# Loss Functions for Binary Class Probability Estimation and Classification: Structure and Applications

Andreas Buja [1] Werner Stuetzle [2] Yi Shen [3]

November 3, 2005

**Abstract**

What are the natural loss functions or fitting criteria for binary class probability estimation? This question has a simple answer: so-called "proper scoring rules", that is, functions that score probability estimates in view of data in a Fisher-consistent manner. Proper scoring rules comprise most loss functions currently in use: log-loss, squared error loss, boosting loss, and as limiting cases cost-weighted misclassification losses. Proper scoring rules have a rich structure:

• Every proper scoring rules is a mixture (limit of sums) of cost-weighted misclassification losses. The mixture is specified by a weight function (or measure) that describes which misclassification cost weights are most emphasized by the proper scoring rule.

• Proper scoring rules permit Fisher scoring and Iteratively Reweighted LS algorithms for model fitting. The weights are derived from a link function and the above weight function.

• Proper scoring rules are in a 1-1 correspondence with information measures for tree-based classification.

• Proper scoring rules are also in a 1-1 correspondence with Bregman distances that can be used to derive general approximation bounds for cost-weighted misclassification errors, as well as generalized bias-variance decompositions.

We illustrate the use of proper scoring rules with novel criteria for 1) Hand and Vinciotti's (2003) localized logistic regression and 2) for interpretable classification trees. We will also discuss connections with exponential loss used in boosting.

**Keywords:** Boosting, stagewise regression, machine learning, proper scoring rules, proper score functions, information measures, entropy, Gini index, Bregman distances, link functions, binary response data, stumps, tree-based classification, CART, logistic regression, Fisher scoring, iteratively reweighted least squares, stagewise fitting.

[1]Statistics Department, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104-6302; www.wharton.upenn.edu/~buja. Part of the work performed while on the technical staff of AT&T Labs.

[2]Department of Statistics, University of Washington, Seattle, WA 98195-4322; wxs@stat.washington.edu. Research partially supported by NSF grant DMS-9803226. Part of the work performed on sabbatical leave at AT&T Labs.

[3]Statistics Department, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104-6302; www.wharton.upenn.edu/~shenyi.

# 1 Introduction

Consider predictor-response data with a binary response $y$ representing the observation of classes $y = 1$ and $y = 0$. Such data are thought of as realizations of a Bernoulli random variable $Y$ with $\eta = P[Y = 1]$. The class 1 probability $\eta$ is interpreted as a function of predictors $\mathbf{x}$: $\eta = \eta(\mathbf{x})$. If the predictors are realizations of a random vector $\mathbf{X}$, then $\eta(\mathbf{x})$ is the conditional class 1 probability given $\mathbf{x}$: $\eta(\mathbf{x}) = P[Y = 1|\mathbf{X} = \mathbf{x}]$. Of interest are two types of problems:

- *Classification*: Estimate a region in predictor space in which class 1 is observed with the greatest possible majority. This amounts to estimating a region of the form $\{\eta(\mathbf{x}) > c\}$.
- *Class probability estimation*: Approximate $\eta(\mathbf{x})$ as well as possible by fitting a model $q(\mathbf{x}, \beta)$ ($\beta$ = parameters to be estimated).

Of the two problems, classification is prevalent in machine learning ("concept learning" in AI), whereas class probability estimation is prevalent in statistics (usually as logistic regression). — The classification problem is peculiar in that estimation of a class 1 region is often based on two kinds of criteria:

- the primary criterion of interest: misclassification loss/rate/error. This is an intrinsically unstable criterion for estimating models, a fact that motivates the use of
- surrogate criteria for estimation, such as log-loss (logistic regression) and exponential loss (boosting). These are just estimation devices and not of primary interest (see, for example, Lugosi and Vayatis 2004).

In a sense that can be made precise the surrogate criteria of classification are exactly the primary criteria of class probability estimation. It seems therefore that classification goes the detour via class probability estimation. This is a contentious issue on which we will comment below. — We turn to two standard examples of surrogate criteria:

- Log-loss: $\mathbf{L}(y|q) = -\log(q^y(1-q)^{1-y}) = -y\,\log(q) - (1-y)\,\log(1-q)$
- Squared error loss : $\mathbf{L}(y|q) = (y-q)^2 = y\,(1-q)^2 + (1-y)\,q^2$

Log-loss is the negative log-likelihood of the Bernoulli model. Its expected value, $-\eta\log(q) - (1-\eta)\log(1-q)$, is called Kullback-Leibler loss or cross-entropy. The equality for squared error loss holds because $y \in \{0, 1\}$. Note that its expected value is $\eta\,(1-q)^2 + (1-\eta)\,q^2$, not $(\eta - q)^2$. — Both losses penalize an estimate $q$ of $\eta$ in light of an observation $y$. They yield Fisher consistent estimates of $\eta$ in the sense that

$$\eta = \text{argmin}_{q\in[0,1]}\ \mathbf{E}_y\,\mathbf{L}(y|q)\ , \quad \text{for}\ \ y \sim \text{Bernoulli}(\eta)\ .$$

Loss functions $\mathbf{L}(y|q)$ with this property have been known as **proper scoring rules**. In subjective probability they are used to judge the quality of probability forecasts by experts, whereas here they are used to judge the quality of class probabilities estimated by automated

procedures. [Proper scoring rules are Fisher consistent for class probability estimation, whereas Y. Lin's (2002) notion of Fisher consistency applies to classification. This is a weaker notion as it only requires "$\text{argmin}_{q\in[0,1]} \mathbf{E}_y \mathbf{L}(y|q) \gtrless 0.5$" iff "$\eta \gtrless 0.5$".]

In light of current interest in boosting we indicate how boosting's exponential loss maps to a proper scoring rule: Friedman, Hastie and Tibshirani (2000) observed that estimates produced by boosting can be mapped to class probability estimates with a certain link function. This link function can be used to transport exponential loss to the probability scale where it turns into a proper scoring rule as follows (see Section 4):

$$\mathbf{L}(y|q) \;=\; y \left(\frac{1-q}{q}\right)^{1/2} + (1-y) \left(\frac{q}{1-q}\right)^{1/2} \quad .$$

We will refer to this novel proper scoring rule as "**boosting loss**".

Proper scoring rules have a simple structure in terms of an integral representation due to Shuford, Albert and Massengill (1966), Savage (1971), and in its most general form by Schervish (1989). We give this representation an interpretation that lends it new meaning. One way to introduce this matter is in geometric terms: proper scoring rules form a non-negative convex cone that is closed if one includes what we call "non-strict" proper scoring rules. It is then natural to expect that the cone elements have integral representations in terms of extremal elements. It turns out that the extremal elements are exactly the cost-weighted misclassification errors:

$$\mathbf{L}_c(y|q) \;=\; y\,(1-c)\cdot 1_{[1-q\geq 1-c]} \;+\; (1-y)\,c\cdot 1_{[q>c]} \;,$$

where w.l.o.g. we chose the costs $c,\ 1-c \in [0,1]$ to sum to one. The conjunction "$y = 0$ & $q > c$" describes "false positives" and "$y = 1$ & $q \leq c$" "false negatives", and the values $c$ and $1-c$ are the respective costs. The Shuford-Albert-Massengil-Savage-Schervish theorem is then nothing other than a Choquet-type integral representation of proper scoring rules in terms of extremal elements:

$$\mathbf{L}(y|q) \;=\; \int_0^1 \mathbf{L}_c(y|q)\,\omega(dc) \;.$$

This representation characterizes proper scoring rules in terms of measures $\omega(dc)$. If $\omega(dc)$ is absolutely continuous w.r.t. Lebesgue measure, we write by abuse of notation $\omega(dc) = \omega(c)\,dc$. The weight function $\omega()$ is fundamental to theory, methodology, and algorithms:

- **Algorithms**: A unifying feature of proper scoring rules is that they can all be minimized with Fisher scoring and Iteratively Reweighted Least Squares (IRLS) algorithms. The IRLS weights derive directly from the weight function $\omega()$ (Section 9).
- **Methodology**: The wealth of loss functions granted by the integral representation can be used for tailoring losses to specific classification problems with cut-offs other than 1/2 of $\eta(\mathbf{x})$. Such tailoring is achieved by designing suitable weight functions $\omega()$ (Section 12). Because the integral representation carries over to information measures, tailoring can also be applied to classification trees.

- **Theory**: The integral representation, when carried over to Bregman distances (Section 19-22), lends itself to the derivation of bounds on cost-weighted misclassification losses in terms of $\omega()$, thereby generalizing approximation theorems such as Zhang's (2004) and Bartlett et al.'s (2003; 2004 p. 87) to estimation with arbitrary proper scoring rules and classification with arbitrary cost weights $c$.

An immediate benefit of the weight functions is their interpretability: Locations where $\omega(c)$ puts most mass...

- ...correspond to the cutoffs $c$ for which the estimated class probabilities attempt the most accurate classification, that is, the most accurate estimation of $\{\eta(\mathbf{x}) > c\}$;
- ...determine the observations that get the most weight in the minimization of the proper scoring rule. This can be seen from the IRLS iterations in which observations are upweighted when $\omega(q)$ is large.

For example, from the form of the weight function for log-loss, which is $\omega(q) = (q(1-q))^{-1}$, one infers immediately a heavy reliance on extreme probability estimates. This has indeed been an issue in logistic regression; see Hand and Vinciotti (2003) for a detailed discussion. Surprisingly boosting loss is even more lopsided than log-loss in its reliance on extreme probability estimates: $\omega(q) = (q(1-q))^{-3/2}$. If boosting loss looks more problematic than log-loss, why would boosting be so successful in practice? The answer is most likely that boosting is successful not because but in spite of its loss function. The loss function is benign if used for classification based on non-parametric models (as in boosting), but boosting loss is certainly not more successful than log-loss if used for fitting linear models as in linear logistic regression.

What guidance can one give for choosing proper scoring rules? One can follow Hand and Vinciotti's (2003) lead and choose weight functions $\omega()$ that concentrate mass around the classification threshold $c$ of interest. Hand and Vinciotti (2003) achieved this effect by modifying the IRLS algorithm and upweighting observations with estimates $q$ near the threshold $c$. This, however, is nothing other than minimizing a proper scoring rule whose weight function $\omega()$ concentrates mass around the threshold $c$. Adapting loss functions in this way will be called "**tailoring the proper scoring rule to the classification threshold** $c$". We will illustrate successful tailoring with Hand and Vinciotti's (2003) artificial data example and with a well-known real dataset (Section 14).

Tailoring proper scoring rules should be most effective for models that are likely to have substantial bias for class probability estimation, such as linear models, while more flexible nonparametric models are less likely to gain. Among the latter are boosting approaches that rely on very flexible function classes such as sums of shallow trees. Although the present work was motivated by boosting, the main benefit of tailored proper scoring rules may be not so much in boosting as in classical linear models, but also in tree-based classification as we indicate next.

Most well-known tree algorithms rely on one of two splitting criteria: entropy or the Gini index. These criteria derive directly from proper scoring rules (Section 17): entropy

from log-loss, and the Gini index from squared error loss. In fact, every proper scoring rule defines an information measure that can be used as a splitting criterion, hence tailoring can be applied to tree-growing as well. Trees, however, form a highly flexible class of fits, so one would expect little benefit from tailoring the criterion. Yet, tailoring for trees turns out to be useful when focusing one-sidedly on extreme probabilities, for example those near one. To this end one can design proper scoring rules that put progressively more weight on larger and larger probabilities. This produces unusually interpretable trees that layer the data in highly unbalanced, cascading trees. This form of trees was proposed by Buja and Lee (2001), but the splitting criteria used there were heuristic and did not derive from information measures.

Finally a word on the relation of this work to boosting: Although motivated by Friedman, Hastie and Tibshirani's (2000) interpretation of boosting, the present analysis is of greater use for techniques other than boosting. One of our contributions to boosting consists of showing that IRLS can be tweaked for fitting stagewise additive models with arbitrary proper scoring rules. For those interested in the original interpretation of boosting as reweighting, we observe that IRLS schemes often produce weights that depend on the response values $y_n$ only through the estimated class-1 probabilities $q(\mathbf{x}_n)$, not the $y_n$'s directly. This is the case when IRLS is used to implement Fisher scoring as opposed to Newton steps, and even for Newton steps if the link function is *canonical* with regard to the weight function $\omega(q)$. The notion of "canonical" is the same as in generalized linear models but it applies to all proper scoring rules. Unlike logistic loss, exponential loss does not decompose into a pair of canonical link and weight functions.

There has been recent theory in the context of boosting that may not directly explain why boosting is successful but may nevertheless have merit on its own terms (for example, Bartlett et al. 2003, and the "Three Papers on Boosting": Jiang; Lugosi and Vayatis; Zhang; 2004). As Bartlett et al. (2004, p. 86f) note, a common first step of such theories consists of "comparison theorems" that bound the excess misclassification risk by the excess surrogate risk, where "excess" refers to excess over the best achievable risk at the population. To date such bounds have been derived individually for special choices of surrogate loss functions and unweighted misclassification losses. We contribute to these theories by deriving bounds for *arbitrary cost-weighted misclassification losses* in terms of surrogate losses that are *arbitrary proper scoring rules*. The only case we do not cover is hinge loss used in support vector machines because it is related to absolute deviation loss $\mathbf{L}(y|\eta) = |y - \eta|$ which is not a proper scoring rule. Thus SVMs seem to be the only case that truly bypasses estimation of class probabilities and directly aims at classification (see Freund and Schapire (2004, p. 114) along similar lines). Our contribution, however, is not about the pros and cons of class probability estimation but about its structural connection with cost-weighted classification.

**Acknowledgments**: We thank J.H. Friedman, D. Madigan, D. Mease and A.J. Wyner for discussions that resulted in several clarifications, and J. Berger and L. Brown for pointing us to the literature on proper scoring rules. We owe a special debt to two articles: Friedman, Hastie and Tibshirani (2000) and Hand and Vinciotti (2003). In what follows we refer to these articles as **FHT (2000)** and **HV (2003)**, respectively.

## 2 Definition of proper scoring rules

Given predictor-response data $(\mathbf{x}_n, y_n)$ with binary response $y_n \in \{0, 1\}$, we are interested in fitting a model $q(\mathbf{x})$ for the conditional class 1 probabilities $\eta(\mathbf{x}) = P[Y = 1|\mathbf{x}]$. (For now it is immaterial whether the model is parametric or nonparametric, which is why we write $q(\mathbf{x})$ without unknown parameters.) This problem would be approached by most statisticians with maximum likelihood with regard to the conditional Bernoulli model, but there exist many other possibilities, and their exploration is the purpose of this article.

We assume the model $q(\mathbf{x})$ takes on values in $[0, 1]$ for estimating $\eta(\mathbf{x})$. The model is to be fitted to the values 0 and 1 of the response $y$ by minimizing a loss function which we take to be the sample average of losses of individual observations:

$$\mathcal{L}(q()) \;=\; \frac{1}{N}\sum_{n=1}^{N} \mathbf{L}(y_n|q_n) \;. \tag{1}$$

Assuming $q()$ varies in a sufficiently rich function class, and in view of the standard population identity

$$\mathbf{E}\; \mathcal{L}(q()) \;=\; \mathbf{E}_{\mathbf{X},Y}\; \mathbf{L}(Y|\, q(\mathbf{X})) \;=\; \mathbf{E}_{\mathbf{X}}\left(\mathbf{E}_{Y|\mathbf{X}}\; \mathbf{L}(Y|\, q(\mathbf{X}))\right) \;,$$

minimization of $\mathcal{L}$ creates Fisher consistent estimates $q(\mathbf{x})$ of $\eta(\mathbf{x})$ if Fisher consistency holds pointwise:

$$\text{argmin}_{q\in[0,1]}\; \mathbf{E}_{Y\sim Bernoulli(\eta)}\, \mathbf{L}(Y|\, q) \;=\; \eta \;, \qquad \forall\; \eta \in [0, 1]. \tag{2}$$

Fisher consistency is the condition that underlies everything that follows; it will be shown to be the defining property of "proper scoring rules".

We next re-express condition (2) by making use of Bernoulli-related simplifications. Because $Y$ takes on only two values, $\mathbf{L}(y|q)$ consists of only two functions of $q$: $\mathbf{L}(1|q)$ and $\mathbf{L}(0|q)$, which we call "partial losses". We anticipate $\mathbf{L}(0|q)$ as increasing in $q$ because values of $q$ closer to $y = 0$ are a better fit, and similarly we anticipate $\mathbf{L}(1|q)$ as a decreasing function of $q$. Because we prefer to express both in terms of increasing functions, we define

$$L_1(1 - q) = \mathbf{L}(1|q) \;, \qquad L_0(q) = \mathbf{L}(0|q) \;. \tag{3}$$

The pointwise empirical loss for the response $y$ and the estimate $q$ can now be written as

$$\mathbf{L}(y|q) \;=\; y\, L_1(1 - q) + (1 - y)\, L_0(q) \;,$$

and the pointwise expected loss or risk as

$$\mathbf{R}(\eta|q) \;=\; \mathbf{E}_Y \mathbf{L}(Y|q) \;=\; \eta\, L_1(1 - q) + (1 - \eta)\, L_0(q) \;. \tag{4}$$

Requirement (2) for Fisher consistency becomes

$$\text{argmin}_q\, \mathbf{R}(\eta|q) \;=\; \eta \;.$$

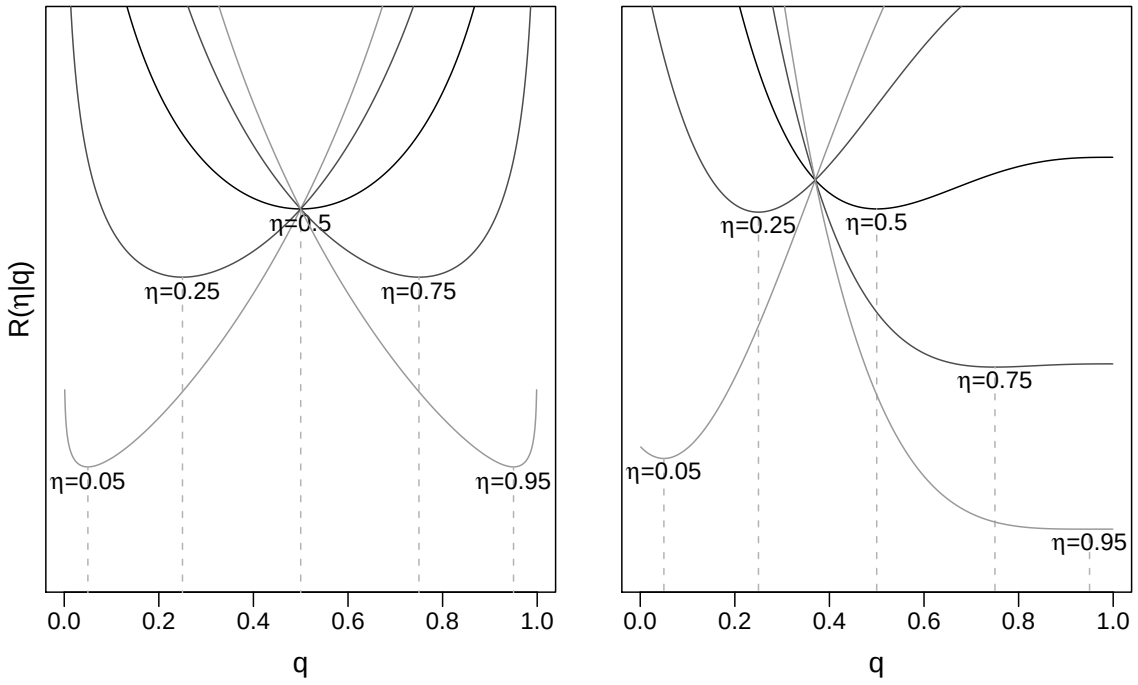


Figure 1: *Depiction of two proper scoring rules. The curves represent sections of expected loss,* $q \mapsto \mathbf{R}(\eta\,|q)$ *for a few fixed values of* $\eta$*. By the definition of proper scoring rules the minimizing value* $q$ *for given* $\eta$ *is* $q = \eta$*.*
*Left: log-loss or Beta loss with* $\alpha = \beta = 0$*; see Section 11. This loss is symmetric about 0.5. Right: Beta loss with* $\alpha = 1,\ \beta = 3$*. This loss is not symmetric about 0.5. It is tailored for classification with misclassification cost* $c = \frac{\alpha}{\alpha+\beta} = 0.25$ *(see Section 12).*

This condition is known in the literature on subjective probability and probability forecasting from which we adopt the following definitions:

**Definition:** *If the expected loss* $\mathbf{R}(\eta|q)$ *is minimized w.r.t.* $q$ *by* $q = \eta\ \forall\eta \in (0,1)$*, we call the loss function a "proper scoring rule". If moreover the minimum is unique, we call it a "strict" proper scoring rule, otherwise "non-strict".*

For a visual illustration of proper scoring rules, see Figure 1. The peculiar expression "proper scoring rule" stems from a large literature on subjective probability, economics, forecasting, and meteorology. See the following references and citations therein: Shuford, Albert and Massengill (1966), Savage (1971), DeGroot and Fienberg (1983), Murphy and Daan (1985), Schervish (1989), Winkler (1993). In subjective probability proper scoring rules are economic incentive systems that elicit a subject's true belief with regard to a probability $\eta$. In probability forecasting proper scoring rules are used as scores for measuring the quality of predictions such as "the chance of rain tomorrow is 60%". In the present article, proper scoring rules are used as loss functions for fitting models to binary response data.

