Surrogate Regret Bounds for Proper Losses

Mark D. Reid The Australian National University
Robert C. Williamson The Australian National University & NICTA

Wednesday, 17 June

ICML 2009

Overview

Introduction

Aims

To better understand loss functions through:

- Translation: Make work on risk from other fields ML-friendly
- Unification: Find key concepts underpinning existing results
- Generalisation: Propose generalisation of existing results

Aims

To better understand loss functions through:

- Translation: Make work on risk from other fields ML-friendly
- Unification: Find key concepts underpinning existing results
- Generalisation: Propose generalisation of existing results

This approach led to:

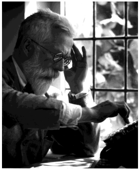
- Simpler proofs of some existing results
- A new type of surrogate regret bound:
  - Symmetric and non-symmetric surrogate losses
  - Bounds on cost-weighted misclassification loss (of which 0-1 loss is a special case)

Key Concepts

Two elementary concepts underpin all the results in this talk:

Fisher Consistency

A loss is Fisher consistent for probability estimation if its point-wise risk is minimised by the true point-wise probability.



Taylor's Theorem - Integral Form

Given a function $f : [x_0, x] \to \mathbb{R}$ then

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \int_{x_0}^{x} f'(t)(x - t)\, dt$$

What is a loss?

A loss ℓ assigns a *penalty* $\ell(y, h)$ to a *prediction* $h \in \mathbb{R}$ relative to a *label* y.

What is a loss?

A loss ℓ assigns a *penalty* $\ell(y, h)$ to a *prediction* $h \in \mathbb{R}$ relative to a *label* y.

Traditionally, losses in machine learning are margin losses:

$$\ell(y, h) = \phi(yh)$$

where $y \in \{-1, 1\}$ and $\phi : \mathbb{R} \to \mathbb{R}$.

These are *necessarily symmetric* in that

$$\ell(-1, h) = \ell(1, -h).$$

Composite Losses

We study a general class of composite losses:

$$\ell^{\psi}(y, h) = \ell(y, \psi^{-1}(h))$$

where $\psi : [0, 1] \to \mathbb{R}$ is an invertible link function that allows predictions $h \in \mathbb{R}$ to be interpreted as probability estimates

$$\hat{\eta} = \psi^{-1}(h).$$

Composite Losses

We study a general class of composite losses:

$$\ell^{\psi}(y, h) = \ell(y, \psi^{-1}(h))$$

where $\psi : [0, 1] \to \mathbb{R}$ is an invertible link function that allows predictions $h \in \mathbb{R}$ to be interpreted as probability estimates

$$\hat{\eta} = \psi^{-1}(h).$$

We focus on the loss for probability estimation rather than the link.

Loss

A loss is a function $\ell : \{0, 1\} \times [0, 1] \to \mathbb{R}$ such that

$$\ell(0, 0) = \ell(1, 1) = 0$$

which assigns a penalty $\ell(y, \hat{\eta})$ for predicting that the probability that $y = 1$ is $\hat{\eta} \in [0, 1]$ when the true label is y.

Risk

Aim is to find an *estimator* $\hat{\eta} : \mathcal{X} \rightarrow [0,1]$ that minimises the risk w.r.t. some unknown distribution $\mathbb{P}$

$$\begin{aligned} \mathbb{L}(Y, \hat{\eta}(X)) &= \mathbb{E}_{(X,Y)\sim\mathbb{P}}[\ell(Y, h(X))] \\ &= \mathbb{E}_X[\mathbb{E}_{Y\sim\eta(X)}[\ell(Y, \hat{\eta}(X))]] \end{aligned}$$

