Journal of Machine Learning Research 17 (2016) 1-52 Submitted 7/14; Revised 5/16; Published 12/16

Composite Multiclass Losses

Robert C. Williamson
Australian National University and Data61 BOB.WILLIAMSON@ANU.EDU.AU

Elodie Vernet
Centre for Mathematical Sciences
University of Cambridge EV315@CAM.AC.UK

Mark D. Reid
Australian National University and Data61 MARK.REID@ANU.EDU.AU

Editor: Nicolas Vayatis

Abstract

We consider loss functions for multiclass prediction problems. We show when a multiclass loss can be expressed as a “proper composite loss”, which is the composition of a proper loss and a link function. We extend existing results for binary losses to multiclass losses. We subsume results on “classification calibration” by relating it to properness. We determine the stationarity condition, Bregman representation, order-sensitivity, and quasi-convexity of multiclass proper losses. We then characterise the existence and uniqueness of the composite representation for multiclass losses. We show how the composite representation is related to other core properties of a loss: mixability, admissibility and (strong) convexity of multiclass losses which we characterise in terms of the Hessian of the Bayes risk. We show that the simple integral representation for binary proper losses can not be extended to multiclass losses but offer concrete guidance regarding how to design different loss functions. The conclusion drawn from these results is that the proper composite representation is a natural and convenient tool for the design of multiclass loss functions.

Keywords: Proper losses, Multiclass losses, Link Functions, Convexity and quasi-convexity of losses, Margin losses, Classification calibration, Parametrisations and representations of loss functions, Admissibility, Mixability, Minimaxity, Superprediction set

1. Introduction

Machine learning is done for a purpose. The performance of a machine learning solution is judged by means of a loss function. Different choices of loss function will lead to different solutions. The theory of binary losses (i.e. losses suitable for binary prediction problems) is well understood. This paper extends that understanding to multiclass losses and aids the choice of a suitable loss function by exploring the parametrisations available and the implications of different choices. It does so by systematically exploring a decomposition of a multiclass loss into two components, one which affects the statistical performance, and one which affects the computational optimisation of models.

The problem setting is where one is given a bag $\lbag (x_i, y_i) \rbag_i$ of pairs of points x_i and their accompanying labels $y_i \in [n] := \{1, \ldots, n\}$, drawn from a finite set of size n. The task can be either predict a label for an unseen instance, or predict the probability that a label takes on a particular value. These two problems are called multiclass *classification* and *probability estimation* respectively.

Proper composite losses are the composition of a *proper loss* and and *invertible link* (both defined formally below). This representation makes the understanding of multiclass losses easier because, crucially, it seperates two distinct concerns: the statistical and the computational. The statistical properties are controlled by the proper loss, while the link function is essentially just a parametrisation. Choice of a suitable link can help—for example, a nonconvex proper loss can be made convex (and thus more amenable to numerical optimisation) by choice of the appropriate link. For prediction purposes it is desirable to use an *admissible* loss (one where every possible prediction is uniquely optimal for some underlying distribution). It turns out that every proper composite loss is admissible; in fact proper composite losses satisfy a stronger adequacy property than admissibility.

We characterise when a multiclass loss has a proper composite representation and when such representations are unique. We consider integral representations (whereby the proper component can be expressed as a weighted combination of elementary proper losses). We show the suprising result that there is a fundamental difference between $n = 2$ and $n > 2$ in terms of the simplicitly of the parametrisation of the class of elementary proper losses. It has been known for some time that proper losses are characterised by their conditional Bayes risks (or entropy functions). It has already been shown how important properties of a loss that control the performance of certain learning tasks can be expressed directly in terms of the Bayes risk. In this paper we extend results due to Reid and Williamson (2010) (for $n = 2$) to general n and characterise the convexity of a proper loss in terms of the associated Bayes risk.

We also illuminate the connection between classification and probability estimation by characterising the relationship between the cruicial property that a loss should have for each of these: *classification calibrated* (which we first generalise to make sense in the more general setting we consider) and *properness*. We explain the relationship between these two concepts, which captures the idea behind the probing reduction from classification to class probability estimation.

We also show how the results of the paper can provide tools to help with the design of multiclass losses, putting this on firmer ground than in the past.

1.1 Previous Work

With some exceptions, existing work on multiclass loss functions attempts to work directly with $\ell \colon \mathcal{V} \to \mathbb{R}^n_+$. As we shall show this conflates two seperate concerns—the design of the *statistical* properties of the loss, those that affect statistical performance, with the aspects that affect the computational properties that control the ease with which empirical averages of the loss are minimized. The proper composite representation is not new—in hindsight the observation of Grünwald and Dawid (2004) that every loss induces a proper scoring rule is tantamount to the proper composite representation. Furthermore, its components (link functions and proper losses) have a long history. The novelty of the present work is to systematically use these two components as a canonical parametrisation of loss functions. Key differences between the present paper and previous work are tabulated in Table 1.

Attribute	**Previous Work**	**Present Paper**	**Ref.**
Structure and Semantics	None—just a function; possibly convex in parameters	Clear seperation of concerns and meaning for λ and ψ. Gives meaning to predictions v as transformed probabilities.	Fig. 1
Classification versus probability estimation	Little insight in the multiclass case; confer recent works such as (Reid and Williamson, 2010, 2011; Narasimhan and Agarwal, 2013; Menon and Williamson, 2014) for the binary case	Clear connection via a characterisation relating classification calibrated, prediction calibrated and proper losses	§3
Effect of choice of loss function on performance	Margin based. Only a sufficient condition and only for statistical batch setting. Mixes up statistical fundamentals ($\underline{L}$) with parametrization (ψ). Strong convexity for speed of convergence in online setting; cf. (Abernethy et al., 2009).	Mixability and Stochastic Mixability. Characterisation in online setting. Both online worst-case and statistical batch settings. Parametrisation ψ automatically ignored.	§6.1
Admissibility	Not considered explicitly. Ensured however by assuming ℓ is convex.	All proper composite losses admissible. All continuous Bayes losses have a proper composite representation.	§6.2
Quasi-convexity and Minimaxity	Guaranteed by assumimg ℓ is convex.	Quasi-convexity guaranteed for all continuous proper losses; minimaxity for all continuous proper composite losses.	§6.4
Convexifiability	No principled way to convexify a loss; can make convex surrogate approximations.	All continuous proper losses convexifiable (using the canonical link).	§6.4
Design principles and parametrisation	No guidance; choose ℓ or margin function ϕ, in which case symmetry imposed.	Principled; general asymmetric losses possible; parametrise via $(\underline{\Delta}, \Psi)$; separation of concerns.	§8.3
Connections to divergences	Many to one for margin losses in binary case. (Nguyen et al., 2009)	Explicit 1:1 correspondence for binary and multiclass case (Reid and Williamson, 2011; García-García and Williamson, 2012).	§9

Table 1: Comparison of present paper to previous works on loss functions.

Proper losses are the natural losses to use for probability estimation. They have been studied in detail when $n = 2$ (the “binary case”) where there is a nice integral representation (Buja et al., 2005; Gneiting and Raftery, 2007; Reid and Williamson, 2011), and characterization (Reid and Williamson, 2010) when differentiable. The proper composite representation for binary losses has proved very illuminating in the study of bipartite ranking problems (Menon and Williamson, 2014). Classification calibrated losses are an analog of proper losses for the problem of classification (Bartlett et al., 2006). The relationship between classification calibration and properness was determined by Reid and Williamson (2010) for $n = 2$. Most of these results have had no multiclass analogue until now. Whilst there is much work on classification problems, it is now widely understood that there are often advantages in being able to predict probabilities, rather than just labels (Bennett, 2003; Cohen and Goldszmidt, 2004).

The *theory* of loss functions makes it clear how one ideally chooses a loss—one takes account of one's utility concerning various incorrect predictions (Kiefer, 1987), (Berger, 1985, Section 2.4). The *practice* rarely involves such a step, primarily, we conjecture, because there is no adequate understanding of the way one can parametrise losses effectively, especially in the multiclass case. There is little guidance in the literature concerning how to choose a loss function; typically heuristic arguments are used for the choice—confer e.g. (Ighodaro et al., 1982; Nayak and Naik, 1989). An early approach to multiclass losses is simply reduction to binary (Allwein et al., 2001; Beygelzimer et al., 2007; Crammer and Singer, 2001; Dietterich and Bakiri, 1995; Zadrozny and Elkan, 2002). Related approaches are pairwise coupling or Bradley-Terry models (Hastie and Tibshirani, 1998; Wu and Weng, 2004; Huang et al., 2006) where certain relationships are assumed to hold between the pairwise probabilities and the multivariate probability of interest.

The design of losses for multiclass prediction has received recent attention (Zhang, 2004; Hill and Doucet, 2007; Tewari and Bartlett, 2007; Liu, 2007; Santos-Rodríguez et al., 2009; Zou et al., 2008; Zhang et al., 2009) although none of these papers developed the connection to proper losses, and most restrict consideration to margin losses (which imply certain symmetry conditions). Zou et al. (2005) proposed a multiclass generalisation of "admissible losses" (their name for classification calibration) for multiclass margin classification. Liu (2007) considered several multiclass generalisations of hinge loss (suitable for multiclass SVMs) and showed some of them were and others were not Fisher consistent, and when they were not it was shown how the training algorithm could be modified to make the losses behave consistently. Shi et al. (2010) have investigated the relationship between classification calibration of multiclass losses and losses for structure prediction, and have proposed an extension of classification calibration which they call parametric consistency, which attempts to take account of the function class used (classification calibration is, like all the results in this paper, concerned with behaviour *per point*; in practice one typically optimises over restricted classes of functions). Multiclass losses have also been considered in the development of multiclass boosting (e.g. Zhu et al., 2009; Mukherjee and Schapire, 2013; Wu and Lange, 2010).

1.2 Outline

The rest of the paper is organised as follows. In §2: we set up the problem formally and state some purely mathematical results we will need; §3: we relate properness, classification calibration, and the notion used by Tewari and Bartlett (2007) which we rename "prediction calibrated"; §4: we provide a novel characterization of multiclass properness; §5: we study composite proper losses (the composition of a proper loss with an invertible link) and characterise when a given loss has such a representation and when the representation is unique; §6: we develop a number of interesting implications of the representation and the characterisation results in terms of mixability (§6.1), admissibility (§6.2) and convexity (§6.4), where we give a complete characterisation of the (strong) convexity of composite multiclass proper losses in terms of the Bayes risk; §7: we present a (somewhat surprising) negative result concerning the integral representation of proper multiclass losses; §8: we outline how the above results can aid in the design of proper losses, especially by use of a (new) multiclass extension of the "canonical link"; finally, §9 summarises the key contributions and outlines some future directions.

2. Formal Setup

Suppose $\mathscr{X}$ is some set and $\mathscr{Y} = [n] = \{1,\ldots,n\}$ is a set of labels. (Throughout the paper n is an integer greater than or equal to 2.) We suppose we are given data $S = \wr(x_i, y_i)\int_{i\in[m]}$ such that $y_i \in \mathscr{Y}$ is the label corresponding to $x_i \in \mathscr{X}$. These data follow a joint distribution $\mathbb{P}_{\mathscr{X},\mathscr{Y}}$ on $\mathscr{X} \times [n]$. We denote by $\mathbb{E}_{\mathscr{X},\mathscr{Y}}$ and $\mathbb{E}_{\mathscr{Y}|\mathscr{X}}$ respectively, the expectation and the conditional expectation with respect to $\mathbb{P}_{\mathscr{X},\mathscr{Y}}$. Given a new observation x we want to predict the probability $p_i := \mathbb{P}(Y = i|X = x)$ of x belonging to class i, for $i \in [n]$. *Multiclass classification* requires the learner to predict the most likely class of x; that is to find $\hat{y} \in \arg\max_{i\in[n]} p_i$.

A loss measures the quality of prediction. Let $\Delta^n := \{(p_1,\ldots,p_n)\colon \sum_{i\in[n]} p_i = 1, \text{and } 0 \leq p_i \leq 1,\ \forall i \in [n]\}$ denote the *n-simplex*. For multiclass probability estimation, $\ell\colon \Delta^n \to \mathbb{R}^n_+$. The *partial losses* ℓ_i are the components of $\ell(q) = (\ell_1(q),\ldots,\ell_n(q))'$ and $\ell_i(q)$ is the loss incurred by predicting $q \in \Delta^n$ when $y = i$. A commonly used loss for probability estimation is the *log loss* $\ell^{\log}$ defined by $\ell_i^{\log}(q) := -\log q_i$ for $i \in [n]$. Other examples of multiclass losses we will refer to in this paper include the *square loss* $\ell_i^{\text{sq}}(q) := \sum_{j\in[n]}([\![i = j]\!] - q_j)^2$, the *absolute loss* $\ell_i^{\text{abs}}(q) := \sum_{j\in[n]} |[\![i = j]\!] - q_j|$ and the *0-1 loss* $\ell_i^{01}(q) := [\![i \in \arg\max_{j\in[n]} q_j]\!]$. Here, $[\![P]\!]$ denotes the function that is 1 when P is true and 0 otherwise.

Throughout the paper, A' denotes transpose of the matrix or vector A, except when applied to a real-valued function where it denotes derivative. We denote matrix multiplication of compatible matrices A and B by $A \cdot B$, so the inner product of two vectors $x, y \in \mathbb{R}^n$ is $x' \cdot y$. The *conditional risk* L associated with a loss ℓ is the function

$$L\colon \Delta^n \times \Delta^n \ni (p,q) \mapsto L(p,q) = \mathbb{E}_{\mathsf{Y}\sim p}\ell_{\mathsf{Y}}(q) = p' \cdot \ell(q) = \sum_{i\in[n]} p_i \ell_i(q) \in \mathbb{R}_+,$$

where $\mathsf{Y} \sim p$ means Y is drawn according to a multinomial distribution with parameter $p \in \Delta^n$. In a typical learning problem one will construct an estimate $q\colon \mathscr{X} \to \Delta^n$. The *full risk* is $\mathbb{L}(q) = \mathbb{E}_{\mathscr{X}}\mathbb{E}_{\mathscr{Y}|\mathscr{X}}\ell_{\mathsf{Y}}(q(\mathsf{X}))$. Minimizing $\mathbb{L}(q)$ over $q\colon \mathscr{X} \to \Delta^n$ is equivalent to minimizing $L(p(x), q(x))$ over $q(x) \in \Delta^n$ for all $x \in \mathscr{X}$ where $p(x) = (p_1(x),\ldots,p_n(x))'$, and $p_i(x) = \mathbb{P}(\mathsf{Y} = i|\mathsf{X} = x)$. Thus it suffices to only consider the conditional risk; confer (Reid and Williamson, 2011).

If one is interested in estimating probabilities ($\ell\colon \Delta^n \to \mathbb{R}^n_+$) it is natural to require the associated conditional risk is minimized when estimating the true underlying probability. Such a loss is called *proper* (formally: if $L(p,p) \leq L(p,q),\ \forall p,q \in \Delta^n$). It is *strictly proper* if the inequality is strict when $p \neq q$ (so it is uniquely minimised by predicting the correct probability). The *conditional Bayes risk* is defined by

$$\underline{L}\colon \Delta^n \ni p \mapsto \inf_{q\in\Delta^n} L(p,q).$$

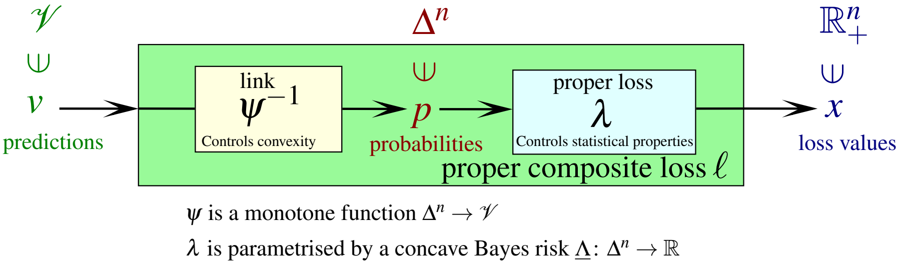
$\mathcal{V}$
Δ^n
$\mathbb{R}^n_+$
v
predictions
link
ψ^{-1}
Controls convexity
p
probabilities
proper loss
λ
Controls statistical properties
proper composite loss ℓ
x
loss values
ψ is a monotone function $\Delta^n \to \mathcal{V}$
λ is parametrised by a concave Bayes risk $\underline{\Lambda}\colon \Delta^n \to \mathbb{R}$

Figure 1: The idea of a proper composite loss.

This function is always concave (Gneiting and Raftery, 2007). If ℓ is proper, then $\underline{L}(p) = L(p,p) = p' \cdot \ell(p)$. Strictly proper losses induce *Fisher consistent* estimators of probabilities: if ℓ is strictly proper, $p = \arg\min_q L(p,q)$. By considering when the derivatives $\frac{\partial}{\partial q_i} L(p,q)$ are zero it is straight-forward to show that, of the example losses introduced above, the log loss, square loss, and 0-1 loss are proper, while absolute loss is not. Furthermore, both log loss and square loss are strictly proper while 0-1 loss is proper but not strictly proper. Using the fact that, for proper losses, the Bayes risk $\underline{L}(p) = L(p,p)$ we see that $\underline{L}^{\log}(p) = -\sum_{i\in[n]} p_i \log p_i$ (*i.e.*, Shannon entropy); $\underline{L}^{sq}(p) = 1 - \sum_{i\in[n]} p_i^2$; and $\underline{L}^{01}(p) = \min_i\{1-p_i\}$.

The losses above are defined on the simplex Δ^n since the argument (a predictor) represents a probability vector. However it is sometimes desirable to use another set $\mathcal{V}$ of predictions. For example if one wishes to use linear predictors, their natural range is $\mathbb{R}^n$. One can consider losses $\ell\colon \mathcal{V} \to \mathbb{R}^n_+$. Suppose there exists an invertible function $\psi\colon \Delta^n \to \mathcal{V}$. Then ℓ can be written as a composition of a loss λ defined on the simplex with ψ^{-1}. That is, $\ell(v) = \lambda^{\psi}(v) := \lambda(\psi^{-1}(v))$. Such a function λ^{ψ} is a *composite loss*. If λ is proper, we say ℓ is a *proper composite loss*, with *associated proper loss* λ and *link* ψ; see Figure 1. Many commonly used multiclass losses are composite losses, even though they are not often expressed as such; see the example in §8.4.

Throughout the paper, ℓ is a general loss defined on $\mathcal{V}$, where $\mathcal{V}$ may equal Δ^n, and λ is always a loss defined on Δ^n, which may be proper. For such a loss $\lambda : \Delta^n \to \mathbb{R}^n_+$, its corresponding conditional risk is denoted $\Lambda(p,q)$ and its conditional Bayes risk is $\underline{\Lambda}(p)$.

In order to differentiate the losses we project the n-simplex into a subset of $\mathbb{R}^{n-1}$. Let

$$\tilde{\Delta}^n := \left\{ (p_1,\ldots,p_{n-1})'\colon\, p_i \geq 0,\ \forall i \in [n],\ \sum_{i=1}^{n-1} p_i \leq 1 \right\}$$

denote the *"bottom" of the n-simplex*. We denote by

$$\Pi_\Delta\colon \Delta^n \ni p = (p_1,\ldots,p_n)' \mapsto \tilde{p} = (p_1,\ldots,p_{n-1})' \in \tilde{\Delta}^n,$$

the projection of the Δ^n, and

$$\Pi_\Delta^{-1} : \tilde{\Delta}^n \ni \tilde{p} = (\tilde{p}_1,\ldots,\tilde{p}_{n-1})' \mapsto p = (\tilde{p}_1,\ldots,\tilde{p}_{n-1}, 1 - \sum_{i=1}^{n-1} \tilde{p}_i)' \in \Delta^n$$

its inverse. For convenience, we will often use $\tilde{n} := n-1$ to denote the dimension of the set $\tilde{\Delta}^n$.

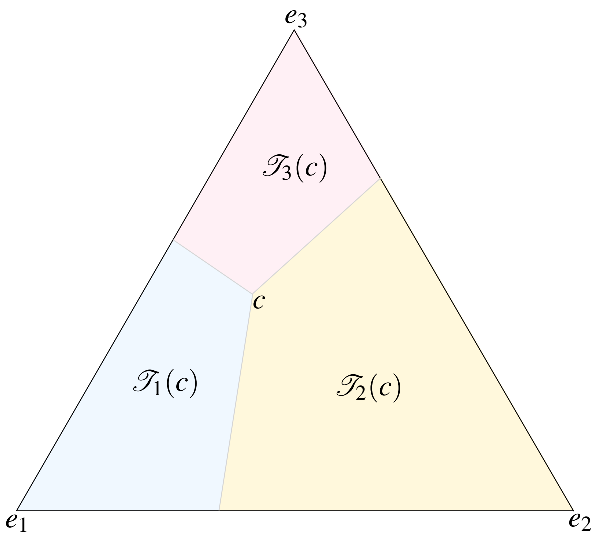

Figure 2: A partitioning of the 3-simplex by regions $\mathscr{T}_i(c)$, $i = 1,2,3$, where $c = (.35,.2,.45)$ as viewed from the direction $(1,1,1)$.

We use the following notation. The kth unit vector e_k is the n vector with all components zero except the kth which is 1. The n-vector $\mathbb{1}_n := (1,\ldots,1)'$. The (relative) interior of the simplex is $\mathring{\Delta}^n := \{(p_1,\ldots,p_n) \colon \sum_{i\in[n]} p_i = 1, \text{and } 0 < p_i < 1,\ \forall i \in [n]\}$ and the boundary is $\partial\Delta^n := \Delta^n \setminus \mathring{\Delta}^n$. We also adopt notation from Magnus and Neudecker (1999). For the reader's convenience we list the essential notations and conventions in Appendix A.

3. Relating Properness to Classification Calibration

Properness is an attractive property of a loss for the task of class probability estimation. However if one is merely interested in *classifying* (predicting $\hat{y} \in [n]$ given $x \in \mathscr{X}$) then it is stronger than one needs. In this section we relate *classification calibration* (the analog of properness for classification problems) to properness.

Suppose $c \in \mathring{\Delta}^n$. We cover Δ^n with n subsets each representing one class:

$$\mathscr{T}_i(c) := \{p \in \Delta^n \colon \forall j \neq i\ p_i c_j \geq p_j c_i\},\quad i \in [n].$$

Observe that for $i \neq j$, the sets $\mathscr{R}_{ij}(c) := \{p \in \Delta^n \colon\ p_i c_j = p_j c_i\}$ are subsets of dimension $n-2$ through c and all e_k such that $k \neq i$ and $k \neq j$. These subsets partition $\mathbb{R}^n$ into two parts. The set $\mathscr{R}_{ij}(c)$ is the intersection of Δ^n and the subspaces delimited by the precedent $(n-2)$-subspace and in the same side as e_i. An example of this partition is shown graphically in Figure 2. We will make use of the following properties of $\mathscr{T}_i(c)$.

Lemma 1 *Suppose $c \in \mathring{\Delta}^n$, $i \in [n]$. Then the following hold:*

1. *For all $p \in \Delta^n$, there exists i such that $p \in \mathscr{T}_i(c)$.*

2. *Suppose $p \in \Delta^n$. $\mathscr{T}_i(c) \cap \mathscr{T}_j(c) \subseteq \{p \in \Delta^n \colon\ p_i c_j = p_j c_i\}$, a subset of a subspace of dimension $n-2$.*

3. *Suppose $p \in \Delta^n$. If $p \in \bigcap_{i=1}^n \mathscr{T}_i(c)$ then $p = c$.*

4. *For all $p, q \in \Delta^n$, $p \neq q$, there exists $c \in \mathring{\Delta}^n$, and $i \in [n]$ such that $p \in \mathscr{T}_i(c)$ and $q \notin \mathscr{T}_i(c)$.*

The proof is deferred to Appendix B.1.

Classification calibrated losses have been developed and studied under some different definitions and names (Zhang, 2004; Bartlett et al., 2006). Below we generalise the notion of c-calibration which was proposed for $n = 2$ by Reid and Williamson (2010) and developed by Scott (2011, 2012) as a generalisation of the notion of classification calibration of Bartlett et al. (2006); confer also Steinwart (2007).

