Composite Multiclass Losses

Elodie Vernet
ENS Cachan
evernet@ens-cachan.fr

Robert C. Williamson
ANU and NICTA
Bob.Williamson@anu.edu.au

Mark D. Reid
ANU and NICTA
Mark.Reid@anu.edu.au

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 determine the stationarity condition, Bregman representation, order-sensitivity, existence and uniqueness of the composite representation for multiclass losses. We subsume existing results on "classification calibration" by relating it to properness and show that the simple integral representation for binary proper losses can not be extended to multiclass losses.

1 Introduction

The motivation of this paper is to understand the intrinsic structure and properties of suitable loss functions for the problem of multiclass prediction, which includes *multiclass probability estimation*. Suppose we are given a data sample $S := (x_i, y_i)_{i\in[m]}$ where $x_i \in \mathscr{X}$ is an observation and $y_i \in \{1,..,n\} =: [n]$ is its corresponding class. We assume the sample S is drawn iid according to some distribution $\mathbb{P} = \mathbb{P}_{\mathscr{X},\mathscr{Y}}$ on $\mathscr{X} \times [n]$. 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 $\hat{y} = \arg\max_{i\in[n]} p_i$.

A loss measures the quality of prediction. Let $\Delta^n := \{(p_1, \ldots, p_n) \colon \sum_{i\in[n]} p_i = 1, \text{and } 0 \le p_i \le 1, \forall i \in [n]\}$ denote the *n-simplex*. For multiclass probability estimation, $\ell \colon \Delta^n \to \mathbb{R}^n_+$. For classification, the loss $\ell \colon [n] \to \mathbb{R}^n_+$. The *partial losses* ℓ_i are the components of $\ell(q) = (\ell_1(q), \ldots, \ell_n(q))'$.

Proper losses are particularly suitable for probability estimation. They have been studied in detail when $n = 2$ (the "binary case") where there is a nice integral representation [1, 2, 3], and characterization [4] when differentiable. Classification calibrated losses are an analog of proper losses for the problem of classification [5]. The relationship between classification calibration and properness was determined in [4] for $n = 2$. Most of these results have had no multiclass analogue until now.

The design of losses for multiclass prediction has received recent attention [6, 7, 8, 9, 10, 11, 12] although none of these papers developed the connection to proper losses, and most restrict consideration to margin losses (which imply certain symmetry conditions). Glasmachers [13] has shown that certain learning algorithms can still behave well when the losses do not satisfy the conditions in these earlier papers because the requirements are actually stronger than needed.

Our contributions are: We relate properness, classification calibration, and the notion used in [8] which we rename "prediction calibrated" §3; we provide a novel characterization of multiclass properness §4; we study composite proper losses (the composition of a proper loss with an invertible link) presenting new uniqueness and existence results §5; we show how the above results can aid in the design of proper losses §6; and we present a (somewhat surprising) negative result concerning the integral representation of proper multiclass losses §7. Many of our results are characterisations. Full proofs are provided in the extended version [14].

2 Formal Setup

Suppose $\mathscr{X}$ is some set and $\mathscr{Y} = \{1,\dots,n\} = [n]$ is a set of labels. We suppose we are given data $(x_i, y_i)_{i\in[m]}$ such that $Y_i \in \mathscr{Y}$ is the label corresponding to $x_i \in \mathscr{X}$. These data follow a joint distribution $\mathbb{P}_{\mathscr{X},\mathscr{Y}}$. 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 $\mathbb{P}_{\mathscr{X},\mathscr{Y}}$.

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 a typical learning problem one will make 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 $p(x) = (p_1(x),\dots,p_n(x))'$, p' is the transpose of p, and $p_i(x) = \mathbb{P}(\mathsf{Y}=i|\mathsf{X}=x)$. Thus it suffices to only consider the conditional risk; confer [3].

A loss $\ell\colon \Delta^n\to\mathbb{R}^n_+$ is *proper* if $L(p,p)\le L(p,q)$, $\forall p,q\in\Delta^n$. It is *strictly proper* if the inequality is strict when $p\neq q$. The *conditional Bayes risk* $\underline{L}\colon \Delta^n\ni p\mapsto \inf_{q\in\Delta^n} L(p,q)$. This function is always concave [2]. 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, $p = \arg\min_q L(p,q)$.

In order to differentiate the losses we project the n-simplex into a subset of $\mathbb{R}^{n-1}$. We denote by $\Pi_\Delta\colon \Delta^n\ni p = (p_1,\dots,p_n)'\mapsto \tilde{p} = (p_1,\dots,p_{n-1})'\in\tilde{\Delta}^n := \{(p_1,\dots,p_{n-1})'\colon p_i\ge 0,\ \forall i\in[n],\ \sum_{i=1}^{n-1} p_i\le 1\}$, the projection of the n-simplex Δ^n, and $\Pi_\Delta^{-1}:\tilde{\Delta}^n\ni\tilde{p} = (\tilde{p}_1,\dots,\tilde{p}_{n-1})\mapsto p = (\tilde{p}_1,\dots,\tilde{p}_{n-1},1-\sum_{i=1}^{n-1}\tilde{p}_i)'\in\Delta^n$ its inverse.

The losses above are defined on the simplex Δ^n since the argument (an estimator) represents a probability vector. However it is sometimes desirable to use another set $\mathscr{V}$ of predictions. One can consider losses $\ell\colon\mathscr{V}\to\mathbb{R}^n_+$. Suppose there exists an invertible function $\psi\colon\Delta^n\to\mathscr{V}$. 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.

