}$ can all be obtained from one another by permuting entries. Therefore, by Corollary 19, we also have

$$
\mathrm{CCdim}(\mathbf{L}^{\mathrm{PD}}) = \mathrm{CCdim}(\widetilde{\mathbf{L}}^{\mathrm{PD}}) \geq \frac{r(r-1)}{2} - 2 .
$$

Informally, this implies that a convex surrogate that achieves calibration w.r.t. $\mathbf{L}^{\mathrm{PD}}$ over the full probability simplex must effectively 'estimate' all edge weights. Formally, this strengthens previous results of Duchi et al. (2010) and Calauzènes et al. (2012). In particular, Duchi et al. (2010) showed that certain popular r-dimensional convex surrogates are not

calibrated for the PD loss, and conjectured that such convex calibrated surrogates (in r dimensions) do not exist; Calauzènes et al. (2012) showed that indeed there do not exist any r-dimensional convex surrogates that use argsort as the predictor and are calibrated for the PD loss. The above result allows us to go further and conclude that in fact, one cannot design convex calibrated surrogates for the PD loss in any prediction space of less than $\frac{r(r-1)}{2} - 2$ dimensions (regardless of the predictor used).

5.3 Mean Average Precision (MAP)

Here the label space $\mathcal{Y}$ is the set of all (non-zero) r-dimensional binary relevance vectors, $\mathcal{Y} = \{0,1\}^r \setminus \{\mathbf{0}\}$, and the prediction space $\widehat{\mathcal{Y}}$ is again the set of permutations of r objects, $\widehat{\mathcal{Y}} = \Pi_r$. The loss on predicting a permutation $\sigma \in \Pi_r$ when the true label is $\mathbf{y} \in \{0,1\}^r \setminus \{\mathbf{0}\}$ is given by

$$
\begin{aligned}
\ell^{\mathrm{MAP}}(\mathbf{y}, \sigma) &= 1 - \frac{1}{\|\mathbf{y}\|_1} \sum_{i:y_i=1} \frac{1}{\sigma(i)} \sum_{j=1}^{\sigma(i)} y_{\sigma^{-1}(j)} \\
&= 1 - \frac{1}{\|\mathbf{y}\|_1} \sum_{i=1}^{r} \frac{y_i}{\sigma(i)} \sum_{j=1}^{\sigma(i)} y_{\sigma^{-1}(j)} \\
&= 1 - \frac{1}{\|\mathbf{y}\|_1} \sum_{i=1}^{r} \sum_{j=1}^{i} \frac{y_{\sigma^{-1}(i)}\, y_{\sigma^{-1}(j)}}{i} \\
&= 1 - \frac{1}{\|\mathbf{y}\|_1} \sum_{i=1}^{r} \sum_{j=1}^{i} \frac{y_i\, y_j}{\max(\sigma(i), \sigma(j))}
\end{aligned}
\tag{12}
$$

Thus the MAP loss can be viewed as a multiclass loss matrix $\mathbf{L}^{\mathrm{MAP}} \in \mathbb{R}_+^{(2^r-1) \times r!}$. Clearly, $\mathrm{affdim}(\mathbf{L}^{\mathrm{MAP}}) \leq \frac{r(r+1)}{2}$, and therefore by Theorem 12, we have

$$
\mathrm{CCdim}(\mathbf{L}^{\mathrm{MAP}}) \leq \frac{r(r+1)}{2}.
$$

One can also show the following lower bound on the rank of $\mathbf{L}^{\mathrm{MAP}}$:

Proposition 21 $\mathrm{rank}(\mathbf{L}^{\mathrm{MAP}}) \geq \frac{r(r-1)}{2} - 2$.

Again, it is easy to see that the columns of $\mathbf{L}^{\mathrm{MAP}}$ can all be obtained from one another by permuting entries, and therefore by Corollary 19, we have

$$
\mathrm{CCdim}(\mathbf{L}^{\mathrm{MAP}}) \geq \frac{r(r-1)}{2} - 4.
$$

This again strengthens a previous result of Calauzènes et al. (2012), who showed that there do not exist any r-dimensional convex surrogates that use argsort as the predictor and are calibrated for the MAP loss. As with the PD loss, the above result allows us to go further and conclude that in fact, one cannot design convex calibrated surrogates for the MAP loss in any prediction space of less than $\frac{r(r-1)}{2} - 4$ dimensions (regardless of the predictor used).

6. Conclusion

We have developed a unified framework for studying consistency properties of surrogate risk minimization algorithms for general multiclass learning problems, defined by a general multiclass loss matrix. In particular, we have introduced the notion of *convex calibration dimension* (CC dimension) of a multiclass loss matrix, a fundamental quantity that measures the smallest 'size' of a prediction space in which it is possible to design a convex surrogate that is calibrated with respect to the given loss matrix, and have used this to analyze consistency properties of surrogate losses for various multiclass learning problems.

Our study both generalizes previous results and sheds new light on various multiclass losses. For example, our analysis shows that for the n-class 0-1 loss, any convex calibrated surrogate must necessarily entail learning at least $n-1$ real-valued functions, thus showing that the calibrated multiclass surrogate of Lee et al. (2004), whose minimization entails learning n real-valued functions, is essentially not improvable (in the sense of the number of real-valued functions that need to be learned). Another implication of our study is to the pairwise disagreement (PD) and mean average precision (MAP) losses for subset ranking: while previous results have shown that for subset ranking problems with r documents per query, there do not exist r-dimensional convex calibrated surrogates for the PD and MAP losses, our analysis shows that (a) these losses do admit convex calibrated surrogates in higher dimensions, and (b) to obtain such convex calibrated surrogates for these losses, one needs to operate in an $\Omega(r^2)$-dimensional surrogate prediction space (i.e. one needs to learn $\Omega(r^2)$ real-valued functions, rather than just r real-valued 'scoring' functions).

As discussed in Section 4.3, while the upper and lower bounds we have obtained on the CC dimension are tight (up to an additive constant of 1) for certain classes of loss matrices, they can be quite loose in general. An important open direction is to obtain a characterization of the CC dimension in more general settings. It would also be useful to develop methods for deriving explicit surrogate regret bounds for general calibrated surrogates, through which one can relate the excess target risk to the excess surrogate risk for any multiclass loss and corresponding calibrated surrogate. Finally, another interesting direction would be to develop a generic procedure for designing convex calibrated surrogates operating on a 'minimal' space according to the CC dimension of a given loss matrix. There has been some recent progress in this direction in (Ramaswamy et al., 2013), where a general method is described for designing convex calibrated surrogates in a surrogate space with dimension at most the rank of the given loss matrix. However, while the rank forms an upper bound on the CC dimension of the loss matrix, as discussed above, this bound is not always tight, giving rise to the possibility of designing convex calibrated surrogates in lower-dimensional spaces for certain losses. Resolving these issues will contribute significantly to our understanding of the conditions under which convex calibrated surrogates can be designed for a given multiclass learning problem.

7. Proofs

7.1 Proof of Theorem 8

The proof uses the following technical lemma:

Lemma 22 *Let $\mathbf{L} \in \mathbb{R}_+^{n \times k}$ and $\psi : \mathcal{C} \to \mathbb{R}_+^n$. Suppose there exist $r \in \mathbb{N}$ and $\mathbf{z}_1, \ldots, \mathbf{z}_r \in \mathcal{R}_\psi$ such that $\bigcup_{j=1}^r \mathcal{N}^\psi(\mathbf{z}_j) = \Delta_n$ and for each $j \in [r]$, $\exists t \in [k]$ such that $\mathcal{N}^\psi(\mathbf{z}_j) \subseteq \mathcal{Q}_t^{\mathbf{L}}$. Then any element $\mathbf{z} \in \mathcal{S}_\psi$ can be written as $\mathbf{z} = \mathbf{z}' + \mathbf{z}''$ for some $\mathbf{z}' \in \mathrm{conv}(\{\mathbf{z}_1, \ldots, \mathbf{z}_r\})$ and $\mathbf{z}'' \in \mathbb{R}_+^n$.*

Proof (Proof of Lemma 22)

Let $\mathcal{S}' = \{\mathbf{z}' + \mathbf{z}'' : \mathbf{z}' \in \mathrm{conv}(\{\mathbf{z}_1, \ldots, \mathbf{z}_r\}), \mathbf{z}'' \in \mathbb{R}_+^n\}$, 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 Gallier (2009), 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 $\bigcup_{j=1}^r \mathcal{N}^\psi(\mathbf{z}_j) = \Delta_n$, $\exists j \in [r]$ such that $\mathbf{q} \in \mathcal{N}^\psi(\mathbf{z}_j)$. 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 8)

We will show $\mathbf{L}$-calibration of ψ via Lemma 2. For each $j \in [r]$, let

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

by assumption, $T_j \neq \emptyset \ \forall j \in [r]$. By Lemma 22, 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 $\mathrm{pred}' : \mathcal{S}_\psi \to [k]$ as

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

We will show pred' satisfies the condition for ℓ-calibration.

