Convexity of Proper Composite Binary Losses

Mark D. Reid
The Australian National University
and NICTA

Robert C. Williamson
The Australian National University
and NICTA

Abstract

A composite loss assigns a penalty to a real-valued prediction by associating the prediction with a probability via a link function then applying a class probability estimation (CPE) loss. If the risk for a composite loss is always minimised by predicting the value associated with the true class probability the composite loss is proper. We provide a novel, explicit and complete characterisation of the convexity of any proper composite loss in terms of its link and its "weight function" associated with its proper CPE loss.

1 INTRODUCTION

The study of convex loss functions is central to the practicality of many machine learning techniques such as boosting and stochastic gradient descent. In this paper we provide a characterisation of the convexity of a large class of losses for probability estimation: the proper composite losses. A *loss* function is the means by which a learning algorithm's performance is judged. A *binary* loss function is a loss for a supervised prediction problem where there are two possible labels associated with the examples. A *composite* loss is the composition of a loss for class probability estimation (defined below) and a link function (also defined below). A *convex* loss is one which is convex in the prediction made by the learning algorithm; they are much more amenable to numerical optimisation. We characterise the convexity of *proper* composite binary losses – a natural class of losses for probability estimation and thus good surrogates for classification. We expect this characterisation will help in the choice of practical surrogate losses for binary classification problems.

Appearing in Proceedings of the 13^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9.

Informally, proper losses are well-calibrated losses for class probability estimation, that is for the problem of not only predicting a binary classification label, but providing an estimate of the probability that an example will have a positive label. Link functions are used to map the outputs of a real-valued predictor to the interval $[0, 1]$ so they can be interpreted as probabilities. Having such probabilities is often important in applications, and there has been considerable interest in understanding how to get accurate probability estimates (Gneiting and Raftery, 2007) and understanding the implications of requiring loss functions that do so (Bartlett and Tewari, 2007).

Much previous work in the machine learning literature has focussed on *margin losses* which intrinsically treat classes symmetrically. However it is now well understood that it is important to deal with the non-symmetric case (Bach et al., 2006; Buja et al., 2005). A key goal of the present work is to consider composite losses in the general (non-symmetric) situation.

Having the flexibility to choose a loss function is important to "tailor" the solution to a machine learning problem; cf. (Hand, 1994; Hand and Vinciotti, 2003; Buja et al., 2005). Understanding the structure of the set of loss functions and having natural parametrisations of them is useful for this purpose. Even when using a loss as a surrogate for the loss one would ideally like to minimise, it is helpful to have an easy to use parametrisation — see the discussion of "surrogate tuning" in the Conclusion.

Our main result is Theorem 11 which characterises when a composite loss is convex (what is called a "nice-pair" in (Cesa-Bianchi and Lugosi, 2006, p.302)). This characterisation is in terms of what seems to be the most natural and intrinsic parametrisation of composite losses.

2 LOSSES AND RISKS

We write $[\![p]\!] = 1$ if p is true and $[\![p]\!] = 0$ otherwise. The generalised function $\delta(\cdot)$ is defined by

$\int_a^b \delta(x)f(x)dx = f(0)$ when f is continuous at 0 and $a < 0 < b$. Random variables are written in sans-serif font: X, Y.

Given a set of labels $\mathcal{Y} := \{-1, 1\}$ and a set of prediction values $\mathcal{V}$ we will say a *loss* is any function[1] $\ell : \mathcal{Y} \times \mathcal{V} \to [0, \infty)$. We interpret such a loss as giving a penalty $\ell(y, v)$ when predicting the value v with an observed label y. We can always write an arbitrary loss in terms of its *partial losses* $\ell_1 := \ell(1, \cdot)$ and $\ell_{-1} := \ell(-1, \cdot)$ using

$$\ell(y, v) = [\![y = 1]\!]\ell_1(v) + [\![y = -1]\!]\ell_{-1}(v). \quad (1)$$

Our definition of a loss function covers all commonly used *margin losses* (*i.e.* those which can be expressed as $\ell(y, v) = \phi(yv)$ for some function $\phi : \mathbb{R} \to [0, \infty)$) such as the *0-1 loss* $\ell(y, v) = [\![yv > 0]\!]$, the *hinge loss* $\ell(y, v) = \max(1 - yv, 0)$, the *logistic loss* $\ell(y, v) = \log(1 + e^{yv})$, and the *exponential loss* $\ell(y, v) = e^{-yv}$ commonly used in boosting. It also covers *class probability estimation losses* where the predicted values $\hat{\eta} \in \mathcal{V} = [0, 1]$ are directly interpreted as probability estimates.[2] We will use $\hat{\eta}$ instead of v as an argument to indicate losses for class probability estimation and use the shorthand *CPE losses* to distinguish them from general losses. For example, *square loss* has partial losses $\ell_{-1}(\hat{\eta}) = \hat{\eta}^2$ and $\ell_1(\hat{\eta}) = (1 - \hat{\eta})^2$, the *log loss* $\ell_{-1}(\hat{\eta}) = \log(1 - \hat{\eta})$ and $\ell_1(\hat{\eta}) = \log(\hat{\eta})$, and the family of *cost-weighted misclassification losses* parametrised by $c \in (0, 1)$ with

$$\ell_c(-1, \hat{\eta}) = c[\![\hat{\eta} \geq c]\!] \text{ and } \ell_c(1, \hat{\eta}) = (1-c)[\![\hat{\eta} < c]\!]. \quad (2)$$

2.1 Conditional and Full Risks

Suppose we have random examples X with associated labels $\mathsf{Y} \in \{-1, 1\}$. The joint distribution of (X, Y) is denoted $\mathbb{P}$ and the marginal distribution of X is denoted M. Let the observation conditional density $\eta(x) := \Pr(\mathsf{Y} = 1 | \mathsf{X} = x)$. Thus one can specify an experiment by either $\mathbb{P}$ or (η, M).

