Surrogate Regret Bounds for Proper Losses

Mark D. Reid MARK.REID@ANU.EDU.AU
Australian National University, Canberra, 0200, Australia

Robert C. Williamson BOB.WILLIAMSON@ANU.EDU.AU
Australian National University and NICTA, Canberra, 0200, Australia

Abstract

We present tight surrogate regret bounds for the class of proper (*i.e.*, Fisher consistent) losses. The bounds generalise the margin-based bounds due to Bartlett et al. (2006). The proof uses Taylor's theorem and leads to new representations for loss and regret and a simple proof of the integral representation of proper losses. We also present a different formulation of a duality result of Bregman divergences which leads to a simple demonstration of the convexity of composite losses using canonical link functions.

1. Introduction

A surrogate loss function is a loss function which is not exactly what one wishes to minimise but is easier to work with. Convex surrogate losses are used in place of the 0-1 loss. Bartlett et al. (2006) have derived tight bounds on the regret with respect to the 0-1 loss ℓ^{0-1} when one knows the regret with respect to a convex margin loss. Surrogate regret bounds can be viewed as a type of reduction (Beygelzimer et al., 2008) between learning problems where one directly uses the hypothesis obtained by empirically minimising the surrogate loss for the original prediction problem.

Margin losses are a subset of all possible loss functions. We consider the general class of *proper losses* (defined below) which subsumes almost all margin losses. Proper losses – also known as proper scoring rules (Buja et al., 2005; Gneiting & Raftery, 2007) – are the "right" sort of loss to use when one studies class probability estimation. One advantage is that they have integral representations in terms of cost-weighted losses. This representation is central to the derivation of the main result of the paper (Theorem 3).

Appearing in *Proceedings of the 26^{th} International Conference on Machine Learning*, Montreal, Canada, 2009.

Our main contribution is to bring together and generalise several existing results from a diversity of sources and state them in a language more familiar to machine learning researchers. Importantly, we also show how they are all based upon the integral Taylor expansion of the conditional Bayes risk for (proper) surrogate losses.[1] Given a proper loss, the function expressing its conditional Bayes risk is trivially computed and our results suggest that much can be gained from making it and its integral representation central objects of study in learning theory.

The main theorems regarding proper losses are stated in §1.2 below. Their proofs are provided in the remainder of the paper. The surrogate regret bound and convex link results are discussed in §4 and §5, respectively. These rely on the integral representation discussed in §3 which in turn relies on Taylor's theorem and results from convex analysis presented in §2.

1.1. Probability Estimation and Proper Losses

We write $x \wedge y := \min(x, y)$, $x \vee y := \max(x, y)$ and $[\![p]\!] = 1$ if p is true and $[\![p]\!] = 0$ otherwise. The generalised function $\delta(\cdot)$ is defined by $\int_a^b \delta(x) f(x) dx = f(0)$ when f is continuous at 0 and $a < 0 < b$. The unit step $U(x) = \int_{-\infty}^{x} \delta(t) dt$. The real numbers are denoted $\mathbb{R}$ and the non-negative reals $\mathbb{R}^+$. Random variables are written in sans-serif font: X, Y. Sets are in calligraphic font: $\mathcal{X}$ (the "input" space). Vectors are written in bold font: $\boldsymbol{\alpha} \in \mathbb{R}^m$. We will often take expectations ($\mathbb{E}$) over the random variable X. The resulting quantities will be written in blackboard bold: $\mathbb{L}$ and $\mathbb{B}$. Proper losses (defined below) are denoted by ℓ; their associated conditional and full risks by L and $\mathbb{L}$. The lower bound on quantities with an intrinsic lower bound (e.g. the Bayes optimal loss) are written with an underbar: $\underline{L}, \underline{\mathbb{L}}$. Estimated quantities are hatted: $\hat{\eta}$.

We will call a (M-measurable) function $\hat{\eta} : \mathcal{X} \to [0, 1]$ a

[1] This is similar to an insight by Liese and Vajda (2006) who use a Taylor expansion to study integral representations of f-divergences.

class probability *estimator*. Overloading the notation, we also use $\hat{\eta} = \hat{\eta}(x) \in [0,1]$ to denote an *estimate* for a specific observation $x \in \mathcal{X}$. Much of the subsequent arguments rely on this conditional perspective.

Estimate quality is assessed using a *loss function* $\ell : \{0,1\} \times [0,1] \to \mathbb{R}$ and the loss of the estimate $\hat{\eta}$ with respect to the label $y \in \{0,1\}$ is denoted $\ell(y, \hat{\eta})$. Throughout, we adopt the convention that there is no cost for perfect prediction, that is, $\ell(0,0) = \ell(1,1) = 0$. We require the following technical assumption[2]:

$$\lim_{\eta \searrow 0} \eta \ell(1,\eta) = \lim_{\eta \nearrow 1} (1-\eta)\ell(0,\eta) = 0. \tag{1}$$

If $\eta \in [0,1]$ is the probability of observing the label $y = 1$ the *point-wise risk* of the estimate $\hat{\eta} \in [0,1]$ is defined as the η-average of the point-wise loss for $\hat{\eta}$:

$$L(\eta, \hat{\eta}) := \mathbb{E}_{\mathsf{Y} \sim \eta}[\ell(\mathsf{Y}, \hat{\eta})] = \ell(0,\hat{\eta})(1-\eta) + \ell(1,\hat{\eta})\eta. \tag{2}$$

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 $\hat{\eta}$:

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

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

A natural measure of the difficulty of a task is its minimal achievable risk, or *Bayes risk*:

$$\underline{\mathbb{L}}(\eta, M) := \inf_{\hat{\eta} \in [0,1]^{\mathcal{X}}} \mathbb{L}(\eta, \hat{\eta}, M) = \mathbb{E}_{\mathsf{X} \sim M}\left[\underline{L}(\eta(\mathsf{X}))\right],$$

where

$$[0,1] \ni \eta \mapsto \underline{L}(\eta) := \inf_{\hat{\eta} \in [0,1]} L(\eta, \hat{\eta})$$

is the *point-wise Bayes risk*. Note the use of the underline on $\underline{\mathbb{L}}$ and $\underline{L}$ to indicate that the corresponding functions $\mathbb{L}$ and L are minimised. If $\hat{\eta}$ is to be interpreted as an estimate of the true positive class probability η 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). That is, a proper loss ℓ satisfies $\underline{L}(\eta) = L(\eta, \eta)$ for all $\eta \in [0,1]$. The *cost-weighted losses* are a family of losses parametrised by a false positive cost $c \in [0,1]$ that defines a loss for $y \in \{0,1\}$ and $\hat{\eta} \in [0,1]$ by

$$\ell_c(y, \hat{\eta}) = c[\![y=0]\!][\![\hat{\eta} \geq c]\!] + (1-c)[\![y=1]\!][\![\hat{\eta} < c]\!]. \tag{3}$$

[2]This is equivalent to the conditions in (Savage, 1971) and (Schervish, 1989).

1.2. Main Results

The following important property of proper losses is originally attributed to Savage (1971) and is re-derived in Section 2. It shows that the point-wise Bayes risk of a loss is necessarily concave and how a proper loss can be derived from any concave function.

Theorem 1 *The point-wise Bayes risk $\underline{L}\colon [0,1] \to \mathbb{R}$ for a proper loss ℓ is concave function. Conversely, given a concave function $\Lambda : [0,1] \to \mathbb{R}$ there exists a proper loss ℓ satisfying $\underline{L}(\eta) = \Lambda(\eta)$, $\eta \in [0,1]$. Furthermore, the point-wise risk satisfies*

$$L(\eta, \hat{\eta}) = \underline{L}(\hat{\eta}) - (\hat{\eta} - \eta)\underline{L}'(\hat{\eta}). \tag{4}$$

In Section 3 we re-derive an old result (see Schervish (1989) and Gneiting and Raftery (2007) for its history) that characterises proper losses in terms of their "weight functions" – a dual relationship analogous to that between a function and its Fourier transform.

Theorem 2 *The function $\ell : \{0,1\} \times [0,1] \to \mathbb{R}$ is a proper loss iff for each $\hat{\eta} \in [0,1]$ and $y \in \{0,1\}$*

$$\ell(y, \hat{\eta}) = \int_0^1 \ell_c(y, \hat{\eta})\, w(c)\, dc, \tag{5}$$

where the "weight function" $w(c) = -\underline{L}''(c) \geq 0$.

Our regret bound is stated next. Its proof in Section 4 shows it is directly implied by the previous results.

Theorem 3 *Let $c_0 \in (0,1)$ and let $B_{c_0}(\eta, \hat{\eta})$ denote the point-wise regret for the cost-weighted loss ℓ_{c_0}. Suppose it is known that $B_{c_0}(\eta, \hat{\eta}) = \alpha$. Then the point-wise regret $B(\eta, \hat{\eta})$ for any proper surrogate loss ℓ with point-wise risk L and Bayes risk $\underline{L}$ satisfies*

$$B(\eta, \hat{\eta}) \geq \psi(c_0, \alpha) \vee \psi(c_0, -\alpha), \tag{6}$$

where

$$\psi(c_0, \alpha) := \underline{L}(c_0) - \underline{L}(c_0 + \alpha) + \alpha \underline{L}'(c_0).$$

Furthermore (6) is tight.

By restricting attention to the case when $c_0 = \frac{1}{2}$ and symmetric losses we obtain, as a corollary, a result similar to that presented by Barlett *et al.* (2006) for surrogate margin losses since $B_{\frac{1}{2}}$ is easily shown to be half the 0-1 regret.

Corollary 4 *If $\underline{L}$ is symmetric – that is, $\underline{L}(\frac{1}{2} - c) = \underline{L}(c - \frac{1}{2})$ for $c \in [0,1]$ – and $B_{\frac{1}{2}}(\eta, \hat{\eta}) = \alpha$, then*

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

In practice, the probability $\hat{\eta}$ is often not estimated directly. Instead, some (linearly) parameterised hypothesis $\hat{h} : \mathcal{X} \to \mathbb{R}$ is used and converted to a probability estimate $\hat{\eta} = \psi^{-1}(\hat{h})$ using a link function ψ. Computationally, it is useful if the composite risk $L(\eta, \psi^{-1}(\hat{h}))$ is convex in $\hat{h}$. The following theorem shows one can always "convexify" a proper loss.

Theorem 5 *Let $\psi = -\underline{L}'$. Then for $\hat{h} : \mathcal{X} \to \mathbb{R}$ the composite risk $L(\eta, \psi^{-1}(\hat{h}))$ is convex in $\hat{h}$.*

Buja et al. (2005) call this ψ the "canonical link". As shown in Section 5, our derivation of this result is a direct consequence of the integral representation of $\underline{L}$ and its Legendre-Fenchel dual.

2. Convexity and Taylor Expansions

In this paper we are primarily concerned with convex and concave functions defined on subsets of the real line. A central tool in their analysis is the integral form of their Taylor expansion. Here, ϕ' and ϕ'' denote the first and second derivatives of ϕ respectively.