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

$$J_{\mathbf{p}} = \left\{ j \in [r] : \mathbf{p} \in \mathcal{N}^\psi(\mathbf{z}_j) \right\};$$

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

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

$$\forall j \notin J_{\mathbf{p}} : \mathbf{p}^\top \mathbf{z}_j > \inf_{\mathbf{z} \in \mathcal{S}_\psi} \mathbf{p}^\top \mathbf{z} \tag{14}$$

Moreover, from definition of T_j, we have

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

Thus we get

$$\forall j \in J_{\mathbf{p}} : \quad T_j \subseteq \operatorname{argmin}_{t'} \mathbf{p}^\top \boldsymbol{\ell}_{t'} . \tag{15}$$

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

$$\inf_{\mathbf{z} \in \mathcal{S}_\psi : \operatorname{pred}'(\mathbf{z}) \notin \operatorname{argmin}_{t'} \mathbf{p}^\top \boldsymbol{\ell}_{t'}} \mathbf{p}^\top \mathbf{z} \;=\; \inf_{\mathbf{z} \in \mathcal{S}_\psi : \operatorname{pred}'(\mathbf{z}) \notin \operatorname{argmin}_{t'} \mathbf{p}^\top \boldsymbol{\ell}_{t'}} \sum_{j=1}^{r} \alpha_j^{\mathbf{z}} \mathbf{p}^\top \mathbf{z}_j + \mathbf{p}^\top \mathbf{u}^{\mathbf{z}} \quad (16)$$

$$\geq \inf_{\boldsymbol{\alpha} \in \Delta_r : \alpha_j \geq \frac{1}{r} \text{ for some } j \notin J_{\mathbf{p}}} \sum_{j=1}^{r} \alpha_j \mathbf{p}^\top \mathbf{z}_j \tag{17}$$

$$\geq \min_{j \notin J_{\mathbf{p}}} \inf_{\alpha_j \in [\frac{1}{r}, 1]} \alpha_j \mathbf{p}^\top \mathbf{z}_j + (1 - \alpha_j) \inf_{\mathbf{z} \in \mathcal{S}_\psi} \mathbf{p}^\top \mathbf{z} \tag{18}$$

$$> \inf_{\mathbf{z} \in \mathcal{S}_\psi} \mathbf{p}^\top \mathbf{z} , \tag{19}$$

where the last inequality follows from Eq. (14). Since the above holds for all $\mathbf{p} \in \Delta_n$, by Lemma 2, we have that ψ is $\mathbf{L}$-calibrated. ∎

7.2 Proof of Lemma 11

Proof For each $\mathbf{u} \in \mathcal{C}$, define $\mathbf{p}^{\mathbf{u}} = \left(1 - \sum_{j=1}^{n-1} u_j \right) \in \Delta_n$. Define $\operatorname{pred} : \mathcal{C} \to [k]$ as

$$\operatorname{pred}(\mathbf{u}) \;=\; \min \left\{ t \in [k] : \mathbf{p}^{\mathbf{u}} \in Q_t^{\mathbf{L}} \right\} .$$

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

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

$$\mathbf{p}^\top \boldsymbol{\psi}(\mathbf{u}) \;=\; \sum_{j=1}^{n-1} \left(p_j (u_j - 1)^2 + (1 - p_j) u_j^2 \right) .$$

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

$$\inf_{\mathbf{u} \in \mathcal{C}} \mathbf{p}^\top \boldsymbol{\psi}(\mathbf{u}) \;=\; \mathbf{p}^\top \boldsymbol{\psi}(\mathbf{u}^*) \;=\; \sum_{j=1}^{n-1} p_j (1 - p_j) .$$

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

$$\operatorname{regret}_{\mathbf{p}}^{\mathbf{L}}(t) \;\stackrel{\triangle}{=}\; \mathbf{p}^\top \boldsymbol{\ell}_t - \min_{t' \in [k]} \mathbf{p}^\top \boldsymbol{\ell}_{t'} .$$

Clearly, $\operatorname{regret}_{\mathbf{p}}^{\mathbf{L}}(t) = 0 \Longleftrightarrow \mathbf{p} \in \mathcal{Q}_t^{\mathbf{L}}$. Note also that $\mathbf{p}^{\mathbf{u}^*} = \mathbf{p}$, and therefore $\operatorname{regret}_{\mathbf{p}}^{\mathbf{L}}(\operatorname{pred}(\mathbf{u}^*)) = 0$. Let

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

29

Then we have

$$\inf_{\mathbf{u}\in\mathcal{C}:\text{pred}(\mathbf{u})\notin\text{argmin}_t \mathbf{p}^\top\boldsymbol{\ell}_t} \mathbf{p}^\top\boldsymbol{\psi}(\mathbf{u}) \quad = \quad \inf_{\mathbf{u}\in\mathcal{C}:\text{regret}_\mathbf{p}^\mathbf{L}(\text{pred}(\mathbf{u}))\geq\epsilon} \mathbf{p}^\top\boldsymbol{\psi}(\mathbf{u}) \tag{20}$$

$$= \quad \inf_{\mathbf{u}\in\mathcal{C}:\text{regret}_\mathbf{p}^\mathbf{L}(\text{pred}(\mathbf{u}))\geq\text{regret}_\mathbf{p}^\mathbf{L}(\text{pred}(\mathbf{u}^*))+\epsilon} \mathbf{p}^\top\boldsymbol{\psi}(\mathbf{u}) . \tag{21}$$

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

$$\|\mathbf{u} - \mathbf{u}^*\| < \delta \implies \text{regret}_\mathbf{p}^\mathbf{L}(\text{pred}(\mathbf{u})) - \text{regret}_\mathbf{p}^\mathbf{L}(\text{pred}(\mathbf{u}^*)) < \epsilon .$$

This gives

$$\inf_{\mathbf{u}\in\mathcal{C}:\text{regret}_\mathbf{p}^\mathbf{L}(\text{pred}(\mathbf{u}))\geq\text{regret}_\mathbf{p}^\mathbf{L}(\text{pred}(\mathbf{u}^*))+\epsilon} \mathbf{p}^\top\boldsymbol{\psi}(\mathbf{u}) \quad \geq \quad \inf_{\mathbf{u}\in\mathcal{C}:\|\mathbf{u}-\mathbf{u}^*\|\geq\delta} \mathbf{p}^\top\boldsymbol{\psi}(\mathbf{u}) \tag{22}$$

$$> \quad \inf_{\mathbf{u}\in\mathcal{C}} \mathbf{p}^\top\boldsymbol{\psi}(\mathbf{u}) , \tag{23}$$

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

$$\inf_{\mathbf{u}\in\mathcal{C}:\text{pred}(\mathbf{u})\notin\text{argmin}_t \mathbf{p}^\top\boldsymbol{\ell}_t} \mathbf{p}^\top\boldsymbol{\psi}(\mathbf{u}) \quad > \quad \inf_{\mathbf{u}\in\mathcal{C}} \mathbf{p}^\top\boldsymbol{\psi}(\mathbf{u}) . \tag{24}$$

Since this holds for all $\mathbf{p} \in \Delta_n$, we have that $\boldsymbol{\psi}$ is $\mathbf{L}$-calibrated. ∎

7.3 Proof of Theorem 16

The proof will require the lemma below, which relates the feasible subspace dimensions of different trigger probability sets at points in their intersection; we will also make critical use of the notion of ϵ-subdifferentials of convex functions (Bertsekas et al., 2003), the main properties of which are also recalled below.