Proper scoring rules have obvious extensions to arbitrary statistical models. In the foundations of Bayes theory they are utilities that are defined on probability distributions and

whose optimization results in "honest" inference reporting policies. See for example Bernardo and Smith (2000, Def. 2.21 and 3.16). For more background and modern uses of proper scoring rules the reader should consult a wide-ranging discussion by Gneiting and Raftery (2004).

# 3 Commonly used proper scoring rules

For most loss functions the partial losses $L_1(1-q)$ and $L_0(q)$ are smooth, hence the proper scoring rule property implies the stationarity condition

$$0 \;=\; \frac{\partial}{\partial q}\bigg|_{q=\eta} \mathbf{R}(\eta|q) \;=\; -\,\eta\, L_1'(1-\eta) + (1-\eta)\, L_0'(\eta)\ ,$$

that is,

$$\eta\, L_1'(1-\eta) \;=\; (1-\eta)\, L_0'(\eta)\ , \tag{5}$$

which is what one uses to check the property in the first two examples below.

- The most common proper scoring rule in statistics is *log-loss*, with the following associated fitting criterion, partial losses and expected loss:

$$\begin{aligned}
\mathcal{L}(q()) \;&=\; \frac{1}{N}\sum_{n=1}^{N} \left[-y_n \log(q_n) - (1-y_n)\log(1-q_n)\right]\ ,\\
L_1(1-q) &= -\log(q)\ ,\quad L_0(q) = -\log(1-q)\ ,\\
\mathbf{R}(\eta|q) \;&=\; -\,\eta\log(q) - (1-\eta)\log(1-q)\ ,
\end{aligned}$$

  where $q_n = \psi(\mathbf{b}^T\mathbf{x}_n)$ and $\psi()$ is the logistic link. Log-loss is sometimes called Kullback-Leibler loss or the cross-entropy term of the Kullback-Leibler divergence.

- Another common proper scoring rule is *squared error loss*:

$$\begin{aligned}
\mathcal{L}(q()) \;&=\; \frac{1}{N}\sum_{n=1}^{N} \left[y_n(1-q_n)^2 + (1-y_n)q_n^2\right]\ ,\\
L_1(1-q) &= (1-q)^2\ ,\quad L_0(q) = q^2\ ,\\
\mathbf{R}(\eta|q) \;&=\; \eta\,(1-q)^2 + (1-\eta)\,q^2\ .
\end{aligned}$$

  The fitting criterion $\mathcal{L}(q())$ in this case is just the residual sum of squares with binary response values. In the literature on proper scoring rules, squared error loss is known as Brier score (Brier 1950). [Squared error loss is usually used with the identity link function which often leads to estimates $q$ outside of $[0,1]$. It would be more meaningful to use here, too, a non-trivial link function such as the logistic.]

- A third example is *misclassification loss*:

$$\begin{aligned}
\mathcal{L}(q()) \;&=\; \frac{1}{N}\sum_{n=1}^{N} \left[y_n 1_{[q_n\le 0.5]} + (1-y_n)1_{[q_n>0.5]}\right]\ ,\\
L_1(1-q) &= 1_{[q\le 0.5]}\ ,\quad L_0(q) = 1_{[q>0.5]}\ ,\\
\mathbf{R}(\eta|q) \;&=\; \eta\, 1_{[q\le 0.5]} + (1-\eta)\, 1_{[q>0.5]}\ .
\end{aligned}$$

While the previous two examples are strict proper scoring rules, misclassification loss is non-strict. The definition requires some convention for $q = 0.5$, but the particular choice is irrelevant for most purposes.

# 4 A proper scoring rule derived from exponential loss

We derive a new type of proper scoring rule that is implicit in boosting. As this is not the place to discuss the motivations and details of boosting, we refer the reader to FHT (2000) for an introduction that is accessible to statisticians. The one detail we need here is that boosting can be interpreted as minimization of so-called "exponential loss", defined as

$$\frac{1}{N}\sum_{n=1}^{N} e^{-(2y_n-1)\,F(\mathbf{x}_n)} \;=\; \frac{1}{N}\sum_{n=1}^{N}\left[y_n\, e^{-F(\mathbf{x}_n)} + (1-y_n)e^{F(\mathbf{x}_n)}\right] \;, \tag{6}$$

where $F(\mathbf{x})$ is *not* a prediction of the conditional class probability $\eta(\mathbf{x})$, but of a transformation thereof, just as logistic regression predicts the logit and not the class probability directly. Classification is performed by predicting class 1 when $F(\mathbf{x}) > 0$ and class 0 otherwise. In analogy to the terminology of Section 2, we define the pointwise observed loss as

$$y\, e^{-F} + (1-y)\, e^{F}$$

and the pointwise expected loss as

$$\eta\, e^{-F} + (1-\eta)\, e^{F} \;. \tag{7}$$

The $F$-values are unconstrained on the real line and, as mentioned, cannot be interpreted as probabilities. However, FHT (2000, p. 345f) have in effect shown that the following transformation maps $F$-values to Fisher-consistent estimates of probabilities $\eta$:

$$q = \frac{1}{1+e^{-2\,F}} \;.$$

The reason is that, for given $\eta$, the minimizer of the expected exponential loss (7) is half the logit of $\eta$:

$$F_{min}(\eta) = \frac{1}{2}\log\frac{\eta}{1-\eta} \;. \tag{8}$$

Using $F(q) = \frac{1}{2}\log\frac{q}{1-q}$ as a link function, it is now natural to map the exponential loss to the probability scale, a step not taken by FHT (2000). Substituting $F(q)$ in the expected exponential loss (7) results in

$$\mathbf{R}(\eta|q) = \eta\,\left(\frac{1-q}{q}\right)^{1/2} + (1-\eta)\,\left(\frac{q}{1-q}\right)^{1/2} \;,$$

which by construction is a proper scoring rule: If $F = F_{min}(\eta)$ is the minimizer of (7) with regard to $F$, then $q = \eta$ is the minimizer of $\mathbf{R}(\eta|q)$ with regard to $q$. We will therefore refer to the proper scoring rule defined by

$$L_1(1-q) = \left(\frac{1-q}{q}\right)^{1/2} \quad \text{and} \quad L_0(q) = \left(\frac{q}{1-q}\right)^{1/2}$$

as "boosting loss". Together with log-loss, squared error loss, and misclassification loss, this will be one of the standard examples in what follows.

# 5 Counterexamples of proper scoring rules

**Wald's decision theory** has a notion of loss function whose purpose is to relate estimates (more generally: decisions) *to true parameter values*. Therefore, $|q-\eta|$ is a valid loss function in the sense of Wald, with $\eta$ being the unknown true parameter and $q$ the estimate (decision). Observed losses $\mathbf{L}(y|q)$, however, are estimation devices whose purpose is to relate estimates $q$ to observations $y$. The only connection is that expected losses $\mathbf{R}(\eta|q) = \eta L_1(1-q) + (1-\eta)L_0(q)$ form a proper subset of Wald loss functions, namely those that are affine in $\eta$. Hence the Wald loss function $|q-\eta|$ is not an expected loss because it is not affine in $\eta$.

**Absolute deviation**, $\mathbf{L}(y|q) = |y-q|$ qualifies as an observed loss but it is not a proper scoring rule. It is identical to $\mathbf{L}(y|q) = y\,(1-q) + (1-y)\,q$ because $y \in \{0,1\}$. The corresponding expected loss is *not* $|q-\eta|$ but $\mathbf{R}(\eta|q) = \eta\,(1-q) + (1-\eta)\,q$. Yet this is not a proper scoring rule because $\mathbf{R}(\eta|q)$ is minimized by $q = 1$ for $\eta > 1/2$ and $q = 0$ for $\eta < 1/2$ and arbitrary $q \in [0,1]$ for $\eta = 1/2$. Absolute deviation loss must therefore be considered as a classification loss: it provides Fisher consistent estimates of the majority class, not the class probability. It is similar to misclassification loss of Section 3 in that both have Bayes risk as minimum expected loss:

$$\min_q \mathbf{R}(\eta|q) \;=\; \min(\eta, 1-\eta)\ .$$

Yet absolute deviation and misclassification loss are distinct in that the latter is indifferent to the value of $q$ as long as it is on the same side of $1/2$ as $\eta$. Misclassification loss is a proper scoring rule because it is minimized, although not uniquely, by the true class probability $q = \eta$, whereas absolute deviation is not. — Absolute deviation loss underlies classification with support vector machines; see for example Zhang (2004).

**Power losses** are generalizations of absolute deviation loss:

$$\begin{aligned}
&\mathbf{L}(y|q) \;=\; |y-q|^r \;=\; y\,(1-q)^r + (1-y)\,q^r\ ,\\
&L_1(1-q) \;=\; (1-q)^r\ , \qquad L_0(q) \;=\; q^r\ ,\\
&\mathbf{R}(\eta|q) \;=\; \eta\,(1-q)^r + (1-\eta)\,q^r
\end{aligned}$$

for positive exponent $r$. For the first identiy one makes again use of $y \in \{0,1\}$. The stationarity condition (5) for a proper scoring rule specializes to

$$r\,(1-q)^{r-2} \;=\; r\,q^{r-2}\ ,$$

which is violated for all choices of $r$ except $r = 2$. Hence in the power family only squared error is a proper scoring rule.

**Power divergences of Cressie and Read** (1984), specialized to the two class case:

$$\mathbf{R}(\eta|q) \;=\; \frac{2}{\lambda(\lambda+1)}\left(\eta\left[\left(\frac{\eta}{q}\right)^{\lambda}-1\right]+(1-\eta)\left[\left(\frac{1-\eta}{1-q}\right)^{\lambda}-1\right]\right)\;.$$

These quantities are meant as goodness-of-fit criteria for empirical probabilities $\eta$ and $1-\eta$ vis-à-vis hypothetical probabilities $q$ and $1-q$. This type of power divergences does not result in proper scoring rules because they are not affine in $\eta$ and cannot be made affine, unlike the very different power divergences proposed by Basu, Harris, Hjort and Jones (1998) (see Section 19).

# 6 The structure of proper scoring rules

We treat the simple case of smooth proper scoring rules first and generalize subsequently. The following proposition, in a slightly different form, goes back to Shuford, Albert and Massengill (1966).

**Proposition:** *Assume $L_1(1-q)$ and $L_0(q)$ are differentiable. They form a proper scoring rule iff they satisfy*

$$L_1'(1-q) \;=\; \omega(q)\,(1-q)\;,\qquad L_0'(q) \;=\; \omega(q)\,q\;, \tag{9}$$

*for some weight function $\omega(q)\geq 0$ on $(0,1)$ that satisfies $\int_\epsilon^{1-\epsilon}\omega(t)dt<\infty\;\forall\epsilon>0$. The proper scoring rule is strict iff $\omega(q)>0$ almost everywhere on $(0,1)$.*

The proof follows by writing the stationarity condition (5), $\partial/\partial q|_{q=\eta}\mathbf{R}(\eta|q)=0$, as

$$\frac{L_1'(1-\eta)}{(1-\eta)} \;=\; \frac{L_0'(\eta)}{\eta}\;, \tag{10}$$

and defining this to be $\omega(\eta)$. An elementary calculation shows:

$$\frac{\partial}{\partial q}\mathbf{R}(\eta|q) \;=\; (q-\eta)\;\omega(q)\;, \tag{11}$$

which proves that the local minimum is unique and global when $\omega(q)>0$ in a neighborhood of $\eta$. A few remarks:

- Assuming more smoothness one obtains:

$$\frac{\partial^2}{\partial q^2}|_{q=\eta}\mathbf{R}(\eta|q) \;=\; \omega(\eta)\;. \tag{12}$$

  Together with stationarity $\frac{\partial}{\partial q}|_{q=\eta}\mathbf{R}(\eta|q)=0$, one arrives at the following second order approximation:

$$\mathbf{R}(\eta|q) \;\approx\; \mathbf{R}(\eta|\eta)+\frac{1}{2}\omega(\eta)(q-\eta)^2\;. \tag{13}$$

- The weight function $\omega(q)$ does not need to be globally integrable on $(0,1)$, which is the same as saying that $L_0$ and $L_1$ can be unbounded near 0 and 1.
- $L_1$ and $L_0$ are bounded below iff $\int_0^1 t\,(1-t)\,\omega(t)\,dt < \infty$, which limits power tail weights of $\omega(t)$ to $t^\gamma$ and $(1-t)^\gamma$ with $\gamma > -2$.
- Assuming $L_1$ and $L_0$ bounded below and hence w.l.o.g. normalized to $L_1(0) = L_0(0) = 0$, the proposition yields the following integral representation:

$$L_1(1-q) \;=\; \int_q^1 (1-t)\,\omega(t)\,dt\;, \qquad L_0(q) \;=\; \int_0^q t\,\omega(t)\,dt\;. \tag{14}$$

A larger universe of proper scoring rules can be characterized by the following construction which is a subset of a more general theorem by Schervish (1989, Theorem 4.2).

**Theorem 1:** *Let $\omega(dt)$ be a positive measure on $(0,1)$ that is finite on intervals $(\epsilon,\ 1-\epsilon)$ $\forall\ \epsilon > 0$. Then the following defines a proper scoring rule:*

$$L_1(1-q) \;=\; \int_q^{f_1} (1-t)\,\omega(dt)\;, \qquad L_0(q) \;=\; \int_{f_0}^q t\,\omega(dt)\;. \tag{15}$$

*The proper scoring rule is strict iff $\omega(dt)$ has non-zero mass on every open interval of $(0,1)$.*

Theorem 4.2 of Schervish (1989) has an "only if" part which shows that essentially every proper scoring rule has an integral representation. The formulation of the theorem is intentionally kept informal. Here are a few details to be filled in:

- When the measure $\mu(dt)$ has point masses one needs a convention for integration limits. For example, excluding the upper limit and including the lower limit makes $L_1(1-q)$ and $L_0(q)$ left-continuous, while the reverse convention makes them right-continuous. Averaging these two conventions produces further conventions. The convention is irrelevant as long as it is applied consistently.
- The fixed integration limits, $f_0$ and $f_1$, are somewhat arbitrary because of the fact that if $L_1(1-q)$, $L_0(q)$ form a proper scoring rule, so do $L_1(1-q)+C_1$, $L_0(q)+C_0$ for all constants $C_1$ and $C_0$, and these proper scoring rules are all equivalent.
- Under certain conditions, however, there are restrictions on $f_0$ and $f_1$: we need $f_1 < 1$ if $\int_{1-\epsilon}^1 (1-t)\omega(dt) = \infty$, and consequently $L_1(1-q)$ is unbounded below near $q = 1$. Similarly we need $f_0 > 0$ if $\int_0^\epsilon t\omega(dt) = \infty$, which makes $L_0(q)$ unbounded below near $q = 0$.
- If $\int_0^\epsilon \omega(dt) = \infty$, then $L_l(1-q)$ is unbounded above near 0, and similarly if $\int_{1-\epsilon}^1 \omega(dt) = \infty$, then $L_0(q)$ is unbounded above near 1.
- While two important proper scoring rules are unbounded above, log-loss and boosting loss, we don't know of any proposals that are unbounded below. Thus all common proper scoring rules seem to satisfy $\int_0^1 t(1-t)\omega(dt) < \infty$.

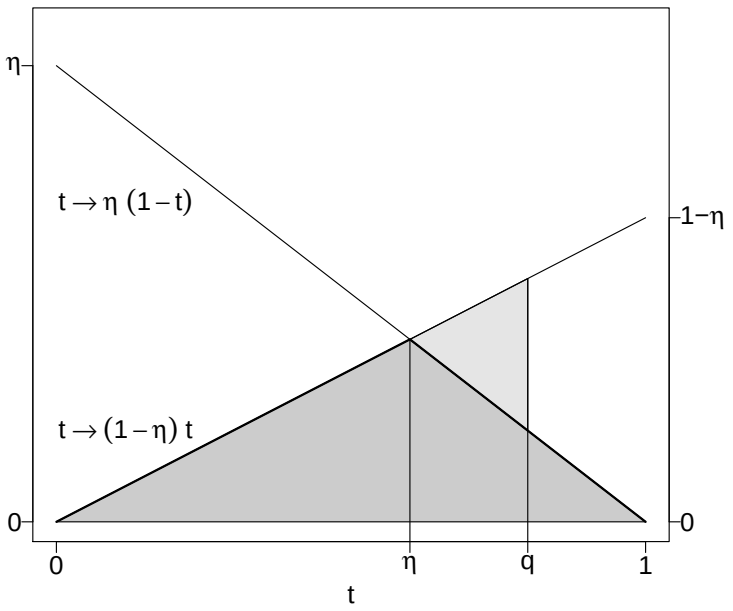


Figure 2: *Illustration for the proof of Theorem 1. The union of shaded areas represents the integrand in Equation (17). It is the lowest for $q = \eta$ when the lightly shaded area vanishes.*

- As in the continuous case, if $\int_0^1 t(1-t)\omega(dt) < \infty$, then both $L_1(1-q)$ and $L_0(q)$ are bounded below and can be normalized to $L_1(0) = L_0(0) = 0$ so that

$$L_1(1-q) \;=\; \int_q^1 (1-t)\,\omega(dt)\;, \qquad L_0(q) \;=\; \int_0^q t\,\omega(dt)\;. \tag{16}$$

The **proof** of Theorem 1 is very short, but to avoid notational difficulties we only give it under the assumption $\int_0^1 t(1-t)\omega(dt) < \infty$ so that we can choose $f_1 = 1$ and $f_0 = 0$. We also use the left-continuous convention for the integrals:

$$\mathbf{R}(\eta|q) \;=\; \eta\,L_1(1-q) + (1-\eta)\,L_0(q) \;=\; \int \big[\,\eta\,(1-t)\,1_{[t\geq q]} + (1-\eta)\,t\,1_{[t<q]}\,\big]\,\omega(dt) \tag{17}$$

The integrand on the right hand side is depicted in Figure 2 as a function of $t$. It is immediate that $q = \eta$ is a minimum. The minimum is unique iff the open intervals $(\eta, \eta+\epsilon)$ and $(\eta-\epsilon, \eta)$ have non-zero mass for all $\epsilon > 0$. — The following is a direct consequence of Figure 2:

**Corollary:** *Any proper scoring rule $q \mapsto \mathbf{R}(\eta|q)$ is non-ascending to the left of $\eta$ and non-descending to the right of $\eta$. If $\omega(dt)$ puts non-zero mass on all open intervals, then the function is strictly descending/ascending to the left/right of $\eta$, and $q = \eta$ is the unique minimum for all $\eta$.*

From the above follows that there exists a wealth of proper scoring rules. Here are the standard examples with their weight functions $\omega(t)$ or measures $\omega(dt)$, respectively:

- Boosting loss:
$$\omega(q) = \frac{1}{[\, q(1-q)\,]^{3/2}} \tag{18}$$
- Log-loss:
$$\omega(q) = \frac{1}{q(1-q)} \tag{19}$$
- Squared error loss:
$$\omega(q) = 1 \tag{20}$$
- Misclassification loss:
$$\omega(dq) = 2\,\delta_{\frac{1}{2}}(dq) \tag{21}$$

In the last example, $\delta_{\frac{1}{2}}$ denotes a point mass at $\frac{1}{2}$.