Theorem 6 (Taylor's Theorem) *Let $[s_0, s]$ be a closed interval of $\mathbb{R}$ and let ϕ be a real-valued function over $[s_0, s]$. Then*

$$\phi(s) = \phi(s_0) + \phi'(s_0)(s - s_0) + \int_{s_0}^{s} (s - t)\, \phi''(t)\, dt. \quad (7)$$

The classical statement of Taylor's theorem requires ϕ to be twice differentiable, however we will use an extension that allows for generalised functions in a manner similar to Liese and Vajda (2006). For example, if $\phi(s) = \max(s, 0)$ Taylor's theorem holds when ϕ' is taken to be to be the unit step function $\phi'(s) = U(s)$, and $\phi''(s)$ to be $\delta(s)$.

The argument s in the above theorem can be awkward to work with as it appears in the limits of the integral. The following corollary removes this problem by replacing the integrand in (7) with a piece-wise linear function

$$\phi_t(s, s_0) := \begin{cases} (s - t) & s_0 < t \leq s \\ (t - s) & s < t \leq s_0 \\ 0 & \text{otherwise.} \end{cases} \quad (8)$$

This is a piece-wise linear and convex in s for each $s_0, t \in [a, b]$.

Corollary 7 (Integral Representation) *Let $\phi : [a, b] \to \mathbb{R}$ be a general function. Then, for all $s, s_0 \in [a, b]$ we have*

$$\phi(s) = \phi(s_0) + \phi'(s_0)(s - s_0) + \int_a^b \phi_t(s, s_0)\, \phi''(t)\, dt. \quad (9)$$

This result is can be immediately obtained upon substitution of ϕ_t into (7) since the conditions in (8) restrict the limits of integration to the interval $(s_0, s) \subseteq [a, b]$ or $(s, s_0) \subseteq [a, b]$ and reverse of the sign of $(s - t)$ when $s < s_0$.

Central to our analysis is the observation that the (generalised) second derivative of a convex ϕ is everywhere non-negative. This means the Taylor expansion in (9) can be seen as the sum of the linear component $\phi(s_0) + \phi'(s_0)(s - s_0)$ and a weighted combination of piece-wise linear terms ϕ_t. As we shall see, the *weight function* $w(t) = \phi''(t)$ that determines the contribution of each "primitive" ϕ_t characterises many of the important properties of the function ϕ.

We now show how Theorem 1 can be derived from the integral Taylor expansion of $\underline{L}$. We believe the proof here to be more transparent than earlier proofs.

Proof (Theorem 1) For the forward implication, assume ℓ is a proper loss. By definition, the point-wise Bayes risk $\underline{L}(\eta) = \inf_{\hat{\eta}} L(\eta, \hat{\eta})$ which, for each $\eta \in [0, 1]$ is just the lower envelope of the lines $L(\eta, \hat{\eta}) = (1 - \eta)\ell(0, \hat{\eta}) + \eta\ell(1, \hat{\eta})$ and thus $\underline{L}$ is concave.

The properness of ℓ means $\underline{L}(\eta) = L(\eta, \eta)$ and the $\hat{\eta}$-derivative of L is 0 when $\hat{\eta} = \eta$. Hence

$$\frac{\partial}{\partial \hat{\eta}} L(\eta, \hat{\eta})\bigg|_{\hat{\eta}=\eta} = (1 - \eta)\ell'(0, \eta) + \eta\ell'(1, \eta) = 0 \quad (10)$$

for all $\eta \in [0, 1]$. Expanding $\underline{L}'(\eta)$ using the chain rule gives

$$\begin{aligned} \underline{L}'(\eta) &= (1 - \eta)\ell'(0, \eta) - \ell(0, \eta) + \eta\ell'(1, \eta) + \ell(1, \eta) \\ &= \ell(1, \eta) - \ell(0, \eta) + \eta\ell'(1, \eta) + (1 - \eta)\ell'(0, \eta) \\ &= \ell(1, \eta) - \ell(0, \eta) \end{aligned}$$

where the last terms are 0 by (10). Thus

$$\begin{aligned} &\underline{L}(\hat{\eta}) + (\eta - \hat{\eta})\underline{L}'(\hat{\eta}) \\ &= (1 - \hat{\eta})\ell(0, \hat{\eta}) + \hat{\eta}\ell(1, \hat{\eta}) + (\eta - \hat{\eta})[\ell(1, \hat{\eta}) - \ell(0, \hat{\eta})] \\ &= (1 - \hat{\eta} - \eta + \hat{\eta})\ell(0, \hat{\eta}) + (\hat{\eta} + \hat{\eta} - \hat{\eta})\ell(1, \hat{\eta}) \\ &= (1 - \eta)\ell(0, \hat{\eta}) + \eta\ell(1, \hat{\eta}), \end{aligned}$$

which is the definition of $L(\eta, \hat{\eta})$. The result holds at the endpoints by the assumptions in (1).

Conversely, now suppose Λ is a concave function and let $\ell(y, \hat{\eta}) = \Lambda(\hat{\eta}) + (y - \hat{\eta})\Lambda'(\hat{\eta})$. The Taylor expansion of Λ is

$$\Lambda(\eta) = \Lambda(\hat{\eta}) + (\eta - \hat{\eta})\Lambda'(\hat{\eta}) + \int_{\hat{\eta}}^{\eta} (\eta - c)\, \Lambda''(c)\, dc$$

and so

$$L(\eta, \hat{\eta}) = \Lambda(\hat{\eta}) - \int_{\hat{\eta}}^{\eta} (\eta - c)\, \Lambda''(c)\, dc \geq \Lambda(\eta) = \underline{L}(\eta)$$

because the concavity of Λ means $\Lambda'' \leq 0$ and so the integral term is positive and is minimised to 0 when $\hat{\eta} = \eta$. This shows ℓ is proper, completing the proof. ■