Lemma 23 *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 $\mathbf{p} \in \mathcal{Q}_{t_1}^\mathbf{L} \cap \mathcal{Q}_{t_2}^\mathbf{L}$),*

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

Proof (Proof of Lemma 23)

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

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

where $\mathbf{e}_n$ denotes the $n \times 1$ all ones vector. Moreover, we have $-\mathbf{p} < \mathbf{0}$, and $(\boldsymbol{\ell}_{t_1} - \boldsymbol{\ell}_t)^\top \mathbf{p} = 0$ iff $\mathbf{p} \in \mathcal{Q}_t^{\mathbf{L}}$. Let $\{t \in [k] : \mathbf{p} \in \mathcal{Q}_t^{\mathbf{L}}\} = \{\widetilde{t}_1, \ldots, \widetilde{t}_r\}$ for some $r \in [k]$. Then by Lemma 15, we have

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

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

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

where $\mathbf{A}_2 \in \mathbb{R}^{(r+1)\times n}$ is a matrix containing r rows of the form $(\boldsymbol{\ell}_{t_2} - \boldsymbol{\ell}_{\widetilde{t}_j})^\top, j \in [r]$ 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 $\boldsymbol{\ell}_{\widetilde{t}_1}, \ldots, \boldsymbol{\ell}_{\widetilde{t}_r}$. Thus both $\mathbf{A}_1$ and $\mathbf{A}_2$ have the same row space and hence the same null space and nullity, and therefore $\mu_{\mathcal{Q}_{t_1}^{\mathbf{L}}}(\mathbf{p}) = \mu_{\mathcal{Q}_{t_2}^{\mathbf{L}}}(\mathbf{p})$. ∎

ϵ-Subdifferentials of a Convex Function. For any $\epsilon > 0$, the ϵ-subdifferential of a convex function $\phi : \mathbb{R}^d \to \overline{\mathbb{R}}$ at a point $\mathbf{u}_0 \in \mathbb{R}^d$ is defined as follows (Bertsekas et al., 2003):

$$\partial_\epsilon \phi(\mathbf{u}_0) = \left\{ \mathbf{w} \in \mathbb{R}^d : \phi(\mathbf{u}) - \phi(\mathbf{u}_0) \geq \mathbf{w}^\top(\mathbf{u} - \mathbf{u}_0) - \epsilon \ \ \forall \mathbf{u} \in \mathbb{R}^d \right\}.$$

We recall some important properties of ϵ-subdifferentials below:

- $\mathbf{0} \in \partial_\epsilon \phi(\mathbf{u}_0) \iff \phi(\mathbf{u}_0) \leq \inf_{\mathbf{u} \in \mathbb{R}^d} \phi(\mathbf{u}) + \epsilon$.

- For any $\lambda > 0$, $\partial_\epsilon(\lambda \phi(\mathbf{u}_0)) = \lambda \partial_{(\epsilon/\lambda)} \phi(\mathbf{u}_0)$.

- If $\phi = \phi_1 + \ldots + \phi_n$ for some convex functions $\phi_i : \mathbb{R}^d \to \overline{\mathbb{R}}$, then

$$\partial_\epsilon \phi(\mathbf{u}_0) \subseteq \partial_\epsilon \phi_1(\mathbf{u}_0) + \ldots + \partial_\epsilon \phi_n(\mathbf{u}_0) \subseteq \partial_{n\epsilon} \phi(\mathbf{u}_0).$$

- $\epsilon_1 \leq \epsilon_2 \implies \partial_{\epsilon_1} \phi(\mathbf{u}_0) \subseteq \partial_{\epsilon_2} \phi(\mathbf{u}_0)$.

We are now ready to give the proof of the lower bound on the CC dimension.

Proof (Proof of Theorem 16)

Let $d \in \mathbb{Z}_+$ be such that there exists a convex set $\mathcal{C} \subseteq \mathbb{R}^d$ and surrogate loss $\boldsymbol{\psi} : \mathcal{C} \to \mathbb{R}_+^n$ such that $\boldsymbol{\psi}$ is $\mathbf{L}$-calibrated. We will show that $d \geq \|\mathbf{p}\|_0 - \mu_{\mathcal{Q}_t^{\mathbf{L}}}(\mathbf{p}) - 1$. We consider two cases:

Case 1: $\mathbf{p} \in \text{relint}(\Delta_n)$.

In this case $\|\mathbf{p}\|_0 = n$. We will show that there exist $\mathcal{H} \subseteq \Delta_n$ and $t_0 \in [k]$ satisfying the following three conditions:

(a) $\mathbf{p} \in \mathcal{H}$;
(b) $\mu_{\mathcal{H}}(\mathbf{p}) = n - d - 1$; and
(c) $\mathcal{H} \subseteq \mathcal{Q}_{t_0}^{\mathbf{L}}$.

Clearly, conditions (a) and (c) above imply $\mathbf{p} \in \mathcal{Q}_{t_0}^{\mathbf{L}}$. Conditions (b) and (c) will then give

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

Further, by Lemma 23, we will then have that

$$\mu_{\mathcal{Q}_t^{\mathbf{L}}}(\mathbf{p}) \ = \ \mu_{\mathcal{Q}_{t_0}^{\mathbf{L}}}(\mathbf{p}) \ \geq \ n - d - 1 \,,$$

thus proving the claim.

We now show how to construct $\mathcal{H}$ and t_0 satisfying the above conditions. Let $\{\mathbf{u}_m\}$ be a sequence in $\mathcal{C}$ such that

$$\mathbf{p}^{\top}\boldsymbol{\psi}(\mathbf{u}_m) \ \longrightarrow \ \inf_{\mathbf{z} \in \mathcal{S}_\psi} \mathbf{p}^{\top}\mathbf{z} = \inf_{\mathbf{u} \in \mathcal{C}} \mathbf{p}^{\top}\boldsymbol{\psi}(\mathbf{u}) \,.$$

Let

$$\epsilon_m = \mathbf{p}^{\top}\boldsymbol{\psi}(\mathbf{u}_m) - \inf_{\mathbf{u} \in \mathcal{C}} \mathbf{p}^{\top}\boldsymbol{\psi}(\mathbf{u}) \,.$$

Then clearly $\epsilon_m \to 0$. Now, for each m, we have

$$\begin{aligned}
\mathbf{0} \ &\in \ \partial_{\epsilon_m}(\mathbf{p}^{\top}\boldsymbol{\psi}(\mathbf{u}_m)) \\
&\subseteq \ \partial_{\epsilon_m}(p_1\psi_1(\mathbf{u}_m)) + \ldots + \partial_{\epsilon_m}(p_n\psi_n(\mathbf{u}_m)) \\
&= \ p_1\,\partial_{(\epsilon_m/p_1)}(\psi_1(\mathbf{u}_m)) + \ldots + p_n\,\partial_{(\epsilon_m/p_n)}(\psi_n(\mathbf{u}_m)) \,.
\end{aligned}$$

Therefore $\exists \mathbf{w}_y^m \in \partial_{(\epsilon_m/p_y)}(\psi_y(\mathbf{u}_m)) \ \forall y \in [n]$ such that

$$\sum_{y=1}^{n} p_y \mathbf{w}_y^m = \mathbf{0} \,.$$

Let

$$\mathbf{A}^m = \begin{bmatrix} \mathbf{w}_1^m \ldots \mathbf{w}_n^m \end{bmatrix} \in \mathbb{R}^{d \times n} \,,$$

and define

$$\begin{aligned}
\mathcal{H}_m \ &= \ \left\{ \mathbf{q} \in \Delta_n : \mathbf{A}^m \mathbf{q} = \mathbf{0} \right\} \\
&= \ \left\{ \mathbf{q} \in \mathbb{R}^n : \mathbf{A}^m \mathbf{q} = \mathbf{0} \,, \mathbf{e}_n^{\top}\mathbf{q} = 1 \,, -\mathbf{q} \leq \mathbf{0} \right\} \,,
\end{aligned}$$

where $\mathbf{e}_n$ denotes the $n \times 1$ all ones vector. Clearly, $\mathbf{p} \in \mathcal{H}_m$ and $-\mathbf{p} < \mathbf{0}$; therefore by Lemma 15, we have

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

This means that there exist $(n - d - 1)$ orthonormal vectors $\mathbf{v}_1^m, \ldots, \mathbf{v}_{n-d-1}^m \in \mathbb{R}^n$ whose span $\mathcal{V}^m = \text{span}(\{\mathbf{v}_1^m, \ldots, \mathbf{v}_{n-d-1}^m\})$ is contained in $\mathcal{F}_{\mathcal{H}_m}(\mathbf{p}) \cap (-\mathcal{F}_{\mathcal{H}_m}(\mathbf{p}))$. It can be verified that this in turn implies

$$\{\mathbf{p} + \mathbf{v} : \mathbf{v} \in \mathcal{V}^m\} \cap \Delta_n \ \subseteq \ \mathcal{H}_m \,.$$

Now, $\{(\mathbf{v}_1^m, \ldots, \mathbf{v}_{n-d-1}^m)\}$ is a bounded sequence in $(\mathbb{R}^n)^{n-d-1}$ and must therefore have a convergent subsequence, say with indices $r_1, r_2, \ldots$, converging to some limit point $(\mathbf{v}_1, \ldots, \mathbf{v}_{n-d-1}) \in (\mathbb{R}^n)^{n-d-1}$. It can be verified that $\mathbf{v}_1, \ldots, \mathbf{v}_{n-d-1}$ must also form an orthonormal set of vectors. Let $\mathcal{V} = \text{span}(\{\mathbf{v}_1, \ldots, \mathbf{v}_{n-d-1}\})$, and define

$$\mathcal{H} \; = \; \{\mathbf{p} + \mathbf{v} : \mathbf{v} \in \mathcal{V}\} \cap \Delta_n \,.$$

Clearly $\mathbf{p} \in \mathcal{H}$, and moreover, since $\mathbf{p} \in \text{relint}(\Delta_n)$, we have $\mu_{\mathcal{H}}(\mathbf{p}) = n - d - 1$, thus satisfying conditions (a) and (b) above.