Risk

Aim is to find an *estimator* $\hat{\eta} : \mathcal{X} \to [0,1]$ that minimises the risk w.r.t. some unknown distribution $\mathbb{P}$

$$\begin{aligned}\mathbb{L}(Y, \hat{\eta}(X)) &= \mathbb{E}_{(X,Y)\sim\mathbb{P}}[\ell(Y, h(X))] \\ &= \mathbb{E}_X[\mathbb{E}_{Y\sim\eta(X)}[\ell(Y, \hat{\eta}(X))]]\end{aligned}$$

Point-wise Risk

The point-wise risk of ℓ under $Y \sim \eta$ is

$$L(\eta, \hat{\eta}) = \mathbb{E}_{Y\sim\eta}[\ell(Y, \hat{\eta})]$$

Risk

Aim is to find an *estimator* $\hat{\eta} : \mathcal{X} \to [0,1]$ that minimises the risk w.r.t. some unknown distribution $\mathbb{P}$

$$
\begin{aligned}
\mathbb{L}(Y, \hat{\eta}(X)) &= \mathbb{E}_{(X,Y)\sim\mathbb{P}}[\ell(Y, h(X))] \\
&= \mathbb{E}_X[\mathbb{E}_{Y\sim\eta(X)}[\ell(Y, \hat{\eta}(X))]]
\end{aligned}
$$

Point-wise Risk

The point-wise risk of ℓ under $Y \sim \eta$ is

$$L(\eta, \hat{\eta}) = \mathbb{E}_{Y\sim\eta}[\ell(Y, \hat{\eta})]$$

Point-wise Bayes Risk

The point-wise Bayes risk is the minimal point-wise risk

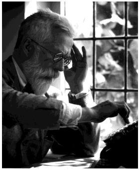
$$\underline{L}(\eta) = \inf_{\hat{\eta}\in\mathbb{R}} L(\eta, \hat{\eta})$$

Key Concepts: Fisher Consistency

Fisher Consistency

A loss $\ell(y, \hat{\eta})$ is Fisher consistent if

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

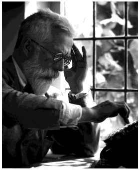


Key Concepts: Fisher Consistency

Fisher Consistency

A loss $\ell(y, \hat{\eta})$ is Fisher consistent if

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



Proper Loss

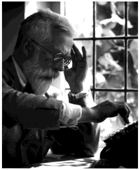
A loss is said be proper if it is Fisher consistent.

Key Concepts: Fisher Consistency

Fisher Consistency

A loss $\ell(y, \hat{\eta})$ is Fisher consistent if

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



Proper Loss

A loss is said be proper if it is Fisher consistent.

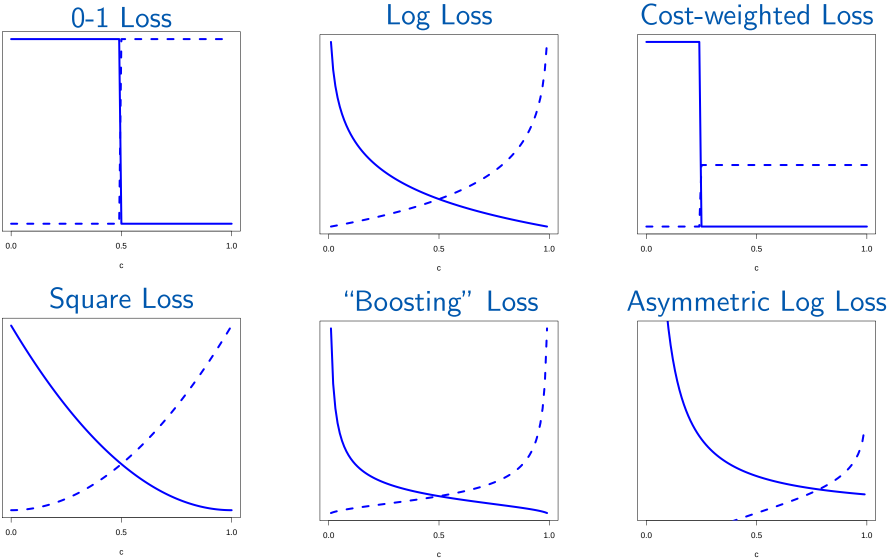
Computing the point-wise Bayes risk of proper losses is easy.