If $\eta \in [0, 1]$ is the probability of observing the label $y = 1$ the *point-wise risk* (or *conditional risk*) of the estimate $v \in \mathcal{V}$ is defined as the η-average of the point-wise risk for v: $L(\eta, v) := \mathbb{E}_{\mathsf{Y} \sim \eta}[\ell(\mathsf{Y}, v)] = \eta\ell_1(v) + (1 - \eta)\ell_{-1}(v)$. Here, $\mathsf{Y} \sim \eta$ is a shorthand for labels being drawn from a Bernoulli distribution with parameter η. When $\eta : \mathcal{X} \to [0, 1]$ is an observation-conditional density, taking the M-average of the point-wise risk gives the *(full) risk* of the estimator v, now interpreted as a function $v : \mathcal{X} \to \mathcal{V}$:

$$\mathbb{L}(\eta, v, M) := \mathbb{E}_{\mathsf{X} \sim M}[L(\eta(\mathsf{X}), v(\mathsf{X}))].$$

The convention of using ℓ, L and $\mathbb{L}$ for the loss, point-wise and full risk is used throughout this paper. The *Bayes risk* is the minimal achievable value of the risk and is denoted

$$\underline{\mathbb{L}}(\eta, M) := \inf_{v \in \mathcal{V}^{\mathcal{X}}} \mathbb{L}(\eta, v, M) = \mathbb{E}_{\mathsf{X} \sim M}\left[\underline{L}(\eta(\mathsf{X}))\right],$$

where $[0, 1] \ni \eta \mapsto \underline{L}(\eta) := \inf_{v \in \mathcal{V}} L(\eta, v)$ is the *point-wise* or *conditional Bayes risk*.

3 LOSSES FOR CLASS PROBABILITY ESTIMATION

We begin by considering CPE losses, that is, functions $\ell : \{-1, 1\} \times [0, 1] \to [0, \infty)$ and briefly summarise a number of important existing structural results for *proper losses*.

3.1 Proper and Fair Losses

If $\hat{\eta}$ is to be interpreted as an estimate of the true positive class probability η (*i.e.*, when $y = 1$) then it is desirable to require that $L(\eta, \hat{\eta})$ be minimised by $\hat{\eta} = \eta$ for all $\eta \in [0, 1]$. Losses that satisfy this constraint are said to be *Fisher consistent* and are known as *proper losses* (Buja et al., 2005; Gneiting and Raftery, 2007). That is, a proper loss ℓ satisfies $\underline{L}(\eta) = L(\eta, \eta)$ for all $\eta \in [0, 1]$. A *strictly proper* loss is a proper loss for which the minimiser of $L(\eta, \hat{\eta})$ over $\hat{\eta}$ is unique.

We will say a loss is *fair* whenever $\ell_{-1}(0) = \ell_1(1) = 0$. That is, there is no loss incurred for perfect prediction. Fairness is relied upon in the integral representation of Theorem 4 where it is used to omit some constants of integration.

3.2 The Structure of Proper Losses

A key result in the study of proper losses is originally due to Shuford et al. (1966) though our presentation follows that of Buja et al. (2005). It characterises proper losses for probability estimation via a constraint on the relationship between its partial losses.

Theorem 1 *Suppose $\ell : \{-1, 1\} \times [0, 1] \to \mathbb{R}$ is a CPE loss and that its partial losses ℓ_1 and ℓ_{-1} are both differentiable. Then ℓ is a proper loss if and only if for all $\hat{\eta} \in (0, 1)$*

$$\frac{-\ell_1'(\hat{\eta})}{1 - \hat{\eta}} = \frac{\ell_{-1}'(\hat{\eta})}{\hat{\eta}} = w(\hat{\eta}) \quad (3)$$

[1] Restricting the output of a loss to $[0, \infty)$ is equivalent to assuming the loss has a lower bound and then translating its output.

[2] These are also known as *scoring rules* (Gneiting and Raftery, 2007).

for some weight function $w : (0,1) \to \mathbb{R}^+$ *such that* $\int_\epsilon^{1-\epsilon} w(c)\,dc < \infty$ *for all* $\epsilon > 0$.

This simple characterisation of the structure of proper losses has a number of interesting implications. Observe from (3) that if ℓ is proper, given ℓ_1 we can determine ℓ_{-1} or vice versa. Also, the partial derivative of the conditional risk can be seen to be the product of a linear term and the weight function:

Corollary 2 *If ℓ is a differentiable proper loss then for all $\eta \in [0,1]$*

$$\frac{\partial}{\partial \hat{\eta}} L(\eta, \hat{\eta}) = (1-\eta)\ell'_{-1}(\hat{\eta}) + \eta \ell'_1(\hat{\eta}) = (\hat{\eta} - \eta) w(\hat{\eta}). \quad (4)$$

Another corollary, observed by Buja et al. (2005), is that the weight function is related to the curvature of the conditional Bayes risk $\underline{L}$.