# 7 Proper scoring rules are mixtures of cost-weighted misclassification losses

The goal of this section is to interpret Theorem 1 as an integral representation of proper scoring rules in terms of cost-weighted misclassification losses. Assume that the two types of misclassification entail differing costs (e.g., the cost of failing to detect a disease versus the cost of falsely detecting a disease). W.l.o.g. we can assume the sum of costs to add to 1 and the costs of correct classification to be zero. The two misclassification costs are therefore

- $c_0 = c$ : the cost of a false positive, that is, of misclassification of $y = 0$ as class 1. Its expected cost is: $\quad c\,P[Y=0] = c\,(1-\eta)$.
- $c_1 = 1 - c$ : the cost of a false negative, that is, of misclassification of $y = 1$ as class 0. Its expected cost is: $\quad (1-c)\,P[Y=1] = (1-c)\,\eta$.

The optimal classification is therefore class 1 iff $\eta\,(1-c) \geq (1-\eta)\,c$, that is, $\eta \geq c$. This shows that cost-weighted classification with costs $c$ and $1-c$ is equivalent to "quantile classification" at the $c$-quantile. Standard classification is median classification or, equivalently, cost-weighted classification with equal costs $c = 1 - c = \frac{1}{2}$.

In the absence of knowledge of $\eta$ but availability of an estimate $q$, we classify as class 1 when $q \geq c$. The cost-weighted misclassification losses (empirical, partial and pointwise expected) can be written as follows:

$$\mathcal{L}_c(\eta|q()) \;=\; \frac{1}{N}\sum_{n=1}^{N} y_n\,(1-c)\cdot 1_{[q_n \leq c]} \;+\; (1-y_n)\,c\cdot 1_{[q_n > c]} \tag{22}$$

$$L_{1,c}(1-q) \;=\; (1-c)\cdot 1_{[q\leq c]}\;, \qquad L_{0,c}(q) \;=\; c\cdot 1_{[q>c]} \tag{23}$$

$$\mathbf{R}_c(\eta|q) \;=\; \eta\,(1-c)\cdot 1_{[q\leq c]} \;+\; (1-\eta)\,c\cdot 1_{[q>c]} \tag{24}$$

There is arbitrariness at $q = c$, but the choice is immaterial. We restate Theorem 1:

**Theorem 1':** *If* $\int_0^1 t(1-t)\,\omega(dt) < \infty$*, the proper scoring rules of Theorem 1 are mixtures of cost-weighted misclassification losses with* $\omega(dc)$ *as the mixing measure:*

$$\mathcal{L}(\eta|q()) \;=\; \int_0^1 \mathcal{L}_c(\eta|q())\,\omega(dc) \tag{25}$$

$$L_1(1-q) \;=\; \int_0^1 L_{1,c}(1-q)\,\omega(dc)\;, \qquad L_0(q) \;=\; \int_0^1 L_{0,c}(q)\,\omega(dc)\;, \tag{26}$$

$$\mathbf{R}(\eta|q) \;=\; \int_0^1 \mathbf{R}_c(\eta|q)\,\omega(dc)\;. \tag{27}$$

The **proof** is immediate by substituting the definitions (23) in the right hand sides of (26) to arrive at the right hand sides of (16). The rest is using the definitions (4) and (1). — Some practical implications of the theorem are as follows:

- **Interpretation of proper scoring rules:** The infinite mass placed near zero and one by the weight functions (18) and (19) of boosting loss and of log-loss indicates an emphasis on getting classification right for extreme misclassification costs of both classes. We will see below (Section 14) that this eagerness to perform well on both ends of the cost axis can lead to situations in which classification performance is low on both ends.

- **Robustness to misspecified costs:** In practice one can often argue that misclassification costs are not known precisely. The relative indeterminacy of costs suggests that classification results should be insensitive to small changes in $c$ or, more precisely, it should be accurate across a range of indeterminacy of $c$. It would then seem natural not to seek optimality for a single misclassification loss but for an average of misclassification losses, that is, for a proper scoring rule. The proper scoring rule should localize by peaking its weight function $\omega(q)$ at costs in the range of interest.

- **Class probability estimation versus classification:** The mixture representation of proper scoring rules gives simple evidence that classification is easier than class probability estimation when the model is biased. While $\eta(\mathbf{x})$ may not be well-approximated by the model $q(\mathbf{x})$, it may still be possible for each cost $c$ to approximate $\{\eta(\mathbf{x}) > c\}$ well with $\{q(\mathbf{x}) > c\}$, but each $c$ requiring a separate model fit $q(.)$. This is expressed by the following simple inequality:

$$\min_{q()}\; \mathcal{L}_\omega(q()) \;\geq\; \int\, \min_{q()} \mathcal{L}_c(q())\;\omega(dc)\;,$$

where $\mathcal{L}_\omega() = \int \mathcal{L}_c()\omega(dc)$.

# 8 Proper scoring rules = negative quasi-likelihoods

Identity (11) can be used to show that smooth proper scoring rules are exactly the (negative) quasi-likelihoods for binary observations $y$. To this end re-write this identity for observations $y$ instead of probabilities $\eta$ as follows:

$$\frac{\partial}{\partial q}\mathbf{L}(y|q) \;=\; -\,(y-q)\;\omega(q)\;. \tag{28}$$

By comparison, the definition of quasi-likelihoods $\mathcal{Q}(y|q)$ as found, for example, in McCullagh and Nelder (1989, Section 9.2.2), with proper translation of notation, reads as follows:

$$\frac{\partial}{\partial q}\;\mathcal{Q}(y|q) \;=\; \frac{y-q}{V(q)}$$

Thus the two equations are structurally identical with the following correspondences:

$$\mathbf{L} = -\mathcal{Q}\;, \qquad \omega(q) = \frac{1}{V(q)}\;.$$

The main purpose of quasi-likelihoods is efficient estimation in the presence of over- or under-dispersion. This can be achieved by specifying a variance function $V(q)$ that deviates from the variance of the exponential model at hand, here the Bernoulli model with $V(q) = q(1-q)$. Because over/under-dispersion is not meaningful in ungrouped Bernoulli data, the interpretation of proper scoring rules as quasi-likelihoods is therefore only a mathematical and structural observation, not a substantive one. Yet there is an intuitive aspect to the interpretation in that it indicates what "precision" a proper scoring rule implicitly assigns a Bernoulli observation $y$ with $P[y=1] = q$. The three standard examples have the following implied variance functions:

- Boosting loss: $V_{boost}(q) \;=\; [\,q(1-q)\,]^{3/2}$
- Log-loss: $V_{log}(q) \;=\; q(1-q)$
- Squared error loss: $V_{squ}(q) \;=\; 1$

Thus one could say that boosting loss considers extreme observations with $q$ near 0 and 1 as more precise than the Bernoulli model, whereas squared error loss considers them as much less reliable. Overall, though, the concern with variance and efficiency in the quasi-likelihood interpretation can be misguided because the real problem with ungrouped Bernoulli data is bias from misspecification of models.

# 9 Fitting linear models with proper scoring rules

We describe fitting linear models with Newton iterations and Fisher scoring and their implementation as Iteratively Reweighted Least Squares (IRLS) algorithms. Assuming an inverse link function q($F$) and a linear model $F(\mathbf{x}) = \mathbf{x}^T\mathbf{b}$, the goodness-of-fit criterion associated with a proper scoring rule is:

$$\mathcal{L}(\mathbf{b}) \;=\; \frac{1}{N}\sum_{n=1..N}\left[\,y_n\,L_1(1-q(\mathbf{x}_n^T\mathbf{b})) + (1-y_n)\,L_0(q(\mathbf{x}_n^T\mathbf{b}))\right]\;. \tag{29}$$

Newton updates have the form

$$\mathbf{b}_{new} \;=\; \mathbf{b}_{old} - \left(\partial^2_{\mathbf{b}}\mathcal{L}(\mathbf{b})\right)^{-1}\left(\partial_{\mathbf{b}}\mathcal{L}(\mathbf{b})\right) \;.$$

The equivalent IRLS form of these updates is as solutions to normal equations with a suitable design matrix $X$, a diagonal weight matrix $W$, and response vector $\mathbf{z}$, the latter two changing from update to update:

$$\mathbf{b}_{new} \;=\; \mathbf{b}_{old} + \left(X^T W X\right)^{-1}\left(X^T W \mathbf{z}\right) \;,$$

We write the IRLS update in an incremental form whereas in the form usually shown in the literature one artifically absorbs $\mathbf{b}_{old}$ in the current response $\mathbf{z}$. The incremental form leds itself more directly to boosting-like stagewise fitting (Section 10). Abbreviating

$$q_n = q(\mathbf{x}_n^T\mathbf{b}) \;, \qquad q'_n = q'(\mathbf{x}_n^T\mathbf{b}) \;,$$

we have $\partial_{\mathbf{b}} q_n = q'_n\,\mathbf{x}_n$. Making use of the proposition of Section 6 the ingredients for Newton updates become

$$\partial_{\mathbf{b}}\mathcal{L}(\mathbf{b}) \;=\; -\frac{1}{N}\sum_{n=1..N}(y_n - q_n)\,\omega(q_n)\,q'_n\,\mathbf{x}_n \;, \tag{30}$$

$$\partial^2_{\mathbf{b}}\mathcal{L}(\mathbf{b}) \;=\; \frac{1}{N}\sum_{n=1..N}\Big(\omega(q_n)\,{q'_n}^2 \;-\; (y_n - q_n)\,[\,\omega(q_n)\,q'_n\,]'\Big)\;\mathbf{x}_n\mathbf{x}_n^T \;. \tag{31}$$

Equating $X^TWX = \partial^2_{\mathbf{b}}\mathcal{L}(\mathbf{b})$ and $X^TW\mathbf{z} = \partial_{\mathbf{b}}\mathcal{L}(\mathbf{b})$, the current weights $W = \mathrm{diag}(w_n)$ and current responses $\mathbf{z} = (z_n)$ for IRLS are:

$$w_n \;=\; \omega(q_n)\,{q'_n}^2 \;-\; (y_n - q_n)\,[\,\omega(q_n)\,q'_n\,]' \;, \tag{32}$$

$$z_n \;=\; (y_n - q_n)\,\omega(q_n)\,q'_n/w_n \;. \tag{33}$$

The weights $w_n$ can be negative when the loss $\mathbf{L}(y|q(F))$ is not strictly convex as a function of $F$. — For Fisher scoring one replaces the Hessian $\partial^2_{\mathbf{b}}\mathcal{L}(\mathbf{b})$ with its expectation, assuming the current estimates of the conditional probabilities of the classes to be the true ones: $\mathbf{P}[Y = 1|X = x_n] = q_n$. This implies $\mathbf{E}_{y_n}(y_n - q_n) = 0$, hence the weights and the current response simplify as follows:

$$w_n \;=\; \omega(q_n)\,{q'_n}^2 \;, \tag{34}$$

$$z_n \;=\; (y_n - q_n)/q'_n \;. \tag{35}$$

In general the response values $y_n$ affect the Newton weights (32) directly, but the Fisher scoring weights (34) depend on the responses only through the current class probability estimates $q_n$. This contrasts with a commonly held intuition about why boosting works: observations should be up- or down-weighted according to whether they have been misclassified in the previous update. No such thing seems to be necessary for a successful iterative reweighting scheme. Below (Section 16) we indicate that even Newton steps do not directly depend on the response value for certain "canonical" pairings of link function and proper scoring rule.

# 10 Boosting as stagewise fitting of additive models with proper scoring rules

In FHT's (2000) interpretation, boosting is a form of stagewise forward regression for building a possibly very large additive model. While Freund and Schapire conceived boosting as a weighted majority vote among a collection of weak learners, FHT see it as the thresholding of a linear combination of basis functions, in other words, of an additive model. We adopt FHT's view of boosting and adapt the IRLS iterations to stagewise forward additive fitting.

For additive modeling one needs a set $\mathcal{F}$ of admitted basis functions $f(\mathbf{x})$. For Real AdaBoost, $\mathcal{F}$ is a constrained set of trees such as stumps (trees of depth one), and for Discrete AdaBoost $\mathcal{F}$ is the set of indicator functions obtained by thresholding trees. These examples explain why $\mathcal{F}$ is not assumed to form a linear space. An additive model is a linear combination of elements of $\mathcal{F}$: $F(x) = \sum_k b_k f_k(\mathbf{x})$.

Stagewise fitting is the successive acquisition of terms $f^{(K)} \in \mathcal{F}$ given that terms $f^{(1)}, \ \ldots, \ f^{(K-1)}$ for a model $F^{(K-1)} = \sum_{k=1..K-1} b_k \, f^{(k)}$ have already been acquired. In what follows we write $F_{old}$ for $F^{(K-1)}$, $f$ for $f^{(K)}$, $b$ for $b_K$, and finally $F_{new} = F_{old} + b\, f$.

Given a training sample $\{(\mathbf{x}_n, \, y_n)\}_{n=1..N}$, the empirical loss is

$$\mathcal{L}(F_{old} + b\, f) \ = \ \frac{1}{N} \sum_{n=1..N} \mathbf{L}(y_n \,|\, q(\, F_{old}(\mathbf{x}_n) + b\, f(\mathbf{x}_n)\,)) \ .$$

A stage in stagewise fitting amounts to finding

$$(f, \, b) \ = \ \text{argmin}_{f \in \mathcal{F}, \ b \in \mathbb{R}} \, \mathcal{L}(\, F_{old} + b\, f) \ .$$

Often, as when $\mathcal{F}$ is a set of trees (real AdaBoost), the stepsize $b$ is absorbed in $f$, but when $F$ is an indicator functions (discrete AdaBoost), $b$ has to be found separately. Friedman (2001 suggested that $b$ should not be optimized; instead, it should be chosen small to form what he calls a "slow learner". Slow learning is generally benefial because it avoids greediness, in particular in the first stages of fitting, when the stagewise optimal choice of $b$ introduces too much of the initial basis functions into the model.

In detail, stagewise fitting proceeds by producing a new term $f$, a stepsize $b$ (if desired), and an update $F_{new} \ = \ F_{old} + b\, f$ based on the current IRLS weights $w_n$, estimated class 1 probabilities $q_n$, and current responses $z_n$. Using the abbreviations $F_n \ = \ F_{old}(\mathbf{x}_n)$, $q_n \ = \ q(F_n)$, and $q'_n = q'(F_n)$, these are for a Newton step:

$$\begin{aligned} w_n \ &= \ \omega(q_n)\, {q'_n}^2 \ - \ (y_n - q_n)\, [\, \omega'(q_n)\, q'_n \,]' \ , \\ z_n \ &= \ (y_n - q_n)\, \omega(q_n)\, q'_n / w_n \ , \end{aligned}$$

and for a Fisher scoring step:

$$w_n \ = \ \omega(q_n)\, {q'_n}^2 \ , \qquad z_n \ = \ (y_n - q_n)\, / q'_n \ .$$

The latter specializes to LogitBoost for log-loss (FHT 2000). In practice, $f$ is the result of some heuristic search procedure such as a greedily grown tree or an indicator function of a thresholded tree.

# 11 The Beta family of proper scoring rules

We introduce a continuous 2-parameter family of scoring rules that is sufficiently rich to encompass most commonly used losses, including boosting loss, log-loss, squared error loss, and misclassification loss in the limit. It will later serve us to "tailor" the weight function $\omega(q)$ to the desired cost for cost-weighted classification. The family is modeled after the Beta densities, but we permit weight functions that are not integrable and hence cannot be normalized to densities:

$$\omega(q) \;=\; q^{\alpha-1}\,(1-q)^{\beta-1} \tag{36}$$

$$L_1'(1-q) \;=\; q^{\alpha-1}(1-q)^{\beta} \qquad L_0'(q) \;=\; q^{\alpha}(1-q)^{\beta-1}$$

The partial losses $L_1$ and $L_0$ are bounded below iff $\alpha,\, \beta > -1$. As the following list suggests, the lowest values ever proposed (at least implicitly) are $\alpha = \beta = -\frac{1}{2}$:

- $\alpha = \beta = -\frac{1}{2}$: Boosting loss
- $\alpha = \beta = 0$: Log-loss
- $\alpha = \beta = \frac{1}{2}$: A new type of loss, intermediate between log-loss and squared error loss,

$$L_1(1-q) = \arcsin((1-q)^{\frac{1}{2}}) - (q(1-q))^{\frac{1}{2}}\ , \qquad L_0(q) = \arcsin(q^{\frac{1}{2}}) - (q(1-q))^{\frac{1}{2}}\ .$$

- $\alpha = \beta = 1$: Squared error loss
- $\alpha = \beta = 2$: A new loss closer to misclassification than squared error loss,

$$L_1(1-q) = \frac{1}{3}(1-q)^3 - \frac{1}{4}(1-q)^4\ , \qquad L_0(q) = \frac{1}{3}q^3 - \frac{1}{4}q^4\ .$$

- $\alpha = \beta \to \infty$: Misclassification loss

Values of $\alpha$ and $\beta$ that are integer multiples of $\frac{1}{2}$ permit closed formulas for $L_1$ and $L_0$. For other values one needs a numeric implementation of the incomplete Beta function. — Weight functions and partial losses for several symmetric choices of $\alpha = \beta$ are depicted in Figure 3.