Definition 2 *Suppose $\ell \colon \Delta^n \to \mathbb{R}^n_+$ is a loss and $c \in \mathring{\Delta}^n$. We say ℓ is c-calibrated at $p \in \Delta^n$ if for all $i \in [n]$ such that $p \notin \mathscr{T}_i(c)$ then $\forall q \in \mathscr{T}_i(c)$, $\underline{L}(p) < L(p,q)$. We say that ℓ is c-calibrated if $\forall p \in \Delta^n$, ℓ is c-calibrated at p.*

Definition 2 means that if the probability vector q one predicts doesn't belong to the same subset (i.e. doesn't predict the same class) as the real probability vector p, then the loss might be larger than $\underline{L}(p)$.

Classification calibration in the sense used by Bartlett et al. (2006) corresponds to $\frac{1}{2}$-calibrated losses when $n = 2$. If $c_{\mathrm{mid}} := (\frac{1}{n}, \ldots, \frac{1}{n})'$, c_{mid}-calibration induces Fisher-consistent estimates in the case of classification. Furthermore "ℓ is c_{mid}-calibrated and for all $i \in [n]$, and ℓ_i is continuous and bounded below" is equivalent to "ℓ is infinite sample consistent" as defined by Zhang (2004). This is because if ℓ is continuous and $\mathscr{T}_i(c)$ is closed, then $\forall q \in \mathscr{T}_i(c)$, $\underline{L}(p) < L(p,q)$ if and only if $\underline{L}(p) < \inf_{q \in \mathscr{T}_i(c)} L(p,q)$.

The following result generalises the correspondence between binary classification calibration and properness (Reid and Williamson, 2010, Theorem 16) to multiclass losses ($n > 2$).

Proposition 3 *A continuous loss $\ell \colon \Delta^n \to \mathbb{R}^n_+$ is strictly proper if and only if it is c-calibrated for all $c \in \mathring{\Delta}^n$.*

Proof ($\Rightarrow$) Suppose that ℓ is strictly proper. Then for all $c \in \mathring{\Delta}^n$, for all $i \in [n]$ such that $p \notin \mathscr{T}_i(c)$ and for all $q \in \mathscr{T}_i(c)$ then $p \neq q$ and thus $\underline{L}(p) < L(p,q)$ since ℓ is strictly proper.

($\Leftarrow$) Suppose that ℓ is c-calibrated for all $c \in \mathring{\Delta}^n$. Suppose $p, q \in \Delta^n$ and $p \neq q$. By Lemma 1 (part 4) one can partition p and q into two different classes: there exists $c \in \mathring{\Delta}^n$ and $i \in [n]$ such that $q \in \mathscr{T}_i(c)$ and $p \notin \mathscr{T}_i(c)$. Hence $\underline{L}(p) < L(p,q)$ since ℓ is c-calibrated. Since ℓ is continuous and Δ^n is closed, the infimum in the definition of $\underline{L}(p)$ is attained. Since $\underline{L}(p) < L(p,q)$ for all $q \neq p$, we conclude $\underline{L}(p) = L(p,p)$. Thus ℓ is strictly proper. ∎

In particular, a continuous strictly proper loss is c_{mid}-calibrated. Thus for any estimator $\hat{q}_n$ of the conditional probability vector one constructs by minimizing the empirical average of a continuous strictly proper loss, one can build an estimator of the label (corresponding to the largest probability of $\hat{q}_n$) which is Fisher consistent for the problem of classification.

In the binary case, ℓ is classification calibrated if and only if the following implication holds (Bartlett et al., 2006):

$$\left(\mathbb{L}(f_n) \to \min_g \mathbb{L}(g)\right) \Rightarrow \left(\mathbb{P}_{\mathscr{X},\mathscr{Y}}(\mathsf{Y} \neq f_n(\mathsf{X})) \to \min_g \mathbb{P}_{\mathscr{X},\mathscr{Y}}(\mathsf{Y} \neq g(\mathsf{X}))\right). \tag{1}$$

Tewari and Bartlett (2007) have characterised when (1) holds in the multiclass case. Since there is no reason to assume the equivalence between classification calibration and (1) still holds for $n > 2$, we give different names for these two notions. We use *classification calibration* for the notion (Definition 2) linked to Fisher consistency and use *prediction calibrated* (defined below) for the notion of Tewari and Bartlett (equivalent to (1)).

Definition 4 *Suppose $\ell\colon \mathscr{V} \to \mathbb{R}^n_+$ is a loss. Let $\mathscr{C}_\ell := \operatorname{co}(\{\ell(v)\colon v \in \mathscr{V}\})$, the convex hull of the image of $\mathscr{V}$. ℓ is said to be* prediction calibrated *if there exists a prediction function* $\operatorname{pred}\colon \mathbb{R}^n \to [n]$ *such that*

$$\forall p \in \Delta^n\colon \inf_{z\in\mathscr{C}_\ell\colon p_{\operatorname{pred}(z)}<\max_{i\in[n]} p_i} p'\cdot z > \inf_{z\in\mathscr{C}_\ell} p'\cdot z = \underline{L}(p).$$

Suppose that $\ell\colon \Delta^n \to \mathbb{R}^n_+$ is such that ℓ is prediction calibrated and $\operatorname{pred}(\ell(p)) \in \arg\max_i p_i$. Then ℓ is c_{mid}-calibrated almost everywhere.

By introducing a *reference link* $\bar{\psi}$ (which corresponds to the actual link ψ if ℓ is a proper composite loss $\ell = \lambda \circ \psi^{-1}$) we now show how the pred function can be canonically expressed in terms of $\arg\max_i p_i$.

Proposition 5 *Suppose $\ell\colon \mathscr{V} \to \mathbb{R}^n_+$ is a loss. Let $\bar{\psi}\colon \Delta^n \to \mathscr{V}$ satisfy $\bar{\psi}(p) \in \arg\min_{v\in\mathscr{V}} L(p,v)$ and $\lambda = \ell \circ \bar{\psi}$. Then λ is proper. If ℓ is prediction calibrated then* $\operatorname{pred}(\lambda(p)) \in \arg\max_{i\in[n]} p_i$.

Proof We show first that λ is proper. Let $p \in \Delta^n$. Then

$$\Lambda(p,p) = L(p,\bar{\psi}(p)) = L(p,\arg\min_v L(p,v)) = \min_v L(p,v) \leq \min_{q\in\Delta^n} \Lambda(p,q).$$

Thus λ is proper and $\underline{L}(p) = \underline{\Lambda}(p)$. We now assume that ℓ is prediction calibrated. Suppose that $\operatorname{pred}(z = \lambda(p)) \notin \arg\max_i p_i$. Then $p_{\operatorname{pred}(\lambda(p))} < \max_i p_i$, thus $p'\cdot z = \Lambda(p,p) > \underline{L}(p) = \underline{\Lambda}(p)$ which contradicts the properness of λ. ∎

4. Characterizing Properness

We now present some simple (but new) consequences of properness in the multiclass case (Proposition 6). We also build some connections between the properness of multiclass losses and the properness of binary losses that can be derived from them via a restriction of the multiclass loss to a line connecting two points in the n-simplex (Proposition 7). Finally, we show that multiclass proper losses are effectively characterised by their Bayes risks (Proposition 8) and the continuity of losses is intimately tied to the differentiability of their Bayes risks (Proposition 9). An important implication of these last results is that we are able to study the class of multiclass proper losses by focusing our attention on concave functions defined over probabilities.

To state our propositions we need to introduce monotone functions, directional derivatives, and superdifferentials (cf. (Hiriart-Urruty and Lemaréchal, 2001)). We say $f\colon C \subset \mathbb{R}^n \to \mathbb{R}^n$ is *monotone* (resp. *strictly monotone*) on C when for all x and y in C,

$$(f(x)-f(y))'\cdot(x-y) \geq 0 \quad \text{resp.} \quad (f(x)-f(y))'\cdot(x-y) > 0; \qquad (2)$$

confer (Hiriart-Urruty and Lemaréchal, 2001; Rockafellar and Wets, 2004). If a function $f:\mathbb{R}^n \to \mathbb{R}$ is concave then $\lim_{t\downarrow 0} \frac{f(x+td)-f(x)}{t}$ exists, and is called the *directional derivative* of f at x in the direction d and is denoted $\mathsf{D}f(x,d)$. By analogy with the usual definition of *sub*differential for convex functions, we introduce the *superdifferential* $\partial f(x)$ for concave f at x is

$$\begin{aligned}\partial f(x) &:= \left\{s\in\mathbb{R}^n\colon\, s'\cdot y \geq \mathsf{D}f(x,y),\ \forall y\in\mathbb{R}^n\right\}\\ &= \left\{s\in\mathbb{R}^n\colon\, f(y)\leq f(x)+s'\cdot(y-x),\ \forall y\in\mathbb{R}^n\right\}.\end{aligned}$$

Similarly, a vector $s\in\partial f(x)$ is called a *supergradient* of f at x.

Proposition 6 *Suppose $\ell\colon \Delta^n\to\mathbb{R}^n_+$ is a loss. If ℓ is proper, then $-\ell$ is monotone on Δ^n. Furthermore, if ℓ is strictly proper then it is also invertible.*

Proof For all $p,q\in\Delta^n$, $(\ell(p)-\ell(q))'\cdot(p-q) = p'\cdot\ell(p)-q'\cdot\ell(p)+q'\cdot\ell(q)-p'\cdot\ell(q)\leq 0$ since $p'\cdot\ell(p)\leq p'\cdot\ell(q)$. For the strictly proper case, we just have to check that ℓ is injective. By way of contradiction assume ℓ is not invertible. Then there exists $p\neq q$ such that $\ell(p)=\ell(q)$. which means $L(p,p)=L(p,q)$, contradicting the supposed strict properness of ℓ. ∎

The following proposition presents several characterisations of multiclass properness. It shows how the characterisation of properness in the general (not necessarily differentiable) multiclass case can be reduced to the binary case. We also show this is equivalent to testing the properness condition for the loss on all possible line segments joining two distributions within the simplex. This latter characterisation can be viewed as a statement connecting "order sensitivity" and properness: the true class probability minimizes the risk and if the prediction moves away from the true class probability in a line then the risk increases. This property appears convenient for optimisation purposes: if one reaches a local minimum in the second argument of the risk and the loss is strictly proper then it is a global minimum. If the loss is proper, such a local minimum is a global minimum or a constant in an open set. But observe that typically one is minimising the full risk $\mathbb{L}(q(\cdot))$ over functions $q\colon \mathscr{X}\to\Delta^n$. We note that order sensitivity of ℓ does *not* imply this optimisation problem is well behaved; one needs convexity of $q\mapsto L(p,q)$ for all $p\in\Delta^n$ to ensure convexity of the functional optimisation problem; we characterise when that holds in section 6.4.

Proposition 7 *Suppose $\ell\colon\Delta^n\to\mathbb{R}^n_+$ is a loss. We define the binary loss*

$$\tilde{\ell}^{p,q}\colon [0,1]\ni\eta\mapsto\begin{pmatrix}\tilde{\ell}^{p,q}_1(\eta)\\ \tilde{\ell}^{p,q}_{-1}(\eta)\end{pmatrix} = \begin{pmatrix} q'\cdot\ell(p+\eta(q-p))\\ p'\cdot\ell(p+\eta(q-p))\end{pmatrix}.$$

The following statements are equivalent:

1. *ℓ is proper;*

2. *$\tilde{\ell}^{p,q}$ is proper for all $p,q\in\partial\Delta^n$;*

3. *$\forall p,q\in\Delta^n$, $\forall 0\leq h_1\leq h_2$, $L(p,p+h_1(q-p))\leq L(p,p+h_2(q-p))$; and*

4. *there exists a concave function $f : \Delta^n \to \mathbb{R}$ and $\forall q \in \Delta^n$, there exists a supergradient $A(q) \in \partial f(q)$ such that $\forall p, q \in \Delta^n$, $p' \cdot \ell(q) = L(p,q) = f(q) + (p-q)' \cdot A(q)$.*

The proof is deferred to Appendix B.3.

Characterisation (2) shows that in order to check if a loss is proper one need only check the properness in each line. One could use the easy characterization of properness for differentiable binary losses ($\ell\colon [0,1] \to \mathbb{R}^2_+$ is proper if and only if $\forall \eta \in [0,1]$, $\frac{-\ell_1'(\eta)}{1-\eta} = \frac{\ell_{-1}'(\eta)}{\eta} \geq 0$, (Reid and Williamson, 2010)). However this needs to be checked for all lines defined by $p, q \in \partial \Delta^n$. The above result can also been seen as a generalisation of a result by Lambert (2010) who proved that properness is equivalent to the fact that the further your prediction is from reality, the larger the loss (hence the name "order sensitivity"); also confer the results on monotonicity due to Nau (1985). His result relied upon on the total order of $\mathbb{R}$. In the multiclass case, there does not exist such a total order. Yet, as the above result shows, one can compare two predictions if they are in the same line as the true real class probability.

Characterisation (4) is a restatement of the well known Bregman representation of proper losses; Cid-Sueiro and Figueiras-Vidal (2001) presented the differentiable case, and Gneiting and Raftery (2007, Theorem 3.2) the general case. This last property gives us the form of the proper losses associated with a given Bayes risk. Suppose $\underline{L}\colon \Delta^n \to \mathbb{R}_+$ is concave. The proper losses whose Bayes risk is equal to $\underline{L}$ are

$$\ell\colon \Delta^n \ni q \mapsto \Big(\underline{L}(q) + (e_i - q)' \cdot A(q)\Big)_{i=1}^n \in \mathbb{R}^n_+, \ \forall A(q) \in \partial \underline{L}(q). \tag{3}$$

This result suggests that some information is lost by representing a proper loss via its Bayes risk (when the last is not differentiable). The next proposition elucidates this by showing that proper losses which have the same Bayes risk are equal almost everywhere.

Proposition 8 *Two proper losses $\ell^1, \ell^2 \colon \Delta^n \to \mathbb{R}^n_+$ have the same conditional Bayes risk function $\underline{L}$ if and only if $\ell^1 = \ell^2$ almost everywhere. If $\underline{L}$ is differentiable, $\ell^1 = \ell^2$ everywhere.*

Proof A concave function is differentiable almost everywhere (Hiriart-Urruty and Lemaréchal, 2001, theorem 4.2.3). Thus (3) proves that two proper losses ℓ^1 and ℓ^2 which have the same Bayes risk are equal almost everywhere. Suppose now that two proper losses are equal almost everywhere. Then their associated Bayes risks $\underline{L}^1$ and $\underline{L}^2$ are equal almost everywhere and continuous (since they are concave). If there exists p such that $\underline{L}^1(p) \neq \underline{L}^2(p)$, then since $\underline{L}^1$ and $\underline{L}^2$ are continuous, there exists $\varepsilon > 0$ such that $\forall q \in B(p,\varepsilon) \cap \Delta^n$, $\underline{L}^1(q) \neq \underline{L}^2(q)$, where $B(p,\varepsilon)$ is a ball of radius ε centred at p. Yet this contradicts the fact that $\underline{L}^1$ and $\underline{L}^2$ are equal almost everywhere. Hence the Bayes risks are equal everywhere. ■

While the previous proposition shows that losses are closely related to their Bayes risks the next proposition also shows how the continuity of a loss is related to the differentiability of its Bayes risk.

Proposition 9 *Suppose $\ell\colon \Delta^n \to \mathbb{R}^n_+$ is a proper loss. Then ℓ is continuous in $\mathring{\Delta}^n$ if and only if $\underline{L}$ is differentiable on $\mathring{\Delta}^n$; ℓ is continuous at $p \in \mathring{\Delta}^n$ if and only if $\underline{L}$ is differentiable at $p \in \mathring{\Delta}^n$.*

The proof of this result can be found in Appendix B.2. This type of relationship is further explored in Section 6.4 where the convexity of a composite loss is related to properties of its Bayes risk.

5. The Proper Composite Representation: Uniqueness and Existence

Many natural predictors have a range other than the simplex (for example those induced by linear functions). It is thus sometimes convenient to define a loss on some set $\mathscr{V}$ rather than Δ^n; confer (Reid and Williamson, 2010). The link function explicates the result of Grünwald and Dawid (2004) that every decision problem induces a decision problem expressed in terms of proper losses; (see van Erven et al., 2011, section 6, for further explanation).

Traditionally (McCullagh and Nelder, 1989) links are defined only for binary problems (where one is using univariate probabilities). However there is scattered (but seemingly unsystematic) work on multivariate links (Glonek and McCullagh, 1995; Glonek, 1996), primarily from the perspective of probabilistic modelling (as opposed to the design of loss functions). Sometimes multivariate links are constructed from univariate links (Molenberghs and Lesaffre, 1999).

Composite losses (see the definition in §2) are a way of constructing losses on sets other than Δ^n: given a proper loss $\lambda\colon \Delta^n \to \mathbb{R}^n_+$ and an invertible link $\psi\colon \Delta^n \to \mathscr{V}$, one defines $\lambda^\psi\colon \mathscr{V} \to \mathbb{R}^n_+$ as $\lambda^\psi := \lambda \circ \psi^{-1}$. We now consider the question: given a loss $\ell\colon \mathscr{V} \to \mathbb{R}^n_+$, when does ℓ have a *proper composite representation* (whereby ℓ can be written as $\ell = \lambda \circ \psi^{-1}$), and is this representation unique? We first consider the binary case. Here the prediction space $\mathscr{V} \subseteq \mathbb{R}$ is assumed to be either an interval or the entire real line.

5.1 The Binary Case

Our first result shows that if you can write a binary loss as a proper composite loss, the proper loss defined on the simplex is unique. Furthermore, as soon as the loss is not constant the link function is also unique. If the loss is constant on an interval, then you can choose any value of the link function on this interval which keeps the link function continuous and invertible and still obtain a composite proper loss. The proof can be found in Appendix B.4. As is common in the literature, we write the binary labels as $\{-1,+1\}$ and so the partial losses are ℓ_{-1} and ℓ_{+1}.

Proposition 10 *Suppose $\ell = \lambda \circ \psi^{-1}\colon \mathscr{V} \to \mathbb{R}^2_+$ is a proper composite loss and that the proper loss λ is differentiable and the link function ψ is differentiable and invertible. Then the proper loss λ is unique. Furthermore ψ is unique if $\forall v_1, v_2 \in \mathscr{V}$, $\exists v \in [v_1, v_2]$, $\ell_1'(v) \neq 0$ or $\ell_{-1}'(v) \neq 0$. If there exists $\bar{v}_1, \bar{v}_2 \in \mathscr{V}$ such that $\ell_1'(v) = \ell_{-1}'(v) = 0\ \forall v \in [\bar{v}_1, \bar{v}_2]$, one can choose any $\psi|_{[\bar{v}_1,\bar{v}_2]}$ such that ψ is differentiable, invertible and continuous in $[\bar{v}_1, \bar{v}_2]$ and obtain $\ell = \lambda \circ \psi^{-1}$, and ψ is uniquely defined where ℓ is invertible.*

We now determine necessary and sufficient conditions for a binary loss to be expressed as a proper composite loss. Once again, the proof is deferred to Section B.5.

Proposition 11 *Suppose $\ell\colon \mathscr{V} \to \mathbb{R}^2_+$ is a differentiable binary loss such that $\forall v \in \mathscr{V}$, $\ell_{-1}'(v) \neq 0$ or $\ell_1'(v) \neq 0$. Then ℓ can be expressed as a proper composite loss if and only if the following three conditions hold:*

1. *ℓ_1 is decreasing (increasing);*

2. *ℓ_{-1} is increasing (decreasing); and*

3. *$f\colon \mathscr{V} \ni v \mapsto \frac{\ell_1'(v)}{\ell_{-1}'(v)}$ is strictly increasing (decreasing) and continuous.*

Observe that the last condition is alway satisfied if both ℓ_1 and ℓ_{-1} are convex.

5.2 Binary Margin Losses

Suppose $\varphi\colon \mathbb{R} \to \mathbb{R}_+$ is a function. The loss $\ell_\varphi\colon \mathscr{V} \ni v \mapsto (\ell_{-1}(v), \ell_1(v))' = (\varphi(-v), \varphi(v))' \in \mathbb{R}_+^2$ is called a binary *margin loss*. Binary margin losses are often used for classification problems. We will now show how the previous proposition applies to them.

Corollary 12 *Suppose $\varphi\colon \mathbb{R} \to \mathbb{R}_+$ is differentiable and $\forall v \in \mathbb{R}$, $\varphi'(v) \neq 0$ or $\varphi'(-v) \neq 0$. Then ℓ_φ can be expressed as a proper composite loss if and only if $f\colon \mathbb{R} \ni v \mapsto -\frac{\varphi'(v)}{\varphi'(-v)}$ is strictly monotonic continuous and φ is monotonic.*

If φ is convex or concave then f defined above is monotonic. However not all binary margin losses are composite proper losses. One can even build a smooth margin loss which cannot be expressed as a proper composite loss. Consider $\varphi(x) = \frac{1-\arctan(x-1)}{\pi}$. Then $f(v) = \frac{\varphi'(v)}{\varphi'(-v)} = \frac{x^2+2x+2}{x^2-2x+2}$ which is not invertible. This loss is illustrated in Figure 5, after some additional concepts are introduced.

5.3 The Multiclass Case

Uniqueness of the composite representation remains straightfoward in the multiclass case.

Proposition 13 *Suppose a loss $\ell\colon \mathscr{V} \to \mathbb{R}_+^n$ has two proper composite representations $\ell = \lambda \circ \psi^{-1} = \mu \circ \phi^{-1}$ where λ and μ are proper losses with corresponding Bayes risks $\underline{\Lambda}$ and $\underline{M}$ respectively, and ψ and ϕ are continuous invertible link functions. Then $\lambda = \mu$ almost everywhere.*

If ℓ is continuous and has a composite representation, then the proper loss (in the decomposition) is unique ($\lambda = \mu$ everywhere).

If ℓ is invertible and has a composite representation, then the representation is unique.

Proof $\underline{\Lambda}(p) = \inf_q p' \cdot \lambda(q) = \inf_q p' \cdot \ell(\psi(q)) = \inf_v L(p,v)$ (since ψ is invertible)
$= \inf_v L(p,v) = \inf_v L(p, \phi(q)) = \underline{M}(p)$.

Then λ and μ are two proper losses which have the same Bayes risk, so these two losses are equal almost everywhere.

If moreover ℓ is continuous, $\lambda = \ell \circ \psi$ and $\mu = \ell \circ \phi$ are continuous. So $\lambda = \mu$ everywhere.

If moreover ℓ is invertible, $\psi = \lambda \circ \ell^{-1}$ and $\phi = \mu \circ \ell^{-1}$. So ψ and ϕ are also equal almost everywhere and as they are continuous, they are equal everywhere. So $\lambda = \ell \circ \psi = \ell \circ \phi = \mu$. ∎

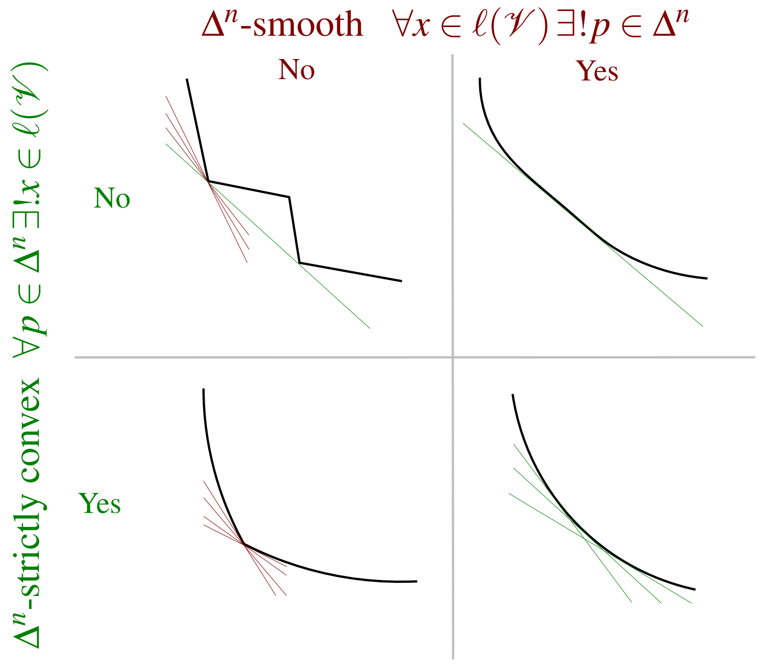
Δ^n-smooth $\forall x \in \ell(\mathscr{V}) \exists! p \in \Delta^n$
No
Yes
Δ^n-strictly convex $\forall p \in \Delta^n \exists! x \in \ell(\mathscr{V})$
No
Yes

Figure 3: Illustration of Δ^n-smoothness and Δ^n-strict convexity. The hyperplanes witness the possession or non-possession of the respective properrties.