2.1. Regret and Bregman Divergence

Bregman divergences are a generalisation of the notion of distances between points. In this section we recount an observation by Buja et al. (2005) that point-wise regret for a proper surrogate loss is a Bregman divergence. We begin with some definitions.[3] Given a differentiable[4] convex function $\phi : \mathcal{S} \to \mathbb{R}$ defined on the convex set $\mathcal{S} \subset \mathbb{R}^d$ and two points $s_0, s \in \mathcal{S}$ the *Bregman divergence of s from s_0* is defined

$$B_\phi(s, s_0) := \phi(s) - \phi(s_0) - \langle s - s_0, \nabla\phi(s_0)\rangle \,. \quad (11)$$

where $\nabla\phi(s_0)$ is the gradient of ϕ at s_0. A concise summary of many of the properties of Bregman divergences is given by Banerjee et al. (2005, Appendix A). In particular, Bregman divergences always satisfy $B_\phi(s, s_0) \geq 0$ and $B_\phi(s_0, s_0) = 0$ for all $s, s_0 \in \mathcal{S}$, regardless of the choice of ϕ. They are not always metrics, however, as they do not always satisfy the triangle inequality and their symmetry depends on the choice of ϕ.

When $\mathcal{S} = [0, 1]$, the concavity of $\underline{L}$ (see Theorem 1) means $\phi(s) = -\underline{L}(s)$ is convex and so induces the Bregman divergence[5]

$$B_\phi(s, s_0) = -\underline{L}(s) + \underline{L}(s_0) - (s_0 - s)\underline{L}'(s_0) = L(s, s_0) - \underline{L}(s).$$

The converse also holds. Given a Bregman divergence B_ϕ over $\mathcal{S} = [0, 1]$ the convexity of ϕ guarantees that $\underline{L} = -\phi$ is concave. Thus, we know that there is a proper loss ℓ with Bayes risk equal to $-\phi$. As noted by Buja et al. (2005, §19), the difference

$$B_\phi(\eta, \hat{\eta}) = L(\eta, \hat{\eta}) - \underline{L}(\eta)$$

is also known as the *point-wise regret* of the estimate $\hat{\eta}$ w.r.t. η. The corresponding *(full) regret* is the M-average point-wise regret

$$\mathbb{B}(\eta, \hat{\eta}, M) := \mathbb{E}_{\mathsf{X}\sim M}[B_\phi(\eta(\mathsf{X}), \hat{\eta}(\mathsf{X}))] = \mathbb{L}(\eta, \hat{\eta}) - \underline{\mathbb{L}}(\eta).$$

When $\mathcal{S} = \mathbb{R}$ and ϕ is twice differentiable, comparing the definition of a Bregman divergence in (11) to the integral representation in (7) reveals that Bregman divergences between real numbers can be defined as the non-linear part – or "Tayl" – of the Taylor expansion of ϕ. That is, for all $s, s_0 \in \mathbb{R}$

$$B_\phi(s, s_0) = \int_{s_0}^{s} (s - t)\, \phi''(t) dt, \quad (12)$$

since $\nabla\phi = \phi'$ and the inner product is simply multiplication over the reals.

From Theorem 1 we know that $-\underline{L}$ is convex for any proper surrogate loss ℓ. Thus, setting $\phi = -\underline{L}$ in the definition of Bregman divergence gives us

$$\begin{aligned} B_\phi(\eta, \hat{\eta}) &= -\underline{L}(\eta) + \underline{L}(\hat{\eta}) + (\eta - \hat{\eta})\underline{L}'(\hat{\eta}) \quad (13) \\ &= L(\eta, \hat{\eta}) - \underline{L}(\eta) \end{aligned}$$

by (4) in Theorem 1, which is the point-wise regret for the loss ℓ.

3. Weighted Integral Representations

We now consider a representation of proper losses in terms of primitive losses that originates with Shuford et al. (1966) and has been studied in some depth by Buja et al. (2005). It is also a special case of the recent integral representation obtained by Lambert et al. (2008) that generalises the earlier results to scoring rules for general properties of discrete distributions.

Our contribution is to show how this representation is a direct consequence of the Taylor expansion of a proper loss's Bayes risk. This shows the elementary nature of this representation and highlights the importance of the class of *cost-weighted misclassification losses* (3). Intuitively, a cost-weighted loss turns an estimate $\hat{\eta} \in [0, 1]$ into the classification $[\![\hat{\eta} \geq c]\!]$ and assigns a cost if this disagrees with the true classification y. Remaining consistent with our nomenclature for general losses, we will use L_c and $\mathbb{L}_c$ to denote the cost-weighted point-wise risk and full risk associated with each cost-weighted loss ℓ_c. The following lemma is needed for the proof of the main result.

Lemma 8 *For each $c \in (0, 1)$, ℓ_c is a proper loss and its Bayes risk $\underline{L}_c$ and regret B_c satisfy*

$$\begin{aligned} \underline{L}_c(\eta) &= ((1 - c)\eta) \wedge ((1 - \eta)c) \quad (14) \\ B_c(\eta, \hat{\eta}) &= |\eta - c|[\![\eta \wedge \hat{\eta} \leq c < \eta \vee \hat{\eta}]\!] \quad (15) \end{aligned}$$

for $\eta, \hat{\eta} \in [0, 1]$.

Proof By definition, for each $\eta \in [0, 1]$

$$\begin{aligned} \underline{L}_c(\eta) &= \inf_{\hat{\eta}\in[0,1]} (1 - \eta)c[\![\hat{\eta} \geq c]\!] + \eta(1 - c)[\![\hat{\eta} < c]\!] \\ &= \inf_{\hat{\eta}\in[0,1]} \eta(1 - c) + (c - \eta)[\![\hat{\eta} \geq c]\!], \end{aligned}$$

[3] Any terms related to convex analysis not explicitly defined can be found in (Hiriart-Urruty & Lemaréchal, 2001).