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 $\mathbb{1}_n := (1,\dots,1)'$. The derivative of a function f is denoted $\mathsf{D}f$ and its Hessian $\mathsf{H}f$. Let $\mathring{\Delta}^n := \{(p_1,\dots,p_n)\colon \sum_{i\in[n]} p_i = 1, \text{and } 0<p_i<1,\ \forall i\in[n]\}$ and $\partial\Delta^n := \Delta^n\setminus\mathring{\Delta}^n$.

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 one requires less. 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:

$$\mathscr{T}_i(c) := \{p\in\Delta^n\colon \forall j\neq i\ p_ic_j\ge p_jc_i\}.$$

Observe that for $i\neq j$, the sets $\{p\in\mathbb{R}\colon p_ic_j = p_jc_j\}$ are subsets of dimension $n-2$ through c and all e_k such that $k\neq i$ and $k\neq j$. These subsets partition Δ^n into two parts, the subspace $\mathscr{T}_i$ is the intersection of the subspaces delimited by the precedent $(n-2)$-subspace and in the same side as e_i. 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 $p\in\Delta^n$. $\mathscr{T}_i(c)\cap\mathscr{T}_j(c)\subseteq\{p\in\Delta^n\colon p_ic_j = p_jc_i\}$, 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 $q\notin\mathscr{T}_i(c)$.*

Classification calibrated losses have been developed and studied under some different definitions and names [6, 5]. Below we generalise the notion of c-calibration which was proposed for $n = 2$ in [4] as a generalisation of the notion of classification calibration in [5].

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.

Classification calibration in the sense used in [5] corresponds to $\frac{1}{2}$-calibrated losses when $n = 2$. If $c_{\text{mid}} := (\frac{1}{n}, \ldots, \frac{1}{n})'$, 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 [6]". 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 [4, 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 $c \in \mathring{\Delta}^n$.*

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 [5]:

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

Tewari and Bartlett [8] 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 keep the name of classification calibration for the notion linked to Fisher consistency (as defined before) and call prediction calibrated the notion of Tewari and Bartlett (equivalent to (1)).

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

$$\forall p \in \Delta^n\colon \inf_{z \in \mathscr{C}_\ell, p_{\text{pred}(z)} < \max_i p_i} p' \cdot z > \inf_{z \in \mathscr{C}_\ell} p' \cdot z = \underline{L}(p).$$

Observe that the class is predicted from $\ell(p)$ and not directly from p (which is equivalent if the loss is invertible). Suppose that $\ell\colon \Delta^n \to \mathbb{R}^n_+$ is such that ℓ is prediction calibrated and $\text{pred}(\ell(p)) \in \arg\max_i p_i$. Then ℓ is c_{mid}-calibrated almost everywhere.

By introducing a reference "link" $\bar{\psi}$ (which corresponds to the actual link if ℓ is a proper composite loss) 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 $\bar{\psi}(p) \in \arg\min_{v \in \mathscr{V}} L(p,v)$ and $\lambda = \ell \circ \bar{\psi}$. Then λ is proper. If ℓ is prediction calibrated then* $\text{pred}(\lambda(p)) \in \arg\max_i p_i$.

4 Characterizing Properness

We first present some simple (but new) consequences of properness. We say $f\colon C \subset \mathbb{R}^n \to \mathbb{R}^n$ is *monotone* on C when for all x and y in C, $(f(x) - f(y))' \cdot (x - y) \geq 0$; confer [15].

Proposition 6 *Suppose $\ell\colon \Delta^n \to \mathbb{R}^n_+$ is a loss. If ℓ is proper, then $-\ell$ is monotone.*

Proposition 7 *If ℓ is strictly proper then it is invertible.*

A theme of the present paper is the extensibility of results concerning binary losses to multiclass losses. The following proposition shows how the characterisation of properness in the general (not necessarily differentiable) multiclass case can be reduced to the binary case. In the binary case, the two classes are often denoted -1 and 1 and the loss is denoted $\ell = (\ell_1, \ell_{-1})'$. We project the 2-simplex Δ^2 into $[0,1]$: $\eta \in [0,1]$ is the projection of $(\eta, 1-\eta) \in \Delta^2$.

Proposition 8 *Suppose $\ell\colon \Delta^n \to \mathbb{R}^n_+$ is a loss. Define*

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

Then ℓ is (strictly) proper if and only if $\tilde{\ell}^{p,q}$ is (strictly) proper $\forall p, q \in \partial\Delta^n$.

This proposition shows that in order to check if a loss is proper one needs only to 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 $\forall \eta \in [0,1]$, $\frac{-\ell_1'(\eta)}{1-\eta} = \frac{\ell_{-1}'(\eta)}{\eta} \geq 0$, [4]). However this needs to be checked for all lines defined by $p, q \in \partial\Delta^n$. We now extend some characterisations of properness to the multiclass case by using Proposition 8.

Lambert [16] proved that in the binary case, properness is equivalent to the fact that the further your prediction is from reality, the larger the loss ("order sensitivity"). The result relied upon on the total order of $\mathbb{R}$. In the multiclass case, there does not exist such a total order. Yet, one can compare two predictions if they are in the same line as the true real class probability. The next result is a generalization of the binary case equivalence of properness and order sensitivity.

Proposition 9 *Suppose $\ell\colon \Delta^n \to \mathbb{R}^n_+$ is a loss. Then ℓ is (strictly) proper if and only if $\forall p, q \in \Delta^n$, $\forall 0 \leq h_1 \leq h_2$, $L(p, p + h_1(q-p)) \leq L(p, p + h_2(q-p))$ (the inequality is strict if $h_1 \neq h_2$).*

"Order sensitivity" tells us more about 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 optimization 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 $q\colon \mathscr{X} \to \Delta^n$. 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.

The order sensitivity along a line leads to a new characterisation of differentiable proper losses. As in the binary case, one condition comes from the fact that the derivative is zero at a minimum and the other ensures that it is really a minimum.