Corollary 3 *Let ℓ be a a twice differentiable[3] proper loss with weight function w defined as in equation (3). Then for all $c \in (0,1)$ its conditional Bayes risk $\underline{L}$ satisfies*

$$w(c) = -\underline{L}''(c). \quad (5)$$

The relationship between a proper loss and its associated weight function is captured succinctly via the following representation of proper losses as a weighted integral of the cost-weighted misclassification losses ℓ_c defined in (2); see (Reid and Williamson, 2009).

Theorem 4 *Let $\ell : \mathcal{Y} \times [0,1] \to \mathbb{R}$ be a fair, proper loss. Then for each $\hat{\eta} \in (0,1)$ and $y \in \mathcal{Y}$*

$$\ell(y, \hat{\eta}) = \int_0^1 \ell_c(y, \hat{\eta})\, w(c)\, dc, \quad (6)$$

where $w = -\underline{L}''$. Conversely, if ℓ is defined by (6) for some weight function $w : (0,1) \to [0,\infty)$ then it is proper.

Buja et al. (2005) show that ℓ is strictly proper if and only if $w(c) > 0$ in the sense that w has non-zero mass on every open subset of $(0,1)$. Some example losses and their associated weight functions are given in Table 1.

4 COMPOSITE LOSSES

General loss functions are often constructed with the aid of a *link function*. For a particular set of prediction values $\mathcal{V}$ this is any continuous mapping $\psi\colon [0,1] \to \mathcal{V}$. In this paper, our focus will be *composite losses* for binary class probability estimation. These are the composition of a CPE loss $\ell\colon \{-1,1\} \times [0,1] \to \mathbb{R}$ and the inverse of a *link function* ψ, an invertible mapping from the unit interval to some range of values. Unless stated otherwise we will assume $\psi\colon [0,1] \to \mathbb{R}$. We will denote a composite loss by

$$\ell^\psi(y, v) := \ell(y, \psi^{-1}(v)). \quad (7)$$

The classical motivation for link functions (McCullagh and Nelder, 1989) is that often in estimating η one uses a parametric representation of $\hat{\eta}\colon \mathcal{X} \to [0,1]$ which has a natural scale not matching $[0,1]$. Traditionally one writes $\hat{\eta} = \psi^{-1}(\hat{h})$ where ψ^{-1} is the "inverse link". The function $\hat{h}\colon \mathcal{X} \to \mathbb{R}$ is the *hypothesis*. Often $\hat{h} = \hat{h}_{\boldsymbol{\alpha}}$ is parametrised linearly in a parameter vector $\boldsymbol{\alpha}$. In such a situation it is computationally convenient if $\ell(\eta, \psi^{-1}(\hat{h}))$ is convex in $\hat{h}$ (which implies it is convex in $\boldsymbol{\alpha}$ when $\hat{h}_{\boldsymbol{\alpha}}$ is linear in $\boldsymbol{\alpha}$).

4.1 Proper Composite Losses

We will call a composite loss ℓ^ψ (7) a *proper composite loss* if ℓ in (7) is a proper loss for class probability estimation. As in the case for losses for probability estimation, the requirement that a composite loss be proper imposes some constraints on its partial losses. Many of the results for proper losses carry over to composite losses with some extra factors to account for the link function.

Theorem 5 *Let $\lambda = \ell^\psi$ be a composite loss with differentiable and strictly monotone link ψ and suppose the partial losses $\lambda_{-1}(v)$ and $\lambda_1(v)$ are both differentiable. Then λ is a proper composite loss if and only if there exists a weight function $w : (0,1) \to \mathbb{R}^+$ such that for all $\hat{\eta} \in (0,1)$*

$$\frac{-\lambda'_1(\psi(\hat{\eta}))}{1-\hat{\eta}} = \frac{\lambda'_{-1}(\psi(\hat{\eta}))}{\hat{\eta}} = \frac{w(\hat{\eta})}{\psi'(\hat{\eta})} =: \rho(\hat{\eta}). \quad (8)$$

Furthermore, $\rho(\hat{\eta}) \geq 0$ for all $\hat{\eta} \in (0,1)$.

Proof This is a direct consequence of Theorem 1 for proper losses for probability estimation and the chain rule applied to $\ell_y(\hat{\eta}) = \lambda_y(\psi(\hat{\eta}))$. Since ψ is assumed to be strictly monotonic we know $\psi' > 0$ and so, since $w \geq 0$ we have $\rho \geq 0$. ■

As we shall see, the ratio $\rho(\hat{\eta})$ is a key quantity in the analysis of proper composite losses. For example, Corollary 2 has natural analogue in terms of ρ that will be of use later. It is obtained by letting $\hat{\eta} = \psi^{-1}(v)$ and using the chain rule.

[3] The restriction to differentiable losses can be removed in most cases if generalised weight functions — that is, possibly infinite but defining a measure on $(0,1)$ — are permitted. For example, the weight function for the 0-1 loss is $w(c) = \delta(c - \frac{1}{2})$.

$w(c)$	$\ell_{-1}(\hat{\eta})$	$\ell_1(\hat{\eta})$	Loss
$2\delta\left(\frac{1}{2}-c\right)$	$[\![\hat{\eta} > \frac{1}{2}]\!]$	$[\![\hat{\eta} \leq \frac{1}{2}]\!]$	0-1
1	$\hat{\eta}^2/2$	$(1-\hat{\eta})^2/2$	Square
$\frac{1}{(1-c)c}$	$-\ln(1-\hat{\eta})$	$-\ln(\hat{\eta})$	Log
$\frac{1}{[(1-c)c]^{3/2}}$	$2\sqrt{\frac{\hat{\eta}}{1-\hat{\eta}}}$	$2\sqrt{\frac{1-\hat{\eta}}{\hat{\eta}}}$	Boosting

Table 1: Weight functions and associated partial losses.