Example (Square Loss)

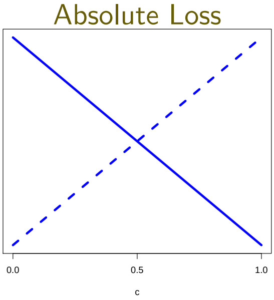
$L(\eta, \hat{\eta}) = (1 - \eta)\hat{\eta}^2 + \eta(1 - \hat{\eta})^2$ so its Bayes risk is

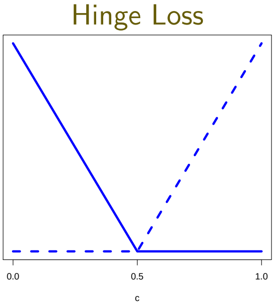
$$\underline{L}(\eta) = L(\eta, \eta) = (1 - \eta)\eta$$

Proper Losses: Examples

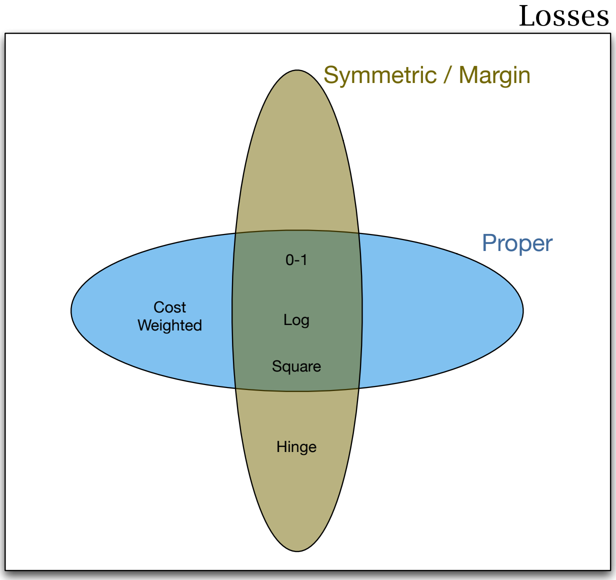


Non-Proper Losses: Examples





Losses


Losses
Symmetric / Margin
Proper
0-1
Cost
Weighted
Log
Square
Hinge

Key Concepts: Taylor's Theorem

Taylor's Theorem - Integral Form

Given a function $f : [x_0, x] \to \mathbb{R}$ then

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \int_{x_0}^{x} f'(t)(x - t)\, dt$$

Taylor's Theorem - Alternative Form

For $x, x_0 \in [a, b]$ and $f : [a, b] \to \mathbb{R}$ suitably differentiable

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \int_{a}^{b} g_c(x, x_0)\, f''(c)\, dc$$

where

$$g_c(x, x_0) = \begin{cases} (x - c) & x_0 < c \leq x \\ (c - x) & x < c \leq x_0 \\ 0 & \text{otherwise} \end{cases}$$

Representations

Savage's Theorem

Theorem (Savage, 1971)

A loss ℓ is proper iff its point-wise Bayes risk $\underline{L}$ is concave and satisfies

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

Savage's Theorem

Theorem (Savage, 1971)

A loss ℓ is proper iff its point-wise Bayes risk $\underline{L}$ is concave and satisfies

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

Proof sketch.

$\Rightarrow$ $\underline{L}(\eta)$ is infimum of $L(\eta, \hat{\eta})$ which is a lower envelope of lines thus concave, and $\underline{L}'(\eta) = \ell(1, \eta) - \ell(0, \eta)$.

□

Savage's Theorem

Theorem (Savage, 1971)

A loss ℓ is proper iff its point-wise Bayes risk $\underline{L}$ is concave and satisfies

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

Proof sketch.