Corollary 10 *Suppose $\ell\colon \Delta^n \to \mathbb{R}^n_+$ is a loss such that $\tilde{\ell} = \ell \circ \Pi_\Delta^{-1}$ is differentiable. Let $M(p) = \mathsf{D}\tilde{\ell}(\Pi_\Delta(p)) \cdot \mathsf{D}\Pi_\Delta(p)$. Then ℓ is proper if and only if*

$$\left.\begin{array}{rcl} p' \cdot M(p) & = & 0 \\ (q-r)' \cdot M(p) \cdot (q-r) & \leq & 0 \end{array}\right\} \forall q, r \in \Delta^n,\ \forall p \in \mathring{\Delta}^n. \tag{2}$$

We know that for any loss, its Bayes risk $\underline{L}(p) = \inf_{q \in \Delta^n} L(p,q) = \inf_{q \in \Delta^n} p' \cdot \ell(q)$ is concave. If ℓ is proper, $\underline{L}(p) = p' \cdot \ell(p)$. Rather than working with the loss $\ell\colon \mathscr{V} \to \mathbb{R}^n_+$ we will now work with the simpler associated conditional Bayes risk $\underline{L}\colon \mathscr{V} \to \mathbb{R}_+$.

We need two definitions from [15]. Suppose $f : \mathbb{R}^n \to \mathbb{R}$ 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, the *superdifferential* $\partial f(x)$ of f at x is

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

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

The next proposition is a restatement of the well known Bregman representation of proper losses; see [17] for the differentiable case, and [2, Theorem 3.2] for the general case.

Proposition 11 *Suppose $\ell\colon \Delta^n \to \mathbb{R}^n_+$ is a loss. Then ℓ is proper if and only if there exists a concave function f and $\forall q \in \Delta^n$, there exists a supergradient $A(q) \in \partial f(q)$ such that*

$$\forall p,q \in \Delta^n,\ p' \cdot \ell(q) = L(p,q) = f(q) + (p-q)' \cdot A(q).$$

Then f is unique and $f(p) = L(p,p) = \underline{L}(p)$.

The fact that f is defined on a simplex is not a problem. Indeed, the superdifferential becomes $\partial f(x) = \{s \in \mathbb{R}^n\colon s' \cdot d \geq \mathsf{D}f(x,d), \forall d \in \Delta^n\} = \{s \in \mathbb{R}^n\colon f(y) \leq f(x) + s' \cdot (y-x),\ \forall y \in \Delta^n\}$. If $\tilde{f} = f \circ \Pi_\Delta^{-1}$ is differentiable at $\tilde{q} \in \tilde{\Delta}^n$, $A(q) = (\mathsf{D}\tilde{f}(\Pi_\Delta(q)), 0)' + \alpha \mathbb{1}'_n$, $\alpha \in \mathbb{R}$. Then $(p-q)' \cdot A(q) = \mathsf{D}\tilde{f}(\Pi_\Delta(q)) \cdot (\Pi_\Delta(p) - \Pi_\Delta(q))$. Hence for any concave differentiable function f, there exists an unique proper loss whose Bayes risk is equal to f (we say that f is differentiable when $\tilde{f}$ is differentiable).

The last property gives us the form of the proper losses associated with a 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

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

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 12 *Two proper losses ℓ^1 and ℓ^2 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.*

We say that $\underline{L}$ is differentiable at p if $\tilde{\underline{L}} = \underline{L} \circ \Pi_\Delta^{-1}$ is differentiable at $\tilde{p} = \Pi_\Delta(p)$.

Proposition 13 *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 $p \in \mathring{\Delta}^n$.*

5 The Proper Composite Representation: Uniqueness and Existence

It is sometimes helpful to define a loss on some set $\mathcal{V}$ rather than Δ^n; confer [4]. Composite losses (see the definition in §2) are a way of constructing such losses: given a proper loss $\lambda : \Delta^n \to \mathbb{R}^n_+$ and an invertible link $\psi\colon \Delta^n \to \mathcal{V}$, one defines $\lambda^\psi\colon \mathcal{V} \to \mathbb{R}^n_+$ using $\lambda^\psi = \lambda \circ \psi^{-1}$. We now consider the question: given a loss $\ell\colon \mathcal{V} \to \mathbb{R}^n_+$, 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 and study the uniqueness of the representation of a loss as a proper composite loss.

Proposition 14 *Suppose $\ell = \lambda \circ \psi^{-1}\colon \mathcal{V} \to \mathbb{R}^2_+$ 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 $\forall v_1, v_2 \in \mathbb{R}$, $\exists v \in [v_1, v_2]$, $\ell'_1(v) \neq 0$ or $\ell'_{-1}(v) \neq 0$. If there exists $\bar{v}_1, \bar{v}_2 \in \mathbb{R}$ 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.*

Proposition 15 *Suppose $\ell\colon \mathcal{V} \to \mathbb{R}^2_+$ is a differentiable binary loss such that $\forall v \in \mathcal{V}$, $\ell'_{-1}(v) \neq 0$ 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) $f\colon \mathcal{V} \ni v \mapsto \frac{\ell'_1(v)}{\ell'_{-1}(v)}$ is strictly increasing (decreasing) and continuous.*

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

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

Corollary 16 *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 $\varphi(x) = 1 - \frac{1}{\pi}\arctan(x-1)$. Then $f(v) = \frac{\varphi'(-v)}{\varphi'(-v)+\varphi'(v)} = \frac{x^2-2x+2}{2x^2+4}$ which is not invertible.