Corollary 6 *Suppose ℓ^ψ is a proper composite loss with conditional risk denoted L^ψ. Then*

$$\frac{\partial}{\partial v} L^\psi(\eta, v) = (\psi^{-1}(v) - \eta)\rho(\psi^{-1}(v)). \tag{9}$$

Loosely speaking then, ρ is a "co-ordinate free" weight function for composite losses where the link function ψ is interpreted as a mapping from arbitrary $v \in \mathcal{V}$ to values which can be interpreted as probabilities.

Another immediate corollary of Theorem 5 shows how properness is characterised by a particular relationship between the choice of link function and the choice of partial composite losses.

Corollary 7 *Let $\lambda := \ell^\psi$ be a composite loss with differentiable partial losses λ_1 and λ_{-1}. Then ℓ^ψ is proper if and only if the link ψ satisfies*

$$\psi^{-1}(v) = \frac{\lambda'_{-1}(v)}{\lambda'_{-1}(v) - \lambda'_1(v)}, \qquad \forall v \in \mathcal{V}. \tag{10}$$

Proof Substituting $\hat{\eta} = \psi^{-1}(v)$ into (8) yields $-\psi^{-1}(v)\lambda'_1(v) = (1-\psi^{-1}(v))\lambda'_{-1}(v)$ and solving this for $\psi^{-1}(v)$ gives the result. ■

These results give some insight into the "degrees of freedom" available when specifying proper composite losses. Theorem 5 shows that the partial losses are completely determined once the weight function w and ψ (up to an additive constant) is fixed. Corollary 7 shows that for a given link ψ one can specify one of the partial losses λ_y but then properness fixes the other partial loss λ_{-y}. Similarly, given an arbitrary choice of the partial losses, equation 10 gives the single link which will guarantee the overall loss is proper.

5 CONVEXITY OF PROPER COMPOSITE LOSSES

We have seen that proper composite losses are defined by the proper loss ℓ and the link ψ. We have further seen from Corollary 6 that it is natural to parametrise proper composite losses in terms of w and ψ', and combine them as ρ. One may wish to choose a weight function w and determine which links ψ lead to a convex loss; or choose a link ψ and determine which weight functions w (and hence proper losses) lead to a convex composite loss. The main result of this section (Theorem 11) answers these questions by characterising convex proper composite losses in terms of (w, ψ') or ρ.

We first establish some convexity results for losses and their conditional and full risks.

Lemma 8 *Let $\ell : \mathcal{Y} \times \mathcal{V} \to [0, \infty)$ denote an arbitrary loss. Then the following are equivalent:*

1. *$v \mapsto \ell(y, v)$ is convex for all $y \in \{-1, 1\}$,*
2. *$v \mapsto L(\eta, v)$ is convex for all $\eta \in [0, 1]$,*
3. *$v \mapsto \hat{\mathbb{L}}(v, S) := \frac{1}{|S|}\sum_{(x,y)\in S} \ell(y, v(x))$ is convex for all finite $S \subset \mathcal{X} \times \mathcal{Y}$.*

Proof $1 \Rightarrow 2$: By definition, $L(\eta, v) = (1-\eta)\ell(-1, v) + \eta\ell(1, v)$ which is just a convex combination of convex functions and hence convex.

$2 \Rightarrow 1$: Choose $\eta = 0$ and $\eta = 1$ in the definition of L.

$1 \Rightarrow 3$: For a fixed (x, y), the function $v \mapsto \ell(y, v(x))$ is convex since ℓ is convex. Thus, $\hat{\mathbb{L}}$ is convex as it is a non-negative weighted sum of convex functions.

$3 \Rightarrow 1$: The convexity of $\hat{\mathbb{L}}$ holds for every S so for each $y \in \{-1, 1\}$ choose $S = \{(x, y)\}$ for some x. In each case $v \mapsto \hat{\mathbb{L}}(v, S) = \ell(y, v(x))$ is convex. ■

The following theorem characterises convexity of proper composite losses with invertible links.

Theorem 9 *Let $\ell^\psi(y,v)$ be a composite loss comprising an invertible link with inverse $q := \psi^{-1}$ and strictly proper loss with weight function w. Assume $q'(\cdot) > 0$. Then $v \mapsto \ell^\psi(y,v)$ is convex for $y \in \{-1,1\}$ if and only if*

$$-\frac{1}{x} \;\leq\; \frac{w'(x)}{w(x)} - \frac{\psi''(x)}{\psi'(x)} \;\leq\; \frac{1}{1-x}, \quad \forall x \in (0,1). \tag{11}$$

This theorem suggests a very natural parametrisation of proper composite losses is via (w, ψ'). Observe that $w, \psi' \colon [0,1] \to \mathbb{R}^+$. (But also see the comment following Theorem 11.)

