Novel Decompositions of Proper Scoring Rules for Classification: Score Adjustment as Precursor to Calibration

Meelis Kull(✉) and Peter Flach

Intelligent Systems Laboratory, University of Bristol, Bristol, UK
{meelis.kull,peter.flach}@bristol.ac.uk

Abstract. There are several reasons to evaluate a multi-class classifier on other measures than just error rate. Perhaps most importantly, there can be uncertainty about the exact context of classifier deployment, requiring the classifier to perform well with respect to a variety of contexts. This is commonly achieved by creating a scoring classifier which outputs posterior class probability estimates. Proper scoring rules are loss evaluation measures of scoring classifiers which are minimised at the true posterior probabilities. The well-known decomposition of the proper scoring rules into calibration loss and refinement loss has facilitated the development of methods to reduce these losses, thus leading to better classifiers. We propose multiple novel decompositions including one with four terms: adjustment loss, post-adjustment calibration loss, grouping loss and irreducible loss. The separation of adjustment loss from calibration loss requires extra assumptions which we prove to be satisfied for the most frequently used proper scoring rules: Brier score and log-loss. We propose algorithms to perform adjustment as a simpler alternative to calibration.

1 Introduction

Classifier evaluation is crucial for building better classifiers. Selecting the best from a pool of models requires evaluation of models on either hold-out data or through cross-validation with respect to some evaluation measure. An obvious choice is the same evaluation measure which is later going to be relevant in the model deployment context.

However, there are situations where the deployment measure is not necessarily the best choice, as in model construction by optimisation. Optimisation searches through the model space to find ways to improve an existing model according to some evaluation measure. If this evaluation measure is simply the error rate, then the model fitness space becomes discrete in the sense that there are improvements only if some previously wrongly classified instance crosses the decision boundary. In this case, surrogate losses such as quadratic loss, hinge loss or log-loss enable SVMs, logistic regression or boosting to converge towards better models.

A. Appice et al. (Eds.): ECML PKDD 2015, Part I, LNAI 9284, pp. 68–85, 2015.
DOI: 10.1007/978-3-319-23528-8_5

The second situation where the choice of evaluation measure is non-trivial is when the exact context of model deployment is unknown during model training. For instance, the misclassification costs or deployment class distribution might be unknown. In such cases a scoring classifier is more versatile than a crisp classifier, because once the deployment context becomes known, the best decision can be made using ROC analysis by finding the optimal score threshold. Particularly useful are scoring classifiers which estimate class probabilities, because these are easiest to adapt to different contexts.

Proper scoring rules are loss measures which give the lowest losses to the ideal model outputting the true class posterior probabilities. Therefore, using a proper scoring rule as model evaluation measure helps to develop models which are good class probability estimators, and hence easy to adapt to different contexts. The best known proper scoring rules are log-loss and Brier score, both of which we are concentrating on in this paper. These two are also frequently used as surrogate losses for optimisation.

In practice it can be hard to decide which proper scoring rule to use. According to one view this choice could be based on the assumptions about the probability distribution over possible deployment contexts. For example, [6] shows that the Brier score can be derived from a particular additive cost model.

Once the loss measure is fixed, the best model has to be found with respect to that measure. The decomposition of expected loss corresponding to any proper scoring rule into calibration loss and refinement loss has facilitated the development of calibration methods (*i.e.* calibration loss reduction methods) which have been shown to be beneficial for classification performance [2]. Another decomposition[1] splits refinement loss into *uncertainty* minus *resolution* [5,9]. Interestingly, none of the decompositions relates to the loss of the optimal model. This inspires our first novel decomposition of any proper scoring rule loss into epistemic loss and irreducible (aleatoric[2]) loss. Irreducible loss is the loss of the optimal model which outputs the true posterior class probability given the instance.

For our second decomposition we introduce a novel *adjustment loss*, which is extra loss due to the difference between the average of estimated scores and the class distribution. For both Brier score and log-loss we propose a corresponding adjustment procedure, which reduces this loss to zero, and hence decreases the overall loss. This procedure uses only the output scores and class distribution and not the feature values. Therefore, it can easily be used in any context, whereas a calibration procedure needs to make extra assumptions about the shape of the calibration map.[3]

Finally, we propose a four-way decomposition by combining the decompositions relating to the notions of optimality, calibration and adjustment. The separation of adjustment loss from calibration loss is specific to the proper scoring

[1] Note that the commonly used bias-variance decompositions apply to the loss of a learning algorithm, whereas we are studying the loss of a particular model.

[2] Our terminology here relates to epistemic and aleatoric uncertainty [10].

[3] In some literature a classifier has been called *calibrated* when it is actually only *adjusted*, a confusion that we hope to help remove by giving a name for the latter.

rule (*i.e.* it relies on the existence of an adjustment procedure) whereas the remainder of the decomposition applies to any proper scoring rule. The decomposition has the following terms: adjustment loss (AL), post-adjustment calibration loss (PCL), grouping loss (GL) and irreducible loss (IL). Grouping loss is the divergence of calibrated probabilities from the true posterior probabilities and intuitively measures the loss due to the model assigning the same score to (*i.e.* grouping together) instances which have different posterior class probabilities (cf. refinement loss is the loss due to the same scores being assigned to instances from different classes). Grouping loss has earlier been introduced in [7] where it facilitated the improvement of probability estimation and classification using reliability maps, which quantify conditional grouping loss given the model output score. Our proposed decompositions aim to provide deeper insight into the causes behind losses and facilitate development of better classification methods.

The structure of the paper is as follows: Section 2 defines proper scoring rules and introduces notation. Section 3 provides two decompositions using ideal scores and calibrated scores, respectively. Section 4 introduces the notion of adjustment and a decomposition using adjusted scores. Section 5 provides two theorems from which all decompositions follow, and provides terminology for the obtained decomposed losses. Section 6 describes two proposed algorithms and the results of convergence experiments. Section 7 discusses related work and Section 8 concludes.

2 Proper Scoring Rules

2.1 Scoring Rules

Consider the task of multi-class classification with k classes. We represent the true class of an instance as a vector $y = (y_1, \ldots, y_k)$ where $y_j = 1$ if the true class is j, and $y_j = 0$ otherwise. Let $p = (p_1, \ldots, p_k)$ be an estimated class probability vector for an instance, *i.e.* $p_j \geq 0$, $j = 1, \ldots, k$ and $\sum_{j=1}^{k} p_j = 1$. A scoring rule $\phi(p, y)$ is a non-negative measure measuring the goodness of match between the estimated probability vector p and the true class y.

