Journal of Machine Learning Research 11 (2010) 2387-2422 Submitted 12/09; Revised 6/10; Published 9/10

Composite Binary Losses

Mark D. Reid* MARK.REID@ANU.EDU.AU
Robert C. Williamson* BOB.WILLIAMSON@ANU.EDU.AU
School of Computer Science, Building 108
Australian National University
Canberra ACT 0200, Australia

Editor: Rocco Servedio

Abstract

We study losses for binary classification and class probability estimation and extend the understanding of them from margin losses to general composite losses which are the composition of a proper loss with a link function. We characterise when margin losses can be proper composite losses, explicitly show how to determine a symmetric loss in full from half of one of its partial losses, introduce an intrinsic parametrisation of composite binary losses and give a complete characterisation of the relationship between proper losses and "classification calibrated" losses. We also consider the question of the "best" surrogate binary loss. We introduce a precise notion of "best" and show there exist situations where two convex surrogate losses are incommensurable. We provide a complete explicit characterisation of the convexity of composite binary losses in terms of the link function and the weight function associated with the proper loss which make up the composite loss. This characterisation suggests new ways of "surrogate tuning" as well as providing an explicit characterisation of when Bregman divergences on the unit interval are convex in their second argument. Finally, in an appendix we present some new algorithm-independent results on the relationship between properness, convexity and robustness to misclassification noise for binary losses and show that all convex proper losses are non-robust to misclassification noise.

Keywords: surrogate loss, convexity, probability estimation, classification, Fisher consistency, classification-calibrated, regret bound, proper scoring rule, Bregman divergence, robustness, misclassification noise

1. Introduction

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 proper loss (defined below) and a link function (also defined below). In this paper we study composite binary losses and develop a number of new characterisation results. Several of these results can be seen as an extension of the work by Buja et al. (2005) applied to an analysis of composite losses by Masnadi-Shirazi and Vasconcelos (2009).

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 often used to map the outputs of a predictor to the interval $[0,1]$ so that they can be interpreted as probabilities. Having such

*. Also at National ICT Australia.

probabilities is often important in applications, and there has been considerable interest in understanding how to get accurate probability estimates (Platt, 2000; Gneiting and Raftery, 2007; Cohen and Goldszmidt, 2004) and understanding the implications of requiring loss functions provide good probability estimates (Bartlett and Tewari, 2007).

Much previous work in the machine learning literature has focussed on *margin losses* which intrinsically treat positive and negative classes symmetrically. However it is now well understood how important it is to be able to deal with the non-symmetric case (Zellner, 1986; Elkan, 2001; Provost and Fawcett, 2001; Buja et al., 2005; Bach et al., 2006; Beygelzimer et al., 2008; Christoffersen and Diebold, 2009). A key goal of the present work is to consider composite losses in the general (non-symmetric) situation. Since our development is for completely general losses, we automatically cover non-symmetric losses. The generalised notion of classification calibration developed in §5 is intrinsically non-symmetric.

1.1 Overview and Contributions

We now provide an overview of the paper's structure, highlighting the novel contributions and how they relate to existing work. Central to this work are the notions of a loss and its associated conditional and full risk. These are introduced and briefly discussed in §2.

In §3 we introduce losses for Class Probability Estimation (CPE), define some technical properties of them, and present some structural results originally by Shuford et al. (1966) and Savage (1971) and recently studied in a machine learning context by Buja et al. (2005) and Masnadi-Shirazi and Vasconcelos (2009). The most important of these are Theorem 4 which gives a representation of proper losses in terms of its associated conditional Bayes risk function, and Theorem 1 which relates a proper loss's partial losses to its "weight function"—the negative second derivative of the conditional Bayes risk (see Corollary 3). We use these to provide a novel characterisation of proper symmetric CPE losses. Specifically, Theorem 9 shows these losses are completely determined by the behaviour of one of its partial losses on half the unit interval.

Learning algorithms often make real-valued predictions that are not directly interpretable as probability estimates but require a link function which maps their output to the interval $[0,1]$. In §4 we define composite losses as the composition of a CPE loss and a link. The new contributions of this section are Theorem 10 which generalises Theorem 1 to composite losses, and Corollaries 12 and 14 which shows how requiring properness completely determines the link function for composite and margin losses. We also introduce a natural and intrinsic parametrisation of proper composite losses that is a generalisation of the weight function and show how it can be used to easily derive gradients for stochastic descent algorithms.

In §5 we generalise the notion of classification calibrated losses (as studied, for example, by Bartlett et al., 2006) so it applies to non-symmetric composite losses (i.e., not just margin losses) and provide a characterisation of it in Theorem 17. We also describe how this new notion of classification calibrated relates to proper CPE and composite losses via its connection with the weight function.

The main results of this paper are found in §6: Theorems 24 and 29 characterise when proper composite losses are convex. These characterisation are in terms of some easily testable constraints relating the losses' weight and link functions. The results also characterise when a Bregman divergence on $[0,1]$ is convex in its second argument (§6.3).

In §7 we study how the above insights can be applied to the problem of choosing a surrogate loss. Here, a *surrogate* loss function is a loss function which is not exactly what one wishes to minimise but is easier to work with algorithmically. This is still a relatively new area of research and our aim here is to open up a discussion rather than have the final word. To do so we define a well founded notion of "best" surrogate loss and show that some convex surrogate losses are incommensurable on some problems. We also consider some other notions of "best" and explicitly determine the surrogate loss that has the best surrogate regret bound in a certain sense.

Finally, in §8 we draw some more general conclusions. In particular, we argue that the weight and link function parametrisation of losses provides a convenient way to work with an entire class of losses that are central to probability estimation and may provide new ways of approaching the problem of "surrogate tuning" (Nock and Nielsen, 2009b).

Appendix C collects several observations which build upon some of the results in the main paper but are digressions from its central themes. In it, we present some new algorithm-independent results on the relationship between properness, convexity and robustness to misclassification noise for binary losses and show that all convex proper losses are non-robust to misclassification noise.

2. Losses and Risks

We write $x \wedge y := \min(x,y)$ and $[\![p]\!] = 1$ if p is true and $[\![p]\!] = 0$ otherwise.[1] 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[2] $\ell : \mathcal{Y} \times \mathcal{V} \to [0, \infty)$. We interpret such a loss as giving a penalty $\ell(y,v)$ when predicting the value v when an observed label is 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

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

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.[3] 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 $c \in (0,1)$ is given by

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

1. This is the Iverson bracket notation as recommended by Knuth (1992).
2. Restricting the output of a loss to $[0,\infty)$ is equivalent to assuming the loss has a lower bound and then translating its output.
3. These are known as *scoring rules* in the statistical literature (Gneiting and Raftery, 2007).

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 loss 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 $v : \mathcal{X} \to \mathcal{V}$:

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

We sometimes write $\mathbb{L}(v, \mathbb{P})$ for $\mathbb{L}(\eta, v, M)$ where (η, M) corresponds to the joint distribution $\mathbb{P}$. We write ℓ, L and $\mathbb{L}$ for the loss, point-wise and full risk 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*.

There has been increasing awareness of the importance of the conditional Bayes risk curve $\underline{L}(\eta)$—also known as "generalized entropy" (Grünwald and Dawid, 2004)—in the analysis of losses for probability estimation (Kalnishkan et al., 2004, 2007; Abernethy et al., 2009; Masnadi-Shirazi and Vasconcelos, 2009). Below we will see how it is effectively the curvature of $\underline{L}$ that determines much of the structure of these losses.

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*—a large, natural class of losses for class probability estimation.

3.1 Proper, Fair, Definite and Regular Losses

There are a few properties of losses for probability estimation that we will require. 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. The main place fairness is relied upon is in the integral representation of Theorem 6 where it is used to get rid of some constants of integration. In order to explicitly construct losses from their associated "weight functions" as shown in Theorem 7, we will require that the loss be *definite*, that is, its point-wise Bayes risk for deterministic events (i.e., $\eta = 0$ or $\eta = 1$) must be bounded from below:

$$\underline{L}(0) > -\infty \, , \, \underline{L}(1) > -\infty.$$

Since properness of a loss ensures $\underline{L}(\eta) = L(\eta, \eta)$ we see that a fair proper loss is necessarily definite since $L(0,0) = \ell_{-1}(0) = 0 > -\infty$, and similarly for $L(1,1)$. Conversely, if a proper loss is definite then the finite values $\ell_{-1}(0)$ and $\ell_1(1)$ can be subtracted from $\ell_{-1}(\cdot)$ and $\ell_1(\cdot)$ to make it fair.

Finally, for Theorem 4 to hold at the endpoints of the unit interval, we require a loss to be *regular*;[4] that is,

$$\lim_{\eta \searrow 0} \eta \ell_1(\eta) = \lim_{\eta \nearrow 1} (1-\eta)\ell_{-1}(\eta) = 0.$$

Intuitively, this condition ensures that making mistakes on events that never happen should not incur a penalty. In most of the situations we consider in the remainder of this paper will involve losses which are proper, fair, definite and regular.

3.2 The Structure of Proper Losses

A key result in the study of proper losses is originally due to Shuford et al. (1966) and Staël von Holstein (1970) (confer Aczel and Pfanzagl, 1967) though our presentation follows that of Buja et al. (2005). The following theorem[5] characterises proper losses for probability estimation via a constraint on the relationship between its partial losses.

Theorem 1 (Shuford *et al.*) *Suppose $\ell : \{-1,1\} \times [0,1] \to \mathbb{R}$ is a 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 $\hat{\eta} \in (0,1)$*

$$\frac{-\ell_1'(\hat{\eta})}{1-\hat{\eta}} = \frac{\ell_{-1}'(\hat{\eta})}{\hat{\eta}} = w(\hat{\eta}) \tag{2}$$

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

The equalities in (2) should be interpreted in the distributional sense.

This simple characterisation of the structure of proper losses has a number of interesting implications. Observe from (2) 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 $\eta \in [0,1]$*

$$\frac{\partial}{\partial \hat{\eta}} L(\eta, \hat{\eta}) = (1-\eta)\ell_{-1}'(\hat{\eta}) + \eta \ell_1'(\hat{\eta}) = (\hat{\eta} - \eta) w(\hat{\eta}). \tag{3}$$

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

4. This is equivalent to the conditions of Savage (1971) and Schervish (1989).
5. This is a restatement of Theorem 1 in Shuford et al. (1966).

Corollary 3 *Let ℓ be a a twice differentiable*[6] *proper loss with weight function w defined as in Equation (2). Then for all $c \in (0,1)$ its conditional Bayes risk $\underline{L}$ satisfies*

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

One immediate consequence of this corollary is that the conditional Bayes risk for a proper loss is always concave. Along with an extra constraint, this gives another characterisation of proper losses (Savage, 1971; Reid and Williamson, 2009a).

Theorem 4 (Savage) *A loss function ℓ is proper if and only if its point-wise Bayes risk $\underline{L}(\eta)$ is concave and for each $\eta, \hat{\eta} \in (0,1)$*

$$L(\eta, \hat{\eta}) = \underline{L}(\hat{\eta}) + (\eta - \hat{\eta})\underline{L}'(\hat{\eta}).$$

Furthermore if ℓ is regular this characterisation also holds at the endpoints $\eta, \hat{\eta} \in \{0,1\}$.

This link between loss and concave functions makes it easy to establish a connection, as Buja et al. (2005) do, between *regret* $\Delta L(\eta, \hat{\eta}) := L(\eta, \hat{\eta}) - \underline{L}(\eta)$ for proper losses and *Bregman divergences*. The latter are generalisations of distances and are defined in terms of convex functions. Specifically, if $f : \mathcal{S} \to \mathbb{R}$ is a convex function over some convex set $\mathcal{S} \subseteq \mathbb{R}^n$ then its associated Bregman divergence[7] is

$$D_f(s, s_0) := f(s) - f(s_0) - \langle s - s_0, \nabla f(s_0) \rangle$$

for any $s, s_0 \in \mathcal{S}$, where $\nabla f(s_0)$ is the gradient of f at s_0. By noting that over $\mathcal{S} = [0,1]$ we have $\nabla f = f'$, these definitions lead immediately to the following corollary of Theorem 4.

Corollary 5 *If ℓ is a proper loss then its regret is the Bregman divergence associated with $f = -\underline{L}$. That is,*

$$\Delta L(\eta, \hat{\eta}) = D_{-\underline{L}}(\eta, \hat{\eta}).$$

Many of the above results can be observed graphically by plotting the conditional risk for a proper loss as in Figure 1. Here we see the two partial losses on the left and right sides of the figure are related, for each fixed $\hat{\eta}$, by the linear map $\eta \mapsto L(\eta, \hat{\eta}) = (1-\eta)\ell_{-1}(\hat{\eta}) + \eta\ell_1(\hat{\eta})$. For each fixed η the properness of ℓ requires that these convex combinations of the partial losses (each slice parallel to the left and right faces) are minimised when $\hat{\eta} = \eta$. Thus, the lines joining the partial losses are tangent to the conditional Bayes risk curve $\eta \mapsto \underline{L}(\eta) = L(\eta, \eta)$ shown above the dotted diagonal. Since the conditional Bayes risk curve is the lower envelope of these tangents it is necessarily concave. The coupling of the partial losses via the tangents to the conditional Bayes risk curve demonstrates why much of the structure of proper losses is determined by the curvature of $\underline{L}$—that is, by the weight function w.