Proof We can write the conditional composite loss as

$$L^\psi(\eta, v) = \eta \ell_1(q(v)) + (1-\eta)\ell_{-1}(q(v))$$

and so by substituting $q = \psi^{-1}$ into (9) we have

$$\frac{\partial}{\partial v} L^\psi(\eta, v) \quad = \quad w(q(v))q'(v)[q(v) - \eta]. \tag{12}$$

A necessary and sufficient condition for $v \mapsto \ell^\psi(y,v) = L^\psi(y,v)$ to be convex for $y \in \{-1,1\}$ is that

$$\frac{\partial^2}{\partial v^2} L^\psi(y,v) \geq 0, \quad \forall v \in \mathbb{R},\ \forall y \in \{-1,1\}.$$

Using (12) the above condition is equivalent to

$$\begin{aligned}[w(q(v))q'(v)]'(q(v) - [\![y=1]\!]) \\ + w(q(v))q'(v)q'(v) \;\geq\; 0,\ \forall v \in \mathbb{R},\end{aligned} \tag{13}$$

where $[w(q(v))q'(v)]' := \frac{\partial}{\partial v} w(q(v))q'(v)$. Inequality (13) is equivalent to (Buja et al., 2005, equation 39). By further manipulations, we can simplify (13) considerably.

Since $[\![y=1]\!]$ is either 0 or 1 we equivalently have the two inequalities

$$\begin{aligned}[w(q(v))q'(v)]'q(v) + w(q(v))(q'(v))^2 \geq 0,\ \forall v \in \mathbb{R}, \\ [w(q(v))q'(v)]'(q(v)-1) + w(q(v))(q'(v))^2 \geq 0,\ \forall v \in \mathbb{R},\end{aligned}$$

which we shall rewrite as the pair of inequalities

$$w(q(v))(q'(v))^2 \geq -q(v)[w(q(v))q'(v)]', \quad \forall v \in \mathbb{R}, \tag{14}$$
$$w(q(v))(q'(v))^2 \geq (1-q(v))[w(q(v))q'(v)]', \forall v \in \mathbb{R}. \tag{15}$$

Observe that if $q(\cdot) = 0$ (resp. $1 - q(\cdot) = 0$) then (14) (resp. (15)) is satisfied anyway because of the assumption on q' and the fact that w is non-negative. It is thus equivalent to restrict consideration to v in the set

$$\{x : q(x) \neq 0 \ \&\ (1-q(x)) \neq 0\} = q^{-1}((0,1)) = \psi((0,1)).$$

Combining (14) and (15) we obtain the equivalent condition

$$\frac{(q'(v))^2}{1-q(v)} \geq \frac{[w(q(v))q'(v)]'}{w(q(v))} \geq \frac{-(q'(v))^2}{q(v)},\ \forall v \in \psi((0,1)), \tag{16}$$

where we have used the fact that $q \colon \mathbb{R} \to [0,1]$ and is thus sign-definite and consequently $-q(\cdot)$ is always negative and division by $q(v)$ and $1-q(v)$ is permissible since as argued we can neglect the cases when these take on the value zero, and division by $w(q(v))$ is permissible by the assumption of *strict* properness since that implies $w(\cdot) > 0$. Now

$$[w(q(\cdot))q'(\cdot)]' = w'(q(\cdot))q'(\cdot)q'(\cdot) + w(q(\cdot))q''(\cdot)$$

and thus (16) is equivalent to

$$\begin{aligned}\frac{(q'(v))^2}{1-q(v)} \quad &\geq \quad \frac{w'(q(v))(q'(v))^2 + w(q(v))q''(v)}{w(q(v))} \\ &\geq \quad \frac{-(q'(v))^2}{q(v)},\ \forall v \in \psi((0,1)).\end{aligned} \tag{17}$$

Now divide all sides of (17) by $(q'(\cdot))^2$ (which is permissible by assumption). This gives the equivalent condition

$$\frac{1}{1-q(v)} \geq \frac{w'(q(v))}{w(q(v))} + \frac{q''(v)}{(q'(v))^2} \geq \frac{-1}{q(v)},\ \forall v \in \psi((0,1)).$$

Let $x = q(v)$ and so $v = q^{-1}(x) = \psi(x)$. Then the above is equivalent to

$$\frac{1}{1-x} \;\geq\; \frac{w'(x)}{w(x)} + \frac{q''(\psi(x))}{(q'(\psi(x)))^2} \;\geq\; \frac{-1}{x}, \quad \forall x \in (0,1).$$

Now $\frac{1}{q'(\psi(x))} = \frac{1}{q'(q^{-1}(x))} = (q^{-1})'(x) = \psi'(x)$. Thus the above equivalent to

$$\frac{1}{1-x} \;\geq\; \frac{w'(x)}{w(x)} + \Phi_\psi(x) \;\geq\; \frac{-1}{x}, \quad \forall x \in (0,1), \tag{18}$$

where

$$\Phi_\psi(x) := q''\left(\psi(x)\right)\left(\psi'(x)\right)^2. \tag{19}$$

All of the above steps are equivalences. We have thus shown that

(18) is true $\Leftrightarrow v \mapsto L^\psi(y,v)$ is convex for $y \in \{-1,1\}$

where the right hand side is equivalent to the assertion in the theorem by Lemma 8.