# 12 Tailoring proper scoring rules for cost-weighted misclassification

We introduce "tailoring" of strict proper scoring rules to cost-weighted classification. Recall that cost-weighted misclassification loss with costs $c$ and $1-c$ for false positives and false negatives, respectively, is given by

$$\omega(dt) \;=\; \delta_c(dt)\ .$$

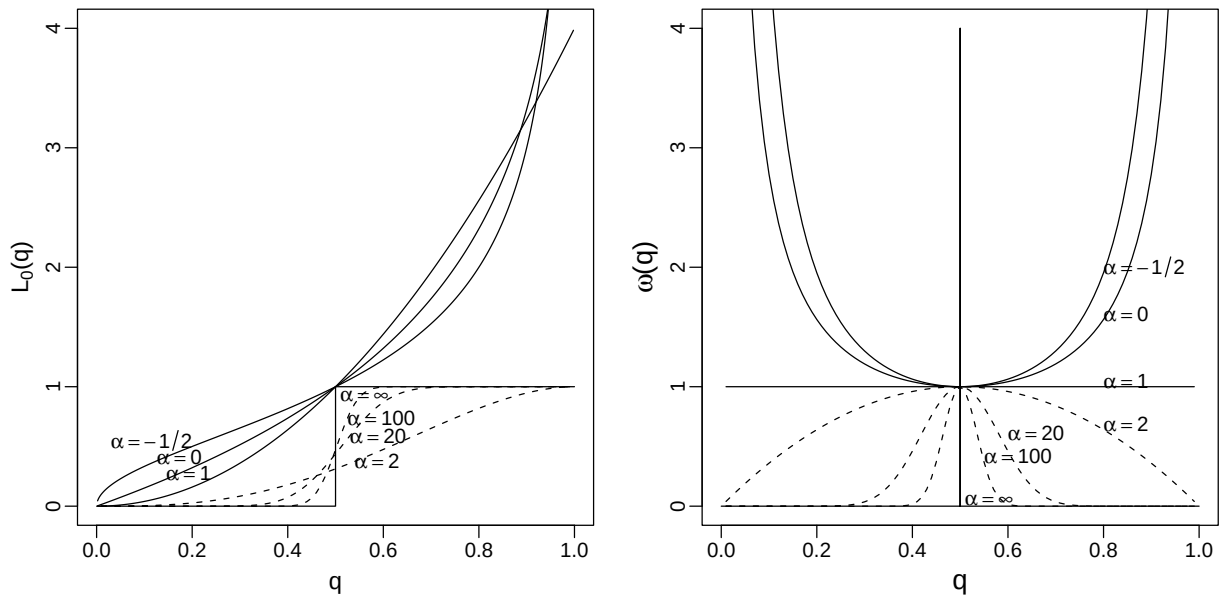


Figure 3: *Loss functions $L_0(q)$ and weight functions $\omega(q)$ for various values of $\alpha = \beta$: exponential loss ($\alpha = -\frac{1}{2}$), log-loss ($\alpha = 0$), squared error loss ($\alpha = 1$), misclassification loss ($\alpha = \infty$). These are scaled to pass through 1 at $q = 0.5$. Also shown are $\alpha = 2$, 20 and 100 scaled to show convergence to the step function and the point mass, respectively.*

These losses are unsuitable for estimation and call for surrogate loss functions in the form of a strict proper scoring rule such as boosting loss or log-loss. These two choices, however, may be going too far by placing extreme weight on costs $c$ near 0 and 1. More plausible choices can be found in the Beta family of strict proper scoring rules which smoothly interpolates the extremes of log-loss or boosting loss on the one hand and cost-weighted misclassification losses on the other hand, and which therefore offers choices that are closer to the latter than the former. To this end we note that the densities ($\alpha, \beta > 0$)

$$\omega_{\alpha,\beta}(q) = \frac{1}{B(\alpha,\beta)} q^{\alpha-1}(1-q)^{\beta-1}$$

converge weakly to $\omega(dq) = \delta_c(dq)$ when

$$\alpha,\ \beta \to \infty\ , \qquad \text{subject to} \quad \frac{\alpha}{\beta} = \frac{c}{1-c}\ , \quad \text{or equivalently} \quad \frac{\alpha}{\alpha+\beta} = c\ .$$

This follows because the mean and variance of the Beta distribution are

$$\mu = \frac{\alpha}{\alpha+\beta} = c \qquad \text{and} \qquad \sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{c(1-c)}{\alpha+\beta+1}\ , \tag{37}$$

so that $\sigma^2 \to 0$ as $\alpha,\ \beta \to \infty$. In a limiting sense, the ratio of the exponents, $\alpha/\beta$, acts as a cost ratio for the classes. — In Section 19 we will need for theoretical purposes a slightly different tailoring convention, using the mode instead of the mean:

$$c \ = \ q_{mode} \ = \ \frac{a-1}{a+b-2}\ . \tag{38}$$

In the limit for large $a$ and $b$ this is equivalent to tailoring at the mean.

These constructions have obvious practical applications: log-loss can be replaced with the proper scoring rules generated by the above weight functions $\omega_{\alpha,\beta}(q)$. In doing so we can possibly achieve improved classification for particular cost ratios or class-probability thresholds when the fitted model is biased but adequate for describing classification boundaries individually. The appropriate degree of peakedness of the weight function could be estimated from the data using cross-validation of cost-weighted misclassification loss. In Section 14 we will illustrate fitting biased linear models with "tailored" proper scoring rules.

# 13 The Hand-Vinciotti reweighting method as minimization of tailored proper scoring rules

HV's (2003) tailoring method based on reweighting can be interpreted as the minimization of certain proper scoring rules. HV devised their scheme algorithmically by modifying the weights of IRLS for Fisher scoring in logistic regression. Their weights are Gaussian kernels:

$$\omega(q) = \phi\left(\frac{q-c}{\sigma}\right) \ ,$$

where $\phi(t)$ is the standard Gaussian density. In light of Section 9 their weights amount to a particular choice of weight function $\omega(q)$ that fully defines a proper scoring rule:

$$L_1(1-q) = \int_q (1-t)\cdot\phi\left(\frac{t-c}{\sigma}\right)\,dt \ , \qquad L_0(q) = \int^q t\cdot\phi\left(\frac{t-c}{\sigma}\right)\,dt \ .$$

The Gaussian standard deviation $\sigma$ can be used as a measure of peakedness although it is not the standard deviation of the weight function because of truncation to the unit interval. — A disadvantage of truncated Gaussians is that they do not include convex weight functions such as the one for log-loss, the efficient choice when the fitted model is accurate. The Gaussian limit for $\sigma \to \infty$ is the constant weight function corresponding to squared error loss. The Beta family is more complete because it includes not only squared error loss but log-loss and even boosting loss.

# 14 A data example for tailoring in linear models

We illustrate the fact that linear models can be unsuitable as global models and yet successful for classification at specific classification quantiles or misclassification costs. To this end we recreate an artificial example of HV (2003) and show that tailoring loss functions to specific classification thresholds allow us to approximate the optimal boundaries. We then show that the HV (2003) scenario is reflected in some well-known real data, the Pima Indian diabetes data (UCI Machine Learning Repository, Newman et al. 1998).

To make their point HV (2003) designed $\eta(\mathbf{x})$ as a surface in the shape of a smooth spiral staircase on the unit square, as depicted in Figure 4. The critical feature of the surface

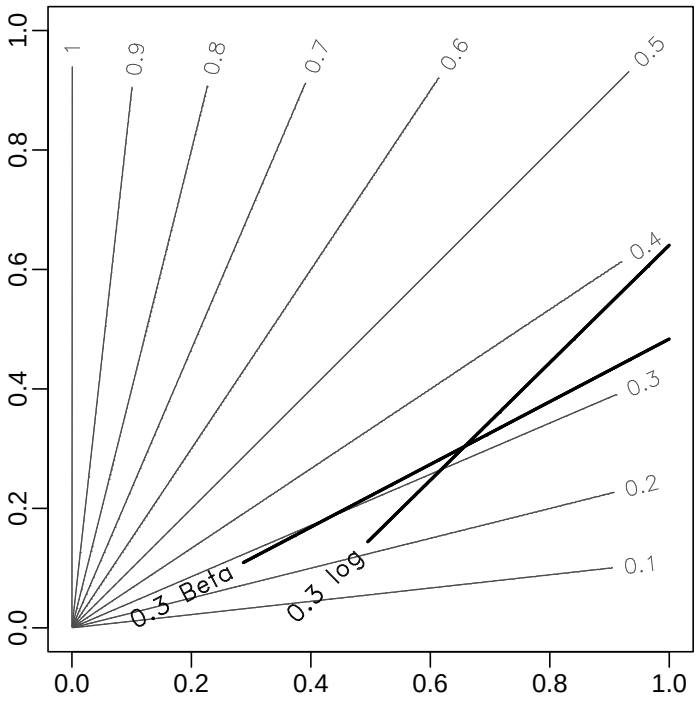


Figure 4: *Hand and Vinciotti's Artificial Data: The class probability function* $\eta(\mathbf{x})$ *has the shape of a smooth spiral ramp on the unit square with axis at the origin. The bold line marked "0.3 log" shows a linear logistic fit thresholded at* $c = 0.3$*. The other bold line, marked "0.3 Beta", shows a thresholded linear fit corresponding to a proper scoring rule in the Beta family with parameters* $\alpha = 6$ *and* $\beta = 14$*.*

is that the optimal classification boundaries $\eta(\mathbf{x}) = c$ are all linear but not parallel. The absence of parallelism renders any linear model $q(\mathbf{b}^T\mathbf{x})$ unsuitable as a global fit, but the linearity of the classification boundaries allows linear models to describe these boundaries, although every level requires a different linear model. The purpose of HV's (2003) and our tailoring schemes is to home in on these level-specific models.

In recreating HV's (2003) example, we simulated 4,000 data points whose two predictors were uniformly distributed in the unit square, and whose class labels had a conditional class 1 probability $\eta(x_1, x_2) = \cos^{-1}(x_1/(x_1^2 + x_2^2)^{1/2})/(\pi/2)$. We fitted a linear model with the logistic function $q(t) = 1/(1 + \exp(-t))$, using a proper scoring rule in the Beta family with $\alpha = 29$ and $\alpha/\beta = 0.3/0.7$ in order to home in on the $c = 0.3$ boundary. Figure 4, which is similar to HV's (2003), shows the success: the estimated boundary is close to the true 0.3-boundary. The figure also shows that the 0.3-boundary estimated with the log-loss of logistic regression is essentially parallel to the 0.5-boundary, which is sensible because logistic regression is bound to find a compromise model which, for reasons of symmetry between the two labels, should be a linear model with level lines roughly parallel to the true 0.5-boundary.

We turn to the Pima Indians Diabetes data to show that non-parallel linear boundaries

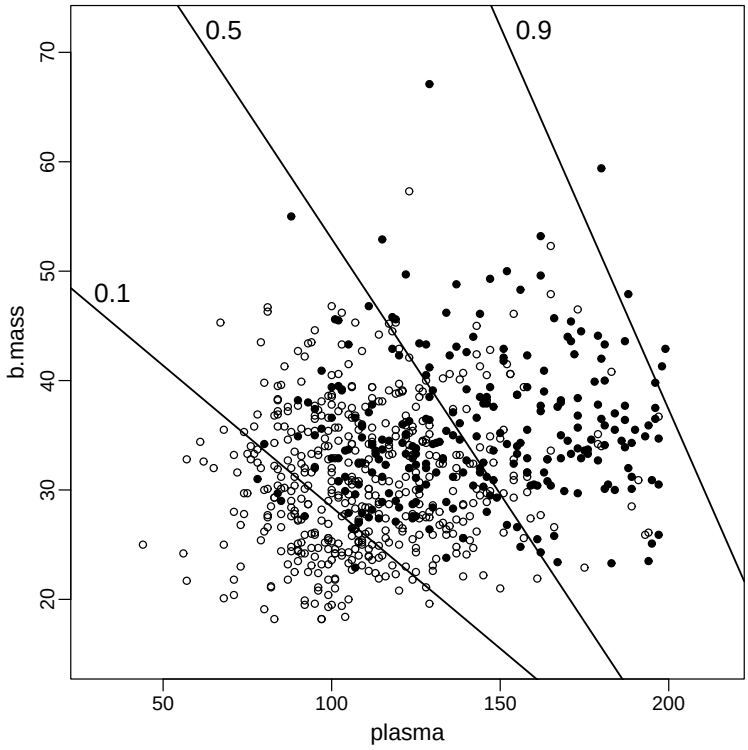


Figure 5: *Illustration of non-parallel linear classification boundaries for different classification thresholds or cost weights: the Pima Indians Diabetes Data, "b.mass" plotted against "plasma", with boundaries for the thresholds c =0.1, 0.5 and 0.9. Glyphs: open squares = no diabetes, class 0; filled circles = diabetes, class 1.*

for differing classification thresholds can produce improvements in real data as well. We start with a graphic illustration similar to Figure 4 by restricting attention to the most powerful predictors, *plasma* and *b.mass*: Figure 5 shows estimated linear classification boundaries for the levels 0.1, 0.5 and 0.9, clearly indicating that at least for level 0.1 the boundary is not parallel to the other two.

We now give quantitative evidence for significant improvements from tailoring using all predictors. [The data have numerous missing values. Some details about their treatment: We removed the cases with zero values on *plasma* (5), *b.press* (35) and *b.mass* (11) for a total of 44, thus reducing the sample size from 768 to 724. For the variables *skin* and *insulin* with 227 and 374 missings, respectively, we introduced two dummy variables for missingness.] To compare tailoring with standard logistic regression, we performed a cross-validation experiment, not in the conventional manner of 10-fold cross-validation, but for greater inferential precision with many more folds as follows: We randomly split the data 400 times into an 80% training set and a 20% test set. On 200 of the splits we fitted a standard linear logistic regression ($\alpha = \beta = 0$) to the training set and evaluated the cost-weighted misclassification errors for $c$ =0.50, 0.80, 0.90 and 0.95 on the test set. On the other 200 splits we fitted a linear model with the logit link and a tailored proper scoring rule with $\alpha = 4.5$ and $\beta = 0.5$ to the training set, and again evaluated the cost-weighted

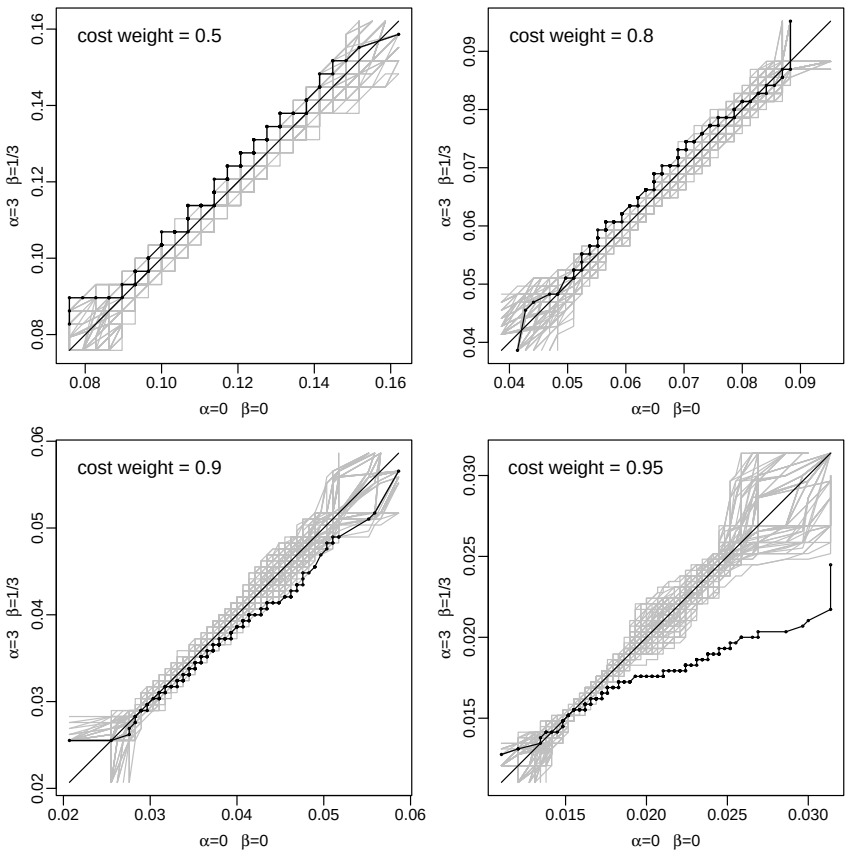


Figure 6: *Pima Indians Diabetes data: Comparison of logistic regression with a proper scoring rule tailored for high class 1 probabilities:* $\alpha = 9,\ \beta = 1$*, hence* $cw = 0.9$*. The black traces with points are empirical Q-Q curves of 200 cost-weighted misclassification costs computed on randomly selected test sets of size 20%. Overplotted in gray are one hundred null traces obtained with random permutations.*

misclassification errors for $c$ =0.50, 0.80, 0.90 and 0.95 on the test set.

We are interested in comparing the 200 misclassification errors of the logistic fit with the 200 of the tailored fit, for all four cost weights $c$. This amounts to four 2-sample comparisons of cost-weighted misclassification errors of one fitting method with those of the other fitting method. Such comparisons are conventionally done in terms of 2-sample $t$-tests for means, but instead we present the results more informatively in terms of 2-sample Q-Q plots, which essentially amount to plots of the two sorted series of 200 values against each other. This is carried out in Figure 6, which shows one plot for each of the four cost weights. The Q-Q traces are shown in black, with the misclassification errors of the tailored Beta rule shown on the vertical axis, those of logistic regression on the horizontal axis.

To inject statistical inference in these plots, we added 100 "null" Q-Q traces overplotted in gray, based on a permutation test idea: by construction of the experiment, the 400 values are exchangeable under the null assumption of identical distributions of the misclassification errors of both methods. Hence we can randomly split the pool of 400 values into two sets of

200 and draw their Q-Q traces, simulating the null hypothesis conditional on the observed values, repeated and overplotted 100 times. Thus, wherever the black trace reaches outside the gray band of null traces, statistical significance is achieved. This is the case in the bottom right hand plot for cost weight $c = 0.95$, and border line in the bottom left plot for $c = 0.90$. In the top left plot we see that the tailored fit performs somewhat worse for cost weight $c = 0.50$, but only marginally so. — In summary, we see evidence that tailoring of proper scoring rules to cost-weights does indeed provide an advantage over standard logistic regression.