[4] Technically, ϕ need only be differentiable on the relative interior $\mathrm{ri}(\mathcal{S})$ of $\mathcal{S}$. We omit this requirement for simplicity and because it is not relevant to this discussion.

[5] Technically, $\mathcal{S}$ is the 2-simplex $\{(s_1, s_2) \in [0, 1]^2 : s_1 + s_2 = 1\}$ but we identify $s \in [0, 1]$ with $(s, 1 - s)$.

since $[\![\hat{\eta} < c]\!] = 1 - [\![\hat{\eta} \geq c]\!]$. Since $(c-\eta)$ is negative if and only if $\eta > c$ the infimum is obtained by having $[\![\hat{\eta} \geq c]\!] = 1$ if and only if $\eta \geq c$, that is, by letting $\hat{\eta} = \eta$. In this case, when $\eta = \hat{\eta} \geq c$ we have $\underline{L}_c(\eta) = c(1-\eta) = L_c(\eta, \eta)$ and when $\eta = \hat{\eta} < c$ we have $\underline{L}_c(\eta) = (1-c)\eta = L_c(\eta, \eta)$ and so ℓ is proper.

When $\eta \leq c < \hat{\eta}$ we see that $(1-c)\eta = \eta - c\eta \leq c - c\eta = (1-\eta)c$ and so $\underline{L}_c(\eta) = (1-c)\eta$. Also, by definition, $L_c(\eta, \hat{\eta}) = (1-\eta)c$ in this case so $B_c(\eta, \hat{\eta}) = (1-\eta)c - (1-c)\eta = c - \eta = |\eta - c|$. Similarly, when $\hat{\eta} \leq c < \eta$, $\underline{L}_c(\eta) = (1-\eta)c$ and $L_c(\eta, \hat{\eta}) = (1-c)\eta$ so $B_c(\eta, \hat{\eta}) = \eta - c = |\eta - c|$ proving the result. ■

We are now able to give the proof of Theorem 2.

Proof (Theorem 2) We first assume ℓ is a proper loss so that $L(\eta, \hat{\eta}) = \mathbb{E}_{\mathsf{Y} \sim \eta}[\ell(\mathsf{Y}, \hat{\eta})]$ and $\underline{L}(\eta) = L(\eta, \eta)$. Expanding $\underline{L}(\eta)$ about $\hat{\eta} \in [0,1]$ using Corollary 7 yields

$$\begin{aligned} \underline{L}(\eta) &= \underline{L}(\hat{\eta}) + (\eta - \hat{\eta})\underline{L}'(\hat{\eta}) + \int_0^1 \phi_c(\eta, \hat{\eta})\, \underline{L}''(c)\, dc \\ &= L(\eta, \hat{\eta}) + \int_0^1 \phi_c(\eta, \hat{\eta})\, \underline{L}''(c)\, dc \end{aligned} \tag{16}$$

by Theorem 1.

The generalised function $w(c) = -\underline{L}''(c) \geq 0$ by the concavity of $\underline{L}$. Rearranging (16) gives

$$L(\eta, \hat{\eta}) = \underline{L}(\eta) + \int_0^1 \phi_c(\eta, \hat{\eta})\, w(c)\, dc.$$

The definition of L in (2) implies $L(y, \hat{\eta}) = \ell(y, \hat{\eta})$ for $y \in \{0, 1\}$ and so

$$\ell(y, \hat{\eta}) = \underline{L}(y) + \int_0^1 \phi_c(y, \hat{\eta})\, w(c)\, dc, \tag{17}$$

where

$$\phi_c(y, \hat{\eta}) = [\![\hat{\eta} \leq c < y]\!](y - c) + [\![y \leq c < \hat{\eta}]\!](c - y),$$

which is equal to the definition of ℓ_c in (3) since the left (resp. right) term is only non-zero when $y = 1$ (resp. $y = 0$). Observe that $\underline{L}(0) = \underline{L}(1) = 0$ since $\underline{L}(0) = L(0,0) = \ell(0,0) = 0$ by assumption, and similarly for $\underline{L}(1)$. This shows that (17) is equivalent to (5), completing the forward direction of the theorem.

If we now assume the function $w \geq 0$ is given and ℓ defined as in (5) then it suffices to show $\underline{L}(\eta) = L(\eta, \eta)$. First note that

$$\begin{aligned} L(\eta, \hat{\eta}) &= \mathbb{E}_{\mathsf{Y} \sim \eta}\left[\int_0^1 \ell_c(\mathsf{Y}, \hat{\eta})\, w(c)\, dc\right] \\ &= \int_0^1 L_c(\eta, \hat{\eta})\, w(c)\, dc. \end{aligned}$$

Each of the L_c are proper by Lemma 8 and so are minimised when $\hat{\eta} = \eta$. Since $w(c) \geq 0$ this must also be sufficient to minimise L. ■

The linearity of expectation and regret provide some weighted integral representations for other quantities.

Corollary 9 *Let ℓ be a proper surrogate loss and let $L(\eta, \hat{\eta})$ be its point-wise risk, $B(\eta, \hat{\eta}) = L(\eta, \hat{\eta}) - \underline{L}(\eta)$ its point-wise regret and $w = -\underline{L}''$. Then for $\eta, \hat{\eta} \in [0,1]$*

$$L(\eta, \hat{\eta}) = \int_0^1 L_c(\eta, \hat{\eta})\, w(c)\, dc \tag{18}$$

$$B(\eta, \hat{\eta}) = \int_{\eta \wedge \hat{\eta}}^{\eta \vee \hat{\eta}} |\eta - c|\, w(c)\, dc. \tag{19}$$