Finally we simplify Φ_ψ. We first compute q'' in terms of $\psi = q^{-1}$. Observe that $q' = (\psi^{-1})' = \frac{1}{\psi'(\psi^{-1}(\cdot))}$. Thus

$$\begin{aligned}q''(\cdot) \quad &= \quad (\psi^{-1})''(\cdot) = \left(\frac{1}{\psi'(\psi^{-1}(\cdot))}\right)' \\ &= \quad \frac{-1}{\left(\psi'(\psi^{-1}(\cdot))\right)^2}\psi''(\psi^{-1}(\cdot))\left(\psi^{-1}(\cdot)\right)' \\ &= \quad \frac{-1}{\left(\psi'(\psi^{-1}(\cdot))\right)^3}\psi''(\psi^{-1}(\cdot)).\end{aligned}$$

Thus by substitution

$$\begin{aligned}\Phi_\psi(\cdot) &= \frac{-1}{\left(\psi'(\psi^{-1}(\psi(\cdot)))\right)^3}\psi''(\psi(\psi^{-1}(\cdot)))\left(\psi'(\cdot)\right)^2 \\ &= \frac{-1}{\left(\psi'(\cdot)\right)^3}\psi''(\cdot)\left(\psi'(\cdot)\right)^2 \\ &= -\frac{\psi''(\cdot)}{\psi'(\cdot)}. \qquad (20)\end{aligned}$$

Substituting the simpler expression (20) for Φ_ψ into (18) completes the proof. ■

Corollary 10 *Proper losses are convex if and only if*

$$-\frac{1}{x} \;\leq\; \frac{w'(x)}{w(x)} \;\leq\; \frac{1}{1-x}, \quad \forall x \in (0,1).$$

5.1 A Simpler Characterisation of Convex Proper Composite Losses

The following theorem more explicitly describes all proper losses that generate a convex composite loss given a particular link function. Noting that loss functions can be multiplied by a scalar without affecting what a learning algorithm will do, it is convenient to normalise them by normalising their weight functions by setting $w(\frac{1}{2}) = 1$. (Observe too that (11) is scale invariant with respect to w.)

Theorem 11 *Consider a proper composite loss ℓ^ψ with invertible link ψ and (strictly proper) weight w normalised such that $w(\frac{1}{2}) = 1$. Then ℓ^ψ is convex if and only if*

$$\frac{\psi'(x)}{x} \;\lesseqgtr\; 2\psi'(\tfrac{1}{2})\, w(x) \;\lesseqgtr\; \frac{\psi'(x)}{1-x}, \quad \forall x \in (0,1), \quad (21)$$

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

Observe that the condition (21) is equivalent to

$$\frac{1}{2\psi'(\frac{1}{2})x} \;\lesseqgtr\; \rho(x) \;\lesseqgtr\; \frac{1}{2\psi'(\frac{1}{2})(1-x)}, \quad \forall x \in (0,1),$$

which reinforces the importance of the function $\rho(\cdot)$.

Proof Observe that if w satisfies (11) then so does αw for all $\alpha \in (0, \infty)$. Thus without loss of generality we will normalise w such that $w(\frac{1}{2}) = 1$. [4] Observing that $\frac{w'(x)}{w(x)} = (\log w)'(x)$ we let $g(x) := \log w(x)$. We have that $g(v) = \int_{\frac{1}{2}}^v g'(x)dx + g(\frac{1}{2})$ and $g(\frac{1}{2}) = \log w(\frac{1}{2}) = 0$. Thus from (11) we obtain

$$-\frac{1}{x} - \Phi_\psi(x) \;\leq\; g'(x) \;\leq\; \frac{1}{1-x} - \Phi_\psi(x).$$

For $v \geq \frac{1}{2}$ we thus have

$$\int_{\frac{1}{2}}^v -\frac{1}{x} - \Phi_\psi(x)dx \;\leq\; g(v) \;\leq\; \int_{\frac{1}{2}}^v \frac{1}{1-x} - \Phi_\psi(x)dx.$$

Conversely, for $v \leq \frac{1}{2}$ we have

$$\int_{\frac{1}{2}}^v -\frac{1}{x} - \Phi_\psi(x)dx \;\geq\; g(v) \;\geq\; \int_{\frac{1}{2}}^v \frac{1}{1-x} - \Phi_\psi(x)dx,$$

and thus

$$\begin{aligned}-\ln v - \ln 2 - \textstyle\int_{\frac{1}{2}}^v \Phi_\psi(x)dx \;\lesseqgtr\; g(v) \;\lesseqgtr\; \\ -\ln 2 - \ln(1-v) - \textstyle\int_{\frac{1}{2}}^v \Phi_\psi(x)dx.\end{aligned}$$

Since $\exp(\cdot)$ is monotone increasing we can apply it to all terms and obtain

$$\begin{aligned}\tfrac{1}{2v}\exp\left(-\textstyle\int_{\frac{1}{2}}^v \Phi_\psi(x)dx\right) \;\lesseqgtr\; w(v) \;\lesseqgtr\; \\ \tfrac{1}{2(1-v)}\exp\left(-\textstyle\int_{\frac{1}{2}}^v \Phi_\psi(x)dx\right). \qquad (22)\end{aligned}$$

Now

$$\begin{aligned}\int_{\frac{1}{2}}^v \Phi_\psi(x)dv &= \int_{\frac{1}{2}}^v -\frac{\psi''(x)}{\psi'(x)}dx = -\int_{\frac{1}{2}}^v (\log\psi')'(x)dx \\ &= -\log\psi'(v) + \log\psi'(\tfrac{1}{2})\end{aligned}$$

and so

$$\exp\left(-\int_{\frac{1}{2}}^v \Phi_\psi(x)dx\right) = \frac{\psi'(v)}{\psi'(\frac{1}{2})}.$$

Substituting into (22) completes the proof. ■

If ψ is the identity (*i.e.* if ℓ^ψ is itself proper) we get the simpler constraints

$$\frac{1}{2x} \;\lesseqgtr\; w(x) \;\lesseqgtr\; \frac{1}{2(1-x)}, \quad \forall x \in (0,1), \qquad (23)$$

which are illustrated as the shaded region in Figure 1. Observe that the (normalised) weight function for squared loss is $w(c) = 1$ which is indeed within the shaded region as one would expect.