For condition (c), we will show that $\mathcal{H} \subseteq \mathcal{N}^\psi(\{\mathbf{z}_m\})$, where $\mathbf{z}_m = \psi(\mathbf{u}_{r_m})$; the claim will then follow from Theorem 7. Consider any point $\mathbf{q} \in \mathcal{H}$. By construction of $\mathcal{H}$, we must have that $\mathbf{q}$ is the limit point of some convergent sequence $\{\mathbf{q}_m\}$ in Δ_n satisfying $\mathbf{q}_m \in \mathcal{H}_{r_m} \; \forall m$, i.e. $\mathbf{A}^{r_m} \mathbf{q}_m = \mathbf{0} \; \forall m$. Therefore for each m, we have

$$\mathbf{0} \; = \; \sum_{y=1}^n q_{m,y} \mathbf{w}_y^{r_m} \;\; \in \;\; \sum_{y=1}^n q_{m,y} \, \partial_{(\epsilon_{r_m}/p_y)}(\psi_y(\mathbf{u}_{r_m}))$$

$$= \;\; \sum_{y=1}^n \partial_{(\epsilon_{r_m} q_{m,y}/p_y)}(q_y^m \, \psi_y(\mathbf{u}_{r_m}))$$

$$\subseteq \;\; \sum_{y=1}^n \partial_{(\epsilon_{r_m}/p_{\min})}(q_{m,y} \, \psi_y(\mathbf{u}_{r_m}))$$

$$\subseteq \;\; \partial_{(n\epsilon_{r_m}/p_{\min})}(\mathbf{q}_m^\top \psi(\mathbf{u}_{r_m})) \,,$$

where $p_{\min} = \min_{y \in [n]} p_y > 0$ (since $\mathbf{p} \in \text{relint}(\Delta_n)$). This gives for each m:

$$\mathbf{q}_m^\top \mathbf{z}_m \; = \; \mathbf{q}_m^\top \psi(\mathbf{u}_{r_m}) \; \leq \; \inf_{\mathbf{u} \in \mathbb{R}^d} \mathbf{q}_m^\top \psi(\mathbf{u}) + \frac{n\epsilon_{r_m}}{p_{\min}} \; = \; \inf_{\mathbf{z} \in \mathcal{S}_\psi} \mathbf{q}_m^\top \mathbf{z} + \frac{n\epsilon_{r_m}}{p_{\min}} \,.$$

Taking limits as $m \to \infty$, we thus get

$$\lim_{m \to \infty} \mathbf{q}_m^\top \mathbf{z}_m \leq \lim_{m \to \infty} \inf_{\mathbf{z} \in \mathcal{S}_\psi} \mathbf{q}_m^\top \mathbf{z} \,. \tag{25}$$

Now, since $\{\mathbf{z}_m\}$ is bounded, we have

$$\lim_{m \to \infty} \mathbf{q}_m^\top \mathbf{z}_m = \lim_{m \to \infty} (\mathbf{q}_m - \mathbf{q})^\top \mathbf{z}_m + \lim_{m \to \infty} \mathbf{q}^\top \mathbf{z}_m = \lim_{m \to \infty} \mathbf{q}^\top \mathbf{z}_m \,. \tag{26}$$

Moreover, since the mapping $\mathbf{p} \mapsto \inf_{\mathbf{z} \in \mathcal{S}_\psi} \mathbf{p}^\top \mathbf{z}$ is continuous over its domain Δ_n (see Lemma 24), we have

$$\lim_{m \to \infty} \inf_{\mathbf{z} \in \mathcal{S}_\psi} \mathbf{q}_m^\top \mathbf{z} = \inf_{\mathbf{z} \in \mathcal{S}_\psi} \mathbf{q}^\top \mathbf{z} \,. \tag{27}$$

Putting together Equations (25), (26) and (27), we therefore get

$$\lim_{m \to \infty} \mathbf{q}^\top \mathbf{z}_m = \inf_{\mathbf{z} \in \mathcal{S}_\psi} \mathbf{q}^\top \mathbf{z} \,. \tag{28}$$

Thus $\mathbf{q} \in \mathcal{N}^\psi(\{\mathbf{z}_m\})$. Since $\mathbf{q}$ was an arbitrary point in $\mathcal{H}$, this gives $\mathcal{H} \subseteq \mathcal{N}^\psi(\{\mathbf{z}_m\})$. The claim follows.

Case 2: $\mathbf{p} \notin \mathrm{relint}(\Delta_n)$.

For each $\mathbf{b} \in \{0,1\}^n \setminus \{\mathbf{0}\}$, define

$$\mathcal{P}^{\mathbf{b}} = \left\{ \mathbf{q} \in \Delta_n : q_y > 0 \iff b_y = 1 \right\}.$$

Clearly, the set $\{\mathcal{P}^{\mathbf{b}} : \mathbf{b} \in \{0,1\}^n \setminus \{\mathbf{0}\}\}$ forms a partition of Δ_n. Moreover, for $\mathbf{b} = \mathbf{e}_n$ (the $n \times 1$ all ones vector), we have

$$\mathcal{P}^{\mathbf{e}_n} = \left\{ \mathbf{q} \in \Delta_n : q_y > 0 \; \forall y \in [n] \right\} = \mathrm{relint}(\Delta_n).$$

Therefore we have $\mathbf{p} \in \mathcal{P}^{\mathbf{b}}$ for some $\mathbf{b} \in \{0,1\}^n \setminus \{\mathbf{0}, \mathbf{e}_n\}$, with $\|\mathbf{p}\|_0 = \|\mathbf{b}\|_0$. Now, define $\boldsymbol{\psi}^{\mathbf{b}} : \mathcal{C} \to \mathbb{R}_+^{\|\mathbf{b}\|_0}$, $\mathbf{L}^{\mathbf{b}} \in \mathbb{R}_+^{\|\mathbf{b}\|_0 \times k}$, and $\mathbf{p}^{\mathbf{b}} \in \Delta_{\|\mathbf{b}\|_0}$ as projections of $\boldsymbol{\psi}$, $\mathbf{L}$ and $\mathbf{p}$ onto the $\|\mathbf{b}\|_0$ coordinates $y : b_y = 1$, so that $\boldsymbol{\psi}^{\mathbf{b}}(\mathbf{u})$ contains the elements of $\boldsymbol{\psi}(\mathbf{u})$ corresponding to coordinates $y : b_y = 1$, the columns $\boldsymbol{\ell}_t^{\mathbf{b}}$ of $\mathbf{L}^{\mathbf{b}}$ contain the elements of the columns $\boldsymbol{\ell}_t$ of $\mathbf{L}$ corresponding to the same coordinates $y : b_y = 1$, and similarly, $\mathbf{p}^{\mathbf{b}}$ contains the strictly positive elements of $\mathbf{p}$. Since $\boldsymbol{\psi}$ is $\mathbf{L}$-calibrated, and therefore in particular is calibrated w.r.t. $\mathbf{L}$ over $\{\mathbf{q} \in \Delta_n : q_y = 0 \; \forall y : b_y = 0\}$, we have that $\boldsymbol{\psi}^{\mathbf{b}}$ is $\mathbf{L}^{\mathbf{b}}$-calibrated (over $\Delta_{\|\mathbf{b}\|_0}$). Moreover, by construction, we have $\mathbf{p}^{\mathbf{b}} \in \mathrm{relint}(\Delta_{\|\mathbf{b}\|_0})$. Therefore by Case 1 above, we have

$$d \geq \|\mathbf{b}\|_0 - \mu_{\mathcal{Q}_t^{\mathbf{L}^{\mathbf{b}}}}(\mathbf{p}^{\mathbf{b}}) - 1.$$

The claim follows since $\mu_{\mathcal{Q}_t^{\mathbf{L}^{\mathbf{b}}}}(\mathbf{p}^{\mathbf{b}}) \leq \mu_{\mathcal{Q}_t^{\mathbf{L}}}(\mathbf{p})$.