The relationship between a proper loss and its associated weight function is captured succinctly by Schervish (1989) via the following representation of proper losses as a weighted integral of the cost-weighted misclassification losses ℓ_c defined in (1). The reader is referred to Reid and Williamson (2009b) for the details, proof and the history of this result.

6. 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 $w(c) = \delta(c - \frac{1}{2})$.
7. A concise summary of Bregman divergences and their properties is given by Banerjee et al. (2005, Appendix A).

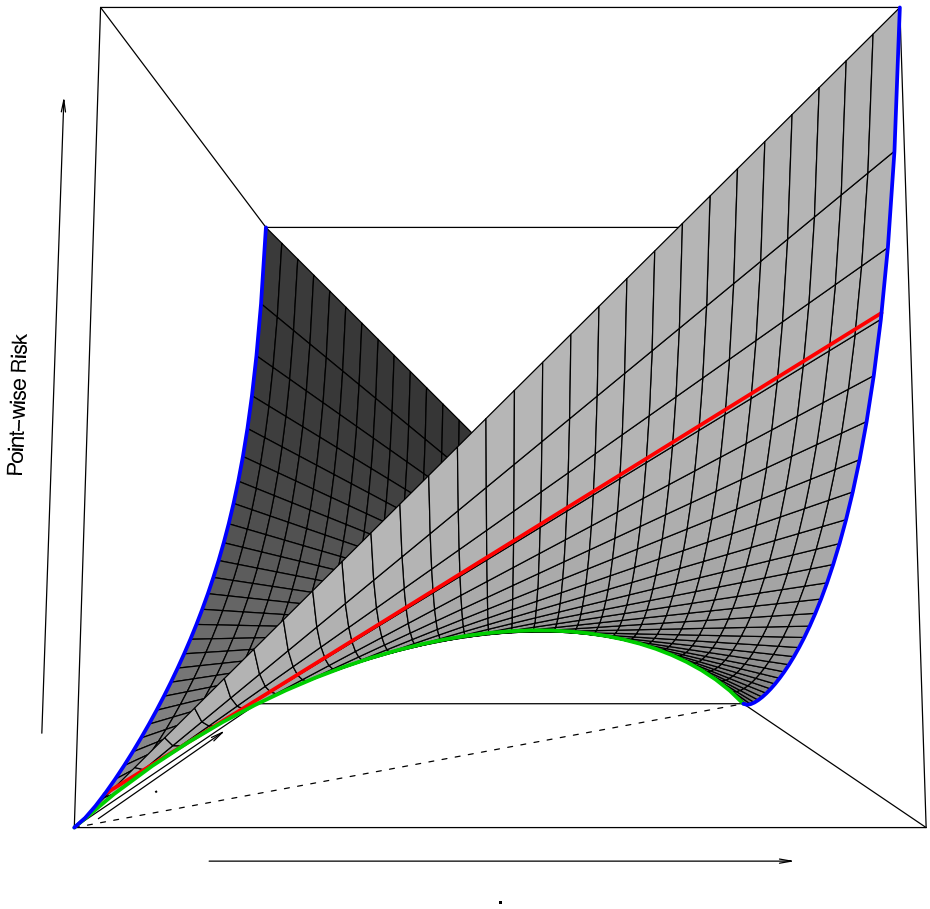

Figure 1: The structure of the conditional risk $L(\eta,\hat{\eta})$ for a proper loss (surface). The loss is log loss and its partials $\ell_{-1}(\hat{\eta}) = -\log(\hat{\eta})$ and $\ell_1(\hat{\eta}) = -\log(1-\hat{\eta})$ shown on the left and right faces of the box. The conditional Bayes risk is the curve on the surface above the dotted line $\hat{\eta} = \eta$. The line connecting points on the partial loss curves shows the conditional risk for a fixed prediction $\hat{\eta}$.

Theorem 6 (Schervish) *Let $\ell : \mathcal{Y} \times [0,1] \to \mathbb{R}$ be a fair, proper loss. Then for each $\hat{\eta} \in (0,1)$ and $y \in \mathcal{Y}$*

$$\ell(y,\hat{\eta}) = \int_0^1 \ell_c(y,\hat{\eta})\, w(c)\, dc, \tag{4}$$

where $w = -\underline{L}''$. Conversely, if ℓ is defined by (4) for some weight function $w : (0,1) \to [0,\infty)$ then it is proper.

Some example losses and their associated weight functions are given in Table 1. 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 $(0,1)$. The following theorem from Reid and Williamson (2009a) shows how to explicitly construct a loss in terms of a weight function.

$w(c)$	$\ell_{-1}(\hat{\eta})$	$\ell_1(\hat{\eta})$	Loss
$2\delta\left(\frac{1}{2}-c\right)$	$[\![\hat{\eta} > \frac{1}{2}]\!]$	$[\![\hat{\eta} \leq \frac{1}{2}]\!]$	0-1
$\delta(c-c_0)$	$c_0[\![\hat{\eta} \geq c_0]\!]$	$(1-c_0)[\![\hat{\eta} < c_0]\!]$	ℓ_{c_0}, $c_0 \in [0,1]$
$\frac{1}{(1-c)^2 c}$	$\left[2\ln(1-\hat{\eta}) + \frac{\hat{\eta}}{1-\hat{\eta}}\right]$	$\left[\ln\frac{1-\hat{\eta}}{\hat{\eta}} - 1\right]$	—
1	$\hat{\eta}^2/2$	$(1-\hat{\eta})^2/2$	Square
$\frac{1}{(1-c)c}$	$-\ln(1-\hat{\eta})$	$-\ln(\hat{\eta})$	Log
$\frac{1}{(1-c)^2 c^2}$	$\left[\ln((1-\hat{\eta})\hat{\eta}) - \frac{1-2\hat{\eta}}{\hat{\eta}}\right]$	$\left[\ln((1-\hat{\eta})\hat{\eta}) + \frac{1-2\hat{\eta}}{\hat{\eta}}\right]$	—
$\frac{1}{[(1-c)c]^{3/2}}$	$2\sqrt{\frac{\hat{\eta}}{1-\hat{\eta}}}$	$2\sqrt{\frac{1-\hat{\eta}}{\hat{\eta}}}$	Boosting

Table 1: Weight functions and associated partial losses.

Theorem 7 (Reid and Williamson) *Given a weight function $w : [0,1] \to [0,\infty)$, let $W(t) = \int^t w(c)\,dc$ and $\overline{W}(t) = \int^t W(c)\,dc$. Then the loss ℓ_w defined by*

$$\ell_w(y,\hat{\eta}) = -\overline{W}(\hat{\eta}) - (y-\hat{\eta})W(\hat{\eta})$$

is a proper loss. Additionally, if $\overline{W}(0)$ and $\overline{W}(1)$ are both finite then

$$\ell_w(y,\hat{\eta}) + (\overline{W}(1) - \overline{W}(0))y + \overline{W}(0)$$

is a fair, proper loss.

Observe that if w and v are weight functions which differ on a set of measure zero then they will lead to the same loss. A simple corollary to Theorem 6 is that the partial losses are given by

$$\ell_1(\hat{\eta}) = \int_{\hat{\eta}}^1 (1-c)w(c)dc \;\text{ and }\; \ell_{-1}(\hat{\eta}) = \int_0^{\hat{\eta}} cw(c)dc. \tag{5}$$

A similar[8] integral representation of the partial losses can also be found in Shuford et al. (1966, Theorem 2) and Staël von Holstein (1970).

3.3 Symmetric Losses

We will say a loss is *symmetric* if $\ell_1(\hat{\eta}) = \ell_{-1}(1-\hat{\eta})$ for all $\hat{\eta} \in [0,1]$. We say a weight function for a proper loss or the conditional Bayes risk is *symmetric* if $w(c) = w(1-c)$ or $\underline{L}(c) = \underline{L}(1-c)$ for all $c \in [0,1]$. Perhaps unsurprisingly, an immediate consequence of Theorem 1 is that these two notions are identical.[9]

8. The weight function h in Theorem 2 of Shuford et al. (1966) is related to the w here by $h(c) = (1-c)w(c)$.
9. The relationship between a symmetric $\underline{L}$ and symmetric behaviour of the loss has been previously recognised by Masnadi-Shirazi and Vasconcelos (2009).

Corollary 8 *A proper loss is symmetric if and only if its weight function is symmetric.*

Proof If ℓ is symmetric, then $\ell'_1(\hat{\eta}) = -\ell'_{-1}(1-\hat{\eta})$ and so Equation (2) implies $w(1-\hat{\eta}) = \frac{\ell'_{-1}(1-\hat{\eta})}{1-\hat{\eta}} = \frac{-\ell'_1(\hat{\eta})}{1-\hat{\eta}} = w(\hat{\eta})$. Conversely, the symmetry of w applied to Equation (5) establishes the symmetry of ℓ. ■

Requiring a loss to be proper and symmetric constrains the partial losses significantly. Properness alone completely specifies one partial loss from the other. Now suppose in addition that ℓ is symmetric. Combining $\ell_1(\hat{\eta}) = \ell_{-1}(1-\hat{\eta})$ with (2) implies

$$\ell'_{-1}(1-\hat{\eta}) = \frac{1-\hat{\eta}}{\hat{\eta}}\ell'_{-1}(\hat{\eta}). \tag{6}$$

This shows that ℓ_{-1} is completely determined by $\ell_{-1}(\hat{\eta})$ for $\hat{\eta} \in [0, \frac{1}{2}]$ (or $\hat{\eta} \in [\frac{1}{2}, 1]$). Thus in order to specify a symmetric proper loss, one needs to only specify one of the partial losses on one half of the interval $[0,1]$. Assuming ℓ_{-1} is continuous at $\frac{1}{2}$ (or equivalently that w has no atoms at $\frac{1}{2}$), by integrating both sides of (6) we can derive an explicit formula for the other half of ℓ_{-1} in terms of that which is specified:

$$\ell_{-1}(\hat{\eta}) = \ell_{-1}(\tfrac{1}{2}) + \int_{\frac{1}{2}}^{\hat{\eta}} \frac{x}{1-x}\ell'_{-1}(1-x)dx, \tag{7}$$

which works for determining ℓ_{-1} on either $[0, \frac{1}{2}]$ or $[\frac{1}{2}, 1]$ when ℓ_{-1} is specified on $[\frac{1}{2}, 1]$ or $[0, \frac{1}{2}]$ respectively (recalling the usual convention that $\int_a^b = -\int_b^a$). We have thus shown:

Theorem 9 *If a loss is proper and symmetric, then it is completely determined by specifying one of the partial losses on half the unit interval (either* $[0, \frac{1}{2}]$ *or* $[\frac{1}{2}, 0]$*) and using (6) and (7).*

We demonstrate (7) with four examples. Suppose that $\ell_{-1}(\hat{\eta}) = \frac{1}{1-\hat{\eta}}$ for $\hat{\eta} \in [0, \frac{1}{2}]$. Then one can readily determine the complete partial loss to be

$$\ell_{-1}(\hat{\eta}) = \frac{[\![\hat{\eta} \leq \frac{1}{2}]\!]}{1-\hat{\eta}} + [\![\hat{\eta} > \tfrac{1}{2}]\!]\left(2 + \log\frac{\hat{\eta}}{1-\hat{\eta}}\right).$$

Suppose instead that $\ell_{-1}(\hat{\eta}) = \frac{1}{1-\hat{\eta}}$ for $\hat{\eta} \in [\frac{1}{2}, 1]$. In that case we obtain

$$\ell_{-1}(\hat{\eta}) = [\![\hat{\eta} \leq \frac{1}{2}]\!]\left(2 + \log\frac{\hat{\eta}}{1-\hat{\eta}}\right) + \frac{[\![\hat{\eta} \geq \frac{1}{2}]\!]}{1-\hat{\eta}}.$$

Suppose $\ell_{-1}(\hat{\eta}) = \frac{1}{(1-\hat{\eta})^2}$ for $\hat{\eta} \in [0, \frac{1}{2}]$. Then one can determine that

$$\ell_{-1}(\hat{\eta}) = \frac{[\![\hat{\eta} < \frac{1}{2}]\!]}{(1-\hat{\eta})^2} + \frac{[\![\hat{\eta} \geq \frac{1}{2}]\!](4 + 2(2\hat{\eta} + \hat{\eta}\log\hat{\eta} - \hat{\eta}\log(1-\hat{\eta}) - 1))}{\hat{\eta}}.$$

Finally consider specifying that $\ell_{-1}(\hat{\eta}) = \hat{\eta}$ for $\hat{\eta} \in [0, \frac{1}{2}]$. In this case we obtain that

$$\ell_{-1}(\hat{\eta}) = [\![\hat{\eta} \leq \tfrac{1}{2}]\!]\hat{\eta} + [\![\hat{\eta} \geq \tfrac{1}{2}]\!](1 - \log 2 - \hat{\eta} - \log(1-\hat{\eta})).$$

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

$$\ell^{\psi}(y,v) := \ell(y,\psi^{-1}(v)). \tag{8}$$

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" (and ψ is of course the forward 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 $\ell(y,\psi^{-1}(\hat{h}))$ is convex in $\hat{h}$ (which implies it is convex in $\boldsymbol{\alpha}$ when $\hat{h}_{\boldsymbol{\alpha}}$ is linear in $\boldsymbol{\alpha}$). The idea of a link function is not as well known as it should be and is thus reinvented—see for example Granger and Machina (2006).