Consider the link $\psi^{\text{logit}}(c) := \log\left(\frac{c}{1-c}\right)$ with corresponding inverse link $q(c) = \frac{1}{1+e^{-c}}$. One can check that $\psi'(c) = \frac{1}{c(1-c)}$. Thus, constraints on the weight function w to ensure a convex composite loss are

$$\frac{1}{8x^2(1-x)} \;\lesseqgtr\; w(x) \;\lesseqgtr\; \frac{1}{8x(1-x)^2}, \quad \forall x \in (0,1).$$

[4] We chose to normalise about $\frac{1}{2}$ for two reasons: symmetry and the fact that w can have non-integrable singularities at 0 and 1; see e.g. (Buja et al., 2005).

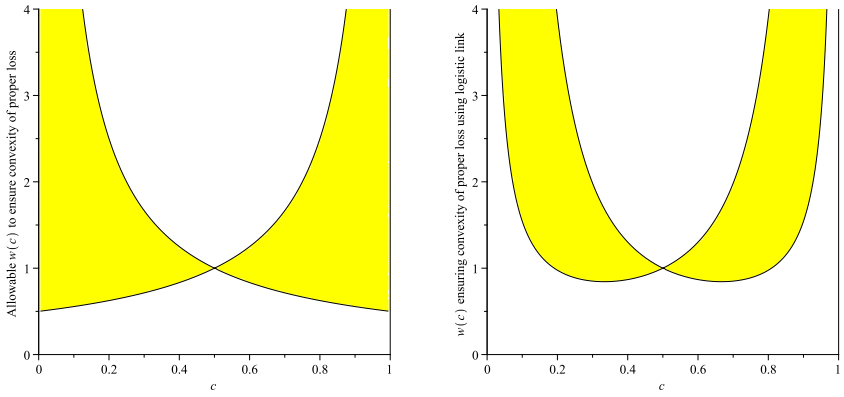

Figure 1: Allowable normalised weight functions to ensure convexity of composite loss functions with identity link (left) and logistic link (right).

This is shown graphically in Figure 1. One can compute similar regions for any link. Two other examples are the Complementary Log-Log link $\psi^{\mathrm{CLL}}(x) = \log(-\log(1-x))$ (cf. (McCullagh and Nelder, 1989)), the "square link" $\psi^{\mathrm{sq}}(x) = x^2$ and the "cosine link" $\psi^{\cos}(x) = 1 - \cos(\pi x)$. All are illustrated in Figure 2.

The reason for considering these last two rather unusual links is to illustrate the following fact. Observing that the allowable region in Figure 1 precludes weight functions that approach zero at the endpoints of the interval, and noting that in order to well approximate the behaviour of 0-1 loss (with its weight function being $w_{0-1}(c) = \delta(c - \frac{1}{2})$) one would like a weight function that does indeed approach zero at the end points, it is natural to ask what constraints are imposed upon a link ψ such that a composite loss with that link and a weight function $w(c)$ such that

$$\lim_{c \searrow 0} w(c) = \lim_{c \nearrow 1} w(c) = 0 \tag{24}$$

is convex. Inspection of (21) reveals it is necessary that $\psi'(x) \to 0$ as $x \to 0$ and $x \to 1$. Such ψ necessarily have bounded range and thus the inverse link ψ^{-1} is only defined on a finite interval and furthermore the gradient of ψ^{-1} will be arbitrarily large. If one wants inverse links defined on the whole real line (such as the logistic link) then one cannot obtain a convex composite link with the associated proper loss having a weight function satisfying (24). Thus one cannot choose an effectively usable link to ensure convexity of a proper loss that is arbitrarily "close to" 0-1 loss in the sense of the closeness of corresponding weight functions.

The requirement that a loss be convex and proper constrains the weight function considerably.

Corollary 12 *If a loss is proper and convex, then it is strictly proper.*

The proof 12 uses the following Gronwall-style lemma (Bainov and Simeonov, 1992, Lemma 1.1.1).

Lemma 13 *Let $b\colon \mathbb{R} \to \mathbb{R}$ be continuous for $t \geq \alpha$. Let $v(t)$ be differentiable for $t \geq \alpha$ and suppose $v'(t) \leq b(t)v(t)$, for $t \geq \alpha$ and $v(\alpha) \leq v_0$. Then for $t \geq \alpha$,*

$$v(t) \leq v_0 \exp\left(\int_\alpha^t b(s)ds\right).$$

Proof (Corollary 12) Observe that the RHS of (11) implies $w'(v) \leq \frac{w(v)}{1-v}$, $v \geq 0$. Suppose $w(0) = 0$. Then $v_0 = 0$ and by setting $\alpha = 0$ the lemma implies

$$w(t) \leq v_0 \exp\left(\int_0^t \frac{1}{1-s} ds\right) = \frac{v_0}{1-t} = 0, \quad t \in (0, 1].$$

So if $w(0) = 0$ then $w(t) = 0$ for all $t \in (0, 1)$. Choosing any other $\alpha \in (0, 1)$ leads to a similar conclusion. Thus if $w(t) = 0$ for some $t \in [0, 1)$, $w(s) = 0$ for all $s \in [t, 1]$. Thus $w(t) > 0$ for all $t \in [0, 1]$ and by the remark following Theorem 4, ℓ is strictly proper. ■