Characterising the existence of a composite representation is more complex in the multiclass case. We need to introduce some definitions: We make use of a set of *hyperplanes* for $p \in \Delta^n$ and $\beta \in \mathbb{R}$,

$$h_p^\beta := \{x \in \mathbb{R}^n : x' \cdot p = \beta\}.$$

A hyperplane h_p^β *supports* a set A at $x \in A$ when $x \in h_p^\beta$ and for all $a \in A$, $a' \cdot p \geq \beta$ or for all $a \in A$, $a' \cdot p \leq \beta$. Given a loss $\ell \colon \mathscr{V} \to \mathbb{R}_+^n$, the *loss image* $\ell(\mathscr{V}) := \{\ell(v) \colon v \in \mathscr{V}\}$.

Definition 14 *Let $\mathfrak{S}(p,x) :=$ "$\ell(\mathscr{V})$ is supported by h_p^β at x for some $\beta \in \mathbb{R}$."*

1. *A loss image $\ell(\mathscr{V})$ is Δ^n-strictly convex if for all $p \in \Delta^n$ there exists a unique $x \in \ell(\mathscr{V})$ such that $\mathfrak{S}(p,x)$.*

2. *A loss image $\ell(\mathscr{V})$ is Δ^n-smooth if for all $x \in \ell(\mathscr{V})$ there exists a unique $p \in \Delta^n$ such that $\mathfrak{S}(p,x)$.*

This definition is illustrated in Figure 3. Dropping the uniqueness requirement in these definitions would drastically change things: since we will require ℓ is continuous, $\ell(\mathscr{V})$ is always closed. Since by assumption $\ell(\mathscr{V}) \subset [0,\infty)^n$ *every* such loss satisfies the weakened version of Δ^n-strict convexity: for all $p \in \Delta^n$ there exists $x \in \ell(\mathscr{V})$ such that $\mathfrak{S}(p,x)$. The weakened version of Δ^n-smoothness requires that for all $x \in \ell(\mathscr{V})$ there exists $p \in \Delta^n$ such that $\mathfrak{S}(p,x)$ is

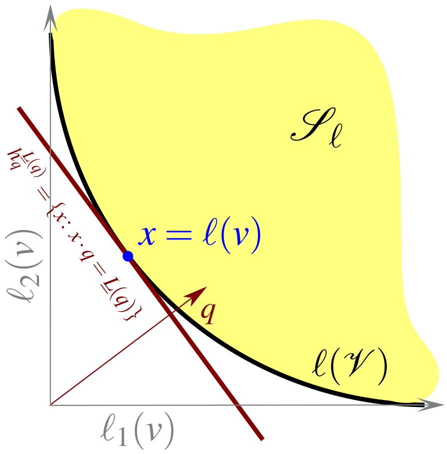

Figure 4: Illustration of geometry of loss functions. The locus of the vector valued loss ℓ is plotted as v varies over $\mathscr{V}$. The superprediction set $\mathscr{S}_\ell$ is the region to the "north-east" of the loss image $\ell(\mathscr{V})$. The hyperplane $h_q^{\underline{L}(q)}$ has normal vector q and offset $\underline{L}(q)$. It supports $\mathscr{S}_\ell$ at the point $x = \ell(v)$ indicating the Bayes risk is achieved at v for the true probability q.

a convexity-like requirement. (Confer the following result (Schneider, 1993, Theorem 1.3.3): *Suppose A is closed set such that $\mathring{A} \neq \varnothing$ and through each boundary point of A there is a support plane to A; then A is convex.*)

The name "Δ^n-strictly convex" is justified by the observation that replacing Δ^n by $B_{l_1^n}$ (the l_1^n unit ball) gives a natural definition of strict convexity of a general set in $\mathbb{R}^n$. We also observe that both Δ^n-strict convexity and Δ^n-smoothness are closely related to the curvature of the Bayes risk $\underline{L}$ by way of the fact that the support function of the set $\ell(\mathscr{V})$ (restricted to Δ^n) is the Bayes risk; confer (Williamson, 2014). Specifically, Δ^n-strict convexity is equivalent to the Hessian $\mathsf{H}\underline{L}(p)$ being non-singular for all $p \in \Delta^n$ while Δ^n-smoothness is implied whenever $\underline{L}(p)$ is continuously differentiable.

Suppose $A, B \subset \mathbb{R}^n$. Then the *Minkowski sum* $A + B := \{a + b \colon a \in A, b \in B\}$.

Definition 15 *Given a loss $\ell \colon \mathscr{V} \to \mathbb{R}_+^n$, we denote by*

$$\mathscr{S}_\ell := \ell(\mathscr{V}) + [0,\infty)^n = \{x \in \mathbb{R}_+^n \colon \exists v \in \mathscr{V},\ \forall i \in [n],\ x_i \geq \ell_i(v)\}$$

the superprediction set of ℓ (Kalnishkan and Vyugin, 2008).

One can characterise the existence of proper composite representations in terms of properties the superprediction set. We start with an old result; confer (Dawid, 2007).

Proposition 16 *Every continuous proper loss has a convex superprediction set.*

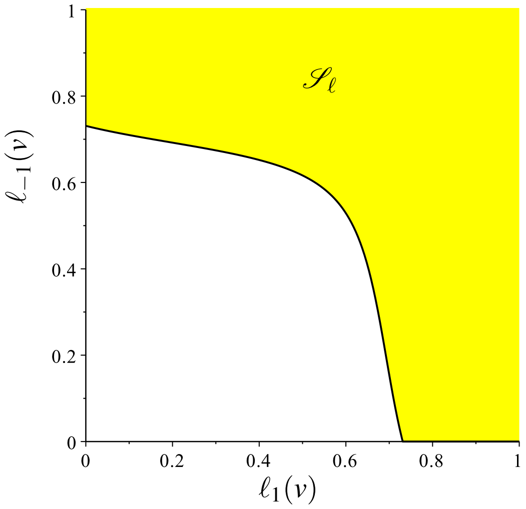

Figure 5: Superprediction set of a binary margin loss which is a not a composite proper loss; See text following Corollary 12.

Proof Suppose ℓ is proper but $\mathscr{S}_\ell$ is not convex. Then there exists $x_0 \in \ell(\Delta^n)$ such that $\ell(\Delta^n)$ is not supported at x_0 by any hyperplane h_p with normal vector $p \in \Delta^n$. Let $q_0 \in \Delta^n$ be such that $\ell(q_0) = x_0$. Then there is a hyperplane h_{q_0} (with normal q_0) that supports $\ell(\Delta^n)$ at some $x_1 \neq x_0$. Thus $q_0'\ell(q)$ is minimised at q_1 and not minimised at q_0 and thus ℓ is can not be proper—a contradiction. ■

The geometry of continuous proper losses is illustrated (for $n = 2$) in Figure 4. The superprediction set of the margin loss discussed following Corollary 12 is *not* convex as can be seen in Figure 5.

Continuous proper losses are quasiconvex, canonically so, as the following result shows.

Proposition 17 *Suppose $\ell\colon \Delta^n \to \mathbb{R}^n_+$ is a continuous proper loss. Then its superprediction set $\mathscr{S}_\ell$ is convex and, for all $p \in \Delta^n$, the function $f_p(q) := L(p,q) = p' \cdot \ell(q)$ is quasi-convex. Conversely, suppose $f_p(q) := p' \cdot \ell(q)$ is quasi-convex in q for all $p \in \Delta^n$. Then there is a unique convex set S such that $\mathscr{T}_\ell = S$ and ℓ is necessarily proper.*

The proof is in Appendix B.6. Some (but not all) proper losses are in addition *convex*; this is studied in more detail in Section 6.4 below.

Working with $\mathscr{S}_\ell$ is problematic for characterising the existence of *strictly* proper composite representations (essentially because while for a strictly proper loss ℓ, $\ell(\Delta^n)$ is Δ^n-strictly convex, $\mathscr{S}_\ell$ is not strictly convex (because of the flat spots at the extremes—bounded losses have superprediction sets with flats parallel to the axes by construction)[1]. We will thus characterise proper and strictly proper composite representations in terms of properties of $\ell(\mathscr{V})$ rather than $\mathscr{S}_\ell$.

[1] It turns out that by *starting* with the superprediction set, and defining the loss in terms of the (super-) gradient of the (concave) support function of the superprediction set, these difficulties can be avoided (Williamson, 2014).

Proposition 18 *Suppose $\ell\colon \mathscr{V} \to \mathbb{R}^n_+$ is a continuous loss. ℓ has a proper composite representation if and only if $\ell(\mathscr{V})$ is Δ^n-smooth. Additionally, ℓ is* strictly *proper composite if and only if $\ell(\mathscr{V})$ is also Δ^n-strictly convex.*

The proof is in Section B.7.

6. Implications: Mixability, Admissibility, Minimaxity and Convexity

We now consider some of the implications that the proper composite representation has for several previously studied properties of loss functions.

6.1 Mixability

Mixability is a fundamental property of a loss function in the study of "prediction with expert advice." In this setting learning takes place in fixed number of sequential rounds. Each round a learner is presented with predictions some finite number of experts. The learner then makes a prediction and the outcome for that round is revealed. The learner's and experts' predictions are assessed using some predefined loss function and the aim of the learner is to incur a total loss not much worse than the best expert – *i.e.*, the one with the smallest total loss. The difference between the learner's total loss and that of the best expert is known as the *regret*. In his seminal work, Vovk (1995) showed that no matter how the experts behave, there exists a strategy for the learner (called the "aggregating algorithm") that guarantees a regret bounded by $\frac{\ln K}{\eta}$ where K is the number of experts and η is a positive number called the *mixability constant* (defined below) that only depends on the loss. Losses for which this constant is defined are called *mixable*. Furthermore, this constant *characterises* when such a constant regret bound is possible. That is, if a loss is not mixable then there is no strategy the learner can use to guarantee a constant regret bound.

Formally, mixability of a loss ℓ is defined in terms of the convexity of a transformation of the loss's superprediction set $\mathscr{S}_\ell$ (see Definition 15). We say that for $\eta > 0$ the *η-exponentiated superprediction set* is the image of $\mathscr{S}_\ell \subset \mathbb{R}^n$ under the mapping $E_\eta\colon \mathbb{R}^n \to \mathbb{R}^n_+$ defined by $E_\eta(x) := (e^{-\eta x_i})_{i=1}^n$. A loss ℓ is said to be *η-mixable* if its η-exponentiated superprediction set is convex. The *mixability of ℓ* is the smallest value of η for which ℓ is η-mixable. For further details, the reader is referred to papers by Vovk (1995); Kalnishkan and Vyugin (2008); Vovk and Zhdanov (2009).

Recently, van Erven et al. (2012b)[2] have shown that the mixability of a loss is related to the curvature of the loss's Bayes risk relative to the curvature of the Bayes risk for log loss. The main result here builds on some of the insights from that work and shows that mixable losses (under mild conditions) always have proper composite representations.

For $\alpha \in (0,1)$ we write $\bar{\alpha} := 1-\alpha$. For $x, y \in \mathbb{R}^n$, $x \leq y \Leftrightarrow (x_i \leq y_i,\ \forall i \in [n])$. We now give a necessary condition for mixability.

[2] An extension of this notion of mixability can be related to a natural convex duality (Reid et al., 2015).

Lemma 19 *Suppose* $\ell\colon \mathscr{V} \to \mathbb{R}^n_+$, $x_0 = \ell(v_0)$, $x_1 = \ell(v_1)$ *with* $x_0 \neq x_1$. *For* $\alpha \in (0,1)$, *define* $x_\alpha := \bar{\alpha}x_0 + \alpha x_1$ *and* $v_\alpha = \bar{\alpha}v_0 + \alpha v_1$. *If for some* α

$$x_\alpha \leq \ell(v_\alpha) \tag{4}$$

then ℓ *is not mixable.*

Proof Pick some $\eta > 0$. Let $f_\eta(a) = e^{-\eta a}$ for $a \in \mathbb{R}$ so that for $x \in \mathbb{R}^n$ we have $E_\eta(x) = (f_\eta(x_i))_{i=1}^n$. Observe that the function f_η is strictly monotone decreasing ($a < b \Rightarrow f_\eta(a) > f_\eta(b)$) and strictly convex ($\bar{\alpha}f_\eta(a) + \alpha f_\eta(b) > f_\eta(\bar{\alpha}a + \alpha b)$). For $i \in [n]$ set $x_{0,i} = \ell_i(v_0)$ and $x_{1,i} = \ell_i(v_1)$. By assumption, we have

$$\bar{\alpha}x_{0,i} + \alpha x_{1,i} \leq \ell_i(\bar{\alpha}v_0 + \alpha v_1),\ \forall i \in [n],$$

which by strict monotonicity

$$\Rightarrow f_\eta(\bar{\alpha}x_{0,i} + \alpha x_{1,i}) \geq f_\eta(\ell_i(\bar{\alpha}v_0 + \alpha v_1)),\ \forall i \in [n],$$

and hence by strict convexity

$$\begin{aligned}&\Rightarrow \bar{\alpha}f_\eta(x_{0,i}) + \alpha f_\eta(x_{1,i}) > f_\eta(\ell_i(\bar{\alpha}v_0 + \alpha v_1)),\ \forall i \in [n]\\ &\Leftrightarrow \bar{\alpha}E_\eta(\ell(v_0)) + \alpha E_\eta(\ell(v_1)) > E_\eta(\ell(\bar{\alpha}v_0 + \alpha v_1))\end{aligned}$$

and thus ℓ is not mixable since we have witnessed the non-convexity of the η-exponentiated superprediction set for ℓ. ■

van Erven et al. (2012b) showed that (under some mild conditions) a proper loss λ and the composite loss λ^ψ obtained via the reference link $\tilde{\psi}$ (see Proposition 5) share the same mixability constant. We now show that mixable losses always have strictly proper composite representations.

Proposition 20 *Suppose* $\ell\colon \mathscr{V} \to \mathbb{R}^n_+$ *is a* Δ^n*-smooth continuous loss. If* ℓ *is mixable then* ℓ *has a strictly proper composite representation.*

Proof We prove the contrapositive. Lack of a strictly proper composite representation is equivalent then to $\ell(\mathscr{V})$ being not Δ^n-strictly convex. Suppose then that $\ell(\mathscr{V})$ is indeed not Δ^n-strictly convex. There are two possibilities to consider:

1. There exists $p \in \Delta^n$ such that there is no $x \in \ell(\mathscr{V})$ such that $\ell(\mathscr{V})$ is supported by h_p^β at x for some $\beta \in \mathbb{R}$; or
2. There exists $p \in \Delta^n$ such that there exists $v_0, v_1 \in \mathscr{V}$, $v_0 \neq v_1$, $x_0 = \ell(v_0)$, $x_1 = \ell(v_1)$, $\exists \beta \in \mathbb{R}$, h_p^β supports $\ell(\mathscr{V})$ at x_1 and x_2.

Since $\ell(\mathscr{V}) \subset [0,\infty)^n$ and ℓ is continuous (and hence $\ell(\mathscr{V})$ is closed), for all $p \in \Delta^n$ there *always* exists $x \in \ell(\mathscr{V})$ such that h_p^β supports $\ell(\mathscr{V})$ at x. Thus under the hypothesis, case 2 must always hold. Then by continuity of ℓ and the definition of a supporting hyperplane, there exists $\alpha \in (0,1)$ such that (4) holds and so ℓ is not mixable. ■

6.2 Admissibility

The above results are strongly related to the classical notion of admissibility (Ferguson, 1967; Chernoff and Moses, 1986; Kiefer, 1987), which is particularly simple in our situation. We adapt the terminology of Ferguson (1967) to be consistent with elsewhere in the present paper.

Definition 21 *Suppose $\ell\colon \mathcal{V} \to \mathbb{R}^n_+$ is a loss. A prediction $v_1 \in \mathcal{V}$ is* better than *$v_2 \in \mathcal{V}$ if $\ell(v_1) \leq \ell(v_2)$ and for some $i \in [n]$, $\ell_i(v_1) < \ell_i(v_2)$. A prediction v_1 is* equivalent *to v_2 if $\ell(v_1) = \ell(v_2)$. A prediction $v \in \mathcal{V}$ is* admissible *if there is no prediction better than v. If a prediction $v \in \mathcal{V}$ is the Bayes-optimal for some distribution p, that is for all $v \in \mathcal{V}$ there exists $p \in \Delta^n$ such that $v \in \arg\min_{\bar{v}\in\mathcal{V}} p' \cdot \ell(\bar{v})$, then we say v is* strongly admissible.

Ferguson (1967, Theorem 1, page 60) states the following (which we present for invertible losses, so that $\ell(v_1) = \ell(v_2) \Rightarrow v_1 = v_2$).

Proposition 22 *Suppose $\ell\colon \mathcal{V} \to \mathbb{R}^n_+$ is invertible and $p \in \Delta^n$. If $v \in \mathcal{V}$ is the unique prediction such that $L(p,v) = \underline{L}(p)$, then v is admissible.*

Proposition 18 then implies the following.

Corollary 23 *Suppose $\ell\colon \mathcal{V} \to \mathbb{R}^n_+$ is continuous and invertible. If ℓ has a strictly proper composite representation then all $v \in \mathcal{V}$ are admissible and strongly admissible.*

Proof If ℓ has a strictly proper composite representation, then $\ell(\mathcal{V})$ is Δ^n-strictly convex and thus for all $p \in \Delta^n$ there exists a unique $x \in \ell(\mathcal{V})$ such that $h_p^{\underline{L}(p)}$ supports $\ell(\mathcal{V})$ at x. Thus by Proposition 22, v such that $\ell(v) = x$ is an admissible prediction. Furthermore, since $\ell(\mathcal{V})$ is Δ^n-smooth, this previous argument actually holds for all $v \in \mathcal{V}$ and thus ℓ is admissible. Furthermore, it follows directly from the definition of Δ^n-smoothness that all v are strongly admissible. ∎

Proposition 24 *If $\ell\colon \mathcal{V} \to \mathbb{R}^n_+$ is continuous and has a proper composite representation then every prediction is admissible.*

Proof We will prove the contrapositive: Suppose a continuous loss $\ell\colon \mathcal{V} \to \mathbb{R}^n_+$ is such that there exist $x_0, x_1 \in \ell(\mathcal{V})$ with x_1 better than x_0. Then ℓ can not have a proper composite representation. Observe that "x_1 is better than x_0" is equivalent to

$$\begin{aligned} &\forall i \in [n],\ e_i' \cdot (x_0 - x_1) \geq 0 \\ &\exists i \in [n],\ e_i' \cdot (x_0 - x_1) > 0. \end{aligned}$$

Consider two mutually exclusive and exhaustive cases:

1. $e_i' \cdot (x_0 - x_1) > 0$, $\forall i \in [n]$. Then for all $p \in \Delta^n$, $p' \cdot (x_0 - x_1) > 0 \Rightarrow p' \cdot x_0 > p' \cdot x_1$ and thus $\ell(\mathcal{V})$ can not be supported at x_0 by h_p^{β} for any $p \in \Delta^n$ and thus $\ell(\mathcal{V})$ is not Δ^n-smooth.

2. Alternatively suppose

$$e_i' \cdot (x_0 - x_1) \begin{cases} = 0, & i \in I \subset [n] \\ > 0, & i \in [n] \setminus I \end{cases}$$

with $1 \leq |I| < n$. Consider the two mutually exclusive subcases over $p \in \Delta^n$:

(a) $p_i > 0$ for *some* $i \in [n] \setminus I$. Then $p' \cdot (x_0 - x_1) > 0$ and $\ell(\mathscr{V})$ can not be supported at x_0 by h_p^β for any $\beta \in \mathbb{R}$.

(b) $p_i - 0$ for *all* $i \in [n] \setminus I$. Then $p' \cdot (x_0 - x_1) = 0$ in which case $\ell(\mathscr{V})$ is supported by h_p^β for some β at *both* x_0 and x_1.

In either of these subcases, the Δ^n-smoothness condition is violated.

Thus in both cases we have shown $\ell(\mathscr{V})$ can not be Δ^n-smooth and by Proposition 18 can not have a proper composite representation. ■

As can be seen in Figure 6, there can be no hope of a converse: mere admissibility of every prediction $x \in \ell(\mathscr{V})$ can not imply that ℓ has a proper composite representation.

However strong admissibility of every prediction implies $\ell(\mathscr{V})$ is Δ^n-smooth and so if ℓ is continuous, strong admissibility of every prediciton implies (via Proposition 18) that ℓ has a proper composite representation.

The relationship between strict convexity of $\mathscr{S}_\ell$ and admissibility is not new (Brown, 1981); but the connection with our characterisation of composite proper losses is new.

We conclude that if ℓ is continuous and invertible and we desire that all predictions are admissible, then it suffices to only consider losses with a proper composite representation. Continuous invertible losses that do not have a proper composite representation are "redundant" in the sense that there are guaranteed to exist predictions that are not Bayes optimal for any true distribution.

6.3 Minimaxity

We say a loss $\ell \colon \mathscr{V} \to \mathbb{R}_+^n$ is *minimax* if its conditional risk $L(p,v) = p' \cdot \ell(v)$ satisfies

$$\max_{p \in \Delta^n} \min_{v \in \mathscr{V}} L(p,v) = \min_{v \in \mathscr{V}} \max_{p \in \Delta^n} L(p,v). \tag{5}$$

Minimaxity of proper losses has been studied in a very general setting by Grünwald and Dawid (2004) who showed the connection between robust Bayes procedures and maximum entropy; confer classical results presented, for example, by Ferguson (1967). In this brief subsection we point out some simple implications of our earlier results. Setting $\mathscr{V} = \Delta^n$, oberve that for all proper losses $\lambda \colon \Delta^n \to \mathbb{R}_+^n$, $p \mapsto \Lambda(p,q) = p' \cdot \lambda(q)$ is linear for all $q \in \Delta^n$, and if λ is also continuous, by Proposition 17 $q \mapsto \Lambda(p,q)$ is quasi-convex for all $p \in \Delta^n$. It thus follows from the minimax theorem of Sion (1958) that all continuous proper losses satisfy

$$\max_{p \in \Delta^n} \min_{q \in \Delta^n} \Lambda(p,q) = \min_{q \in \Delta^n} \max_{p \in \Delta^n} \Lambda(p,q) \tag{6}$$

and are thus minimax.