∎

7.4 Proof of Proposition 20

Proof We will establish $\mathrm{rank}(\widetilde{\mathbf{L}}^{\mathrm{PD}}) \geq \frac{r(r-1)}{2}$ by showing the existence of $\frac{r(r-1)}{2}$ linearly independent rows in $\widetilde{\mathbf{L}}^{\mathrm{PD}}$; the claim will then follow by combining this with the previously stated upper bound $\mathrm{rank}(\widetilde{\mathbf{L}}^{\mathrm{PD}}) \leq \frac{r(r-1)}{2}$.

Consider the $\frac{r(r-1)}{2}$ rows of $\widetilde{\mathbf{L}}^{\mathrm{PD}}$ corresponding to graphs consisting of single directed edges (i,j) with $i < j$. We claim these rows are linearly independent. To see this, suppose for the sake of contradiction that this is not the case. Then one of these rows, say the row corresponding to the graph with directed edge (a,b) for some $a, b \in [r], a < b$, can be written as a linear combination of the other rows:

$$\widetilde{\ell}^{\mathrm{PD}}((a,b),\sigma) = \sum_{1 \leq i < j \leq r,\, (i,j) \neq (a,b)} c_{ij}\, \widetilde{\ell}^{\mathrm{PD}}((i,j),\sigma) \qquad \forall \sigma \in \Pi_r, \tag{29}$$

for some coefficients $c_{ij} \in \mathbb{R}$. Now consider two permutations $\sigma, \sigma' \in \Pi_r$ such that

$$\begin{aligned}
\sigma(a) &= \sigma'(b) \\
\sigma(b) &= \sigma'(a) \\
\sigma(i) &= \sigma'(i) \; \forall i \neq a, b.
\end{aligned}$$

Then applying Eq. (29) to these two permutations gives

$$\widetilde{\ell}^{\mathrm{PD}}((a,b),\sigma) \;=\; \widetilde{\ell}^{\mathrm{PD}}((a,b),\sigma')\,.$$

However from the definition of $\widetilde{\mathbf{L}}^{\mathrm{PD}}$ it is easy to verify that the columns corresponding to these two permutations have identical entries in all rows except for the row corresponding to the graph (a,b), giving

$$
\begin{aligned}
\widetilde{\ell}^{\mathrm{PD}}((i,j),\sigma) &= \widetilde{\ell}^{\mathrm{PD}}((i,j),\sigma') && \forall i < j, (i,j) \neq (a,b)\,; \\
\widetilde{\ell}^{\mathrm{PD}}((a,b),\sigma) &\neq \widetilde{\ell}^{\mathrm{PD}}((a,b),\sigma')\,.
\end{aligned}
$$

This yields a contradiction, and therefore we must have that the $\frac{r(r-1)}{2}$ rows above are linearly independent, giving

$$\mathrm{rank}(\widetilde{\mathbf{L}}^{\mathrm{PD}}) \geq \frac{r(r-1)}{2}\,.$$

The claim follows. ∎

7.5 Proof of Proposition 21

Proof From Eq. (12), we have that $\mathbf{L}^{\mathrm{MAP}} \in \mathbb{R}^{(2^r-1)\times r!}$ can be written as

$$\mathbf{L}^{\mathrm{MAP}} = \mathbf{e}_{(2^r-1)}\mathbf{e}_{r!}^{\top} - \mathbf{A}\mathbf{B}\,,$$

where $\mathbf{e}_{(2^r-1)}$ and $\mathbf{e}_{r!}$ denote the $(2^r-1)\times 1$ and $r!\times 1$ all ones vectors, respectively, and where $\mathbf{A}\in\mathbb{R}^{(2^r-1)\times\frac{r(r+1)}{2}}$ and $\mathbf{B}\in\mathbb{R}^{\frac{r(r+1)}{2}\times r!}$ are given by

$$
\begin{aligned}
A_{\mathbf{y},(i,j)} &= \frac{1}{\|\mathbf{y}\|_1} y_i y_j && \forall \mathbf{y}\in\{0,1\}^r\setminus\{\mathbf{0}\},\ i,j\in[r]:i\leq j\,, \\
B_{(i,j),\sigma} &= \frac{1}{\max(\sigma(i),\sigma(j))} && \forall i,j\in[r]:i\leq j,\ \sigma\in\Pi_r\,.
\end{aligned}
$$

We will show that

$$\mathrm{rank}(\mathbf{A}) \;\geq\; \frac{r(r+1)}{2} - 1 \tag{30}$$

and

$$\mathrm{rank}(\mathbf{B}) \;\geq\; \frac{r(r-1)}{2}\,. \tag{31}$$

The result will then follow, since we will then have

$$
\begin{aligned}
\mathrm{rank}(\mathbf{L}^{\mathrm{MAP}}) &= \mathrm{rank}\big(\mathbf{e}_{(2^r-1)}\mathbf{e}_{r!}^{\top} - \mathbf{A}\mathbf{B}\big) \\
&\geq \mathrm{rank}(\mathbf{A}\mathbf{B}) - 1 \\
&\geq \mathrm{rank}(\mathbf{B}) - 2\,, && \text{since } \mathbf{A} \text{ is away from full (column) rank by at most 1} \\
&\geq \frac{r(r-1)}{2} - 2\,.
\end{aligned}
$$

To see why Eq. (30) is true, consider the 2^r vectors $\mathbf{v}^\alpha \in \mathbb{R}^{2^r}$ defined as

$$v_{\mathbf{y}}^\alpha = \prod_{i \in \alpha} y_i \quad \forall \alpha \subseteq [r], \mathbf{y} \in \{0,1\}^r .$$

It is easy to see that these vectors form a basis in $\mathbb{R}^{2^r}$. The columns of $\mathbf{A}$ can be obtained from the $\frac{r(r+1)}{2}$ vectors $\mathbf{v}^\alpha$ corresponding to subsets $\alpha \subseteq [r]$ of sizes 1 and 2, by removing the element corresponding to $\mathbf{y} = \mathbf{0}$ and dividing all other rows corresponding to $\mathbf{y} \in \{0,1\}^r \backslash \{\mathbf{0}\}$ by $\|\mathbf{y}\|_1$. This establishes the lower bound on rank$(\mathbf{A})$ in Eq. (30).

To see why Eq. (31) is true, let us make the dependence of $\mathbf{B}$ on r explicit by denoting $\mathbf{B}_r = \mathbf{B}$, and observe that $\mathbf{B}_r$ can be decomposed as

$$\mathbf{B}_r = \begin{bmatrix} \mathbf{B}_{r-1} & \mathbf{D} \\ \mathbf{C} & \mathbf{E} \end{bmatrix} ,$$

where the sub-matrix $\mathbf{B}_{r-1} \in \mathbb{R}^{\frac{r(r-1)}{2} \times (r-1)!}$ is obtained by taking the $\frac{r(r-1)}{2}$ rows (i,j) in $\mathbf{B}_r$ with $i \leq j < r$ and the $(r-1)!$ columns σ in $\mathbf{B}_r$ with $\sigma(r) = r$:

	$\Upsilon = \{\sigma \in \Pi_r : \sigma(r) = r\}$	$\Omega = \{\sigma \in \Pi_r : \sigma(r) \neq r\}$
$\Gamma = \{(i,j) \in [r] \times [r] : i \leq j < r\}$	$\mathbf{B}_{r-1}$	$\mathbf{D}$
$\Lambda = \{(i,j) \in [r] \times [r] : i \leq j = r\}$	$\mathbf{C}$	$\mathbf{E}$

Consider the matrix $\mathbf{C} \in \mathbb{R}^{r \times (r-1)!}$. Each entry in this matrix has the form $\frac{1}{\max(\sigma(i),\sigma(j))}$ with $i \leq j = r$ and $\sigma(r) = r$. Thus all entries in $\mathbf{C}$ are equal to $\frac{1}{r}$ and rank$(\mathbf{C}) = 1$.