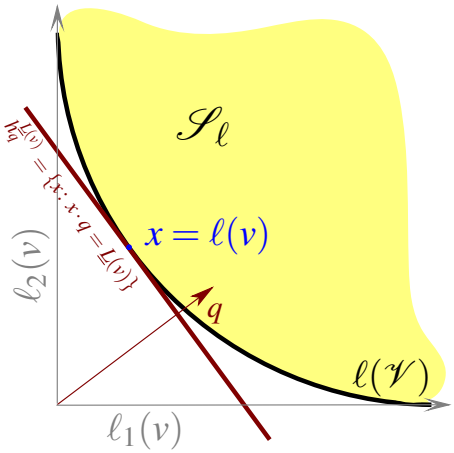
We now generalize the above results to the multiclass case.

Proposition 17 *Suppose ℓ has two proper composite representations $\ell = \lambda \circ \psi^{-1} = \mu \circ \phi^{-1}$ where λ and μ are proper losses and ψ and ϕ are continuous invertible. Then $\lambda = m$ 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.

Given a loss $\ell\colon \mathcal{V} \to \mathbb{R}^n_+$, we denote by $\mathcal{S}_\ell = \ell(\mathcal{V}) + [0,\infty)^n = \{\lambda : \exists v \in \mathcal{V},\ \forall i \in [n],\ \lambda_i \geq \ell_i(v)\}$ the *super-prediction set* of ℓ (confer e.g. [18]). We introduce 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 $\mathscr{A}$ at $x \in \mathscr{A}$ when $x \in h_p^\beta$ and for all $a \in \mathscr{A}$, $a' \cdot p \geq \beta$ or for all $a \in \mathscr{A}$, $a' \cdot p \leq \beta$. We say that $\mathcal{S}_\ell$ is *strictly convex in its inner part* when for all $p \in \Delta^n$, there exists an unique $x \in \ell(\mathcal{V})$ such that there exists a hyperplane h_p^β supporting $\mathcal{S}_\ell$ at x. $\mathcal{S}_\ell$ is said to be *smooth* when for all $x \in \ell(\mathcal{V})$, there exists an unique hyperplane supporting $\mathcal{S}_\ell$ at x. If ℓ is invertible, we can express these two definitions in terms of $v \in \mathcal{V}$ rather than $x \in \ell(\mathcal{V})$. If $\ell\colon \mathcal{V} \to \mathbb{R}^n_+$ is strictly convex, then $\mathcal{S}_\ell$ will be strictly convex in its inner part.



Proposition 18 *Suppose $\ell\colon \mathcal{V} \to \mathbb{R}^n_+$ is a continuous invertible loss. Then ℓ has a strictly proper composite representation if and only if $\mathcal{S}_\ell$ is convex, smooth and strictly convex in its inner part.*

Proposition 19 *Suppose $\ell\colon \mathcal{V} \to \mathbb{R}^n_+$ is a continuous loss. If ℓ has a proper composite representation, then $\mathcal{S}_\ell$ is convex and smooth. If ℓ is also invertible, then $\mathcal{S}_\ell$ is strictly convex in its inner part.*

6 Designing Proper Losses

We now build a family of conditional Bayes risks. Suppose we are given $\frac{n(n-1)}{2}$ concave functions $\{\underline{L}^{i_1,i_2}\colon \Delta^2 \to \mathbb{R}\}_{1 \leq i_1 < i_2 \leq n}$ on Δ^2, and we want to build a concave function $\underline{L}$ on Δ^n which is equal to one of the given functions on each edge of the simplex ($\forall 1 \leq i_1 < i_2 \leq n$, $\underline{L}(0,.,0,p_{i_1},0,.,0,p_{i_2},0,.,0) = \underline{L}^{i_1,i_2}(p_{i_1},p_{i_2})$). 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 infinity of solutions — one can simply add any concave function equal to zero in each edge).

Lemma 20 *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*

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

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

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

Lemma 21 *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

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

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.

7 Integral Representations of Proper Losses

Unlike the natural generalisation of the results from proper binary to proper multiclass losses above, there is one result that does not carry over: the integral representation of proper losses [1]. In the binary case there exists a family of "extremal" loss functions (cost-weighted generalisations of the 0-1 loss) each 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 in [1, 3], given these extremal functions, any proper binary loss ℓ can be expressed as the weighted integral $\ell = \int_0^1 \ell^c\, w(c)\, dc + \textit{constant}$ with $w(c) = -\underline{L}''(c)$. This representation is a special case of a representation from Choquet theory [19] 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 $n > 2$. The rest of this section makes this statement precise.

A *convex cone* $\mathscr{K}$ is a set of points closed under linear combinations of positive coefficients. 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. Theorem 2.2 of [20] shows 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 22 *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 proof of Proposition 22 requires the following lemma which relies upon the correspondence between a proper loss and its Bayes risk (Proposition 11) and the fact that two continuous functions equal almost everywhere are equal everywhere.

Lemma 23 *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 $\mathscr{L}_n$.*

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 24 *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.*

Bronshtein [20] and Johansen [21] showed how to construct a set of extremal convex functions which is dense in $\mathscr{K}(U)$. With a trivial change of sign this leads to a family of extremal proper fair Bayes risks that is dense in the set of Bayes risks in the topology of uniform convergence. This means that it is not possible to have a small set of extremal ("primitive") losses from which one can construct any proper fair loss by linear combinations when $n > 2$.