Often one will choose the loss first (tailoring its properties by the weighting given according to $w(c)$), and *then* choose the link somewhat arbitrarily to map the hypotheses appropriately. An interesting alternative perspective arises in the literature on "elicitability". Lambert et al. (2008)[10] provide a general characterisation of proper scoring rules (i.e., losses) for general *properties* of distributions, that is, continuous and locally non-constant functions Γ which assign a real value to each distribution over a finite sample space. In the binary case, these properties provide another interpretation of links that is complementary to the usual one that treats the inverse link ψ^{-1} as a way of interpreting scores as class probabilities.

To see this, we first identify distributions over $\{-1,1\}$ with the probability η of observing 1. In this case properties are continuous, locally non-constant maps $\Gamma : [0,1] \to \mathbb{R}$. When a link function ψ is continuous it can therefore be interpreted as a property since its assumed invertibility implies it is locally non-constant. A property Γ is said to be *elicitable* whenever there exists a strictly proper loss ℓ for it so that the composite loss ℓ^{Γ} satisfies for all $\hat{\eta} \neq \eta$

$$L^{\Gamma}(\eta,\hat{\eta}) := \mathbb{E}_{Y\sim\eta}[\ell^{\Gamma}(Y,\hat{\eta})] > L^{\Gamma}(\eta,\eta).$$

Theorem 1 of Lambert et al. (2008) shows that Γ is elicitable if and only if $\Gamma^{-1}(r)$ is convex for all $r \in \text{range}(\Gamma)$. This immediately gives us a characterisation of "proper" link functions: those that are both continuous and have convex level sets in $[0,1]$—they are the non-decreasing continuous functions. Thus in Lambert's perspective, one chooses a "property" first (i.e., the invertible link) and *then* chooses the proper loss.

4.1 Proper Composite Losses

We will call a composite loss ℓ^{ψ} (8) a *proper composite loss* if ℓ in (8) 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.

10. See also Gneiting (2009).

Theorem 10 *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 $\hat{\eta} \in (0,1)$*

$$\frac{-\lambda_1'(\psi(\hat{\eta}))}{1-\hat{\eta}} = \frac{\lambda_{-1}'(\psi(\hat{\eta}))}{\hat{\eta}} = \frac{w(\hat{\eta})}{\psi'(\hat{\eta})} =: \rho(\hat{\eta}), \tag{9}$$

where equality is interpreted in the distributional sense. 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.

Corollary 11 *Suppose ℓ^ψ is a proper composite loss with conditional risk denoted L^ψ. Then*

$$\frac{\partial}{\partial v} L^\psi(\eta, v) = (\psi^{-1}(v) - \eta)\rho(\psi^{-1}(v)). \tag{10}$$

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 10 shows how properness is characterised by a particular relationship between the choice of link function and the choice of partial composite losses.

Corollary 12 *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*

$$\psi^{-1}(v) = \frac{\lambda_{-1}'(v)}{\lambda_{-1}'(v) - \lambda_1'(v)}, \quad \forall v \in \mathcal{V}. \tag{11}$$

Proof Substituting $\hat{\eta} = \psi^{-1}(v)$ into (9) 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 10 shows that the partial losses are completely determined once the weight function w and ψ (up to an additive constant) is fixed. Corollary 12 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 11 gives the single link which will guarantee the overall loss is proper.

We see then that Corollary 12 provides us with a way of constructing a *reference link* for arbitrary composite losses specified by their partial losses. The reference link can be seen to satisfy

$$\psi(\eta) = \operatorname*{arg\,min}_{v \in \mathbb{R}} L^\psi(\eta, v)$$

for $\eta \in (0,1)$ and thus *calibrates* a given composite loss in the sense of Cohen and Goldszmidt (2004).

Finally, we make a note of an analogue of Corollary 5 for composite losses. It shows that the regret for an arbitrary composite loss is related to a Bregman divergence via its link.

Corollary 13 *Let ℓ^ψ be a proper composite loss with invertible link. Then for all $\eta, \hat{\eta} \in (0,1)$,*

$$\Delta L^\psi(\eta, v) = D_{-\underline{L}}(\eta, \psi^{-1}(v)). \tag{12}$$

This corollary generalises the results due to Zhang (2004b) and Masnadi-Shirazi and Vasconcelos (2009) who considered only margin losses respectively without and with links.

4.2 Derivatives of Composite Losses

We now briefly consider an application of the parametrisation of proper losses as a weight function and link. In order to implement Stochastic Gradient Descent (SGD) algorithms one needs to compute the derivative of the loss with respect to predictions $v \in \mathbb{R}$. Letting $\hat{\eta}(v) = \psi^{-1}(v)$ be the probability estimate associated with the prediction v, we can use (10) when $\eta \in \{0,1\}$ to obtain the update rules for positive and negative examples:

$$\begin{aligned} \frac{\partial}{\partial v}\ell_1^\psi(v) &= (\hat{\eta}(v) - 1)\,\rho(\hat{\eta}(v)), \\ \frac{\partial}{\partial v}\ell_{-1}^\psi(v) &= \hat{\eta}(v)\,\rho(\hat{\eta}(v)). \end{aligned}$$

Given an arbitrary weight function w (which defines a proper loss via Corollary 2 and Theorem 4) and link ψ, the above equations show that one could implement SGD directly parametrised in terms of ρ without needing to explicitly compute the partial losses themselves.

4.3 Margin Losses

The *margin* associated with a real-valued prediction $v \in \mathbb{R}$ and label $y \in \{-1,1\}$ is the product $z = yv$. Any function $\phi\colon \mathbb{R} \to \mathbb{R}^+$ can be used as a *margin loss* by interpreting $\phi(yv)$ as the penalty for predicting v for an instance with label y. Margin losses are inherently symmetric since $yv = (-y)(-v)$ and so the penalty $\phi(yv)$ given for predicting v when the label is y is necessarily the same as the penalty for predicting $-v$ when the label is $-y$. Margin losses have attracted a lot of attention (Bartlett et al., 2000) because of their central role in Support Vector Machines (Cortes and Vapnik, 1995). In this section we explore the relationship between these margin losses and the more general class of composite losses and, in particular, symmetric composite losses.

Recall that a general composite loss is of the form $\ell^\psi(y,v) = \ell(y, \psi^{-1}(v))$ for a loss $\ell\colon \mathcal{Y} \times [0,1] \to [0,\infty)$ and an invertible link $\psi\colon \mathbb{R} \to [0,1]$. We would like to understand when margin losses are suitable for probability estimation tasks. As discussed above, proper losses are a natural class of losses over $[0,1]$ for probability estimation so a natural question in this vein is the following: given a margin loss ϕ can we choose a link ψ so that there exists a proper loss ℓ such that $\phi(yv) = \ell^\psi(y,v)$? In this case the proper loss will be $\ell(y,\hat{\eta}) = \phi(y\psi(\hat{\eta}))$.

The following corollary of Theorem 10 gives necessary and sufficient conditions on the choice of link ψ to guarantee when a margin loss ϕ can be expressed as a proper composite loss.

Corollary 14 *Suppose $\phi\colon \mathbb{R} \to \mathbb{R}$ is a differentiable margin loss. Then, $\phi(yv)$ can be expressed as a proper composite loss $\ell^\psi(y,v)$ if and only if the link ψ satisfies*

$$\psi^{-1}(v) = \frac{\phi'(-v)}{\phi'(-v)+\phi'(v)}.$$

Proof Margin losses, by definition, have partial losses $\lambda_y(v) = \phi(yv)$ which means $\lambda'_1(v) = \phi'(v)$ and $\lambda'_{-1}(v) = -\phi'(-v)$. Substituting these into (11) gives the result. ■

This result provides a way of interpreting predictions v as probabilities $\hat{\eta} = \psi^{-1}(v)$ in a consistent manner, for a problem defined by a margin loss. Conversely, it also guarantees that using any other link to interpret predictions as probabilities will be inconsistent.[11] Another immediate implication is that for a margin loss to be considered a proper loss its link function must be *symmetric* in the sense that

$$\psi^{-1}(-v) = \frac{\phi'(v)}{\phi'(v)+\phi'(-v)} = 1 - \frac{\phi'(-v)}{\phi'(-v)+\phi'(v)} = 1 - \psi^{-1}(v),$$

and so, by letting $v = \psi(\hat{\eta})$, we have $\psi(1-\hat{\eta}) = -\psi(\hat{\eta})$ and thus $\psi(\frac{1}{2}) = 0$.

Corollary 14 can also be seen as a simplified and generalised version of the argument by Masnadi-Shirazi and Vasconcelos (2009) that a concave minimal conditional risk function and a symmetric link completely determines a margin loss.[12]

We now consider a couple of specific margin losses and show how they can be associated with a proper loss through the choice of link given in Corollary 14. The exponential loss $\phi(v) = e^{-v}$ gives rise to a proper loss $\ell(y,\hat{\eta}) = \phi(y\psi(\hat{\eta}))$ via the link

$$\psi^{-1}(v) = \frac{-e^v}{-e^v - e^{-v}} = \frac{1}{1+e^{-2v}}$$

which has non-zero denominator. In this case $\psi(\hat{\eta}) = \frac{1}{2}\log\left(\frac{\hat{\eta}}{1-\hat{\eta}}\right)$ is just the logistic link. Now consider the family of margin losses parametrised by $\alpha \in (0,\infty)$

$$\phi_\alpha(v) = \frac{\log(\exp((1-v)\alpha)+1)}{\alpha}.$$

This family of differentiable convex losses approximates the hinge loss as $\alpha \to \infty$ and was studied in the multiclass case by Zhang et al. (2009). Since these are all differentiable functions with $\phi'_\alpha(v) = \frac{-e^{\alpha(1-v)}}{e^{\alpha(1-v)}+1}$, Corollary 14 and a little algebra gives

$$\psi^{-1}(v) = \left[1 + \frac{e^{2\alpha}+e^{\alpha(1-v)}}{e^{2\alpha}+e^{\alpha(1+v)}}\right]^{-1}.$$

Examining this family of inverse links as $\alpha \to 0$ gives some insight into why the hinge loss is a surrogate for classification but not probability estimation. When $\alpha \approx 0$ an estimate $\hat{\eta} = \psi^{-1}(v) \approx \frac{1}{2}$ for all but very large $v \in \mathbb{R}$. That is, in the limit all probability estimates sit infinitesimally to the right or left of $\frac{1}{2}$ depending on the sign of v.

11. Strictly speaking, if the margin loss has "flat spots"—that is, where $\phi'(v) = 0$—then the choice of link may not be unique.
12. Shen (2005, Section 4.4) seems to have been the first to view margin losses from this more general perspective.

5. Classification Calibration and Proper Losses

The notion of properness of a loss designed for class probability estimation is a natural one. If one is only interested in classification (rather than estimating probabilities) a weaker condition suffices. In this section we will relate the weaker condition to properness.

5.1 Classification Calibration for CPE Losses

We begin by giving a definition of classification calibration for CPE losses (i.e., over the unit interval $[0,1]$) and relate it to composite losses via a link.

Definition 15 *We say a CPE loss ℓ is* classification calibrated at $c \in (0,1)$ *and write ℓ is CC_c if the associated conditional risk L satisfies*

$$\forall \eta \neq c,\ \underline{L}(\eta) < \inf_{\hat{\eta}:\ (\hat{\eta}-c)(\eta-c)\leq 0} L(\eta,\hat{\eta}). \tag{13}$$

The expression constraining the infimum ensures that $\hat{\eta}$ is on the opposite side of c to η, or $\hat{\eta} = c$.

The condition $\mathrm{CC}_{\frac{1}{2}}$ is equivalent to what is called "classification calibrated" by Bartlett et al. (2006) and "Fisher consistent for classification problems" by Lin (2002) although their definitions were only for margin losses. One situation where this more general CC_c notion is more appropriate is when the false positive and false negative costs for a classification problem are unequal.

One might suspect that there is a connection between classification calibrated at c and standard Fisher consistency for class probability estimation losses. The following theorem, which captures the intuition behind the "probing" reduction (Langford and Zadrozny, 2005), characterises the situation.

Theorem 16 *A CPE loss ℓ is CC_c for all $c \in (0,1)$ if and only if ℓ is strictly proper.*

Proof The loss ℓ is CC_c for all $c \in (0,1)$ is equivalent to

$$\begin{aligned}
& \forall c \in (0,1),\ \forall \eta \neq c \begin{cases} \underline{L}(\eta) < \inf_{\hat{\eta}\geq c} L(\eta,\hat{\eta}), & \eta < c \\ \underline{L}(\eta) < \inf_{\hat{\eta}\leq c} L(\eta,\hat{\eta}), & \eta > c \end{cases} \\
\Leftrightarrow\ & \forall \eta \in (0,1),\ \forall c \neq \eta \begin{cases} \forall c > \eta,\ \underline{L}(\eta) < \inf_{\hat{\eta}\geq c} L(\eta,\hat{\eta}) \\ \forall c < \eta,\ \underline{L}(\eta) < \inf_{\hat{\eta}\leq c} L(\eta,\hat{\eta}) \end{cases} \\
\Leftrightarrow\ & \forall \eta \in (0,1),\ \begin{cases} \underline{L}(\eta) < \inf_{\hat{\eta}\geq c>\eta} L(\eta,\hat{\eta}) \\ \underline{L}(\eta) < \inf_{\hat{\eta}\leq c<\eta} L(\eta,\hat{\eta}) \end{cases} \\
\Leftrightarrow\ & \forall \eta \in (0,1),\ \underline{L}(\eta) < \inf_{(\hat{\eta}>\eta)\ \text{or}\ (\hat{\eta}<\eta)} L(\eta,\hat{\eta}) \\
\Leftrightarrow\ & \forall \eta \in (0,1),\ \underline{L}(\eta) < \inf_{\hat{\eta}\neq\eta} L(\eta,\hat{\eta})
\end{aligned}$$

which means L is strictly proper. ■

The following theorem is a generalisation of the characterisation of $\mathrm{CC}_{\frac{1}{2}}$ for margin losses via $\phi'(0)$ due to Bartlett et al. (2006).