Two well known scoring rules are log-loss ϕ^{LL} (also known as ignorance score) and Brier score ϕ^{BS} (also known as squared loss or quadratic score), defined as follows:

$$\phi^{\mathsf{LL}}(p, y) := -\log p_y \qquad \text{log-loss,}$$

$$\phi^{\mathsf{BS}}(p, y) := \sum_{i=1}^{k} (p_i - y_i)^2 \qquad \text{Brier score}^{4},$$

where by a slight abuse of notation p_y denotes p_j for j such that $y_j = 1$. Both these rules are *proper* in the sense defined in the following subsection. Note that the scoring rules apply for a single instance, application to a dataset is by averaging across all instances.

[4] This Brier score definition agrees with the original definition by Brier [3]. Since it ranges between 0 and 2, sometimes half of this quantity is also referred to as Brier score.

2.2 Divergence, Entropy and Properness

Suppose now that the true class y is being sampled from a distribution q over classes (*i.e.* q is a probability vector). We denote by $\mathsf{s}(p,q)$ the expected score with rule ϕ on probability vector p with respect to the class label drawn according to q:

$$\mathsf{s}(p,q) := \mathbb{E}_{Y\sim q}\phi(p,Y) = \sum_{j=1}^{k} \phi(p,e_j)q_j \ ,$$

where e_j denotes a vector of length k with 1 at position j and 0 everywhere else. We define divergence $\mathsf{d}(p,q)$ of p from q and entropy $\mathsf{e}(q)$ of q as follows:

$$\mathsf{d}(p,q) := \mathsf{s}(p,q) - \mathsf{s}(q,q) \ , \qquad\qquad \mathsf{e}(q) := \mathsf{s}(q,q) \ .$$

A scoring rule ϕ is called proper if the respective divergence is always non-negative, and strictly proper if additionally $\mathsf{d}(p,q) = 0$ implies $p = q$. It is easy to show that both log-loss and Brier score are strictly proper scoring rules.

For the scoring rules ϕ^{LL} and ϕ^{BS} the respective divergence and entropy measures can easily be shown to be the following:

$\mathsf{d}^{\mathsf{LL}}(p,q) = \sum_{j=1}^{k} q_j \log \frac{q_j}{p_j}$	KL-divergence;
$\mathsf{e}^{\mathsf{LL}}(q) = -\sum_{j=1}^{k} q_j \log q_j$	information entropy;
$\mathsf{d}^{\mathsf{BS}}(p,q) = \sum_{j=1}^{k} (p_j - q_j)^2$	mean squared difference;
$\mathsf{e}^{\mathsf{BS}}(q) = \sum_{j=1}^{k} q_j(1-q_j)$	Gini index.

In the particular case where q is equal to the true class label y, divergence is equal to the proper scoring rule itself, *i.e.* $\mathsf{d}(p,y) = \phi(p,y)$. In the following we refer to proper scoring rules as $\mathsf{d}(p,y)$ because this makes the decompositions more intuitive.

2.3 Expected Loss and Empirical Loss

Proper scoring rules define the loss of a class probability estimator on a single instance. In practice, we are interested in the performance of the model on test data. Once the test data are fixed and known, the proper scoring rules provide the performance measure as the average of instance-wise losses across the test data. We refer to this as *empirical loss.* If the test data are drawn randomly from a (potentially infinite) labelled instance space, then the performance measure can be defined as the expected loss on a randomly drawn labelled instance. We refer to this as *expected loss.*

Empirical loss can be thought of as a special case of expected loss with uniform distribution over the test instances and zero probability elsewhere. Indeed, suppose that the generative model is uniformly randomly picking and outputting one of the test instances. The empirical loss on the (original) test data and the expected loss with this generative model are then equal. Therefore, all decompositions that we derive for the expected loss naturally apply to the empirical loss as well, assuming that test data represent the whole population.

Next we introduce our notation in terms of random variables. Let X be a random variable (a vector) representing the attributes of a randomly picked instance, and $Y = (Y_1, \ldots, Y_k)$ be a random vector specifying the class of that instance, where $Y_j = 1$ if X is of class j, and $Y_j = 0$ otherwise, for $j = 1, 2, \ldots, k$. Let now f be a fixed scoring classifier (or class probability estimator), then we denote by $S = (S_1, S_2, \ldots, S_k) = f(X)$ the score vector output by the classifier on instance X. Note that S is now a random vector, as it depends on the random variable X. The expected loss of S with respect to Y under the proper scoring rule d is $\mathbb{E}[\mathsf{d}(S, Y)]$.

Example 1. Consider a binary ($k = 2$) classification test set of 8 instances with 2 features, as shown in column $X^{(i)}$ of Table 1. Suppose the instances with indices 1,2,3,5,6 are positives (class 1) and the rest are negatives (class 2). This information is represented in column $Y_1^{(i)}$, where 1 means 'class 1' and 0 means 'not class 1'.

Suppose we have two models predicting both 0.9 as the probability of class 1 for the first 4 instances, but differ in probability estimates for the remaining 4 instances with 0.3 predicted by the first and 0.4 by the second model. This information is represented in the columns labelled $S_1^{(i)}$ for both models.

Table 1. Example dataset with 2 classes, with information shown for class 1 only. The score for class 1 is $S_1 = 0.3X_1$ by Model 1 and $S_1 = 0.25X_1 + 0.15$ by Model 2, whereas the optimal model is $Q_1 = 0.5X_2$ (or any other model which outputs 1 for first two instances and 0.5 for the rest). Columns $A_{+,1}$, $A_{*,1}$ and C_1 represent additively adjusted, multiplicatively adjusted, and calibrated scores, respectively. The average of each column is presented (mean), as well as log-loss (LL) and Brier score (BS) with respect to the true labels ($Y_1 = 1$ stands for class 1).

	Task			Model 1				Model 2			
i	$X^{(i)}$	$Y_1^{(i)}$	$Q_1^{(i)}$	$S_1^{(i)}$	$A_{+,1}^{(i)}$	$A_{*,1}^{(i)}$	$C_1^{(i)}$	$S_1^{(i)}$	$A_{+,1}^{(i)}$	$A_{*,1}^{(i)}$	$C_1^{(i)}$
1	(3,2)	1	1.0	0.9	0.925	0.914	0.75	0.9	0.875	0.886	0.75
2	(3,2)	1	1.0	0.9	0.925	0.914	0.75	0.9	0.875	0.886	0.75
3	(3,1)	1	0.5	0.9	0.925	0.914	0.75	0.9	0.875	0.886	0.75
4	(3,1)	0	0.5	0.9	0.925	0.914	0.75	0.9	0.875	0.886	0.75
5	(1,1)	1	0.5	0.3	0.325	0.336	0.50	0.4	0.375	0.364	0.50
6	(1,1)	1	0.5	0.3	0.325	0.336	0.50	0.4	0.375	0.364	0.50
7	(1,1)	0	0.5	0.3	0.325	0.336	0.50	0.4	0.375	0.364	0.50
8	(1,1)	0	0.5	0.3	0.325	0.336	0.50	0.4	0.375	0.364	0.50
mean:		0.625	0.625	0.6	0.625	0.625	0.625	0.65	0.625	0.625	0.625
LL:		0	0.520	0.717	0.732	0.715	0.628	0.684	0.673	0.683	0.628
BS:		0	0.375	0.5	0.499	0.491	0.438	0.47	0.469	0.474	0.438

The second model is better according to both log-loss ($0.684 < 0.717$) and Brier score ($0.47 < 0.5$). These can equivalently be considered either as empirical losses (as they are averages over 8 instances) or as expected losses (if the generative model picks one of the 8 instances uniformly randomly). The meaning of the remaining columns in Table 1 will become clear in the following sections.