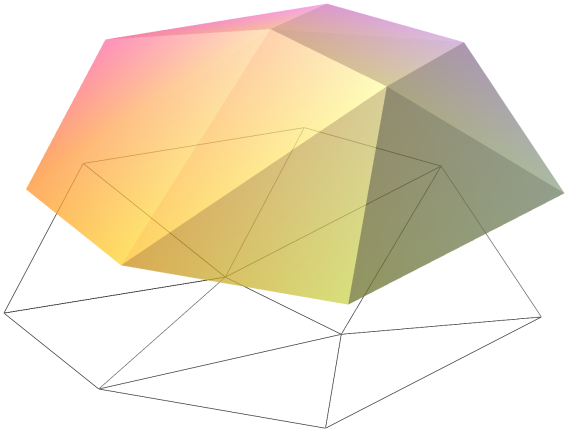


Figure 1: Complexity of extremal concave functions in two dimensions (corresponds to $n = 3$). Graph of an extremal concave function in two dimensions. Lines are where the slope changes. The pattern of these lines can be arbitrarily complex.

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. Theorem 2.1 of [20] 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 $\underline{L}_c(p) = \sum_{i=1}^{n} \frac{p_i}{c_i} \prod_{j \neq i} [\![\frac{p_i}{c_i} \leq \frac{p_j}{c_j}]\!]$ or $\underline{L}_c(p) = \bigwedge_{i \in [n]} \frac{1-p_i}{1-c_i}$.

8 Conclusion

We considered loss functions for multiclass prediction problems and made four main contributions:

- We extended existing results for binary losses to multiclass prediction problems including several characterisations of proper losses and the relationship between properness and classification calibration;
- We related the notion of prediction calibration to classification calibration;
- We developed some new existence and uniqueness results for proper composite losses (which are new even in the binary case) which characterise when a loss has a proper composite representation in terms of the geometry of the associated superprediction set; and
- We showed that the attractive (simply parametrised) integral representation for binary proper losses can *not* be extended to the multiclass case.

Our 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 in [1]. The proper composite representation is used in [22].

Acknowledgements

The work was performed whilst Elodie Vernet was visiting ANU and NICTA, and was supported by the Australian Research Council and NICTA, through backing Australia's ability.

References

[1] 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. http://www-stat.wharton.upenn.edu/~buja/PAPERS/paper-proper-scoring.pdf.

[2] 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.

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

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

[5] 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.

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

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

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

[9] Yufeng Liu. Fisher consistency of multicategory support vector machines. *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, side 289-296, 2007.

[10] 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. http://dx.doi.org/10.1007/s10994-009-5132-8.

[11] 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.

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

[13] Tobias Glasmachers. Universal consistency of multi-class support vector classication. *Advances in Neural Information Processing Systems (NIPS)*, 2010.

[14] Elodie Vernet, Robert C. Williamson and Mark D. Reid. Composite multiclass losses. (with proofs). To appear in NIPS 2011, October 2011. http://users.cecs.anu.edu.au/~williams/papers/P188.pdf.

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

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

[17] 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.

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

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

[20] Efim Mikhailovich Bronshtein. Extremal convex functions. *Siberian Mathematical Journal*, 19:6-12, 1978.

[21] Søren Johansen. The extremal convex functions. *Mathematica Scandinavica*, 34:61-68, 1974.

[22] Tim van Erven, Mark D. Reid and Robert C. Williamson. Mixability is Bayes risk curvature relative to log loss. *Proceedings of the 24th Annual Conference on Learning Theory*, 2011. To appear. http://users.cecs.anu.edu.au/~williams/papers/P186.pdf.

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

9 Proofs for *Composite Multiclass Losses*

Here we present proofs that were omitted from the main body of the paper due to lack of space.

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
$$p \notin \mathscr{T}_{1=j_1}(c) \Rightarrow \exists j_2 \neq j_1 \text{ such that } \frac{p_{j_1}}{c_{j_1}} < \frac{p_{j_2}}{c_{j_2}}$$
$$p \notin \mathscr{T}_{j_2}(c) \Rightarrow \exists j_3 \neq j_2 \text{ such that } \frac{p_{j_2}}{c_{j_2}} < \frac{p_{j_3}}{c_{j_3}}$$
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.

∎

Proof of Proposition 3

($\Rightarrow$) 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. ∎

Proof of Proposition 5

We show first that λ is proper. Let $p \in \Delta^n$, $\Lambda(p,p) = L(p, \bar{\psi}(p)) = L(p, \arg\min_v L(p,v)) = \min_v L(p,v) \leq \min_{q \in \Delta^n} \Lambda(p,q)$. Thus λ is proper and $\underline{L}(p) = \underline{\Lambda}(p)$.

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

Proof of Proposition 6

$(\ell(p) - \ell(q))' \cdot (p - q) = p' \cdot \ell(p) - q' \cdot \ell(p) + q' \cdot \ell(q) - p' \cdot \ell(q) \leq 0$ since $p \cdot \ell(p) \leq p \cdot \ell(q)$. ∎

Proof of Proposition 7

We just have to check that ℓ is injective. If ℓ is not invertible, there exists $p \neq q$ such that $\ell(p) = \ell(q)$. Then, $L(p,p) = L(p,q)$ which contradicts the supposed strict properness of ℓ. ∎

Proof of Proposition 8

($\Rightarrow$) 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 $\tilde{L}^{p,q}(\eta, \hat{\eta}) = \big(\eta q + (1-\eta)p\big)' \cdot \ell\big(p + \hat{\eta}(q-p)\big) = L\big(p + \eta(q-p), p + \hat{\eta}(q-p)\big) \geq L\big(p + \eta(q-p), p + \eta(q-p)\big) = \tilde{L}^{p,q}(\eta, \eta)$. Hence $\tilde{\ell}^{p,q}$ is proper.

($\Leftarrow$) 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 $\tilde{q}$). Then $L(p,q) = \tilde{L}^{\tilde{p},\tilde{q}}(\eta,\hat{\eta}) \geq \tilde{L}^{\tilde{p},\tilde{q}}(\eta,\eta) = L(p,p)$. Hence ℓ is proper. ∎

Proof of Proposition 9

One can easily prove that the second part of the equivalence implies the first one with $h_1 = 0$.