Theorem 17 *Suppose ℓ is a loss and suppose that ℓ'_1 and ℓ'_{-1} exist everywhere. Then for any $c \in (0,1)$ ℓ is CC_c if and only if*

$$\ell'_{-1}(c) > 0 \ \text{ and } \ \ell'_1(c) < 0 \ \text{ and } \ c\ell'_1(c) + (1-c)\ell'_{-1}(c) = 0. \tag{14}$$

Proof Since ℓ'_1 and ℓ'_{-1} are assumed to exist everywhere

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

exists for all $\hat{\eta}$. L is CC_c is equivalent to

$$\begin{aligned}
&\frac{\partial}{\partial\hat{\eta}}L(\eta,\hat{\eta})\bigg|_{\hat{\eta}=c} \begin{cases} >0, & \eta<c<\hat{\eta} \\ <0, & \hat{\eta}<c<\eta \end{cases} \\
\Leftrightarrow\quad &\begin{cases} \forall\eta<c, & \eta\ell'_1(c)+(1-\eta)\ell'_{-1}(c)>0 \\ \forall\eta>c, & \eta\ell'_1(c)+(1-\eta)\ell'_{-1}(c)<0 \end{cases} \qquad (15) \\
\Leftrightarrow\quad &\begin{array}{l} c\ell'_1(c)+(1-c)\ell'_{-1}(c)=0 \\ \text{and } \ell'_{-1}(c)>0 \text{ and } \ell'_1(c)<0, \end{array} \qquad (16)
\end{aligned}$$

where we have used the fact that (15) with $\eta=0$ and $\eta=1$ respectively substituted implies $\ell'_{-1}(c)>0$ and $\ell'_1(c)<0$. ■

If ℓ is proper, then by evaluating (3) at $\eta=0$ and $\eta=1$ we obtain $\ell'_1(\hat{\eta}) = -w(\hat{\eta})(1-\hat{\eta})$ and $\ell'_{-1}(\hat{\eta}) = w(\hat{\eta})\hat{\eta}$. Thus (16) implies $-w(c)(1-c)<0$ and $w(c)c>0$ which holds if and only if $w(c)\neq 0$. We have thus shown the following corollary.

Corollary 18 *If ℓ is proper with weight w, then for any $c\in(0,1)$,*

$$w(c)\neq 0 \;\Leftrightarrow\; \ell \text{ is } \mathrm{CC}_c.$$

The simple form of the weight function for the cost-sensitive misclassification loss ℓ_{c_0} ($w(c) = \delta(c-c_0)$) gives the following corollary (confer Bartlett et al., 2006):

Corollary 19 *ℓ_{c_0} is CC_c if and only if $c_0 = c$.*

5.2 Calibration for Composite Losses

The translation of the above results to general proper composite losses with invertible differentiable link ψ is straight forward. Condition (13) becomes

$$\forall\eta\neq c,\;\; \underline{L}^{\psi}(\eta) < \inf_{v:\,(\psi^{-1}(v)-c)(\eta-c)\leq 0} L^{\psi}(\eta,\psi^{-1}(v)).$$

Theorem 16 then immediately gives:

Corollary 20 *A composite loss $\ell^{\psi}(\cdot,\cdot) = \ell(\cdot,\psi^{-1}(\cdot))$ with invertible and differentiable link ψ is CC_c for all $c\in(0,1)$ if and only if the associated proper loss ℓ is strictly proper.*

Theorem 17 immediately gives:

Corollary 21 *Suppose ℓ^{ψ} is as in Corollary 20 and that the partial losses ℓ_1 and ℓ_{-1} of the associated proper loss ℓ are differentiable. Then for any $c\in(0,1)$, ℓ^{ψ} is CC_c if and only if (14) holds.*

It can be shown that in the special case of margin losses L_{ϕ}, which satisfy the conditions of Corollary 14 such that they are proper composite losses, Corollary 21 leads to the condition $\phi'(0)<0$ which is the same as obtained by Bartlett et al. (2006).

6. Convexity of Composite Losses

We have seen that composite losses are defined by the proper loss ℓ and the link ψ. We have further seen from (14) that it is natural to parametrise 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 is Theorem 29 answers these questions by characterising the convexity of composite losses in terms of (w,ψ') or ρ.

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

Lemma 22 *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 as required. ■

The following theorem generalises the corollary on page 12 of Buja et al. (2005) to arbitrary composite losses with invertible links. It has less practical value than the previous lemma since, in general, sums of quasi-convex functions are not necessarily quasi-convex (a function f is quasi-convex if the set $\{x\colon f(x) \geq \alpha\}$ is convex for all $\alpha \in \mathbb{R}$). Thus, assuming properness of the loss ℓ does not guarantee its empirical risk $\hat{\mathbb{L}}(\cdot,S)$ will not have local minima.

Theorem 23 *If $\ell^\psi(y,v) = \ell(y,\psi^{-1}(v))$ is a composite loss where ℓ is proper and ψ is invertible and differentiable then $L^\psi(\eta,v)$ is quasi-convex in v for all $\eta \in [0,1]$.*

Proof Since ℓ is proper we know by Corollary 11 that the conditional Bayes risk satisfies

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

Since ψ is invertible and $\rho \geq 0$ we see that $\frac{\partial}{\partial v} L^\psi(\eta,v)$ only changes sign at $\eta = \psi^{-1}(v)$ and so L^ψ is quasi-convex as required. ■

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

Theorem 24 *Let $\ell^\psi(y,v)$ be a composite loss comprising an invertible link ψ 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*

$$-\frac{1}{x} \;\leq\; \frac{w'(x)}{w(x)} - \frac{\psi''(x)}{\psi'(x)} \;\leq\; \frac{1}{1-x}, \quad \forall x \in (0,1). \tag{17}$$

This theorem suggests a very natural parametrisation of composite losses is via (w, ψ'). Observe that $w, \psi' \colon [0,1] \to \mathbb{R}^+$. (But also see the comment following Theorem 29.)

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 by substituting $q = \psi^{-1}$ into (10) we have

$$\frac{\partial}{\partial v} L^\psi(\eta, v) \quad = \quad w(q(v))q'(v)[q(v) - \eta]. \tag{18}$$

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 (18) the above condition is equivalent to

$$[w(q(v))q'(v)]'(q(v) - [\![y=1]\!]) + w(q(v))q'(v)q'(v) \ \geq \ 0, \quad \forall v \in \mathbb{R}, \tag{19}$$

where

$$[w(q(v))q'(v)]' := \frac{\partial}{\partial v} w(q(v))q'(v).$$

Inequality (19) is equivalent to (Buja et al., 2005, Equation 39). By further manipulations, we can simplify (19) considerably.

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

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

which we shall rewrite as the pair of inequalities

$$w(q(v))(q'(v))^2 \ \geq \ -q(v)[w(q(v))q'(v)]', \qquad \forall v \in \mathbb{R}, \tag{20}$$

$$w(q(v))(q'(v))^2 \ \geq \ (1-q(v))[w(q(v))q'(v)]', \quad \forall v \in \mathbb{R}. \tag{21}$$

Observe that if $q(\cdot) = 0$ (resp. $1 - q(\cdot) = 0$) then (20) (resp. (21)) 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

$$\{x \colon q(x) \neq 0 \ \text{ and } \ (1-q(x)) \neq 0\} = q^{-1}((0,1)) = \psi((0,1)).$$

Combining (20) and (21) we obtain the equivalent condition

$$\frac{(q'(v))^2}{1-q(v)} \ \geq \ \frac{[w(q(v))q'(v)]'}{w(q(v))} \ \geq \ \frac{-(q'(v))^2}{q(v)}, \quad \forall v \in \psi((0,1)), \tag{22}$$

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 (22) is equivalent to

$$\frac{(q'(v))^2}{1-q(v)} \;\geq\; \frac{w'(q(v))(q'(v))^2 + w(q(v))q''(v)}{w(q(v))} \;\geq\; \frac{-(q'(v))^2}{q(v)}, \quad \forall v \in \psi((0,1)) \tag{23}$$

Now divide all sides of (23) by $(q'(\cdot))^2$ (which is permissible by assumption). This gives the equivalent condition

$$\frac{1}{1-q(v)} \;\geq\; \frac{w'(q(v))}{w(q(v))} + \frac{q''(v)}{(q'(v))^2} \;\geq\; \frac{-1}{q(v)}, \quad \forall v \in \psi((0,1)). \tag{24}$$

Let $x = q(v)$ and so $v = q^{-1}(x) = \psi(x)$. Then (24) is equivalent to

$$\frac{1}{1-x} \;\geq\; \frac{w'(x)}{w(x)} + \frac{q''(\psi(x))}{(q'(\psi(x)))^2} \;\geq\; \frac{-1}{x}, \quad \forall x \in (0,1). \tag{25}$$

Now $\frac{1}{q'(\psi(x))} = \frac{1}{q'(q^{-1}(x))} = (q^{-1})'(x) = \psi'(x)$. Thus (25) is equivalent to

$$\frac{1}{1-x} \;\geq\; \frac{w'(x)}{w(x)} + \Phi_\psi(x) \;\geq\; \frac{-1}{x}, \quad \forall x \in (0,1), \tag{26}$$

where

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

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

$$\text{(26) is true} \;\Leftrightarrow\; v \mapsto L^\psi(y,v) \text{ is convex for } y \in \{-1,1\}$$

where the right hand side is equivalent to the assertion in the theorem by Lemma 22.

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

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

Thus by substitution

$$\begin{aligned}
\Phi_\psi(\cdot) &= \frac{-1}{\left(\psi'(\psi^{-1}(\psi(\cdot)))\right)^3}\psi''(\psi(\psi^{-1}(\cdot)))\left(\psi'(\cdot)\right)^2 \\
&= \frac{-1}{\left(\psi'(\cdot)\right)^3}\psi''(\cdot)\left(\psi'(\cdot)\right)^2 \\
&= -\frac{\psi''(\cdot)}{\psi'(\cdot)}.
\end{aligned} \tag{27}$$

Substituting the simpler expression (27) for Φ_ψ into (26) completes the proof. ■

Lemma 25 *If q is affine then $\Phi_\psi = 0$.*

Proof Using (27), this is immediate since in this case $\psi''(\cdot) = 0$. ■

Corollary 26 *Composite losses with a linear link (including as a special case the identity link) are convex if and only if*

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

6.1 Canonical Links

Buja et al. (2005) introduced the notion of a *canonical link* defined by $\psi'(v) = w(v)$. The canonical link corresponds to the notion of "matching loss" as developed by Helmbold et al. (1999) and Kivinen and Warmuth (2001). Note that choice of canonical link implies $\rho(c) = w(c)/\psi'(c) = 1$.

Lemma 27 *Suppose ℓ is a proper loss with weight function w and ψ is the corresponding canonical link, then*

$$\Phi_\psi(x) = -\frac{w'(x)}{w(x)}. \tag{28}$$

Proof Substitute $\psi' = w$ into (27). ■

This lemma gives an immediate proof of the following result due to Buja et al. (2005).

Theorem 28 *A composite loss comprising a proper loss with weight function w combined with its canonical link is always convex.*

Proof Substitute (28) into (17) to obtain

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

which holds for any w. ■

An alternative view of canonical links is given in Appendix B.

6.2 A Simpler Characterisation of Convex Composite Losses

The following theorem prrovides a simpler characterisation of the convexity of composite losses. Noting that loss functions can be multiplied by a scalar without affecting what a learning algorithm will do, it is convenient to normalise them. If w satisfies (17) then so does αw for all $\alpha \in (0,\infty)$. Thus without loss of generality we will normalise w such that $w(\frac{1}{2}) = 1$. 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, for example, Buja et al. (2005).

Theorem 29 *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 is convex if and only if*

$$\frac{\psi'(x)}{x} \;\begin{matrix}\leqq\\ \geqq\end{matrix}\; 2\psi'(\tfrac{1}{2})\,w(x) \;\begin{matrix}\leqq\\ \geqq\end{matrix}\; \frac{\psi'(x)}{1-x}, \quad \forall x \in (0,1), \tag{29}$$

where $\begin{matrix}\leqq\\ \geqq\end{matrix}$ denotes $\leq$ for $x \geq \frac{1}{2}$ and denotes $\geq$ for $x \leq \frac{1}{2}$.