3 Decompositions with Ideal Scores and Calibrated Scores

In this paper, all decompositions of proper scoring rules are built on procedures to map the estimated scores to new scores such that the loss is guaranteed to decrease. We start from an idealistic procedure requiring an optimal model and move towards realistic procedures.

3.1 Ideal Scores Q and the Decomposition $L = EL + IL$

Our first novel decomposition is determined by a procedure which changes the estimated scores into true posterior class probabilities (which is clearly impossible to do in practice). We denote the true posterior probability vector by $Q = (Q_1, Q_2, \ldots, Q_k)$ where $Q_j := \mathbb{E}[Y_j|X]$. Variable Q_j can be interpreted as the true proportion of class j among the instances with feature values X, and hence it is independent of the model. For our running example in Table 1 the true posterior probabilities for class 1 are given in column $Q_1^{(i)}$.

Our decomposition states that the expected loss corresponding to any proper scoring rule is the sum of expected divergence of S from Q and the expected divergence of Q from Y:

$$\mathbb{E}[\mathsf{d}(S, Y)] = \mathbb{E}[\mathsf{d}(S, Q)] + \mathbb{E}[\mathsf{d}(Q, Y)] \ .$$

This can be proved as a direct corollary of Theorem 2 in Section 5. As all these expected divergences are non-negative (due to properness of the scoring rule) and Q is the same regardless of the scoring model S, it immediately follows that $S := Q$ is the optimal model with respect to any proper scoring rule (it is a model because it is a function of X). This justifies the following terminology:

- **Epistemic Loss** $EL = \mathbb{E}[\mathsf{d}(S, Q)]$ is the extra loss due to the model not being optimal, and equals zero if and only if the model is optimal. The term relates to epistemic uncertainty (as opposed to aleatoric uncertainty) [10] and is due to our mistreatment of the evidence X with respect to the ideal model.
- **Irreducible Loss** $IL = \mathbb{E}[\mathsf{d}(Q, Y)]$ is the loss due to inherent uncertainty in the classification task, the loss which is the same for all models. This type of uncertainty is called aleatoric [10] so the loss could also be called *aleatoric loss*. It is the loss of the optimal model and equals zero only if the attributes of the instance X provide enough information to uniquely determine the right label Y (with probability 1).

For our running example the epistemic log-loss EL^{LL} for the two models is 0.198 and 0.164 (not shown in Table 1) and the (model-independent) irreducible log-loss is $IL^{\mathsf{LL}} = 0.520$, which (as expected) sum up to the total expected log-loss of 0.717 and 0.684, respectively (with the rounding effect in the last digit of 0.717). For Brier score the decomposition for the two models is $0.5 = 0.125 + 0.375$ and $0.47 = 0.095 + 0.375$, respectively.

3.2 Calibrated Scores C and the Decomposition $L = CL + RL$

The second, well-known decomposition [5] is determined by a procedure which changes the estimated scores into calibrated probabilities. We denote the calibrated probability vector by $C = (C_1, C_2, \dots, C_k)$ where $C_j := \mathbb{E}[Y_j|S]$. Variable C_j can be interpreted as the true proportion of class j among the instances for which the model has output the same estimate S, and hence calibration is model-dependent. For our running example in Table 1 the calibrated probabilities of class 1 for the two models are given in columns $C_1^{(i)}$. Note that the columns for the two models are identical. This is only because for any two instances in our example, the first model gives them the same estimate if and only if the second model does so.

The standard calibration-refinement decomposition [4] states[5] that the expected loss according to any proper scoring rule is the sum of expected divergence of S from C and the expected divergence of C from Y:

$$\mathbb{E}[\mathsf{d}(S, Y)] = \mathbb{E}[\mathsf{d}(S, C)] + \mathbb{E}[\mathsf{d}(C, Y)] \ .$$

This is another direct corollary of Theorem 2 in Section 5. The standard terminology is as follows:

- **Calibration Loss** $CL = \mathbb{E}[\mathsf{d}(S, C)]$ is the loss due to the difference between the model output score S and the proportion of positives among instances with the same output (calibrated score).
- **Refinement Loss** $RL = \mathbb{E}[\mathsf{d}(C, Y)]$ is the loss due to the presence of instances from multiple classes among the instances with the same estimate S.

For our running example the calibration loss for Brier score CL^{BS} for the two models is 0.062 and 0.033 (not shown in Table 1) and the refinement loss is for both equal to $RL^{\mathsf{BS}} = 0.438$, which sum up to the total expected Brier scores of 0.5 and 0.47, respectively (with the rounding effect in the last digit, we omit this comment in the following cases). For log-loss the decomposition for the two models is $0.717 = 0.090 + 0.628$ and $0.684 = 0.056 + 0.628$, respectively.

In practice, calibration has proved to be an efficient way of decreasing proper scoring rule loss [2]. Calibrating a model means learning a calibration mapping from the model output scores to the respective calibrated probability scores. Calibration is simple to perform if the model has only a few possible output

[5] Actually, in [4] the calibration-refinement decomposition is stated as $\mathbb{E}[\mathsf{s}(S, Y)] = \mathbb{E}[\mathsf{d}(S, C)] + \mathbb{E}[\mathsf{e}(C)]$ but this can easily be shown to be equivalent to our statement.

scores, each covered by many training examples. Then the empirical class distribution among training instances with the same output scores can be used as calibrated score vector. However, in general, there might be a single or even no training instances with the same score vector as the model outputs on a test instance. Then the calibration procedure needs to make additional assumptions (inductive bias) about the shape of the calibration map, such as monotonicity and smoothness.

Regardless of the method, calibration is almost never perfect. Even if perfectly calibrated on the training data, the model can suffer some calibration loss on test data. In the next section we propose an adjustment procedure as a precursor of calibration. Adjustment does not make any additional assumptions and is guaranteed to decrease loss if the test class distribution is known exactly.

4 Adjusted Scores A and the Decomposition $L = AL + PL$

Ideal scores cannot be obtained in practice, and calibrated scores are hard to obtain, requiring extra assumptions about the shape of the calibration map. Here we propose two procedures which take as input the estimated scores and output *adjusted scores* such that the mean matches with the given target class distribution. As opposed to calibration, no labels are required for learning how to adjust, only the scores and target class distribution are needed. We prove that *additive adjustment* is guaranteed to decrease Brier score, and *multiplicative adjustment* is guaranteed to decrease log-loss. In both cases we can decompose the expected loss in a novel way.