Thanks to proposition 1 of [16], we know that 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 $\Gamma : p$ distribution on $\{0,1\} \to \mathbb{E}(p)$) . We also know that if ℓ is proper then $\forall p,q \in \partial\Delta^n$, $\tilde{\ell}^{p,q}$ (introduced in Proposition 8) 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 $\eta_2 = \eta + h_2(\mu - \eta)$. And 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. ∎

Proof of Corollary 10

If ℓ is proper then $\forall p \in \mathring{\Delta}^n$, $q \mapsto L(p,q) = p' \cdot (\tilde{\ell} \circ \Pi_\Delta)(q)$ reaches its minimum at p and thus $p' \cdot M(p) = 0$. Define $f_{p,q} \colon \eta \mapsto L(p, p + \eta(q-p)) = p' \cdot (\tilde{\ell} \circ \Pi_\Delta)(p + \eta(q-p))$. By Proposition 9 f is decreasing for $\eta < 0$ and increasing for $\eta > 0$. However, $f'_{p,q}(\eta) = p' \cdot M(p + \eta(q-p)) \cdot (q-p)$ is negative if $\eta < 0$ and positive if $\eta > 0$. Let $r = p + \eta(q-p)$. Then $f'_{p,q}(\eta) = (r - \eta(q-p))' \cdot M(r)(q-p) = -\eta(q-p)' \cdot M(r) \cdot (q-p)$, since $r' \cdot M(r) = 0$, Which proves the second part of the first implication. To prove the other implication, it suffices to show the order sensitivity property using $f_{p,q}$ and appeal to Proposition 9. ∎

Proof of Proposition 11

($\Rightarrow$) 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 $p' \cdot \ell(q) = f(q) + (p-q)' \cdot A(q)$.

($\Leftarrow$) 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, $L(p,p) - L(p,q) = f(p) - f(q) + (p-q)' \cdot A(q) \geq 0$. Then ℓ is proper. ∎

Proof of Proposition 12

A concave function is differentiable almost everywhere [15, theorem 4.2.3]. Thus (3) proves that two proper losses 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 f and g are equal almost everywhere and continuous (since they are concave). If there exists x such that $f(x) \neq g(x)$, then since f and g are continuous, there exists $\varepsilon > 0$ such that $\forall y \in \mathscr{B}(x,\varepsilon) \cap \Delta^n$, $f(y) \neq g(y)$. Yet this contradicts the fact that f and g are equal almost everywhere. Hence the Bayes risks are equal everywhere. ∎

Proof of Proposition 13

Observe that

$$\partial\underline{L}(p) = \{(s',0)' + \alpha\mathbb{1},\ s \in \partial\tilde{\underline{L}}(\tilde{p}),\ \alpha \in \mathbb{R}\}. \tag{4}$$

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 from [15, 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 p, $\exists \tilde{\delta} > 0$, such that

$$\forall \tilde{q} \in \mathscr{B}(\tilde{p},\tilde{\delta}),\ \forall A(\tilde{q}) \in \partial\tilde{\underline{L}}(\tilde{q}),\ ||A(\tilde{q}) - \mathsf{D}\tilde{\underline{L}}(\tilde{p})|| \leq \varepsilon. \tag{5}$$

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 (5), $\forall i \in [n]$, $\forall q \in \mathscr{B}(p,\delta)$, for $\alpha_1, \alpha_2 \in \mathbb{R}$,

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

where $A(\tilde{q}) \in \partial\tilde{\underline{L}}(\tilde{q})$, and $\gamma = -(e_i - q)' \cdot \alpha_1 \mathbb{1} + (e_i - p)' \cdot \alpha_2 \mathbb{1} = -\alpha_1 + \alpha_1 q' \cdot \mathbb{1} + \alpha_2 - \alpha_2 p' \cdot \mathbb{1} = -\alpha_1 + \alpha_1 + \alpha_2 - \alpha_2 = 0$,

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

By continuity of $\underline{L}$, $||\underline{L}(q) - \underline{L}(p)|| < \varepsilon$ for small enough δ. Furthermore by (5), $||A(\tilde{q}) - \mathsf{D}\tilde{\underline{L}}(\tilde{p})|| \leq 0$ 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 $\forall i \in [n]$,

$$(e_i - p)' \cdot ((A(\tilde{p})', 0)' + \alpha_1 \mathbb{1}) \leq (e_i - p)' \cdot ((B(\tilde{p})', 0)' + \alpha_2 \mathbb{1}), \ \ \alpha_1, \alpha_2 \in \mathbb{R}. \tag{6}$$

Thus $\forall q \in \Delta^n$,

$$\begin{aligned}&\sum_{i \in [n]} q_i (e_i - p)' \cdot ((A(\tilde{p})', 0)' + \alpha_1 \mathbb{1}) \ \leq \ \sum_{i \in [n]} q_i (e_i - p)' \cdot ((B(\tilde{p})', 0)' + \alpha_2 \mathbb{1}), \ \forall q \in \Delta^n, \ \alpha_1, \alpha_2 \in \mathbb{R} \\ \Leftrightarrow\ &(q - p)' \cdot ((A(\tilde{p})', 0)' + \alpha_1 \mathbb{1}) \ \leq \ (q - p)' \cdot ((B(\tilde{p})', 0)' + \alpha_2 \mathbb{1}), \quad \forall q \in \Delta^n, \alpha_1, \alpha_2 \in \mathbb{R} \\ \Leftrightarrow\ &(\tilde{q} - \tilde{p})' \cdot A(\tilde{p}) \ \leq \ (\tilde{q} - \tilde{p})' \cdot B(\tilde{p}), \ \ \forall \tilde{q} \in \tilde{\Delta}^n. \end{aligned} \tag{7}$$