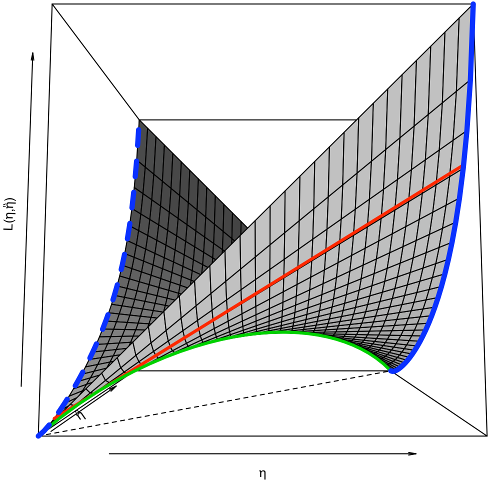
$\Rightarrow$ $\underline{L}(\eta)$ is infimum of $L(\eta, \hat{\eta})$ which is a lower envelope of lines thus concave, and $\underline{L}'(\eta) = \ell(1, \eta) - \ell(0, \eta)$.

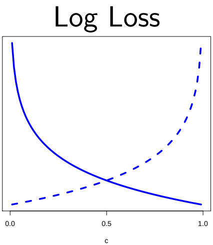
$\Leftarrow$ Taylor expansion of $\Lambda(\eta)$ about $\hat{\eta}$ gives

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

and since $-\Lambda'' \geq 0$, $L = \Lambda + B$ is min when $\hat{\eta} = \eta$ thus proper. $\square$

Savage's Theorem: Example





$\ell(0, \hat{\eta}) = -\log(1 - \hat{\eta})$
$\ell(1, \hat{\eta}) = -\log(\hat{\eta})$

$\eta \mapsto L(\eta, 0.14)$

$\eta \mapsto L(\eta, \eta)$

Bregman Divergence

Definition (Bregman Divergence)

Given a convex function $\phi : \mathbb{R} \to \mathbb{R}$ its Bregman Divergence is

$$B_\phi(s, s_0) = \phi(s) - \phi(s_0) - \langle s - s_0, \nabla\phi(s_0)\rangle$$

Bregman Divergence

Definition (Bregman Divergence)

Given a convex function $\phi : \mathbb{R} \to \mathbb{R}$ its Bregman Divergence is

$$B_\phi(s, s_0) = \phi(s) - \phi(s_0) - \langle s - s_0, \nabla\phi(s_0)\rangle$$

The Savage result immediately shows the following

Corollary

If ℓ is a proper loss then its point-wise regret

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

is a Bregman divergence B_ϕ with $\phi = -\underline{L}$

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

Integral Representation

Theorem (Schervish, 1989 and others)

Given a proper loss $\ell : \mathcal{Y} \times [0,1] \to \mathbb{R}$ there exists a (general) weight function $w(c)$ such that

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

Cost-weighted misclassification losses:

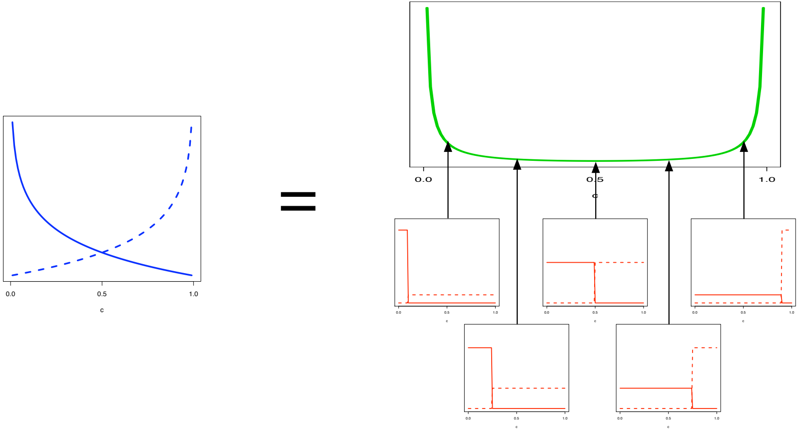
$$\ell_c(y, \hat{\eta}) = \begin{cases} c & y = 0, \hat{\eta} \geq c \quad \text{False Positve} \\ (1 - c) & y = 1, \hat{\eta} < c \quad \text{False Negative} \end{cases}$$