4.1 Adjustment

Suppose we are given the class distribution of the test data, represented as a vector π of length k, with non-negative entries and adding up to 1. It turns out that if the average of the model output scores on the test data does not match with the given distribution then for both log-loss and Brier score it is possible to adjust the scores with guaranteed reduction of loss. First we define what we mean by adjusted scores.

Definition 1. *Let π be a class distribution with k classes and A be a random real-valued vector of length k. If $\mathbb{E}[A_j] = \pi_j$ for $j = 1, \ldots, k$, then we say that A* is adjusted to the class distribution *π.*

If the scores are not adjusted, then they can be adjusted using one of the two following procedures.

Additive (score) adjustment is a procedure applying the following function α_+:

$$\alpha_+(s) = (s_1 + b_1, \ldots, s_k + b_k) \qquad \forall s \in \mathbb{R}^k \ ,$$

where $b_j = \pi_j - \mathbb{E}[S_j]$, for $j = 1, \dots, k$. Hence, the function is different depending on what the model output scores and class distribution are. It is easy to prove that the scores $\alpha_+(S)$ are adjusted: $\mathbb{E}[S_j + b_j] = \mathbb{E}[S_j] + b_j = \mathbb{E}[S_j] + \pi_j - \mathbb{E}[S_j] = \pi_j$, for $j = 1, \dots, k$.

Multiplicative (score) adjustment is a procedure applying the function α_*:

$$\alpha_*(s) = \left(\frac{w_1 s_1}{\sum_{j=1}^{k} w_j s_j}, \dots, \frac{w_k s_k}{\sum_{j=1}^{k} w_j s_j} \right) \qquad \forall s \in \mathbb{R}^k \ ,$$

where w_j are suitably chosen non-negative weights such that $\alpha_*(S)$ is adjusted to π. It is not obvious that such weights exist because of the required renormalisation, but the following theorem gives this guarantee.

Theorem 1 (Existence of weights for multiplicative adjustment). *Let π be a class distribution with $k \geq 2$ classes and S be a random positive real vector of length k. Then there exist non-negative weights $w_1, \dots, w_k$ such that $\mathbb{E}\left[\frac{w_i S_i}{\sum_{j=1}^{k} w_j S_j}\right] = \pi_i$ for $i = 1, \dots, k$.*

Proof. All the proofs are in the Appendix and the extended proofs are available at http://www.cs.bris.ac.uk/~flach/Kull_Flach_ECMLPKDD2015_Supplementary.pdf.

For our running example in Table 1 the additively adjusted and multiplicatively adjusted scores for class 1 are shown in columns $A_{+,1}^{(i)}$ and $A_{*,1}^{(i)}$, respectively. The shift b for additive adjustment was $(+0.025, -0.025)$ for Model 1 and $(-0.025, +0.025)$ for Model 2. The weights w for multiplicative adjustment were $(1.18, 1)$ for Model 1 and $(1, 1.16)$ for Model 2. For example, for Model 1 the scores $(0.9, 0.1)$ (of first four instances) become $(1.062, 0.1)$ after weighting and $(0.914, 0.086)$ after renormalising (dividing by $1.062 + 0.1 = 1.162$). The average score for class 1 becomes 0.625 for both additive and multiplicative adjustment and both models, confirming the correctness of these procedures.

4.2 The Right Adjustment Procedure Guarantees Decreased Loss

The existence of multiple adjustment procedures raises a question of which one to use. As seen from the losses after adjustment in Table 1, multiplicative adjustment achieves a lower loss for Model 1 and additive adjustment achieves a lower loss for Model 2, for both log-loss and Brier score. This shows that neither procedure is better than the other across all models.

Further inspection of Table 1 shows that for Model 1 the log-loss increased after additive adjustment and for Model 2 the Brier score increased after multiplicative adjustment. Interestingly, we can guarantee decreased loss if the right adjustment procedure is used: multiplicative adjustment always decreases log-loss, and additive adjustment always decreases Brier score. Of course, the exception is when the scores are already adjusted, in which case there is no change in the loss. The guarantee is due to the following novel loss-specific decompositions (and non-negativity of divergence):

$$\mathbb{E}[\mathsf{d}^{\mathsf{BS}}(S,Y)] = \mathbb{E}[\mathsf{d}^{\mathsf{BS}}(S,A_+)] + \mathbb{E}[\mathsf{d}^{\mathsf{BS}}(A_+,Y)] \ ,$$
$$\mathbb{E}[\mathsf{d}^{\mathsf{LL}}(S,Y)] = \mathbb{E}[\mathsf{d}^{\mathsf{LL}}(S,A_*)] + \mathbb{E}[\mathsf{d}^{\mathsf{LL}}(A_*,Y)] \ ,$$

where $A_+ = \alpha_+(S)$ and $A_* = \alpha_*(S)$ are obtained from the scores S using additive and multiplicative adjustment, respectively. Note that the additive adjustment procedure can produce values out of the range $[0,1]$ but Brier score is defined for these as well. The decompositions follow from Theorem 4 in Section 5, which provides a unified decomposition:

$$\mathbb{E}[\mathsf{d}(S,Y)] = \mathbb{E}[\mathsf{d}(S,A)] + \mathbb{E}[\mathsf{d}(A,Y)]$$

under an extra assumption which links the adjustment method and the loss measure. Due to this unification we propose the following terminology for the losses:

- **Adjustment Loss** $AL = \mathbb{E}[\mathsf{d}(S,A)]$ is the loss due to the difference between the mean model output $\mathbb{E}[S]$ and the overall class distribution $\pi := \mathbb{E}[Y]$. This loss is zero if the scores are adjusted.
- **Post-adjustment Loss** $PL = \mathbb{E}[\mathsf{d}(A,Y)]$ is the loss after adjusting the model output with the method corresponding to the loss measure.

For our running example the adjustment log-loss AL^{LL} for the two models is 0.0021 and 0.0019 (not shown in Table 1) and the respective post-adjustment losses PL^{LL} are 0.7154 and 0.6822, which sum up to the total expected log-loss of 0.7175 and 0.6841, respectively. For Brier score the decomposition for the two models is $0.5 = 0.00125 + 0.49875$ and $0.47 = 0.00125 + 0.46875$, respectively.

In practice, the class distribution is usually not given, and has to be estimated from training data. Therefore, if the difference between the average output scores and class distribution is small (*i.e.* adjustment loss is small), then the benefit of adjustment might be subsumed by class distribution estimation errors. Experiments about this remain as future work.

So far we have given three different two-term decompositions of expected loss: epistemic loss plus irreducible loss, calibration loss plus refinement loss, and adjustment loss plus post-adjustment loss. In the following section we show that these can all be obtained from a single four-term decomposition, and provide more terminology and intuition.