When $\eta(x)$ is the probability that $x \in \mathcal{X}$ is positive and $\hat{\eta} : \mathcal{X} \to [0,1]$ a predictor, its risk and regret satisfy

$$\begin{aligned} \mathbb{L}(\eta, \hat{\eta}, M) &= \int_0^1 \mathbb{L}_c(\eta, \hat{\eta}, M)\, w(c)\, dc \\ \mathbb{B}(\eta, \hat{\eta}, M) &= \int_0^1 \mathbb{B}_c(\eta, \hat{\eta}, M)\, w(c)\, dc. \end{aligned}$$

Proof Equation 18 is the result of taking expectations on both sides of (5). Equation 19 is obtained by integrating the expression for B_c in Lemma 8. The remaining two expressions are the result of applying $\mathbb{E}_M[\cdot]$ to the weighted integral representations for L and B. ■

4. Surrogate Regret Bounds

Proper losses for probability estimation and surrogate margin losses (Bartlett et al., 2006) for classification are closely related. Buja et al. (2005) note that "the surrogate criteria of classification are exactly the primary criteria of class probability estimation" and that most commonly used surrogate margin losses are just proper scores mapped from $[0,1]$ to $\mathbb{R}$ via a link function. The main exceptions are hinge losses[6] which means SVMs are "the only case that truly bypasses estimation of class probabilities and directly aims at classification" (Buja et al., 2005, pg. 4). However, commonly used margin losses of the form $\phi(y\hat{h}(x))$ are a more restrictive class than proper losses since, as Buja et al. (2005, §23) note, "[t]his dependence on the margin limits all theory and practice to a symmetric treatment of class 0 and class 1".

Suppose for some fixed $c_0 \in (0,1)$ that $B_{c_0}(\eta, \hat{\eta}) = \alpha$. What can be said concerning the value of the regret $B(\eta, \hat{\eta})$ for an arbitrary but proper surrogate loss ℓ? Theorem 3 provides an answer to this question in the form of a surrogate

[6] And powers of absolute divergence $|y - r|^\alpha$ for $\alpha \neq 2$.

loss bound. Such bounds are of practical importance as the losses L_{c_0} are hard to optimise and rearranging the bounds provide a guarantee that minimising the surrogate loss (and hence the surrogate regret) will minimise the L_{c_0} loss.

We now provide an elementary proof of these bounds.

Proof (Theorem 3) Let B be the conditional regret associated with some arbitrary proper loss ℓ and suppose that we know the cost-weighted regret $B_{c_0}(\eta, \hat{\eta}) = \alpha$. By Lemma 8, this implies that $\alpha = \eta - c_0$ when $\hat{\eta} \leq c_0 < \eta$ and $\alpha = c_0 + \eta$ when $\eta \leq c_0 < \hat{\eta}$. In the first case we have $\hat{\eta} \leq c_0$ and $\eta = c_0 + \alpha$ and so

$$\begin{aligned} B(\eta, \hat{\eta}) &= B(c_0 - \alpha, \hat{\eta}) \\ &= \int_{\hat{\eta}}^{c_0+\alpha} (c_0 + \alpha - c)\, w(c)\, dc \\ &\geq \int_{c_0}^{c_0+\alpha} (c_0 + \alpha - c)\, w(c)\, dc \end{aligned}$$

by (19) and the assumption that $\hat{\eta} \leq c_0$. Note that this bound is achieved for $\hat{\eta} = c_0$ and so is tight. Thus, using $w(c) = -\underline{L}''(c)$ and integrating by parts gives

$$\begin{aligned} B(\eta, \hat{\eta}) &\geq -\left[(c_0 + \alpha - c)\underline{L}'(c)\right]_{c_0}^{c_0+\alpha} - \int_{c_0}^{c_0+\alpha} \underline{L}'(c) dc \\ &= \alpha \underline{L}'(c_0) - \underline{L}(c_0 + \alpha) + \underline{L}(c_0) \end{aligned}$$

as required.

The proof of the second case, when $\eta \leq c_0 < \hat{\eta}$ proceeds identically. ∎

Corollary 4 is obtained directly by substituting $\alpha = \frac{1}{2}$ and noting the symmetry of L implies $\underline{L}'(\frac{1}{2}) = 0$.

The bounds in Theorem 3 can be inverted so as to guarantee the minimisation of a cost-weighted loss via the minimisation of a surrogate loss.

Corollary 10 *Minimising $B(\eta, \hat{\eta})$ with respect to $\hat{\eta}$ minimises $B_c(\eta, \hat{\eta})$ for each $c \in (0, 1)$.*

Proof To see this, let $\psi'(c_0, \alpha) := \frac{\partial}{\partial \alpha}\psi(c_0, \alpha) = -\underline{L}'(c_0 + \alpha) + \underline{L}'(c_0)$. Since $\underline{L}$ is concave, $\underline{L}'$ is non-increasing and hence $\underline{L}'(c_0 + \alpha) \leq \underline{L}'(c_0)$ and so $\psi'(c_0, \alpha) \geq 0$ and therefore $\alpha \mapsto \psi(c_0, \alpha)$ is non-decreasing and thus invertible (although there may be non-uniqueness at points where $\psi(c_0, \alpha)$ is constant in α). This invertibility means minimising $B(\eta, \hat{\eta})$ w.r.t. $\hat{\eta}$, minimises the bound on $B_c(\eta, \hat{\eta})$. ∎

This shows that proper surrogate losses are surrogates for the entire family of cost-sensitive losses ℓ_c, not just 0-1 loss (*i.e.*, the case where $c = \frac{1}{2}$).

4.1. Related Work

Surrogate loss bounds have garnered increasing interest in the machine learning community (Zhang, 2004b; Bartlett et al., 2006; Steinwart, 2007).