Next, consider the matrix $\mathbf{E} \in \mathbb{R}^{r \times (r!-(r-1)!)}$. We will show that there are $r - 1$ linearly independent columns in $\mathbf{E}$. In particular, consider any permutations $\sigma^1, \sigma^2, \ldots, \sigma^{r-1}$ in the set Ω such that $\sigma^j(j) = r$ and $\sigma^j(r) = r - 1$ (such permutations clearly exist). The sub-matrix of $\mathbf{E}$ corresponding to these columns is given by

	$\begin{array}{c}\sigma^1(1) = r \\ \sigma^1(r) = r-1\end{array}$	$\begin{array}{c}\sigma^2(2) = r \\ \sigma^2(r) = r-1\end{array}$	$\ldots$	$\begin{array}{c}\sigma^{r-1}(r-1) = r \\ \sigma^{r-1}(r) = r-1\end{array}$
$(1,r)$	$\frac{1}{r}$	$\frac{1}{r-1}$	$\ldots$	$\frac{1}{r-1}$
$(2,r)$	$\frac{1}{r-1}$	$\frac{1}{r}$	$\ldots$	$\frac{1}{r-1}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$(r-1,r)$	$\frac{1}{r-1}$	$\frac{1}{r-1}$	$\ldots$	$\frac{1}{r}$
(r,r)	$\frac{1}{r-1}$	$\frac{1}{r-1}$	$\ldots$	$\frac{1}{r-1}$

Thus excluding the last row of $\mathbf{E}$, one gets a square $(r-1) \times (r-1)$ matrix with diagonal entries equal to $\frac{1}{r}$ and off-diagonal entries equal to $\frac{1}{r-1}$. The last row of $\mathbf{E}$ has all entries equal to $\frac{1}{r-1}$. Clearly, this gives rank$(\mathbf{E}) = r - 1$. Moreover, the span of the $r - 1$ column vectors of $\mathbf{E}$ does not intersect with column space of $\mathbf{C}$ non-trivially, since it does not contain the all ones vector. This implies that the $r - 1$ columns of $\mathbf{B}_r$ corresponding to the permutations $\sigma^1, \sigma^2, \ldots, \sigma^{r-1} \in \Omega$ (which yield the linearly independent columns of $\mathbf{E}$),

together with the columns of $\mathbf{B}_r$ corresponding to permutations $\sigma \in \Upsilon$ that yield linearly independent columns of $\mathbf{B}_{r-1}$, are all linearly independent. Therefore, we have

$$\text{rank}(\mathbf{B}_r) \geq \text{rank}(\mathbf{B}_{r-1}) + r - 1 \,.$$

Trivially, $\text{rank}(\mathbf{B}_1) \geq 0$. Expanding the above recursion therefore gives

$$\text{rank}(\mathbf{B}_r) \geq \frac{r(r-1)}{2} \,.$$

This establishes the lower bound on $\text{rank}(\mathbf{B})$ in Eq. (31). ∎

Acknowledgments

Appendix A. Proofs of Lemma 2 and Theorem 3

A.1 Proof of Lemma 2

Proof Let $\boldsymbol{\psi} : \mathcal{C} \to \mathbb{R}_+^n$. We will show that $\exists$ pred $: \mathcal{C} \to [k]$ satisfying the condition in Definition 1 if and only if $\exists$ pred$' : \mathcal{S}_\psi \to [k]$ satisfying the stated condition.

('if' direction) First, suppose $\exists$ pred$' : \mathcal{S}_\psi \to [k]$ such that

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

Define pred $: \mathcal{C} \to [k]$ as follows:

$$\text{pred}(\mathbf{u}) = \text{pred}'(\boldsymbol{\psi}(\mathbf{u})) \quad \forall \mathbf{u} \in \mathcal{C} \,.$$

Then for all $\mathbf{p} \in \mathcal{P}$, we have

$$
\begin{aligned}
\inf_{\mathbf{u} \in \mathcal{C} : \text{pred}(\mathbf{u}) \notin \text{argmin}_t \mathbf{p}^\top \boldsymbol{\ell}_t} \mathbf{p}^\top \boldsymbol{\psi}(\mathbf{u}) &= \inf_{\mathbf{z} \in \mathcal{R}_\psi : \text{pred}'(\mathbf{z}) \notin \text{argmin}_t \mathbf{p}^\top \boldsymbol{\ell}_t} \mathbf{p}^\top \mathbf{z} \\
&\geq \inf_{\mathbf{z} \in \mathcal{S}_\psi : \text{pred}'(\mathbf{z}) \notin \text{argmin}_t \mathbf{p}^\top \boldsymbol{\ell}_t} \mathbf{p}^\top \mathbf{z} \\
&> \inf_{\mathbf{z} \in \mathcal{S}_\psi} \mathbf{p}^\top \mathbf{z} \\
&= \inf_{\mathbf{u} \in \mathcal{C}} \mathbf{p}^\top \boldsymbol{\psi}(\mathbf{u}) \,.
\end{aligned}
$$

Thus $\boldsymbol{\psi}$ is $(\mathbf{L}, \mathcal{P})$-calibrated.

('only if' direction) Conversely, suppose $\boldsymbol{\psi}$ is $(\mathbf{L}, \mathcal{P})$-calibrated, so that $\exists$ pred $: \mathcal{C} \to [k]$ such that

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

By Caratheodory's Theorem (e.g. see Bertsekas et al. (2003)), we have that every $\mathbf{z} \in \mathcal{S}_\psi$ can be expressed as a convex combination of at most $n+1$ points in $\mathcal{R}_\psi$, i.e. for every $\mathbf{z} \in \mathcal{S}_\psi$, $\exists \boldsymbol{\alpha} \in \Delta_{n+1}, \mathbf{u}_1, \ldots, \mathbf{u}_{n+1} \in \mathcal{C}$ such that $\mathbf{z} = \sum_{j=1}^{n+1} \alpha_j \boldsymbol{\psi}(\mathbf{u}_j)$; w.l.o.g., we can assume $\alpha_1 \geq \frac{1}{n+1}$. For each $\mathbf{z} \in \mathcal{S}_\psi$, arbitrarily fix a unique such convex combination, i.e. fix $\boldsymbol{\alpha}^\mathbf{z} \in \Delta_{n+1}, \mathbf{u}_1^\mathbf{z}, \ldots, \mathbf{u}_{n+1}^\mathbf{z} \in \mathcal{C}$ with $\alpha_1^\mathbf{z} \geq \frac{1}{n+1}$ such that

$$\mathbf{z} = \sum_{j=1}^{n+1} \alpha_j^\mathbf{z} \boldsymbol{\psi}(\mathbf{u}_j^\mathbf{z}) .$$

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

$$\mathrm{pred}'(\mathbf{z}) = \mathrm{pred}(\mathbf{u}_1^\mathbf{z}) \quad \forall \mathbf{z} \in \mathcal{S}_\psi .$$

Then for any $\mathbf{p} \in \mathcal{P}$, we have

$$
\begin{aligned}
\inf_{\mathbf{z} \in \mathcal{S}_\psi : \mathrm{pred}'(\mathbf{z}) \notin \mathrm{argmin}_t \mathbf{p}^\top \boldsymbol{\ell}_t} \mathbf{p}^\top \mathbf{z} \;&=\; \inf_{\mathbf{z} \in \mathcal{S}_\psi : \mathrm{pred}(\mathbf{u}_1^\mathbf{z}) \notin \mathrm{argmin}_t \mathbf{p}^\top \boldsymbol{\ell}_t} \sum_{j=1}^{n+1} \alpha_j^\mathbf{z} \mathbf{p}^\top \boldsymbol{\psi}(\mathbf{u}_j^\mathbf{z}) \\
&\geq\; \inf_{\boldsymbol{\alpha} \in \Delta_{n+1}, \mathbf{u}_1, \ldots, \mathbf{u}_{n+1} \in \mathcal{C} : \alpha_1 \geq \frac{1}{n+1}, \mathrm{pred}(\mathbf{u}_1) \notin \mathrm{argmin}_t \mathbf{p}^\top \boldsymbol{\ell}_t} \sum_{j=1}^{n+1} \alpha_j \mathbf{p}^\top \boldsymbol{\psi}(\mathbf{u}_j) \\
&\geq\; \inf_{\boldsymbol{\alpha} \in \Delta_{n+1} : \alpha_1 \geq \frac{1}{n+1}} \sum_{j=1}^{n+1} \inf_{\mathbf{u}_j \in \mathcal{C} : \mathrm{pred}(\mathbf{u}_1) \notin \mathrm{argmin}_t \mathbf{p}^\top \boldsymbol{\ell}_t} \alpha_j \mathbf{p}^\top \boldsymbol{\psi}(\mathbf{u}_j) \\
&\geq\; \inf_{\alpha_1 \in [\frac{1}{n+1}, 1]} \alpha_1 \inf_{\mathbf{u} \in \mathcal{C} : \mathrm{pred}(\mathbf{u}) \notin \mathrm{argmin}_t \mathbf{p}^\top \boldsymbol{\ell}_t} \mathbf{p}^\top \boldsymbol{\psi}(\mathbf{u}) + (1 - \alpha_1) \sum_{j=2}^{n+1} \inf_{\mathbf{u} \in \mathcal{C}} \mathbf{p}^\top \boldsymbol{\psi}(\mathbf{u}) \\
&>\; \inf_{\mathbf{u} \in \mathcal{C}} \mathbf{p}^\top \boldsymbol{\psi}(\mathbf{u}) \\
&=\; \inf_{\mathbf{z} \in \mathcal{S}_\psi} \mathbf{p}^\top \mathbf{z} .
\end{aligned}
$$