Suppose $\ell = \lambda^\psi = \lambda \circ \psi^{-1} \colon \mathscr{V} \to \mathbb{R}_+^n$ is a proper composite loss, with conditional risk $L(p,v) = \Lambda(p, \psi^{-1}(v))$. Since ψ^{-1} is invertible,

$$\max_{p \in \Delta^n} \min_{v \in \mathscr{V}} L(p,v) = \max_{p \in \Delta^n} \min_{q \in \Delta^n} \Lambda(p,q), \tag{7}$$

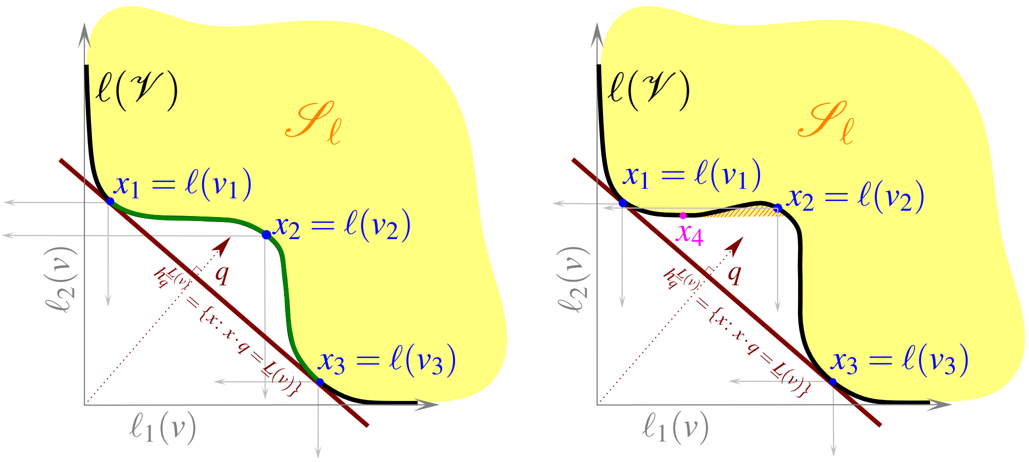

Figure 6: *Left:* Illustration of a continuous loss ℓ (which can be presumed invertible) with a non-convex superprediction set. For true probability q, v_1 and v_2 both are Bayes optimal since $q' \cdot \ell(v_1) = q' \cdot \ell(v_2) = \underline{L}(v)$; thus $h_q^{\underline{L}(v_1)} = h_q^{\underline{L}(v_3)}$ supports $\ell(\mathscr{V})$ at both x_1 and x_3. The point x_2 is never a member of a supporting hyperplane of $\ell(\mathscr{V})$ and is thus never the Bayes optimal prediction for any q and so not strongly admissible. The green line indicates the set of predictions that are not strongly admissible—they will never be Bayes optimal for *any* $q \in \Delta^n$. Such predictions are, however, admissible, as can be seen by the grey translated negative orthants centred at x_1, x_2 and x_3 (each orthant does not contain any other predictions "better than" them). All the other predictions whose image lies in the black line are both admissible and strongly admissible. The loss image $\ell(\mathscr{V})$ is not Δ^n-smooth because there exist no $p \in \Delta^n$ that supports $\ell(\mathscr{V})$ at x_2. Hence by Proposition 18, ℓ can not have a proper composite representation. *Right:* Similar to the figure on the left, except there are now some predictions, such as x_2, which are not admissible: x_4 is *better than* x_2 as can be seen since x_4 is contained in the interior of the shifted negative orthant centred at x_2. Note in this case the boundary of the super-prediction set $\mathscr{S}_\ell$ does not equal $\ell(\mathscr{V})$ (see the part of $\mathscr{S}_\ell$ cross-hatched in red). This loss can not have a proper composite representation by Proposition 24.

where by the relationship between q and v, $\arg\min_{v \in \mathscr{V}} L(p,v) = \psi\left(\arg\min_{q \in \Delta^n} \Lambda(p,q)\right)$. Similarly,

$$\min_{v \in \mathscr{V}} \max_{p \in \Delta^n} L(p,v) = \min_{q \in \Delta^n} \max_{p \in \Delta^n} \Lambda(p,q). \tag{8}$$

Since λ is proper, Λ satisfies (6) which combined with (7) and (8) proves the following.

Proposition 25 *Every continuous proper composite loss is minimax.*

Note that this alone does not imply that all continuous proper composite losses are quasi-convex, which would follow if ψ mapped convex sets to convex sets; however this can not be true in general in $\mathbb{R}^n$ because convexity preserving mappings must be affine (Webster, 1994, Theorem 7.3.7); confer (Meyer and Kay, 1973). Recall Proposition 17 showed the quasi-convexity of all proper losses.

Proposition 25 means that the use of the classical minimax theorem by Abernethy et al. (2009) in order to prove their main result for *convex* losses can be foregone; their result also holds for arbitrary continuous proper composite losses.

6.4 Convexity

In order to computationally optimise models with respect to a loss function it is convenient if the loss is convex. In this subsection we develop conditions for the convexity of multiclass composite proper losses. We assume throughout this section that the loss and link are twice differentiable. We start by proving some identities for their first and second derivatives.

6.4.1 TECHNICAL PRELIMINARIES

Suppose $\ell = \lambda \circ \psi^{-1}$ is composed of the proper loss $\lambda\colon \Delta^n \to \mathbb{R}^n_+$ and the inverse of the link $\psi\colon \Delta^n \to \mathscr{V}$. In order to simplify the calculation of derivatives for the function $\ell\colon \mathscr{V} \to \mathbb{R}^n_+$ we will assume the set $\mathscr{V}$ is a flat, $(n-1)$-dimensional, convex subset of $\mathbb{R}^n_+$. We do so since if $\mathscr{V}$ were some arbitrary manifold the extra definitions required to make sense of convexity (*e.g.*, in terms of geodesics) and derivatives on manifolds would obscure the gist of the results below. Furthermore, little is lost either practically or theoretically by assuming a simple $\mathscr{V}$. In practice, predictions are usually vectors in $\mathbb{R}^n_+$, and in theory one could always choose a parametrisation of $\mathscr{V}$ in terms of some simpler space $\mathscr{U}$ and redefine the link via composition with that parametrisation. Alternatively, since links must be invertible, a composite loss could be defined by a choice of loss and choice of *inverse link* $\psi^{-1}\colon \mathscr{V} \to \Delta^n$ for a $\mathscr{V}$ assumed to be flat, *etc*.

Recalling the convention that $\tilde{n} := n-1$, let $v \in \mathscr{V}$ fixed but arbitrary with corresponding $\tilde{p} = \tilde{\psi}^{-1}(v)$ where $\tilde{\psi}(\tilde{p}) := \psi((\tilde{p}_1, \ldots, \tilde{p}_{\tilde{n}}, p_n)')$ with $p_n := \sum_{i=1}^{\tilde{n}} \tilde{p}_i$ is the induced function from $\tilde{\Delta}^n$ to $\mathscr{V}$. By the chain rule and the inverse function theorem, the derivatives for each of the partial losses ℓ_i satisfy

$$\mathsf{D}\ell_i(v) = \mathsf{D}\left[\lambda_i(\tilde{\psi}^{-1}(v))\right] = \mathsf{D}\lambda_i(\tilde{p}) \cdot [\mathsf{D}\tilde{\psi}(\tilde{p})]^{-1}. \tag{9}$$

We use e_i^n to denote the ith n-dimensional unit vector, $e_i^n = (0,\ldots,0,1,0,\ldots,0)'$ when $i \in [n]$, and define $e_i^n = 0_n$ when $i > n$. We can now write $\mathsf{D}\lambda_i(\tilde{p})$ in terms of the $n \times \tilde{n}$ matrix $\mathsf{D}\lambda(\tilde{p})$ using $\mathsf{D}\lambda_i(\tilde{p}) = (e_i^n)' \cdot \mathsf{D}\lambda(\tilde{p})$. Now $\mathsf{D}\lambda(\tilde{p}) = (\mathsf{D}\tilde{\lambda}(\tilde{p})', \mathsf{D}\lambda_n(\tilde{p})')'$, where $\tilde{\lambda}(\tilde{p}) = (\lambda_1(\tilde{p}), \ldots, \lambda_{\tilde{n}}(\tilde{p}))'$, and so

$$\mathsf{D}\lambda_i(\tilde{p}) = (e_i^n)' \cdot \mathsf{D}\lambda(\tilde{p}) = (e_i^n)' \cdot \begin{pmatrix} \mathsf{D}\tilde{\lambda}(\tilde{p}) \\ \mathsf{D}\lambda_n(\tilde{p}) \end{pmatrix}. \tag{10}$$

Furthermore, since λ is proper, Lemma 6 of (van Erven et al., 2012b) means we can use the relationship between a proper loss and its projected Bayes risk $\tilde{\underline{L}} := \underline{L} \circ \Pi_\Delta^{-1}$ to write

$$\mathsf{D}\tilde{\lambda}(\tilde{p}) = W(\tilde{p}) \cdot \mathsf{H}\tilde{\underline{L}}(\tilde{p}) \tag{11}$$

$$\mathsf{D}\lambda_n(\tilde{p}) = y(\tilde{p})' \cdot \mathsf{D}\tilde{\lambda}(\tilde{p}) \tag{12}$$

where $W(\tilde{p}) := I_{\tilde{n}} - \mathbb{1}_{\tilde{n}} \cdot \tilde{p}'$ and where $y(\tilde{p}) := -\tilde{p}/p_n(\tilde{p})$ and $p_n(\tilde{p}) := 1 - \sum_{i \in [\tilde{n}]} p_i$.

Thus, combining (10–12) we have for all $i \in [\tilde{n}]$

$$\begin{aligned} \mathsf{D}\lambda_i(\tilde{p}) &= (e_i^{\tilde{n}})' \cdot W(\tilde{p}) \cdot \mathsf{H}\underline{\tilde{L}}(\tilde{p}) \\ &= ((e_i^{\tilde{n}})' - (e_i^{\tilde{n}})' \cdot \mathbb{1}_{\tilde{n}} \cdot \tilde{p}') \cdot \mathsf{H}\underline{\tilde{L}}(\tilde{p}) \\ &= (e_i^{\tilde{n}} - \tilde{p})' \cdot \mathsf{H}\underline{\tilde{L}}(\tilde{p}) \end{aligned} \tag{13}$$

and

$$\begin{aligned} \mathsf{D}\lambda_n(\tilde{p}) &= y(\tilde{p})' \cdot W(\tilde{p}) \cdot \mathsf{H}\underline{\tilde{L}}(\tilde{p}) \\ &= \frac{-1}{p_n(\tilde{p})} \tilde{p}' \cdot (I_{\tilde{n}} - \mathbb{1}_{\tilde{n}} \cdot \tilde{p}') \cdot \mathsf{H}\underline{\tilde{L}}(\tilde{p}) \\ &= \frac{-1}{p_n(\tilde{p})} (\tilde{p}' - (1 - p_n(\tilde{p}))\tilde{p}') \cdot \mathsf{H}\underline{\tilde{L}}(\tilde{p}) \\ &= -\tilde{p}' \cdot \mathsf{H}\underline{\tilde{L}}(\tilde{p}). \end{aligned} \tag{14}$$

Finally, noting that by definition $e_n^{\tilde{n}} = 0$, (14) and (13) can be merged and combined with (9) to obtain the following proposition.

Proposition 26 *For all $i \in [n]$, $\tilde{p} \in \mathring{\tilde{\Delta}}^n$ (the relative interior of $\tilde{\Delta}^n$), and $v = \tilde{\psi}(\tilde{p})$,*

$$D\ell_i(v) = -\left(e_i^{\tilde{n}} - \tilde{p}\right)' \cdot \kappa(\tilde{p}) \tag{15}$$

where

$$\kappa(\tilde{p}) := -\mathsf{H}\underline{\tilde{L}}(\tilde{p}) \cdot [\mathsf{D}\tilde{\psi}(\tilde{p})]^{-1}. \tag{16}$$

Using the definition of the Hessian $\mathsf{H}\ell_i = \mathsf{D}[(\mathsf{D}\ell_i)']$ and the product rule (31) gives

$$\begin{aligned} \mathsf{D}\left[(\mathsf{D}\ell_i(v))'\right] =& \mathsf{D}_v[\overbrace{\left[\mathsf{D}\tilde{\psi}(\tilde{p})'\right]^{-1} \cdot \mathsf{H}\underline{\tilde{L}}(\tilde{p})'}^{f(\tilde{p})} \cdot \overbrace{\left(e_i^{\tilde{n}} - \tilde{p}\right)}^{g(\tilde{p})}] \\ =& \left(\left(e_i^{\tilde{n}} - \tilde{p}\right)' \otimes I_{\tilde{n}}\right) \cdot \mathsf{D}_v[f(\tilde{p})' + (I_1 \otimes f(\tilde{p})) \cdot \mathsf{D}\left(e_i^{\tilde{n}} - \tilde{\psi}^{-1}(v)\right) \\ =& \left(\left(e_i^{\tilde{n}} - \tilde{p}\right)' \otimes I_{\tilde{n}}\right) \cdot \mathsf{D}_v\left[\mathsf{H}\underline{\tilde{L}}(\tilde{p}) \cdot [\mathsf{D}\tilde{\psi}(\tilde{p})]^{-1}\right] - \left(\left[\mathsf{D}\tilde{\psi}(\tilde{p})'\right]^{-1} \mathsf{H}\underline{\tilde{L}}(\tilde{p})'\right) \cdot [\mathsf{D}\tilde{\psi}(\tilde{p})]^{-1}, \end{aligned}$$

where D_v is used to indicate that the derivative is with respect to v even when the terms inside the derivative are expressed using $\tilde{p}$. We have now established the following proposition.

Proposition 27 *For all $i \in [n]$, $\tilde{p} \in \mathring{\tilde{\Delta}}^n$, and $v = \tilde{\psi}(\tilde{p})$,*

$$\mathsf{H}\ell_i(v) = -\left(\left(e_i^{\tilde{n}} - \tilde{p}\right)' \otimes I_{\tilde{n}}\right) \cdot \mathsf{D}\left[\kappa\left(\tilde{\psi}^{-1}(v)\right)\right] + \left(\kappa(\tilde{p})'\right) \cdot [\mathsf{D}\tilde{\psi}(\tilde{p})]^{-1},$$

where $\kappa(\tilde{p})$ is defined in (16).

The product $\kappa(\tilde{p}) := -\mathsf{H}\underline{\tilde{L}}(\tilde{p})\,[\mathsf{D}\tilde{\psi}(\tilde{p})]^{-1}$ that appears in both propositions above can be interpreted as the curvature of the Bayes risk function $\underline{\tilde{L}}$ relative to the rate of change of the link function $\tilde{\psi}$. When the link function is the identity $\tilde{\psi}(\tilde{p}) = \tilde{p}$ (*i.e.* when we have a proper loss directly) the expressions for the derivative and Hessian of each ℓ_i simplify to

$$\mathsf{D}\ell_i(\tilde{p}) = (e_i^{\tilde{n}} - \tilde{p})' \cdot \mathsf{H}\underline{\tilde{L}}(\tilde{p}) \tag{17}$$

$$\mathsf{H}\ell_i(\tilde{p}) = \left(\left(e_i^{\tilde{n}} - \tilde{p}\right)' \otimes I_{\tilde{n}}\right) \cdot \mathsf{D}\left[\mathsf{H}\underline{\tilde{L}}(\tilde{p})\right] - \mathsf{H}\underline{\tilde{L}}(\tilde{p})'. \tag{18}$$

The form of κ as the product of $\mathsf{H}\underline{\tilde{L}}$ and $\mathsf{D}\tilde{\psi}$ suggests another simplification.

Definition 28 *The canonical link function for a loss λ with Bayes risk $\underline{L}$ is defined via*

$$\tilde{\psi}_\lambda(\tilde{p}) := -\mathsf{D}\underline{\tilde{L}}(\tilde{p})'. \tag{19}$$

We will show in section 8.1 that (19) is indeed guaranteed to be a legitimate link. The term κ simplifies to $\kappa(\tilde{p}) = I_{\tilde{n}}$ since $\mathsf{D}\tilde{\psi}(\tilde{p}) = -\mathsf{D}(\mathsf{D}\underline{\tilde{L}}(\tilde{p})') = -\mathsf{H}\underline{\tilde{L}}(\tilde{p})$. For this choice of link function, the first and second derivatives become considerably simpler.

Proposition 29 *If $\lambda : \Delta^n \to \mathbb{R}_+^n$ is a proper loss and $\tilde{\psi}_\lambda$ is its associated canonical link then, for all $i \in [n]$, $\tilde{p} \in \mathring{\tilde{\Delta}}^n$, and $v = \tilde{\psi}_\lambda(\tilde{p})$, the composite loss $\ell = \lambda \circ \tilde{\psi}$ satisfies*

$$\mathsf{D}\ell_i(v) = (e_i^{\tilde{n}} - \tilde{p}) \tag{20}$$

$$\mathsf{H}\ell_i(v) = \left[\mathsf{H}\underline{\tilde{L}}(\tilde{p})\right]^{-1}. \tag{21}$$

The simplified form of the Hessian above is established by noting that since $\kappa(\tilde{p}) = I_{\tilde{n}}$ we have $\mathsf{D}[\kappa(\tilde{\psi}^{-1}(v))] = 0$ for all $v \in \mathscr{V}$ in Proposition 27.

The above propositions hold for any number of classes n. It is instructive (both here and later in the paper) to examine the binary case where $n = 2$. In this case, Proposition 26 and Proposition 27 reduce to

$$\ell_1'(v) = -(1-\tilde{p})\kappa(\tilde{p}) \quad ; \quad \ell_2'(v) = \tilde{p}\kappa(\tilde{p}) \tag{22}$$

$$\ell_1''(v) = \frac{-(1-\tilde{p})\kappa'(\tilde{p}) + \kappa(\tilde{p})}{\tilde{\psi}'(\tilde{p})} \tag{23}$$

$$\ell_2''(v) = \frac{\tilde{p}\kappa'(\tilde{p}) + \kappa(\tilde{p})}{\tilde{\psi}'(\tilde{p})} \tag{24}$$

where $\kappa(\tilde{p}) = -\frac{\underline{\tilde{L}}''(\tilde{p})}{\tilde{\psi}'(\tilde{p})} \geq 0$ and so $\frac{d}{dv}\kappa(\tilde{\psi}^{-1}(v)) = \frac{\kappa'(\tilde{p})}{\tilde{\psi}'(\tilde{p})}$.

6.4.2 CONDITIONS FOR CONVEXITY OF MULTICLASS COMPOSITE PROPER LOSSES

We will now consider when multiclass proper losses are convex, and give a characterisation in terms of the corresponding Bayes risk which as we have seen is the natural way to parametrise a loss. The results below are the multiclass generalisation of the characterisation of convexity of binary composite losses (Reid and Williamson, 2010). In fact we obtain more general results even in the binary case because here we consider *strongly* convex losses. We will also show how any non-convex proper loss can be made convex by suitable choice of a link function (the canonical link)[3].

For a convex set $C \subseteq \mathbb{R}^n$, a loss $\ell \colon C \to \mathbb{R}^n_+$ is said to be *convex* if for all $p \in \Delta^n$, the map $C \ni v \mapsto L(p,v) = p' \cdot \ell(v)$ is convex. That is, a loss is convex if, under any distribution p over outcomes $i \in [n]$, the expected loss $\mathbb{E}_{\mathsf{i} \sim p}[\ell_{\mathsf{i}}(v)]$ is convex in v. It is easy to see that ℓ is convex if and only if $\ell_i \colon C \to \mathbb{R}_+$ is convex for all $i \in [n]$. (The "if" part follows since a sum of convex functions is convex; the "only if" follows by considering $p = e_i$, for $i \in [n]$.)

Definition 30 *Suppose $C \subseteq \mathbb{R}^n$ is convex. A function $f \colon C \to \mathbb{R}$ is strongly convex on C with modulus $c \geq 0$ if for all $x, x_0 \in C$, $\forall \alpha \in (0,1)$,*

$$f(\alpha x + (1-\alpha)x_0) \leq \alpha f(x) + (1-\alpha) f(x_0) - \frac{1}{2} c\alpha(1-\alpha)\|x - x_0\|^2.$$

When $c = 0$ in the above definition, f is convex. The function f is strongly convex on C with modulus c if and only if $x \mapsto f(x) - \frac{c}{2}\|x\|^2$ is convex on C (Hiriart-Urruty and Lemaréchal, 2001, page 73). Therefore, the maps $v \mapsto \ell_i(v)$ are c-strongly convex if and only if $\mathsf{H}\ell_i(v) \succcurlyeq cI_{\tilde{n}}$. By applying Proposition 27 we obtain the following characterisation of the c-strong convexity of the loss ℓ.

Proposition 31 *A proper composite loss $\ell = \lambda \circ \psi^{-1}$ is strongly convex with modulus $c \geq 0$ if and only if for all $\tilde{p} \in \mathring{\tilde{\Delta}}^n$ and for all $i \in [n]$*

$$\left(\left(e_i^{\tilde{n}} - \tilde{p}\right) \otimes I_{\tilde{n}}\right) \cdot \mathsf{D}\left(\kappa\left(\tilde{\psi}^{-1}(v)\right)\right) \preccurlyeq \kappa(\tilde{p})' \cdot \left[\mathsf{D}\tilde{\psi}(\tilde{p})\right]^{-1} - cI_{\tilde{n}}. \tag{25}$$

We now consider the implications of Proposition 31 in two special cases: in the multiclass case with canonical link, and in the binary case with the identity link.

Recall that the canonical link $\tilde{\psi}_\ell$ is chosen so that $\tilde{\psi}(\tilde{p}) = -\mathsf{D}\underline{\tilde{L}}(\tilde{p})'$. This simplifies $\kappa(\tilde{p})$ to the identity matrix $I_{\tilde{n}}$ so $\mathsf{D}\kappa(\tilde{p}) = 0$. In this case the above proposition reduces to the following corollary.

Corollary 32 *If $\ell = \lambda \circ \psi^{-1}$ is defined so that $\tilde{\psi} = -\mathsf{D}\underline{\tilde{L}}'$ then each map $v \mapsto \ell_i(v)$ is c-strongly convex if and only if $\left[-\mathsf{H}\underline{\tilde{L}}(\tilde{p})\right]^{-1} \succcurlyeq cI_{\tilde{n}}$, or equivalently $-\mathsf{H}\underline{\tilde{L}}(\tilde{p}) \preccurlyeq \frac{1}{c} I_{\tilde{n}}$.*

An immediate consequence of this result is obtained by observing that the definiteness constraint is always met when $c = 0$ since $\underline{\tilde{L}}$ is always a concave function. Thus, *using a canonical link guarantees a proper composite loss is convex.*

There is an upper definiteness condition analogous to that for strong convexity that has implications for rates of convergence in numerical optimisation. Boyd and Vandenberghe (2004, §9.1.2) show that if a twice differentiable function $f \colon \mathscr{X} \to \mathbb{R}$ satisfies

$$MI \succcurlyeq \mathsf{H}f(x) \succcurlyeq mI$$

[3] There are problems associated with the domain of definition of such link functions than need to be dealt with (Kamalaruban et al., 2015).

for all $x \in \mathscr{X} \subset \mathbb{R}^n$ then the value $\frac{M}{m}$ is an upper bound on the *condition number* of $\mathsf{H}f$, that is, the ratio of maximum to minimum eigenvalue of $\mathsf{H}f$. This value measures the eccentricity of the sublevel sets of f and controls the rate at which optima of f are approached.

Applying this result to the Hessian of a composite loss ℓ with a canonical link shows that the condition number bound is controlled by the Hessian of the Bayes risk of ℓ. Specifically, if the condition number is to be no more than M/m then $\frac{1}{M} \succcurlyeq -\mathsf{H}\underline{\tilde{L}}(\tilde{p}) \succcurlyeq \frac{1}{m}$ for all $\tilde{p}$. In the case that $M = m$ and the condition number is 1, the only Hessian that satisfies these conditions is $\mathsf{H}\underline{\tilde{L}}(\tilde{p}) = -I_{\tilde{n}}$ which is easily shown to be the Bayes risk for square loss. Thus, square loss is the only canonical composite loss for which a condition number of 1 is possible.