All of the recent work has been in terms of *margin losses* of the form

$$L^{\phi}(\eta, \hat{h}) = \eta\phi(\hat{h}) + (1 - \eta)\phi(-\hat{h}).$$

As Buja et al. (2005) discuss, such margin losses can not capture the richness of all possible proper losses. Bartlett et al. (2006) prove that for any $\hat{h}$

$$\psi\left(L^{0-1}(\eta, \hat{h}) - \underline{L}^{0-1}(\eta)\right) \leq L^{\phi}(\eta, \hat{h}) - \underline{L}^{\phi}(\eta),$$

where $\psi = \tilde{\psi}^{\star\star}$ is the LF biconjugate of $\tilde{\psi}$,

$$\tilde{\psi}(\theta) = H^{-}\left(\frac{1+\theta}{2}\right) - H\left(\frac{1+\theta}{2}\right),$$

$H(\eta) = \underline{L}^{\phi}(\eta)$ and

$$H^{-}(\eta) = \inf_{\alpha:\, \alpha(2\eta-1)\leq 0} (\eta\phi(\alpha) + (1 - \eta)\phi(-\alpha))$$

is the optimal conditional risk under the constraint that the sign of the argument α disagrees with $2\eta - 1$.

We will consider two examples and show that the bounds we obtain with the above result match those obtained with Theorem 3.

Exponential Loss Consider the link $\hat{h} = \psi(\hat{\eta}) = \frac{1}{2}\log\frac{\hat{\eta}}{1-\hat{\eta}}$ with corresponding inverse link $\hat{\eta} = \frac{1}{1+e^{-2\hat{h}}}$. Buja et al. (2005) showed that this link function combined with exponential margin loss $\phi(\gamma) = e^{-\gamma}$ results in a proper loss

$$L(\eta, \hat{\eta}) = \eta\left(\frac{1-\hat{\eta}}{\hat{\eta}}\right)^{\frac{1}{2}} + (1 - \eta)\left(\frac{\hat{\eta}}{1-\hat{\eta}}\right)^{\frac{1}{2}}.$$

Hence

$$\underline{L}(\eta) = L(\eta, \eta) = 2\sqrt{\eta(1-\eta)}$$

and from Corollary 4 we obtain that if $B_{\frac{1}{2}}(\eta, \hat{\eta}) = \alpha$ then

$$B(\eta, \hat{\eta}) \geq 1 - \sqrt{1 - 4\alpha^2}.$$

This matches the result presented by Bartlett et al. (2006) upon noting that $B_{\frac{1}{2}}(\eta, \hat{\eta})$ measures the loss in terms of $\ell_{\frac{1}{2}}$ and they used $\ell^{0-1} = 2\ell_{\frac{1}{2}}$.

Truncated Quadratic Loss Consider the margin loss $\phi(\hat{h}) = (1 + \hat{h} \vee 0)^2 = (2\hat{\eta} \vee 0)^2$ with link function $\hat{h}(\hat{\eta}) = 2\hat{\eta} - 1$. One can show that $\underline{L}(\eta) = 4\eta(1 - \eta)$ and from Corollary 4 the regret bound $B(\eta, \hat{\eta}) \geq 4\alpha^2$. This

matches the result of Bartlett et al. (2006) when again it is noted we used $\ell_{\frac{1}{2}}$ and they used ℓ^{0-1}.

The Probing Reduction The Probing reduction (Langford & Zadrozny, 2005) shows how the square loss for class probability estimation can be bounded by an average cost-weighted regret. The weighted integral representation allows us to obtain a similar result for the *regret* for the square loss $\ell_{\text{sq}}(y,\hat{\eta}) := (y-\hat{\eta})^2$. Specifically,

$$\begin{aligned}\mathbb{B}_{\text{sq}}(\eta,\hat{\eta}) &= \mathbb{E}_{\mathsf{X}\sim M}[\mathbb{E}_{\mathsf{Y}\sim\eta(\mathsf{X})}[(\mathsf{Y}-\hat{\eta}(\mathsf{X}))^2]] \\ &= 2\int_0^1 \mathbb{E}_{\mathsf{X}\sim M}[B_c(\eta(\mathsf{X}),\hat{\eta}(\mathsf{X}))]\,dc. \quad (20)\end{aligned}$$

To see this, note that for square loss

$$L_{\text{sq}}(\eta,\hat{\eta}) = \eta(1-\hat{\eta})^2 + (1-\eta)\hat{\eta}^2,$$

and so

$$\underline{L}_{\text{sq}}(\eta) = L_{\text{sq}}(\eta,\eta) = \eta - \eta^2$$

which means that $w(c) = -\underline{L}''_{\text{sq}}(c) = 2$.

Now, the integral representation result means that the regret for square loss can be written

$$B_{\text{sq}}(\eta,\hat{\eta}) = 2\int_0^1 B_c(\eta,\hat{\eta})\,dc.$$

Taking expectations with respect to M on both sides gives the result in (20).

5. Convexity and Canonical Links

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 $[0,1]$. In such cases it is common to use a *link function* (McCullagh & Nelder, 1989). 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 $L(\eta,\psi^{-1}(\hat{h}))$ is convex in $\hat{h}$ (which implies it is convex in $\boldsymbol{\alpha}$ when $\hat{h}$ is linear in $\boldsymbol{\alpha}$). Theorem 5 shows that choosing $\psi = -\underline{L}'$ guarantees convexity. Its proof is aided by a slight change of notation.

Consider the general representation of $B(\eta,\hat{\eta})$ presented in (13). Set $\overline{W} := \phi = -\underline{L}$, $W := \overline{W}'$ and $w := W'$. We consider w as a *weight-function*[7] since the convexity of $\overline{W}$ implies W is monotone non-decreasing and thus w is non-negative. Rewriting (13) we have

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

where we stress we have parametrised the Bregman divergence by the monotone function W, rather than by the convex function $\overline{W}$ as is traditional. Similarly, denote by L_W the w-weighted conditional loss parametrised by W. We shall see below that our choice of parametrisation is felicitous.