5 Decomposition Theorems and Terminology

In the previous sections we had the following decompositions of expected loss using a proper scoring rule (with extra assumptions for the last decomposition):

$$\mathbb{E}[\mathsf{d}(S,Y)] = \mathbb{E}[\mathsf{d}(S,Q)] + \mathbb{E}[\mathsf{d}(Q,Y)] = \mathbb{E}[\mathsf{d}(S,C)] + \mathbb{E}[\mathsf{d}(C,Y)] = \mathbb{E}[\mathsf{d}(S,A)] + \mathbb{E}[\mathsf{d}(A,Y)]$$

All these decompositions follow a pattern $\mathbb{E}[\mathsf{d}(S,Y)] = \mathbb{E}[\mathsf{d}(S,V)] + \mathbb{E}[\mathsf{d}(V,Y)]$ for some random variable V. In this section we generalise further, and introduce decompositions $\mathbb{E}[\mathsf{d}(V_1,V_3)] = \mathbb{E}[\mathsf{d}(V_1,V_2)] + \mathbb{E}[\mathsf{d}(V_2,V_3)]$ for some random variables V_1, V_2, V_3. The random variables will always be from the list S, A, C, Q, Y, and always in the same order. Actually, we will prove that the decomposition holds for any subset of 3 variables out of these 5, as long as the ordering is preserved. For decompositions involving adjusted scores A there is an extra assumption required, this is introduced in Section 5.2. First we provide decompositions without A.

5.1 Decompositions with S, C, Q, Y

Theorem 2. *Let (X, Y) be random variables representing features and labels for a k-class classification task, f be a scoring classifier, and* d *be the divergence function of a strictly proper scoring rule. Denote $S = f(X)$, $C_j = \mathbb{E}[Y_j|S]$, and $Q_j = \mathbb{E}[Y_j|X]$ for $j = 1, \ldots, k$. Then for any subsequence V_1, V_2, V_3 of the random variables S, C, Q, Y the following holds:*

$$\mathbb{E}[\mathsf{d}(V_1,V_3)] = \mathbb{E}[\mathsf{d}(V_1,V_2)] + \mathbb{E}[\mathsf{d}(V_2,V_3)]\ .$$

This theorem proves the decompositions of Section 3 but adds two more:

$$\mathbb{E}[\mathsf{d}(S,Q)] = \mathbb{E}[\mathsf{d}(S,C)] + \mathbb{E}[\mathsf{d}(C,Q)]\ , \qquad EL = CL + GL\ ;$$
$$\mathbb{E}[\mathsf{d}(C,Y)] = \mathbb{E}[\mathsf{d}(C,Q)] + \mathbb{E}[\mathsf{d}(Q,Y)]\ , \qquad RL = GL + IL\ .$$

These decompositions introduce the following new quantity:

– **Grouping Loss** $GL = \mathbb{E}[\mathsf{d}(C,Q)]$ is the loss due to many instances being grouped under the same estimate S while having different true posterior probabilities Q.

The above decompositions together imply the following three-term decomposition:

$$\mathbb{E}[\mathsf{d}(S,Y)] = \mathbb{E}[\mathsf{d}(S,C)] + \mathbb{E}[\mathsf{d}(C,Q)] + \mathbb{E}[\mathsf{d}(Q,Y)]\ , \quad L = CL + GL + IL\ .$$

5.2 Decompositions with S, A, C, Q, Y and Terminology

As discussed in Section 4, the decomposition of expected loss into adjustment loss and post-adjustment loss requires a link between the adjustment procedure and loss measure. The following definition presents the required link formally.

Definition 2. *Let (X, Y) be random variables representing features and labels for a k-class classification task, f be a scoring classifier, and ϕ be a strictly proper scoring rule. Denote $S = f(X)$. Let $\alpha = (\alpha_1, \ldots, \alpha_k)$ be a vector function with $\alpha_j : \mathbb{R} \to \mathbb{R}$ and let us denote $A = (A_1, \ldots, A_k)$ with $A_j = \alpha_j(S)$. We say that α* provides coherent adjustment of S for proper scoring rule d *if A is* *adjusted to the class distribution $\mathbb{E}[Y]$ and the following quantity is a constant (not a random variable), depending on i, j only:*

$$\phi(A, e_i) - \phi(A, e_j) - \phi(S, e_i) + \phi(S, e_j) = const_{i,j} \tag{1}$$

where e_m is a vector of length k with 1 at position m and 0 everywhere else.

Intuitively, (1) requires α to apply in some sense the same adjustment to different scores, with respect to the scoring rule. In Appendix we prove the following theorem:

Theorem 3. *Additive adjustment is coherent with Brier score and multiplicative adjustment is coherent with log-loss.*

Now we are ready to present our most general decomposition theorem:

Theorem 4. *Let (X, Y) be random variables representing features and labels for a k-class classification task, f be a scoring classifier, and d be the divergence function of a strictly proper scoring rule. Denote $S = f(X)$, $C_j = \mathbb{E}[Y_j|S]$, and $Q_j = \mathbb{E}[Y_j|X]$ for $j = 1, \ldots, k$. Let $A = \alpha(S)$ where α provides coherent adjustment of S for proper scoring rule d. Then for any subsequence V_1, V_2, V_3 of the random variables S, A, C, Q, Y the following holds:*

$$\mathbb{E}[\mathsf{d}(V_1, V_3)] = \mathbb{E}[\mathsf{d}(V_1, V_2)] + \mathbb{E}[\mathsf{d}(V_2, V_3)] \ .$$

Note that coherent adjustment might not exist for all proper scoring rules: then the decompositions involving A do not work, falling back to Theorem 2. Theorem 4 proves the decompositions in Section 4 and also provides the following decompositions:

$$\begin{aligned}
\mathbb{E}[\mathsf{d}(S, C)] &= \mathbb{E}[\mathsf{d}(S, A)] + \mathbb{E}[\mathsf{d}(A, C)] \ , \qquad & CL &= AL + PCL \ ; \\
\mathbb{E}[\mathsf{d}(S, Q)] &= \mathbb{E}[\mathsf{d}(S, A)] + \mathbb{E}[\mathsf{d}(A, Q)] \ , \qquad & EL &= AL + PEL \ ; \\
\mathbb{E}[\mathsf{d}(A, Q)] &= \mathbb{E}[\mathsf{d}(A, C)] + \mathbb{E}[\mathsf{d}(C, Q)] \ , \qquad & PEL &= PCL + GL \ ; \\
\mathbb{E}[\mathsf{d}(A, Y)] &= \mathbb{E}[\mathsf{d}(A, Q)] + \mathbb{E}[\mathsf{d}(Q, Y)] \ , \qquad & PL &= PEL + IL \ ,
\end{aligned}$$

which introduce new quantities PCL and PEL.