In the binary case, when $n = 2$, (23) and (24) and the positivity of $\tilde{\psi}'$ simplify (25) to the two conditions:

$$\left.\begin{aligned}(1-\tilde{p})\kappa'(\tilde{p}) &\leq \kappa(\tilde{p}) - c\tilde{\psi}'(\tilde{p}) \\ -\tilde{p}\kappa'(\tilde{p}) &\leq \kappa(\tilde{p}) - c\tilde{\psi}'(\tilde{p})\end{aligned}\right\}, \quad \forall \tilde{p} \in (0,1).$$

Further assuming that $\tilde{\psi}$ is the identity link ($\tilde{\psi}(v) = v$) and letting $w(\tilde{p}) := -\underline{\tilde{L}}''(\tilde{p})$ gives

$$\begin{aligned}&\left.\begin{aligned}w'(\tilde{p}) &\leq \tfrac{1}{1-\tilde{p}}(w(\tilde{p})-c)) \\ w'(\tilde{p}) &\geq \tfrac{-1}{\tilde{p}}(w(\tilde{p})-c)\end{aligned}\right\}, \quad \forall \tilde{p} \in (0,1) \\ \Leftrightarrow\ & -\frac{1}{\tilde{p}} \leq \frac{w'(\tilde{p})}{w(\tilde{p})-c} \leq \frac{1}{1-\tilde{p}}, \quad \forall \tilde{p} \in (0,1).\end{aligned} \tag{26}$$

The last equivalence is achieved by dividing through by $w(\tilde{p}) - c$ which must necessarily be positive since if it were not the final pair of inequalities would imply $-\frac{1}{\tilde{p}} \geq \frac{1}{1-\tilde{p}}$, a contradiction given that $\tilde{p} \in [0,1]$. Note that (26) reduces to (Reid and Williamson, 2010, Corollary 26) for $c = 0$.

Observe that if $g(\tilde{p}) := \log(w(\tilde{p}) - c)$ then $g'(\tilde{p}) = \frac{w'(\tilde{p})}{w(\tilde{p})-c}$ is the middle term in (26). This allows a simplification of the inequality. Specifically, if we assume $w(\frac{1}{2}) = 1$ then

$$\begin{aligned}& -\frac{1}{\tilde{p}} \leq g'(\tilde{p}) \leq \frac{1}{1-\tilde{p}}, \ \forall \tilde{p} \in (0,1) \\ \Rightarrow\ & \int_{\frac{1}{2}}^{q} -\frac{1}{\tilde{p}} d\tilde{p} \lesseqgtr \int_{\frac{1}{2}}^{q} g'(\tilde{p}) d\tilde{p} \lesseqgtr \int_{\frac{1}{2}}^{q} \frac{1}{1-\tilde{p}} d\tilde{p}, \ \forall q \in (0,1) \\ \Leftrightarrow\ & -\log(q) - \log(2) \lesseqgtr g(q) - \log(1-c) \\ & \lesseqgtr -\log(2) - \log(1-q), \ \forall q \in (0,1) \\ \Leftrightarrow\ & \frac{1}{2q} \lesseqgtr e^{g(q)-\log(1-c)} \lesseqgtr \frac{1}{2(1-q)}, \ \forall q \in (0,1)\end{aligned} \tag{27}$$

which gives the following proposition purely in terms of $w(\tilde{p})$, rather than $w(\tilde{p})$ and its derivative.

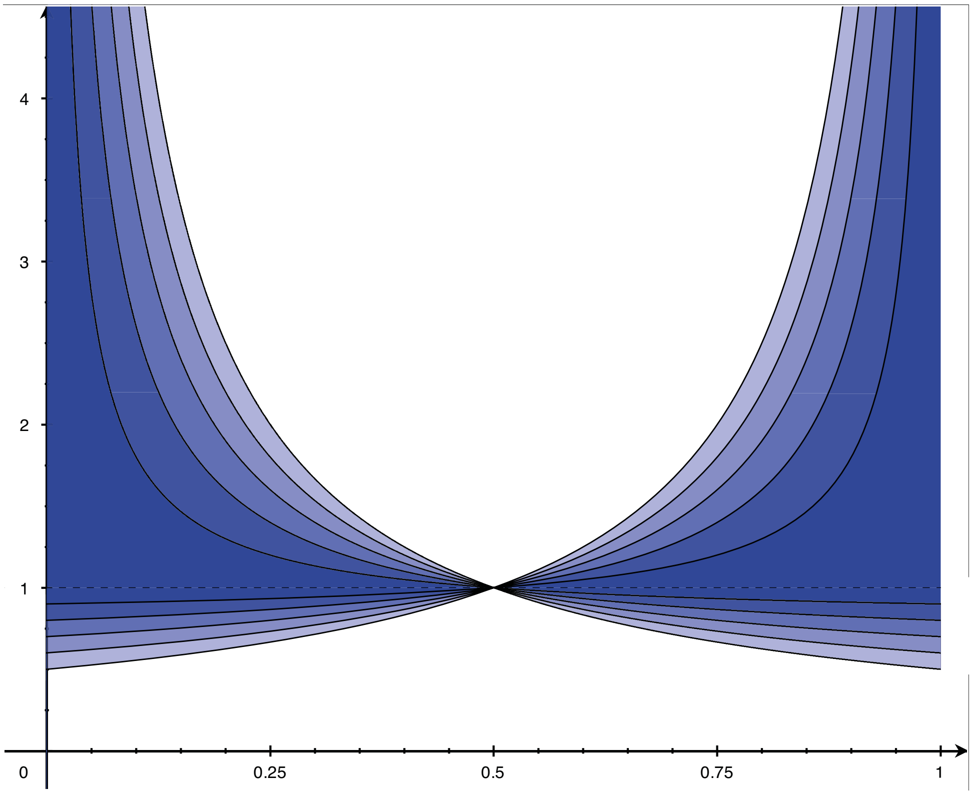

Figure 7: Graph of $w(\tilde{p}) = -\tilde{\underline{L}}''(\tilde{p})$ as a function of $\tilde{p}$ necessary for a suitably normalised binary proper loss to be strongly convex with modulus $c \in \{0, \frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}, 1\}$. The regions R_c are nested by subsethood so that $R_0 \supset R_{1/5} \supset R_{2/5} \supset R_{3/5} \supset R_{4/5} \supset R_1$, where R_1 is simply the dotted line (containing only the function $w(c) = 1$, $\forall c \in [0,1]$, which is the weight function corresponding to squared loss). The palest shaded region corresponds to R_0, the allowable range of $w(c)$ necessary for the corresponding proper loss to be convex, and the darkest corresponds to $R_{4/5}$.

Proposition 33 *Let $w(\tilde{p}) = -\mathsf{H}\tilde{\underline{L}}(\tilde{p}) = -\tilde{\underline{L}}''(\tilde{p})$ and assume $w(1/2) = 1$. A proper binary loss $\ell\colon \Delta^2 \to \mathbb{R}^2_+$ is strongly convex with modulus $c \in [0,1]$ only if*

$$\frac{1}{2\tilde{p}} \lesseqgtr \frac{w(\tilde{p}) - c}{1-c} \lesseqgtr \frac{1}{2(1-\tilde{p})}, \quad \forall \tilde{p} \in (0,1), \tag{28}$$

where $\lesseqgtr$ denotes $\leq$ for $\tilde{p} \geq \frac{1}{2}$ and denotes $\geq$ for $\tilde{p} \leq \frac{1}{2}$.

When $c = 0$ (corresponding to ℓ being convex) this is equivalent to an expression by Reid and Williamson (2010, Equation 31), where it was incorrectly claimed this condition was also sufficient. Inequation 28 is illustrated in Figure 7.

The above proposition only gives a necessary condition for strong convexity. (In addition to w belonging to the specified region, $w'(\tilde{p})$ also needs to be suitably controlled). A sufficient condition is useful for designing strongly convex proper losses. Observe that if

$$w(\tilde{p}) = \exp\left(\int_{1/2}^{\tilde{p}} u(t)dt + K\right) + c$$

where $u\colon [0,1] \to \mathbb{R}$ and $K, c \in \mathbb{R}$, then $\frac{\partial}{\partial \tilde{p}} \log(w(\tilde{p}) - c) = u(\tilde{p})$. We require $w(1/2) = 1$ and so $\exp\left(\int_{1/2}^{1/2} u(t)dt + K\right) + c = 1$ and so $e^K = 1 - c$ and

$$w(\tilde{p}) = (1-c)\exp\left(\int_{1/2}^{\tilde{p}} u(t)dt\right) + c \tag{29}$$

satisfies (26) if

$$-\frac{1}{\tilde{p}} \leq u(\tilde{p}) \leq \frac{1}{1-\tilde{p}}, \quad \forall \tilde{p} \in (0,1), \tag{30}$$

and hence the loss with weight function w is strongly convex with modulus c. Thus by choosing u to satisfy (30) and constructing w via (29) one can design strongly convex proper binary losses.

One can ask whether equation (25) can be simplified in the $n > 2$ case by using a matrix version of the logarithmic derivative trick in a manner similar to that used above when $n = 2$. Such a result does exist (Horn and Johnson, 1991, Section 6.6.19) but it requires that $(\mathsf{H}\tilde{\underline{L}}(\tilde{p}))^{-1}$ and $\mathsf{D}(\mathsf{H}\tilde{\underline{L}}(\tilde{p}))$ commute for all $\tilde{p} \in \tilde{\Delta}^n$, which is not generally the case.

7. Integral Representations of Proper Losses

Binary proper losses have an attractive integral representation that provides substantial insight and is a useful tool for both designing losses and understanding the implications of different choices of loss. Specifically, there exists a family of "extremal" loss functions (cost-weighted generalisations of the 0-1 loss) parametrised by $c \in [0,1]$ and defined for all $\eta \in [0,1]$ by $\ell^c_{-1}(\eta) := c[\![\eta \geq c]\!]$ and $\ell^c_1 := (1-c)[\![\eta < c]\!]$. As shown by Buja et al. (2005) and Reid and Williamson (2011), given these extremal functions, any proper binary loss ℓ can be expressed as the weighted integral

$$\ell = \int_0^1 \ell^c\, w(c)\, dc + constant$$

with "weight function" $w(c) = -\tilde{\underline{L}}''(c)$. This representation is a special case of a representation from Choquet theory (Phelps, 2001; Simon, 2011) which characterises when every point in some set can be expressed as a weighted combination of the "extremal points" of the set. Although there *is* such a representation when $n > 2$, the difficulty is that the set of extremal points is *much* larger and this rules out the existence of a nice small set of "primitive" proper losses when $n > 2$, and consequently rules out an easy-to-work-with weight function parameterizing all possible multiclass lossses in a manner analogous to the binary case. The rest of this section makes this statement precise.

A *convex cone* $\mathscr{K}$ is a set of points closed under positive linear combinations. That is, $\mathscr{K} = \alpha\mathscr{K} + \beta\mathscr{K}$ for any $\alpha, \beta \geq 0$. A point $f \in \mathscr{K}$ is *extremal* if $f = \frac{1}{2}(g+h)$ for $g, h \in \mathscr{K}$

implies $\exists \alpha \in \mathbb{R}_+$ such that $g = \alpha f$. That is, f cannot be represented as a non-trivial combination of other points in $\mathscr{K}$. The set of extremal points for $\mathscr{K}$ will be denoted $\operatorname{ex}\mathscr{K}$. Suppose U is a bounded closed convex set in $\mathbb{R}^d$, and $\mathscr{K}_b(U)$ is the set of convex functions on U bounded by 1, then $\mathscr{K}_b(U)$ is compact with respect to the topology of uniform convergence. Bronshtein (1978, Theorem 2.2) showed that the extremal points of the convex cone $\mathscr{K}(U) = \{\alpha f + \beta g : f, g \in \mathscr{K}_b(U), \alpha, \beta \geq 0\}$ are dense (w.r.t. the topology of uniform convergence) in $\mathscr{K}(U)$ when $d > 1$. This means for any function $f \in \mathscr{K}(U)$ there is a sequence of functions $(g^i)_i$ such that for all i $g^i \in \operatorname{ex}\mathscr{K}(U)$ and $\lim_{i\to\infty} \|f - g^i\|_\infty = 0$, where $\|f\|_\infty := \sup_{u\in U} |f(u)|$. We use this result to show that the set of extremal Bayes risks is dense in the set of Bayes risks when $n > 2$.

In order to simplify our analysis, we restrict attention to fair proper losses. A loss is *fair* if each partial loss is zero on its corresponding vertex of the simplex ($\ell_i(e_i) = 0$, $\forall i \in [n]$). A proper loss is *fair* if and only if its Bayes risk is zero at each vertex of the simplex (in this case the Bayes risk is also called fair). One does not lose generality by studying fair proper losses since any proper loss is a sum of a fair proper loss and a constant vector.

The set of fair proper losses defined on Δ^n form a closed convex cone, denoted $\mathscr{L}_n$. The set of concave functions which are zero on all the vertices of the simplex Δ^n is denoted $\mathscr{F}_n$ and is also a closed convex cone.

Proposition 34 *Suppose $n > 2$. Then for any fair proper loss $\ell \in \mathscr{L}_n$ there exists a sequence $(\ell^i)_i$ of extremal fair proper losses ($\ell^i \in \operatorname{ex}\mathscr{L}_n$) which converges almost everywhere to ℓ.*

The implication of this proposition is that the set of extremal multiclas proper losses is very large. Some intuition can be gleaned from Figure 8 from which it is apparent that there is a qualitative difference between the complexity of the set of extremal concave functions in one dimension (corresponding to $n = 2$) and higher dimensions ($n > 2$). The proof of Proposition 34 requires the following lemma which relies upon the correspondence between a proper loss and its Bayes risk (Proposition 8) and the fact that two continuous functions equal almost everywhere are equal everywhere.

Lemma 35 *If $\ell \in \operatorname{ex}\mathscr{L}_n$ then its corresponding Bayes risk $\underline{L}$ is extremal in $\mathscr{F}_n$. Conversely, if $\underline{L} \in \operatorname{ex}\mathscr{F}_n$ then all the proper losses ℓ with Bayes risk equal to $\underline{L}$ are extremal in $\mathscr{L}_n$.*

Proof We suppose that $\ell \in \operatorname{ex}\mathscr{L}_n$ and denote its Bayes risk by $\underline{L}(p) = p' \cdot \ell(p)$. Let $\underline{F}, \underline{G} \in \mathscr{F}_n$ so that $\underline{L} = \frac{1}{2}(\underline{F} + \underline{G})$. Suppose f and g are proper losses whose Bayes risks are respectively equal to $\underline{F}$ and $\underline{G}$, then $\forall p \in \Delta^n$ and almost everywhere in q (more precisely where $\underline{L}$, $\underline{F}$ and $\underline{G}$ are differentiable), $L(p,q) = \frac{1}{2}(G(p,q) + F(p,q))$. Then $\ell = \frac{1}{2}(g + f)$ almost everywhere, so there exists α such as $g = \alpha\ell$ almost everywhere, hence $\underline{G} = \alpha\underline{L}$ almost everywhere and then everywhere by continuity. So $\underline{L}$ is extremal in $\mathscr{F}_n$.

Now suppose that the concave function $\underline{L}$ is extremal and let ℓ be a proper loss whose Bayes risk is $\underline{L}$. Then $L(p,q) = p' \cdot \ell(q) = \underline{L}(q) + (p-q)' \cdot A(q)$ where $A(q) \in \partial\underline{L}(q)$. Suppose that there exist $f, g \in \mathscr{L}_n$ so that $\ell = \frac{1}{2}(f + g)$ almost everywhere, and have associated Bayes risks $\underline{F}$ and $\underline{G}$, respectively. Then $\underline{L}(p) = p' \cdot \ell(p) = p' \cdot \frac{1}{2}(f(p) + g(p)) = \frac{1}{2}(\underline{F} + \underline{G})$ almost everywhere so $\underline{L} = \frac{1}{2}(\underline{F} + \underline{G})$ everywhere by continuity. Since $\underline{L}$ is extremal we must have $\underline{F} = \alpha\underline{L}$ and so $f = \alpha\ell$ where $\underline{L}$ is differentiable (and so almost everywhere). Thus ℓ is extremal in $\mathscr{L}_n$. ∎

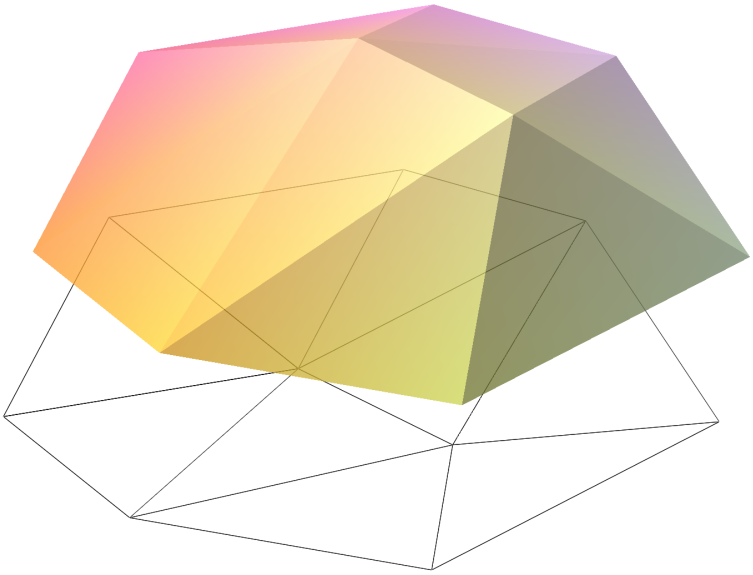

Figure 8: Complexity of extremal concave functions in two dimensions (corresponds to $n = 3$). The figure shows the graph of an extremal concave function in two dimensions. The lines below indicate where the slope changes. The pattern of these lines can be made arbitrarily complex. This illustrates the fact (Proposition 34) that the set of extremal concave functions is very large.

We also need a correspondence between the uniform convergence of a sequence of Bayes risk functions and the convergence of their associated proper losses.

Lemma 36 *Suppose $\underline{L}, \underline{L}^i \in \mathscr{F}_n$ for $i \in \mathbb{N}$ and suppose ℓ and ℓ^i, $i \in \mathbb{N}$ are associated proper losses. Then $(\underline{L}^i)_i$ converges uniformly to $\underline{L}$ if and only if $(\ell^i)_i$ converges almost everywhere to ℓ.*

Proof We require two facts from convex analysis; confer (Hiriart-Urruty and Lemaréchal, 2001, Theorems B.3.1.4 and D.6.2.7). If a sequence $(f^i)_i$ of convex functions f^i converges pointwise to f then: 1) the sequence converges uniformly on any compact domain; and 2) $\forall \varepsilon > 0$, $\partial f^i(x) \subset \partial f(x) + \mathscr{B}(0, \varepsilon)$ for i large enough. Then the reverse implication of the lemma is a direct consequence of the first result and the forward implication is obtained by considering the set $\{x : \forall n, \underline{L}^i$ and $\underline{L}$ are differentiable at $x\}$ which is of measure 1. ∎

Proof (Proposition 34) When $n > 2$ the simplex Δ^n is isomorphic to a subset of $\mathbb{R}^d$ for $d > 1$. Since $\mathscr{F}_n$ is a convex cone associated with the set of bounded concave functions (*i.e.*, the fair Bayes risks), (Bronshtein, 1978, Theorem 2.2) guarantees (with an appropriate change from concavity to convexity) that $\operatorname{ex} \mathscr{F}_n$ is dense in $\mathscr{F}_n$ w.r.t. the topology of uniform convergence. Therefore, if $\ell \in \mathscr{L}_n$ there exists a sequence $(f^i)_i$ with $f^i \in \operatorname{ex} \mathscr{F}_n$ which converges uniformly to the Bayes risk $\underline{L}$ of ℓ and so by Lemma 36 there is a corresponding sequence $(\ell^i)_i$ of fair proper losses that converges almost everywhere to ℓ. Lemma 35 guarantees that each ℓ^i is extremal in $\mathscr{L}_n$ since each $f^i \in \operatorname{ex}\mathscr{F}_n$ and so we have shown there exists a sequence $(\ell^i)_i$ with $\ell^i \in \operatorname{ex}\mathscr{L}_n$ which converges to an ℓ which was arbitrary. ∎

8. Tools for Designing Losses

In this section we show how the results developed above could be used to design losses for particular purposes. There is no question about how this should be done "in principle" (Berger, 1985, Section 2.4). And even when not made explicit at the outset, *all* inference ultimately has an *implicit* loss function that captures what matters to the end user, even if the original purpose was to merely "gather information", simply because in the end the "information" is acted upon (DeGroot, 1962). Until now, the lack of convenient and canonical parametrisations of multiclass loss functions has made the comparison of different loss functions, and their tuning for specific applications, difficult.

In subsection 8.1, we show how to construct a parametric family of valid link functions from a finite number of "base" links by effectively taking their convex combination. By composing each link with a fixed proper loss, this immediately allows for the specification of a family of losses with a fixed Bayes risk. This construction enables the creation of losses with a range of optimisation characteristics (*e.g.*, convexity, robustness) but a common statistical basis (*i.e.*, the same Bayes risk).

In subsection 8.2 we show how it is possible to build losses by building them up from constraints on their Bayes risk curves on the edges of the simplex. This allows a loss to be constructed by effectively specifying its behaviour on pairs of outcomes. We show how this observation can be used to create piecewise linear, proper losses for cost-sensitive misclassification.

Finally (subsection 8.3) we observe how link functions are in fact themselves very similar to loss functions, and (subsection 8.4) we present some examples of proper composite losses from the literature (where they were not expressed in the proper composite parametrisation)

8.1 Families of Losses with Fixed Bayes Risk

The theory developed above suggests that each choice of proper loss λ and link function ψ results in an overall loss function with properties (*e.g.*, convexity) that depend entirely on their relationship to each other. Given these two "knobs" for parameterising a loss function, we can begin to ask what kind of practical trade-offs are involved when selecting a composite loss as a surrogate loss for a particular problem.

We now propose a simple scheme for constructing families of losses with the same Bayes risk. This is achieved by fixing a choice of proper loss λ and creating a parameterised family (described below) of link functions ψ_α for parameters $\alpha \in A$. Since the Bayes risk is entirely determined by λ any composite loss $\lambda \circ \psi_\alpha^{-1}$ for $\alpha \in A$ will have Bayes risk $\underline{L}(p) = p' \cdot \lambda(p)$. Thus, we are able to examine the effect different choices of composite loss can have on a problem *without changing the essential underlying problem.*[4]

[4] Of course, this argument only holds in a point-wise analysis. That is, where choices for estimates $p(x)$ can be made independently. Once a restricted hypothesis class for the functions p is introduced the choice of link can

In order to construct a parametric family of links we first choose some set of inverse link functions $\mathscr{I} = \{\psi_1^{-1}, \ldots, \psi_B^{-1}\}$ with a common domain, that is, $\psi_b^{-1} : \mathscr{V} \to \Delta^n$ for a common n and $\mathscr{V}$. This collection will be called the *basis set* of link functions. We then take the convex hull of $\mathscr{I}$ to form a set of inverse link functions $\Psi^{-1} = \mathrm{co}(\mathscr{I})$. Each $\psi^{-1} \in \Psi^{-1}$ is then identified with the unique $\alpha \in A = \Delta^B$ such that $\sum_{b=1}^{B} \alpha_b \psi_b^{-1} = \psi^{-1}$. For this construction to be valid, it it necessary to show that every such $\psi^{-1} \in \Psi^{-1}$ is indeed an inverse link function, that is, it is invertible.

The following proposition shows that it suffices to assume that all of the basis functions are *strictly monotone* (see Equation 2).

Proposition 37 *Every function ψ^{-1} in the set $\Psi^{-1} = \mathrm{co}(\mathscr{I})$ is invertible whenever each basis function in $\mathscr{I}$ is strictly monotone.*

This result is a consequence of: 1) strict monotonicity being preserved under convex combination; and 2) strict monotonicity implies invertibility. The first claim is established by considering strictly monotone f and g and some $\alpha \in [0,1]$ and noting that if $h = \alpha f + (1-\alpha)g$ then $(h(u) - h(v))'(u-v) = \alpha(f(u) - f(v))'(u-v) + (1-\alpha)(g(u) - g(v))'(u-v) > 0$. A strictly monotone function f that is not invertible is impossible since if we have $(f(u)-f(v))'(u-v) > 0$ for all u, v then a $u \neq v$ such that $f(u) = f(v)$ would lead to a contradiction.

Strictly monotone basis functions are easily obtained via canonical links for strictly proper losses. By definition, a canonical link satisfies $\tilde{\psi} = -\mathsf{D}\tilde{\underline{L}}$ for some Bayes risk function. Strict properness guarantees $\tilde{\underline{L}}$ is strictly concave (van Erven et al., 2012b, Lemma 1). Kachurovskii's theorem (Hiriart-Urruty and Lemaréchal, 2001, Theorem 4.1.4) states that the derivative of a function is (strictly) monotone if and only if the function is (strictly) convex. Since $(f(f^{-1}(u)) - f(f^{-1}(v)))'(f^{-1}(u) - f^{-1}(v)) = (u-v)'(f^{-1}(u) - f^{-1}(v))$ we see that strictly monotone functions have strictly monotone inverses and we have established the following proposition.