Thus pred' satisfies the stated condition. $\blacksquare$

A.2 Proof of Theorem 3

The proof is similar to that for the multiclass 0-1 loss given by Tewari and Bartlett (2007). We will make use of the following two lemmas; the first is a straightforward generalization of a similar lemma in (Tewari and Bartlett, 2007), and the second follows directly from Lemma 2.

Lemma 24 *The map* $\mathbf{p} \mapsto \inf_{\mathbf{z} \in \mathcal{S}_\psi} \mathbf{p}^\top \mathbf{z}$ *is continuous over* Δ_n.

Lemma 25 *Let* $\mathbf{L} \in \mathbb{R}_+^{n \times k}$. *A surrogate* $\psi : \mathcal{C} \to \mathbb{R}^n$ *is* $\mathbf{L}$-*calibrated if and only if there exists a function* $\mathrm{pred}' : \mathcal{S}_\psi \to [k]$ *such that the following holds: for all* $\mathbf{p} \in \Delta_n$ *and all sequences* $\{\mathbf{z}_m\}$ *in* $\mathcal{S}_\psi$ *such that* $\lim_{m \to \infty} \mathbf{p}^\top \mathbf{z}_m = \inf_{\mathbf{z} \in \mathcal{S}_\psi} \mathbf{p}^\top \mathbf{z}$, *we have* $\mathbf{p}^\top \boldsymbol{\ell}_{\mathrm{pred}'(\mathbf{z}_m)} = \min_{t \in [k]} \mathbf{p}^\top \boldsymbol{\ell}_t$ *for all large enough* m.

Proof (Proof of Theorem 3)

Let $\boldsymbol{\psi} : \mathcal{C} \rightarrow \mathbb{R}_+^n$.

('only if' direction) First, suppose $\boldsymbol{\psi}$ is **L**-calibrated. Then by Lemma 2, $\exists \, \text{pred}' : \mathcal{S}_\psi \rightarrow [k]$ such that

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

Now, for each $\epsilon > 0$, define

$$H(\epsilon) = \inf_{\mathbf{p} \in \Delta_n, \mathbf{z} \in \mathcal{S}_\psi : \mathbf{p}^\top \boldsymbol{\ell}_{\text{pred}'(\mathbf{z})} - \min_{t \in [k]} \mathbf{p}^\top \boldsymbol{\ell}_t \geq \epsilon} \left\{ \mathbf{p}^\top \mathbf{z} - \inf_{\mathbf{z} \in \mathcal{S}_\psi} \mathbf{p}^\top \mathbf{z} \right\} .$$

We claim that $H(\epsilon) > 0 \; \forall \epsilon > 0$. Assume for the sake of contradiction that $\exists \epsilon > 0$ for which $H(\epsilon) = 0$. Then there must exist a sequence $(\mathbf{p}_m, \mathbf{z}_m)$ in $\Delta_n \times \mathcal{S}_\psi$ such that

$$\mathbf{p}_m^\top \boldsymbol{\ell}_{\text{pred}'(\mathbf{z}_m)} - \min_{t \in [k]} \mathbf{p}_m^\top \boldsymbol{\ell}_t \geq \epsilon \quad \forall m \tag{32}$$

and

$$\mathbf{p}_m^\top \mathbf{z}_m - \inf_{\mathbf{z} \in \mathcal{S}_\psi} \mathbf{p}_m^\top \mathbf{z} \to 0 \, . \tag{33}$$

Since $\mathbf{p}_m$ come from a compact set, we can choose a convergent subsequence (which we still call $\{\mathbf{p}_m\}$), say with limit $\mathbf{p}$. Then by Lemma 24, we have $\inf_{\mathbf{z} \in \mathcal{S}_\psi} \mathbf{p}_m^\top \mathbf{z} \longrightarrow \inf_{\mathbf{z} \in \mathcal{S}_\psi} \mathbf{p}^\top \mathbf{z}$, and therefore by Eq. (33), we get

$$\mathbf{p}_m^\top \mathbf{z}_m \; \longrightarrow \; \inf_{\mathbf{z} \in \mathcal{S}_\psi} \mathbf{p}^\top \mathbf{z} \, .$$

Now we show that $\mathbf{z}_m$ is a sequence such that $\mathbf{p}^\top \mathbf{z}_m \longrightarrow \inf_{\mathbf{z} \in \mathcal{S}_\psi} \mathbf{p}^\top \mathbf{z}$. Without loss of generality, we assume that the first a coordinates of $\mathbf{p}$ are non-zero and the rest are zero. Hence the first a coordinates of $\mathbf{z}_m$ are bounded for sufficiently large m, and we have

$$\limsup_m \mathbf{p}^\top \mathbf{z}_m = \limsup_m \sum_{y=1}^a p_{m,y} z_{m,y} \leq \lim_{m \to \infty} \mathbf{p}_m^\top \mathbf{z}_m = \inf_{\mathbf{z} \in \mathcal{S}_\psi} \mathbf{p}^\top \mathbf{z} \, .$$

By Lemma 25, we therefore have $\mathbf{p}^\top \boldsymbol{\ell}_{\text{pred}'(\mathbf{z}_m)} = \min_{t \in [k]} \mathbf{p}^\top \boldsymbol{\ell}_t$ for all large enough m, which contradicts Eq. (32) as $\mathbf{p}_m$ converges to $\mathbf{p}$. Thus we must have $H(\epsilon) > 0 \; \forall \epsilon > 0$; the rest of the proof then follows from (Zhang, 2004a).

('if' direction)

Conversely, suppose $\boldsymbol{\psi}$ is not **L**-calibrated. Consider any $\text{pred} : \mathcal{C} \rightarrow [k]$. Then $\exists \mathbf{p} \in \Delta_n$ such that

$$\inf_{\mathbf{u} \in \mathcal{C} : \text{pred}(\mathbf{u}) \notin \text{argmin}_t \, \mathbf{p}^\top \boldsymbol{\ell}_t} \mathbf{p}^\top \boldsymbol{\psi}(\mathbf{u}) \; = \; \inf_{\mathbf{u} \in \mathcal{C}} \mathbf{p}^\top \boldsymbol{\psi}(\mathbf{u}) \, .$$

In particular, this means there exists a sequence of points $\{\mathbf{u}_m\}$ in $\mathcal{C}$ such that

$$\text{pred}(\mathbf{u}_m) \notin \text{argmin}_t \, \mathbf{p}^\top \boldsymbol{\ell}_t \quad \forall m$$

and

$$\mathbf{p}^\top \boldsymbol{\psi}(\mathbf{u}_m) \longrightarrow \inf_{\mathbf{u} \in \mathcal{C}} \mathbf{p}^\top \boldsymbol{\psi}(\mathbf{u}) .$$

Now consider a data distribution $D = D_X \times D_{Y|X}$ on $\mathcal{X} \times [n]$, with D_X being a point mass at some $x \in \mathcal{X}$ and $D_{Y|X=x} = \mathbf{p}$. Let $\mathbf{f}_m : \mathcal{X} \to \mathcal{C}$ be any sequence of functions satisfying $\mathbf{f}_m(x) = \mathbf{u}_m \ \forall m$. Then we have

$$\mathrm{er}_D^\psi[\mathbf{f}_m] = \mathbf{p}^\top \boldsymbol{\psi}(\mathbf{u}_m) ; \quad \mathrm{er}_D^{\psi,*} = \inf_{\mathbf{u} \in \mathcal{C}} \mathbf{p}^\top \boldsymbol{\psi}(\mathbf{u})$$

and

$$\mathrm{er}_D^{\mathbf{L}}[\mathrm{pred} \circ \mathbf{f}_m] = \mathbf{p}^\top \boldsymbol{\ell}_{\mathrm{pred}(\mathbf{u}_m)} ; \quad \mathrm{er}_D^{\mathbf{L},*} = \min_t \mathbf{p}^\top \boldsymbol{\ell}_t .$$