- **Post-adjustment Calibration Loss** $PCL = \mathbb{E}[\mathsf{d}(A, C)]$ is the loss due to the remaining calibration loss after perfect adjustment.
- **Post-adjustment Epistemic Loss** $PEL = \mathbb{E}[\mathsf{d}(A, Q)]$ is the loss due to the remaining epistemic loss after perfect adjustment.

Now we have introduced all pairwise divergences between two variables from the ordered list S, A, C, Q, Y. Table 2 summarises our proposed terminology.

A direct corollary from Theorem 4 is that if we choose 4 or 5 out of 5 variables from S, A, C, Q, Y, then we get a 3- or 4-term decomposition, respectively. In particular, the full 4-term decomposition involving all 5 variables is as follows:

$$\mathbb{E}[\mathsf{d}(S,Y)]{=}\mathbb{E}[\mathsf{d}(S,A)]{+}\mathbb{E}[\mathsf{d}(A,C)]{+}\mathbb{E}[\mathsf{d}(C,Q)]{+}\mathbb{E}[\mathsf{d}(Q,Y)] \ , \quad L{=}AL{+}PCL{+}GL{+}IL \ .$$

Table 2. Proposed terminology

	Definition	Visual	Name	Description
L	$\mathbb{E}[\mathsf{d}(S,Y)]$	`S...Y`	Loss	total expected loss
AL	$\mathbb{E}[\mathsf{d}(S,A)]$	`SA...`	Adjustment Loss	loss due to lack of adjustment
PCL	$\mathbb{E}[\mathsf{d}(A,C)]$	`.AC..`	Post-adjustment Calibration Loss	calibration loss after adjustment
GL	$\mathbb{E}[\mathsf{d}(C,Q)]$	`..CQ.`	Grouping Loss	loss due to grouping
IL	$\mathbb{E}[\mathsf{d}(Q,Y)]$	`...QY`	Irreducible Loss	loss of the optimal model
CL	$\mathbb{E}[\mathsf{d}(S,C)]$	`S.C..`	Calibration Loss	loss due to lack of calibration
PEL	$\mathbb{E}[\mathsf{d}(A,Q)]$	`.A.Q.`	Post-adjustment Epistemic Loss	epistemic loss after adjustment
RL	$\mathbb{E}[\mathsf{d}(C,Y)]$	`..C.Y`	Refinement Loss	loss after calibration
EL	$\mathbb{E}[\mathsf{d}(S,Q)]$	`S..Q.`	Epistemic Loss	loss due to non-optimal model
PL	$\mathbb{E}[\mathsf{d}(A,Y)]$	`.A..Y`	Post-adjustment Loss	loss after adjustment

Table 3. The decomposed losses (left) and their values for model 1 of the running example using log-loss (middle) and Brier score (right).

	S	A	C	Q	Y
S	0	AL	CL	EL	L
A		0	PCL	PEL	PL
C			0	GL	RL
Q				0	IL
Y					0

LL	S	A_*	C	Q	Y
S	0	0.002	0.090	0.198	0.717
A_*		0	0.088	0.196	0.715
C			0	0.108	0.628
Q				0	0.520
Y					0

BS	S	A_+	C	Q	Y
S	0	0.001	0.062	0.125	0.5
A_+		0	0.061	0.124	0.499
C			0	0.062	0.438
Q				0	0.375
Y					0

Table 3 provides numerical values for all 10 losses of Table 2 for Model 1 in our running example data (Table 1). The 4-term decomposition proves that the numbers right above the main diagonal (AL, PCL, GL, IL) add up to the total loss at the top right corner (L). All other decompositions can be checked numerically from the table (taking into account the accumulating rounding errors).

6 Algorithms and Experiments

We have proposed two new procedures in the paper: additive and multiplicative adjustment. Here we provide algorithms to perform these procedures. Both procedures first require estimation of the parameter vectors: b for additive and w for multiplicative adjustment. If the test instances are all given together in batch, then the scores of the model on test data can be used to estimate these parameter vectors. Otherwise, these need to be estimated on training (or validation) data.

Additive adjustment is algorithmically very easy. Parameter b_j is the difference of proportion π_j of class j and the mean $\mathbb{E}[S_j]$, calculated as the average output score for class j over all instances. This is exact if test data are given in batch and π_j is the true proportion, and it is approximate if π_j is estimated from training data. Finally, adjusted scores can be calculated by adding b to the model output scores, for each test instance.

Table 4. Average number of rounds to convergence of multiplicative adjustment across 10000 synthetic tasks with k classes and n instances. The number in parentheses shows the count of failures to converge out of 10000.

	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 10$	$k = 20$	$k = 30$	$k = 50$
$n = 10$	1.00 (21)	3.66 (9)	3.95 (3)	3.97 (3)	3.88 (3)	3.66 (4)	3.49 (23)	3.25 (99)
$n = 100$	1.00 (4)	3.64 (2)	3.95 (2)	3.97 (0)	3.86 (1)	3.63 (0)	3.44 (2)	3.22 (48)
$n = 1000$	1.00 (6)	3.64 (0)	3.95 (1)	3.96 (0)	3.85 (0)	3.62 (0)	3.44 (4)	3.22 (43)

For multiplicative adjustment the hard part is to obtain the parameter (weight) vector w, whereas applying adjustment using the weights is straightforward. The weight vector w can be obtained by the coordinate descent optimisation algorithm where for coordinate j the task is to minimise the difference between $\mathbb{E}[w_j s_j / \sum_{i=1}^{k} w_i s_i]$ and π_j, by changing only w_j. The minimisation in one coordinate can be done by binary search, since the expected value is monotonically increasing with respect to w_j. It is clear that if coordinate descent algorithm converges, then the obtained w is the right one. However, the algorithm can fail to converge.

We have performed experiments with synthetic tasks with $k = 2, 3, 4, 5, 10, 20, 50$ classes and $n = 10, 100, 1000$ instances to check convergence. Each task is a pair of a $n \times k$ model score matrix and class distribution vector of length k, all filled with uniformly random entries between 0 and 1, and each row is normalised to add up to 1. Table 4 reports the number of cycles through the coordinates to convergence, averaged over 10000 tasks for each k, n pair. As expected, the results have almost no dependence on the number of instances. The maximal number of rounds to convergence was 6. However, on average in 10 out of 10000 times there was no convergence. Further improvement of this result remains as future work.

7 Related Work

Proper scoring rules have a long history of research, with Brier score introduced in 1950 in the context of weather forecasting [3], and the general presentation of proper scoring rules soon after, see *e.g.* [11]. The decomposition of Brier score into calibration and refinement loss (which were back then called reliability and resolution) was introduced by Murphy [8] and was generalised for proper scoring rules by DeGroot and Fienberg [5]. The decompositions with three terms were introduced by Murphy [9] with uncertainty, reliability and resolution (Murphy reused the same name for a different quantity), later generalised to all proper scoring rules as well [4]. In our notation these can be stated as $\mathbb{E}[\mathsf{d}(S, Y)] = REL + UNC - RES = \mathbb{E}[\mathsf{d}(S, C)] + \mathbb{E}[\mathsf{d}(\pi, Y)] - \mathbb{E}[\mathsf{d}(\pi, C)]$. This can easily be proved by taking into account that the last term can be viewed as calibration loss for constant estimator π but segmented in the same way as S.