6 CONCLUSIONS

We have characterised the convexity of composite binary losses in terms of the weight function associated with the proper loss, and the link function. The parametrisation of a composite loss in terms (w, ψ') (or ρ) has advantages over using (ϕ, ψ) (for margin losses) or $(\underline{L}, \psi)$. As Masnadi-Shirazi and Vasconcelos (2009) explain, the representation in terms of (ϕ, ψ) is in general not unique. The representation in terms of $\underline{L}$ is harder to intuit: whilst the conditional Bayes risk for squared loss and 0-1 loss are "close" (compare graphs of $c \mapsto c(1-c)$ and $c \mapsto c \wedge (1-c)$) their weight functions they are seen to be very different ($w(c) = 1$ versus

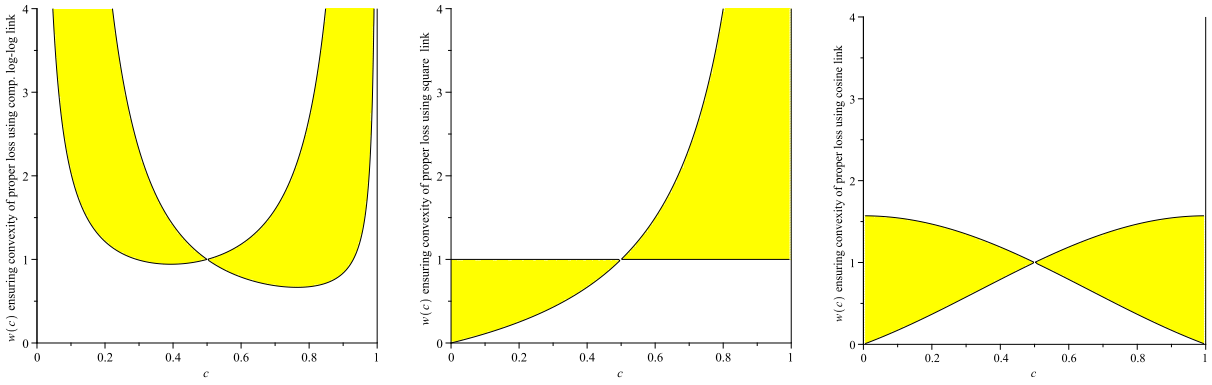

Figure 2: Allowable normalised weight functions to ensure convexity of loss functions with complementary log-log, square and cosine links.

$w(c) = 2\delta(c - \frac{1}{2})$). We have also seen that on the basis of Theorem 9, the parametrisation (w, ψ') is perhaps the most natural — there is a nice symmetry between the loss and the link as they are both parametrised in terms of non-negative functions on $[0, 1]$.

The parametrisation (w, ψ') arising from our characterisation suggests an implementation of a novel inductive principle known as *surrogate tuning* (Nock and Nielsen, 2009). The idea of surrogate tuning is simple: noting that the best surrogate loss depends on the problem at hand, adapt the surrogate loss you are using to the problem. To do so it is important to have a good parametrisation of the loss as is given by the weight function view of Theorem 11. It would be easy to develop low dimensional parametrisations of w that satisfy the conditions of this theorem. This would allow a learning algorithm to explore the space of convex losses. One could (taking care with the subsequent multiple hypothesis testing problem) regularly *evaluate* the 0-1 loss of the hypotheses so obtained.

Surrogate tuning differs from loss *tailoring* (Hand, 1994; Hand and Vinciotti, 2003; Buja et al., 2005) which adapts the loss to what is important rather than adjusting a surrogate for computational reasons (which is why convexity is desirable in the first place).

Acknowledgements

This work was supported by the Australian Research Council and NICTA; an initiative of the Commonwealth Government under Backing Australia's Ability.

References

Bach, F., Heckerman, D., and Horvitz, E. (2006). Considering Cost Asymmetry in Learning Classifiers. *Journal of Machine Learning Research*, 7:1713–1741.

Bainov, D. and Simeonov, P. (1992). *Integral Inequalities and Applications*. Kluwer, Dordrecht.

Bartlett, P. and Tewari, A. (2007). Sparseness vs Estimating Conditional Probabilities: Some Asymptotic Results. *The Journal of Machine Learning Research*, 8:775–790.

Buja, A., Stuetzle, W., and Shen, Y. (2005). Loss functions for binary class probability estimation and classification: Structure and applications. Technical report, University of Pennsylvania.

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, Learning, and Games*. Cambridge University Press.

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

Hand, D. (1994). Deconstructing Statistical Questions. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 157(3):317–356.

Hand, D. and Vinciotti, V. (2003). Local Versus Global Models for Classification Problems: Fitting Models Where it Matters. *The American Statistician*, 57(2):124–131.

Masnadi-Shirazi, H. and Vasconcelos, N. (2009). On the design of loss functions for classification: theory, robustness to outliers, and savageboost. In *Advances in Neural Information Processing Systems 21*, pages 1049–1056.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman & Hall/CRC.

Nock, R. and Nielsen, F. (2009). On the efficient minimization of classification calibrated surrogates. In *Advances in Neural Information Processing Systems 21*, pages 1201–1208. MIT Press.

Reid, M. D. and Williamson, R. C. (2009). Information, divergence and risk for binary experiments. arXiv preprint arXiv:0901.0356v1.

Shuford, E., Albert, A., and Massengill, H. (1966). Admissible probability measurement procedures. *Psychometrika*, 31(2):125–145.