# 15 Stability of tailoring: convexity and asymptotics

Tailoring may raise fears of unstable fits because for peaked weight functions the composite loss function $\mathbf{L}(y|q(F))$ loses convexity rather quickly. If $\mathbf{L}(y|q(F))$ is sufficiently smooth in $F$, a sufficient condition for convexity is a non-negative second derivative, which according to (32) amounts to

$$\omega(q)\, q'^2 \;-\; (y-q)\,[\,\omega(q)\, q'\,]' \;\geq\; 0\;. \tag{39}$$

This requirement results in two inequalities, one for $y = 1$ and one for $y = 0$, which we can summarize as follows:

**Proposition:** *Composite losses $\mathbf{L}(y|q(F))$ are convex in $F$ if $\omega(q)$ and $q'$ are strictly positive, both are smooth, and they satisfy*

$$\frac{q'}{1-q} \;\geq\; \frac{[\,\omega\, q'\,]'}{\omega\, q'} \;\geq\; -\,\frac{q'}{q}\;.$$

For the combination of proper scoring rules in the Beta family, $\omega(q) \;=\; q^{\alpha-1}(1-q)^{\beta-1}$, and scaled logistic links,

$$q(F) \;=\; \frac{1}{1+e^{-F/\sigma}}\;, \qquad F(q) \;=\; \sigma\,\log\frac{q}{1-q}\;, \qquad q' \;=\; \frac{1}{\sigma}\,q\,(1-q)\;,$$

the convexity condition becomes

$$q \;\geq\; \alpha\,(1-q) - \beta\, q \;\geq\; -\,(1-q)\;,$$

which implies:

**Corollary:** *Proper scoring rules in the Beta family and scaled logistic links result in convex composite losses iff $\alpha,\ \beta \in [-1, 0]$ and $\sigma > 0$.*

In particular, exponential loss ($\alpha = \beta = -1/2$, $\sigma = 1/2$) and log-loss with the logistic link ($\alpha = \beta = 0$, $\sigma = 1$) are convex, but even squared error loss is not convex in combination with any scaled logistic link. This seems to spell trouble for tailoring because there we typically choose $\alpha,\ \beta > 0$ and hence forfeit the benefits of convexity. The situation is not as dire,

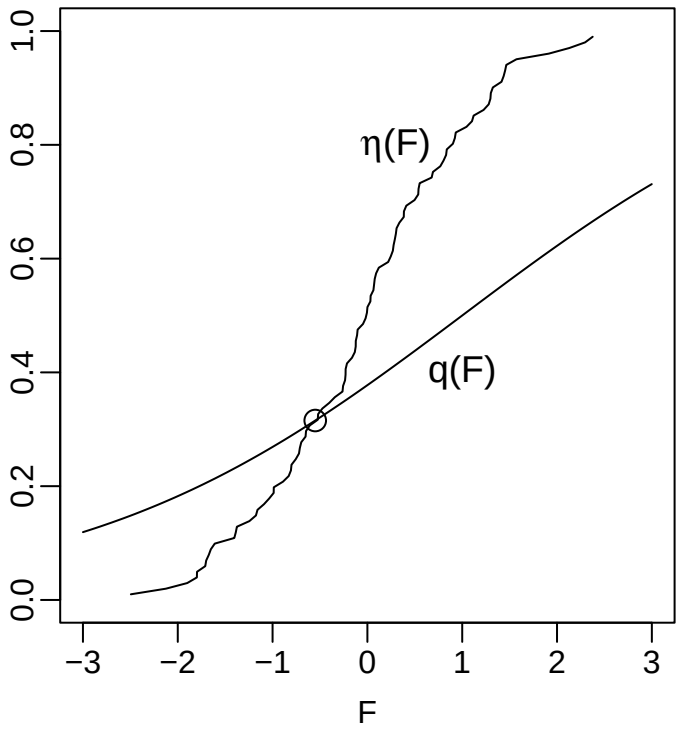


Figure 7: *A stable situation for tailored estimation, shown for a single predictor: $q(F)$ has positive bias for $q(F) < c$ and negative bias for $q(F) > c$.*

however, in view of a simple asymptotic argument which is as follows: Standard formulas for asymptotic variances in estimating equations (for example, Shao 2003, p. 369) yield

$$\mathrm{AVar}(\hat{\mathbf{b}}) \;\approx\; N \left(\mathbf{E}\,\partial^2\mathcal{L}(\mathbf{b})\right)^{-1} \mathrm{Var}\left(\partial\mathcal{L}(\mathbf{b})\right) \left(\mathbf{E}\,\partial^2\mathcal{L}(\mathbf{b})\right)^{-1}$$

From Section 9, Equations (30) and (31), we obtain

$$\mathrm{Var}\;\partial\mathcal{L}(\mathbf{b}) \;=\; \frac{1}{N^2}\sum_{n=1..N} \eta_n(1-\eta_n)\,[\,\omega(q_n)\,q_n'\,]^2\;\mathbf{x}_n\mathbf{x}_n^T\;, \tag{40}$$

$$\mathbf{E}\,\partial^2\mathcal{L}(\mathbf{b}) \;=\; \frac{1}{N}\sum_{n=1..N} \Big(\,\omega(q_n)\,{q_n'}^2 \;-\; (\eta_n - q_n)\,[\,\omega(q_n)\,q_n'\,]'\,\Big)\;\mathbf{x}_n\mathbf{x}_n^T\;. \tag{41}$$

We allow that the model $q(\mathbf{b}^T\mathbf{x})$ is biased, meaning that it differs from the truth, $\eta(\mathbf{x}) = P[Y = 1|\mathbf{X} = \mathbf{x}]$. The parameter $\mathbf{b}$ only indicates the best approximation $q(\mathbf{b}^T\mathbf{x})$ to $\eta(\mathbf{x})$.

We note that making $\mathbf{E}\,\partial^2\mathcal{L}(\mathbf{b})$ large makes the asymptotic variance small (both in the sense of the ordering of symmetric matrices), hence from (41) we see that we can again focus on an expression similar to (39) above which determined convexity of the loss function:

$$\omega(q_n)\,{q_n'}^2 \;-\; (\eta_n - q_n)\,[\,\omega(q_n)\,q_n'\,]' \;\gg\; 0\;.$$

Of the two terms the first is positive because we assume a strict proper scoring rule ($\omega(q) > 0$) and a strictly increasing link function ($q'(F) > 0$). Problematic is the second term which can

reduce $\mathbf{E}\,\partial^2\mathcal{L}(\mathbf{b})$ and hence inflate the asymptotic variance for some linear combinations of $\mathbf{b}$. On the other hand, this same term can also contribute to stability by enlarging $\mathbf{E}\,\partial^2\mathcal{L}(\mathbf{b})$ and hence diminishing the asymptotic variance, namely, when

$$-(\eta_n - q_n)\,[\,\omega(q_n)\,q_n'\,]' \;\geq\; 0\;. \tag{42}$$

To see the meaning of this condition, we specialize it to the Beta family and the logistic link, as used for tailoring in Section 12, and in particular we consider $\alpha,\ \beta > 0$, which is possibly problematic because of non-convexity. Making use of $q' = q(1-q)$ for the logistic link, we see that (42) is equivalent to

$$-(\eta_n - q_n)\,[\,\alpha(1-q_n) - \beta q_n\,] \;\geq\; 0$$

or

$$\eta_n \;\gtreqless\; q_n \quad \text{when} \quad q_n \;\gtreqless\; \frac{\alpha}{\alpha+\beta}\;. \tag{43}$$

Recall that for tailoring to misclassification cost $c = \alpha/(\alpha+\beta)$ we use weights $\omega(q)$ that concentrate mass around $c$. Assuming a situation in which tailoring works, that is, $q(\mathbf{b}^T\mathbf{x}) > c$ approximates $\eta(\mathbf{x}) > c$ well even in the absence of a good fit of $q(\mathbf{b}^T\mathbf{x})$ to $\eta(\mathbf{x})$, we may find a fair degree of stability as long as the model bias has a certain structure:

> negative bias for $\eta(\mathbf{x}) > c$ and positive bias for $\eta(\mathbf{x}) < c$,

which is just (43) re-expressed. The situation is illustrated in Figure 7 for a single predictor. Can this type of structured bias be counted on in applications? Not in general, but it can be helped by biasing the model artificially: One can flatten $q(\hat{\mathbf{b}}^T\mathbf{x})$ by shrinking $\hat{\mathbf{b}}$, or more precisely: shrinking all regression coefficients except the intercept. Not much shrinking may be needed to create the situation depicted in Figure 7, with the ensuing insurance against inflation of the asymptotic variance. The intercept must not be shrunk so as to allow the weight function $\omega(q)$ to match the level of $q(\hat{\mathbf{b}}^T\mathbf{x})$ to that of $\eta(\mathbf{x})$ in the neighborhood of the target cost $c$, as expressed by the stationarity condition $\partial\,\mathcal{L}(\mathbf{b}) = 0$:

$$\frac{1}{N}\sum_{n=1..N} y_n\,\omega(q_n)\,q_n'\,\mathbf{x}_n \;=\; \frac{1}{N}\sum_{n=1..N} q_n\,\omega(q_n)\,q_n'\,\mathbf{x}_n\;.$$

In summary, tailoring may not necessarily incur the feared costs of non-convexity and asymptotic variance inflation unless the fitted model overfits, resulting in a situation opposite to that of (43) and of Figure 7.

# 16 Canonical links and weights

The above discussion implies that the term

$$(y_n - q_n)\,[\,\omega(q_n)\,q_n'\,]' \tag{44}$$

in the Hessian $\partial^2 \mathcal{L}(\mathbf{b})$ is a potential source of problems: non-convexity and asymptotic variance inflation. One may therefore wonder whether this term could not be made to disappear altogether. In fact, it can, and the condition it generates constrains the weight function $\omega(q)$ and the link function $q(F)$ to each other in a "canonical" way that coincides with the meaning of "canonical link functions" in generalized linear models. As implied by the previous section, canonicity has benefits: the loss function $\mathbf{L}(y|q(F))$ is convex, Newton interations and Fisher scoring are the same, the second derivative and hence the degree of convexity is not directly affected by the response values $y_n$, and the ingredients of the asymptotic variance take on a simple and robust form:

$$\begin{aligned} \text{Var } \partial\mathcal{L}(\mathbf{b}) &= \frac{1}{N^2} \sum_{n=1..N} \eta_n(1-\eta_n)\, \mathbf{x}_n \mathbf{x}_n^T \ , \\ \mathbf{E}\, \partial^2\mathcal{L}(\mathbf{b}) &= \frac{1}{N} \sum_{n=1..N} q'_n\, \mathbf{x}_n \mathbf{x}_n^T \ . \end{aligned}$$

Canonical pairs of weights and links are minimax in a sense: they avoid the non-convexity and asymptotic variance inflation by killing the term that may cause these problems.

The structure of canonical pairs of weights and links is easily determined: The non-canonical term (44) disappears when $[\,\omega(q)\,q'\,]' = 0$, that is, when $\omega(q)\,q'$ is a constant, which we can choose to be 1 (assuming the fitted models are closed under multiplication with constants, as are linear, additive and tree models):

$$\omega(q)\,q' \ = \ 1 \ . \tag{45}$$

Now $q = q(F)$ is the inverse of the link function $F = F(q)$, hence $F'(q) = 1/q'(F(q))$ and from (45)

$$F'(q) \ = \ \omega(q) \ = \ L'_1(1-q) + L'_0(q) \ , \tag{46}$$

where the second equality follows from $L'_1(1-q) + L'_0(q) = (1-q)\,\omega(q) + q\,\omega(q)$, which is the sum of Equations (9). The solution is unique up to an irrelevant additive constant (assuming the fitted models are closed under addition of constants). Two equivalent solutions are

$$F(q) \ = \ \int^q \omega(q)\,dq \qquad \text{and} \qquad F(q) \ = \ L_0(q) - L_1(1-q) \ . \tag{47}$$

If $\omega(q)$ happens to be a probability density function, then $F(q)$ can be chosen to be its cumulative distribution function, a relatively uninteresting case because it limits the link function to a finite range (the unit interval).

The converse holds also: prescribing a link function as canonical determines the proper scoring rule, as is immediately seen from Equation (46). We summarize:

**Proposition:** *For any strict proper scoring rule there exists a canonical link function $F(q)$ which is unique up to addition and multiplication with constants. Conversely, any link function is canonical for a unique proper scoring rule.*

Here are the standard examples in the Beta family for $\alpha = \beta = 1$, 0 and -1/2:

- *Squared error loss*: $L_1(1-q) = (1-q)^2$ and $L_0(q) = q^2$, hence the inverse link is essentially the identity transform:

$$F(q) \;=\; 2q - 1 \;.$$

- *Log-loss*: $L_1(1-q) = -\log(q)$ and $L_0(q) = -\log(1-q)$, hence, as is well-known, the natural link is the logit,

$$F(q) \;=\; \log \frac{q}{1-q} \;,$$

its inverse $q(F)$ is the logistic function and the composite loss is

$$L_0(q(F)) \;=\; \log(1 + \exp(F)) \;.$$

- *Boosting loss*: $L_1(1-q) = ((1-q)/q)^{1/2}$ and $L_0(q) = (q/(1-q))^{1/2}$, hence the canonical link is

$$F(q) \;=\; \left(\frac{q}{1-q}\right)^{1/2} - \left(\frac{1-q}{q}\right)^{1/2} \;=\; \frac{2\,q-1}{[q(1-q)]^{1/2}} \;,$$

Its inverse is

$$q(F) \;=\; \frac{1}{2}\left(\frac{F/2}{((F/2)^2+1)^{1/2}} \;+\; 1\right) \;,$$

and the composite loss is

$$L_0(q(F)) \;=\; \left((F/2)^2 \;+\; 1\right)^{1/2} \;+\; F/2 \;.$$

Thus the decomposition of the exponential loss into a proper scoring rule and a link function (Section 4) does *not* result in a canonical pair.

In the last two examples we observe that the composite loss function $L_0(q(F))$ is convex and grows less than linear for $F \to +\infty$. This is a general fact for canonical pairs:

**Corollary:** *For canonical pairs of proper scoring rules and links, the composite loss function $L(F) = L_0(q(F))$ is non-decreasing, convex, but has growth limited by 1: $L'(F) \leq 1$.*

This is immediately seen from $L_0'(q) = q\,\omega(q)$ (9) and $\omega(q)\,q' = 1$ (45) which yield

$$L'(F) = L_0'(q(F))\,q'(F) = q(F) \;\leq 1 \;.$$

This characteristic of canonical pairs immediately rules out the possibility that exponential loss $L(F) = \exp(F)$ is the composite of a proper scoring rule and its canonical link.

# 17 Information measures for tree estimation

Proper scoring rules and information measures such as the Gini index and entropy are in a one-to-one relationship. Given an expected loss $\mathbf{R}(\eta|q) = \eta L_1(1-q) + (1-\eta)L_0(q)$, the associated information measure is the minimal achievable loss, which we write as

$$\mathbf{H}(\eta) \;=\; \min_q \mathbf{R}(\eta|q) \;=\; \mathbf{R}(\eta|\eta) \;.$$

Information measures are of technical interest; see for example Zhang (2004; Def. 2.1, Lemma 2.1), Lugosi and Vayatis (2004; Lemma 4), Bartlett, Jordan and McAuliffe (2003; Section 2). — Information measures are also of practical interest in algorithms for tree-based classification: CART (Breiman et al. 1984) uses the Gini index, and C4.5 (Quinlan 1993) and the original tree functions in the S language (Clark and Pregibon 1992) use entropy. The former is the information measure derived from squared error loss, the latter from log-loss. Tree algorithms agree with each other in that they all estimate local conditional class probabilities with simple proportions, but they differ in how they judge the fit of these proportions in terms of information measures.

Here are the usual examples from the Beta family: They are also contained in a similar list given by Zhang (2004, after Definition 2.1):

- $\alpha = \beta = -1/2$: Boosting loss leads to a semi-circle criterion,

$$\mathbf{H}(q) = 2 \cdot [q(1-q)]^{1/2} \;.$$

- $\alpha = \beta = 0$: Log-loss leads to entropy,

$$\mathbf{H}(q) = -q \log(q) - (1-q) \log(1-q) \;.$$

- $\alpha = \beta = 1$: Squared error loss leads to the Gini index,

$$\mathbf{H}(q) = q(1-q) \;.$$

- $\alpha,\ \beta \to \infty,\ \frac{\alpha}{\beta} \to \frac{c}{1-c}$ : Cost-weighted misclassification loss leads to cost-weighted Bayes risk:

$$\mathbf{H}_c(q) = \min((1-c)\,q,\, c\,(1-q)) \;. \tag{48}$$

The semi-circle criterion derived from boosting loss was proposed by Kearns and Mansour (1996) as a criterion for tree construction.

The following proposition shows that every information measure determines essentially a unique proper scoring rule, modulo some arbitrary choices for those $q$ for which the function $\mathbf{H}(q)$ does not have a unique tangent, as for cost-weighted misclassification losses at the location $q = c$. The situation is best understood in terms of Figure 8.

**Proposition:** *The information measure* $\mathbf{H}(q)$ *is the concave lower envelope of its proper scoring rule* $\mathbf{R}(\eta|q)$*, with upper tangents* $\eta \mapsto \mathbf{R}(\eta|q)$*. Conversely, if* $\mathbf{H}(q)$ *is an arbitrary*

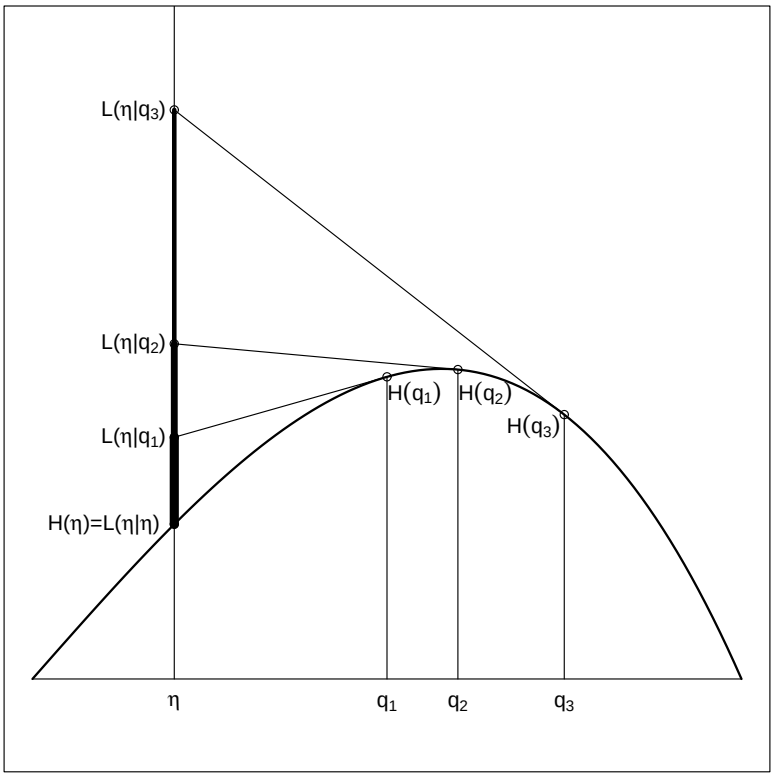


Figure 8: *Envelopes and proper scoring rules: For a true $\eta$ and any estimate $q$, the proper scoring rule can be read off the tangent of the envelope $\mathbf{H}()$ at $q$ evaluated at the location $\eta$. Trivially, the minimum of $\mathbf{R}(\eta|q)$ w.r.t. $q$ is taken on at $q = \eta$. The solid vertical segments at $\eta$ rising from $\mathbf{H}(\eta)$ indicate the Bregman distances between $\eta$ and $q_i$: $\mathbf{B}(\eta|q_i) = \mathbf{R}(\eta|q_i) - \mathbf{H}(\eta)$ (Section 19).*

*concave function, it determines a proper scoring rule $\mathbf{R}(\eta|q)$ which is unique except at locations where there are multiple upper tangents, in which case any of the tangents can be selected as part of the definition of the proper scoring rule.*

**Corollary:** *If $\mathbf{H}(q)$ is concave and smooth, its associated proper scoring rule is*

$$L_1(1-q) \;=\; \mathbf{H}(q) + \mathbf{H}'(q)\,(1-q)\;, \qquad L_0(q) \;=\; \mathbf{H}(q) - \mathbf{H}'(q)\,q\;. \tag{49}$$

This proposition in some form goes back to Savage (1971). A generalization from the binary case to general probability spaces and a careful analysis for the finite case is given by Gneiting and Raftery (2004, Theorems 2.1, 3.2 and 3.4). For a **proof**, note that $\mathbf{R}(\eta|q)$ is affine in $\eta$ and $\mathbf{R}(\eta|q) \geq \mathbf{H}(\eta)$, which makes $\eta \mapsto \mathbf{R}(\eta|q)$ an upper tangent for all $q$ and hence $\mathbf{H}(q)$ the concave lower envelope. Conversely, given a concave function $\mathbf{H}(q)$, there exist upper tangents at every $q$. They are affine functions and can be written $\eta \mapsto \eta\, T_1(q) + (1-\eta)\, T_0(q)$, so one can define $L_1(1-q) = T_1(q)$ and $L_0(q) = T_0(q)$. For the corollary, the upper tangent at $q$ is unique and can be written as an affine function $\eta \mapsto \mathbf{H}'(q)(\eta - q) + \mathbf{H}(q)$. Hence $\mathbf{R}(\eta|q) = \mathbf{H}'(q)(\eta - q) + \mathbf{H}(q)$ defines the proper scoring rule: $L_1(1-q) = \mathbf{R}(1|q)$ and $L_0(q) = \mathbf{R}(0|q)$.