This gives

$$\mathrm{er}_D^\psi[\mathbf{f}_m] \longrightarrow \mathrm{er}_D^{\psi,*}$$

but

$$\mathrm{er}_D^{\mathbf{L}}[\mathrm{pred} \circ \mathbf{f}_m] \not\longrightarrow \mathrm{er}_D^{\mathbf{L},*} .$$

This completes the proof. ∎

Appendix B. Calculation of Trigger Probability Sets for Figure 2

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

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

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

By symmetry,

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

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

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

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

By symmetry,

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

Finally,

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

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

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

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

By symmetry,

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

Finally,

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

Appendix C. Calculation of Positive Normal Sets for Figure 3

For $n = 3$, the Crammer-Singer surrogate $\boldsymbol{\psi}^{\mathrm{CS}} : \mathbb{R}^3 \to \mathbb{R}^3_+$ is given by

$$
\begin{aligned}
\psi_1^{\mathrm{CS}}(\mathbf{u}) &= \max(1 + u_2 - u_1, 1 + u_3 - u_1, 0) \\
\psi_2^{\mathrm{CS}}(\mathbf{u}) &= \max(1 + u_1 - u_2, 1 + u_3 - u_2, 0) \\
\psi_3^{\mathrm{CS}}(\mathbf{u}) &= \max(1 + u_1 - u_3, 1 + u_2 - u_3, 0) \quad \forall \mathbf{u} \in \mathbb{R}^3 .
\end{aligned}
$$

Clearly, $\boldsymbol{\psi}^{\mathrm{CS}}$ is a convex function. Let $\mathbf{u}_1 = (1,0,0)^\top$, $\mathbf{u}_2 = (0,1,0)^\top$, $\mathbf{u}_3 = (0,0,1)^\top$, $\mathbf{u}_4 = (0,0,0)^\top$, and let

$$
\begin{aligned}
\mathbf{z}_1 &= \boldsymbol{\psi}^{\mathrm{CS}}(\mathbf{u}_1) = (0,2,2)^\top \\
\mathbf{z}_2 &= \boldsymbol{\psi}^{\mathrm{CS}}(\mathbf{u}_2) = (2,0,2)^\top \\
\mathbf{z}_3 &= \boldsymbol{\psi}^{\mathrm{CS}}(\mathbf{u}_3) = (2,2,0)^\top \\
\mathbf{z}_4 &= \boldsymbol{\psi}^{\mathrm{CS}}(\mathbf{u}_4) = (1,1,1)^\top .
\end{aligned}
$$

We apply Lemma 9 to compute the positive normal sets of $\boldsymbol{\psi}^{\mathrm{CS}}$ at the 4 points $\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{z}_4$ above. In particular, to see that $\mathbf{z}_4$ satisfies the conditions of Lemma 9, note that by Danskin's Theorem (Bertsekas et al., 2003), we have that

$$
\partial \psi_1^{\mathrm{CS}}(\mathbf{u}_4) = \mathrm{conv}\left(\begin{bmatrix} -1 \\ +1 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \\ +1 \end{bmatrix} \right) ;
$$

$$
\partial \psi_2^{\mathrm{CS}}(\mathbf{u}_4) = \mathrm{conv}\left(\begin{bmatrix} +1 \\ -1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \\ +1 \end{bmatrix} \right) ;
$$

$$
\partial \psi_3^{\mathrm{CS}}(\mathbf{u}_4) = \mathrm{conv}\left(\begin{bmatrix} +1 \\ 0 \\ -1 \end{bmatrix}, \begin{bmatrix} 0 \\ +1 \\ -1 \end{bmatrix} \right) .
$$

We can therefore use Lemma 9 to compute $\mathcal{N}^{\mathrm{CS}}(\mathbf{z}_4)$. Here $s = 6$, and

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

By Lemma 9 (and some algebra), this gives

$$
\begin{aligned}
\mathcal{N}^{\mathrm{CS}}(\mathbf{z}_4) &= \big\{ \mathbf{p} \in \Delta_3 : \mathbf{p} = (q_1 + q_2, q_3 + q_4, q_5 + q_6) \text{ for some } \mathbf{q} \in \Delta_6, \\
& \qquad q_1 + q_2 = q_3 + q_5, \ q_3 + q_4 = q_1 + q_6, \ q_5 + q_6 = q_2 + q_4 \big\} \\
&= \big\{ \mathbf{p} \in \Delta_3 : p_1 \le \tfrac{1}{2}, p_2 \le \tfrac{1}{2}, p_3 \le \tfrac{1}{2} \big\} .
\end{aligned}
$$

It is easy to see that $\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3$ also satisfy the conditions of Lemma 9; similar computations then yield

$$
\begin{aligned}
\mathcal{N}^{\mathrm{CS}}(\mathbf{z}_1) &= \big\{ \mathbf{p} \in \Delta_3 : p_1 \ge \tfrac{1}{2} \big\} \\
\mathcal{N}^{\mathrm{CS}}(\mathbf{z}_2) &= \big\{ \mathbf{p} \in \Delta_3 : p_2 \ge \tfrac{1}{2} \big\} \\
\mathcal{N}^{\mathrm{CS}}(\mathbf{z}_3) &= \big\{ \mathbf{p} \in \Delta_3 : p_3 \ge \tfrac{1}{2} \big\} .
\end{aligned}
$$

References

Shivani Agarwal. Surrogate regret bounds for bipartite ranking via strongly proper losses. *Journal of Machine Learning Research*, 15:1653–1674, 2014.

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

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

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.

Clément Calauzènes, Nicolas Usunier, and Patrick Gallinari. On the (non-)existence of convex, calibrated surrogate losses for ranking. In *Neural Information Processing Systems*, 2012.

Stéphan Clémençon and Nicolas Vayatis. Ranking the best instances. *Journal of Machine Learning Research*, 8:2671–2699, 2007.

Stéphan Clémençon, Gábor Lugosi, and Nicolas Vayatis. Ranking and empirical minimization of U-statistics. *Annals of Statistics*, 36:844–874, 2008.

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

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

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

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.

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

Kalervo Järvelin and Jaana Kekäläinen. Ir evaluation methods for retrieving highly relevant documents. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2000.

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

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

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

Yoonkyung Lee, Yi Lin, and Grace Wahba. Multicategory support vector machines: Theory and application to the classification of microarray data. *Journal of the American Statistical Association*, 99(465):67–81, 2004.

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

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

Bernardo Á. Pires, Csaba Szepesvari, and Mohammad Ghavamzadeh. Cost-sensitive multiclass classification risk bounds. In *International Conference on Machine Learning*, 2013.

Harish G. Ramaswamy and Shivani Agarwal. Classification calibration dimension for general multiclass losses. In *Neural Information Processing Systems*, 2012.

Harish G. Ramaswamy, Shivani Agarwal, and Ambuj Tewari. Convex calibrated surrogates for low-rank loss matrices with applications to subset ranking losses. In *Neural Information Processing Systems*, 2013.

Harish G. Ramaswamy, Ambuj Tewari, and Shivani Agarwal. Consistent algorithms for multiclass classification with a reject option. arXiv:1505.04137, 2015.

Pradeep Ravikumar, Ambuj Tewari, and Eunho Yang. On NDCG consistency of listwise ranking methods. In *International Conference on Artificial Intelligence and Statistics*, 2011.

Mark D. Reid and Robert C. Williamson. Composite binary losses. *Journal of Machine Learning Research*, 11:2387–2422, 2010.

Clayton Scott. Calibrated asymmetric surrogate losses. *Electronic Journal of Statistics*, 6:958–992, 2012.

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

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

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

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

Jason Weston and Chris Watkins. Support vector machines for multi-class pattern recognition. In *7th European Symposium On Artificial Neural Networks*, 1999.

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.

Ming Yuan and Marten Wegkamp. Classification methods with reject option based on convex risk minimization. *Journal of Machine Learning Research*, 11:111–130, 2010.

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

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