In machine learning proper scoring rules are often treated as surrogate loss functions, which are used instead of the 0-1 loss to facilitate optimisation [1]. An important question in practice is which proper scoring rule to use. One possible viewpoint is to assume a particular distribution over anticipated deployment contexts and derive the expected loss from that assumption. Hernández-Orallo *et al.* have shown that the Brier score can be derived from a particular additive cost model [6].

8 Conclusions

This paper proposes novel decompositions of proper scoring rules. All presented decompositions are sums of expected divergences between original scores S, adjusted scores A, calibrated scores C, true posterior probabilities Q and true labels Y. Each such divergence stands for one part of the total expected loss. Calibration and refinement loss are known losses of this form, the paper proposes names for the other 7 losses and provides underlying intuition. In particular, we have introduced adjustment loss, which arises from the difference between mean estimated scores and true class distribution. While it is a part of calibration loss, it is easier to eliminate or decrease than calibration loss. We have proposed first algorithms for additive and multiplicative adjustment, which we prove to be coherent with (decomposing) Brier score and log-loss, respectively. More algorithm development is needed for multiplicative adjustment, as the current algorithm can sometimes fail to converge. An open question is whether there are other, potentially better coherent adjustment procedures for these losses. We hope that the proposed decompositions provide deeper insight into the causes behind losses and facilitate development of better classification methods, as knowledge about calibration loss has already delivered several calibration methods, see *e.g.* [2].

Acknowledgments. This work was supported by the REFRAME project granted by the European Coordinated Research on Long-term Challenges in Information and Communication Sciences & Technologies ERA-Net (CHIST-ERA), and funded by the Engineering and Physical Sciences Research Council in the UK under grant EP/K018728/1.

References

1. Bartlett, P.L., Jordan, M.I., McAuliffe, J.D.: Convexity, Classification, and Risk Bounds. Journal of the American Statistical Association **101**(473), 138–156 (2006)
2. Bella, A., Ferri, C., Hernández-Orallo, J., Ramírez-Quintana, M.J.: On the effect of calibration in classifier combination. Applied Intelligence **38**(4), 566–585 (2012)
3. Brier, G.W.: Verification of forecasts expressed in terms of probability. Monthly weather review **78**(1), 1–3 (1950)
4. Bröcker, J.: Reliability, sufficiency, and the decomposition of proper scores. Quarterly Journal of the Royal Meteorological Society **135**(643), 1512–1519 (2009)
5. De Groot, M.H., Fienberg, S.E.: The Comparison and Evaluation of Forecasters. Journal of the Royal Statistical Society. Series D (The Statistician) **32**(1/2), 12–22 (1983)

6. Hernández-Orallo, J., Flach, P., Ferri, C.: A unified view of performance metrics: translating threshold choice into expected classification loss. The Journal of Machine Learning Research **13**(1), 2813–2869 (2012)
7. Kull, M., Flach, P.A.: Reliability maps: a tool to enhance probability estimates and improve classification accuracy. In: Calders, T., Esposito, F., Hüllermeier, E., Meo, R. (eds.) ECML PKDD 2014, Part II. LNCS, vol. 8725, pp. 18–33. Springer, Heidelberg (2014)
8. Murphy, A.H.: Scalar and vector partitions of the probability score: Part I. Two-state situation. Journal of Applied Meteorology **11**(2), 273–282 (1972)
9. Murphy, A.H.: A new vector partition of the probability score. Journal of Applied Meteorology **12**(4), 595–600 (1973)
10. Senge, R., Bösner, S., Dembczynski, K., Haasenritter, J., Hirsch, O., Donner-Banzhoff, N., Hüllermeier, E.: Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty. Information Sciences **255**, 16–29 (2014)
11. Winkler, R.L.: Scoring Rules and the Evaluation of Probability Assessors. Journal of the American Statistical Association **64**(327), 1073–1078 (1969)

Appendix: Proofs of the Theorems

Here we prove the theorems presented in the paper, extended proofs are available at http://www.cs.bris.ac.uk/~flach/Kull_Flach_ECMLPKDD2015_Supplementary.pdf.

Proof of Theorem 1: If there are any zeros in the vector π, then we can set the respective positions in the weight vector also to zero and solve the problem with the remaining classes. Therefore, from now on we assume that all entries in π are positive.

Let $\mathbb{W}$ denote the set of all non-negative (weight) vectors of length k with at least one non-zero component. We introduce functions $t_i : \mathbb{W} \to \mathbb{R}$ with $t_i(w) = \mathbb{E}[w_i S_i / \sum_{j=1}^{k} w_j S_j]$. Then we need to find w^* such that $t_i(w^*) = \pi_i$ for $i = 1, \ldots, k$. For this we prove the existence of increasingly better functions $h_0, h_1, \ldots, h_{k-1} : \mathbb{W} \to \mathbb{W}$ such that for $m = 0, \ldots, k-1$ the function h_m satisfies $t_i(h_m(w)) = \pi_i$ for $i = 1, \ldots, m$ for any w. Then $w^* = h_{k-1}(w)$ is the desired solution, where $w \in \mathbb{W}$ is any weight vector, such as the vector of all ones. Indeed, it satisfies $t_i(w^*) = \pi_i$ for $i = 1, \ldots, k-1$ and hence for $i = k$.

We choose h_0 to be the identity function and prove the existence of other functions h_m by induction. Let h_m for $m < k-1$ be such that for any w the vector $h_m(w)$ does not differ from w in positions $m+1, \ldots, k$ and $t_i(h_m(w)) = \pi_i$ for $i = 1, \ldots, m$. For a fixed w it is now sufficient to prove the existence of w' such that it does not differ from w in positions $m+2, \ldots, k$ and $t_i(w') = \pi_i$ for $i = 1, \ldots, m+1$. We search for such w' among the vectors $h_m(w[m+1 : x])$ with $x \in [0, \infty)$ where $w[m+1 : x]$ denotes the vector w with the element at position $m+1$ changed into x. The chosen form of w' guarantees that it does not differ from w in positions $m+2, \ldots, k$ and $t_i(w') = \pi_i$ for $i = 1, \ldots, m$. It only remains to choose x such that $t_{m+1}(w') = \pi_{m+1}$. For this we note that for $x = 0$ we have $t_{m+1}(h_m(w[m+1 : 0])) = 0$ because the weight at position $m+1$ is zero. In the limit process $x \to \infty$ we have $t_{m+1}(h_m(w[m+1:x])) \to 1-\sum_{i=1}^{m} \pi_i$ because the weight x at position $m+1$ will dominate over weights at $m+2,\ldots,k$, whereas the weights at $1,\ldots,m$ ensure that $t_i(h_m(w[m+1:x])) = \pi_i$ for $i = 1,\ldots,m$. Since $0 < \pi_{m+1} < 1-\sum_{i=1}^{m} \pi_i$ then according to the intermediate value theorem there exists x such that $t_{m+1}(h_m(w[m+1:x])) = \pi_{m+1}$. By this we have proved the existence of a suitable function h_{m+1}, proving the step of induction, which concludes the proof. □