Observe that the condition (29) is equivalent to

$$\frac{1}{2\psi'(\frac{1}{2})x} \;\begin{matrix}\leqq\\ \geqq\end{matrix}\; \rho(x) \;\begin{matrix}\leqq\\ \geqq\end{matrix}\; \frac{1}{2\psi'(\frac{1}{2})(1-x)}, \quad \forall x \in (0,1),$$

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

Proof Observing that $\frac{w'(x)}{w(x)} = (\log w)'(x)$ we let $g(x) := \log w(x)$. Observe that $g(v) = \int_{\frac{1}{2}}^{v} g'(x)dx + g(\frac{1}{2})$ and $g(\frac{1}{2}) = \log w(\frac{1}{2}) = 0$. From Theorem 24, we know that ℓ is convex iff (17) holds. Using the newly introduced notation, this is equivalent to

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

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

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

Similarly, 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

$$-\ln v - \ln 2 - \int_{\frac{1}{2}}^{v} \Phi_\psi(x)dx \;\begin{matrix}\leqq\\ \geqq\end{matrix}\; g(v) \;\begin{matrix}\leqq\\ \geqq\end{matrix}\; -\ln 2 - \ln(1-v) - \int_{\frac{1}{2}}^{v} \Phi_\psi(x)dx.$$

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

$$\frac{1}{2v}\exp\left(-\int_{\frac{1}{2}}^{v} \Phi_\psi(x)dx\right) \;\begin{matrix}\leqq\\ \geqq\end{matrix}\; w(v) \;\begin{matrix}\leqq\\ \geqq\end{matrix}\; \frac{1}{2(1-v)}\exp\left(-\int_{\frac{1}{2}}^{v} \Phi_\psi(x)dx\right). \tag{30}$$

Now

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

and so

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

Substituting into (30) completes the proof. ■

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

$$\frac{1}{2x} \;\begin{matrix}\leqq\\ \geqq\end{matrix}\; w(x) \;\begin{matrix}\leqq\\ \geqq\end{matrix}\; \frac{1}{2(1-x)}, \quad \forall x \in (0,1), \tag{31}$$

which are illustrated as the shaded region in Figure 2. Observe that the (normalised) weight function for squared loss is $w(c) = 1$ which is indeed within the shaded region as one would expect.

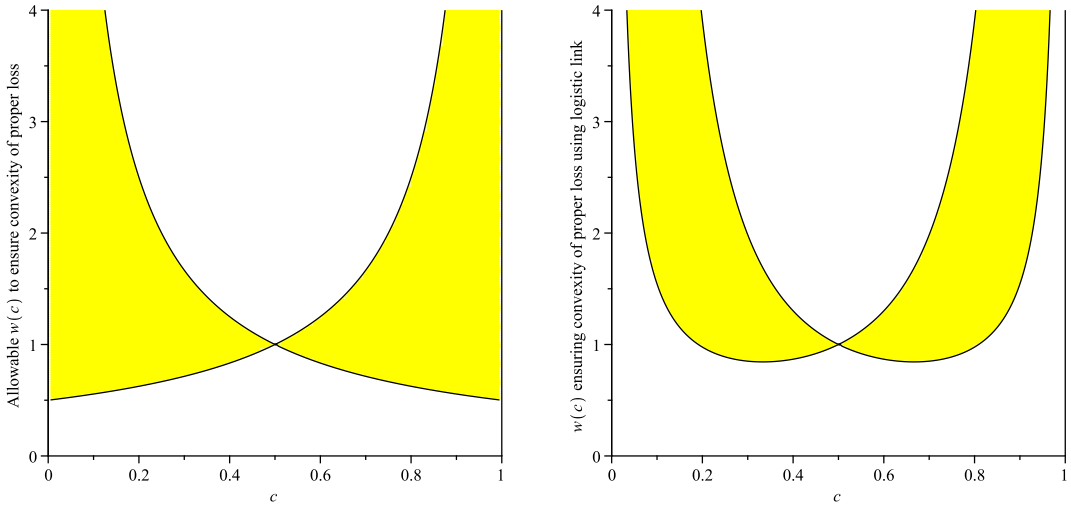

Figure 2: Allowable normalised weight functions to ensure convexity of composite loss functions with identity link (left) and logistic link (right).

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 the constraints on the weight function w to ensure convexity of the composite loss are

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

This is shown graphically in Figure 2. One can compute similar regions for any link. Two other examples are the Complementary Log-Log link $\psi^{\text{CLL}}(x) = \log(-\log(1-x))$ (confer McCullagh and Nelder, 1989), the "square link" $\psi^{\text{sq}}(x) = x^2$ and the "cosine link" $\psi^{\cos}(x) = 1-\cos(\pi x)$. All of these are illustrated in Figure 3. The reason for considering these last two rather unusual links is to illustrate the following fact. Observing that the allowable region in Figure 2 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

$$\lim_{c \searrow 0} w(c) = \lim_{c \nearrow 1} w(c) = 0 \tag{32}$$

is convex. Inspection of (29) 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 can not obtain a convex composite link with the associated proper loss having a weight function satisfying (32). Thus one can not 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 corresponding weight functions.

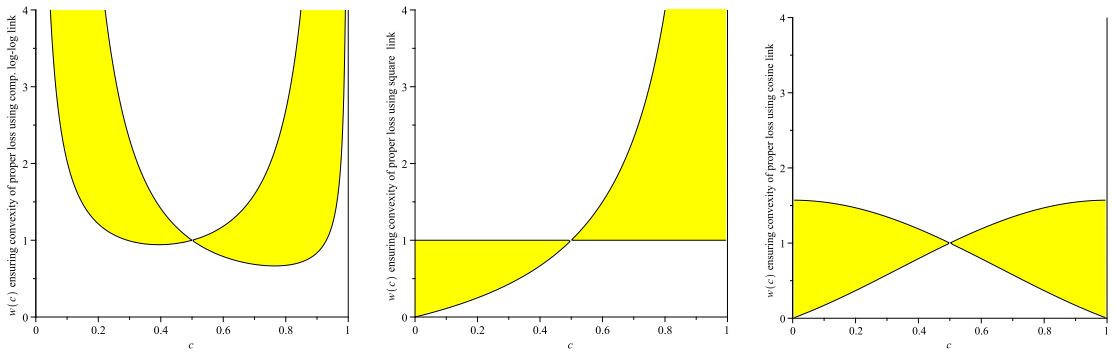

Figure 3: Allowable normalised weight functions to ensure convexity of loss functions with complementary log-log, square and cosine links.

Corollary 30 *If a loss is proper and convex, then it is strictly proper.*

The proof of Corollary 30 makes use of the following special case of the Gronwall style Lemma 1.1.1 of Bainov and Simeonov (1992).

Lemma 31 *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 30) Observe that the RHS of (17) implies

$$w'(v) \leq \frac{w(v)}{1-v},\ v \geq 0.$$

Suppose $w(0) = 0$. Then $v_0 = 0$ and the setting $\alpha = 0$ the lemma implies

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

Thus 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 $s \in [t,1]$. Hence $w(t) > 0$ for all $t \in [0,1]$ and hence by the remark immediately following Theorem 6 ℓ is strictly proper. ■

6.3 Convexity of Bregman Divergences in their Second Argument

Bregman divergences are always convex in the first argument but only sometimes in their second. Corollary 5 and Equation 31 together characterise when the Bregman divergence $D_\phi(\eta, \hat{\eta})$ defined on $(0,1) \times (0,1)$ is convex in $\hat{\eta}$, providing a more direct result that that in Bauschke and Borwein (2001): Setting $\phi = -\underline{L}$ we immediately obtain that $\hat{\eta} \mapsto D_\phi(\eta, \hat{\eta})$ is convex for all $\eta \in (0,1)$ iff (31) holds, where $w(c) = \phi''(c)$.

7. Choosing a Surrogate Loss

A *surrogate* loss function is a loss function which is not exactly what one wishes to minimise but is easier to work with algorithmically. Convex surrogate losses are often used in place of the 0-1 loss which is not convex.

Surrogate losses have garnered increasing interest in the machine learning community (Zhang, 2004b; Bartlett et al., 2006; Steinwart, 2007; Steinwart and Christmann, 2008). Some of the questions considered to date are bounding the regret of a desired loss in terms of a surrogate ("surrogate regret bounds"—see Reid and Williamson, 2009b and references therein), the relationship between the decision theoretic perspective and the elicitability perspective (Masnadi-Shirazi and Vasconcelos, 2009), and efficient algorithms for minimising convex surrogate margin losses (Nock and Nielsen, 2009b,a).

Typically convex surrogates are used because they lead to convex, and thus tractable, optimisation problems. To date, work on surrogate losses has focussed on margin losses which necessarily are symmetric with respect to false positives and false negatives (Buja et al., 2005). In line with the rest of this paper, our treatment will not be so restricted.

The aim here is put forward some plausible definitions of what it might mean to select a "best" surrogate from a class of losses—for example, the class of proper, convex composite losses. We make use of the weight function perspective and the convexity results given in the previous section to investigate some new definitions for "best" surrogate and put forward some conjectures regarding them.

7.1 The "Best" Surrogate Loss

There are many choices of surrogate loss one can choose. A natural question is thus "which is best?". In order to do this we need to first define how we are evaluating losses as surrogates. To do this we require notation to describe the set of minimisers of the conditional and full risk associated with a loss. Given a loss $\ell\colon \{-1,1\}\times\mathcal{V}\to\mathbb{R}$ its *conditional minimisers at* $\eta\in[0,1]$ is the set

$$H(\ell,\eta) := \{v\in\mathcal{V}\colon L(\eta,v) = \underline{L}(\eta)\}. \tag{33}$$

Given a set of hypotheses $\mathcal{H}\subseteq\mathcal{V}^{\mathcal{X}}$, the (constrained) Bayes optimal risk is

$$\underline{\mathbb{L}}_{\mathcal{H}} := \inf_{h\in\mathcal{H}} \mathbb{L}(h,\mathbb{P}).$$

The *(full) minimisers over* $\mathcal{H}$ *for* $\mathbb{P}$ is the set

$$\mathcal{H}(\ell,\mathbb{P}) := \{h\in\mathcal{H}\colon \mathbb{L}(h) = \underline{\mathbb{L}}_{\mathcal{H}}\},$$

where $\mathcal{H}\subseteq\mathcal{V}^{\mathcal{X}}$ is some restricted set of functions and $\mathbb{L}(h) := \mathbb{E}_{(\mathsf{X},\mathsf{Y})\sim\mathbb{P}}[\ell(\mathsf{Y},h(\mathsf{X}))]$ and the expectation is with respect to $\mathbb{P}$. Given a *reference loss* ℓ_{ref}, we will say the ℓ_{ref}*-surrogate penalty* of a loss ℓ over the function class $\mathcal{H}$ on a problem (η,M) (or equivalently $\mathbb{P}$) is

$$S_{\ell_{\text{ref}}}(\ell,\eta,M) = S_{\ell_{\text{ref}}}(\ell,\mathbb{P}) := \inf_{h\in\mathcal{H}(\ell,\mathbb{P})} \mathbb{L}_{\text{ref}}(h),$$

where it is important to remember that $\mathbb{L}$ is with respect to $\mathbb{P}$. That is, $S_{\ell_{\text{ref}}}(\ell,\mathbb{P})$ is the minimum ℓ_{ref} risk obtainable by a function in $\mathcal{H}$ that minimises the ℓ risk.

Given a fixed experiment $\mathbb{P}$, if $\mathcal{L}$ is a class of losses then the *best surrogate losses in* $\mathcal{L}$ for the reference loss ℓ_{ref} are those that minimise the ℓ_{ref}-surrogate penalty. This definition is motivated by the manner in which surrogate losses are used—one minimizes $\mathbb{L}(h)$ over h to obtain the minimiser h^* and one hopes that $\mathbb{L}_{\text{ref}}(h^*)$ is small. Clearly, if the class of losses contains the reference loss (i.e., $\ell_{\text{ref}} \in \mathcal{L}$) then ℓ_{ref} will be a best surrogate loss. Therefore, the question of best surrogate loss is only interesting when $\ell_{\text{ref}} \notin \mathcal{L}$. One particular case we will consider is when the reference loss is the 0-1 loss and the class of surrogates $\mathcal{L}$ is the set of convex proper losses. Since 0-1 loss is not convex the question of which surrogate is best is non-trivial.

It would be nice if one could reason about the "best" surrogate loss using the conditional perspective (that is working with L instead of $\mathbb{L}$) and in a manner independent of $\mathcal{H}$. It is simple to see why this can not be done. Since all the losses we consider are proper, the minimiser over $\hat{\eta}$ of $L(\eta, \hat{\eta})$ is η. Thus any proper loss would lead to the same $\hat{\eta} \in [0,1]$. It is only the introduction of the restricted class of hypotheses $\mathcal{H}$ that prevents this reasoning being applied for $\mathbb{L}$: restrictions on $h \in \mathcal{H}$ prevent $h(x) = \eta(x)$ for all $x \in \mathcal{X}$. We conclude that the problem of best surrogate loss only makes sense when one both takes expectations over X *and* restricts the class of hypotheses h to be drawn from some set $\mathcal{H} \subsetneq [0,1]^{\mathcal{X}}$.

This reasoning accords with that of Nock and Nielsen (2009b,a) who examined which surrogate to use and proposed a data-dependent scheme that tunes surrogates for a problem. They explicitly considered proper losses and said that "minimizing any [lower-bounded, symmetric proper] loss amounts to the *same* ultimate goal" and concluded that "the crux of the choice of the [loss] relies on data-dependent considerations".