Weight function:

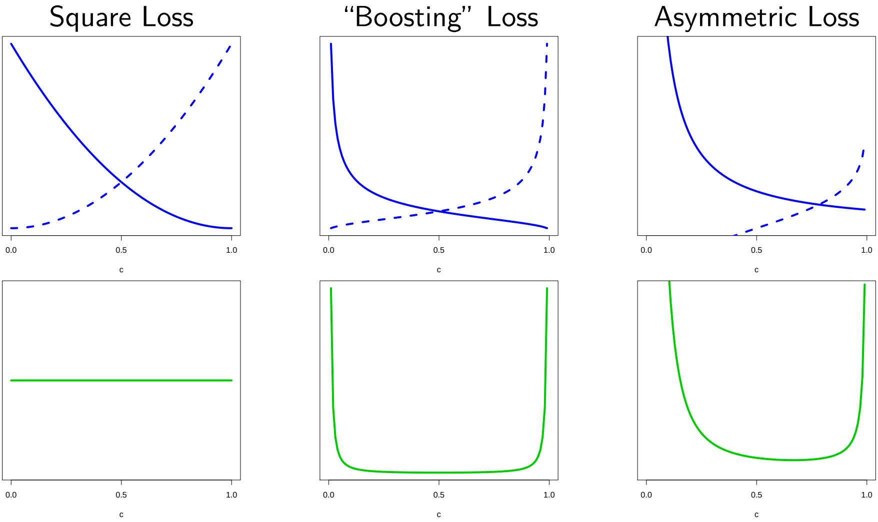
$$w(c) = -\underline{L}''(c)$$

Integral Representation: Example

$$\begin{aligned} \ell(1,\hat{\eta}) &= -\log(\hat{\eta}) \\ \ell(0,\hat{\eta}) &= -\log(1-\hat{\eta}) \end{aligned} \implies w(c) = \frac{1}{(1-c)c}$$



Integral Representation: Examples



Integral Representation: Proof Sketch

Proof Sketch.

Taylor's theorem on $\underline{L}$ gives

$$\begin{aligned}
\underline{L}(\eta) &= \underbrace{\underline{L}(\hat{\eta}) + (\eta - \hat{\eta})\underline{L}'(\hat{\eta})}_{L(\eta,\hat{\eta})} + \int_0^1 g_c(\eta, \hat{\eta})\, \underline{L}''(c)\, dc \\
L(\eta, \hat{\eta}) &= \underline{L}(\eta) - \int_0^1 g_c(\eta, \hat{\eta})\, \underline{L}''(c)\, dc \\
\ell(y, \hat{\eta}) &= \underline{L}(y) + \int_0^1 g_c(y, \hat{\eta})\, w(c)\, dc
\end{aligned}$$

where $w(c) = -\underline{L}''(c)$ since $L(y, \hat{\eta}) = \ell(y, \hat{\eta})$ for $y \in \{0, 1\}$.
Letting $\ell_c = g_c$ and recalling $\underline{L}(0) = \underline{L}(1) = 0$ gives result. □

Integral Representation: Corollaries

Point-wise Risk

$$L(\eta, \hat{\eta}) = \mathbb{E}_\eta[\ell(Y, \hat{\eta})] = \int_0^1 L_c(\eta, \hat{\eta})\, w(c)\, dc$$

where $L_c(\eta, \hat{\eta}) = \mathbb{E}_\eta[\ell_c(Y, \hat{\eta})] = \min((1-\eta)c, (1-c)\eta)$.