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 (7) 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 (6) 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

$$\exists i \in [n], \ (\tilde{e}_i - \tilde{p})' \cdot A(\tilde{p}) > (\tilde{e}_i - \tilde{p})' \cdot B(\tilde{p}). \tag{8}$$

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 $\eta > 0$,

$$C(\tilde{p}_\eta)' \cdot (\tilde{e}_i - \tilde{p}) \ \leq \ B(\tilde{p})' \cdot (\tilde{p} - \tilde{e}_i). \tag{9}$$

Now $\ell_i(p_\eta) = \underline{L}(p_\eta) + (\tilde{e}_i - \tilde{p}_\eta)' \cdot C(\tilde{p}_\eta)$. Hence (9) 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 $\underline{L}$,

$$\begin{aligned}\lim_{\eta \searrow 0} \underline{L}(p_\eta) + (\tilde{e}_i - \tilde{p})' \cdot B(\tilde{p}) \ &= \ \underline{L}(p) + (\tilde{e}_i - \tilde{p})' \cdot B(\tilde{p}) \\ &< \ \underline{L}(p) + (\tilde{e}_i - \tilde{p})' \cdot A(\tilde{p}) \\ &= \ \ell_i(p) \text{ by (8).}\end{aligned}$$

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

Proof of Proposition 14

The proposition is a direct consequence of the characterization of differential binary proper loss. A differential 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

$$\frac{-\lambda_1'(\eta)}{1-\eta} = \frac{\lambda_{-1}'(\eta)}{\eta}, \ \forall \eta \in (0,1) \tag{10}$$

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

$$\Leftrightarrow \psi'(\psi^{-1}(v)) = 0 \text{ or } \ell_{-1}'(v) = \ell_1'(v) = 0 \text{ or } \psi^{-1}(v) = \frac{\ell_{-1}'(v)}{\ell_{-1}'(v) - \ell_1'(v)}, \ \forall v \in \mathscr{V}. \tag{11}$$

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 $I = \mathbb{R}$ 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 $\psi|_{[v_1,v_2]}$ differentiable invertible 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. ∎

Proof of Proposition 15

The loss λ is proper if and only if (10) and $-\lambda_1'(\eta) \geq 0$ and $\lambda_{-1}'(\eta) \geq 0$. This is equivalent to there exists an invertible ψ such that (11) holds and

$$-\psi'(\psi^{-1}(v))\ell_1'(v) \geq 0 \text{ and } \psi'(\psi^{-1}(v))\ell_{-1}'(v) \geq 0, \ \forall v \in \mathscr{V}. \tag{12}$$

($\Rightarrow$) 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 (11) 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 (12) holds. The other case is analogous. ∎

Proof of Proposition 17

$\underline{\Lambda}(p) = \inf_q p' \cdot \lambda(q) = \inf_q p' \ell(\psi(q)) = \inf_v L(p,v)$ (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 $\lambda = \ell \circ \psi = \ell \circ \phi = \mu$. ∎

Proof of Proposition 18

($\Leftarrow$) Let $p \in \Delta^n$. By strict convexity of the inner part of $\mathscr{S}_\ell$, there exists an unique $v \in \mathscr{V}$ such that there exists a hyperplane $h_p^{\beta^*}$ supporting $\mathscr{S}_\ell$ at $\ell(v)$. Define ψ such that for all $p \in \Delta^n$, $\psi(p)$ is thus unique previous v. Since $h_p^{\beta^*}$ supports $\mathscr{S}_\ell$, $\beta^* = \inf\{\beta : h_p^\beta \cap \mathscr{S}_\ell \neq \varnothing\} = p' \cdot \ell(v) = p' \cdot \ell(\psi(p))$.

By smoothness of $\mathscr{S}_\ell$, ψ is invertible. Indeed one can build the inverse which for all v, associated the normalized normal vector to the hyperplane supporting $\mathscr{S}_\ell$ at $\ell(v)$.

By continuity of ℓ and strict convexity of $\mathscr{S}_\ell$ in its inner part, ψ is continuous. Let $\lambda = \ell \circ \psi$. Then $p' \cdot \lambda(p) = p' \cdot (\ell \circ \psi)(p) = \inf_{v \in \mathscr{V}} p' \cdot \ell(v)$ and by invertibility of ψ, $p' \cdot \lambda(p) = \inf_{q \in \Delta^n} p' \cdot \lambda(q)$. Thus λ is proper and since there exists an unique point where $h_p^{\underline{\Lambda}(p)}$ supports $\mathscr{S}_\ell$ (due to strict convexity of the inner part), then λ is strictly proper and thus ℓ has a strictly proper composite representation.

($\Rightarrow$) We make use of the following result [23, Theorem 1.3.3]. Suppose A is closed set such that $\mathring{A} \neq \varnothing$ and such that through each boundary point of A there is a support plane to A. Then A is convex.

Suppose that ℓ has a strictly composite representation $\ell = \lambda \circ \psi^{-1}$. Observe that $\ell(\mathscr{V}) = \lambda(\Delta^n)$. By invertibility and continuity of ℓ and ψ, λ is also invertible and continuous. Thus to each points x of the image of ℓ there corresponds an unique v and p such that $x = \ell(v) = \lambda(p)$. Hence it is equivalent to prove the properties on $\mathscr{S}_\lambda$. We now prove two auxiliary claims that hold for all $p \in \Delta^n$:

1. $\forall \beta < \underline{\Lambda}(p)$, $h_p^\beta \cap \mathscr{S}_\lambda = \varnothing$.