We demonstrate the difficulty of finding a universal best surrogate loss in by constructing a simple example. One can construct experiments (η_1, M) and (η_2, M) and proper losses ℓ_1 and ℓ_2 such that

$$S_{\ell_{0-1}}(\ell_1, (\eta_1, M)) > S_{\ell_{0-1}}(\ell_2, (\eta_1, M)) \text{ but } S_{\ell_{0-1}}(\ell_1, (\eta_2, M)) < S_{\ell_{0-1}}(\ell_2, (\eta_2, M)).$$

(The examples we construct have weight functions that "cross-over" each other; the details are in Appendix A.) However, this does not imply there can not exist a particular convex ℓ^* that minorizes all proper losses in this sense. Indeed, we conjecture that, in the sense described above, there is no best proper, convex surrogate loss.

Conjecture 32 *Given a proper, convex loss ℓ there exists a second proper, convex loss $\ell^* \neq \ell$, a hypothesis class $\mathcal{H}$, and an experiment $\mathbb{P}$ such that $S_{\ell_{0-1}}(\ell^*, \mathbb{P}) < S_{\ell_{0-1}}(\ell, \mathbb{P})$ for the class $\mathcal{H}$.*

To prove the above conjecture it would suffice to show that for a fixed hypothesis class and any pair of losses one can construct two experiments such that one loss minorises the other loss on one experiment and *vice versa* on the other experiment.

Supposing the above conjecture is true, one might then ask for a best surrogate loss for some reference loss ℓ_{ref} in a minimax sense. Formally, we would like the loss $\ell^* \in \mathcal{L}$ such that the worst-case penalty for using ℓ^*,

$$\Upsilon_{\mathcal{L}}(\ell^*) := \sup_{\mathbb{P}} \left\{ S_{\ell_{\text{ref}}}(\ell^*, \mathbb{P}) - \inf_{\ell \in \mathcal{L}} S_{\ell_{\text{ref}}}(\ell, \mathbb{P}) \right\}$$

is minimised. That is, $\Upsilon_{\mathcal{L}}(\ell^*) \leq \Upsilon_{\mathcal{L}}(\ell)$ for all $\ell \in \mathcal{L}$.

7.2 The "Minimal" Symmetric Convex Proper Loss

Theorem 29 suggests an answer to the question "What is the proper convex loss closest to the 0-1 loss?" A way of making this question precise follows. Since ℓ is presumed proper, it has a weight function w. Suppose w.l.o.g. that $w(\frac{1}{2}) = 1$. Suppose the link is the identity. The constraints in (17) imply that the weight function that is most similar to that for 0-1 loss meets the constraints. Thus from (31)

$$w^{\text{minimal}}(c) = \frac{1}{2}\left(\frac{1}{c} \wedge \frac{1}{1-c}\right) \tag{34}$$

is the weight for the convex proper loss closest to 0-1 loss in this sense. It is the weight function that forms the lower envelope of the shaded region in the left diagram of Figure 2. Using (5) one can readily compute the corresponding partial losses explicitly

$$\ell_{-1}^{\text{minimal}}(\hat{\eta}) = \frac{1}{2}\left([\![\hat{\eta} < \tfrac{1}{2}]\!](-\hat{\eta} - \ln(1-\hat{\eta})) + [\![\hat{\eta} \geq \tfrac{1}{2}]\!](\hat{\eta} - 1 - \ln(\tfrac{1}{2}))\right) \tag{35}$$

and

$$\ell_{1}^{\text{minimal}}(\hat{\eta}) = \frac{1}{2}\left([\![\hat{\eta} < \tfrac{1}{2}]\!](-\hat{\eta} - \log(\tfrac{1}{2})) + [\![\hat{\eta} \geq \tfrac{1}{2}]\!](\hat{\eta} - 1 - \ln\hat{\eta})\right). \tag{36}$$

Observe that the partial losses are (in part) linear, which is unsurprising as linear functions are on the boundary of the set convex functions. This loss is also best in another more precise (but ultimately unsatisfactory) sense, as we shall now show.

Surrogate regret bounds are theoretical bounds on the regret of a desired loss (say 0-1 loss) in terms of the regret with respect to a surrogate. Reid and Williamson (2009b) have shown the following (we only quote the simpler symmetric case here):

Theorem 33 *Suppose ℓ is a proper loss with corresponding conditional Bayes risk $\underline{L}$ which is symmetric about $\frac{1}{2}$: $\underline{L}(\frac{1}{2} - c) = \underline{L}(\frac{1}{2} + c)$ for $c \in [0, \frac{1}{2}]$. If the regret for the $\ell_{\frac{1}{2}}$ loss $\Delta L_{\frac{1}{2}}(\eta, \hat{\eta}) = \alpha$, then the regret ΔL with respect to ℓ satisfies*

$$\Delta L(\eta, \hat{\eta}) \geq \underline{L}(\tfrac{1}{2}) - \underline{L}(\tfrac{1}{2} + \alpha). \tag{37}$$

The bound in the theorem can be inverted to upper bound $\Delta L_{\frac{1}{2}}$ given an upper bound on $\Delta L(\eta, \hat{\eta})$. Considering all symmetric proper losses normalised such that $w(\frac{1}{2}) = 1$, the right side of (37) is maximised and thus the bound on $\Delta L_{\frac{1}{2}}$ in terms of ΔL is minimised when $\underline{L}(\frac{1}{2} + \alpha)$ is maximised (over all losses normalised as mentioned). But since $w = -\underline{L}''$, that occurs for the pointwise minimiser of w (subject to $w(\frac{1}{2}) = 1$). Since we are interested in convex losses, the minimising w is given by (34). In this case the right hand side of (37) can be explicitly determined to be $(\frac{\alpha}{2} + \frac{1}{4})\log(2\alpha + 1) - \frac{\alpha}{2}$, and the bound can be inverted to obtain the result that if $\Delta L^{\text{minimal}}(\eta, \hat{\eta}) = x$ then

$$\Delta L_{\frac{1}{2}}(\eta, \hat{\eta}) \leq \frac{1}{2}\exp\left(\text{LambertW}\left(\frac{(4x-1)}{\mathrm{e}}\right) + 1\right) - \frac{1}{2} \tag{38}$$

which is plotted in Figure 4.[13]

The above argument does *not* show that the loss given by (35,36) is the *best* surrogate loss. The reason is that the above is optimising a *bound* on the regret, not the *actual* regret; the argument in

13. The LambertW function is the real-valued solution of $x \mapsto W(x)e^{W(x)}$. It is commonly found in solutions to differential equations, has no closed form. Its details are not relevant to this discussion except for computing Figure 4.

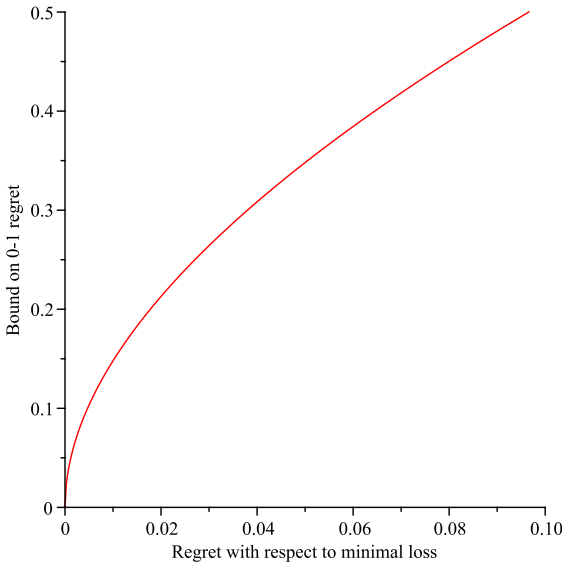

Figure 4: Upper bound on the 0-1 regret in terms of $\Delta L^{\text{minimal}}$ as given by (38).

Appendix A demonstrates there can in general be no universally best surrogate loss (independent of the underlying distribution). Nevertheless it does suggest it is at least worth considering using ℓ^{minimal} as a convex proper surrogate binary loss.

We conjecture that ℓ^{minimal} is somehow special in the class of proper convex losses in some way other than being the pointwise minimiser of weights (and the normalised loss with smallest regret bound with respect to ℓ^{0-1}), but the exact nature of the specialness still eludes us. Perhaps it is optimal in some weaker (minimax) sense. The reason for this suggestion is that it is not hard to show that for reasonable $\mathbb{P}$ there exists $\mathcal{H}$ such that $c \mapsto \mathbb{L}_c(h, \mathbb{P})$ takes on all possible values within the constraints

$$0 \leq \mathbb{L}_c(h, \mathbb{P}) \leq \max(c, 1-c)$$

which follows immediately from the definition of cost-sensitive misclassification loss. Furthermore the example in the appendix below seems to require loss functions whose corresponding weight functions cross over each other and there is no weight function corresponding to a convex proper loss that crosses over w^{minimal}.

8. Conclusions

Composite losses are widely used. As outlined in §1.1, we have characterised a number of aspects of them: their relationship to margin losses, the connection between properness and classification calibration, the constraints symmetry imposes, when composite losses are convex, and natural ways to parametrise them. We have also considered the question of the "best" surrogate loss.

The parametrisation of a composite loss in terms (w, ψ') (or ρ) has advantages over using (ϕ, ψ) or $(\underline{L}, \psi)$. As explained by Masnadi-Shirazi and Vasconcelos (2009), the representation in terms of (ϕ, ψ) is in general not unique. The representation in terms of $\underline{L}$ is harder to intuit: whilst indeed the Bayes risk for squared loss and 0-1 loss are "close" (compare the graph of $c \mapsto c(1-c)$ with that of

$c \mapsto c \wedge (1-c)$), by examining their weight functions they are seen to be very different ($w(c) = 1$ versus $w(c) = 2\delta(c - \frac{1}{2})$). We have also seen that on the basis of Theorem 24, the parametrisation (w, ψ') is perhaps the most natural—there is a pleasing symmetry between the loss and the link as they are in this form both parametrised in terms of non-negative weight functions on $[0,1]$. Recall too that the canonical link sets ψ' equal to w.

The observation suggests an alternate inductive principle known as *surrogate tuning*, which seems to have been first suggested by Nock and Nielsen (2009b).[14] The idea of surrogate tuning is simple: noting that the best surrogate depends on the problem, adapt the surrogate you are using to the problem. In order to do so it is important to have a good parametrisation of the loss. The weight function perspective does just that, especially given Theorem 29. It would be straight forward to develop low dimensional parametrisations of w that satisfy the conditions of this theorem which would thus allow a learning algorithm to explore the space of convex losses. One could (taking due care with the subsequent multiple hypothesis testing problem) regularly *evaluate* the 0-1 loss of the hypotheses so obtained. The observations made in Section 4 regarding stochastic gradient descent algorithms may be of help in this regard.

Acknowledgments

This work was motivated in part by a question due to John Langford. Thanks to Fangfang Lu for discussions and finding several bugs in an earlier version. Thanks to Ingo Steinwart for pointing out the η_α trick. Thanks to Tim van Erven and the anonymous reviewers for comments and corrections. This work was supported by the Australian Research Council and NICTA through Backing Australia's Ability.

Appendix A. Example Showing Incommensurability of Two Proper Surrogate Losses

We consider $\mathcal{X} = [0,1]$ with M being uniform on $\mathcal{X}$, and consider the two problems that are induced by

$$\eta_1(x) = x^2 \quad \text{and} \quad \eta_2(x) = \frac{1}{3} + \frac{x}{3}.$$

We use a simple linear hypothesis class

$$\mathcal{H} := \{h_\alpha(x) := \alpha x \colon \alpha \in [0,1]\},$$

with identity link function and consider the two surrogate proper losses ℓ_1 and ℓ_2 with weight functions

$$w_1(c) = \frac{1}{c}, \qquad w_2(c) = \frac{1}{1-c}.$$

These weight functions correspond to the two curves that construct the left diagram in Figure 2. The corresponding conditional losses can be readily calculated to be

$$\begin{aligned} L_1(\eta, h) &:= \eta(h - 1 - \log(h)) + (1-\eta)h \\ L_2(\eta, h) &:= \eta(1-h) + (1-\eta)(-h - \log(1-h)). \end{aligned}$$

14. Surrogate tuning differs from loss *tailoring* (Hand, 1994; Hand and Vinciotti, 2003; Buja et al., 2005) which involves adapting the loss to what you really think is important.

One can numerically compute the parameters for the constrained Bayes optimal for each problem and for each surrogate loss:

$$\begin{aligned}
\alpha^*_{1,1} &= \arg\min_{\alpha\in[0,1]} \mathbb{L}_1(\eta_1, h_\alpha, M) = 0.66666667 \\
\alpha^*_{2,1} &= \arg\min_{\alpha\in[0,1]} \mathbb{L}_2(\eta_1, h_\alpha, M) = 0.81779259 \\
\alpha^*_{1,2} &= \arg\min_{\alpha\in[0,1]} \mathbb{L}_1(\eta_2, h_\alpha, M) = 1.00000000 \\
\alpha^*_{2,1} &= \arg\min_{\alpha\in[0,1]} \mathbb{L}_2(\eta_2, h_\alpha, M) = 0.77763472.
\end{aligned}$$

Furthermore

$$\begin{aligned}
&\mathbb{L}_{0-1}(\eta_1, h_{\alpha^*_{1,1}}, M) = 0.3580272, \qquad \mathbb{L}_{0-1}(\eta_1, h_{\alpha^*_{2,1}}, M) = 0.3033476, \\
&\mathbb{L}_{0-1}(\eta_2, h_{\alpha^*_{1,2}}, M) = 0.4166666, \qquad \mathbb{L}_{0-1}(\eta_2, h_{\alpha^*_{2,2}}, M) = 0.4207872.
\end{aligned}$$

Thus for problem η_1 the surrogate loss L_2 has a constrained Bayes optimal hypothesis $h_{\alpha^*_{2,1}}$ which has a lower 0-1 risk than the constrained Bayes optimal hypothesis $h_{\alpha^*_{1,1}}$ for the surrogate loss L_1. Thus for problem η_1 surrogate L_2 is better than surrogate L_1. However for problem η_2 the situation is reversed: surrogate L_2 is *worse* than surrogate L_1.