**Proposition:** *If* $\mathbf{H}(q)$ *is concave and sufficiently smooth, the weight function* $\omega(q)$ *and the canonical link function of the associated proper scoring rule are:*

$$\omega(q) \;=\; -\,\mathbf{H}''(q)\;, \qquad F(q) \;=\; -\,\mathbf{H}'(q)\;. \tag{50}$$

The proof is a simple calculation. The relation links concavity and non-negative weights, as well as strict concavity and positive weights. The proposition indicates that information measures with the same second derivative are equivalent. In other words, information measures that differ only in an affine function are equivalent: $\mathbf{H}(q) + C_1\, q + C_0\,(1-q)$. This fact follows for all information measures, not only smooth ones, from the equivalence of proper scoring rules that differ in two constants only, $L_1(1-q) + C_1$ and $L_0(q) + C_0$.

The mixture proposition of Section 7 for proper scoring rules has an immediate analog for information measures:

**Proposition:** *Under conditions analogous to Theorem 1, information measures are mixtures of cost-weighted Bayes risks* (48) *with* $\omega(dc)$ *as the mixing measure:*

$$\mathbf{H}(q) \;=\; \int_0^1 \mathbf{H}_c(q)\,\omega(dc)\;. \tag{51}$$

**Tailored classification trees:** In light of the above facts it is natural to apply the tailoring ideas of Section 12 to information measures used in tree-based classification. The idea is again to put weight on the class probabilities that are of interest. "Weight" is meant literally in terms of the weight function $\omega(q)$. In practice, areas of interest are often the extremes with highest or lowest class probabilities. In the Pima Indians Diabetes data (Section 18), for example, this may mean focusing on the cases with an estimated probability of diabetes of 0.9 or greater. It would then be reasonable to use an information measure derived from a weight function that puts most of its mass on the right end of the interval (0,1). In Section 18 we will show experiments with weights that are simple power functions of $q$. In terms of the Beta family of weights (Section 11) this could mean using $\beta = 1$ and $\alpha > 1$: $\omega(q) \sim q^{\alpha-1}$. An associated information measure is $\mathbf{H}(q) \sim -q^{\alpha+1}$, but taking advantage of the indeterminacy of information measures modulo affine functions, an equivalent but more pleasing choice is $\mathbf{H}(q) = (1-q^{\alpha})q$, which is normalized such that $\mathbf{H}(0) = \mathbf{H}(1) = 0$.

# 18 A data example for tailoring in classification trees

"Tailoring" means putting weight $\omega(q)$ on the class probabilities of greatest interest. In practice, areas of interest are often the extremes with highest or lowest class probabilities. In the Pima Indians Diabetes data (UCI ML Repository, Newman et al. 1998), for example, this may mean focusing on the cases with an estimated probability of diabetes of 0.9 or greater. We interpret this as meaning that increasing values of $q$ are of increasing interest, implying an ascending $\omega(q)$. In our experiments we used members of the Beta family that are simple power functions of $q$: $\omega(q) \sim q^{\alpha-1}$ ($\beta = 1$ and $\alpha > 1$). An associated information

measure is $\mathbf{H}(q) \sim -q^{\alpha+1}$, but taking advantage of the indeterminacy of information measures modulo affine functions, an equivalent but more pleasing choice is $\mathbf{H}(q) = (1 - q^{\alpha})q$, which is normalized such that $\mathbf{H}(0) = \mathbf{H}(1) = 0$.

Our experiments use again the Pima Indians Diabetes data (UCI ML Repository, Newman et al. 1998). Figure 18 shows a conventional tree grown with the Gini index ($\omega(q) =$ const). The depth to which it is grown and the quality of the fit is not of concern; instead, we focus on interpretability. This tree shows the usual relative balance whereby most splits are not more lopsided than about 1:2 in bucket size. Overall, interpretation is not easy, at least in comparison to the trees shown in the following two figures. The tree in Figure 18 is based on the parameters $\alpha = 16$ and $\beta = 1$, which means a strong focus on large class 1 probabilities (more correctly: class 1 frequencies). By contrast, Figure 18 is based on $\alpha = 1$ and $\beta = 31$, hence a strong focus on small class 1 probabilities.

The focus on the upper and the lower end of probabilities in these two trees amounts to prioritizing the splits according to decreasing and increasing class 1 probabilities from the top down. Figure 18 therefore shows a highly unbalanced tree that peels off small terminal nodes with high class 1 probabilities from the top down. The result is a cascading tree that layers the data as best as possible from highest to lowest class 1 probabilities. The dual effect is seen in Figure 18.

While it may be true that the lopsided focus of the criterion is likely to lead to suboptimal trees for prediction, cascading trees are vastly more powerful in terms of interpretation: they are able, for example, to express monotone dependences between predictors and responses. This is illustrated by both cascading trees: "plasma", the most frequent splitting variable, appears to correlate positively with the probability of diabetes; as we descend the trees and as the splitting values on "plasma" decrease in Figure 18 and increase in Figure 18, the class 1 probabilities decrease and increase, respectively. In Figure 18 the variable "b.mass" asserts itself as the second most frequent predictor, and again a positive association with class 1 probabilities can be gleaned from the tree. Similarly, Figure 18 exhibits positive dependences for "age" and "pedigree" as well.

A second aspect that helps the interpretation of cascading trees is the fact that repeat appearances of the same predictors lower the complexity of the description of low hanging nodes. For example, in Figure 18 the right most leaf at the bottom with the highest class 1 probability $q = 0.91$ is of depth nine, yet it does not require nine inequalities to describe it. Instead, the following four suffice: "$plasma > 155$", "$pedigree > 0.3$", "$b.mass > 29.9$" and "$age > 24$". Interestingly the tree of Figure 18 that peels off low class 1 probabilities ends up with a crisper high-probability leaf than the tree of Figure 18 whose greedy search for high class 1 probabilities results only in an initial leaf with $q = 0.86$ characterized by the single inequality "$plasma > 166$".

We reiterate and summarize: the extreme focus on high or low class 1 probabilities cannot be expected to perform well in terms of conventional prediction, but it may have merit in producing trees with vastly greater interpretability. The present approach to interpretable trees produces results that are similar to an earlier proposal by Buja and Lee (2001) based on splitting that maximizes the larger of the two class 1 probabilities. The advantage of the tailored trees introduced here is that there actually exists a criterion that is being minimized,

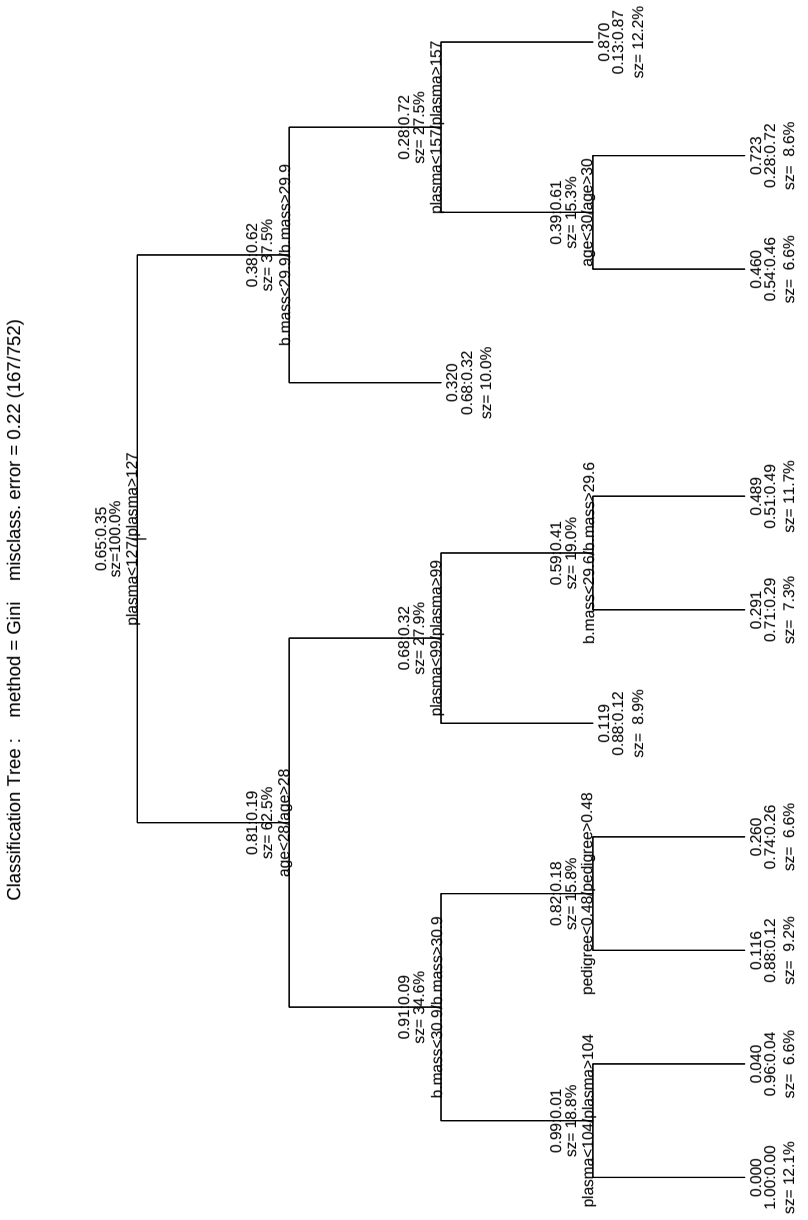


Figure 9: *Tree based on the Gini criterion, as in CART. Each split shows the predictor variable and the threshold that were used. Each node shows the ratio of class 0 to class 1 and the size of the node. The terminal nodes redundantly show the final fitted class 1 probability.*

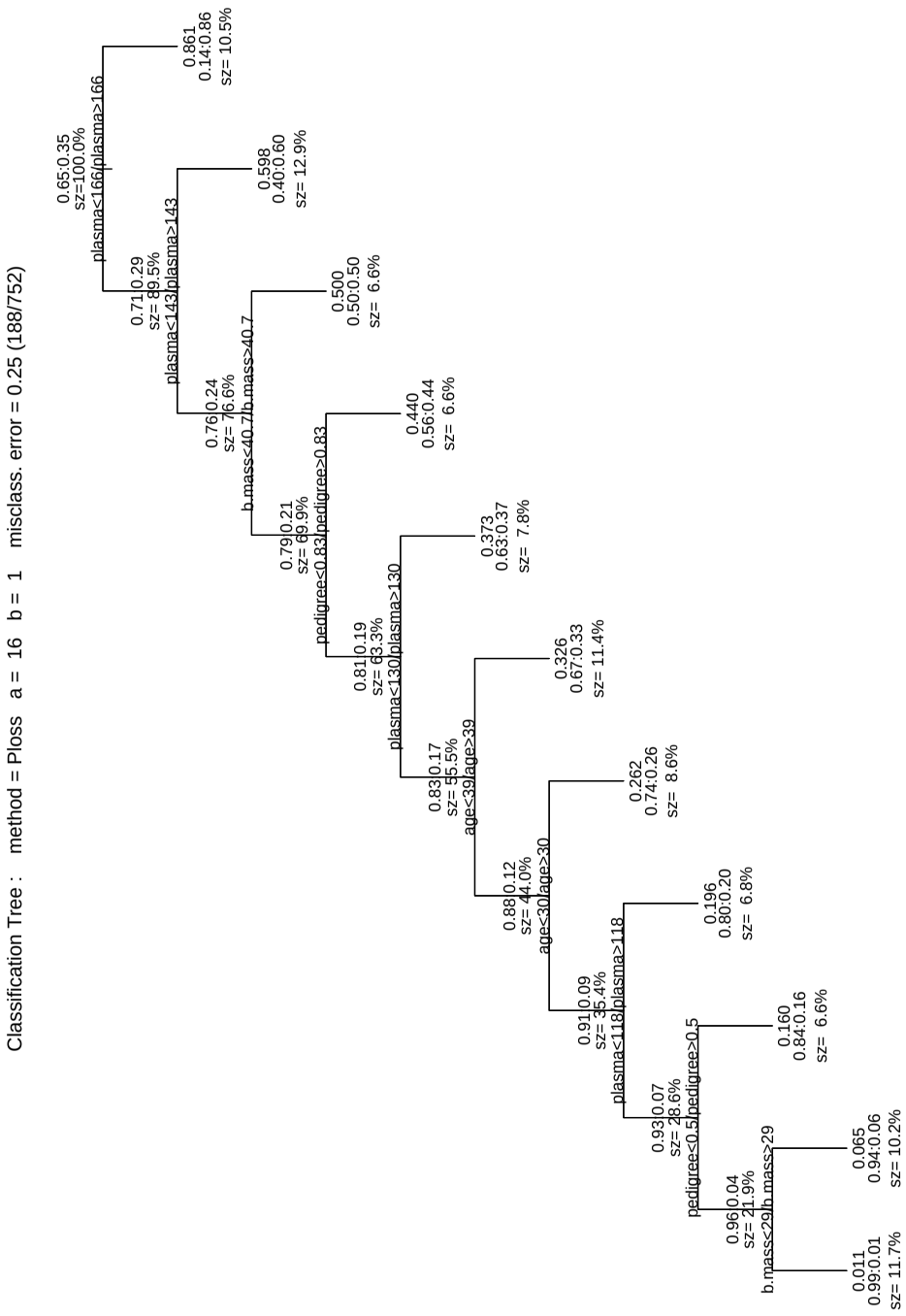


Figure 10: *Tree that focuses on large class 1 probabilities ($\alpha = 16$, $\beta = 1$).*

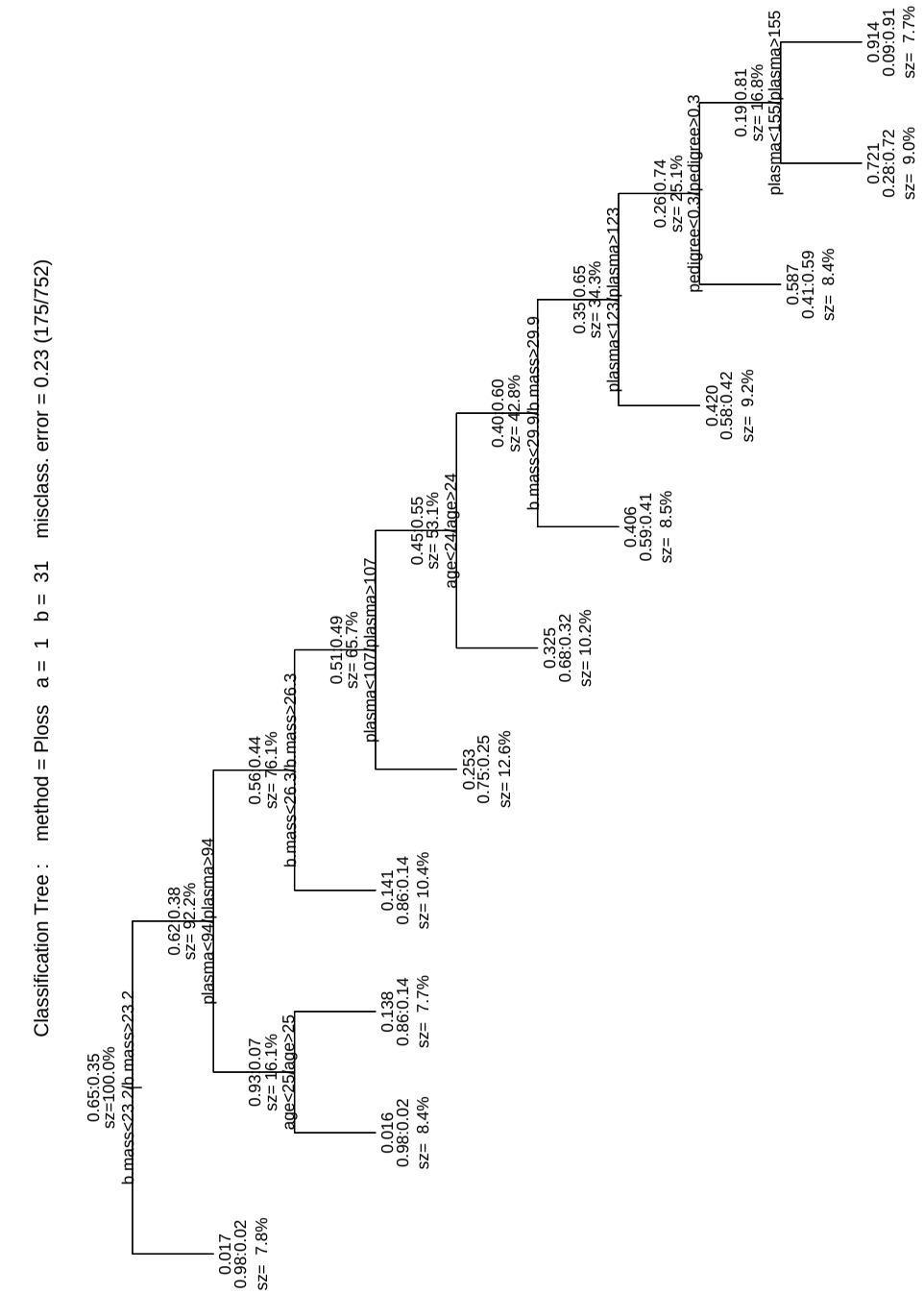


Figure 11: *Tree that focuses on small class 1 probabilities ($\alpha = 1$, $\beta = 31$).*

and tree-performance can be measured in terms of this criterion.

# 19 Proper scoring rules and their Bregman distances

Any proper scoring rule has an associated "Bregman distance" (Bregman 1967) defined as

$$\mathbf{B}(\eta|q) \;=\; \mathbf{R}(\eta|q) - \mathbf{H}(\eta) \;. \tag{52}$$

See again Figure 8. A Bregman distance is the expected loss of a proper scoring rule normalized such that a value zero obtains for the minimum at $q = \eta$:

$$\inf_q \mathbf{B}(\eta|q) \;=\; \mathbf{B}(\eta|\eta) \;=\; 0 \;.$$

Because $\mathbf{B}$ and $\mathbf{R}$ differ only in a function of $\eta$, minimizing one or the other w.r.t. $q$ is equivalent. Because $\mathbf{R}(\eta|q)$ is affine in $\eta$ and $\mathbf{H}(\eta)$ is concave, Bregman distances as defined here are convex in the first argument $\eta$.

There is a division of labor between proper scoring rules and Bregman distances in that the latter are technically useful in proofs whereas proper scoring rules are practically useful in estimation. The technical usefulness of Bregman distances stems from the fact that they describe the excess of the proper scoring rule of an estimate $q$ above the best achievable value of a proper scoring rule. If a procedure can be shown to have zero Bregman distance in the limit, it means that it achieves the minimal possible value of the proper scoring rule. Bregman distances are used in proving convergence of boosting algorithms (Lafferty et al. 1997, Collins et al. 2002) and in proving consistency of boosting-inspired algorithms as class probability estimation methods (Lugosi and and Vayatis 2004, Zhang 2004).