Integral Representation: Corollaries

Point-wise Risk

$$L(\eta, \hat{\eta}) = \mathbb{E}_\eta[\ell(Y, \hat{\eta})] = \int_0^1 L_c(\eta, \hat{\eta})\, w(c)\, dc$$

where $L_c(\eta, \hat{\eta}) = \mathbb{E}_\eta[\ell_c(Y, \hat{\eta})] = \min((1-\eta)c, (1-c)\eta)$.

Point-wise Regret

$$B_c(\eta, \hat{\eta}) = \begin{cases} |\eta - c| & \min(\eta, \hat{\eta}) < c \leq \max(\eta, \hat{\eta}) \\ 0 & \text{otherwise} \end{cases}$$

and so

$$B(\eta, \hat{\eta}) = \int_0^1 B_c(\eta, \hat{\eta})\, w(c)\, dc = \int_{\min(\eta,\hat{\eta})}^{\max(\eta,\hat{\eta})} |\eta - c|\, w(c)\, dc$$

Results

Surrogate Regret Bounds: Theorem

Theorem (Theorem 3 in Paper)

Suppose $B_{c_0}(\eta, \hat{\eta}) = \alpha$ for a $c_0 \in (0, 1)$.
Then for any proper loss ℓ the following tight bound holds:

$$B(\eta, \hat{\eta}) \geq \max\{\beta_{c_0}(\alpha), \beta_{c_0}(-\alpha)\}$$

where $\beta_{c_0}(\alpha) = B(c_0 + \alpha, c_0)$.

Surrogate Regret Bounds: Theorem

Theorem (Theorem 3 in Paper)

Suppose $B_{c_0}(\eta, \hat{\eta}) = \alpha$ for a $c_0 \in (0, 1)$.
Then for any proper loss ℓ the following tight bound holds:

$$B(\eta, \hat{\eta}) \geq \max\{\beta_{c_0}(\alpha), \beta_{c_0}(-\alpha)\}$$

where $\beta_{c_0}(\alpha) = B(c_0 + \alpha, c_0)$.

Proof.

When $\hat{\eta} \leq c_0 < \eta$ we have $B_{c_0}(\eta, \hat{\eta}) = \eta - c_0 = \alpha$ and so
$\hat{\eta} \leq c_0 < \eta = c_0 + \alpha$. Thus,

$$B(\eta, \hat{\eta}) = B(c_0 + \alpha, \hat{\eta}) \geq B(c_0 + \alpha, c_0) = \beta_{c_0}(\alpha).$$

Similarly for $\eta \leq c_0 < \eta$. $\square$

Surrogate Regret Bounds: Corollary

We say a loss is symmetric if, for all $\hat{\eta} \in [0,1]$ $\ell(1,\hat{\eta}) = \ell(0, 1-\hat{\eta})$.
All margin losses are symmetric.

Corollary

If ℓ is symmetric and $B(\eta,\hat{\eta}) = \alpha$ then

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

Surrogate Regret Bounds: Corollary

We say a loss is symmetric if, for all $\hat{\eta} \in [0,1]$ $\ell(1,\hat{\eta}) = \ell(0, 1-\hat{\eta})$.
All margin losses are symmetric.