Appendix B. An Alternate View of Canonical Links

This appendix contains an alternate approach to understanding canonical links using convex duality. In doing so we present an improved formulation of a result on the duality of Bregman divergences that may be of independent interest.

The *Legendre-Fenchel* (LF) dual $\phi^\star$ of a function $\phi\colon \mathbb{R} \to \mathbb{R}$ is a function defined by

$$\phi^\star(s^\star) := \sup_{s\in\mathbb{R}} \{\langle s, s^\star\rangle - \phi(s)\}.$$

The LF dual of any function is convex.

When $\phi(s)$ is a function of a real argument s and the derivative $\phi'(s)$ exists, the Legendre-Fenchel conjugate $\phi^\star$ is given by the *Legendre transform* (Rockafellar, 1970; Hiriart-Urruty and Lemaréchal, 2001)

$$\phi^\star(s) = s \cdot (\phi')^{-1}(s) - \phi\left((\phi')^{-1}(s)\right). \tag{39}$$

Thus (writing $\partial f := f'$) $f' = (\partial f^\star)^{-1}$. Thus with w, W, and $\overline{W}$ defined as above,

$$W = (\partial(\overline{W}^\star))^{-1}, \quad W^{-1} = \partial(\overline{W}^\star), \quad \overline{W}^\star = \int W^{-1}. \tag{40}$$

Let w, W, $\overline{W}$ be as in Theorem 7. Denote by L_W the w-weighted conditional loss parametrised by $W = \int w$ and let ΔL_W be the corresponding regret (we can interchange ΔL and D here by (12) since $\psi_L = \mathrm{id}$.

$$D_w(\eta, \hat{\eta}) = \overline{W}(\eta) - \overline{W}(\hat{\eta}) - (\eta - \hat{\eta})W(\hat{\eta}). \tag{41}$$

We now further consider D_w as given by (41). It will be convenient to parametrise D by W instead of w. Note that the standard parametrisation for a Bregman divergence is in terms of the convex function $\overline{W}$. Thus will write $D_{\overline{W}}$, D_W and D_w to all represent (41). The following theorem is known (e.g., Zhang, 2004a) but as will be seen, stating it in terms of D_W provides some advantages.

Theorem 34 *Let w, W, $\overline{W}$ and D_W be as above. Then for all $x,y \in [0,1]$,*

$$D_W(x,y) = D_{W^{-1}}(W(y),W(x)). \tag{42}$$

Proof Using (39) we have

$$\begin{aligned} & \overline{W}^\star(u) = u \cdot W^{-1}(u) - \overline{W}(W^{-1}(u)) \\ \Rightarrow \quad & \overline{W}(W^{-1}(u)) = u \cdot W^{-1}(u) - \overline{W}^\star(u). \end{aligned} \tag{43}$$

Equivalently (using (40))

$$\overline{W}^\star(W(u)) = u \cdot W(u) - \overline{W}(u). \tag{44}$$

Thus substituting and then using (43) we have

$$\begin{aligned} D_W(x,W^{-1}(v)) &= \overline{W}(x) - \overline{W}(W^{-1}(v)) - (x - W^{-1}(v)) \cdot W(W^{-1}(v)) \\ &= \overline{W}(x) + \overline{W}^\star(v) - vW^{-1}(v) - (x - W^{-1}(v)) \cdot v \\ &= \overline{W}(x) + \overline{W}^\star(v) - x \cdot v. \end{aligned} \tag{45}$$

Similarly (this time using (44) we have

$$\begin{aligned} D_{W^{-1}}(v,W(x)) &= \overline{W}^\star(v) - \overline{W}^\star(W(x)) - (v - W(x)) \cdot W^{-1}(W(x)) \\ &= \overline{W}^\star(v) - xW(x) + \overline{W}(x) - v \cdot x + xW(x) \\ &= \overline{W}^\star(v) + \overline{W}(x) - v \cdot x \end{aligned} \tag{46}$$

Comparing (45) and (46) we see that

$$D_W(x,W^{-1}(v)) = D_{W^{-1}}(v,W(x))$$

Let $y = W^{-1}(v)$. Thus subsitituting $v = W(y)$ leads to (42). ■

The weight function corresponding to $D_{W^{-1}}$ is $\frac{\partial}{\partial x} W^{-1}(x) = \frac{1}{w(W^{-1}(x))}$.

Theorem 35 *If the inverse link $\psi^{-1} = W^{-1}$ (and thus $\hat{\eta} = W^{-1}(\hat{h})$) then*

$$\begin{aligned} D_W(\eta,\hat{\eta}) &= D_W(\eta,W^{-1}(\hat{h})) = \overline{W}(\eta) + \overline{W}^\star(\hat{h}) - \eta \cdot \hat{h} \\ L_W(\eta,\hat{\eta}) &= L_W(\eta,W^{-1}(\hat{h})) = \overline{W}^\star(\hat{h}) - \eta \cdot \hat{h} + \eta(\overline{W}(1) + \overline{W}(0)) - \overline{W}(0) \\ \frac{\partial}{\partial \hat{h}} L_W(\eta,W^{-1}(\hat{h})) &= \hat{\eta} - \eta \end{aligned}$$

and furthermore $D_W(\eta,W^{-1}(\hat{h}))$ and $L_W(\eta,W^{-1}(\hat{h}))$ are convex in $\hat{h}$.

Proof The first two expressions follow immediately from (45) and (46) by substitution. The derivative follows from calculation: $\frac{\partial}{\partial \hat{h}} L_W(\eta,W^{-1}(\hat{h})) = \frac{\partial}{\partial \hat{h}} (\overline{W}^\star(\hat{h}) - \eta \cdot \hat{h}) = W^{-1}(\hat{h}) - \eta = \hat{\eta} - \eta$. The convexity follows from the fact that $\overline{W}^\star$ is convex (since it is the LF dual of a convex function $\overline{W}$) and the overall expression is the sum of this and a linear term, and thus convex. ■

Buja et al. (2005) call W the *canonical link*. We have already seen (Theorem 27) that the composite loss constructed using the canonical link is convex.

Appendix C. Convexity and Robustness

In this appendix we show how the characterisation of the convexity of proper losses (Theorem 29) allows one to make general algorithm independent statements about the robustness of convex proper losses to random mis-classification noise.

Long and Servedio (2008) have shown that boosting with convex potential functions (i.e., convex margin losses) is not robust to random class noise.[15] That is, they are susceptible to random class noise. In particular they present a very simple learning task which is "boostable"—can be perfectly solved using a linear combination of base classifiers—but for which, in the presence of any amount of label noise, idealised, early stopping and L_1 regularised boosting algorithms will learn a classifier with only 50% accuracy.

This has led to the recent proposal of boosting algorithms that use non-convex margin losses and experimental evidence suggests that these are more robust to class noise than their convex counterparts. Freund (2009) recently described RobustBoost, which uses a parameterised family of non-convex surrogate losses that approximates the 0-1 loss as the number of boosting iterations increases. Experiments on a variant of the task proposed by Long and Servedio (2008) show that RobustBoost is very insensitive to class noise. Masnadi-Shirazi and Vasconcelos (2009) presented SavageBoost, a boosting algorithm built upon a non-convex margin function. They argued that even when the margin function is non-convex the conditional risk may still be convex. We elucidate this via our characterisation of the convexity of composite losses. Although all these results are suggestive, it is not clear from these results whether the robustness or not is a property of the loss function, the algorithm or a combination. We study that question by considering robustness in an algorithm-independent fashion.

For $\alpha \in (0, \frac{1}{2})$ and $\eta \in [0,1]$ we will define

$$\eta_\alpha := \alpha(1-\eta) + (1-\alpha)\eta$$

as the α-*corrupted version* of η. This captures the idea that instead of drawing a positive label for the point x with probability $\eta(x)$ there is a random class flip with probability α. This might be done on purpose in order to avoid problems with losses (e.g., log loss) that assign infinite penalty to 0 or 1 valued probability predictions. Since η_α is a convex combination of α and $1-\alpha$ it follows that $\eta_\alpha \in [\alpha, 1-\alpha]$. The effect of α-corruption on the conditional risk of a loss can be seen as a transformation of the loss (Steinwart, 2009).

Lemma 36 *If ℓ^Ψ is any composite loss then its conditional risk satisfies*

$$L^\Psi(\eta_\alpha, v) = L^\Psi_\alpha(\eta, v), \quad \eta \in [0,1], \ \ v \in \mathcal{V},$$

where $\ell^\Psi_\alpha(y,v) = (1-\alpha)\ell^\Psi(y,v) + \alpha\ell^\Psi(-y,v)$.

15. We define exactly what we mean by robustness below. The notion that Long and Servedio (2008) examine is akin to that studied for instance by Kearns (1998). There are many other meanings of "robust" which are different to that which we consider. The classical notion of robust statistics (Huber, 1981) is motivated by robustness to contamination of additive observation noise (some heavy-tail noise mixed in with the Gaussian noise often assumed in designing estimators). There are some results about particular machine learning algorithms being robust in that sense (Schölkopf et al., 2000). "Robust" is also used to mean robustness with respect to random attribute noise (Trafalis and Gilbert, 2006), robustness to unknown prior class probabilities (Provost and Fawcett, 2001), or a Huber-style robustness to attribute noise ("outliers") for classification (Fidler et al., 2006). We only study robustness in the sense of random label noise.

Proof By simple algebraic manipulation we have

$$\begin{aligned} L^{\psi}(\eta_\alpha, v) &= (1-\eta_\alpha)\ell^{\psi}(-1,v) + \eta_\alpha \ell^{\psi}(1,v) \\ &= [(1-\alpha)(1-\eta)+\alpha\eta]\ell^{\psi}(-1,v) + [\alpha(1-\eta)+(1-\alpha)\eta]\ell^{\psi}(1,v) \\ &= (1-\eta)[(1-\alpha)\ell^{\psi}(-1,v)+\alpha\ell^{\psi}(1,v)] + \eta[\alpha\ell^{\psi}(-1,v)+(1-\alpha)\ell^{\psi}(1,v)] \\ &= (1-\eta)\ell^{\psi}_\alpha(-1,v) + \eta\ell^{\psi}_\alpha(1,v) \\ &= L^{\psi}_\alpha(\eta, v) \end{aligned}$$

proving the result. ■

In particular, if ℓ is strictly proper then ℓ_α cannot be proper because the minimiser of $L(\eta_\alpha, \cdot)$ is η_α and so $\eta_\alpha \neq \eta$ must also be the minimiser of $L_\alpha(\eta, \cdot)$. This suggests that strictly proper losses are not robust to any class noise.

C.1 Robustness Implies Non-convexity

We now define a general notion of robustness for losses for class probability estimation.

Definition 37 *Given an $\alpha \in [0, \frac{1}{2})$, we will say a loss $\ell\colon \{-1,1\} \times [0,1] \to \mathbb{R}$ is* α-robust at η *if the set of minimisers of the conditional risk for η and the set of minimisers of the conditional risk for η_α have some common points.*

That is, a loss is α-robust for a particular η if minimising the noisy conditional risk can potentially give an estimate that is also a minimiser of the non-noisy conditional risk. Formally, ℓ *is* α*-robust at* η when

$$H(\ell, \eta_\alpha) \cap H(\ell, \eta) \neq \emptyset,$$

where $H(\ell, \eta)$ is defined in (33). Due to the equivalence of α-corruption of data and a transformed loss, another way to think about this type of robustness is the following: under what conditions can using non-proper losses still lead to the recovery of accurate conditional probability estimates?

Label noise is symmetric about $\frac{1}{2}$ and so the map $\eta \mapsto \eta_\alpha$ preserves the side of $\frac{1}{2}$ on which the values η and η_α are found. That is, $\eta \leq \frac{1}{2}$ if and only if $\eta_\alpha \leq \frac{1}{2}$ for all $\alpha \in [0, \frac{1}{2})$. This means that 0-1 misclassification loss or, equivalently, $\ell_{\frac{1}{2}}$ is α-robust for all η and for all α. For other c, the range of η for which ℓ_c is α-robust is more limited.

Theorem 38 *For each $c \in (0,1)$, the loss ℓ_c is α-robust at η if and only if*

$$\eta \notin \left[\frac{c-\alpha}{1-2\alpha}, c\right) \text{ for } c < \tfrac{1}{2} \quad \text{or} \quad \eta \notin \left[c, \frac{c-\alpha}{1-2\alpha}\right) \text{ for } c \geq \tfrac{1}{2}.$$

Proof By the definition of L_c and $[\![\hat{\eta} < c]\!] = 1 - [\![\hat{\eta} \geq c]\!]$ we have

$$L_c(\eta, \hat{\eta}) = (1-\eta)c[\![\hat{\eta} \geq c]\!] + \eta(1-c)[\![\hat{\eta} < c]\!] = \eta(1-c) + (c-\eta)[\![\hat{\eta} \geq c]\!].$$

Since $c - \eta$ is positive iff $c > \eta$ we see $L_c(\eta, \hat{\eta})$ is minimised for $\eta < c$ when $\hat{\eta} < c$ and for $\eta \geq c$ when $\hat{\eta} \geq c$. So $H(\ell_c, \eta) = [0, c)$ for $\eta < c$ and $H(\ell_c, \eta) = [c, 1]$ for $\eta \geq c$. Since $[0, c)$ and $[c, 1]$

are disjoint for all $c \in [0,1]$ we see that $H(\ell_c, \eta)$ and $H(\ell_c, \eta_\alpha)$ coincide if and only if $\eta, \eta_\alpha < c$ or $\eta, \eta_\alpha \geq c$ and are disjoint otherwise.