Bregman distances are non-negative but *not* generally symmetric in their arguments. They are therefore not metrics in the usual sense, yet by convention the term "distance" is used, although the variant "Bregman divergence" is in use also. The closest thing to squared distances is given by the following second order approximation which, however, exhibits asymmetry again:

$$\mathbf{B}(\eta|q) \;\approx\; \frac{1}{2}\omega(\eta)(q-\eta)^2 \;. \tag{53}$$

This follows from Equation (13). Below it will be used repeatedly in informal arguments.

In what follows we give a mixture representation for Bregman distances that derives from those for proper scoring rules and information measures. The new representation is useful because it allows us to derive simple bounds for cost-weighted misclassification losses.

We start with the usual examples from the Beta family which yield the following Bregman distances:

- $\alpha = \beta = -1/2$: Boosting loss leads to

$$\mathbf{B}(\eta|q) \;=\; \left[\frac{\eta}{q} + \frac{1-\eta}{1-q}\right] [q(1-q)]^{1/2} - 2 \cdot [\eta(1-\eta)]^{1/2} \;. \tag{54}$$

- $\alpha = \beta = 0$: Log-loss leads to Kullback-Leibler divergence,

$$\mathbf{B}(\eta|q) \;=\; -\,\eta \log \frac{q}{\eta} - (1-\eta) \log \frac{1-q}{1-\eta}\;. \tag{55}$$

- $\alpha = \beta = 1$: Squared error loss leads to squared deviation,

$$\mathbf{B}(\eta|q) \;=\; (\eta - q)^2\;. \tag{56}$$

Further examples are obtained from Basu, Harris, Hjort and Jones (1998) who proposed a family of power divergences. Its specialization to the binary case is a one-parameter family of Bregman distances that can be written as follows:

$$\begin{aligned}\mathbf{B}(\eta|q) \;&=\; \frac{1}{\gamma+1}\left(q^{\gamma+1} \;+\; (1-q)^{\gamma+1}\right) \;-\; \frac{1}{\gamma}\left(q^{\gamma}\,\eta \;+\; (1-q)^{\gamma}\,(1-\eta)\right) \\ &+\; \frac{1}{\gamma(\gamma+1)}\left(\eta^{\gamma+1} \;+\; (1-\eta)^{\gamma+1}\right)\end{aligned}$$

This version differs from Basu et al.'s orginal in that we divided by $\gamma+1$, which makes these well-defined Bregman distances for all real $\gamma$, with limiting cases for $\gamma \to 0,\; -1$ easily filled in. The associated proper scoring rule is (up to constants):

$$\begin{aligned}L_1(1-q) \;&=\; \frac{1}{\gamma+1}\left(\,q^{\gamma+1} + (1-q)^{\gamma+1}\right) \;-\; \frac{1}{\gamma}\,q^{\gamma} \\ L_0(q) \;&=\; \frac{1}{\gamma+1}\left(\,q^{\gamma+1} + (1-q)^{\gamma+1}\right) \;-\; \frac{1}{\gamma}\,(1-q)^{\gamma}\end{aligned}$$

Its weight function is

$$\omega(q) \;=\; q^{\gamma-1} + (1-q)^{\gamma-1}\;.$$

This shows immediately that the limiting case $\gamma \to 0$ results in $\omega(q) = 1/[q(1-q)]$, that is, log-loss. For $\gamma = 1$ one obtains $\omega(q) = \text{const}$, that is, squared error loss. Basu et al. (1998) use this family to interpolate these two cases with $0 \le \gamma \le 1$, with $\gamma = 0$ being the extreme of efficiency when the model is true and $\gamma = 1$ the extreme of robustness. The advantage of this family over the Beta family of Section 11 is that the losses $L_1(1-q)$ and $L_0(q)$ are given by explicit formulas, whereas the Beta family requires a numerical implementation of the partial Beta function.

# 20 Bregman distances are mixtures of excess-over-Bayes

The calculation of loss in excess of the best achievable loss can be made not only for strict proper scoring rules, but for the non-strict case of misclassification losses as well. If $c$ and $1-c$ are the costs of false positives and false negatives, respectively, then the associated Bregman semi-distance is the excess cost-weighted misclassification loss over Bayes risk, or *"excess-over-Bayes"* for short:

$$\mathbf{B}_c(\eta|q) \;=\; \mathbf{L}_c(\eta|q) \;-\; \mathbf{H}_c(\eta)\;. \tag{57}$$

The meaning of $\mathbf{B}_c(\eta|q)$ is easily seen from Figure 2 where the lightly shaded area is bounded by $(1-\eta)t$ and $\eta(1-t)$ so that the difference is $|(1-\eta)t - \eta(1-t)| = |\eta - t|$. One infers immediately the following facts:

$$\eta \leq c: \quad \mathbf{B}_c(\eta|q) \;=\; (c-\eta)\, 1_{[c<q]} \;, \qquad c < \eta: \quad \mathbf{B}_c(\eta|q) \;=\; (\eta - c)\, 1_{[q \leq c]} \;, \tag{58}$$

This can be summarized as follows:

**Lemma:**

$$\mathbf{B}_c(\eta|q) \;=\; |\eta - c| \cdot 1_{[\min(\eta,q) \leq c < \max(\eta,q)]} \tag{59}$$

This lemma is a pointwise version and generalization of an identity that is well-known in one form or another (see, e.g., Koltchinskii 2004, identity (1)).

The mixture representations for proper scoring rules and information measures immediately imply corresponding representations for Bregman distances:

**Proposition:** *Under the lower-bounding condition* $\int_0^1 t(1-t)\omega(dt) < \infty$*, the Bregman distance for* $\omega(dt)$ *is a mixture of cost-weighted excesses-over-Bayes:*

$$\mathbf{B}(\eta|q) \;=\; \int_0^1 \mathbf{B}_c(\eta|q)\, \omega(dc) \;.$$

From this proposition and the preceding lemma we immediately obtain:

**Corollary:**

$$\mathbf{B}(\eta|q) \;=\; \int_{\min(\eta,q)}^{\max(\eta,q)} |\eta - c|\, \omega(dc) \;. \tag{60}$$

This corollary is easily interpreted in terms of Figure 2 where the Bregman distance is essentially the integral of the lightly shaded area with regard to $\omega(dt)$.

# 21 A bias-variance decomposition for Bregman distances

The results of the previous section can be used to derive a "bias-variance" type of decomposition for Bregman distances. To this end, we assume $q$ to be a random variable with values in $[0,1]$. In any application $q$ will be a conditional estimate $q(\mathbf{x})$ of class probability 1 given a fixed predictor vector $\mathbf{x}$. The estimate is random due to fitting on a training dataset. Therefore, the expected value $\mathbf{E}q$, which we will need below, refers to averaging across training sets. For expected values of Bregman distances we write $\mathbf{E}_q\mathbf{B}(\eta|q)$ to indicate that only $q$ is considered random. If $\mathbf{E}_q\mathbf{B}(\eta|q)$ is a generalized mean squared error, then we can define the minimizing $\eta$ to be the generalized first order moment:

$$\eta_m \;=\; \mathrm{argmin}_\eta\; \mathbf{E}_q\, \mathbf{B}(\eta|q) \;.$$

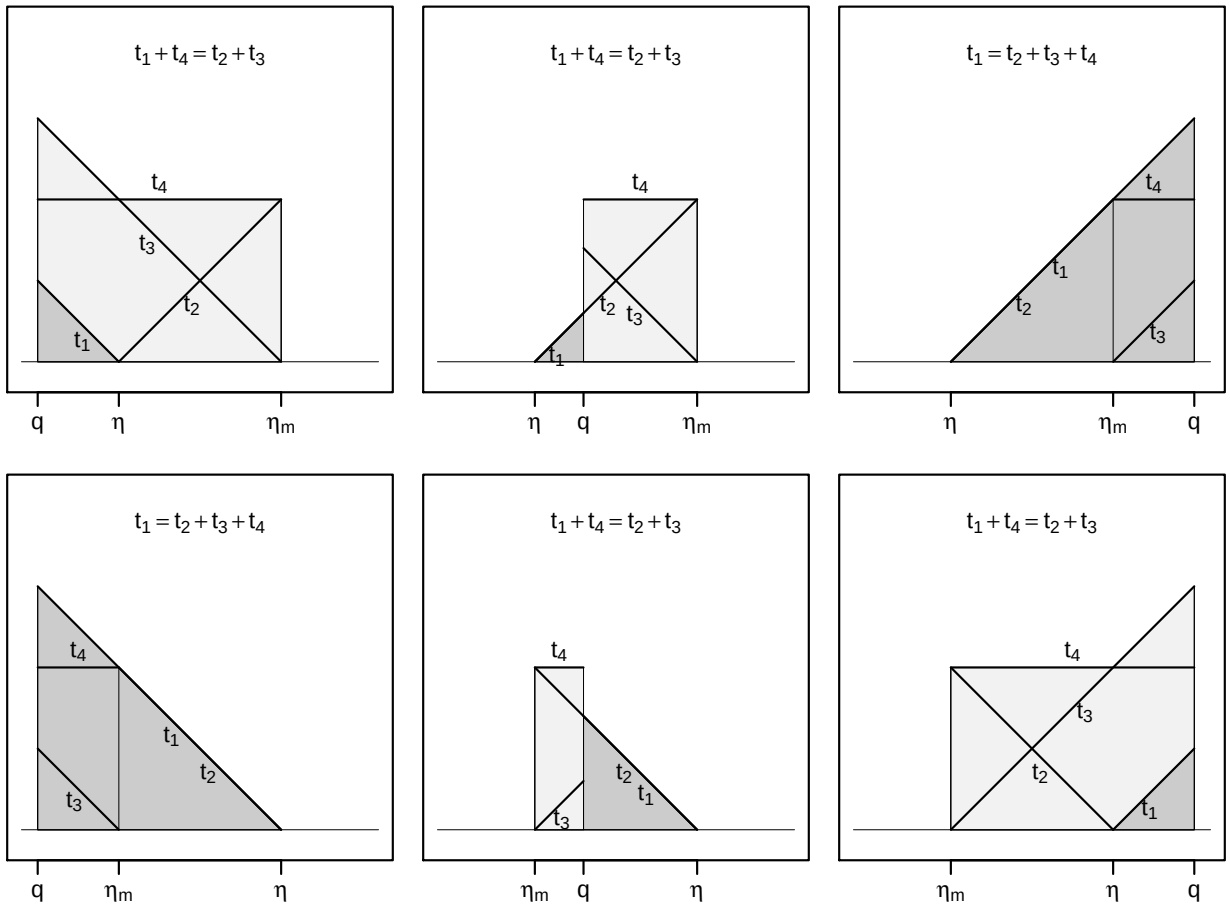


Figure 12: *Proof of the decomposition (62) for excess-over-Bayes: $t_1, ..., t_4$ are the magnitudes of the four terms of Equation (62), $t_1 = \mathbf{B}_c(\eta|q)$, $t_2 = \mathbf{B}_c(\eta|\eta_m)$, $t_3 = \mathbf{B}_c(\eta_m|q)$ and $t_4 = |\mathbf{S}_c(\eta, \eta_m|q)|$. Dark-gray indicates the first term $t_1$, which corresponds to the left hand side of (62). Note that the height of $t_4$ is $|\eta - \eta_m|$. The six frames correspond to all possible orderings of $q$, $\eta$ and $\eta_m$. For each ordering, the four terms add up as indicated at the top in each frame. These equalities are identical to (62) in each case, which completes the proof.*

With a stationariy condition it is easily seen that the idea of $\eta_m$ as a generalized moment is correct:

$$\frac{\partial}{\partial \eta} \mathbf{E}_q \mathbf{B}(\eta|q) \;=\; \mathbf{E}F(q) - F(\eta) \;=\; 0 \;, \tag{61}$$

which in turn follows from

$$\frac{\partial}{\partial \eta} \mathbf{B}(\eta|q) \;=\; F(q) - F(\eta) \;,$$

where $F(t) = \int^t \omega(c)dc$ is the canonical link function associated with $\omega(c)$, assuming that $\omega(dc) = \omega(c)dc$ is absolutely continuous. Equation (61) shows that

$$F(\eta_m) \;=\; \mathbf{E}\, F(q) \;.$$

Thus, if $\mathbf{E}_q\mathbf{B}(\eta|q)$ is a generalized mean squared error, then $\mathbf{E}_q\mathbf{B}(\eta_m|q)$ is a generalized variance, and consequently $\mathbf{B}(\eta|\eta_m)$ would be a generalized squared bias. It will turn out that these three terms will add up to form a generalized bias-variance decomposition.

We proceed with its derivation. The main idea is to prove a version for excess-over-Bayes and then use the mixture property of Bregman distances. The derivation relies on the piecewise linearity of the integrand $c \mapsto \mathbf{B}_c(\eta|q)$, which as a graph represents a triangle. The decomposition will amount to a jigsaw puzzle of triangles and rectangles, which is why we also need a piecewise constant function to represent rectangles, specifically defined as follows:

$$c \;\mapsto\; \mathbf{S}_c(\eta, \eta_m|q) \;=\; -\,(\eta - \eta_m)\,\mathrm{sign}(q - \eta_m)\,1_{[\min(q,\eta_m)\le c<\max(q,\eta_m)]}\;.$$

We use the symbols $q$, $\eta$ and $\eta_m$ which have specific meaning, but in this definition they are just three arbitrary values in $[0, 1]$. The definition is arranged such that $\mathbf{S}_c(\eta, \eta_m|q)$ takes on a positive sign exactly when $\eta_m$ is between $\eta$ and $q$. This property is necessary to obtain the following identity for arbitrary $q$, $\eta$ and $\eta_m$:

$$\mathbf{B}_c(\eta|q) \;=\; \mathbf{B}_c(\eta|\eta_m) \;+\; \mathbf{B}_c(\eta_m|q) \;+\; \mathbf{S}_c(\eta, \eta_m|q)\;. \tag{62}$$

This identity is informally proven in the legend of Figure 21. Next we integrate both sides of the identity w.r.t. $\omega(dc)$ and generalize it from excess-over-Bayes to general Bregman distances. Assuming absolute continuity, $\omega(dc) = \omega(c)dc$, we note from the definition of $\mathbf{S}_c(\eta, \eta_m|q)$ that

$$\int \mathbf{S}_c(\eta, \eta_m|q)\,\omega(dc) \;=\; -\,(\eta - \eta_m)\,(F(q) - F(\eta_m))\;,$$

where $F(q) = \int^q \omega(c)dc$ is the canonical link function associated with $\omega(dt)$ (Section 16). Therefore:

$$\mathbf{B}(\eta|q) \;=\; \mathbf{B}(\eta|\eta_m) \;+\; \mathbf{B}(\eta_m|q) \;-\; (\eta - \eta_m)\,(F(q) - F(\eta_m))\;. \tag{63}$$

This identity is reminiscent of an inner product expansion $\|a-b\|^2 = \|a\|^2+\|b\|^2-2\langle a, b\rangle$, and indeed it can be used to prove generalized Pythagorean theorems (see for example, Csiszár 1995; also Lafferty et al. 1997, Collins et al. 2002, Murata et al. 2004). In Pythagorean theorems one uses some sort of orthogonality between $\eta_m - \eta$ and $F(q) - F(\eta_m)$, for example, with regard to an inner product formed by integrating over $\mathbf{x}$ (Murata et al. 2004). Our goal, however, is a generalized bias-variance decomposition, which is why we now need the assumption that $q$ is a random variable. Unlike Pythagorean theorems, we construct $\eta_m$ in such a way that the term $F(q) - F(\eta_m)$ disappears:

$$F(\eta_m) \;=\; \mathbf{E}F(q)\;,$$

that is, we choose $\eta_m$ to be the generalized first moment with regard to $\mathbf{E}_q\mathbf{B}(\eta|q)$. Therefore:

**Theorem:** *If we assume the canonical link function $F$ associated with $\omega(c)$ is invertible and the expectation of $F(q)$ exists, and if $\eta_m$ satisfies $F(\eta_m) = \mathbf{E}F(q)$ or equivalently*

$\mathbf{E}_q \mathbf{B}(\eta_m|q) = \min_\eta \mathbf{E}_q \mathbf{B}(\eta|q)$, *we obtain the following generalized bias-variance decomposition:*

$$\mathbf{E}_q \, \mathbf{B}(\eta|q) \;=\; \mathbf{B}(\eta|\eta_m) \;+\; \mathbf{E}_q \, \mathbf{B}(\eta_m|q) \; . \tag{64}$$

We note that the moment condition $F(\eta_m) = \mathbf{E}F(q)$ is simple only when the canonical link is used to form the composite loss function. In this case $F(q(\mathbf{x})) = \hat{\mathbf{b}}^T\mathbf{x}$ for a linear model, say, hence $F(\eta_m) = \mathbf{E}\hat{\mathbf{b}}^T\mathbf{x}$.

Examples: For squared error loss, $F()$ is linear, hence $\eta_m = \mathbf{E}\, q$, and the decomposition specializes to the actual bias-variance decomposition because the Bregman distance is the square of the difference (56):

$$\mathbf{E}_q \, [(q-\eta)^2] \;=\; (\eta_m - \eta)^2 \;+\; \mathrm{Var}\, q \; .$$

For log loss, $\eta_m$ is defined by $\log(\eta_m/(1-\eta_m)) = \mathbf{E}\, \log(q/(1-q))$, and the decomposition is for the Kullback-Leibler divergence (55). For boosting loss, the definition of $\eta_m$ is *not* the same because (half) the logit is the actual link function used the composite exponential loss, but it is not the canonical link function. The bias-variance decomposition is for the Bregman distance (54).

# 22 Bounds on cost-weighted misclassification loss

We can use the mixture representation to derive bounds on $\mathbf{B}_c(\eta|q)$ in terms of $\mathbf{B}(\eta|q)$. We use the following facts:

- A Bregman distance as a function of $q$, $q \mapsto \mathbf{B}(\eta|q)$, is descending for $q < \eta$ and ascending for $q > \eta$. (This follows from $\mathbf{B}(\eta|q) = \mathbf{R}(\eta|q) - \mathbf{H}(\eta)$ and the corollary at the end of Section 6.)
- Excess-over-Bayes as a function of $q$, $q \mapsto \mathbf{B}_c(\eta|q)$, is a step function with a step at $q = c$. (This follows from (58) above.)

Thus, if we standardize both a Bregman distance and excess-over-Bayes such that their graphs pass through 1 at $q = c$, the former will dominate the latter, as depicted in Figure 13:

$$\left( \frac{\mathbf{B}_c(\eta|q)}{|\eta - c|} \right)^k \;\leq\; \frac{\mathbf{B}(\eta|q)}{\mathbf{B}(\eta|c)} \; , \qquad \text{for arbitrary} \;\; k > 0 \; . \tag{65}$$

The power is permitted because it is vacuous as applied here to a 0-1 step function. We obtain the following initial bound:

**Lemma:**

$$\mathbf{B}_c(\eta|q)^k \;\leq\; \frac{|\eta - c|^k}{\mathbf{B}(\eta|c)} \, \mathbf{B}(\eta|q) \; . \tag{66}$$

Note that the ratio factor on the right hand side is independent of $q$; it is a pure function of the class probabilities $\eta$ and the chosen classification threshold $c$, but not the estimate $q$.