Proposition 38 *If λ is a strictly proper loss then its canonical link $\tilde{\psi}_\lambda = -\mathsf{D}\tilde{\underline{L}}$ has a strictly monotone inverse.*

This result means that a set of basis links can be defined via a choice of strictly concave Bayes risk functions. As an example, the class of Fisher-consistent margin losses proposed by Zou et al. (2008) provides a flexible starting point for designing sets of link functions as described above. They give explicit formulae for the inverse link for a composite loss defined by a choice of convex function $\phi : \mathbb{R} \to \mathbb{R}$. Specifically, if the loss for predicting $v \in \mathscr{V} = \{v \in \mathbb{R}^n : \sum_i v_i = 0\}$ is given by $\ell(v) = \phi(v_j)$ then its inverse link is $\psi_\phi^{-1}(v) = \frac{1}{Z_\phi(v)}\left([\phi'(v_i)]^{-1}\right)_{i=1}^{n}$ where $Z_\phi(v)$ normalises the vector to lie in Δ^n. Each choice of strictly convex ϕ gives a valid inverse link which can be used as a basis function.

8.2 Piecewise Linear Multiclass Losses

We now build a family of conditional Bayes risks. Suppose we are given $\frac{n(n-1)}{2}$ concave functions $\{\underline{L}^{i_1,i_2} : \Delta^2 \to \mathbb{R}\}_{1 \le i_1 < i_2 \le n}$ on Δ^2, and we want to build a concave function $\underline{L}$ on Δ^n

affect the minimal achievable risk. The interaction between the hypothesis class and the loss function is complex (van Erven et al., 2015).

which is equal to one of the given functions on each edge of the simplex ($\forall 1 \leq i_1 < i_2 \leq n$, $\underline{L}(0,\ldots,0,p_{i_1},0,\ldots,0,p_{i_2},0,\ldots,0) = \underline{L}^{i_1,i_2}(p_{i_1},p_{i_2})$). This is equivalent to choosing a binary loss function, knowing that the observation is in the class i_1 or i_2. The result below gives one possible construction. (There exists an infinite number of solutions—one can simply add any concave function equal to zero in each edge).

Lemma 39 *Suppose we have a family of concave functions* $\{\underline{L}^{i_1,i_2}\colon \Delta^2 \to \mathbb{R}\}_{1\leq i_1<i_2\leq n}$, *then*

$$\underline{L}\colon \Delta^n \ni p \mapsto \underline{L}((p_1,\ldots,p_n)') = \sum_{1\leq i_1<i_2\leq n} (p_{i_1}+p_{i_2})\underline{L}^{i_1,i_2}\left(\left(\frac{p_{i_1}}{p_{i_1}+p_{i_2}}, \frac{p_{i_2}}{p_{i_1}+p_{i_2}}\right)'\right)$$

is concave and $\forall 1 \leq i_1 < i_2 \leq n$, $\underline{L}((0,\ldots,0,p_{i_1},0,\ldots,0,p_{i_2},0,\ldots,0)') = \underline{L}^{i_1,i_2}((p_{i_1},p_{i_2})')$.

Proof In order to show that $\underline{L}$ is concave it suffices to show that for $g : \Delta^2 \to \mathbb{R}$ concave, $f : p \in \Delta^n \to f(p) = (p_1+p_2)g\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\right)$ is concave, since a sum of concave functions is concave. Let $\gamma \in [0,1]$, $p,q \in \Delta^n$. Since g is concave, $\forall \alpha \in [0,1]$, $\forall p,q \in \Delta^2$, $g\big(\frac{\alpha}{p_1+p_2}p + \frac{1-\alpha}{q_1+q_2}q\big) \geq \alpha g\big(\frac{p}{p_1+p_2}\big) + (1-\alpha)g\big(\frac{q}{q_1+q_2}\big)$. Then with $\alpha = \frac{\gamma(p_1+p_2)}{\gamma(p_1+p_2)+(1-\gamma)(q_1+q_2)}$, we get $f(\gamma p + (1-\gamma)q) \geq \gamma f(p) + (1-\gamma)f(q)$.

Moreover, $\underline{L}((0,\ldots,0,p_{i_1},0,\ldots,0,p_{i_2},0,\ldots,0)') = \sum_{i\notin\{i_1,i_2\}} (p_{i_1}\times 0 + p_{i_2}\times 0)$ $+ (p_{i_1}+p_{i_2})\underline{L}^{i_1,i_2}\left(\left(\frac{p_{i_1}}{p_{i_1}+p_{i_2}}, \frac{p_{i_2}}{p_{i_1}+p_{i_2}}\right)'\right) = \underline{L}^{i_1,i_2}((p_{i_1},p_{i_2})')$, ($p \in \Delta^n$, so $p_{i_1}+p_{i_2} = 1$). ■

Using this family of Bayes risks, one can build a family of proper losses.

Lemma 40 *Suppose we have a family of binary proper losses* $\ell^{i_1,i_2} : \Delta^2 \to \mathbb{R}^2$. *Then*

$$\ell\colon \Delta^n \ni p \mapsto \ell(p) = \left(\sum_{i=1}^{j-1} \ell_{-1}^{i,j}\left(\frac{p_i}{p_i+p_j}\right) + \sum_{i=j+1}^{n} \ell_1^{i,j}\left(\frac{p_j}{p_i+p_j}\right)\right)_{j=1}^{n} \in \mathbb{R}_+^n$$

is a proper n-class loss such that

$$\ell_i((0,\ldots,0,p_{i_1},0,\ldots,0,p_{i_2},0,\ldots,0)') = \begin{cases} \ell_1^{i_1,i_2}(p_{i_1}) & i = i_1 \\ \ell_{-1}^{i_1,i_2}(p_{i_1}) & i = i_2 \\ 0 & \textit{otherwise} \end{cases}.$$

Proof Use the correspondence between Bayes risk and proper losses and Lemma 39. ■

Observe that it is much easier to work at first with the Bayes risk and then using the correspondence between Bayes risks and proper losses.

We have already seen (Section 7) that it is not possible to parametrise *all* extremal concave functions in a tractable manner. However, for the sake of offering a range of knobs to the designer to design losses, it could often suffice to use a subset of extremal losses. These will all have polyhedral forms. A convex *polytope* is a compact convex intersection of a finite set of half-spaces and is therefore the convex hull of its vertices. Let $\{a_i\}_i$ be a finite family of affine functions defined on Δ^n. Now define the convex *polyhedral function* f by $f(x) := \max_i a_i(x)$. The set $K := \{P_i = \{x \in \Delta^n\colon f(x) = a_i(x)\}\}$ is a covering of Δ^n by polytopes. Bronshtein (1978, Theorem 2.1) shows that for f, P_i and K so defined, f is extremal if the following two conditions are satisfied: 1) for all polytopes P_i in K and for every face F of P_i, $F \cap \Delta^n \neq \varnothing$ implies F has a vertex in Δ^n; 2) every vertex of P_i in Δ^n belongs to n distinct polytopes of K. The set of all such f is dense in $\mathscr{K}(U)$.

Using this result it is straightforward to exhibit some sets of extremal fair Bayes risks $\{\underline{L}_c(p) \colon c \in \Delta^n\}$. Two examples are when

$$\underline{L}_c(p) = \sum_{i=1}^{n} \tfrac{p_i}{c_i} \prod_{j \neq i} [\![\tfrac{p_i}{c_i} \leq \tfrac{p_j}{c_j}]\!]$$

or

$$\underline{L}_c(p) = \bigwedge_{i \in [n]} \tfrac{1-p_i}{1-c_i}.$$

Any convex combination of either of these families will be the Bayes risk of a proper fair multiclass loss. Thus the convex combination of the elementary losses induced by such $\underline{L}_c(p)$ will also be proper fair multiclass losses.

8.3 Parametrisation of Composite Losses

A composite loss $\ell = \lambda \circ \psi^{-1} \colon \mathscr{V} \to \mathbb{R}^n_+$ is directly parametrised by the proper loss $\lambda \colon \Delta^n \to \mathbb{R}^n_+$ and the invertible link $\psi \colon \Delta^n \to \mathbb{R}^n$. However we have seen (Section 4) that proper losses λ are more nicely parametrised by their concave conditional Bayes risk $\underline{\Lambda} \colon \Delta^n \to \mathbb{R}$, which being scalar valued, are simpler objects to work with than λ. Although not *every* invertible function ψ can be written as the gradient of an analogous *convex* function $\Psi \colon \Delta^n \to \mathbb{R}$, by Kachurovskii's theorem (see Section 8.1) if for some $\Psi \colon \Delta^n \to \mathbb{R}$, $\psi = \mathsf{D}\Psi$, then ψ is monotone (resp. strictly monotone) if and only if Ψ is convex (resp. strictly convex). A link ψ is a gradient if and only if $\mathsf{D}\psi$ is symmetric (so that $\mathsf{H}\Psi$ is symmetric) as a Hessian needs to be.

Thus if one were willing to restrict oneself to links such that $\mathsf{D}\psi$ is symmetric, then a composite loss ℓ can be parametrised by $(\underline{\Lambda}, \Psi)$, which are concave (resp. strictly convex) functions from Δ^n to $\mathbb{R}$. The parametrisation of ψ via Ψ allows the specification of the canonical link as that satisfying $\Psi = -\underline{\Lambda}$.

8.4 Examples from Related Work

In this subsection we look at some existing candidate multiclass losses from the perspective of proper composite representations. Not all such losses as the generalisation of hinge loss are so representable, a prominent example being those introduced by Crammer and Singer (2001).

Zou et al. (2008) presented multi-category losses of the form $\ell_y(f) = \phi(f_y)$ for f such that $\sum_y f_y = 0$ and $\phi'(0) < 0$ and $\phi''(t) \geq 0$ so that we have Fisher consistency and the inverse link is $\frac{1/\phi'(f_j)}{\sum_i 1/\phi'(f_i)}$. As their examples of this class show, it is not always possible to write the link in closed form, even if the inverse link can be (e.g., logit loss $\phi(t) = \log(1+e^t)$).

The *coherence functions* of Zhang et al. (2009) are a separate class of surrogate functions that emphasise the margin of a prediction:

$$\ell_y(v) = T \log \left[1 + \sum_{i \neq y} \exp\left(T^{-1}(1 + v_i - v_y) \right) \right].$$

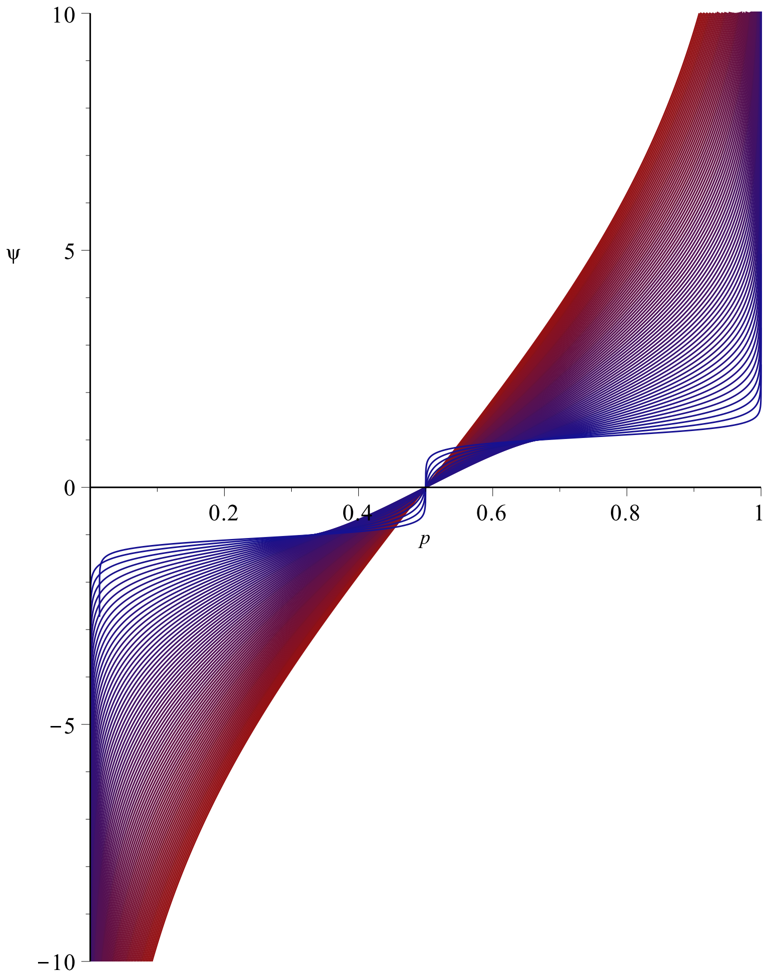

Figure 9: Illustration of the link function in the proper composite representation of the binary coherence loss for $T \in [0.1,4]$. Blue corresponds to $T = 0.1$ and red to $T = 4$, with there being 80 equal increments of T plotted.

We will illustrate some aspects of the present paper with reference to this parametrised family of losses. For ease of calculation, we consider only $n = 2$ below, but the conclusions we draw below hold for $n > 2$ also. In the binary case, this corresponds to a parametric family of margin losses with margin function

$$\phi_T(z) := T \log\left(1 + \exp\left(\frac{1-z}{T}\right)\right),$$

and thus $\phi'_T(z) = \frac{e^{(1-z)/T}}{1+e^{(1-z)/T}}$ and one can check that $g_T(v) := -\frac{\phi'_T(v)}{\phi'_T(-v)}$ is strictly monotone continuous and ϕ_T is monotone for all $T > 0$ and thus by Corollary 12 there is a strictly proper composite representation. Identifying $\ell_{-1}(v)$ with $\phi(-v)$ and $\ell_1(v)$ with $\phi(v)$ we can find for a

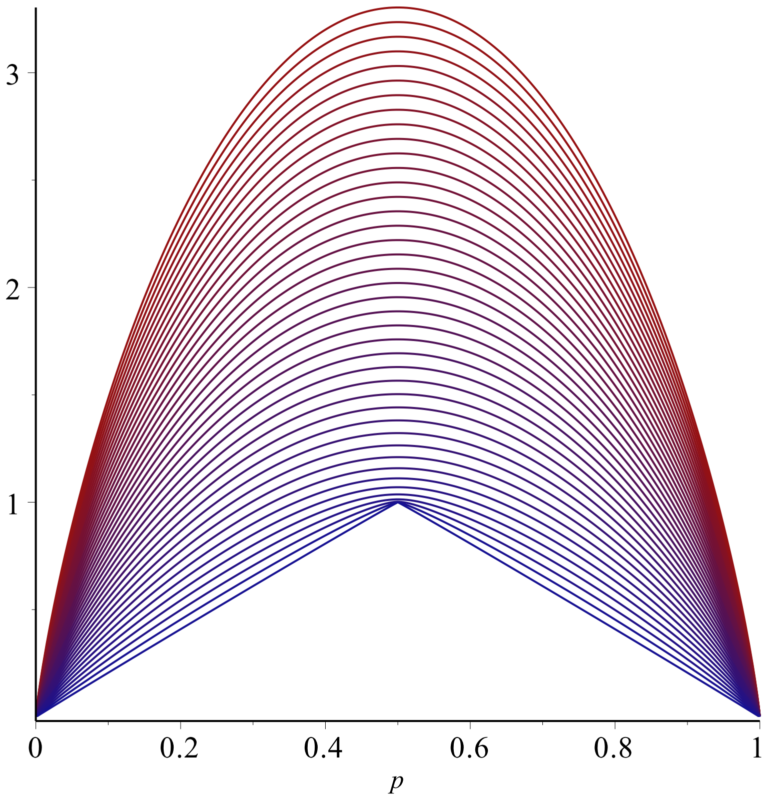

Figure 10: Illustration of the conditional Bayes risk corresponding to the the binary coherence loss for $T \in [0.02, 4]$ with blue corresponding to $T = 0.02$ and red to $T = 4$.

given p the v that minimises $L(p, v)$ by solving

$$\frac{\partial}{\partial v} L(p, v) = (1-p)\phi_T'(-v) + p\phi_T'(v) = 0$$

and obtain $\frac{1-p}{p} = -g_T(v)$. Solving this for v (using Maple) we find the link function component of the proper composite representation:

$$\psi_T(p) = T\left(\ln\left(\frac{1}{2(1-p)}\left(2\,p e^{T^{-1}} - e^{T^{-1}} + \sqrt{4\,e^{2\,T^{-1}} p^2 - 4\,p e^{2\,T^{-1}} - 4\,p^2 + e^{2\,T^{-1}} + 4\,p}\right)\right)\right).$$

This is illustrated in Figure 9. Zou et al. (2008, Theorem 1) effectively compute ψ_T^{-1} for general n. One can determine the proper component as

$$\underline{L}_T(p) = p\,\phi_T\left(\psi_T(p)\right) + (1-p)\,\phi_T\left(-\psi_T(p)\right)$$

which is plotted in Figure 10. One can glean further insight by considering the corresponding weight function $w_T(p) := -\underline{L}_T''(p)$, which is plotted in Figure 11.

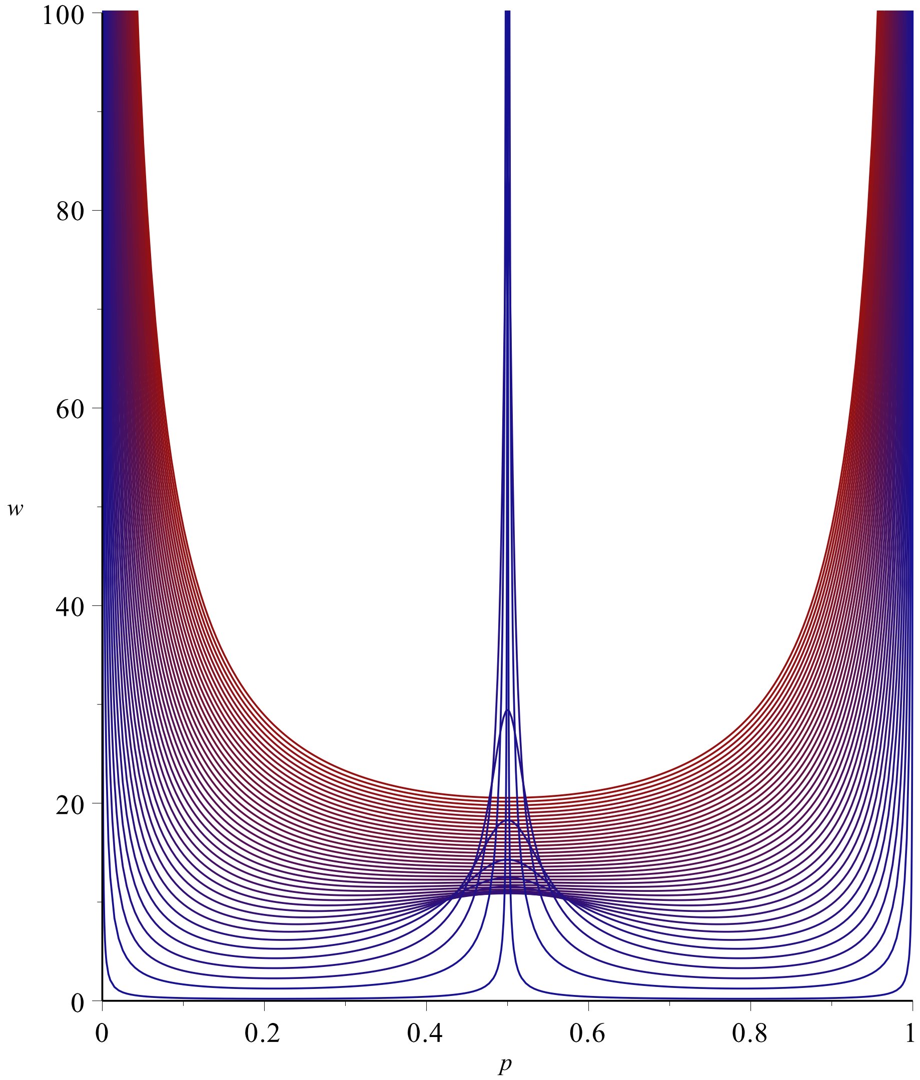

Figure 11: Weight function of the proper loss component of the proper composite representation of the binary coherence loss for $T \in [0.02, 4]$, where blue corresponds to $T = 0.02$ and red to $T = 4$.

The weight function view makes it clear how the proper component of the loss approaches 0-1 loss as $T \to 0$. (The weight function for 0/1 loss is $w(c) = \delta(c - \frac{1}{2})$; note the convergence in Figure 11 is not uniform, given the behaviour at 0 and 1.) Observe too that as well as the proper loss varying with T, the associated link also varies (in a complex way—see Figure 9). Thus not only is one varying the statistical properties of the loss (in a substantial way—when T is small, the weight is centered near $p = \frac{1}{2}$, whereas for large T, a series expansion of $w_T(p)$ shows that $w_T(p) \approx T\left(\frac{1}{p} + \frac{1}{1-p}\right) + 4$. An alternative to this class of losses, would be to fix a link, and then vary the proper component. For a given model or hypothesis class $\mathscr{F}$ this has the advantage that the effective hypothesis class $\{\psi^{-1} \circ f \colon f \in \mathscr{F}\}$ remains fixed, and only the (proper) loss varies. The $(\underline{\Lambda}, \Psi)$ parametrisation of section 8.3 offers a convenient way to do this.

9. Conclusions

We have systematically studied multiclass composite losses. The results of the paper (summarised below) show that this is an attractive parametrisation of multiclass losses. If one desires all predictions be strongly admissible, then there is nothing lost in using the proper composite representation. Since the link is only a reparametrisation, this means one still has the relationship between losses and divergences as described by García-García and Williamson (2012).

The proper composite representation leads to a desirable separation of concerns, where the inferential properties of the loss (such as its mixability) are governed by the proper loss, and the convexity (necessary for numerical optimisation) is controlled by the link function. It thus seems to be *the* best way to parametrise loss functions.

The key technical contributions of the paper are as follows.

- Relationship between prediction calibration and classification calibration, showing that the latter can be seen as a "pointwise" version of the former (Section 3);
- Characterisation of multiclass proper losses in terms of their binary restrictions (Proposition 7);
- Every (multiclass) proper loss is quasi-convex (Proposition 17);
- Characterisation of which binary margin losses have a proper composite representation (Corollary 12);
- Characterisation of when a multiclass loss has a proper composite representation and when the representation is unique (Section 5.3);
- Relationship between the proper composite representation, mixability and admissibility (Sections 6.1 and 6.2);
- Necessary conditions for strong convexity of multiclass proper losses in terms of their corresponding Bayes risks (Proposition 31);
- Canonical links always convexify proper losses, and outline how this can help in the design of losses (Proposition 32);
- The attractive (simply parametrised) integral representation for binary proper losses can *not* be extended to the multiclass case (Section 7) ;

These results suggest that in order to design losses for multiclass prediction problems it is helpful to use the composite representation, and design the proper part via the Bayes risk as suggested for the binary case by Buja et al. (2005). The link function can be tuned to control the optimisation properties of the loss. Merely requiring the loss to be convex confounds two seperate aspects of a loss: the *shape* of $\ell(\mathscr{V})$ which controls the predictive performance, and the *parametrization* of $\ell(\mathscr{V})$ which affects the numerical optimisation of a loss.

There remain open questions. Perhaps the most practically important is the interaction between the loss and restricted hypothesis classes: typically one does not optimise conditionally, one optimises the full expected risk with respect to a restricted function class $\mathscr{F} \subset \mathscr{Y}^{\mathscr{X}}$. The question of how knowledge of $\mathscr{F}$ should influence the design of a loss remains open; some initial work along these lines is the notion of "stochastic mixability" (van Erven et al., 2012a, 2015).