Corollary

If ℓ is symmetric and $B(\eta,\hat{\eta}) = \alpha$ then

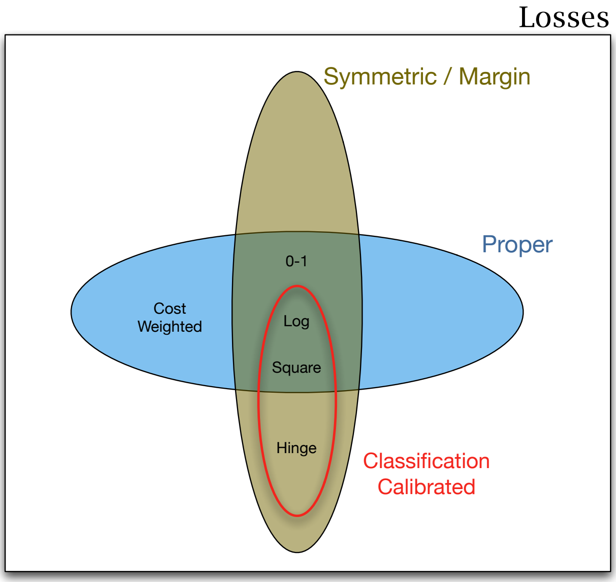
$$B(\eta,\hat{\eta}) \geq \underline{L}(\tfrac{1}{2}) - \underline{L}(\tfrac{1}{2} + \alpha).$$

Example (Square Loss Bound)

For square loss $\underline{L}(\eta) = (1-\eta)\eta$ so

$$\begin{aligned} B(\eta,\hat{\eta}) &\geq \tfrac{1}{4} - [1 - (\tfrac{1}{2} + B_{\frac{1}{2}}(\eta,\hat{\eta}))(\tfrac{1}{2} + B_{\frac{1}{2}}(\eta,\hat{\eta}))] \\ \iff B_{\frac{1}{2}}(\eta,\hat{\eta}) &\leq \sqrt{B(\eta,\hat{\eta})} \end{aligned}$$

Losses


Losses
Symmetric / Margin
Proper
0-1
Cost
Weighted
Log
Square
Hinge
Classification
Calibrated

Convex Composite Proper Losses

Theorem (Theorem 5 in Paper)

Let ℓ be a proper loss and ψ a link. Then the composite risk $L(\eta, \psi^{-1}(h))$ is convex in h when $\psi = -\underline{L}'$.

Convex Composite Proper Losses

Theorem (Theorem 5 in Paper)

Let ℓ be a proper loss and ψ a link. Then the composite risk $L(\eta, \psi^{-1}(h))$ is convex in h when $\psi = -\underline{L}'$.

Proof.

Let $\hat{\eta}_h = \psi^{-1}(h)$ and use Savage and inverse function theorems

$$\begin{aligned} \frac{\partial}{\partial h} L(\eta, \hat{\eta}_h) &= (\eta - \hat{\eta}_h) \frac{\underline{L}''(\hat{\eta}_h)}{\psi'(\hat{\eta}_h)} \\ &= (\hat{\eta}_h - \eta) \end{aligned}$$

since $\psi' = -\underline{L}''$. So

$$\frac{\partial^2}{\partial h^2} L(\eta, \hat{\eta}_h) = \frac{1}{\psi'(\hat{\eta}_h)} = \frac{1}{-\underline{L}''(\hat{\eta}_h)} \geq 0$$

since $\underline{L}$ is concave. □

Conclusions

Conclusions

Proper losses are the “right” loss for probability estimation and make for good surrogates for classification.

- Point-wise Bayes risk is easy to analyse
- Rich structure via Savage’s Theorem and integral representation

Conclusions

Proper losses are the "right" loss for probability estimation and make for good surrogates for classification.

- Point-wise Bayes risk is easy to analyse
- Rich structure via Savage's Theorem and integral representation

The weight functions characterise proper losses.

- Can interpret as which probabilities are important
- Large $w(\eta)$ means "must estimate η well"

Conclusions

Proper losses are the “right” loss for probability estimation and make for good surrogates for classification.

- Point-wise Bayes risk is easy to analyse
- Rich structure via Savage’s Theorem and integral representation

The weight functions characterise proper losses.

- Can interpret as which probabilities are important
- Large $w(\eta)$ means “must estimate η well”

Future work:

- Principled ways of choosing good surrogate losses?
- Better characterisation of convexity for losses?

Thank You!

Psst! Looking for a Post-Doc position?
Come speak to Bob Williamson or myself after the talk...