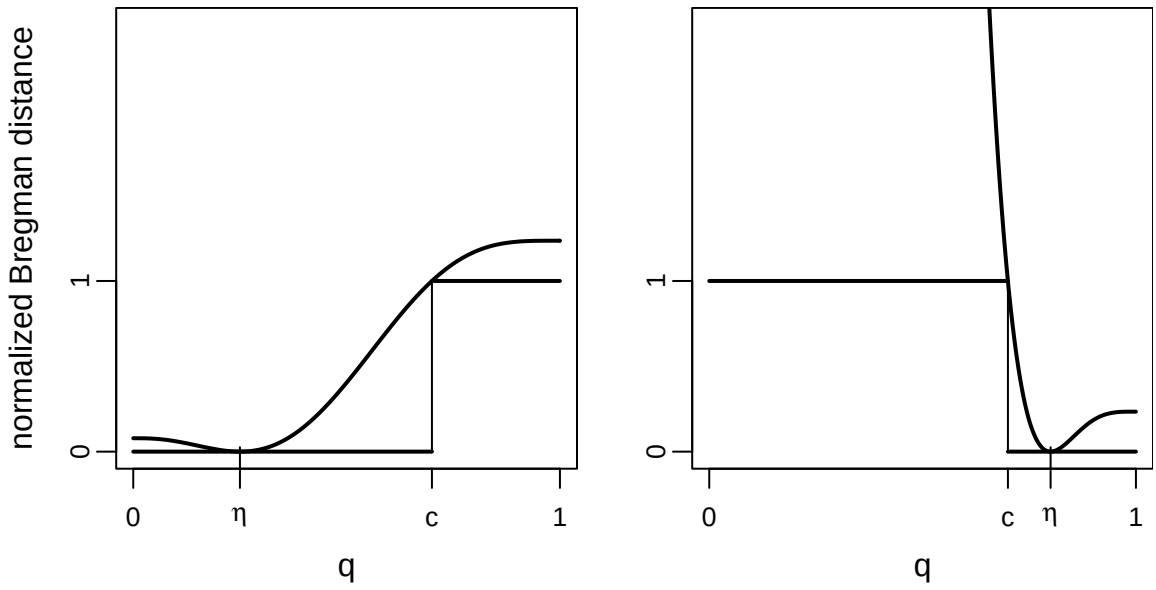


Figure 13: *Bregman distance as a bound on excess-over-Bayes: The curves show sections of a normalized Bregman distance as a function of $q$ for fixed $\eta$: $q \mapsto \mathbf{B}(\eta|q)/\mathbf{B}(\eta|c)$ (right hand side of Inequality* (65)*). The normalization has the sections go through 1 at a chosen classification threshold $c$. Also shown is a step function representing excess-over-Bayes as a function of $q$, $q \mapsto \mathbf{B}_c(\eta|q)/|\eta - c|$ equally normalized to 1 at $c$ (left hand side of Inequality* (65)*). The two plots reflect the situations $\eta < c$ and $\eta > c$. The bound is rather loose for $\eta$ near $c$ as the right hand plot shows.*

The problem with the term is that it is possibly ill-behaved because the denominator vanishes at $\eta = c$, unless the power $k$ in the numerator is large enough to cause cancellation. This can indeed be achieved for $k = 2$, as is intuitively clear from Equation (53): $\mathbf{B}(\eta|c) \approx \frac{1}{2}\omega(\eta)(\eta - c)^2$. A rigorous quadratic lower bound on the denominator $\mathbf{B}(\eta|c)$ is as follows, making use of the mixture representation (60):

$$\begin{aligned}
\mathbf{B}(\eta|c) &= \int_{\min(\eta,c)}^{\max(\eta,c)} |\eta - c'|\, \omega(c')\, dc' \\
&\geq \int_{\min(\eta,c)}^{\max(\eta,c)} |\eta - c'|\, dc' \cdot \min_{\min(\eta,c) \leq c" < \max(\eta,c)} \omega(c") \\
&= \frac{1}{2}|\eta - c|^2 \cdot \min_{\min(\eta,c) \leq c" < \max(\eta,c)} \omega(c")
\end{aligned}$$

If we apply this bound with $k = 2$ in Equation (66) we obtain:

$$\mathbf{B}_c(\eta|q)^2 \;\leq\; \frac{2}{\min_{\min(\eta,c) \leq c" < \max(\eta,c)} \omega(c")}\, \mathbf{B}(\eta|q)\ . \tag{67}$$

We consider the application to two cases:

- Assume $\omega(q) \geq 1$ for all $q$. Then a bound is:

$$\mathbf{B}_c(\eta|q)^2 \;\leq\; 2\, \mathbf{B}(\eta|q)\ . \tag{68}$$

The assumption is true, for example, for the members of the Beta family for $\alpha,\ \beta \leq 1$, including squared error loss ($\alpha = \beta = 1$), log-loss ($\alpha = \beta = 0$), and boosting loss ($\alpha = \beta = -1/2$).

- Assume the weight function attains its maximum at $c$ and descends with increasing distance from $c$. Then a bound is:

$$\mathbf{B}_c(\eta|q)^2 \ \leq \ \frac{2}{\omega(\eta)}\, \mathbf{B}(\eta|q)\ . \tag{69}$$

The assumption is satisfied for the members of the Beta family for $\alpha,\ \beta > 1$ when tailored for the threshold $c$ by matching the mode of the weight function $\omega()$ according to (38) above: $(\alpha - 1)/(\beta - 1) = c/(1 - c)$, so that the maximum of $\omega()$ is attained at $c$. Because of (53) above, the right hand side of (69) is of order $\approx (\eta - q)^2$.

A note on Zhang (2004): The above bounds explain why Zhang in his Theorem 2.1 required a condition which in our notation says $|\eta - 0.5|^k \leq C \cdot \mathbf{B}(\eta|0.5)$. Equally, it becomes clear why in all of his examples he ended up with $k = 2$.

Bartlett, Jordan, McAuliffe (2003) developed a technique for tightly bounding misclassification loss with equal costs by convex losses such as exponential loss. We can reconstruct their technique in terms of Bregman distances associated with proper scoring rules. This reconstruction applies to any proper scoring rule and to classification with any cost weight $c$ without convexity assumptions.

**Lemma:** *Let $\psi_c(t)$ be a convex function that satisfies $\psi_c(|\eta - c|) \ \leq \ \mathbf{B}(\eta, c)$, $\forall \eta \in (0, 1)$. Then $\psi_c(\mathbf{B}_c(\eta|q) \ \leq \ \mathbf{B}(\eta|q)$.*

**Proof:** In the following chain of inequalities, the first is by assumption while the second is due to the fact that $\eta$ and $q$ are on opposite sides of $c$ and $q \mapsto \mathbf{B}(\eta|q)$ is descending to the left of $\eta$ and ascending to the right of $\eta$.

$$\begin{aligned}
\psi_c(\mathbf{B}_c(\eta|q) \ &= \ \psi_c(|\eta - c|) \cdot 1_{[\min(\eta,q) \leq c < \max(\eta,q)]} \\
&\leq \ \mathbf{B}(\eta|c) \cdot 1_{[\min(\eta,q) \leq c < \max(\eta,q)]} \\
&\leq \ \mathbf{B}(\eta|q) \cdot 1_{[\min(\eta,q) \leq c < \max(\eta,q)]} \\
&\leq \ \mathbf{B}(\eta|q) \qquad \square
\end{aligned}$$

Now the question is how to find functions $\psi_c(t)$. One could follow the lead of Bartlett et al. (2003) and construct $\psi_c(t)$ as the largest convex function dominated by both $t \mapsto \mathbf{B}(c + t|c)$ (for $t \leq 1 - c$) and $t \mapsto \mathbf{B}(c - t|c)$ (for $t \leq c$). However, again in light of (53), $\mathbf{B}(\eta|c) \approx \frac{1}{2}\omega(\eta)(\eta - c)^2$, one cannot expect better than $\psi_c(t) \sim (t - c)^2$. Thus this technique cannot improve on the bounds found earlier in this section.

At this point one could execute a program similar to that carried out by Bartlett et al. (2003) and Zhang (2004) who assembled a body of approximation and consistency theory for a small collection of loss functions. The above exercise in bounding excess-over-Bayes in

terms of Bregman distances could be used to generalize much of their programs to proper scoring rules in combination with arbitrary link functions. To this end one would consider estimates of class probabilities, $\hat{q}_N(X)$, and true class probabilities $\eta(X)$ as functions on predictor space and examine the limiting behavior of $\mathbf{E}\,\mathbf{B}(\eta(X)|\hat{q}_N(X))$. We will not carry this program forward but end by comparing Zhang's and our analyses:

- The assumption of convexity on the modeling scale, made by Zhang (2004) and others, can be replaced by the more natural assumption that the loss function is decomposable into a link function and a proper scoring rule. We thereby exclude some cases that do not confine probability estimates to $[0, 1]$, but we benefit from the full information geometry that links Bregman distances and Bayes risks, even for combinations of proper scoring rules and link functions that result in non-convex losses, such as those for tailoring to specific costs $c$ (Section 12) and implied by HV (2003) (Section 13).
- Much of Zhang's (2004, Section 2 and 3) interpretations hinge on bounds on misclassification loss for $c = 0.5$. We extend these bounds to arbitrary cost weights $c \in (0, 1)$ and non-convex losses, but our interpretations flow straighforwardly from the mixture representations: the costs $c$ on which the weight function $\omega(c)$ puts the most mass are those for which the proper scoring rule attempts classification with the greatest precision.

# 23 Summary and Discussion

We developed a theory of binary class probability estimation based on Fisher-consistent loss functions, known as "proper scoring rules". We showed that proper scoring rules $\mathbf{L}(y|q)$ have a simple structure in that they are mixtures of cost-weighted misclassification losses: $\mathbf{L}(y|q) = \int \mathbf{L}_c(y|q)\,\omega(c)dc$, where $c$ and $1-c$ are the costs of false positives and false negatives, respectively. The weight function $\omega(c)$ has an immediate interpretation: the proper scoring rule emphasizes cost weights $c$ to the degree that $\omega(c)$ puts mass on the costs $c$. This provides the heuristic for "tailoring" of proper scoring rules to arbitrary costs $c$ and $1 - c$, in particular costs other than 0.5. Tailoring addresses situations in which false positives and false negatives have different cost implications and where the class of fits succeeds for classification but fails for global class probability estimation. We showed that any proper scoring rule combined with any link function can be used to fit linear and additive class probability models to binary response data. The IRLS algorithm used to this end can be given a stagewise additive and hence boosting-like form, with LogitBoost (FHT 2000) as a particular instance. Furthermore, information measures associated with proper scoring rules can equally be tailored and used for estimating tree-based models in novel ways.

Although we see our study as having practical implications mostly for linear and tree models, it was initially motivated by boosting (FHT 2000) and it does indeed address two conceptual limitations of the boosting literature: One limitation has to do with the notion of "large-margin classifiers", where margin is defined as $-y^*\,F(\mathbf{x})$ ($y^* = \pm 1$) and loss functions $L(-y^*\,F(\mathbf{x}))$ are assumed monotone increasing and convex in the margin. This dependence on the margin limits all theory and practice to a symmetric treatment of class 0 and class 1,

and in particular to the equal-cost case $c = 1 - c = 0.5$. The possibility of asymmetric treatment is apparent from the identity $L(-y^* F(\mathbf{x})) = yL(-F(\mathbf{x})) + (1-y)L(F(\mathbf{x}))$ which could be freed of the constraint of equal partial losses $L()$ in the two terms. — The second limitation of the boosting literature has to do with the idea that $L()$ should be convex. We gave evidence that convexity is not a structural necessity for large-margin classifiers and that a more plausible requirement is decomposability of the loss function into a (inverse) link function $q(F)$ and a proper scoring rule $\mathbf{L}(y|q)$:

$$\mathbf{L}(y|q(F(\mathbf{x}))) = yL_1(1 - q(F(\mathbf{x}))) + (1-y)L_0(q(F(\mathbf{x}))) \ .$$

This approach makes it plain that margin is linked to class probabilities and that the loss function is meant to estimate class probabilities in a Fisher consistent manner. Link functions and class probability estimation have been lurking in the literature, in particular in the "Three Papers on Boosting" (Jiang; Lugosi and Vayatis; Zhang; 2004). This leaves us with the old question of why boosting so often classifies well even though it vastly overfits in terms of class probabilities (Mease et al. 2005). The answer is that boosting does rely on class probability gradients without attempting to estimate class probability with any precision. This explains what should not be a surprise: boosting can improve misclassification error out-of-sample even after it has dropped to zero in-sample — the surrogate criterion keeps borrowing strength from data points with estimated probabilities above and below the cut-off, $c = 0.5$, even if these estimates run off to 0 and 1 due to overfit. The degree of reliance on these estimates is revealed by the weight function $\omega(c)$.

# References


[1] BARTLETT, P. L., JORDAN, M. I., and MCAULIFFE, J. D. (2003). Large margin classifiers: convex loss, low noise, and convergence rates. In: *Advances in Neural Information Processing Systems* **16**.

[2] BARTLETT, P. L., JORDAN, M. I., and MCAULIFFE, J. D. (2004). Discussion of Jiang (2004), Lugosi and Vayatis (2004) and Zhang (2004). *The Annals of Statistics*, **32**, 85-91.

[3] BASU A., HARRIS, I. R., HJORT, N.L., and JONES, M.C. (1998), Robust and efficient estimation by minimising a density power divergence, Biometrika, 85, 549-559.

[4] BERNARDO, J. M., and SMITH, A. F. M. (2000). *Bayesian Theory*. Chichester, UK: Wiley and Sons.

[5] BREGMAN, L. M. (1967). The Relaxation Method of Finding the Common Point of Convex Sets and its Applications to the Solution of Problems in Convex Programming. *U.S.S.R. Computational Mathematics and Mathematical Physics* **7** (1), 200-217.

[6] BREIMAN, L. (1996). Bias, variance and arcing classifiers. Technical Report 460, Statistics Department, University of California at Berkeley, Berkeley, CA.

[7] BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R., and STONE, C. J. (1984). *Classification and Regression Trees*, Belmont, California: Wadsworth.

[8] BRIER, G. W. (1950). Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*, **78**, 1-3.

[9] BUJA, A., SHEN, Y., and STUETZLE, W. (2005). *Cost-Weighted Misclassification and Class Probability Estimation* Technical Report, http://www-stat.wharton.upenn.edu/˜buja/paper-cost-weighting.pdf

[10] BUJA, A., and LEE, Y.-S. (2001). *Data Mining Criteria for Tree-Based Regression and Classification*, Proceedings of KDD 2001, 27-36.

[11] CLARK, L. A., and PREGIBON, D. (1992). Tree-Based Models, in *Statistical Models in S*, edited by J. M.CHAMBERS and T. J. HASTIE, Pacific Grove, CA: Wadsworth & Brooks/Cole, 377-419.

[12] COLLINS, M., SCHAPIRE, R. E., and SINGER, Y. (2002). Logistic Regression, AdaBoost and Bregman Distances, *Machine Learning*, **48** (1/2/3).

[13] CRESSIE, N. and READ, T. R. C. (1984). Multinomial Goodness-of-fit Tests, *J. R. Statist. Soc.* B, **46**, No. 3, 440-464.

[14] CSISZÁR, I. (1995). Maxent, mathematics, and information theory. In *Proceedings of the Fifteenth International Workshop on Maximum Entropy and Bayesian Methods*, 35-50.

[15] DEGROOT, M., and FIENBERG, S.E. (1983). The Comparison and Evaluation of Probability Forecasters. *The Statistician* **32**, 12-22.

[16] GNEITING, T., and RAFTERY, A. E. (2004). Strictly Proper Scoring Rules, Prediction, and Estimation. Technical Report no. 463, Dept. of Statistics, Univ. of Washington. Preprint available from http://www.stat.washington.edu/www/research/reports/2004/tr463.pdf

[17] FREUND, Y., and SCHAPIRE, R. (1996). Experiments with a New Boosting Algorithm. In: *Machine Learning: Proceedings of the Thirteenth International Conference*, 148-156.

[18] FREUND, Y., and SCHAPIRE, R. (2004). Discussion of Jiang (2004), Lugosi and Vayatis (2004) and Zhang (2004). *The Annals of Statistics*, **32**, 113-117.

[19] FRIEDMAN, J.H., HASTIE, T., and TIBSHIRANI, R. (2000). [Abbreviated FHT.] Additive Logistic Regression: A Statistical View of Boosting, *The Annals of Statistics* **28**, 337-407.

[20] FRIEDMAN, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* **29**, 1189-1232.

[21] HAND, D. J., and VINCIOTTI, V. (2003). [Abbreviated HV.] Local Versus Global Models for Classification Problems: Fitting Models Where it Matters. *The American Statistician* **57**, 124-131.

[22] JIANG, W. (2004). Process Consistency for AdaBoost. *The Annals of Statistics* **32**, 13-29.

[23] KEARNS, M., and MANSOUR, Y. (1996). On the boosting ability of top-down decision tree learning algorithms. *Proceedings of the Annual ACM Symposium on the Theory of Computing*, 459-468.

[24] KOLTCHINSKII, V. (2004). "Discussion" of "Three Papers on Boosting". *The Annals of Statistics* **32**, 107-113.

[25] LAFFERTY, J. D., DELLA PIETRA S., and DELLA PIETRA V. (1997). Statistical Learning Algorithms based on Bregman Distances, in: *Proceedings of the Canadian Workshop on Information Theory 1997.*

[26] LUGOSI, G., and VAYATIS, N. (2004). On the Bayes-Risk Consistency of Regularized Boosting Methods, *The Annals of Statistics* **32**, 30-55.

[27] MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models* (2nd edition), London: Chapman & Hall/CRC.

[28] MEASE, D., WYNER, A. J., and BUJA, A. (2005). Boosted Classication Trees and Probability/Quantile Estimation. Submitted.

[29] MURATA, N., TAKENOUCHI, T., KANAMORI, T., and EGUCHI, S. (2004). Information Geometry of U-Boost and Bregman Divergence. *Neural Computation*, **16**, 1437-1481.

[30] MURPHY, A.H., and DAAN, H. (1985). Forecast Evaluation, in: *Probability, Statistics and Decision Making in the Atmospheric Sciences*, eds. Murphy, A.H. and Katz, P.W.

[31] QUINLAN, J.R. (1993). *C4.5: Programs for Machine Learning.* San Francisco, CA: Morgan Kaufmann Publishers.

[32] SAVAGE, L.J. (1971). Elicitation of Personal Probabilities and Expectations, *J. of the American Statistical Association* **66**, No. 336, 783-801.

[33] SCHERVISH, M.J. (1989). A General Method for Comparing Probability Assessors, *The Annals of Statistics* **17**, 1856-1879.

[34] SHUFORD, E. H., ALBERT, A., and MASSENGILL, H. E. (1966). Admissible Probability Measurement Procedures, *Psychometrika* **31**, 125-145.

[35] SCHAPIRE, R. E., and SINGER, Y. (1998). Improved Boosting Algorithms Using Confidence-Rated Predictions, in *Proceedings of the Eleventh Annual Conference on Computational Learning Theory.*

[36] SCHAPIRE, R. E. (2002). The Boosting Approach to Machine Learning: An Overview. In: *MSRI Workshop on Nonlinear Estimation and Classification*.

[37] SHAO, J. (2003). *Mathematical Statistics*, New York: Springer.

[38] NEWMAN, D. J., HETTICH, S., BLAKE, C. L., and MERZ, C.J. (1998). UCI Repository of machine learning databases [http://www.ics.uci.edu/˜mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science.

[39] WINKLER, R. L (1993). Evaluating Probabilities: Asymmetric Scoring Rules, *Management Science* **40**, No. 11, 1395-1405.

[40] ZHANG, T. (2004). Statistical Behavior and Consistency of Classification Methods based on Convex Risk Minimization. *The Annals of Statistics* (to appear).