Acknowledgments

The research reported here was performed whilst Elodie Vernet was a student at ENS Cachan and visiting ANU and NICTA, and was supported by the Australian Research Council and NICTA, which was funded by the Australian government through the ICT centre of excellence program. An earlier version of some of these results appeared in NIPS2011 and ICML2012. The work benefited from discussions with Jake Abernethy, Tim van Erven, Rafael Frongillo, and Dario García-García, and comments from the referees who we thank. Thanks also to Harish Guruprasad for identifying a flaw in an earlier proof of quasi-convexity of proper losses.

Appendix A. Matrix Calculus

If $A = [a_{ij}]$ is an $n \times m$ matrix, vec$A$ is the vector of columns of A stacked on top of each other. The *Kronecker product* of an $m \times n$ matrix A with a $p \times q$ matrix B is the $mp \times nq$ matrix

$$A \otimes B := \begin{pmatrix} A_{1,1}B & \cdots & A_{1,n}B \\ \vdots & \ddots & \vdots \\ A_{m,1}B & \cdots & A_{m,n}B \end{pmatrix}.$$

We use the following properties of Kronecker products (Magnus and Neudecker, 1999, Chapter 2): $(A \otimes B)(C \otimes D) = (AC \otimes BD)$ for all appropriately sized A, B, C, D, and $I_1 \otimes A = A$.

If $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at c then the *partial derivative* of f_i with respect to the jth coordinate at c is denoted $\mathsf{D}_j f_i(c)$. The $m \times n$ matrix of partial derivatives of f is the *Jacobian* of f and denoted

$$(\mathsf{D}f(c))_{i,j} := \mathsf{D}_j f_i(c) \quad \text{for } i \in [m], j \in [n].$$

If F is a matrix valued function $\mathsf{D}F(X) := \mathsf{D}f(\operatorname{vec} X)$ where $f(X) = \operatorname{vec} F(X)$.

We will require the *product rule* for matrix valued functions (Vetter, 1970; Fackler, 2005): Suppose $f \colon \mathbb{R}^n \to \mathbb{R}^{m \times p}$, $g \colon \mathbb{R}^n \to \mathbb{R}^{p \times q}$ so that $(f \times g) \colon \mathbb{R}^n \to \mathbb{R}^{m \times q}$. Then

$$\mathsf{D}(f \times g)(x) = (g(x)' \otimes I_m) \cdot \mathsf{D}f(x) + (I_q \otimes f(x)) \cdot \mathsf{D}g(x). \tag{31}$$

The *Hessian* at $x \in \mathscr{X} \subseteq \mathbb{R}^n$ of a real-valued function $f : \mathscr{X} \to \mathbb{R}$ is the $n \times n$ real, symmetric matrix of second derivatives at x

$$(\mathsf{H}f(x))_{j,k} := \mathsf{D}_{k,j} f(x) = \frac{\partial^2 f}{\partial x_k \partial x_j}.$$

Note that the derivative $D_{k,j}$ is in row j, column k. It is easy to establish that the Jacobian of the transpose of the Jacobian of f is the Hessian of f. That is,

$$\mathsf{H}f(x) = \mathsf{D}\left((\mathsf{D}f(x))'\right) \tag{32}$$

(Magnus and Neudecker, 1999, Chapter 10). If $\mathscr{X} \subset \mathbb{R}^n$ and $f\colon \mathscr{X} \to \mathbb{R}^m$ is a vector-valued function, then the Hessian of f at $x \in \mathscr{X}$ is the $mn \times n$ matrix that consists of the Hessians of the functions f_i stacked vertically:

$$\mathsf{H}f(x) := \begin{pmatrix} \mathsf{H}f_1(x) \\ \vdots \\ \mathsf{H}f_m(x) \end{pmatrix}.$$

Appendix B. Deferred Proofs

This appendix contains proofs of results in the main text that, due to their length or technicality, are better presented outside the flow of the main text.

B.1 Proof of Lemma 1

1. We prove this by contradiction. Suppose $p \in \Delta^n$ such that for all $i \in [n]$, $p \notin \mathscr{T}_i(c)$. Then

$$p \notin \mathscr{T}_{j_1}(c) \Rightarrow \exists j_2 \neq j_1 \text{ such that } \frac{p_{j_1}}{c_{j_1}} < \frac{p_{j_2}}{c_{j_2}}$$
$$p \notin \mathscr{T}_{j_2}(c) \Rightarrow \exists j_3 \neq j_2 \text{ such that } \frac{p_{j_2}}{c_{j_2}} < \frac{p_{j_3}}{c_{j_3}}$$

   and hence by repeating this argument

$$p \notin \mathscr{T}_{j_n}(c) \Rightarrow \exists j_{n+1} \neq j_n \text{ such that } \frac{p_{j_n}}{c_{j_n}} < \frac{p_{j_{n+1}}}{c_{j_{n+1}}}.$$

   Thus we have $n+1$ indices $j_1, \ldots, j_{n+1}$ belonging to $[n]$ and therefore one is repeated (j_k) and $\frac{p_{j_k}}{c_{j_k}} < \frac{p_{j_k}}{c_{j_k}}$ which is a contradiction.

2. Obvious.

3. If $p \in \bigcap_{i=1}^n \mathscr{T}_i(c)$, then for all $j \in [n]$, $c_j = \sum_i p_i c_j = \sum_i p_j c_i = p_j$. Thus $p = c$.

4. We prove this by contradiction. Suppose $p \neq q$ such that for all c if $p \in \mathscr{T}_i(c)$ then $q \in \mathscr{T}_i(c)$. Observe that $\forall j \in [n]$, $p \in \mathscr{T}_j(p)$, and so $q \in \bigcap_{j=1}^n \mathscr{T}_j(q)$, and hence $q = p$, a contradiction.

B.2 Proof of Proposition 9

Observe that

$$\partial \underline{L}(p) = \{(s', 0)' + \alpha \mathbb{1},\ s \in \partial \underline{\tilde{L}}(\tilde{p}),\ \alpha \in \mathbb{R}\}. \tag{33}$$

Indeed $(\tilde{q}-\tilde{p})'\cdot s=(q-p)'\cdot((s',0)'+\alpha\mathbb{1})$.

($\Leftarrow$) We first assume that $\underline{L}$ is differentiable at p. We use the following result (Hiriart-Urruty and Lemaréchal, 2001, page 203): *If f is a convex function, then $\forall\varepsilon>0$, $\exists\delta>0$, $y\in\mathscr{B}(x,\delta)\Rightarrow\partial f(y)\subset\partial f(x)+\mathscr{B}(0,\varepsilon)$.*

Assume $\varepsilon>0$, then since $\underline{L}$ is differentiable at $\tilde{p}$, $\exists\tilde{\delta}>0$, such that

$$\forall\tilde{q}\in\mathscr{B}(\tilde{p},\tilde{\delta}),\ \forall A(\tilde{q})\in\partial\underline{\tilde{L}}(\tilde{q}),\ ||A(\tilde{q})-\mathsf{D}\underline{\tilde{L}}(\tilde{p})||\leq\varepsilon. \tag{34}$$

Then there exists δ such that $q\in\mathscr{B}(p,\delta)$ implies $\tilde{p}\in\mathscr{B}(\tilde{p},\tilde{\delta})$. Thus using (3) and (34), $\forall i\in[n]$, $\forall q\in\mathscr{B}(p,\delta)$, for $\alpha_1,\alpha_2\in\mathbb{R}$,

$$\begin{aligned}\ell_i(q)-\ell_i(p)=&\,\underline{L}(q)+(e_i-q)'\cdot((A(\tilde{q})',0)'+\alpha_1\mathbb{1})-\\&\quad(\underline{L}(p)+(e_i-p)'\cdot((\mathsf{D}\underline{L}(\tilde{p})',0)'+\alpha_2\mathbb{1})),\\=&\,\underline{L}(q)-\underline{L}(p)+(\tilde{e}_i-\tilde{q})'\cdot A(\tilde{q})-(\tilde{e}_i-\tilde{p})'\cdot\mathsf{D}\underline{L}(\tilde{p})+\gamma,\ \ \forall A(\tilde{q})\in\partial\underline{\tilde{L}}(\tilde{q}),\end{aligned}$$

where $A(\tilde{q})\in\partial\underline{\tilde{L}}(\tilde{q})$, and

$$\begin{aligned}\gamma&=-(e_i-q)'\cdot\alpha_1\mathbb{1}+(e_i-p)'\cdot\alpha_2\mathbb{1}=-\alpha_1+\alpha_1q'\cdot\mathbb{1}+\alpha_2-\alpha_2p'\cdot\mathbb{1}\\&=-\alpha_1+\alpha_1+\alpha_2-\alpha_2=0\end{aligned}$$

and so

$$\ell_i(q)-\ell_i(p)=\underline{L}(q)-\underline{L}(p)+(\tilde{e}_i-\tilde{q})'\cdot(A(\tilde{q})-\mathsf{D}\underline{L}(\tilde{p}))+(\tilde{p}-\tilde{q})'\cdot\mathsf{D}\underline{L}(\tilde{p}).$$

By continuity of $\underline{L}$, $||\underline{L}(q)-\underline{L}(p)||<\varepsilon$ for small enough δ. Furthermore by (34), $||A(\tilde{q})-\mathsf{D}\underline{\tilde{L}}(\tilde{p})||\leq 0$ and $||\tilde{p}-\tilde{q}||\leq\varepsilon$. Hence $||\ell_i(q)-\ell_i(p)||\leq\varepsilon+\varepsilon+\delta$ which can be made arbitrarily small by suitable choice of ε. Thus ℓ_i is continuous for all $i\in[n]$ and so ℓ is continuous.

($\Rightarrow$) Assume that $\underline{L}$ is not differentiable at $p\in\mathring{\Delta}^n$. Thus there exists two different supergradients at p: $A(\tilde{p})$ and $B(\tilde{p})$. Assume that one of these supergradients, $A(\tilde{p})$, is the one associated to the loss ℓ in the sense that for all $i\in[n]$ $\ell_i(p)=\underline{L}(p)+(\tilde{e}_i-\tilde{p})'\cdot A(\tilde{p})$.

Suppose that $\forall i\in[n]$,

$$(e_i-p)'\cdot((A(\tilde{p})',0)'+\alpha_1\mathbb{1})\leq(e_i-p)'\cdot((B(\tilde{p})',0)'+\alpha_2\mathbb{1}),\ \ \alpha_1,\alpha_2\in\mathbb{R}. \tag{35}$$

Thus

$$\begin{aligned}\sum_{i\in[n]}q_i(e_i-p)'\cdot((A(\tilde{p})',0)'+\alpha_1\mathbb{1})\leq\sum_{i\in[n]}q_i(e_i-p)'\cdot((B(\tilde{p})',0)'+\alpha_2\mathbb{1}),\ \forall q\in\Delta^n,\ \alpha_1,\alpha_2\in\mathbb{R}\\\Leftrightarrow\ (q-p)'\cdot((A(\tilde{p})',0)'+\alpha_1\mathbb{1})\ \leq\ (q-p)'\cdot((B(\tilde{p})',0)'+\alpha_2\mathbb{1}),\quad\forall q\in\Delta^n,\alpha_1,\alpha_2\in\mathbb{R}\\\Leftrightarrow\ (\tilde{q}-\tilde{p})'\cdot A(\tilde{p})\ \leq\ (\tilde{q}-\tilde{p})'\cdot B(\tilde{p}),\ \ \forall\tilde{q}\in\tilde{\Delta}^n.\end{aligned} \tag{36}$$

Since $p\in\mathring{\Delta}^n$ we can choose $\tilde{q}_1$ and $\tilde{q}_2\in\tilde{\Delta}^n$ such that $\tilde{q}_1-\tilde{p}=\tilde{p}-\tilde{q}_2$ and so the only way (36) can hold is if

$$(\tilde{q}-\tilde{p})'\cdot A(\tilde{p})=(\tilde{q}-\tilde{p})'\cdot B(\tilde{p}).$$

Since $p \in \mathring{\Delta}^n$ is arbitrary, we obtain that $A(\tilde{p}) = B(\tilde{p})$, a contradiction and so (35) must be false.

Thus there exists $i \in [n]$ such that

$$(e_i - p)' \cdot ((A(\tilde{p})', 0)' + \alpha_1 \mathbb{1}) > (e_i - p)' \cdot ((B(\tilde{p})', 0)' + \alpha_2 \mathbb{1}), \ \alpha_1, \alpha_2 \in \mathbb{R}.$$

Thus

$$\exists i \in [n], \ (\tilde{e}_i - \tilde{p})' \cdot A(\tilde{p}) > (\tilde{e}_i - \tilde{p})' \cdot B(\tilde{p}). \tag{37}$$

Let $p_\eta := p + \eta(e_i - p)$ and denote by $C(\tilde{p}_\eta)$ the supergradient associated with ℓ at p_η (that is, $\ell_i(p_\eta) = \underline{L}(p_\eta) + (\tilde{e}_i - \tilde{p}_\eta)' \cdot C)\tilde{p}_\eta)$). By definition of the supergradient,

$$\underline{L}(p_\eta) \leq \underline{L}(p) + (\tilde{p}_\eta - \tilde{p})' \cdot B(\tilde{p}) \ \text{ and } \ \underline{L}(p) \leq \underline{L}(p_\eta) + (\tilde{p} - \tilde{p}_\eta)' \cdot C(\tilde{p}_\eta).$$

Thus

$$\begin{aligned} & \underline{L}(p_\eta) \ \leq \ \underline{L}(p_\eta) + C(\tilde{p}_\eta)' \cdot (\tilde{p} - \tilde{p}_\eta) + B(\tilde{p})' \cdot (\tilde{p}_\eta - \tilde{p}) \\ \Rightarrow \ & C(p_\eta)' \cdot (\tilde{p}_\eta - \tilde{p})' \ \leq \ B(\tilde{p})' \cdot (\tilde{p}_\eta - \tilde{p})'. \end{aligned}$$

But by definition of p_η, $\tilde{p}_\eta - \tilde{p} = \tilde{p} + \eta(\tilde{e}_i - \tilde{p}) - \tilde{p} = \eta(\tilde{e}_i - \tilde{p})$. Thus for $\eta > 0$,

$$C(\tilde{p}_\eta)' \cdot (\tilde{e}_i - \tilde{p}) \ \leq \ B(\tilde{p})' \cdot (\tilde{p} - \tilde{e}_i). \tag{38}$$

Now $\ell_i(p_\eta) = \underline{L}(p_\eta) + (\tilde{e}_i - \tilde{p}_\eta)' \cdot C(\tilde{p}_\eta)$. Hence (38) implies

$$\ell_i(p_\eta) \leq \underline{L}(p_\eta) + (\tilde{e}_i - \tilde{p})' \cdot B(\tilde{p}).$$

However $\lim_{\eta \searrow 0} p_\eta = p$ and by continuity of $\underline{L}$,

$$\begin{aligned} \lim_{\eta \searrow 0} \underline{L}(p_\eta) + (\tilde{e}_i - \tilde{p})' \cdot B(\tilde{p}) \ &= \ \underline{L}(p) + (\tilde{e}_i - \tilde{p})' \cdot B(\tilde{p}) \\ &< \ \underline{L}(p) + (\tilde{e}_i - \tilde{p})' \cdot A(\tilde{p}) \\ &= \ \ell_i(p) \text{ by (37).} \end{aligned}$$

Thus $\lim_{\eta \searrow 0} \ell_i(p_\eta) < \ell_i(p)$ and so ℓ_i is not continuous at p and thus ℓ is not continuous at p.

B.3 Proof of Proposition 7

The proof shows the equivalence of statements 1 and 2 and, separately the equivalence of 1 and 3 and 1 and 4.

$1 \Rightarrow 2$: Suppose that ℓ is proper and $p, q \in \partial\Delta^n$. Let $\tilde{L}^{p,q}$ denote the conditional risk associated with $\tilde{\ell}^{p,q}$. Then

$$\begin{aligned} & \tilde{L}^{p,q}(\eta, \hat{\eta}) = \big(\eta q + (1 - \eta)p\big)' \cdot \ell\big(p + \hat{\eta}(q - p)\big) = L\big(p + \eta(q - p), p + \hat{\eta}(q - p)\big) \\ \geq & L\big(p + \eta(q - p), p + \eta(q - p)\big) = \tilde{L}^{p,q}(\eta, \eta). \end{aligned}$$

Hence $\tilde{\ell}^{p,q}$ is proper.

1 ⇐ 2: Suppose that $\tilde{\ell}^{p,q}$ is proper $\forall p,q \in \partial\Delta^n$. Suppose $p,q \in \Delta^n$. Then there exists $\tilde{p}$ and $\tilde{q} \in \partial\Delta^n$ such that $p = \tilde{p} + \eta(\tilde{q} - \tilde{p})$ and $q = \tilde{p} + \hat{\eta}(\tilde{q} - \tilde{p})$, where $\eta, \hat{\eta} \in [0,1]$ (the line passing through p and q cuts $\partial\Delta^n$ at $\tilde{p}$ and $\tilde{q}$; see Figure 12). Then

$$L(p,q) = \tilde{L}^{\tilde{p},\tilde{q}}(\eta,\hat{\eta}) \geq \tilde{L}^{\tilde{p},\tilde{q}}(\eta,\eta) = L(p,p).$$

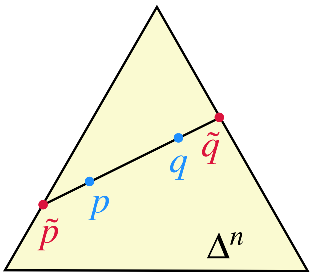

Figure 12: Illustration of proof of Proposition 7.

Hence ℓ is proper.

One can easily prove that 3 ⇒ 1 by taking $h_1 = 0$.

For 3 ⇐ 1 we use a result of Lambert (2010, Proposition 1), which tells us a binary probability estimation loss ℓ_b is proper if and only if $\forall \eta \leq \eta_1 \leq \eta_2$ or $\eta \geq \eta_1 \geq \eta_2$, $L_b(\eta,\eta_1) \leq L_b(\eta,\eta_2)$ (the assumptions on the statistic are checked in the binary case with the statistic function $\Gamma\colon \Delta^2 \ni p \mapsto \mathbb{E}(p) \in [0,1]$). We also know that if ℓ is proper then $\forall p,q \in \partial\Delta^n$, $\tilde{\ell}^{p,q}$ (introduced in Proposition 7) is proper. We assume that ℓ is proper, $\forall p,q \in \Delta^n$, $\forall 0 \leq h_1 \leq h_2$, we introduce the projections $\tilde{p},\tilde{q} \in \partial\Delta^n$ of p and q, then there exists η and μ such that $p = \tilde{p} + \eta(\tilde{q} - \tilde{p})$ and $q = \tilde{p} + \mu(\tilde{q} - \tilde{p})$. We denote $\eta_1 = \eta + h_1(\mu - \eta)$ and $\eta_2 = \eta + h_2(\mu - \eta)$. Then the result of Lambert applied to $\tilde{\ell}^{p,q}$ gives us $L(p, p + h_1(q - p)) \leq L(p, p + h_2(q - p))$. One can adapt the proof in the case of strict properness.

1 ⇒ 4: If ℓ is proper, $p' \cdot \ell(q) = q' \cdot \ell(q) + (p - q)' \cdot \ell(q) = \underline{L}(q) + (p - q)' \cdot \ell(q)$. Thus $\forall q \in \Delta^n$ there exists $A(q)$ such as $L(p,q) = \underline{L}(q) + (p - q)' \cdot A(q)$. Since ℓ is proper, $\forall p \in \Delta^n$, $0 \leq L(p,p) - L(p,q) = \underline{L}(q) - \underline{L}(p) + (p - q)' \cdot A(q)$. Then $A(q)$ is a supergradient of $\underline{L} = f$ (which is concave) at q, and $p' \cdot \ell(q) = f(q) + (p - q)' \cdot A(q)$.

4 ⇒ 1: If there exists a function f concave and $\forall q \in \Delta^n$, there exists a supergradient $A(q) \in \partial f(q)$ such that $\forall p,q \in \Delta^n$, $p' \cdot \ell(q) = f(q) + (p - q)' \cdot A(q)$. Then, $L(p,p) - L(p,q) = f(p) - f(q) + (p - q)' \cdot A(q) \geq 0$. Hence ℓ is proper.

B.4 Proof of Proposition 10

The proposition is a direct consequence of the characterization of differentiable binary proper losses (Reid and Williamson, 2010). A differentiable binary loss λ is proper if and only if $\frac{-\lambda_1'(\eta)}{1-\eta} = \frac{\lambda_{-1}'(\eta)}{\eta} \geq 0$, $\forall \eta \in (0,1)$.

Suppose the loss ℓ can be expressed as a proper composite loss: $\ell = \lambda^{\psi} = \lambda \circ \psi^{-1}$ and so $\lambda = \ell \circ \psi$. Therefore for $y \in \{-1,1\}$, $\lambda_y'(\eta) = \psi'(\eta)\ell_y'(\psi(\eta))$. Then λ is proper and thus

$$\frac{-\lambda_1'(\eta)}{1-\eta} = \frac{\lambda_{-1}'(\eta)}{\eta}, \ \forall \eta \in (0,1) \tag{39}$$

$$\Leftrightarrow -\frac{\psi'(\psi^{-1}(v))}{1-\psi^{-1}(v)}\ell_1'(v) = \frac{\psi'(\psi^{-1}(v))}{\psi^{-1}(v)}\ell_{-1}'(v), \ \forall v \in \mathscr{V}$$

$$\Leftrightarrow \psi'(\psi^{-1}(v)) = 0 \text{ or } \ell_{-1}'(v) = \ell_1'(v) = 0 \text{ or } \psi^{-1}(v) = \frac{\ell_{-1}'(v)}{\ell_{-1}'(v) - \ell_1'(v)}, \ \forall v \in \mathscr{V}. \tag{40}$$

Since ψ is differentiable and invertible, ψ' cannot equal zero on an interval. By continuity, ψ^{-1} is uniquely defined on an interval I when $\forall v_1, v_2 \in I$, $\exists v \in [v_1, v_2]$, $\ell_1'(v) \neq 0$ or $\ell_{-1}'(v) \neq 0$. If $I = \mathscr{V}$ then ψ is unique and thus $\lambda = \ell \circ \psi$ is unique.

If $\ell_1'(v) = \ell_{-1}'(v) = 0$, $\forall v \in [v_1, v_2]$ then one can choose any $\psi|_{[v_1,v_2]}$ which is differentiable, invertible and such that ψ is continuous in v_1 and v_2 and as ℓ_1 and ℓ_{-1} are constant on $[v_1, v_2]$, $\lambda(\eta) = \ell(\psi(\eta))$ does not depend on ψ and so in any case λ is unique.

B.5 Proof of Proposition 11

The loss λ is proper if and only if (39) and $-\lambda_1'(\eta) \geq 0$ and $\lambda_{-1}'(\eta) \geq 0$. This is equivalent to there exists an invertible ψ such that (40) holds and

$$-\psi'(\psi^{-1}(v))\ell_1'(v) \geq 0 \text{ and } \psi'(\psi^{-1}(v))\ell_{-1}'(v) \geq 0,\ \forall v \in \mathscr{V}. \tag{41}$$

($\Rightarrow$) Suppose ℓ has a composite representation with ψ strictly increasing and thus $\psi'(v) > 0$ for all $v \in \mathscr{V}$ and thus $-\ell_1'(v) \geq 0$ and $\ell_{-1}'(v) \geq 0$. Hence ℓ_1 is decreasing and ℓ_{-1} is increasing. By hypothesis, $\ell_{-1}'(v) \neq 0$ or $\ell_1'(v) \neq 0$. Furthermore $\psi'(v)$ can not equal zero except at isolated points. Thus (40) implies $\psi^{-1}(v) = \frac{\ell_{-1}'(v)}{\ell_{-1}'(v) - \ell_1'(v)} = \frac{1}{1-f(v)}$ and thus f is strictly increasing. (If instead ψ was strictly decreasing, we can run the same argument to conclude ℓ_1 is increasing, ℓ_{-1} is decreasing and f is strictly decreasing.)

($\Leftarrow$) Suppose ℓ_1 is decreasing, ℓ_{-1} is increasing and f is strictly increasing. By setting $\psi^{-1}(v) = \frac{1}{1-f(v)}$, ψ^{-1} is invertible and (41) holds. The other case is analogous.