Lemma 1. *Let V_1, V_2, V_3, W be real-valued random vectors with length k where $V_{2,j} = \mathbb{E}[V_{3,j}|W]$ for $j = 1,\ldots,k$, and V_1 is functionally dependent on W. If d is divergence of a proper scoring rule, then $\mathbb{E}[\mathsf{d}(V_1,V_3)] = \mathbb{E}[\mathsf{d}(V_1,V_2)] + \mathbb{E}[\mathsf{d}(V_2,V_3)]$.*

Proof. Due to the law of total expectation it is enough to prove that $\mathbb{E}[\mathsf{d}(V_1,V_3)|W] = \mathbb{E}[\mathsf{d}(V_1,V_2)|W] + \mathbb{E}[\mathsf{d}(V_2,V_3)|W]$. After expressing each d as a difference of two s terms, all obtained terms are sums over $j = 1,\ldots,k$ and it is enough to prove that for each j the equality holds. Also, as we are conditioning on W, all terms that do not involve V_3 are constants with respect to conditional expectation. Therefore, we need to prove that $\phi(V_1,e_j)\mathbb{E}[V_{3,j}|W] - \mathbb{E}[\mathsf{s}(V_3,V_3)|W]$ equals $\phi(V_1,e_j)V_{2,j} - \phi(V_2,e_j)V_{2,j} + \phi(V_2,e_j)\mathbb{E}[V_{3,j}|W] - \mathbb{E}[\mathsf{s}(V_3,V_3)|W]$. This holds due to $\mathbb{E}[V_{3,j}|W] = V_{2,j}$. □

Proof of Theorem 2: We consider the following two possibilities:

1. $V_2 = C$. Let us take $W = S$ in Lemma 1. Then $V_1 = S$ and it is functionally dependent on itself, W. Also, $V_{2,j} = \mathbb{E}[V_{3,j}|W]$ regardless of whether $V_3 = Y$ or $V_3 = Q$ because $C_j = \mathbb{E}[Y_j|S] = \mathbb{E}[\mathbb{E}[Y_j|X,S]|S] = \mathbb{E}[Q_j|S]$, where the second equality is due to the law of iterated expectations. The result now follows from Lemma 1.

2. $V_2 = Q$. Then $V_3 = Y$ and the result follows from Lemma 1 with $W = X$ because $V_{2,j} = Q_j = \mathbb{E}[Y_j|X] = \mathbb{E}[V_{3,j}|W]$ and both candidates S and C for V_1 are functionally dependent on $W = X$. □

Proof of Theorem 3: In Section 4 we proved that both methods provide adjusted scores, so we only need to prove Eq.(1). For log-loss we need to prove that $-\log A_i + \log A_j + \log S_i - \log S_j$ is a constant. For this it is enough to show that $(A_j/A_i)/(S_j/S_i)$ is constant. According to the definition of multiplicative adjustment this quantity equals $((w_j S_j)/(w_i S_i))/(S_j/S_i) = w_j/w_i$ which is a constant, proving that multiplicative adjustment is coherent with log-loss. For Brier score we need to prove that

$$\sum_{m=1}^{k}(A_m - \delta_{mi})^2 - \sum_{m=1}^{k}(A_m - \delta_{mj})^2 - \sum_{m=1}^{k}(S_m - \delta_{mi})^2 + \sum_{m=1}^{k}(S_m - \delta_{mj})^2 = const_{ij} ,$$

where δ_{mi} is 1 if $m = i$ and 0 otherwise. For $m \notin \{i,j\}$ the respective terms in the first and second sums and in the third and fourth sums are equal and therefore cancel each other. For $m = i$ the respective terms together give

$(A_i - 1)^2 - A_i^2 - (S_i - 1)^2 + S_i^2$, for additive adjustment this equals to the constant $-2b_i$ due to $A_i = S_i + b_i$. A similar argument holds for $m = j$ and as a result we have proved that the requirement (1) holds and additive adjustment is coherent with Brier score. □

Proof of Theorem 4: If none of V_1, V_2, V_3 is A, then the result follows from Theorem 2. If $V_1 = A$, then the result follows from Theorem 2 with $f^{NEW} = \alpha \circ f$ because then $S^{NEW} = A$, $C^{NEW} = C$, $Q^{NEW} = Q$. It remains to prove the result for the case where $V_1 = S$ and $V_2 = A$. Denote $\beta_j = \phi(A, e_1) - \phi(A, e_j) - \phi(S, e_1) + \phi(S, e_j)$ for $j = 1, \dots, k$, then β_j are all constants. Now it is enough to prove that the following quantity is zero:

$$\mathbb{E}[\mathsf{d}(S, V_3)] - \mathbb{E}[\mathsf{d}(S, A)] - \mathbb{E}[\mathsf{d}(A, V_3)] =$$

$$= \mathbb{E}\Big[\sum_{j=1}^{k}\Big(\phi(S, e_j)V_{3,j} - \phi(S, e_j)A_j + \phi(A, e_j)A_j - \phi(A, e_j)V_{3,j}\Big) - \mathsf{s}(V_3, V_3) + \mathsf{s}(V_3, V_3)\Big]$$

$$= \mathbb{E}\big[\sum_{j=1}^{k}\big(\phi(S, e_j) - \phi(A, e_j)\big)\big(V_{3,j} - A_j\big)\big] = \mathbb{E}\big[\sum_{j=1}^{k}\big(\beta_j + \phi(S, e_1) - \phi(A, e_1)\big)\big(V_{3,j} - A_j\big)\big]$$

$$= \sum_{j=1}^{k}\beta_j\Big(\mathbb{E}[V_{3,j}] - \mathbb{E}[A_j]\Big) + \mathbb{E}\Big[\Big(\phi(S, e_1) - \phi(A, e_1)\Big)\Big(\sum_{j=1}^{k}V_{3,j} - \sum_{j=1}^{k}A_j\Big)\Big] .$$

The first term is equal to zero regardless of whether V_3 is Y or Q or C since $\mathbb{E}[A_j] = \mathbb{E}[Y_j] = \mathbb{E}[Q_j] = \mathbb{E}[C_j]$. The second term is equal to zero because both $V_{3,j}$ and A_j for $j = 1, \dots, k$ add up to 1. □