   Indeed if there exists $z \in \mathscr{S}_\lambda \cap h_p^\beta$, there exists $q \in \Delta^n$ such that for all $i \in [n]$, $\lambda_i(q) \leq z_i$. And $\underline{\Lambda}(p) > \beta = p' \cdot z = \sum_{i \in [n]} p_i z_i \geq \sum p_i \lambda_i(q) = p'.\lambda(q)$ — a contradiction.

2. $h_p^{\underline{\Lambda}(p)} \cap \mathscr{S}_\lambda = \{\lambda(p) + \sum \alpha_i e_i [\![p_i = 0]\!],\ \alpha_i \geq 0\}$.

   ($\supseteq$) $p' \cdot \lambda(p) = \underline{\Lambda}(p)$, and so $\lambda(p) \in h_p^{\underline{\Lambda}(p)} \cap \mathscr{S}_\lambda$. Consequently,

   $$p' \cdot \Big(\sum_{i \in [n]} \alpha_i e_i [\![p_i = 0]\!] + \lambda(p)\Big) = \sum_{i \in [n]} \alpha_i p_i [\![p_i = 0]\!] + \underline{\Lambda}(p) = \underline{\Lambda}(p).$$

   Thus $\sum \alpha_i e_i [\![p_i = 0]\!] + \lambda(p) \in h_p^{\underline{\Lambda}(p)} \cap \mathscr{S}_\lambda$.

   ($\subseteq$) If $z \in h_p^{\underline{\Lambda}(p)} \cap \mathscr{S}_\lambda$, there exists $q \in \Delta^n$, $\alpha_i \geq 0$ such that $z = \lambda(q) + \sum \alpha_i e_i$. By strict properness of λ, $p = q$, indeed $\underline{\Lambda}(p) = p' \cdot z \geq p' \cdot \lambda(q)$. Thus $\alpha_i = 0$ if $p_i \neq 0$ because otherwise, $p' \cdot z > p' \cdot \lambda(p) = \underline{\Lambda}(p)$ which would be a contradiction.

Hence there is one hyperplane supporting $\mathscr{S}_\lambda$ at each of the points $\lambda(p) + \sum \alpha_i e_i [\![p_i = 0]\!]$, $p \in \Delta^n$, $\alpha_i \geq 0$. These points belong to the boundary of $\mathscr{S}_\lambda$. Since λ is continuous and by definition of $\mathscr{S}_\lambda$, the last points are the only points of th boundary of $\mathscr{S}_\lambda$. So for each point of the boundary, there exists a suporting hyperplane, then $\mathscr{S}_\lambda$ is convex. The second point give the strict convexity of $\mathscr{S}_\lambda$ in its inner part.

Since λ is continuous, the associated Bayes risk is differentiable. Indeed $\mathscr{S}_\lambda$ is smooth because the differentiability of the support function is equivalent to the fact that in each points of the boundary there exists an unique hyperplane supporting the set. ∎

Proof of Proposition 19

This proof is very similar to the preceeding proof. Assume that ℓ is continuous and has a proper composite representation. Then the proper loss associated with $\lambda = \ell \circ \psi$ is also continuous and $\ell(\mathscr{V}) = \lambda(\Delta^n)$ so the convexity and smoothness of $\mathscr{S}_\ell$ is equivalent to the convexity and smoothness of $\mathscr{S}_\lambda$. Then the first point of the last proof still holds and so $h_p^{\underline{\Lambda}(p)} \cap \mathscr{S}_\lambda \supseteq \{\lambda(p) + \sum \alpha_i e_i [\![p_i = 0]\!],\ \alpha_i \geq 0\}$. These last points belong to the boundary of $\mathscr{S}_\lambda$. By continuity of λ, the points of the boundary of $\mathscr{S}_\lambda$ are still $\lambda(p) + \sum \alpha_i e_i [\![p_i = 0]\!],\ \alpha_i \geq 0$. Thus by the result quoted at the beginning of the second part of the proof of the Proposition 18, $\mathscr{S}_\lambda$ is convex. The continuity of λ still implies the differentiability of the Bayes risk so the smoothness of $\mathscr{S}_\lambda$. Thus for each point $\lambda(p)$, there exists an unique hyperplane supporting $\mathscr{S}_\lambda$.

If moreover ℓ is invertible, λ is also invertible. Assume there exists two points v and w such that a hyperplane h_r^β supports $\mathscr{S}_\ell$ at $\ell(v)$ and $\ell(w)$. Then h_r^β supports $\mathscr{S}_\lambda$ at $\lambda(\psi^{-1}(v))$ and $\lambda(\psi^{-1}(w))$. Yet $h_{\psi^{-1}(v)}^{\underline{\Lambda}(\psi^{-1}(v))}$ and $h_{\psi^{-1}(w)}^{\underline{\Lambda}(\psi^{-1}(w))}$ are two hyperplanes supporting $\mathscr{S}_\lambda$ at these points. So $v = w$ and thus

$\mathscr{S}_\ell$ is strictly convex in its inner part. ∎

Proof of Lemma 20

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

Proof of Lemma 21

Use the correspondence between Bayes risk and proper losses and the preceeding lemma. ∎

Proof of Lemma 23

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 $\ell=(g+f)$ 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 $\underline{L}(p)=p'\cdot\ell(p)=p'\frac{1}{2}(f(p)+g(p))=\frac{1}{2}(\underline{F}+\underline{G})$ 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 $\mathscr{L}_n$. ∎

Proof of Lemma 24

We require two facts from convex analysis (cf. Theorems B.3.1.4 and D.6.2.7 of [15]). 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 of Proposition 22

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), Theorem 2.2 of [20] 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 24 there is a corresponding sequence $(\ell^i)_i$ of fair proper losses that converges almost everywhere to ℓ. Lemma 23 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. ∎