B.6 Proof of Proposition 17

Fix an arbitrary $p \in \Delta^n$. The function f_p is quasi-convex if its α sublevel sets

$$F_p^\alpha := \{q \in \Delta^n \colon p'\ell(q) \leq \alpha\}$$

are convex for all $\alpha \in \mathbb{R}$ (Greenberg and Pierskalla, 1971). Fix an arbitrary $\alpha > \underline{L}(p)$i, and thus $F_p^\alpha \neq \emptyset$. Let

$$Q_p^\alpha := \{x \in \mathbb{R}^n \colon p'x \leq \alpha\}$$

so $F_p^\alpha = \{q \in \Delta^n \colon \ell(q) \in Q_p^\alpha\}$. Denote by

$$h_q^\beta := \{x \colon x' \cdot q = \beta\}$$

the hyperplane in direction $q \in \Delta^n$ with offset $\beta \in \mathbb{R}$ and by

$$H_q^\beta := \{x \colon x' \cdot q \geq \beta\}$$

the corresponding half-space. Since ℓ is proper, $\mathscr{S}_\ell$ is supported at $x = \ell(q)$ by the hyperplane $h_q^{\underline{L}(q)}$ and furthermore since $\mathscr{S}_\ell$ is convex, $\mathscr{S}_\ell = \bigcap_{q \in \Delta^n} H_q^{\underline{L}(q)}$.

Let

$$V_p^\alpha := \bigcap_{x \in \ell(\Delta^n) \cap Q_p^\alpha} H_{\ell^{-1}(x)}^{\underline{L}(\ell^{-1}(x))} = \bigcap_{q \in F_p^\alpha} H_q^{\underline{L}(q)}$$

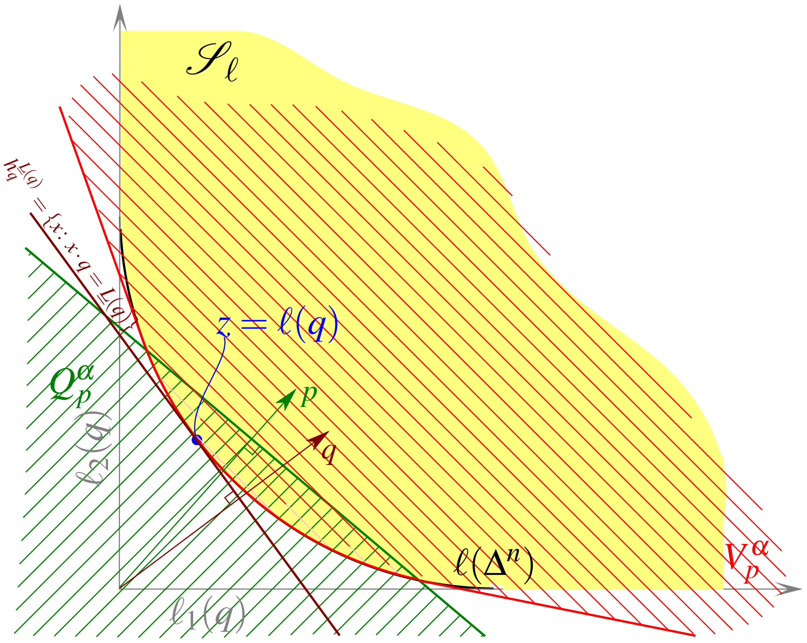

Figure 13: Illustration of proof of quasi-convexity of continuous proper losses (see text).

(see figure 13). Since V_p^α is the intersection of halfspaces it is convex. Note that a given half-space $H_q^{\underline{L}(q)}$ is supported by exactly one hyperplane, namely $h_q^{\underline{L}(q)}$. Thus the set of hyperplanes that support V_p^α is $\{h_q^{\underline{L}(q)} : q \in F_p^\alpha\}$ If $u \in F_p^\alpha$ then there is a hyperplane in direction u that supports V_p^α and its offset is given by

$$\sigma_{V_p^\alpha}(u) := \inf_{x \in V_p^\alpha} u' \cdot x = \underline{L}(p) > -\infty$$

whereas if $u \notin F_p^\alpha$ then for all $\beta \in \mathbb{R}$, h_u^β does not support V_p^α and hence $\sigma_{V_p^\alpha}(u) = -\infty$. Thus we have shown

$$\left(u \notin W_p^\alpha\right) \Leftrightarrow \left(\sigma_{V_p^\alpha}(u) = -\infty\right).$$

Observe that $\sigma_{V_p^\alpha}(u) = -s_{V_p^\alpha}(-u)$ where $s_C(u) = \sup_{x \in C} u' \cdot x$ is the support function of a set C. It is known (Valentine, 1964, Theorem 5.1) that the "domain of definition" of a support function $\{u \in \mathbb{R}^n : s_C(u) < +\infty\}$ for a convex set C is always convex. Thus $G_p^\alpha := \{u \in \Delta^n : \sigma_{V_p^\alpha}(u) > -\infty\} = \{u \in \mathbb{R}^n : \sigma_{V_p^\alpha}(u) > -\infty\} \cap \Delta^n$ is always convex because it is the intersection of convex sets. Finally by observing that

$$G_p^\alpha = \{p \in \Delta^n : \ell(p) \in \ell(\Delta^n) \cap Q_p^\alpha\} = F_p^\alpha$$

we have shown that F_p^α is convex. Since $p \in \Delta^n$ and $\alpha \in \mathbb{R}$ were arbitrary we have thus shown that f_p is quasi-convex for all $p \in \Delta^n$.

For the converse, the convexity of S comes from the contruction in the first part of this proof. All that needs to be shown is uniqueness. Suppose then, for the sake of a contradiction, that a given quasi-convex f_p corresponds to two distinct S_1, S_2. Then the corresponding support functions $s_{S_1} \neq s_{S_2}$ and thus there exists u such that $s_{S_1}(u) \neq s_{S_2}(u)$. Let x_i be the point of support of S_i by a hyperplane with normal u. By assumption $x_1 \neq x_2$. Suppose x_1 and x_2 do not lie on a single hyperplane with normal p. Then there exists α such that h_p^α seperates x_1 and x_2. Thus $u \in \operatorname{dom} s_{S_1 \cap Q_p^\alpha}$ but $u \notin \operatorname{dom} s_{S_2 \cap Q_p^\alpha}$ (or vice versa). Since $F_p^\alpha = -\operatorname{dom} s_{S_1 \cap Q_p^\alpha}$ and $F_p^\alpha = -\operatorname{dom} s_{S_2 \cap Q_p^\alpha}$ we have a contradiction. In case x_1, x_2 do lie on a single hyperplane $h_p^{\alpha^*}$, one can argue (by the continuity of s_S that there must exist a u^*, a convex combination of u and p which induces support points that can be seperated as above.

B.7 Proof of Proposition 18

Δ^n-smooth $\Rightarrow$ proper composite ℓ: Suppose $\ell(\mathscr{V})$ is Δ^n-smooth. Pick some $x \in \ell(\mathscr{V})$ with $x = \ell(v)$ for some $v \in \mathscr{V}$ (not necessarily unique because we have not assumed ℓ is invertible). By assumption, there exists a unique $p \in \Delta^n$ such that h_p^β supports $\ell(\mathscr{V})$ at x for some $\beta \in \mathbb{R}$. Define ψ to be the function such that $\psi(p) = v$. Since the corresponding p is unique, ψ is invertible. Now

$$\beta = \inf\{b \in \mathbb{R} : h_p^b \cap \ell(\mathscr{V}) \neq \emptyset\} = p' \cdot \ell(v) = p' \cdot \ell(\psi(p)) = p' \cdot \lambda(p).$$

Let $\lambda := \ell \circ \psi$. We will show λ is proper. Observe that for each $p \in \Delta^n$

$$\inf\{b\colon h_p^b \cap \ell(\mathscr{V}) \neq \emptyset\} = \inf\{p' \cdot \ell(v)\colon h_p^{p' \cdot \ell(v)} \cap \ell(\mathscr{V}) \neq \emptyset,\ v \in \mathscr{V}\}.$$

Thus $p' \cdot \lambda(p) = p' \cdot (\ell \circ \psi)(p) = \inf_{v \in \mathscr{V}} p' \cdot \ell(v)$.

Since $\psi\colon \Delta^n \to \mathscr{V}$ is invertible,

$$\inf_{v \in \mathscr{V}} p' \cdot \ell(v) = \inf_{v \in \mathscr{V}} p' \cdot (\lambda \circ \psi^{-1})(v) = \inf_{q \in \Delta^n} p' \cdot \lambda(q)$$

which we have shown above equals $p' \cdot \lambda(p)$. Thus λ is proper and ℓ has a proper composite representation.

Proper composite $\ell \Rightarrow \Delta^n$**-smooth**: Suppose $\ell = \lambda \circ \psi^{-1}$, where λ is proper. We need to show that for all $x \in \ell(\mathscr{V})$ there is a unique $p \in \Delta^n$ such that $\ell(\mathscr{V})$ is supported by h_p^β at x for some $\beta \in \mathbb{R}$.

Now pick an arbitrary $v \in \mathscr{V}$ which induces an arbitrary $x = \ell(v) \in \ell(\mathscr{V})$. Let $p = \psi^{-1}(v)$. Then $h_p^{\underline{L}(p)}$ supports $\ell(\mathscr{V})$ at x since λ is proper. Suppose there was another $q \neq p$, $q \in \Delta^n$ such that $h_q^{\underline{L}(q)}$ supports $\ell(\mathscr{V})$ at x. But that would require that $v = \psi(q)$ which is impossible since $v = \psi(p)$ and ψ is invertible. Thus p is unique and hence $\ell(\mathscr{V})$ is Δ^n-smooth.

Δ^n-strict convexity $\Rightarrow$ strict proper composite ℓ: Let $p \in \Delta^n$. By invertibility of ℓ and Δ^n-strict convexity of $\ell(\mathscr{V})$ there exists a unique $v \in \mathscr{V}$ such that there exists a hyperplane h_p^β supporting $\ell(\mathscr{V})$ at $\ell(v)$. Define ψ such that for all $p \in \Delta^n$, $\psi(p)$ is this unique v. Since h_p^β supports $\ell(\mathscr{V})$,

$$\beta = \inf\{b\colon h_p^b \cap \ell(\mathscr{V}) \neq \emptyset\} = p' \cdot \ell(v) = p' \cdot \ell(\psi(p)).$$

By Δ^n-smoothness of $\ell(\mathscr{V})$ for all $v \in \mathscr{V}$ there is a unique $p^* \in \Delta^n$ such that $h_{p^*}^\beta$ supports $\ell(\mathscr{V})$ at $\ell(v)$. By continuity of ℓ and Δ^n-strict convexity of $\ell(\mathscr{V})$, ψ is continuous. Let $\lambda := \ell \circ \psi$.

Observe that

$$\inf\{\beta\colon h_p^\beta \cap \ell(\mathscr{V}) \neq \emptyset\} = \inf\{p' \cdot \ell(v)\colon h_p^{p'\cdot\ell(v)} \cap \ell(\mathscr{V}) \neq \emptyset,\ v \in \mathscr{V}\}.$$

Thus $p' \cdot \lambda(p) = p' \cdot (\ell \circ \psi)(p) = \inf_{v \in \mathscr{V}} p' \cdot \ell(v)$. By Δ^n-smoothness of $\ell(\mathscr{V})$, for all $v \in \mathscr{V}$ there exists a unique $p \in \Delta^n$ such that $\psi(p) = v$ and thus ψ is invertible. Hence $p' \cdot \lambda(p) = \inf_{q \in \Delta^n} p' \cdot \lambda(q)$ and thus λ is proper. Since $\ell(\mathscr{V})$ is Δ^n-strictly convex there exists a unique point where $h_p^{\Lambda(p)}$ supports $\ell(\mathscr{V})$. Hence λ is strictly proper and we have shown that ℓ has a strictly proper composite representation.

Strictly proper composite $\ell \Rightarrow \Delta^n$-strictly convex: Suppose ℓ has a strictly proper composite representation $\ell(v) = \lambda(\psi^{-1}(v))$. Pick $p \in \Delta^n$. By assumption, there exists $v \in \mathscr{V}$ such that $\psi^{-1}(v) = p$. Since λ is strictly proper, there is a unique $q \in \Delta^n$ which minimises $q \mapsto p' \cdot \lambda(q)$. By invertibility of ψ, there thus exists a unique $v \in \mathscr{V}$ that minimises $v \mapsto p' \cdot \ell(v)$ and so there is a unique x at which h_p^β supports $\ell(\mathscr{V})$ for some $\beta \in \mathbb{R}$. Thus $\ell(\mathscr{V})$ is Δ^n-strictly convex.

Now pick an arbitrary $v \in \mathscr{V}$ which induces an arbitrary $x = \ell(v) \in \ell(\mathscr{V})$. Let $p = \psi^{-1}(v)$. Then $h_p^{L(p)}$ supports $\ell(\mathscr{V})$ at x since λ is proper. Suppose there was another $q \neq p$, $q \in \Delta^n$ such that $h_q^{L(q)}$ supports $\ell(\mathscr{V})$ at x. But that would require that $v = \psi(q)$ which is impossible since $v = \psi(p)$ and ψ is invertible. Thus p is unique and $\ell(\mathscr{V})$ is Δ^n-smooth.

References

Jacob Abernethy, Alekh Agarwal, Peter L. Bartlett, and Alexander Rakhlin. A stochastic view of optimal regret through minimax duality. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.

Erin L. Allwein, Robert E. Schapire, and Yoram Singer. Reducing multiclass to binary: A unifying approach for margin classifiers. *The Journal of Machine Learning Research*, 1:113–141, 2001.

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

Paul N. Bennett. Using asymmetric distributions to improve text classifier probability estimates. In *Proceedings of SIGIR'03*, pages 111–118, 2003.

James O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, New York, 1985.

Alina Beygelzimer, John Langford, and Pradeep Ravikumar. Multiclass classification with filter trees. Preprint, June 2007. URL http://hunch.net/~jl/projects/reductions/mc_to_b/invertedTree.pdf.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Efim Mikhailovich Bronshtein. Extremal convex functions. *Siberian Mathematical Journal*, 19: 6–12, 1978.

Lawrence D. Brown. A complete class theorem for statistical problems with finite sample spaces. *The Annals of Statistics*, 9(6):1289–1300, 1981.

Andreas Buja, Werner Stuetzle, and Yi Shen. Loss functions for binary class probability estimation and classification: Structure and applications. Technical report, University of Pennsylvania, November 2005. URL http://www-stat.wharton.upenn.edu/~buja/PAPERS/paper-proper-scoring.pdf.

Hermann Chernoff and Lincoln E. Moses. *Elementary Decision Theory*. Dover, 1986.

Jesús Cid-Sueiro and Aníbal R. Figueiras-Vidal. On the structure of strict sense Bayesian cost functions and its applications. *IEEE Transactions on Neural Networks*, 12(3):445–455, May 2001.

Ira Cohen and Moises Goldszmidt. Properties and benefits of calibrated classifiers. Technical Report HPL-2004-22(R.1), HP Laboratories, Palo Alto, July 2004. URL http://www.hpl.hp.com/techreports/2004/HPL-2004-22R1.pdf.

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

A. Philip Dawid. The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, 59(1):77–93, March 2007.

Morris H. DeGroot. Uncertainty, Information, and Sequential Experiments. *The Annals of Mathematical Statistics*, 33(2):404–419, 1962.

Thomas G. Dietterich and Ghulum Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2:263–286, 1995.

Paul K. Fackler. Notes on matrix calculus. North Carolina State University, 2005. URL http://www4.ncsu.edu/~pfackler/MatCalc.pdf.

Thomas S. Ferguson. *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press, New York, 1967.

Dario García-García and Robert C. Williamson. Divergences and risks for multiclass experiments. In *Conference on Learning Theory (JMLR: W&CP)*, volume 23, pages 28.1–28.20, 2012.

Gary F.V. Glonek. A class of regression models for multivariate categorical responses. *Biometrika*, 83(1):15–28, 1996.

Gary F.V. Glonek and Peter McCullagh. Multivariate logistic models. *Journal of the Royal Statistical Society, Series B*, 57(3):533–546, 1995.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, March 2007.

Harvey J. Greenberg and William P. Pierskalla. A review of quasi-convex functions. *Operations Research*, 19(7):1553–1570, November 1971.

Peter D. Grünwald and A. Philip Dawid. Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *The Annals of Statistics*, 32(4):1367–1433, 2004.

Trevor Hastie and Robert Tibshirani. Classification by pairwise coupling. *The Annals of Mathematical Statistics*, 26(2):451–471, 1998.

Simon I. Hill and Arnaud Doucet. A framework for kernel-based multi-category classification. *Journal of Artificial Intelligence Research*, 30:525–564, 2007.

Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of Convex Analysis*. Springer, Berlin, 2001.

Roger A. Horn and Charles A. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1991.

Tzu-Kuo Huang, Ruby C. Weng, and Chih-Jen Lin. Generalized Bradley-Terry models and multi-class probability estimates. *Journal of Machine Learning Research*, 7:85–115, 2006.

Ayodele Ighodaro, Thomas Santner, and Lawrence Brown. Admissibility and complete class results for the multinomial estimation problem with entropy and squared error loss. *Journal of Multivariate Analysis*, 12:469–479, 1982.

Yuri Kalnishkan and Michael V. Vyugin. The weak aggregating algorithm and weak mixability. *Journal of Computer and System Sciences*, 74:1228–1244, 2008.

Parameswaran Kamalaruban, Robert C. Williamson, and Xinhua Zhang. Exp-concavity of proper composite losses. In *JMLR Workshop and Conference Proceedings (Proceedings COLT 2015)*, volume 40, 2015.

Jack Carl Kiefer. *Introduction to Statistical Inference*. Springer-Verlag, New York, 1987.

Nicolas S. Lambert. Elicitation and evaluation of statistical forecasts. Technical report, Stanford University, March 2010. URL http://www.stanford.edu/~nlambert/lambert_elicitation.pdf.

Yufeng Liu. Fisher consistency of multicategory support vector machines. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, pages 289–296, 2007.

Jan R. Magnus and Heinz Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics (revised edition)*. John Wiley & Sons, 1999.

Peter McCullagh and John A. Nelder. *Generalized Linear Models*. Chapman & Hall/CRC, 1989.

Aditya K. Menon and Robert C. Williamson. Bipartite ranking: risk, optimality, and equivalences. *Journal of Machine Learning Research*, July 2014. URL http://users.cecs.anu.edu.au/~williams/papers/P194.pdf. Submitted.

Walter Meyer and David C. Kay. A convexity structure admits but one real linearization of dimension greater than one. *Journal of the London Mathematical Society (2)*, 7:124–130, 1973.

Geert Molenberghs and Emmanuel Lesaffre. Marginal modelling of multivariate categorical data. *Statistics in Medicine*, 18:2237–2255, 1999.

Indraneel Mukherjee and Robert E Schapire. A theory of multiclass boosting. *The Journal of Machine Learning Research*, 14(1):437–497, 2013.

Harikrishna Narasimhan and Shivani Agarwal. On the relationship between binary classification, bipartite ranking, and binary class probability estimation. In *Advances in Neural Information Processing Systems*, pages 2913–2921, 2013.

Robert F. Nau. Should scoring rules be 'effective'? *Management Science*, 31(5):527–535, May 1985.

Tapan K. Nayak and Dayanand N. Naik. Estimating multinomial cell probabilities under quadratic loss. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 38(1): 3–10, 1989.

XuanLong Nguyen, Martin J. Wainwright, and Michael I. Jordan. On surrogate loss functions and f-divergences. *Annals of Statistics*, 37:876–904, 2009.

Robert R. Phelps. *Lectures on Choquet's Theorem*, volume 1757 of *Lecture Notes in Mathematics*. Springer, 2nd edition, 2001.

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

Mark D. Reid and Robert C. Williamson. Information, divergence and risk for binary experiments. *Journal of Machine Learning Research*, 12:731–817, March 2011.

Mark D. Reid, Rafael M. Frongillo, Robert C. Williamson, and Nishant A. Mehta. Generalized mixability via entropic duality. *JMLR: Workshop and Conference Proceedings (COLT2015)*, 40:1–22, 2015.

R. Tyrrell Rockafellar and Roger J-B. Wets. *Variational Analysis*. Springer-Verlag, Berlin, 2004.

Raúl Santos-Rodríguez, Alicia Guerrero-Curieses, Rocío Alaiz-Rodriguez, and Jesús Cid-Sueiro. Cost-sensitive learning based on Bregman divergences. *Machine Learning*, 76:271–285, 2009.

Rolf Schneider. *Convex Bodies: The Brunn-Minkowski Theory*. Cambridge University Press, 1993.

Clayton Scott. Surrogate losses and regret bounds for cost-sensitive classificationwith example-dependent costs. In *Proc. of the 28th International Conference on Machine Learning (ICML)*, 2011.

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

Qinfeng Shi, Mark Reid, and Tiberio Caetano. Conditional random fields and support vector machines: A hybrid approach. arXiv:1009:3346v1, September 2010. URL http://arxiv.org/PS_cache/arxiv/pdf/1009/1009.3346v1.pdf.

Barry Simon. *Convexity: An Analytic Viewpoint*. Cambridge University Press, 2011.

Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.

Ingo Steinwart. How to compare different loss functions. *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.

Frederick A. Valentine. *Convex Sets*. McGraw-Hill, New York, 1964.

Tim van Erven, Mark D. Reid, and Robert C. Williamson. Mixability is Bayes risk curvature relative to log loss. In *Proceedings of the 24th Annual Conference on Learning Theory*, 2011.

Tim van Erven, Peter Grünwald, Mark D Reid, and Robert C Williamson. Mixability in statistical learning. In *Advances in Neural Information Processing Systems*, pages 1691–1699, 2012a.

Tim van Erven, Mark D. Reid, and Robert C. Williamson. Mixability is Bayes risk curvature relative to log loss. *Journal of Machine Learning Research*, 13:1639–1663, May 2012b.

Tim van Erven, Peter D. Grünwald, Nishant A. Mehta, Mark D. Reid, and Robert C. Williamson. Fast rates in statistical and online learning. *Journal of Machine Learning Research*, 16:1793–1861, 2015.

William J. Vetter. Derivative operations on matrices. *IEEE Transactions on Automatic Control*, 15(2):241–244, April 1970.

Volodya Vovk. A game of prediction with expert advice. In *Proceedings of the Eighth Annual Conference on Computational Learning Theory*, pages 51–60. ACM, 1995.

Volodya Vovk and Fedor Zhdanov. Prediction with expert advice for the Brier game. *Journal of Machine Learning Research*, 10:2445–2471, 2009.

Roger Webster. *Convexity*. Oxford University Press, 1994.

Robert C. Williamson. The geometry of losses. In *Conference on Learning Theory (JMLR: W&CP)*, volume 35, pages 1078–1108, 2014.

Ting-Fan Wu and Ruby C. Weng. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5:975–1005, 2004.

Tong Tong Wu and Kenneth Lange. Multicategory vertex discriminant analysis for high-dimensional data. *The Annals of Applied Statistics*, 4:1698–1721, 2010.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of SIGKDD*, 2002.

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

Zhihua Zhang, Michael I. Jordan, Wu-Jun Li, and Dit-Yan Yeung. Coherence functions for multicategory margin-based classification methods. In *Proceedings of the Twelfth Conference on Artificial Intelligence and Statistics (AISTATS)*, 2009.

Ji Zhu, Hui Zou, Saharaon Rosset, and Trevor Hastie. Multi-class AdaBoost. *Statistics and its Interface*, 2:349–360, 2009.

Hui Zou, Ji Zhu, and Trevor Hastie. The margin vector, admissible loss and multi-class margin-based classifiers. Preprint, 2005. URL http://www-stat.stanford.edu/~hastie/Papers/margin.pdf.

Hui Zou, Ji Zhu, and Trevor Hastie. New multicategory boosting algorithms based on multicategory Fisher-consistent losses. *The Annals of Applied Statistics*, 2(4):1290–1306, 2008.