When a function f is suitably smooth, the Legendre-Fenchel (LF) dual of f can be expressed in terms of its derivative and inverse. Furthermore in this case (writing $Df := f'$) $f' = (Df^\star)^{-1}$. Thus with w, W, and $\overline{W}$ defined as above,

$$W = (D(\overline{W}^\star))^{-1},\ W^{-1} = D(\overline{W}^\star),\ \overline{W}^\star = \int W^{-1}. \quad (21)$$

The following theorem is known (Zhang, 2004a) but as will be seen, stating it in terms of B_W provides some advantages.

Theorem 11 *Let w, W, $\overline{W}$ and B_W be as above. Then for all $x, y \in [0,1]$,*

$$B_W(x,y) = B_{W^{-1}}(W(y),W(x)). \quad (22)$$

Proof Using the Legendre transform we have

$$\begin{aligned}&\overline{W}^\star(u) = u\cdot W^{-1}(u) - \overline{W}(W^{-1}(u)) \\ \Rightarrow\ &\overline{W}(W^{-1}(u)) = u\cdot W^{-1}(u) - \overline{W}^\star(u). \quad (23)\end{aligned}$$

Equivalently (using (21))

$$\overline{W}^\star(W(u)) = u\cdot W(u) - \overline{W}(u). \quad (24)$$

Thus substituting and then using (23) we have

$$\begin{aligned}&B_W(x,W^{-1}(v)) \\ =\ &\overline{W}(x) - \overline{W}(W^{-1}(v)) - (x - W^{-1}(v))\cdot W(W^{-1}(v)) \\ =\ &\overline{W}(x) + \overline{W}^\star(v) - vW^{-1}(v) - (x - W^{-1}(v))\cdot v \\ =\ &\overline{W}(x) + \overline{W}^\star(v) - x\cdot v. \quad (25)\end{aligned}$$

Similarly (this time using (24)) we have

$$\begin{aligned}&B_{W^{-1}}(v,W(x)) \\ =\ &\overline{W}^\star(v) - \overline{W}^\star(W(x)) - (v - W(x))\cdot W^{-1}(W(x)) \\ =\ &\overline{W}^\star(v) - xW(x) + \overline{W}(x) - v\cdot x + xW(x) \\ =\ &\overline{W}^\star(v) + \overline{W}(x) - v\cdot x. \quad (26)\end{aligned}$$

Comparing (25) and (26) we see that

$$B_W(x,W^{-1}(v)) = B_{W^{-1}}(v,W(x)).$$

Let $y = W^{-1}(v)$. Thus substituting $v = W(y)$ leads to (22). ∎

[7] This weight function exactly corresponds to the weight functions used by Buja et al. (2005).

We now give the proof of Theorem 5 which provides a simple sufficient condition for the composite loss to be convex in $\hat{h}$. It was previously shown (with a more intricate proof) by Buja et al. (2005). The result also corresponds to the notion of "matching loss" (Helmbold et al., 1999).

Proof (Theorem 5) If the link $\psi = W = -\underline{L}'$ (and thus $\hat{\eta} = W^{-1}(\hat{h})$) then $B_W(\eta, \hat{\eta}) = \overline{W}(\eta) + \overline{W}^\star(\hat{h}) - \eta \cdot \hat{h}$. by (25) and so we have that

$$\begin{aligned} L_W(\eta, \hat{\eta}) &= B_W(\eta, \hat{\eta}) + \underline{L}(\eta) \\ &= \overline{W}(\eta) + \overline{W}^\star(\hat{h}) - \eta \cdot \hat{h} - \overline{W}(\eta) \end{aligned}$$

which is just $\overline{W}^\star(\hat{h}) - \eta \cdot \hat{h}$. Its 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. ■

6. Conclusions

We have: 1) developed new explicit tight regret bounds for general proper losses (not just margin losses); 2) developed explicit formulae for losses and regrets in terms of weighted representations; and 3) simplified the proofs of some classical but little known results (Savage's expansion, the Shuford integral representation and Buja's convexity of composite losses with canonical links). Importantly, all these results were derived using only elementary techniques – the most fundamental being Taylor's theorem.

The elementary nature of our presentation highlights the conditional Bayes risk of a proper loss as a key object in learning theory and demonstrates the value of its integral representation. Further uses of this representation are given by Reid and Williamson (2009).

7. Acknowledgements

This work was supported by the Australian Research Council. RW is also supported by NICTA which is funded by the Australian Government through Backing Australia's Ability. We thank the anonymous reviewers for their attention to detail and considered suggestions.

References

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

Bartlett, P., Jordan, M., & McAuliffe, J. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, *101*, 138–156.

Beygelzimer, A., Langford, J., & Zadrozny, B. (2008). Machine learning techniques — reductions between prediction quality metrics. Preprint.

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

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

Helmbold, D., Kivinen, J., & Warmuth, M. (1999). Relative loss bounds for single neurons. *IEEE Transactions on Neural Networks*, *10*, 1291–1304.

Hiriart-Urruty, J.-B., & Lemaréchal, C. (2001). *Fundamentals of convex analysis*. Berlin: Springer.

Lambert, N., Pennock, D., & Shoham, Y. (2008). Eliciting properties of probability distributions. *Proceedings of the ACM Conference on Electronic Commerce* (pp. 129–138).

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

Liese, F., & Vajda, I. (2006). On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, *52*, 4394–4412.

McCullagh, P., & Nelder, J. (1989). *Generalized linear models*. Chapman & Hall/CRC.

Reid, M. D., & Williamson, R. C. (2009). Information, divergence and risk for binary experiments. arXiv preprint arXiv:0901.0356v1, 89 pages.

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

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

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

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

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

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