We proceed by cases. First, suppose $c < \frac{1}{2}$. For $\eta < c < \frac{1}{2}$ it is easy to show $\eta_\alpha \geq c$ iff $\eta \geq \frac{c-\alpha}{1-2\alpha}$ and so ℓ_c is not α-robust for $\eta \in [\frac{c-\alpha}{1-2\alpha}, c)$. For $c \leq \eta$ we see ℓ_c must be α-robust since $\eta_\alpha < c$ iff $\eta < \frac{c-\alpha}{1-2\alpha}$ but $\frac{c-\alpha}{1-2\alpha} < c$ for $c < \frac{1}{2}$ which is a contradiction. Thus, for $c < \frac{1}{2}$ we have ℓ_c is α-robust iff $\eta \notin [\frac{c-\alpha}{1-2\alpha}, c)$.

For $c > \frac{1}{2}$ the main differences are that $\frac{c-\alpha}{1-2\alpha} > c$ for $c > \frac{1}{2}$ and $\eta_\alpha < \eta$ for $\eta > \frac{1}{2}$. Thus, by a similar argument as above we see that ℓ_c is α-robust iff $\eta \notin [c, \frac{c-\alpha}{1-2\alpha})$. ■

This theorem allows us to characterise the robustness of arbitrary proper losses by appealing to the integral representation in (4).

Lemma 39 *If ℓ is a proper loss with weight function w then $H(\ell, \eta) = \bigcap_{c:\, w(c)>0} H(\ell_c, \eta)$ and so*

$$H(\ell, \eta) \cap H(\ell, \eta_\alpha) = \bigcap_{c:\, w(c)>0} H(\ell_c, \eta) \cap H(\ell_c, \eta_\alpha).$$

Proof We first show that $H(\ell, \eta) \subseteq \bigcap_{c:\, w(c)>0} H(\ell_c, \eta)$ by contradiction. Assume there is an $\hat{\eta} \in H(\ell, \eta)$ but for which there is some c_0 such that $w(c_0) > 0$ and $\hat{\eta} \notin H(\ell_{c_0}, \eta)$. Then there is a $\hat{\eta}' \in H(\ell_{c_0}, \eta)$ and $\hat{\eta}' \in H(\ell_c)$ for all other c for which $w(c) > 0$ (otherwise $H(\ell, \eta) = \{\hat{\eta}\}$). Thus, $L_{c_0}(\eta, \hat{\eta}') < L_{c_0}(\eta, \hat{\eta}')$ and so $\int_0^1 L_c(\eta, \hat{\eta}')\, w(c)\, dc < \int_0^1 L_c(\eta, \hat{\eta})\, w(c)\, dc$ since $w(c_0) > 0$.

Now suppose $\hat{\eta} \in \bigcap_{c:\, w(c)>0} H(\ell_c, \eta)$. That is, $\hat{\eta}$ is a minimiser of $L_c(\eta, \cdot)$ for all c such that $w(c) > 0$ and therefore must also be a minimiser of $L(\eta, \cdot) = \int_0^1 L_c(\eta, \cdot)\, w(c)\, dc$ and is therefore in $H(\ell, \eta)$, proving the converse. ■

One consequence of this lemma is that if $w(c) > 0$ and ℓ_c is not α-robust at η then, by definition, $H(\ell_c, \eta) \cap H(\ell_c, \eta_\alpha) = \emptyset$ and so ℓ cannot be α-robust at η. This means we have established the following theorem regarding the α-robustness of an arbitrary proper loss in terms of its weight function.

Theorem 40 *If ℓ is a proper loss with weight function w then it is not α-robust for any*

$$\eta \in \bigcup_{c:\, w(c)>0} \left[\frac{c-\alpha}{1-2\alpha}, c\right) \cup \left[c, \frac{c-\alpha}{1-2\alpha}\right).$$

By Corollary 30 we see that convex proper losses are strictly proper and thus have weight functions which are non-zero for all $c \in [0,1]$ and so by Theorem 40 we have the following corollary.

Corollary 41 *If a proper loss is convex, then for all $\alpha \in (0, \frac{1}{2})$ it is not α-robust at any $\eta \in [0,1]$.*

At a high level, this result—"convexity implies non-robustness"—appears to be logically equivalent to Long and Servedio's result that "robustness implies non-convexity". However, there are a few discrepancies that mean they are not directly comparable. The definitions of robustness differ. We focus on the point-wise minimisation of conditional risk as this is, ideally, what most risk minimisation approach try to achieve. However, this means that robustness of ERM with regularisation or restricted function classes is not directly captured with our definition whereas Long and Servedio

analyse this latter case directly. In our definition the focus is on probability estimation robustness while the earlier work is focussed on classification accuracy. Our work could be extended to this case by analysing $H(\ell,\eta)\cap H(\ell_{\frac{1}{2}},\eta)$.

Additionally, their work restricts attention to the robustness of boosting algorithms that use convex potential functions whereas our analysis is not tied to any specific algorithm. By restricting their attention to a specific learning task and class of functions they are able to show a very strong result: that convex losses for boosting lead to arbitrarily bad performance with arbitrarily little noise. Also, our focus on proper losses excludes some convex losses (such as the hinge loss) that is covered by Long and Servedio's results.

Finally, it is worth noting that there are non-convex loss functions that are strictly proper and so are not robust in the sense we use here. That is, the converse of Corollary 41 is not true. For example, any loss with weight function that sits above 0 but outside the shaded region in Figure 2 will be non-convex and non-robust. This suggests that the arguments made by Masnadi-Shirazi and Vasconcelos (2009); Freund (2009) for the robustness of non-convex losses need further investigation.

References

J.D. Abernethy, A. Agarwal, P.L. Bartlett, and A. Rakhlin. A stochastic view of optimal regret through minimax duality. March 2009. URL http://arxiv.org/abs/0903.5328.

J. Aczel and J. Pfanzagl. Remarks on the measurement of subjective probability and information. *Metrika*, 11(1):91–105, December 1967.

F.R. Bach, D. Heckerman, and E. Horvitz. Considering cost asymmetry in learning classifiers. *Journal of Machine Learning Research*, 7:1713–1741, 2006.

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

A. Banerjee, S. Merugu, I.S. Dhillon, and J. Ghosh. Clustering with bregman divergences. *The Journal of Machine Learning Research*, 6:1705–1749, 2005.

P.J. Bartlett, B. Schölkopf, D. Schuurmans, and A.J. Smola, editors. *Advances in Large-Margin Classifiers*. MIT Press, 2000.

P.L. Bartlett and A. Tewari. Sparseness vs estimating conditional probabilities: Some asymptotic results. *The Journal of Machine Learning Research*, 8:775–790, 2007.

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

H.H. Bauschke and J.M. Borwein. Joint and separate convexity of the bregman distance. In Dan Butnariu, Yair Censor, and Simeon Reich, editors, *Inherently Parallel Algorithms in Feasibility and Optimization and their Applications*, volume 8 of *Studies in Computational Mathematics*, pages 23–36. North-Holland, 2001.

A. Beygelzimer, J. Langford, and B. Zadrozny. Machine learning techniques — reductions between prediction quality metrics. In Z. Liu and C.H. Xia, editors, *Performance Modeling and Engineering*, pages 3–28. Springer US, April 2008. URL http://hunch.net/~jl/projects/reductions/tutorial/paper/chapter.pdf.

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

P.F. Christoffersen and F.X. Diebold. Optimal prediction under asymmetric loss. *Econometric Theory*, 13(06):808–817, 2009.

I. Cohen and M. Goldszmidt. Properties and benefits of calibrated classifiers. Technical Report HPL-2004-22(R.1), HP Laboratories, Palo Alto, July 2004.

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

C. Elkan. The foundations of cost-sensitive learning. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, volume 17, pages 973–978, 2001.

S. Fidler, D. Skocaj, and A. Leonardis. Combining reconstructive and discriminative subspace methods for robust classification and regression by subsampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(3):337–350, 2006.

Y. Freund. A more robust boosting algorithm. arXiv:0905.2138v1 [stat.ML], May 2009. URL `http://arxiv.org/abs/0905.2138`.

T. Gneiting. Evaluating point forecasts. arXiv:0912.0902v1, December 2009.

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

C.W.J. Granger and M.J. Machina. Forecasting and decision theory. In G. Elliot, C.W.J. Granger, and A. Timmermann, editors, *Handbook of Economic Forecasting*, volume 1, pages 82–98. North-Holland, Amsterdam, 2006.

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

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

D.J. Hand and V. Vinciotti. Local versus global models for classification problems: Fitting models where it matters. *The American Statistician*, 57(2):124–131, 2003.

D.P. Helmbold, J. Kivinen, and M.K. Warmuth. Relative loss bounds for single neurons. *IEEE Transactions on Neural Networks*, 10:1291–1304, 1999.

J.-B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of Convex Analysis*. Springer, Berlin, 2001.

P.J. Huber. *Robust Statistics*. Wiley, New York, 1981.

Y. Kalnishkan, V. Vovk, and M.V. Vyugin. Loss functions, complexities, and the legendre transformation. *Theoretical Computer Science*, 313(2):195–207, 2004.

Y. Kalnishkan, V. Vovk, and M.V. Vyugin. Generalised entropy and asymptotic complexities of languages. In *Learning Theory*, volume 4539/2007 of *Lecture Notes in Computer Science*, pages 293–307. Springer, 2007.

M. Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM*, 45(6): 983–1006, November 1998.

J. Kivinen and M.K. Warmuth. Relative loss bounds for multidimensional regression problems. *Machine Learning*, 45:301–329, 2001.

D.E. Knuth. Two notes on notation. *American Mathematical Monthly*, pages 403–422, 1992.

N. Lambert, D. Pennock, and Y. Shoham. Eliciting properties of probability distributions. In *Proceedings of the ACM Conference on Electronic Commerce*, pages 129–138, 2008.

J. Langford and B. Zadrozny. Estimating class membership probabilities using classifier learners. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics (AISTAT'05)*, 2005.

Y. Lin. A note on margin-based loss functions in classification. Technical Report 1044, Department of Statistics, University of Wisconsin, Madison, February 2002.

P.M. Long and R.A. Servedio. Random classification noise defeats all convex potential boosters. In William W. Cohen, Andrew McCallum, and Sam T. Roweis, editors, *ICML*, pages 608–615, 2008. doi: 10.1145/1390156.1390233.

H. Masnadi-Shirazi and N. Vasconcelos. On the design of loss functions for classification: theory, robustness to outliers, and savageboost. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1049–1056. 2009.

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

R. Nock and F. Nielsen. Bregman divergences and surrogates for learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009a. To appear.

R. Nock and F. Nielsen. On the efficient minimization of classification calibrated surrogates. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1201–1208. MIT Press, 2009b.

J. Platt. Probabilities for sv machines. In A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 61–71. MIT Press, 2000.

F. Provost and T. Fawcett. Robust classification for imprecise environments. *Machine Learning*, 42 (3):203–231, 2001.

M.D. Reid and R.C. Williamson. Surrogate regret bounds for proper losses. In *Proceedings of the International Conference on Machine Learning*, pages 897–904, 2009a.

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

R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

L.J. Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.

M.J. Schervish. A general method for comparing probability assessors. *The Annals of Statistics*, 17 (4):1856–1879, 1989.

B. Schölkopf, A. Smola, R. C. Williamson, and P. L. Bartlett. New support vector algorithms. *Neural Computation*, 12:1207–1245, 2000.

Y. Shen. *Loss Functions for Binary Classification and Class Probability Estimation*. PhD thesis, Department of Statistics, University of Pennsylvania, October 2005.

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

C.-A. S. Staël von Holstein. *Assessment and evaluation of subjective probability distributions*. Economic Research Institute, Stockholm School of Economics, Stockholm, 1970.

I. Steinwart. How to compare different loss functions and their risks. *Constructive Approximation*, 26(2):225–287, August 2007.

I. Steinwart. Two oracle inequalities for regularized boosting classifiers. *Statistics and Its Interface*, 2:271–284, 2009.

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, New York, 2008.

T.B. Trafalis and R.C. Gilbert. Robust classification and regression using support vector machines. *European Journal of Operational Research*, 173(3):893–909, 2006.

A. Zellner. Bayesian estimation and prediction using asymmetric loss functions. *Journal of the American Statistical Association*, 81(394):446–451, June 1986.

J. Zhang. Divergence function, duality, and convex analysis. *Neural Computation*, 16(1):159–195, 2004a.

T. Zhang. Statistical behaviour and consistency of classification methods based on convex risk minimization. *Annals of Mathematical Statistics*, 32:56–134, 2004b.

Z. Zhang, M. I. Jordan, W. J. Li, and D. Y. Yeung. Coherence functions for multicategory margin-based classification methods. In *Proceedings of the Twelfth Conference on Artificial Intelligence and Statistics (AISTATS)*, 2009.